Chapter 20

Comparing Two Numerical Response Populations: Paired Data

This chapter is an extension of Chapter 16. In Chapter 16 we considered populations in which each population member or trial yields two dichotomous responses. In the current chapter each population member or trial yields two numbers. In other ways, however, this chapter also extends the work we did in Chapters 17–19.

20.1 Subject Reuse

I will introduce you to the idea of subject reuse with an artificial study of drug therapy for tension headaches. We will compare two different ways to design a study. I cannot use a real scientific study to make my comparisons because, to my knowledge, medical researchers select a design and use it. They do not investigate a medical issue twice, with two different designs, just to make me happy!

I am interested in studying drug therapies for a fairly mild health ailment, tension headaches. As you will see shortly, it is important that I have chosen an ailment that is both nonlethal and recurrent. I want to compare two drug therapies for the treatment of a tension headache. I will refer to the two therapies as drug A (treatment 1 and population 1) and drug B (treatment 2 and population 2).

We need a response that is a number. Each subject is given the following instructions:

The next time you experience a tension headache, take the drug we have given to you. Wait 20 minutes. Write down your assessment of your pain on a scale from 0 (no pain) to 10 (worst pain ever).

How can I study this? Going all the way back to Chapter 1, I can use a completely randomized design. Following Chapter 19, I can perform population-based inference on the data I obtain from my completely randomized design. In particular, I can compare the mean of population 1 (drug A), $\mu_1$, to the mean of population 2 (drug B), $\mu_2$. I can estimate $\mu_1 - \mu_2$ with confidence and test the
null hypothesis that $\mu_1 = \mu_2$. In order to choose an alternative, we need more information about the drugs. Three scenarios come to mind, listed below:

- Drug A is a placebo and drug B supposedly is beneficial. In this situation, remembering that smaller responses are preferred to larger responses, my alternative would be $>$.  
- Drug A is the extra-strength version of drug B. In this situation, my alternative would be $<$.  
- Drugs A and B are different active drugs. In this situation, my alternative would be $\neq$.

Suppose now that I have 32 subjects available for study and I am willing to pretend that they are a random sample from my superpopulation of interest. I decide to use a balanced design. Thus, I will use the online randomizer to assign 16 subjects to each treatment.

The artificial data for my CRD on the 32 subjects is given in Table 20.1. The data have been separated by treatments and sorted within each treatment. You can verify the following values of summary statistics (or trust me if you don’t need additional practice on these computations):

$$\bar{x} = 5.500, s_1 = 2.366, \bar{y} = 4.000, s_2 = 2.251 \text{ and } n_1 = n_2 = 16.$$  

Next, I calculate

$$s_p^2 = \frac{(2.366)^2 + (2.251)^2}{2} = 5.3325 \text{ and } s_p = \sqrt{5.3325} = 2.309.$$  

The 95% confidence interval estimate of $\mu_1 - \mu_2$ is (see Formula 19.9 on page 501):

$$(5.50 - 4.00) \pm 2.042(2.309)\sqrt{2/16} = 1.50 \pm 2.042(0.8164) = 1.50 \pm 1.67 = [-0.17, 3.17].$$  

This interval is inconclusive because it contains both positive and negative numbers. For future reference, note that the half-width of this interval is 1.67.

For a test of hypotheses, from Equation 19.11 on page 502 the observed value of the test statistic is

$$t = \frac{1.50}{0.8164} = 1.837.$$  

With the help of our website calculator,

we find that the area under the t-curve with $df = 16 + 16 − 2 = 30$ to the right of 1.837 is equal to 0.0381. Thus, the approximate P-value for the alternative $>$ is 0.0381 and the approximate P-value for the alternative $\neq$ is $2(0.0381) = 0.0762$.

Let’s look at the data in Table 20.1 again. In the drug A row, two subjects gave a response of 2—not much pain—and two gave a response of 9—a great deal of pain. In words, for drug A there is a large amount of subject-to-subject variation. The same is true for drug B. The idea behind the randomized pairs design (RPD) is to attempt to reduce this subject-to-subject variation.

I mentioned above that it is important that tension headaches are nonlethal and recurrent. Recurrence is important because if each subject has a headache (which is necessary in the CRD for us to obtain a response from each subject) then the subject will have a second headache. The RPD we learn about below will use responses from two headaches per subject, compared to the CRD which looked at one headache per subject. Nonlethal is important because—and I don’t mean to be insensitive—in order to have a second headache the subject must survive the first one.

Admittedly, I am ignoring studies that would involve looking at 3, 4, 5 or more headaches per subject. I must draw the line somewhere!

You can now see the reason for the term subject reuse. We reuse each subject and, thus, obtain two responses per subject. And, somewhat obviously, because our goal is to compare the two treatments, for each subject we obtain a response from both treatments. Thus, for example, subject Sally gives us two numbers: her pain with drug A and her pain with drug B.

My next step is to provide you with artificial headache pain data from an RPD. My goal is to compare my RPD to my CRD for the artificial headache pain study. What is a fair way to do this? Well, my CRD had 32 subjects, with one response per subject, yielding a total of 32 observations. I could have 32 subjects in my RPD, but that would yield $32 \times 2 = 64$ observations. This strikes me as an unfair comparison. Thus, instead, my RPD below has only 16 subjects; with each subject giving two responses, I will have a total of $16 \times 2 = 32$ observations, the same as I had in my CRD. In fact, my RPD has exactly the same 32 observations as my CRD did. The data for my RPD is given in Table 20.2. Let’s take a moment to make sure we can read this table correctly.

I have 16 subjects in my RPD and they have been labeled, for ease of reference, in the first row. If you compare the Drug A row of Table 20.2 with the Drug A row of Table 20.1, you can easily verify (because of the sorting in both tables) that the 16 responses to drug A are the same for the two data sets. You can also verify that the 16 responses to drug B are the same for the two data sets, but it’s a lot easier to trust me on this! Let’s look at subjects labeled 1 and 16. Subject 1 gives small responses (2 and 3) for both drugs, and subject 16 gives large responses (9 and 7) for both drugs. In words, subjects 1 and 16 vary a great deal in their responses to drug A and they vary a great deal in their responses to drug B.

Table 20.2 contains a row of numbers unlike any we have seen previously. For each subject I have calculated the difference in the subject’s responses: response to A minus response to B. In symbols, the difference $d$ is equal to $x − y$ for each subject. Let’s look at subjects 1 and 16 again, but now let’s look at their values of $d$. For subject 1, $d = −1$, and for subject 16, $d = 2$. Remembering that smaller values of $x$ and $y$ are better, a negative value of $d$ indicates that the subject responded better to A than to B, whereas a positive value of $d$ indicates that the subject responded better to B than to A. The $d$’s for subjects 1 and 16 are much closer to each other (a
Table 20.2: First set of artificial data from an RPD on headache pain.

<table>
<thead>
<tr>
<th>Subject</th>
<th>A(x)</th>
<th>B(y)</th>
<th>Difference(d)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>−1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>4</td>
<td>−1</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>6</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>8</td>
<td>7</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>7</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>10</td>
<td>8</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>11</td>
<td>8</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>9</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>9</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>14</td>
<td>9</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>15</td>
<td>9</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>16</td>
<td>9</td>
<td>7</td>
<td>2</td>
</tr>
</tbody>
</table>

The difference of 3) than either their values on A (a difference of 7) or B (a difference of 4). Thus, at least for these two subjects, there is less subject-to-subject variation on $d$ than on both $x$ and $y$. In fact, I reported earlier that the standard deviations of the $x$’s and $y$’s are, respectively, 2.366 and 2.251. By comparison, you may verify that the standard deviation of the $d$’s is 1.592. Thus, by examining the three standard deviations we arrive at the same conclusion we have from looking at subjects 1 and 16: the subject-to-subject variation is smaller for the differences than it is for both drugs.

