Introduction into Biostatistics

Anna Poetsch, Biotechnology Center, TU Dresden

Hypotheses in the statistical sense

Innocent until proven guilty!

	0-Hypothesis:
Defendant is innocent	**There is no difference between the distributions**
Decision based on factual evidence.	Decision based only on data from one experiment.
Guilty when the evidence is inconsistent, otherwise not guilty.	Statistically significant when the P value is small enough, otherwise not statistically significant
Verdict of not guilty when the evidence is consistent with the presumption of innocence	Does not have to be convinced that the null hypothesis is true

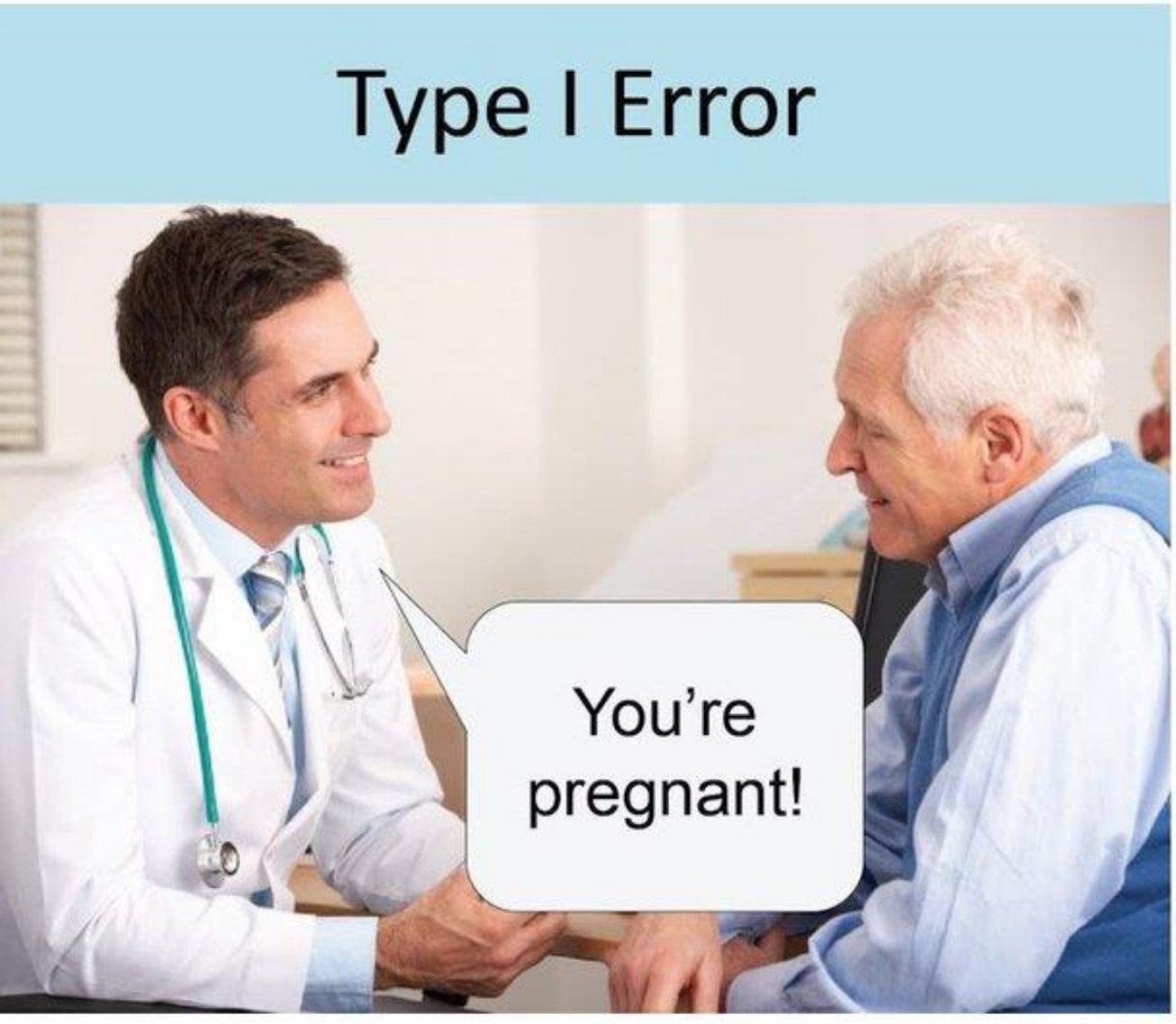
How to reject H_0

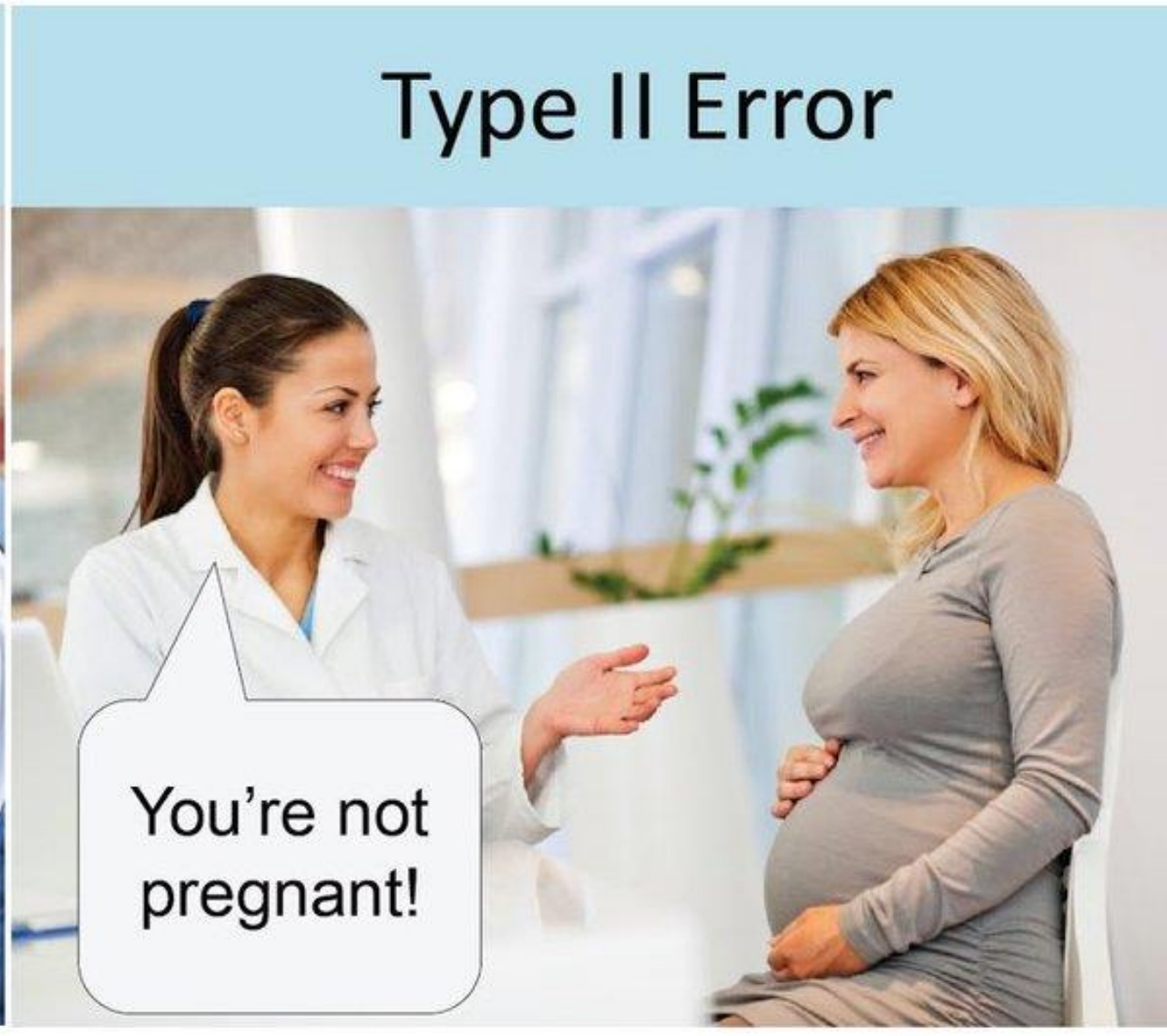
How much probability do you allow yourself to be wrong?

