So what do you do if the data are not normal and you still need to perform a test?

Remember, if your $n$ is reasonably large, don’t bother doing anything. Your $\bar{y}$ will have a distribution that’s approximately normal.

What is large? Depends on how not normal your original data are, but somewhere between 20 and 30 is often adequate.

But suppose you only have small sample sizes, and your data are not normal. Then what???

The Mann-Whitney $U$-test.

(Note: Many statisticians (and several computer packages) will call this the Wilcoxon rank sum test. It's mathematically equivalent).

The only assumption for this test is that the data are random.

You still need to make sure you collected the data randomly!

What are you testing? That the two distributions are the same!

In other words, you test the following:

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

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

In symbols: $H_1: D_1 \neq D_2$

This is the original form of the Mann-Whitney $U$-test, but often we want to say something about the means (or medians) as this is a little easier to explain.

If we assume that the distributions are only different in location, but not in shape (e.g., two binomials, two uniforms, etc.), then we can use this test to test for equal means!
We will make this assumption in this class and then we can test the usual hypotheses:

\[ H_0: \mu_1 = \mu_2 \]

\[ H_1: \mu_1 \neq \mu_2 \]

(Incidentally, if we make the assumption of equal distributions except for locations we could test for equal medians as well.

The test is identical; we just change \( H_0 \) and \( H_1 \) to use medians instead of means.)

So now we proceed as usual:

Figure out our \( \alpha \).

Calculate our test statistic (see example on board and then below).

As you might guess, this is called “\( U \)”, not “\( t \)”

i) rank your data, from smallest value to largest.

ii) for each data point, write down the number of values that are smaller IN THE OTHER SAMPLE next to it (use \( \frac{1}{2} \) if a value is perfectly tied).

iii) Add up these numbers for both samples

vi) check your work - the two sums should add up to the product of the sample sizes (see the example below).

v) Choose the larger of these two sums, that’s our \( U^* \).
Compare your $U^*$ with the tabulated value of $U$:

Use the correct table for your value of $\alpha$

The find the correct values of $n_1$ and $n_2$:

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

Otherwise, “fail to reject” $H_0$.

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 (exaggerated) results (already sorted):

<table>
<thead>
<tr>
<th>Decaffeinated Coffee</th>
<th>Regular Coffee</th>
</tr>
</thead>
<tbody>
<tr>
<td>42</td>
<td>74</td>
</tr>
<tr>
<td>67</td>
<td>78</td>
</tr>
<tr>
<td>68</td>
<td>79</td>
</tr>
<tr>
<td>69</td>
<td>81</td>
</tr>
<tr>
<td>70</td>
<td>96</td>
</tr>
<tr>
<td>73</td>
<td>124</td>
</tr>
<tr>
<td>93</td>
<td></td>
</tr>
<tr>
<td>$\bar{y}$</td>
<td>68.9</td>
</tr>
</tbody>
</table>

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

![Q-Q Plots](image-url)
The q-q plots show that the data are seriously not normal (you should always do q-q plots before doing a t-test!)

So a t-test is not appropriate here, and we'll do the MWU test instead.

Let's set up the hypotheses (we'll assume equal distributions except for location - which might not be entirely true looking at the q-q plots):

\[ H_0: \text{the true mean heart rate for people drinking decaffeinated coffee is the same as the true mean heart rate for people drinking regular coffee.} \]

\[ H_1: \text{the true mean heart rate for people drinking decaffeinated coffee is NOT the same as the true mean heart rate for people drinking regular coffee.} \]

Let's pick \( \alpha = 0.05 \)

Now we count the smaller data points in the other sample (doing both columns at once):

<table>
<thead>
<tr>
<th>( Y_1 )</th>
<th>( Y_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>42</td>
</tr>
<tr>
<td>0</td>
<td>67</td>
</tr>
<tr>
<td>0</td>
<td>68</td>
</tr>
<tr>
<td>0</td>
<td>69</td>
</tr>
<tr>
<td>0</td>
<td>70</td>
</tr>
<tr>
<td>0</td>
<td>73</td>
</tr>
<tr>
<td>74</td>
<td>6</td>
</tr>
<tr>
<td>78</td>
<td>6</td>
</tr>
<tr>
<td>79</td>
<td>6</td>
</tr>
<tr>
<td>81</td>
<td>6</td>
</tr>
<tr>
<td>4</td>
<td>93</td>
</tr>
<tr>
<td>96</td>
<td>7</td>
</tr>
<tr>
<td>124</td>
<td>7</td>
</tr>
</tbody>
</table>

\[ K_1 = 4 \quad K_2 = 38 \]

You don't have to arrange things this way, but it's probably a little easier if you sort both columns together like this.

The sums of our counts are referred to as \( K_1 \) and \( K_2 \).

We check our work:

\[ n_1 \times n_2 = K_1 + K_2 \]

\[ K_1 = 4 \quad K_2 = 38 \]

And we have: \( 7 \times 6 = 38 + 4 = 42 \) (correct!)
We pick the larger of $K_1$ and $K_2$ and set this equal to $U^*$:

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

Now look this up in the tables:

$$n = 7, \ m = 6, \ \alpha = 0.05$$

(remember, we're using $n$ as the larger sample size and $m$ as the smaller sample size)

So we get: $U_{\text{table}} = 36$, and since:

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

we reject $H_0$ and conclude that the means are different.

Some theory on this test (or “why does it work?”)

