Introduction into Biostatistics

Anna Poetsch, Biotechnology Center, TU Dresden

Organisation

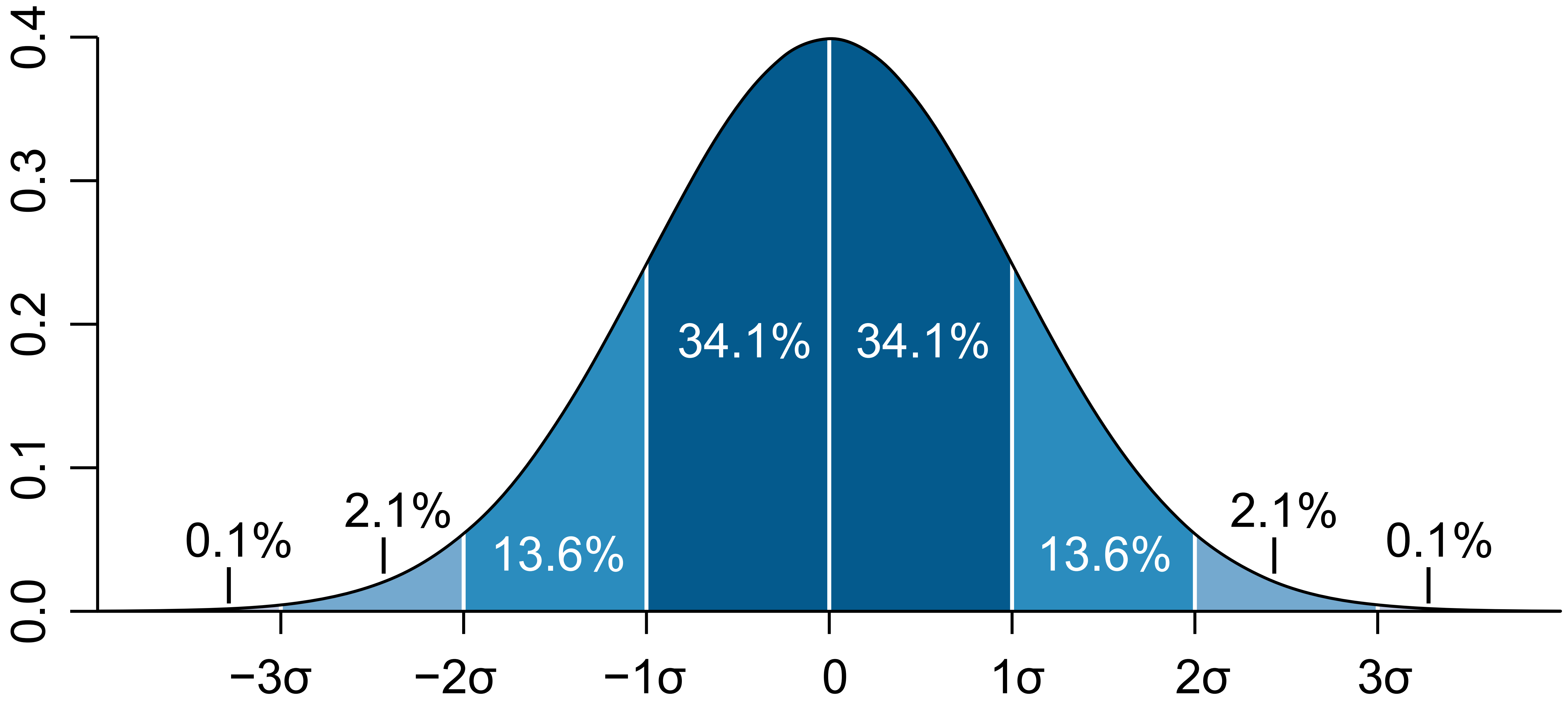
- **9.5. Introduction to biostatistics**
- **16.5. Descriptive statistics**
- **23.5. Hypothesis testing**
- 6.6. Introduction machine learning (Robert)
- 13.6. Unsupervised Machine learning (Melissa)
- 20.6. Supervised machine learning/ deep learning (Melissa)
- 21.6. Introduction into genomics data
- 4.7. Multimodal machine learning
- 11.7. Summary (all)

Evaluation

https://befragung.zqa.tu-dresden.de/uz/de/sl/ZHf1mF00C1dP

Normal distribution

Gaussian distribution, bell-shaped distribution



The result of general imprecision: weighing, pipetting, randomness

Therefore also: height, weight

Density defined by mean and standard deviation

The graph is adapted from: M. W. Toews - Own work, based (in concept) on figure by Jeremy Kemp, on 2005-02-09, CC BY 2.5, https://commons.wikimedia.org/w/index.php?curid=1903871

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

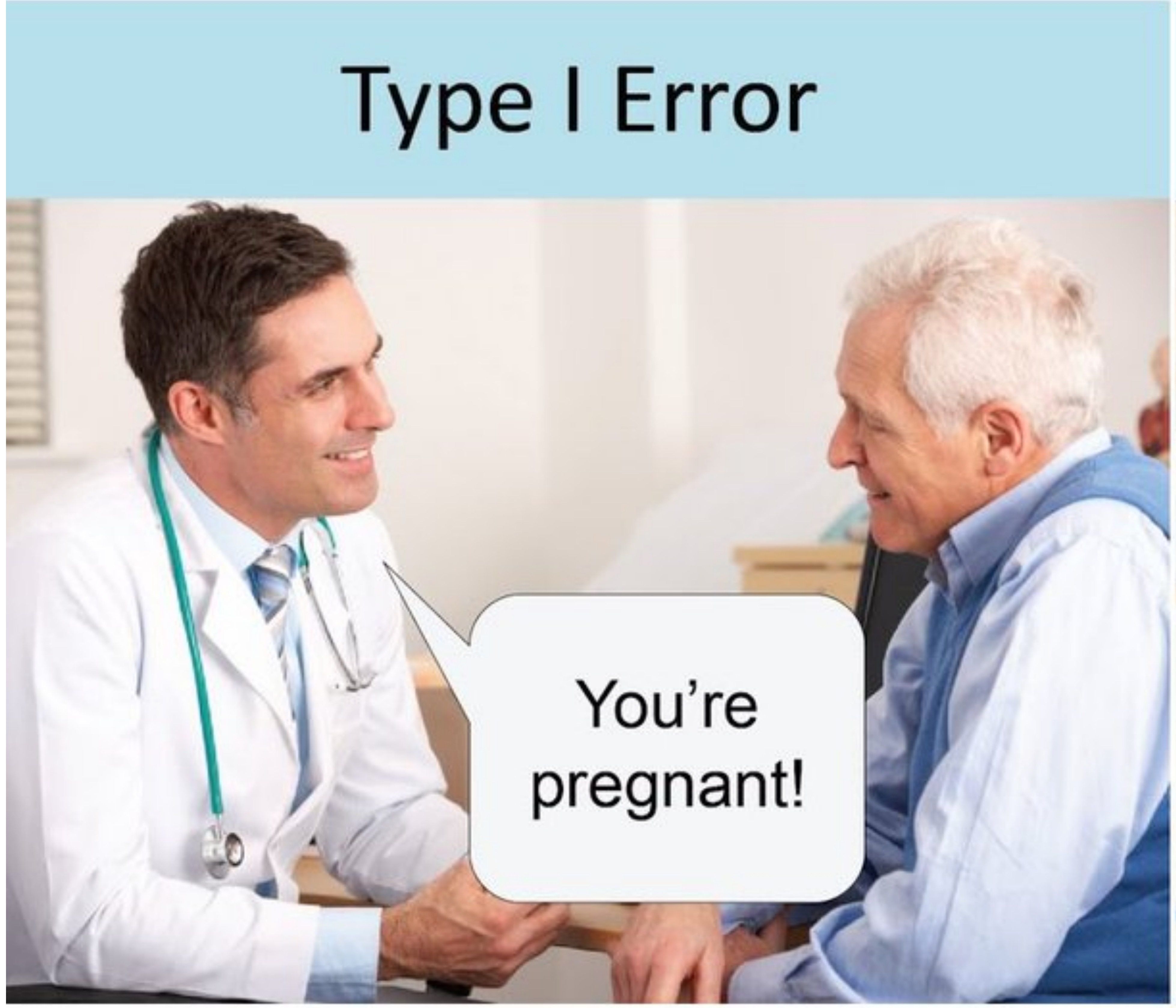
How to reject H_0

How much probability do you allow yourself to be wrong?

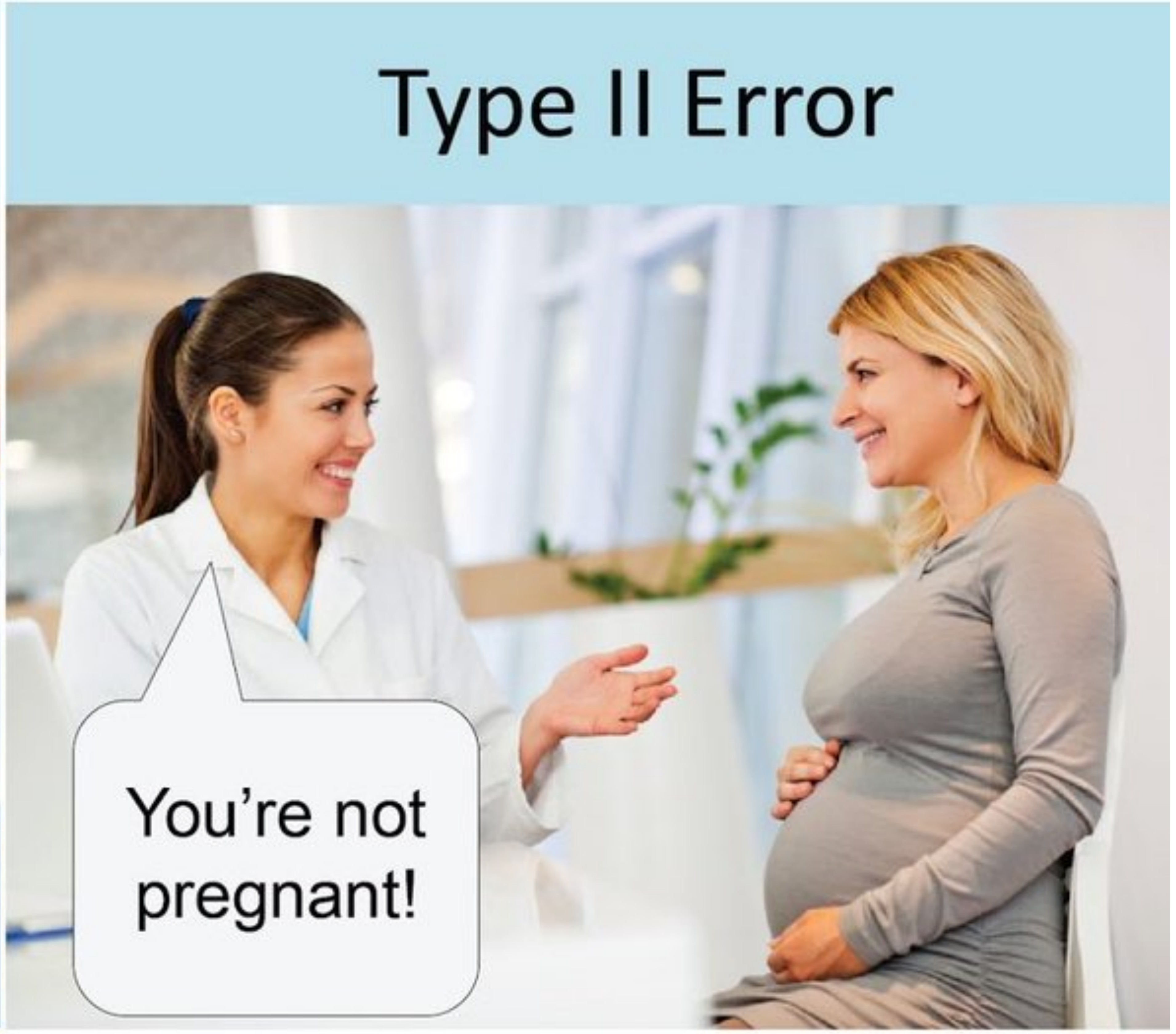
- last line chemotherapy treatment: Every bit of hope counts

- 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.

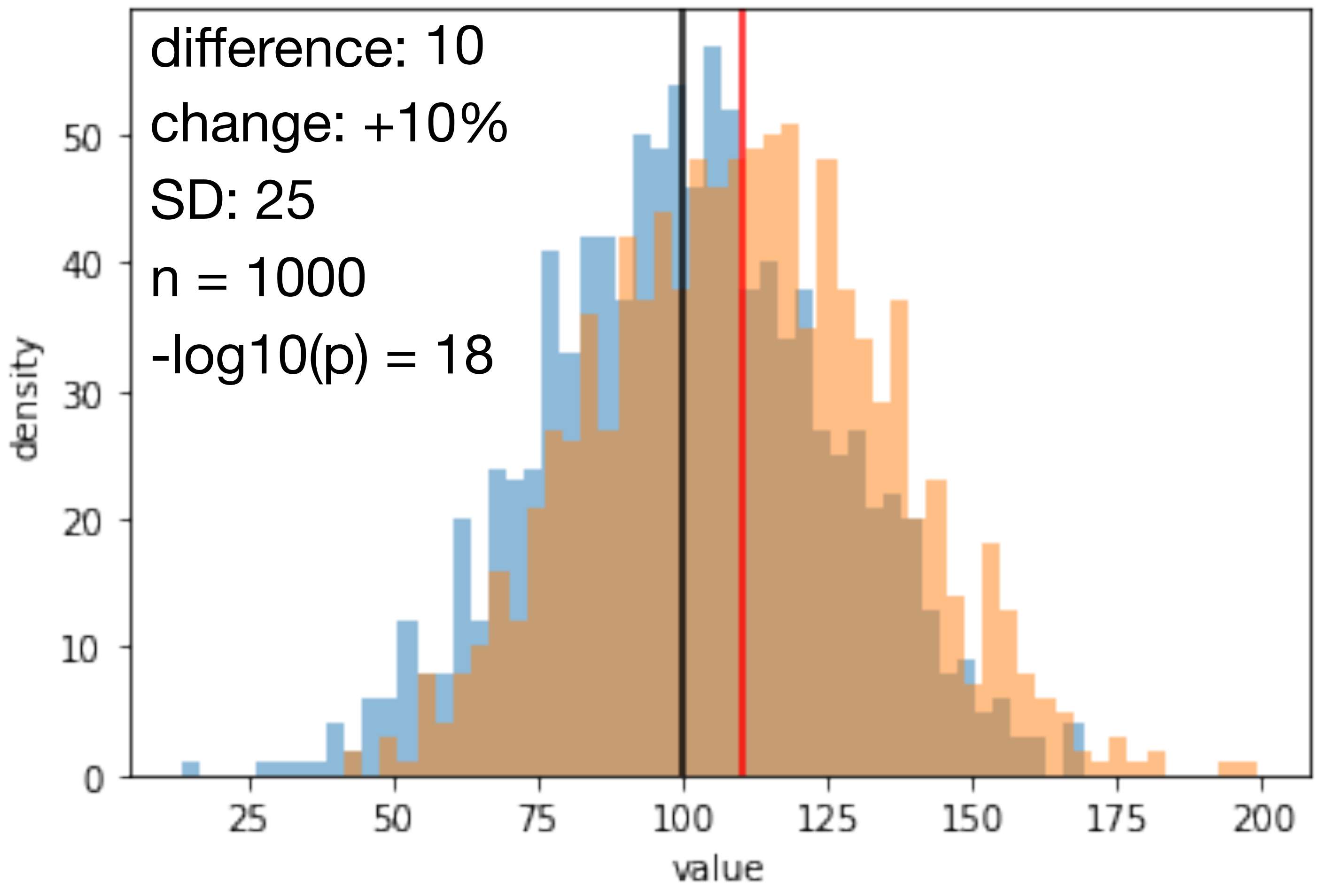
Misconceptions about P-values

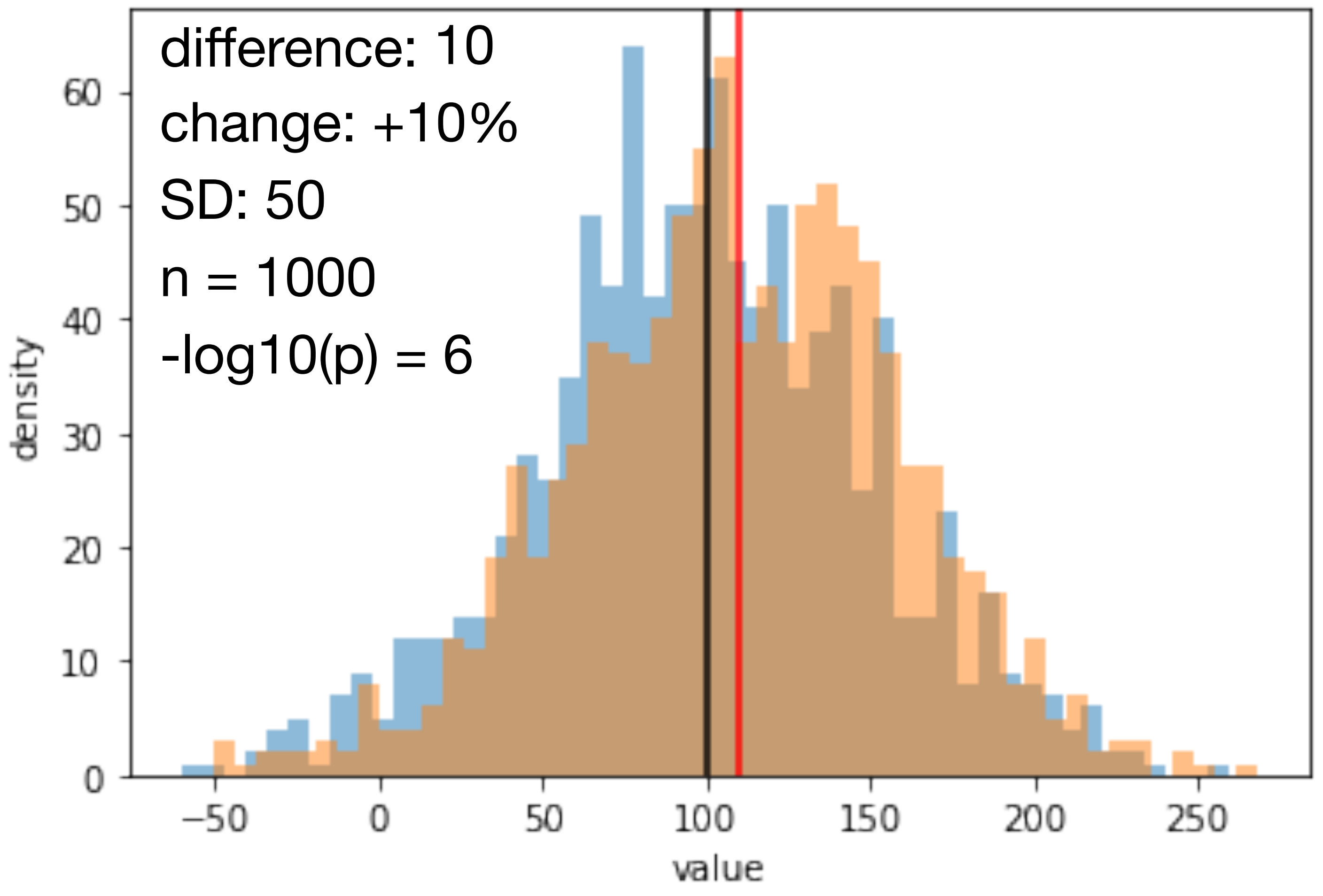
The logic is not reversible:

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.

