Chapter 8

Dichotomous Responses; Critical Regions

8.1 Introduction and Notation

In all previous studies in these notes, the response has been either a numerical variable or an ordered categorical variable with at least three categories. For a numerical response we compared the treatments by comparing the means or the ranks of their responses. For an ordered categorical response we compared the treatments by comparing the ranks of their responses.

In this chapter we consider studies that have a dichotomous response—a categorical response with two categories. We begin with four examples.

Example 8.1 (Therese’s infidelity study.) Therese studied 20 of her adult female friends. The women were divided into two treatment groups, both of size 10, by randomization. Women assigned to the first treatment group read the following question:

- You are friends with a married couple and are equally fond of the man and the woman. You discover that the husband is having an affair. The wife suspects that something is going on and asks you if you know anything about her husband having an affair. Do you tell?

Women assigned to the second treatment group read the following question:

- You are friends with a married couple and are equally fond of the man and the woman. You discover that the wife is having an affair. The husband suspects that something is going on and asks you if you know anything about his wife having an affair. Do you tell?

Each subject was instructed to respond either yes or no.

Example 8.2 (Ruth’s prisoner study.) Ruth’s subjects were 50 male inmates at a minimum security federal prison camp in Wisconsin. All of the men were first-time nonviolent criminal offenders serving two or more years of prison time. The men were divided into two treatment groups of 25 each by randomization. Men assigned to the first treatment group were given the following question:
• The prison is beginning a program in which inmates have the opportunity to volunteer for community service with developmentally disabled adults. *Inmates who volunteer will receive a sentence reduction.* Would you participate?

Men assigned to the second treatment group were given the following question:

• The prison is beginning a program in which inmates have the opportunity to volunteer for community service with developmentally disabled adults. Would you participate?

Each subject responded either yes or no.

**Example 8.3 (Thomas’s golf putting study.)** Thomas wanted to investigate the difference in difficulty between four and eight foot putts in golf. He performed a balanced study with randomization and a total of 50 putts. The first treatment was putting from four feet on a level surface and the second treatment was putting from eight feet on a level surface. Each putt was either made or missed.

**Example 8.4 (The artificial Headache Study-2 (HS-2).)** A researcher has 100 persons available for study. Each person routinely suffers mild tension headaches (not migraines). The researcher wants to compare two active drugs, call them A and B, for the treatment of mild headaches. The 100 subjects are divided into two groups of size 50 each by randomization. Each subject is given the following instructions:

> The next time you have a mild headache take the drug we have given you. Fifteen minutes later answer the following question with a response of either yes or no: Has your headache pain diminished?

When the response is a dichotomy, there are technical names for the two possible responses: one is called a **success** and the other is called a **failure**. The methods we learn will focus on **counting successes**. We use the following method for deciding which possible outcome gets the distinction of being called a success.

1. If one of the possible responses is very rare (admittedly vague), then it is labeled the success.
2. If neither possible response is very rare, then the more desirable response is labeled the success.
3. If neither of the previous two scenarios applies, then the researcher arbitrarily assigns the label success to one of the possible responses.

Here is the idea behind the first rule. Every time I drive a car I have the potential to be involved in a traffic accident. Fortunately, in my 47 years of driving I have been in only one accident and it was very minor. When something occurs only rarely, it is much easier to keep track of how many times it happens rather than how many times it fails to happen.

In our examples above, the researchers labeled as successes: telling, agreeing to volunteer, making a putt and reporting that the pain has diminished. Tables 8.1–8.4 present and summarize the data for each of our four studies.

168
Table 8.1: The $2 \times 2$ contingency table of observed counts for Therese’s infidelity study.

<table>
<thead>
<tr>
<th>Cheater was:</th>
<th>Tell?</th>
<th>Row Proportions</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
<td>Total</td>
</tr>
<tr>
<td>The Husband</td>
<td>7</td>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td>The Wife</td>
<td>4</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>Total</td>
<td>11</td>
<td>9</td>
<td>20</td>
</tr>
</tbody>
</table>

Table 8.2: The $2 \times 2$ contingency table of observed counts for Ruth’s prisoner study.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
<td>Total</td>
</tr>
<tr>
<td>Sentence Reduction</td>
<td>18</td>
<td>7</td>
<td>25</td>
</tr>
<tr>
<td>No Sentence Reduction</td>
<td>23</td>
<td>2</td>
<td>25</td>
</tr>
<tr>
<td>Total</td>
<td>41</td>
<td>9</td>
<td>50</td>
</tr>
</tbody>
</table>

Table 8.3: $2 \times 2$ Contingency table of observed counts for Thomas’s golf putting study.

<table>
<thead>
<tr>
<th>Distance:</th>
<th>Putt was</th>
<th>Row Prop.</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Made</td>
<td>Missed</td>
<td>Total</td>
</tr>
<tr>
<td>Four feet</td>
<td>18</td>
<td>7</td>
<td>25</td>
</tr>
<tr>
<td>Eight feet</td>
<td>10</td>
<td>15</td>
<td>25</td>
</tr>
<tr>
<td>Total</td>
<td>28</td>
<td>22</td>
<td>50</td>
</tr>
</tbody>
</table>

Table 8.4: $2 \times 2$ contingency table of observed counts for the artificial Headache Study-2 (HS-2).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
<td>Total</td>
</tr>
<tr>
<td>A</td>
<td>29</td>
<td>21</td>
<td>50</td>
</tr>
<tr>
<td>B</td>
<td>21</td>
<td>29</td>
<td>50</td>
</tr>
<tr>
<td>Total</td>
<td>50</td>
<td>50</td>
<td>100</td>
</tr>
</tbody>
</table>
Table 8.5: General notation for a $2 \times 2$ contingency table of observed counts for a CRD with a dichotomous response.

<table>
<thead>
<tr>
<th>Treatment :</th>
<th>Response</th>
<th>Row Proportions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$S$</td>
<td>$F$</td>
</tr>
<tr>
<td>1</td>
<td>$a$</td>
<td>$b$</td>
</tr>
<tr>
<td>2</td>
<td>$c$</td>
<td>$d$</td>
</tr>
<tr>
<td>Total</td>
<td>$m_1$</td>
<td>$m_2$</td>
</tr>
</tbody>
</table>

Table [8.5] presents our general notation for a CRD with a dichotomous response. When I develop ideas below it will be convenient to use the general notation. First, a few comments:

1. The orientation for these tables in these notes will follow the four examples above; namely, the rows distinguish between treatments and the columns identify the possible responses. Many materials (texts, research papers, etc.) reverse this orientation. Thus, when reading other materials, be careful to identify which orientation is being used.

2. We summarize the table of counts by computing the **row proportions**: the $\hat{p}$’s and $\hat{q}$’s given above. There is a great deal of redundancy in these; namely, in each row the sum of its $\hat{p}$ and $\hat{q}$ is always one. Thus, after you get more familiar with these ideas I usually will suppress the $\hat{q}$’s.

3. In these tables, I do **not** calculate the row proportions for the **Total** row because in a CRD these numbers typically are not of interest.

There is a very simple, but useful, connection between the treatment (row) proportions of successes and the means ($\bar{x}$ and $\bar{y}$) of our earlier work for a CRD with a numerical response. I will illustrate the connection with Therese’s data; the interested reader can show easily that the connection is also true for the general case.