Let’s gather together our various summary statistics:

<table>
<thead>
<tr>
<th>Source</th>
<th>Symbol</th>
<th>Mean</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment 1</td>
<td>$x$</td>
<td>5.50</td>
<td>2.366</td>
</tr>
<tr>
<td>Treatment 2</td>
<td>$y$</td>
<td>4.00</td>
<td>2.251</td>
</tr>
<tr>
<td>Difference</td>
<td>$d$</td>
<td>1.50</td>
<td>1.592</td>
</tr>
</tbody>
</table>

Notice that

$$\bar{x} - \bar{y} = 5.50 - 4.00 = 1.50 = \bar{d}.$$  

On reflection, we realize that this is true for any set of data; calculating two means and subtracting gives the same answer as first subtracting then finding the mean of the differences. It is obvious that this argument can be extended to an entire finite population. Let $\mu_d$ denote the mean of the population of differences. Then:

$$\mu_d = \mu_1 - \mu_2. \quad (20.1)$$

Equation [20.1] is also true for populations for trials. (The argument is a bit trickier; I recommend that you simply believe me.) This leads to the following very important realization:

- Inference for $\mu_1 - \mu_2$—estimation or testing—is equivalent to inference for $\mu_d$.

In particular, if we are willing to assume that our $X$’s are a random sample from a population, then our $Y$’s are also a random sample from the same population, although the two samples (the $X$’s and the $Y$’s) are not independent samples. If we let

$$D_1, D_2, D_3 \ldots D_m$$

denote the random variables that yield the observed values

$$d_1, d_2, d_3 \ldots d_m,$$
then it also follows that the $D$’s are a random sample from the population of differences. Because the $D$’s are a random sample from a single population, the methods of Chapters 17 and 18 may be used to analyze them.

In particular, Gosset’s confidence interval estimate of $\mu$ in Chapter 17, Formula 17.6, yields—after we change the symbols—the following result.

**Result 20.1** Gosset’s confidence interval estimate of $\mu_d = \mu_1 - \mu_2$ is:

$$\bar{d} \pm t^*(s_d/\sqrt{m}).$$  \hspace{1cm} (20.2)

In this formula, $\bar{d}$ and $s_d$ are the sample mean and standard deviation, respectively, of the differences. The number of pairs is denoted by $m$ and the degrees of freedom for $t^*$ is $(m - 1)$.

I will illustrate the use of Formula 20.2 for our first set of artificial data from an RPD on headache pain, given in Table 20.2. For $df = m - 1 = 16 - 1 = 15$, you can verify that $t^*$ for the 95% confidence level is 2.131. Thus, the 95% confidence interval estimate of $\mu_d = \mu_1 - \mu_2$ is

$$1.50 \pm 2.131(1.592/\sqrt{16}) = 1.50 \pm 2.131(0.398) = 1.50 \pm 0.85 = [0.65, 2.35].$$

This interval is **conclusive**; the mean pain on drug A is between 0.65 and 2.35 units larger than the mean pain on drug B.

Recall that when we had exactly the same data from a CRD, the confidence interval estimate was:

$$1.50 \pm 1.67.$$ 

Thus, the confidence interval is—approximately—one-half as wide for the RPD as it is for the CRD. Subject reuse is effective! More accurately, I have given you artificial data that made subject reuse effective.

Recall also that, as a very rough guide, we must quadruple the number of subjects to reduce the half-width of a confidence interval by a factor of two. (This is rough because with more data the value of $t^*$ will definitely be reduced and the various sample standard deviations will likely change.) Thus, very roughly, I would need $4 \times 32 = 128$ subjects on a CRD to obtain the same precision that I get from an RPD with 16 subjects! As the expression goes, “Work smarter, not harder!”

We can also perform a test of hypotheses for data from an RPD. For the null hypothesis that $\mu_d = 0$, we rewrite the observed value of the test statistic—using the notation of this chapter—given in Equation 18.2 on page 470:

$$t = \frac{\bar{d}}{s_d/\sqrt{m}} = \sqrt{m}(\bar{d}/s_d).$$  \hspace{1cm} (20.3)

The three rules for finding the P-value are summarized in the following result.

**Result 20.2** For the null hypothesis that $\mu_d = 0$, and $t$ given in Equation 20.3, the rules for finding the P-value are below. In these rules, areas are computed under the t-curve with $df = m - 1$. 

517
1. For the alternative $\mu_d > 0$, the P-value equals the area to the right of $t$. Equivalently, the P-value equals the area to the left of $-t$.

2. For the alternative $\mu_d < 0$, the P-value equals the area to the left of $t$. Equivalently, the P-value equals the area to the right of $-t$.

3. For the alternative $\mu_d \neq 0$, the P-value equals twice the area to the right of $|t|$. Equivalently, the P-value equals twice the area to the left of $-|t|$.

In the above result, if the population of differences is a Normal pdf, then the P-values are exact; otherwise, they are approximations and the comments from Chapters 17 and 18 regarding their accuracy are relevant.

For the data in Table 20.2, the observed value of the test statistic is 

$$t = \sqrt{16(1.50/1.592)} = 3.769.$$  

Using the website, http://stattrek.com/online-calculator/t-distribution.aspx, we find that the area under the t-curve with $df = m - 1 = 16 - 1 = 15$ to the right of 3.769 equals 0.0009. Thus, the approximate P-value for $>$ is 0.0009 and the approximate P-value for $\neq$ is $2(0.0009) = 0.0018$. For comparison, the P-value for $>$ for a CRD with the same responses was shown earlier to equal 0.0381. Thus, for alternative $>$ or $\neq$, the approximate P-value from the RPD value is more than 38 times smaller than the approximate P-value from the CRD!

Think about the question: Have I convinced you that an RPD is better than a CRD for a study of headache pain? I hope not; all of my data are artificial. What I have shown you is that it is possible that an RPD can be better than a CRD. In the name of basic fairness, I should show you that the opposite also can be true.

Table 20.3 provides a second set of artificial data for an RPD on headache pain. As with Table 20.2, the data in Table 20.3 are the same responses for both drugs as given in the original CRD, Table 20.1. For these new data, it can be shown that $\bar{d} = 1.50$ and $s_d = 3.162$. The first of these summaries is no surprise; because the $x$’s and $y$’s have not changed, $\bar{x} = 5.50$, $\bar{y} = 4.00$ and, perforce, $\bar{d} = \bar{x} - \bar{y} = 1.50$. Note, however, that for these new data, $s_d$ is much larger than both $s_1$ and $s_2$ and, hence, $s_p$ too.

I will evaluate Formula 20.2 with these new data. Gosset’s 95% confidence interval estimate of $\mu_d$ is:

$$1.50 \pm 2.131(3.162/\sqrt{16}) = 1.50 \pm 2.131(0.7905) = 1.50 \pm 1.68 = [-0.18, 3.18].$$

Notice that the half-width of this new interval, 1.68, is larger, but only slightly larger, than the half-width of the interval for the CRD data, 1.67.

We have seen that for the first set of artificial data, the RPD gives a much more sensitive analysis than the CRD. For the second set of artificial data, however, the analysis is virtually the same for the RPD and CRD. What is it about these data sets that is causing this difference? Well,
Table 20.3: Second set of artificial data from an RPD on headache pain.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Drug $A(x)$</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>
<th>$11$</th>
<th>$12$</th>
<th>$13$</th>
<th>$14$</th>
<th>$15$</th>
<th>$16$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$B(y)$</td>
<td>$8$</td>
<td>$3$</td>
<td>$2$</td>
<td>$4$</td>
<td>$7$</td>
<td>$0$</td>
<td>$5$</td>
<td>$3$</td>
<td>$1$</td>
<td>$3$</td>
<td>$7$</td>
<td>$6$</td>
<td>$4$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Difference($d$)</td>
<td>$-6$</td>
<td>$-1$</td>
<td>$1$</td>
<td>$-1$</td>
<td>$2$</td>
<td>$0$</td>
<td>$-2$</td>
<td>$5$</td>
<td>$1$</td>
<td>$1$</td>
<td>$4$</td>
<td>$6$</td>
<td>$5$</td>
<td>$1$</td>
<td>$3$</td>
<td>$5$</td>
</tr>
</tbody>
</table>

the simple answer is that the second set has a much larger value of $s_d$ than the first data set. In particular, the half-width of the confidence interval for $\mu_d$ is

$$t^* \left( \frac{s_d}{\sqrt{m}} \right);$$

clearly, as $s_d$ increases, the half-width increases and the analysis becomes less sensitive.

