Chapter 3

Randomization, Probability and Sampling Distributions

3.1 Assignments and Randomization

Recall that Dawn’s study of her cat Bob was presented in Chapter 1. Table 3.1 presents her data. Reading from this table, we see that in Dawn’s study, the chicken-flavored treats were presented to Bob on days (trials):

1, 5, 7, 8, 9, 11, 13, 15, 16 and 18.

Why did she choose these days? How did she choose these days? It will be easier to begin with the ‘How’ question.

We have been looking at the data collected by Dawn. We have listed the observations; separated them by treatment; sorted them within treatment; and, within treatments, drawn dot plots and computed means, medians, variances and standard deviations. But now we need to get into our time machine and travel back in time to before Dawn collected her data. We go back to when Dawn had her study largely planned: treatments selected; trials defined; response specified; and the decision to have a balanced study with 20 trials. We are at the point where Dawn pondered, “Which 10 trials should have chicken-flavored treats assigned to them? How should I decide?”

Table 3.1: Dawn’s data on Bob’s consumption of cat treats. ‘C’ [‘T’] is for chicken [tuna] flavored.

<table>
<thead>
<tr>
<th>Day:</th>
<th>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>Flavor:</td>
<td>C</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>C</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>T</td>
</tr>
<tr>
<td>Number Consumed:</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>0</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>Day:</td>
<td>11</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>15</td>
<td>16</td>
<td>17</td>
<td>18</td>
<td>19</td>
<td>20</td>
</tr>
<tr>
<td>Flavor:</td>
<td>C</td>
<td>T</td>
<td>C</td>
<td>T</td>
<td>C</td>
<td>T</td>
<td>C</td>
<td>T</td>
<td>T</td>
<td></td>
</tr>
<tr>
<td>Number Consumed:</td>
<td>6</td>
<td>3</td>
<td>7</td>
<td>1</td>
<td>3</td>
<td>6</td>
<td>3</td>
<td>8</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>
The answer is that Dawn did this by using a process called *randomization*. I will explain *what* randomization is by showing you three equivalent ways to randomize.

First, some terminology. We call the list of 10 trials above an *assignment* of treatments to trials. It tells us which trials were assigned to the first treatment (chicken). It also implies which trials were assigned to the second treatment (tuna); namely, all of the trials *not* listed above. If we are going to study assignments—and we are—it is easier if we make our assignments as simple to display as possible. **Thus, an assignment will be presented by listing the trials that it assigns to treatment 1.**

A natural question is: How many different assignments were possible for Dawn’s study? The answer is 184,756. I will give a brief digression into how I obtained this number.

You might recall from math the expression $m!$, which is read *em-factorial*. If $m$ is a positive integer, then this expression is defined as:

$$m! = m(m-1)(m-2)\cdots 1 \quad (3.1)$$

Thus, for example,

$$1! = 1; \quad 2! = 2(1) = 2; \quad 3! = 3(2)(1) = 6; \quad \text{and so on.}$$

By special definition (which will allow us to write more easily certain formulas that will arise later in these notes), $0! = 1$. Finally, for any other value of $m$ (negatives, non-integers), the expression $m!$ is not defined.

We have the following result. You don’t need to worry about proving it; it is a given in these notes.

**Result 3.1 (The number of possible assignments.)** For a total of $n = n_1 + n_2$ units, the number of possible assignments of two treatments to the units, with $n_1$ units assigned to treatment 1 and the remaining $n_2$ units assigned to treatment 2, is

$$\frac{n!}{n_1!n_2!} \quad (3.2)$$

I will evaluate Equation (3.2) for three of the studies presented in Chapters 1 and 2.

- For Cathy’s study, $n = 6$ and $n_1 = n_2 = 3$. Thus, the number of possible assignments is

$$\frac{6!}{3!3!} = \frac{6(5)(4)}{3(2)(1)} = 20.$$ 

Notice that it is *always* possible to reduce the amount of arithmetic we do by canceling some terms in the numerator and denominator. In particular, the $6!$ in the numerator can be written as

$$6(5)(4)3!$$

and its $3!$ cancels a $3!$ in the denominator.
• For Kymn’s study, \( n = 10 \) and \( n_1 = n_2 = 5 \). Thus, the number of possible assignments is

\[
\frac{10!}{5!5!} = \frac{10(9)(8)(7)(6)}{5(4)(3)(2)(1)} = 252.
\]

• For Dawn’s study, \( n = 20 \) and \( n_1 = n_2 = 10 \). Thus, the number of possible assignments is

\[
\frac{20!}{10!10!} = 184,756.
\]

Notice that for Cathy’s and Kymn’s study, I determined the answer by hand because the numbers are small enough to handle easily. Dawn’s study is trickier. Many of you, perhaps most, perhaps all, will consider it easy to determine the answer: 184,756. But I will not require you to do so. As a guide, I will never have you evaluate \( m! \) for any \( m > 10 \).

Sara’s study is a real challenge. The number of possible assignments is

\[
\frac{80!}{40!40!}.
\]

This answer, to four significant digits, is

\[
1.075 \times 10^{23}.
\]

Don’t worry about how I obtained this answer. If this issue, however, keeps you awake at night, then send me an email and I will tell you. If enough people email me, then I will put a brief explanation in the next version of these Course Notes.

I now will describe three ways—two physical and one electronic—that Dawn could have performed her randomization.

1. **A box with 20 cards.** Take 20 cards of the same size, shape, texture, etc. and number them 1, 2, \ldots, 20, with one number to each card. Place the cards in a box; mix the cards thoroughly and select 10 cards at random without replacement. The numbers on the cards selected denoted the trials that will be assigned treatment 1.

