Introduction into Biostatistics

Anna Poetsch, Biotechnology Center, TU Dresden

Organisation

- 1.6. Introduction into biostatistics
- 8.6. Descriptive statistics
- 22.6. Hypothesis testing
- **29.6. Non-parametric testing, multiple testing, correlation statistics**

Hypotheses in the statistical sense

Innocent until proven guilty!

-> at first sight counterintuitive…

0-Hypothesis:

“The Astra Zeneca vaccine does not protect from COVID-19 in > 65 yo”

Test: Can we reject it?

A few months ago: No

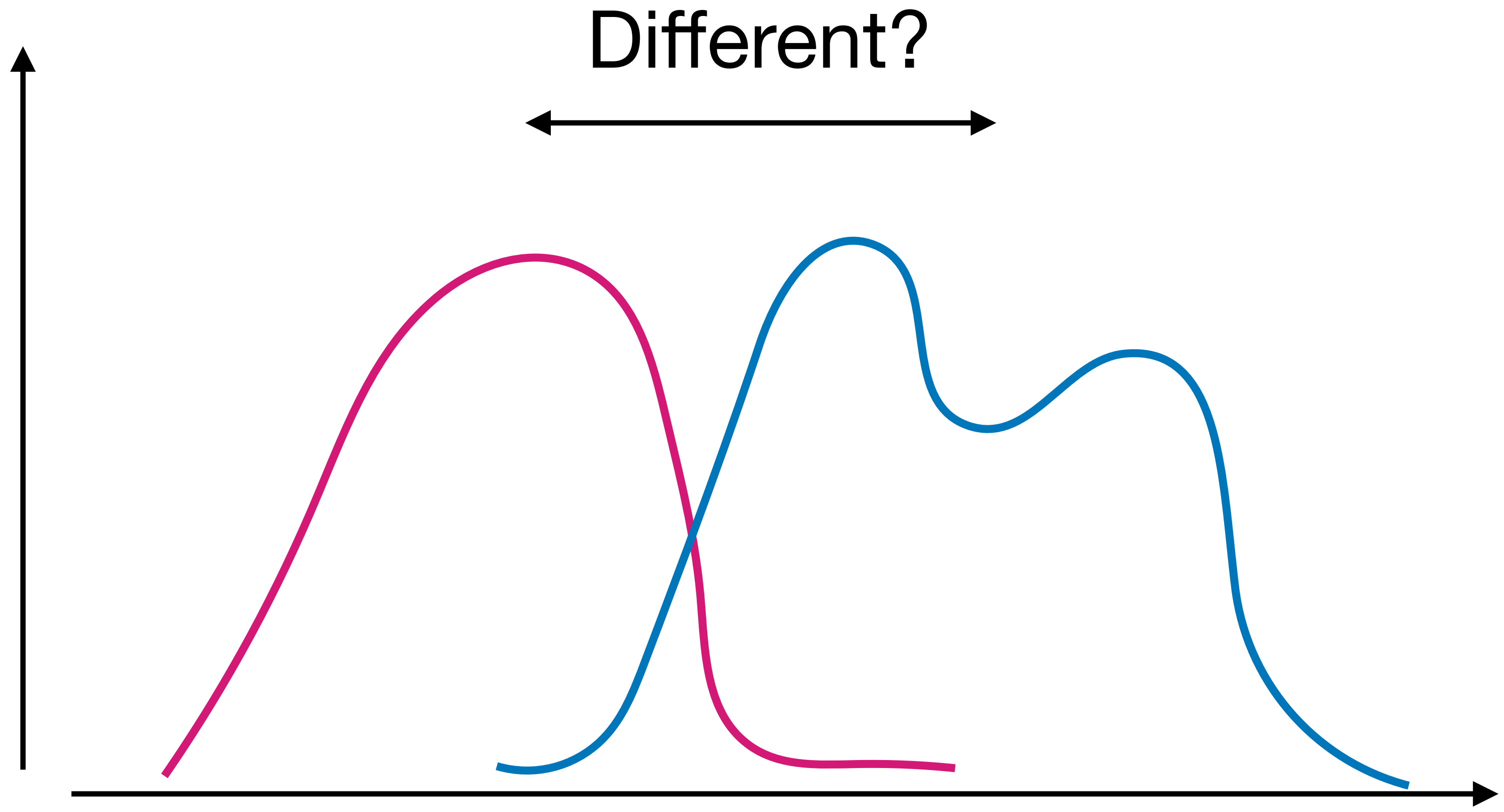
Does it mean that it is not protective? No - we just don’t know!

A few months later, H_0 can be rejected

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

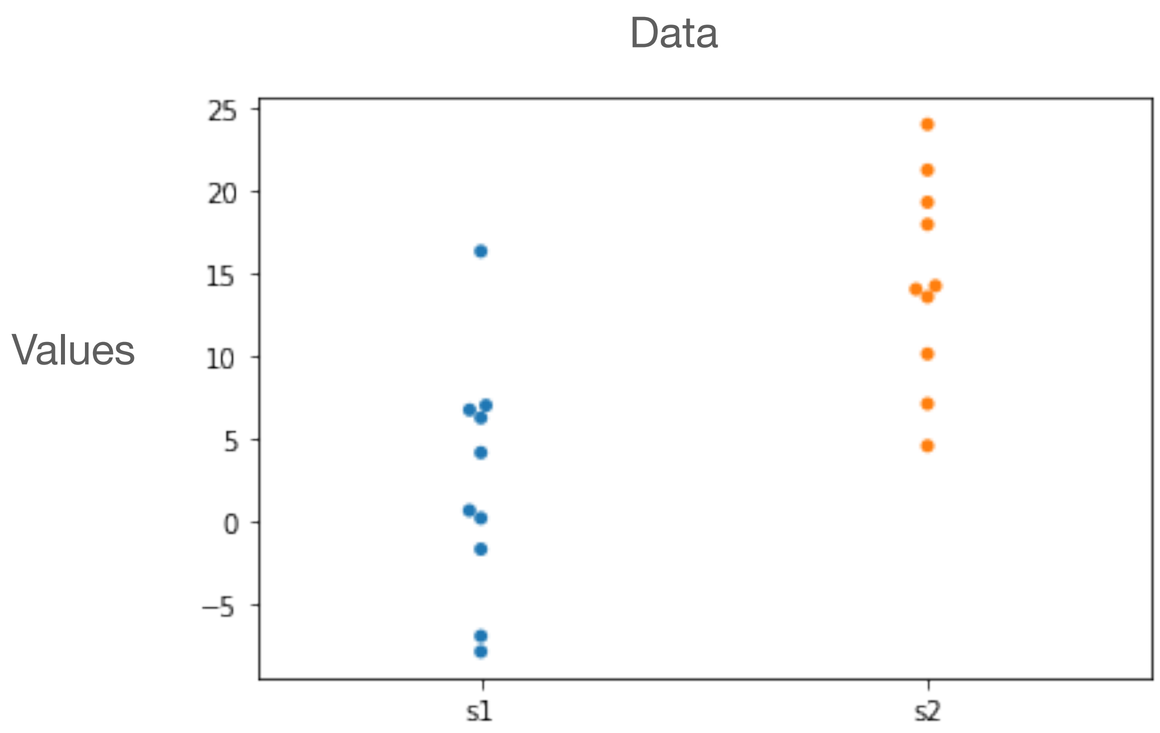
- Whenever you know your distributions and none of the assumptions are violated, go with parametric tests
- Outliers are the most important issue in this regard!
- With lower numbers you will always have more power with a parametric test
- Bootstrapping is a good alternative to rank based non-parametric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

