Chapter 7

Visualizing a Sampling Distribution

Let’s review what we have learned about sampling distributions. We have considered sampling distributions for the test of means (test statistic is $U$) and the sum of ranks test (test statistic is $R_1$). We have learned, in principle, how to find an exact sampling distribution. I say in principle because if the number of possible assignments is large, then it is impractical to attempt to obtain an exact sampling distribution.

We have learned an excellent way to approximate a sampling distribution, namely a computer simulation experiment with $m = 10,000$ runs. We can calculate a nearly certain interval to assess the precision of any given approximation and, if we are not happy with the precision, we can obtain better precision simply by increasing the value of $m$. Computer simulations are a powerful tool and I am more than a bit sad that they were not easy to perform when I was a student many decades ago. (We had to walk uphill, through the snow, just to get to the large building that housed the computer and then we had to punch zillions of cards before we could submit our programs.)

Before computer simulations were practical, or even before computers existed, statisticians and scientists obtained approximations to sampling distributions by using what I will call fancy math techniques. We will be using several fancy math methods in these notes.

Fancy math methods have severe limitations. For many situations they give poor approximations and, unlike a computer simulation, you cannot improve a fancy math approximation simply by increasing the value of $m$; there is nothing that plays the role of $m$ in a fancy math approximation. Also, there is nothing like the nearly certain interval that will tell us the likely precision of a fancy math approximation.

Nevertheless, fancy math approximations are very important and can be quite useful; here are two reasons why:

1. Do not think of computer simulations and fancy math as an either/or situation. We can, and often will, use them together in a problem. For example, a simple fancy math argument will often show that one computer simulation experiment can be applied to many—sometimes an infinite number of—situations. We will see many examples of this phenomenon later in these Course Notes.

2. Being educated is not about acquiring lots and lots of facts. It is more about seeing how lots and lots of facts relate to each other or reveal an elegant structure in the world. Computer
simulations are very good at helping us acquire facts, whereas fancy math helps us see how these facts fit together.

Fancy math results can be very difficult to prove and these proofs are not appropriate for this course. Many of these results, however, can be motivated with pictures. This begs the question: Which pictures? The answer: Pictures of sampling distributions.

Thus, our first goal in this chapter is to learn how to draw a particular picture, called the probability histogram, of a sampling distribution.

### 7.1 Probability Histograms

As the name suggest, a probability histogram is similar to the histograms we learned about in Chapter 2. For example, just like a histogram for data, a probability histogram is comprised of rectangles on the number line. There are some important differences, however. First, a motivation for our histograms in Chapter 2 was to group data values in order to obtain a better picture. By contrast, we never group values in a probability histogram. Second, without grouping, we don’t need an endpoint convention for a probability histogram and, as a result, we will have a new way to place/locate its rectangles.

The total area of the rectangles in a probability histogram equals 1, which is a feature shared by density histograms of Chapter 2. The reason? Density histograms use area to represent relative frequencies of data; hence, their total area is one. Probability histograms use area to represent probabilities; hence, their total area equals the total probability, one.

Table 7.1 presents the sampling distribution of $R_1$ for Cathy’s study of running. (Remember: There were no ties in Cathy’s six response values.) This table was presented in Chapter 6. Its probability histogram is presented in Figure 7.1. Look at it briefly and then read my description below of how it was created.

First, some terminology. Thus far in these Course Notes our sampling distributions have been for test statistics, either $U$ or $R_1$. In general, we talk about a sampling distribution for a random variable $X$, with observed value $x$. Here is the idea behind the term random variable. We say variable because we are interested in some feature that has the potential to vary. We say random because the values that the feature might yield are described by probabilities. Both of our test

<table>
<thead>
<tr>
<th>$r_1$</th>
<th>$P(R_1 = r_1)$</th>
<th>$r_1$</th>
<th>$P(R_1 = r_1)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>0.05</td>
<td>11</td>
<td>0.15</td>
</tr>
<tr>
<td>7</td>
<td>0.05</td>
<td>12</td>
<td>0.15</td>
</tr>
<tr>
<td>8</td>
<td>0.10</td>
<td>13</td>
<td>0.10</td>
</tr>
<tr>
<td>9</td>
<td>0.15</td>
<td>14</td>
<td>0.05</td>
</tr>
<tr>
<td>10</td>
<td>0.15</td>
<td>15</td>
<td>0.05</td>
</tr>
</tbody>
</table>
statistics are special cases of random variables and, hence, are covered by the method described below.

1. On a horizontal number line, mark all possible values, \( x \), of the random variable \( X \).

For the sampling distribution in Table 7.1 these values of \( x \) (\( r_1 \)) are 6, 7, 8, \ldots 15 and they are marked in Figure 7.1.

2. Determine the value of \( \delta \) (lower case Greek delta) for the random variable of interest. The number \( \delta \) is the smallest distance between any two consecutive values of the random variable.

For the sampling distribution in Table 7.1 the distance between consecutive values is always 1; hence, \( \delta = 1 \).

3. Above each value of \( x \), draw a rectangle, with its center at \( x \), its base equal to \( \delta \) and its height equal to \( P(X = x) / \delta \).

In the current example, \( \delta = 1 \), making the height of each rectangle equal to the probability of its center value.

For a probability histogram the area of a rectangle equals the probability of its center value, because:

\[
\text{Area of rectangle centered at } x = \text{Base} \times \text{Height} = \delta \times \frac{P(X = x)}{\delta} = P(X = x).
\]

In the previous chapter we studied the sum of ranks test with test statistic \( R_1 \). In all ways mathematical, this test statistic is much easier to study if there are no ties in the data. I will now show how the presence of one tie affects the probability histogram.

**Example 7.1 (A small CRD with two values tied.)** Table 7.2 presents the sampling distribution of \( R_1 \) for a balanced CRD with a total of \( n = 6 \) units with one particular pair of tied observations: the two smallest observations are tied and the other four observations are not. Thus, the ranks are: 1.5, 1.5, 3, 4, 5 and 6. If you feel a need to have more practice at determining sampling distributions, you may verify the entries in this table. Otherwise, I suggest you trust me on it. The probability histogram for this sampling distribution is presented in Figure 7.2. I will walk you through the three steps to create this picture.
Table 7.2: The sampling distribution of $R_1$ for a balanced CRD with a total of $n = 6$ units with ranks 1.5, 1.5, 3, 4, 5 and 6.

<table>
<thead>
<tr>
<th>$r_1$</th>
<th>$P(R_1 = r_1)$</th>
<th>$r_1$</th>
<th>$P(R_1 = r_1)$</th>
<th>$r_1$</th>
<th>$P(R_1 = r_1)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.0</td>
<td>0.05</td>
<td>9.5</td>
<td>0.10</td>
<td>12.5</td>
<td>0.10</td>
</tr>
<tr>
<td>7.0</td>
<td>0.05</td>
<td>10.5</td>
<td>0.20</td>
<td>13.0</td>
<td>0.05</td>
</tr>
<tr>
<td>8.0</td>
<td>0.05</td>
<td>11.5</td>
<td>0.10</td>
<td>14.0</td>
<td>0.05</td>
</tr>
<tr>
<td>8.5</td>
<td>0.10</td>
<td>12.0</td>
<td>0.05</td>
<td>15.0</td>
<td>0.05</td>
</tr>
<tr>
<td>9.0</td>
<td>0.05</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1. On a horizontal number line, mark all possible values, $x$, of the random variable $X$.

   The 13 possible values of $R_1$ in Table 7.2 are marked, but not always labeled, in Figure 7.2.