2. **A deck of 20 ordinary playing cards.** (This method is especially suited for units that are trials.) We need to have 10 black cards (spades or clubs) and 10 red cards (diamonds or hearts). We don’t care about the rank (ace, king, 3, etc.) of the cards. The cards are thoroughly shuffled and placed in a pile, face down. Before each trial select the top card from the pile; if it is a black card, then treatment 1 is assigned to the trial; if it is a red card, then treatment 2 is assigned to the trial. The selected card is set aside and the above process is repeated for the remaining trials.

3. **Using a website.** This method will be explained in Section 3.5 later in this chapter.

Are you familiar with the term **black box**? I like the definition in Wikipedia

\[\text{http://en.wikipedia.org/wiki/Black_box}\]
which is:

In science and engineering, a black box is a device, system or object which can be viewed solely in terms of its input, output and transfer characteristics without any knowledge of its internal workings, that is, its implementation is "opaque" (black). Almost anything might be referred to as a black box: a transistor, an algorithm, or the human mind.

Our website for randomization is a black box. It executes a computer program that supposedly is mathematically equivalent to my two methods of randomization that involve using cards. I say it's a black box because we aren't really interested in the computer code of the program.

Thus, if you want to think about how randomization works, I recommend you think of the cards. For my purposes of instruction the most convenient method for me is to use the website to obtain examples for you. If I were to replicate Dawn’s study on my cat Buddy I would use the second card method above. But that’s me. If you perform a project for this class, you may randomize however you please, as long as you use one of the three methods above.

Before I get back to Dawn’s study, I want to deal with an extremely common misconception about randomization. Randomization is a process or a method that is fair in the sense that every possible assignment has the same chance of being selected. Randomization does not guarantee that the assignment it yields will look random. (Among the many issues involved is how one decides what it means for an assignment to look random.) Later in these Course Notes, we will discuss designs that involve restricted randomization; in particular, we will learn about the Randomized Pairs Design.

I used the website and obtained the following assignment for Dawn’s study:

1, 2, 4, 7, 9, 10, 11, 14, 15, 18.

Note that this assignment is different from the one that Dawn used. Given that there are 184,756 possible assignments I would have been very surprised if I had obtained the same assignment as Dawn!

Here is the commonality of our three methods of randomization: Before we select an assignment, all 184,756 possible assignments for Dawn’s study are equally likely to be selected. For the first method of randomizing, this fact is conveyed by saying that the cards are indistinguishable; they are thoroughly mixed; and 10 cards are selected at random. For the second method of randomizing, this fact is conveyed by saying that the cards are shuffled thoroughly. Finally, whereas the electronic methods’ operations are a total mystery to us, the programmer claims that it makes all assignments equally likely. In this class we will use the website randomizer for a variety of purposes (not just randomization) and we will accept without question its claim of making every assignment equally likely.

The most important thing to remember about probability is that it is always about a future uncertainty.

Now, back to Dawn’s study of Bob. Remember, we are at a place in time before Dawn selected an assignment. Thus, it makes sense to talk about probabilities. Because all assignments are
equally likely to be selected, we assign each of them the same probability of being selected, as given below.

\[ P(\text{Any particular assignment}) = \frac{1}{\text{Total number of possible assignments}} \quad (3.3) \]

For Dawn’s study, the probability of any particular assignment is \(1/184,756\). For Cathy’s study, the probability of any particular assignment is \(1/20 = 0.05\) because, recall (Equation 3.2) that for \(n_1 = n_2 = 3\), there are 20 possible assignments of treatments to units.

We will be interested in combining assignments (perhaps with some common feature) into a collection of assignments. Let \(A\) denote any collection of assignments. Such a collection is called an **event**. Before we select an assignment via randomization, it makes sense to ask whether or not the assignment that will be obtained (note the future tense) is one of the assignments belonging to the event \(A\). If it is, we will say that the event \(A\) has occurred. We have the following definition of the probability that an event \(A\) will occur.

\[ P(A) = \frac{\text{Number of assignments in } A}{\text{Total number of possible assignments}} \quad (3.4) \]

An important example of a collection of assignments is the collection of all possible assignments; it is called the **Sample Space** and is denoted by \(S\) (an upper case ess with an attitude). From Equation (3.4) we see that the probability of the sample space occurring is 1 or 100%.

The sample space is called the certain event. Why? Well, by definition it contains all possible assignments; thus, it is certain to occur. Thus, we see that certainty corresponds to a probability of 1. An empty collection of assignments is impossible to occur and has probability equal to 0. Thus, probability is a quantification of uncertainty for which 0 corresponds to impossible and 1 corresponds to certainty.

### 3.2 The Skeptic’s Argument

Recall my earlier discussion of within- and between-variation. Let’s return to within variation. Consider, for example, Dawn’s sorted chicken data: 1, 3, 4, 5, 5, 6, 6, 6, 7 and 8. These numbers vary. This fact cannot be debated. In words, some days Bob consumed a large number of chicken treats and other days he consumed very few. Bob exhibited a large amount of day-to-day variation in his eating of chicken treats. In somewhat more picturesque language, some days Bob was very hungry; other days hungry; other days not so hungry and other days barely hungry at all.

Now consider between-variation. I originally stated that the chicken data were, as a group, larger than the tuna data. I need to be more precise; I need to replace the vague expression *as a group* by something else. Well, I have already talked about this; I will summarize a data set by calculating its mean (or median, but let’s stick to the mean now for simplicity). Thus, the between-variation is reflected in the mean for chicken being 2.2 treats larger than the mean for tuna.

