The Mann-Whitney U test

The T test is actually quite flexible, and providing you have a large enough sample size, can be used even if your data are not normally distributed. But there are times when you may have small sample sizes and find that your data are not normal. You still want to analyze your data and possibly perform some type of two sample test.

The Mann-Whitney U test. This test is part of a large group of tests known as *non-parametric* or *distribution-free*. Tests like the Mann-Whitney U (MWU) test do not require your data to have a particular distribution (e.g, the normal).

The Mann-Whitney U test actually goes by several different names. Some statisticians (and computer packages like R) will call this the Wilcoxon rank sum test. Sometimes it's even called the Wilcoxon-Mann-Whitney U test. The reason for some of this confusion has to do with who published what when, and who provided the (slightly) more useful version of this test.

So why is this test useful? Because it doesn't require *any* assumptions except that the data are random (and, of course, independent for a two sample test). So we no longer have to worry about the normal distribution assumption.

One downside is that in it's original form, the test does not test for equal means. Instead, it tests for equal distributions:

H_0: The distribution of sample 1 is the same as the distribution of sample 2.

Or in symbols: $H_0 : D_1 = D_2$

H_1: The distribution of sample 1 is *not* the same as the distribution of sample 2.

Or in symbols: $H_1 : D_1 \neq D_2$

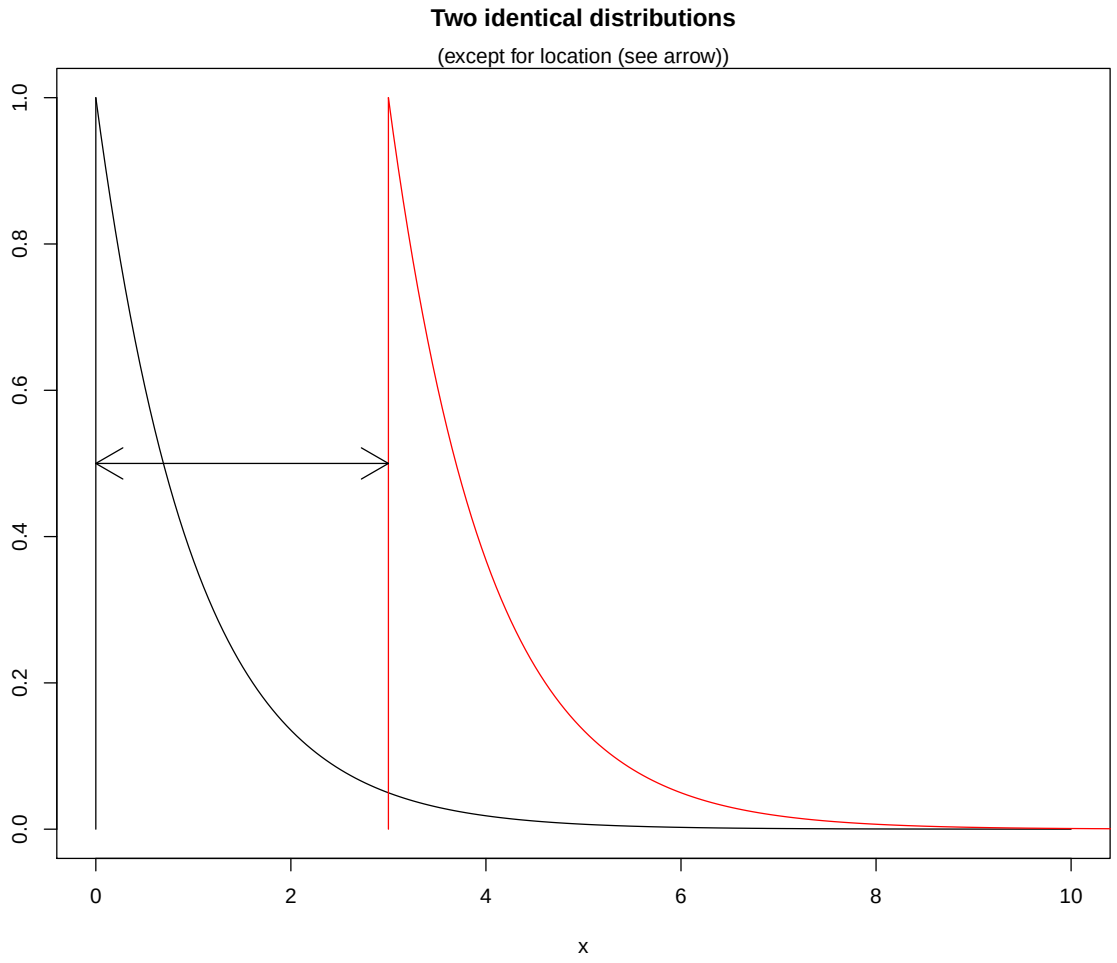
If this is the version of the MWU test that we use, we're pretty much ready to go. The only problem is that sometimes it can be a little difficult to explain to a non-statistician what we mean by saying "the two distributions are not the same". Telling someone that the means are not the same, is usually easier to understand.

