Chapter 6

The Sum of Ranks Test

Thus far in these Course Notes we have considered CRDs with a numerical response. In Chapter 5 we learned how to perform a statistical test of hypotheses to investigate whether the Skeptic’s Argument is correct. Every test of hypotheses has a test statistic; in Chapter 5 we chose the test statistic $U$ which has observed value $u = \bar{x} - \bar{y}$. For rather obvious reasons, this test using $U$ is referred to a test of means or a test of comparing means.

We learned in Chapter 1 that the mean is a popular way to summarize a list of numbers. Thus, it is not surprising to learn that comparing means, by subtraction, is a popular way to compare two treatments and, hence, the test of Chapter 5 seems sensible. But we also learned in Chapter 1 that the median is another popular way to summarize a list of numbers. Thus, you might guess that another popular choice for a test statistic would be the one whose observed value is $v = \tilde{x} - \tilde{y}$. If you make this guess, you would be wrong, but close to the truth.

Recall from Chapter 1 that the distinction between the mean and the median can be viewed as the distinction between focusing on arithmetic versus position. The median, recall, is the number at the center position of a sorted list—for an odd sample size—or the average of the values at the two center positions—for an even sample size. Thus, the value $v$ in the previous paragraph compares two sorted lists by comparing the numbers in their center positions. This comparison ignores a great deal of information! In those situations in which, for whatever reasons, we prefer to focus on positions rather than arithmetic, it turns out that using ranks, defined below, is superior to using medians in order to compare two sets of numbers.

In this chapter we will consider an option to using $U$: i.e., we will present a test that compares the two sets of data by comparing their ranks. When we study power in a later chapter, we will see that sometimes the test that compares ranks is better than the test that compares means. The last section of this chapter presents an additional advantage of using a test based on ranks; it can be used when the response is ordinal, but not numerical.

6.1 Ranks

We begin by doing something that seems quite odd: We combine the data from the two treatments into one set of data and then we sort the $n = n_1 + n_2$ response values. For example, for Dawn’s
Table 6.1: Dawn’s 20 sorted response values, with ranks.

<table>
<thead>
<tr>
<th>Position</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Response</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Rank</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>5</td>
<td>7.5</td>
<td>7.5</td>
<td>7.5</td>
<td>7.5</td>
<td>10.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Position</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Response</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>7</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>Rank</td>
<td>10.5</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>18.5</td>
<td>18.5</td>
<td>20</td>
</tr>
</tbody>
</table>

study of her cat, the 20 sorted response values are given in Table 6.1. You can verify these numbers from the data presented in Table 1.3 on page 7, but I recommend that you just trust me on this. We note that Dawn’s 20 numbers consist of nine distinct values. Her number of distinct values is smaller than 20 because several of the responses are tied; for example, four responses are tied with the value 3. Going back to Chapter 1, we talk about the 20 positions in the list in Table 6.1. As examples: position 1 has the response 0; position 20 has the response 8; and positions 6–9 all have the response 3.

If the \( n \) numbers in our list are all distinct, then the rank of each response is its position. This is referred to as the no-ties situation and it makes all of the computations below much simpler. Sadly, in practice, data with ties are commonplace. Whenever there are ties, all tied responses receive the same rank, which is equal to the mean of their positions. Thus, for example, all four of the responses equal to 3 receive the rank of 7.5 because they occupy positions 6 through 9 and the mean of 6, 7, 8 and 9 is 7.5. It is tedious to sum these four numbers to find their mean; here is a shortcut that always works: simply compute the mean of the smallest (first) and largest (last) positions in the list. For example, to find the mean of 6, 7, 8 and 9, simply calculate \( (6+9)/2 = 7.5 \). When we consider ordinal data in Section 6.5, we will have occasion to find the mean of 43, 44, \ldots, and 75.

Summing these 33 numbers is much more tedious than simply computing \( (43+75)/2 = 59 \). Finally, for any responses in the list that is not tied with another—responses 0, 2 and 8 in Dawn’s data—its rank equals its position.

The basic idea of our test based on ranks is that we analyze the ranks, not the responses. For example, I have retyped Table 6.1 in Table 6.2 (dropping the two Position rows) with the added feature that the responses from treatment 1 (chicken) and their ranks are in bold face type. For the test statistic \( U \) we performed arithmetic on the responses to obtain the means for each treatment and then we subtracted. We do the same arithmetic now, but we use the ranks instead of the responses. For example, let \( R_1 \) denote the sum of the ranks for treatment 1 and let \( r_1 \) denote its observed value. For Dawn’s data we get:

\[
r_1 = 3 + 7.5 + 10.5 + 13 + 13 + 16 + 16 + 16 + 18.5 + 20 = 133.5.
\]
Table 6.2: Dawn’s 20 sorted responses, with ranks. The responses from treatment 1, and their ranks, are in bold-faced type.

<table>
<thead>
<tr>
<th>Response</th>
<th>Rank:</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 1 1 2 3 3 3 4</td>
<td>1 3 3 5 7.5 7.5 7.5 10.5</td>
</tr>
<tr>
<td>4 5 5 6 6 6 7 8</td>
<td>10.5 13 13 16 16 18.5 18.5 20</td>
</tr>
</tbody>
</table>

Similarly, let $R_2$ denote the sum of the ranks for treatment 2 and let $r_2$ denote its observed value. For Dawn’s data we get:

$$r_2 = 1 + 3 + 3 + 5 + 7.5 + 7.5 + 7.5 + 10.5 + 13 + 18.5 = 76.5.$$ 

In order to compare the treatments’ ranks descriptively, we calculate the mean of the ranks for each treatment:

$$\bar{r}_1 = r_1/n_1 = 133.5/10 = 13.35 \text{ and } \bar{r}_2 = r_2/n_2 = 76.5/10 = 7.65,$$

which show that, based on ranks, the responses on treatment 1 are larger than the responses on treatment 2.

The next obvious step is that we define

$$v = \bar{r}_1 - \bar{r}_2 = r_1/n_1 - r_2/n_2,$$

to be the observed value of the test statistic $V$. I say that this step is obvious because it is analogous to our definition of $u = \bar{x} - \bar{y}$; i.e., $v$ is for ranks what $u$ is for responses. **Except** we don’t do the obvious. Here is why.

As the legend goes, as a child, Carl Friedrich Gauss (1777–1855), discovered that for any positive integer $n$:

$$1 + 2 + 3 + \ldots + n = n(n + 1)/2.$$ 

In our current chapter, this says that the sum of the positions for the combined set of $n$ responses equals $n(n + 1)/2$. Because of the way ranks are defined above, it follows that the sum of the $n$ ranks also equals $n(n + 1)/2$. If this is a bit abstract, note that for $n = 20$ Gauss showed that the sum of the ranks is $20(21)/2 = 210$, which agrees with our findings for Dawn’s data:

$$r_1 + r_2 = 133.5 + 76.5 = 210.$$

As a result, given the values of $n_1$ and $n_2$—which the researcher will always know—knowledge of the value of $r_1$ immediately gives us the value of $v$. (If you like to see such things explicitly, for Dawn’s study:

$$v = r_1/n_1 - (210 - r_1)/n_2.$$ 

Now, I don’t want to spend my time doing messy arithmetic, converting back-and-forth between $v$ and $r_1$. Messy arithmetic is not my point! My point is that we are free to use either $v$ or $r_1$ as the observed value of our test statistic. In these Course Notes, we will use $r_1$ as the observed value of the test statistic $R_1$. There are several advantages to using $r_1 [R_1]$ instead of $v [V]$:
1. The value of $r_1$ is always a positive integer, whereas $v$ will typically be a decimal and can be negative.