Therese assigned 10 friends to her first treatment, the husband having an affair; seven responded *yes* (success) and three responded *no* (failure) giving $\hat{p}_1 = 7/10 = 0.70$. Alternatively, we can make Therese’s response a number: 1 for *yes* and 0 for *no*. With this latter viewpoint, Therese’s data consist of seven 1’s and three 0’s. Clearly, the sum of her 10 numbers is 7—which is the total number of successes. The mean of these 10 numbers, which we call $\bar{x}$, is $7/10 = 0.70$. In other words,

$$\hat{p}_1 = \bar{x} \text{ and, similarly, } \hat{p}_2 = \bar{y}. $$

This identification is important because:

It shows that the Skeptic’s Argument and all that follows from it—the Advocate, the hypotheses, the test of hypotheses, the rules for computing the P-value and so on—can be immediately adapted to a dichotomous response.
For convenience, we will call the observed value of our test statistic \( x = \hat{p}_1 - \hat{p}_2 \) which implies that we refer to our test statistic as \( X \). We could, of course, call it \( U \), but I fear that would create confusion. Let’s keep \( U \) for comparing means and \( X \) for comparing proportions, even though they coincide. By the way, \( X \) is also mathematically equivalent to \( R_1 \), but you don’t need to understand the details of the argument. You are welcome to ignore this latter connection or rejoice in it.

### 8.2 The Test of Hypotheses: Fisher’s Test

Let me be precise about the test of hypotheses in this chapter. Define \( p_1 \) to be the proportion of successes that would be obtained if all units were assigned to treatment 1. Similarly, \( p_2 \) is the proportion of successes that would be obtained if all units were assigned to treatment 2. If the Skeptic is correct, then \( p_1 = p_2 \) because the treatment does not matter; i.e., some units will yield successes and others will yield failures, regardless of the treatment they receive. For example, if the Skeptic is correct for Therese’s study of infidelity, then some of her friends are tellers, no matter the sex of the person having the affair, and the remaining friends keep quiet, again no matter the sex of the person having the affair.

If we could perform the clone-enhanced study we would know whether the Skeptic is correct. If the Skeptic is incorrect, the clone-enhanced study would reveal this fact as well as the values of \( p_1 \) and \( p_2 \).

For testing, the null hypothesis is that the Skeptic is correct. As with our test based on comparing means, there are three options for the alternative:

- \( H_1 : p_1 > p_2 \);
- \( H_1 : p_1 < p_2 \); and
- \( H_1 : p_1 \neq p_2 \).

As before, you may choose whichever alternative you like, provided you make your choice before collecting data. Also as before, I recommend using the Inconceivable Paradigm to make this choice.

In addition, when the response is a dichotomy there is a big bonus: On the assumption the Skeptic is correct, we don’t need to perform a computer simulation experiment to approximate the P-value; there is a *simple mathematical formula* that allows us to calculate the exact P-value.

When I say a *simple mathematical formula* in the previous sentence, I am being a bit tricky. It is a simple formula to write down, but tedious to compute by hand. Fortunately, there is a website that, with minimal effort from us, will produce the exact P-value for each of the three possible alternative hypotheses. The website is

http://www.langsrud.com/fisher.htm

Note that I am deviating from my usual method of presenting material. In the previous chapters, the computational methods had their own section, named *Computing*. I believe that this chapter
will be easier for you to follow if I incorporate the website now. (As always, let me know if you think I am making a tactical or strategic error.)

The first thing I notice when I look at this website is the label **Fisher’s Exact Test** in large, bold-faced type in its upper left corner. This test is named in honor of Sir Ronald A. Fisher (1890–1962), the most famous statistician of the first half of the twentieth century. Fisher also did well-respected work in genetics. Among Fisher’s many important works, he was a strong advocate of the importance of randomization in scientific studies, which is part of the reason this test bears his name.

Immediately below the name of the test, you will find:

1. A box labeled ‘COMPUTE.’
2. A rectangle containing four small windows. The windows contain the default data: 3, 1, 1 and 3.
3. A box labeled ‘CLEAR TABLE.’
4. A box labeled ‘CLEAR OUTPUT.’

I will now illustrate the use of this website for Ruth’s data (see Table 8.2). Just follow the steps below.

1. Click on the box ‘CLEAR TABLE.’ This action will result in the default data disappearing from the four small windows.
2. Enter Ruth’s counts, 18, 7, 23, and 2, in the four small windows in the same order that they appear in the contingency table.
3. Click on the box labeled ‘COMPUTE.’ Three P-values will appear in the output window on the right side of the screen. For Ruth’s data, your display should read:

   • Left: p-value = 0.0691666077613616.
   • Right: p-value = 0.9883922891274275.
   • 2-Tail: p-value = 0.1383332155227232.

This website was **not** written by me; thus, unsurprisingly, it does not follow my notation. In particular, note the following:

• Left on the website gives the P-value for our alternative <.
• Right on the website gives the P-value for our alternative >.
• 2-Tail on the website gives the P-value for our alternative ≠.
Note that for Ruth’s data, the P-value for $\neq$ is twice the smaller of the other two P-values. This is no surprise. We know from our earlier work that whenever a study is balanced—i.e., whenever $n_1 = n_2$—then the exact sampling distribution of $U$—and, hence, $X$—is symmetric about 0. If $X = 0$, however, this doubling does not work. The P-value for $\neq$ is 1 and the P-values for the other two alternatives are equal and both exceed 0.5000.

Before we discuss the meaning of these P-values, I want you to get some additional practice using this website. In particular,

1. Go to the website and enter the counts for Therese’s infidelity study: 7, 3, 4 and 6. You should obtain the following P-values:
   - Left: $p$-value = 0.9651107406525374.
   - Right: $p$-value = 0.18492498213860464.
   - 2-Tail: $p$-value = 0.3698499642772093.

2. Go to the website and enter the counts for Thomas’s golf putting study: 18, 7, 10 and 15. You should obtain the following P-values:
   - Left: $p$-value = 0.9952024757062247.
   - Right: $p$-value = 0.02250227023961592.
   - 2-Tail: $p$-value = 0.04500454047923211.

3. Go to the website and enter the counts for HS-2: 29, 21, 21 and 29. You should obtain the following P-values:
   - Left: $p$-value = 0.964328798255893.
   - Right: $p$-value = 0.08060057614938207.
   - 2-Tail: $p$-value = 0.16120115229876414.

Let’s look at one of these P-values, say, the ‘Right’ P-value for Thomas’s golf putting study: 0.02250227023961592. This is a pretty absurd answer. Correct and absurd. It is absurd because it is so precise. Do I really care that the 17th digit after the decimal point is a 2? Would my interpretation change if it were a 7? Of course not! As a result, in these notes I will report a Fisher’s Exact Test P-value to four digits after the decimal, with the exception noted below. The exception can be illustrated by what I will call the really big and phony golf putting study (RBPGP).

I don’t want to spend much time on the RBPGP study, because, as its name suggests, it is make-believe. Consider the actual putting study performed by Thomas, but now suppose that his counts are all multiplied by ten, becoming: 180, 70, 100 and 150. I entered these counts into the Fisher website and obtained the following P-values:
   - Left: $p$-value = 0.9999999999999133.
Look at the P-value for Right (>). Reading left-to-right we begin with a curious result: the P-value equals 3.42 . . . which cannot be correct because it is larger than 1! This alerts us to keep reading. We are rewarded for our persistence when we find the e-13 at the end of the P-value. This is the website’s way of writing scientific notation. The correct way to interpret this P-value is that it equals 3.42 × 10^{−13}, or, if you prefer, 0.00000000003424, a really small P-value!