This leads to a very important realization. Whereas the within variation is definitely real, I cannot be sure that the between variation is real. Indeed, for dramatic purposes I am going to invent a person I call the Skeptic. The Skeptic is the originator of what we will call the **Skeptic’s Argument**, stated below:
The flavor of the treat is irrelevant. The number of treats that Bob consumed on any given day was determined by how hungry Bob was on that day.

The Skeptic’s Argument is so central to what we do in this course that it behooves us to spend some time thinking about it.

Consider the day when Bob was offered chicken and he consumed eight treats. According to the Skeptic, the flavor was irrelevant. Bob consumed eight treats because he was very hungry that day. If Bob had been offered tuna that day, the Skeptic believes that Bob would have consumed eight treats. Similarly, there was a day that Bob was offered tuna and he refused to eat. To the Skeptic, this means that on that day Bob was not hungry at all and, if he had been offered chicken treats, he would have eaten none of them.

Please note the following two items:

1. As mere mortals, it is impossible for us to determine with certainty whether or not the Skeptic’s Argument is correct.

2. You are certainly allowed to have your own opinion as to whether the Skeptic is correct. I am not going to tell you what to believe, just as I would not tell you whether you like chocolate. You need to learn, however, how statisticians evaluate the Skeptic’s Argument.

Regarding the second item above, there are many details to learn about how statisticians work. The method we advocate has both strengths and weaknesses and we will learn about both categories of features.

It will be useful to invent an adversary for the Skeptic; I will call it the Advocate. I want to avoid making this current chapter too long; thus, I will delay our consideration of the Advocate’s Argument until Chapter 5.

Please note the following. For ease of exposition, on occasion I will refer to the Skeptic and Advocate as real people. I want these characters to be gender-free; hence, when I refer to either of them with a pronoun—as I do the Advocate in the previous paragraph—I will use the pronoun it rather than he or she.

In the remainder of this chapter we will combine the Skeptic’s Argument with our notions of probability to obtain the sampling distribution of the test statistic. The importance of this sampling distribution will be presented in Chapter 5.

### 3.3 The Sampling Distribution of the Test Statistic for Cathy’s Study

Recall that Cathy’s running study was introduced on page 40. For convenience, I have presented her data again in Table 3.2. Also, her treatment means are:

\[
\bar{x} = (530 + 521 + 539)/3 = 530 \quad \text{and} \quad \bar{y} = (528 + 520 + 527)/3 = 525.
\]

We see that Cathy’s randomization assigned trials 1, 2 and 5 to the high school and the remaining
Table 3.2: Cathy’s times, in seconds, to run one mile. HS means she ran at the high school and P means she ran through the park.

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

Table 3.3: The 20 possible assignments for Cathy’s CRD.

1,2,3 1,2,4 1,2,5 1,2,6 1,3,4 1,3,5 1,3,6 1,4,5 1,4,6 1,5,6
2,3,4 2,3,5 2,3,6 2,4,5 2,4,6 2,5,6 3,4,5 3,4,6 3,5,6 4,5,6

trials to the park. In order to save space below, I will refer to this as assignment 1,2,5.

The number of possible assignments for Cathy’s study is quite small; using Equation 3.2 on page 56 we obtain:

\[
\frac{6!}{3!3!} = \frac{6(5)(4)}{3(2)} = 20
\]

Thus, it is relatively easy to list all possible assignments; they are presented in Table 3.3.

As I show above, Cathy’s mean time on the high school route is 5 seconds larger than her mean time on the park route, because, \( \bar{x} - \bar{y} = 530 - 525 = 5 \) seconds. It is helpful now, and necessary in Chapter 5, for me to introduce additional notation and terminology.

Because I am comparing treatments by subtracting means, I will give the difference a symbol; let \( u = \bar{x} - \bar{y} \). For Cathy’s data, \( u = 5 \). We call \( u \) the **observed value** of the test statistic \( U \). Admittedly, this language and notation is confusing. But it is standard and I am unable to change it because, frankly, I am not the tsar of the Statistics world!

I suggest that you think of these ideas in the following way. Before anyone collects data, the test statistic, \( U \), is a **rule** or **plan** or **protocol** for what will be done with the data. In this chapter (and two more to follow) the rule is: we will collect data, separate data by treatments, compute means and compare means by subtracting. When the rule is applied, the result of all that work will be a number, which we denote by \( u \).

The reason we need the above is because of the interesting way (peculiar way?) statisticians analyze data. Above we see that Cathy’s study led to \( u = 5 \) as the observed value of the test statistic. The question we face is: How do we interpret \( u = 5 \)? Should we be impressed? Should we be unimpressed? Is it real? (Whatever that means!)

This leads us to one of the big ideas in Statistics:

We evaluate what **actually happens** in a study by comparing it to **everything that could have happened** in the study.

So, the first thing I do now is I start adding the adjective **actual** in places that were previously modifier free. In particular, Cathy’s actual assignment was 1,2,5 which led to her actual \( x \)’s of 530,
521 and 539 and her actual y’s of 528, 520 and 527, and her actual means of 530 and 525 and, most importantly, her actual u of 5.

Next, let’s look at the statement, “Everything that could have happened.” Do not take an expansive view of this statement! It could have happened that Cathy chose a different topic; or that Cathy did not take my class; or that I never became a statistician; or that the dinosaurs were never wiped out by an asteroid.

Statisticians take an extremely narrow view of the statement, “Everything that could have happened.” To us, what happened is that Cathy obtained assignment 1,2,5; everything that could have happened refers to the 19 other possible assignments.

Therefore, the fundamental quest of a statistician looking at Cathy’s data—and indeed most studies we ever consider—is to determine what u would have been obtained for each of the other (19, in Cathy’s case) possible assignments. Nineteen is a pretty small number; thus, examining 19 assignments seems manageable. Let’s give it a try!