2. If computing by hand, one can obtain $r_1$ faster than one can obtain $v$.

3. For given values of $n_1$ and $n_2$, in the old days, it was more elegant to have tables of exact P-values as a function of the positive integer $r_1$ rather than the fraction $v$.

Admittedly, the last two of these advantages are less important in the computer age. On the other hand, positive integers have been our friends since early childhood and we all have early bad memories of decimals and negatives!

There are two main disadvantages of working with $r_1$ rather than $v$:

1. The value of $r_1$ alone does not tell us how the treatments compare descriptively.

2. Because $r_1$ is always positive, when we have symmetry (see below) it is no longer around 0, which makes our rule for the P-value for $\neq$ a bit more difficult to remember.

I want to introduce you to the **Mann Whitney (Wilcoxin) Rank Sum Test**. (The tribute to Wilcoxin, a chemist by training, is often suppressed because there is another test called the Wilcoxin Signed Rank Test.) We will call it the **sum of ranks test**, or, occasionally, the Mann Whitney test. The obvious reason for either name is: the observed value of the test statistic is obtained by summing ranks.

We will not spend a great deal of time in these *Course Notes* on procedures based on ranks. A big problem is interpretation. I can understand what $u = 2.2$ signifies: on average, Bob consumed 2.2 more chicken treats than tuna treats. I do not have a clear idea of how to interpret the difference in mean ranks for Dawn’s data:

$$133.5/10 - 76.5/10 = 13.35 - 7.65 = 5.70.$$ 

For the—admittedly narrow—goal of deciding whether or not to reject the null hypothesis that the Skeptic is correct, the sum of ranks test can be useful.

### 6.2 The Hypotheses for the Sum of Ranks Test

The Skeptic’s Argument is exactly the same as it was earlier for the difference of means test: The treatment is irrelevant; the response to any unit would have remained the same if the other treatment had been applied. The null hypothesis is, as before, that the Skeptic is correct.

In order to visualize the alternative, we need to remember the imaginary **clone-enhanced** study, first introduced on page 87. There is a total of $n = n_1 + n_2$ units being studied. With the clone-enhanced study, each of these units would yield two responses. Thus, the combined data for the clone-enhanced study would consist of $2n$ observations, with $n$ observations from each treatment. Thus, note that even if the CRD is unbalanced the clone-enhanced study is balanced because each unit gives a response to both treatments.
Table 6.3: Case 2 of the clone-enhanced studies of Chapter 5: A constant treatment effect of \( c = 6 \).
The actual data are in bold-faced type.

<table>
<thead>
<tr>
<th>Unit:</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>( \rho_i )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Response on Treatment 1:</td>
<td>18</td>
<td>9</td>
<td>24</td>
<td>18</td>
<td>12</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>Ranks:</td>
<td>10</td>
<td>3.5</td>
<td>12</td>
<td>10</td>
<td>6</td>
<td>8</td>
<td>8.25</td>
</tr>
<tr>
<td>Response on Treatment 2:</td>
<td>12</td>
<td>3</td>
<td>18</td>
<td>12</td>
<td>6</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>Ranks:</td>
<td>6</td>
<td>1</td>
<td>10</td>
<td>6</td>
<td>2</td>
<td>3.5</td>
<td>4.75</td>
</tr>
</tbody>
</table>

This presentation is getting quite abstract; thus, we will look at a numerical example. Table 6.3 reproduces Case 2 of Table 5.3 on page 90. In HS-1, \( n_1 = n_2 = 3 \) giving \( n = 6 \) and \( 2n = 12 \) total responses in the clone-enhanced study. You may verify the ranks given in Table 6.3 for practice. Or not; your choice.

Define \( \rho_1 \) (pronounced ‘roh’) to be the mean of the ranks on treatment 1 in the clone-enhanced study. Similarly, let \( \rho_2 \) be the mean of the ranks on treatment 2 in the clone-enhanced study. If the Skeptic is correct then the two sets of data in the clone-enhanced study will be identical, which implies that \( \rho_1 = \rho_2 \). Although I won’t give you details (and, thus, don’t worry about it) it is possible for the Skeptic to be incorrect and yet \( \rho_1 = \rho_2 \).

For the clone-enhanced data in Table 6.3 we see that \( \rho_1 = 8.25 \) is larger than \( \rho_2 = 4.75 \) which is in the same direction as our earlier computation that for Case 2, \( \mu_1 = 16 \) is larger than \( \mu_2 = 10 \). Thus, sometimes (often?) looking at ranks gives a similar answer to looking at means, but, as we will see below, not always.

For the sum of ranks test, the three options for the alternative are given below:

- \( H_1 : \rho_1 > \rho_2 \).
- \( H_1 : \rho_1 < \rho_2 \).
- \( H_1 : \rho_1 \neq \rho_2 \).

As a practical matter, just remember that \( > [<] \) means that—in terms of ranks—treatment 1 tends to give larger [smaller] responses than treatment 2; and that \( \neq \) means that treatment 1 tends to give either larger or smaller responses than treatment 2.

Earlier in these notes I mentioned that \( \bar{x} [\bar{y}] \) can be viewed as our point estimate of \( \mu_1 [\mu_2] \). The relationships between \( r_1, r_2, \rho_1 \) and \( \rho_2 \) are problematic. All that I can safely say is that

\[
r_1/n_1 > r_2/n_2 \text{ provides evidence that } \rho_1 > \rho_2.
\]

I will defer further discussion of this issue until we examine population-based inference for the sum of ranks test. (For example, performing the AT-1 study would give us the value of \( \mu_1 \). Thus, even though we need to imagine the fanciful clone-enhanced study to obtain both \( \mu_1 \) and \( \mu_2 \), it is always possible to determine one of these with a study. By contrast, with ranks, the AT-1 study tells us nothing about \( \rho_1 \); think about why this is true.)

Let’s stop and take a breath. We have completed Step 1 our sum of ranks test; we have specified the null and alternative hypotheses. We can now move to Steps 2 and 3.
Table 6.4: Cathy’s times, in seconds, to run one mile. HS means she ran at the high school and P means she ran through the park. The responses and ranks of the high school data are in bold-face.

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

6.3 Step 2: The Test Statistic and Its Sampling Distribution

The test statistic will be $R_1$ with observed value $r_1$. We have three options for finding the sampling distribution and then using it to find our P-value. Two of these options are familiar and one is new. The familiar ones are:

1. Determine the exact sampling distribution of $R_1$. In these notes, this option is practical only for very small studies.

2. Use a computer simulation experiment to approximate the sampling distribution of $R_1$.

The third option is called the Normal curve approximation of the sampling distribution of $R_1$; it will be presented in Chapter 7.

I will begin by determining the exact sampling distribution of $R_1$ for Cathy’s study of routes for running. Table 6.4 is a reproduction of our earlier table of Cathy’s data, with ranks now added. Again, make sure you are able to assign ranks correctly. You will need this skill for exams and homework.