To summarize the above, I usually will report my P-values to four digits after the decimal point unless the result is close to 0.0000, in which case I usually will use scientific notation with four significant digits. (As Oscar Wilde said, “Consistency is the last refuge of the unimaginative.”)

Let us now look at Ruth’s prisoner study again. Ruth’s data, as you may have noted, display a remarkable pattern, which we will get to shortly. Let’s go through Ruth’s reasoning that led to her choice of alternative. The Skeptic’s Argument for Ruth’s study is that some men would volunteer to work with developmentally disabled adults and some would not, but that the treatment was irrelevant. Ruth decided that the only conceivable alternative is that the offer of a sentence reduction would increase the number of men who would volunteer. Thus, Ruth felt that the < alternative was inconceivable and she opted for the alternative >. Now look again at Ruth’s data: her \( \hat{p}_2 \) is larger than her \( \hat{p}_1 \) by 0.20, twenty percentage points. In words, her data support the alternative she labeled as inconceivable!

You might be wondering why I have included Ruth’s study in these notes. Let me assure you, my motive is not to ridicule Ruth. In fact, I admire Ruth quite a lot. Not many of my students traveled to a prison to collect data! (Ruth’s major was Social Work.) If I had been asked to select the alternative—remember, we do this before we collect data—I definitely would have made the same choice that Ruth did.

Thus, why have I included Ruth’s somewhat embarrassing study? First, to illustrate that inconceivable is not the same as impossible. Yes, it might be embarrassing, but I conjecture that every scientist will occasionally obtain data that support the inconceivable.

When data support the inconceivable, it is natural to wonder: What went wrong? So, take a minute and think about Ruth’s study. Can you think of any reason(s) why the mention of a sentence reduction would lead to less participation in the volunteer program? Below, I have listed some ideas of mine. Do any of these match your reason(s)? Do any of these seem particularly clever or stupid?

1. Perhaps the experience of being a prisoner makes the men not trust prison officials—or a researcher—to the extent that the promise of a reward has a negative impact.

2. Prisons are very expensive to operate. Perhaps it is routine for first-time nonviolent criminal offenders serving two or more years to receive a sentence reduction for good behavior. In this case, the prisoner might feel like he is being tricked; being offered a reward that he would likely obtain anyways. Thus, perhaps the question should have read,
Inmates who volunteer will receive a sentence reduction *in addition to any other sentence reductions.*

I am not going to *obsess* about possible explanations; my reason will become clear after we examine Ruth’s three P-values.

As we have seen earlier, Ruth’s exact P-value for her selected alternative > (Right) is 0.9884. This incredibly large P-value—remember the maximum possible value is 1—reflects the fact that, for her chosen alternative, Ruth obtained almost the *weakest possible evidence.* But look at her other two P-values: 0.0692 for < and 0.1383 for ≠. Under the usual classical approach to tests of hypotheses, *neither* of these P-values is small enough to reject the null hypothesis. Thus, the pattern in Ruth’s data, while surprising, is within the usual bounds of chance behavior, computed, of course, under the assumption the null hypothesis is correct. You might argue that 0.0692 is very close to the magic threshold of 0.05; but I would respond that, in my opinion and experience, it would be very difficult to convince many people that it is reasonable—before collecting data—to argue that > is *inconceivable* for Ruth’s study.

For the golf putting study, Thomas chose the alternative >. (Why does this make sense, before collecting the data?) Thus, his P-value is 0.0225 and the classical approach says to reject the null in favor of the alternative. In words, his data convinced Thomas that he was better at making a four foot putt rather than an eight foot putt.

Now, you might be thinking, “Of course, a shorter putt is easier. What a waste of time!” I have two comments to make on this attitude:

1. Quite often, especially in hindsight, research confirms the obvious. For example, it might seem *obvious* that smoking tobacco is harmful to a person’s health, but this conclusion was not reached easily, in part because it would have been unethical to use randomization.

2. Don’t fall into the trap of thinking you *know* how the world works. Subtle issues can operate in a study. For example, consider a highly skilled basketball player shooting baskets in a practice setting. Let treatment 1 be shooting a free throw and let treatment 2 be shooting from 12 inches in front of the free throw line. Now, you could argue that treatment 2 *must* be easier than treatment 1 because it is a shorter shot. I am not sure about this. Treatment 2 is a shorter shot, but my guess is that the player has previously spent a huge amount of time practicing treatment 1 and virtually no time practicing treatment 2. Thus, the player *might* actually be better from the longer distance. I do admit that this basketball argument seems, to me, not to apply to golf putts.

I will end this chapter with a few comments about the HS-2. I see three possible scenarios for this study:

1. If drug A is an active drug and drug B is, in fact, a placebo, then the alternative > seems obvious and the P-value is 0.0806.

2. If drug B is an active drug and drug A is, in fact, a placebo, then the alternative < seems obvious and the P-value is 0.9643. This would be a strange and alarming result: drug B seems worthless for headache relief and, indeed, is borderline contraindicated.
3. If both drugs A and B are reasonable treatments for headaches (see the variety of products available at a pharmacy) then $\neq$ is likely the alternative of choice, giving a P-value of 0.1612. Unless, for example, A was an *extra strength* version of B, in which case I would opt for the alternative $\geq$.

### 8.3 The Critical Region of a Test

#### 8.3.1 Motivation: Comparing P-values

Our focus on tests of hypotheses has been to find the P-value of a test. We can compare P-values *within* tests and *between* tests. Let me explain what I mean by these two forms of comparison.

Consider Dawn’s study of her cat Bob with the alternative $\geq$. Recall that for the test of means, the observed value of the test statistic is $u = 2.2$. Table 4.1 on page 76 presents an approximation to the sampling distribution of $U$ based on a simulation experiment with $m = 10,000$ reps. This approximation yields 0.0198 as the approximate P-value for Dawn’s data and the alternative $\geq$. Thus, the actual value of $u$, 2.2, provides fairly strong evidence in support of the alternative $\geq$.

I argued that a value of $u$ that is *larger than* the actual 2.2 would yield even stronger evidence for the alternative $\geq$. In particular, you can verify from Table 4.1 that if the observed value of the test statistic had been $u = 2.6$, then the approximate P-value for the alternative $\geq$ would have been:

\[
\text{Rel. Freq.}(U \geq 2.6) = 0.0023 + 0.0017 + 0.0008 + 0.0003 = 0.0051.
\]

The above is what I mean by a *within test* comparison of P-values. The value of $u$ which provides stronger evidence for $\geq$ (2.6 versus 2.2) yields the smaller P-value. This comparison is pretty noncontroversial: Within a study, stronger evidence yields a smaller P-value. By contrast, a *between test* comparison of P-values raises important issues.

