STAT 621 Lecture Notes
The Wilcoxon / Mann-Whitney Test

We've seen that the both the Sign Test and the Wilcoxon Rank Sum test may be used to make an inference about a single population median, θ. This is often used in the context where pairs of dependent observations are measured, and interest is in determining whether the population distributions are located in the same place.

Next we will look at a procedure that also compares the location of two population distributions. This is a version of the Rank Sum test for independent samples and is credited to both Wilcoxon and Mann-Whitney. This time the two samples are independent, for example, measurements taken on one group of subjects given a treatment, and a separate unrelated group given a placebo. As before, the objective is to draw inferences about the presence or size of a treatment effect through a location parameter (i.e. the median).

Data: Observations $X_1, \ldots, X_m$ from Population 1 and $Y_1, \ldots, Y_n$ from Population 2, for a total of $N = m + n$ observations.

Assumptions:

1.

2.

3.

Hypothesis Test: Suppose $X_i \sim F$ and $Y_j \sim G$ for $i = 1, \ldots n$ and $j = 1, \ldots m$. The test of interest is that for all t

$$H_0 : G(t) = F(t) \qquad \text{vs.} \qquad H_1 : G(t) = F(t - \Delta).$$

What do these hypotheses imply? How do they compare to the one-sample problem?

We can rewrite the above hypotheses in a more familiar form. Suppose that $X_i \sim F(t)$ and $Y_j \sim G(t) = F(t - \Delta)$. Then we test, for example,

$$H_0 : \Delta = 0 \qquad \text{vs.} \qquad H_A : \Delta > 0.$$

Test Statistic: First combine the data from the two samples and order from smallest to largest. Assign ranks $R_1, \ldots, R_n$ to the Y observations in the joint ordering. The test statistic is the sum of these ranks,

$$W = \sum_{j=1}^{n} R_j.$$

Rejection Rules:

What value of W would we expect under the null? What would we see instead if $\Delta < 0$? $\Delta > 0$? How would we find critical values or p-values?

Note: Critical values can be found with the R function `qwilcox` and p-values for the test can be found with the `pwilcox` function. These function report properties of W only indirectly, as is illustrated in an example later on.

Example: The following data come from a study of stove insulation. Ovens having two types of insulation were tested. Insulation A is the standard type and Insulation B is a newer and cheaper type. Ovens were heated to 400 degrees C and the time it took to cool to 350 degrees C was measured.

Y=Insulation A	15.7	14.8	14.2	16.1	15.3	13.9	17.2	14.9	
X=Insulation B	13.7	14.1	14.7	15.4	15.6	14.4	12.9	15.1	14.0

Does Insulation B hold heat as well as Insulation A, or does it loose heat more rapidly? Write some hypotheses. Compute the test statistic.

Sampling Distribution of W

First the basic idea: what can we say about the likely values of W if H_0 is true and the X and Y samples come from the same distribution?

Let's consider a simple example. Suppose that m=4 and n=3. Then we need to assign ranks to 3 of the $N = 7$ observations. Think of this as choosing 3 places to put the Y's in a sequence. For example, one outcome is $YYYXXXX$ resulting in ranks 1, 2 and 3 so that W=6. Under H_0 each sequence is equally likely. What is the probability $W = 6$ if H_0 is true? What is the probability the $W = 9$?

By considering all possible sequences in this way, the probability of each possible value of W can be computed. This is the sampling distribution of W under the null hypothesis. Values of W that are improbable under the null but support the alternative lead to rejection. Clearly for large sample sizes, finding the exact distribution if W is computationally intensive. A large-sample version of the test follows from the asymptotic normality of W.

Large-Sample Approximation

For relatively large sample sizes, standardized statistic W^* follows an approximate $N(0,1)$ distribution under H_0,

$$W^* = \frac{W - E(W)}{V(W)}$$

It can be shown that under the null,

$$E(W) = \frac{n(N+1)}{2} \qquad \text{and} \qquad \frac{mn(N+1)}{12}$$

So, a large-sample test could be performed by.....

The Mann-Whitney Statistic

An equivalent procedure was described by Mann and Whitney around the same time. The Mann-Whitney test statistic for the hypothesis $H_0 : \Delta = 0$ is given by

$$U = \sum_{i=1}^{m} \sum_{j=1}^{n} \phi(X_i, Y_j)$$

where $\phi(X_i, Y_j)$=1 if $X_i < Y_j$ and 0 otherwise. What is U measuring? Describe what you would expect if H_0 were true? If each of the alternatives were true?

It can be shown that

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

Prove or explain why this is true.

Find U for the insulation data set.

Many software programs, including R, report only the Mann-Whitney version of this hypothesis test. The test statistic and p-values are computed with the `wilcox.test` function and the `paired=FALSE` (default) option. The function `qwilcox` reports critical values for the test, based on the probability distribution of U. These can be translated into critical values for W if desired (how would you do that?). Similarly p-values for U are found with `pwilcox`.

```
> qwilcox(.05,9,8,lower.tail=F)
[1] 53
> pwilcox(51,9,8,lower.tail=F)
[1] 0.06939531
```

Note: Be careful specifying alternatives for one-sided tests in the `wilcox.test` function. For example, the following command

```
wilcox.test(x, y, alternative = "greater", paired = FALSE)
```

specifies the alternative that the distribution of x is shifted to the right of the distribution of y.

Below are the results for the insulation data set using this function.

```
A=c(15.7, 14.8, 14.2, 16.1, 15.3, 13.9, 17.2, 14.9)
B=c(13.7, 14.1, 14.7, 15.4, 15.6, 14.4, 12.9, 15.1, 14.0)
wilcox.test(A,B,alternative="greater",paired=F)

        Wilcoxon rank sum test

data:  A and B W = 52, p-value = 0.0694 alternative hypothesis: true
location shift is greater than 0
```

Note that specifying the option `exact=FALSE` computes the large-sample approximate test. By default R uses a continuity correction for the p-value in the large-sample version of the test. This can be suppressed with the `correct=F` option.

Ties

Tied values in a data set change slightly the sampling distributions of both versions of our test statistic. Use the following recommendations in the presence of ties.

(1) As described for the one sample problem, replace the ranks of tied values with average ranks. Note that the resulting test will have *approximate* level α.

(2) Large-Sample: Tied values reduce the variance of W for calculating W^*, under H_0. The correct variance to use is

$$V(W) = \frac{mn(N+1)}{12} - \left\{ \frac{mn}{12N(N-1)} \sum_{j=1}^{g} (t_j - 1)t_j(t_j+1) \right\}$$

where g is the number of tied groups and t_j is the number of ranks in the jth tied group.

Example: The data below are points scored by the University of Iowa basketball team in games played in 1997. Is there a difference in the number of points scored at Home vs. Away games?

Home	72	78	76	82	75	66	80	84	81
Away	69	59	71	51	51	67	69	48	75

```
Home=c(72,78,76,82,75,66,80,84,81)
Away=c(69,59,71,51,51,67,69,48,75)


wilcox.test(Home,Away,alternative="two.sided",paired=F,exact=T)

   Wilcoxon rank sum test with continuity correction

   data:  Home and Away
   W = 74.5,   p-value = 0.003049
   alternative hypothesis: true location shift is not equal to 0

   Warning message:
   In wilcox.test.default(Home, Away, alternative = "two.sided", paired = F,  :
     cannot compute exact p-value with ties
```