

## **CHAPTER 14 – NONPARAMETRIC TESTS**

Everything that we have done up until now in statistics has relied heavily on one major fact: that our data is normally distributed. We have been able to make inferences about population means (one-sample, two-sample z and t tests and analysis of variance), but in each case we assumed that our population was normal. What happens when we want to perform a test on our data, but we have no idea what its true distribution is, and therefore can't assume that our data are normally distributed? In this case, we use what are called **nonparametric tests. These tests do not require any specific form for the distribution of the population.**

If we know what distribution we are dealing with, it is much more practical and useful to use a particular test that is designed for your specific purpose and conditions. If we know we are dealing with a normally distributed population, it is more beneficial to use a Z, t, or F test when performing an inference about means, and our results will be more accurate.

When we don't know what population we are dealing with, it is more beneficial to use a test that can work for any type of distribution, because that way you can be sure that you will be 'prepared' for any condition. At the same time, though, if your population really was normally distributed and you used a nonparametric test, then your results won't be as accurate had you used a Z, t or F test.

When our data is normally distributed, the mean is equal to the median and we use the mean as our measure of center. However, if our data is skewed, then the median is a much better measure of center. Therefore, just like the Z, t and F tests made inferences about the **population mean(s)**, nonparametric tests make inferences about the **population median(s)**.

We are going to be focusing on nonparametric tests which are **rank tests**. In these types of tests, we rank (or place in order) each observation from our data set. Although we aren't specifying which distribution our data is from, it must be from a continuous distribution. That is, each distribution must be described by a density curve that allows observations to take on any value in some interval.

The following is a table which identifies a particular normal test and its nonparametric (or rank) counterpart.

Setting	Normal test	Rank test
One sample	One-sample $t$ test Section 7.1	Wilcoxon signed rank test Section 14.2
Matched pairs	Apply one-sample test to differences within pairs	
Two independent samples	Two-sample $t$ test Section 7.2	Wilcoxon rank sum test Section 14.1
Several independent samples	One-way ANOVA $F$ test Chapter 12	Kruskal-Wallis test Section 14.3

## **WILCOXON RANK SUM TEST**

The Wilcoxon Rank Sum test is used to test for a **difference between two samples**. It is the nonparametric counterpart to the two-sample Z or t test. Instead of comparing two population means, we compare two population medians.

Draw an SRS of size  $n_1$  from population 1, and then draw an independent SRS of size  $n_2$  from population 2. So the **total number of observations is  $N = n_1 + n_2$** . The next step in this test is to **rank** our set of observations. Although we are dealing with two samples, when we rank the observations, we rank them as if they came from one large group. When  $N$  is equal to our total sample size, our **smallest observation receives a rank of 1**, and the **largest observation receives a rank of  $N$** . Working with ranks instead of numerical outcomes, allows us to abandon specific assumptions about the shape of the distribution.

The sum of the ranks of the first sample is  **$W$ , the Wilcoxon Rank Sum test statistic**. If one sample is truly bigger than the other, we'd expect its ranks to be higher than the others. So after we have ranked all of the observations, we sum up the ranks for each of the two samples and we can then **compare the two rank sums**. If there is no difference between our two samples and our sample sizes are equal, then we'd expect  $W$  to be roughly half of  $[N(N + 1)]/2$  (or  $[N(N + 1)]/4$ ). If our sample sizes are different, then we'd expect  $W$  to be roughly equal to its mean/expected value ( $\mu_w = \frac{n_1(N + 1)}{2}$ ).

When there truly is a difference between the two samples, then  $W$  would be a value far from its mean.

If both of our samples come from the same continuous distribution, then  $W$  has:

- **Mean**       $\mu_w = \frac{n_1(N + 1)}{2}$

- **Standard Deviation**       $\sigma_w = \sqrt{\frac{n_1 n_2 (N + 1)}{12}}$

We generally use words, rather than symbols to express the null and alternative hypotheses. As usual, our null hypothesis is that there is no difference between our two populations, and our alternative hypothesis specifies how we think the two populations are different (whether one-sided or two-sided).