Recall Sara’s study of golf, with data presented in Table 2.2 on page 29. For the alternative $\geq$ and the test based on means, I stated in Table 4.3 on page 77 that the approximate P-value—based on a simulation experiment with 10,000 reps—is 0.0903. By contrast, in Example 6.4 on page 127 I stated that the approximate P-value for the alternative $\geq$ and the sum of ranks test is 0.0293. Thus, for Sara’s data and the alternative $\geq$ the approximate P-value for the sum of ranks test is much smaller than the approximate P-value for the test that compares means; this is what I mean by a *between test* comparison of P-values.

I am tempted to say (and I have heard many scientists and statisticians in similar situations say) that for Sara’s data the sum of ranks test is better than the test that compares means because its P-value is much smaller. There are two reasons I resist saying this:

1. It is not literally true, even though arguably it is suggestive.

2. It sounds like I am cheating. (Perhaps) I prefer to reject the null hypothesis in favor of the alternative; thus, I state that the test that does what I want is better than the test that doesn’t!

The difficulty is that we have not, as yet, developed the tools needed to compare tests. We will do so in Chapter 9 when I introduce the notion of the power of a test. *Power* is a difficult topic and I have chosen to begin, in this chapter, the journey to your understanding power.
Table 8.6: A partial $2 \times 2$ contingency table of observed counts for a balanced CRD.

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Response</th>
<th>S</th>
<th>F</th>
<th>Total</th>
<th>$\hat{p}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>$a$</td>
<td>$b$</td>
<td>27</td>
<td>$a/27$</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>$c$</td>
<td>$d$</td>
<td>27</td>
<td>$c/27$</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>31</td>
<td>23</td>
<td>54</td>
<td></td>
</tr>
</tbody>
</table>

8.3.2 From P-values to Critical Regions

If a researcher has a CRD with a dichotomous response, then

http://www.langsrud.com/fisher.htm

is an incredibly useful site because it provides exact P-values with almost no effort from the researcher. As a teacher, however, the site is disappointing because it does not provide the entire sampling distribution of $X$. With some work, however, the site can be used to obtain the entire sampling distribution of $X$.

Example 8.5 (A balanced CRD on a total of 54 units that yields a total of 31 successes.) The table of data described in the name of this example is given in Table 8.6. Note that I am not telling you the actual cell counts, $a$, $b$, $c$ and $d$. My goal is to examine the sampling distribution of $X$ for this table; for this goal, as you recall, we don’t need the actual data, just the marginal totals.

A natural question for you to have is:

Bob, why did you choose a balanced study with 27 units on each treatment? Twenty-seven makes the arithmetic messy, to say the least.

My answer is twofold:

- We are going to avoid any arithmetic that involves dividing by 27; thus its messiness is irrelevant; and
- These data give me a number for a particular P-value that I like; I will reveal why soon.

From Table 8.6 we see that, given the cell counts:

\[
\hat{p}_1 = a/27 \text{ and } \hat{p}_2 = c/27; \text{ thus, } x = \hat{p}_1 - \hat{p}_2 = (a - c)/27
\]

is the observed value of the test statistic $X$. Also, $a + c = 31$ or $a - 31 = -c$. Thus,

\[
x = (a - c)/27 = (a + a - 31)/27 = (2a - 31)/27.
\]

My point is that the value of $a$ in Table 8.6 determines the value of $x$; as it must, because the value of $a$, given the marginal totals, determines the values of $b$, $c$ and $d$.

Let’s go to the site
and investigate what happens when I plug-in the value $a = 21$, which implies that $b = 6$, $c = 10$ and $d = 17$—otherwise, the margins in Table 8.6 would be wrong. The observed value of the test statistic is

$$x = (a - c)/27 = (21 - 10)/27 = 11/27.$$

I obtain the following three P-values, after rounding to four digits after the decimal point:

- **Left (<):** 0.9996; i.e., $P(X \leq 11/27) = 0.9996$.
- **Right (>):** 0.0027; i.e., $P(X \geq 11/27) = 0.0027$.
- **2-Tail($\neq$):** 0.0054; i.e., by symmetry:

$$P(|X| \geq 11/27) = 2P(X \geq 11/27) = 2(0.0027) = 0.0054.$$
1. Table 8.7 does not actually present all possible values of \( a \) as I claimed above. Why not? Well, for \( a < 8 \), equivalently \( x < -15/27 \),
\[
P(X \leq x) = 0.0000, \ P(X \geq x) = 1.0000 \text{ and } P(|X| \geq |x|) = 0.0000.
\]
In words, for \( x < -15/27 \), the P-value is really small for both < and \( \neq \). For my current purposes, we won’t be concerned with how small.

Similarly, for \( a > 23 \), equivalently \( x > 15/27 \),
\[
P(X \leq x) = 1.0000, \ P(X \geq x) = 0.0000 \text{ and } P(|X| \geq |x|) = 0.0000.
\]
In words, for \( x > 15/27 \), the P-value is really small for both > and \( \neq \).

2. For \( x < 0 \), the P-value for \( \neq \) is twice the P-value for <, except for possible round-off error. Similarly, for \( x > 0 \), the P-value for \( \neq \) is twice the P-value for >, except for possible round-off error. These relationships are no surprise; they follow from the fact that, because the study is balanced, the sampling distribution of \( X \) is symmetric around zero.

Recall the classical approach to interpreting a P-value, given in Chapter 5 on page 99 and reproduced below:

Reject the null hypothesis in favor of the alternative if, and only if, the P-value is less than or equal to 0.05.

Let’s consider the alternative >. Look at Table 8.7 again and note that the P-value is 0.05 or smaller, if, and only if, \( x \geq 7/27 \) (\( a \geq 19 \)). As a result, if my primary interest is on whether or not I reject the null hypothesis, then I have the rule

\[
\text{Reject the null hypothesis if, and only if, } X \geq 7/27.
\]

Being lazy, statisticians summarize the above by saying that the critical region for the test is
\[
(X \geq 7/27).
\]

(I will add parentheses, when deemed necessary, to set off the formula for a critical region from its neighboring text.) Thus, the null hypothesis is rejected if, and only if, the observed value of the test statistic falls in the critical region. The point being that it is easier to say the critical region is \( (X \geq 7/27) \) instead of the whole if, and only if, stuff.

By similar reasoning, the critical region for the alternative < is \( (X \leq -7/27) \). Finally, the critical region for the alternative \( \neq \) is \((|X| \geq 9/27)\).

Let’s pause for a moment. You might be wondering, “Why are we learning about critical regions?” This is a fair question. Sadly, you won’t see the answer until we have finished Chapter 9. Thus, please be patient. As Andre Gide said:

One does not discover new lands without consenting to lose sight of the shore for a very long time.
8.3.3 Two Types of Errors

Please look at Table 8.8. Statisticians find Table 8.8 to be a very useful way to think about a test of hypotheses. I will attempt to explain it to you.

The table consists of two rows and two columns, but unlike such tables earlier in this chapter, this table is not for data presentation.

1. The two rows of the table correspond to the two possible actions, or conclusions, available to the researcher:
   - The researcher can fail to reject the null hypothesis; or
   - The researcher can reject the null hypothesis.

2. The two columns correspond to the two possibilities for reality:
   - The null hypothesis can be correct; or
   - The alternative can be correct.

We see that the columns are an idealization; as we saw in Chapter 5, it is possible for the Skeptic (null hypothesis) to be incorrect and also for every alternative—even ≠—to be incorrect. For the current purposes, we are not concerned with this possibility. Thus, for example, for ease of exposition, I will sometimes refer to the second column as corresponding to the null hypothesis being false.