- last line chemotherapy treatment: Every bit of hope counts

- vaccination side-effects: Even rare events can be too much

What can go wrong?





False positive

False negative

P-values

The probability that you reject H_0 by chance

Other ways to phrase it:

The probability that two samples **are declared different** although they belong to the **same population**.

The probability of observing **a difference** as large as you see it (or larger), if the samples are indeed from the **same population**.

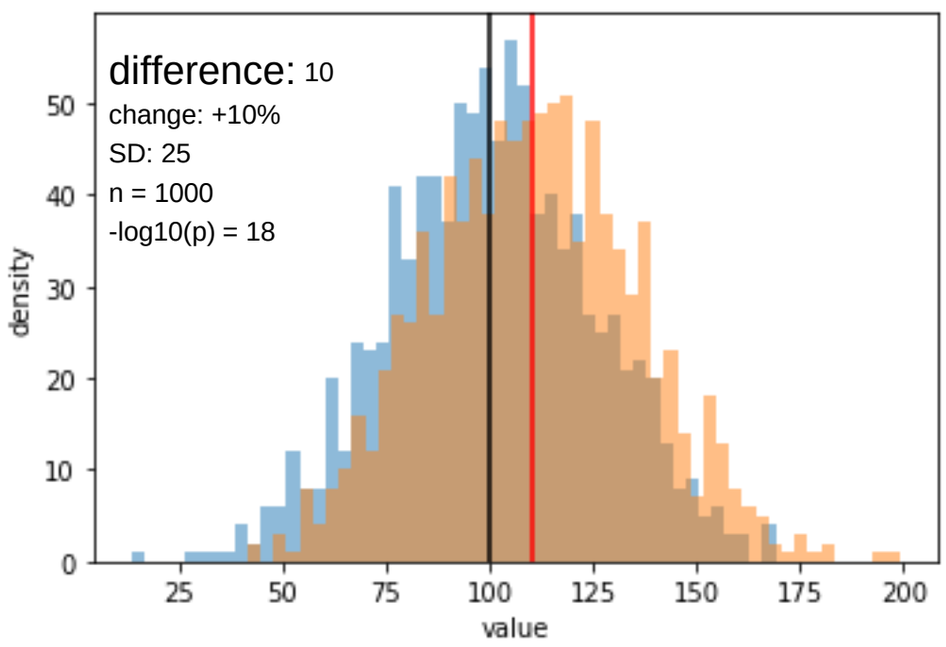
Misconceptions about P-values

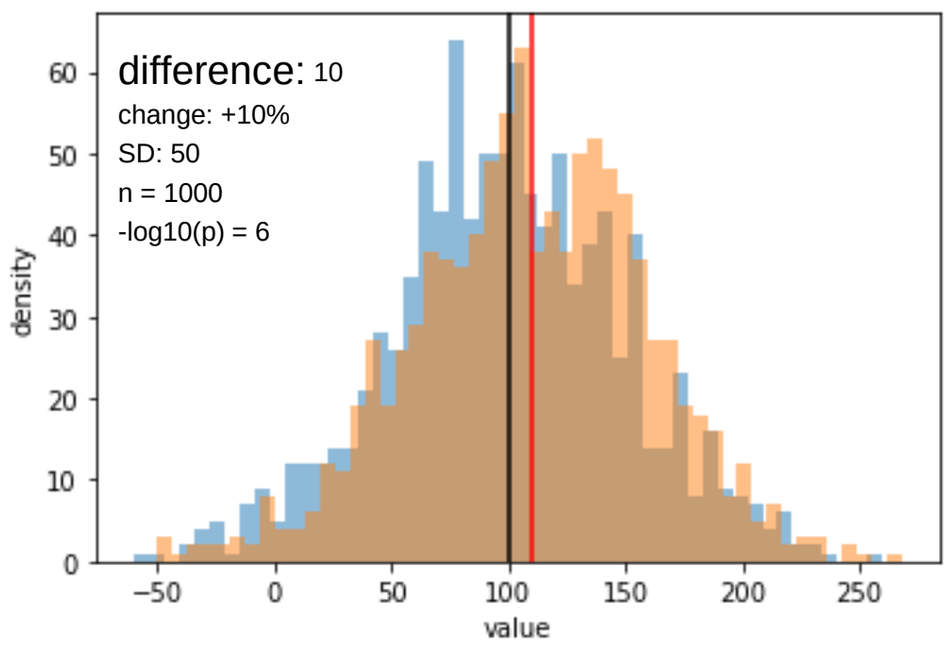
The logic is **not reversible**: A p-value of 0.05 means a 5% chance concluding on a difference by chance. Don't try to interpret the 95%!

You **cannot determine** whether H_0 **is true**.

Reproducibility of p-values is inherently very poor.

A p-value is **not appropriate** to conclude about the **magnitude of a difference**





How to put a threshold alpha for P-values

1. Not at all
2. At the next “pleasant number” (0.05, 0.001….)
3. At a threshold that is custom in the field (0.05, 5 sigma,…)

Whatever seems appropriate for your specific setup*

*also consider multiple testing correction

How to perform the actual hypothesis testing

Do we know the distribution?

-> Parametric testing

-> Fit a distribution to our data and compare whether the two groups are sufficiently different

Do we not know the distribution?