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





Misconceptions about P-values

Reproducibility of p-values is inherently very poor.

For measures of reproducibility of an effect the appropriate measure is the effect size, e.g. the actual difference or ratio.

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

Summary parameters

1 2 2 5 5 5 10 30

Min value: 1

Max value: 30

Parametric measures

$$\mathrm{Var}(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

https://en.wikipedia.org/wiki/Variance

Summary parameters

1 2 2 5 5 5 10 30

Min value: 1

Max value: 30

Parametric measures

Mean (μ): (1+2+2+5+5+5+10+30)/8 = 7.5

$$\mathrm{Var}(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

https://en.wikipedia.org/wiki/Variance

Summary parameters

1 2 2 5 5 5 10 30

Min value: 1

Max value: 30

Parametric measures

Mean (μ): (1+2+2+5+5+5+10+30)/8 = 7.5

Variance: $\mathrm{Var}(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$

$((1-7.5)^2+(2-7.5)^2+(2-7.5)^2+(5-7.5)^2+(5-7.5)^2+(5-7.5)^2+(10-7.5)^2+(30-7.5)^2)/8 = 90.57143$

https://en.wikipedia.org/wiki/Variance

Summary parameters

1 2 2 5 5 5 10 30

Min value: 1

Max value: 30

Parametric measures

Mean (μ): (1+2+2+5+5+5+10+30)/8 = 7.5

Variance: $\mathrm{Var}(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$

$((1-7.5)^2+(2-7.5)^2+(2-7.5)^2+(5-7.5)^2+(5-7.5)^2+(5-7.5)^2+(10-7.5)^2+(30-7.5)^2)/8 = 90.57143$

SD: square_root (variance) = 9.516902

https://en.wikipedia.org/wiki/Variance

Summary parameters

1 2 2 5 5 5 10 30

Min value: 1

Max value: 30

Parametric measures

Mean (μ): (1+2+2+5+5+5+10+30)/8 = 7.5

Variance:

$$\mathrm{Var}(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

$((1-7.5)^2+(2-7.5)^2+(2-7.5)^2+(5-7.5)^2+(5-7.5)^2+(5-7.5)^2+(10-7.5)^2+(30-7.5)^2)/8 = 90.57143$

SD: square_root (variance) = 9.516902

SD = standard deviation = sigma

https://en.wikipedia.org/wiki/Variance

non-parametric measures:

	1	2	2	5	5	5	10	30
Ranks:	1	2	2	4	4	4	7	8

non-parametric measures:

1 2 2 5 5 5 10 30

Ranks: 1 2 2 4 4 4 7 8

Median: the central value: 5

non-parametric measures:

1 2 2 5 5 5 10 30

Ranks: 1 2 2 4 4 4 7 8

Median: the central value: 5

Quartiles: the value of the lower and upper quarter: 2, 6.25

non-parametric measures:

1 2 2 5 5 5 10 30

Ranks: 1 2 2 4 4 4 7 8

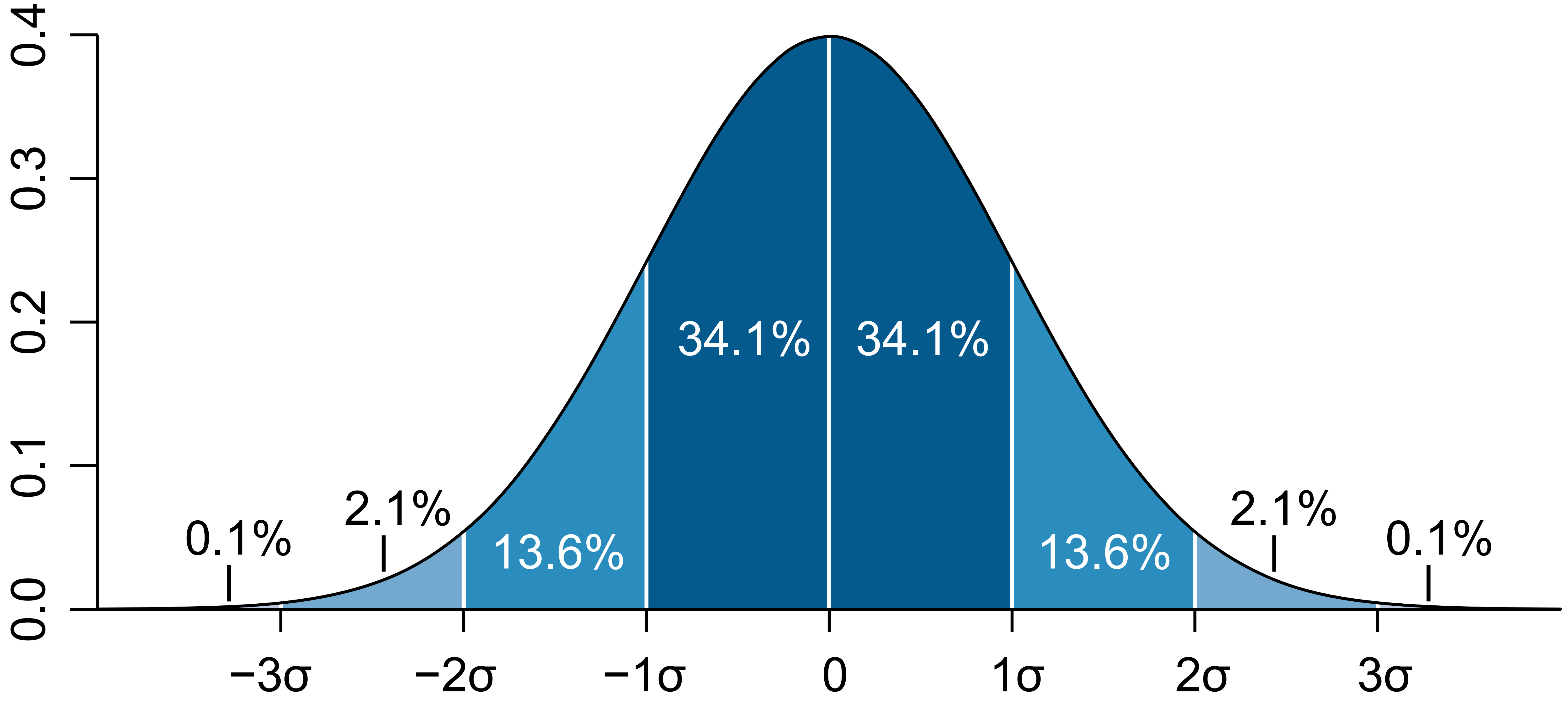
Median: the central value: 5

Quartiles: the value of the lower and upper quarter: 2, 6.25

Inter quartile range (IQR): 6.25-2

Normal distribution

Gaussian distribution, bell-shaped distribution



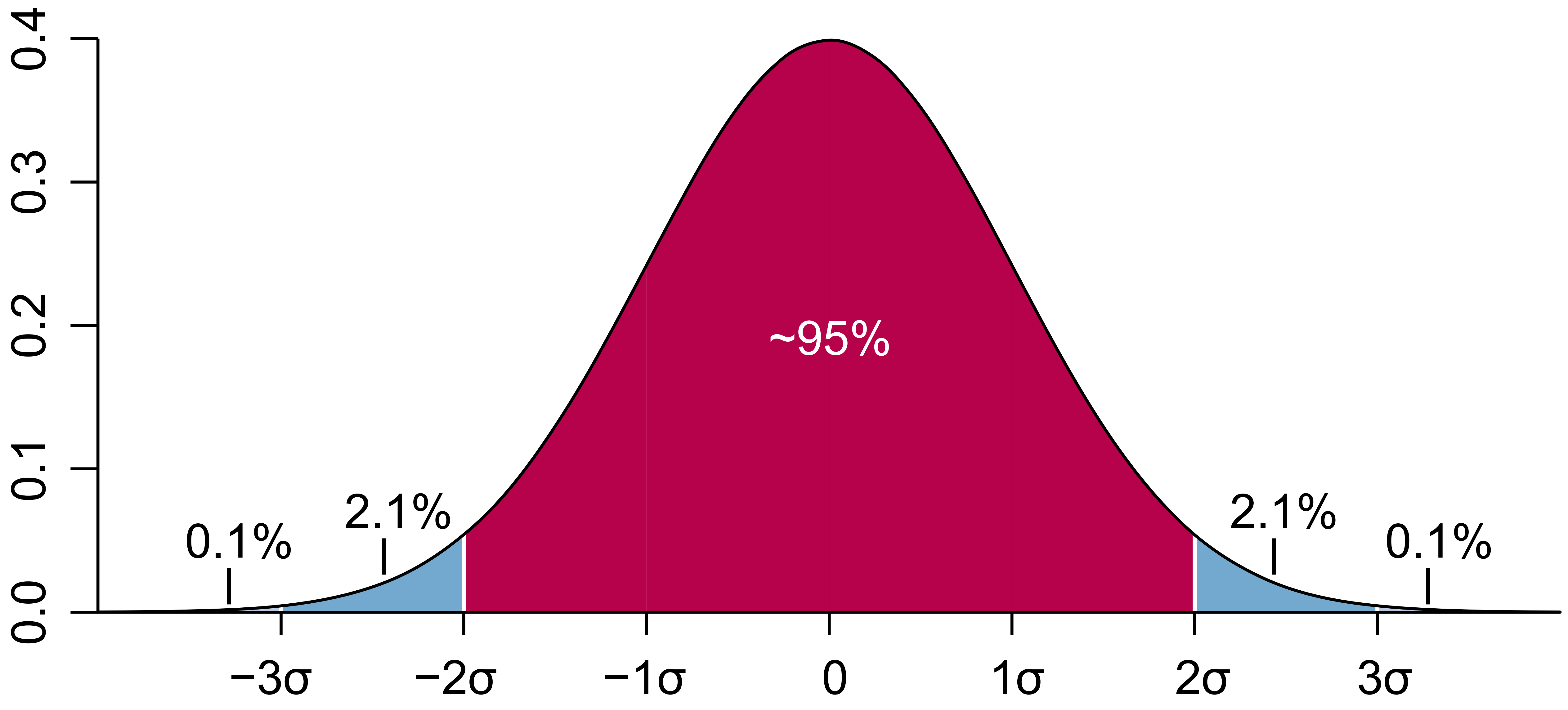
The result of general imprecision: weighing, pipetting, randomness

Therefore also: height, weight

Density defined by mean and standard deviation

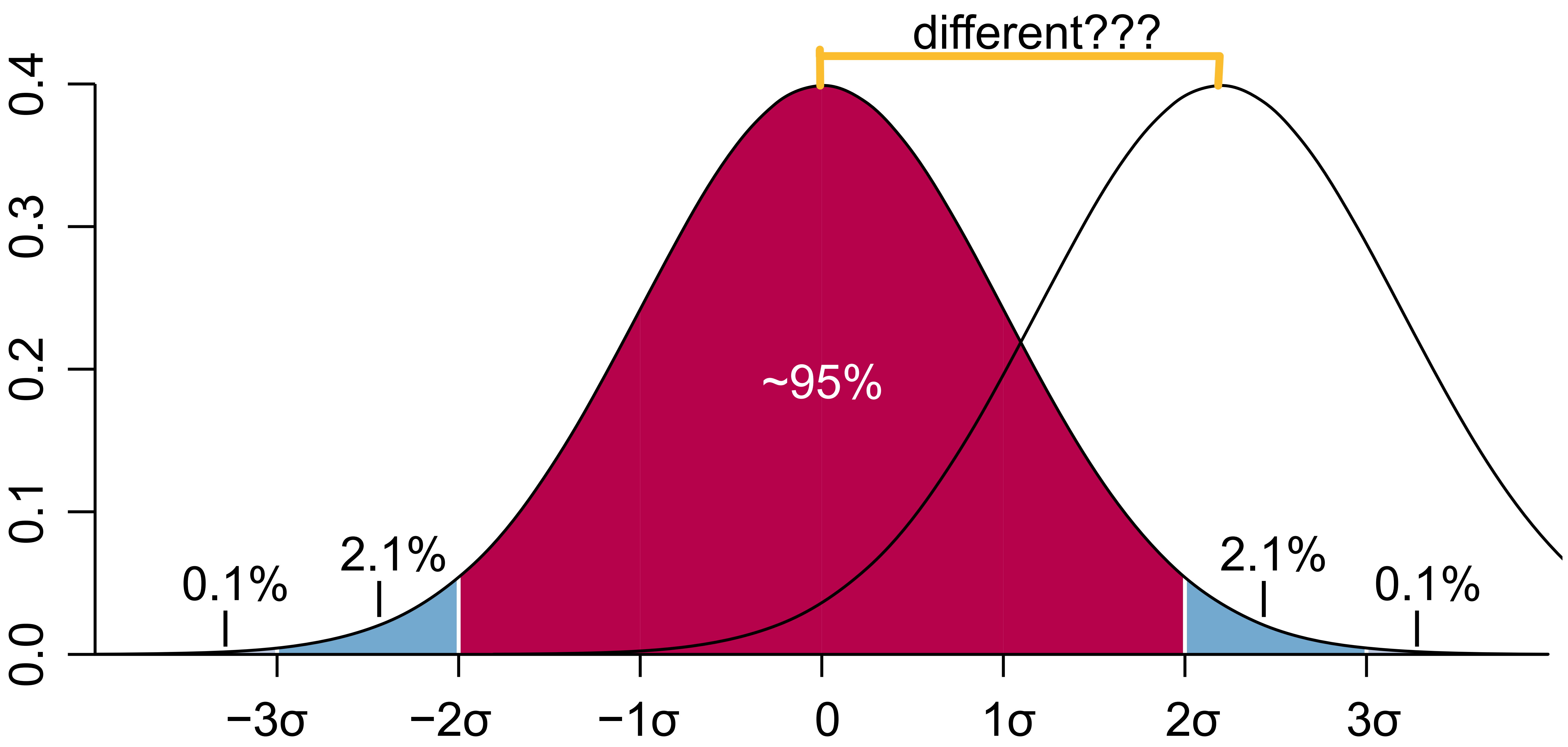
The graph is adapted from: M. W. Toews - Own work, based (in concept) on figure by Jeremy Kemp, on 2005-02-09, CC BY 2.5, https://commons.wikimedia.org/w/index.php?curid=1903871

Normal distribution



The graph is adapted from: M. W. Toews - Own work, based (in concept) on figure by Jeremy Kemp, on 2005-02-09, CC BY 2.5, https://commons.wikimedia.org/w/index.php?curid=1903871

Assumptions for unpaired parametric statistical testing

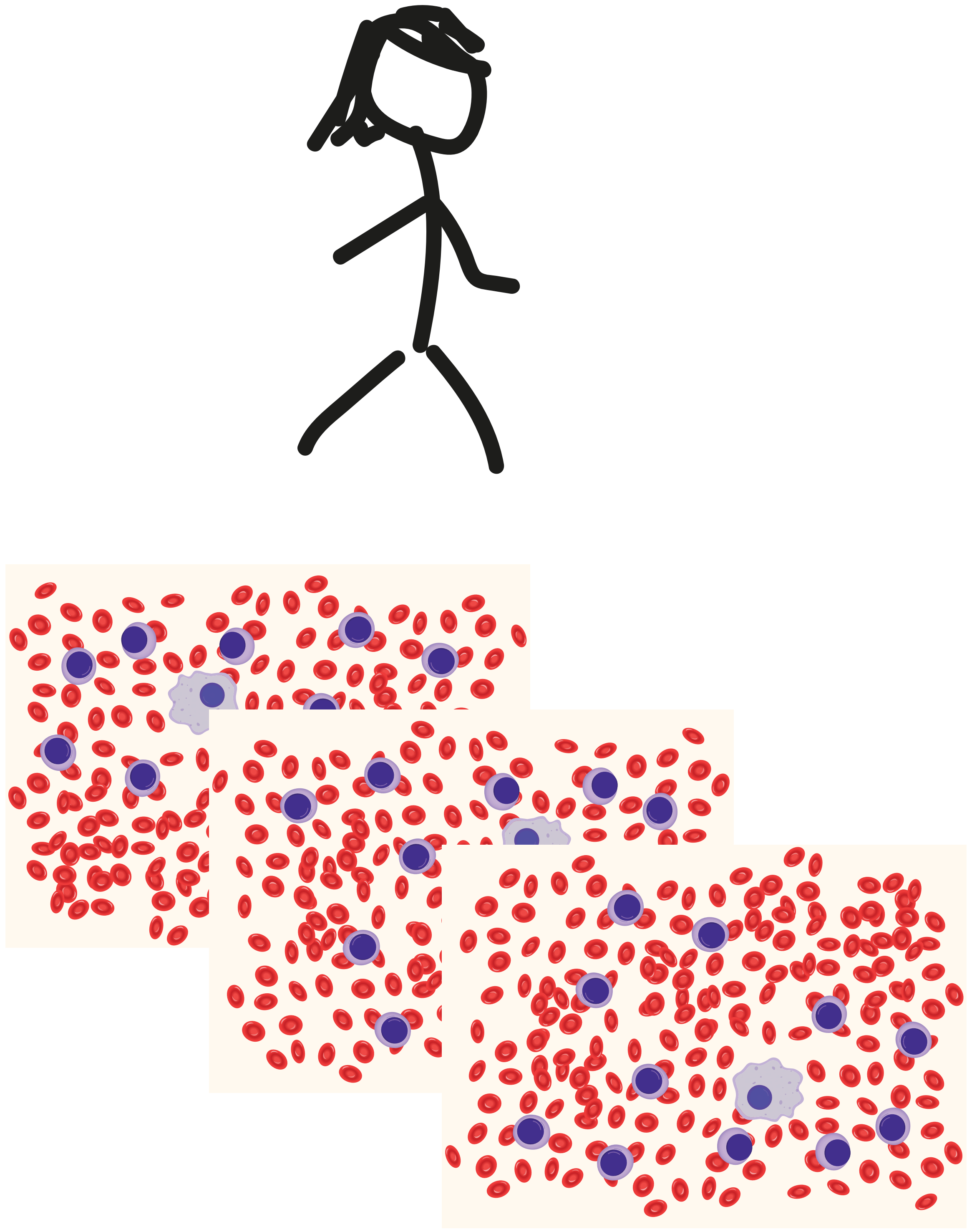


Assumptions:

- Our data follow a certain distribution
- They are representative samples
- Independent observations
- Accurate data

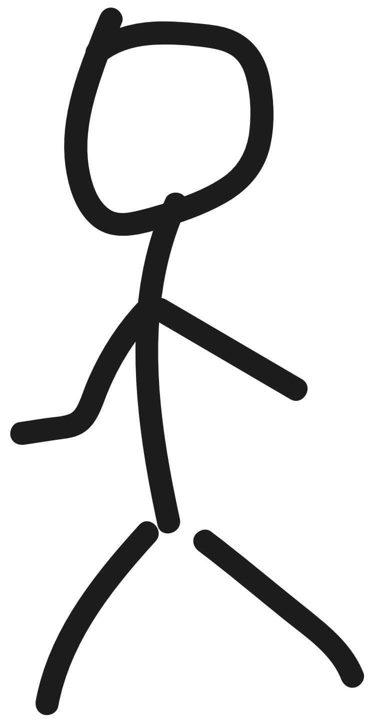
The graph is adapted from: M. W. Toews - Own work, based (in concept) on figure by Jeremy Kemp, on 2005-02-09, CC BY 2.5, https://commons.wikimedia.org/w/index.php?curid=1903871

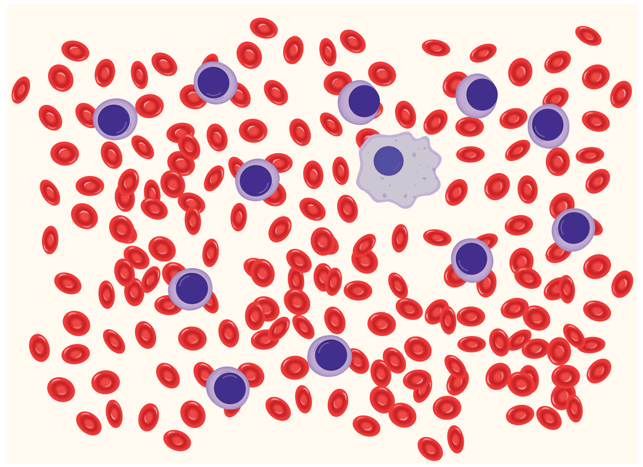
Assumptions for statistical testing



Abnormal white

blood cell count?



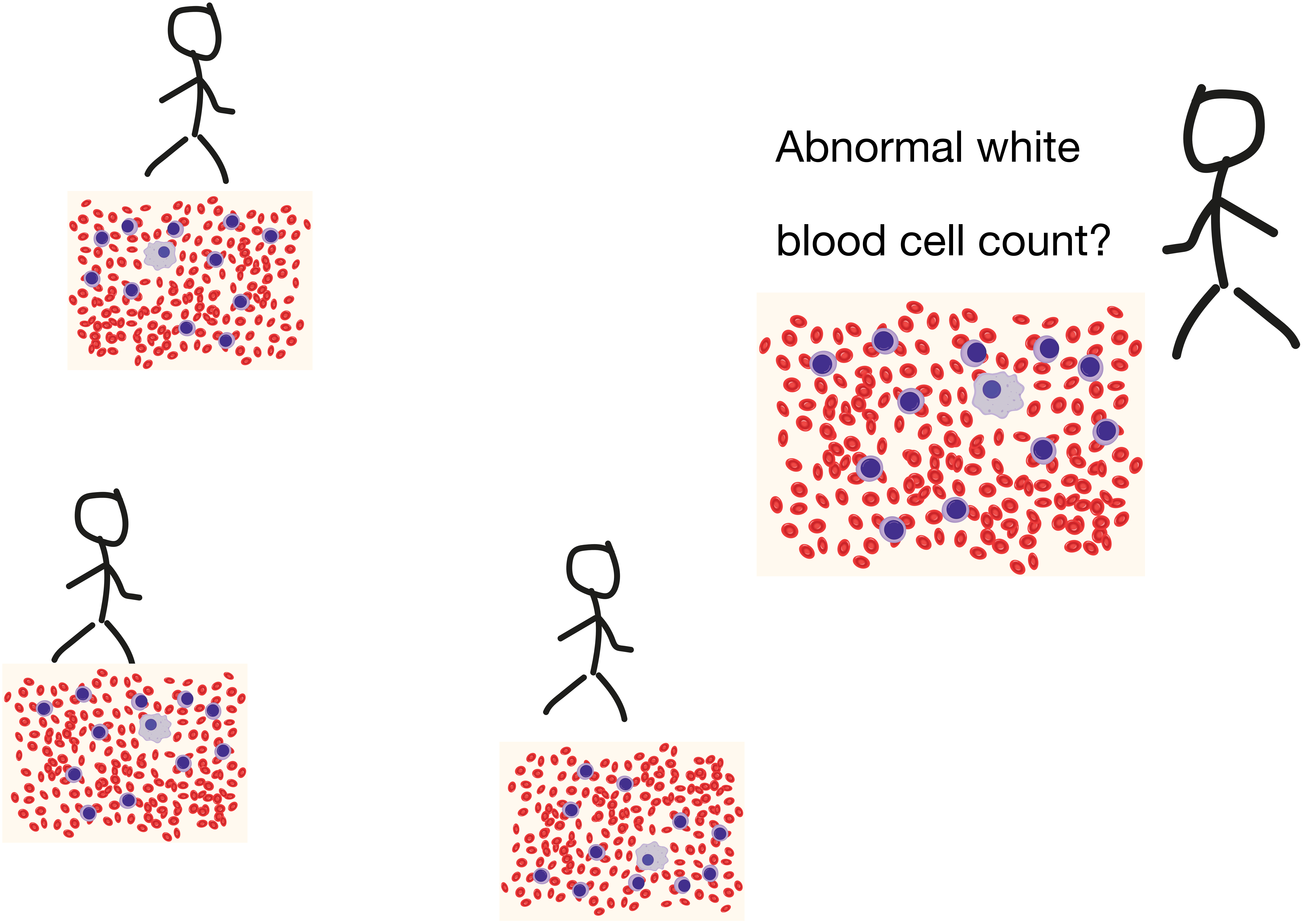


If all control cell counts are from one individual,

they are not independent!!!

Adapted from A. Rad and M. Häggström. CC-BY-SA 3.0 license

Assumptions for statistical testing

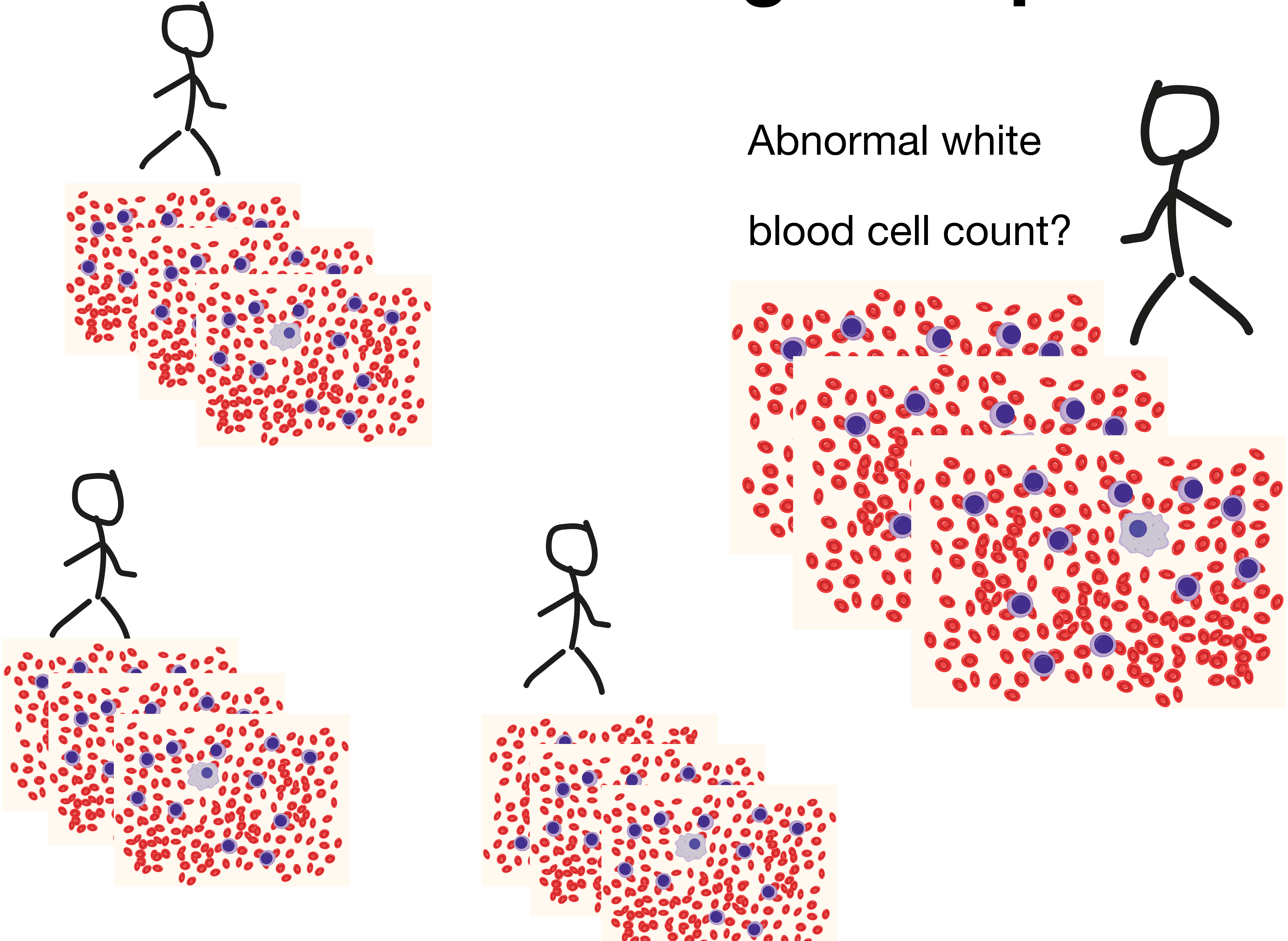


Abnormal white

blood cell count?

Adapted from A. Rad and M. Häggström. CC-BY-SA 3.0 license

Technical and biological replicates



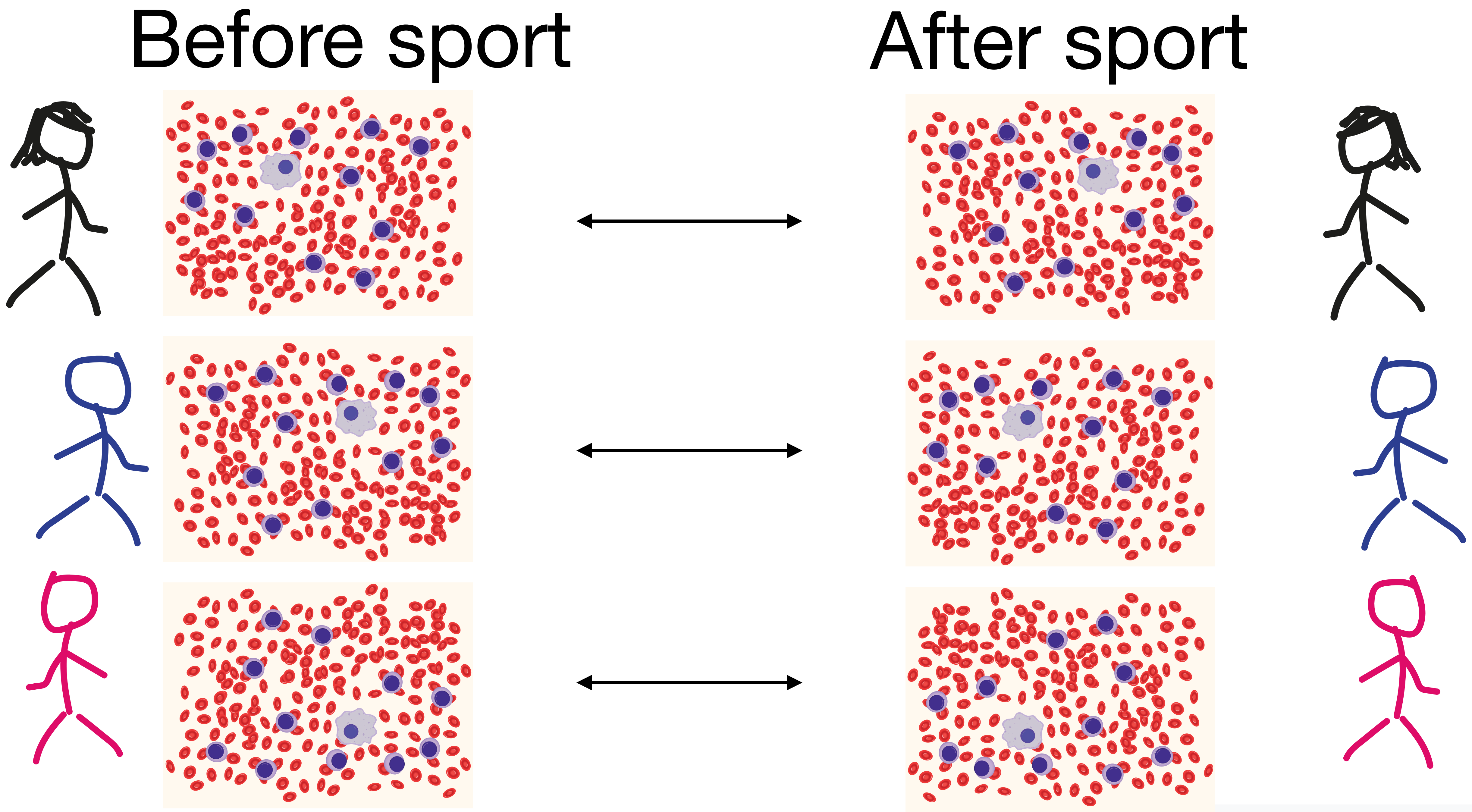
Abnormal white

blood cell count?

Special considerations

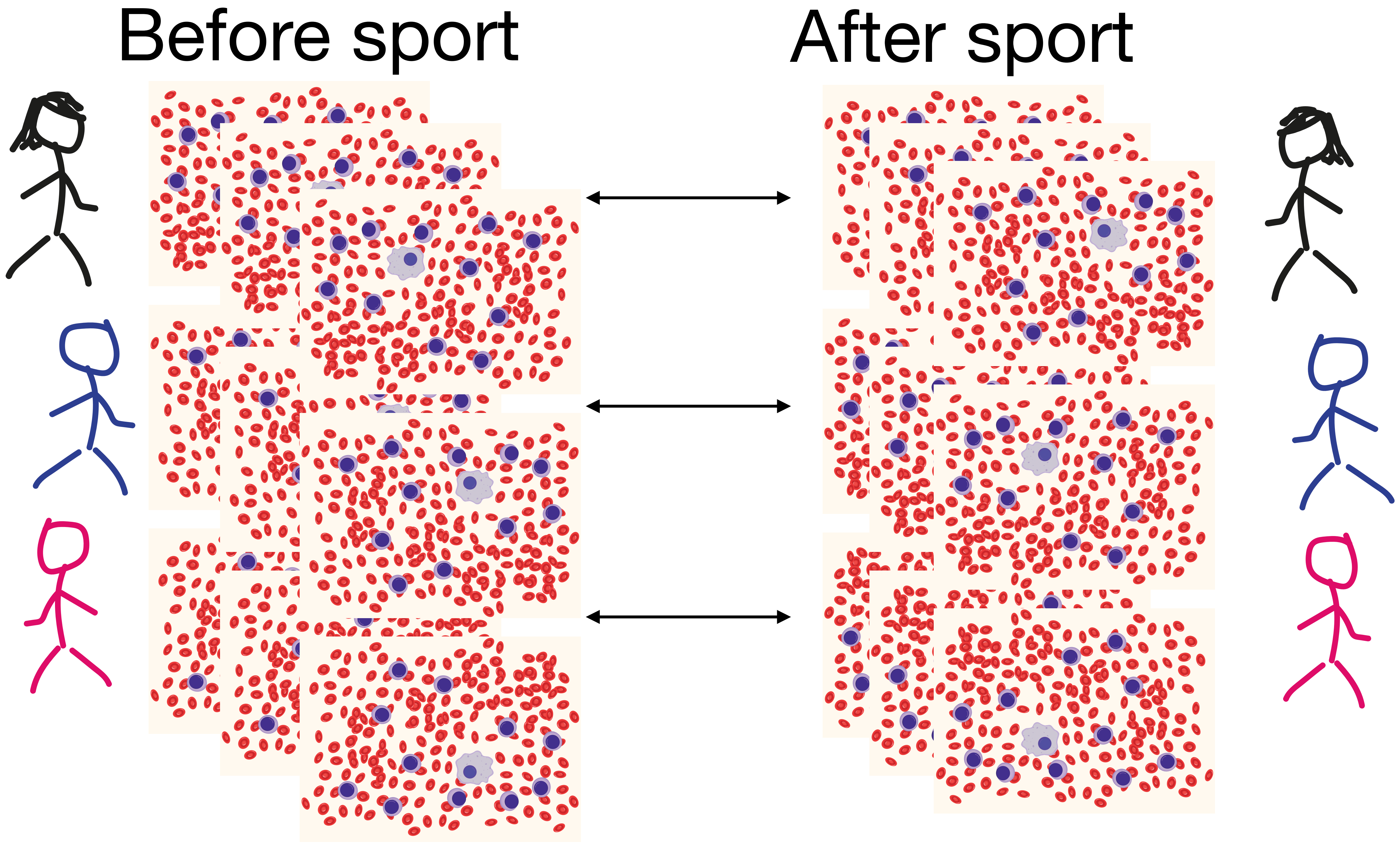
Does our hypothesis have a clear direction?
-> consider a **one-sided** test
i.e. we don’t state in H0: There no difference
But: There no increase (or decrease)

Is our data paired?



