7.4: Non-Parametric Significance Tests

The significance tests described in Chapter 7.2 assume that we can treat the individual samples as if they are drawn from a population that is normally distributed. Although often a reasonable assumption, there are times when this is a poor assumption, such as when there is a likely outlier that we are not inclined to remove. Non-parametric significance tests allow us to compare data sets, but without making implicit assumptions about our data's distribution. In this section we will consider two non-parametric tests, the Wicoxson signed rank test, which we can use in place of a paired *t*-test, and the Wilcoxon rank sum test, which we can use in place of an unpaired *t*-test.

Wilcoxson Signed Rank Test

When we use paired data we first calculate the difference, d_i, between each sample's paired values. We then subtract the expected difference from each d_i and then sort these adjusted differences from smallest-to-largest without considering the sign. We then assign each difference a rank (1, 2, 3, ...) and add back its sign. If two or more entries have the same absolute difference, then we average their ranks. Finally, we add together the positive ranks and add together the negative ranks. If there is no difference in the two data sets, then we expect that these two sums should be similar in value. If the smaller of the two ranks is less than a critical value, then there is reason to believe that the two data sets are significantly different from each other; see Appendix 6 for a table of critical values.

Example 7.4.1

Marecek et. al. developed a new electrochemical method for the rapid determination of the concentration of the antibiotic monensin in fermentation vats [Marecek, V.; Janchenova, H.; Brezina, M.; Betti, M. *Anal. Chim. Acta* **1991**, *244*, 15–19]. The standard method for the analysis is a test for microbiological activity, which is both difficult to complete and time-consuming. Samples were collected from the fermentation vats at various times during production and analyzed for the concentration of monensin using both methods. The results, in parts per thousand (ppt), are reported in the following table. This is the same data as in Example 7.2.6.

Sample	Microbiological	Electrochemical
1	129.5	132.3
2	89.6	91.0
3	76.6	73.6
4	52.2	58.2
5	110.8	104.2
6	50.4	49.9
7	72.4	82.1
8	141.4	154.1
9	75.0	73.4
10	34.1	38.1
11	60.3	60.1

Is there a significant difference between the methods at $\alpha = 0.05$?

Solution

Defining the difference between the methods as

$$d_i = (X_{\text{elect}})_i - (X_{\text{micro}})_i$$

we calculate the difference for each sample.

sample	1	2	3	4	5	6	7	8	9	10	11
d_i	2.8	1.4	–3.0	6.0	–6.6	–0.5	9.7	12.7	–1.6	4.0	–0.2

Next, we order the individual differences from smallest-to-largest without considering the sign

d_i	–0.2	–0.5	1.4	–1.6	2.8	–3.0	4.0	6.0	–6.6	9.7	12.7

We then assign each individual difference a rank, retaining the sign; thus

d_i	–1	–2	3	–4	5	–6	7	8	–9	10	11

The sum of the negative ranks is 22 and the sum of the positive ranks is 44. The critical value for 11 samples and $\alpha = 0.05$ is 10. As the smaller of our two ranks, 22, is greater than 10, there is no evidence to suggest that there is a difference between the two methods.

Wilcoxson Rank Sum Test

The Wilcoxon rank sum test (also know as the Mann-Whitney U test) is used to compare two unpaired data sets. The values in the two data sets are sorted from smallest-to-largest, maintaining sample identity. After sorting, each value is assigned a rank (1, 2, 3, ...), again, maintaining sample identity. If two or more entries have the same absolute difference, then their ranks are averaged. Next, we add up the ranks for each sample. If there is no difference in the two data sets, then we expect that the positive and negative ranks should be similar in value. To account for differences in the size of each sample, we subtract

$$\frac{n_i(n_i+1)}{2}$$

from each sum where n_i is the size of the sample. If the smaller of the two ranks is less than a critical value, then there is reason to believe that the two data sets are significantly different from each other; see Appenidx 7 for a table of critical values.

Example 7.4.2

To compare two production lots of aspirin tablets, you collect samples from each and analyze them, obtaining the following results (in mg aspirin/tablet).

Lot 1: 256, 248, 245, 244, 248, 261

Lot 2: 241, 258, 241, 256, 254

Is there any evidence at $\alpha = 0.05$ that there is a significant difference between these two sets of results?

Solution

First, we sort the results from smallest-to-largest. To distinguish between the two samples, those from Lot 1 are shown in bold.

241, 241, **244**, **245**, **248**, **248**, 254, **256**, 256, 258, **261**

Next we assign ranks, identifying those samples from Lot 1 by underlying them.

1.5, 1.5, **3**, **4**, **5.5**, **5.5**, 7, **8.5**, 8.5, 10, **11**

The sum of the ranks for Lot 1 is 37.5 and the sum of the ranks for Lot 2 is 28.5. After adjusting for the size of each sample, we have

$$37.5 - \frac{6(6+1)}{2} = 16.5$$

for Lot 1 and

$$28.5 - \frac{(5)(5+1)}{2} = 13.5$$

for Lot 2. From Appendix 7, the critical value for $\alpha = 0.05$ is 3. As the smaller of our two ranks, 13.5, is greater than 3, there is no evidence to suggest that there is a difference between the two methods.