# STAT 135 Lab 8
# Hypothesis Testing Review, Mann-Whitney Test by Normal Approximation, and Wilcoxon Signed Rank Test.

Rebecca Barter

March 30, 2015

# Mann-Whitney Test

# Mann-Whitney Test

Recall that the Mann-Whitney test is a test for the difference between **two independent populations**, and can be conducted as follows

1. Concatenate the $X_i$ and $Y_j$ into a single vector $Z$
2. Let $n_1$ be the sample size of the smaller sample
3. Compute $R =$ sum of the ranks of the smaller sample in $Z$
4. Compute $R' = n_1(m + n + 1) - R$
5. Compute $R^* = \min(R, R')$
6. Compare the value of $R'$ to critical values in a table: if the value is less than or equal to the tabulated value, reject the null that $F = G$

# Mann-Whitney test: the normal approximation

The **Mann-Whitney** test tests the null hypothesis $H_0 : F = G$. This approach is based on the fact that when $H_0$ is true, we should have

$$\pi = P(X < Y) = \frac{1}{2}$$

i.e. that it is equally likely that $X < Y$ and that $Y < X$.

# Mann-Whitney test: the normal approximation

Under $H_0 : F = G$, we have $\pi = P(X < Y) = \frac{1}{2}$.

Estimate the observed value of $\pi$ by

- ranking the combined observations, $X_1, \ldots, X_n, Y_1, \ldots, Y_m$, from smallest (rank $= 1$) to largest (rank $= n + m$)
- look at the number of times that we have elements of $X$ that are smaller than elements of $Y$.

Then our observed proportion is:

$$\hat{\pi} = \frac{\#\text{ times we see } X_i < Y_j \text{ over all } i, j}{nm}$$

so if we observe a value of $\hat{\pi}$ that is significantly different from $\frac{1}{2}$, then this provides evidence against $H_0$.

# Mann-Whitney test: the normal approximation

Suppose, for example, that we had

$$X = (6.2, 3.7, 4.7, 1.3)$$

$$Y = (1.2, 0.8, 1.4, 2.5, 1.1)$$

Then our ranks are

$$rank(X) = (9, 7, 8, 4)$$

$$rank(Y) = (3, 1, 5, 6, 2)$$

We only see elements of $X$ being less than elements of $Y$ twice:

$$X_4 = 1.3 < Y_3 = 1.4$$

$$X_4 = 1.3 < Y_4 = 2.5$$

so $\hat{\pi} = \frac{2}{nm} = \frac{2}{4 \times 5} = 0.1$, which is much less than the expected 0.5!

# Mann-Whitney test: the normal approximation

It turns out that

$$\boxed{\hat{\pi} = \frac{1}{mn}\left(R' - \frac{m(m+1)}{2}\right)}$$

where $R'$ is the sum of the ranks of the $Y_i$'s, $n$ is the sample size for the $X_i$'s and $m$ is the sample size for the $Y_i$'s.

# Mann-Whitney test: the normal approximation

Thus, for our example, where

$$X = (6.2, 3.7, 4.7, 1.3) \qquad rank(X) = (9, 7, 8, 4)$$

$$Y = (1.2, 0.8, 1.4, 2.5, 1.1) \qquad rank(Y) = (3, 1, 5, 6, 2)$$

we can verify the given formula:

$$\begin{aligned}
\hat{\pi} &= \frac{1}{mn}\left(R' - \frac{m(m+1)}{2}\right) \\
&= \frac{1}{4 \times 5}\left((3 + 1 + 5 + 6 + 2) - \frac{5 \times 6}{2}\right) \\
&= 0.1
\end{aligned}$$

which is the same as we got before!

# Mann-Whitney test: the normal approximation

Thus the Mann-Whitney test can be constructed as follows:

Define our test statistic to be

$$\boxed{U_Y = R' - \frac{m(m+1)}{2}}$$

and recall that $U_Y = nm\hat{\pi}$, where under $H_0$, we have $\hat{\pi} = 0.5$.

- To calculate an exact $p$-value for the test, we would need to know the distribution of $U_Y$ under $H_0$.
- Unfortunately, we don't know the exact distribution, but we can use the normal approximation to compute an approximate $p$-value.

