7. Some non-parametric tests

Remember that you use statistics to make inferences about populations from samples. Most of the techniques statisticians use require that two assumptions are met. First, the population that the sample comes from is normal. Second, whenever means and variances were computed, the numbers in the data are "cardinal" or "interval", meaning that the value given an observation not only tells you which observation is larger or smaller, but how much larger or smaller. There are many situations when these assumptions are not met, and using the techniques developed so far will not be appropriate. Fortunately, statisticians have developed another set of statistical techniques, non-parametric statistics, for these situations. Three of these tests will be explained in this chapter. These three are the Mann-Whitney U-Test, which tests to see if two independently chosen samples come from populations with the same location; the Wilcoxon Rank Sum Test, which tests to see if two paired samples come from populations with the same location; and Spearman's Rank Correlation, which tests to see if two variables are related.

What does "non-parametric" mean?

To a statistician, a parameter is a measurable characteristic of a population. The population characteristics that usually interest statisticians are the location and the shape. Non-parametric statistics are used when the parameters of the population are not measurable or do not meet certain standards. In cases when the data only orders the observations, so that the interval between the observations is unknown, neither a mean nor a variance can be meaningfully computed. In such cases you need to use non-parametric tests. Because your sample does not have cardinal, or interval, data you cannot use it to estimate the mean or variance of the population, though you can make other inferences. Even if your data is cardinal, the population must be normal before the shape of the many sampling distributions are known. Fortunately, even if the population is not normal, such sampling distributions are usually close to the known shape if large samples are used. In that case, using the usual techniques is acceptable. However, if the samples are small and the population is not normal, you have to use non-parametric statistics. As you know, "there is no such thing as a free lunch". If you want to make an inference about a population without having cardinal data, or without knowing that the population is normal, or with very small samples, you will have to give up something. In general, non-parametric statistics are less precise than parametric statistics. Because you know less about the population you are trying to learn about, the inferences you make are less exact.

When either (1) the population is not normal and the samples are small, or (2) when the data is not cardinal, the same non-parametric statistics are used. Most of these tests involve ranking the members of the sample, and most involve comparing the ranking of two or more samples. Because we cannot compute meaningful sample statistics to compare to a hypothesized standard, we end up comparing two samples.

Do these populations have the same location? The Mann-Whitney U test.

In the chapter “T-test”, you learned how to test to see if two samples came from populations with the same mean by using the t-test. If your samples are small and you are not sure if the original populations are normal, or if your data does not measure intervals, you cannot use that t-test because the sample t-scores will not follow the sampling distribution in the t-table. Though there are two different data problems that keep you from using the t-test, the solution to both problems is the same, the non-parametric Mann-Whitney U test. The basic idea behind the test is to put the samples together, rank the members of the combined sample, and then see if the two samples are mixed together in the common ranking.

Once you have a single ranked list containing the members of both samples, you are ready to conduct a Mann-Whitney U test. This test is based on a simple idea. If the first part of the combined ranking is largely made up of members from one sample, and the last part is largely made up of members from the other sample, then the two samples are probably from populations with different ”averages” and therefore different locations. You can test to see if the members of one sample are lumped together or spread through the ranks by adding up the ranks of each of the two groups and comparing the sums. If these "rank sums" are about equal, the two groups are mixed together. If these ranks sums are far from equal, each of the samples is lumped together at the beginning or the end of the overall ranking.

Willy Senn works for Old North Gadgets, a maker and marketer of computer peripherals aimed at scientists, consultants, and college faculty. Old North's home office and production facilities are in a small town in the US state of Maine. While this is a nice place to work, the firm wants to expand its sales and needs a sales office in a location closer to potential customers and closer to a major airport. Willy has been given the task of deciding where that office should be. Before he starts to look at office buildings and airline schedules, he needs to decide if Old North's potential customers are in the east or the west. Willy finds an article in Fortune magazine that lists the best cities for finding "knowledge workers", Old North's customers. That article lists the ten best cities in the United States.

