Lecture 3

Caden Hewlett

2024-01-15

The Rank Sum Test

The Wilcoxon rank sum test is a test for the difference in Distributions, namely, in a property called *distribution shift.*

This property is also known as "slippage." Essentially, we consider F_X, which generates $x_1, x_2, \dots x_m$ and F_Y, which generates $y_1, y_2, \dots y_n$, where n may not equal m. We then test the hypotheses:

$$H_0 : F_X = F_Y$$

Against the $\neq$, $>$, $<$ alternatives.

Rank Sum Test: Overview

We then consider the rank function, which essentially will quite literally provide the rankings for a list of numbers.

```
x = c(3, 6, 7, 2, 5)
y = c(8, 9, 7.1)
rank(x)
```

```
## [1] 2 4 5 1 3
```

Note that the smallest number gets a rank of 1, and the largest gets the rank of n, where n is the number of observations.

Finding W_X.

To conduct a Wilcoxon Rank Sum test, we will let $\vec{x} = x_1, \dots x_m$ and $\vec{y} = y_1, \dots y_n$.

1.) The first thing we do is combine these two sets of observations into a single vector, which I will call $\vec{z}$.

```
x = cbind(x, "x"); y = cbind(y, "y")
z = data.frame(rbind(x, y))
colnames(z) = c("source", "name")
```

2.) Then, we take the rank of $\vec{z}$, which we will call R.

$$R = \text{rank}(\vec{z}), \text{ where } |R| = m + n$$

```
z$R = rank(z$source)
t(z$R)
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]
## [1,]    2    4    5    1    3    7    8    6
```

3.) Now, we find W_X, the sum of the ranks of the x observations. We could have just as easily found W_Y

$$W_X = \sum_{i=1}^{m} R_i^X$$

Where R_i^X is the i-th x value in the ranks of combined readings.

```
    wx = sum(z$R[z$name == "x"])
    wx
```

```
## [1] 15
```

4.) Then, for n and m sufficiently large,

$$W_X \sim N\Big(\mu = \frac{1}{2}n(n+m+1), \sigma^2 = \frac{1}{12}nm(n+m+1)\Big)$$

However, this is a non-parametric test so we normally use small data sizes. The exact method for this will be explained later.

Rank Sum Test: Intuition

The intuition behind this test, while confusing at first, isn't too bad to understand. If we assume that H_0 is true; that is, the data generators F_X and F_Y are the same, we would expect the ranking sums to be relatively similar.

This is due to the fact that the arrangement of the rank vector, R, will be a random "jumble" of x and y observations. For example,

$$R_{\text{names}} = \{x, y, x, x, y, x, y, x\}$$

We would want the p-value of a test like this to be **small.**

In contrast, if the data generators F_X and F_Y are vastly different, there will be an apparent order to the rankings, i.e.:

$$R_{\text{names}} = \{x, x, x, x, x, y, y, y\}$$

We would want the p-value of this to be very **large.**

So the question remains: given a combined vector of rankings R, how likely are we to see the specific sum of ranks W_X?

Rank Sum Test: Combinatorics

The next step is to consider all possible combinations of organizations of ranks, and the corresponding sums of ranks. Then, we can get some intuition on how rare our particular W_X is.

In this case, there are ${}_nC_r$ options, where $n = m + n$ and $r = m$. Since we normally have small sample sizes, there aren't an incredibly large number of combinations.

```
m = length(x); n = length(y)
choose(m + n, m)
```

```
## [1] 8008
```

We will denote the following R matrix to be the set of all possible combinations of our observations.

For example,

$$R' = \begin{pmatrix} \{x_1, x_2, x_3, \ldots, x_m, y_1, y_2, \ldots, y_n\}, \\ \{x_2, x_1, x_3, \ldots, x_m, y_1, y_2, \ldots, y_n\}, \\ \vdots \\ \{x_1, x_2, \ldots, x_k, y_1, y_2, \ldots y_\ell, x_{k+1}, \ldots, x_m, y_{\ell+1}, \ldots, y_n\}, \\ \vdots \\ \{y_1, y_2, \ldots, y_n, x_1, x_2, x_3, \ldots, x_m\} \end{pmatrix}$$

Even in our small case, there are many combinations.