STAT 515 – Section 14.3 Supplement

Brian Habing – University of South Carolina
Last Updated: June 12, 2017

S14.3 - More on Nonparametric Methods for Two Independent Samples

Section 14.3 in the on-line chapter of the text book introduces the Wilcoxon rank sum test. This suplement fills in a few details about this test, including: other names it goes by, what null hypothesis it actually tests, how the p-values are calculated, and how it compares to the t-test in terms of power. It also introduces the Randomization (aka Permutation or Resampling) tests.

S14.3.1 - The Mann-Whitney-Wilcoxon rank sum test

The Wilcoxon rank sum test was proposed by Wilcoxon in 1945 with an example where the two samples have equal sample sizes. Mann and Whitney developed the test more generally in 1947, and so it is now known by many names: the rank sum test, Wilcoxon rank sum test, Mann-Whitney *U* test, Mann-Whitney Wilcoxon test, and Wilcoxon, Mann-Whitney test! Kruskal in 1957 noted that the German mathematician Deuchler came up with the idea way back in 1914, but it doesn't seem to have been widely noticed – and much of his subsequent work wasn't published during his lifetime. Adding a third name doesn't seem to have caught on, and we'll just abbreviate it as the MWW test.

Wilcoxon originally proposed using the statistic our text uses where *T*= the smaller sum of ranks. (although Wilcoxon doesn't use the name *T*, it is given to it by Mann and Whitney). Mann and Whitney prefer to use the statistic :

$$U = n_1 n_2 + \frac{n_2(n_2+1)}{2} - T_{2nd}$$

where T_{2nd} is the sum of the ranks of the second sample and *U* counts the number of times observations in the second sample are less than an observation in the first sample. For example, if the two samples were (1, 3, 7) and (2, 5) then 2 is less than 3 and 7, and 5 is less than 7, giving $U = 3$. $T=2+4=6$ in this case and you can check that $U=3\cdot2+2(2+1)/2 - 6$ does indeed equal 3.

The reason they choose to go with this U statistic is that it is related to what the chances are that an observation that you randomly pick from the second population will be less than one you randomly pick from the first. A small value of U means that the first group "tends to be smaller". This *U* statistics is the one reported by the R function `Wilcox.test` (and it is called *W* in the output). Because either the *T* or *U* statistic can be used to solve for the other, it doesn't matter in practice which is used.

S14.3.2 – What does "Shifted" mean?

Statistics texts often take different approaches to describing the hypothesis tested by the MWW test. The reasons for this are hinted at by the note and foot-note on page 14-12.

The null hypothesis that the MWW test actually uses is that the probability a randomly chosen observation X from the first sample is larger than a randomly chosen observation Y from the second

sample, is equal to the chance that X is less than Y. That is, H_0: P(X>Y) = P(X<Y). If we imagine we were choose one person from each of the two populations and compare their heights, the null hypothesis would mean that each of the groups would "win" 50% of the time. The three possible alternate hypotheses are then:

H_A: P(X>Y) < P(X<Y) or "X tends to be smaller than Y",

H_A: P(X>Y) > P(X<Y) or "X tends to be larger than Y",

or H_A: P(X>Y) ≠ P(X<Y) or "either X tends to be larger than Y,
or Y tends to be larger than X".

In general, if we're using assumption that we have two independent i.i.d. continuous distributions (like on page 14-12) the MWW isn't necessarily testing anything about the mean or median. Books that write the alternate hypotheses in terms of the means or medians need to make another assumption – that the two distributions are either the same or that one of them is simply shifted over from the other (and they have the same shape).

S14.3.3 – How are the p-values calculated?

Consider a small example where we have two populations X and Y where we want to test that X is shifted to the left of Y (or X tends to be smaller than Y). That is, H_0: P(X>Y) = P(X<Y) and
H_A: P(X>Y) < P(X<Y).

Assume our data consists of two independent simple random samples:

X	0	2	
Y	1	4	7

Putting the ranks in gives:

X	0^1	2^3	
Y	1^2	4^4	7^5

and the possible test statistics are T_1=1+3=4 and T_2=2+4+5=11. The smaller of these two values is the value 4 and is the one that goes with X.