Adapted from A. Rad and M. Häggström. CC-BY-SA 3.0 license

Technical and biological replicates



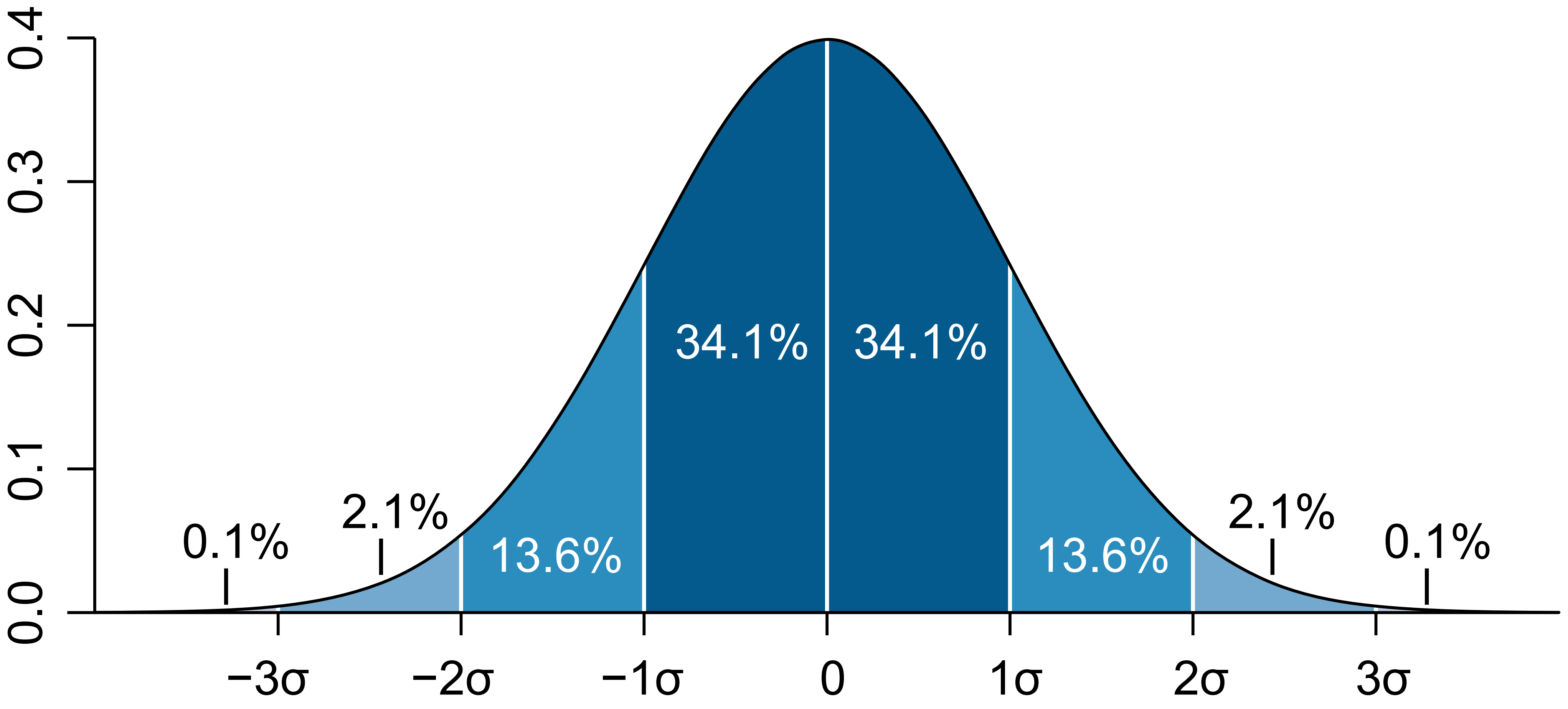
Comparing Two Means

Or: The t-test

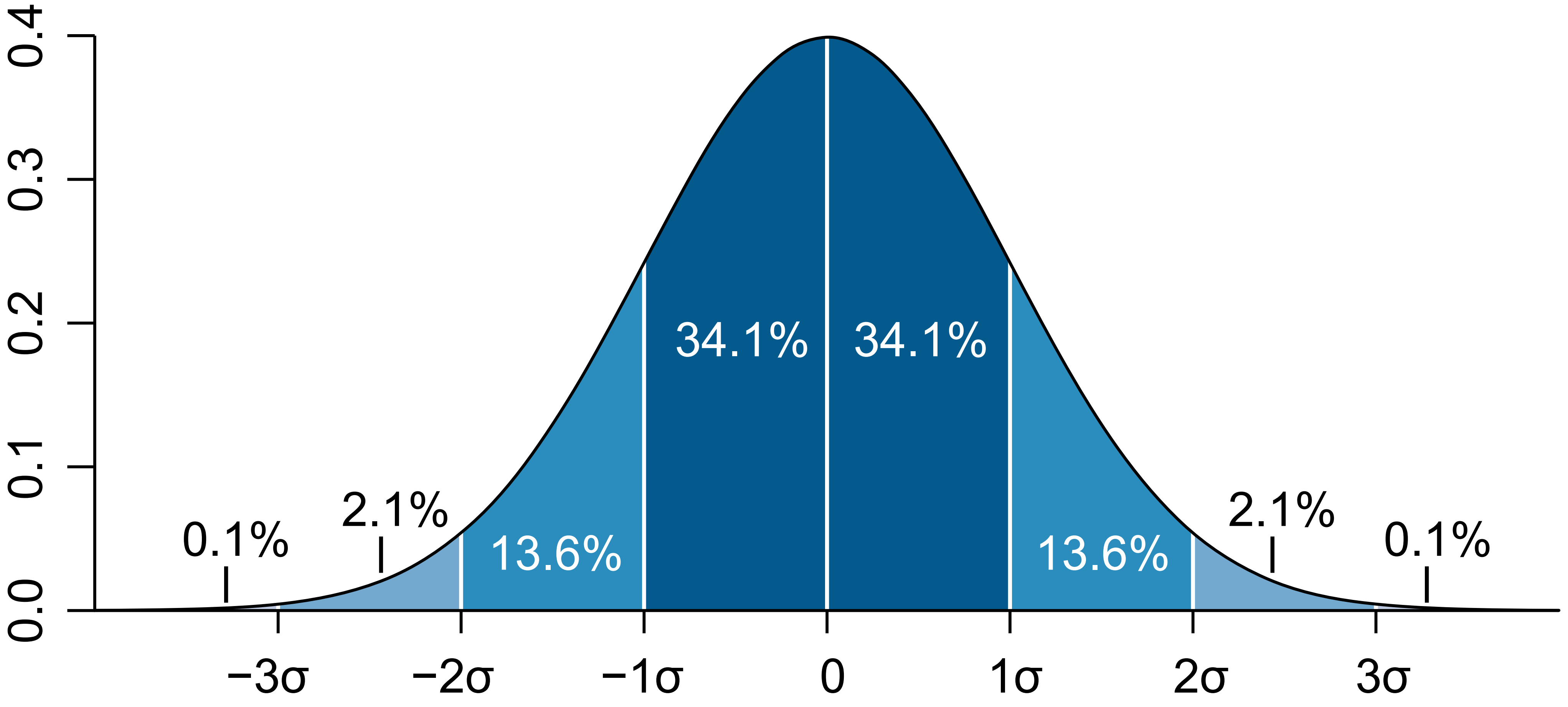
Assumptions:

- 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)

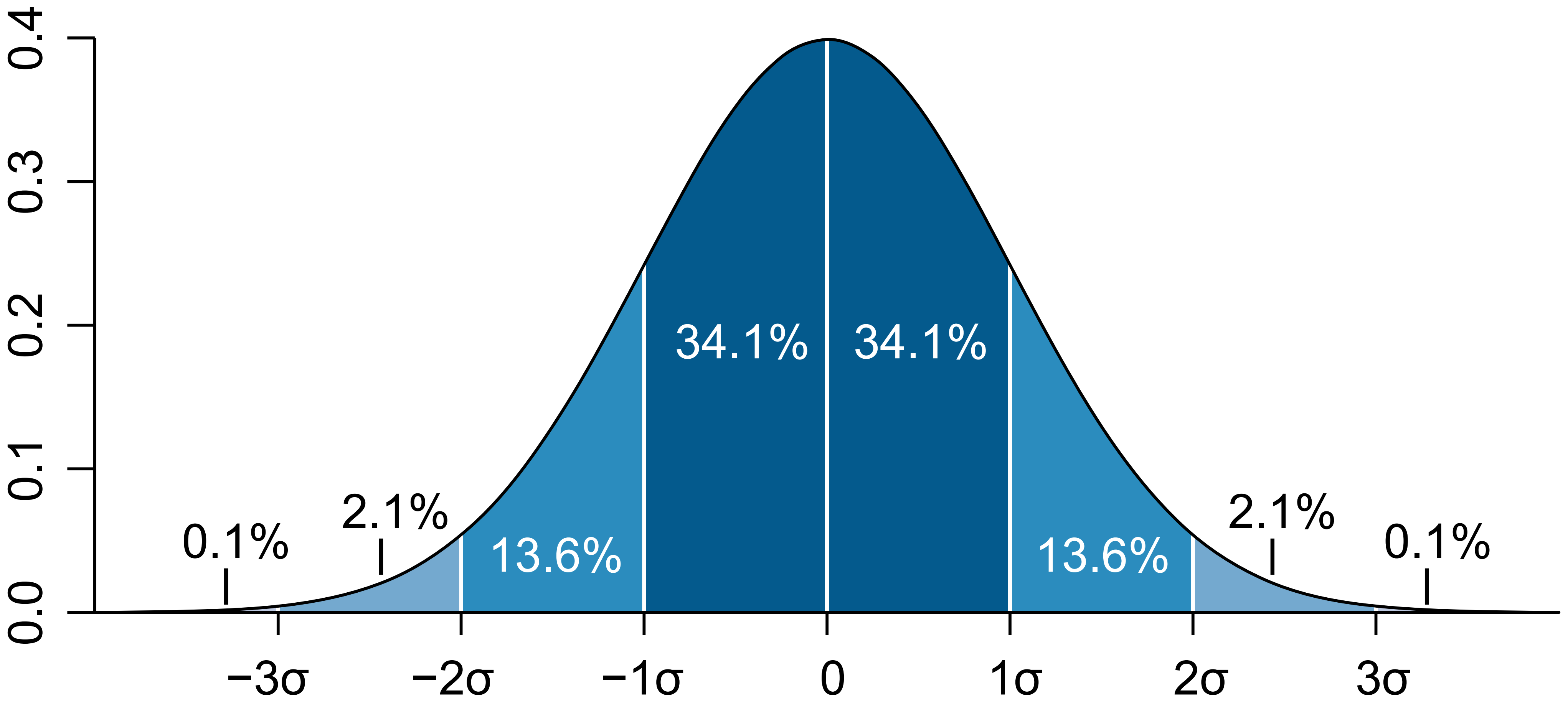


The Standard Error of the Mean (SEM)



SEM = SD/square_root(n)

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?

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?

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?

Why?

What are the alternatives?

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?

What are the alternatives?

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?

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

What we have covered

- Raising a hypothesis
- Types of errors
- P-values
- Assumptions for testing
- Comparing two means

-> Jupyter Notebook

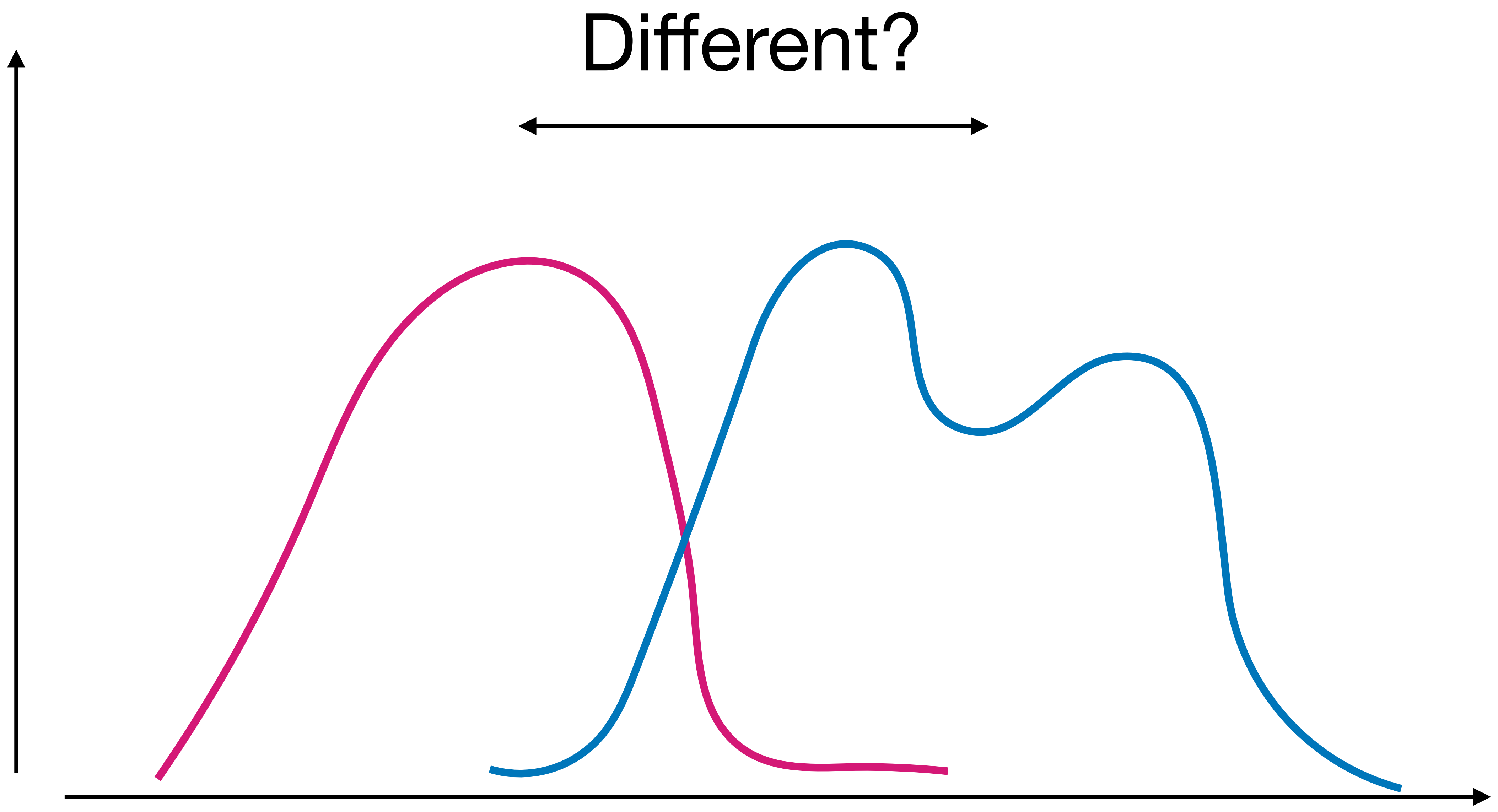
Non-parametric testing

- 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

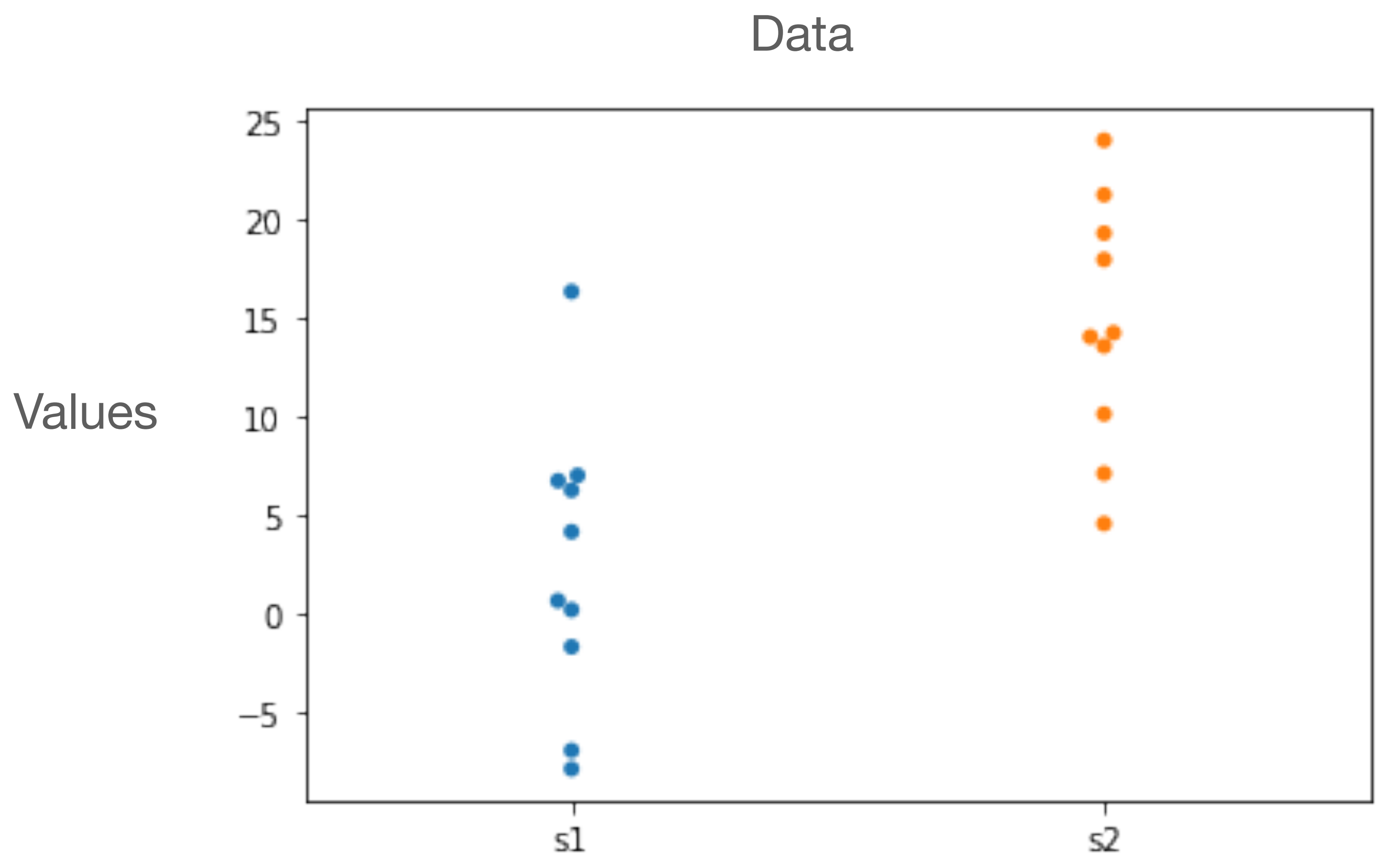
- Bootstrapping is a good alternative to rank based non-parametric tests, but it can get computationally very intense and they are not really custom in molecular biology (yet)