Rank	Metro Area	Region
1	Raleigh-Durham	East
2	New York	East
3	Boston	East
4	Seattle	West
5	Austin	West
6	Chicago	East
7	Houston	West
8	San Jose	West
9	Philadelphia	East
10	Minnesota-St Paul	East

Exhibit 11: Data for Willy's problem. From Kenneth Labich, "The Best Cities for Knowledge Workers," Fortune, 128:12, Nov. 15, 1993, pp. 50 ff.

Six of the top ten are in the east and four are in the west, but these ten represent only a sample of the market. It looks like the eastern places tend to be higher in the top ten, but is that really the case? If you add up the ranks, the six eastern cities have a "rank sum" of 31 while the western cities have a rank sum of 24, but there are more eastern cities and even if there were the same number would that difference be due to a different "average" in the rankings, or is it just due to sampling? The Mann-Whitney U test can tell you if the rank sum of 31 for the eastern cities is significantly less than would be expected if the two groups really were about the same and six of the ten in the sample happened to be from one group. The general formula for computing the Mann-Whitney U for the first of two groups is:

$$U_1 = n_1 n_2 + [n_1(n_1+1)]/2 - T_1$$

where:

T1 = the sum of the ranks of group 1.

n1 = the number of members of the sample from group 1

n2 = the number of members of the sample from group 2.

This formula seems strange at first, but a little careful thought will show you what is going on. The last third of the formula, $-T_1$, subtracts the rank sum of the group from the rest of the formula. What is the first two-thirds of the formula? The bigger the total of your two samples, and the more of that total that is in the first group, the bigger you would expect T_1 to be, everything else equal. Looking at the first two-thirds of the formula, you can see that the only variables in it are n_1 and n_2, the sizes of the two samples. The first two-thirds of the formula depends on the how big the total group is and how it is divided between the two samples. If either n_1 or n_2 gets larger, so does this part of the formula. The first two-thirds of the formula is the maximum value for T_1, the rank sum of group 1. T_1 will be at its maximum if the members of the first group were all at the bottom of the rankings for the combined samples. The U_1 score then is the difference between the actual rank sum and the maximum possible. A bigger U_1

means that the members of group 1 are bunched more at the top of the rankings and a smaller U_1 means that the members of group 1 are bunched near the bottom of the rankings so that the rank sum is close to its maximum. Obviously, a U-score can be computed for either group, so there is always a U_1 and a U_2. If U_1 is larger, U_2 is smaller for a given n_1 and n_2 because if T_1 is smaller, T_2 is larger.

What should Willy expect if the best cities are in one region rather than being evenly distributed across the country? If the best cities are evenly distributed, then the eastern group and the western group should have U's that are close together since neither group will have a T that is close to either its minimum or its maximum. If the one group is mostly at the top of the list, then that group will have a large U since its T will be small, and the other group will have a smaller U since its T will be large. $U_1 + U_2$ is always equal to n_1n_2 so either one can be used to test the hypothesis that the two groups come from the same population. Though there is always a pair of U-scores for any Mann-Whitney U-test, the published tables only show the smaller of the pair. Like all of the other tables you have used, this one shows what the sampling distribution of U's is like.

The sampling distribution, and this test, were first described by HB Mann and DR Whitney in 1947.[1] While you have to compute both U-scores, you only use the smaller one to test a two-tailed hypothesis. Because the tables only show the smaller U, you need to be careful when conducting a one-tail test. Because you will accept the alternative hypothesis if U is very small, you use the U computed for that sample which H_a: says is farther down the list. You are testing to see if one of the samples is located to the right of the other, so you test to see if the rank sum of that sample is large enough to make its U small enough to accept H_a:. If you learn to think through this formula, you will not have to memorize all of this detail because you will be able to figure out what to do.

Let us return to Willy 's problem. He needs to test to see if the best cities in which to locate the sales office, the best cities for finding "knowledge workers", are concentrated in one part of the country or not. He can attack his problem with a hypothesis test using the Mann-Whitney U-test. His hypotheses are:

H_o: The distributions of eastern and western city rankings among the "best places to find knowledge workers" are the same.

