Lecture 17

Review Exam #2

Non-parametric tests
Sometimes these tests are referred to as distribution free, or non-parametric tests. They try to analyze claims of a hypothesis test without the use of specific distributions (like the normal distribution or F distribution) as we've done to date. Some parameters don't have clear theoretical distributions. We've seen briefly how we can analyze these computationally using simulations and bootstrapping (and we'll return to more than before the end of this course), but these tests use other procedures, like rank, to test the claims and thus are easier to do without the aid of a computer.

Recall that rank is the position of an observation in an ordered list. Percentile expresses a similar idea but as a percentage of the list, where the 0^{th} percentile is the minimum and the 100^{th} percentile is the maximum regardless of the length of the list of observations. So, it is equivalent to a relative rank.

For our tests, we will generally adopt the strategy that if two observations are the same value, then both observations will receive the average of the two ranks. If we deviate from that practice, it will be noted.

The first test we are going to look at is the **Wilcoxon signed-rank test**. This test makes a slightly weaker assumption than our previous tests: we don't assume normality, but we do assume that the distribution is symmetric. It is a test of the median of our data. Specifically, we are going to test whether or not the median is zero, i.e. $H_0: \tilde{\mu} = 0, H_a: \tilde{\mu} \neq 0$.

This test can be used for a single sample, or it can be used in place of a paired t-test. The test uses the W-distribution, which we haven't encountered before, but which approximates the normal distribution when the sample size is bigger than 10. To conduct the test, if the mean or median we are comparing the data to is not zero, we can translate the data by subtracting the mean or median of the null hypothesis from all the observations, and then carry out the test from that point.

Step 0: For the paired test, find the differences.

Step 1: (single sample data starts here). Take the absolute value of all observations (assuming that you are comparing to 0 in the null hypothesis). Then rank the results.

Step 2: Collect the sum of the ranks of observations that were originally positive. Collect the ranks of the observations that were originally negative. These give you W_+ and W_- (or you will see different notations for these sums, sometimes with s for sum).

Double check your sums. Adding the two values back together should give you $\frac{n(n+1)}{2}$ where n is the sample size.

When the sample sizes are small, you can list out all the possible combinations of combinations of signs and calculate the probability by matching up your sum to the other possible ways of obtaining the same sum. When the sample size is bigger than 10, then we can use the approximate normality to estimate the rest of the test. There is an expected mean for the given sample size and standard deviation. We must make an adjustment to the standard deviation if there are ties.

The mean and standard deviation formulas depend on the sample size.

$$\mu_w = \frac{n(n+1)}{4}, \ \sigma_w = \sqrt{\frac{n(n+1)(2n+1)}{24}}$$

The adjustment for ties is

$$\frac{t^3 - t}{48}$$

Where t is the rank where the tie occurred. So, if there are multiple ties, we will need to make multiple adjustments.

Using the normal approximation, we have a test statistic of $z = \frac{Max(W_+, W_-) - \mu_w}{\sqrt{\sigma_w^2 - adj_{ties}}}$.
This version of the formula has you choose the maximum of your two sums, but you can just pick either side to do for the two-tailed test. It's also worth noting that the adjustment for ties is usually small and won't affect the outcome much unless your P-value is close to the significance level, or unless there are a lot of ties.

Example. A manufacturer of electric irons, wishing to test the accuracy of the thermostat control at the 500°F setting, instructs a test engineer to obtain actual temperatures at that setting for 15 irons using a thermocouple. The resulting measurements are as follows: 494.6, 510.8, 487.5, 493.2, 502.6, 485.0, 495.9, 498.2, 501.6, 497.3, 492.0, 504.3, 499.2, 493.5, 505.8.

The manufacturer is going to conduct a test of $H_0: \mu = 500, H_a: \mu \neq 500$.

Recall that since the mean is not zero, we subtract off 500 from all the observations. Then rank the absolute values.

Absolute Magnitude	.8	1.6	1.8	2.6	2.7	4.1	4.3	5.6	5.8	6.5	6.8	8.0	10.8	12.5	15.0
Rank	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Sign	−	+	−	+	−	−	+	−	+	−	−	−	+	−	−

Collect the ranks that were originally less than the mean, and those that were more than the mean, then add them up. Calculate your test statistic. You won't need to do all this by hand fortunately, since this is something R can do for us. You can either subtract the null hypothesis off your data yourself, or just specify the mean or median value in your null in the wilcox.test() function.

R produces the following output:

Wilcoxon signed rank exact test

data: x

V = 35, p-value = 0.1688
alternative hypothesis: true location is not equal to 500

Here, R tells us that the P-value is greater than 0.05 and so we fail to reject the null hypothesis. There is not good reason to think the mean temperature is not 500.

A good question to ask at this point is why have a test like this at all? One possible reason is the weaker standard for symmetry rather than normality. Another is that it is capable of testing medians, which we could not do before using non-computer-based methods. Another might be computational efficiency. Calculating a standard deviation for the sample requires a lot of steps that might be avoided with a test like this.