It turns out that we *can* use the MWU test to test for equal means, but to do this, we need to make an assumption. This, in a way, ruins some of the benefits of using the MWU test, but let's take a look at the assumption anyway:

We *assume*: the distributions are the same except for location.

In other words, the shapes of the two distributions need to be the same, but their locations

can be different. Let's illustrate with an example:



If you make this assumption, you can now test the usual hypotheses about means:

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

Incidentally, if we make this assumption, we can also use the MWU test to test for equal medians.

So how do we carry out the MWU test? Regardless of which set of hypotheses we want to use (distributions, means, or medians) the math is the same. We proceed as follows:

1. Set up your hypotheses.
2. Pick your value for α.

3. Sort your data in *each* sample from smallest to largest.

4. Now for each data point, look at the other sample. Count how many values are *smaller* than the data point you're looking at and write this number down next to your data point (use 1/2 if a value is perfectly tied - see below under ties).

5. Add up the numbers you wrote down for each sample. This will give you two sums which we will label K_1 and K_2.

6. Check your work. If you did it correctly, then $n_1 \times n_2 = K_1 + K_2$. This doesn't guarantee you did it right, but if this check is not true, you definitely made a mistake.

7. Look at the two sums (K_1 and K_2) and pick the larger of the two. This is your test statistic, U^*. In other words, $U^* = max(K_1, K_2)$.

8. Compare your U^* with the tabulated value of U from the tables. Use your sample sizes, n_1 and n_2 to get the correct value of U_{table}. Also make sure you use the correct table for your value of α.

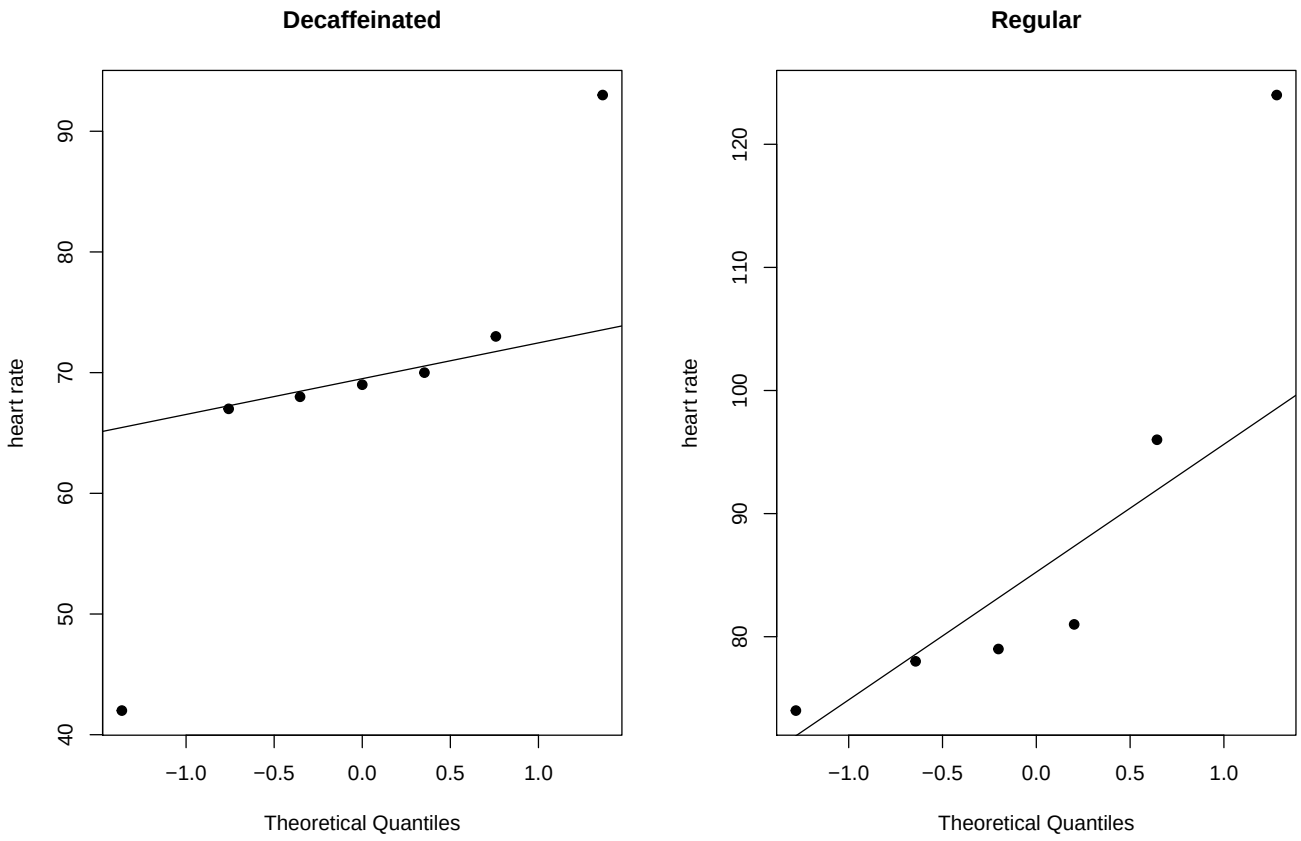
8. Finally, if $U^* \geq U_{\text{table}}$, *reject* H_0, otherwise *fail to reject* H_0.