H_a: The distributions are different.

Looking at the table of Mann-Whitney values, he finds the following if one of the n's is 6:

	n_1			
U_0	**1**	**2**	**3**	**4**
0	0.1429	0.0357	0.0119	0.0005
1	0.2857	0.0714	0.0238	0.0095
2	0.4286	0.1429	0.0476	0.0190
3	0.5714	0.2143	0.0833	0.0333
4		0.4286	0.1310	0.0571
5		0.5714	0.1905	0.0857
6			0.2738	0.1286
7			0.3571	0.1762
8			0.4524	0.2381

1 ["On a test of whether one or two random variables is stochastically larger than the other." *Annals of Mathematical Statistics*, 18, 50-60.].

9		0.5476	0.3048
10			0.3810

Exhibit 12: Some lower-tail values for the Mann Whitney U statistic

The values in the table show what portion of the sampling distribution of U-statistics is in the lower tail, below the U value in the first column, if the null hypothesis is true. Willy decides to use an a = .10. Since he will decide that the data supports H_a if either the east or the west has a small U, Willy has a two-tail test and needs to divide his a between the two tails. He will choose H_a if either U is in the lowest .05 of the distribution. Going down the column for the other n equal to 4, Willy finds that if the null hypothesis is true, the probability that the smaller of the two U-scores will be 4 or less is only .0571, and probability that the lower U-score will be 3 or less is .0333. His half a of .05 is between these two, so he decides to be conservative and use as a decision rule to conclude that the data supports H_a: The distributions are different, if his sample U is less than 3 and that the data supports H_o: the distributions are the same, if his U is greater than or equal to 3. Now he computes his U, finding both U_e and U_w.

Remembering the formula from above, he finds his two U values::

For the eastern cities:

$$U_e = 6 \times 4 + \frac{6 \times 7}{2} - 31 = 14$$

For the western cities:

$$U_w = 6 \times 4 + \frac{4 \times 5}{2} - 24 = 10$$

The smaller of his two U-scores is U_w = 10.Because 10 is larger than 3, his decision rule tells him that the data supports the null hypothesis that eastern and western cities rank about the same. Willy decides that the sales office can be in either an eastern or western city, at least based on locating the office close to near large numbers of knowledge workers. The decision will depend on office cost and availability and airline schedules.

Testing with matched pairs: the Wilcoxon signed ranks test

During your career, you will often be interested in finding out if the same population is different in different situations. Do the same workers perform better after a training session? Do customers who used one of your products prefer the "new improved" version? Are the same characteristics important to different groups? When you are comparing the same group in two different situations, you have "matched pairs". For each member of the population or sample you have what happened under two different sets of conditions.

There is a non-parametric test using matched pairs that allows you to see if the location of the population is different in the different situations. This test is the Wilcoxon Signed Ranks Test. To understand the basis of this test, think about a group of subjects who are tested under two sets of conditions, A and B. Subtract the test score under B from the test score under A for each subject. Rank the subjects by the absolute size of that difference, and look to see if those who scored better under A are mostly lumped together at one end of your ranking. If most of the biggest absolute differences belong to subjects who scored higher under one of the sets of conditions, then the subjects probably perform differently under A than under B.

The details of how to perform this test were published by Frank Wilcoxon in 1945[2]. Wilcoxon found a method to find out if the subjects who scored better under one of the sets of conditions were lumped together or not. He also

2 "Individual comparisons by ranking methods", Biometrics,1,80-83

found the sampling distribution needed to test hypotheses based on the rankings. To use Wilcoxon's test, collect a sample of matched pairs. For each subject, find the difference in the outcome between the two sets of conditions and then rank the subjects according to the absolute value of the differences. Next, add together the ranks of those with negative differences and add together the ranks of those with positive differences. If these rank sums are about the same, then the subjects who did better under one set of conditions are mixed together with those who did better under the other condition, and there is no difference. If the rank sums are far apart, then there is a difference between the two sets of conditions.