The four cells in the table represent all possible combinations of the two rows with the two columns.

1. The cells on the main diagonal correspond to correct actions, or conclusions:
   - The researcher fails to reject a correct null hypothesis; or
   - The researcher rejects a false null hypothesis.

2. The cells on the off diagonal correspond to incorrect actions, or conclusions:
   - The researcher rejects a true null hypothesis, called a Type 1 error; or
   - The researcher fails to reject a false null hypothesis, called a Type 2 error.
Rather obviously, an honest researcher prefers to take a correct action and to avoid our two types of errors.

Before data are collected, the researcher is uncertain about which action will be taken (which row will occur). In addition the researcher is always uncertain about the truth. The uncertainty in the rows (actions) will be dealt with using the ideas of probability, as you will see. The uncertainty in the columns, however, is a different matter. I will explain the popular method and then make a few comments about a different possible approach.

The popular approach is to condition on one of the columns being true. This, indeed, is what I did in Chapters 3–5 and, in fact, what I will do for all of the tests in these Course Notes. In particular, all of our analyses to date have been conditional on the assumption that the Skeptic is correct; i.e., that the null hypothesis is true. (In Chapter 9 we will consider what happens if we condition on the second column being correct.)

Look at the first column in Table 8.8 again. Conditional on the null hypothesis being correct, there are only two possible actions: the correct action—failing to reject the null hypothesis—and the incorrect action—rejecting the null hypothesis and making a Type 1 error. Thus, if you read that a test has, say, an 8% chance of making a Type 1 error, literally this means that on the assumption that the null hypothesis is true there is an 8% probability that the test will mistakenly reject the null hypothesis. The usually unstated implication is, of course, that there is a 92% probability that the test will correctly fail to reject the null hypothesis. Note that we know nothing about how the test performs if the null hypothesis is false—this will be the topic of Chapter 9.

There is another approach which I will briefly describe; view the material in the remainder of this subsection as optional enrichment. If you have heard of Bayesian analysis or the Bayesian approach and are curious about it, then you might want to continue reading.

The other approach is called the Bayesian approach. Before collecting data, the researcher specifies his/her personal probability that the null hypothesis is true. For example, Bert the researcher might state, “The probability that the null hypothesis is true is equal to 10%.” After the data are collected, Bert can update his personal probability. Updating is achieved by applying something called Bayes’ formula, which you will learn about later in these Course Notes. As a result of the updating, after analyzing the data, Bert might say, “My new probability that the null hypothesis is true is equal to 20%.”

Bayes’ formula for updating probabilities is great, but I don’t like applying it to test of hypotheses. There are two reasons I feel this way.

1. It can be difficult to obtain widespread acceptance of personal probabilities. For example, suppose that another researcher—call her Sally—is investigating the same phenomenon as Bert. She states, “The probability that the null hypothesis is true is equal to 90%.” Without going into details, you will likely agree that it seems reasonable that Bert and Sally could draw wildly different conclusions from the same data.

2. This is my stronger reason. I feel that, to a certain extent, the probability that the null hypothesis is correct is not that interesting. Why? Because I believe in Occam’s razor. A main idea of tests of hypotheses is to discard a simpler explanation only if it appears to be inadequate; whether the simpler explanation is likely to be true, to me, is largely irrelevant.
8.3.4 The Significance Level of a Test

Let’s return to my earlier example of Fisher’s test with margins $n_1 = n_2 = 27$, $m_1 = 31$ and $m_2 = 23$. I proposed the following critical regions for the three possible alternatives:

- For the alternative $>$, the critical region is $(X \geq 7/27)$ which gives 0.0489 as the probability that the test will make a Type 1 error.
- For the alternative $<$, the critical region is $(X \leq -7/27)$ which gives 0.0489 as the probability that the test will make a Type 1 error.
- For the alternative $\neq$, the critical region is $(|X| \geq 9/27)$ which gives 0.0267 as the probability that the test will make a Type 1 error.

The probability of a Type 1 also is called the **significance level** of a test and is denoted by $\alpha$ (the lower case Greek letter ‘alpha’). Thus, for the critical regions given above, $\alpha = 0.0489$ for the alternatives $>$ or $<$ and $\alpha = 0.0267$ for the alternative $\neq$.

Many textbooks state, “Researchers usually take $\alpha = 0.05$,” which begs the question, “Why didn’t I use $\alpha = 0.05$?” This is an easy question to answer: It is impossible to have $\alpha = 0.05$ for the margins that I have given you. Indeed, there are only a finite number of possible choices for $\alpha$ for this study; in particular, the possible values of $\alpha$ for the alternative $>$ coincide with the entries in the ‘$>$’ column in Table 8.7, e.g., $\alpha$ could be 0.0133 or 0.0489 or 0.1355 to name three possibilities, but it cannot be 0.05. (In fact, my strange choice of margins—27, 27, 31 and 23—reflected my desire to have a significance level that is close to 0.05; I discovered these happy margins by trial-and-error.)

8.4 Two Final Remarks

In this section I will examine two issues that frequently arise while using Statistics in scientific problems. In my experience, there is a great deal of confusion about these issues, but both may be resolved quite easily, thanks to the concept of critical regions.

8.4.1 Choosing the Alternative after Looking at the Data: Is it Really Cheating?

Yes, it is cheating, as I will now demonstrate. To keep this presentation brief, I will illustrate my conclusion with only one example. I hope that you will see that this example can be generalized; if not, I hope that you will trust me on this.

Refer to Example 8.5 on page 177 and its marginal counts, presented in Table 8.7 on page 178. Recall that we found critical regions and significance levels ($\alpha$’s) for the three possible alternatives:

- For the alternative $>$, the critical region is $(X \geq 7/27)$ and the significance level is $\alpha = 0.0489$. 

• For the alternative $<$, the critical region is $(X \leq -7/27)$ and the significance level is $\alpha = 0.0489$.

• For the alternative $\neq$, the critical region is $(|X| \geq 9/27)$ and the significance level is $\alpha = 0.0267$.

Next, consider a hypothetical researcher who I will call Hindsight Hank (Hank, for short). Hank wants to have $\alpha$ smaller than 0.05, but as close to 0.05 as possible. Thus, once Hank chooses his alternative, he should use one of the critical regions listed above.

Hank, however, refuses to select his alternative before collecting data and, instead, proceeds as follows.

• If the observed value of the test statistic, $x$, is greater than 0, then Hank says, “Well, obviously, $<$ is inconceivable. Thus, my alternative is $>$, my critical region is $(X \geq 7/27)$ and $\alpha = 0.0489$.”

• If the observed value of the test statistic, $x$, is smaller than 0, then Hank says, “Well, obviously, $>$ is inconceivable. Thus, my alternative is $<$, my critical region is $(X \leq -7/27)$ and $\alpha = 0.0489$.”

(Note that for the sampling distribution given in Table 8.7, the observed value of the test statistic cannot equal zero. For situations in which zero is a possible value of the test statistic, the argument below needs to be modified, but the basic important idea remains the same.)

We see that Hank’s actual critical region is $(|X| \geq 7/27)$; i.e., if the observed value $x$ satisfies:

$$x \geq 7/27 \text{ or } x \leq -7/27,$$

then Hank will reject the null hypothesis. From Table 8.7 we can see that Hank’s actual $\alpha$ is equal to 0.0978.

