Objectives

This time, we'll discuss some nonparametric difference tests. If you recall, nonparametric tests generally have lower power than parametric tests, but make fewer assumptions about the distribution of the data, so they may be valid when parametric tests are not. These are the rough equivalents of the parametric and nonparametric tests we've looked at:

Parametric test	Equivalent nonparametric test
Two-sample unpaired t test	Mann–Whitney *U* test (≡ Wilcoxon rank-sum test)
Two-sample paired t test	Wilcoxon signed-rank test with matched pairs
One-sample t test	Wilcoxon signed-rank test, pairing data with a fixed value

They assume that the variable is measured on at least an ordinal scale. (That's it.)

Stuff with a solid edge, like this, is important.

But remember — you can totally ignore stuff with single / double wavy borders.

4.1 Background

Nonparametric tests often operate on the **rank** order of a set of numbers, rather than on the numbers themselves. This also means that nonparametric tests are less affected by **outliers** (a few extreme scores) than parametric tests. Outliers may make parametric tests *less* powerful (they increase the variance as well as distorting the mean), sometimes less powerful than the nonparametric equivalent.

It should be obvious how ranking 'removes' information about the distribution. The scores {2,8,10,12,14,24} might have come from a normal distribution and the scores {1,2,3,100,101,102} might have come from a bimodal distribution, but both reduce to the ranks {1,2,3,4,5,6}.

How to rank data (repeated from Handout 2)

Suppose we have ten measurements (e.g. test scores) and want to rank them. First, place them in ascending numerical order:

5	8	9	12	12	15	16	16	16	17

Then start assigning them ranks. When you come to a tie, give each value the mean of the ranks they're tied for — for example, the 12s are tied for ranks 4 and 5, so they get the rank 4.5; the 16s are tied for ranks 7, 8, and 9, so they get the rank 8:

X:	5	8	9	12	12	15	16	16	16	17
rank:	1	2	3	4.5	4.5	6	8	8	8	10

4.2 The Mann–Whitney *U* test (for two independent samples)

This is a nonparametric analogue of a two-sample unpaired *t* test. Its null hypothesis is that the two samples were drawn from identical populations (rather than the *t* test's null hypothesis that the two samples were drawn from populations with the same means). So a 'significant' Mann–Whitney result might be due to a difference between the *central tendency* of the two populations (like a 'significant' *t* test) but it might also have been due to some other difference, such as a difference in the *distributions* of the populations. If we assume the distributions are similar, a significant Mann–Whitney test suggests that the **medians** of the two populations are different.

Basic logic of the test

Let's suppose we have two samples with n_1 and n_2 observations in each ($n_1 + n_2 = N$ observations in total). We can rank them, lowest to highest, from 1 to N. If the two samples come from identical populations, the sum of the ranks of 'sample 1' scores is likely to be about the same as the sum of the ranks of 'sample 2' scores. If, on the other hand, sample 1 comes from a population with generally much lower values then sample 2, then the sum of the ranks of 'sample 1' scores will be lower than the sum of the ranks of 'sample 2' scores.

Calculating the Mann–Whitney U statistic

1. Call the smaller group 'group 1', and the larger group 'group 2', so $n_1 < n_2$. (If $n_1 = n_2$, ignore this step.)
2. Calculate the sum of the ranks of group 1 (= R_1) and group 2 (= R_2).
3. $U_1 = R_1 - \frac{n_1(n_1+1)}{2}$
4. $U_2 = R_2 - \frac{n_2(n_2+1)}{2}$
5. The Mann–Whitney statistic U is the smaller of U_1 and U_2.

Check your sums: verify that $U_1 + U_2 = n_1 n_2$ and $R_1 + R_2 = \frac{(n_1+n_2)(n_1+n_2+1)}{2}$.

It doesn't matter which numbers you call U_1 and U_2, since all you do is take the smaller. Incidentally, why this formula for $R_1 + R_2$? Because if you add consecutive numbers from 1 to x, the total is $\sum_{i=1}^{x} i = \frac{x(x+1)}{2}$.

