Introduction into Biostatistics

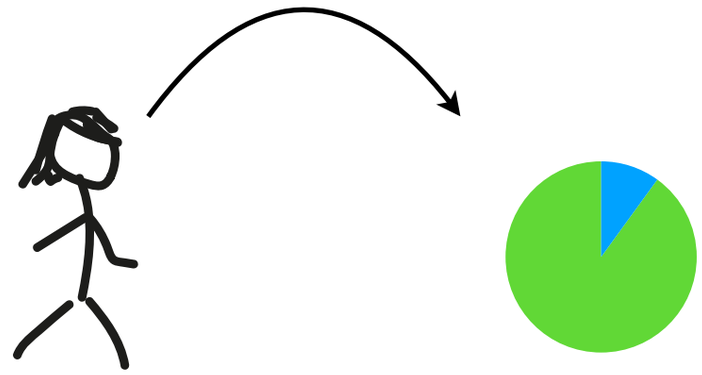
Anna Poetsch, Biotechnology Center, TU Dresden

Organisation

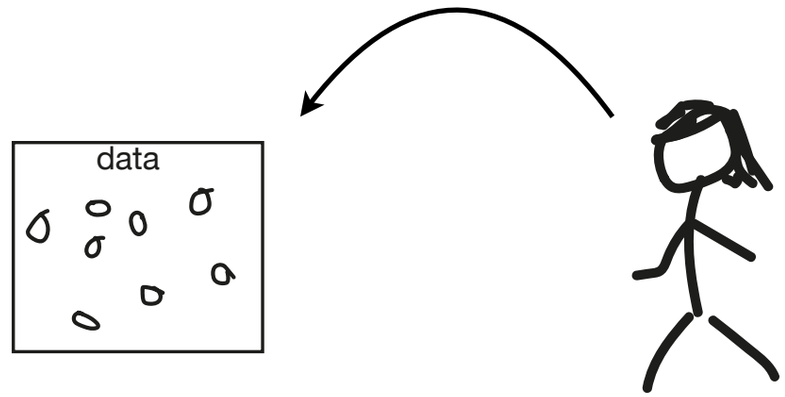
- 16.5. Introduction to biostatistics
- **14.6. Distributions and hypothesis testing**
- 21.6. Non-parametric testing and multiple comparisons
- 28.6. Correlations and dimensionality reduction

