Introduction into Biostatistics

Anna Poetsch, Biotechnology Center, TU Dresden

Organisation

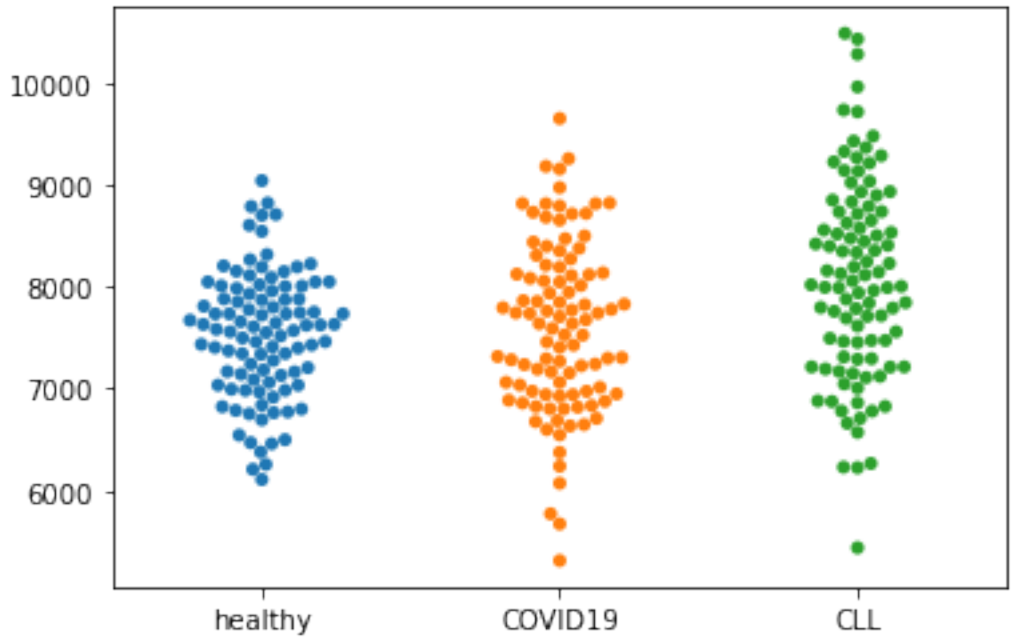
- 16.5. Introduction to biostatistics
- 14.6. Distributions and hypothesis testing
- 21.6. Non-parametric testing and multiple comparisons
- **28.6. Correlations and dimensionality reduction**

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?

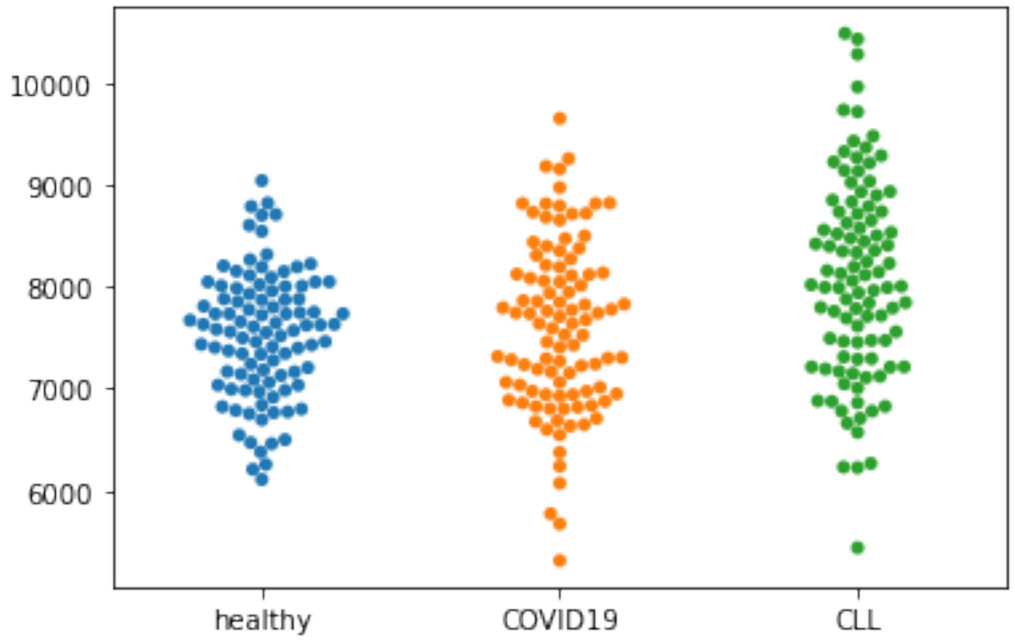
-> Analysis of Variance, one-way ANOVA (= multi-sample-t-test)