Note that if the Skeptic is correct, then trials 1–6 will always yield the responses (ranks) 5, 2, 4, 1, 6, and 3, respectively. We see that the actual observed value of $R_1$ for Cathy’s study is $r_1 = 5 + 2 + 6 = 13$. Although we don’t need it, I note that the actual observed value of $R_2$ is $r_2 = 4 + 1 + 3 = 8$. Thus, the mean of the ranks at the high school ($\bar{r}_1 = 13/3$) is larger than the mean of the ranks through the park ($\bar{r}_2 = 8/3$) because, as a group, Cathy’s times were larger (worse) at the high school.

Table 6.5 is analogous to Table 3.4 on page 63. In the earlier table, we determined the values of $\bar{x}$, $\bar{y}$ and $u = \bar{x} − \bar{y}$ for each of the 20 possible assignments of trials to treatments. In the current chapter, our task is much easier; for each assignment we need calculate only the value of $r_1$. You should make sure that you follow the reasoning behind Table 6.5 (A similar problem is on the homework and might be on an exam.) For example, reading the first row of this table, we see that we are interested in assignment 1, 2, 3. Reading from Table 6.4 we see that the ranks are 5, 2 and 4, giving $r_1 = 5 + 2 + 4 = 11$. The information in Table 6.5 is summarized in Table 6.6 the sampling distribution of $R_1$ for Cathy’s study. Again, make sure you can create this latter table from the former.

In fact, any balanced CRD with $n = 6$ units and no tied responses will yield the sampling distribution for $R_1$ given in Table 6.6. This is easy to see because, without ties, the ranks will be 1,
Table 6.5: The values of $r_1$ for all possible assignments for Cathy’s CRD.

<table>
<thead>
<tr>
<th>Assignment</th>
<th>Ranks for Treatment 1</th>
<th>$r_1$</th>
<th>Assignment</th>
<th>Ranks for Treatment 1</th>
<th>$r_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1, 2, 3</td>
<td>5, 2, 4</td>
<td>11</td>
<td>2, 3, 4</td>
<td>2, 4, 1</td>
<td>7</td>
</tr>
<tr>
<td>1, 2, 4</td>
<td>5, 2, 1</td>
<td>8</td>
<td>2, 3, 5</td>
<td>2, 4, 6</td>
<td>12</td>
</tr>
<tr>
<td>1, 2, 5</td>
<td>5, 2, 6</td>
<td>13</td>
<td>2, 3, 6</td>
<td>2, 4, 3</td>
<td>9</td>
</tr>
<tr>
<td>1, 2, 6</td>
<td>5, 2, 3</td>
<td>10</td>
<td>2, 4, 5</td>
<td>2, 1, 6</td>
<td>9</td>
</tr>
<tr>
<td>1, 3, 4</td>
<td>5, 4, 1</td>
<td>10</td>
<td>2, 4, 6</td>
<td>2, 1, 3</td>
<td>6</td>
</tr>
<tr>
<td>1, 3, 5</td>
<td>5, 4, 6</td>
<td>15</td>
<td>2, 5, 6</td>
<td>2, 6, 3</td>
<td>11</td>
</tr>
<tr>
<td>1, 3, 6</td>
<td>5, 4, 3</td>
<td>12</td>
<td>3, 4, 5</td>
<td>4, 1, 6</td>
<td>11</td>
</tr>
<tr>
<td>1, 4, 5</td>
<td>5, 1, 6</td>
<td>12</td>
<td>3, 4, 6</td>
<td>4, 1, 3</td>
<td>8</td>
</tr>
<tr>
<td>1, 4, 6</td>
<td>5, 1, 3</td>
<td>9</td>
<td>3, 5, 6</td>
<td>4, 6, 3</td>
<td>13</td>
</tr>
<tr>
<td>1, 5, 6</td>
<td>5, 6, 3</td>
<td>14</td>
<td>4, 5, 6</td>
<td>1, 6, 3</td>
<td>10</td>
</tr>
</tbody>
</table>

Table 6.6: The sampling distribution of $R_1$ for Cathy’s CRD.

<table>
<thead>
<tr>
<th>$r_1$</th>
<th>$P(R_1 = r_1)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>0.05</td>
</tr>
<tr>
<td>7</td>
<td>0.05</td>
</tr>
<tr>
<td>8</td>
<td>0.10</td>
</tr>
<tr>
<td>9</td>
<td>0.15</td>
</tr>
<tr>
<td>10</td>
<td>0.15</td>
</tr>
</tbody>
</table>

2, 3, 4, 5 and 6. Thus, we have the following **huge difference** between the test based on $U$ and the test based on $R_1$:

In the no-ties situation, for any given values of $n_1$ and $n_2$ there is only one sampling distribution for $R_1$, but, as we can imagine, there are an infinite number of sampling distributions for $U$. (Just change any response and you will likely obtain a different sampling distribution for $U$.)

I will follow my **pattern of presentation** that is becoming familiar to you. For Cathy’s very small study—only 20 possible assignments—we are able to obtain the exact sampling distribution of our test statistic quite easily; tedious perhaps, but easy. Next, we move to a larger study: Kymn’s study of rowing with its 252 possible assignments. This number of assignments is easily manageable for me (even though I am not particularly good at this), but I would never require you to examine so many assignments.

As I did in Chapter 3—because I hate looking at the results from dividing by 252—Table 6.7 is not quite the sampling distribution we want; to obtain the sampling distribution we need to divide each of the table’s frequencies by 252.
Table 6.7: Frequency table for the observed values $r_1$ of $R_1$ for the 252 possible assignments for Kynn’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>15.0</td>
<td>1</td>
<td>20.5</td>
<td>2</td>
<td>24.0</td>
<td>6</td>
<td>28.0</td>
<td>5</td>
<td>31.5</td>
<td>10</td>
</tr>
<tr>
<td>16.0</td>
<td>1</td>
<td>21.0</td>
<td>6</td>
<td>24.5</td>
<td>10</td>
<td>28.5</td>
<td>14</td>
<td>32.0</td>
<td>6</td>
</tr>
<tr>
<td>17.0</td>
<td>2</td>
<td>21.5</td>
<td>4</td>
<td>25.0</td>
<td>6</td>
<td>29.0</td>
<td>5</td>
<td>32.5</td>
<td>6</td>
</tr>
<tr>
<td>18.0</td>
<td>3</td>
<td>22.0</td>
<td>6</td>
<td>25.5</td>
<td>14</td>
<td>29.5</td>
<td>14</td>
<td>33.0</td>
<td>6</td>
</tr>
<tr>
<td>19.0</td>
<td>4</td>
<td>22.5</td>
<td>6</td>
<td>26.0</td>
<td>5</td>
<td>30.0</td>
<td>6</td>
<td>33.5</td>
<td>4</td>
</tr>
<tr>
<td>19.5</td>
<td>2</td>
<td>23.0</td>
<td>6</td>
<td>26.5</td>
<td>14</td>
<td>30.5</td>
<td>10</td>
<td>34.0</td>
<td>6</td>
</tr>
<tr>
<td>20.0</td>
<td>5</td>
<td>23.5</td>
<td>10</td>
<td>27.0</td>
<td>5</td>
<td>31.0</td>
<td>6</td>
<td>34.5</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>27.5</td>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Total</td>
<td>252</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Dawn’s and Sara’s studies are too large for me to obtain the exact sampling distribution of $R_1$. Instead, for each study I performed a simulation experiment with 10,000 runs. I will report on the results of these simulations in the next section.