Let’s begin with assignment 1,2,3 which means, of course, that Cathy would run at the high school on trials 1, 2 and 3, and through the park on trials 4, 5 and 6. What would have happened?

Well, I can give you a partial answer quite easily. In fact, I can list several things that I know to be true:

- Trial 1: In the actual study, trial 1 was assigned to the high school and Cathy obtained a response of 530. Thus, because assignment 1,2,3 assigns trial 1 to the high school, the response would have been 530.
- Trial 2: For the same reason I give above for trial 1, the response would have been 521.
- Trial 4: In the actual study, trial 4 was assigned to the park and Cathy obtained a response of 520. Thus, because assignment 1,2,3 assigns trial 4 to the park, the response would have been 520.
- Trial 6: For the same reason I give above for trial 4, the response would have been 527.
- Trial 3: In the actual study, trial 3 was assigned to the park and Cathy obtained a response of 528. Assignment 1,2,3 assigns trial 3 to the high school. As a result, nobody can say what the response would have been!
- Trial 5: As with trial 3, nobody can say what the response would have been.

It seems that we are at an impasse. I cannot say, with certainty, what would have happened if Cathy had used assignment 1,2,3. In fact, a variation of the above argument can be made for every one of the 19 assignments that were not used in the actual study.

So, what do we do? We add an assumption. We add the assumption that the Skeptic’s Argument is correct. Now, let’s revisit my argument above for trial 3.

- Trial 3: In the actual study, trial 3 was assigned to the park and Cathy obtained a response of 528. Assignment 1,2,3 assigns trial 3 to the high school.
Table 3.4: The values of $u$ for all possible assignments for Cathy’s CRD.

<table>
<thead>
<tr>
<th>Assignment</th>
<th>$x$ values</th>
<th>$y$ values</th>
<th>$u$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,2,3</td>
<td>530,521,528</td>
<td>520,539,527</td>
<td>-2.33</td>
</tr>
<tr>
<td>1,2,4</td>
<td>530,521,520</td>
<td>528,539,527</td>
<td>-7.67</td>
</tr>
<tr>
<td>1,2,5</td>
<td>530,521,539</td>
<td>528,520,527</td>
<td>+5.00 (Actual)</td>
</tr>
<tr>
<td>1,2,6</td>
<td>530,521,527</td>
<td>528,520,539</td>
<td>-3.00</td>
</tr>
<tr>
<td>1,3,4</td>
<td>530,528,520</td>
<td>521,539,527</td>
<td>-3.00</td>
</tr>
<tr>
<td>1,3,5</td>
<td>530,528,539</td>
<td>521,520,527</td>
<td>+9.67</td>
</tr>
<tr>
<td>1,3,6</td>
<td>530,528,527</td>
<td>521,520,539</td>
<td>+1.67</td>
</tr>
<tr>
<td>1,4,5</td>
<td>530,520,539</td>
<td>521,528,527</td>
<td>+4.33</td>
</tr>
<tr>
<td>1,4,6</td>
<td>530,520,527</td>
<td>521,528,539</td>
<td>-3.67</td>
</tr>
<tr>
<td>1,5,6</td>
<td>530,539,527</td>
<td>521,528,520</td>
<td>+9.00</td>
</tr>
<tr>
<td>2,3,4</td>
<td>521,528,520</td>
<td>530,539,527</td>
<td>-9.00</td>
</tr>
<tr>
<td>2,3,5</td>
<td>521,528,539</td>
<td>530,520,527</td>
<td>+3.67</td>
</tr>
<tr>
<td>2,3,6</td>
<td>521,528,527</td>
<td>530,520,539</td>
<td>-4.33</td>
</tr>
<tr>
<td>2,4,5</td>
<td>521,520,539</td>
<td>530,528,527</td>
<td>-1.67</td>
</tr>
<tr>
<td>2,4,6</td>
<td>521,520,527</td>
<td>530,528,539</td>
<td>-9.67</td>
</tr>
<tr>
<td>2,5,6</td>
<td>521,539,527</td>
<td>530,528,520</td>
<td>+3.00</td>
</tr>
<tr>
<td>3,4,5</td>
<td>528,520,539</td>
<td>530,521,527</td>
<td>+3.00</td>
</tr>
<tr>
<td>3,4,6</td>
<td>528,520,527</td>
<td>530,521,539</td>
<td>-5.00</td>
</tr>
<tr>
<td>3,5,6</td>
<td>528,539,527</td>
<td>530,521,520</td>
<td>+7.67</td>
</tr>
<tr>
<td>4,5,6</td>
<td>520,539,527</td>
<td>530,521,528</td>
<td>+2.33</td>
</tr>
</tbody>
</table>

Because the Skeptic is correct, the treatment does not matter. Cathy obtained a response of 528 on trial 3 because that reflected her energy, enthusiasm, the weather, whatever, on trial 3. If she had run at the high school, her response would have been 528.

With a similar argument, we see that by adding the assumption that the Skeptic is correct, we know that Cathy’s time on trial 5 would have been 539 whether she ran at the high school (as she actually did) or through the park, as assignment 1,2,3 would have told her to do.

To summarize: for assignment 1,2,3, Cathy’s $x$’s would have been: 530, 521 and 528; her $y$’s would have been 520, 539 and 527. (You can find these easily by looking at Table 3.2 and simply ignoring the listed treatments; after all, according to the Skeptic, the treatments don’t matter.)

Continuing, for assignment 1,2,3, Cathy’s $\bar{x} = 526.33$, $\bar{y} = 528.66$ and $u = -2.33$. I can repeat this analysis for the remaining 18 possible assignments. My work is summarized in Table 3.4. You don’t need to verify all of the entries in this table, but you should verify enough to convince yourself that you understand the method.

