Introduction into Biostatistics

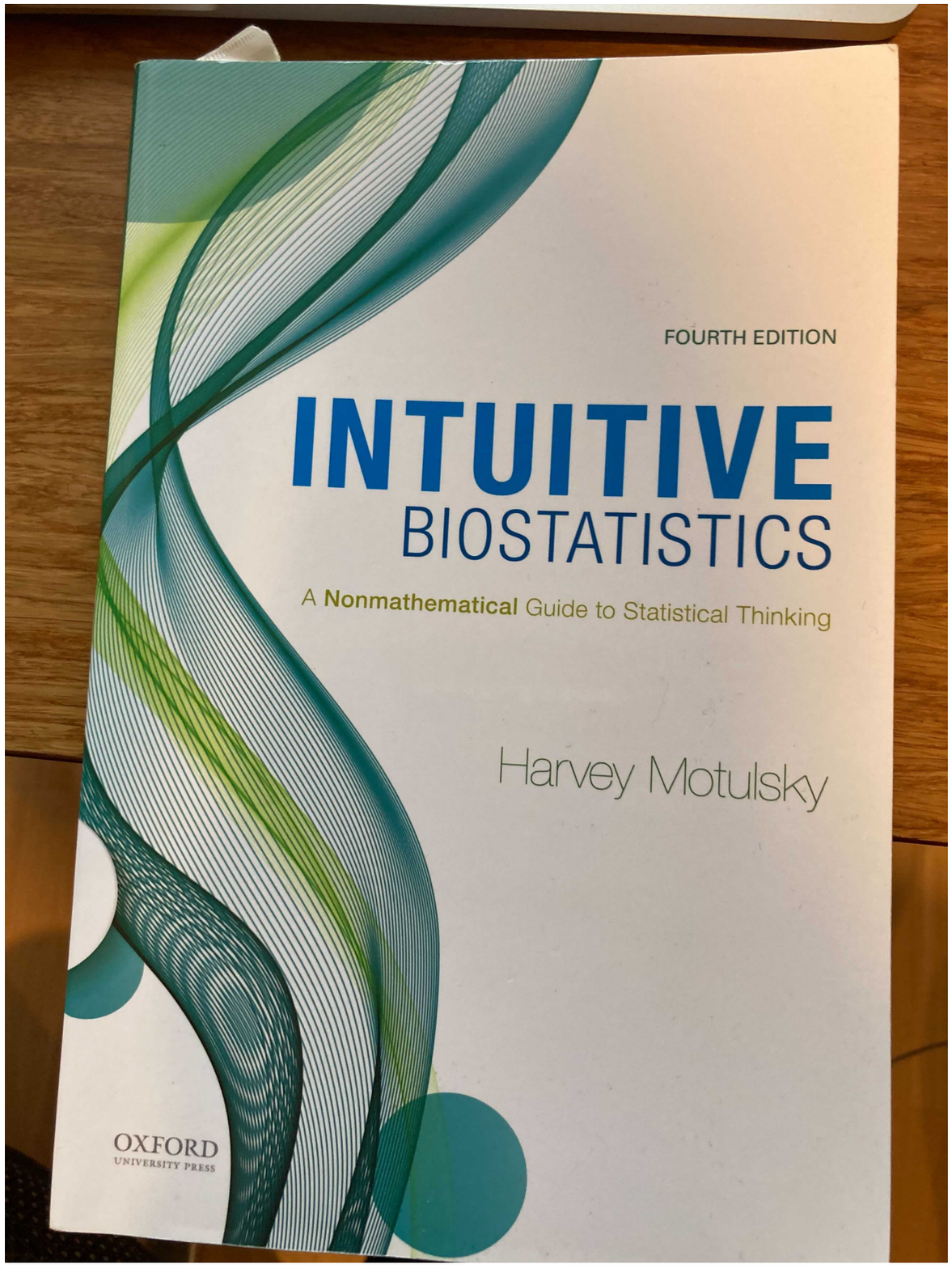
Anna Poetsch, Biotechnology Center, TU Dresden

Organisation

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

by Melissa Sanabria

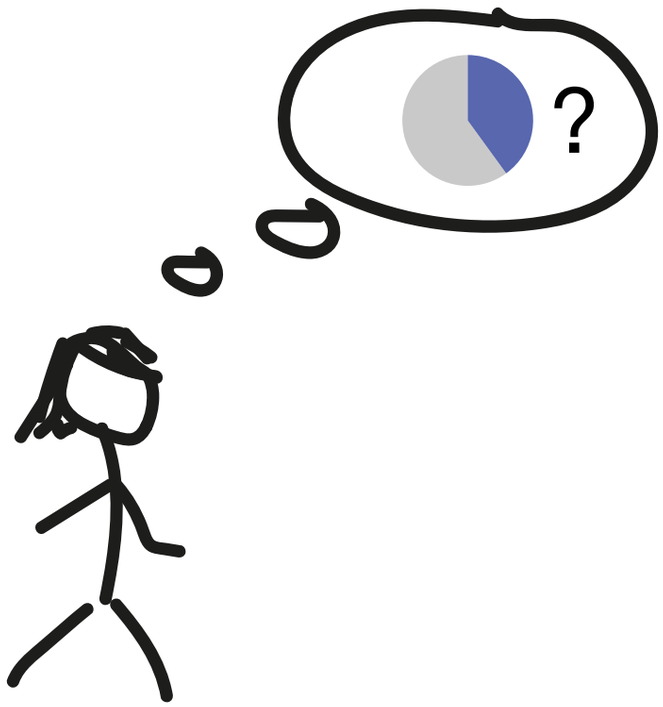
Sources

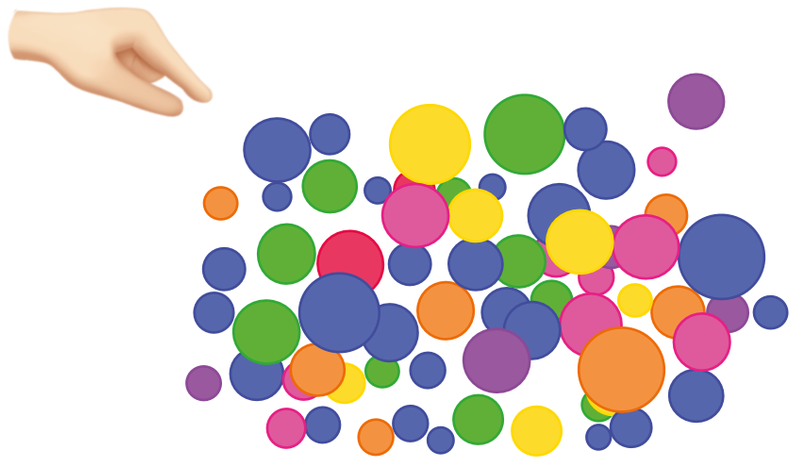


& “the internet”

Recap on confidence of one probability measurement

- A random (or representative) sample!
- They are independent observations!
- The data are accurate!





Recap on confidence of one probability measurement

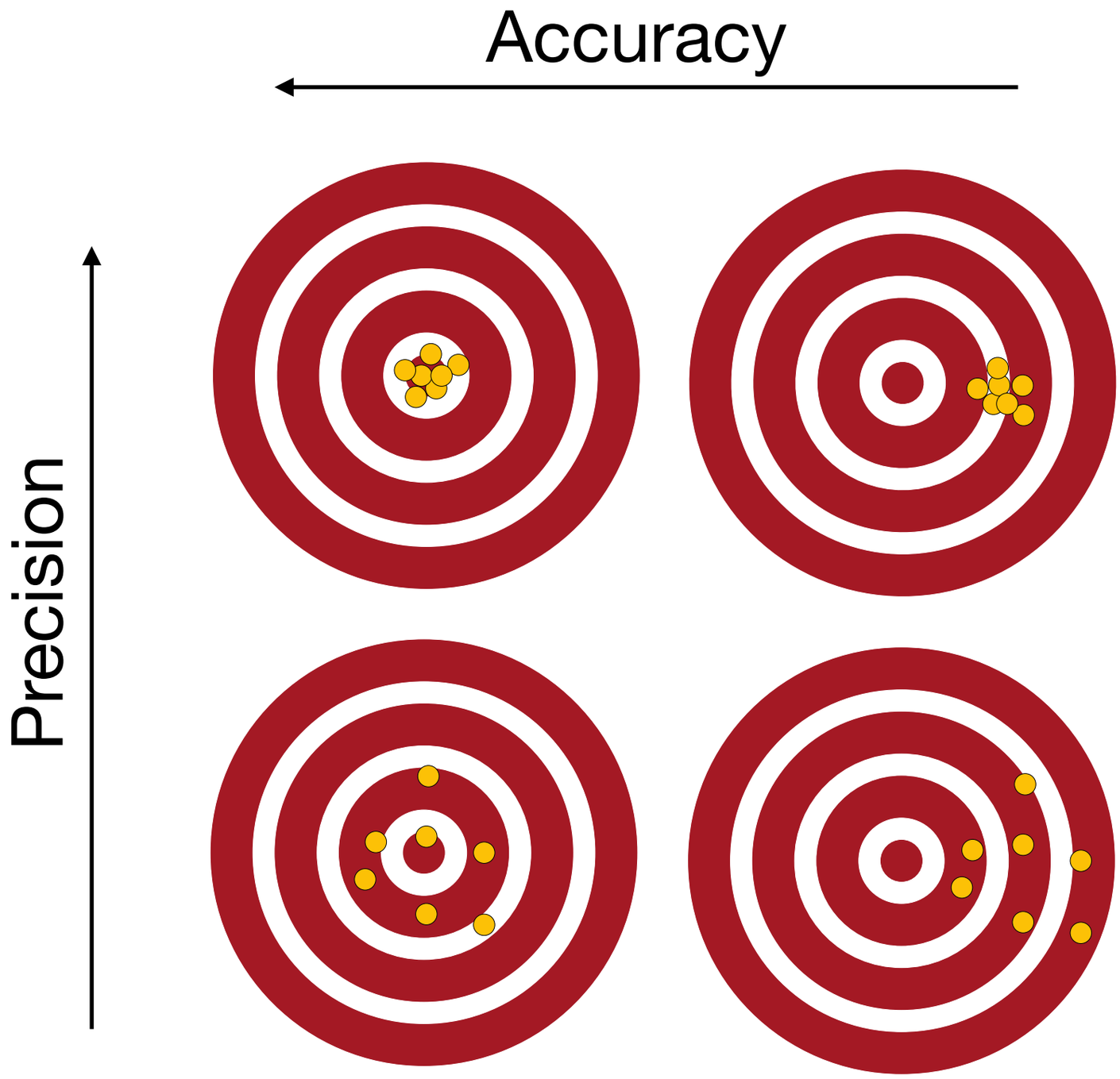
Please answer in the chat:

Does the confidence interval get bigger, if you…

… increase n?