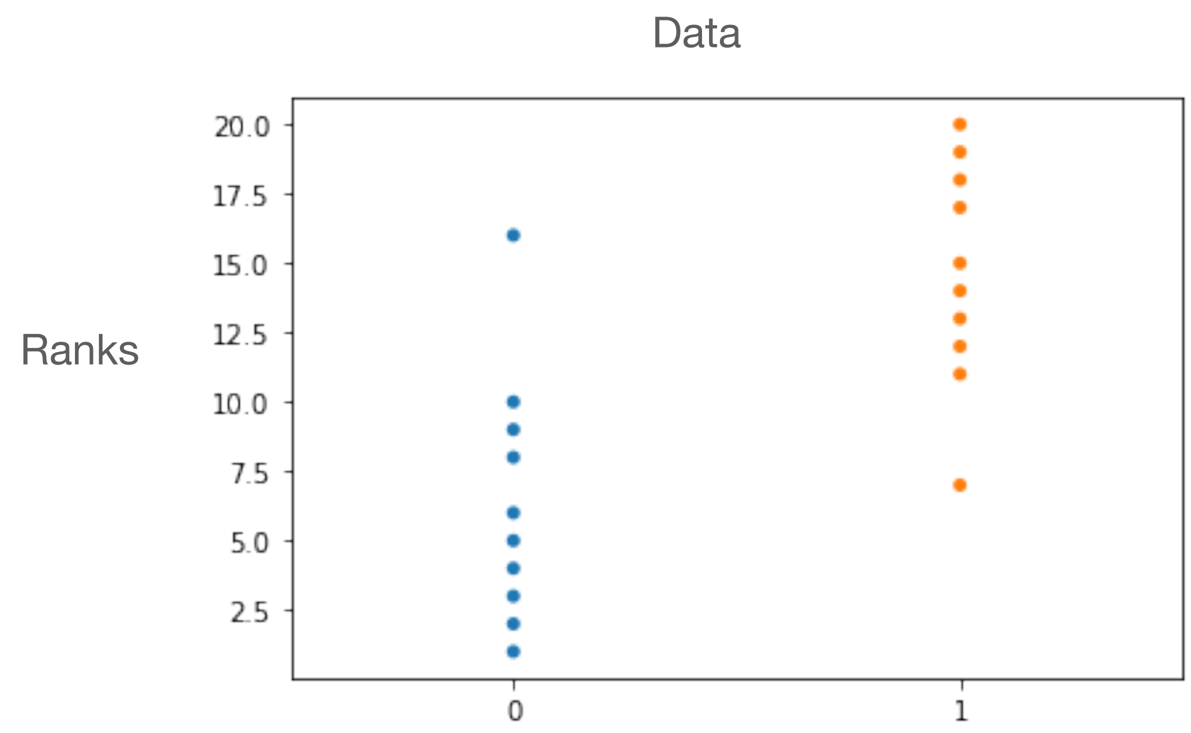
The test's names

- There tends to be a bit of confusion on how to call them….
- 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