This may sound a little intimidating or confusing (particularly steps 4 & 5), but it's really not that difficult. Let's do an example.

We want to find out if caffeine affects heart rate. We take seven volunteers and measure their heart rate after drinking decaffeinated coffee. We take another six volunteers and measure their heart rate after drinking regular coffee. we get the following (somewhat exaggerated) results, which have already been sorted:

	Decaffeinated Coffee	Regular Coffee
	42	74
	67	78
	68	79
	69	81
	70	96
	73	124
	93	
$\bar{y}$	68.9	88.7

If we look at these data, we'd be tempted to conclude that a T test should be able to find a difference here (there's a pretty big difference in the means. But let's look a the QQ plots first:



The QQ plots show that the data are seriously not normal. The data for decaffeinated coffee show long tails, and the data for regular coffee are skewed right. This means that a T test is not appropriate here. (You should *always* do QQ plots before doing a T test).

So let's set up the hypotheses. We'll assume equal distributions except for location, which might not be true looking at the QQ plots (long tailed and skewed indicates they're different).

> H_0: The true mean heart rate of people drinking decaffeinated coffee is the same as the true mean heart rate for people drinking regular coffee ($\mu_d = \mu_r$).
>
> H_1: The true mean heart rate of people drinking decaffeinated coffee is *not* the same as the true mean heart rate for people drinking regular coffee ($\mu_d \neq \mu_r$).

We'll pick $\alpha = 0.05$.

Now let's start counting the number of data values that are less than the one we're looking at in the *other sample*. Notice also that the data have been re-arranged just a bit to make it easier to do this:

		Decaffeinated Coffee	Regular Coffee		
no values less than					
42 in other sample →	0	42			
	0	67			
	0	68			
	0	69			
	0	70			
	0	73			six values less than
			74	6	← 74 in other sample
			78	6	
			79	6	
			81	6	
	4	93			
			96	7	
			124	7	
$K_1 =$	4			38	$= K_2$

The sums of our counts are given by K_1 and K_2. Before we go on, let's check our work:

$$n_1 \times n_2 = K_1 + K_2 \text{ and we have:}$$

$$K_1 = 4, K_2 = 38 \text{ so:}$$

$$7 \times 6 = 38 + 4 = 42 \text{ correct!}$$

So now we can move on and get U^*. This is the larger of the two values, K_1, K_2, or using mathematical notation:

$$U^* = \max(K_1, K_2) = \max(4, 38) = 38$$

Now let's go into our U tables. We make sure to use the correct U table by choosing the one that is labeled "two sided probability = 0.05". Now we go into this table using $n_1 = 7$ and $n_2 = 6$ (the tables are set up so it's irrelevant which is n_1 and n_2). We find that $U_{\text{table}} = 36$.

From this we can conclude that since,

$$U^* = 38 \geq U_{\text{table}} = 36$$

is true, we *reject* our H_0 and conclude that caffeine in coffee does affect (increase) heart rate.

Some theory on the MWU test. We can't dig deep into the theory of the MWU test in an introductory class, but we do at least want to have some understanding of *why* it

works.