I am not really satisfied with the above answer. The value of $s_d$ is somewhat of a mystery; what makes it larger or smaller? It turns out that it is possible to understand better what is happening if we draw a picture of the data.

### 20.2 The Scatterplot

I want to introduce you to a very important picture in Statistics, one I actually used—without much explanation—in Figure 17.9 on page 454 and Figure 17.10 on page 455. Figure 20.1 presents two scatterplots, one for each our sets of artificial data from an RPD. Let’s take a few minutes to examine these pictures. Let’s look at the top picture, the scatterplot of $y$ versus $x$ for the first set of artificial data from an RPD on headache pain.

The scatterplot begins with the familiar coordinate system from childhood, with the vertical axis corresponding to $y$ (response to Drug B in this plot) and the horizontal axis corresponding to $x$ (response to Drug A in this plot). We will refer to this as a plot of $y$ versus $x$. Next, both axes are given scales that are sufficient to include the entire plot, as described below. I have given each axis the values 0 through 10.

At this point you may have noticed something untoward about my scatterplot: even though both axes measure the same thing—subjective assessment of pain—I have used different scales on the two axes. In particular, in my picture the $x$ values are stretched out a bit compared to the $y$ values. Why did I do this? I give you two reasons.

1. Scatterplots are used extensively in Statistics, most notably in regression analysis, which we will study in the next two chapters of these notes. Overwhelmingly the norm in these applications, especially regression, is that the $y$ and $x$ features are like apples and oranges; i.e. there is no natural relationship between the features. For example, if a unit is an adult male human, then $x$ could be his height in inches and $y$ could be his weight in pounds. There is neither a natural nor obvious way to choose the same scale for $y$ and $x$. As a result, when statisticians choose scales we mostly are concerned with the next item.
Figure 20.1: Scatterplots of the response to drug B versus the response to drug A for the 16 Subjects in the two artificial data sets from RPDs on headache pain.
2. Statisticians generally prefer scatterplots for which the width is greater than its height. Such a picture is deemed to be more aesthetically pleasing than a square. (If you are interested in this topic, see the Wikipedian entry for the golden rectangle:


or read The Da Vinci Code!)

My scatterplot displays the pair of values \(x\) and \(y\) for each of the 16 subjects in my RPD. For example, consider subject 1; its values are \(x = 2\) and \(y = 3\). Subject 1 appears in the scatterplot as an ‘O’ at the location (ordered pair) \((x, y) = (2, 3)\). (Make sure that you can locate this ‘O.’) Thus, because there are 16 subjects in my RPD, there are 16 O’s in Figure 20.1. Except that there aren’t; there are actually only 14 O’s in the scatterplot and a numeral ‘2.’ The numeral 2 is located at \((x, y) = (5, 2)\); it indicates that there should be two O’s at this point because subjects 7 and 8 both had \(x = 5\) and \(y = 2\).

Now, look at the scatterplot. What do you see? We will look at many scatterplots when we study regression, so I am going to keep this brief. First, the relationship between \(x\) and \(y\) is not deterministic; it is not a mathematical function—the value of \(x\) does not determine the value of \(y\) (nor does the value of \(y\) determine the value of \(x\)). I mention this because, I conjecture, you have had a great deal of experience with deterministic relationships in your various mathematics classes. In Statistics, however, we almost always study relationships that are not deterministic.

As a statistician I look at the scatterplot in Figure 20.1 and I see two main features: the relationship between \(x\) and \(y\) is increasing and it looks linear. Increasing is self-explanatory. But if it’s not: as we move from unit to unit in a way in which the values of \(x\) are increasing—in every day language, scan the scatterplot from left to right—the values of \(y\) tend to become larger. Linear is my subjective assessment that the pattern in the scatterplot can be described reasonably well by a straight line and does not require a curve to describe it. Again, let me remind you that we will consider these issues in greater detail when we study regression later in these notes.

Whenever the relationship between \(x\) and \(y\) looks linear it is reasonable to summarize the relationship by calculating the correlation coefficient, denoted by \(r\). Figure 20.1 tells us that the correlation coefficient of its first scatterplot is \(r = 0.763\). We will learn a great deal about the correlation coefficient when we study regression. Let me just say for now that the possible values of the correlation coefficient fall between \(-1\) and \(1\) inclusive: \(-1 \leq r \leq 1\). Also, an increasing [decreasing] relationship between \(x\) and \(y\) makes \(r\) positive [negative]. For the purpose of an RPD, the correlation coefficient plays a role in a mathematical relationship that exists between our three standard deviations (for the \(x\)’s, for the \(y\)’s and for the \(d\)’s), The relationship is

\[
s_d^2 = s_1^2 + s_2^2 - 2rs_1s_2. \tag{20.4}
\]

This equation can be illustrated with the values of the three standard deviations and \(r = 0.763\):

\[
s_d^2 = (1.592)^2 = 2.534464, \text{ and } s_1^2 + s_2^2 - 2rs_1s_2 = (2.366)^2 + (2.251)^2 - 2(0.763)(2.366)(2.251) = \]

521
which are the same, except for round-off error.

We saw earlier that as the value of \( s_d \) increases the half-width of the confidence interval for \( \mu_d \) also increases. As a result, as \( s_d \) increases, subject reuse becomes less useful. We can see from Equation [20.4] that as the value of \( r \) increases, the value of \( s^2_d \) and, hence \( s_d \), decreases. **Thus, the effectiveness of subject reuse grows with the value of \( r \).** This is important because as you learn more about how \( r \) relates to a scatterplot, you will be better at deciding whether subject reuse is effective.

For example, the second scatterplot in Figure 20.1 is for our second set of artificial data from an RPD. In this picture I see only a very weak increasing relationship between \( x \) and \( y \). My visual assessment agrees with the value of \( r = 0.063 \) which is barely larger than 0. (Again, we will learn more about this in the next chapter.)

Let’s now go back in time to before we collected our data. Imagine that I am a researcher who knows a great deal about headache pain. I know that if my scatterplot of values of \( x \) and \( y \) looks like the second scatterplot in Figure 20.1 then pairing won’t be any better than a CRD. If, indeed, my scatterplot provides a smaller value of \( r \)—including negative values—then pairing is less effective than a CRD. If, however, my scatterplot yields an \( r \) substantially larger than 0.063, then pairing would be effective, possibly extremely effective. Based on my expertise as a headache pain researcher, I am convinced that there will be an increasing relationship between \( x \) and \( y \) and that the relationship will be substantially stronger than one that yields \( r = 0.063 \). (Does this make sense to you medically? Why or why not?) Thus, given my expert opinion, I would definitely opt for pairing over independent samples.

### 20.3 Putting the ‘R’ in RPD

I have talked (well, keyboarded) a great deal about the ‘P’ in an RPD, but have said nothing about the ‘R;’ I will do so now.

Randomization occurs at each pair in an RPD. In general, let \( m \) denote the number of pairs in an RPD. This means that the data will consist of \( m \) values each of \( x \)'s, \( y \)'s and \( d \)'s. For my two headache RPDs, \( m = 16 \). At each pair there are two choices for the assignment of treatments to members of the pair; they are:

- Assign the first member of the pair to treatment 1 and the second member of the pair to treatment 2. We denote this possibility as 1.
- Assign the first member of the pair to treatment 2 and the second member of the pair to treatment 1. We denote this possibility as 2.