2. Determine the value of $\delta$ for the random variable of interest. The number $\delta$ is the smallest distance between any two consecutive values of the random variable.

   This is trickier than it was in our first example. Sometimes the distance between consecutive values is 1 and sometimes it is 0.5. Thus, $\delta = 0.5$.

3. Above each value of $x$, draw a rectangle, with its center at $x$, its base equal to $\delta$ and its height equal to the $P(X = x)/\delta$.

   In the current example, $\delta = 0.5$, making the height of each rectangle equal to twice the probability of its center value.

Now that we have the method of constructing a probability histogram for a random variable, let’s look at the two pictures we have created, Figures 7.1 and 7.2. Both probability histograms are symmetric with point of symmetry at 10.5. (It can be shown that the sampling distribution for $R_1$ is symmetric if the CRD is balanced and/or there are no ties in the data set.)

Figure 7.1 is, in my opinion, much more well-behaved than Figure 7.2. Admittedly, well-behaved sounds a bit subjective; here is what I mean. Figure 7.1 has one peak—four rectangles wide, but still only one peak—and no gaps. By contrast, Figure 7.2 has one dominant peak (at 10.5), eight lesser peaks (at 6, 7, 8.5, 9.5, 11.5, 12.5, 14 and 15) and six gaps (at 6.5, 7.5, 10.0, 11.0, 13.5 and 14.5).

In general, for the sum of ranks test, if there are no ties in the combined data set, the probability histogram for the sampling distribution of $R_1$ is symmetric, with one peak and no gaps. If there is as few as one pair of tied values in the combined data set, the probability histogram for the sampling distribution of $R_1$ might not be very nice! Or it might be, as our next example shows.

**Example 7.2 (A small CRD with three values tied.)** I have a balanced CRD with a total of six units. My six ranks are: 2, 2, 2, 4, 5 and 6. (An exercise for you: create data that would yield these ranks.) The sampling distribution for $R_1$ is given in Table 7.3 and its probability histogram is presented in Figure 7.3.
If you compare this table and figure to the table and figure for Cathy’s no-ties data—Table 7.1 and Figure 7.1—you see that the effect of the three tied observations is quite minimal. Two of the 20 values of $r_1$ in Cathy’s distribution, a 7 and a 14, are replaced by an 8 and a 13 in the current study.

To summarize, when there are no ties in the combined data set, our probability histograms are of a type: they are symmetric with one peak and no gaps. We have seen one such probability histogram in Figure 7.1 and will see another one later in this chapter in Figure 7.5.

We have seen two probability histograms for the situation in which there is at least one tie in the combined data set. The variety of such pictures (these two as well as other possible ones) is too great for us to devote more time to the topic. Thus, I propose the following agreement between you and me. I will focus on the no-ties pictures to motivate the fancy math approximate method I present later in this chapter. The fancy math method can be used when there are ties; in fact, there is an explicit adjustment term in the method that comes into play only if there are ties. As with all approximation methods in Statistics, the method’s utility depends on how close the approximate
answers are to the exact answers; this utility does not depend on how cleverly—or clumsily—I motivate it.

Before we proceed further, let me acknowledge a point. There is no unique way to draw a picture of a sampling distribution. You might be able to create a picture that you prefer to the probability histogram. Thus, I do not present a probability histogram as the correct picture for a sampling distribution. You will see soon, however, that our probability histogram is very good at motivating the fancy math approximation of this chapter. Indeed, our definition of a probability histogram is very good at motivating many of the the fancy math approximations in these Course Notes.

Look at the rectangle centered at \( r_1 = 10 \) in Figure 7.1. Its area is its base times its height: \( 1 \times 0.15 = 0.15 \). Thus, because area equals probability, \( P(R_1 = 10) = 0.15 \). Here is my point: the probability belongs to the value 10, but the probability histogram visually spreads the probability over the entire base of the rectangle, from 10.5 to 11.5. This spreading has no real meaning, but, as we will see, it does help motivate our fancy math approximation.

I have one more point to make before we move on to the next section. Knowing the sampling distribution of a test statistic (more generally, a random variable) is mathematically equivalent to knowing its probability histogram. Thus, approximating a sampling distribution is equivalent to approximating its probability histogram. Thus, below we will learn how to approximate a probability histogram.

### 7.2 The Mean and Standard Deviation of \( R_1 \)

We learned in Chapter 1 that a set of numbers can be summarized—admittedly, sometimes poorly—by its mean. We also learned that the mean of a set of numbers can be visualized as the center of gravity of its dot plot. It is also possible to summarize the sampling distribution of a random variable by its mean. For example, consider the sampling distribution for \( R_1 \) given in Table 7.1. Recall that there are 20 possible assignments for the CRD in question and that each assignment yields a value of \( r_1 \). From the sampling distribution, we can infer that these 20 values are, after sorting:

\[ 6, 7, 8, 8, 9, 9, 10, 10, 10, 11, 11, 11, 12, 12, 12, 13, 13, 14, 15. \]
(For example, \( P(R_1 = 8) = 0.10 \); thus, 10\% of the 20 assignments—two—yield the value \( r_1 = 8 \).) If we sum these 20 values and divide by 20, we find that the mean of these 20 values is 10.5. For a sampling distribution, we refer to the mean by the Greek letter \( \mu \) (pronounced ‘mu’ with a long ‘u’). Thus, \( \mu = 10.5 \) for the sampling distribution in Table 7.1.

Technical Note: For a random variable, \( X \), with observed value \( x \), we have the following formula for \( \mu \):

\[
\mu = \sum x P(X = x),
\]

where the sum is taken over all possible values \( x \) of \( X \). We won’t ever use Equation 7.1 so, thus, you may safely ignore it. I have included it for completeness only.

For our method of creating a probability histogram, it can be shown that \( \mu \) is equal to the center of gravity of the probability histogram. (Remember: our probability histogram, unlike the histograms for data introduced in Chapter 2, do not group values into categories.) Thus, we can see immediately from Figures 7.1—7.3 that all three probability histograms—and, hence, all three sampling distributions—have \( \mu = 10.5 \).

The following mathematical result is quite useful later in this chapter. I will not prove it and you need not be concerned with its proof.

**Result 7.1 (Mean of \( R_1 \)).** Suppose that we have a CRD with \( n_1 \) units assigned to the first treatment and a total of \( n \) units. The mean, \( \mu \), of the sampling distribution of \( R_1 \) is given by the following equation.

\[
\mu = n_1(n + 1)/2
\]  

Note that this result is true whether or not there are ties in the combined data set. For example, suppose that \( n_1 = 3 \) and \( n = 6 \) in Equation 7.2. We find that

\[
\mu = 3(6 + 1)/2 = 10.5,
\]

in agreement with what we could see in Figures 7.1—7.3.

Here is the important feature in Equation 7.2. Even though—as we have seen repeatedly—it can be difficult to obtain exact probabilities for \( R_1 \), it is very easy to calculate the mean of its sampling distribution.

In order to use our fancy math approximation, we also need to be able to calculate the variance of a sampling distribution. You recall that the word variance was introduced in Chapter 1. For a set of data, the variance measures the amount of spread, or variation, in the data. Let’s review how we obtained the variance in Chapter 1.