Lets suppose for a minute that *all* of the values for regular coffee had been higher than the any of the values for decaffeinated coffee. What would have changed? We would have had six 7's for regular coffee (and seven 0's for decaffeinated coffee) giving us $K_1 = 0$ and $K_2 = 42$, and so $U^* = 42$.

Larger values of U^* indicate that the samples are further apart. In this case, a value of 42 is as far apart as the data in our two samples can get (U^* can't get larger than 42). That's pretty good evidence that there's a difference in our heart rates. But this doesn't give us a probability.

Or, in other words, if the null hypothesis (H_0) is correct, what's the probability of getting $U^* = 42$? We actually know how to do this! Let's write out our probability a bit more formally and extend it to the values for K_1 and K_2.

$$Pr\{U^* = 42\} \;=\; Pr\{K_1 = 0 \text{ and } K_2 = 42\} \;\; \textbf{or} \;\; Pr\{K_1 = 42 \text{ and } K_2 = 0\}$$

To calculate the probability that $U^* = 42$ we need to look at both ways that we can get 42. Either all the values in sample 1 are larger than any of the values in sample 2, or all the values in sample 1 are smaller than any of the values in sample 2.

Let's see if we can deal with $Pr\{K_1 = 0 \text{ and } K_2 = 42\}$. Basically what the MWU test does is figure out how many different ways can we arrange the values in our two samples and use that to figure out our probability. We have a total of 13 data values in our example above. We have seven values for decaffeinated coffee and six for regular coffee.

We can look at this as a problem of choosing seven people drinking decaffeinated coffee out of 13 total, so we can use the binomial coefficient to calculate the number of different ways of getting 7 people out of 13:

$$\binom{13}{7} = 1,716$$

So now we have 1,716 different possible arrangements for the data set we have above. If $K_2 = 42$, this means (as mentioned) that all the values in sample 2 are larger than those in sample 1. This is *one* arrangement out of 1,716. Or, simply:

$$Pr\{K_1 = 0 \text{ and } K_2 = 42\} = \frac{1}{1,716} = 0.0005828$$

It should be obvious that the probability of $K_1 = 42$ is the same, so now we can do:

$$\begin{aligned} Pr\{U^* = 42\} \;&=\; Pr\{K_1 = 0 \text{ and } K_2 = 42\} \;\; \textbf{or} \;\; Pr\{K_1 = 42 \text{ and } K_2 = 0\} \\ &= 0.0005828 + 0.0005828 = 0.0011656 \end{aligned}$$

And finally we can say that the p-value = 0.0011656. Since our p-value $\leq \alpha$, we could reject our null hypothesis if we had gotten $U^* = 42$.

The actual calculations our slightly more difficult (mostly "tedious"), but essentially this is what makes the MWU test works.

The problem with really small sample sizes. One problem with the MWU test is that really small sample sizes sometimes don't give you small enough probabilities. What does this mean? Let's suppose we have two small samples, with $n_1 = 4$ and $n_2 = 4$. Following the rationale above, we notice that $n_1 \times n_2 = 4 \times 4 = 16$. So the largest possible value for $U^* = 16$. What's the probability of getting $U^* = 16$?

We have a total of eight data points, and we want to pick four of them, so:

$$\binom{8}{4} = 70$$

Which implies:

$$\begin{aligned} Pr\{U^* = 16\} &= Pr\{K_1 = 0 \text{ and } K_2 = 16\} \text{ or } Pr\{K_1 = 16 \text{ and } K_2 = 0\} \\ &= \frac{1}{70} + \frac{1}{70} \\ &= 0.0142857 + 0.0142857 = 0.0285714 \end{aligned}$$

So the probability of getting $U^* = 16$ is 0.02585714. If $\alpha = 0.05$, we could reject. But what if $\alpha = 0.01$?

If $\alpha = 0.01$, we could *never* reject H_0 ! No matter how different our two samples, we could never get a probability (= p-value) small enough to reject our null hypothesis. We need a larger sample size before we can probabilities small enough. If you look up the U_{table} value for $\alpha = 0.01$ with $n_1 = 4$ and $n_2 = 4$, you'll find N/A, which in the table means you can *not* possibly reject because you can't get a p-value smaller than α.

Think of it this way. You toss a coin twice and get two heads. Can you claim that the coin is unfair? The probability of two heads (if the coin is fair) is 0.25. This isn't small enough for you to say anything about the coin being unfair. If you toss a coin four times and get four heads, the probability is 0.0625. If you're using $\alpha = 0.10$, you could claim the coin is unfair. If you want to reject at $\alpha = 0.05$, you'd have to toss the coin at least five times (p = 0.03125). You'd have to toss it seven times and get seven heads before you could reject at $\alpha = 0.01$.