Because the sum of the rank sums is always equal to [N(N-1)]/2], if you know the rank sum for either the positives or the negatives, you know it for the other. This means that you do not really have to compare the rank sums, you can simply look at the smallest and see if it is very small to see if the positive and negative differences are separated or mixed together. The sampling distribution of the smaller rank sums when the populations the samples come from are the same was published by Wilcoxon. A portion of a table showing this sampling distribution is in Exhibit 3. See below.

one-tail significance	0.05	0.025	0.01
two-tail significance	0.1	0.05	0.02
number of pairs, N			
5	0		
6	2	0	
7	3	2	0
8	5	3	1
9	8	5	3
10	10	8	5

Exhibit 13: Sampling distribution

Wendy Woodruff is the President of the Student Accounting Society at the University of North Carolina at Burlington (UNC-B). Wendy recently came across a study by Baker and McGregor ["Empirically Assessing the Utility of Accounting Student Characteristics", unpublished, 1993] in which both accounting firm partners and students were asked to score the importance of student characteristics in the hiring process. A summary of their findings is in Exhibit 11.

ATTRIBUTE	**Mean: student rating**	**Mean: big firm rating**
High Accounting GPA	2.06	2.56
High Overall GPA	0.08	-0.08
Communication Skills	4.15	4.25
Personal Integrity	4.27	7.5
Energy, drive, enthusiasm	4.82	3.15
Appearance	2.68	2.31

Exhibit 14: Data on importance of student attributes. From Baker and McGregor.

Wendy is wondering if the two groups think the same things are important. If the two groups think that different things are important, Wendy will need to have some society meetings devoted to discussing the differences. Wendy

has read over the article, and while she is not exactly sure how Baker and McGregor's scheme for rating the importance of student attributes works, she feels that the scores are probably not distributed normally. Her test to see if the groups rate the attributes differently will have to be non-parametric since the scores are not normally distributed and the samples are small.Wendy uses the Wilcoxon Signed Ranks Test.

Her hypotheses are:

H_o: There is no true difference between what students and Big 6 partners think is important.

H_s: There is a difference.

She decides to use a level of significance of .05. Wendy's test is a two-tail test because she wants to see if the scores are different, not if the Big 6 partners value these things more highly. Looking at the table, she finds that, for a two-tail test, the smaller of the two sum of ranks must be less than or equal to 2 to accept H_a:.

Wendy finds the differences between student and Big 6 scores, and ranks the absolute differences, keeping track of which are negative and which are positive. She then sums the positive ranks and sum the negative ranks. Her work is shown below:

ATTRIBUTE	Mean student rating	Mean big firm rating	Difference	Rank
High Accounting GPA	2.06	2.56	-0.5	-4
High Overall GPA	0.08	-0.08	0.16	2
Communication Skills	4.15	4.25	-0.1	-1
Personal Integrity	4.27	7.5	-2.75	-6
Energy, drive, enthusiasm	4.82	3.15	1.67	5
Appearance	2.68	2.31	0.37	3

sum of positive ranks = 4+5+3=10

sum of negative ranks = 4+1=6=11

number of pairs=6

Exhibit 15: The worksheet for the Wilcoxon Signed Ranks Test

Her sample statistic, T, is the smaller of the two sums of ranks, so T=10. According to her decision rule to accept H_a: if T < 2, she decides that the data supports H_o: that there is no difference in what students and Big 6 firms think is important to look for when hiring students. This makes sense, because the attributes that students score as more important, those with positive differences, and those that the Big 6 score as more important, those with negative differences, are mixed together when the absolute values of the differences are ranked. Notice that using the rankings of the differences rather than the size of the differences reduces the importance of the large difference between the importance students and Big 6 partners place on Personal integrity. This is one of the costs of using non-parametric statistics. The Student Accounting Society at UNC-B does not need to have a major program on what accounting firms look for in hiring. However, Wendy thinks that the discrepancy in the importance in hiring placed on Personal Integrity by Big 6 firms and the students means that she needs to schedule a speaker on that subject. Wendy wisely tempers her statistical finding with some common sense.