Thus, at each pair our **randomizer** must give us either a 1 or a 2, with these options being equally likely to occur. Also, the decisions at different pairs must be statistically independent.

There are a number of physical devises that will allow us to randomize for an RPD. Instead, I will focus on an electronic method using the randomizer we learned about in Chapter 2. We begin by going to the website
This site asks you to provide input information. I will walk you through the choices.

- The first question is: **How many sets of numbers do you want to generate?**
  We want an assignment for one RPD; thus, leave it at the default value of 1.

- The second question is: **How many numbers per set?**
  Enter \( m \) which equals 16 for our headache pain study.

- Next, you need to specify: **Number range.**
  For an RPD, this will always be from 1 to 2.

- Next, we have another question: **Do you wish each number in a set to remain unique?**
  Answer: No.

- Next, we have another question: **Do you wish to sort the numbers that are generated?**
  Answer: No.

- You may ignore the final question; i.e. we are happy with the default response.

- You are now ready to click on the box: **Randomize Now!**

I operated our randomizer with the choices above and obtained:

<table>
<thead>
<tr>
<th>Pairs</th>
<th>1–4</th>
<th>5–8</th>
<th>9–12</th>
<th>13–16</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2,1,1,1</td>
<td>2,1,1,1</td>
<td>2,2,1,1</td>
<td>2,1,2,1</td>
</tr>
</tbody>
</table>

(I have added some headings and spacings above to make the string of 1’s and 2’s easier to read.) In particular, we see that

- In pairs (subjects) 2, 3, 4, 6, 7, 11, 12, 14 and 16 treatment 1 (Drug A) is assigned to the first headache and treatment 2 (Drug B) is assigned to the second headache.

- In pairs (subjects) 1, 5, 8, 9, 10, 13 and 15 treatment 2 (Drug B) is assigned to the first headache and treatment 1 (Drug A) is assigned to the second headache.

Let me say a bit about why we randomize the order of the treatments within each pair. There are two main reasons:

1. If we performed randomization-based inference—we won’t because of time limitations—the process of randomization becomes the basis for our inference; in particular, the P-value is obtained by looking at all possible assignments in an RPD.

2. For scientific **validity**.
Regarding this second reason: As an honorable scientist you strive to learn things that are, indeed, true. But you also want the scientific community to take your work seriously.

For example, you might decide that randomizing is silly and a waste of effort. Instead you decide to have every subject take treatment 1 first and then treatment 2. At a personal level this might lead you to conclusions that are false. As a global matter I would be amazed if the scientific community paid much attention to your conclusions. The issue is that there might be an order effect in your study. What do I mean by this?

Imagine a situation in which all, or nearly all, subjects would give a lower response to their first headache than to their second headache, even if the pain levels were, indeed, identical. Or imagine the opposite pattern, where responses are systematically lower on the second headache compared to the first. In either of these situations, a decision to always give treatment 1 first would bias the study. The possibility of such an order effect causes the scientific community to discount, or even ignore, your findings.

Let’s look at the randomization I obtained above for the headache RPD. In nine pairs drug A is taken before drug B, and in only seven pairs drug B is taken before drug A. Thus, if there is indeed an order effect, one of the drugs (I can’t tell which one without knowing the direction of the order effect) has a slight advantage over the other.

There is available to a researcher a design that is a bit more complicated than an RPD. It is called the crossover design and it has two features that are not present in an RPD:

1. A crossover design forces balance between what I earlier called ‘1’ and ‘2.’ More precisely, the number of pairs that have treatment 1 first (what I called ‘1’) is exactly equal to the number of pairs that have treatment 2 first (what I called ‘2’). Thus, unlike my RPD above which had nine 1’s and seven 2’s, the crossover design would, perforce, have eight of each. This makes obvious a modest limitation on a crossover design: the number of pairs must be an even number.

2. For a crossover design, the analysis of the data explicitly incorporates—and estimates—the order effect, as compared to an RPD—the population-based method is given above—that ignores a possible order effect in the analysis. As a result, the analysis of a crossover design is more complicated than the analysis of an RPD. If, indeed, there is a large order effect (admittedly, large is vague here) then a crossover design can be more powerful than the corresponding RPD. Sadly, because these notes cannot cover every topic in Statistics, I will not show you how to analyze data from a crossover design.

20.4 Other Ways to Form Pairs

Thus far, I have discussed subject reuse as the only way to obtain paired data. Other methods are possible. I am going to be very cautious in my presentation of this material.

1. I will show you a situation other than subject reuse for which pairing is valid.

2. I will show you a situation other than subject reuse for which pairing gives wildly invalid results.
I will give you a rule that helps distinguish between these situations, but there will be holes in my rule; i.e., my rule does not necessarily cover every situation that could arise in science. Why? My standard reason: we cannot cover everything in a one semester course.

I begin with a situation in which pairing is valid.

### 20.4.1 Forming Pairs from Adjacent Trials

Let’s return to the game of Tetris. I want to focus (again) on an entire game as a trial. Many years ago, I enjoyed playing Tetris. My game had a feature that allowed the player to see or not see the next shape while manipulating the current shape. (Seeing was the default.) It seemed to me that selecting the default, preview, option would lead to much higher scores. So, I decided to collect data to investigate this matter.

A game is a trial and the response is the number of lines I completed before the game ended. I decided to perform 20 trials, with 10 on each setting. I was very worried that fatigue or boredom would affect my later scores, so I formed pairs out of consecutive trials: 1 and 2; 3 and 4; and so on. I will slow down and present these ideas carefully. Please refer to Table 20.4.

Find the rows that begin with **Trial**. The first such row lists trials 1–10 and the second such row lists trials 11–20. I would prefer it if these 20 trials were physically all in the same row, but our paper isn’t wide enough.

Find trials 1 and 2; in the row immediately above, these trials are identified as the trials that form pair 1. Next, you see that trials 3 and 4 form pair 2; trials 5 and 6 form pair 3; and so on; and trials 19 and 20 form pair 10.

Next, I went to the randomizer—details not shown—and it gave me the following assignment:

<table>
<thead>
<tr>
<th>Pair</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>Randomizer gives</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

The randomizer gave ‘1’ to pairs 2–5, 7 and 9. This means that within these pairs, the first game was played on treatment 1 (preview) and the second game was played on treatment 2 (no preview). The randomizer gave ‘2’ to pairs 1, 6, 8 and 10. This means that within these pairs, the first game was played on treatment 2 (no preview) and the second game was played on treatment 1 (preview). This explanation I have just given can be seen in the **Treatment** rows of Table 20.4.

After all of this work, it was time for me to have fun! I finally was able to play my 20 games of Tetris. In the first game, I set the machine to no preview—treatment 2—and obtained a score of 84. In the second game, I set the machine to preview—treatment 1—and obtained a score of 106. And so on, as displayed in the **Response** rows of Table 20.4.

Table 20.4 provides an accurate description of how my data were collected, but it needs to be rewritten to facilitate a statistical analysis. Table 20.5 rewrites my data in a form that is ready for analysis. (You should check to make sure you understand how I used Table 20.4 to create Table 20.5. You don’t need to check every entry, just make sure that you understand the process.)

Not surprisingly, and obviously from even a quick glance at the data, I was a much better player with the preview option. It is not so clear that pairing was beneficial; we shall explore this issue below.
Table 20.4: The RPD to compare the preview and no preview options in Tetris.

<table>
<thead>
<tr>
<th>Pair</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trial</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Treatment</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Response</td>
<td>84</td>
<td>106</td>
<td>112</td>
<td>93</td>
<td>118</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Pair</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trial</td>
<td>11</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>15</td>
</tr>
<tr>
<td>Treatment</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Response</td>
<td>88</td>
<td>110</td>
<td>130</td>
<td>108</td>
<td>91</td>
</tr>
</tbody>
</table>

Table 20.5: Paired data to compare the preview and no preview options in Tetris.