Recap on probability

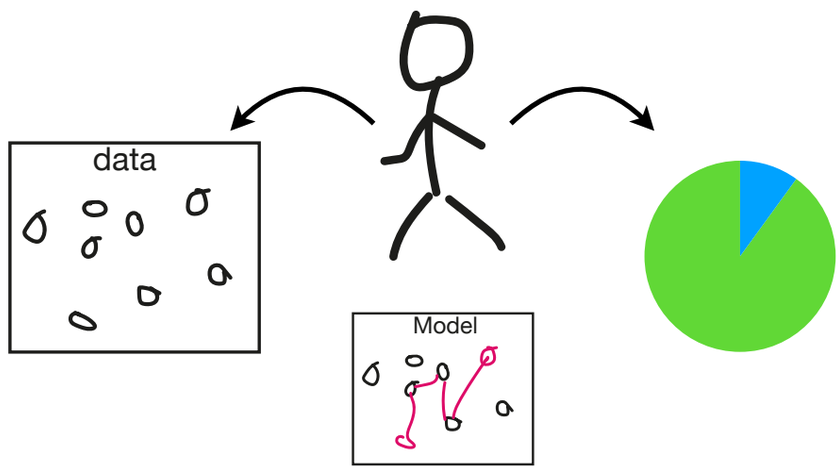
A model



Data



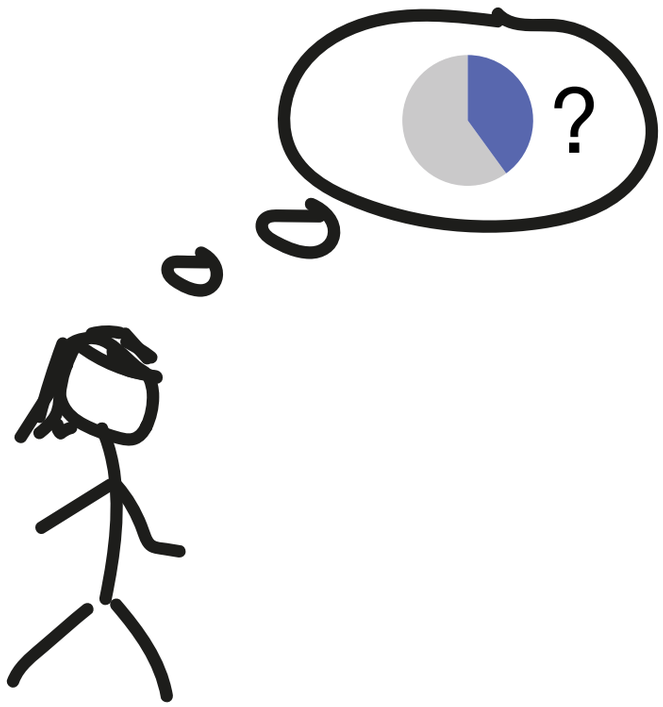
A model based on data

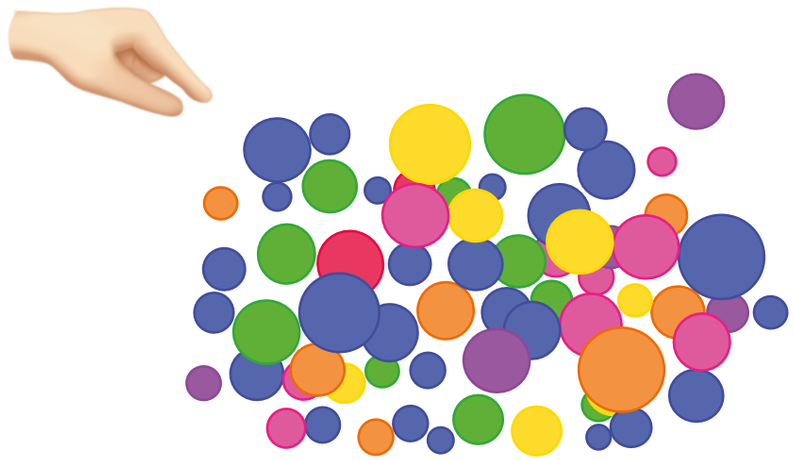


! **Estimating probabilities can only be as good as your assumptions/ data**

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

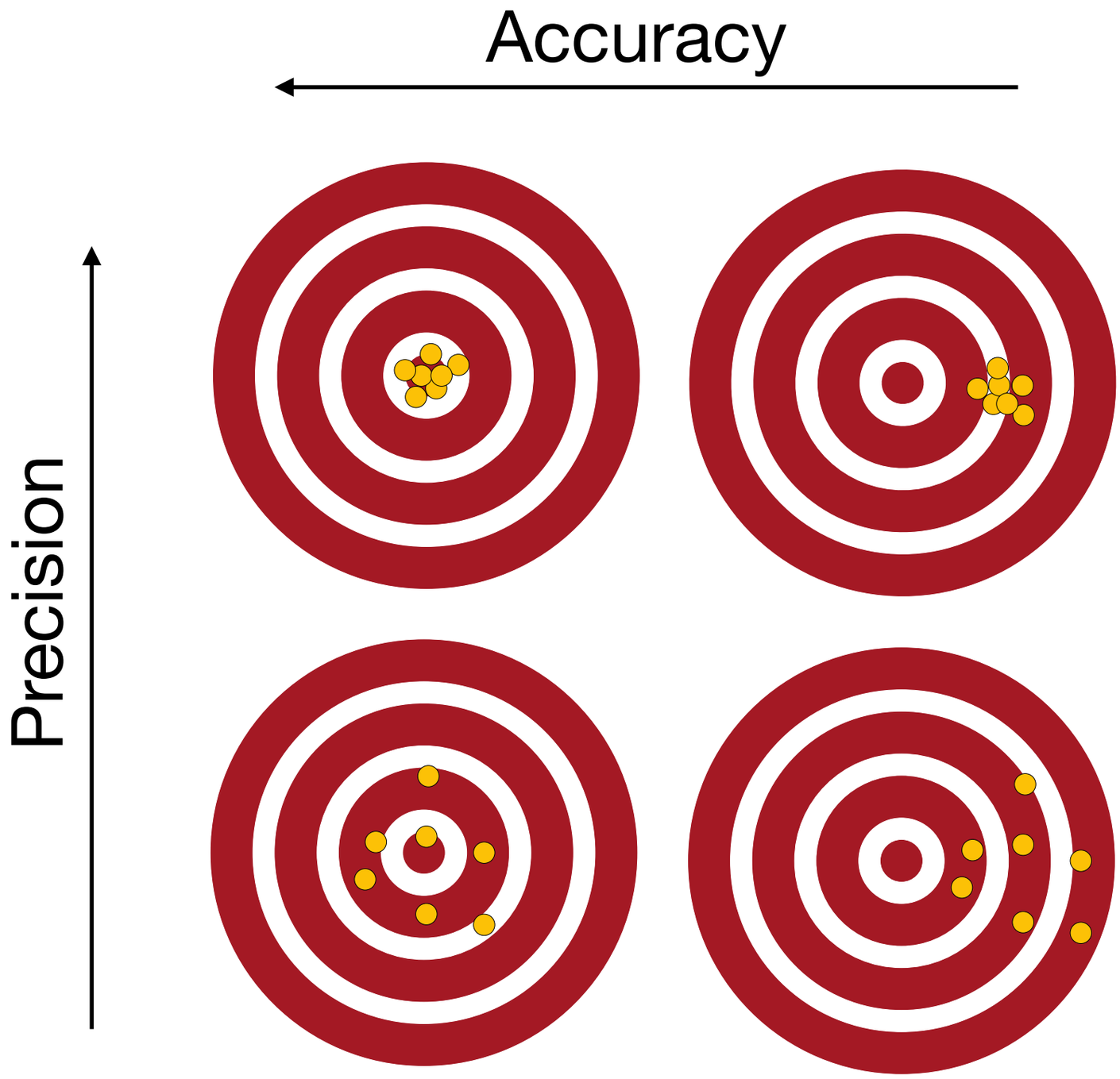
Please answer in the chat:

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

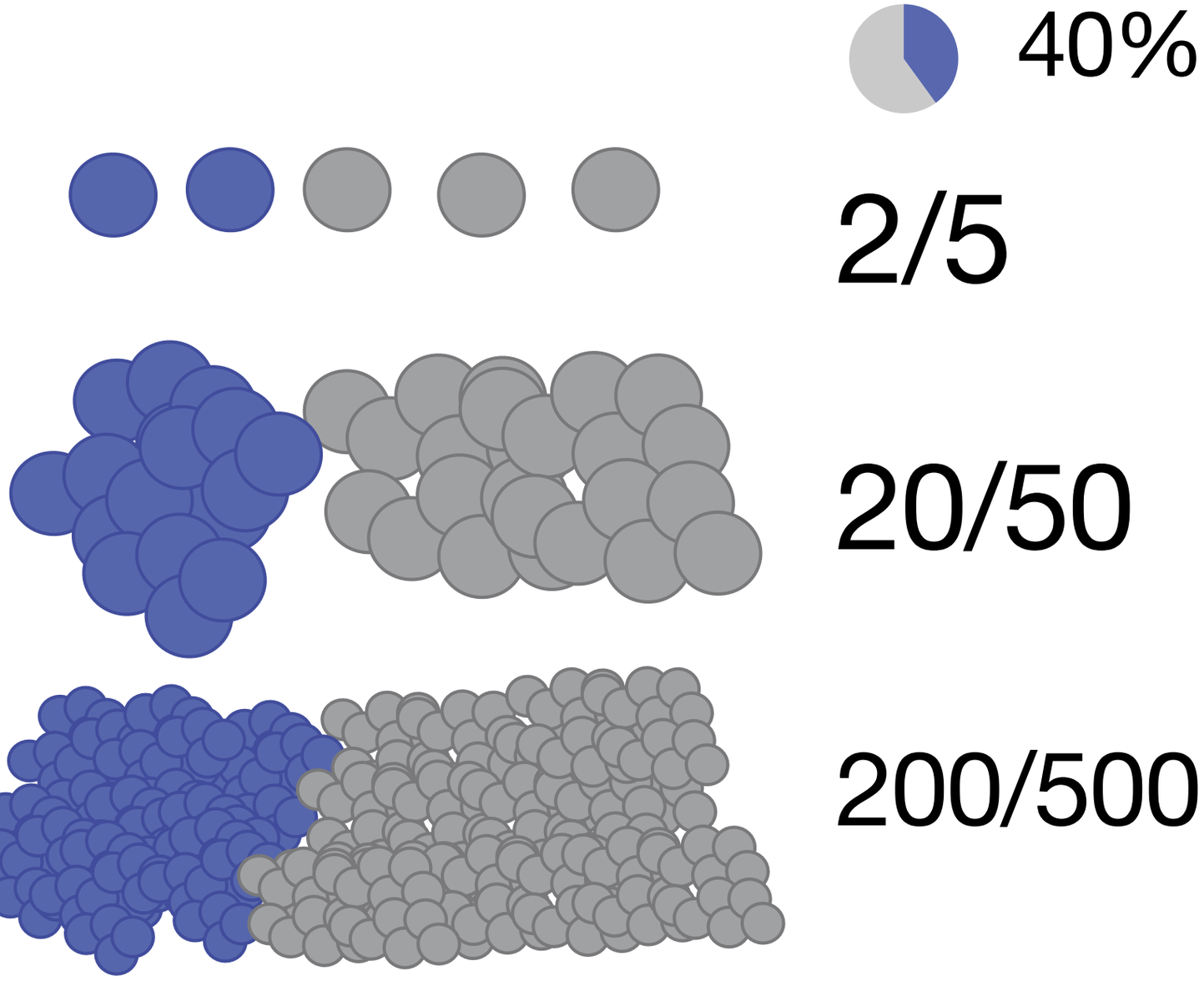


How do these relate to confidence intervals?

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

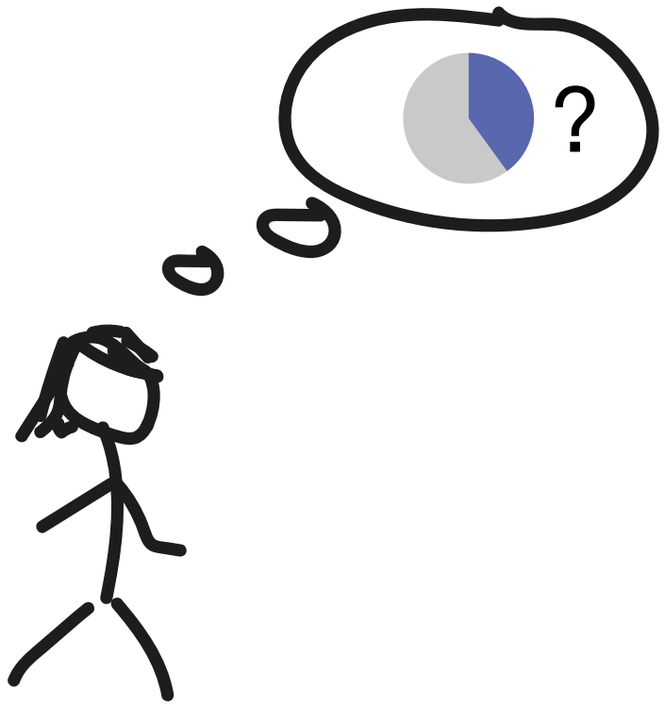
Data types and distributions