… increase the confidence level, e.g. from 95% to 99%?

Recap on accuracy and precision

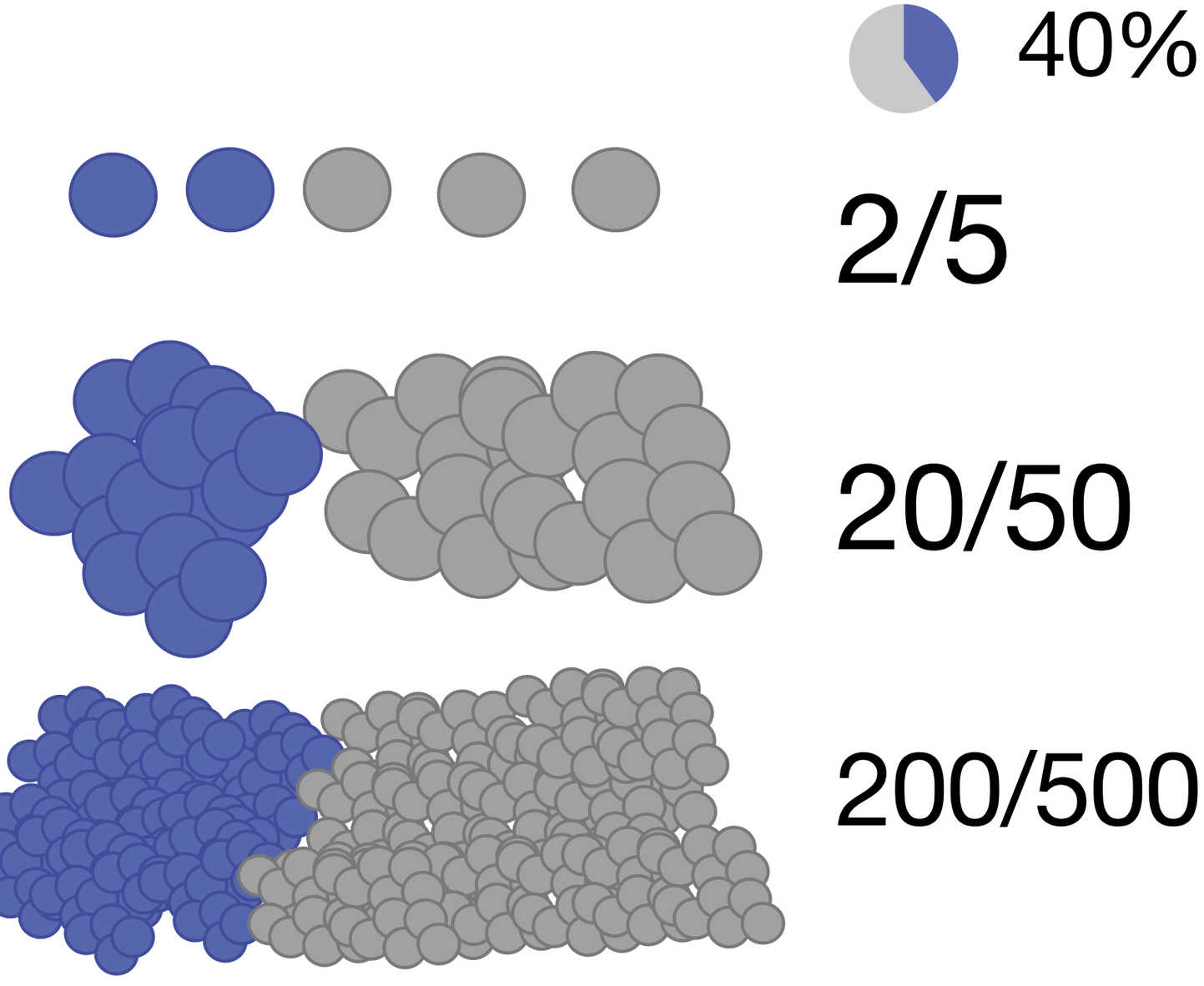


How do these relate to confidence intervals?

Does the confidence interval get bigger, if you increase n?

Know your problem, know your distribution!

Binomial



40%

2/5

20/50

200/500

Confidence increases with n

Normal

Confidence does not increase with n

Descriptive statistics

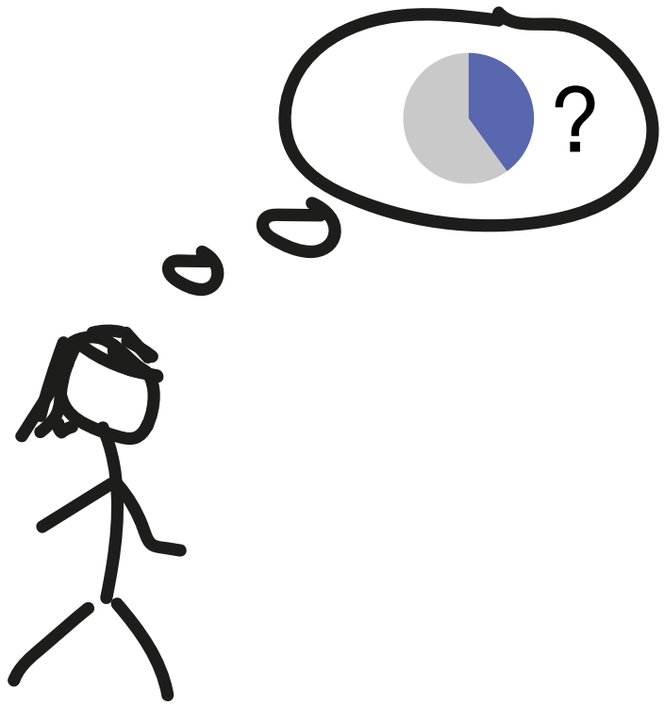
Data types and distributions

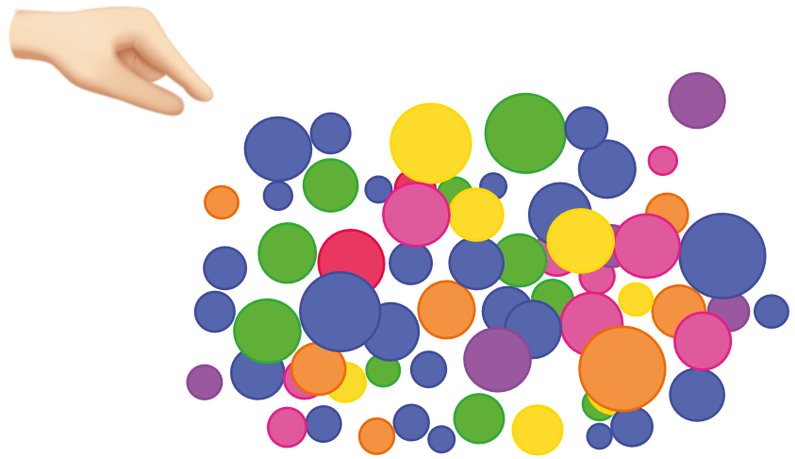
- Probability data (binomial distribution)
- Counted data (Poisson distribution)
- Normal distribution

Summary statistics

- Min, max, mean
- Mode
- Median and quartiles
- Confidence intervals

Probability data/ Binominal distribution





-> Jupyter Notebook

Counted data/ Poisson distribution

- Radioactive decay
- Raisins in a Dresdner Stollen
- Mutations in a genome

-> Jupyter Notebook

Discrete variables

Ordinal variables

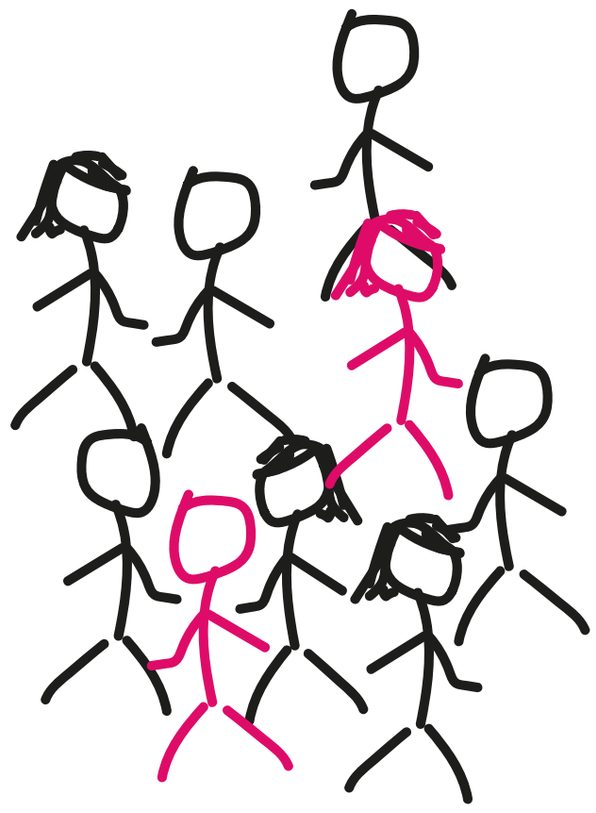
- limited set of discrete values with order

e.g. scale from 1-10

Nominal, binomial variables

- limited set of discrete values without order

e.g. responder <-> non responder



Continuous variables

Interval variables

- continuous value, for which intervals make sense, but no ratios

e.g. °C

Ratio variables

- continuous value, for which ratios make sense

e.g. height, weight, enzyme activity, Kelvin

Summary parameters

1 2 2 5 5 5 10 30

Summary parameters

1 2 2 5 5 5 10 30

Min value: 1

Max value: 30

Summary parameters

1 2 2 5 5 5 10 30

Min value: 1

Max value: 30

Parametric measures

Mean: (1+2+2+5+5+5+10+30)/8 = 7.5

Variance: (1+4+4+25+25+25+100+900)/8 = 90.57143

SD: square_root (variance) = 9.516902

SD = standard deviation = sigma

Summary parameters

1 2 2 5 5 5 10 30

Min value: 1

Max value: 30

Parametric measures

Mean: (1+2+2+5+5+5+10+30)/8 = 7.5

Variance: (1+4+4+25+25+25+100+900)/8 = 90.57143

SD: square_root (variance) = 9.516902

SD = standard deviation = sigma

non-parametric measures:

1 2 2 5 5 5 10 30

Ranks: 1 2 2 4 4 4 7 8

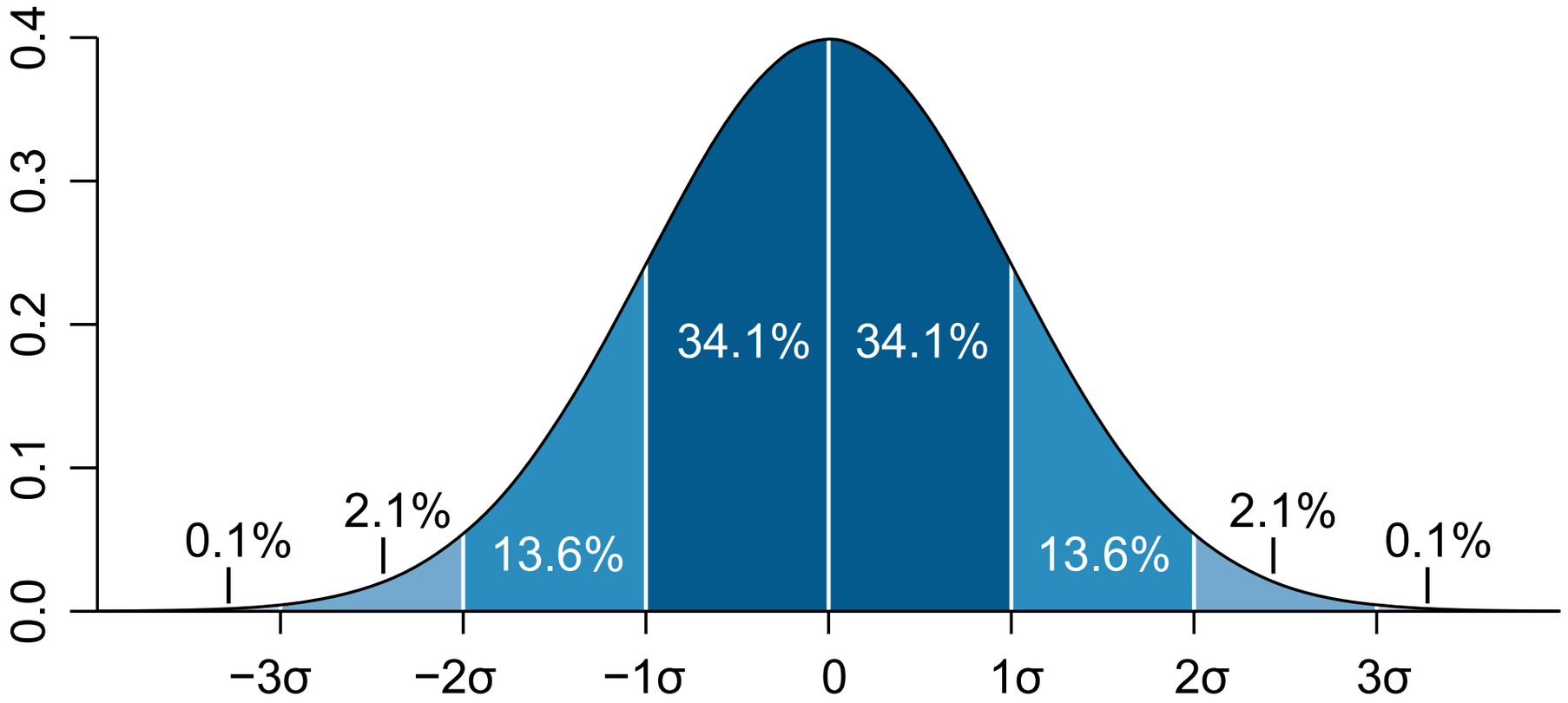
Median: the central value: 5

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

Summary

- Probability data (Binomial distribution)
- Count data (Poisson distribution)
- Categorical and continuous data types
- Normal distributions
- Describing a distribution (mean, median, standard deviation, mode, error)

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

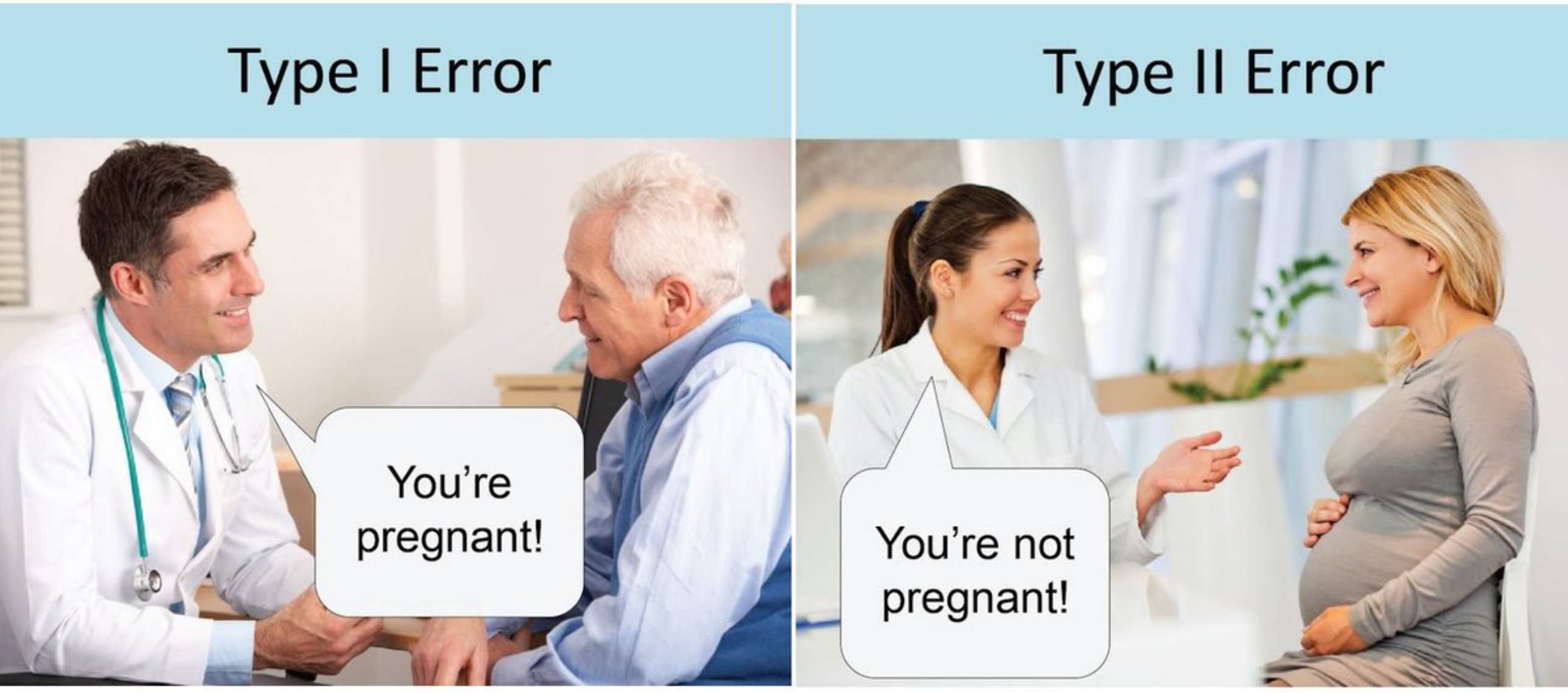
How to reject H_0

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.