In Chapter 1 we began by associating with each response value, e.g., \( x_i \) for data from treatment 1, its deviation: \((x_i - \bar{x})\). For a sampling distribution, the deviation associated with possible value \( x \) is \((x - \mu)\). Next, in Chapter 1 we squared each deviation; we do so again, obtaining the squared deviation \((x - \mu)^2\). Next, we sum the squared deviations over all possible assignments, and finally, for sampling distributions, we follow the method mathematicians use for data and divide the sum of squared deviations by the total number of possible assignments; i.e., unlike with data, we do not subtract one. (Trust me, the reason for this is not worth the time needed to explain it.) The result, the mean of these squared deviations, is called the variance of the sampling distribution.
and is denoted by $\sigma^2$. (Note: $\sigma$ is the lower case Greek letter sigma.) Following the ideas of Chapter 1, the positive square root of the variance, $\sigma$, is called the standard deviation of the sampling distribution.

Technical Note: For a random variable, $X$, with observed value $x$, we have the following formula for the variance $\sigma^2$:

$$
\sigma^2 = \sum_x (x - \mu)^2 P(X = x),
$$

(7.3)

where the sum is taken over all possible values $x$ of $X$. We won’t ever use Equation 7.3, thus, you may safely ignore it. I have included it for completeness only.

Statisticians prefer to focus on the formula for variance, as we do in Equation 7.3, rather than a formula for the standard deviation because the latter will have that nasty square root symbol. In our fancy math approximation, however, we will need to use the standard deviation. We always calculate the variance first; then take its square root to obtain the standard deviation.

I will present two formulas for the variance of the sampling distribution of $R_1$: the first is for the situation when there are no ties in the combined data set. If there are ties, then the second formula should be used.

**Result 7.2 (Variance of $R_1$ when there are no ties.)** Suppose that we have a CRD with $n_1$ units assigned to the first treatment, $n_2$ units assigned to the second treatment and a total of $n$ units. Suppose, in addition, that there are no tied values in our combined list of $n$ observations. The variance of the sampling distribution of $R_1$ is given by the following equation.

$$
\sigma^2 = \frac{n_1 n_2 (n + 1)}{12}.
$$

(7.4)

For example, suppose we have a balanced CRD with $n = 6$ units and no ties; for example, Cathy’s study. Using Equation 7.4 we see that the variance of $R_1$ is

$$
\sigma^2 = \frac{3(3)(7)}{12} = 5.25 \text{ and } \sigma = \sqrt{5.25} = 2.291.
$$

Next, suppose we have a balanced CRD with $n = 10$ units and no ties. Using Equation 7.4 we see that the variance of $R_1$ is

$$
\sigma^2 = \frac{5(5)(11)}{12} = 22.917 \text{ and } \sigma = \sqrt{22.917} = 4.787.
$$

When there are ties in the combined data set, then we need to use the following more complicated formula for the variance. Note that I will define the symbol $t_i$ below the Result.

**Result 7.3 (Variance of $R_1$ when there are ties.)** Suppose that we have a CRD with $n_1$ units assigned to the first treatment, $n_2$ units assigned to the second treatment and a total of $n$ units. Suppose, in addition, that there is at least one pair of tied values in the combined list of $n$ observations. The variance of the sampling distribution of $R_1$ is given by the following equation.

$$
\sigma^2 = \frac{n_1 n_2 (n + 1)}{12} - \frac{n_1 n_2 \sum (t_i^3 - t_i)}{12n(n - 1)}.
$$

(7.5)
In this equation, the \( t_i \)'s require a bit of explanation. As we will see, each \( t_i \) is a positive count. Whenever a \( t_i = 1 \) then the term in the sum, \((t_i^3 - t_i)\) will equal 0. This fact has two important consequences:

- All of the values of \( t_i = 1 \) can (and should) be ignored when evaluating Equation 7.5.
- If all \( t_i = 1 \) then Equation 7.5 is exactly the same as Equation 7.4. As we will see soon, saying that all \( t_i = 1 \) is equivalent to saying that there are no ties in the combined data set.

It will be easier for me to explain the \( t_i \)'s with an example.

Recall Kymn’s study. She performed a balanced CRD with \( n = 10 \) trials. In her combined data set she had nine distinct numbers: Eight of these numbers occurred once in her combined data set and the remaining number occurred twice. This means that Kymn’s \( t_i \)'s consisted of eight 1’s and one 2. Thus, the variance of \( R_1 \) for Kymn’s study is

\[
\sigma^2 = \frac{5(5)(11)}{12} - \frac{5(5)(2^3 - 2)}{12(10)(9)} = 22.917 - \frac{150}{1080} = 22.778 \text{ and } \sigma = \sqrt{22.778} = 4.773.
\]

For this example, the presence of one pair of tied values has very little impact on the value of the variance. Indeed, the standard deviation for Kymn’s study is only 0.29% smaller than the standard deviation when there are no ties (which, recall, is 4.787). You can well understand why many statisticians ignore a small number of ties in a combined data set. When the response is ordinal and categorical, however, ties can have a notable impact on the standard deviation, as our next computation shows.

Recall the data in Table 6.8 on page 129. These artificial data come from a balanced CRD with \( n = 100 \) subjects and only three response values (categories). Recall that 42 subjects gave the first response, 33 subjects gave the second response and 25 subjects gave the third response. Thus, the \( t_i \)'s are 42, 33 and 25. I will plug these values into Equation 7.5. First, I will calculate the value of the expression containing the \( t_i \)'s:

\[
\sum(t_i^3 - t_i) = (42^3 - 42) + (33^3 - 33) + (25^3 - 25) = 74,046 + 35,904 + 15,600 = 125,550.
\]

Thus, the variance is:

\[
\sigma^2 = \frac{50(50)(101)}{12} - \frac{50(50)(125,550)}{12(100)(99)} = 21,041.67 - 2,642.05 = 18,399.62.
\]

The correct standard deviation—taking into account ties—is \( \sqrt{18,399.62} = 135.65 \). It is smaller than the answer one would obtain by ignoring the effect of ties: \( \sqrt{21,041.67} = 145.06 \). The former is 6.49% smaller than the latter and this difference has an important impact on our fancy math approximation, as you will see later in this chapter.

### 7.3 The Family of Normal Curves

Do you recall \( \pi \), the famous number from math? It is the ratio of the circumference to the diameter of a circle. Another famous number from math is \( e \), which is the limit as \( n \) goes to infinity of
As decimals, $\pi = 3.1416$ and $e = 2.7183$, both approximations. If you want to learn more about $\pi$ or $e$, go to Wikipedia. If you do not want to learn more about them, that is fine too.

Let $\mu$ denote any real number—positive, zero or negative. Let $\sigma$ denote any positive real number. In order to avoid really small type, when $t$ represents a complicated expression, we write $e^t$ as $\exp(t)$. Consider the following function.

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}}\exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right), \text{ for all real numbers } x. \quad (7.6)$$

The graph of the function $f$ is called the **Normal curve** with parameters $\mu$ and $\sigma$; it is pictured in Figure 7.4. By allowing $\mu$ and $\sigma$ to vary, we generate the family of Normal curves. We use the terminology:

the N($\mu$, $\sigma$) curve

to designate the Normal curve with parameters $\mu$ and $\sigma$. Thus, for example, the N(25,8) curve is the Normal curve with mean $\mu = 25$ and standard deviation $\sigma = 8$.

Below is a list of important properties of Normal curves.

1. The total area under a Normal curve is one.

2. A Normal curve is symmetric about the number $\mu$. Clearly, $\mu$ is the center of gravity of the curve, so we call it the mean of the Normal curve.

3. It is possible to talk about the spread in a Normal curve just as we talked about the spread in a sampling distribution or its probability histogram. In fact, one can define the standard deviation as a measure of spread for a curve and if one does, then the standard deviation for a Normal curve equals its $\sigma$.