-> Non-parametric testing

-> Determine the ranks of our datapoint and look whether the ranks are sufficiently unbalanced

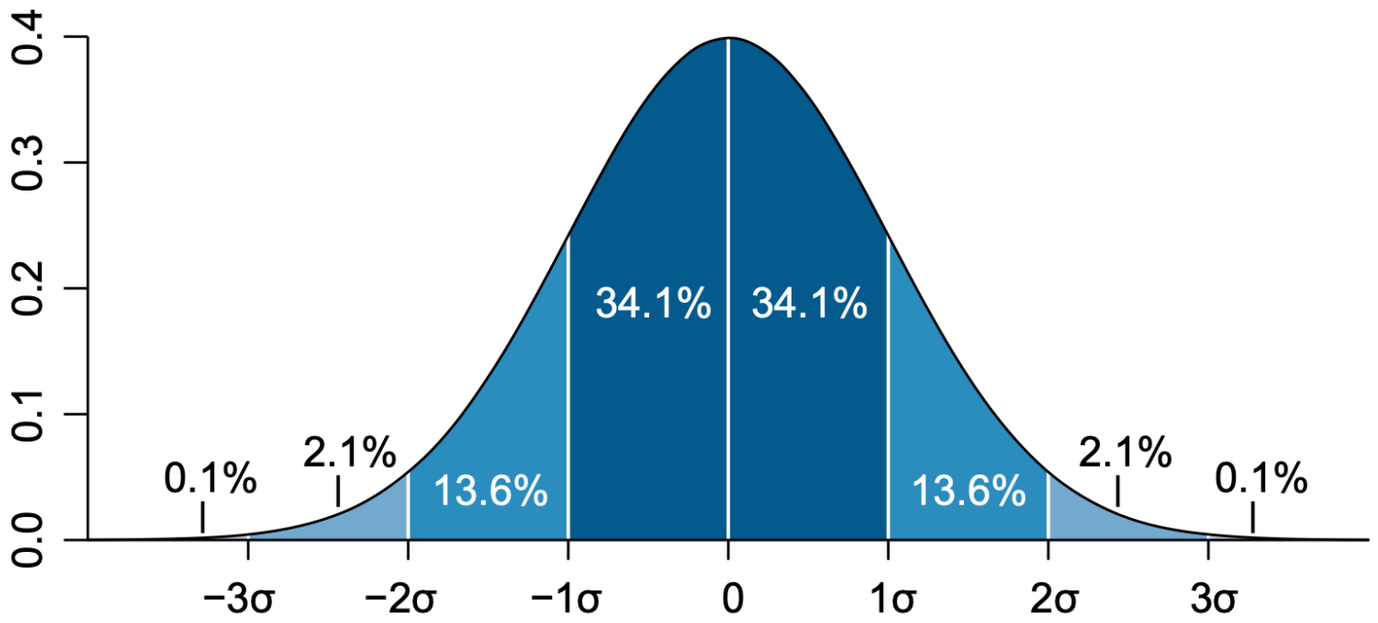
Comparing Two Means

Or: The t-test

Assumptions:

- Our data follow a distribution that can be approximated by the mean
- Equal standard deviation between samples
- They are representative samples
- Independent observations
- Accurate data

The Standard Error of the Mean (SEM)



SEM = SD/square_root(n)

The t-test calculates the **standard error of the difference between two means**

From this, the t-ratio is generated, the **difference of the means divided by the standard error of that difference**

The p-value is computed from this t-ratio and total sample size.

What does the p-value from the t-test tell us?

The probability that we are wrong, if we consider the two distributions to be different.

What is different in a paired test?

The difference of each pair is used to compute the standard error and thus the p-value.

When is a t-test inappropriate?

When any of the assumptions is violated, especially the assumption about that the mean needs to be a good approximation of the distribution.

Why?

When using t-tests inappropriately, outliers become very powerful and misleading!!!

What are the alternatives?

- If data are supposed to meet the criteria theoretically, find the source of your issues
- Assume a different distribution
- Change to non-parametric testing

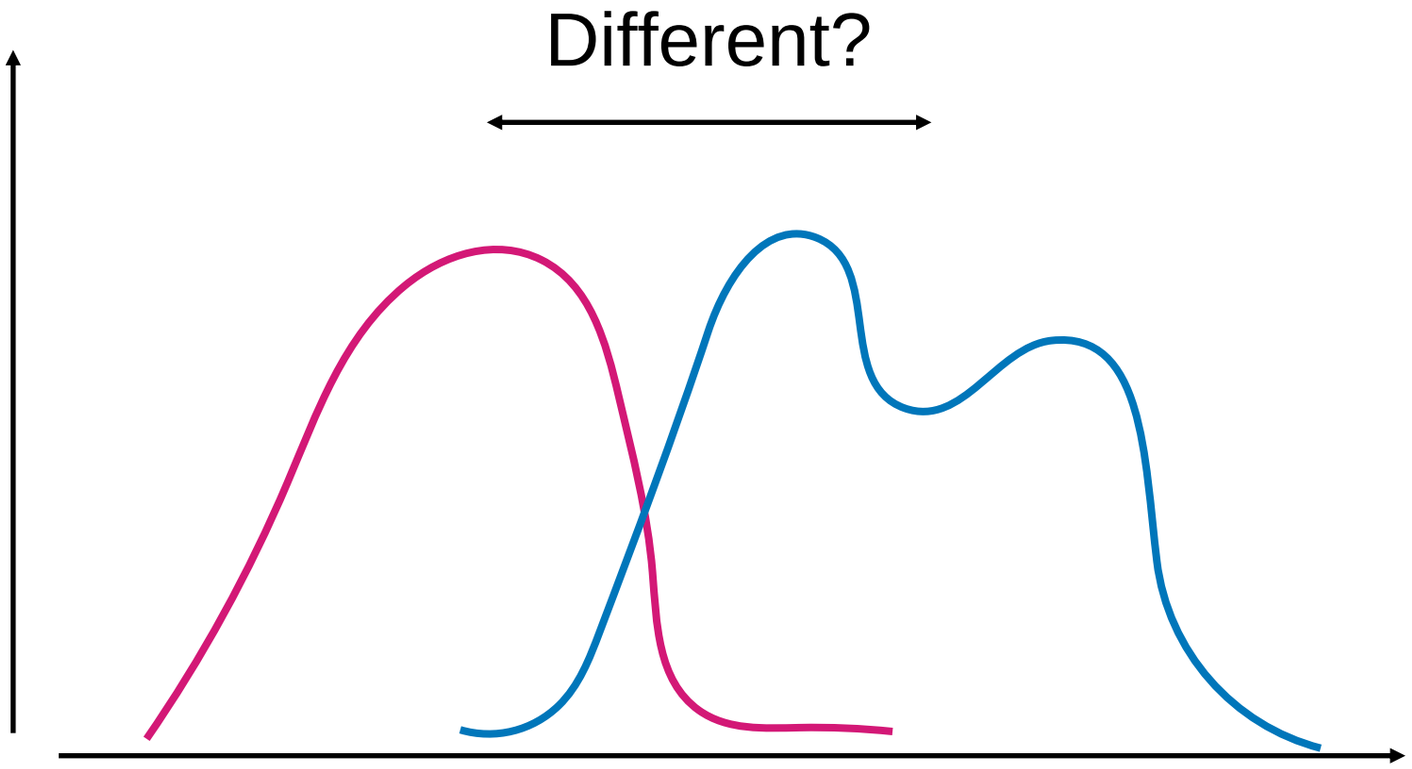
Organization

- 16.5. Introduction to biostatistics
- 14.6 Hypothesis testing
- **21.6. Multiple comparisons and correlations** by Melissa Sanabria
- 28.6. Big data, clustering, dimensionality reduction

Non-parametric testing

- Non-parametric refers to testing based on ranks not on a known distribution.
- Non-parametric can also mean to determine a distribution through resampling (bootstrapping)
- Parametric tests assume a specific distribution (normal, Poisson,…)

