CAS Applied Data Science – Module 2 – Day 1

Statistical Inference for Data Science

Dr. Anja Mühlemann

31. August 2021

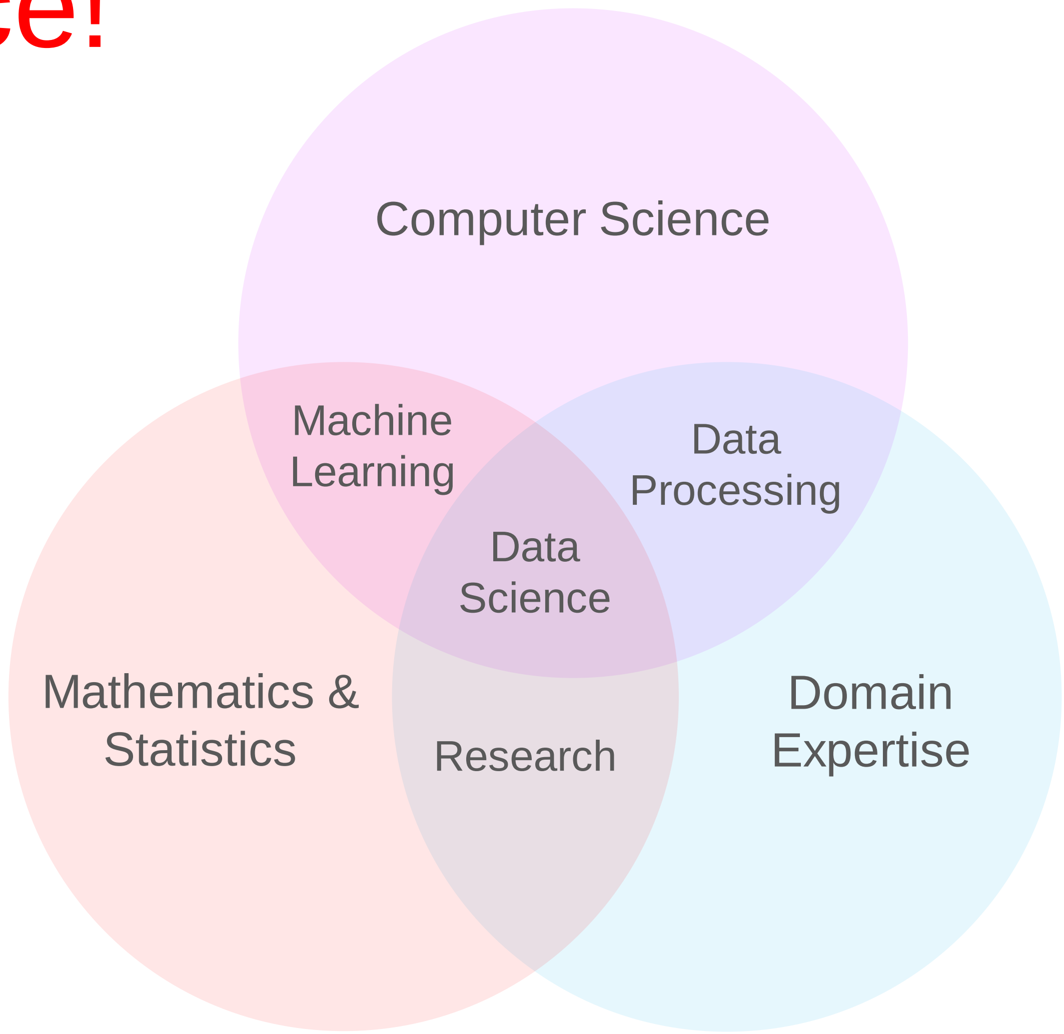
b
UNIVERSITÄT
BERN

Welcome to Data Science!

Data Science uses

- Mathematics and Statistics
- Computer Science
- Domain expertise

on data to build information and extract knowledge.