4. You can now see why we use the symbols $\mu$ and $\sigma$ for the parameters of a Normal curve: $\mu$ is the mean of the curve and $\sigma$ is its standard deviation.

5. A Normal curve has points of inflection at $\mu + \sigma$ and $\mu - \sigma$. If you don’t know what a point of inflection is, here goes: it is a point where the curve changes from ‘curving downward’ to ‘curving upward.’ I only mention this because: If you see a picture of a Normal curve you can immediately see $\mu$, its point of symmetry. You can also see its $\sigma$ as the distance between $\mu$ and either point of inflection.

6. The Normal curve with $\mu = 0$ and $\sigma = 1$—that is, the N(0,1) curve—is called the **Standard Normal curve**.

Statisticians often want to calculate areas under a Normal curve. Fortunately, there exists a website that will calculate areas for us; I will present instructions on how to use this site in Section 7.4.
7.3.1 Using a Normal Curve to obtain a fancy math approximation

Table 7.4 presents the exact sampling distribution for $R_1$ for a balanced CRD with $n = 10$ units and no tied values. (Trust me; it is no fun to verify this!) Figure 7.5 presents its probability histogram. In this figure, I have shaded the rectangles centered at 35, 36, ..., 40. If we remember that in a probability histogram, area equals probability, we see that the total area of these six shaded rectangles equals $P(R_1 \geq 35)$. If, for example, we performed a CRD and the sampling distribution is given by Table 7.4 and the actual value of $r_1$ is 35, then $P(R_1 \geq 35)$ would be the P-value for the alternative $>$. I.e., the area of the shaded region in Figure 7.5 would be the P-value for the alternative $>$. Now look at Figures 7.5 and 7.4. Both pictures are symmetric with a single peak and are bell-shaped. These similarities suggest that using a Normal curve to approximate the probability histogram might work well. There is no reason to argue my visual interpretation; let’s try it and see what happens.

The idea is to use one of the members of the family of Normal curves to approximate the probability histogram in Figure 7.5. Which one? Because the probability histogram is symmetric around 27.5, we know that its mean is 27.5. We could also obtain this answer by using Equation 7.2. As we found on page 148, the standard deviation of this probability histogram is 4.787. It seems sensible that we should use the Normal curve with $\mu = 27.5$ and $\sigma = 4.787$; i.e., the Normal curve with the same center and spread as the probability histogram.

I am almost ready to show you the fancy math approximation to $P(R_1 \geq 35)$. Look at the shaded rectangles in Figure 7.5. The left boundary of these rectangles is at 34.5 and the right boundary is at 40.5. The approximation I advocate is to find the area to the right of 34.5 under the Normal curve with $\mu = 27.5$ and $\sigma = 4.787$. There are two things to note about my approximation:
Table 7.4: The sampling distribution of $R_1$ for the 252 possible assignments of a balanced CRD with $n = 10$ units and 10 distinct response values.

<table>
<thead>
<tr>
<th>$r_1$</th>
<th>$P(R_1 = r_1)$</th>
<th>$r_1$</th>
<th>$P(R_1 = r_1)$</th>
<th>$r_1$</th>
<th>$P(R_1 = r_1)$</th>
<th>$r_1$</th>
<th>$P(R_1 = r_1)$</th>
<th>$r_1$</th>
<th>$P(R_1 = r_1)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>1/252</td>
<td>20</td>
<td>7/252</td>
<td>25</td>
<td>18/252</td>
<td>31</td>
<td>16/252</td>
<td>36</td>
<td>5/252</td>
</tr>
<tr>
<td>16</td>
<td>1/252</td>
<td>21</td>
<td>9/252</td>
<td>26</td>
<td>19/252</td>
<td>32</td>
<td>14/252</td>
<td>37</td>
<td>3/252</td>
</tr>
<tr>
<td>17</td>
<td>2/252</td>
<td>22</td>
<td>11/252</td>
<td>27</td>
<td>20/252</td>
<td>33</td>
<td>11/252</td>
<td>38</td>
<td>2/252</td>
</tr>
<tr>
<td>18</td>
<td>3/252</td>
<td>23</td>
<td>14/252</td>
<td>28</td>
<td>20/252</td>
<td>34</td>
<td>9/252</td>
<td>39</td>
<td>1/252</td>
</tr>
<tr>
<td>19</td>
<td>5/252</td>
<td>24</td>
<td>16/252</td>
<td>29</td>
<td>19/252</td>
<td>35</td>
<td>7/252</td>
<td>40</td>
<td>1/252</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>100%</td>
</tr>
</tbody>
</table>

- We begin at the value 34.5, not 35. This adjustment is called a **continuity correction**.
- Instead of ending the approximation at the value 40.5 (the right extreme of the rectangles), we take the approximation all the way to the right. As Buzz Lightyear from Toy Story would say, “To infinity and beyond.” Well, statisticians don’t attempt to go beyond infinity.

The area I seek: the area to the right of 34.5 under the Normal curve with $\mu = 27.5$ and $\sigma = 4.787$ is 0.0718. When you read Section 7.4 you will learn how to obtain this area. I don’t show you now because I don’t want to interrupt the flow of the narrative.

Is this approximation any good? Well, we can answer this question because it is possible to compute the exact $P(R_1 \geq 35)$ from Table 7.4. Reading from this table, we find:

$$P(R_1 \geq 35) = (7 + 5 + 3 + 2 + 1 + 1)/252 = 19/252 = 0.0754.$$ 

Let me make three comments about this approximation.

1. The approximation (0.0718) is pretty good; it is smaller than the exact probability (0.0754) by 0.0036. Math theory tells us that this approximation tends to get better as the total number of trials, $n$, becomes larger. This is a remarkably good approximation for an $n$ that is so small that we don’t really need an approximation (because it was ‘easy’ to find the exact probability).

2. This example shows that the continuity correction is very important. If we find the area under the Normal curve to the right of 35.0—i.e., if we do not adjust to 34.5—the area is 0.0586, a pretty bad approximation of 0.0754.

3. If we find the area between 34.5 and 40.5—the region suggested by the rectangles, instead of simply the area to the right of 34.5 which I advocate—the answer is 0.0685; which is a worse approximation than the one I advocate.

We have spent a great deal of effort looking at $P(R_1 \geq 35)$. Table 7.5 looks at several other examples. Again, after you read Section 7.4 you will be able to verify the approximations in this table, if you desire to do so.
There is a huge amount of information in this table. I don’t want you to stare at it for 15 minutes trying to absorb everything in it! Instead, please note the following features.

1. If you scan down the ‘Exact’ and ‘Normal’ columns, you see that the numbers side-by-side are reasonably close to each other. Thus, the approximation is never terrible.

2. Interestingly, the best approximation is the value $0.0473$ which differs from the exact answer, $0.0476$, by only $0.0003$. We know from the definition of statistical significance in Chapter 5 that statisticians are particularly interested in P-values that are close to 0.05. Thus, in a sense, for this example the approximation is best when we especially want it to be good.

3. This table—and a bit additional work—illustrates how difficult it is to compare different methods of approximating a probability. I am a big fan of the continuity correction and want you to always use it. Intellectual honesty, however, requires me to note the following. From our table we see that for the event $(R_1 \geq 40)$ the exact probability is $1/252 = 0.0040$. This is considerably smaller (in terms of ratio) than the approximate probability of $0.0061$. For this event, we actually get a better approximation if we don’t use the continuity correction: the area under the Normal curve to the right of 40 is $0.0045$. This is a much better answer than the one in the table.