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

How a rank based test works

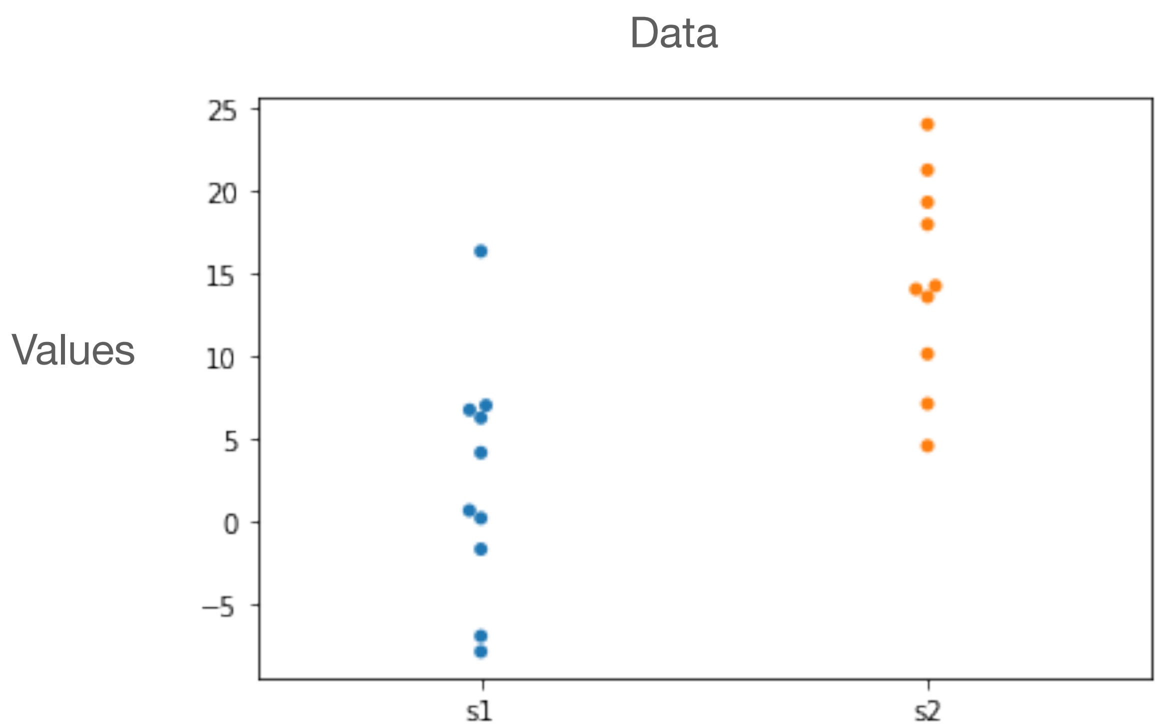


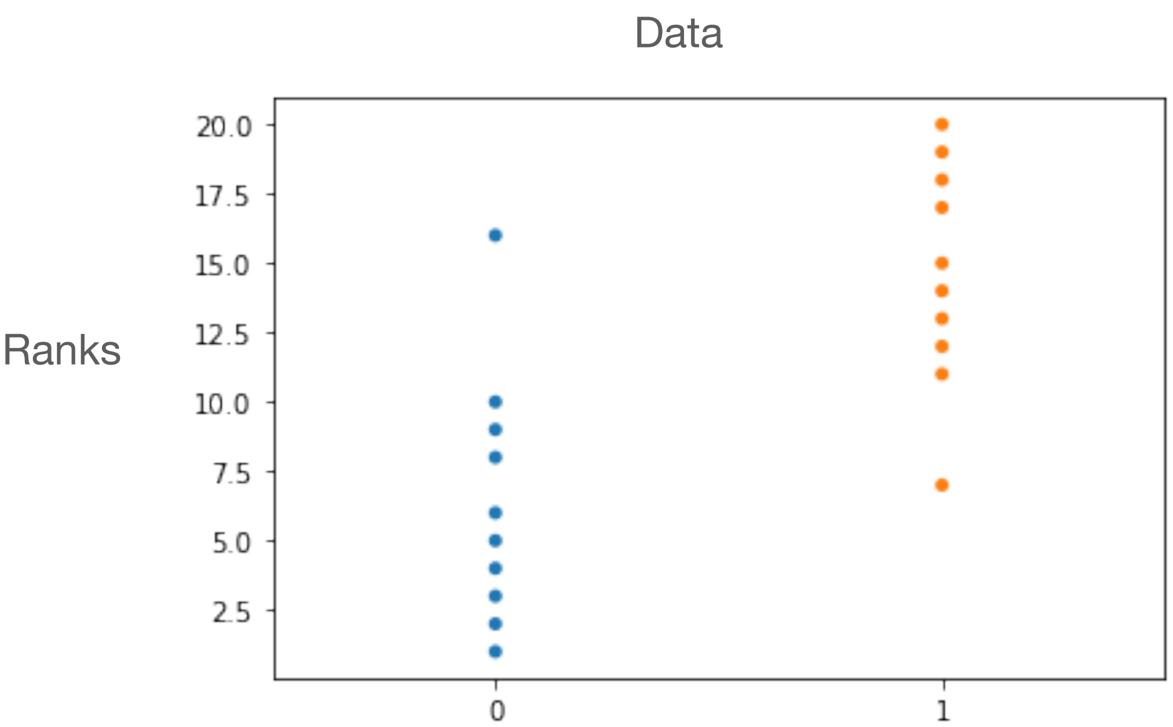


The absolute information is lost and only the ranks are compared.