P-values are calculated by software, or using a normal approximation. We can form another  $Z$  statistic by standardizing  $W$ . Then once we have our  $Z$  statistic, we can find our p-value using the same methods as Stat I.

$$W \text{ is standardized by: } Z = \frac{W - \mu_w}{\sigma_w} = \frac{W - \frac{n_1(N+1)}{2}}{\sqrt{\frac{n_1 n_2 (N+1)}{12}}}$$

### **Example #1**

Many states are considering lowering the blood-alcohol level at which a driver is designated as driving under the influence (DUI) of alcohol. An investigator for a legislative committee designed the following test to study the effect of alcohol on reaction time. Ten participants consumed a specified amount of alcohol. Another group of ten participants consumed the same amount of a nonalcoholic drink, a placebo. The two groups did not know whether they were receiving alcohol or the placebo. The twenty participants' average reaction times (in seconds) to a series of simulated driving situations are reported in the following table. A boxplot of the two samples show that the population distributions are skewed right. Does it appear that alcohol consumption increases reaction time? Perform a significance test and clearly state your hypotheses, test statistic, p-value and conclusion.

Placebo      0.90 0.37 1.63 0.83 0.95 0.78 0.86 0.61 0.38 1.97

Alcohol      1.46 1.45 1.76 1.44 1.11 3.07 0.98 1.27 2.56 1.32

Solution:

The following is the same table as above with each observation's rank in parentheses:

Placebo	0.90 (7)	0.37 (1)	1.63 (16)	0.83 (5)	0.95 (8)	0.78 (4)	0.86 (6)	0.61 (3)	0.38 (2)	1.97 (18)
Alcohol	1.46 (15)	1.45 (14)	1.76 (17)	1.44 (13)	1.11 (10)	3.07 (20)	0.98 (9)	1.27 (11)	2.56 (19)	1.32 (12)

Let  $W$  be the sum of the ranks of the group that consumed alcoholic beverages. So  $W = 15 + 14 + 17 + 13 + 10 + 20 + 9 + 11 + 19 + 12 = 140$ .

For our test, we have:

$H_0$ : There is no difference between the alcohol and placebo populations on reaction time.  
 $H_a$ : The alcohol population has a greater reaction time than does the placebo population.

Test Statistic: Note that  $\mu_w = (10 \times 21) / 2 = 105$  and  $\sigma_w = \sqrt{[(10 \times 10 \times 21) / 12]} = 13.23$ . So,

$$Z = (140 - 105) / 13.23 = 2.65$$

This Z-statistic corresponds to a p-value of  $\Pr(Z > 2.65) = 0.004$ . This is evidence to reject the null hypothesis that the populations are equal in reaction times.

**TIES**

When we find ties in our group of observations (two or more observations have the same value) how do we decide which gets the higher rank? We don't, and instead we assign the same rank to all of our ties, so that each one contributes the same amount to the sum of the ranks for its group. Therefore, we assign the **average of the all of the ranks that the ties occupy** to them.

For example,

Observation:	153	155	158	158	161	164
Rank:	1	2	3.5	3.5	5	6

So in this case, had the two 158s been different, they would have received ranks of 3 and 4. Since they are equal, they each are assigned a rank of the average of 3 and 4 (3.5) and the remaining numbers continue being ranked as if 3 and 4 were assigned.

**Example #2**

Rank this set of observations.

Sample 1

5    7    8    9

Sample 2

5    6    9    12

Solution:

The ranks are listed below the observations in parentheses:

Sample 1

5    7    8    9  
(1.5) (4)    (5) (6.5)

Sample 2

5    6    9    12  
(1.5) (3)    (6.5) (8)

Since 5 is the first rank, and it is listed twice, each observation is assigned the average of 1 and 2 (1.5). Also, 9 is listed twice. Since both nines come in the 6 and 7 ranks, we assign 6.5 as the ranks of these observations.