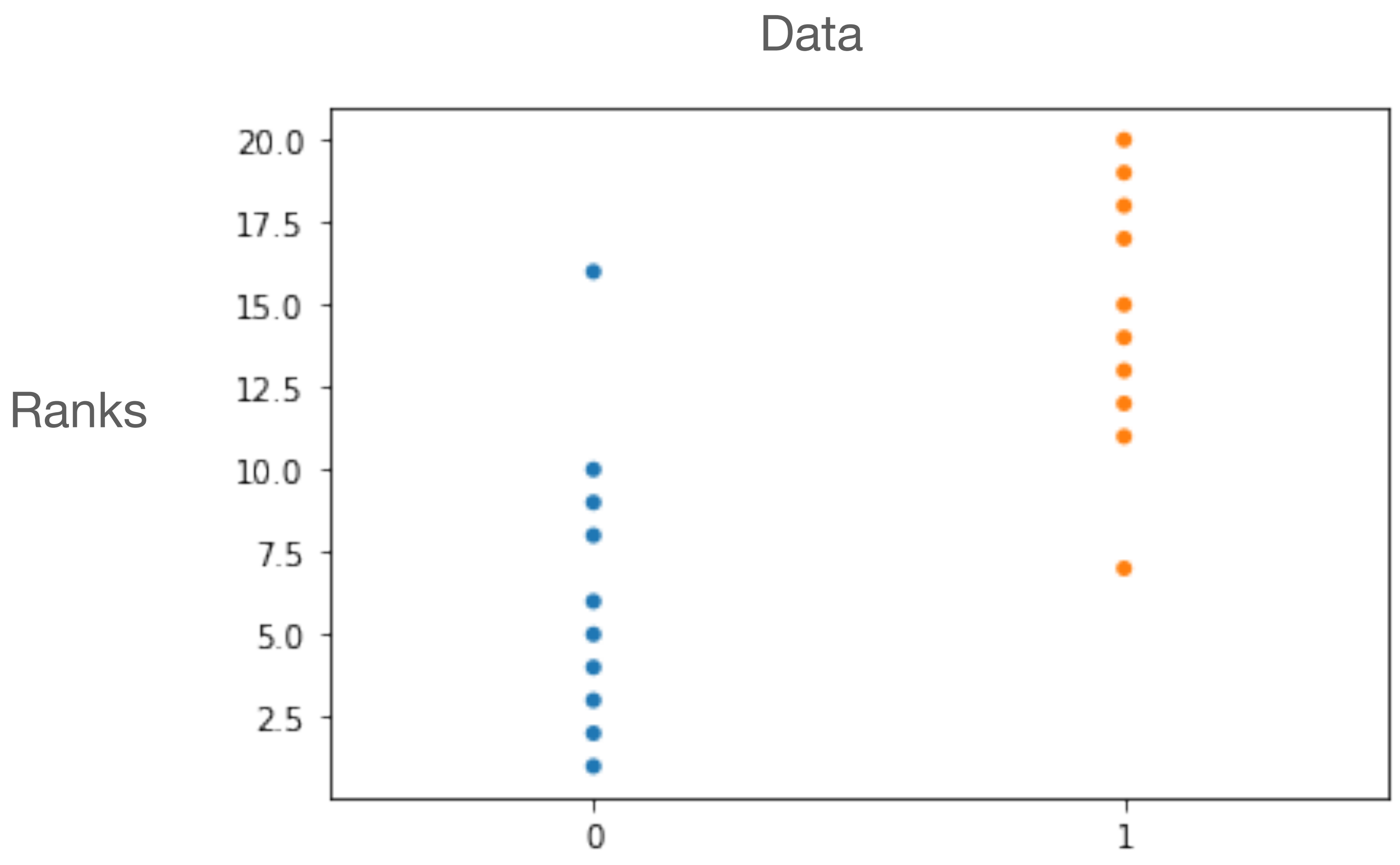
The test's names

- 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

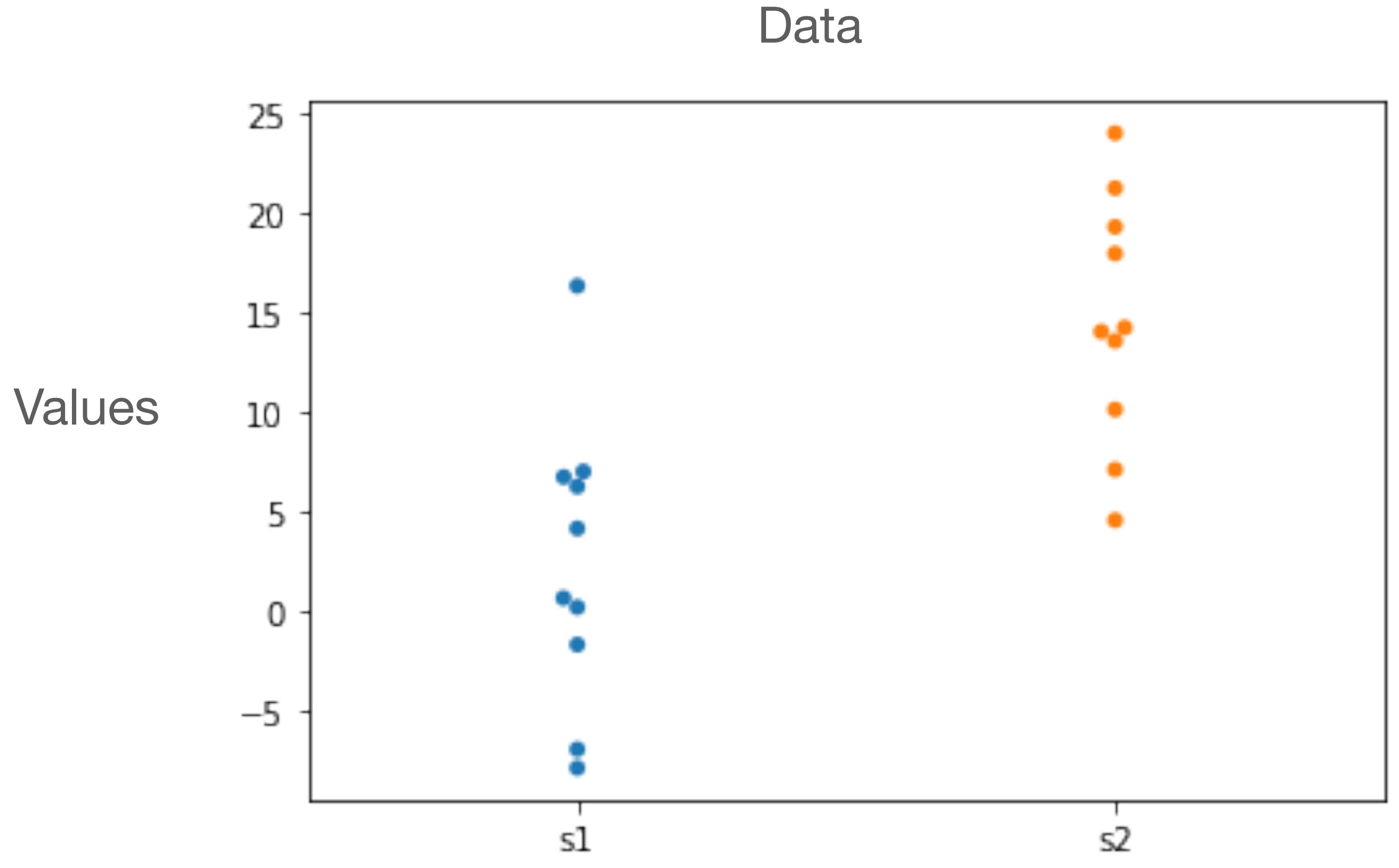


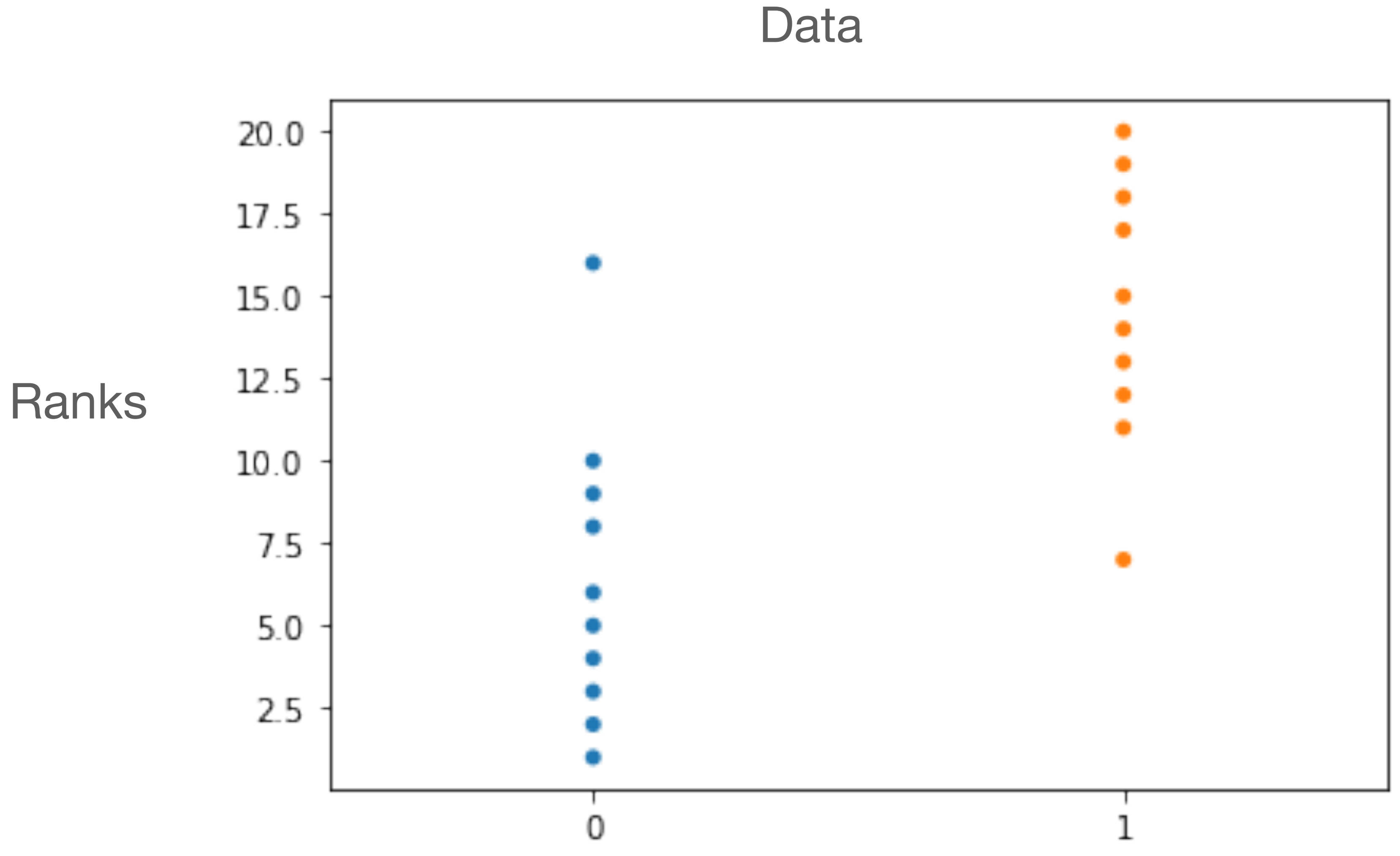


The absolute information is lost and only the ranks are compared.