Are these two variables related? Spearman's rank correlation

Are sales higher in those geographic areas where more is spent on advertising? Does spending more on preventive maintenance reduce down-time? Are production workers with more seniority assigned the most popular jobs? All of these questions ask how the two variables move up and down together; when one goes up, does the other also rise? when one goes up does the other go down? Does the level of one have no effect on the level of the other? Statisticians measure the way two variables move together by measuring the **correlation coefficient** between the two.

Correlation will be discussed again in the next chapter, but it will not hurt to hear about the idea behind it twice. The basic idea is to measure how well two variables are tied together. Simply looking at the word, you can see that it means co-related. If whenever variable X goes up by 1, variable Y changes by a set amount, then X and Y are perfectly tied together, and a statistician would say that they are perfectly correlated. Measuring correlation usually requires interval data from normal populations, but a procedure to measure correlation from ranked data has been developed. Regular correlation coefficients range from -1 to +1. The sign tells you if the two variables move in the same direction (positive correlation) or in opposite directions (negative correlation) as they change together. The absolute value of the correlation coefficient tells you how closely tied together the variables are; a correlation coefficient close to +1 or to -1 means they are closely tied together, a correlation coefficient close to 0 means that they are not very closely tied together. The non-parametric Spearman's Rank Correlation Coefficient is scaled so that it follows these same conventions.

The true formula for computing the Spearman's Rank Correlation Coefficient is complex. Most people using rank correlation compute the coefficient with a computer program, but looking at the equation will help you see how Spearman's Rank Correlation works. It is:

$$r_s = 1 - \left(\frac{6}{n(n^2-1)}\right)\left(\sum d^2\right)$$

where:

n = the number of observations

d = the difference between the ranks for an observation

Keep in mind that we want this non-parametric correlation coefficient to range from -1 to +1 so that it acts like the parametric correlation coefficient. Now look at the equation. For a given sample size, n, the only thing that will vary is Σd^2. If the samples are perfectly positively correlated, then the same observation will be ranked first for both variables, another observation ranked second for both variables, etc. That means that each difference in ranks, d, will be zero, the numerator of the fraction at the end of the equation will be zero, and that fraction will be zero. Subtracting zero from one leaves one, so if the observations are ranked in the same order by both variables, the Spearman's Rank Correlation Coefficient is +1. Similarly, if the observations are ranked in exactly the opposite order by the two variables, there will many large d^2's, and Σd^2 will be at its maximum. The rank correlation coefficient should equal -1, so you want to subtract 2 from 1 in the equation. The middle part of the equation, $6/n(n^2-1)$, simply scales Σd^2 so that the whole term equals 2. As n grows larger, Σd^2 will grow larger if the two variables produce exactly opposite rankings. At the same time, $n(n^2-1)$ will grow larger so that $6/n(n^2-1)$ will grow smaller.

Colonial Milling Company produces flour, corn meal, grits, and muffin, cake, and quickbread mixes. They are considering introducing a new product, Instant Cheese Grits mix. Cheese grits is a dish made by cooking grits, combining the cooked grits with cheese and eggs, and then baking the mixture. It is a southern favorite in the United States, but because it takes a long time to cook, is not served much anymore. The Colonial mix will allow someone to prepare cheese grits in 20 minutes in only one pan, so if it tastes right, it should be a good-selling product in the South. Sandy Owens is the product manager for Instant Cheese Grits, and is deciding what kind of cheese flavoring to use. Nine different cheese flavorings have been successfully tested in production, and samples made with each of those nine flavorings have been rated by two groups: first, a group of food experts, and second, a group of potential customers. The group of experts was given a taste of three dishes of "homemade" cheese grits and ranked the samples according to how well they matched the real thing. The customers were given the samples and asked to rank them according to how much they tasted like "real cheese grits should taste". Over time, Colonial has found that using experts is a better way of identifying the flavorings that will make a successful product, but they always check the experts' opinion against a panel of customers. Sandy must decide if the experts and customers basically agree. If they do, then she will use the flavoring rated first by the experts. The data from the taste tests is in Exhibit 13.