<table>
<thead>
<tr>
<th>Pair</th>
<th>Treatment</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>1: Preview ((x))</td>
<td>106</td>
<td>112</td>
<td>118</td>
<td>102</td>
<td>112</td>
<td>110</td>
<td>130</td>
<td>110</td>
<td>127</td>
<td>138</td>
<td></td>
</tr>
<tr>
<td>2: No preview ((y))</td>
<td>84</td>
<td>93</td>
<td>86</td>
<td>86</td>
<td>94</td>
<td>88</td>
<td>108</td>
<td>91</td>
<td>79</td>
<td>91</td>
<td></td>
</tr>
<tr>
<td>Difference ((d = x - y))</td>
<td>12</td>
<td>19</td>
<td>32</td>
<td>16</td>
<td>18</td>
<td>22</td>
<td>22</td>
<td>19</td>
<td>48</td>
<td>47</td>
<td></td>
</tr>
</tbody>
</table>

I calculated the following summary statistics:

\[ \bar{x} = 116.5, \ s_1 = 11.56, \ \bar{y} = 90.0, \ s_2 = 7.77, \ \bar{d} = 26.5, \ s_d = 11.87 \text{ and } m = 10. \]

With \(df = 9\), the value needed for the 95% confidence interval is \(t^* = 2.262\). Thus, the 95% confidence interval for \(\mu_d\) is

\[ 26.50 \pm 2.262(11.87/\sqrt{10}) = 26.50 \pm 2.262(3.754) = 26.50 \pm 8.49 = [18.01, 34.99]. \]

At the 95% confidence level, my population mean score with the preview feature is between 18 and 35 lines larger than my population mean score without the preview feature.

I can also perform a test of hypotheses on my Tetris data. Using the *Inconceivable Paradigm*, I select \(\mu_d > 0\) as my alternative. The observed value of the test statistic is

\[ t = 26.50/3.754 = 7.059. \]

The area under the t-curve with \(df = m - 1 = 9\) to the right of 7.059 is (using Minitab) \(0.0000296\), just smaller than 3 in one-hundred-thousand. This is the approximate P-value for \(>\). Note that if the alternative had been \(\neq\), then the approximate P-value would be twice as large, just smaller than 6 in one-hundred-thousand.

For comparison, we will now pretend that the data come from independent random samples. First,

\[ s_p^2 = [(11.56)^2 + (7.77)^2]/2 = 97.00. \]

526
Thus, \( s_p = \sqrt{97} = 9.85 \). Moreover, for \( df = 16 + 16 - 2 = 30 \), we get \( t^* = 2.101 \). Thus, the 95% confidence interval for \( \mu_1 - \mu_2 \) is

\[
(116.5 - 90.0) \pm 2.101(9.85)\sqrt{1/10 + 1/10} = 26.50 \pm 9.25 = [17.25, 35.75].
\]

The half-width for the RPD interval, 8.49, is 8.2% smaller than the half-width, 9.25, for the pretend CRD. Thus, pairing seems to have been effective, but not as dramatically as it was in my first set of artificial data on headache pain.

### 20.4.2 When is it Valid to do a Paired Data Analysis?

The methods given above—confidence intervals and tests of hypotheses using the differences as data—are valid in the following situations. Of course, implicit is the notion that we are willing to assume we have i.i.d. random variables.

1. It is valid if the units (subjects or trials) are reused. Scientifically, it is much better if one is able to use randomization, but it’s not necessary for statistical validity. In my experience, the most common type of unit reuse in an observational study is a before/after study.

2. If one forms pairs of units (trials or subjects) by matching different units based on some feature—often prognosis in medicine, then the analysis is valid in two situations:
   
   (a) Within each pair, units are assigned to treatments by randomization.
   
   (b) For an observational study on two finite populations of subjects, pairs are formed at the population level. If pairs are formed at the sample level, regardless of how, then a paired data analysis is invalid and, indeed, can be grossly misleading. The notions of population level and sample level are discussed below.

I will now provide some examples of the ideas listed above.

First, let’s look at an example of a before and after study. Suppose we have \( m = 50 \) subjects who are interested in losing weight. A study might proceed as follows. Each person is weighed at the beginning of the study. Each person then follows a rigorous program of diet and exercise for, say, three months. at which time each person is weighed again. If \( x [y] \) is a subject’s weight at the beginning [end] of the study, then \( d = x - y \) is the amount of weight the subject lost during the study. (Keep in mind that \( d < 0 \) means that the subject’s weight increased.)

Is this weight loss study really paired data?

- Yes, because we get two numbers from each subject and it is meaningful to calculate their difference.
- No, because we can view the data as one response, the difference in weights.

In my opinion, it does not matter which of these viewpoints you adopt, the data are analyzed the same way and the scientific interpretation is unchanged.
Too often in a before and after study, researchers forget the need for a control group, as illustrated in the following example. **Full disclosure:** I found this example years ago in a textbook on designing experiments in the social sciences. I don’t have a reference; and I can’t swear that the authors were being honest!

Anyways, in August, 1939, students at an American university were given a pre-test to measure their attitudes towards the government of Nazi Germany. Then they took a four-week course that presented that government in a positive light. At the end of the course, the students were given a post-test to determine the extent to which the course influenced the students’ attitudes. There was, however, an unforeseen difficulty: On September 1, 1939, while the course was still in session, Germany invaded Poland, starting World War 2. As a result, I sincerely doubt that the differences between pre-test and post-test scores were due to the course! A control group would have improved this study greatly, but I suspect it was doomed in any event.

Regarding item 2(a) in our earlier list: forming pairs of different units and then randomizing the assignment of unit to treatment within each pair. I advocate this method for trials—as I demonstrate above for my Tetris study—but am not a fan of this method for subjects. For subjects, I believe it is better to form blocks of subjects, as I describe in Chapter 4 of my textbook, *Statistics: Learning in the Presence of Variation*. In addition, if you do form pairs this way, I believe that randomization-based inference is valid, but not population-based inference. Not everybody, however, agrees with me. Sadly, we have time for neither this topic nor a presentation on blocks.

The remainder of this subsection is devoted to item 2(b) in my list: forming pairs at the population level and sample level.

Are husbands taller than their wives? Are husbands older than their wives? Personally, I have never been interested in either of these questions, but I must admit that during my long life, I have heard many people talk about them. More pragmatically, I can’t think of an example of pairing at the population level other than one involving husbands and wives.

First, a disclaimer. At the time of my typing these words, I live in Wisconsin, a state in which a legal marriage consists of exactly two people, one of each sex. The fact that my example is restricted to such pairs should not be interpreted in any way politically, etc.

Let’s focus on height. There is a population of husbands in Wisconsin and there is a population of wives in Wisconsin. Let \( \mu_1 \) denote the mean height of the husbands and \( \mu_2 \) denote the mean height of the wives. My goal is to estimate \( \mu_1 - \mu_2 \). Here are two statistically valid ways for me to sample these populations:

1. I could select a random sample of \( m \) men from the population of husbands. I could select a random sample of \( m \) women from the population of wives. I would have my samples be independent of each other. I would determine the height of each of the \( 2m \) persons in my study.

2. I could select a random sample of \( m \) women from the population of wives. I would determine the height of each of the \( m \) women in my study as well as the heights of their husbands.

With the first method, I would analyze the data with the methods of Chapter 19, independent samples. For the second method, I would analyze the data with the methods of this chapter, paired data. Based on my many years of observation of married couples in Wisconsin, I conjecture that
there is a pretty strong positive correlation between the heights of husbands and wives; thus, I would use the second method of sampling. If my conjecture is correct, my paired data analysis will be more efficient than independent samples would have been. If my conjecture is wrong, my paired data will still be valid, but it won’t be as efficient as independent samples would have been.