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

Advantages of a rank based test

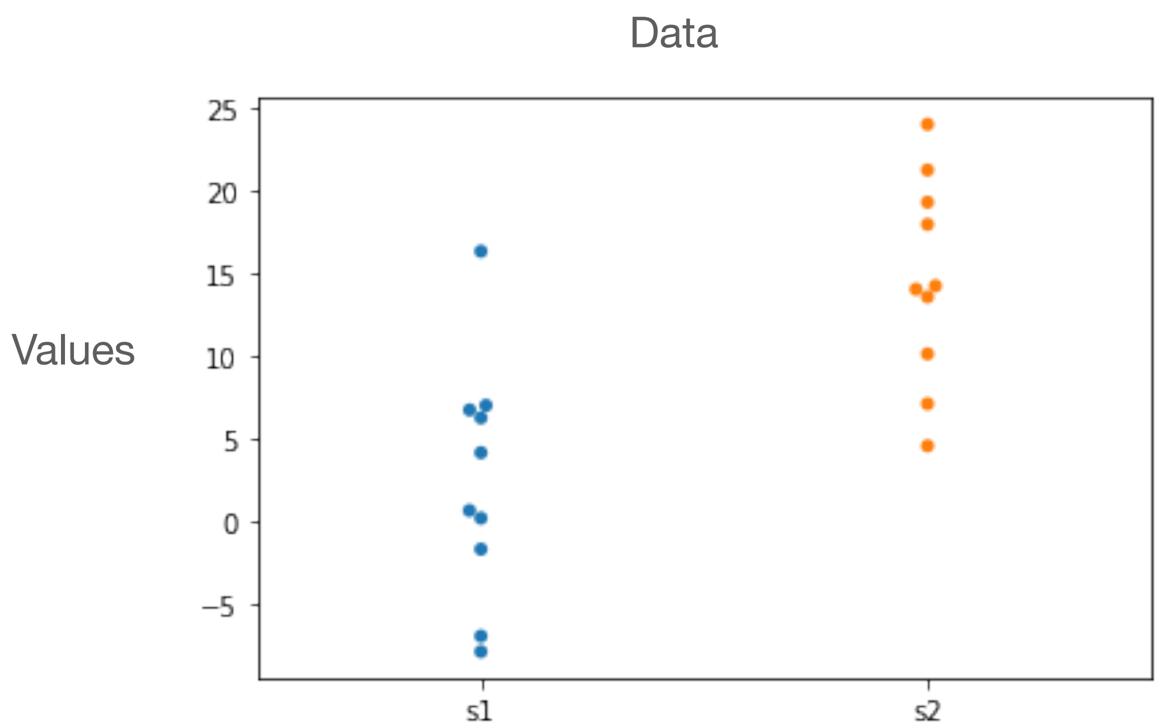


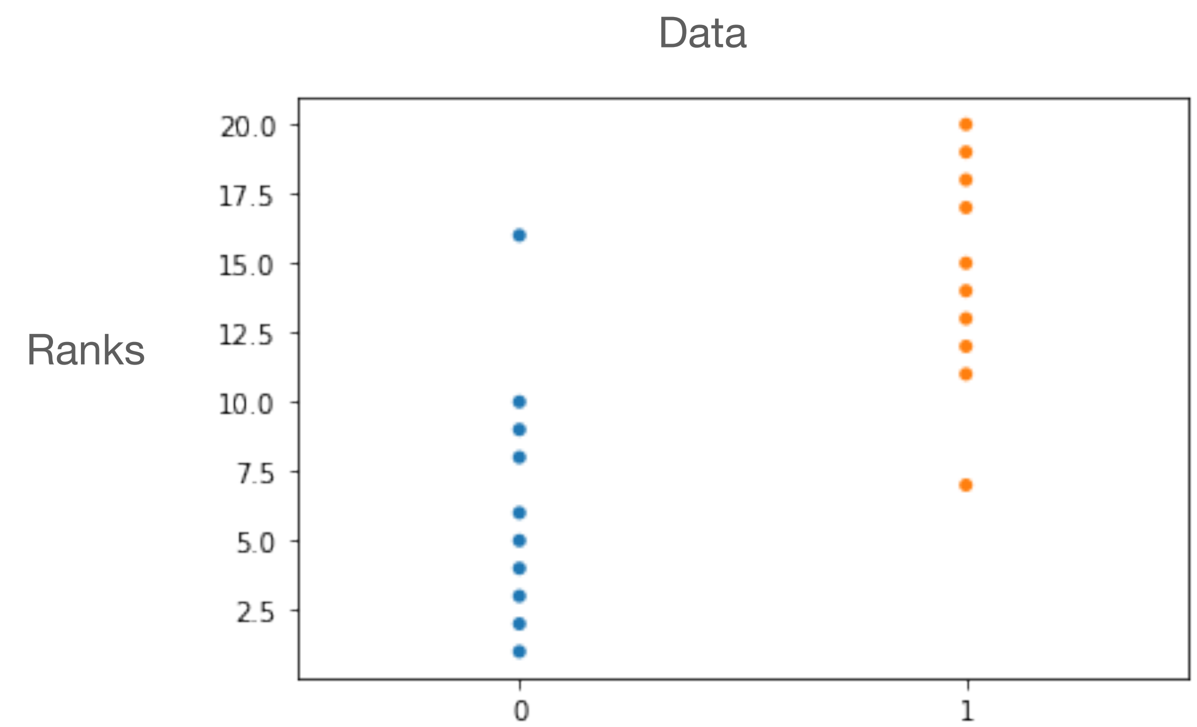


Outliers have limited influence on the outcome

We are not dependent on a distribution

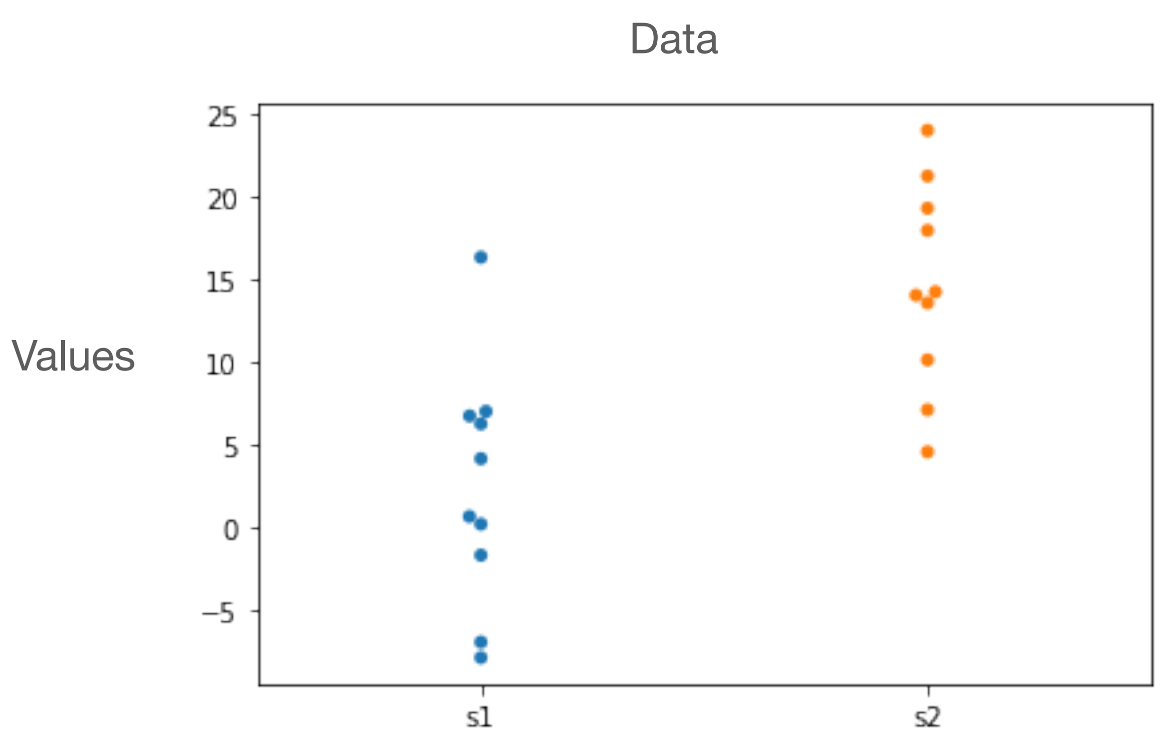
Disadvantages of a rank based test

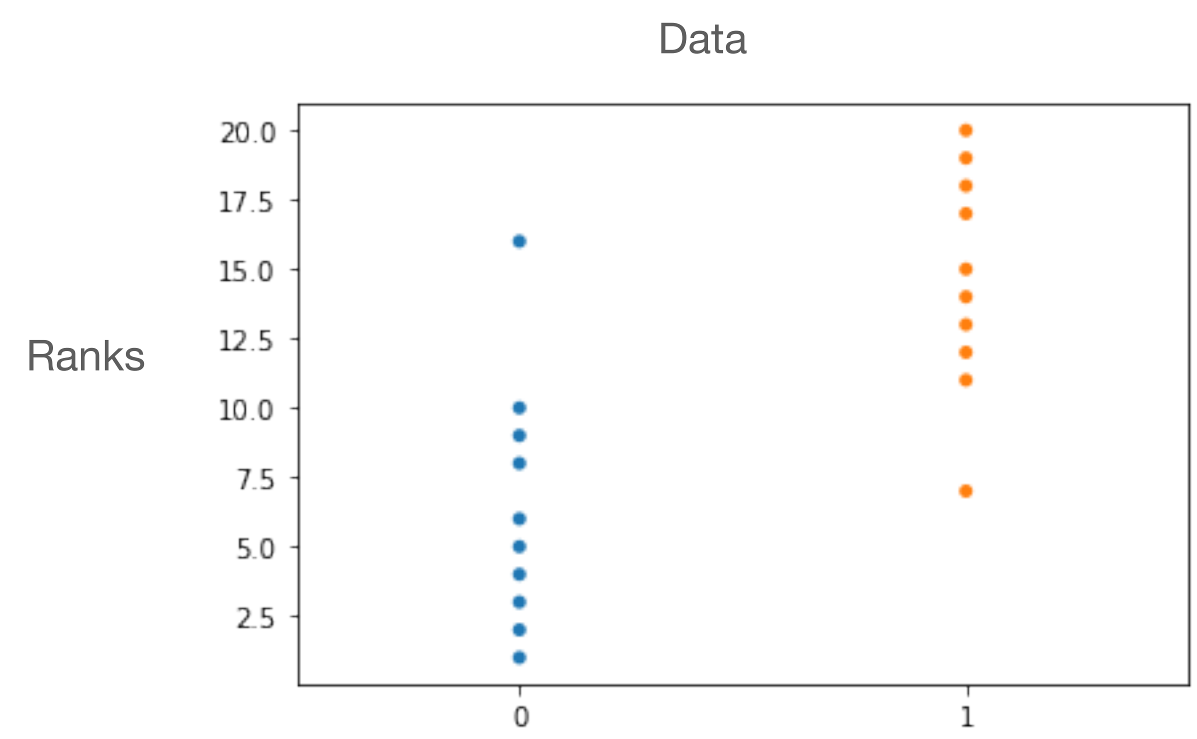




- We are loosing power
- Confidence intervals are more tricky
- Limited with more complex use-cases (regression models)