	Expert ranking	**Consumer ranking**
Flavoring		
NYS21	7	8
K73	4	3
K88	1	4
Ba4	8	6
Bc11	2	5
McA A	3	1
McA A	9	9
WIS 4	5	2
WIS 43	6	7

Exhibit 16: Data from two taste tests of cheese flavorings

Sandy decides to use the SAS statistical software that Colonial has purchased. Her hypotheses are:

H_o: The correlation between the expert and consumer rankings is zero or negative.

H_a: The correlation is positive.

Sandy will decide that the expert panel does know best if the data supports H_a: that there is a positive correlation between the experts and the consumers. She goes to a table that shows what value of the Spearman's Rank Correlation Coefficient will separate one tail from the rest of the sampling distribution if there is no association in the population. A portion of such a table is in Exhibit 12.

n	α=.05	α=.025	α=.10
5	0.9		
6	0.829	0.886	0.943
7	0.714	0.786	0.893
8	0.643	0.738	0.833
9	0.6	0.683	0.783
10	0.564	0.648	0.745
11	0.523	0.623	0.736
12	0.497	0.591	0.703

Exhibit 17: Some one-tail critical values for Spearman's Rank Correlation Coefficient

Using a = .05, going across the n = 9 row in Exhibit 12, Sandy sees that if H_o: is true, only .05 of all samples will have an r_s greater than .600. Sandy decides that if her sample rank correlation is greater than .600, the data supports the alternative, and flavoring K88, the one ranked highest by the experts, will be used. She first goes basck to the two sets of rankings and finds the difference in the rank given each flavor by the two groups, squares those differences and adds them together:

	Expert ranking	Consumer ranking	difference	d^2
Flavoring				
NYS21	7	8	-1	1
K73	4	3	1	1
K88	1	4	-3	9
Ba4	8	6	2	4
Bc11	2	5	-3	9
McA A	3	1	2	4
McA A	9	9	0	0
WIS 4	5	2	3	9
WIS 43	6	7	-1	1
			sum =	38

Exhibit 18: Sandy's worksheet

Then she uses the formula from above to find her Spearman rank correlation coefficient:

$1- [6/(9)(9^2-1)][38] = 1 - .3166 = .6834$

Her sample correlation coefficient is .6834, greater than .600, so she decides that the experts are reliable, and decides to use flavoring K88. Even though Sandy has ordinal data that only ranks the flavorings, she can still perform a valid statistical test to see if the experts are reliable. Statistics has helped another manager make a decision.

Summary

Though they are less precise than other statistics, non-parametric statistics are useful. You will find yourself faced with small samples, populations that are obviously not normal, and data that is not cardinal. At those times, you can still make inferences about populations from samples by using non-parametric statistics.

Non-parametric statistical methods are also useful because they can often be used without a computer, or even a calculator. The Mann-Whitney U, and the t-test for the difference of sample means, test the same thing. You can usually perform the U-test without any computational help, while performing a t-test without at least a good calculator can take a lot of time. Similarly, the Wilcoxon Signed Ranks test and Spearman's Rank Correlation are easy to compute once the data has been carefully ranked. Though you should proceed on to the parametric statistics when you have access to a computer or calculator, in a pinch you can use non-parametric methods for a rough estimate.

Notice that each different non-parametric test has its own table. When your data is not cardinal, or your populations are not normal, the sampling distributions of each statistic is different. The common distributions, the t, the χ^2, and the F, cannot be used.

Non-parametric statistics have their place. They do not require that we know as much about the population, or that the data measure as much about the observations. Even though they are less precise, they are often very useful.