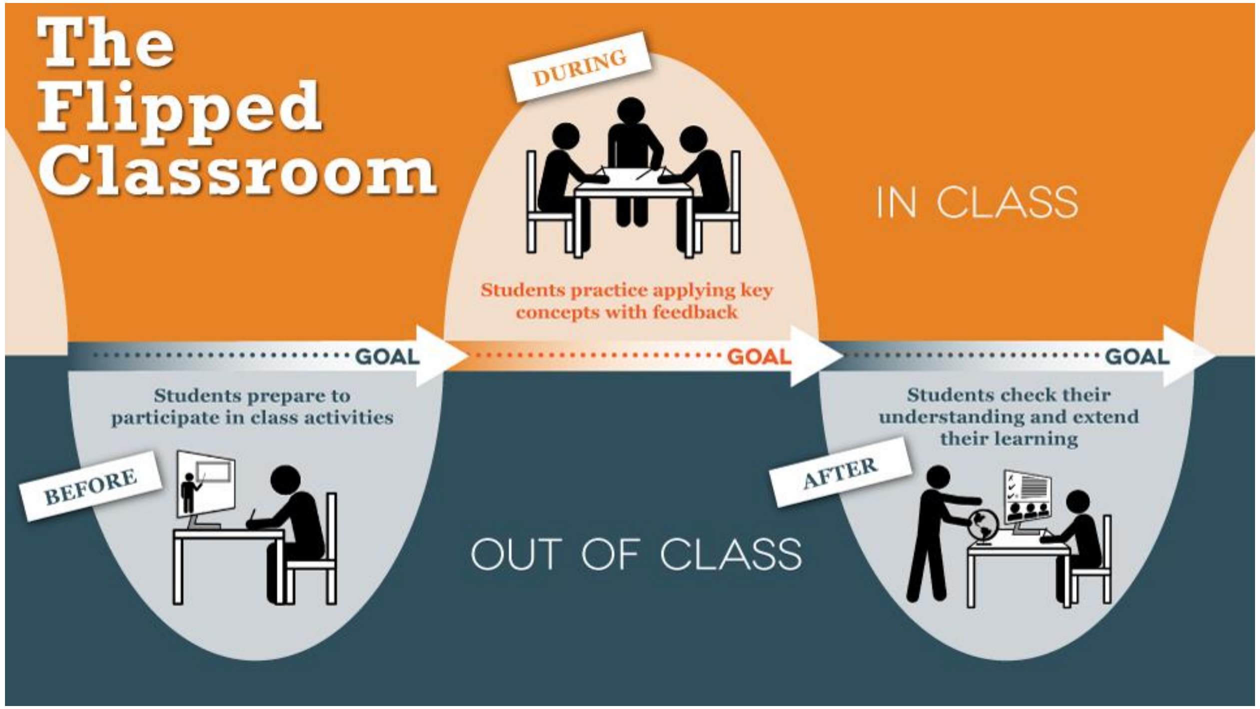


Module 2

Day 1	Descriptive Statistics and Probability
Day 2	Parameter estimation
Day 3	Hypothesis testing
Day 4	Putting it all together
Project	Presenttation session (date to be fixed!)

Inverted classroom

- Introduction lectures
- In-depth study of the content with notebooks
- Discussion sessions based on your questions and comments
- Project: poster with poster presentation
- 1-2 questions per day