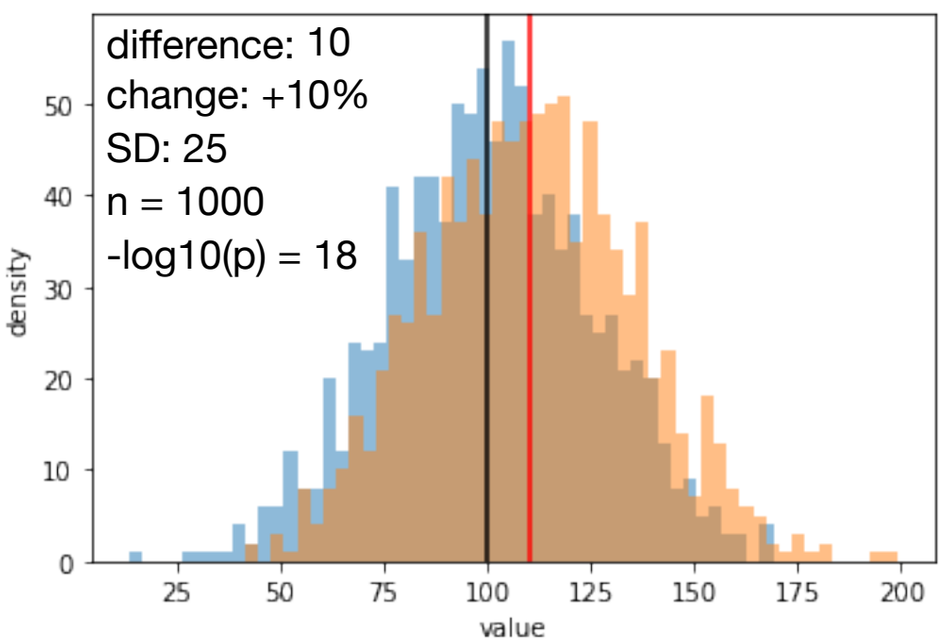
Misconceptions about P-values

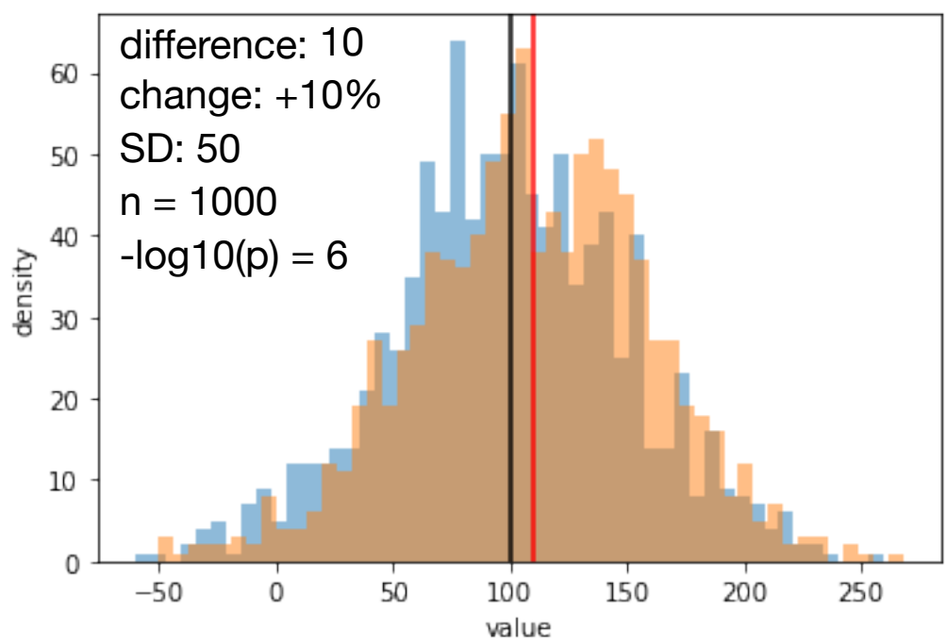
The logic is not reversible:

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

Summary parameters

1 2 2 5 5 5 10 30

Min value: 1

Max value: 30

Parametric measures

Mean: (1+2+2+5+5+5+10+30)/8 = 7.5

Variance: (1+4+4+25+25+25+100+900)/8 = 90.57143

SD: square_root (variance) = 9.516902

SD = standard deviation = sigma

Parametric measures

1 2 2 5 5 5 10 30

Mean: (1+2+2+5+5+5+10+30)/8 = 7.5

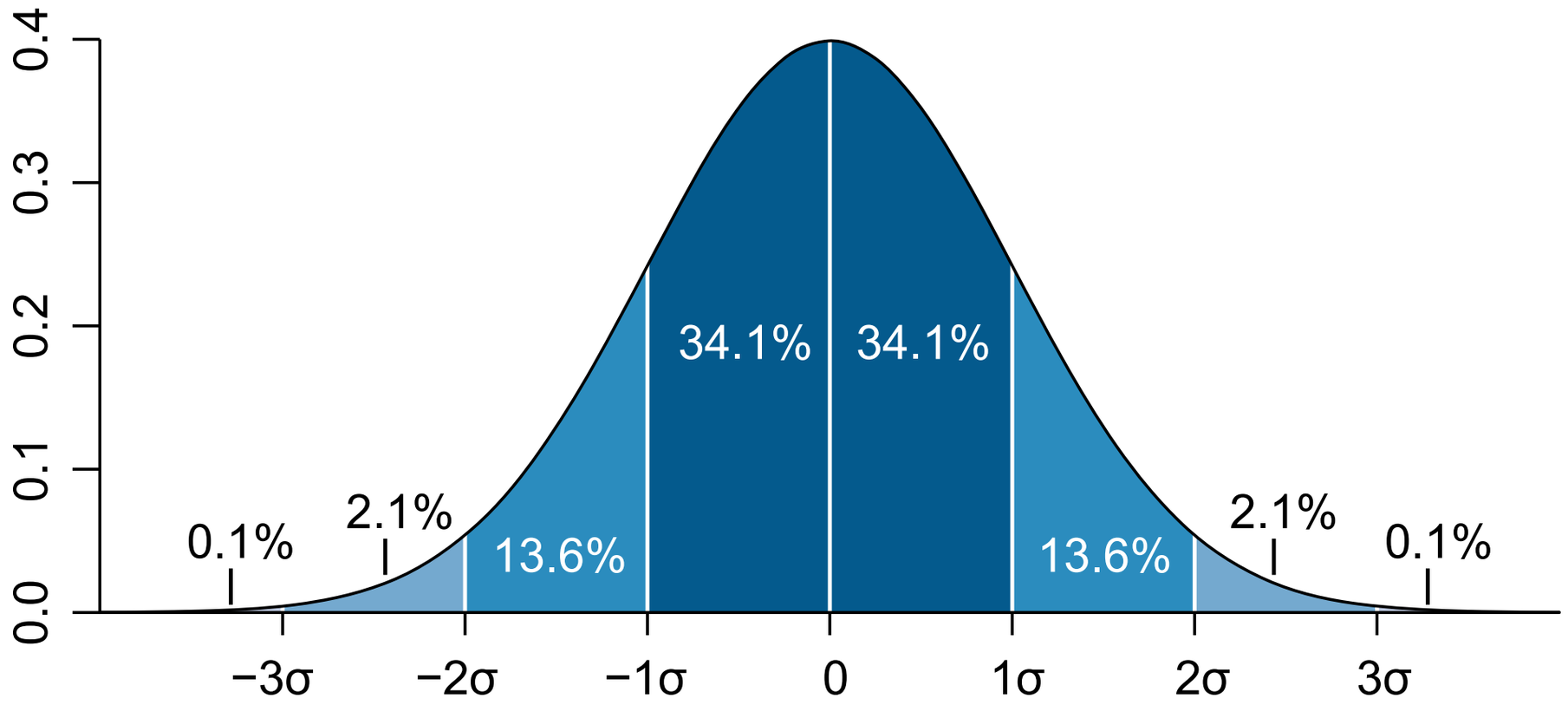
Variance: (1+4+4+25+25+25+100+900)/8 = 90.57143

