Wilcoxon Test Theory Notes

Charles J. Geyer

September 19, 2007

1 Introduction

These are class notes for Stat 5601 (nonparametrics) taught at the University of Minnesota, Spring 2006. This not a theory course, so the bit of theory we do here is very simple, but very important, since without it we cannot understand the duality of tests and confidence intervals for the two Wilcoxon tests (signed rank and rank sum).

2 Wilcoxon Signed Rank Test

Why are the two alternative forms of the test statistic for the Wilcoxon signed rank test

(a) the sum of the positive ranks

(b) the number of Walsh averages greater than the hypothesized μ

equal?

First we need to say that they aren't necessarily when there are ties, hence thoughout these notes (for both Wilcoxon tests) we assume

- the distribution of the data is continuous

This implies, in particular, for the signed rank test

- no ties occur among the Walsh averages (with probability one)

This fact follows directly from the sum of continuous random variables being continuous, which we will not try to prove (it follows directly from the convolution formula for the density of a sum of random variables, which should be taught in any theory of statistics course).

So what this section will prove is that, assuming no ties among the Walsh averages, the two forms (a) and (b) are equal.

The way proofs like this go is by mathematical induction. First we show they are equal for some simple case. Then we show that any change to the data keeps them equal. Hence we conclude they are equal no matter what the data are (as long as there are no ties).

2.1 Base of the Induction

The simple case we start with is when the hypothesized value of the parameter of interest (population center of symmetry μ) is greater than all data points, hence greater than all Walsh averages.

Then if Z_i are the data, all of the $Z_i - \mu$ are negative hence all of the ranks are negative and the sum of the positive ranks is zero.

Similarly all of the Walsh averages

$$\frac{Z_i + Z_j}{2}, \qquad i \le j \tag{1}$$

are less than μ, hence (b) is zero too.

That finishes the start of the induction. We know (a) and (b) are equal for this case (all of the data points below the hypothesized μ).

2.2 Induction Steps

Now we consider moving μ keeping the data fixed from past the upper end of the data (where we now know the values of the alternative forms of the test statistic) to past the lower end of the data, where symmetry considerations tell us both forms take on their maximal value $n(n + 1)/2$.

Clearly, (b) changes value when and only when μ moves past a Walsh average.

Changes of value of (a) are more complicated. It changes value when

- the ranks of some of the $|Z_i - \mu|$ change, or
- the signs attached to some of the ranks change.

Latter first. Clearly the sign of $Z_i - \mu$ changes when μ moves past Z_i, which is a Walsh average, the case $i = j$ in (1).

The rank of $|Z_i - \mu|$ cannot change by itself. Some other rank, say that of $|Z_j - \mu|$ must change too. If we are looking at a very small change in μ causing the change, as these move past each other, they must be equal.

Furthermore they cannot have the same sign, because decreasing μ cannot change the order of $Z_i - \mu$ and $Z_j - \mu$. Thus when they are equal we must have

$$(Z_i - \mu) = -(Z_j - \mu) \tag{2}$$

which implies $\mu = (Z_i + Z_j)/2$. Hence in this case too, (a) changes when μ moves past a Walsh average.

To summarize this section, we have proved: (a) and (b) change only when μ moves past a Walsh average (of one kind or the other).

2.2.1 Induction Step at a Data Point

When $i = j$ in (1), then the Walsh average is just Z_i. In this section we consider what happens when μ moves past Z_i (going from right to left on the number line)

Clearly, (b) is increased by one. There is exactly one Walsh average equal to Z_i and when μ moves past that, there is one more Walsh average greater than μ.

Clearly, (a) is also increased by one. Exactly one signed rank changes, that associated with Z_i, when μ moves past Z_i. When μ is very near Z_i, in which case $|Z_i - \mu|$ is the smallest of all the $|Z_j - \mu|$, hence the i-th rank is one, and the sign is negative when $\mu > Z_i$ and positive when $\mu < Z_i$.

2.2.2 Induction Step at other Walsh Averages

The word “other” in the heading means those not yet considered. Suppose $i \neq j$ and consider what happens when μ moves past the Walsh average $w = (Z_i + Z_j)/2$ (going from right to left on the number line). As μ moves past w, at some point it is equal to w, and $\mu = w$ implies (2).

Since Z_i and Z_j are not tied, we must have one greater and one less than w, say (without loss of generality) $Z_i < w < Z_j$.