Determining a significance level from U

If n_2 is small, look up the critical value of U in tables — **values of U smaller than the critical value are significant. If $n_2 > 20$,** the U statistic is approximately normally distributed; mean $\mu = \frac{n_1 n_2}{2}$, variance $\sigma^2 = \frac{n_1 n_2 (n_1+n_2+1)}{12}$. So we can calculate a ***Z* score:**

$$z = \frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1+n_2+1)}{12}}}$$

and test that in the usual way (see Handout 1).

Example

Borrowing an example from Howell (1997, p. 651), suppose we imagine that we collect information on birth weights of babies whose mothers received prenatal care either from the first trimester onwards or from the third trimester onwards. Suppose these birthweights, in kg, were {1.68, 3.83, 3.11, 2.76, 1.70, 2.79, 3.05, 2.66, 1.40, 2.775} for the first trimester group and {2.94, 3.38, 4.90, 2.81, 2.80, 3.21, 3.08, 2.95} for the third trimester group. If we chose to calculate a Mann–Whitney test on these data, we would calculate the ranks as {2, 17, 14, 5, 3, 7, 12, 4, 1, 6} for the first trimester group (n = 10, rank sum = 2 + 17 + 14 + … = 71) and {10, 16, 18, 9, 8, 15, 13, 11} for the third trimester group (n = 8, rank sum = 100). We'd therefore call the third trimester group 'group 1', because it's the smaller, and the first trimester group 'group 2'. So we have $n_1 = 8$, $n_2 = 10$, $R_1 = 100$, $R_2 = 71$. From this we can calculate $U_2 = 64$, $U_1 = 16$. The Mann–Whitney U is the smaller of these, i.e. 16.

From tables we can find that the critical value of U for these values of n and a two-tailed test at $\alpha = 0.05$ is 17. Our U is less than this, so it's significant; we reject the null hypothesis, and say that there's a difference between the birthweights of our two sets of babies, $p < 0.05$.

If our ns had been larger, we could have calculated a Z score. Pretending for a mo-

ment that our *n*s were larger, for these data $z = -2.13$, corresponding to $p = 0.033$.

For the same data, a two-sample unequal-variance *t* test would have given $p = 0.063$, and a two-sample equal-variance *t* test would have given $p = 0.066$. This is an example when a nonparametric test has more power because the assumptions of the parametric test — in this case normality of the underlying distribution — were not met.

4.3 The Wilcoxon matched-pairs signed-rank test (for two related samples)

This is a nonparametric test for **paired** scores. It's the nonparametric analogue of the *t* test for related samples (the paired *t* test). The null hypothesis is that the distribution of differences between the pairs of scores is symmetric about zero. (Since the median and the mean of a symmetric population are the same, the null hypothesis can be restated either as 'the differences between the pairs of scores are symmetric with a mean and a median of zero'.)

Let's do this as a worked example (borrowed from Howell, 1997, p. 653). Suppose 10 subjects have their systolic blood pressure measured (BP_1), engage in a running program for 6 months, and then have their systolic blod pressure measured again (BP_2). We can calculate the difference for each subject as $BP_2 - BP_1$. If there's no difference between the 'before' and 'after' scores, there should be about as many differences that are positive as there are differences that are negative…

Calculating the Wilcoxon matched-pairs signed-rank statistic, T

The procedure is:

1. Calculate the difference scores.
2. Ignore any differences that are zero.
3. Rank the difference scores, *ignoring their sign* (+ or –).
4. Add up all the ranks for difference scores that were positive; call this T^+.
5. Add up all the ranks for difference scores that were negative; call this T^-.
6. The Wilcoxon matched-pairs statistic *T* is the smaller of T^+ and T^-.

Check your sums: verify that $T^+ + T^- = \frac{n(n+1)}{2}$.

Here's a worked example:

Before (BP_1):	130	148	170	125	170	130	130	145	119	160
After (BP_2):	120	148	163	120	135	143	136	144	119	120
Difference ($BP_1 - BP_2$):	10	0	7	5	35	–13	–6	1	0	40
Rank of difference (ignoring zero differences and sign):	5		4	2	7	6	3	1		8
'Signed rank'	*5*		*4*	*2*	*7*	*–6*	*–3*	*1*		*8*
Ranks of positive differences:	5		4	2	7			1		8
Ranks of negative differences:						6	3			

The 'signed rank' row is what gives the test its name; it's what you get when you put the signs (+ or –) from the difference scores back on the ranks you calculated by ignoring those signs. But you don't need to do this to calculate *T*.