-> 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 200 experiments with a significance threshold of 0.05, how many do you expect to be “significant” by chance?

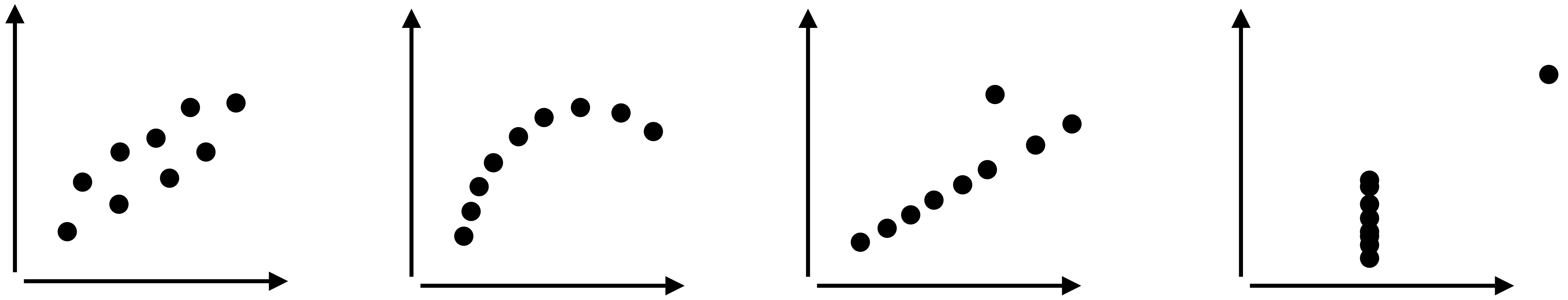
Correlations

What for?

To compare paired data in a population.

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

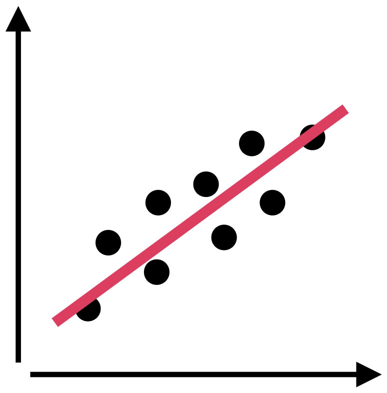
Main rule for any correlation analysis: **Look at your data first!**



These would all roughly have the same correlation coefficient!