Non-parametric testing



Choosing between tests

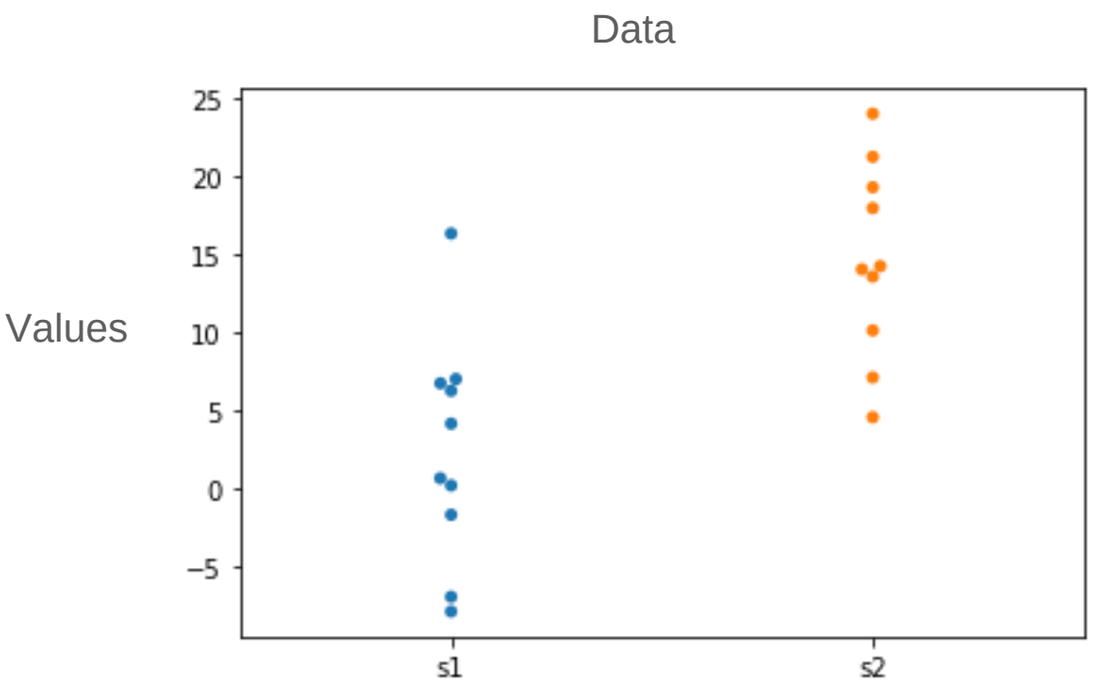
- Whenever you know your distributions and none of the assumptions are violated, go with parametric tests
- With lower numbers you will always have more power with a parametric test
- Bootstrapping is a good alternative to rank based non-parametric tests, but it can get computationally very intense and they are not really custom in molecular biology (yet)

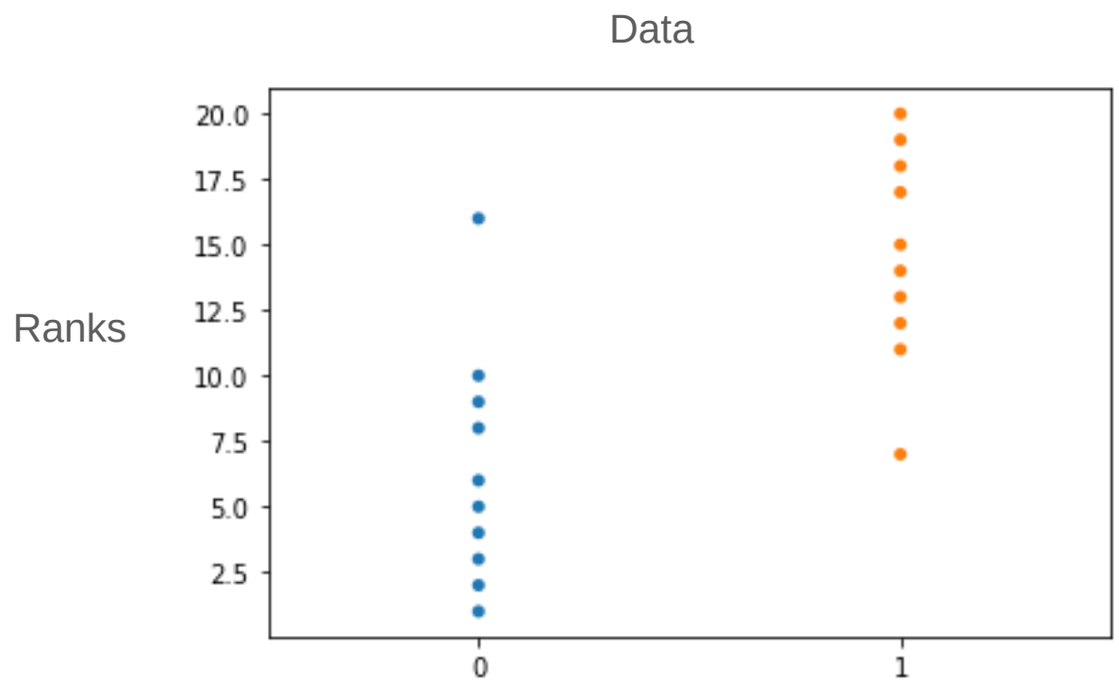
The test’s names

- Comparing two unpaired groups: Mann-Whitney test
- Comparing two paired groups: Wilcoxon matched-pairs signed-rank test
- Comparing multiple samples (i.e. the non-parametric version of ANOVA): Kruskal-Wallis test
- There tends to be a bit of confusion on how to call them….

The names are frequently interchanged, e.g. Mann-Whitney is frequently called “unpaired Wilcoxon”!