### 6.4 Step 3: The Three Rules for Calculating the P-value

You will recall that in Chapter 5, I presented lengthy arguments in order to derive the three rules (one for each alternative) for computing the P-value. I could type similar arguments below for the sum of ranks test. I could, but I won’t. Why not?

1. If we go to the earlier arguments and replace references to the observed value of the test statistic $U$ by references to the observed value of the test statistic $R_1$ the arguments—with very minor modifications—remain valid.

2. In view of the item above, and the amount of material I want to cover in these notes, I have made the executive decision that while there is substantial educational benefit to your seeing the arguments once, seeing similar arguments repeatedly in these notes is not warranted.

Thus, without further ado, I give you the three rules for finding the P-value for our sum of ranks test.
Result 6.1 (The P-values for the sum of ranks test.) In the rules below, remember that \( r_1 \) is the sum of the ranks of treatment 1 data for the actual data.

- For the alternative \( > \), the P-value is equal to:
  \[
P(R_1 \geq r_1)
  \]
  (6.1)

- For the alternative \( < \), the P-value is equal to:
  \[
P(R_1 \leq r_1)
  \]
  (6.2)

- For the alternative \( \neq \), compute
  \[
c = n_1(n + 1)/2.
  \]
  
  - If \( r_1 = c \) then the P-value equals 1.
  
  - If \( r_1 > c \) then the P-value equals:
    \[
P(R_1 \geq r_1) + P(R_1 \leq 2c - r_1)
    \]
    (6.3)
  
  - If \( r_1 < c \) then the P-value equals:
    \[
P(R_1 \leq r_1) + P(R_1 \geq 2c - r_1)
    \]
    (6.4)

Before I illustrate the use of these rules, let me make a brief comment about the value \( c \) (short for center) in the rule for the alternative \( \neq \). The value \( c \) plays the same role that 0 played in our rule for the test statistic \( U \). The appearance of \( c \) in the current rule is a direct result of my choosing to have the test statistic be \( R_1 \) rather than the difference of the mean ranks. If our test statistic was the difference of the mean ranks, then \( c \) would be replaced by 0 in our rule. I made the executive decision that it is better to make the test statistic simple and the rule—for the two-sided alternative only—complicated.

I will illustrate these rules with our four studies from Chapters 1 and 2.

Example 6.1 (Cathy’s study.) In the answers below, please refer to the sampling distribution in Table 6.6. Recall from above that Cathy’s actual \( r_1 \)=13. For the alternative \( \neq \) only, we also need the values:

  \[
c = n_1(n + 1)/2 = 3(7)/2 = 10.5 \text{ and } 2c - r_1 = 2(10.5) - 13 = 8.
  \]

For the alternative \( > \) her P-value is:

  \[
P(R_1 \geq 13) = 0.10 + 0.05 + 0.05 = 0.20.
  \]

For the alternative \( < \) her P-value is:

  \[
P(R_1 \leq 13) = 1 - 0.05 - 0.05 = 0.90.
  \]
For the alternative $\neq$ her P-value is:

$$P(R_1 \geq 13) + P(R_1 \leq 8) = 0.20 + 0.20 = 0.40.$$  

We found in Chapter 4 that with the test statistic equal to the difference of the means, Cathy’s P-values are: 0.20 for $>$; 0.85 for $<$; and 0.40 for $\neq$. Thus, the two tests give the same P-values for $>$ and $\neq$.

**Example 6.2 (Kymn’s study.)** In the answers below, please refer to Table 6.7. It can be shown that for Kymn’s actual data, $r_1 = 40$. For the alternative $\neq$ only, we also need the values:

$$c = n_1(n + 1)/2 = 5(11)/2 = 27.5 \text{ and } 2c - r_1 = 2(27.5) - 40 = 15.$$  

For the alternative $>$ her P-value is:

$$P(R_1 \geq 40) = 1/252 \approx 0.0040.$$  

For the alternative $<$ her P-value is:

$$P(R_1 \leq 40) = 1.$$  

For the alternative $\neq$ her P-value is:

$$P(R_1 \geq 40) + P(R_1 \leq 15) = 2/252 = 0.0080.$$  

All three of these P-values are exactly the same as the ones we found in Chapter 5 with the test statistic equal to the difference of the means.

**Example 6.3 (Dawn’s study.)** For brevity, I will restrict attention to the alternative $>$. As shown earlier in this chapter, the observed value of $R_1$ is $r_1 = 133.5$. Also, recall that for Dawn’s actual data, $u = \bar{x} - \bar{y} = 5.1 - 2.9 = 2.2$. Thus, there are two possibilities for the P-value:

- $P(R_1 \geq 133.5)$; and
- $P(U \geq 2.2)$.

With 184,756 possible assignments, I am not going to compute these exact probabilities! Instead, I performed a simulation experiment with 10,000 runs. Each run, of course, selected an assignment at random from the possible assignments. Then, for the given assignment I computed both of the values $r_1$ and $u$. Below are the results I obtained:

- The relative frequency of $(R_1 \geq 133.5)$ was 0.0127; thus, the approximate P-value using ranks is 0.0127.
- The relative frequency of $(U \geq 2.2)$ was 0.0169; thus, the approximate P-value using the difference of means is 0.0169.
How should we interpret the fact that the two tests give different approximate P-values for Dawn’s data? The two tests summarize the data differently, so it should be no surprise that they give somewhat different answers. Also, in my opinion, the difference between an approximation of 0.0127 and 0.0169 is not very important. The difference does suggest, however, that the test based on ranks is a bit better than the test based on comparing means. This is a tricky point, so don’t worry if you don’t believe me; we will see a better way to look at this issue when I introduce you to the technical concept of the power of a test.

I did something subtle in my computer simulation experiment. Did you spot it? I could have performed separate simulations for each test statistic. Indeed, I previously showed you the results from two different simulation experiments for Dawn’s data (a simulation with 10,000 runs and then another with 100,000 runs). Thus, I could have used either (or both) of my earlier simulations for $U$ and combined that with a new simulation for $R_1$. Instead, I performed one new simulation experiment and in this new experiment for every assignment I evaluated both test statistics. As we will see later in these notes, the method I used gives much more precision for comparing P-values than using two separate simulations.

Why is one simulation better than two? We will see the details later, but here is the intuition. In CRDs we have been comparing treatments. It can be difficult to reach a conclusion because of variation from unit-to-unit. For example, it seems that the appetite of Bob the cat varied a great deal from day-to-day. This variation makes it difficult to see which treatment is preferred by Bob. A common theme in science and Statistics is that we learn better (more validly and more efficiently) if we can reduce variation.

In the context of computer simulations, the role of unit-to-unit variation in a CRD is played by assignment-to-assignment variation. Thus, as we will see later in these notes, by comparing $U$ to $R_1$ on the same assignments we obtain a much more precise comparison of the two tests. In the vernacular, we avoid the possibility that one test gets lucky and is evaluated on better assignments.