The MWU test works the same way; if you want to use small values of α you need to make sure your sample sizes are large enough so that you can use the value of α that you want.

This is pretty simple to do - just check to see if there's a number in the table for the α you want to use with your sample sizes.

The problem of ties. The MMU test, just like the T test is designed for continuous data. If the data are truly continuous, then the probability of a value in one sample being tied with a value in the other sample is exactly 0.

Suppose I take a sample of heights of male students and another sample of heights of female students. What is the probability that the height of a male student is *exactly* the same as the height of a female student? The answer is actually trivial if we look at it this way:

$$Pr\{Y_m = 67.000000000000... \text{ and } Y_f = 67.000000000000...\} = 0$$

In other words, the MWU test assumes no ties, because they shouldn't exist. But obviously, they do exist. So how do we deal with ties? For calculations, it's not too difficult. We just use ½ for *each* of the values in the other sample that are tied with the value we're looking at. It's easiest to illustrate this with an example:

		Y_1	Y_2		
	0	1			
			3	1	
	1	4			
one value is less than 3 in →	3	5	5	2.5	← two values are less than 5 in
the other sample; four values			5	2.5	the other sample, one value
are tied (so $1 + 4(1/2) = 3$)			5	2.5	(5) is tied (so $2 + 1/2 = 2.5$)
			5	2.5	
	6	7	6	3	
	6	7			
	6				
$K_1 =$	22			14	$= K_2$

The 5 in the first sample (Y_1) gets a value of 3:

It is tied with four 5's from sample 2 (Y_2), so $4 \times 1/2 = 2$.

One value in sample 2 is less than 5 (the 3), so that gets a 1.

$2 + 1 = 3$

Each of the 5's in the second sample (Y_2) gets a value of 2.5:

Each 5 has two values in sample 1 that are smaller, so they get a 2 for that.

Each 5 is tied with one 5 from sample 1, so that gets ½.

$2 + 1/2 = 2.5$

Incidentally, the check still works:

$$n_1 \times n_2 = 6 \times 6 = 36 = K_1 + K_2 = 22 + 14$$

So we know how to deal with ties. But there's another issue. As mentioned, the MWU test assumes ties shouldn't exist. If your data have a lot of ties (don't worry if you only have a few ties), then you may need to look for other methods to deal with ties as the MWU test may not do so well.

The MWU test versus the T test.

We now have two tests for two samples. Both can be used to test for equal means (although we need to make an assumption for the MWU test). Which is better?

To answer that question we need to remember our discussion of power from the lecture on hypothesis tests. In short, we always try to use the most powerful test. Unfortunately, the test that is most powerful can change based on the distribution of our data.

For example, if the data in both samples are normally distributed, the T test will have the most power. It will do the best job of disproving a false null hypothesis, and will clearly outperform the MWU test. Note that this is true *regardless* of sample size. If you have a small sample with data that are normally distributed the T test *will* do better.

But if the data are *not* normally distributed and you have small samples, the MWU test can become more powerful and do a better job. How well the MWU test does compared to the T test depends on what kind of distribution the data have. A very general comment might be that the less normal the data are, the better the MWU test will do.

Something else that is a little confusing is that the MWU test is a perfectly valid test to use when you have normally distributed data, it's just not the best test to use. But, having said that, the power of the MWU test is actually pretty good compared to the T test if the data are normal.

Incidentally, the opposite is not true - it is incorrect to use the T test with non-normal data and small sample sizes.

So where does that leave us? It's always best to think about things and not just jump in, but here are some general rules that might be helpful.

> If you have very large sample sizes (e.g., each sample has over 75 data points) just use the T test. the CLT will take of any problems.
>
> If you have smaller sample sizes it depends on how badly not normal your data are.

If your data are badly not normal (very long tails, badly skewed, etc.), then you may need sample sizes of 50 or even 75 before you can rely on the CLT and use the T test.

If your data are not too badly not normal (short tails, slight skew, etc.) then you might be able to use a T test with sample sizes as small as 20 or 25.

If you can't use a T test, then the MWU test is most likely the appropriate test to use.

Finally, suppose you need to do a two sample test in a few years and forget (you shouldn't!) the details of when to use a T test versus the MWU test. What do you do? If you need one recommendation for a test that is always valid and doesn't do too badly even if it's not the best test to use, use the MWU test. It always works.

(Of course, what you really should do is look up when to use which test, check the assumptions, etc., and proceed to do the best possible analysis).