How a rank based test works

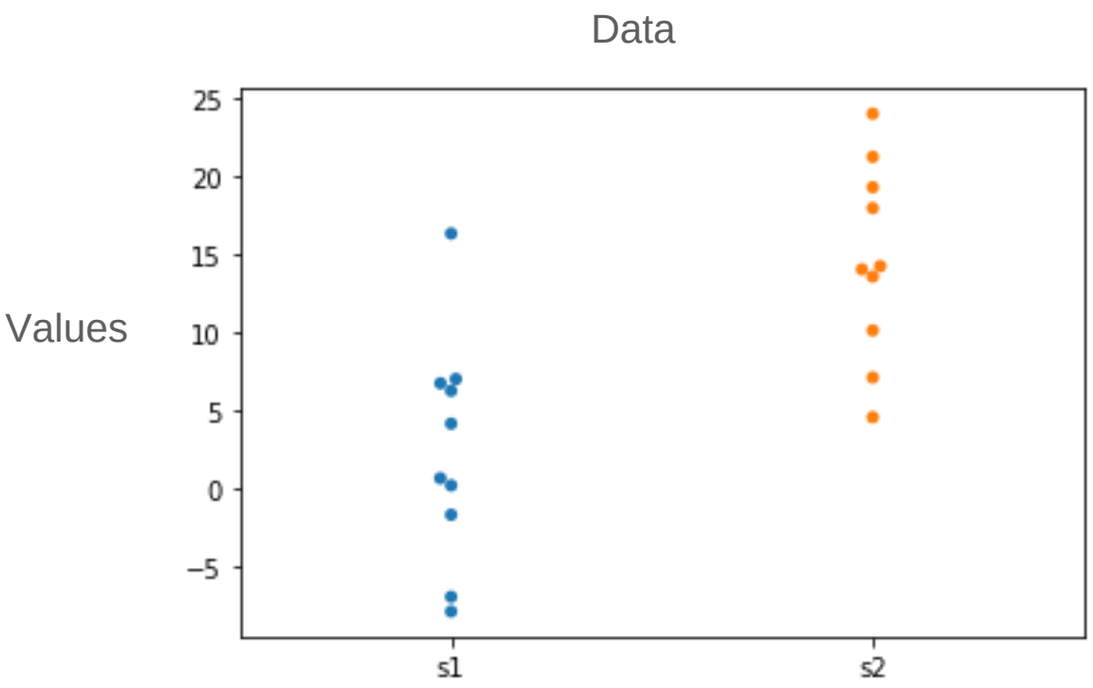


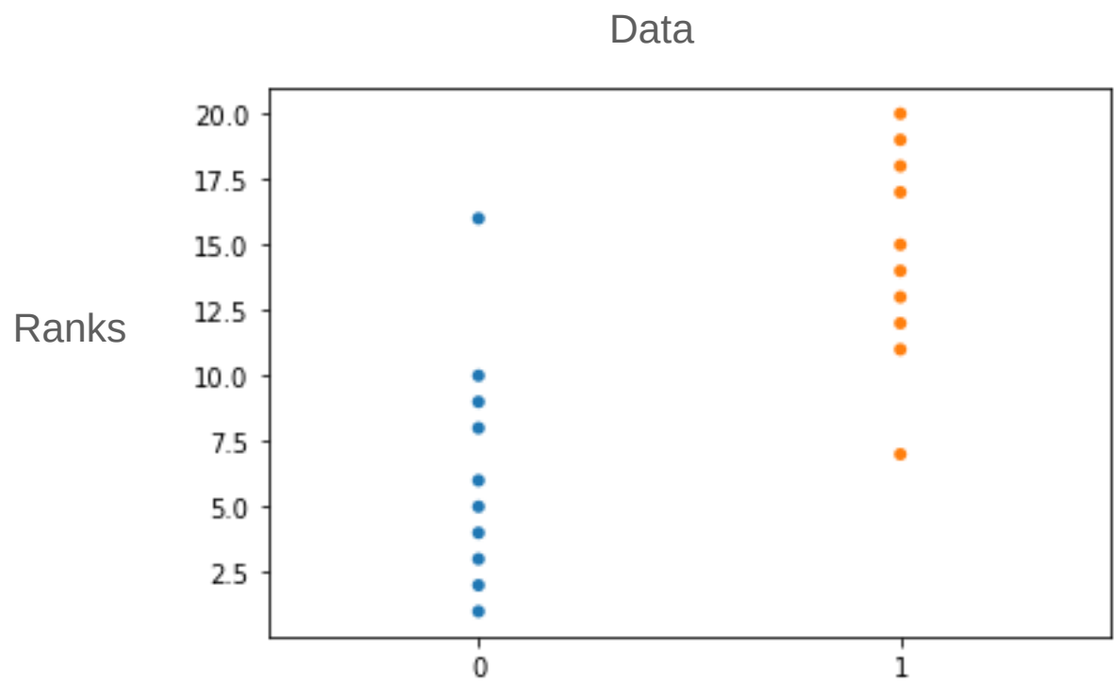


The absolute information is lost and only the ranks are compared.

The p-value describes the probability that the test considers the ranks non-random although they are.