Since we're worried about X being smaller, the p-value would be P(observing $T_1 \leq 4 \mid H_0$ is true). Since H_0 is that a randomly chosen X and Y have equal chance of being biggest, H0 is that the ranks of the Xs and Ys are random. That is, each of the following sets of sorted rankings for X and Y should be equally likely:

X ranks		Y ranks			P(ranking)	*T* statistic for group X	W
1	2	3	4	5	$\frac{1}{10} = 0.1$	3	0
1	3	2	4	5	$\frac{1}{10} = 0.1$	4	1
1	4	2	3	5	$\frac{1}{10} = 0.1$	5	2
2	3	1	4	5	$\frac{1}{10} = 0.1$	5	2
1	5	2	3	4	$\frac{1}{10} = 0.1$	6	3
2	4	1	3	5	$\frac{1}{10} = 0.1$	6	3
2	5	1	3	4	$\frac{1}{10} = 0.1$	7	4
3	4	1	2	5	$\frac{1}{10} = 0.1$	7	4
3	5	1	2	4	$\frac{1}{10} = 0.1$	8	5
4	5	1	2	3	$\frac{1}{10} = 0.1$	9	6

We can then use this to make the distribution of *T* under the assumption that the null hypothesis is true.

T	3	4	5	6	7	8	9
P(*T*)	0.1	0.1	0.2	0.2	0.2	0.1	0.1

The p-value is P(observing $T_1 \leq 4 \mid H_0$ is true) = 0.1+0.1=0.2 from the above table. This agrees with the output from R:

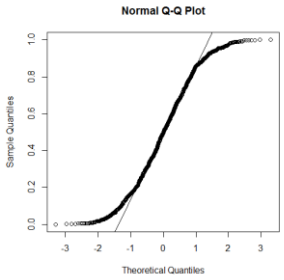
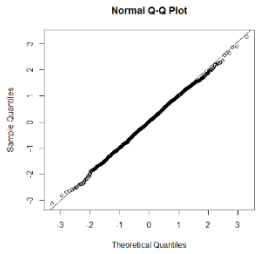
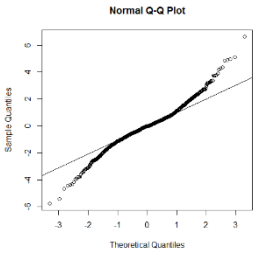
```
> x<-c(0,2)
> y<-c(1,4,7)
> wilcox.test(x,y,alternative="less")

        Wilcoxon rank sum test

data:  x and y
W = 1, p-value = 0.2
alternative hypothesis: true location shift is less than 0
```

For large samples, or if there are ties, these calculations can be a lot more complicated, and various other methods are used to approximate the p-values.

S14.3.4 – If you have a choice, should you use the t-test or the MWW?

One question that comes up in nonparametric statistics is, how do the t-test and Wilcoxon test compare? In STAT 513 and 518 this is discussed using a technical concept called asymptotic relative efficiency. The basic idea is that the relative efficiency of test 1 to test 2 is how many times bigger does the sample for test 2 need to be to have the same power as test 1. A relative efficiency greater than one means test one is better, a relative efficiency less than one means that test two is better. (The asymptotic means that that the values will only be rough guidelines in real situations).

Hodges and Lehmann (1956) and Lehmann (1998) report the asymptotic relative efficiency of the two sample t-test to the MWW test for some symmetric distributions are given below:

Distributions	Q-Q plot	Asymptotic relative efficiency of MWW to *t*
Uniform (a light tailed distribution)		1
Normal		$3/\pi \approx 0.955$
Double Exponential (a heavy tailed distribution)		1.5

This means for somewhat light tailed distributions you would want about the same sample sizes for the MWW and *t* to have the same power. For the normal distribution you can get away with about 5% fewer observations for the *t*-test. But for a heavy tailed distribution like the double exponential, you would need about 50% more observtions if you were using the t-distribution. Hodges and Lehmann show that the worst the MWW performs for any symmetric distributions with a variance is to have an asymptotic relative efficiency of 0.864, and the best it can perform is ∞!

Thus, if your data is symmetric – the case where the mean makes the most sense as a measure of center – then the MWW seems like the safe choice. At worst for the MWW, the *t*-test could get away with a sample size around 15% smaller, but at best for it the *t*-test would need infinitely more. In the case of non-symmetric distributions, the relative efficiency argument isn't clear cut, but the MWW still has the benefit of having a null and alternate hypothesis that make sense, even if the two populations have the same distribution.