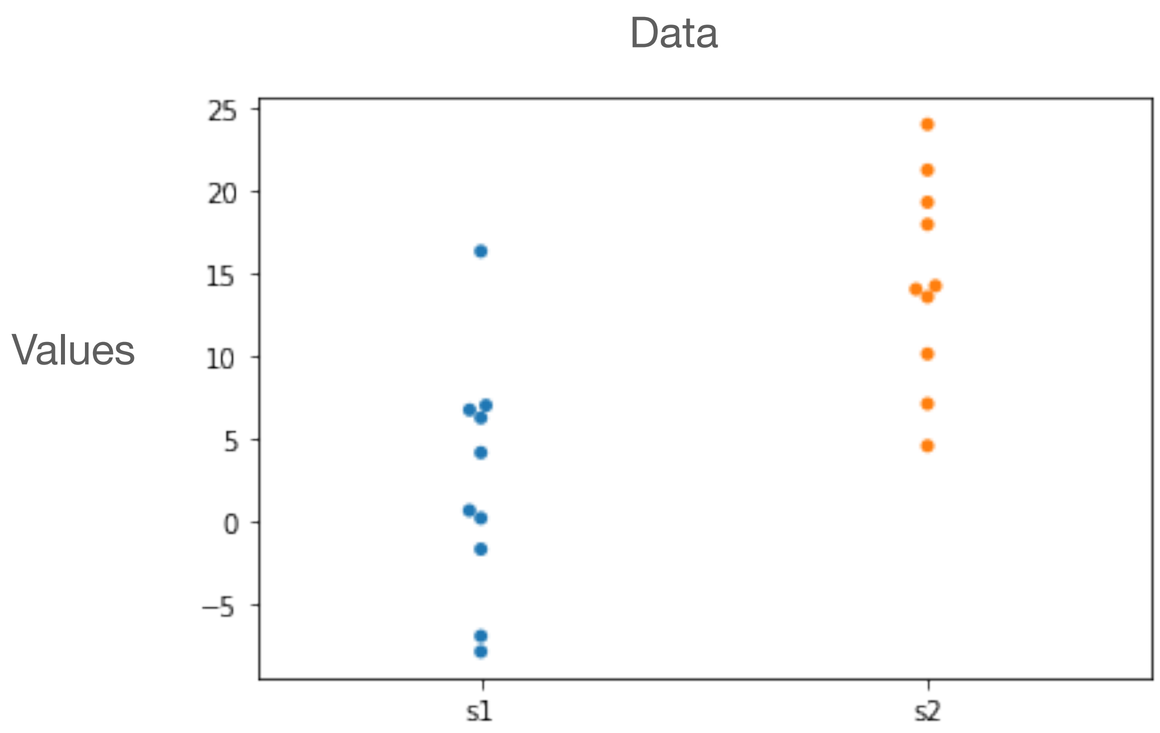
Assumptions

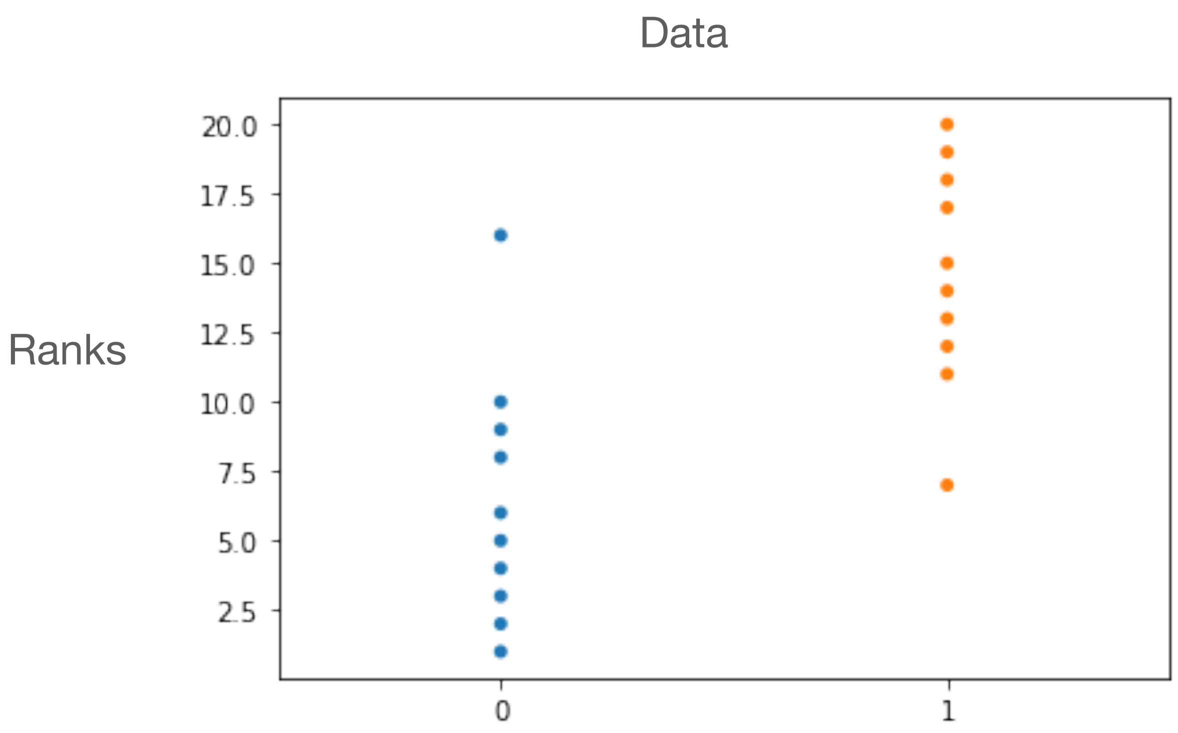




- Random sampling
- Each value is obtained independently

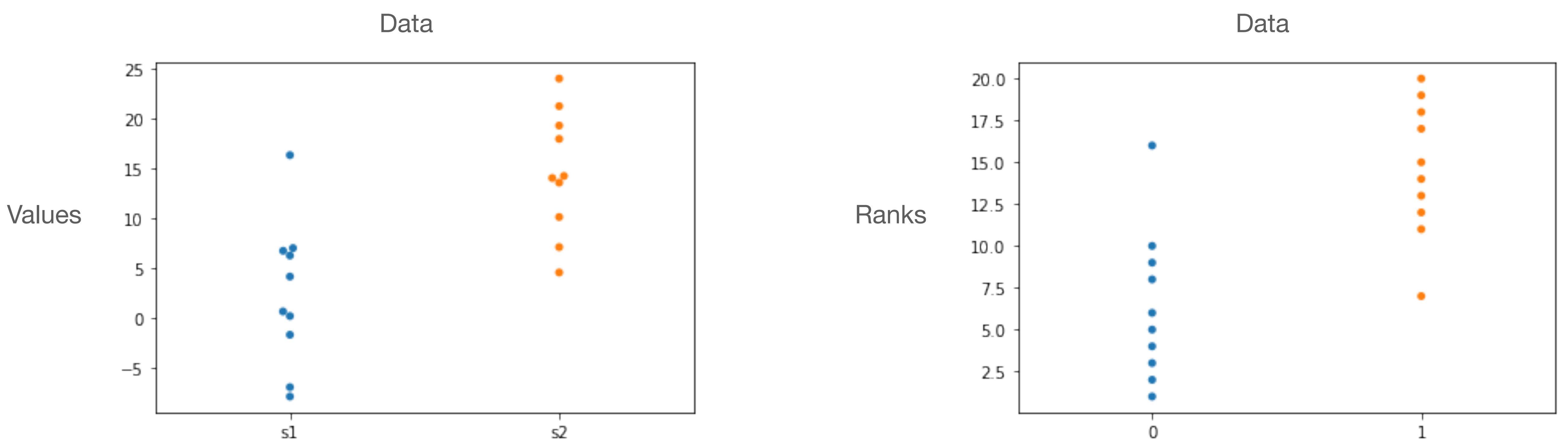
Sample sizes





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.