Advantages of a rank based test

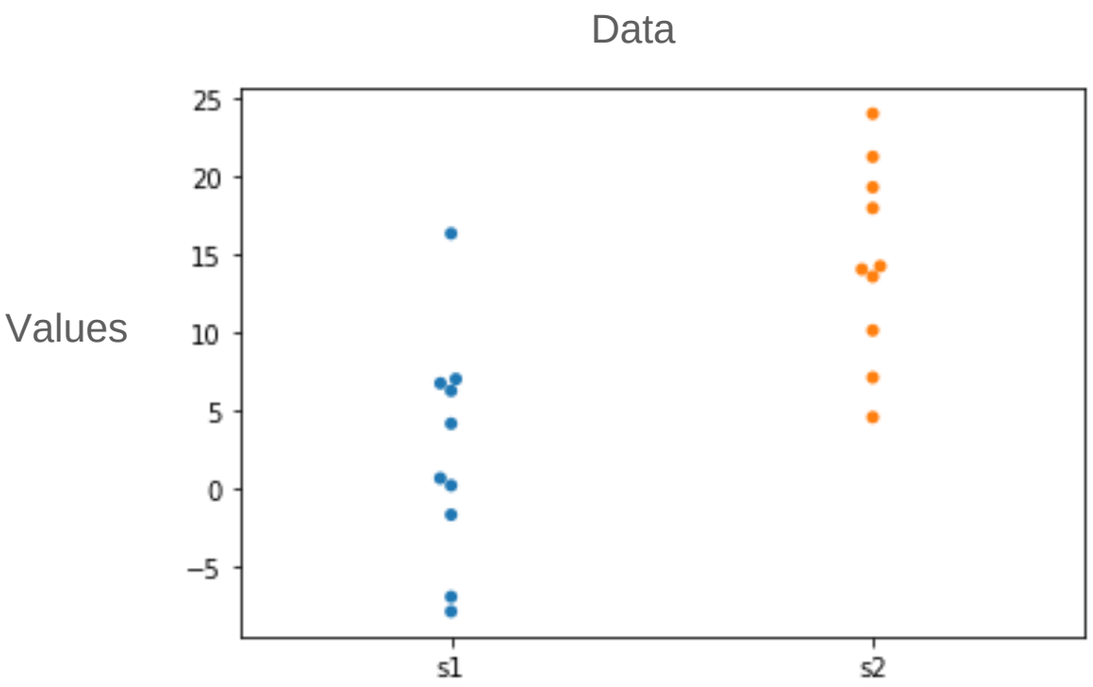


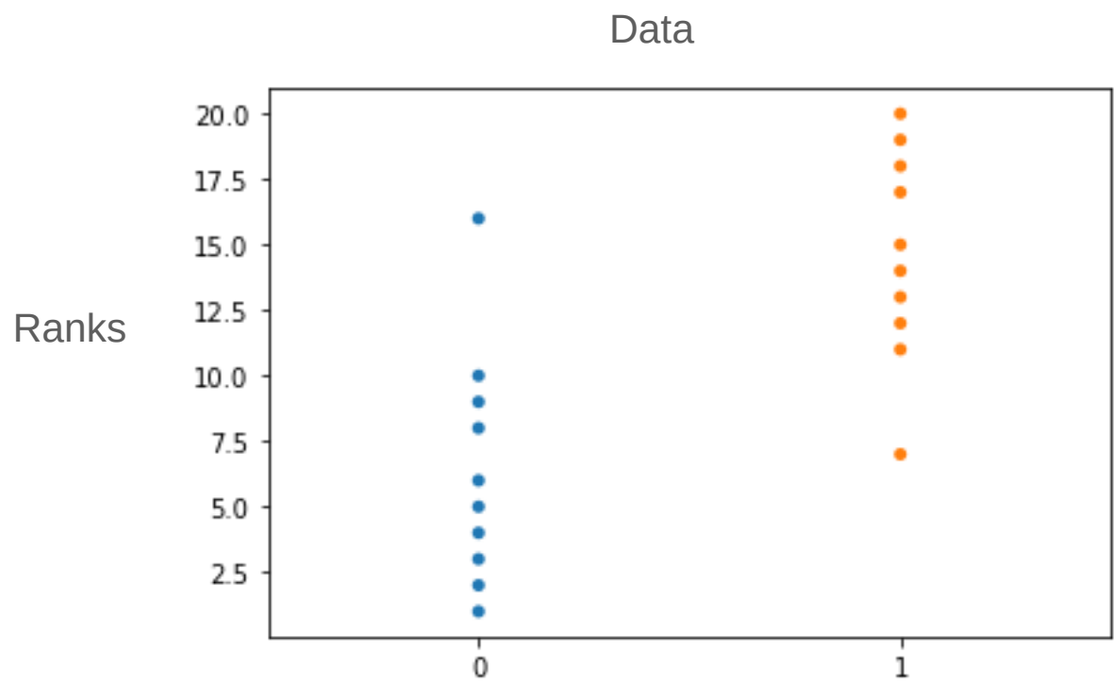


Outliers have limited influence on the outcome

We are not dependent on a distribution

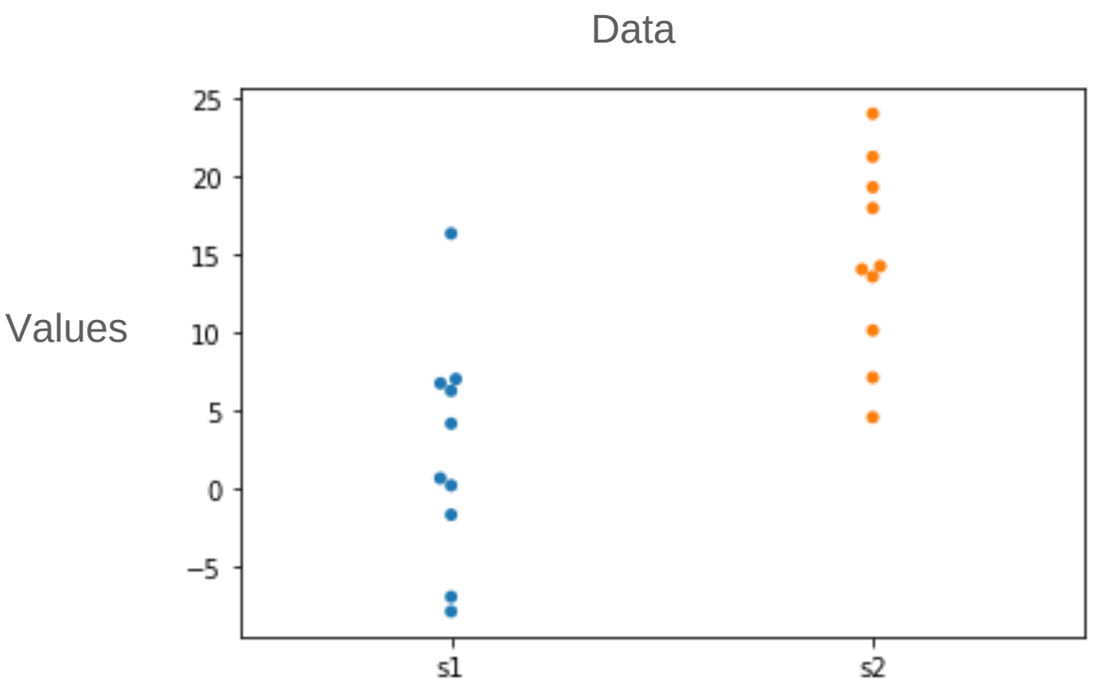
Disadvantages of a rank based test

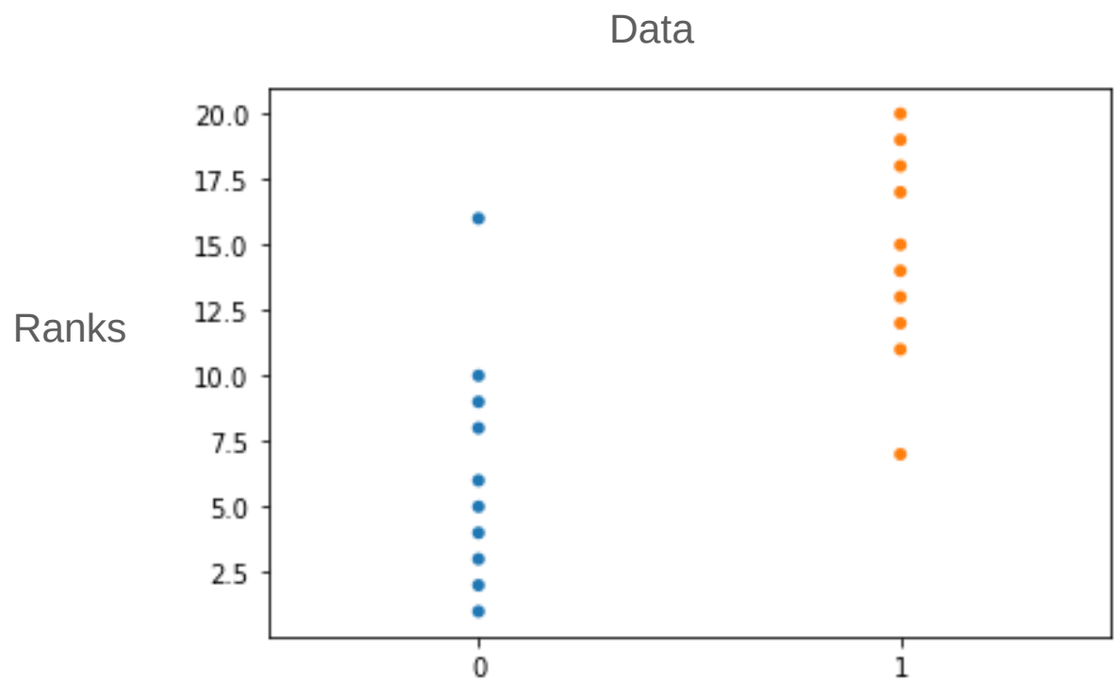




- We are loosing power
- Confidence intervals are more tricky
- Limited with more complex use-cases (regression models)

