Lecture 6

Nonparametric procedures for comparing two distributions

In this lecture we will discuss nonparametric comparisons of two distributions. Nonparametric tests and confidence intervals often involve inference about the median rather than the mean. They also often do things like replace the values of the data with their ranks which minimizes the influence of outliers. However, it is not fair to say that these procedures have no assumptions. For the procedure discussed in this lecture, under some circumstances, the conclusions about the median are only valid if the two population distributions have the same shape (in particular, the same variance - a rather restrictive assumption.) Nonetheless, nonparametric procedures provide an alternative to t-test that are useful when the t-tests are unlikely to give correct results. Since Welch's t-test is fairly robust, the circumstances we are talking about typically involve relatively small sample sizes and relatively large amounts of skewness.

Mann-Whitney Test

The Mann-Whitney test, also called the Mann-Whitney-Wilcoxon test or the Wilcoxon rank sum test), is a nonparametric test of the equality of medians under the assumption that the two population distributions otherwise have the same shape (that is, the same variance and the same skewness, etc...) The test replaces the data with their ranks based on the pool data. If the medians are equal, each set of data should get their relative proportion of the ranks. If either data set has a sum of ranks that is too large (or too small), it indicates that the medians are not equal.

Let your two data sets be denoted using x and y (they do not have to be the same size).

Data:

X	$\rightarrow$	Ranks	Y	$\rightarrow$	Ranks
x_1	$\rightarrow$	R_1^x	y_1	$\rightarrow$	R_1^y
x_2	$\rightarrow$	R_2^x	y_2	$\rightarrow$	R_2^y
$\vdots$		$\vdots$	$\vdots$		$\vdots$
x_{n_x}	$\rightarrow$	$R_{n_x}^x$	$\vdots$		$\vdots$
			y_{n_y}	$\rightarrow$	$R_{n_y}^y$

The ranks are based on the pooled data and ties are given the average rank for the group of tied values. Mann-Whitney can handle some ties, but the test will not be valid if there are too many of them. The following is an example to illustrate the assignment of ranks:

Data:

X	$\rightarrow$	Ranks	Y	$\rightarrow$	Ranks
13	$\rightarrow$	3	2	$\rightarrow$	1
20	$\rightarrow$	4.5	20	$\rightarrow$	4.5
4	$\rightarrow$	2			

The statistic can be the sum of the ranks of either column. Let $T_x = \sum R_i^x$. Then the expectation (or mean) and the standard error of this statistic are give by the formulas:

$$\mathrm{E}(T_x) = \frac{n_x(n_x+n_y+1)}{2} \qquad \mathrm{S.E.}(T_x) = \sqrt{\frac{n_x n_y(n_x+n_y+1)}{12}}$$

and the normalized statistic approximately follows a standard normal distribution

$$\frac{T_x - \mathrm{E}(T_x)}{\mathrm{S.E.}(T_x)} \approx Z(0,1)$$

Sometimes, there is a continuity correction in the numerator as follows which works well for small sample sizes:

$$\frac{|T_x - \mathrm{E}(T_x)| - 0.5}{\mathrm{S.E.}(T_x)} \approx |Z(0,1)|$$

For the mathematically inclined, the above formulas are developed by considering the random variable

$$Z_{i,j} = \begin{cases} 1 \text{ if } X_i < Y_j \\ 0 \text{ otherwise} \end{cases}$$

and noting that $T_x = \sum Z_{i,j} + \frac{n_x(n_x+1)}{2}$, and that, under the null hypothesis of equal medians, $EZ_{i,j} = 1/2$, $\mathrm{Var}(Z_{i,j}) = 1/12$, and

$$\mathrm{Cov}(Z_{i,j}, Z_{k,l}) = \begin{cases} 0 \text{ if } i \neq k, j \neq l \\ 1/12 \text{ if } i = k \text{ and } j \neq l \text{ or if } i \neq k \text{ and } j = l \\ 1/4 \text{ if } i = k, j = l \end{cases}$$

Suppose that you wanted to test that the median of X was d more than the median of Y. You could subtract d from all the X data and then use the above method to test if the medians were then the same.

Confidence intervals can also be developed based on the ranks; however, exactly 95% (or any other specified percent) coverage may not be achievable. The software will come as close a possible and provide the percent coverage actually achieved. A 95% confidence interval is the set of d so that the test of whether the difference in medians is d or not is not rejected at the 5% level. There are ranges of d that won't change the ranks. Then when d crosses a boundary that does change the ranks, the significance of the test jumps. Thus, an exact, pre=specified level of coverage may not be obtained.

Remember, the Mann-Whitney procedure tests the hypothesis that the two distributions are equal. One can conclude that the medians are not likely to be equal ONLY if everything else about the two distributions (variance, skewness, etc...) are the same. We will explore how robust the test is the violations of equal shapes below.