One of the great mysteries of statistics is why the Mann-Whitney-Wilcoxon rank sum test isn't a lot more popular. One reason for this could be that it is harder to come up with the meaure of the difference between the distributions (what do you use if you can't use μ_1-μ_2?) and that the confidence interval is a bit harder to explain. Those are two of the topics discussed in STAT 518. Another reason could be that the t-test naturally grows into the methods discussed in STAT 516, and it can be a lot harder to do similar things using ranks like the MWW does.

S14.3.5 – Resampling Tests

The Mann-Whitney-Wilcoxon rank sum test can be thought of as an example of a class of methods sometimes called permutation tests, randomization tests, or resampling tests. These methods assume the groups each observation belongs to are randomly assigned if H_0 is true (keeping the sample sizes the same), and then asks what the chances are that the test statistic you calculated for your data would be that extreme if H_0 was true. The test statistic used by the MWW test is the sum of the ranks, and this choice allows for some easy calculations for large sample sizes (as on page 14-14 of the text).

There are other possibilities besides the sum of ranks that could be used. For example we might want to use the difference of the means (say $\bar{x} - \bar{y}$) like in the two sample *t*-test instead. This gives us the advantage of having an easily interpretable test statistic – it is simply the difference of the sample means!

Consider the small example in S14.4.3 where we want to test that X tends to be smaller than Y. Our data was two independent simple random samples:

X	0	2	
Y	1	4	7

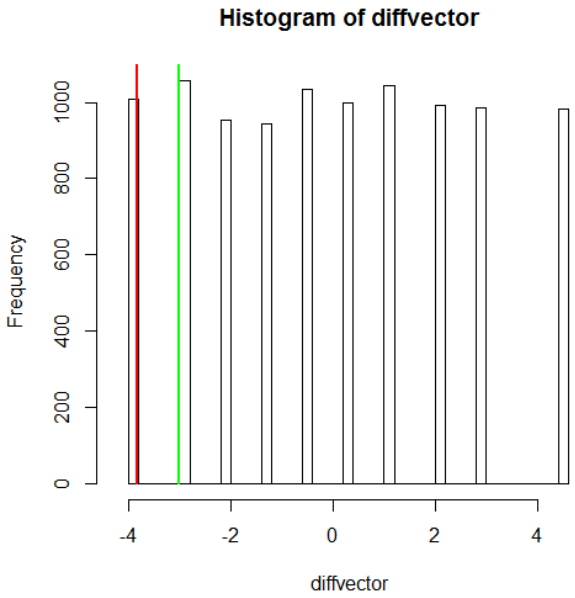
In this case, our test statistic $\bar{x} - \bar{y}$ is equal to -3 for this data set. The p-value would be the probability of observing a test statistic this small or smaller if H_0 were true. If the null hypothesis is true, then the two observations in X were randomly chosen from the five total observations. We could use this idea to make a table of all of the possible ways the observations could be allocated to the two groups.

X observations		Y observations			P(allocation)	$\bar{x} - \bar{y}$
0	1	2	4	7	$\frac{1}{10} = 0.1$	1/2-13/3 ≈ - 3.83
0	2	1	4	7	$\frac{1}{10} = 0.1$	1-4 = - 3
1	2	0	4	7	$\frac{1}{10} = 0.1$	3/2 – 11/3 ≈ - 2.17
0	4	1	2	7	$\frac{1}{10} = 0.1$	2 – 10/3 ≈ - 1.33
1	4	0	2	7	$\frac{1}{10} = 0.1$	5/2 – 3 = -0.5
2	4	0	1	7	$\frac{1}{10} = 0.1$	3 – 8/3 ≈ 0.33
0	7	1	2	4	$\frac{1}{10} = 0.1$	7/2 – 7/3 ≈ - 1.17
1	7	0	2	4	$\frac{1}{10} = 0.1$	4 – 2= 2
2	7	0	1	4	$\frac{1}{10} = 0.1$	9/2 – 5/3 ≈ 2.83
4	7	0	1	2	$\frac{1}{10} = 0.1$	11/2 – 1 = 4.5

The p-value is the P(observing $\bar{x} - \bar{y} \leq$ our test statistic value | H_0 is true). In this case that is P($\bar{x} - \bar{y} \leq$ -3 | H_0 is true) = 0.2 from the above table.

