**The Mann-Whitney Test**

In a laboratory, organisms are grown in glass flasks.
These flasks can either be left stationary, or aerated by continuously being shaken.
To compare the effects of these two growing conditions, the generation times of individual creatures, in hours, are measured:

| | | | | | | |
|---|---|---|---|---|---|---|
| Aerated | 5.5 | 4.8 | 3.7 | 8.1 | 5.2 | |
| Stationary | 8.2 | 5.6 | 7.9 | 5.4 | 7.1 | 6.5 |

We wish to know whether aerating the flasks has a shortening effect on the generation time of organisms.

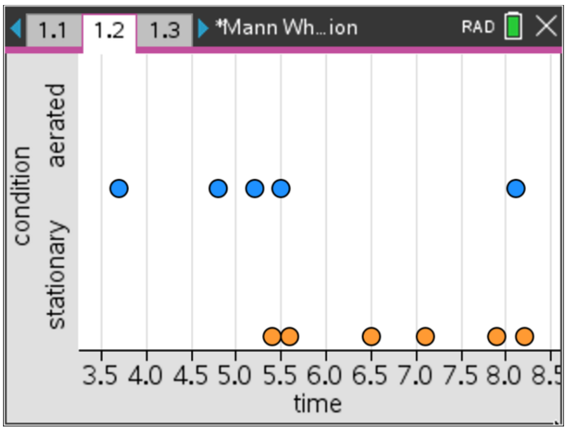
---

We shall assume that this laboratory's results are a random sample from the population of similarly conducted experiments.

We have non-paired data, so a Mann-Whitney test is potentially appropriate.

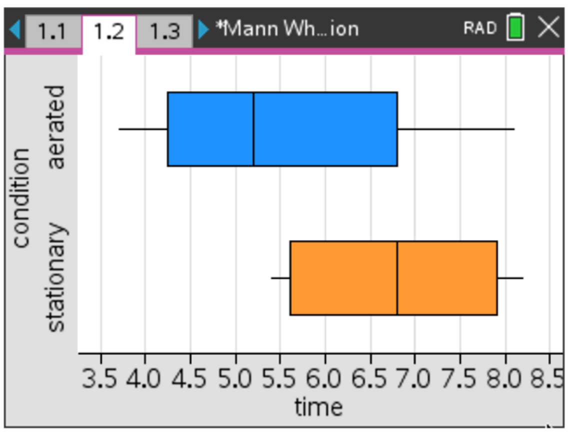
For the Mann-Whitney test we have to **assume** that the population distributions from which the samples have come from have the **same shape and spread**. This allows us to perform a test on the population medians of the two samples.

We can check the plausibility of this assumption by visualisations of the samples:




| aerated | | stationary |
|---|---|---|
| 7 | 3 | |
| 8 | 4 | |
| 5 2 | 5 | 4 6 |
| | 6 | 5 |
| | 7 | 1 9 |
| 1 | 8 | 2 |

key: 6|5 = 6.5 hours




The dotplot, boxplot and stem-and-leaf diagram all show that our assumption about the distributions of the two populations, from which these samples have come, is plausible.

We now proceed with the hypothesis test....

the null hypothesis is:

$H_0$: population median time for stationary = population median time for aerated

the alternative hypothesis is:

$H_1$: population median time for stationary > population median time for aerated

We shall assume $H_0$ to be true.

We shall decide the significance level of our test, $\alpha$ = 5%

We confirm that this is a 1 tailed-test (because $H_1$ has a '>' or a '<' in it, and not '$\neq$')

We note that aerated flasks have the smaller sample size, so $m$ = 5
Therefore the stationary flasks have the larger sample size, so $n$ = 6.