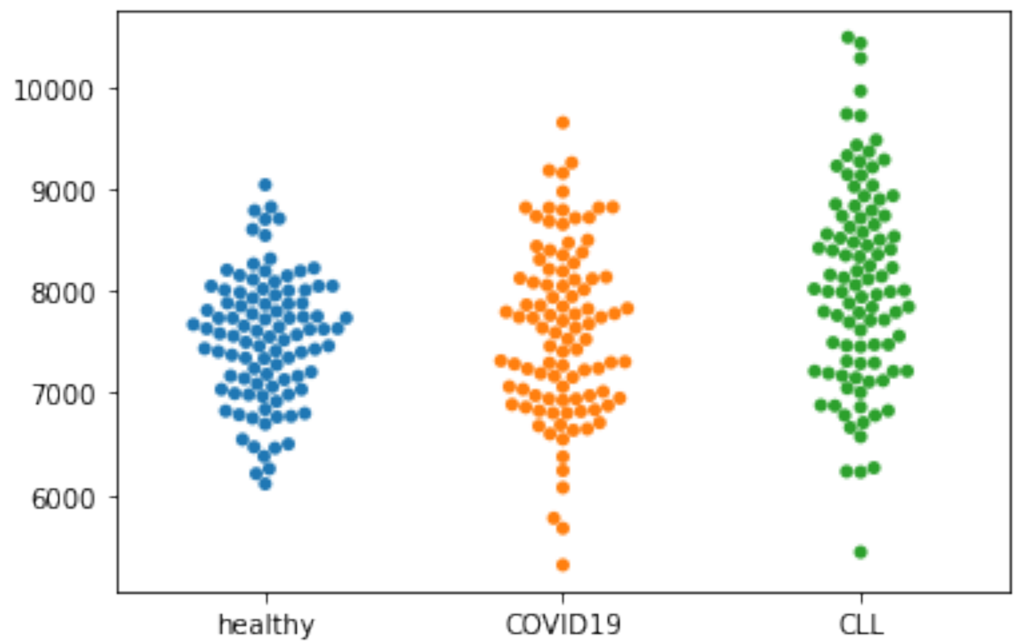
Don’t do both and pick the best!!!!!!!!!!!!

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?

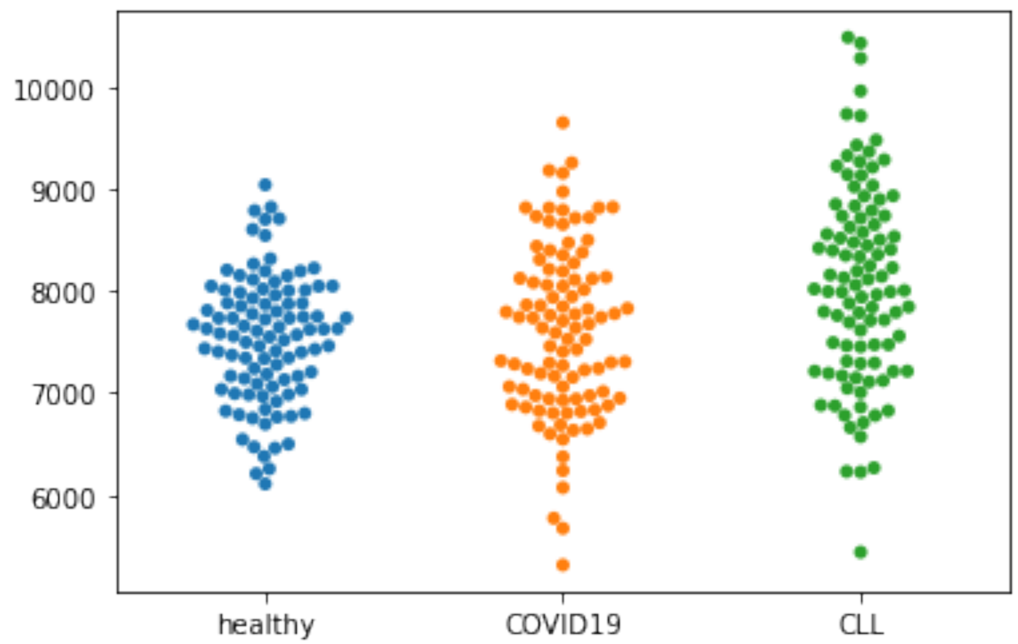
-> Analysis of Variance, one-way ANOVA (= multi-sample-t-test)

-> 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)

What happens, if you don’t control for multiple testing?

If you do 50 experiments with a significance threshold of 0.05, how many do you expect to be “significant” by chance?

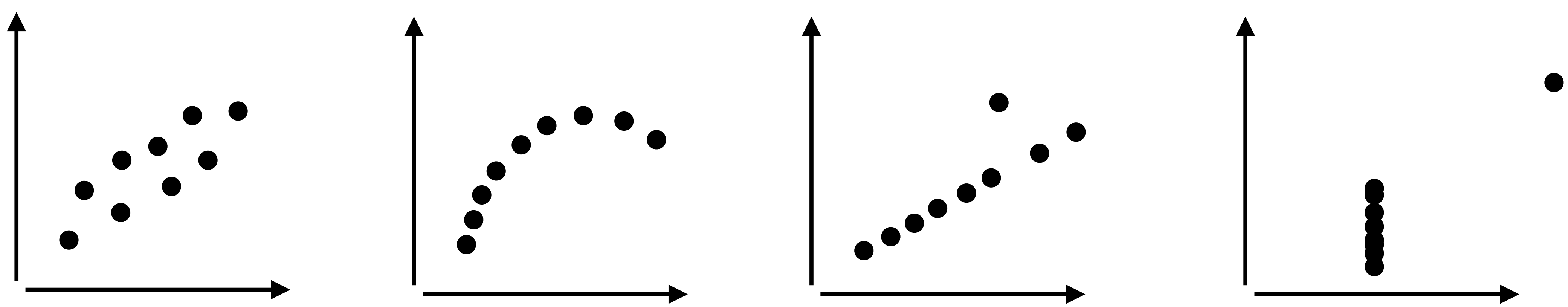
Correlations

What for?

To compare paired data in a population.

Correlations are defined by a correlation coefficient (R) and a p-value

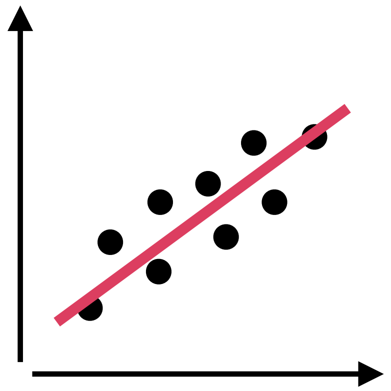
Main rule for any correlation analysis: **Look at your data first!**



These would all roughly have the same correlation coefficient!