There is no nice way to say this: The results in Table 3.4 are a mess! The 20 possible as-
Table 3.5: The sampling distribution of $U$ for Cathy’s CRD.

<table>
<thead>
<tr>
<th>$u$</th>
<th>$P(U = u)$</th>
<th>$u$</th>
<th>$P(U = u)$</th>
<th>$u$</th>
<th>$P(U = u)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$-9.67$</td>
<td>0.05</td>
<td>$-3.00$</td>
<td>0.10</td>
<td>$3.67$</td>
<td>0.05</td>
</tr>
<tr>
<td>$-9.00$</td>
<td>0.05</td>
<td>$-2.33$</td>
<td>0.05</td>
<td>$4.33$</td>
<td>0.05</td>
</tr>
<tr>
<td>$-7.67$</td>
<td>0.05</td>
<td>$-1.67$</td>
<td>0.05</td>
<td>$5.00$</td>
<td>0.05</td>
</tr>
<tr>
<td>$-5.00$</td>
<td>0.05</td>
<td>$1.67$</td>
<td>0.05</td>
<td>$7.67$</td>
<td>0.05</td>
</tr>
<tr>
<td>$-4.33$</td>
<td>0.05</td>
<td>$2.33$</td>
<td>0.05</td>
<td>$9.00$</td>
<td>0.05</td>
</tr>
<tr>
<td>$-3.67$</td>
<td>0.05</td>
<td>$3.00$</td>
<td>0.10</td>
<td>$9.67$</td>
<td>0.05</td>
</tr>
</tbody>
</table>

Table 3.6: Kymn’s times, in seconds, to row 2000 meters on an ergometer. Treatment 1 is the small gear with the vent closed; and treatment 2 is the large gear with the vent open.

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

Assignments lead to 18 different values of $u$! Before I give in to despair, I will summarize this information in Table 3.5. In this new table, I have taken the 18 different values of $u$ and sorted them from smallest to largest. I then divided their frequencies of occurrence (16 of which are one and two of which are two) by the total number of assignments, 20, to obtain the probabilities given in the table. This table is a representation of the sampling distribution of the test statistic $U$. Sometimes, the sampling distribution is represented by an equation; in any event it contains all possible values of $U$, each matched with its probability of occurring. As we will see in Chapter 5, the sampling distribution is a critical component of how we learn from a CRD. Before our next example, the last one of this chapter, I will make an obvious comment: Even for a CRD with a very small number of possible assignments—20 in this case—it is extremely tedious and messy to determine the sampling distribution of the test statistic $U$.

### 3.4 The Sampling Distribution of $U$ for Kymn’s CRD

Kymn’s study and data were presented in Chapter 2. Her data are reproduced in Table 3.6.

Kymn’s study will be manageable for me because, as shown earlier, there are only 252 possible assignments of trials to treatments. Note from Table 3.6 that her actual assignment was 2, 3, 4, 7 and 10, which yielded

$$\bar{x} = 491.4, \bar{y} = 484.2 \text{ and } u = 491.4 - 484.2 = 7.2.$$ 

I used the randomizer to obtain a second possible assignment: I obtained 4, 5, 7, 9 and 10.
Table 3.7: Frequency table for \( u \) for the 252 possible assignments for Kymn’s study.

<table>
<thead>
<tr>
<th>( u )</th>
<th>Freq.</th>
<th>( u )</th>
<th>Freq.</th>
<th>( u )</th>
<th>Freq.</th>
<th>( u )</th>
<th>Freq.</th>
<th>( u )</th>
<th>Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>−7.2</td>
<td>1</td>
<td>−4.8</td>
<td>3</td>
<td>−2.4</td>
<td>10</td>
<td>0.4</td>
<td>12</td>
<td>2.8</td>
<td>10</td>
</tr>
<tr>
<td>−6.8</td>
<td>1</td>
<td>−4.4</td>
<td>5</td>
<td>−2.0</td>
<td>8</td>
<td>0.8</td>
<td>10</td>
<td>3.2</td>
<td>8</td>
</tr>
<tr>
<td>−6.4</td>
<td>1</td>
<td>−4.0</td>
<td>8</td>
<td>−1.6</td>
<td>14</td>
<td>1.2</td>
<td>13</td>
<td>3.6</td>
<td>6</td>
</tr>
<tr>
<td>−6.0</td>
<td>1</td>
<td>−3.6</td>
<td>6</td>
<td>−1.2</td>
<td>13</td>
<td>1.6</td>
<td>14</td>
<td>4.0</td>
<td>8</td>
</tr>
<tr>
<td>−5.6</td>
<td>3</td>
<td>−3.2</td>
<td>8</td>
<td>−0.8</td>
<td>10</td>
<td>2.0</td>
<td>8</td>
<td>4.4</td>
<td>5</td>
</tr>
<tr>
<td>−5.2</td>
<td>4</td>
<td>−2.8</td>
<td>10</td>
<td>−0.4</td>
<td>12</td>
<td>2.4</td>
<td>10</td>
<td>4.8</td>
<td>3</td>
</tr>
<tr>
<td>0.0</td>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Total 252**

Referring to Table 3.6, you can verify the following:

\[ \bar{x} = \frac{492 + 483 + 490 + 486 + 493}{5} = 488.8, \quad \bar{y} = \frac{485 + 493 + 489 + 488 + 479}{5} = 486.8 \]

and \( u = 488.8 - 486.8 = 2.0 \). I continued the above process, examining all 252 possible assignments and determining the value of \( u = \bar{x} - \bar{y} \) for each of them. (Warning: Do not attempt this! It was no fun and quite tedious; although, in fairness, I should acknowledge that many teachers are actually quite good at doing this. I am not.) The 252 assignments result in 37 possible values of \( u \); my results are presented in Table 3.7. If we divide these frequencies by 252 we have the sampling distribution for \( U \). (Admittedly, we would need to change the column headings.) But I do not enjoy dividing by 252 and such divisions do not give pleasant decimals, so I won’t do it. You get the idea without this extra pain. We will return to Kymn’s sampling distribution in Chapter 5.