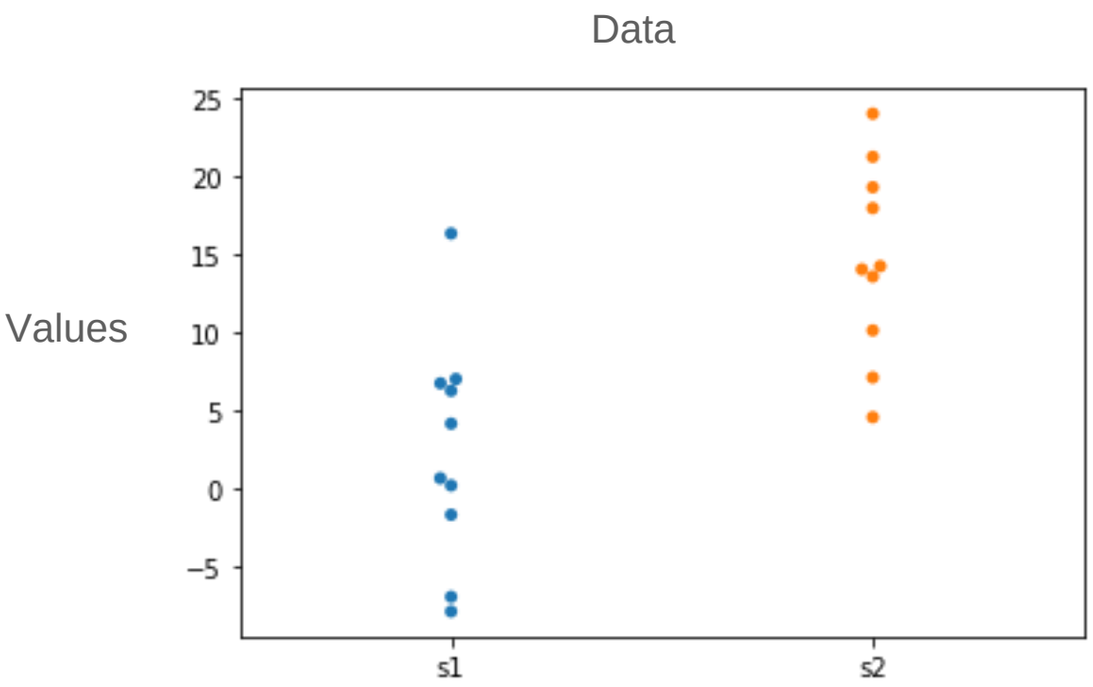
Assumptions

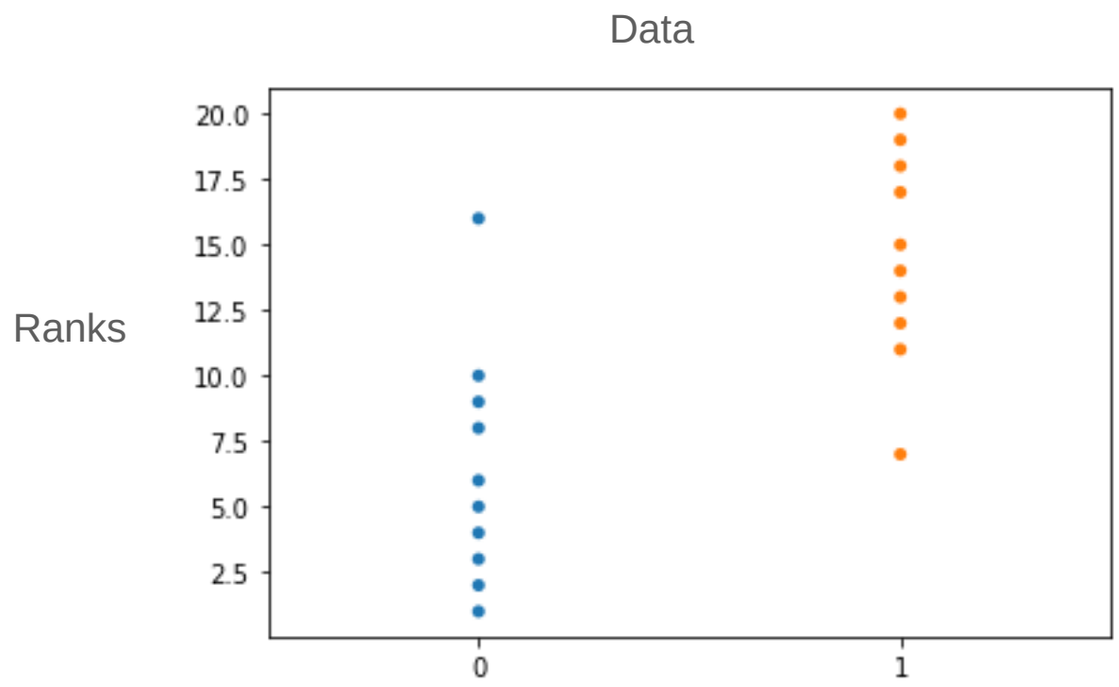




- Random sampling
- Each value is obtained independently

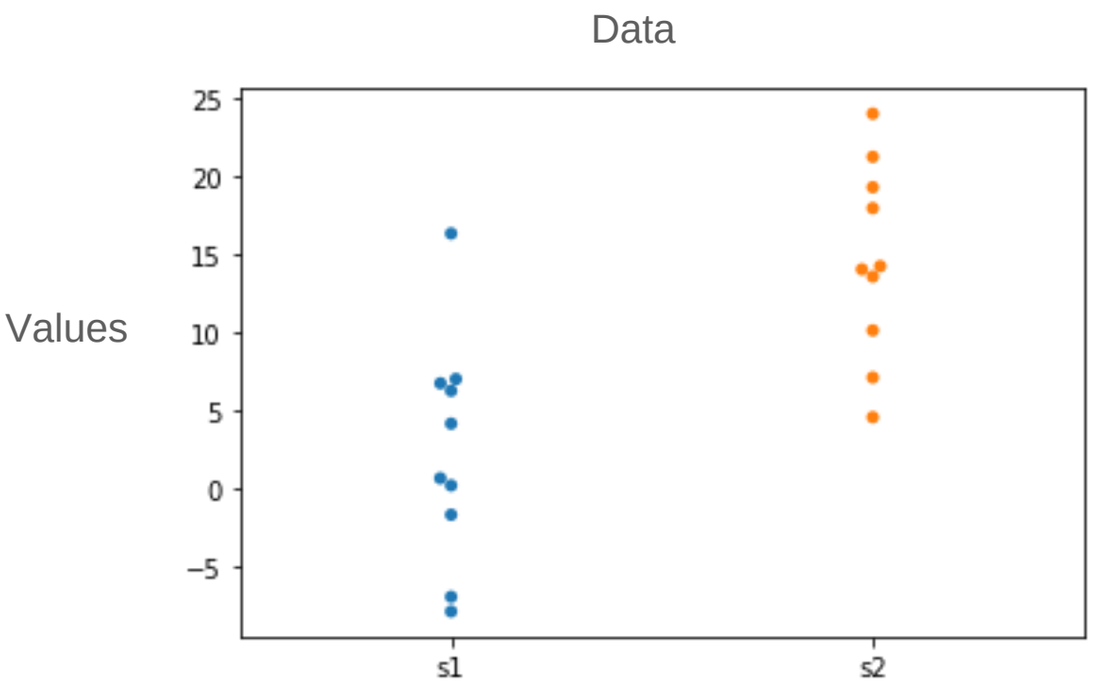
Sample sizes

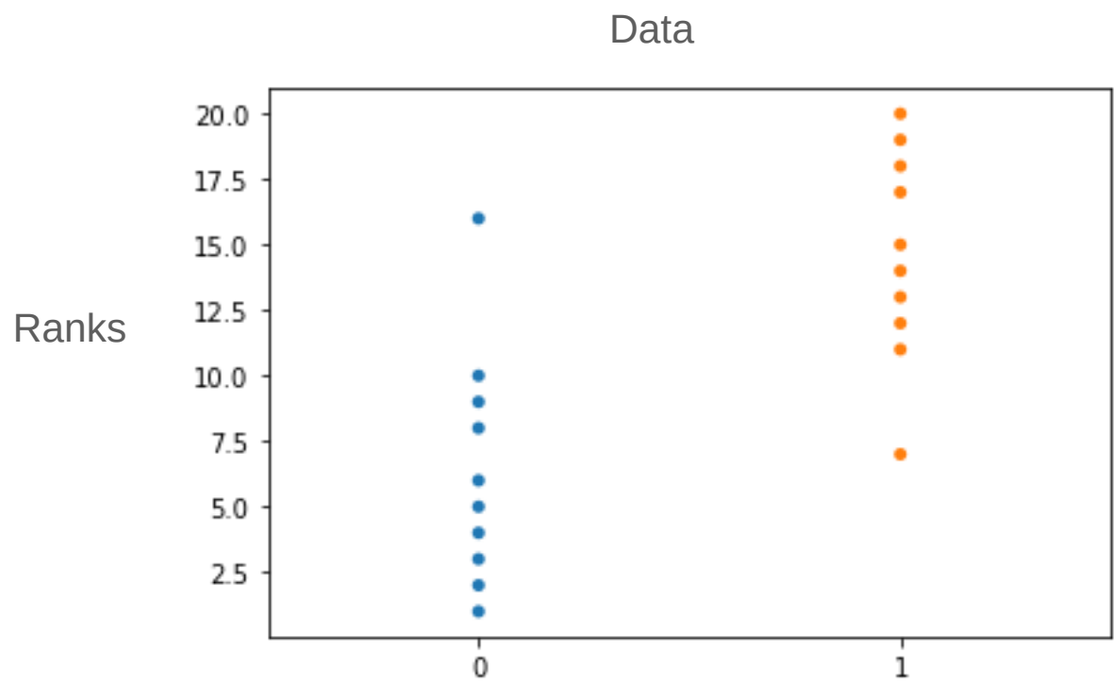




Do you think we need more or fewer samples for a non-parametric test?

Sample sizes





Do you think we need more or fewer samples for a non-parametric test?

It depends… but as a rule of thumb one can estimate the same as a parametric test + 15%

Dos and Don’ts

Do think about your assumptions of your test before you do it.

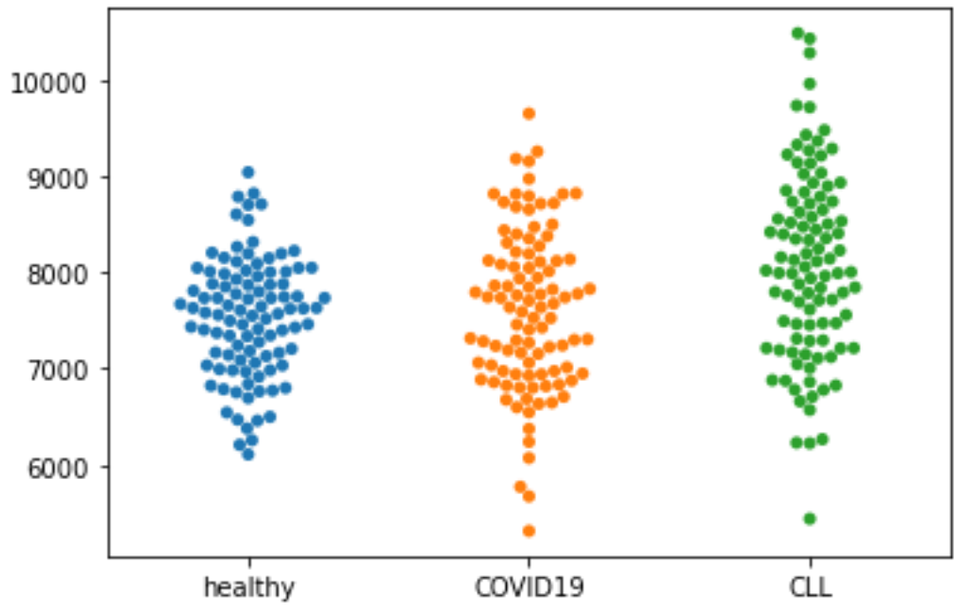
Don’t do both and pick the best!!!!!!!!!!!!

Multiple testing corrections

Why do we need it?

The more comparisons you do, the more likely you are to hit your significance level by chance.

What do you do, if you want to do multiple comparisons?



Do the assumptions for “comparison of means” (t-test) apply?