Correlations

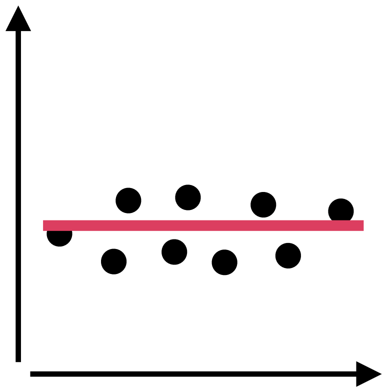
Positive



R = 0.7

p = 0.01

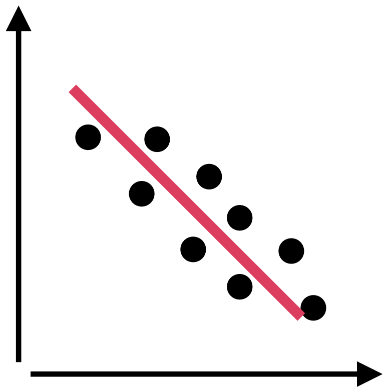
None



R = 0.05

p = 0.01

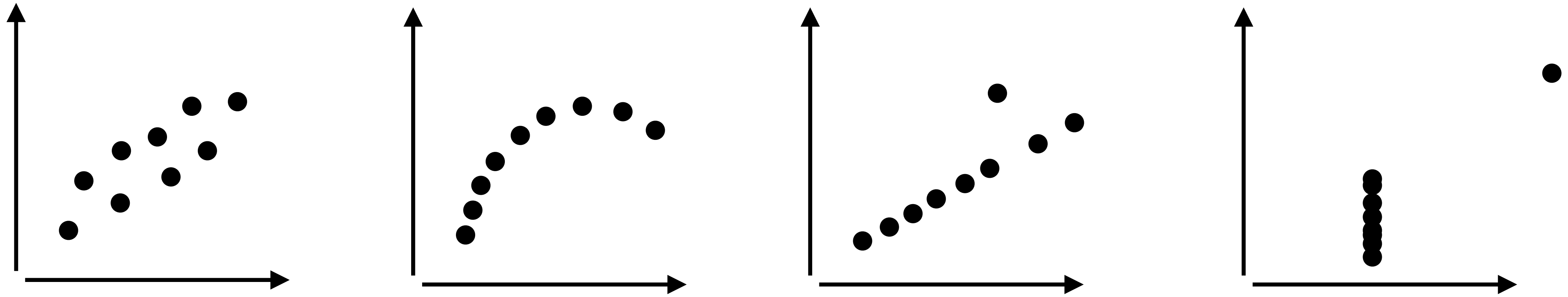
Negative



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

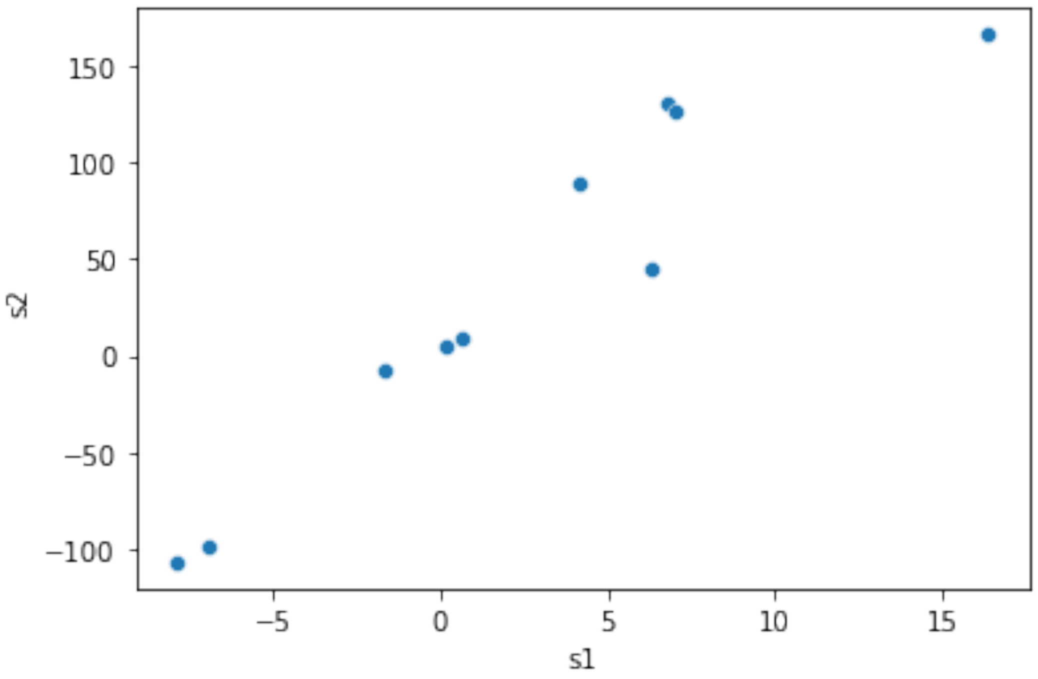
Specifically for parametric:

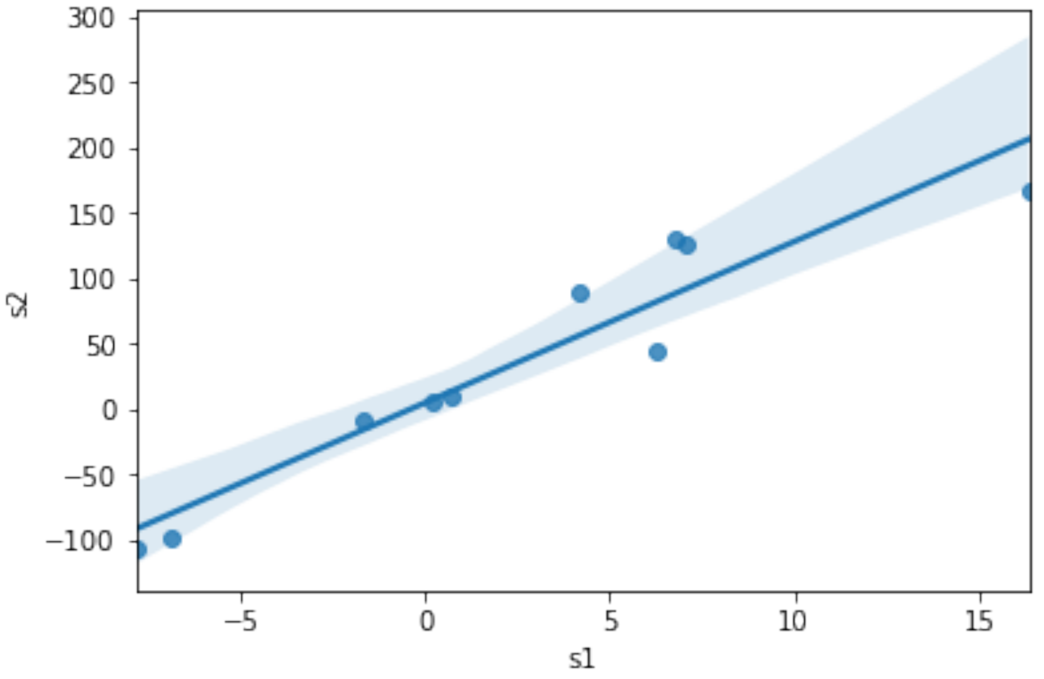
- Approximate normal distribution
- All covariation is linear
- No outliers !!!!