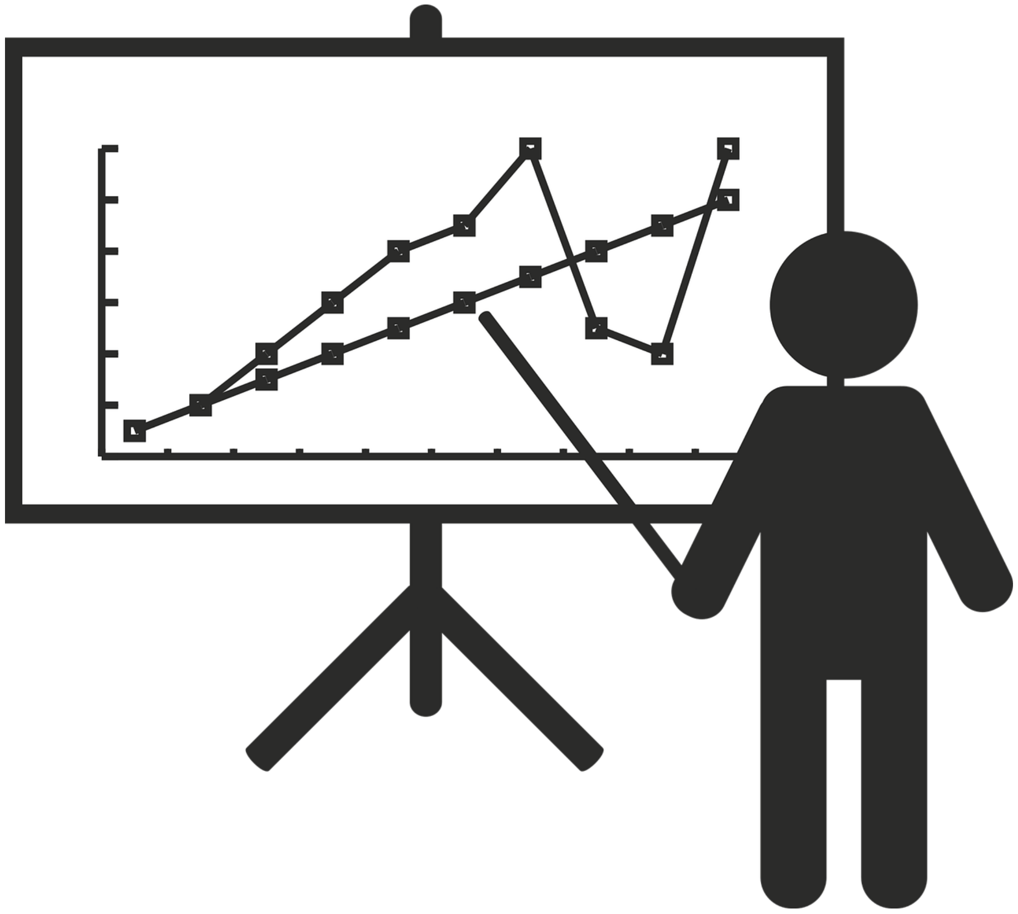
Project

Formal

- Group of 2 people
- 15min presentation, 15min discussion
- Half-day presence on two dates (to be fixed)

Content

- Choose your own data set
- answer research questions using statistics



Iris data set

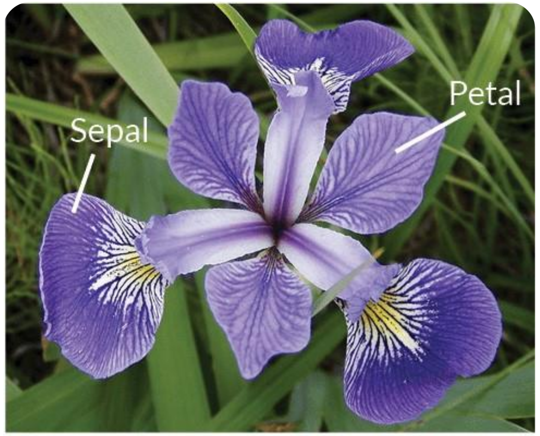
3 classes: versicolor, setosa, virginica

4 characteristics

- petal: *length, width*
- sepal: *length, width*


Iris Virginica


Iris Setosa



Iris Versicolor

Foretaste (Project, 4th day)