Here's an outline of what's happening:

Suppose $n_1 = 5$ and $n_2 = 4$, then $K_1 + K_2$ must be 20 (our check gives us $5 \times 4 = 20$).

So, suppose there is no overlap at all (the data is least compatible with $H_0$).

This implies that either $K_1$ or $K_2 = 20$, and therefore $U^* = 20$.

But if the data overlap perfectly, then $K_1 = K_2 = 10$, and $U^* = 10$.

This implies that large values of $U^*$ are incompatible with $H_0$, and this is the basis of the test.

So where does $U$ come from? Better, how is $U$ distributed?

Remember, $t$ has a $t$-distribution, $z$ a normal, etc.

Note that $U$ is discrete.

We can calculate the probabilities associated with any given arrangement of the two samples.

For our example here ($n_1 = 5$ and $n_2 = 4$) the probability of getting

$$Pr\{K_1 = 0 \text{ and } K_2 = 20\} = .008$$

This is one possible arrangement out of 126 different possibilities, or $1/126 = 0.008$

How do we know there are 126 different possible arrangements?

We can use the binomial coefficient. We have a total of $4 + 5 = 9$ items, we want to choose $5 \ (= n_1)$ of them. In other words, how many ways can we pick 5 items from 9?
\[ \binom{9}{5} = 126 \]

So we just go through and figure out the probabilities for all possible combinations of \( K_1 \) and \( K_2 \), for as many different values of \( n_1 \) and \( n_2 \) as we want.

Incidentally, since we're doing a two sided test, we also need to calculate

\[ Pr\{K_1 = 20 \text{ and } K_2 = 0\} \]

which is also (of course), 0.008

So for a two sided test we need:

\[ Pr\{U^* = 20\} = Pr\{K_1 = 0 \text{ and } K_2 = 20\} + Pr\{K_1 = 20 \text{ and } K_2 = 0\} \]

\[ = 0.008 + 0.008 = 0.016 \]

Finally, note again that for this example, \( Pr\{U^* = 20\} = .016 \).

So what happens if we choose \( \alpha = .01 \)?

We loose! \( U \) is discrete, and can’t take on values greater than 20 (if \( n_1 = 5 \) and \( n_2 = 4 \)).

It is impossible to reject \( H_0 \) if \( n_1 = 5 \) and \( n_2 = 4 \), and we pick \( \alpha = 0.01 \) (and we're doing a two sided test).

\( U^* = 20 \) is the most extreme outcome, and this probability is not small enough to reject at \( \alpha = 0.01 \)

If you have really small sample sizes, you should probably check in the table first to make sure you can actually use the value of \( \alpha \) that you want.

An assumption of the Mann-Whitney \( U \)-test that's not obvious.

Just like the \( t \)-test, the Mann-Whitney \( U \)-test is designed for continuous data.

It may not do well with discrete data. In particular, it doesn't like ties.

If data are truly continuous, the probability of any one value in sample 1 being the same as any one value in sample 2 is exactly 0.

But before we look at this further, how do you deal with ties?

When you have data that are tied, you use \( \frac{1}{2} \) for EACH of the data points in the other sample that are tied with the data point you’re looking at.

Probably the easiest way to describe this with an example or two:
Example 1:

<table>
<thead>
<tr>
<th></th>
<th>$Y_1$</th>
<th>$Y_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>2.5</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>2.5</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>2.5</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>8</td>
<td></td>
</tr>
</tbody>
</table>

Sum 22 14

Note that the 5 in sample 1 ($Y_1$) gets a value of 3.

This is because it is tied with four 5's in sample 2, and $4 \times 1/2 = 2$

Any number less than our number is counted as usual, so overall we have:

$$1 + (4 \times 1/2) = 3.$$  

The 5's in sample 2 ($Y_2$) each get 2.5 (there are two numbers less than 5 in sample 1; the 5 in sample 1 gets 1/2):

$$2 + 1/2 = 2.5$$

Incidentally, the check still works ($n_1 \times n_2 = 6 \times 6 = 36$, and $K_1 + K_2 = 36$)

Example 2 (just the results this time, skipping the details):

<table>
<thead>
<tr>
<th></th>
<th>$Y_1$</th>
<th>$Y_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>1.5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>1.5</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>3.5</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>3.5</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>3.5</td>
</tr>
<tr>
<td>5</td>
<td>7</td>
<td></td>
</tr>
</tbody>
</table>

Sum 11.5 13.5

And yes, the check still works.
The Mann-Whitney U test versus the t-test.

You now have two tests that can each be used to test for equal means.

Which is better? The answer depends on the power of the test.

If the data are normal or the n's are large, the t-test is better.

If the data are not normal, and the n's are small, the Mann-Whitney U-test is better.

Sometimes, particularly if you have very long tails, it is much better.

Incidentally, if the data are normal, the MWU test is still valid, just not as good as the t-test.

So how do you decide which test to use? Follow the guidelines given above.

If \( n_1 \) and \( n_2 \) are both less than about 20 to 30, and the data are badly not normal, use the MWU test.

Otherwise, you can probably use the t-test.

Of course, remember, the less normal your data are, the bigger the n's should be.

If you can’t check your assumptions, or forget in a couple of years what to do to make sure the t-test works, use the MWU test.

Even if the data are normal, it actually has reasonable power (though obviously not as good as the t-test).

It’s almost always valid, unless the data are not random.