This result reflects a general pattern. In the extremes of sampling distributions, the Normal curve approximation often is better without the continuity correction. The continuity correction, however, tends to give much better approximations for probabilities in the neighborhood of 0.05. Because statisticians care so much about 0.05 and its neighbors, I advocate using the continuity correction. But I cannot say that it is always better to use the continuity correction, because sometimes it is not.

I have two final extended comments before I leave this section.

1. I have focused on approximating ‘≥’ probabilities, such as $P(R_1 \geq 35)$. In these situations we obtain the continuity correction by subtracting 0.5 from the value of interest, in the previous sentence, 35. The thing to remember is that our motivation for subtracting 0.5
Table 7.5: Several Normal curve approximations for the sampling distribution in Table 7.4.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$R_1 \geq 29$</td>
<td>0.4206</td>
<td>0.4173</td>
<td>$R_1 \geq 33$</td>
<td>0.1548</td>
<td>0.1481</td>
<td>$R_1 \geq 37$</td>
<td>0.0278</td>
<td>0.0300</td>
</tr>
<tr>
<td>$R_1 \geq 30$</td>
<td>0.3452</td>
<td>0.3380</td>
<td>$R_1 \geq 34$</td>
<td>0.1111</td>
<td>0.1050</td>
<td>$R_1 \geq 38$</td>
<td>0.0159</td>
<td>0.0184</td>
</tr>
<tr>
<td>$R_1 \geq 31$</td>
<td>0.2738</td>
<td>0.2654</td>
<td>$R_1 \geq 35$</td>
<td>0.0754</td>
<td>0.0718</td>
<td>$R_1 \geq 39$</td>
<td>0.0079</td>
<td>0.0108</td>
</tr>
<tr>
<td>$R_1 \geq 32$</td>
<td>0.2103</td>
<td>0.2017</td>
<td>$R_1 \geq 36$</td>
<td>0.0476</td>
<td>0.0473</td>
<td>$R_1 \geq 40$</td>
<td>0.0040</td>
<td>0.0061</td>
</tr>
</tbody>
</table>

comes from the probability histogram in Figure 7.5. We see that the rectangle centered at 35 has a left boundary of 34.5. Thus, we subtract 0.5 to move from 35 to 34.5 and include all of the rectangle. Note that we don’t need to actually draw the probability histogram; from its definition we know that $\delta = 1$ is the base of the rectangle. Hence, to move from the center, 35, to its left boundary, we move one-half of $\delta$ to the left.

Suppose, however, that we want to find an approximate P-value for the alternative $< \neq$. In either of these cases we would need to deal with a ‘$\leq$’ probability, say $P(R_1 \leq 21)$. Now we want the rectangle centered at 21 and all rectangles to its left. Thus, the continuity correction changes 21 to 21.5.

You have a choice as to how to remember these facts about continuity corrections: you may memorize that ‘$\geq$’ means subtract and that ‘$\leq$’ means add. Personally, I think it is better to think of the picture and deduce the direction by reasoning.

2. In many, but not all, of our Normal curve approximations in these Course Notes, $\delta = 1$ and the size of the continuity correction is 0.5. But as we saw above, if there are ties in our combined data set, then for $R_1$ we could have $\delta = 0.5$ or we could have $\delta = 1$. In addition, when there are ties, the probability histogram has gaps and it’s not clear (trust me on this!) how, if at all, gaps should affect a continuity correction. For simplicity, I declare that whenever we want to use a Normal curve to approximate the sampling distribution of $R_1$, we will always adjust by 0.5, whether $\delta$ equals 1 or 0.5.

7.4 Computing

7.4.1 Areas Under any Normal Curve

My objective in this subsection is to introduce you to a website that can be used to obtain areas under any Normal curve.

Go the website:

http://davidmlane.com/hyperstat/z_table.html

Click on this site, please, and I will walk you through its use.
As you visually scan down the page, the first thing you will see is a graph of the N(0,1) curve. (I know this is the standard Normal curve—N(0,1)—because a bit farther down the page I see the entries 0 for Mean and 1 for SD.) For now, ignore the shaded region under the graph.

Immediately below the graph are two circles that let you choose the type of problem you wish to solve. The options are:

- Area from a value; and
- Value from an area.

The site’s default is the first of these options. Do not change the default. (We will return to the second option in Chapter 9.)

Next, the site allows you to specify the Normal curve of interest to you. The default values are Mean ($\mu$) equal to 0 and SD ($\sigma$) equal to 1. For now, let’s leave the default values unchanged.

Immediately below SD is a vertical display of four circles, labeled: Above, Below, Between and Outside. To the right of each circle is a rectangle—or two—in which you may enter a number—or numbers. The default selects the circle Above with the numerical entry 1.96.

The website author chose the word Above, whereas I would prefer either Greater than or To the right of. Similarly, instead of Below I would prefer Less than or To the left of. Just so you don’t think I am a Mr. Negative, I can’t think of any way to improve upon either of the labels Between or Outside.

Now, move your eyes up a bit to the picture of the Standard Normal curve. You will note that the area Above 1.96 (when in Rome . . .) is shaded black. Finally, below the display of circles and rectangles you will find the display: Results: Area (probability) 0.025. The norm in Statistics is to report this area rounded to four digits; for the current problem the area is 0.0250, and the site drops the trailing 0. (I would include it, but it’s not my site!)

Let’s do four quick examples of the use of this site.

1. In the rectangle next to Above, replace the value 1.96 by 2.34 and click on the Recalculate box, which is located immediately below the word Area. I did this and obtained 0.0096 as the answer from the site. Note that the site is behaving as I said; the answer is reported to four digits after the decimal point. Also, note that the area to the right of 2.34 is shaded in the site’s graph.

2. I enter 3.89 in the rectangle next to Above and click on Recalculate; the answer is 0.0001.

3. I enter 3.90 in the rectangle next to Above and click on Recalculate; the answer is 0. I don’t like this answer. This is not zero as in, “My Detroit Lions have won zero Super Bowls.” This 0 means that the area, rounded to the fourth digit after the decimal point (nearest ten-thousandth, if you prefer), is 0.0000. If this were my site, it would definitely report the answer as 0.0000.

4. Finally, click on the circle next to Below; place $-0.53$ in its box and click on Recalculate. You should obtain the answer 0.2981.
Next, I will use this site to obtain the Normal Approximation entries in Table 7.5. As explained earlier, I want $\mu = 27.5$ and $\sigma = 4.787$; thus, I enter these numbers in Mean and SD, respectively.

Let’s begin with $r_1 = 29$. I remember the continuity correction, which tells me to replace 29 by 28.5. I select the circle for Above and type 28.5 into its rectangle. I click on Recalculate and obtain the answer 0.4173, as reported in Table 7.5.

Here is one more example. For $r_1 = 33$ I entered 32.5 in the rectangle next to Above. The site responds with 0.1481, as reported in Table 7.5.

### 7.4.2 Using a Website to Perform the Sum of Ranks Test

In Chapter 6 you learned how to obtain the exact P-value for the sum of ranks test provided I give you the exact sampling distribution for $R_1$. In the absence of the exact sampling distribution, you learned how to use the vassarstats website to obtain an approximate P-value via a computer simulation. The computer simulation gives us two P-values: for the two-sided alternative ($\neq$) and for the one-sided alternative that is supported by the data: $> [\leq]$ if $r_1 > \mu$ [$r_1 < \mu$], with $\mu$ given by Equation 7.2. Also, if $r_1 = \mu$, then the website should not be used because: we know that the exact P-value for the two-sided alternative must be one and the website is wrong for the one-sided alternative.