Correlations

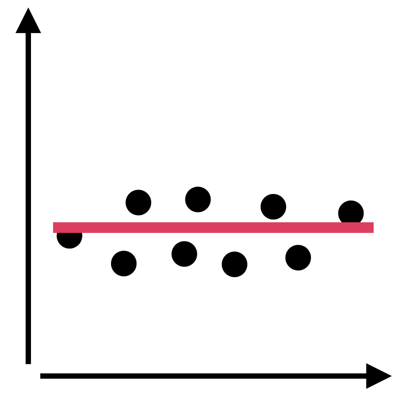
Positive



R = 0.7

p = 0.01

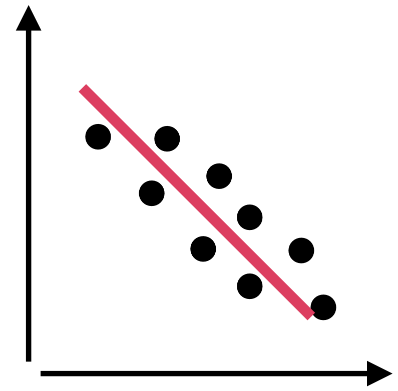
None



R = 0.05

p = 0.01

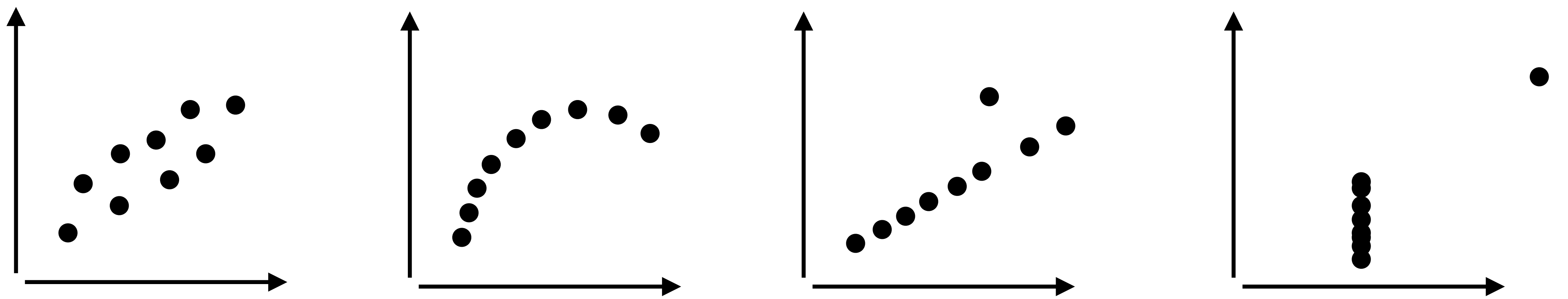
Negative



R = -0.7

p = 0.01

Assumptions



- Random sample
- Paired samples
- Sampled from one populations
- Independent observations
- X-values are not used to compute y-values
- Values are not experimentally controlled

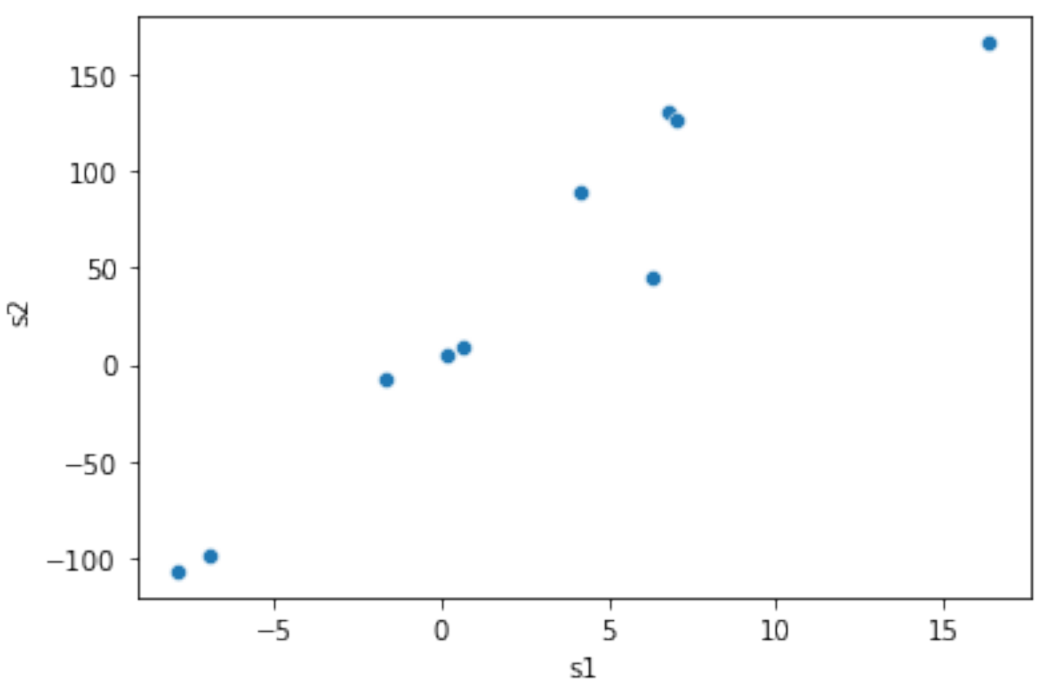
Specifically for parametric:

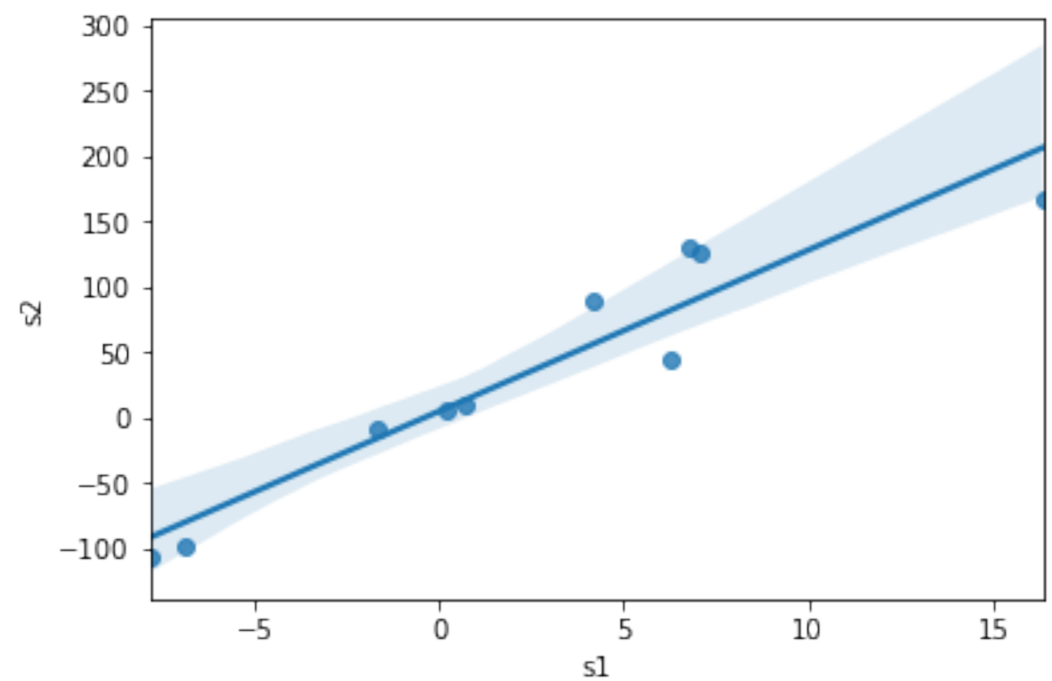
- Approximate normal distribution
- All covariation is linear
- No outliers !!!!