**WILCOXON SIGNED RANK TEST**

The Wilcoxon Signed Rank test is the nonparametric equivalent to the one-sample Z or t test and the matched pairs test. It is used when we want to make inferences about the **mean of one population** or the **mean difference between two populations** in a matched pairs setting.

Draw an SRS of size n from a population for a matched pairs study and for each pair find the difference between the two responses. Then rank the absolute value of the differences. Then group all of the positive differences and the negative differences separately. The sum of the ranks of the positive differences is W+, the Wilcoxon Signed Rank test statistic.

W+ has mean  $\mu_{w+} = \frac{n(n+1)}{4}$  and standard deviation  $\sigma_{w+} = \sqrt{\frac{n(n+1)(2n+1)}{24}}$ .

If there is a difference between our matched pairs, then we'd expect W+ to be far from its mean (or the expected value). To find the p-value, we need to standardize W+, which is

done by finding Z, where  $Z = \frac{W_+ - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}$ . Once we have Z, we can find the

appropriate p-value. When ties are present among the pairs, we assign average ranks. However, ties that occur within pairs, giving a difference equal to 0 for that pair, don't add anything to our test statistic and are therefore dropped from our sample.

**Example #3**

Eight subjects were asked to perform a simple puzzle assembly under normal conditions and under conditions of stress. During the stress condition the subjects were told that a mild shock would be delivered 3 minutes after the start of the experiment and every 30 seconds thereafter until the task was completed. Blood pressure readings were taken under both conditions. Data in the accompanying table represent the highest reading during the experiment. Do the data present sufficient evidence to indicate higher blood pressure readings during conditions of stress? Perform a significance test and clearly state your hypotheses, test statistic, p-value and conclusion.

Subject	1	2	3	4	5	6	7	8	
Normal		126	117	115	118	118	128	125	120
Stress		130	118	125	120	121	125	130	120

Solution:

Subject	1	2	3	4	5	6	7	8	
Normal		126	117	115	118	118	128	125	120
Stress		130	118	125	120	121	125	130	120
Difference		-4	-1	-10	-2	-3	3	-5	0
Abs Value		4	1	10	2	3	3	5	0
Rank		5	1	7	2	3.5	3.5	6	drop

Now, group the positive differences and the negative differences together:

Positive (rank in parentheses): 3 (3.5)

Negative: -4 (5), -1 (1), -10 (7), -2 (2), -3 (3.5), -5 (6)

Thus,  $W^+ = 3.5$ . Note that  $\mu_{w^+} = (8 \times 9) / 4 = 18$  and  $\sigma_{w^+} = \sqrt{[(8 \times 9 \times 17) / 24]} = 7.14$ . So, for our test, we have:

$H_0$ : There is no difference between normal and stress in indicating blood pressure

$H_a$ : The stress conditions promote higher blood pressure (normal conditions less)

Test Statistic:  $Z = (3.5 - 18) / 7.14 = -2.03$ . This Z-statistic corresponds to a p-value of  $\Pr(Z < -2.03) = 0.02$ .

This is evidence that the stress conditions promote higher blood pressure.

## **KRUSKAL-WALLIS TEST**

When we can assume that our data is normally distributed and that the population standard deviations are equal, we can test for a difference among several populations by using the One-way ANOVA F test. However, when our data is not normal, or we aren't sure if it is, we can use the nonparametric Kruskal-Wallis test to **compare more than two populations** as long as our data come from a continuous distribution.

In the One-way ANOVA F test, we are testing to see if our population means are equal. Since our data might not necessarily be symmetric in the nonparametric setting, it is better to use the median as the measure of center, and so in the Kruskal-Wallis test we are testing to see if our **population medians are equal**.

Recall the analysis of variance idea: we write the total observed variation in the responses as the sum of two parts, one measuring variation **among** the groups (sum of squares for groups, SSG) and one measuring variation among individual observations **within** the same group (sum of squares for error, SSE). The ANOVA F test rejects the null hypothesis that the mean responses are equal in all groups if SSG is large relative to SSE.