The vassarstats site also gives two P-values based on the Normal curve approximation of this chapter. First, go to

http://vassarstats.net

About one-half way down the list of topics in the left margin, click on Ordinal Data. This action takes you to a page that offers you several options; click on the second option, the Mann-Whitney Test. (Recall that Mann Whitney (Wilcoxin) is the official name of our sum of ranks test.)

The site asks you to enter your values of $n_1$ and $n_2$, which the site call $n_a$ and $n_b$, respectively. **Warning:** The site says that if $n_a \neq n_b$, then you must have $n_a > n_b$; thus, you might need to relabel your treatments in order to use the site. Remember: If you relabel your treatments, you need to reverse any one-sided alternative.

After entering your sample sizes, vassarstats is ready to receive your data. A nice feature of the Mann-Whitney Test site is that you can cut-and-paste your data. If you import your data—i.e., you cut-and-paste—then you must click on Import data to data cells or the site won’t work. The effect of this clicking is that the site will move your data to the Raw Data for portion of the page.

Next, you click on Calculate from Raw Data. The site gives us some useful information and a lot of information that you should ignore. We are given the mean ranks, from which we could obtain the value $r_1$ for our data. The site provides something called $U_A$ which is a function of $r_1$ and I recommend that you ignore it. The site gives two proportions, labeled $P_{(1)}$ and $P_{(2)}$. The first of these is the approximate P-value for the one-sided alternative that is supported by the data. The second of these is the approximate P-value for the two-sided alternative. Both approximations are based on using the Normal curve approximation of this chapter. As best I can tell, the site uses the continuity correction with an adjustment of 0.5.

**Warning:** Whether or not there are ties in the data set, this Mann-Whitney Test site calculates the variance of $R_1$ using the no ties formula, Equation 7.4. For numerical data with a few ties (I
know, this is vague), this variance oversight by vassarstats does not seriously affect the quality of the approximation. (More on this topic in the Practice Problems.) Note, however, that this site should not be used for ordinal categorical data. As we have seen, with ordinal categorical data the value of the variance is changed substantially by the presence of (many) ties.

7.4.3 Reading Minitab Output; Comparing Approximation Methods

This is a new topic for the Computing sections of these Course Notes. I have previously told you that I use Minitab to perform what I call full-blown simulation experiments. In those situations I took the Minitab output—which isn’t pretty—and presented it to you in a more understandable format. This subsection, however, is different.

Several times in these notes I will show you the output of a Minitab statistical analysis without showing you how to create such output. In this subsection I will show you the output that Minitab creates for our sum of ranks test. Finally, this subsection ends with a comparison of methods for Dawn’s study of her cat, Sara’s study of golf and the data from an artificial study of a serious disease presented in Table 6.8 on page 129.

Example 7.3 (Dawn’s study of her cat Bob.) Recall that Dawn’s sorted data are presented in Table 1.3 on page 7. I will find the Normal curve approximate P-value for the sum of ranks test and the alternative > for Dawn’s data. I will do this three ways: using Minitab; using the vassarstats website; and by hand. These three methods better give the same answer!

I entered these data into Minitab and ran the Mann command on it. The edited output is below. (I eliminated output that we don’t use.)

Mann-Whitney Test: C1, C2

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>10</td>
<td>5.500</td>
</tr>
<tr>
<td>C2</td>
<td>10</td>
<td>3.000</td>
</tr>
</tbody>
</table>

W = 133.5

Test of \( = \) vs \( > \) is significant at 0.0171
The test is significant at 0.0163 (adjusted for ties)

Let me walk you through this output, although it is fairly self-explanatory.

- Minitab reminds me that I have placed my data into Minitab columns C1 and C2.
- Minitab reminds me that the number of trials on both treatments is 10; and that the median of the chicken [tuna] data is 5.5 [3.0].
- Minitab tells me that \( r_1 = 133.5 \), but calls it \( W \).
- I told Minitab that I wanted the alternative >. Minitab gives me two Normal curve approximation P-values: the better one—adjusting for ties—0.0163; and the inferior one—no adjustment—0.0171. Even though it seems to me that there are a lot of ties in the data—the values of \( t_i \)’s are three 1’s, two 2’s, three 3’s and a 4—the presence of ties has very little influence on the P-value.
I ran these data through the vassarstats website and it gave me 0.017 for the approximate P-value based on the Normal curve; i.e., the same answer—to three digits—as the Minitab answer that does not adjust for ties.

Finally, by hand, I obtained $\mu = 105$ and $\sigma = 13.1089$, adjusted for ties. As noted above, Dawn’s data give $r_1 = 133.5$. I went to

http://davidmlane.com/hyperstat/z_table.html
to find my approximate P-value. I entered the above values of $\mu$ and $\sigma$ and asked for the area above $133.5 - 0.5 = 133$ (remember the continuity correction). The site gave me 0.0163, which agrees with Minitab’s adjusted answer.

In Example 6.3 on page 126, I use a simulation experiment with 10,000 runs to obtain an approximate P-value for $R_1$ and the alternative $>$. I obtained the value 0.0127 which gives a nearly certain interval of $[0.0093, 0.0161]$. (Details not shown.) The Normal approximation—adjusted for ties—P-value, 0.0163, is very close to being in this interval. Thus, I conclude that the Normal approximation—adjusted for ties—is a pretty good approximation for Dawn’s data, although it appears to be a bit too large.

**Example 7.4 (Sara’s study of golf.)** Recall that Sara’s sorted data are presented in Table 2.2 on page 29. I entered Sara’s data into Minitab and ran the Mann command on it with the alternative $>$. The edited output is below.

Mann–Whitney Test: C1, C2

<table>
<thead>
<tr>
<th>C1</th>
<th>N = 40</th>
<th>Median = 112.00</th>
</tr>
</thead>
<tbody>
<tr>
<td>C2</td>
<td>N = 40</td>
<td>Median = 99.50</td>
</tr>
</tbody>
</table>

$W = 1816.0$

Test of $= vs >$ is significant at 0.0300

The test is significant at 0.0299 (adjusted for ties)

The sample sizes and treatment medians agree with our earlier work. We see that adjusting for ties has only an unimportant impact on our approximate P-value. Thus, the vassarstats site would have been fine for analyzing these data.

If you refer to Example 6.4 in Chapter 7, you will recall that our approximate P-value for $>$ based on a 10,000 run simulation experiment is 0.0293, with nearly certain interval:

$$0.0293 \pm 3\sqrt{\frac{0.0293(0.9707)}{10,000}} = 0.0293 \pm 0.0051 = [0.0242, 0.0344].$$

I know that I can trust computer simulations. Thus, I conclude that the Normal curve approximation—0.0293—appears to be excellent for this data set.

**Example 7.5 (An artificial study of a serious disease.)** Example 6.5 on page 128 introduced an artificial study with an ordinal categorical response. Its data are in Table 6.8 on page 129.

I entered these data into Minitab and ran the Mann command on it with the alternative $<$. The edited output is below.
Mann-Whitney Test: C4, C5

C4  N = 50  Median = 2.0000
C5  N = 50  Median = 2.0000
W = 2340.0
Test of = vs < is significant at 0.1017
The test is significant at 0.0869 (adjusted for ties)

As discussed earlier in these notes, computing medians for ordinal categorical data often is nearly worthless, as it is for these data. We see that it is important to use the correct—adjusted for ties—formula for the variance because the two P-values are quite different.

Finally, Example 6.5 reported the results of a simulation experiment with 10,000 runs. It gave 0.0946 as the approximate P-value for the alternative <. Its nearly certain interval is

\[0.0946 \pm 3\sqrt{\frac{0.0946(0.9054)}{10,000}} = 0.0946 \pm 0.0088 = [0.0858, 0.1034].\]

The Normal approximation P-value, 0.0869, lies in this interval, but just barely.