Pearson Correlation

With regression line and

confidence interval





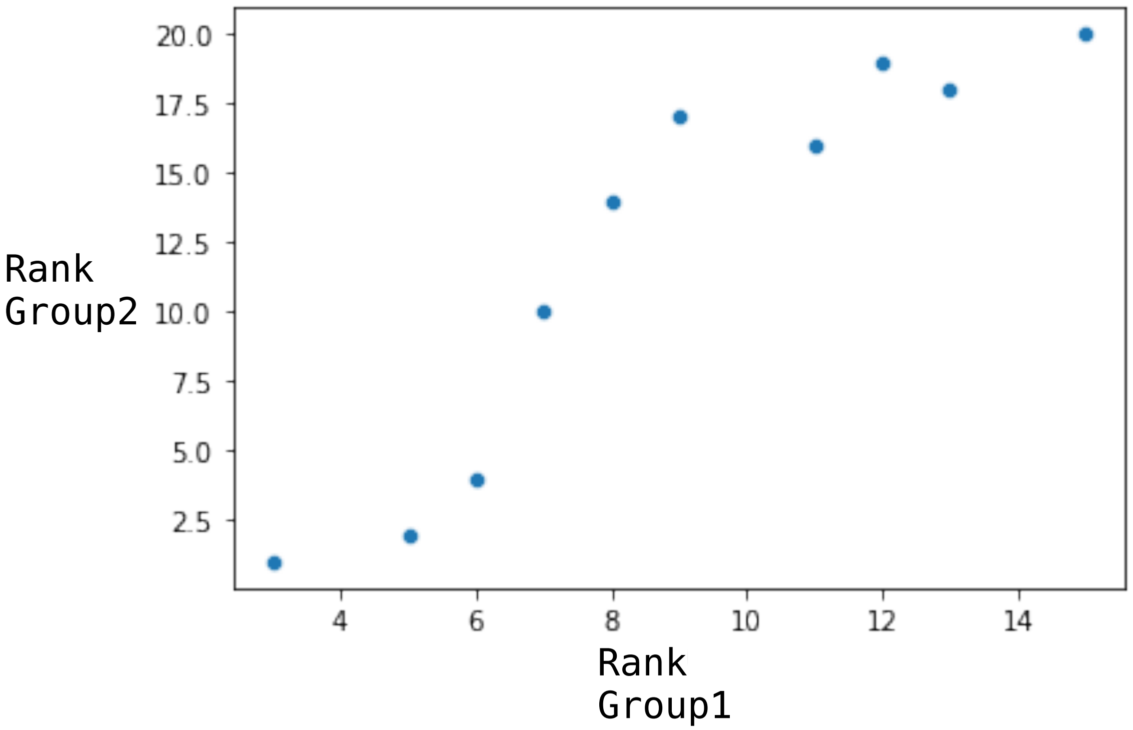
Parametric correlation statistics

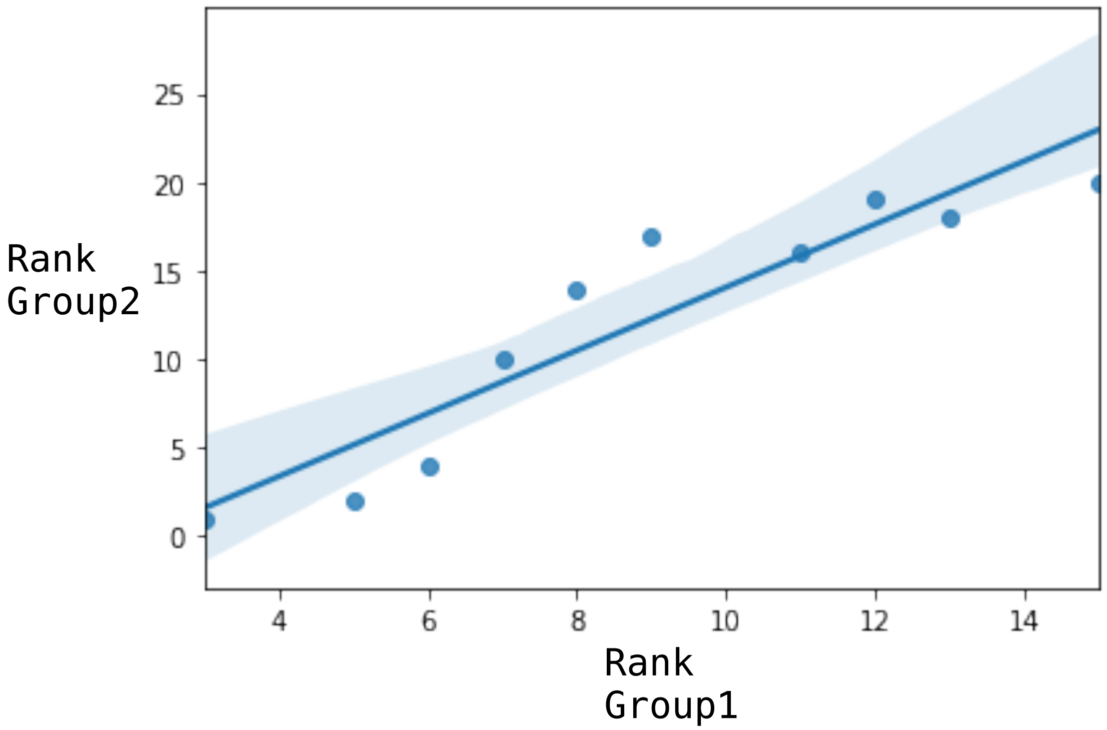
```
R = 0.95
p = 2.6e−05
```

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!

-> Jupyter Notebook