In summary, I label Hank’s behavior to be cheating because he claims to have $\alpha = 0.0489$, but his actual $\alpha$ is twice as large!

### 8.4.2 The Two-Sided Alternative Revisited

Suppose that you choose the alternative $\neq$ and have the sampling distribution given in Table 8.7. You choose the critical region $(|X| \geq 9/27)$, which gives $\alpha = 0.0267$. After data are collected, your observed value of the test statistic, $x$, equals or exceeds $9/27$. Thus, the action is to reject the null hypothesis. Here is the question I address in this subsection:

Which of the following is correct?

1. The scientific conclusion is that $p_1 > p_2$.
2. The scientific conclusion is that $p_1 \neq p_2$.

183
(If the answer seems obvious, please bear with me.) This can be an important issue for a scientist. For example, if the treatments are different medical protocols and a success is preferred to a failure, then the first conclusion indicates that that treatment 1 is preferred to treatment 2, whereas the second conclusion simply indicates that the treatments differ.

I have witnessed rather heated discussions over which conclusion is correct. Thus, I will spend a few minutes explaining why the first conclusion is correct; i.e., it is proper to conclude that treatment 1 is better than treatment 2. (Remember that successes are preferred. Also, remember that our conclusions in Part I of these notes are quite limited; I am saying that the proper conclusion is that for the units being studied treatment 1 is superior to treatment 2.)

My argument is actually quite simple, once I reveal a fact to you.

You have noted, no doubt, that I allow only two hypotheses in our tests of hypotheses. There are two reasons that all introductory statistics texts make this restriction.

1. Having exactly two hypotheses is the norm in scientific work.
2. Allowing for three or more hypotheses will result, for the most part, in more work for you with little benefit.

The current situation, however, is an exception to the for the most part disclaimer.

To a professional statistician, the pair of hypothesis:

\[ H_0: \text{The Skeptic is correct; and } H_1: p_1 \neq p_2, \]

is shorthand for the three hypothesis problem:

\[ H_0: \text{The Skeptic is correct; } H_1: p_1 > p_2; \text{ and } H_2: p_1 < p_2. \]

The critical region is actually:

If \( X \geq 9/27 \), then reject \( H_0 \) in favor of \( H_1 \); and if \( X \leq -9/27 \), then reject \( H_0 \) in favor of \( H_2 \).

We can see that the probability of a Type 1 error (rejecting a true null hypothesis for either conclusion) is:

\[ P(X \geq 9/27) + P(X \leq -9/27) = P(|X| \geq 9/27) = 0.0267. \]

In summary, if one selects the two-sided alternative and rejects the null hypothesis, then the proper scientific conclusion is:

- \( p_1 > p_2 \) if \( x > 0 \); or
- \( p_1 < p_2 \) if \( x < 0 \).

Although the example of this subsection is for a dichotomous response and Fisher’s test, the conclusion remains the same for the other two tests of Part I of these notes and, indeed, for analogous situations in population-based inference in Part II of these Course Notes.
8.5 Summary

In this chapter we consider CRDs with a dichotomous response. When the response is dichotomous, one possible response is labeled a success and the other a failure. Our suggested three rules for assigning these labels are given on page 168. We summarize our data by computing \( \hat{p}_1 \) [\( \hat{p}_2 \)], which is the proportion of successes on treatment 1 [2] in the data. Often, we also compute the (redundant) \( \hat{q}_1 \) [\( \hat{q}_2 \)], which is the proportion of failures on treatment 1 [treatment 2] in the data. Note that I say redundant because:

\[
\hat{p}_1 + \hat{q}_1 = 1 \text{ and } \hat{p}_2 + \hat{q}_2 = 1.
\]

If we identify the number ‘1’ with a success and the number ‘0’ with a failure, we see that we can match our current notation with our earlier notation for a numerical response, namely:

\[ \hat{p}_1 = \bar{x} \text{ and } \hat{p}_2 = \bar{y}. \]

This identification is very helpful because it implies that all of our earlier work on:

- the Skeptic’s argument;
- specify the hypotheses;
- the test statistic \( U \);
- sampling distributions;
- and the rules for computing P-values

apply immediately to the studies of this chapter. In particular, we define \( p_1 \) [\( p_2 \)] (note, no hat) to be the proportion of successes the researcher would have obtained if the All Treatment-1 [2] study had been performed. If the clone-enhanced study could be performed, then we would know the values of both \( p_1 \) and \( p_2 \).

As before, the null hypothesis is that the Skeptic is correct. The three options for the alternative are given on page 171. With a dichotomous response, our test is called Fisher’s test.

Finally, there is a wonderful website, [http://www.langsrud.com/fisher.htm](http://www.langsrud.com/fisher.htm), that is easy to use and gives us the exact P-value for Fisher’s test for every choice of alternative. When using this website, recall that:

- Left represents the alternative \(<\);
- Right represents the alternative \(>\); and
- 2-Tail represents the alternative \(\neq\).

It is convenient to use the symbol \( X \) to represent the test statistic for Fisher’s test, with observed value

\[ x = \hat{p}_1 - \hat{p}_2. \] (8.1)

We don’t actually compute \( x \) to obtain our P-value; the website above takes the counts \( a, b, c \) and \( d \) as input, saving us from the tedium of the dividing and subtracting needed to obtain \( x \).

The critical region of a test consists of the collection of all values of the test statistic that would result in the rejection of the null hypothesis.

Table 8.8 displays the four possible combinations of action (by the researcher) and truth (known only to Nature) that could occur in a test of hypotheses. It is reproduced below:
<table>
<thead>
<tr>
<th>Action (by researcher)</th>
<th>Truth (Only Nature knows)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$H_0$ is correct</td>
</tr>
<tr>
<td></td>
<td>$H_1$ is correct</td>
</tr>
<tr>
<td>Fails to reject $H_0$</td>
<td>Correct action</td>
</tr>
<tr>
<td>Rejects $H_0$</td>
<td>Type 1 Error</td>
</tr>
<tr>
<td></td>
<td>Correct action</td>
</tr>
</tbody>
</table>

After data are collected, the researcher rejects the null hypothesis if, and only if, the observed value of the test statistic lies in the critical region. Thus, the data determine which action and, hence, row of the table is relevant. The columns of the table are more problematic. The standard approach—which we will follow in these Course Notes—is to study the columns one-at-a-time. In particular, first we condition on the null hypothesis being true. You are familiar with this idea; we used it to obtain the sampling distributions for our three tests to date, comparisons of: means, mean ranks and proportions.

When statisticians talk about the probability of a Type 1 error, they really mean the probability of making a Type 1 error *conditional* on the assumption that the null hypothesis is true. (Conditional probabilities will be discussed more carefully in Chapter 16.) The probability of a Type 1 error is called the *significance level* of the test and is denoted by $\alpha$.

Typically, before collecting data a researcher selects a target value for $\alpha$; usually—but not exclusively—the target is 0.05. After the exact or approximate sampling distribution is obtained, using trial-and-error the researcher determines a critical region which gives a value of $\alpha$ which is close to the target value. I have shown you one example of this search for a critical region and will show you others in the Practice Problems below.

