1 Wilcoxon Rank Sum Test

In practice, very few variable have a true normal distribution. The tests we have used are considered to be *robust* (that is, not sensitive to moderate lack of normality). If your data is not normal, then we need to use other test.

The *Wilcoxon Rank Sum Test* will replace the t test if the data is not normal.

Start with two independent samples, with size n_1 and n_2, from two independent populations.

For any two samples (regardless of shape or normality), the hypotheses are

$$H_0 : \text{The two distributions are the same.}$$
$$H_a : \text{One has values that are systematically larger.}$$

(A more exact statement of the "systematically larger" alternative hypothesis is a bit tricky, so we wont try to give it here.) We can always use these hypotheses. Note that the hypotheses do not involve any specific parameter such as the mean or median. We say that it is a "nonparametric" test.

In the rare case that both populations have distributions of the same shape (maybe both are skewed to the right), our hypotheses become

$$H_0 : \text{Median 1} - \text{Median 2} = 0$$
$$H_a : \text{Median 1} - \text{Median 2} > 0$$

You can color code the two sets (make one blue and the other red) because we are going to combine the numbers into one larger set of size $N = n_1 + n_2$. Order the larger set and apply *ranks* to each number. That is, label the smallest number 1, the next smallest number 2, ..., and the largest number N, the big sample size. You can color code these values to match the sample colors. Note that

$$1 + 2 + \cdots + N = \frac{N(N+1)}{2}.$$

What do we do if we have repeated values? We will say that repeated values are *tied*. Tied values will be asigned the average of the corresponding ranks.

1. If the 3rd and 4th values are tied (the 2nd and 5th values are different), then assign each a rank of 3.5. The 5th value will get a rank of 5.

$$\text{Ranks} : 1, 2, 3.5, 3.5, 5$$

2. If the 4th, 5th, and 6th values are tied, then assign each a rank of 5 (average of the rankings 4, 5, and 6). The 7th value will get a rank of 7.

$$\text{Ranks} : 1, 2, 3, 5, 5, 5, 7$$

After you take care of all the ties, the sum of the ranks will still be $N(N+1)/2$.

If the two distributions are identical, then samples of the same size should have roughly the same number of small values, the same number of medium

values, and the same number of large values. Thus, the *sums of ranks* for each sample should be roughly the same.

If instead, the sample sizes differ, then they should have proportionally similar sums of ranks.

Now, count the rankings of the first sample (remember that we color coded them). The *Sum W of the ranks for the first sample* is the Wilcoxon rank sum statistic.

If the two populations have the same continuous distribution, then W has

$$\text{Mean}_W = \frac{n_1 (N+1)}{2}$$

and standard deviation

$$SD_W = \sqrt{\frac{n_1 n_2 (N+1)}{12}}.$$

The *Wilcoxon rank sum test* rejects the hypothesis that the two populations have identical distributions when the rank sum W is far from its mean.

Ex: The test scores shown below were recorded by two different professors for two sections of the same course. Test to determine if the distributions are identical or not.

Professor A	74	78	68	72	76	69	71	74
Professor B	75	80	87	81	72	73	80	76

Here, $n_1 = 8$, $n_2 = 8$, and $N = N_1 + n_2 = 16$. We see that

$$1 + 2 + \cdots + 16 = \frac{16\,(16+1)}{2} = 136.$$

If the grade distributions were the same, we would expect the sum of the ranks in either group to be 68 (half of 136).

We start by combining the samples into one big data set. We then order the

set. Next we assign rankings and keep track of values that are tied.

Data Values	Rankings
68	1
69	2
71	3
72	4.5
72	4.5
73	6
74	7.5
74	7.5
75	9
76	10.5
76	10.5
78	12
80	13.5
80	13.5
81	15
87	16

Looking at the chart above, we think that maybe the values for Professor B are systematically larger than the values for Professor A. Our hypotheses are

H_0 : The two distributions of grades are the same.

H_a : The grade values for Professor B are systematically larger than the grade values for Professor A.

The sum of professor A's ranks (the blue numbers) is $W = 48$. We calculate the mean

$$\text{Mean}_W = \frac{n_1\,(N+1)}{2} = \frac{(8)\,((16)+1)}{2} = 68$$

and standard deviation

$$SD_W = \sqrt{\frac{n_1 n_2\,(N+1)}{12}} = \sqrt{\frac{(8)(8)\,((16)+1)}{12}} = 9.52.$$

Standardizing gives the *Wilcoxon test statistic*

$$z = \frac{W - \text{Mean}_W}{SD_W} = \frac{48-68}{9.52} = -2.10.$$

We look up the area to the left of $z = -2.10$ since our first set is for Professor A and Professor A'a values are "smaller" than Professor B's values. The p-value that corresponds to the left of $z = -2.10$ is about 1.785%.

Since the p-value of 1.785% is less than 5%, there is strong evidence that the two professors have different grade distributions.

Mail to rstephens@colgate.edu