I will note that Kymn’s actual \( u = 7.2 \) was the largest possible value of \( u \); that is, every one of the other 251 possible assignments gives a value for \( u \) that is smaller than 7.2. This is no great surprise; Kymn’s actual data had the five largest values on treatment 1 and the five smallest values on treatment 2. Thus, not a surprise, but as we will see, very important.

In Chapter 5 we will learn why the sampling distribution of a test statistic is important. But first we must deal with the following issue. When there are trillions or billions or even thousands of possible assignments, the method I used for Cathy’s and Dawn’s studies—basically, a tedious enumeration of every possibility—simply is not practical, even with the help of a clever computer program. Therefore, in Chapter 4 we will learn how to approximate a sampling distribution.

I end this section with the following useful result on symmetry.

**Result 3.2 (Balance and symmetry.)** For a balanced CRD, the sampling distribution of \( U \) is symmetric around 0.

**Proof:** You don’t need to read this proof, but it is so elegant that math-ophiles might enjoy it.

Let \( a_1 \) denote any assignment. Let \( -a_1 \) denote its mirror image; i.e., the units assigned to treatment 1 [2] by \( a_1 \) are assigned to treatment 2 [1] by \( -a_1 \). Thus, for example, if \( n = 6 \), the mirror image of assignment 1,2,3 is assignment 4,5,6.

Let \( b \) be any positive number. Consider the collection, \( A \), of assignments, if any, that give \( u = b \). Let \( B \) denote the collection of assignments that are mirror images of the assignments in \( A \).
Clearly, every assignment in B gives \( u = -b \). As a result, \( P(U = b) = P(U = -b) \) for every \( b > 0 \). Hence, the result follows.

The symmetry promised by Result 3.2 can be seen in Tables 3.5 and 3.7.

### 3.5 Computing

In this section you will learn how to use a website to perform randomization. Please click onto the following site:

[http://www.randomizer.org/form.htm](http://www.randomizer.org/form.htm)

If all is working well, you will still be able to see this page plus another window that has opened up and is at the above site. Let’s look at your new web-window.

You will see that there are eight rectangular boxes. When you use this site to obtain an assignment via randomization you will be able to enter your values of interest in the top seven of these boxes. Currently, these seven boxes contain the default values provided by the site’s programmer. After you have changed any, all or none of the entries in these seven boxes you should click on the bottom box, which contains the words:

Randomize Now!

My guess is that the exclamation point is to remind you how exciting it is to be doing Statistics! Personally, I am so excited to see how this works, I will click on the bottom box without any changes to the default settings. I did so and obtained:

11, 49, 18, 27, 50

You try it. I bet that you did not obtain the same numbers I obtained above.

Let’s now look at the seven boxes and discover why they exist. The first box sits to the right of the following question:

**How many sets of numbers do you want to generate?**

As we shall see below, the “sets of numbers” are assignments. Thus, the question is: How many assignments do you want to generate? If I am a researcher seeking an assignment for my study, I would want only one assignment and I would not change the default value, which is 1. If a Homework question asks you to generate, say, five assignments, then you would place the number 5 in the box. Well, you could leave the 1 in the box and run the site five times, but that would be tedious and waste time.

The second box sits to the to the right of the question:

**How many numbers per set?**

For our purpose—obtaining an assignment for CRD—enter the value \( n_1 \) in the box.

The third and fourth boxes sit to the right of a directive to specify the range of numbers you desire. Calculate the value of \( n = n_1 + n_2 \), the total number of units in the CRD, and enter:
From: 1 (i.e., you may leave the default value)
To: \( n \)

The fifth box sits to the right of the question:

Do you wish each number in a set to remain unique?

For randomization we need \( n_1 \) different units selected for assignment to treatment 1. Thus, we opt for the default answer: Yes.

The sixth box sits to the right of the question:

Do you wish to sort the numbers that are generated?

For randomization we don’t care the order in which the \( n_1 \) units are selected. Thus—and because it is easier to work with a sorted list—you should change the default to

Yes: Least to Greatest.

Finally, the seventh box sits to the right of the question:

How do you wish to view your random numbers?

I find the proffered place markers to be very annoying. Thus, I recommend using the default: Place Markers Off.

Problem 1 in the Practice Problems will provide you with practice using this site.
3.6 Summary

In a CRD, the researcher has the authority to assign units to levels of the study factor and exercises this authority by assigning units using the process of randomization. The three equivalent methods of randomizing are listed on page 57. The first two of these involve using a collection of cards. If you ever need to randomize (for example, in your own research or for a project in this class) it is fine to use cards, but most people find it more convenient to use an electronic method. In other words, the cards give us a way to visualize—if not perform—randomization. The electronic method was detailed in Section 3.5.

For a CRD with \( n = n_1 + n_2 \) units, of which \( n_1 \) are assigned to treatment 1, the number of possible assignments equals

\[
\frac{n!}{n_1!n_2!},
\]

where factorials are defined in Equation 3.1 on page 56.

Probability is a quantification of uncertainty about the future. At this stage in these notes, we are interested in applying the idea of probability to the process of randomization. In particular, before we randomize, we believe that every possible assignment is equally likely to be the assignment chosen by randomization. Thus, we assign the same probability to every possible assignment:

\[
P(\text{Any particular assignment}) = \frac{1}{\text{Total number of possible assignments}}.
\]

For a CRD with \( n = n_1 + n_2 \) units, of which \( n_1 \) are assigned to treatment 1, the number of possible assignments equals

\[
\frac{n!}{n_1!n_2!},
\]

where factorials are defined in Equation 3.1 on page 56.