# Mann-Whitney test: the normal approximation

Under $H_0 : F = G$, we have that

$$E_{H_0}(U_Y) = \frac{nm}{2}$$

$$Var_{H_0}(U_Y) = \frac{nm(n+m+1)}{12}$$

and it is known that, asymptotically, under $H_0$, the standardized statistic tends to a normal distribution:

$$\frac{U_Y - E_{H_0}(U_Y)}{\sqrt{Var_{H_0}(U_Y)}} \sim N(0,1)$$

# Exercise

# Exercise: Mann-Whitney

Researchers have asked several smokers how many cigarettes they had smoked in the previous day. The data is as follows

| Women | Men |
|---|---|
| 4 | 2 |
| 7 | 3 |
| 20 | 5 |
| 21 | 6 |
| | 8 |
| | 16 |

Why might we prefer to use a non-parametric (Mann-Whitney) test rather than a $t$-test? Is there enough evidence to conclude that there is a difference between the number of cigarettes smoked per day between the sexes?

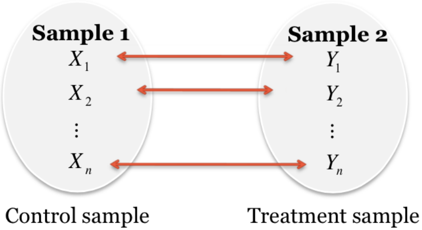
# Hypothesis testing for comparing <u>paired samples</u>

# Comparing samples from paired populations

So far we have focused on testing hypotheses for

- a parameter from a single population ($H_0 : \mu = 2$)
- comparing a parameter from two different, but arbitrary and independent, populations ($H_0 : \mu_1 = \mu_2$)

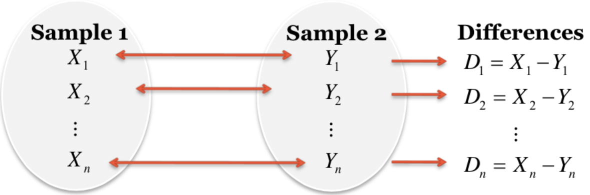
We now want to compare a parameter from *paired* populations.




e.g. the $(X_i, Y_i)$'s might be independent pairs of twins.

# Comparing samples from paired populations

We can treat paired tests as a one-sample problem:




where if we assume that $D_i = X_i - Y_i$ has true mean $\mu_D$, we are testing

$$H_0 : \mu_D = 0$$

versus $\quad H_1 : \mu_D > 0 \quad , \quad H_1 : \mu_D < 0 \quad , \quad H_1 : \mu_D \neq 0$

by observing whether $\overline{D}$ is far from 0

Can you see why is this different to observing $\overline{X} - \overline{Y}$?

# Hypothesis testing for comparing paired samples

The paired Z-test

# Comparing samples from paired populations

If we **know the true variance** $\sigma_D^2$, then we can use a $Z$-test:

$$H_0 : \mu_D = 0 \quad H_1 : \mu_D < 0$$

corresponds to $p$-value:

$$P_{H_0}(\overline{D} \leq \overline{d}) = P_{H_0}\left(\frac{\overline{D}}{\sigma_D/\sqrt{n}} \leq \frac{\overline{d}}{\sigma_D/\sqrt{n}}\right) \overset{CLT}{\approx} \Phi\left(\frac{\overline{d}}{\sigma_D/\sqrt{n}}\right)$$

# Comparing samples from paired populations

If we **know the true variance** $\sigma_D^2$, then we can use a $Z$-test:

$$H_0 : \mu_D = 0 \quad H_1 : \mu_D > 0$$

corresponds to $p$-value:

$$P_{H_0}(\overline{D} \geq \overline{d}) = P_{H_0}\left(\frac{\overline{D}}{\sigma_D/\sqrt{n}} \geq \frac{\overline{d}}{\sigma_D/\sqrt{n}}\right) \overset{CLT}{\approx} 1 - \Phi\left(\frac{\overline{d}}{\sigma_D/\sqrt{n}}\right)$$

# Comparing samples from paired populations

If we **know the true variance** $\sigma_D^2$, then we can use a $Z$-test:

$$H_0 : \mu_D = 0 \quad H_1 : \mu_D \neq 0$$

corresponds to $p$-value:

$$P_{H_0}(|\overline{D}| \geq |\overline{d}|) = 2P_{H_0}\left(\frac{\overline{D}}{\sigma_D/\sqrt{n}} \geq \left|\frac{\overline{d}}{\sigma_D/\sqrt{n}}\right|\right) \overset{CLT}{\approx} 2\left(1 - \Phi\left(\left|\frac{\overline{d}}{\sigma_D/\sqrt{n}}\right|\right)\right)$$

# Hypothesis testing for comparing paired samples

The paired t-test

# Comparing samples from paired populations

If we **don't know the true variance** $\sigma_D^2$, but the data comes from **normal distributions**, then we can use a $t$-test:

$$H_0 : \mu_D = 0 \quad H_1 : \mu_D < 0$$

corresponds to $p$-value:

$$P_{H_0}(\overline{D} \leq \overline{d}) = P_{H_0}\left(\frac{\overline{D}}{s_d/\sqrt{n}} \leq \frac{\overline{d}}{s_d/\sqrt{n}}\right) = P\left(t_{n-1} \leq \frac{\overline{d}}{s_d/\sqrt{n}}\right)$$

where $s_d$ is the sample standard deviation from our observed differences $d_1, ..., d_n$.

and similarly for the other alternative hypotheses.