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

Advantages of a rank based test

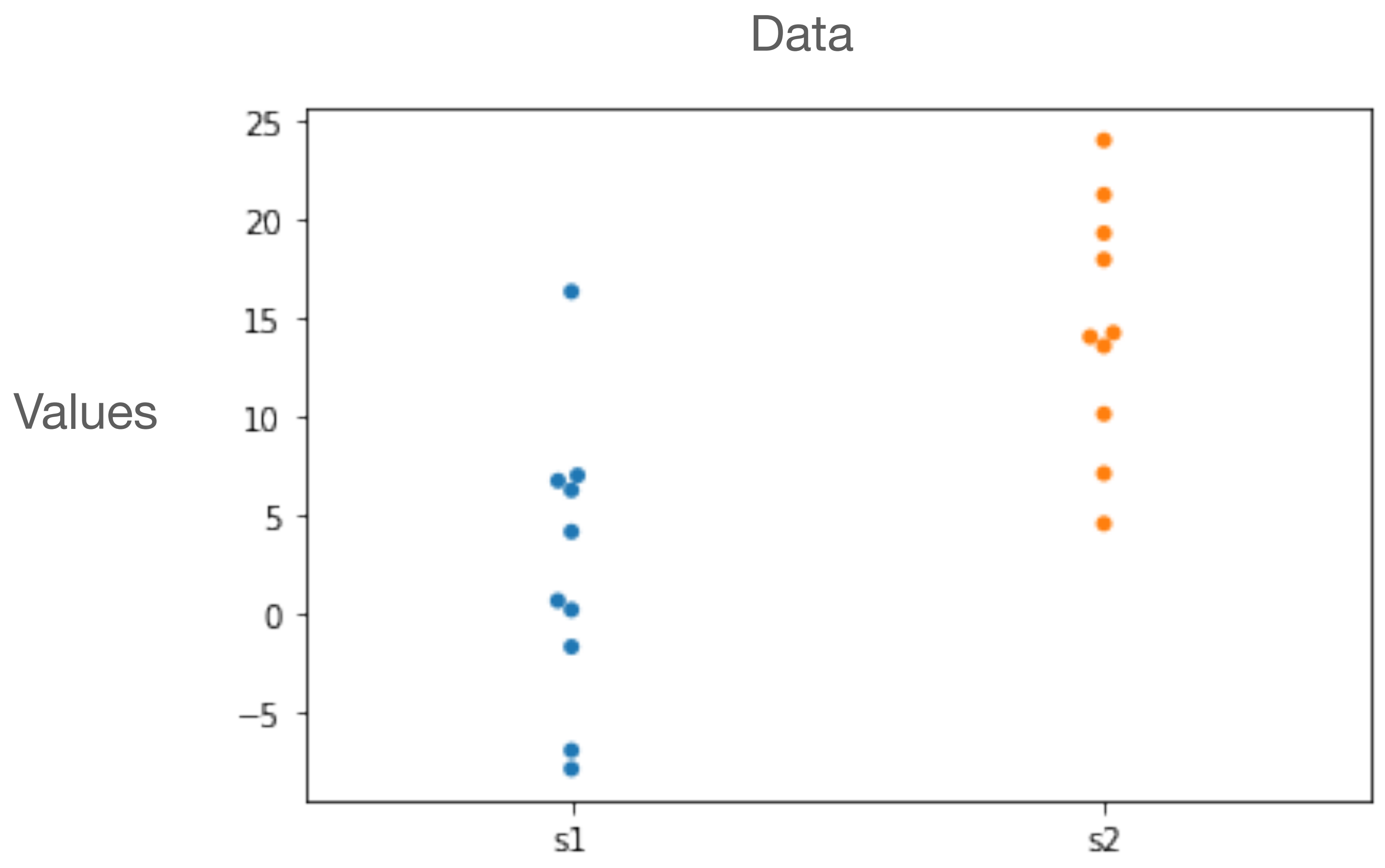


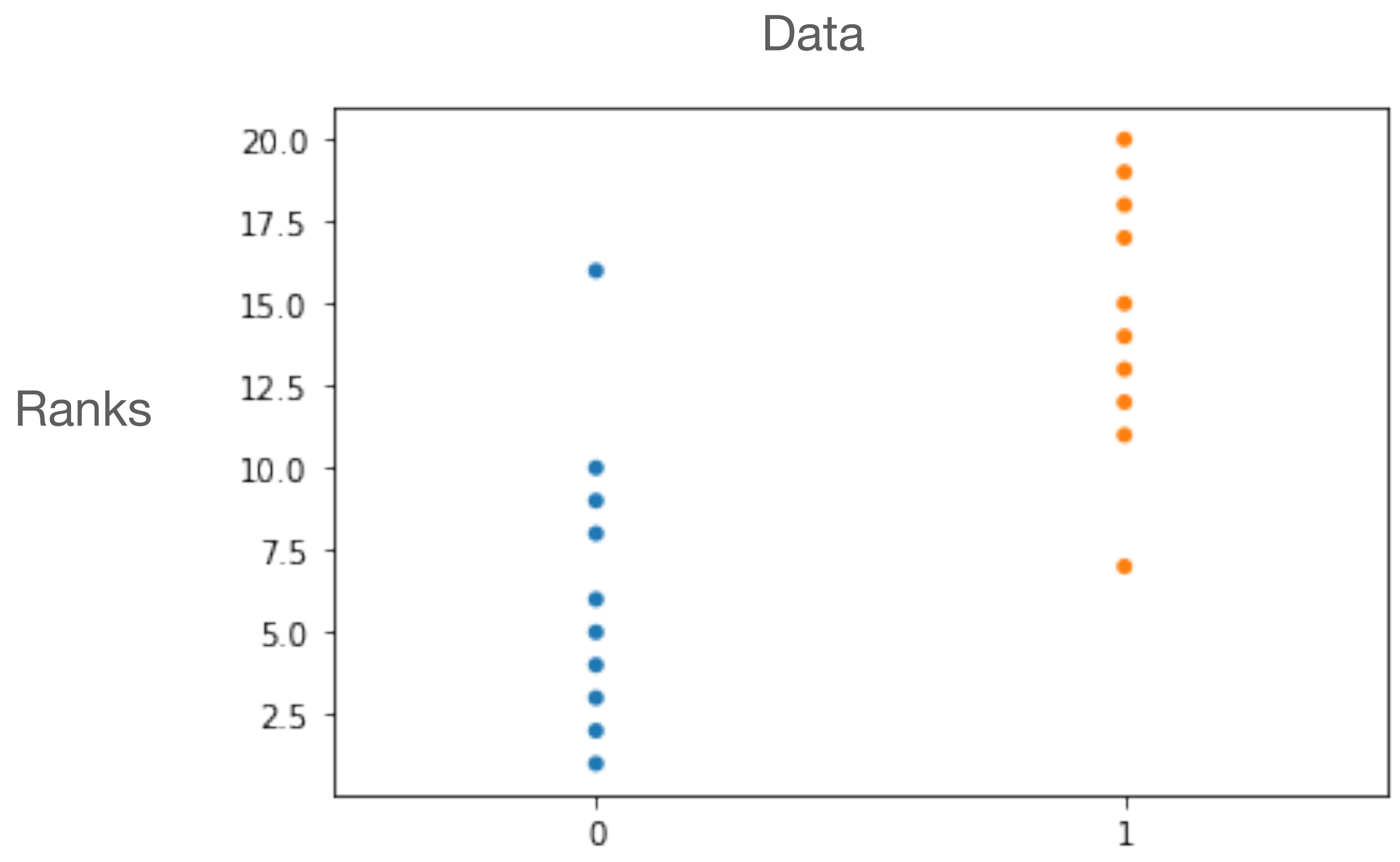


Outliers have limited influence on the outcome

We are not dependent on a distribution

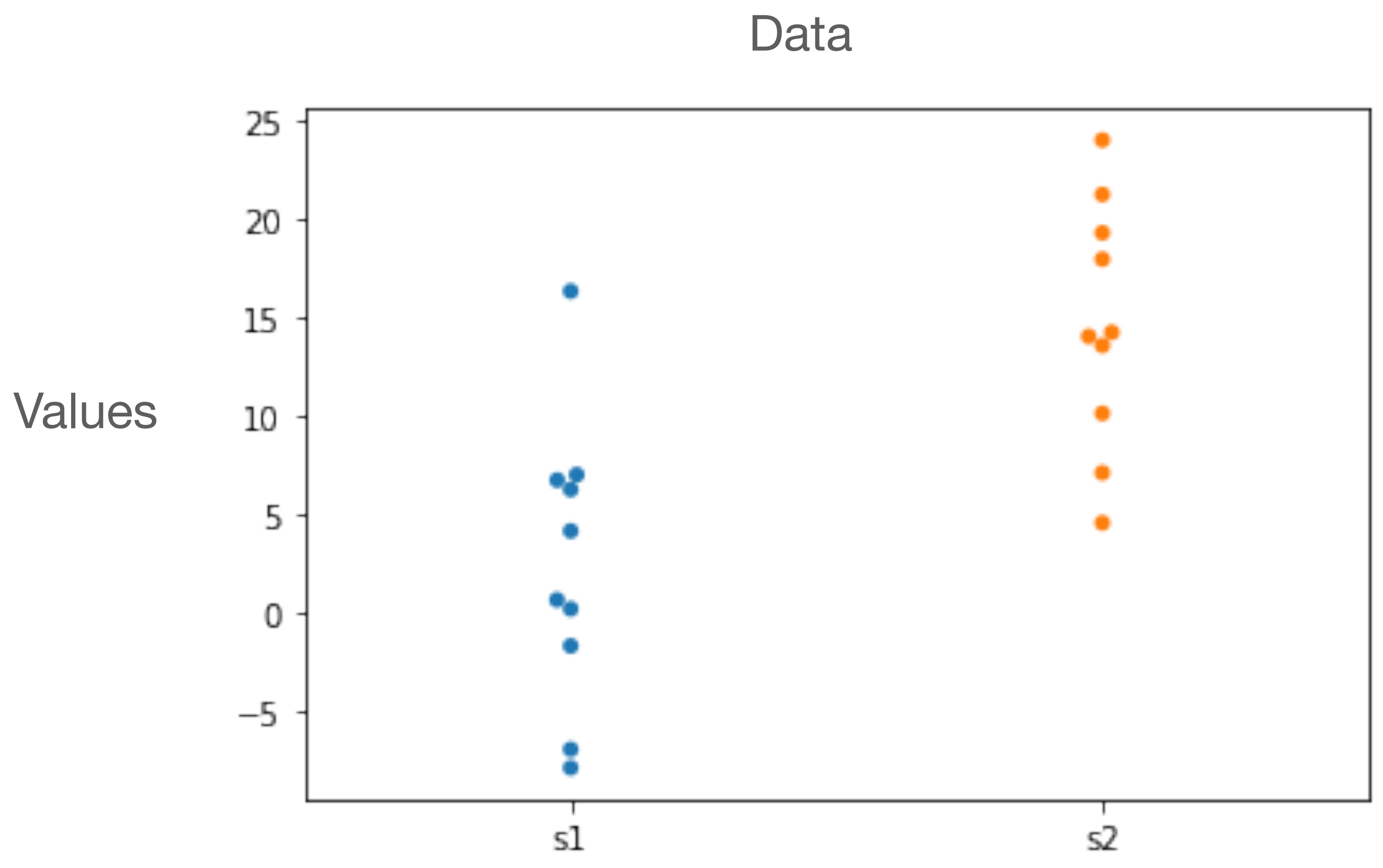
Disadvantages of a rank based test

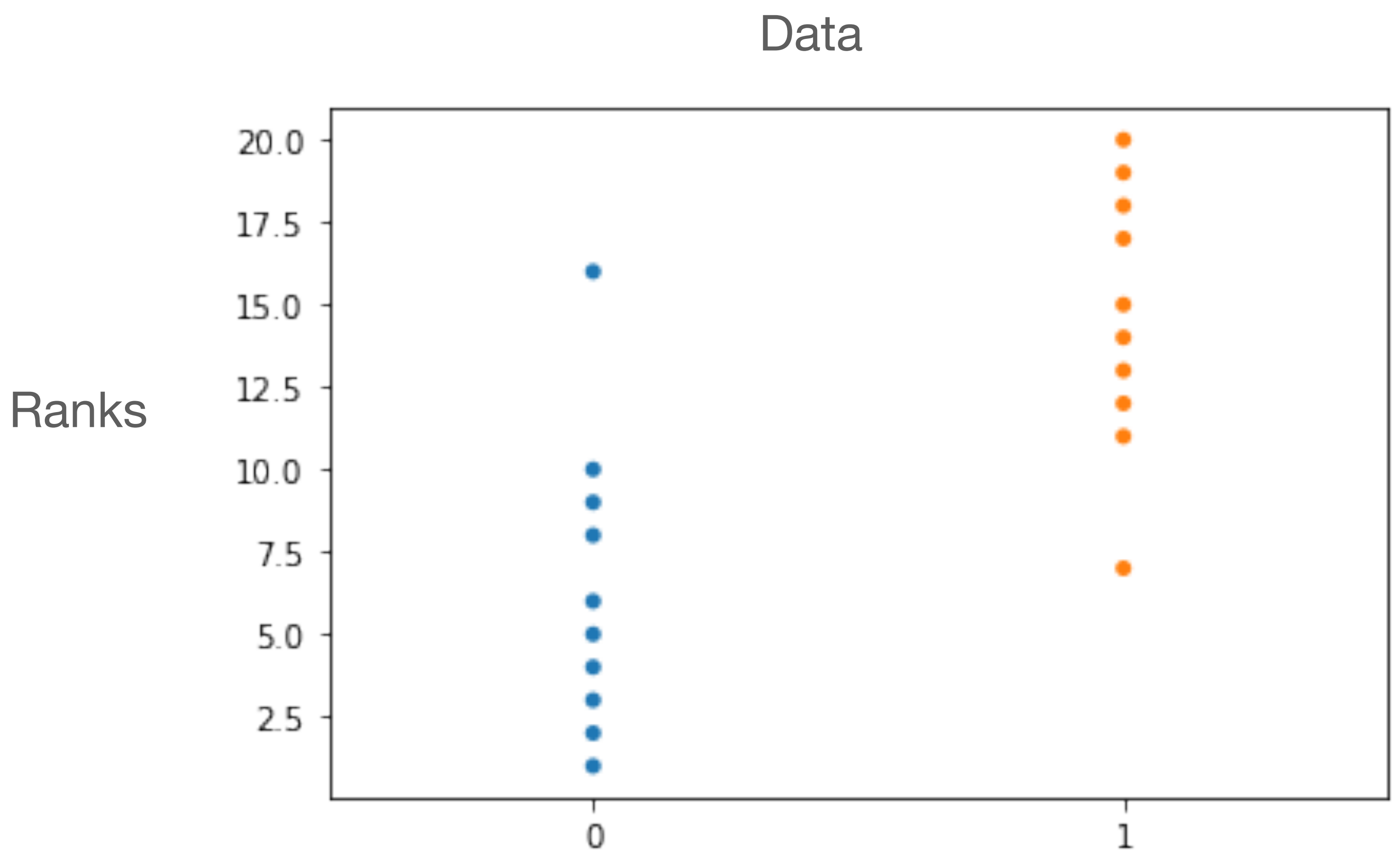




- We are loosing power
- Confidence intervals are more tricky
- Limited with more complex use-cases (regression models)