**Example 6.4 (Sara’s study.)** For brevity, I will restrict attention to the alternative $>$. It can be shown that for Sara’s actual data, the observed value of $R_1$ is $r_1 = 1816$. Also, recall that for Sara’s actual data, $u = \bar{x} - \bar{y} = 106.875 - 98.175 = 8.700$. Thus, there are two possibilities for the P-value:

- $P(R_1 \geq 1816)$; and
- $P(U \geq 8.700)$.

I performed a simulation experiment with 10,000 runs. Each run, of course, selected an assignment at random from the possible assignments. Then, for the given assignment I computed both of the values $r_1$ and $u$. Below are the results I obtained:

- The relative frequency of $(R_1 \geq 1816)$ was 0.0293; thus, the approximate P-value using ranks is 0.0293.
- The relative frequency of $(U \geq 8.700)$ was 0.0960; thus, the approximate P-value using the difference of means is 0.0960.

127
Our P-values for Sara’s study are dramatically different than what we found earlier. For Cathy’s study, $U$ and $R_1$ give exactly the same P-values for all but the alternative not supported by the data. For Kymn’s study, the two tests give exactly the same P-values for all three possible alternatives. For Dawn’s study the tests give different P-values for the alternative supported by the data (and also for $\neq$, although I did not show you this), but the difference is not dramatic. For Sara’s study, however, the two P-values for $>$ are dramatically different. All scientists would agree that a P-value of 0.0960 is importantly different than a P-value of 0.0293.

Why are the P-values so different for Sara’s data? Sadly, I must leave this question unanswered until we learn about the power of a test.

### 6.5 Ordinal Data

Thus far in this chapter I have presented the sum of ranks test as an alternative to the test of means. For example, for the studies of Dawn, Kymn, Sara, Cathy and others mentioned in the homework, one could use either of these tests to investigate the Skeptic’s Argument. Later in these Course Notes when we study power we will investigate the issue of when each test is better than the other—alas, neither is universally better for all scientific problems—as well as why the idea of doing both tests is tricky. In this section, I introduce you to a class of scientific problems for which the sum of ranks test can be used, but the test of means should not be used.

I will introduce the ideas with a type of medical study. The data are artificial.

**Example 6.5 (Artificial study of a serious disease.)** A collection of hospitals serves many patients with a particular serious disease. There are two competing methods of treating these patients, call them treatment 1 and treatment 2. One hundred patients are available for study and 50 are assigned to each treatment by randomization. Each patient is treated until a response is obtained and after all 100 responses are obtained the data will be analyzed. The response is **categorical** with three possibilities:

1. The patient is cured after a short period of treatment.
2. The patient is cured after a long period of treatment.
3. The patient dies.

The (artificial) data are presented in Table 6.8.

Before I proceed, I want to acknowledge that this example is a simplification that ignores some serious issues of medical ethics. I am not qualified to discuss—or even identify—these issues, so I won’t try. For the sake of this presentation let’s all agree to the following ideas.

1. The three response categories are naturally **ordered**: a fast cure is preferred to a slow cure and any cure is preferred to death.
2. **For convenience** I will assign numbers to these categories: 1 for fast cure, 2 for slow cure and 3 for death.
Table 6.8: Data from an artificial study of a serious disease.

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Fast Cure</th>
<th>Slow Cure</th>
<th>Death</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>24</td>
<td>16</td>
<td>10</td>
<td>50</td>
</tr>
<tr>
<td>2</td>
<td>18</td>
<td>17</td>
<td>15</td>
<td>50</td>
</tr>
<tr>
<td>Total</td>
<td>42</td>
<td>33</td>
<td>25</td>
<td>100</td>
</tr>
</tbody>
</table>

3. **We should never** compute a mean for these responses. It would be outrageous to say that a fast cure coupled with a death is the same as two slow cures. (The mean of 1 and 3 is 2.)

4. It is fine to compute medians, but with so few possible responses, medians are not very helpful. For example, in our data, the median for each treatment is 2, which suggests they are equivalent, but clearly treatment 1 is performing better than treatment 2 in these data.

5. As we will see below, positions and ranks make sense for these data, but there are a lot of ties! (Forty-two patients respond 1; and so on.)

I will now proceed to perform the sum of ranks test on these data. Recall that first we combine the 50 responses from treatment 1 with the 50 responses from treatment 2 to obtain a total of 100 responses. Next, we sort these 100 numbers and assign a rank to each response. It is quite easy to assign these ranks because of the many ties necessitated by the presence of only three possible responses. In particular, reading from the **Total row** of Table 6.8, we find:

- The same number, 1, appears in positions 1–42. Thus, each of these 42 numbers is assigned a rank equal to the mean of these positions: 
  \[
  \frac{1 + 42}{2} = 21.5.
  \]

- The same number, 2, appears in positions 43–75. Thus, each of these 33 numbers is assigned a rank equal to the mean of these positions: 
  \[
  \frac{43 + 75}{2} = 59.
  \]

- The same number, 3, appears in positions 76–100. Thus, each of these 25 numbers is assigned a rank equal to the mean of these positions: 
  \[
  \frac{76 + 100}{2} = 88.
  \]

From the above three bullets and reading from the treatment 1 row of Table 6.8, we find:

\[
r_1 = 24(21.5) + 16(59) + 10(88) = 2340.
\]

We may now use Formulas 6.1–6.4 on page 125 to obtain the expressions for the P-values for the three possible alternatives. They are:

- For the alternative $>$, the P-value equals
  \[
P(R_1 \geq 2340).
  \]
• For the alternative $<$, the P-value equals
  \[ P(R_1 \leq 2340). \]

• For the alternative $\neq$, first we compute
  
  \[ c = n_1(n+1)/2 = 50(101)/2 = 2525 \text{ and } 2c-r_1 = 2(2525) - 2340 = 5050 - 2340 = 2710. \]
  
  Thus, from Formula 6.4 the P-value equals
  \[ P(R_1 \leq 2340) + P(R_1 \geq 2710). \]

I do not know the exact sampling distribution for $R_1$. Therefore, I performed a computer simulation experiment with 10,000 runs on Minitab. The simulation gave me the following approximate P-values.

• For the alternative $>$, the approximate P-value equals
  \[ \text{Rel. Freq. } (R_1 \geq 2340) = 0.9196. \]

• For the alternative $<$, the approximate P-value equals
  \[ \text{Rel. Freq. } (R_1 \leq 2340) = 0.0946. \]

• For the alternative $\neq$, the approximate P-value equals
  \[ \text{Rel. Freq. } (R_1 \leq 2340) + \text{Rel. Freq. } (R_1 \geq 2710) = 0.0946 + 0.0919 = 0.1865. \]

For this study, smaller responses are better. Thus, the smallest P-value is obtained for the alternative $<$, which is the alternative supported by the data. (In the data treatment 1 performed better than treatment 2.)

### 6.6 Computing

In Section 5.4, beginning on page 104, we learned how to use the vassarstats website, [http://vassarstats.net](http://vassarstats.net) to obtain a simulation study for our test of means with test statistic given by $U$. I have very good news for you! If we enter the ranks into the vassarstats website for comparing means, we can obtain a valid simulation for the sum of ranks test, following the same steps you learned in Chapter 5. Here is why, but you don’t need to know this: As I argued earlier in this chapter, a test based on $R_1$ is the same as a test based on the difference of the mean ranks; hence, if we enter ranks into the vassarstats website, it works!