SD: square_root (variance) = 9.516902

SD = standard deviation = sigma

Normal distribution

Gaussian distribution, bell-shaped distribution

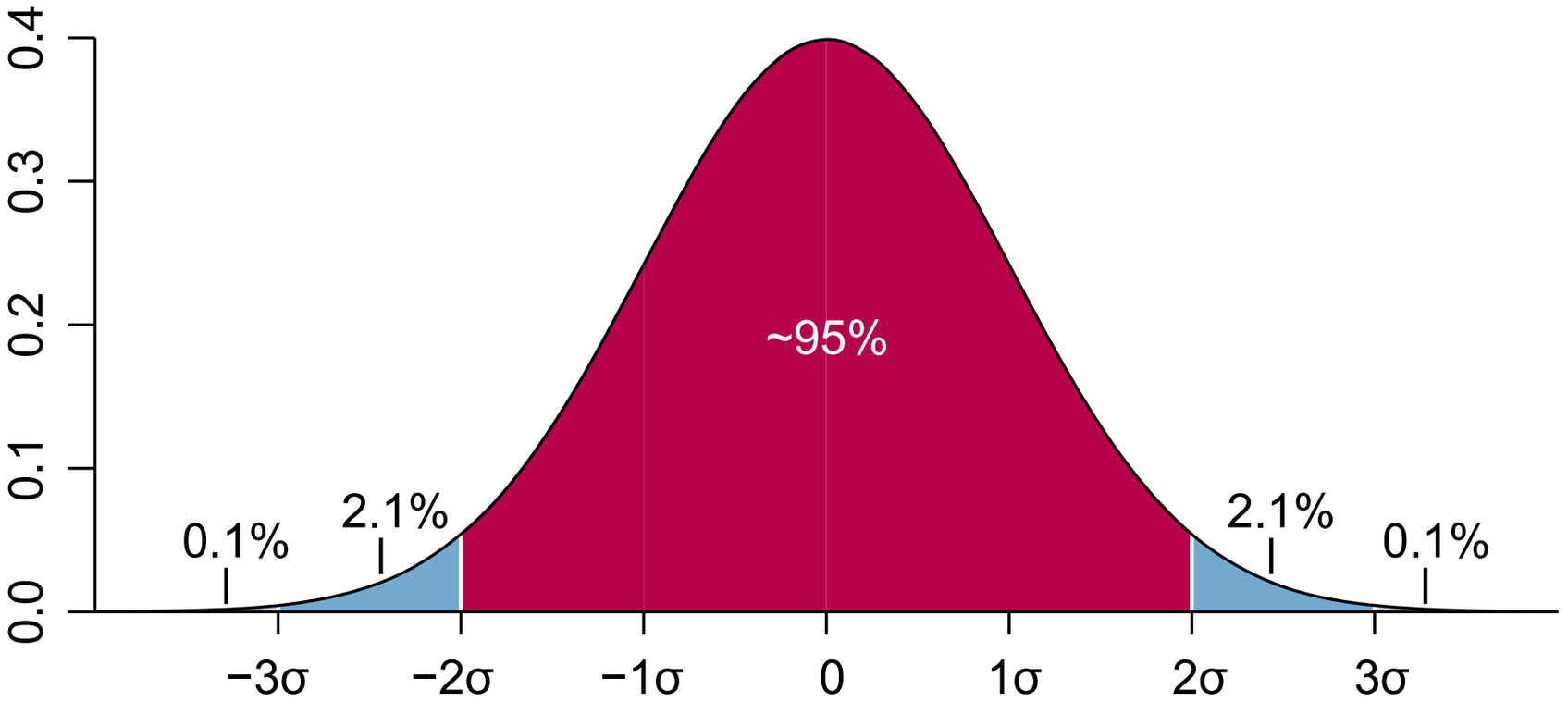


The result of general imprecision: weighing, pipetting, randomness

Therefore also: height, weight

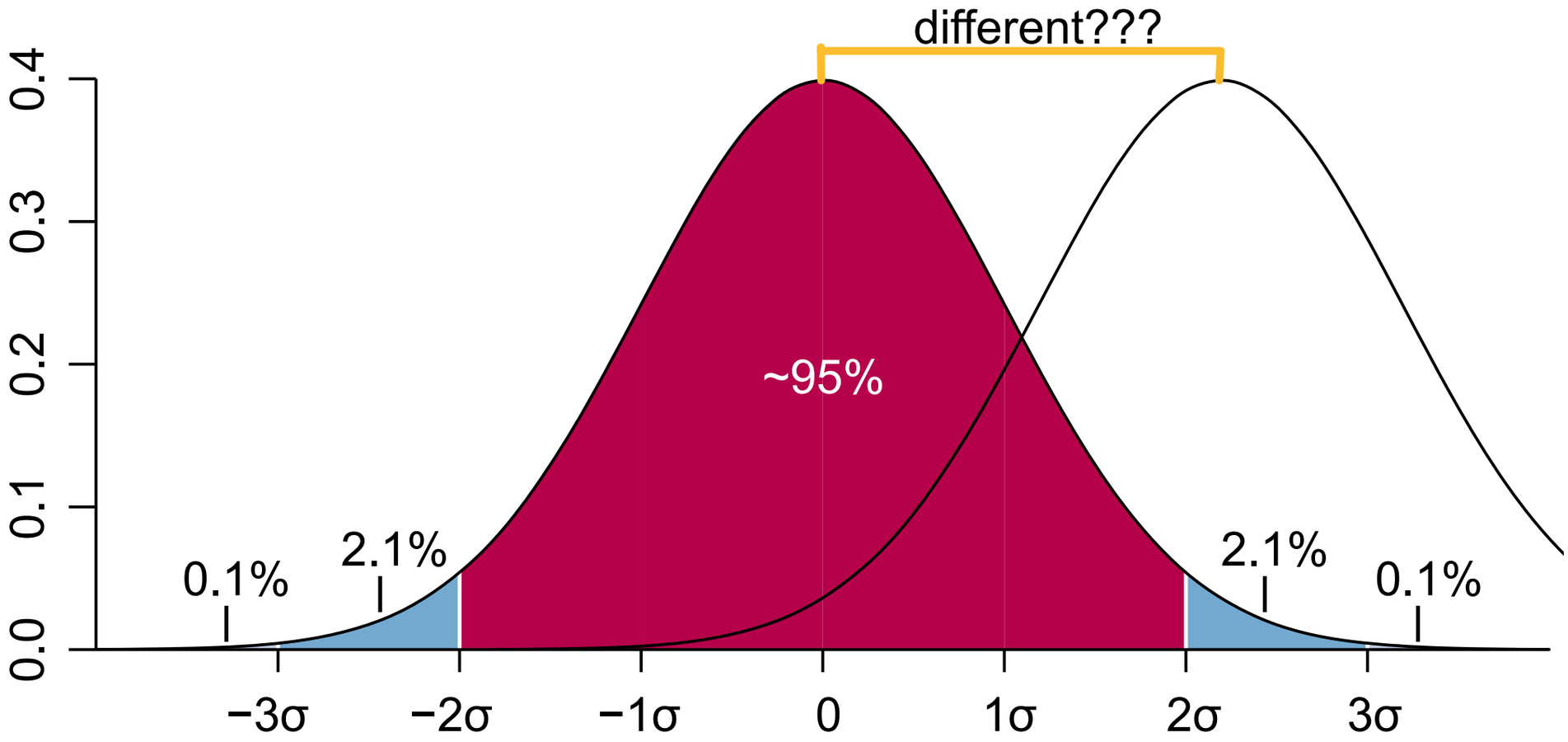
Density defined by mean and standard deviation

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

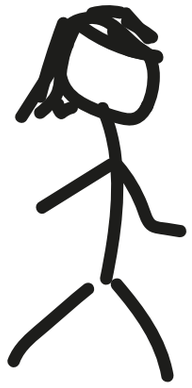


Assumptions:

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

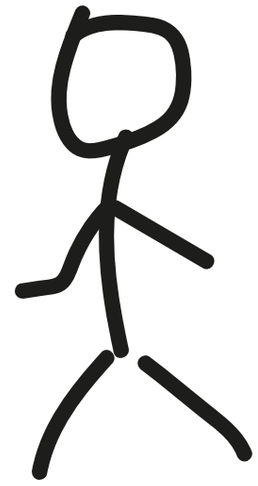
The graph is adapted from: M. W. Toews - Own work, based (in concept) on figure by Jeremy Kemp, on 2005-02-09, CC BY 2.5, https://commons.wikimedia.org/w/index.php?curid=1903871

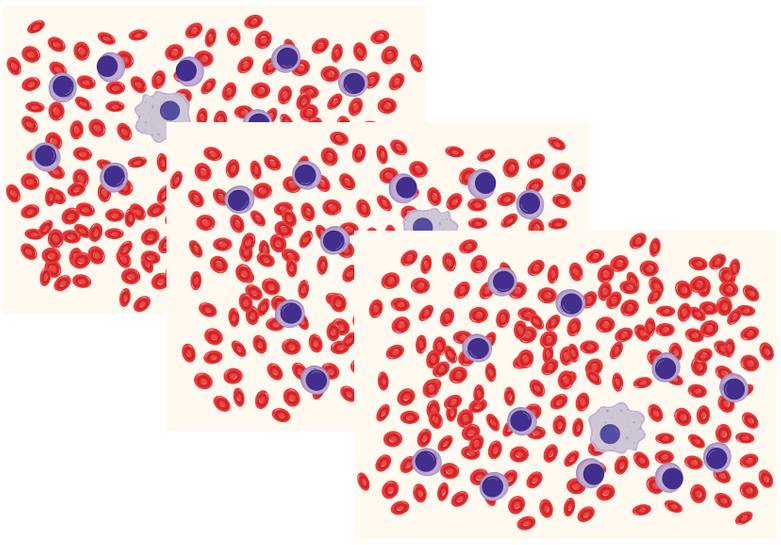
Assumptions for statistical testing

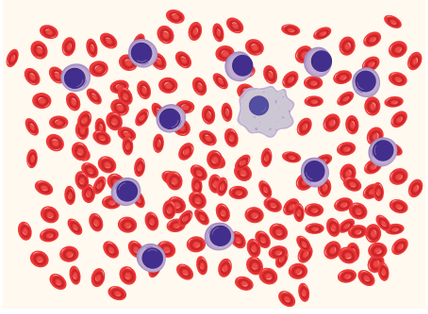


Abnormal white

blood cell count?



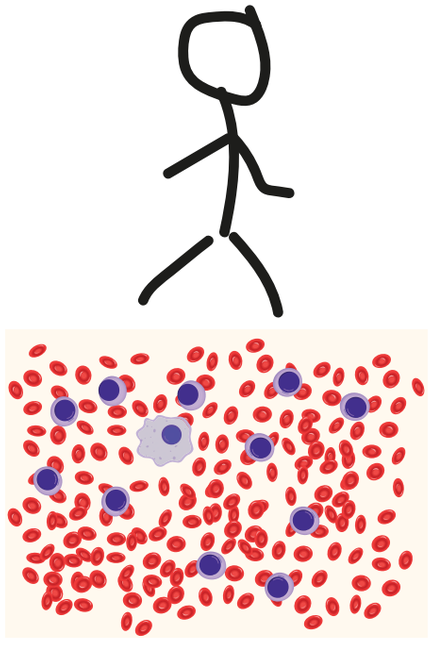


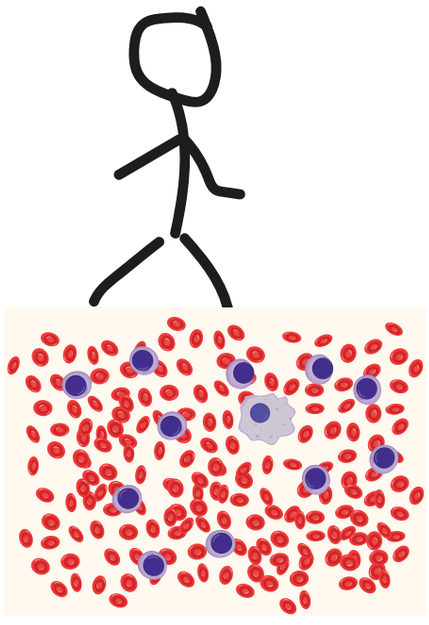


If all control cell counts are from one individual,

they are not independent!!!