Sometimes a researcher specifies that the actual $\alpha$ should be as close as possible to the target, *without exceeding the target*. Thus, for example, if the target is 0.05, such a researcher would prefer $\alpha = 0.0450$ over, say, $\alpha = 0.0520$, even though the latter is closer to the target. I call this approach The Price is Right paradigm, but this name is unlikely to become widely used. (See also Practice Problem 4.)

Section 8.4 shows why it is considered to be cheating to look at one's data before selecting the alternative hypothesis. Also, Section 8.4 shows that if one rejects the null hypothesis after selecting a two-sided alternative, the proper scientific conclusion is the one-sided alternative supported by the data.

Finally, Chapter 9 will address the issue of determining the probability of a Type 2 error of a test.
Table 8.9: The $2 \times 2$ contingency table of observed counts for Sara’s golf study. A response is a success if, and only if, it equals or exceeds 100 yards.

<table>
<thead>
<tr>
<th>Club:</th>
<th>Response</th>
<th>Total</th>
<th>Proportions</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$S$</td>
<td>$F$</td>
<td>$S$</td>
<td>$F$</td>
</tr>
<tr>
<td>3-Wood</td>
<td>31</td>
<td>9</td>
<td>40</td>
<td>0.775</td>
</tr>
<tr>
<td>3-Iron</td>
<td>20</td>
<td>20</td>
<td>40</td>
<td>0.500</td>
</tr>
<tr>
<td>Total</td>
<td>51</td>
<td>29</td>
<td>80</td>
<td></td>
</tr>
</tbody>
</table>

Table 8.10: The $2 \times 2$ contingency table of observed counts for Reggie’s dart study. A response is a success if, and only if, it equals or exceeds 200 points.

<table>
<thead>
<tr>
<th>Distance:</th>
<th>Response</th>
<th>Total</th>
<th>Proportions</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$S$</td>
<td>$F$</td>
<td>$S$</td>
<td>$F$</td>
</tr>
<tr>
<td>10 Feet</td>
<td>9</td>
<td>6</td>
<td>15</td>
<td>0.60</td>
</tr>
<tr>
<td>12 Feet</td>
<td>5</td>
<td>10</td>
<td>15</td>
<td>0.33</td>
</tr>
<tr>
<td>Total</td>
<td>14</td>
<td>16</td>
<td>30</td>
<td></td>
</tr>
</tbody>
</table>

8.6 Practice Problems

1. Refer to Sara’s golf data that are presented in Table 2.2 on page 29. Suppose that, being a novice golfer, Sara is mostly interested in not embarrassing herself. Thus, she decides that a response of 100 yards or more is a success and a response of less than 100 yards is a failure. With this definition, Sara’s data are presented in Table 8.9. (I recommend that you verify the counts in this table.)

   (a) Suppose that Sara chooses the alternative $>$. Explain what this means in terms of the *Inconceivable Paradigm*.

   (b) Find the exact P-value for Sara’s data for Fisher’s test and the alternative $>$. 

   (c) Recall that, based on simulation experiments, approximate P-values for Sara’s data and the alternative $>$ are:

   - 0.0293 for the sum of ranks test; and
   - 0.0903 for the test that compares means.

   Compare these P-values to the answer you obtained in (b) and comment.

2. Refer to Reggie’s dart data that are presented in the Chapter 1 Homework Problems on page 25. Suppose that Reggie decided that a response of 200 or more points is a success
and a response of fewer than 200 points is a failure. With this definition, Reggie’s data are presented in Table 8.10. (I recommend that you verify the counts in this table.)

(a) Suppose that Reggie chooses the alternative $>$. Explain what this means in terms of the Inconceivable Paradigm.

(b) Find the exact P-value for Reggie’s data for Fisher’s test and the alternative $>$. 

(c) Based on simulation experiments, approximate P-value for Reggie’s data and the alternative $>$ are:
   - 0.0074 for the sum of ranks test; and
   - 0.0050 for the test that compares means.

   Compare these P-values to the answer you obtained in (b) and comment.

3. In this problem I want to reinforce the ideas of finding a critical region for Fisher’s test. Suppose that you want to perform a Fisher’s test for a table with the following marginal totals:

   $n_1 = n_2 = m_1 = m_2 = 20$.

   (a) Use trial-and-error to find the critical region for the alternative $>$ and $\alpha$ as close to 0.05 as possible.

   (b) Use trial-and-error to find the critical region for the alternative $>$ and $\alpha$ as close to 0.10 as possible.

   (c) Use trial-and-error to find the critical region for the alternative $<$ and $\alpha$ as close to 0.05 as possible.

   (d) Use trial-and-error to find the critical region for the alternative $\neq$ and $\alpha$ as close to 0.05 as possible.

4. This problem introduces the idea of finding the critical region for the test that compares means. In this problem I will use an exact sampling distribution. In Chapter 9 we will consider using an approximate sampling distribution.

   Table 8.11 presents the frequency distribution of $u$ for the 252 possible assignments for Kymn’s study of rowing. (If this table seems familiar to you it’s because you saw it in Table 3.7 on page 65)

   (a) Find the critical region for the alternative $>$ for $\alpha$ as close to the target value 0.05 as possible.

   (b) Many statisticians are fans of the television show The Price is Right. (I don’t actually know this; it’s just my segue to the following idea.) In particular, they want $\alpha$ to be as close to the target as possible without exceeding it. Repeat part (a) with this The Price is Right paradigm.

   (c) Find the critical region for the alternative $<$ for $\alpha$ as close to the target value 0.05 as possible.
Table 8.11: 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></td>
<td>0.0</td>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>252</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(d) Find the critical region for the alternative $\neq$ for $\alpha$ as close to the target value 0.05 as possible.

5. Refer to the sampling distribution given in the previous problem, see Table 8.11. Recall **Hindsight Hank**, who was introduced in Section 8.4. Hank decides to proceed as follows:

- If $x \geq 0$, he declares $>$ to be his alternative and $(U \geq 4.8)$ to be his critical region.
- If $x < 0$, he declares $<$ to be his alternative and $(U \leq -4.8)$ to be his critical region.

(a) According to Hank, what is his value of $\alpha$?

(b) What is Hank’s actual value of $\alpha$?

6. Recall that in Kymn’s actual study, $u = 7.2$. Suppose that she had chosen the alternative $\neq$; what would be her appropriate scientific conclusion?

8.7 **Solutions to Practice Problem**

1. (a) Sara decided that the alternative $<$ was inconceivable. This inconceivable alternative states that if Sara could have performed the clone-enhanced study, then there would have been more successes with the 3-Iron than with the 3-Wood.

(b) I entered the values 31, 9, 20 and 20 in the Fisher’s test website and obtained 0.0096 as the P-value for the alternative $>$. 

(c) The sum of ranks test gives a much smaller P-value than the test that compares means. In addition, Fisher’s test gives a much smaller P-value than the sum of ranks test.

2. (a) Reggie decided that the alternative $<$ was inconceivable. This inconceivable alternative states that if Reggie could have performed the clone-enhanced study, then there would have been more successes from 12 feet than from 10 feet. In short, Reggie felt that it was inconceivable for the greater distance to result in better accuracy.
(b) I entered the values 9, 6, 5 and 10 in the Fisher’s test website and obtained 0.1362 as the P-value for the alternative $>$. 