We order all the sample data in ascending order, keeping tracking of which group each piece of data comes from, and then we rank them in order of size:

| hours | 3.7 | 4.8 | 5.2 | 5.4 | 5.5 | 5.6 | 6.5 | 7.1 | 7.9 | 8.1 | 8.2 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| group | A | A | A | S | A | S | S | S | S | A | S |
| rank | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |

the rank sum for aerated, Wa = 1 + 2+ 3 + 5 + 10 = 21
the rank sum for stationary, Ws = 4 + 6 + 7 + 8 + 9 + 11 = 45

We use the rank sum from the **smallest sample size**, which is $m$ = 5, the aerated data.
So we proceed and use Wa = 21, which is known as a 'test statistic'.

Now, Wa being 21 is just one possible value of the whole range of possible values that the rank sum of a group of 5 objects from a group of 11 objects could take.

We need to find out whether 21 is an unusually low or high value, if the null hypothesis, $H_0$, really is true.

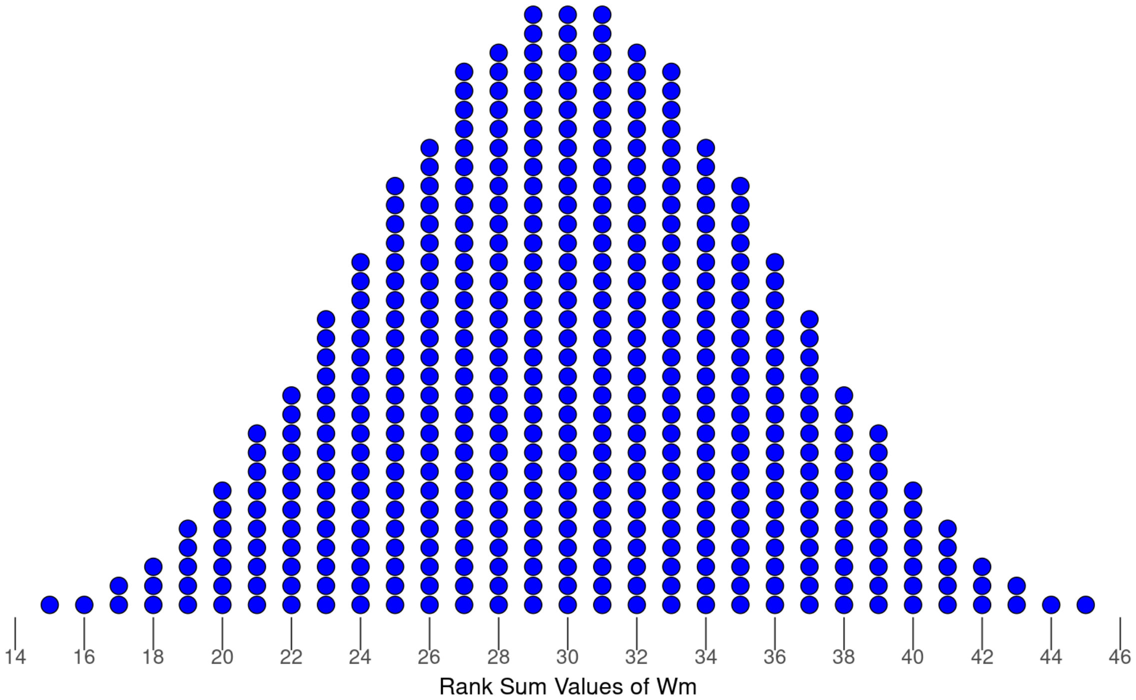
We therefore build up the theoretical distribution of all possible values that Wa could take, where there is a group of 5 numbers and a different group of 6 numbers.

If we had different data values in our samples, then the smallest possible value that Wa could take is 1 + 2 + 3 + 4 + 5 = 15.
The largest possible value that Wa could take is 7 + 8 + 9 + 10 +11 = 45
We need to find all of the other possible values that Wa could take, adding up different combinations of 5 numbers, from 11 numbers.