In summary, the Normal curve approximation gives reasonably accurate P-values for each of these three studies. If one has access to Minitab, the analysis is quite easy to obtain. For all but ordinal categorical data, the vassarstats website usually will give an accurate approximate P-value.
7.5 Summary

A probability histogram is our picture of a sampling distribution. So far in these notes, we have had sampling distributions for our two test statistics: $U$ and $R_1$. For more generality, we talk about a random variable $X$ with observed value denoted by $x$. The random variable $X$ has a sampling distribution and we obtain its probability histogram by following the three steps given on page 143.

We next turn to the question of finding an approximation of a probability histogram (which also provides an approximation of the sampling distribution represented by the probability histogram). Earlier in these Course Notes we learned how to obtain an approximation of a sampling distribution by using a computer simulation experiment. In the current chapter, we learn our first example of what I call a fancy math approximation. To that end, we need to summarize a probability histogram (equivalently, a sampling distribution) by determining its center (mean) and spread (standard deviation).

There is a simple formula for the mean of the sampling distribution of $R_1$ (Equation 7.2):

$$
\mu = n_1(n + 1)/2,
$$

and two formulas for its variance. The first formula (Equation 7.4) is appropriate if the $n$ observations in the combined data set are all distinct numbers (called the no-ties situation):

$$
\sigma^2 = \frac{n_1 n_2 (n + 1)}{12}.
$$

The second formula (Equation 7.5) should be used whenever there are ties in the combined data set:

$$
\sigma^2 = \frac{n_1 n_2 (n + 1)}{12} - \frac{n_1 n_2 \sum (t_i^3 - t_i)}{12n(n - 1)}.
$$

If you don’t recall the meaning of the $t_i$’s in this formula, review the material in the text of this chapter immediately following Equation 7.5 on page 148.

Next, we learned about the family of Normal curves, pictured in Figure 7.4. A particular Normal curve is characterized by its center of gravity (mean) $\mu$ and its spread (standard deviation) $\sigma$; it is sometimes denoted as the $N(\mu, \sigma)$ curve. The Standard Normal curve corresponds to $\mu = 0$ and $\sigma = 1$.

There is a website that calculates areas under Normal curves and its use is demonstrated.

The main idea of this chapter is to use a Normal curve as an approximation of the probability histogram of $R_1$. We use the Normal curve that matches $R_1$ on its values of $\mu$ and $\sigma$ (whichever version—ties or no ties—of $\sigma$ is appropriate).

We always use the continuity correction when we obtain the Normal approximation to $R_1$. If we want to approximate $P(R_1 \geq r_1)$, we calculate the area under the Normal curve to the right of $(r_1 - 0.5)$. If we want to approximate $P(R_1 \leq r_1)$, we calculate the area under the Normal curve to the left of $(r_1 + 0.5)$.

Section 7.4 shows that the vassarstats website for the Mann-Whitney test will calculate the Normal approximation P-value for us. The site use the no-ties formula for the variance, which works reasonably well for a small number of ties. For a large number of ties—in particular, for an ordinal categorical response—the Mann-Whitney test site yields P-values that are a bit too large.

Also, in Section 7.4, we learned how to read Minitab output for the sum of ranks test.
7.6 Practice Problems

1. I have a balanced CRD with a total of six units. My six ranks are: 1, 2, 4, 4, 4 and 6.

   (a) Find six response values that would yield these ranks. Note that because ranks are obtained for the combined data set, you don’t need to match observations to treatments to answer this question. Also note that there are an infinite number of correct answers.

   (b) Determine the exact sampling distribution of \( R_1 \) for these six ranks.

   (c) Draw the probability histogram of the sampling distribution of \( R_1 \) for these six ranks.

   (d) Show that the \( \mu = 10.5 \) with two different arguments.

   (e) Compute \( \sigma^2 \) and \( \sigma \).

2. Refer to Practice Problem 3 on page 134 in Chapter 6; it presents a balanced CRD with \( n = 50 \) units and a four-category ordinal response.

   Using the same assignment of numbers to categories I used in Chapter 6, I entered my data into Minitab and directed Minitab to perform the sum of ranks test for the alternative \(<\). I executed the Mann command and obtained the output below.

   **Mann-Whitney Test: C1, C2**

<table>
<thead>
<tr>
<th></th>
<th>N = 25</th>
<th>Median =</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>2.000</td>
<td></td>
</tr>
<tr>
<td>C2</td>
<td>3.000</td>
<td></td>
</tr>
</tbody>
</table>

   \( W = 548.5 \)

   Test of \( = \) vs \( < \) is significant at 0.0430

   The test is significant at 0.0380 (adjusted for ties)

   In the Chapter 6 Practice Problems I reported that a 10,000 run computer simulation experiment gave 0.0354 as an approximate \( P \)-value for the alternative \(<\). Use this relative frequency to obtain the nearly certain interval (Formula 4.1 in Chapter 4) for the exact \( P \)-value. Comment on the Minitab approximate \( P \)-values.

3. Imagine that we have a CRD with \( n_1 = 15 \) units assigned to the first treatment and \( n_2 = 20 \) units assigned to the second treatment. Also, assume that there are no ties in the combined list of 35 response values.

   (a) Calculate the mean, variance and standard deviation of the sampling distribution of \( R_1 \).

   (b) Assume that the alternative is \( >\). Use the Normal curve website to find the approximate \( P \)-values for the following actual values of \( r_1 \): 300, 330 and 370. Remember to use the continuity correction.

   (c) Assume that the alternative is \( \neq \). Use the Normal curve website to find the approximate \( P \)-values for the following actual values of \( r_1 \): 300, 330 and 370. Remember to use the continuity correction.
(d) Assume that the alternative is <. Use the Normal curve website to find the approximate P-values for the following actual values of \( r_1 \): 250, 220 and 180. Remember to use the continuity correction.

7.7 Solutions to Practice Problems

1. (a) There is an infinite number of possible response values. I will proceed as follows. The three ranks equal to 4 mean that there are three tied values; call them 15, 15 and 15. There must be one observation larger than 15, which will have the rank 6, and two non-tied observations smaller than 15 to receive the ranks 1 and 2. I select observation values of 11, 14 and 17. Thus, my combined data set consists of 11, 14, 15, 15, 15 and 17.

(b) There is one assignment that puts all three 4’s on treatment 1, giving \( r_1 = 12 \).

There are three assignments that put two 4’s with one non-4 on treatment 1. The two 4’s can be matched with 1, 2 or 6. Thus, this gives us nine more values of \( r_1 \): three each of 9, 10 and 14.

There are three assignments that put one 4 with two non-4’s on treatment 1. The one 4 can be matched with: 1 and 2; 1 and 6; or 2 and 6. Thus, this gives us nine more values of \( r_1 \): three each of 7, 11 and 12.

Finally, there is one assignment that places all 4’s on treatment 2. This assignment gives \( r_1 = 1 + 2 + 6 = 9 \).

Combining the above, we see that three assignments each give the following values for \( r_1 \): 7, 10, 11 and 14. Also, four assignments each give the following values for \( r_1 \): 9 and 12. Thus,

\[
\begin{align*}
\cdot P(R_1 = r_1) &= 0.15 \text{ for } r_1 = 7, 10, 11, 14. \\
\cdot P(R_1 = r_1) &= 0.20 \text{ for } r_1 = 9, 12.
\end{align*}
\]

(c) The probability histogram is in Figure 7.6. I will explain its derivation.