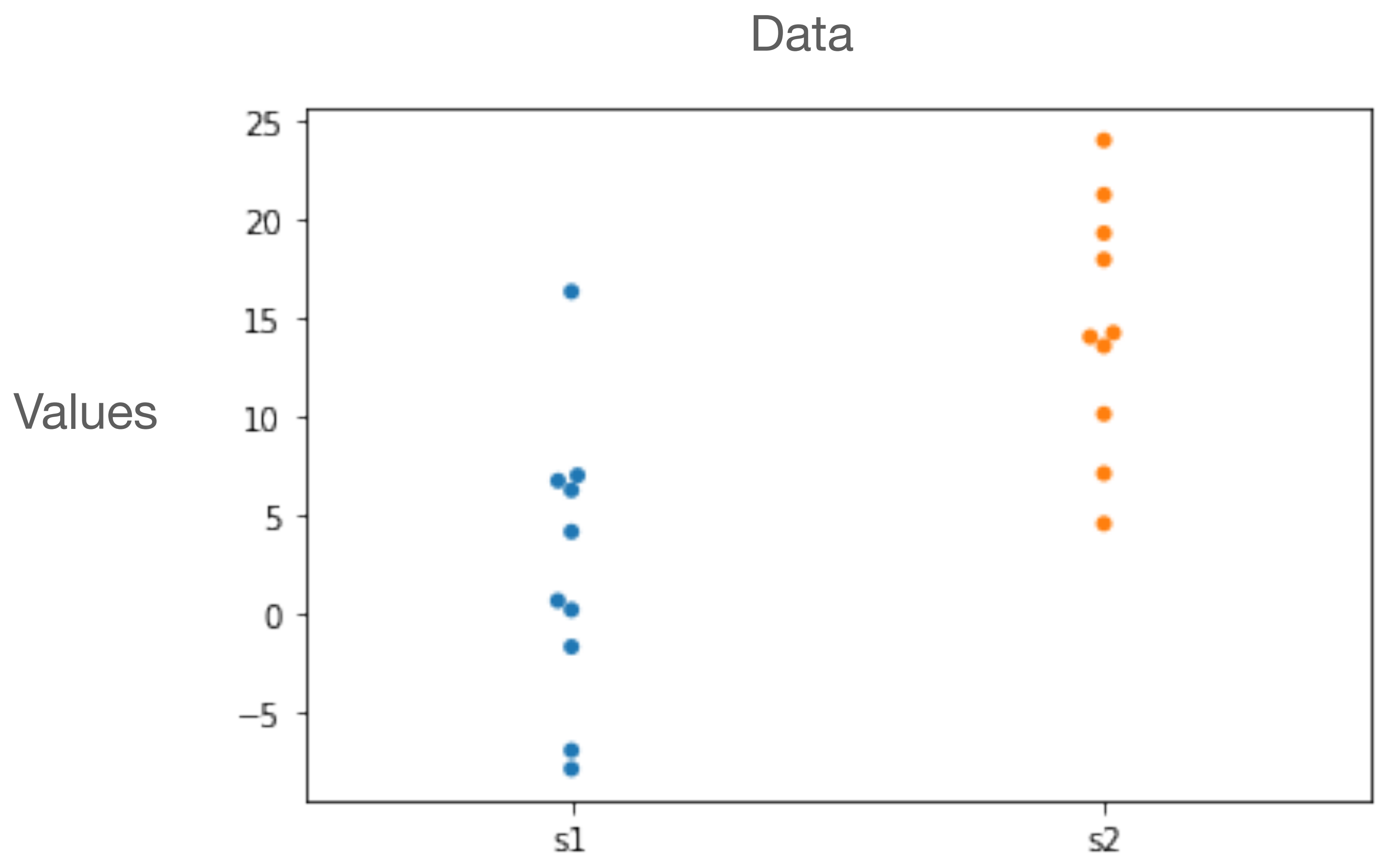
Assumptions

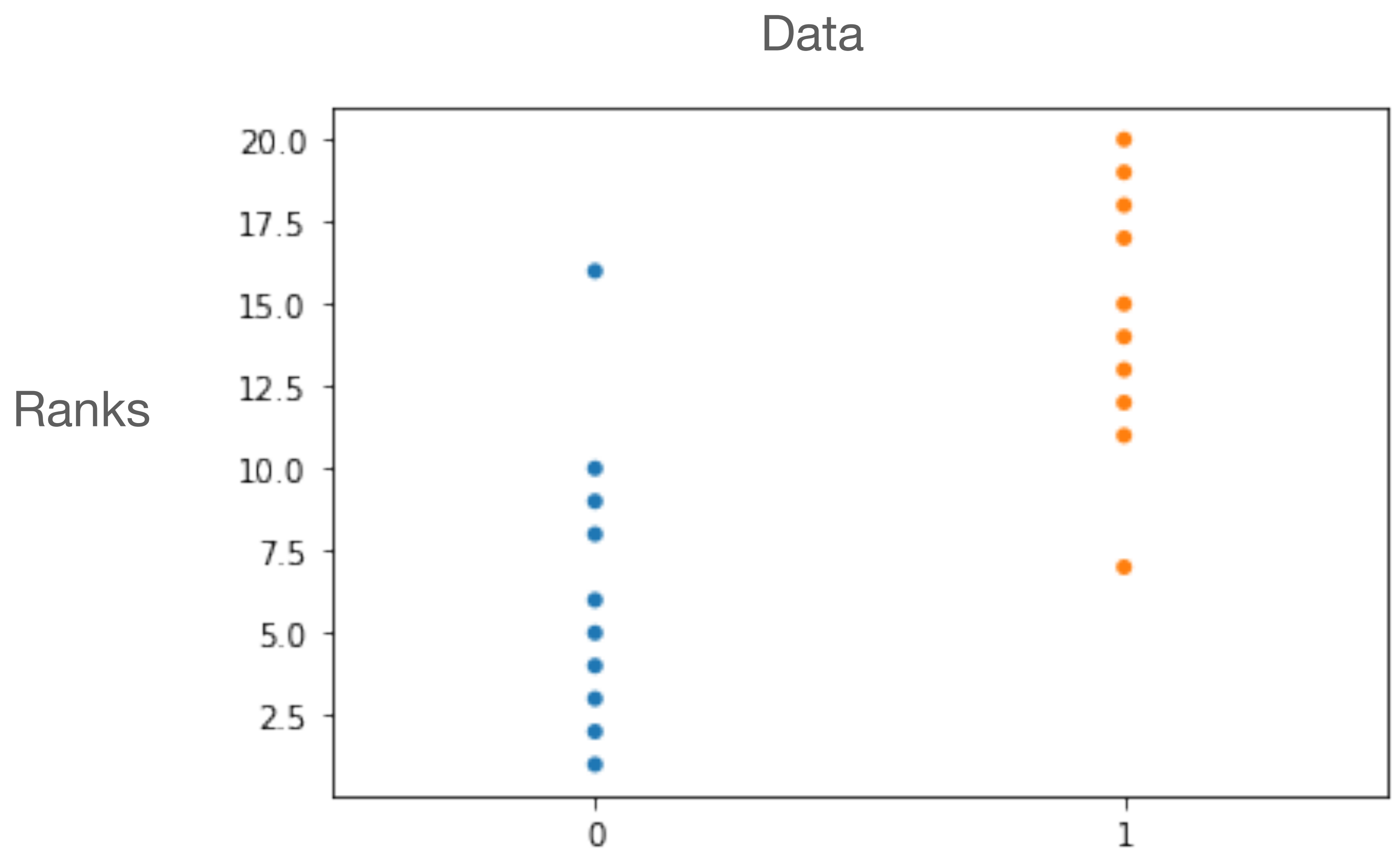




- Random sampling
- Each value is obtained independently

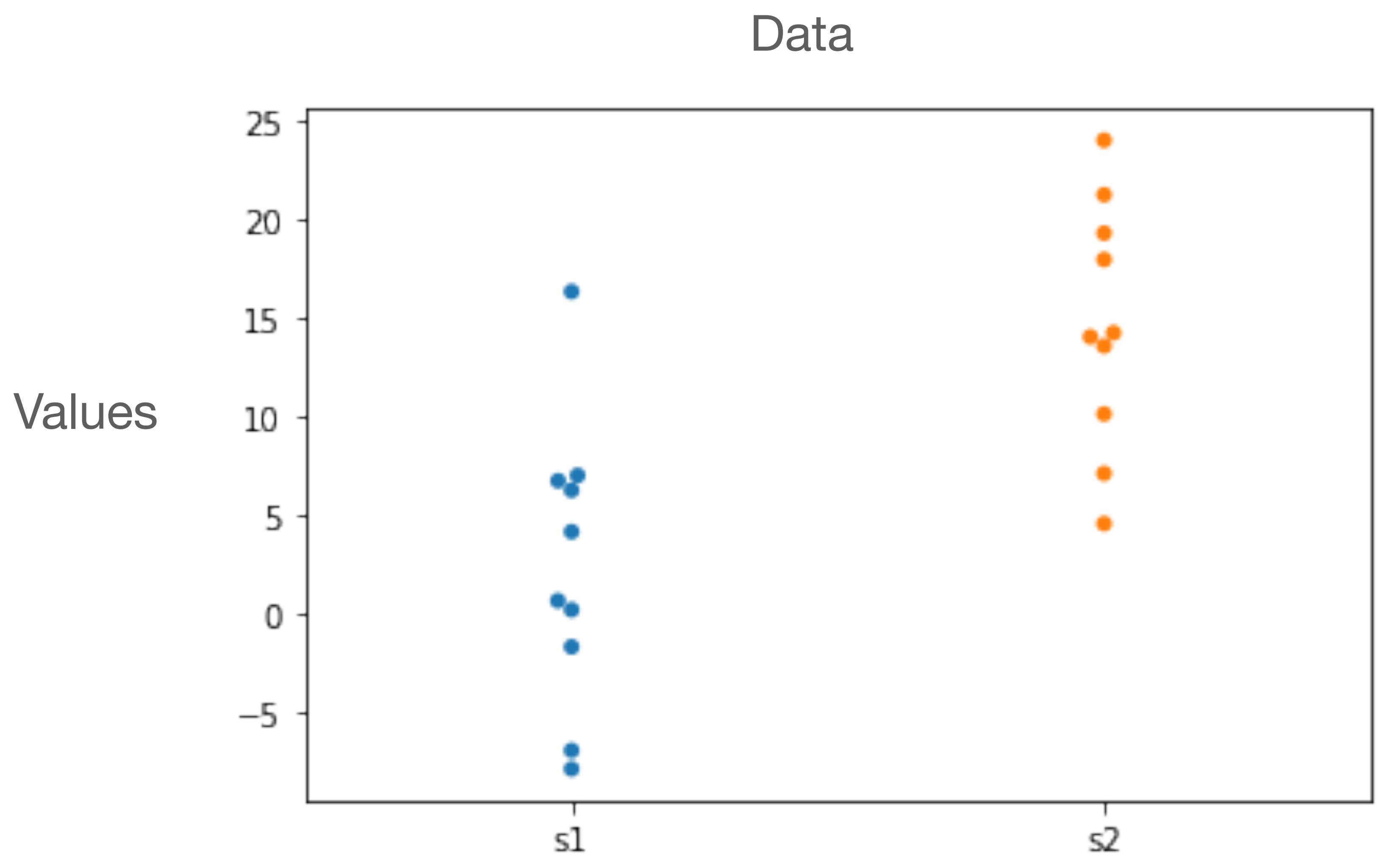
Sample sizes

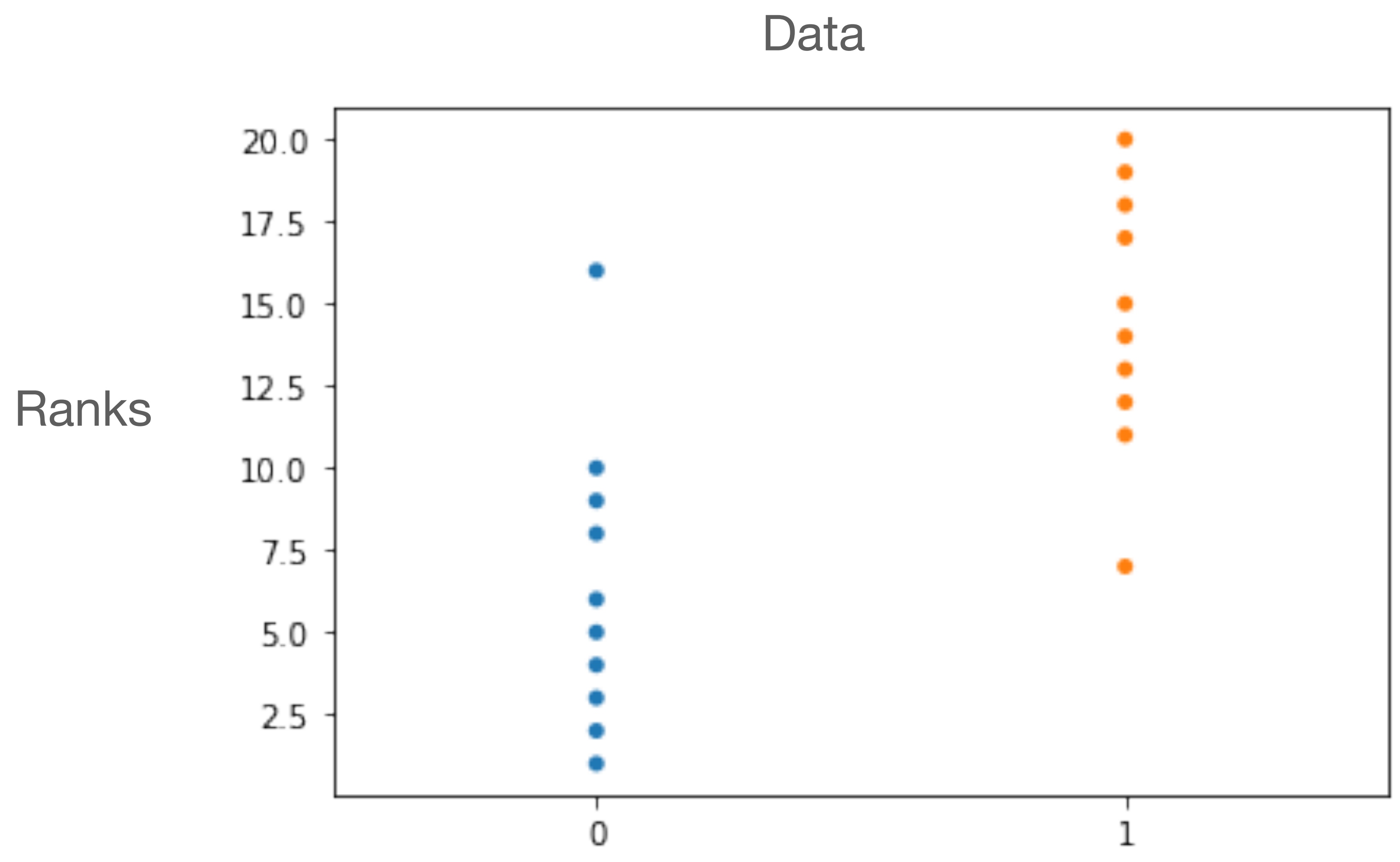




Do you think we need more or fewer samples for a non-parametric test?

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!!!!!!!!!!!!

What we have covered

- Non-parametric testing
- The advantages and disadvantages of ignoring a distribution
- The choice between a parametric and nonparametric test

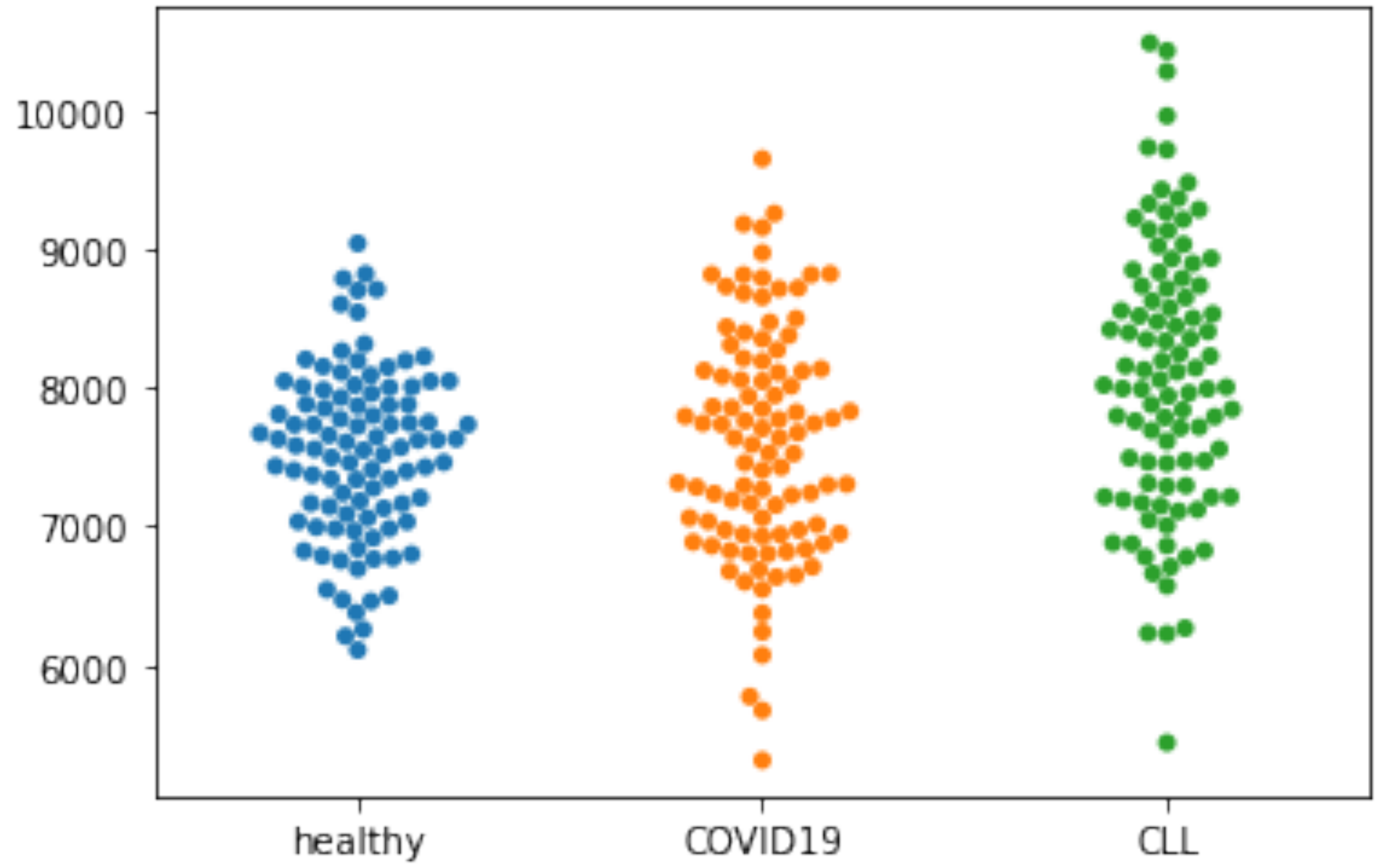
-> Jupyter Notebook

Multiple testing correction

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?

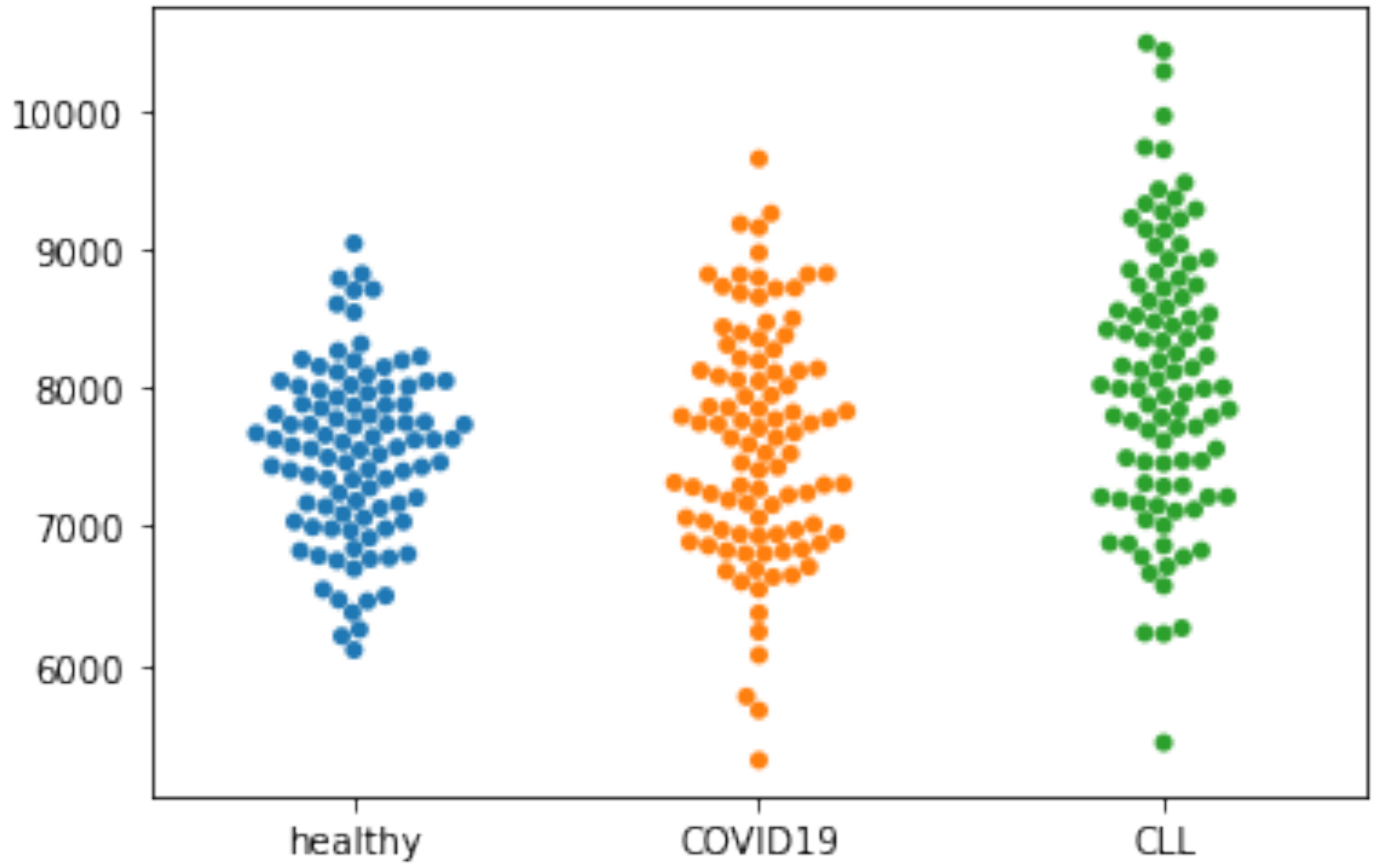
-> Analysis of Variance, one-way ANOVA (= multi-sample-t-test)

-> 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?

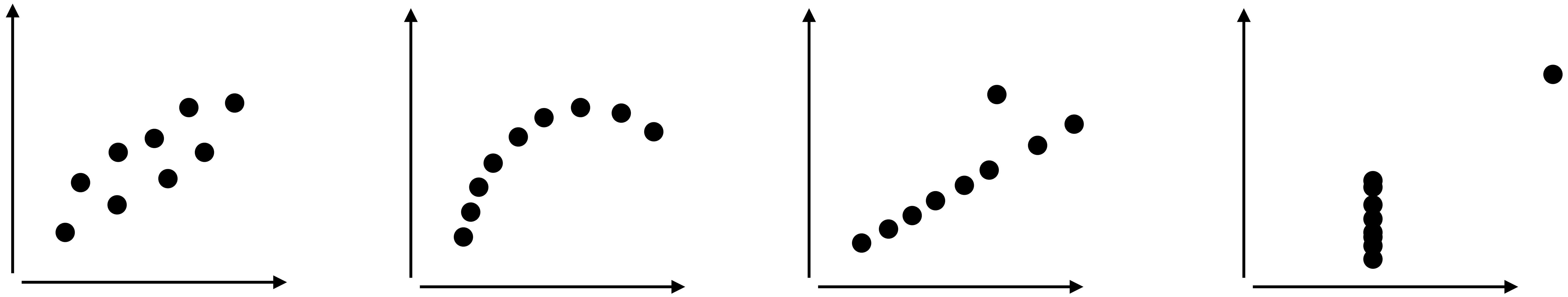
Correlations

What for?

To compare paired data in a population.

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

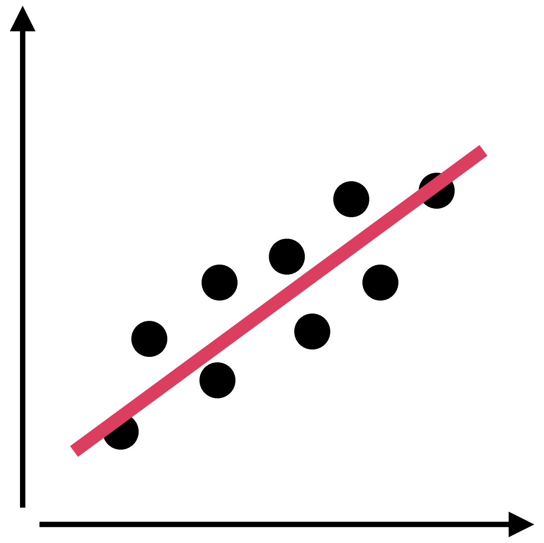
Main rule for any correlation analysis: **Look at your data first!**



These would all roughly have the same correlation coefficient!