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 15 | ✓ | ✓ | ✓ | ✓ | ✓ | | | | | | |
| 16 | ✓ | ✓ | ✓ | ✓ | | ✓ | | | | | |
| 17 | ✓ | ✓ | ✓ | ✓ | | | ✓ | | | | |
| 17 | ✓ | ✓ | ✓ | | ✓ | ✓ | | | | | |
| 18 | ✓ | ✓ | ✓ | ✓ | | | | ✓ | | | |
| 18 | ✓ | ✓ | ✓ | | ✓ | | ✓ | | | | |
| 18 | ✓ | ✓ | | ✓ | ✓ | ✓ | | | | | |

etc etc

Visit https://nhost.shinyapps.io/MannWhitneyWilcoxonDistributions/
Select 'Mann-Whitney Distributions'
Select $m$ = 5 and n = 6
Select 'lower tail'




Assuming the null hypothesis, $H_0$, is true, this theoretical distribution of rank sums shows that:

1 way to obtain a rank sum of 15
1 way to obtain a rank sum of 16
2 ways to obtain a rank sum of 17
3 ways to obtain a rank sum of 18
5 ways to obtain a rank sum of 19
7 ways to obtain a rank sum of 20
10 ways to obtain a rank sum of 21

So, this gives a total of 1+1+2+3+5+7+10 = 29 ways to obtain a rank sum of 21, or less.

How many possible combinations of rank sums of 5 values from 11 numbers are there?
*(this is the total number of dots in the above diagram)*

${}^{11}C_5 = 462$

So, assuming that the null hypothesis, $H_0$, is true, the probability that we can expect to obtain a rank sum of 21 or less is

$$P(Wa \leq 21) = \frac{29}{462} \approx 0.062771$$

This probability is greater than 5%, and so our rank sum value of Wa = 21 is <u>not</u> in the most extreme 5% of possible values that it could take. It's not an unexpected number to observe.

We therefore do not reject $H_0$.
We conclude that we do not have evidence to suggest that the population median times for stationary flasks are not equal to the population median times for aerated flasks, and so there does not appear to be a shortening effect on the generation times when aerating.

## Variation 1 - using critical values

Return to https://nhost.shinyapps.io/MannWhitneyWilcoxonDistributions/
Select 'lower tail'
Select a significance level of '5%'

It displays '5% critical value for lower tail = 20'

This means that the most extreme lower 5% of possible values are those whose rank sums are 20, or less.

We measured Wa = 21, which is not in this 5% critical region and we have no evidence to reject $H_0$.

This is a consistent result with the first conclusion.

## Variation 2 - changing significance level

Return to https://nhost.shinyapps.io/MannWhitneyWilcoxonDistributions/
Select a significance level of '10%'

It now displays '10% critical value for lower tail = 22', and the region that is shaded red has now become larger.

This means that the most extreme lower 10% of possible values are those whose rank sums are 22, or less.

We measured Wa = 21, which **is** in the 10% critical region and we now **do** have evidence to reject $H_0$.

This means that if we had chosen a 10% significance level at the very beginning of this analysis, our conclusion would now read as either of the following:

$P(Wa \leq 21) = \frac{29}{462} \approx 0.06 < 0.10$ and so our rank sum value of Wa = 21 is in the most extreme 10% of possible values that it could take.

We therefore reject $H_0$, in favour of $H_1$.
We conclude that we have evidence to suggest that the population median times for stationary flasks **are greater than** the population median times for aerated flasks, and so there **does** appear to be a shortening effect on the generation times when aerating.

or

As Wa = 21 $\leq$ 22, the 10% critical value, we reject $H_0$, in favour of $H_1$.
We conclude that we have evidence to suggest that the population median times for stationary flasks **are greater than** the population median times for aerated flasks, and so there **does** appear to be a shortening effect on the generation times when aerating.

## Variation 3 - exploring other situations

Return to https://nhost.shinyapps.io/MannWhitneyWilcoxonDistributions/

Change the controllers for **m**, **n**, the **Alternative Hypothesis** and the **Significance Level** to see the effects that each of these can have on the critical values.

Your teacher will show you how these are recorded in the AH Statistical Formulae and Tables Booklet.