Below I give you examples of using the site for three of our studies. Note that if you replicate what I do below, you will likely obtain different, but similar, approximate P-values. Also, be aware that some of the examples below are fairly tedious because they involve typing lots of data values into the site.
1. I performed a 10,000 run simulation of Dawn’s data on *vassarstats* and obtained an approximate P-value of 0.0130 for the alternative $>$, which agrees quite well with the 0.0127 I obtained with Minitab.

2. I performed a 10,000 run simulation of Sara’s data on *vassarstats* and obtained an approximate P-value of 0.0253 for the alternative $>$, which agrees reasonably well with the 0.0293 I obtained with Minitab.

3. Finally, for the artificial ordinal data in Table 6.8, I performed a 10,000 run simulation on *vassarstats* and obtained an approximate P-value of 0.0927 for the alternative $<$, which agrees quite well with the 0.0946 I obtained with Minitab.
Table 6.9: An Example of how to obtain $r_1$.

Data, by treatment:

<table>
<thead>
<tr>
<th>Treatment 1:</th>
<th>11</th>
<th>14</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment 2:</td>
<td>14</td>
<td>15</td>
<td>12</td>
</tr>
</tbody>
</table>

Data combined, sorted, assigned ranks; observations from treatment 1 are bold-faced:

<table>
<thead>
<tr>
<th>Position:</th>
<th>1 2 3 4 5 6 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observation:</td>
<td>11</td>
</tr>
<tr>
<td>Rank:</td>
<td>1.5</td>
</tr>
</tbody>
</table>

$r_1 = 1.5 + 1.5 + 5 = 8$

6.7 Summary

In Chapter 3 we learned about the Skeptic’s Argument which states that the treatment level in a CRD is irrelevant. In Chapter 5 we learned how to investigate the validity of the Skeptic’s Argument by using a statistical test of hypotheses. The test statistic in Chapter 5 is $U$, which tells us to compare the two sets of data by comparing their means. In this chapter we propose an alternative to test statistic $U$: the test statistic based on summing ranks, $R_1$.

The obvious question is: What are ranks? In a CRD, combine the data from the two treatments into one list. Next, sort the data (from smallest to largest, as we always do in Statistics). An example of these ideas is presented in Table 6.9 artificial data for a CRD with $n_1 = 3$ and $n_2 = 4$. In this table we have the sorted combined data, which consists of the seven numbers: 11, 11, 12, 14, 14, 14 and 15. We assign ranks to these seven numbers as follows:

- When a value is repeated (or tied; 11’s and 14’s in our list) all occurrences of the value receive the same rank. This common rank is the mean of the positions of these values. Thus, the 11’s reside in positions 1 and 2; hence, their ranks are both $(1+2)/2 = 1.5$. Also, the 14’s reside in positions 4, 5 and 6; hence, their common rank is $(4+5+6)/3 = (4+6)/2 = 5$.

- For each non-repeated (non-tied) value, its rank equals its position; hence, rank 3 [7] for the observation 12 [15].

The observed value, $r_1$ of the test statistic $R_1$ is obtained by summing the ranks of the data that came from treatment 1. For our current table, this means that we sum the ranks in bold-faced type:

$r_1 = 1.5 + 1.5 + 5 = 8$.

We can also obtain the sum of the ranks of the data from treatment 2:

$r_2 = 3 + 5 + 5 + 7 = 20$.  

132
We can reduce the time we spend summing ranks if we remember that for a total of \( n \) units in a CRD, the sum of all ranks is \( n(n + 1)/2 \). For our artificial data in Table 6.9, \( n = 7 \); thus, the sum of all ranks is \( 7(8)/2 = 28 \) which agrees with our earlier \( r_1 + r_2 = 8 + 20 = 28 \).

For descriptive purposes, we should compare the mean ranks, which for our artificial data are:

\[
\bar{r}_1 = r_1/n_1 = 8/3 = 2.67 \quad \text{and} \quad \bar{r}_2 = r_2/n_2 = 20/4 = 5.
\]

In words, the data from treatment 2 tend to be larger than the data from treatment 1.

We are now ready to consider the sum of ranks test. The null hypothesis is that the Skeptic is correct. There are three possible choices for the alternative: abbreviated by \( > \), \( < \) and \( \neq \). As a practical matter, just remember that \( > [<] \) means that—in terms of ranks—treatment 1 tends to give larger [smaller] responses than treatment 2; and that \( \neq \) means that treatment 1 tends to give either larger or smaller responses than treatment 2.

In principle, the sampling distribution of \( R_1 \) is obtained exactly like the sampling distribution of \( U \): for every possible assignment we calculate its value of \( r_1 \), on the assumption, of course, that the Skeptic is correct. For studies with a small number of possible assignments, we can obtain the exact sampling distribution of \( R_1 \). For studies with a large number of possible assignments, we can use a computer simulation experiment to obtain an approximation to the sampling distribution of \( R_1 \). In addition, in Chapter 7 we will obtain a fancy math approximation to the sampling distribution of \( R_1 \). By fancy math I mean a result based on clever theorems that have been proven by professional mathematicians.

For Sara’s data we found that the P-value for the test statistic \( U \) is very different from the P-value for the test statistic \( R_1 \). This issue will be explored later when we consider the power of a test.

Finally, the sum of ranks test can be used for ordinal data. The test of means should not be used for ordinal data because a mean is not an appropriate summary of ordinal data.
6.8 Practice Problems

1. An unbalanced CRD with \( n_1 = 4 \) and \( n_2 = 5 \) yields the following data:

<table>
<thead>
<tr>
<th>Treatment 1:</th>
<th>4</th>
<th>12</th>
<th>9</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment 2:</td>
<td>6</td>
<td>3</td>
<td>25</td>
<td>12</td>
</tr>
</tbody>
</table>

Use these data to answer the following questions.

(a) Assign ranks to these observations.

(b) Calculate the values of \( r_1, r_2 \) and the mean rank for each treatment. Briefly interpret the two mean ranks.

(c) Calculate \( \bar{x} \) and \( \bar{y} \) for these data. Briefly interpret these means. Compare this interpretation to your interpretation of the mean ranks in part(b). Comment.

2. The purpose of this problem is to give you practice at computing exact P-values. Suppose that you conduct a balanced CRD with a total of \( n = 10 \) units and that your sampling distribution is given by the frequencies (divided by 252) in Table 6.7.

(a) Find the exact P-values for the alternatives \( > \) and \( \neq \) for each of the following actual values of \( r_1 \): \( r_1 = 37.0; r_1 = 33.5; \) and \( r_1 = 31.5 \).

(b) Find the exact P-values for the alternatives \( < \) and \( \neq \) for each of the following actual values of \( r_1 \): \( r_1 = 19.0; r_1 = 20.5; \) and \( r_1 = 22.0 \).