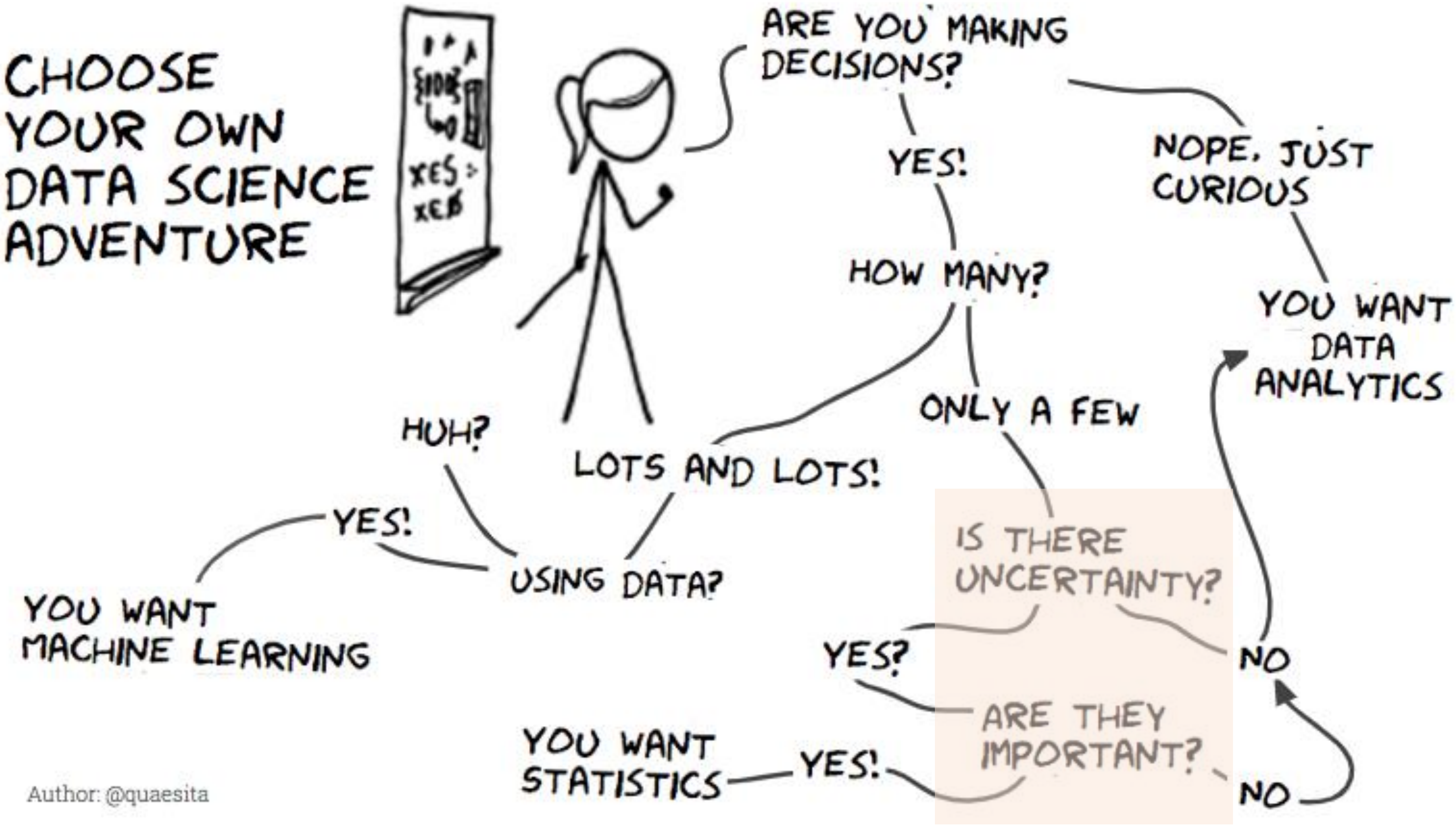
Some new company recently sequenced the genes of the Iris species Setosa and patented it, apparently in order to preserve this species because it is so beautiful. Due this patent it is not allowed to change the plant.

A big farmer and hater of Iris and with a field where Iris is a disturbing weed, has been using a new product from Sonte Manto for a couple of years. The product is supposed to effectively kill Iris plants.

A big Iris lover collected a sample of Iris plants from the farmer's field and thinks the Iris Setosa setal leaves are bigger than normal. She sent the sample to the company, which in turn came to the conclusion that Setosa must have mutated due to the product from Sonte Manto.

So the company sued Sonte Manto with the claim that they have changed the plant with their product. Sonte Manto may risk to pay a billion dollars.

The court is asking you to give a neutral and scientific advice.


CHOOSE YOUR OWN DATA SCIENCE ADVENTURE
ARE YOU MAKING DECISIONS?
YES!
NOPE, JUST CURIOUS
HOW MANY?
YOU WANT DATA ANALYTICS
ONLY A FEW
HUH?
LOTS AND LOTS!
YES!
USING DATA?
IS THERE UNCERTAINTY?
YOU WANT MACHINE LEARNING
YES?
NO
ARE THEY IMPORTANT?
YOU WANT STATISTICS
YES!
NO
Author: @quaesita

General Procedure

1. Planning
2. Data Collection & Preprocessing
3. Descriptive Statistics
4. Classification, Clustering,…
5. Inferential Statistics

Describing Data

Why?