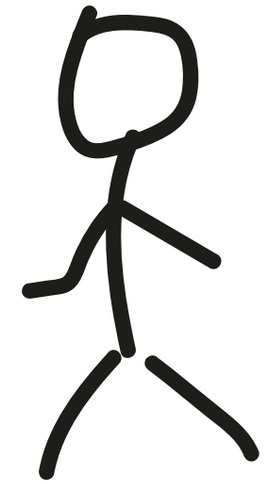
Assumptions for statistical testing

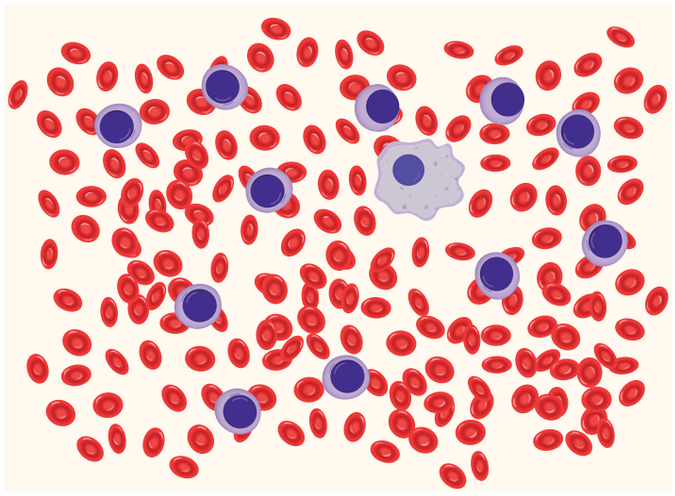


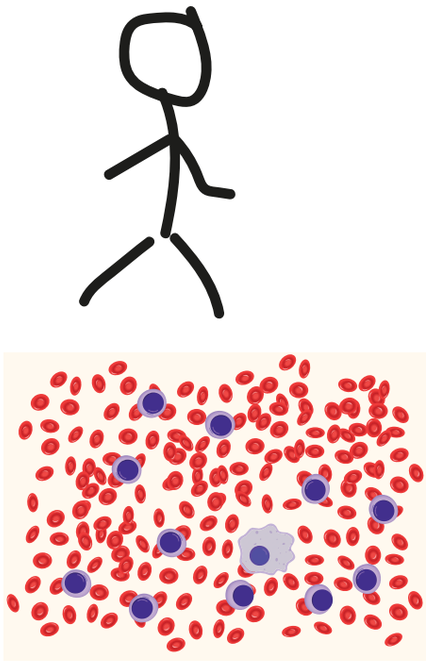


Abnormal white

blood cell count?



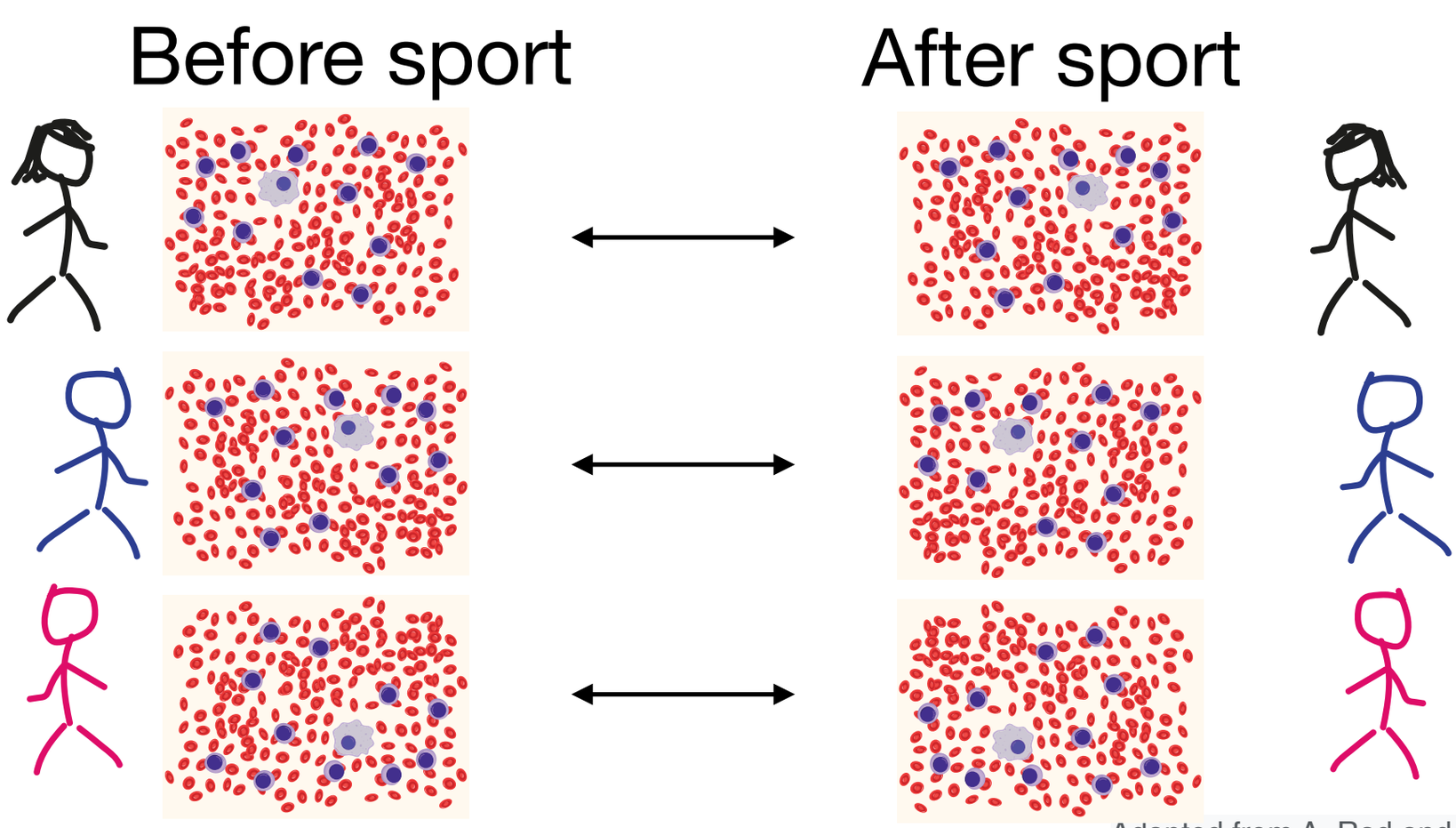




Special considerations

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?