The difference scores don't appear to be anything like normally distributed, so we want to use a distribution-free (nonparametric) test. We can calculate $n = 8$ (since we ignore zero differences), $T^+ = 5 + 4 + 2 + 7 + 1 + 8 = 27$, and $T^- = 6 + 3 = 9$. Therefore the Wilcoxon statistic $T = 9$.

Determining a significance level from T

For small *n*, look up the critical value of *T* in tables — **values of *T* smaller than the critical value are significant. If *n* > 20,** the *T* statistic is approximately normally distributed; mean $\mu = \frac{n(n+1)}{4}$, variance $\sigma^2 = \frac{n(n+1)(2n+1)}{24}$. So we can calculate a ***Z* score:**

$$z = \frac{T - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}$$

and test that in the usual way (see Handout 1).

4.4 Using the Wilcoxon signed-rank test as a one-sample test

The Wilcoxon signed-rank test may also be used to test whether the median of one group of scores is significantly difference from some expected value *M*. In this case, the null hypothesis is that the median is equal to *M*. Calculate a difference score **(*x* – *M*)** for each score *x*, and proceed as above.

4.5 Supplementary and / or advanced material

The Wilcoxon rank-sum test (not the same as the Wilcoxon signed-rank test!)

There are actually *two* tests based on the logic used for the Mann–Whitney *U* test: they are the Mann–Whitney *U* test itself and the *Wilcoxon rank-sum test*. They're directly equivalent: both will give the same *p* value. (Some people even mix the names up, calling *U* a Wilcoxon rank-sum statistic, which confuses everybody.) The Mann–Whitney *U* test is more popular and has a name that's not so easily confused with the Wilcoxon signed-rank test. However, the Wilcoxon rank-sum calculations make it a bit clearer how we get a statistic out of the sums of a set of ranks, so I've included it here **only** in case you want to understand how the two tests work.

1. Call the smaller group 'group 1', and the larger group 'group 2', so $n_1 < n_2$. (If $n_1 = n_2$, ignore this step.)
2. Calculate the sum of the ranks of group 1 (= R_1) and group 2 (= R_2).
3. If $n_1 < n_2$, then $W_S = R_1$. If $n_1 = n_2$, then W_S = whichever of R_1 and R_2 is smaller.
4. Calculate $W_S' = n_1(n_1 + n_2 + 1) - W_S$.

Now we evaluate W_S and W_S' using tables. The smaller W_S is, the more likely it is to be significant. W_S will be significant (small) if the smaller group (group 1) contains significantly smaller-than-average ranks, or if the larger group (group 2) contains significantly larger-than-average ranks, i.e. if group 1 < group 2. W_S' is the sum of the ranks we would have found if we reversed our ranking and ranked from high to low; it will be significant (small) if group 2 > group 1. Normally we want to test for a two-tailed difference between groups; we'd then pick whichever of W_S and W_S' is the smaller and look up the critical values in tables (doubling α if the table gives one-tailed values).

Two other ways of calculating the Mann–Whitney U statistic

This shows the equivalence of the Wilcoxon rank-sum and Mann–Whitney tests:

1. Compute W_S and W_S' as above. Let W_S'' be whichever of the two is larger.
2. The Mann–Whitney statistic $U = \frac{n_1(n_1 + 2n_2 + 1)}{2} - W_S''$

A third method is this:

1. For each observation in group 1, count the number of observations in group 2 that exceed it (score 0.5 for equality). Sum these values to obtain U_1.
2. Do the same for group 2 to obtain U_2.
3. The Mann–Whitney statistic *U* is the smaller of U_1 and U_2.

Bibliography

Howell, D. C. (1997). *Statistical Methods for Psychology*. Fourth edition, Wadsworth, Belmont, California.