- Get an overview
- Patterns
- Outliers
- Quality
- Learn about distibutions

How?

- Tables
- Plots
- Words
- Statistics

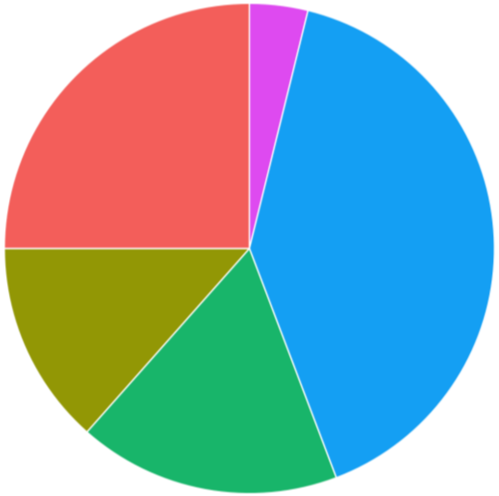
➡ Good description is the basis for good inference

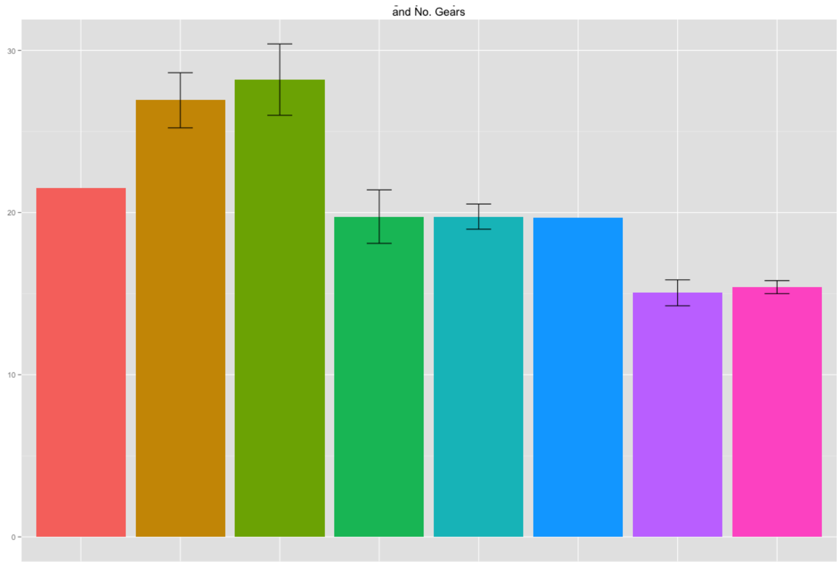
Descriptive Statistics

The two main tasks of descriptive statistics are

- the quantitative description and summary, and
- the graphical representation of data

What tools are suitable depends on the type of the variable we want to describe.





Categorical Variables

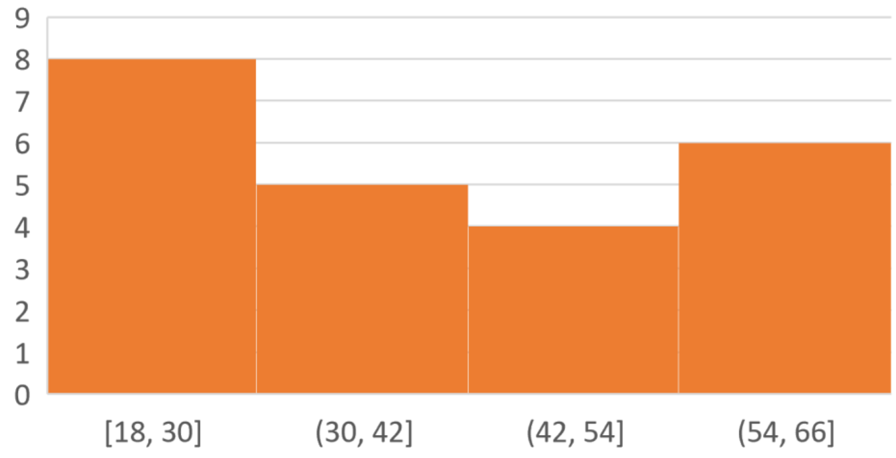
- Absolute frequency (eg. Number of female participants)
- Relative frequence (eq. Number of female participants divided by the sample size)