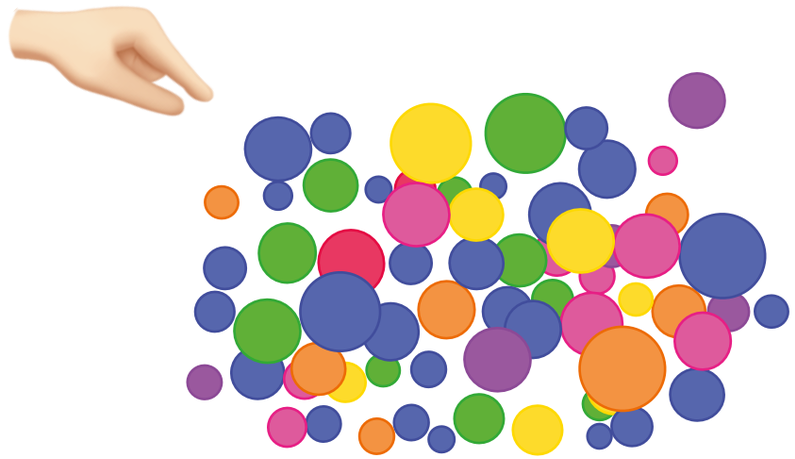
- Probability data (binomial distribution)
- Counted data (Poisson distribution)
- Normal distribution

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

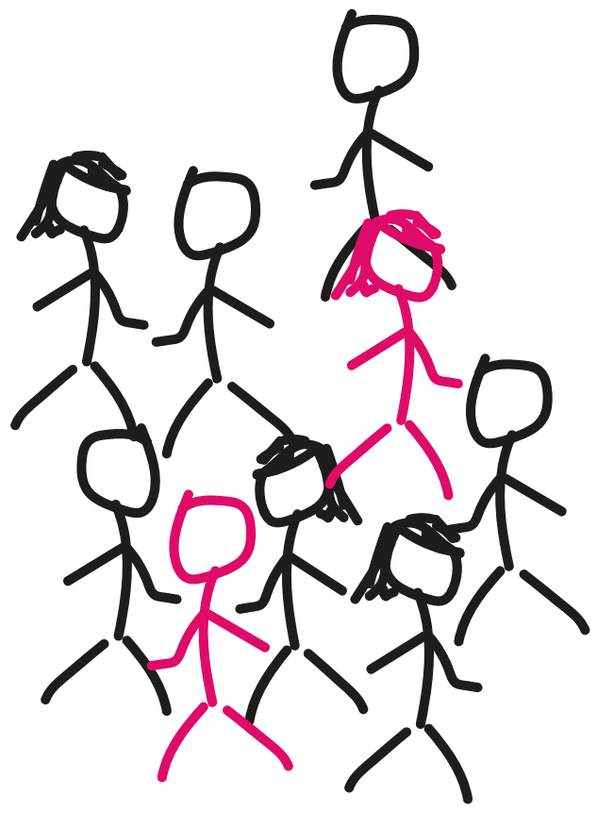
- limited set of discrete values with order