Another situation we might wonder about is whether it makes sense to choose this non-parametric test when the distribution is normal? Is there any loss from using the signed rank test? We could frame this as asking is there a higher or lower (or the same) chance of a Type II error? Alternatively, we could ask if the underlying distribution is clearly not normal, why kind of improvement is achieved with this test?

Unfortunately, these questions are not easy to answer, because obtaining β for any underlying distribution is not well-established mathematically. The mean can be substantially impacted distributions with heavy tails (they can make the mean more unstable).

One method of comparing the two tests is called asymptotic relative efficiency (ARE). This method uses the limiting ratio of the sample sizes necessary to obtain identical error probabilities for the two tests. The ARE results are such that if the underlying distribution is normal, the Wilcoxon test compared to the t-test is about 0.95. For other distributions, the comparison could be much higher than 1. Thus, it appears that Wilcoxon will outperform the t-test more when the distribution is very different from normal. When it is normal, they are approximately similar.

Another test we are going to consider is the **Wilcoxon rank-sum test**, or also known as the Mann-Whitney test. While these tests are basically the same, the **Mann-Whitney** test notation can be different. This test also applies to a large number of distributions, and thus is considered to be distribution-free.

In this test we are dealing with two independent samples with similarly shaped distributions with the main difference between them being in their means. While the general null hypothesis for this test is $H_0: \mu_1 - \mu_2 = \Delta_0$, test procedures recommend testing with $\Delta_0 = 0$ initially. We will consider for the 0-difference test. If there is a fixed difference being tested, then as with the previous test, subtract that difference from the set with the larger mean before proceeding.

Combine the data in both samples into a single set and then find the ranks of the observations of the combined sample. Then add up the ranks in each of the two original samples. This gives our (two) test statistic(s), W. There is a theoretical minimum and maximum values of W. If there is no overlap in the data at all, then all the smallest ranks will be in one sample, and the largest ones in the other. If there are m observations in the first sample, then the smallest possible sum is $\frac{m(m+1)}{2}$. If the largest samples are all in set two, and there are n observations in the second sample, the largest possible sum is $\frac{(m+n)(m+n+1)}{2} - \frac{m(m+1)}{2}$. If the sample sizes are very small, one calculates the probability by listing out all the possible outcomes, and finding the probability of obtaining the outcome you have (or more

extreme) and then dividing by the number of total possible outcomes. As the sample sizes grow ($n, m > 8$), then the distribution of the test statistic can be approximated as normal. The mean and variance of the W statistic are given by

$$\mu_W = \frac{m(m+n+1)}{2}$$

$$\sigma_W = \sqrt{\frac{mn(m+n+1)}{12}}$$

$$z = \frac{W - \mu_W}{\sqrt{\frac{mn(m+n+1)}{12}}}$$

As with our previous test, if there are ties, the σ will need to be adjusted to account for the reduced variability.

$$\sigma_W = \sqrt{\frac{mn(m+n+1)}{12} - \frac{mn}{12(m+n)(m+n-1)}\sum(t^3 - t)}$$

Where t is the positions where the ties occurred. This will not have a large impact on the test statistic unless there are a lot of ties. This is much less likely in continuous data than in discrete data.

The efficiency of this test is similar to other Wilcoxon test discussed and so when the distributions are very non-normal, this test is to be preferred, but will never have an ARE of less than 0.86 even when the distributions are both normal with equal variances.

Example.
A study of guinea pigs tests the effects of orange juice vs. synthetic ascorbic acid on odontoblasts. The data is below.

Orange Juice	8.2	9.4	9.6	9.7	10.0	14.5	15.2	16.1	17.6	21.5
Ascorbic Acid	4.2	5.2	5.8	6.4	7.0	7.3	10.1	11.2	11.3	11.5

We'll conduct a Wilcoxon Rank-Sum test using R. It's the same wilcox.test() function we used before, but include 2 variables x and y, and set paired=FALSE. We want to know if the effect is different, so we can use the default null hypothesis value of 0 difference.

We get the following output:

Wilcoxon rank sum exact test

data: x and y
W = 80, p-value = 0.02323
alternative hypothesis: true location shift is not equal to 0

If we use a standard significance level of 0.05, we can reject the null hypothesis. If we wanted a more rigorous significance level of 0.01, we would fail to reject the null.

Next time we'll talk more about confidence intervals and other types of non-parametric tests.

References:

1. https://faculty.ksu.edu.sa/sites/default/files/probability_and_statistics_for_engineering_and_the_sciences.pdf
2. https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/how-to-conduct-the-wilcox-sign-test/
3. http://www.r-tutor.com/elementary-statistics/non-parametric-methods/wilcoxon-signed-rank-test
4. https://onlinestatbook.com/2/distribution_free_tests/benefits.html
5. https://onlinestatbook.com/2/distribution_free_tests/rank_two.html