Numerical Variables

Summary Tables

Age	Nr. of People
18-30	8
30-42	5
42-54	4
54-66	6

Histograms



Location

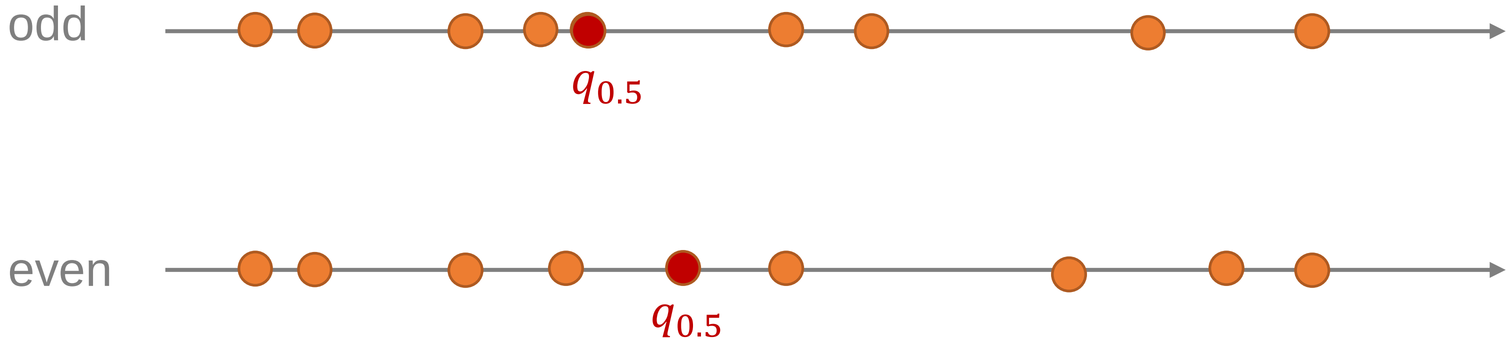
(Numerical Variables)

What are typical values for the variable X?

- Sample Mean:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

- Sample Median: «center of the observations»