For μ just a little above w we have the i-th rank greater than the j-th and for for μ just a little below w we have the i-th rank less than the j-th. Moreover the i-th is a negative rank (because $Z_i - \mu$ is negative) and the j-th is a positive rank (because $Z_j - \mu$ is positive). Finally the i-th and j-th absolute ranks are consecutive integers because as $|Z_i - \mu|$ and $|Z_j - \mu|$ change places in the rank order there is no room between them for other $Z_m - \mu$. Thus as μ passes w going from right to left the j-th rank (which is positive) increases by one, the i-th rank decreases by one (but is negative and does not count), and none of the other ranks or signs change. Hence the form (a) increases by one.

And that finishes the proof. In all cases, no matter where μ is, (a) and (b) change in sync and hence are equal no matter where μ is.

3 Wilcoxon Rank Sum Test

The situation for the rank sum test is similar. There are two alternative test statistics, which, although not identical, differ only by a constant. If the data are X_1, …, X_m and Y_1, …, Y_n and the hypothesized value of the shift is μ, meaning that under the null hypothesis we assume that the X_i and the $Y_j - \mu$ have the same distribution, then the test statistics are

(a) the sum of the y ranks

(b) the number of $Y_j - X_i$ differences greater than the hypothesized μ.

The former (a) we call the Wilcoxon statistic W. The latter (b) we call the Mann-Whitney statistic U. To be more precise, to calculate W we assign ranks to the $m + n$ numbers of the form X_i and $Y_j - \mu$ and W is the sum of the ranks of the $Y_j - \mu$.

In this section we prove two things.

- W and U differ by a constant, which we identify.
- The distributions W and U are symmetric about the midpoints of their ranges, which we also identify.

As with the theory for the signed rank test, the proof of the first proceeds by induction. Throughout we assume no ties in the data.

3.1 Base of the Induction

The proof by induction moves the hypothesized value μ from right to left along the number line (this is just like the proof for the signed rank test). As μ decreases, all of the $Y_j - \mu$ values increase in synchrony, moving from below all the X_i values (when μ is very large and positive) to above all the X_i values (when μ is very large and negative).

When μ is very large and positive, so all of the $Y_j - \mu$ are to the left of all the X_i, U is zero (all of the $Y_j - X_i$ are less than μ) and the Y ranks are the numbers from 1 to n, so $W = n(n+1)/2$ and

$$W = U + \frac{n(n+1)}{2}. \tag{3}$$

3.1.1 Induction Step

We now need to show that as μ moves from left to right along the number line U and W change in synchrony. Clearly W changes whenever some $Y_j - \mu$ is equal to some X_j. Clearly U changes whenever some $Y_j - X_i$ is equal to μ. Clearly both of these are the same event. Write

$$Z_{ij} = Y_j - X_i.$$

As μ moves from just above to just below Z_{ij} the ranks assigned to X_i and $Y_j - \mu$ swap. The one for X_j decreasing by one and the one for $Y_j - \mu$ increasing by one, so W increases by one. As μ moves from just above to just below Z_{ij} the number of Z_{ij} greater than μ increases by one, so U increases by one.

And that concludes the proof; (3) always holds.

3.2 Symmetry

Symmetry is easier to see for the Mann-Whitney test statistic. The way to make it easy is to envisage the X's to be sampled before the Y's and to envisage the Y's arriving one at a time. So consider m X's and one Y (take the null hypothesis to be zero, so $Y_j - \mu$ is just Y_j).

Under the null hypothesis, these $m + 1$ variables are IID, hence the Y is equally probable to be any place in the sorted order; from 1 to $m + 1$. Thus the number of $Y_j - X_i$ pairs that exceed $\mu = 0$ ranges from 0 to m and each of these $m + 1$ numbers is equally probable. In short the contribution to U from all the X's and one Y has the discrete uniform distribution on the set $\{0, \ldots, m\}$. Since the contributions of the Y's are IID (the Y's being IID and independent of the X's), we see that U is the sum of n IID discrete uniform $\{0, \ldots, m\}$ random variables.

We take as a fact from theory (not difficult to prove) that the IID sum of symmetric random variables is symmetric. The discrete uniform distribution is symmetric, thus the distribution of U is symmetric. This discrete uniform distribution has center of symmetry $m/2$, which is also the mean. The mean of the sum of n IID random variable is n times the mean of one of them, thus

$$E(U) = \frac{mn}{2}$$

and this is also the center of symmetry and the median of the distribution of U. The range of U is from 0 to mn, which also makes it obvious the the center of symmetry, which must be the midpoint of the range, is $mn/2$.

From (3) we know that

$$E(W) = E(U) + \frac{n(n+1)}{2} = \frac{n(m+n+1)}{2}$$

and this is also the center of symmetry and the median of the distribution of W. The range of W is found by adding $n(n+1)/2$ to the range of U, giving lower end

$$\frac{n(n+1)}{2}$$

and upper end

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