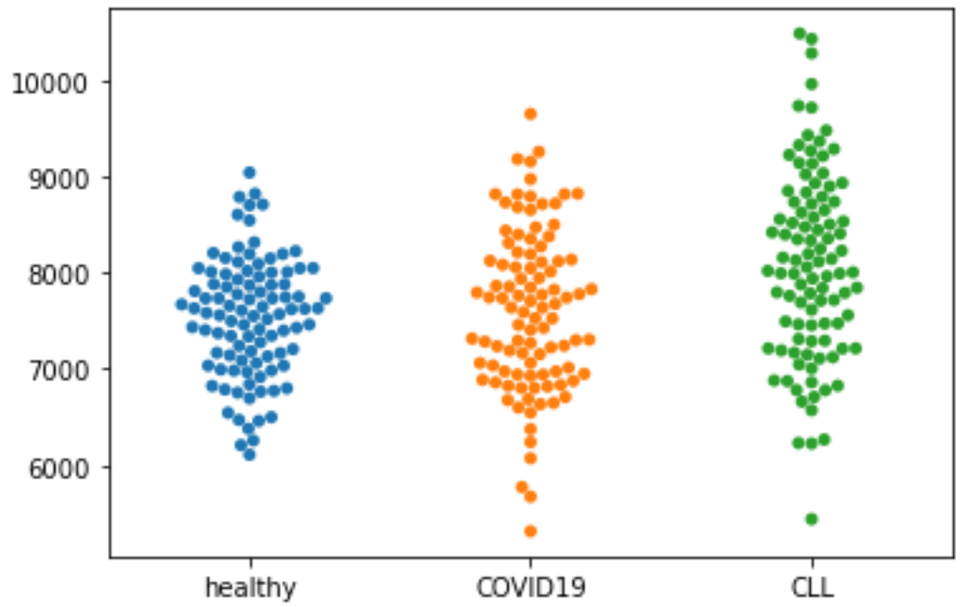
-> Analysis of Variance, one-way ANOVA (= multi-sample-t-test)

-> repeated-samples ANOVA (= multi-sample-paired-t-test)

0-Hypothesis: The mean is identical in all three samples

-> one p-value as output!

But we want to know which one is different!



To extract the p-values for multiple comparisons with corrections, we can take Tukey's Multiple comparisons test, which takes the differences of the means for each comparing pair and corrects for the number of comparisons.

Is Tukey always the best choice?

- No, it is the best choice after an ANOVA, because it takes the other comparisons into account, which makes it very powerful

- Alternatives for any other situation are:

- **Bonferroni**, which is used a lot in genetics, i.e. divide the p-value by the numbers of comparisons

- **Benjamini-Hochberg**: Controlling the false-discovery rate (FDR)