→ median ist more robust than the mean

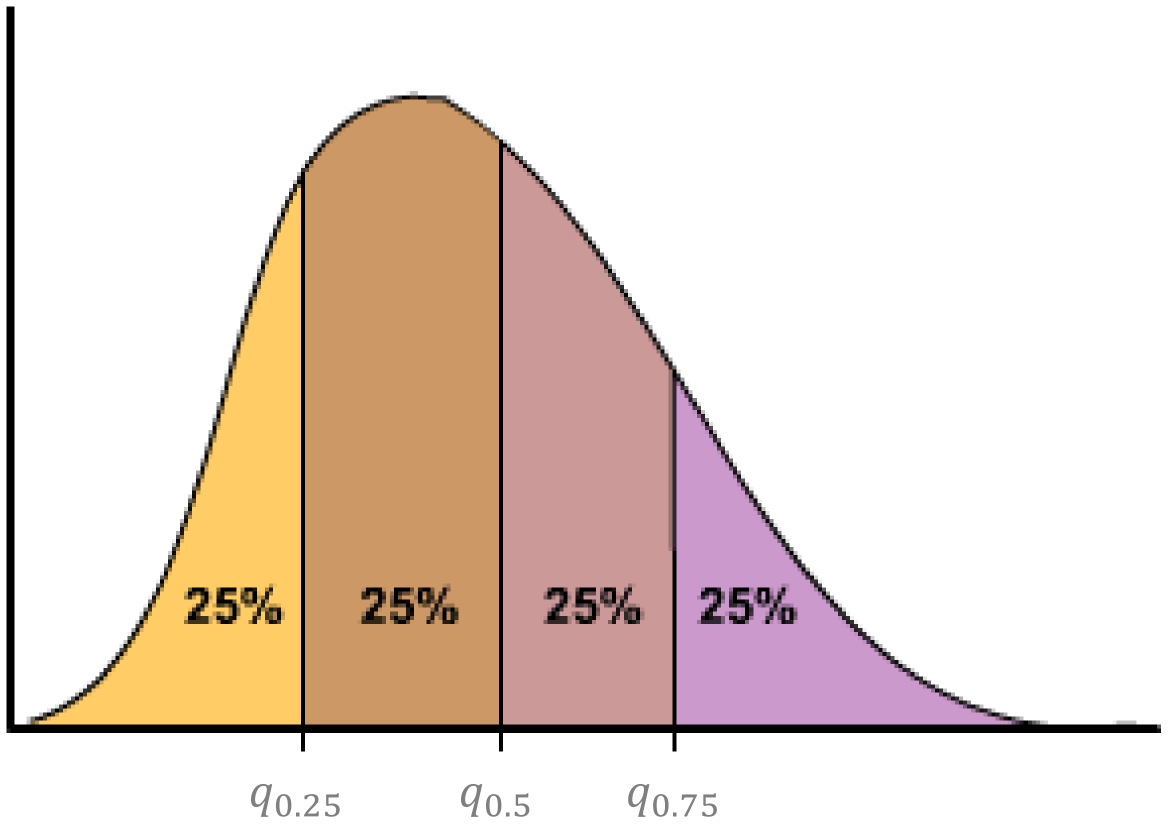
Quantiles

(Numerical Variables)

Generalizing the idea of the median to other fractions.

Typical for descripitve analyses: $q_{0.25}, q_{0.5}, q_{0.75}$

Typical for hypothesis testing: $q_{0.01}, q_{0.05}, q_{0.95}, q_{0.99}$



Spread

(Numerical Variables)

How strong is the deviation from the center?

- Sample standard deviation:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- IQR (inter quartile range):

$$IQR = q_{0.75} - q_{0.25}$$

$S = 4.05, IQR = 5.93$

$S = 1.16, IQR = 1.34$

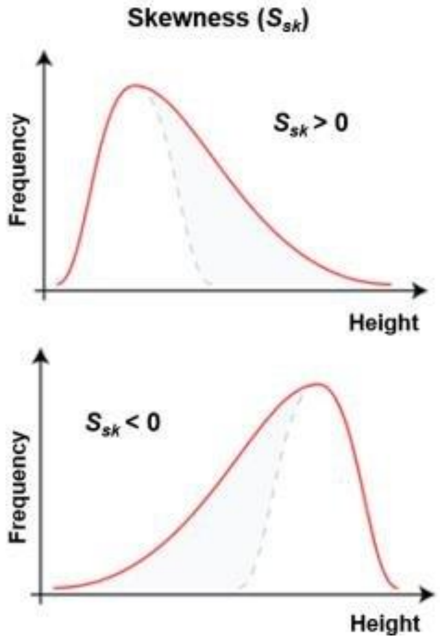
Shape

(Numerical Variables)

Is the distribution symmetric?

- Skewness:

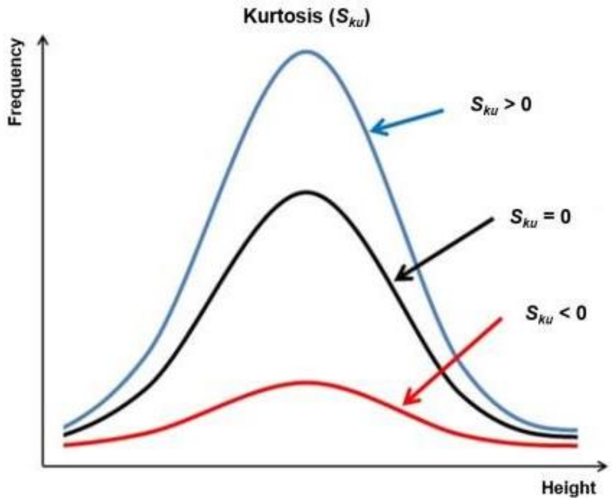
$$skewness = \frac{1}{n}\sum_{i=1}^{n}\frac{(x_i - \bar{x})^3}{s^3}$$



Does the distribution look like a bell curve?