Probability is a quantification of uncertainty about the future. At this stage in these notes, we are interested in applying the idea of probability to the process of randomization. In particular, before we randomize, we believe that every possible assignment is equally likely to be the assignment chosen by randomization. Thus, we assign the same probability to every possible assignment:

\[
P(\text{Any particular assignment}) = \frac{1}{\text{Total number of possible assignments}}.
\]

An event is any specified collection of assignments. For any event \( A \) we calculate its probability as:

\[
P(A) = \frac{\text{Number of assignments in } A}{\text{Total number of possible assignments}}.
\]

Typically—but not always—in a CRD the two treatment means calculated from the data will be different numbers. Statisticians want to decide whether this observed difference between treatments is real. This is not an easy question to formulate or answer. The first step is to consider the Skeptic’s Argument which states that the two treatments are the same in the following sense:

For any given unit, if it were possible to assign the unit to both treatments, the two responses obtained would be identical.

In a CRD, because it is impossible to assign a unit to both treatments, a researcher cannot say, with certainty, whether the Skeptic’s Argument is correct. In Chapter 5 we will learn how to examine the validity of the Skeptic’s Argument using the ideas of a statistical test of hypotheses. Learning all the facets of a statistical test of hypotheses is a lengthy and difficult endeavor; as a result, we will break the process into manageable sized pieces. The first piece is the specification of the test statistic and the derivation of its sampling distribution.

We summarize the data from a CRD by computing the mean response for each treatment: \( \bar{x} [\bar{y}] \) for treatment 1 [2]. Define \( u = \bar{x} - \bar{y} \); \( u \) is the difference of the treatment means. The number \( u \) is called the observed value of the test statistic \( U \).

One of the big ideas in Statistics is:
We evaluate what actually happens in a study by comparing it to everything that could have happened in the study.

Statisticians take a rather modest view of the notion of everything that could have happened; this means the following. On the one hand we have the evidence from the actual data: the actual value of $u$. On the other hand, we have the distribution of the values of $u$. The distribution is obtained by looking at every possible assignment—including the assignment that was actually used—and for each assignment determining the value of $u$. As illustrated in the text, such a distribution cannot be obtained without an additional assumption. The assumption we will explore in the next few chapters is the assumption that the Skeptic is correct. With this assumption, the distribution is called the sampling distribution of the test statistic $U$.

In this chapter we have seen via two examples that if the number of possible assignments is small—an admittedly vague term—then the exact sampling distribution can be determined. In the next chapter, we will learn how a computer can be used to obtain an approximate sampling distribution. These ideas will come together in Chapter 5 when we learn how scientists and statisticians use either the exact or an approximate sampling distribution to investigate the issue of whether the observed treatment difference is real.

Much later in this Part I we will see that it is possible to obtain a distribution of values of $u$ when the Skeptic is incorrect. This will allow us to investigate what is called the power of a test of hypotheses and we will learn why the power is very important to a scientist.

Result 3.2 on page 65 states that for a balanced CRD, the sampling distribution of $U$ is symmetric around 0.
3.7 Practice Problems

1. Use the website discussed in Section 3.5 to obtain five assignments for Kynn’s rowing study using the process of randomization. Recall that Kynn performed a balanced CRD with a total of 10 trials.

2. Table 3.8 presents artificial data for a balanced CRD with a total of $n = 6$ units.
   (a) What is the actual assignment that the researcher used?
   (b) Calculate the actual values of $\bar{x}$, $\bar{y}$ and $u = \bar{x} - \bar{y}$.

3. Refer to the previous question. It is too tedious for you to determine the entire sampling distribution for $U$, but I want to make sure you understand the steps. Thus, please answer the questions below.
   (a) Assuming that the Skeptic is correct, calculate the values of $\bar{x}$, $\bar{y}$ and $u = \bar{x} - \bar{y}$ for assignment 1,3,6.
   (b) Assuming that the Skeptic is correct, calculate the values of $\bar{x}$, $\bar{y}$ and $u = \bar{x} - \bar{y}$ for assignment 1,4,5.
   (c) Assuming that the Skeptic is correct, calculate the largest possible value of $u$; which assignment gives this value?

4. Table 3.9 presents an artificial data set with $n_1 = 2$ and $n_2 = 4$.
   (a) How many possible assignments are there?
   (b) List all possible assignments. Identify the actual assignment.
   (c) Calculate the actual values of $\bar{x}$, $\bar{y}$ and $u$.
   (d) Assuming that the Skeptic is correct, calculate the values of $\bar{x}$, $\bar{y}$ and $u = \bar{x} - \bar{y}$ for assignment 5,6.
   (e) Assuming that the Skeptic is correct, calculate the values of $\bar{x}$, $\bar{y}$ and $u = \bar{x} - \bar{y}$ for assignment 1,4.
   (f) The sampling distribution of $U$ is given in Table 3.10
      i. Is the sampling distribution of $U$ symmetric? Explain your answer.
      ii. Obtain $P(U \geq 6)$; obtain $P(U \leq -6)$; obtain $P(|U| \geq 6)$.
Table 3.8: Artificial data for Practice Problems 2 and 3.

<table>
<thead>
<tr>
<th>Unit</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Response</td>
<td>9</td>
<td>6</td>
<td>30</td>
<td>3</td>
<td>12</td>
<td>15</td>
</tr>
</tbody>
</table>

Table 3.9: Artificial data for Practice Problem 4.