Pearson Correlation

With regression line and

confidence interval





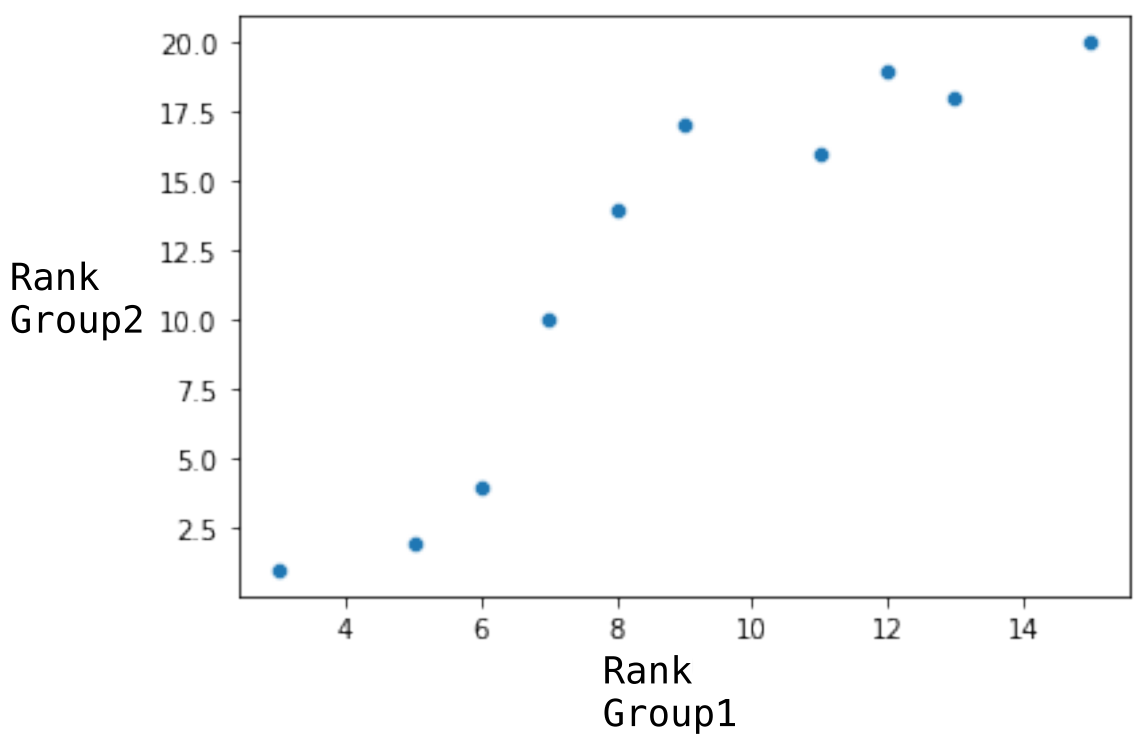
Parametric correlation statistics

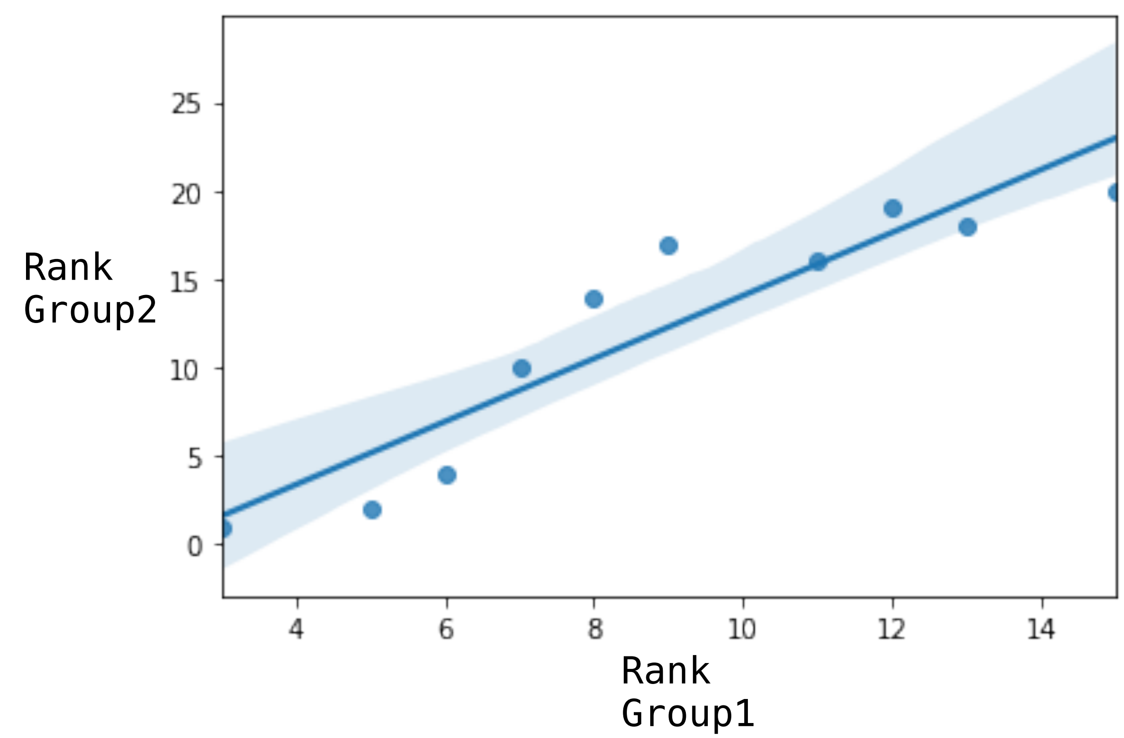
```
R = 0.95
p = 2.6e-05
```

Spearman Correlation

With regression line and

confidence interval





Non-parametric correlation statistics

R = 0.97
p = 1.5e-06

Correlation statistics

Correlation does not mean causation!

Beware your data structure and outliers!

Summary

- Non-parametric testing
- Multiple testing correction
- Correlation statistics

-> Jupyter Notebook