To summarize, I formed pairs of all members of the two populations—which, necessarily, needed to have exactly the same number of members. This is what I mean by forming pairs at the population level.

I end this material with a cautionary tale that, I hope, will convince you to never form pairs at the sample level, but, first, a story from my career.

For part of my career at the University of Wisconsin–Madison, part of my job was to provide statistical advice to graduate students from other departments. One day a student came to me with her data. She was willing to assume that she had independent random samples of size $n_1 = n_2 = 40$ from two populations. (Her advisor said that) She needed to show that the two populations had different means. She knew the methods of Chapter 19 and had applied them to her data. Sadly, following standard statistical reasoning, she could not conclude that the population means were different.

After she explained all of the above to me, we had the following conversation:

**BW:** So, why are you here?

**Student:** Somebody told me that if I had paired data I would get a smaller P-value and, thus, be able to make my advisor happy.

**BW:** That might be true. Do you want to do a new study, one with paired data?

**Student:** No, I want you to pair my data.

**BW:** Huh? I don’t understand.

**Student:** (With exasperation) I want you to take my data, manipulate them into pairs to give me the answer I need.

**BW:** Oh.

I have cited the above exchange many times in my teaching. I then point out that unlike physics, chemistry and mathematics, there is no demon in Statistics. What do I mean by this? Well, if you don’t believe in physics, gravity might kill you. If you don’t believe in chemistry, a mixture of ammonia and bleach might kill you. If you don’t understand fractions, somebody might take all of your money by continually forcing you to make change. You can, however, perform any number of ridiculous statistical analyses and nothing bad will happen to you!

Suppose that you want to determine which university has taller men: UW–Madison or the University of Minnesota–Twin Cities. (I realize that this is silly; bear with me please.) You select independent random samples of sizes $n_1 = n_2 = 40$ from both populations. Denote your observed data from Wisconsin by:

$$x_1, x_2, x_3, \ldots, x_{40}.$$  

Similarly, denote your observed data from Minnesota by:

$$y_1, y_2, y_3, \ldots, y_{40}.$$
Sort each set of data, yielding

\[ x_1 \leq x_2 \leq x_3 \leq \ldots \leq x_{40} \quad \text{and} \quad y_1 \leq y_2 \leq y_3 \leq \ldots \leq y_{40}. \]

Thus, for example, \( x_1 [y_1] \) is the height of the shortest of the 40 men in the Wisconsin [Minnesota] sample; \( x_{40} [y_{40}] \) is the height of the tallest of the 40 men the Wisconsin [Minnesota] sample; and so on.

Next, we form pairs at the sample level: we match \( x_1 \) with \( y_1 \); \( x_2 \) with \( y_2 \); and so on; and we match \( x_{40} \) with \( y_{40} \). In other words, we form pairs based on the value of the response. Finally, I create 40 \( d \)'s for my data set:

\[ d_1 = x_1 - y_1; \quad d_2 = x_2 - y_2; \quad \ldots; \quad d_{40} = x_{40} - y_{40}. \]

I summarize my 40 differences with \( \bar{d} \) and \( s_d \). Finally, I calculate Gosset’s 95% confidence interval for \( \mu_d = \mu_1 - \mu_2 \):

\[ \bar{d} \pm 2.023 \left( \frac{s_d}{\sqrt{40}} \right). \]

What happens if we do this? To answer this question, I need to involve Nature and computer simulations.

Suppose, for example, Nature knows that the two populations are identical and both are the Normal curve with \( \mu = 69 \) inches and \( \sigma = 3 \) inches. Thus, a confidence interval for \( \mu_d = \mu_1 - \mu_2 \) will be correct if, and only if, it contains zero. I performed a simulation experiment with 10,000 reps to investigate the actual performance of this confidence interval for paired data. The results were:

- A total of 3,523 confidence intervals were too large;
- A total of 3,527 confidence intervals were too small; thus,
- A total of 7,050 confidence intervals were incorrect.

Note that there should be approximately 500 incorrect confidence intervals. Seven thousand fifty is quite a bit larger than 500. This simulation study shows convincingly that forming pairs based on the response is invalid! By the way, the above simulation applies to all pairs of Normal curves that are congruent; i.e., the two population means don’t need to be the same number. Similar results will be obtained for noncongruent Normal curves, but they will require a different simulation experiment to discover just how horrible the method performs! Similar results are true for populations that are not Normal curves. In short, this method is always bad!

I have never found a textbook that was shameless enough to propose the above method—sort the data, form pairs, subtract, etc. Alarmingly, however, I did find several textbooks that advocated the following form of experimental design. I will state their suggestion in terms of the above height study.

They do not say, “Form pairs based on the response, height;” instead, they advocate forming pairs based on another feature that is correlated with height, perhaps weight. **This is also invalid!** If you do this the actual confidence level of your nominal 95% confidence interval will be much smaller than 95%.
20.5 An Extended Example

Pairing is a very exciting topic. (I know, exciting is like funny; if it’s really funny, do I need to tell you?) It is exciting because it allows a researcher to use scientific knowledge to improve a study; i.e., it’s not about math or algebra.

When I decide to investigate a topic statistically, after I have a general notion of the response, I always ask myself the following two-part question:

1. What factor(s) do I suspect will cause variation in the responses from unit to unit?

2. Of the factors listed, can I deal with one or more of them by forming pairs?

Are you a fan of major league baseball? Well, sadly, this example will be more interesting if you are. In any event, I will proceed.

One of the charms of major league baseball is that the dimensions of the 30 major league ballparks are not constant. The most famous ballpark (some residents of New York and Chicago might disagree) is Fenway Park in Boston. The distance from home plate to its left field fence is the shortest in the major leagues, partly offset by the fact that said fence is the tallest, measuring 37 feet, two inches high. The second most famous ballpark (again, according to me) is Wrigley Field in Chicago. Wrigley Field is the topic of this example.

Wrigley Field has a reputation for being hitter friendly; in particular, the conventional wisdom is that it is easier to hit a home run in Wrigley Field than in an average ballpark. I will investigate this issue. How might one investigate this issue?

Here is my first attempt. Take for my response the total number of home runs hit in a stadium in a year. Think of this as a 30-population problem, with each year giving us another observation for each of the 30 populations (ballparks).

For example, in the 2013 National League season, a total of 2,145 home runs were hit in its 15 ballparks, for a mean of \( \frac{2145}{15} = 143 \) home runs per ballpark. The four largest responses are: Milwaukee, 185; Cincinnati, 184; Philadelphia, 176; and Chicago, 175. The three smallest responses are: Miami, 84; Pittsburgh 106; and St. Louis, 108. The Wrigley Field data support the conventional wisdom; the number of home runs hit there was well above the mean.

Do you see a weakness in the above discussion? Here is one. To paraphrase the NRA, “Ballparks don’t hit home runs, players do.” It is inarguable that the management of a baseball team considers its ballpark while building its roster of players. Thus, for example, part of the reason there were more home runs hit in Milwaukee than in Miami is that the former’s roster contained more power hitters.

