A Primer on Calculating the Mann-Whitney Statistic

Hello everyone, The Mann-Whitney test is pretty useful. I told my discussion section it's one of the bread-and-butter techniques in nonparametric statistics, because if the data is actually normal, it's almost as efficient (i.e.: almost the same power) as the t-test (which relies on normality), and if the data isn't normal, it's often far more efficient. I wrote this walkthrough a few years ago to explain to a biostatistics class how to calculate the MW U-Statistic the way you discussed in lecture. I hope this helps some of you understand how to compute the statistic.

I will use the dataset from page 284 (example 7.10.2) of Statistics for the Life Sciences, Fourth Edition, by Samuels, Witmer, and Schaffner, and test the hypothesis that the distribution of population 1 is equal to population 2, against the alternative that between the two at $\alpha = 0.05$:

X	64	315	17	170	20	22	190	
Y	13	14	15	6	29	16	18	22

1 Order the Data

Order the data so that the smallest values are first and the largest values are last. This makes it easier to do the later steps. Reproduced below are the two, ordered, samples.

X	17	20	22	64	170	190	315	
Y	6	13	14	15	16	18	22	29

2 Count Number of X Smaller Than Y

Starting with the first sample, X, count how many values in Y are smaller than each value of X. For example, the first value of X is 17. The Y values 6, 13, 14, 15, and 16 are all smaller than 17, so the first number of observations in Y smaller than the first observation of X is 5. The second value of X is 20, and in Y, values 6, 13, 14, 15, 16, and 18 are smaller. This means the number of observations smaller than the second observation is 6.

For the third number, 22, you'll notice there are ties with Y. Instead of giving the usual point value of 1 to a greater number, if you encounter a tie, add .5. Thus, the third observation of X (22) is bigger than 6, 13, 14, 15, 16, 18, and ties with 22, for a score of 6.5. Repeat this process for all remaining values of X. You will get the following numbers of X's greater than Y's: 5, 6, 6.5, 8, 8, 8, 8. Note that the last four observations of X (64, 170, 190, 315) are all greater than the largest number in Y, so the number of observations of Y that are smaller than X will be the same for the last four observations. The K_1 statistic is the sum of all those numbers. $5 + 6 + 6.5 + 8 + 8 + 8 + 8 = 49.5$.

Now, there's a nifty property that shows that if you repeated the procedure but with X and Y flipped, their sum would be the product of the two sample sizes. We will leverage

this fact to avoid having to do this procedure again. If $U_{lower} + U_{upper} = nm$, then $U_{lower} = nm - U_{upper}$. So the lower statistic will be $(7 \cdot 8) - 49.5 = 6.5$.

3 Look Up the Statistic in the Book

The book is going to give you a range of probable lower and upper statistics for a few levels of significance. X has 7 observations and Y has 8, so $n = 7$ and $m = 8$ (if you confuse n and m, it's okay, because it's symmetric–the table gives the same values as $n = 8$ and $m = 7$). The region to we need to hit is $(13, 43)$. If our lower statistic is less than 13 or our upper statistic is bigger than 43, we would have evidence to reject our null hypothesis.

As a note, you only need to check either the upper *or* the lower value, because once you know the sample sizes and just one of the statistics (lower or upper), you can calculate the other one.

Our statistic is 49.5. Since we said we would reject if the upper statistic was bigger than 46, we reject H_0, and conclude that there are differences in the distributions of X and Y.

4 How to do This in R

Get the data into R. Then use the function `wilcox.test`.

```
1 X = c(64, 315, 17, 170, 20, 22, 190)
2 Y = c(13, 14, 15, 6, 29, 16, 18, 22)
3 wilcox.test(X, Y)
```

```
Wilcoxon rank sum test with continuity correction
data:  X and Y
W = 49.5, p-value = 0.015
alternative hypothesis: true location shift is not equal to 0

Warning message:
In wilcox.test.default(X, Y) : cannot compute exact p-value with ties
```

For more information about the Wilcoxon-Mann-Whitney function in R, check out the help page:
`http://stat.ethz.ch/R-manual/R-patched/library/stats/html/wilcox.test.html`.
The page also gives you information about how to specify alternative hypotheses, and whether to use a normal approximation.