(c) The sum of ranks test and the test that compares means give similar and very small approximate P-values. The P-value from Fisher’s test is much larger than the other two P-values.

3. (a) My first guess is $a = 13$, which implies $b = c = 7$ and $d = 13$. The site gives me 0.0564 for the exact P-value for $>$. This is a good start! 

I next try $a = d = 14$ and $b = c = 6$, which yields 0.0128 for the exact P-value for $>$. Next, $a = 13$ gives $x = 13/20 - 7/20 = 6/20 = 0.30$. Thus, the critical region is $(X \geq 0.30)$ which has significance level $\alpha = 0.0564$.

(b) Building on my answer to (a), I next try $a = d = 12$ and $b = c = 8$, which yields 0.1715 for the exact P-value for $>$. Thus, the critical region in part (a) yields the value of $\alpha$ closest to 0.10.

(c) By symmetry, the critical region is $(X \leq -0.30)$ which gives $\alpha = 0.0564$ for the alternative $<$. 

(d) By symmetry and the work above, the critical region $(|X| \geq 0.30)$ gives $\alpha = 2(0.0564) = 0.1128$ for the alternative $\neq$; and the critical region $(|X| \geq 0.40)$ gives $\alpha = 2(0.0128) = 0.0256$ for the alternative $\neq$. Thus, the latter of these is the answer I seek.

4. (a) Because there are 252 equally likely possible assignments, all probabilities for $U$ will be of the form $k/252$, for some value of $k$. My target is to obtain the probability 0.05. Thus, I begin by setting 

$$k/252 = 0.05$$ 

and solving for $k$: $k = 0.05(252) = 12.6$.

By trial-and-error in Table 8.11 I find:

$$P(U \geq 4.8) = 14/252 = 0.0556 \text{ and } P(U \geq 5.2) = 11/252 = 0.0437.$$ 

Thus, the critical region is $(U \geq 4.8)$ and $\alpha = 0.0556$.

(b) The critical region chosen in part (a) violates The Price is Right paradigm because its $\alpha$ exceeds the target, 0.05. Instead, we use the critical region $(U \geq 5.2)$ which gives $\alpha = 0.0437$.

(c) By symmetry, the critical region is $(U \leq -4.8)$ and $\alpha = 0.0556$.

(d) By symmetry, I want to find the number $c$ such that $P(U \geq c)$ is as close as possible to 0.025, one-half of the target value. I begin by setting 

$$k/252 = 0.025$$ 

and solving for $k$: $k = 0.025(252) = 6.3$.

By trial-and-error in Table 8.11 I find:

$$P(U \geq 5.6) = 7/252 = 0.0278$$ 

is the closest value to 0.025. Thus, the critical region is $(|U| \geq 5.6)$ and $\alpha = 2(0.0278) = 0.0556$.
5.  (a) According to Hank, if his $x \geq 0$, he would say:

$$\alpha = P(U \geq 4.8) = \frac{14}{252} = 0.0556.$$  

Alternatively, if his $x < 0$, he would say:

$$\alpha = P(U \leq -4.8) = \frac{14}{252} = 0.0556.$$  

(b) Actually,

$$\alpha = P(U \geq 4.8) + P(U \leq -4.8) = \frac{28}{252} = 0.1111.$$  

6. Kymn would reject the null hypothesis and, because $x = 7.2 > 0$, her appropriate conclusion is that $\mu_1 > \mu_2$.

### 8.8 Homework Problems

1. Mary, a student in my class, conducted a study that she called *Is Pearl Ambidextrous?* The balanced study consisted of a total of 50 trials. Treatment 1 [2] was a canter depart on a left [right] lead. According to Mary—I don’t know about these things—a canter can be executed successfully or not. Pearl, a seven year-old mare, obtained a total of 42 successes, with 22 successes coming on treatment 1.

   (a) Present Mary’s data in a $2 \times 2$ table; calculate the proportions of successes; and comment.

   (b) Find Mary’s exact $P$-value for each of the three possible alternatives. Comment.

2. Robert, a student in my class, enjoyed target shooting with his rifle. He performed a balanced study on a total of 100 trials. Treatment 1 [2] was shooting from the prone [kneeling] position. A shot was labeled a success if it hit a specified region of the target. Robert obtained a total of 67 successes, with 25 coming from the kneeling position.

   (a) Present Robert’s data in a $2 \times 2$ table; calculate the proportions of successes; and comment.

   (b) In his report, Robert stated:

   "Everyone believes that shooting from the prone position is more accurate than shooting from the kneeling position."

   Given this belief, what should Robert use for his alternative hypothesis?

   (c) Find Robert’s exact $P$-value for each of the three possible alternatives. Comment on the $P$-value for your answer in part (b).

3. Suppose that you want to perform a Fisher’s test for a table with the following marginal totals:

   $$n_1 = n_2 = m_1 = m_2 = 25.$$  

   (If you have difficulty answering the questions below, refer to Practice Problem 3.)
Table 8.12: Frequency table for the values $r_1$ of $R_1$ for the 252 possible assignments for a balanced CRD with $n = 10$ units and 10 distinct response values.

<table>
<thead>
<tr>
<th>$r_1$</th>
<th>Freq.</th>
<th>$r_1$</th>
<th>Freq.</th>
<th>$r_1$</th>
<th>Freq.</th>
<th>$r_1$</th>
<th>Freq.</th>
<th>$r_1$</th>
<th>Freq.</th>
<th>$r_1$</th>
<th>Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>1</td>
<td>19</td>
<td>5</td>
<td>23</td>
<td>14</td>
<td>28</td>
<td>20</td>
<td>33</td>
<td>11</td>
<td>37</td>
<td>3</td>
</tr>
<tr>
<td>16</td>
<td>1</td>
<td>20</td>
<td>7</td>
<td>24</td>
<td>16</td>
<td>29</td>
<td>19</td>
<td>34</td>
<td>9</td>
<td>38</td>
<td>2</td>
</tr>
<tr>
<td>17</td>
<td>2</td>
<td>21</td>
<td>9</td>
<td>25</td>
<td>18</td>
<td>30</td>
<td>18</td>
<td>35</td>
<td>7</td>
<td>39</td>
<td>1</td>
</tr>
<tr>
<td>18</td>
<td>3</td>
<td>22</td>
<td>11</td>
<td>26</td>
<td>19</td>
<td>31</td>
<td>16</td>
<td>36</td>
<td>5</td>
<td>40</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Total 252</td>
</tr>
</tbody>
</table>

(a) Use trial-and-error to find the critical region for the alternative $>\alpha$ and $\alpha$ as close to 0.05 as possible.

(b) Use trial-and-error to find the critical region for the alternative $<\alpha$ and $\alpha$ as close to 0.05 as possible.

(c) Use trial-and-error to find the critical region for the alternative $\neq\alpha$ and $\alpha$ as close to 0.05 as possible.

4. Table [8.12] presents the frequency distribution of the values of $r_1$ for the sum of ranks test for a balanced CRD with a total of $n = 10$ units and no tied values. (You saw this table previously in Table [6.11]) If you have difficulty with the questions below, refer to Practice Problem 4.

Find the critical region for the test statistic $R_1$ and the alternative $>\alpha$ that gives $\alpha$ as close as possible to my target value 0.05, **without exceeding the target**.