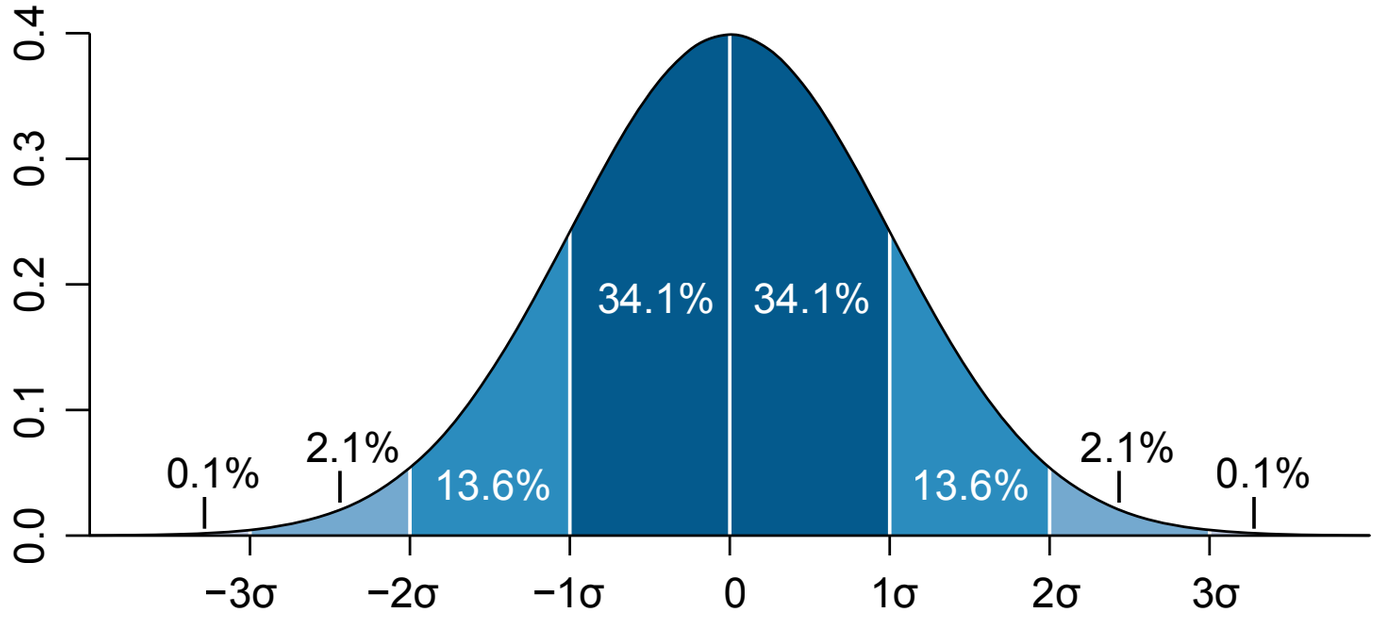
Comparing Two Means

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

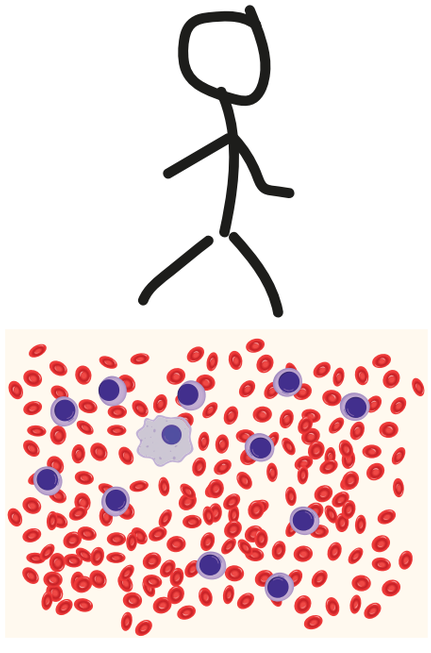
- Assume a different distribution

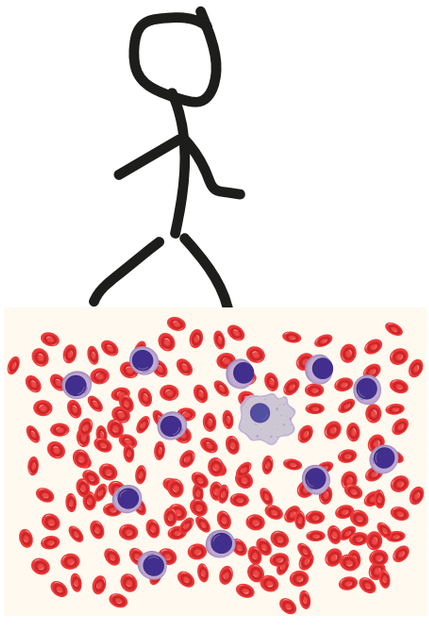
- Change to non-parametric testing

What we have covered

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

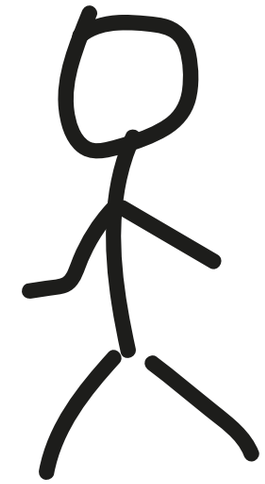
Assumptions for statistical testing

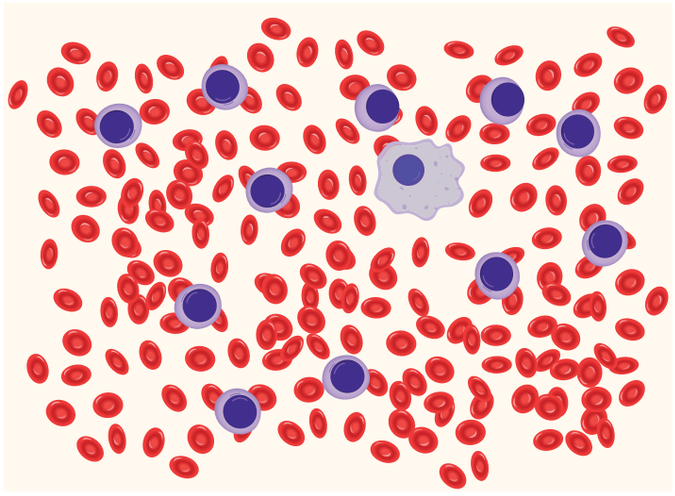


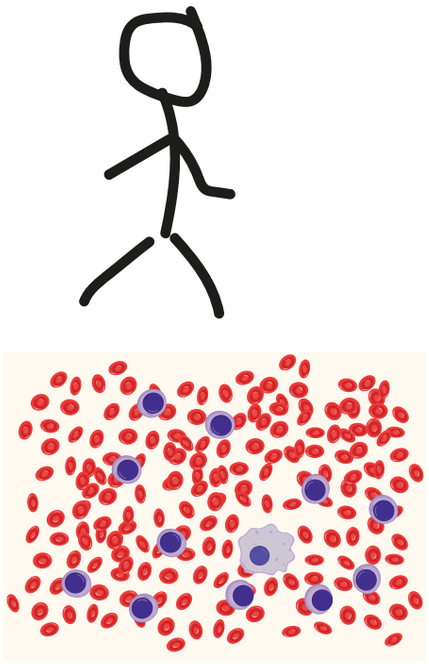


Abnormal white

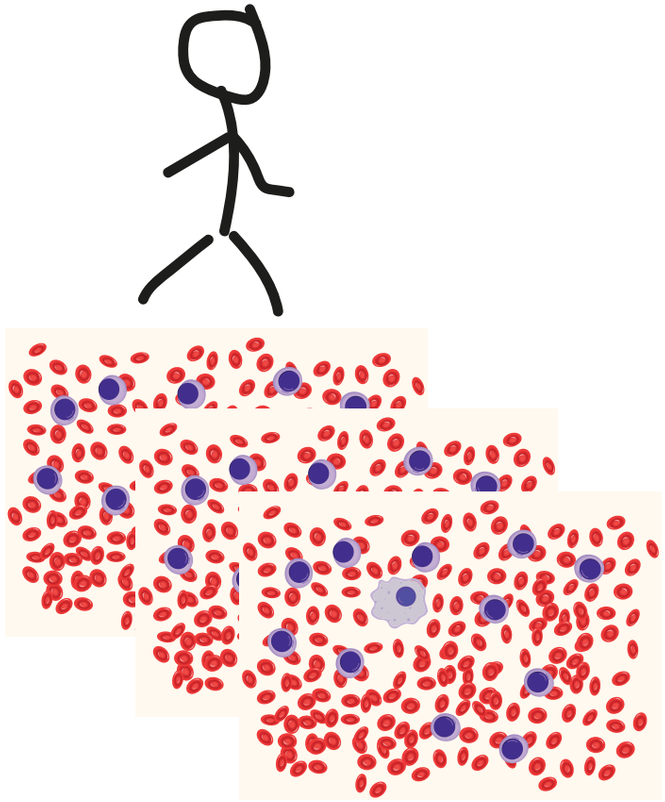
blood cell count?