Correlations

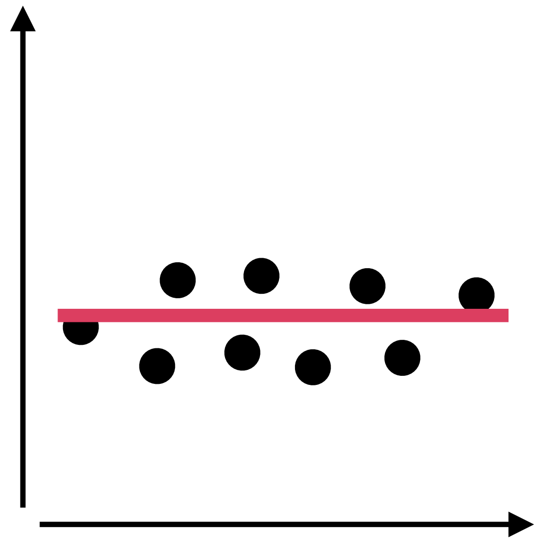
Positive



R = 0.7

p = 0.01

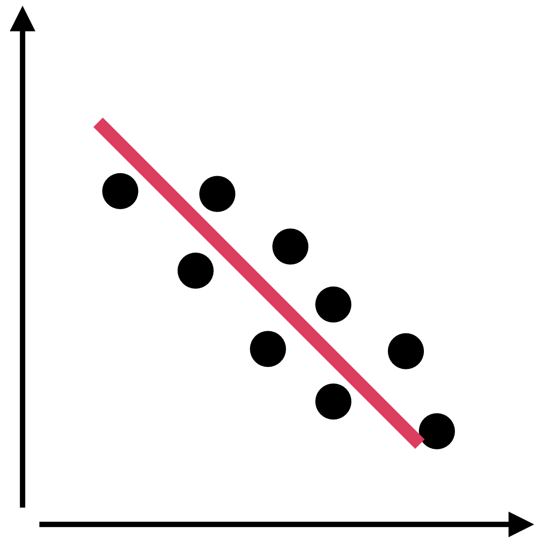
None



R = 0.05

p = 0.01

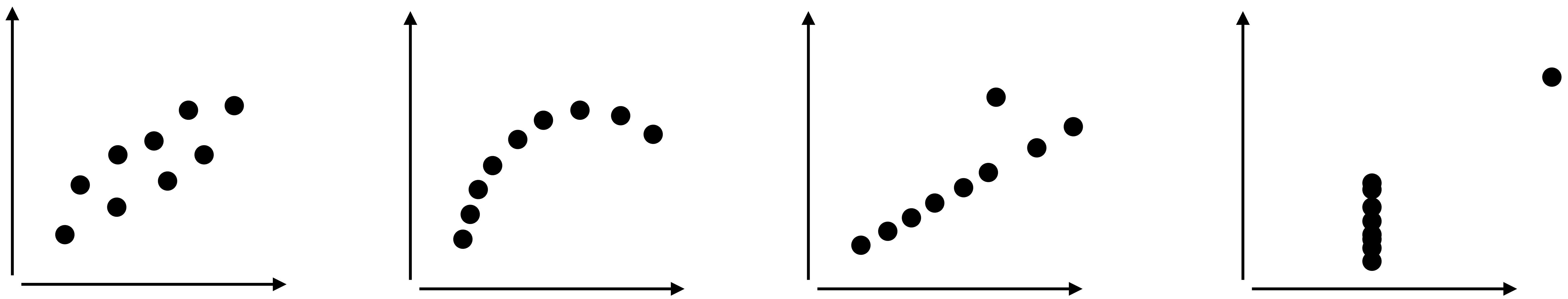
Negative



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

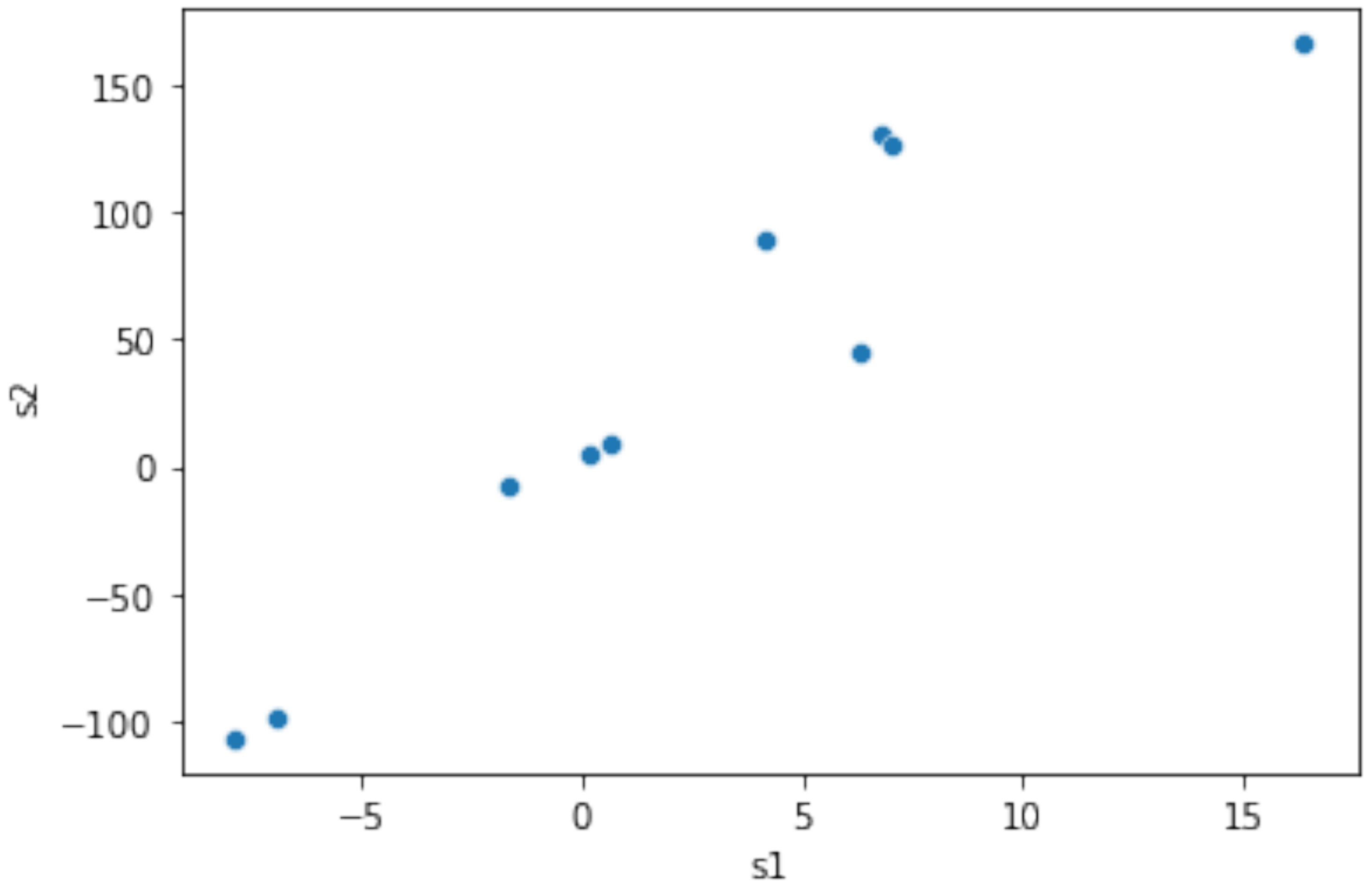
Specifically for parametric:

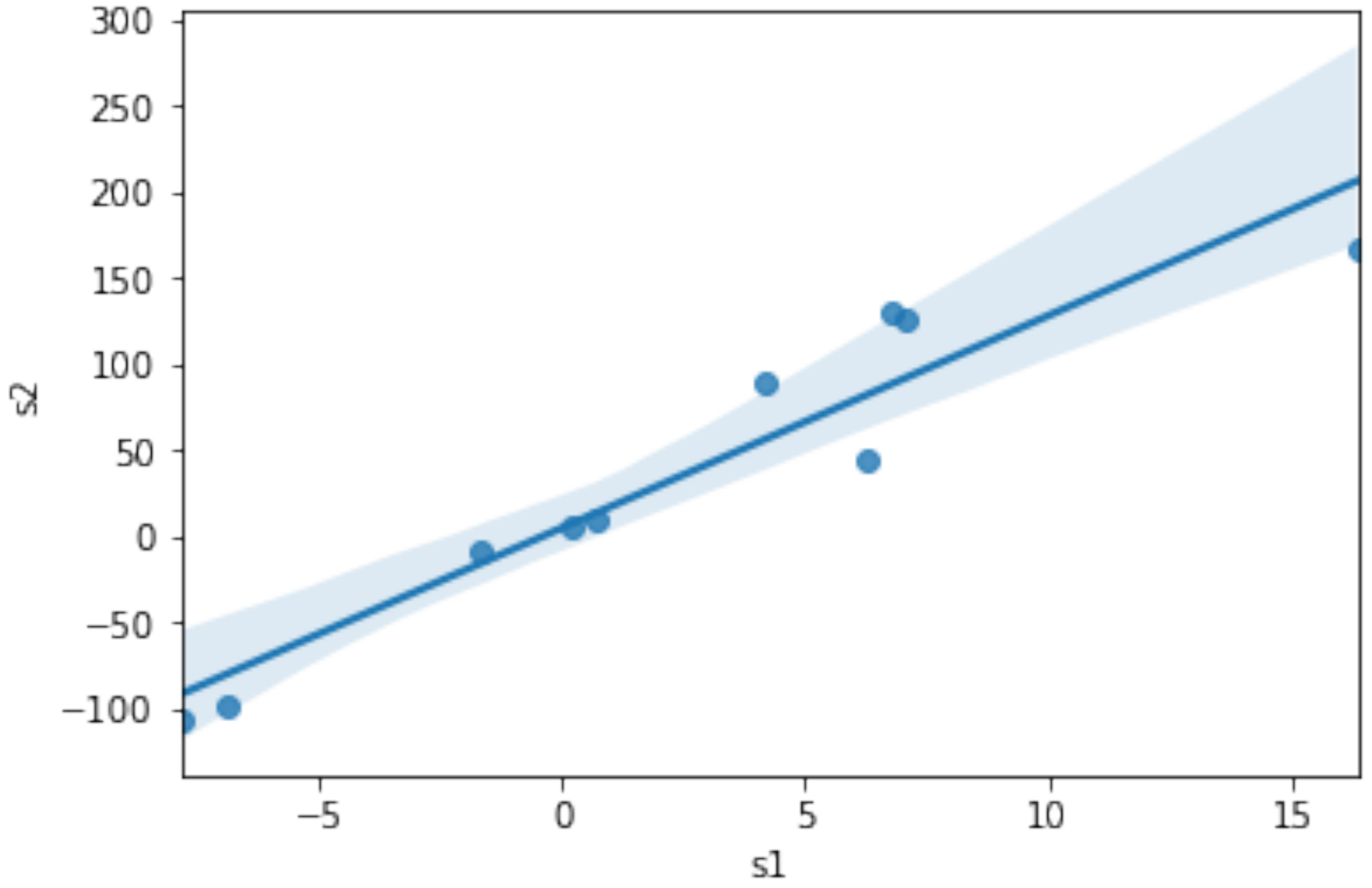
- Approximate normal distribution
- All covariation is linear
- No outliers !!!!

Pearson Correlation

With regression line and

confidence interval

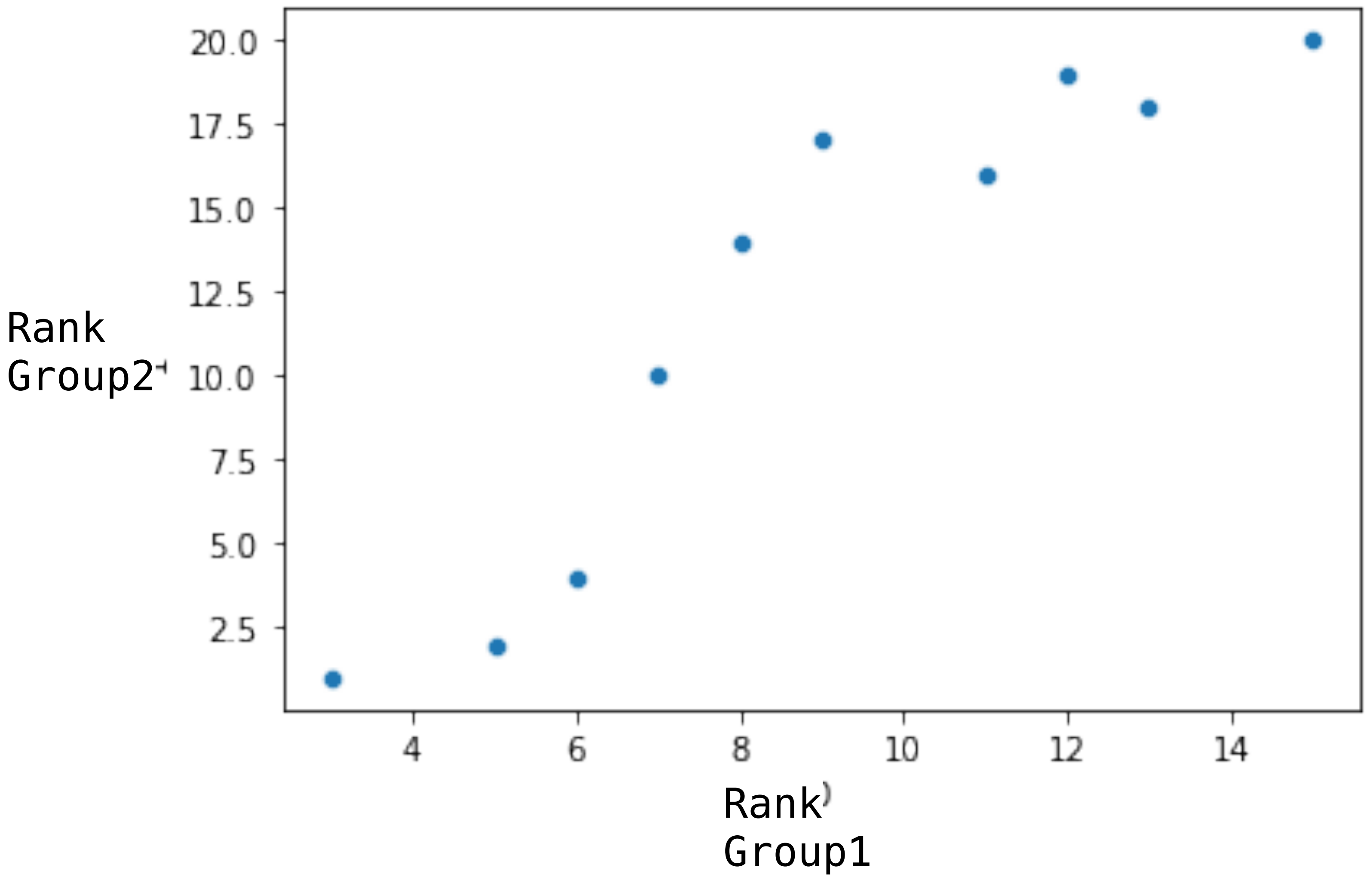




Parametric correlation statistics

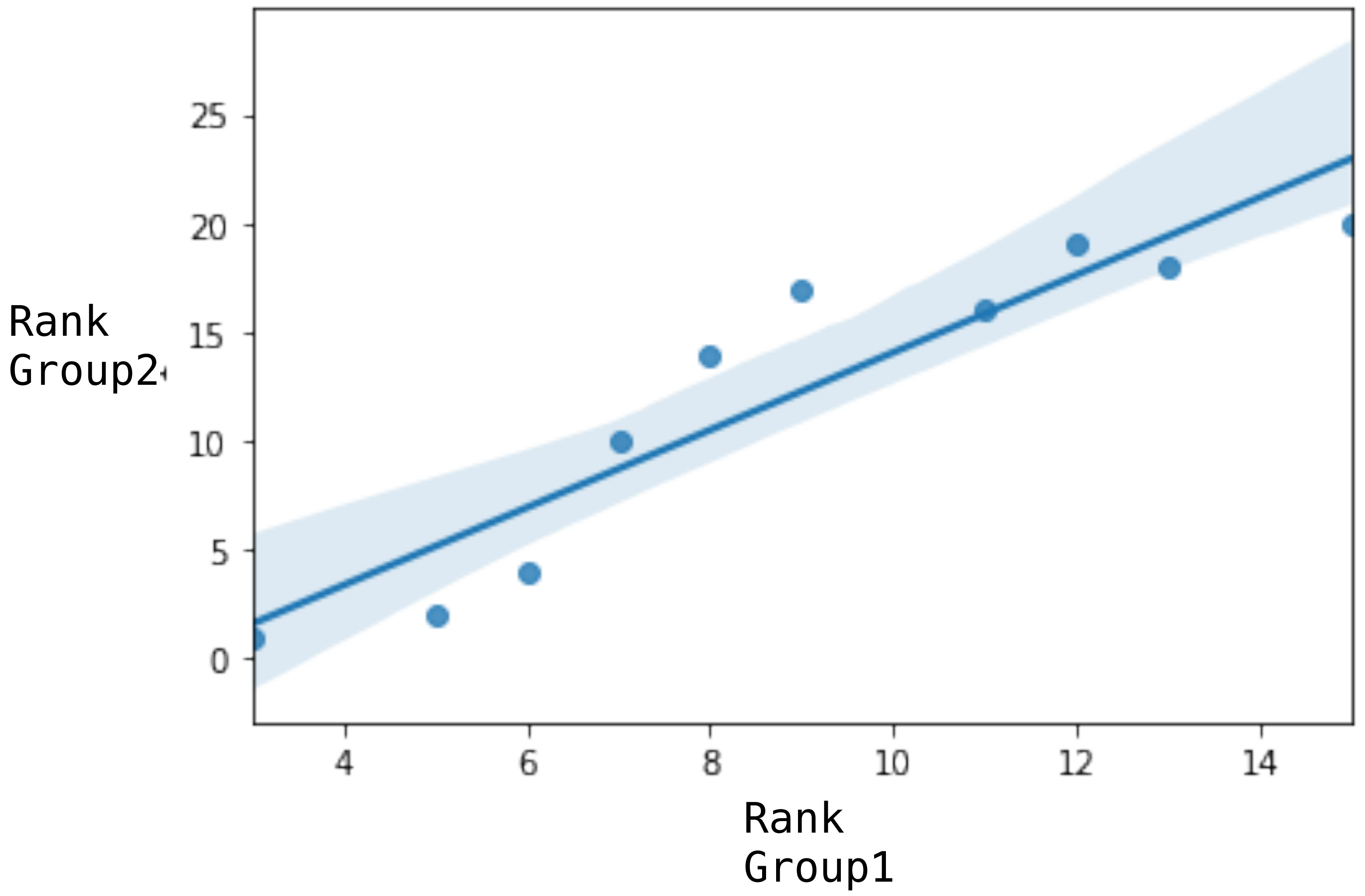
```
R = 0.95
p = 2.6e−05
```

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