Example 1 The following data are the amount of nickel (in units of micrograms per 100 grams of dry weight) in the lungs of Legionnaire victims and in control cases (originally from Chen et al. 1977, used in Biostatistics by van Belle et al.):

Legionnaire	Control
65	12
24	10
52	31
86	6
120	5
82	5
399	29
87	9
139	12

In 1976, there was a mysterious outbreak of illness among American Legionnaire's at their convention in Philadelphia. No one knew what was causing it and a number of Legionnaires died. One thought was that it was caused by heavy metal poisoning and initial analyses supported nickel poisoning. That was wrong - the nickel in the lungs of the Legionnaire patients autopsied was likely due to contamination with metal instruments used in the autopsy. Ultimately, the source of the disease was found - it was a bacterium that had been identified several time previously and was likely transmitted in water vapor in the hotel's air conditioning system. Chen *et al.*'s article confirms an increase in nickel in the lungs of the Legionnaire victims but not in their kidneys and suggests a model for the contamination.

Before applying the Mann-Whitney procedure, we might note that the Legionnaire data looks more skewed than the Control data. However, two analyses lead us to conclude otherwise. First, a straightforward test of equal variances (in Minitab under Stat > Basic Statistics > 2 Variances...) leads us to conclude that there is no evidence of unequal variances using Levene's test. Note that Bartlett's test is completely inappropriate for these data. Also, a log transform restores normality and preserves the ranks. Since there is no evidence whatsoever for unequal variances in the log-transformed data, the illusion of more skewness in the Legionnaire data is only due to the inherent skewness in the distributions leading to outliers.

Mann-Whitney may not be the most appropriate statistic for these data but it is an acceptable one whose assumptions are met. Software will conduct the test and form the confidence interval for us. In Minitab, Mann-Whitney is found under Stat > Nonparametrics > Mann-Whitney...

A write up of the results might look like the following:

We consider the question of whether nickel levels in 9 Legionnaire patients were the same as in 9 control patients or not. The test of equality was rejected with a p-value of 0.0008 (based on a two-sided Mann-Whitney test). A 95.8% confidence interval for the difference in medians indicates that Legionnaire patients have 46.0 to 115.0 μg/100 g dry weight more nickel in their lungs than the control patients. Obviously, the ultimate suggestion that Legionnaire's disease was caused by heavy metal poisoning was false but we do not have enough information to comment on the experimental design ourselves. ■

We can explore how sensitive the Mann-Whitney test is to violations that the two distributions have the same shape as follows. Find 2 distributions with the same median but different shapes. Repeatedly simulate two data sets, one from each distribution. Evaluate the Mann-Whitney test statistic and record whether it is significant or not. Since the distributions have the same median, we want the test to be significant only $\alpha \times 100$ percent of the time when the cut-off is found putting α in the tails. Let's use $\alpha = 0.05$, a typical significance level. We want to reject the null hypothesis (which is making a mistake, a Type I error) 5% of the time. If we reject much more or much less often, the test is not achieving its reported type I error rate and that is a very bad thing.

The following code is a template for a simulation determining robustness. Initially, it is a small simulation with only 100 trials, with two normal populations with the same mean (median) and the same variance and with the same sample size. Unless we have a mistake in the code, the variation we see from 0.05 is simply due to chance. We can vary each of these things and will vary some of them in class. Results further from 0.05 than we saw by chance tell us about the robustness of the test.

Recall that you need to put this code in a file with extension ".MAC" and you need to store it and the Minitab file in the same folder.

```
GMACRO
MannWhitneyRobustness
Name k1 "loop"
Name k2 "SampleSize1"
Name k3 "SampleSize2"
Name k4  "W"
Name k5  "Z"

Let SampleSize1 = 10
Let SampleSize2 = 10

Do k1 = 1:100
Random  SampleSize1  c1;
Normal 0.0 1.0.
Random SampleSize2 c2;
Normal 0.0 1.0.
Stack c1 c2 c4;
subscripts c3.
rank c4 c5
unstack c5 c6 c7;
subscripts c3.
Let W= Sum(c6)
Let Z = abs((W  - SampleSize1*(SampleSize1 + SampleSize2 + 1)/2)) - 0.5
Let Z = Z/Sqrt(SampleSize1*SampleSize2*(SampleSize1 + SampleSize2 + 1)/12)

if Z> 1.96
let c10(k1) = 1
else
```

```
let c10(k1) = 0
endif
enddo
Let c11(1) = sum(c10)/100
ENDMACRO
```

Through exploration and modification of this code, I believe we find that the procedure is robust to violations of equal shapes when the sample sizes are equal and is not robust to such violations when the sample sizes are unequal.

References and Readings

[1] Lawrence K. Altman. In philadelphia 30 years ago, an eruption of illness and fear. *New York Times*, page F1, August 1 2006.

[2] J. R. Chen, R. B. Francisco, and T. E. Miller. Legionnaires' disease: nickel levels. *Science*, 196:906–908, 1977.

[3] Gerald van Belle, Lloyd Fisher, Patrick J. Heagerty, and Thomas Lumley. *Biostatistics: A Methodology for the Health Sciences*. John Wiley and Sons, Inc., 2^{nd} edition, 2004.

Exercises for Lecture 6

1. Is a 2-sample t-test on the log-transformed data another viable option for the analysis of the Legionnaire data? If so, how do the results of this option compare to the results of the Mann-Whitney analysis? Specifically, how well do the confidence intervals for the two methods match up? Be careful about the different interpretation of the two intervals.

2. How do you modify the code given to use 1000 simulations instead of only 100. Hint: The code needs to be modified in two places.

3. How do you modify the code given so that the first sample size is 7 and the second sample size is 14? How do unequal sample sizes affect the robustness?