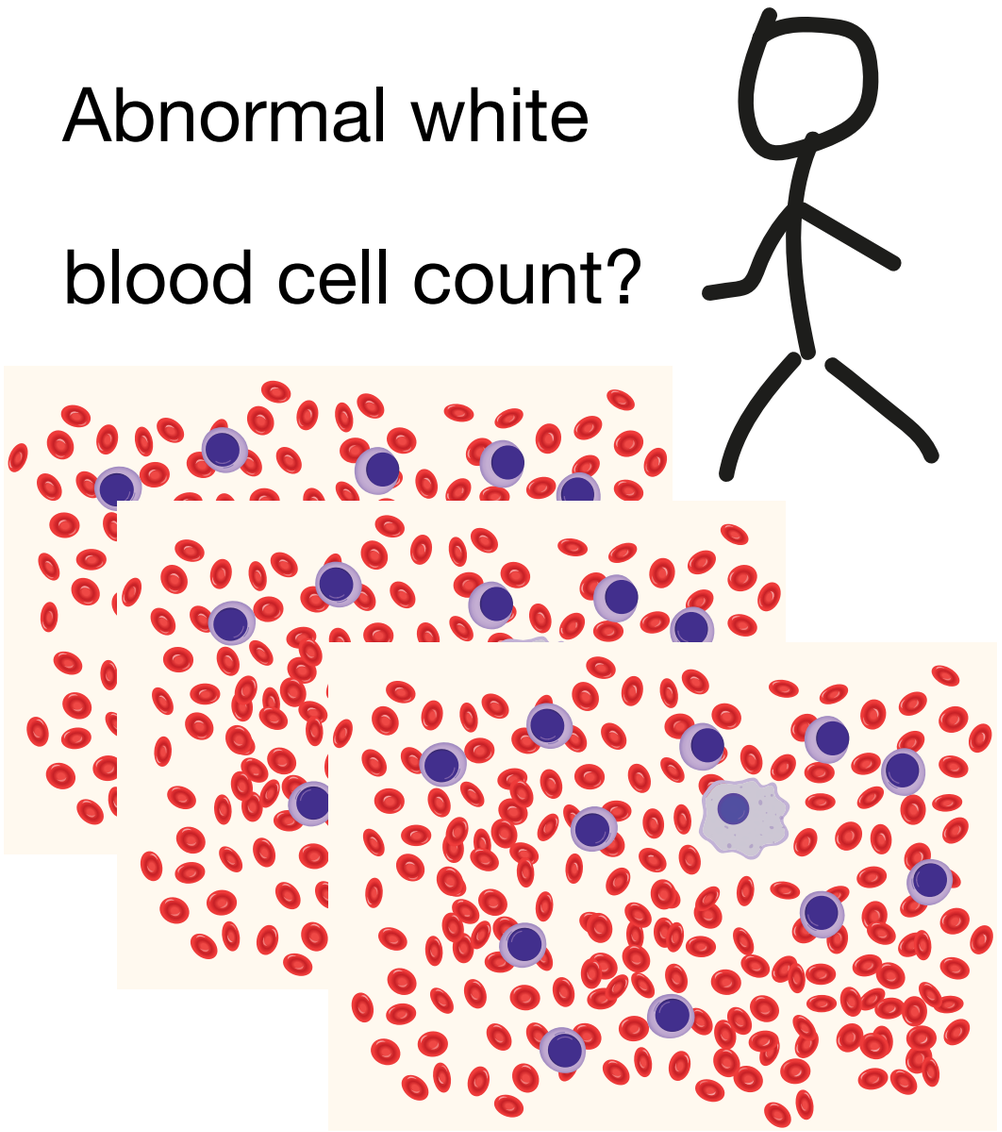


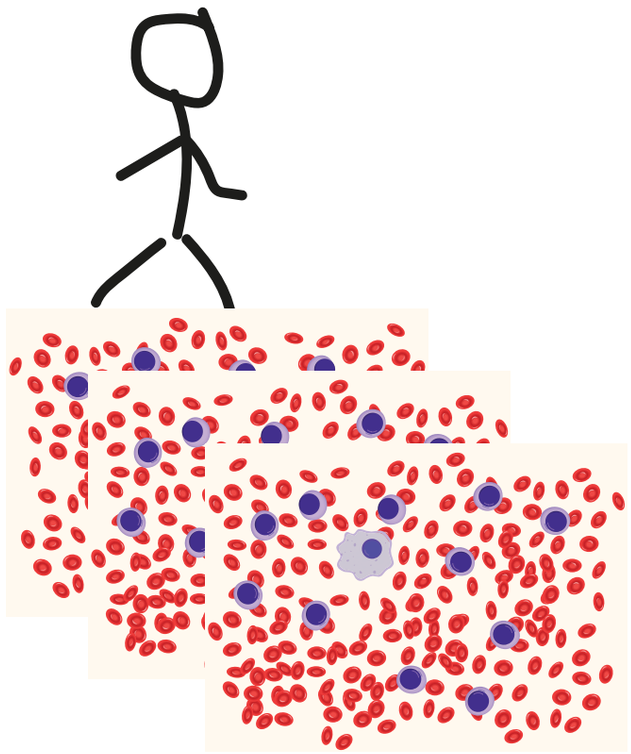
Technical and biological replicates

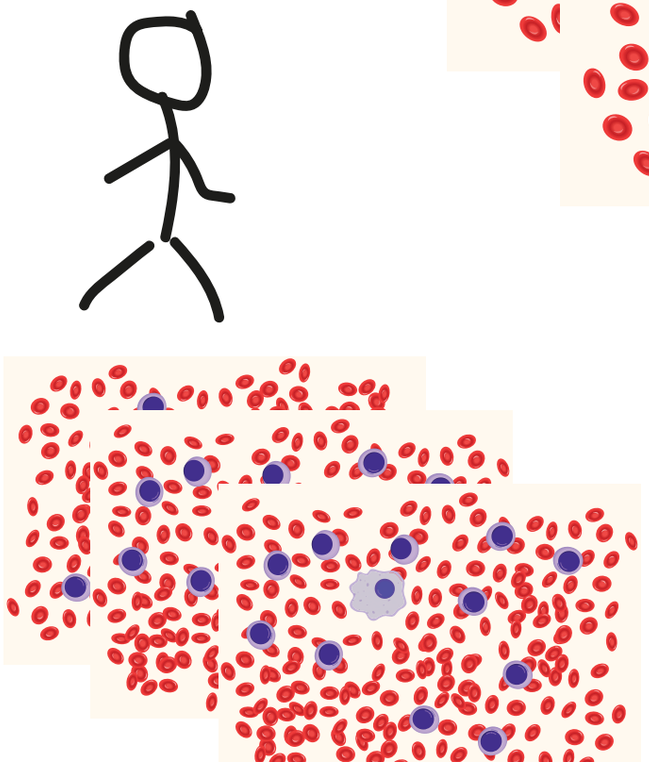


Abnormal white

blood cell count?



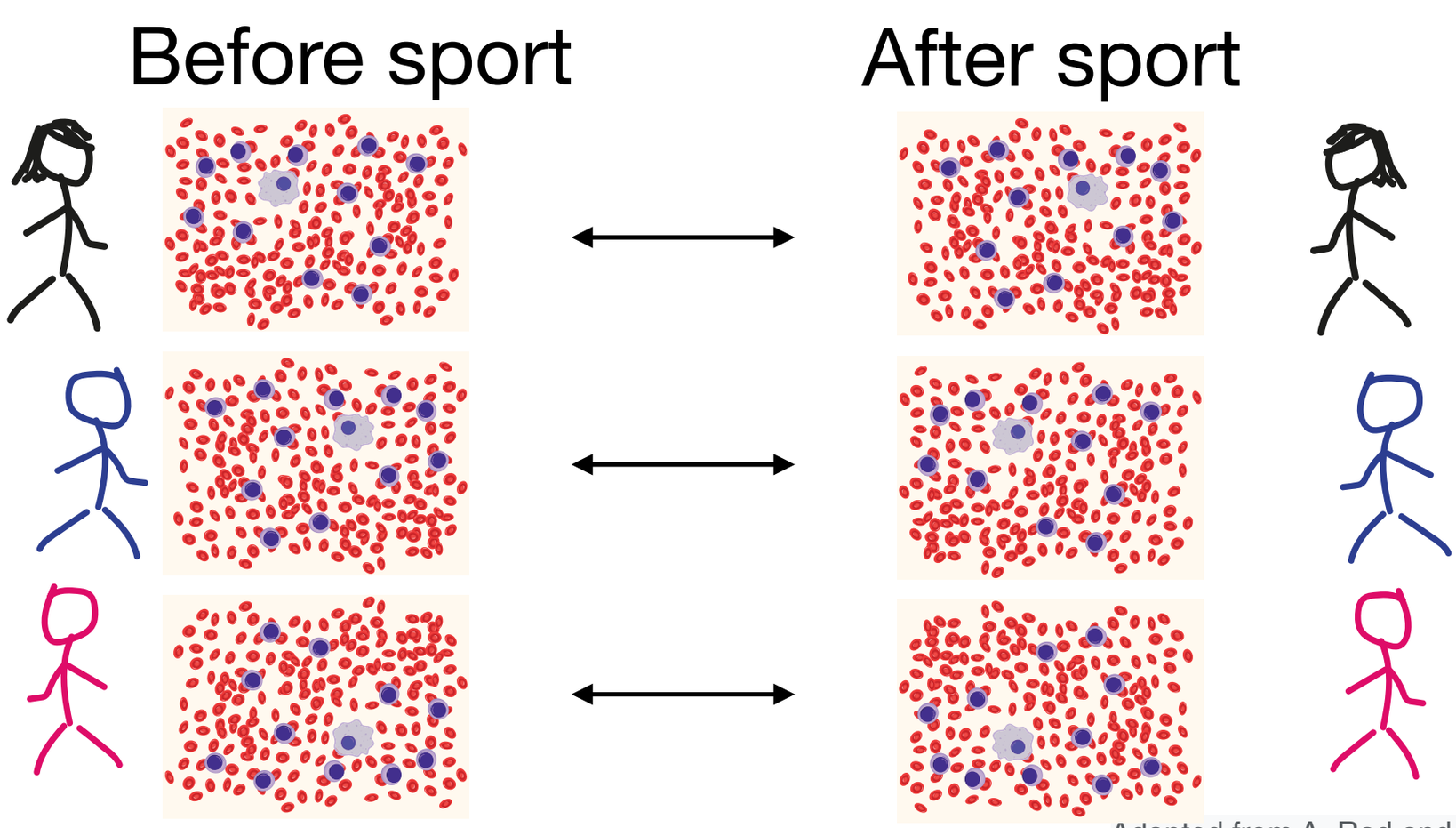




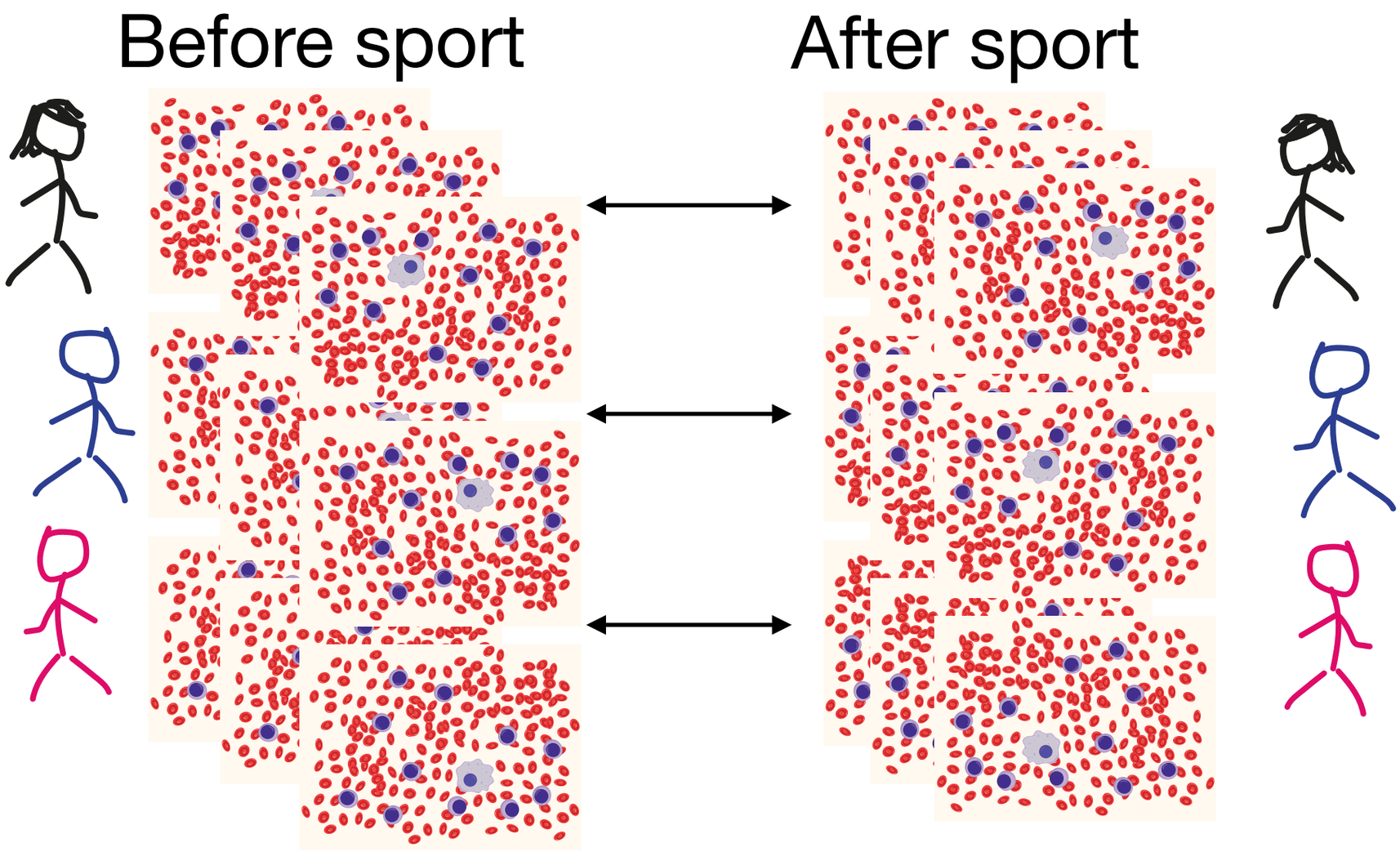
Special considerations

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?



Technical and biological replicates



Next week

- Non-parametric testing
- Multiple comparisons
- Correlations
- A glimpse into big data and visualisation

-> Jupyter Notebook