Many years ago somebody—sorry, I don’t know who gets credit—had a clever idea. Let me show you some data and then explain the idea. Look at Table 20.6. Let’s look at 1967. We see the response value 160 for the Cubs’ (Chicago’s team) home games. This is the total number of home runs hit in Wrigley Field in 1967. Thus, it is the same idea as the response values I gave you earlier for the 2013 season. Here is the twist. We compare this value, 160, to the total number of home runs hit by both teams in all of the Cubs’ away games, 110. By comparing all of the Cubs’ home games with all of their away games, we have—for the most part—removed the effect of rosters.
Table 20.6: Number of home runs, for both teams, in Cubs games, 1967–1987.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Home</td>
<td>160</td>
<td>166</td>
<td>148</td>
<td>201</td>
<td>144</td>
<td>146</td>
<td>138</td>
<td>139</td>
<td>125</td>
<td>155</td>
</tr>
<tr>
<td>Away</td>
<td>110</td>
<td>102</td>
<td>112</td>
<td>121</td>
<td>116</td>
<td>99</td>
<td>107</td>
<td>93</td>
<td>100</td>
<td>73</td>
</tr>
<tr>
<td>Home—Away</td>
<td>50</td>
<td>64</td>
<td>36</td>
<td>80</td>
<td>28</td>
<td>47</td>
<td>31</td>
<td>46</td>
<td>25</td>
<td>82</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Home</td>
<td>151</td>
<td>117</td>
<td>151</td>
<td>116</td>
<td>115</td>
<td>140</td>
<td>156</td>
<td>202</td>
<td>168</td>
<td>204</td>
</tr>
<tr>
<td>Away</td>
<td>88</td>
<td>80</td>
<td>111</td>
<td>100</td>
<td>112</td>
<td>117</td>
<td>79</td>
<td>104</td>
<td>130</td>
<td>164</td>
</tr>
<tr>
<td>Home—Away</td>
<td>63</td>
<td>57</td>
<td>40</td>
<td>16</td>
<td>3</td>
<td>23</td>
<td>77</td>
<td>98</td>
<td>38</td>
<td>40</td>
</tr>
</tbody>
</table>

By the way, here are two summary statistics for the data in Table 20.6:

\[ \bar{d} = 46.20 \text{ and } s_d = 24.42. \]

If I view the 20 seasons of data as the result of observing 20 i.i.d. trials, then I can obtain a 95% confidence interval estimate of \( \mu_d \):

\[ 46.20 \pm 2.093(24.42/\sqrt{20}) = 46.20 \pm 11.43. \]

In words, on average, Wrigley Field increases the mean number of home runs by at least 34.77 and at most 57.63 per season.
20.6 Computing

The extremely versatile and useful *vassarstats* website can be used to analyze paired data. I will illustrate the method for my Tetris data in Table 20.5. Go to the website:

http://vassarstats.net

The left-side of the page lists a number of options; click on *t-Tests & Procedures*. This takes you to a new set of options; click on the top one, *Two-Sample t-Test for Independent or Correlated Samples*. This takes you to a new page. In the *Setup* section, click on *Correlated Samples*. (If you forget to do this, it’s a big problem because *Independent Samples* is the default.) Next, enter the data, by typing or pasting, and click on *Calculate*.

The website gives me the following information:

\[ m = n_A = n_B = 10; \bar{x} = 116.5; \bar{y} = 90.0; \text{ and } \bar{d} = \text{Mean}_a - \text{Mean}_b = 26.5. \]

It also reports that the observed value of the test statistic is \( t = 7.06 \) with approximate P-value \(< 0.0001\) for both of the alternatives \( > \) and \( \neq \). The *vassarstats* testing output is all consistent, though a bit less precise, than what I obtained earlier by hand. Finally, *vassarstats* reports a variety of confidence intervals, including \( 26.5 \pm 8.4846 \) as the 95% confidence interval estimate of \( \mu_d \). This is the same answer I obtained, except for round-off error.

I have found a website that will create a scatterplot and compute the correlation coefficient:

http://www.wessa.net/rwasp_Pregnancy%20and%20cognition.wasp#output

You are not responsible for using this site; I am very grateful that it exists, but it’s a bit tedious to use. (For example, it requires a fair amount of time to delete the site’s default data before you can enter your own data.) If you want to try it out; I suggest that you use my Tetris data. As a partial check, you should obtain \( r = 0.2955 \) for the correlation coefficient. Or you could use either one of my RPDs for headache pain to see whether your output matches the scatterplot in Figure 20.1.
20.7 Summary

This chapter continues the theme of Chapter 19; namely, comparing the means of two populations. Instead of independent samples, in this chapter we have paired data. Each pair gives: a response from population 1, \( x \); and a response from population 2, \( y \). These two numbers can be used to compute \( d = x - y \), which can be viewed as a response from the population of differences.

Here’s another way to view this structure: We have random samples from both populations 1 and 2, but the random samples are not independent of each other. The result is that we have a random sample from the population of differences.

In Chapter 19, when considering population means, we focused on estimating \( \mu_1 - \mu_2 \) with confidence and testing the null hypothesis that \( \mu_1 = \mu_2 \). In the current chapter, these inference problems become estimating \( \mu_d \) with confidence and testing the null hypothesis that \( \mu_d = 0 \).

Mathematically, Chapter 20 reduces to the problem of inference for a single population mean (of the population of differences). This problem was studied in Chapters 17 and 18 and I recommend using Gosset’s procedures, subject to the caveats mentioned in these earlier chapters.

A researcher needs to be careful to avoid misusing the formulas for paired data. In particular, we found that the methods for paired data are appropriate for:

- Unit reuse, with or without randomization;
- Forming pairs of adjacent trials, using randomization to assign one trial of each pair to each treatment; and
- Matching subjects at the population level, as described earlier.

I gave a simple and dramatic example illustrating that paired data methods should never be used for pairing performed at the sample level, again, as described earlier.

Finally, you learned how to create and interpret a scatterplot of pairs of responses. Thinking about the likely pattern in such a scatterplot can help a researcher decide whether to have a design with independent samples or paired data.
Table 20.7: Data for the RPDs described in Practice Problems 1 and 2.

<table>
<thead>
<tr>
<th>Pair</th>
<th>Treatment 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>1</td>
<td>173</td>
<td>175</td>
<td>169</td>
<td>175</td>
<td>180</td>
<td>184</td>
<td>182</td>
<td>186</td>
<td>190</td>
<td>188</td>
</tr>
<tr>
<td>2</td>
<td>163</td>
<td>160</td>
<td>157</td>
<td>165</td>
<td>167</td>
<td>167</td>
<td>159</td>
<td>170</td>
<td>155</td>
<td>164</td>
</tr>
<tr>
<td>3</td>
<td>Difference</td>
<td>10</td>
<td>15</td>
<td>12</td>
<td>10</td>
<td>13</td>
<td>25</td>
<td>12</td>
<td>31</td>
<td>37</td>
</tr>
</tbody>
</table>

20.8 Practice Problems

1. Bascom Hill is a long, steep hill (by Wisconsin standards) in the center of the university campus in Madison. A student in my class, Damion, wondered whether smoking a cigarette affected his climbing of Bascom Hill. He performed an RPD with response equal to the time, measured to the nearest second, he needed to walk from the bottom to the top of the hill. The first treatment consisted of walking while smoking a cigarette and the second consisted of walking while not smoking a cigarette. Damion formed pairs from his trials, exactly as I did for the Tetris study described in this chapter.

Damion’s data are in Table 20.7. Below are various summary statistics for these data.

\[ \bar{x} = 180.2, \bar{y} = 161.3, s_1 = 6.99, s_2 = 5.44 \text{ and } s_d = 9.67. \]

(a) Calculate Gosset’s 95% confidence interval estimate of \( \mu_d \). Write one sentence that interprets your confidence interval.

(b) Find the approximate P-value for the alternative \( \mu_d > 0 \).

(c) Pretend that the data came from a CRD instead of an RPD. Calculate Gosset’s 95% confidence interval estimate of \( \mu_1 - \mu_2 \).

(d) Compare your answers to (a) and (c). In your opinion, which would have been a better way to conduct the study; an RPD or a CRD? Explain your answer.

(e) Use Equation \[20.4\] to determine the value of the correlation coefficient for \( x \) and \( y \).