The idea of the Kruskal-Wallis rank test is to rank all the responses from all groups together and then **apply one-way ANOVA to the ranks rather than to the original observations**. If there are  $N$  observations in all, the ranks are always the whole numbers from 1 to  $N$ . The total sum of squares for the ranks is therefore a fixed number no matter what the data are. So we do not need to look at both SSG and SSE. Although it isn't obvious without some unpleasant algebra, the Kruskal-Wallis test statistic is essentially just SSG for the ranks. When SSG is large, that is evidence that the groups differ.

Draw independent SRSs of sizes  $n_1, n_2, \dots, n_I$  from  $I$  populations. There are  $N$  observations in all. Rank all  $N$  observations and let  $R_i$  be the sum of the ranks for the  $i^{\text{th}}$  sample. The

Kruskal-Wallis statistic is 
$$H = \frac{12}{N(N+1)} \sum \frac{R_i^2}{n_i} - 3(N+1)$$

When the sample sizes  $n_i$  are large and all  $I$  populations have the same continuous distribution,  $H$  has approximately the chi-square distribution with  $I-1$  degrees of freedom. The Kruskal-Wallis test rejects the null hypothesis that all populations have the same distribution when  $H$  is large.

So like the Wilcoxon rank sum statistic, the Kruskal-Wallis test statistic is based on the sums of the ranks for the groups we are comparing. The more different these sums are, the stronger is the evidence that responses are systematically larger in some groups than in others. As usual, we again assign average ranks to tied observations.

**Example #4**

A psychologist is trying to determine if there is a difference in three methods of training six-year-old children to learn a foreign language. A random selection of 10 six-year-old children with similar backgrounds is assigned to each of three different methods. Method 1 uses the traditional teaching format. Method 2 uses repeated listening to tapes of the language along with classroom instruction. Method 3 uses videotapes exclusively. At the end of a 6-week period, the children were given identical, standardized exams. The exams were scored with high scores indicating a better grasp of the language. Because of drop outs, method 1 had 7 students finishing, method 2 had 8, and method 3 only 6. It is, however, important to note that we must assume that children dropped out for reasons unrelated to performance. The data are given in the following table. Please conduct a significance test to determine if there is a difference between the three teaching methods, when we assume our data are not normally distributed.

Teaching Method		
1 - traditional	2 – tapes + classroom	3 - tapes
78	70	60
80	72	70
83	73	71
86	74	72
87	75	74
88	78	76
90	82	
	95	
$n_1=7$	$n_2=8$	$n_3=6$

Solution:

Here is the table with the ranks in parentheses:

Teaching Method		
1 - traditional	2 – tapes + classroom	3 - tapes
78 (12.5)	70 (2.5)	60 (1)
80 (14)	72 (5.5)	70 (2.5)
83 (16)	73 (7)	71 (4)
86 (17)	74 (8.5)	72 (5.5)
87 (18)	75 (10)	74 (8.5)
88 (19)	78 (12.5)	76 (11)
90 (20)	82 (15)	
	95 (21)	
$n_1=7$	$n_2=8$	$n_3=6$

$$R_1 = 12.5 + 14 + 16 + 17 + 18 + 19 + 20 = 116.5$$

$$R_2 = 2.5 + 5.5 + 7 + 8.5 + 10 + 12.5 + 15 + 21 = 82$$

$$R_3 = 1 + 2.5 + 4 + 5.5 + 8.5 + 11 = 32.5$$

$$H = [12 / (21 \times 22)] [(116.5^2/7) + (82^2/8) + (32.5^2/6)] - (3 \times 22) = (12 / 462)(2955.43) - 66 = 10.76$$

$H_0$ : All three populations have the same distribution

$H_a$ : The three populations are not all the same

Test Statistic:  $H = 10.76$  with  $df = 2$

This test statistic corresponds to a p-value of 0.005. We have evidence to reject the null hypothesis that the three populations have the same distribution.