Understanding the Mann-Whitney U Test

This document goes into some of the details of the Mann-Whitney U test that are not covered in the videos of the Introduction to Data Science. Reading this will give a better intuition of how to interpret the values that come out of the test when using statistical software such as scipy.

The **Mann-Whitney U test** (also known as **Wilcoxon rank-sum test**) is a *non-parametric* test that can be used to test, for two populations with unknown distributions, if we draw randomly from each distribution, whether one distribution is more likely to generate a higher value than the other. Stated in mathematical terms, given random draws x from population X and y from population Y, the standard *two-tailed* hypotheses are as follows:

$$H_0 : P(x > y) = 0.5$$

$$H_1 : P(x > y) \neq 0.5$$

Note that this is not a hypothesis test of whether or not two distributions are the same, nor is it a test of whether or not the median of two distributions are equal. While the most common assumption under the null hypothesis is that the distributions being compared are identical, this need not be the case for the null hypothesis to be true. It is for this reason that it is recommended to report additional descriptive statistics, such as median and interquartile range, to supplement the statistics generated from the Mann-Whitney U test.

Calculating U_x and U_y

Let us assume we have n_x items sampled from population X and n_y items from population Y. Arrange all $N = n_x + n_y$ items in order from smallest to largest. Compute the total sum of the ranks for each of the samples, where the smallest item has a rank of 1, the next smallest a rank of 2, and so on, up to the largest with the rank of N. (In the case of ties, an item's rank is equal to the median of the tied items' ranks if they were not tied) Call these rank-sums R_x and R_y. Since the sample sizes may be unequal, we cannot directly compare R_x and R_y. We take into account the size of each sample by subtracting from the rank-sums the worst case scenario where all items in the sample were the smallest overall. Since

$$1 + 2 + ... + k = k(k + 1)/2 \,,$$

$$U_x = R_x - n_x(n_x + 1)/2$$

and

$$U_y = R_y - n_y(n_y + 1)/2 \,.$$

Calculating U

Consider the total sum of ranks R_x + R_y. This will be a constant value, as all items in the full list must be part of the sample from X or the sample from Y; the total sum of ranks is the sum of integers from 1 to N. Using the formulas above and performing some algebra, we verify the same for the sum U_x + U_y:

$$R_x + R_y = N(N + 1)/2 = U_x + n_x(n_x + 1)/2 + U_y + n_y(n_y + 1)/2 \,,$$

$$U_x + U_y = [(n_x + n_y)(n_x \; + n_y + 1) - n_x(n_x + 1) - n_y(n_y + 1)]/2 \,,$$

$$U_x + U_y = n_x n_y \,.$$

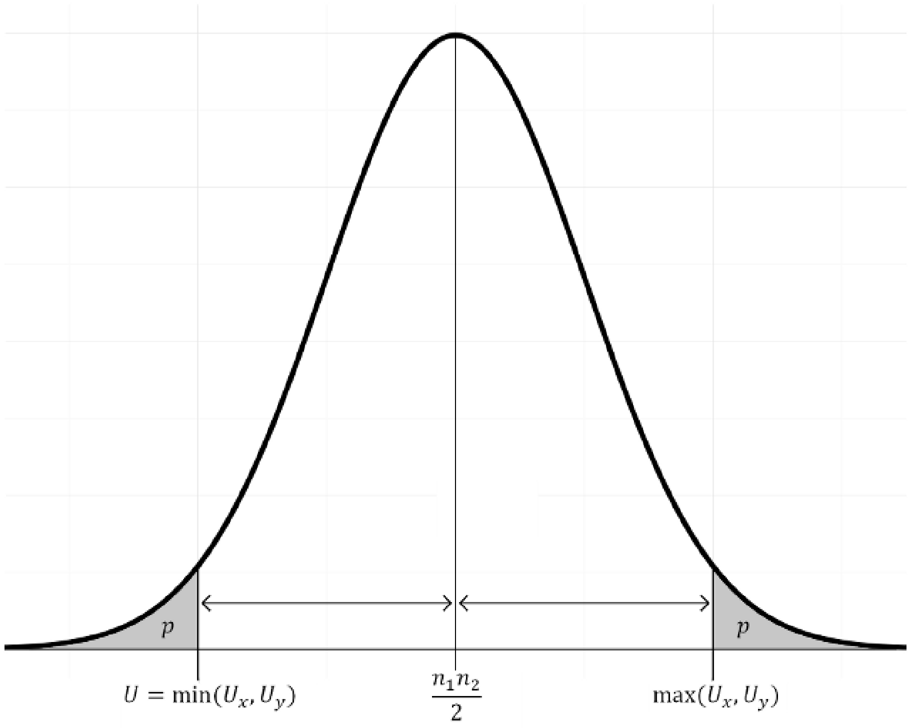
Any change in U_x will be met with an equal and opposite change in the value of U_y. Under the null hypothesis, we will expect the values of U_x and U_y to be close to one another, centered around $n_x n_y/2$. Taking these two statements together, we observe the following:

$$|U_x - n_x n_y/2| = |U_y - n_x n_y/2|$$

Values that are far away from $n_x n_y/2$ are indicative that the null hypothesis is not the true state of the world. By convention, the U statistic is reported as the smallest of these values, $U = min(U_x, U_y)$, and smaller values of U are suggestive of more extreme deviations from the null hypothesis.

Calculating a p-value

When we have small sample sizes, pre-computed tables can be used to determine whether or not U has fallen below critical thresholds of *p*-values (usually p = 0.05 and p = 0.01). If we have larger sample sizes (generally taken as at least 20 observations from each population), we can actually approximate the distribution of the U statistic under the null hypothesis as a normal distribution with mean $n_x n_y/2$ and variance $n_x n_y(n_x+n_y+1)/12$.



Using this normal approximation, we obtain a *one-sided* *p*-value from the area captured below U. Under the standard, *two-sided* formulation of the null hypothesis, we need to double this probability, accounting for the fact that a case where U_x and U_y are reversed is just as extreme as the observed scenario under the null.

Using the Mann-Whitney U test in `scipy`

`scipy` takes as arguments two lists of values representing the samples drawn from the two populations and returns the *U* statistic and a one-sided *p*-value. This *p*-value is based off of the normal approximation to the distribution of the *U* statistic, so `scipy` recommends at least 20 values in each list so that the *p*-value is reasonably accurate.

Take care in the reporting of statistics generated by the `scipy` function. Normally, the formulation of the hypothesis when using the Mann-Whitney U test is *two-tailed*, so be certain to double the *p* generated by the `scipy` function in order to report the proper *p*-value. Even if you have specified a *one-tailed* test, you cannot necessarily report the *p* directly from `scipy`. If the observed relationship is in line with the parameters of your null hypothesis (e.g. your alternative hypothesis suggests that population X takes smaller values than Y, and you observe the opposite in your data), then you will need to report 1-*p* as your *p*-value (which will be larger than 0.5).

In either case, you will want to be explicit about the hypothesis type you are testing (*one-sided* vs. *two-sided*) and report descriptive statistics about each distribution (e.g. median, interquartile range) to support your conclusions. The Mann-Whitney U has a very specific meaning, so use of supplemental statistics will strengthen the arguments you make.