3. A CRD is performed with an ordinal categorical response. The data are below.

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Low</th>
<th>Middle Low</th>
<th>Middle High</th>
<th>High</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>25</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>25</td>
</tr>
<tr>
<td>Total</td>
<td>14</td>
<td>13</td>
<td>12</td>
<td>11</td>
<td>50</td>
</tr>
</tbody>
</table>

Assign numbers 1 (Low), 2 (Middle Low), 3 (Middle High) and 4 (High) to these categories.

(a) Assign ranks to the 50 observations.

(b) Calculate \( r_1, r_2 \) and the two mean ranks. Comment.

(c) Use the vassarstats website to obtain two approximate P-values based on a simulation with 10,000 runs. Identify the alternative for each approximate P-value.

4. I reminded you of Doug’s study of the dart game 301 in Practice Problem 4 in Chapter 5, in Section 5.6 on page 110.

Below are the ranks for Doug’s data on treatment 1 (personal darts):
Below are the ranks for Doug’s data on treatment 2 (bar darts):

\[
\begin{matrix}
2.5 & 7.0 & 9.5 & 9.5 & 12.5 & 12.5 & 12.5 & 16.0 & 19.5 & 25.5 \\
25.5 & 28.5 & 30.5 & 32.5 & 34.5 & 34.5 & 37.0 & 37.0 & 39.0 & 40.0 \\
\end{matrix}
\]

I entered these ranks into the vassarstats website and obtained the following output:

The mean for the first [second] set of ranks is 17.7 [23.3]. The approximate P-values based on 10,000 runs are: 0.0623 for one-tailed and 0.1277 for two-tailed.

(a) We saw earlier that for Doug’s data, \( \bar{x} = 18.6 \) and \( \bar{y} = 21.2 \). Do the means of the ranks tell a similar or different story than the means of the data? Explain.

(b) Match the two P-values given by vassarstats to their alternatives. Compare these two P-values to the approximate P-values I presented in Practice Problem 4 in Chapter 5. Comment.
6.9 Solutions to Practice Problems

1. (a) I combine the data into one set of \( n = 9 \) numbers and sort them:

<table>
<thead>
<tr>
<th>Position</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>6</td>
<td>9</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>25</td>
</tr>
<tr>
<td>Ranks</td>
<td>1.5</td>
<td>1.5</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>9</td>
</tr>
</tbody>
</table>

The two observations equal to 3 reside in positions 1 and 2; hence, they are both assigned the rank of \((1 + 2)/2 = 1.5\). The three observations equal to 12 reside in positions 6–8; hence, they are all assigned the rank of \((6 + 8)/2 = 7\). There are no other tied values. Thus, the rank of each remaining observation equals its position.

(b) The observations on the first treatment, 4, 12, 9 and 12, have ranks 3, 7, 5 and 7; thus, \( r_1 = 3 + 7 + 5 + 7 = 22 \). The sum of all nine ranks is \( 9(10)/2 = 45 \). Thus,

\[
r_2 = 45 - r_1 = 45 - 22 = 23.
\]

Note that we also could obtain \( r_2 \) by summing ranks:

\[
r_2 = 1.5 + 1.5 + 4 + 7 + 9 = 23.
\]

The mean ranks are:

\[
r_1/n_1 = 22/4 = 5.5 \text{ and } r_2/n_2 = 23/5 = 4.6.
\]

The mean of the treatment 1 ranks is larger than the mean of the treatment 2 ranks. This means, in terms of ranks, that the observations on treatment 1 tend to be larger than the observations on treatment 2.

(c) The means are

\[
\bar{x} = (4+9+12+12)/4 = 37/4 = 9.25 \text{ and } \bar{y} = (3+3+6+12+25)/5 = 49/5 = 9.8.
\]

In terms of the means, the observations on treatment 2 are larger than the observations on treatment 1. This interpretation is the reverse of what we found for ranks.

Comment: The one unusually large observation, 25, has a pronounced effect on \( \bar{y} \). In terms of ranks, 25 has the same impact that it would if it were replaced by 13; it would still be the largest observation and have rank of 9.

2. (a) First, note that \( c = n_1(n + 1)/2 = 5(11)/2 = 27.5 \) is smaller than all of the values of \( r_1 \). Thus, in addition to using Formula 6.1 on page 125 for the alternative >, we use Formula 6.3 for the alternative \( \neq \). You should verify the numbers in the following display. (I will verify one of them for you, immediately below this display.)

<table>
<thead>
<tr>
<th>( r_1 )</th>
<th>( P(R_1 \geq r_1) )</th>
<th>( 2c - r_1 )</th>
<th>( P(R_1 \leq 2c - r_1) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>37.0</td>
<td>7/252 = 0.0278</td>
<td>18.0</td>
<td>7/252 = 0.0278</td>
</tr>
<tr>
<td>33.5</td>
<td>30/252 = 0.1190</td>
<td>21.5</td>
<td>30/252 = 0.1190</td>
</tr>
<tr>
<td>31.5</td>
<td>58/252 = 0.2302</td>
<td>23.5</td>
<td>58/252 = 0.2302</td>
</tr>
</tbody>
</table>
For the first entry in the ‘\( r_1 = 37.0 \)’ row: from Table 6.7 we find that
\[
\text{Frequency } (R_1 \geq 37.0) = 3 + 2 + 1 + 1 = 7.
\]
We now have the following P-values:

<table>
<thead>
<tr>
<th>( r_1 )</th>
<th>Alternative</th>
<th>( &gt; )</th>
<th>( \neq )</th>
</tr>
</thead>
<tbody>
<tr>
<td>37.0</td>
<td>0.0278</td>
<td>2(0.0278) = 0.0556</td>
<td></td>
</tr>
<tr>
<td>33.5</td>
<td>0.1190</td>
<td>2(0.1190) = 0.2380</td>
<td></td>
</tr>
<tr>
<td>31.5</td>
<td>0.2302</td>
<td>2(0.2302) = 0.4604</td>
<td></td>
</tr>
</tbody>
</table>

(b) As above, \( c = 27.5 \), which is larger than all of the values of \( r_1 \). Thus, in addition to using Formula 6.2 for the alternative \(<\), we use Formula 6.4 for the alternative \( \neq \). You should verify the numbers in the following display:

<table>
<thead>
<tr>
<th>( r_1 )</th>
<th>( P(R_1 \leq r_1) )</th>
<th>( 2c - r_1 )</th>
<th>( P(R_1 \geq 2c - r_1) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>19.0</td>
<td>11/252 = 0.0437</td>
<td>36.0</td>
<td>11/252 = 0.0437</td>
</tr>
<tr>
<td>20.5</td>
<td>20/252 = 0.0794</td>
<td>34.5</td>
<td>20/252 = 0.0794</td>
</tr>
<tr>
<td>22.0</td>
<td>36/252 = 0.1429</td>
<td>33.0</td>
<td>36/252 = 0.1429</td>
</tr>
</tbody>
</table>

For the first entry in the ‘\( r_1 = 19.0 \)’ row: from Table 6.7 we find that
\[
\text{Frequency } (R_1 \leq 19.0) = 4 + 3 + 2 + 1 + 1 = 11.
\]
We now have the following P-values:

<table>
<thead>
<tr>
<th>( r_1 )</th>
<th>Alternative</th>
<th>( &lt; )</th>
<th>( \neq )</th>
</tr>
</thead>
<tbody>
<tr>
<td>19.0</td>
<td>0.0437</td>
<td>2(0.0437) = 0.0874</td>
<td></td>
</tr>
<tr>
<td>20.5</td>
<td>0.0794</td>
<td>2(0.0794) = 0.1588</td>
<td></td>
</tr>
<tr>
<td>22.0</td>
<td>0.1429</td>
<td>2(0.1429) = 0.2858</td>
<td></td>
</tr>
</tbody>
</table>

3. (a) In the combined list of 50 observations, positions 1–14 contain the response ‘1;’ hence, each ‘1’ is assigned the rank of \((1 + 14)/2 = 7.5\). Positions 15–27 contain the response ‘2;’ hence, each ‘2’ is assigned the rank of \((15 + 27)/2 = 21\). Positions 28–39 contain the response ‘3;’ hence, each ‘3’ is assigned the rank of \((28 + 39)/2 = 33.5\). Finally, positions 40–50 contain the response ‘4;’ hence, each ‘4’ is assigned the rank of \((40 + 50)/2 = 45\).