2. Now suppose that Damion had ended his RPD after the first five pairs were completed. Use the vassarstats website to redo problem 1. For part (e), to save time you may use the following summary statistics, which I obtained from vassarstats:

\[ s_1 = s_2 = 3.9749 \text{ and } s_d = 2.1213. \]

3. Alisa performed an RPD to compare bowling with one hand (the usual method) and bowling two handed (‘granny style;’ her words, not mine). A trial consisted of a game of bowling and the response was Alisa’s score. Below are the results of the study:
Game | Hands | Score | Game | Hands | Score
--- | --- | --- | --- | --- | ---
1 | One | 97 | 6 | One | 110
2 | Two | 85 | 7 | One | 123
3 | Two | 91 | 8 | Two | 96
4 | One | 108 | 9 | One | 125
5 | Two | 95 | 10 | Two | 94

(a) Present these data in a format similar to what I used in Table 20.7. Put one-handed bowling in the first row; i.e., the $x$’s. Don’t analyze these data; I simply want you to make sure you can transform one table into another.

(b) Assuming Alisa used our website randomizer, what output did it give her?

20.9 Solutions to Practice Problems

1. (a) First, 
$$\bar{d} = \bar{x} - \bar{y} = 180.2 - 161.3 = 18.9.$$ 
Next, $t^*$ for $df = 10 - 1 = 9$ is 2.262. Thus, Gosset’s 95% confidence interval estimate of $\mu_d$ is:
$$18.90 \pm 2.262(9.67/\sqrt{10}) = 18.90 \pm 2.262(3.058) = 18.90 \pm 6.92 = [11.98, 25.82].$$

The mean time to walk up the hill while smoking is between 11.98 and 25.82 seconds larger than the mean time to walk up the hill while not smoking.

(b) The observed value of the test statistic is 
$$t = 18.90/3.058 = 6.1805.$$ 

The area under the t-curve with $df = 9$ to the right of 6.1805 equals (using Minitab) 0.00008. This is the approximate P-value.

(c) First, 
$$s_p^2 = \frac{(6.99)^2 + (5.44)^2}{2} = 39.227 \text{ and } s_p = \sqrt{39.227} = 6.263.$$ 
Next, $t^*$ for $df = 10 + 10 - 2 = 18$ is 2.101. Thus, Gosset’s 95% confidence interval estimate of $\mu_1 - \mu_2$ is:
$$18.90 \pm 2.101(6.263)\sqrt{2/10} = 18.90 \pm 5.88 = [13.02, 24.78].$$

(d) The half-width from the pretend CRD, 5.88, is 15.0% narrower than the half-width from the actual RPD, for the same data. This supports the notion that a CRD would have been better, but we don’t really know what would have happened with a CRD.
(e) First,
\[ s_d^2 = (9.67)^2 = 93.5089. \]

Next,
\[ s_1^2 + s_2^2 - 2rs_1s_2 = (6.99)^2 + (5.44)^2 - 2r(6.99)(5.44) = 78.4537 - 76.0512r. \]

Setting these equal to each other, we get:
\[ 76.0512r = 78.4537 - 93.5089 = -15.0552 \text{ or } r = -0.198. \]

2. I entered the data into vassarstats, being careful to specify Correlated Samples and obtained the following relevant summaries:
\[ \bar{x} = 174.4, \bar{y} = 162.4 \text{ and } \bar{d} = 12.0. \]

(a) The website tells me that the 95% confidence interval estimate of \( \mu_d \) is 12.00 ± 2.64. Notice that this interval is much narrower than the interval from all ten pairs!

(b) The website tells me that the observed value of the test statistic is \( t = 12.65 \) with \( df = 4 \) and that the approximate P-value for > is 0.0001125.

(c) I enter the same data into the website, being careful to specify Independent Samples. The site tells me that the 95% confidence interval estimate of \( \mu_1 - \mu_2 \) is 12.00 ± 5.81.

(d) The half-width of the confidence interval for the actual RPD, 2.64, is 54.6% narrower than the half-width for the pretend CRD with the same data. This is a huge difference! We don’t know for sure, however, what would have happened with a CRD, but the RPD does look better.

(e) From Equation 20.4
\[ (2.1213)^2 = (3.9749)^2 + (3.9749)^2 - 2r(3.9749)^2. \]

This becomes:
\[ 4.499914 = 2(15.79983) - 2r(15.79983); \text{ or } 2r(15.79983) = 27.099746; \text{ or } r = 0.858. \]

3. (a) Alisa’s table is below.

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Pair</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>One-handed ((x))</td>
<td>97</td>
<td>108</td>
<td>110</td>
<td>123</td>
<td>125</td>
<td></td>
</tr>
<tr>
<td>Two-handed ((y))</td>
<td>85</td>
<td>91</td>
<td>95</td>
<td>96</td>
<td>94</td>
<td></td>
</tr>
<tr>
<td>Difference ((d = x - y))</td>
<td>12</td>
<td>17</td>
<td>15</td>
<td>27</td>
<td>31</td>
<td></td>
</tr>
</tbody>
</table>

(b) The site gave her: 1, 2, 2, 1, 1.
20.10 Homework Problems

1. Martha and Lisa performed an RPD to investigate Martha’s juggling skills. The first treatment was Martha juggling three tennis balls; the second treatment was Martha juggling three large apples. The response is the length of time, measured to the nearest second, that the three items were in what they called a regular cycle of juggling. Below are selected summary statistics:

\[ \bar{x} = 6.100, \bar{y} = 5.200, s_1 = 3.888, s_2 = 3.517, s_d = 5.826 \text{ and } m = 40. \]

(a) Calculate Gosset’s 95% confidence interval estimate of \( \mu_d \). Write one sentence that interprets your confidence interval.

(b) Find the approximate P-value for the alternative \( \mu_d > 0 \).

(c) Pretend that the data came from a CRD instead of an RPD. Calculate Gosset’s 95% confidence interval estimate of \( \mu_1 - \mu_2 \).

(d) Compare your answers to (a) and (c). In your opinion, which would have been a better way to conduct the study; an RPD or a CRD? Explain your answer.

(e) Use Equation 20.4 to determine the value of the correlation coefficient for \( x \) and \( y \).

2. Deborah’s son Scotty is convinced that his Snowbie sled is slower than his friend Sam’s Sno-Racer. An RPD was conducted to investigate this issue. A trial consisted of a slide down a local hill. The response is the time, measured to the nearest tenth of a second, that Scotty required to complete a slide. The first treatment consists of Scotty riding his Snowbie and the second treatment is Scotty riding Sam’s Sno-Racer. Below are the results of the study.

<table>
<thead>
<tr>
<th>Trial</th>
<th>Treat.</th>
<th>Time</th>
<th>Trial</th>
<th>Treat.</th>
<th>Time</th>
<th>Trial</th>
<th>Treat.</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>11.3</td>
<td>7</td>
<td>2</td>
<td>9.0</td>
<td>13</td>
<td>2</td>
<td>10.1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>12.0</td>
<td>8</td>
<td>1</td>
<td>12.1</td>
<td>14</td>
<td>1</td>
<td>8.8</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>11.3</td>
<td>9</td>
<td>1</td>
<td>8.9</td>
<td>15</td>
<td>2</td>
<td>9.9</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>11.1</td>
<td>10</td>
<td>2</td>
<td>10.7</td>
<td>16</td>
<td>1</td>
<td>10.5</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>10.1</td>
<td>11</td>
<td>2</td>
<td>10.6</td>
<td>17</td>
<td>1</td>
<td>12.2</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>8.4</td>
<td>12</td>
<td>1</td>
<td>9.8</td>
<td>18</td>
<td>2</td>
<td>12.1</td>
</tr>
</tbody>
</table>

(a) Present these data in a format similar to what I used in Table 20.7.

(b) Use the vassarstats website to obtain Gosset’s 95% confidence interval estimate of \( \mu_d \).

(c) Use the vassarstats website to obtain Gosset’s approximate P-value for the alternative \( \mu_d > 0 \).

(d) Assuming Deborah used our website randomizer, what output did it give her?