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

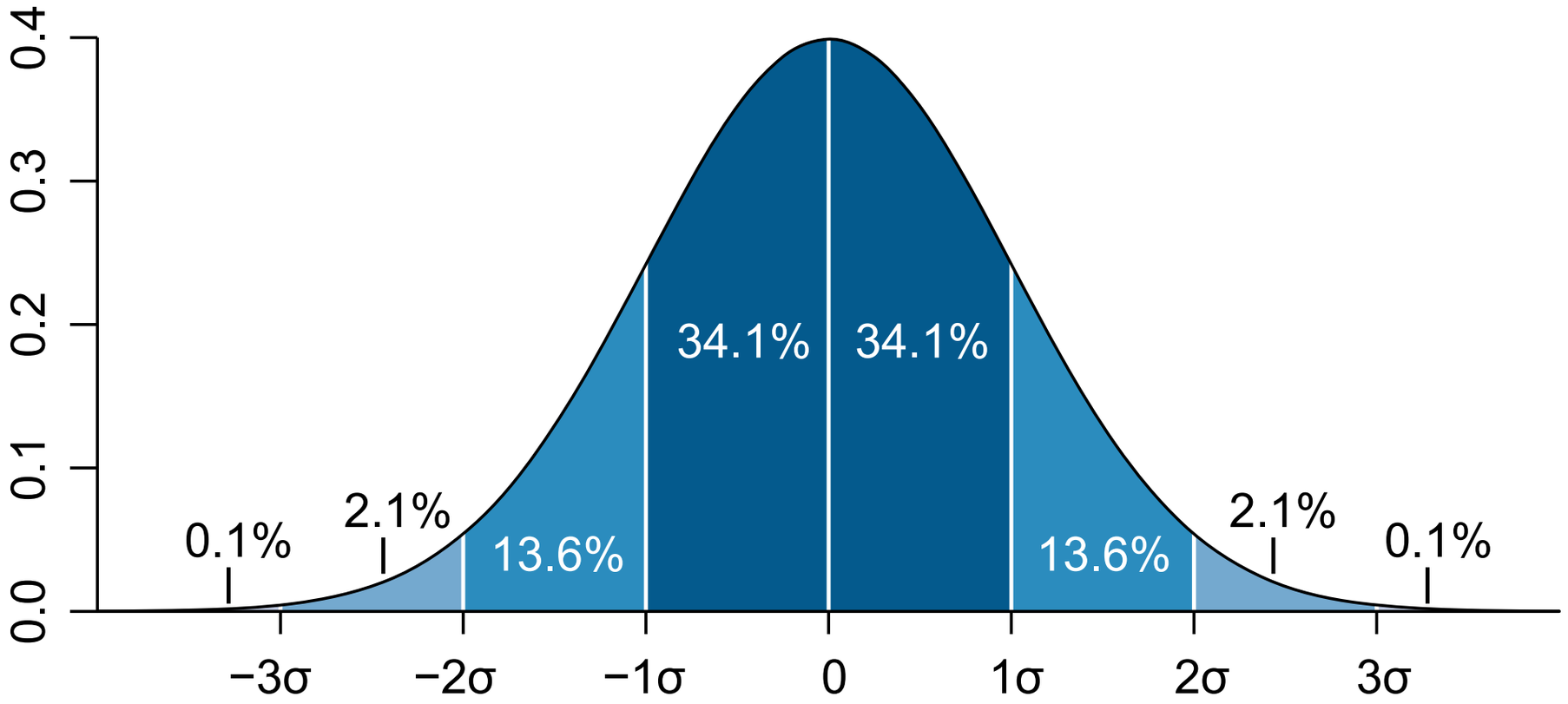
Median: the central value: 5

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

-> Jupiter Notebook

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)