(b) From part(a),
\[
\begin{align*}
\hat{r}_1 &= 10(7.5) + 6(21) + 5(33.5) + 4(45) = 548.5 \text{ and } \\
\hat{r}_2 &= 4(7.5) + 7(21) + 7(33.5) + 7(45) = 726.5.
\end{align*}
\]
As a partial check, the sum of all ranks must be
\[
50(51)/2 = 1275 \text{ and } r_1 + r_2 = 548.5 + 726.5 = 1275.
\]
The mean ranks are

\[ \frac{r_1}{n_1} = \frac{548.5}{25} = 21.94 \quad \text{and} \quad \frac{r_2}{n_2} = \frac{726.5}{25} = 29.06. \]

The observations on treatment 2 tend to be larger than the observations on treatment 1.

(c) I enter 25 and 25 for the sample sizes. Then I enter the ranks as data and click on **Calculate**. The site presents 21.94 and 29.06 as the means; this is a partial check that I entered the ranks correctly! I click on **Resample x 1000** ten times to obtain my 10,000 runs. The relative frequencies I obtained are 0.0354 for one-tailed and 0.0709 for two-tailed. Thus, the approximate P-value for \( \neq \) is 0.0709. The mean ranks of the data agree with the alternative \(<\); hence, the approximate P-value for \(<\) is 0.0354. The site does not give an approximate P-value for \(>\).

4. (a) The mean of the ranks on treatment 1 is smaller than the mean of the ranks on treatment 2; this means that, in terms of position, the data in treatment 1 tend to be smaller than the data in treatment 2. The data support the alternative \(<\), not \(>\). Similarly, the fact that \(\bar{x} < \bar{y}\) supports the alternative \(<\), not \(>\). Thus, unlike in Practice Problem 1 of this section, looking at observations gives the same qualitative conclusion as looking at ranks.

(b) For the sum of ranks test, the approximate P-value for \(<\) is 0.0623. In Chapter 4, I analyzed these data using the test that compares means. I reported on two simulation experiments, each with 10,000 runs (one using Minitab, one using vassarstats). My two approximate P-values for \(<\) were 0.0426 and 0.0433. Both of these are considerably smaller than the value for the sum of ranks test. As we will learn when we study power, this suggests that comparing means is better than comparing mean ranks for Doug’s data. Recall that we had found the opposite pattern for Sara’s golfing data.

For the sum of ranks test, the approximate P-value for \(\neq\) is 0.1277. In Chapter 4, I analyzed these data using the test that compares means. I reported on two simulation experiments, each with 10,000 runs (one using Minitab, one using vassarstats). My two approximate P-values for \(\neq\) were 0.0844 and 0.0836. Both of these are considerably smaller than the value for the sum of ranks test.
Table 6.10: Artificial data for Homework Problem 1.

<table>
<thead>
<tr>
<th>Treatment 1:</th>
<th>7 7 8 8 9 14 14 14</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment 2:</td>
<td>1 3 4 5 7 8 11 11</td>
</tr>
</tbody>
</table>

Table 6.11: 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>
</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>
</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>
</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>
</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>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>27</td>
<td>20</td>
<td>32</td>
<td>14</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Total</td>
<td>252</td>
</tr>
</tbody>
</table>

6.10 Homework Problems

1. Table 6.10 presents artificial data for a balanced CRD with a total of $n = 16$ units.

   (a) Calculate the values of $r_1$, $r_2$ and the two mean ranks. Comment.

   (b) Use the vassarstats website to perform a 10,000 run simulation experiment. Use your results to obtain two approximate P-values and identify the alternative corresponding to each approximate P-value.

2. Table 6.11 presents frequencies for the sampling distribution of $R_1$ for every balanced CRD with $n = 10$ units and 10 distinct response values.

   (a) Find the exact P-values for the alternatives $>$ and $\neq$ for each of the following actual values of $r_1$: $r_1 = 37$; $r_1 = 33$; and $r_1 = 31$.

   (b) Find the exact P-values for the alternatives $<$ and $\neq$ for each of the following actual values of $r_1$: $r_1 = 17$; $r_1 = 21$; and $r_1 = 26$.

3. A CRD is performed with an ordinal categorical response. The data are below.

<table>
<thead>
<tr>
<th>Response</th>
<th>Treatment</th>
<th>Disagree</th>
<th>Neutral</th>
<th>Agree</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>8</td>
<td>7</td>
<td>5</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>7</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>11</td>
<td>12</td>
<td>12</td>
<td>35</td>
</tr>
</tbody>
</table>

Assign numbers 1 (Disagree), 2 (Neutral) and 3 (Agree) to these categories.
(a) Calculate \( r_1, r_2 \) and the two mean ranks. Comment.

(b) Use the \textit{vassarstats} website to obtain two approximate P-values based on a simulation with 10,000 runs. Identify the alternative for each approximate P-value.

4. In Homework Problem 4 of Chapter 5 (Section 5.8 on page 115) you performed a test of means on Reggie’s dart study introduced in Chapter 1. Below are the ranks for Reggie’s data on treatment 1.

\[
6.0 \quad 7.0 \quad 9.0 \quad 13.5 \quad 15.5 \quad 15.5 \quad 18.5 \quad 18.5 \quad 23.5 \quad 23.5 \\
25.5 \quad 27.0 \quad 28.5 \quad 28.5 \quad 30.0
\]

Below are the ranks for Reggie’s data on treatment 2.

\[
1.0 \quad 2.0 \quad 3.0 \quad 4.0 \quad 5.0 \quad 8.0 \quad 10.0 \quad 11.5 \quad 11.5 \quad 13.5 \\
18.5 \quad 18.5 \quad 21.0 \quad 22.0 \quad 25.5
\]

Enter these ranks into the \textit{vassarstats} website and obtain two approximate P-values based on a simulation with 10,000 runs. Use the output to answer the following questions.

(a) What is the mean of the ranks on treatment 1? What is the mean of the ranks on treatment 2? Compare these means and comment.

(b) Match each approximate P-value for the sum of ranks test obtained from \textit{vassarstats} to its alternative.

(c) Compare each sum of ranks test P-value to its corresponding P-value for a comparison of means that you found in doing the Chapter 5 Homework.