# Hypothesis testing for comparing paired samples

The non-parametric Wilcoxon signed rank test

# Comparing samples from paired populations

What if we don't know the true variance, $\sigma_D^2$, and we also don't think our data comes from a normal distribution?

We can use a non-parametric rank test to test the hypothesis. This test is called the **Wilcoxon signed rank test**, and is the paired sample analog to the Mann-Whitney test.

# Wilcoxon signed rank test

Technically, the WSRT is testing whether the $D_i$ come from a symmetric distribution, but think of this as testing the familiar

$$H_0 : \mu_d = 0$$

since if there is no difference between the two paired conditions, we expect about half of the $D_i$ to be positive and half negative.

# Wilcoxon signed rank test

The general procedure is:

- Remove observations with no difference ($D_i = 0$)
- Calculate $R_i$ = the rank of $|D_i|$
- Calculate $W_i = \text{sign}(D_i) \times R_i$
- Compute the test statistic

$$\boxed{W_+ = \sum_{i: W_i > 0} W_i}$$

If $H_0$ is true, then

$$E(W_+) = \frac{n(n+1)}{4} \qquad Var(W_+) = \frac{n(n+1)(2n+1)}{24}$$

so we can use a normal approximation to calculate a $p$-value.

# Wilcoxon signed rank test

If $H_0$ is true, then

$$E(W_+) = \frac{n(n+1)}{4} \qquad Var(W_+) = \frac{n(n+1)(2n+1)}{24}$$

and it turns out that the standardized version of the statistic tends to a normal distribution:

$$\frac{W_+ - E(W_+)}{\sqrt{Var(W_+)}} \sim N(0,1)$$

so our $p$-value for the alternative $H_1 : \mu_d < 0$ is

$$P\left(\frac{W_+ - E(W_+)}{\sqrt{Var(W_+)}} \leq \frac{w_+ - E(W_+)}{\sqrt{Var(W_+)}}\right) = \Phi\left(\frac{w_+ - E(W_+)}{\sqrt{Var(W_+)}}\right)$$

and similarly for the other alternatives.

Note that $w_+$ is the observed value of our test statistic.

# Example: Wilcoxon signed rank test

Suppose we have four pairs of "before" and "after" measurements, and we want to test the hypothesis that the "after" measurements are larger than the "before" measurements. That is

$$H_0 : \mu_d = 0 \qquad H_1 : \mu_d > 0$$

| Before | After | Difference ($D$) | \|Difference\| ($\|D\|$) | Rank ($R$) | Signed Rank ($W$) |
|---|---|---|---|---|---|
| 25 | 27 | 2 | 2 | 2 | 2 |
| 29 | 25 | -4 | 4 | 3 | -3 |
| 60 | 59 | -1 | 1 | 1 | -1 |
| 27 | 37 | 10 | 10 | 4 | 4 |

Thus our observed test statistic is

$$w_+ = 2 + 4 = 6$$

# Example: Wilcoxon signed rank test

$$H_0 : \mu_d = 0 \qquad H_1 : \mu_d > 0$$

Our test statistic is

$$w_+ = 6$$

And

$$E(W_+) = \frac{n(n+1)}{4} = \frac{4 \times 5}{4} = 5$$

$$Var(W_+) = \frac{n(n+1)(2n+1)}{24} = \frac{4 \times 5 \times 9}{24} = \frac{15}{2}$$

So we can estimate our $p$-value using the normal approximation

$$p \approx P\left(Z > \frac{6-5}{\sqrt{15/2}}\right) = 1 - \Phi(0.365) = 0.358$$

And we thus fail to reject the null hypothesis: not enough evidence to show that the "after" measurements are larger.

# Exercise

# Exercise (Rice Exercise 11.6.22)

An experiment was done to compare two methods of measuring the calcium content of animal feeds. The standard method uses calcium oxalate precipitation followed by titration and is quite time-consuming. A new method using flame photometry is faster. Measurements of the percent calcium content made by each method of 118 routine feed samples are contained in the file `calcium.csv`. Analyze the data to see if there is any systematic difference between the two methods. Use both parametric and nonparametric tests.