If the sample sizes are much larger, it gets very time consuming to construct tables like the above, and there is often no short-cut formula like there is for the MWW test. As a substitute, a large number of reallocations are made randomly using the computers random number generation. If, say, 10,000 random divisions are made of assigning two observations to X and three to Y, then the $\bar{x} - \bar{y}$ from those should be pretty close to the values from the table above. The reallocate function in the course's R templates page uses this method and provides the p-value and a graph showing distribution it generated, the test statistic (as a green line), and the rejection region cut-off for you're α-level (as a red line).

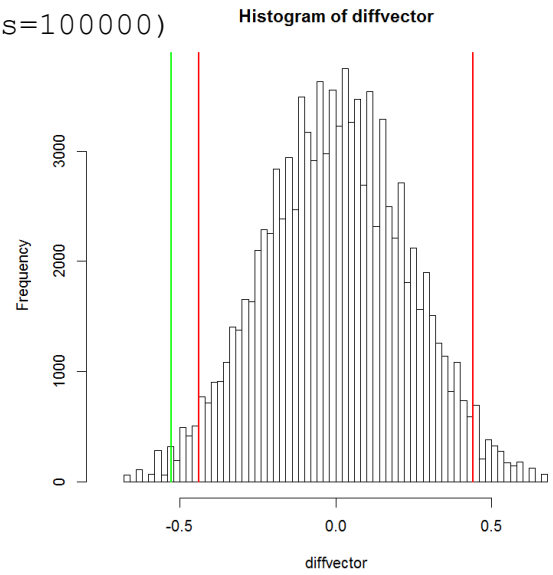
```
> x<-c(0,2)
> y<-c(1,4,7)
> reallocate(x,y,nreallocations=10000,alpha=0.05,alternative="less")
$test.statistic
[1] -3

$critical.value.low
       5%
-3.833333

$critical.value.high
NULL

$pvalue
[1] 0.2062794
```



With only 10,000 random reallocations, the ten bars in the histogram don't contain exactly 10% of the values, but they're close, and so the p-value is approximately equal to the exact 0.20. Given how fast modern computers are, there is no reason not to put a very large number in for the number reallocations to use.

Notice that because the distribution is discrete, it is tricky to decide if you should reject or not using a rejection region. For example, the smallest p-value you can get for this problem is around 0.10 – so if you rejected when the test statistic was equal to -3.833 you would actually have an α-level of 0.10 and not 0.05. If you were using α=0.10, then whether you decided to reject or not could depend on if there was randomly a bit more or a bit less than ten percent of the observations in the first bar. This issue tends to become less important as the two sample sizes increase, because the odds of falling exactly on the border diminish greatly. The output for data in section 14.2 of the text is given below to illustrate this:

```
> reallocate(DrugA,DrugB,nreallocations=100000)
$test.statistic
[1] -0.527381

$critical.value.low
     2.5%
-0.437619

$critical.value.high
    97.5%
0.4414286

$pvalue
[1] 0.01553984
```



Randomization tests can be traced at least as far back as Fisher (1935). The increases in computer access and power in the past decades has led to their becoming very popular in recent years, and they and their related methods are now even found in a few introductory high school and college text books. Some of the reasons they aren't more widely used are similar to the reasons the MWW isn't, in terms of generalizing the methods to the topics of STAT 516 and that a different method (called bootstrapping) has to be used to make confidence intervals. Sheskin (2011, page 550) provides several references about randomization tests, and notes that the ones using a statistic like the difference in means like in the example above, might still have some difficulty in the cases of skewed distributions and outliers.

References:

Fisher, R.A. (1935). *The design of experiments*. Edinburgh: Oliver and Boyd.

Hodges, J.L., and Lehmann (1956). The efficiency of some nonparametric competitors of the t-test. *Annals of Mathematical Statistics, 27* (2), 324-335.

Kruskal, W.H. (1957). Historical notes on the Wilcoxon unpaired two-sample test. *Journal of the American Statistical Association, 52* (279), 356-360.

Lehmann, E.L. (1998). *Nonparametrics: Statistical Methods Based on Ranks*. New Jersey: Prentice Hall.

Mann, H.B., and Whitney, D.R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics, 18* (1), 50-60.

Sheskin, D.J. (2011). *Handbook of parametric and nonparametric statistical procedures*. Boca Raton: CRC Press.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin, 1* (6), 80-83.