The possible values of \( r_1 \) are 7, 9, 10, 11, 12 and 14. They are marked and labeled in my picture. The minimum distance between consecutive values is 1. Because \( \delta = 1 \), the height of each rectangle equals the probability of its center’s location; these probabilities are given in the answer above and presented in the probability histogram.

(d) First, the probability histogram in Figure 7.6 is symmetric. Thus, its mean equals its point of symmetry, 10.5.

Second, we can use Equation 7.2 on page 147

\[
\mu = n_1(n + 1)/2 = 3(6 + 1)/2 = 21/2 = 10.5.
\]

(e) Because there are ties in the data set, we use Equation 7.5

\[
\sigma^2 = \frac{n_1n_2(n + 1)}{12} - \frac{n_1n_2 \sum (t_i^3 - t_i)}{12n(n - 1)}.
\]
I will evaluate these terms individually and then subtract. The first term is:

\[
\frac{n_1n_2(n + 1)}{12} = \frac{3(3)(7)}{12} = 63/12 = 5.25.
\]

For the second term, first note that only one \(t_i\) is larger than 1; it is 3. The value of the second term is:

\[
\frac{n_1n_2 \sum(t_i^3 - t_i)}{12n(n - 1)} = \frac{3(3)(3^3 - 3)}{12(6)(5)} = 216/360 = 0.60.
\]

Thus,

\[
\sigma^2 = 5.25 - 0.60 = 4.65 \quad \text{and} \quad \sigma = \sqrt{4.65} = 2.1564.
\]

2. For \(\hat{r} = 0.0354\) and 10,000 runs, the nearly certain interval is:

\[
0.0354 \pm 3\sqrt{\frac{0.0354(0.9646)}{10000}} = 0.0354 \pm 0.0055 = [0.0299, 0.0409].
\]

Minitab gives us two approximate P-values. Whenever our data set has ties, we should use the adjusted (smaller) P-value. For some data sets, the two P-values are nearly identical, but not in this case. Thus, our Normal approximation is 0.0380. This value falls within our nearly certain interval; thus, the Normal approximation seems to be reasonable.

3. (a) From Equation [7.2]

\[
\mu = n_1(n + 1)/2 = 15(36)/2 = 270.
\]

From Equation [7.4]

\[
\sigma^2 = \frac{n_1n_2(n + 1)}{12} = \frac{15(20)(36)}{12} = 900 \quad \text{and} \quad \sigma = \sqrt{900} = 30.
\]

(b) We go to the website

http://davidmlane.com/hyperstat/z_table.html

Following the instructions given in Section [7.4] we enter 270 in the Mean box and 30 in the SD box. We select the option Above and enter 299.5 in its box. We obtain the area 0.1627; this is the approximate P-value for \(r_1 = 300\).

For the other values of \(r_1\) we repeat the above, first typing 329.5 in the box and then 369.5. For 329.5, we get the area 0.02367, which is the approximate P-value for \(r_1 = 330\). Finally, for 369.5, we get the area 0.0005, which is the approximate P-value for \(r_1 = 370\).

(c) Here is a useful trick. The approximating curve (Normal) is symmetric. As a result, provided the one-sided approximate P-value does not exceed 0.5000, if you simply double it you obtain the approximate P-value for \(\neq\). Hence, our approximate P-values are:
Figure 7.6: The probability histogram for the sampling distribution in Practice Problem 1.

- For $r_1 = 300$: $2(0.1627) = 0.3254$.
- For $r_1 = 330$: $2(0.0237) = 0.0474$.
- For $r_1 = 370$: $2(0.0005) = 0.0010$.

(d) We select the option Below and enter 250.5 in its box. We obtain the area 0.2578; this is the approximate P-value for $r_1 = 250$.

For the other values of $r_1$ we repeat the above, first typing 220.5 in the box and then 180.5. For 220.5, we get the area 0.0495, which is the approximate P-value $r_1 = 220$. Finally, for 180.5, we get the area 0.0014, which is the approximate P-value for $r_1 = 180$. 
7.8 Homework Problems

1. An unbalanced CRD has \( n_1 = 2, n_2 = 3 \) and \( n = 5 \). Thus, there are 10 possible assignments. (You don’t need to verify this fact.) The sorted numbers in the combined data set are:

   \[ 3, 8, 8, 8 \text{ and } 20. \]

   (a) Determine the ranks for the combined data set.

   (b) Complete the following table, the sampling distribution for \( R_1 \).

   \[
   \begin{array}{c|ccc}
   r_1: & 4 & 6 & 8 \\
   P(R_1 = r_1): & & & \\
   \end{array}
   \]

   (c) Draw the probability histogram for your table in (b).

   (d) Calculate the variance and standard deviation of the sampling distribution for \( R_1 \).

2. Refer to Homework Problem 3 on page 139 an unbalanced CRD with an ordinal response. Its data are reproduced below.

   \[
   \begin{array}{c|ccc|c}
   \text{Treatment} & \text{Disagree} & \text{Neutral} & \text{Agree} & \text{Total} \\
   1 & 8 & 7 & 5 & 20 \\
   2 & 3 & 5 & 7 & 15 \\
   \text{Total} & 11 & 12 & 12 & 35 \\
   \end{array}
   \]

   I assign numbers 1 (Disagree), 2 (Neutral) and 3 (Agree) to these categories. I placed these data in Minitab and executed the Mann command. My output is below.

   Mann-Whitney Test: C1, C2

   \[
   \begin{array}{l}
   \text{C1} \quad N = 20 \quad \text{Median} = \quad 2.000 \\
   \text{C2} \quad N = 15 \quad \text{Median} = \quad 2.000 \\
   \text{W} = 318.0 \\
   \text{Test of } = \text{ vs not } = \text{ is significant at 0.1666} \\
   \text{The test is significant at 0.1424 (adjusted for ties)}
   \end{array}
   \]

   Use this output to answer the questions below.

   (a) What is the observed value of \( R_1 \)?

   (b) Which alternative (> , < or ≠) did Minitab use?

   (c) Which P-value is better to use? Why?

3. Recall Reggie’s study of darts, first introduced in Homework Problems 5–7 in Chapter 1 (Section 1.8).
(a) Use the vassarstats website’s Mann-Whitney command to obtain the Normal approximation to the P-value for >. (Using Minitab, I found that ignoring ties, as vassarstats website’s Mann-Whitney does, has no impact on the approximation. Well, it changes the answer by 0.0001, which I am willing to ignore.)

(b) I performed a simulation with 10,000 runs and obtained 0.0075 for my approximate P-value for the alternative >. Compute the nearly certain interval (Formula 4.1) for the exact P-value for the alternative >.

(c) Compare your approximations from (a) and (b); comment.

4. Consider an unbalanced CRD with \( n_1 = 36 \) and \( n_2 = 27 \). Assume that the combined data set contains no ties.

(a) Calculate the mean and standard deviation of the sampling distribution of \( R_1 \).

(b) Use your answers from (a) and the website

http://davidmlane.com/hyperstat/z_table.html

to obtain the approximate P-values for the alternative > and each of the following values of \( r_1 \): 1200, 1250 and 1300.

(c) Use your answers from (a) and the website

http://davidmlane.com/hyperstat/z_table.html

to obtain the approximate P-values for the alternative < and each of the following values of \( r_1 \): 1120 and 1070.

(d) Use your answers from (a) and the website

http://davidmlane.com/hyperstat/z_table.html

to obtain the approximate P-values for the alternative \( \neq \) and each of the following values of \( r_1 \): 1300 and 1070.