<table>
<thead>
<tr>
<th>Unit</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Response</td>
<td>8</td>
<td>4</td>
<td>4</td>
<td>0</td>
<td>20</td>
<td>12</td>
</tr>
</tbody>
</table>

3.8 Solutions to Practice Problems

1. My answer is below. You will no doubt obtain an answer different from mine.

   Below are my responses to the information requested by the website, some with a brief parenthetical explanation.

   • 5 (the number of assignments I requested); 5 (the value of $n_1$); From 1 To 10 (10 is the total number of trials in the study); Yes; Yes, Least to Greatest; and Place Markers Off.

   Below are the assignments I obtained:

   1, 3, 4, 8, 9
   3, 4, 5, 8, 9
   1, 4, 7, 9, 10
   1, 2, 3, 8, 9
   2, 3, 7, 8, 9

2. (a) From Table 3.8 we can see that units 2, 3 and 5 were assigned to treatment 1. Thus, the actual assignment is 2,3,5.

   (b) Again from Table 3.8,

   $$\bar{x} = (6 + 30 + 12)/3 = 16$$ and $$\bar{y} = (9 + 3 + 15)/3 = 9.$$
Thus, \( u = 16 - 9 = 7 \).

3. (a) From Table 3.8

\[
\bar{x} = \frac{(9 + 30 + 15)}{3} = 18 \quad \text{and} \quad \bar{y} = \frac{(6 + 3 + 12)}{3} = 7.
\]

Thus, \( u = 18 - 7 = 11 \).

(b) Again from Table 3.8

\[
\bar{x} = \frac{(9 + 3 + 12)}{3} = 8 \quad \text{and} \quad \bar{y} = \frac{(6 + 30 + 15)}{3} = 17.
\]

Thus, \( u = 8 - 17 = -9 \).

(c) The six response values sum to 75. Thus, for any assignment

\[
\bar{x} + \bar{y} = 75/3 = 25,
\]

as illustrated above repeatedly. Thus,

\[
u = \bar{x} - \bar{y} = \bar{x} - (25 - \bar{x}) = 2\bar{x} - 25.
\]

We get the largest possible value of \( u \) by finding the largest possible value of \( \bar{x} \). The largest possible value of \( \bar{x} \) is \( (30 + 12 + 15)/3 = 19 \) which gives \( \bar{y} = 25 - 19 = 6 \). Thus, \( 19 - 6 = 13 \) is the largest possible value of \( u \). From Table 3.8 we see that this largest value of \( u \) is achieved by assignment 3,5,6.

4. (a) Using Equation 3.2 we obtain:

\[
6!/[2!4!] = [6(5)]/2 = 15.
\]

(b) As usual, I identify an assignment by listing the units that are assigned to the first treatment. I obtain:

1,2; 1,3; 1,4; 1,5; 1,6; 2,3; 2,4; 2,5; 2,6; 3,4; 3,5; 3,6; 4,5; 4,6; and 5,6.

From Table 3.9 the actual assignment is 3,5.

(c) From Table 3.9

\[
\bar{x} = \frac{(4 + 20)}{2} = 12, \quad \bar{y} = \frac{(8 + 4 + 0 + 12)}{4} = 6 \quad \text{and} \quad u = 12 - 6 = 6.
\]

(d) From Table 3.9

\[
\bar{x} = \frac{(20 + 12)}{2} = 16, \quad \bar{y} = \frac{(8 + 4 + 4 + 0)}{4} = 4 \quad \text{and} \quad u = 16 - 4 = 12.
\]

(e) From Table 3.9

\[
\bar{x} = \frac{(8 + 0)}{2} = 4, \quad \bar{y} = \frac{(4 + 4 + 20 + 12)}{4} = 10 \quad \text{and} \quad u = 4 - 10 = -6.
\]
(f)  i. No. There are many ways to justify this answer; here is mine. For symmetry (around any value, not just 0) the largest possible value of $U$, 12 in this case, must have the same probability as the smallest possible value of $U$, $-9$ in this case. They don’t.

ii. The computations are below.

\[
P(U \geq 6) = P(U = 6) + P(U = 9) + P(U = 12) = \frac{2}{15} + \frac{1}{15} + \frac{1}{15} = \frac{4}{15}.
\]

\[
P(U \leq -6) = P(U = -6) + P(U = -9) = \frac{2}{15} + \frac{2}{15} = \frac{4}{15}.
\]

\[
P(|U| \geq 6) = P(U \geq 6) + P(U \leq -6) = \frac{4}{15} + \frac{4}{15} = \frac{8}{15}.
\]
3.9 Homework Problems

1. Use randomization to obtain one assignment for each of the following situations.
   (a) $n_1 = n_2 = 12$.
   (b) $n_1 = 8; n_2 = 15$.
   (c) $n_1 = 15; n_2 = 8$.

2. Table 3.11 presents artificial data from a CRD with $n_1 = 2$ and $n_2 = 3$. Use these data to answer the questions below.
   (a) How many possible assignments are there?
   (b) List all possible assignments. Identify the actual assignment.
   (c) Calculate the actual values of $\bar{x}, \bar{y}$ and $u$.
   (d) Assuming that the Skeptic is correct, calculate the values of $\bar{x}, \bar{y}$ and $u$ for assignment 1,4.
   (e) Assuming that the Skeptic is correct, calculate the values of $\bar{x}, \bar{y}$ and $u$ for assignment 3,5.

Table 3.11: Artificial data for Homework Problem 2.

<table>
<thead>
<tr>
<th>Unit</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Response</td>
<td>12</td>
<td>6</td>
<td>24</td>
<td>0</td>
<td>12</td>
</tr>
</tbody>
</table>