- Kurtosis:

$$kurtosis = \frac{1}{n}\sum_{i=1}^{n}\frac{(x_i - \bar{x})^4}{s^4} - 3$$



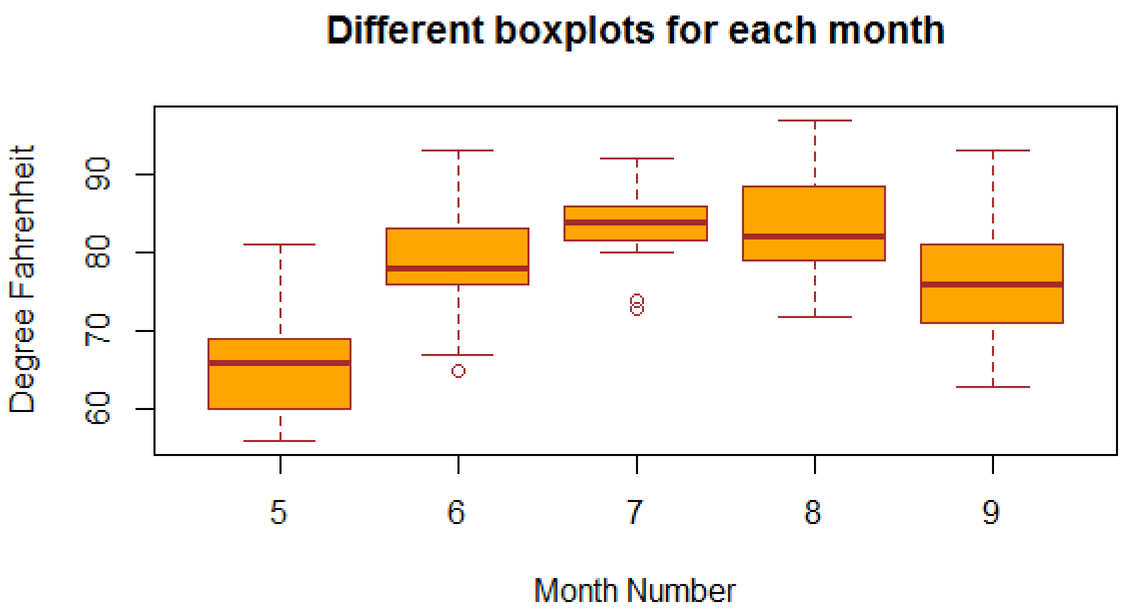
Simultaneous description I

(of two features)

- Contigency table (2 categorical features)

	Male	Female	Total
Blonde	4	8	12
Brunette	7	9	16
Total	11	17	28

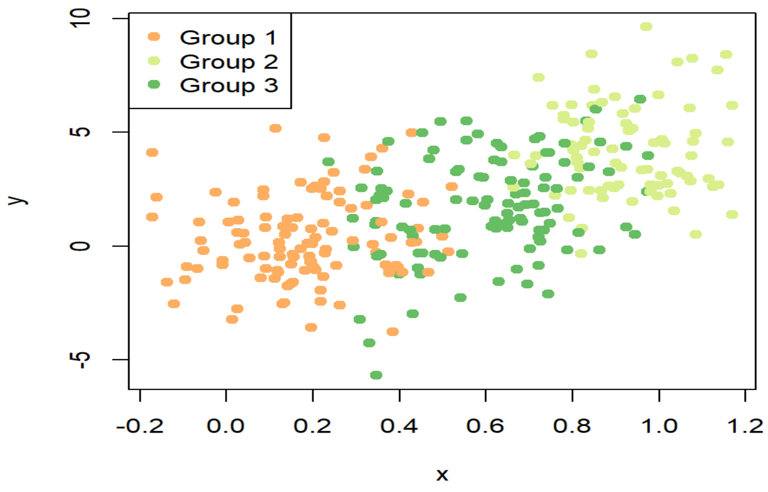
- Boxplots (1 categorical and 1 numerical feature)



Simultaneous description II

(of two features)

- Scatterplot (2 numerical features)



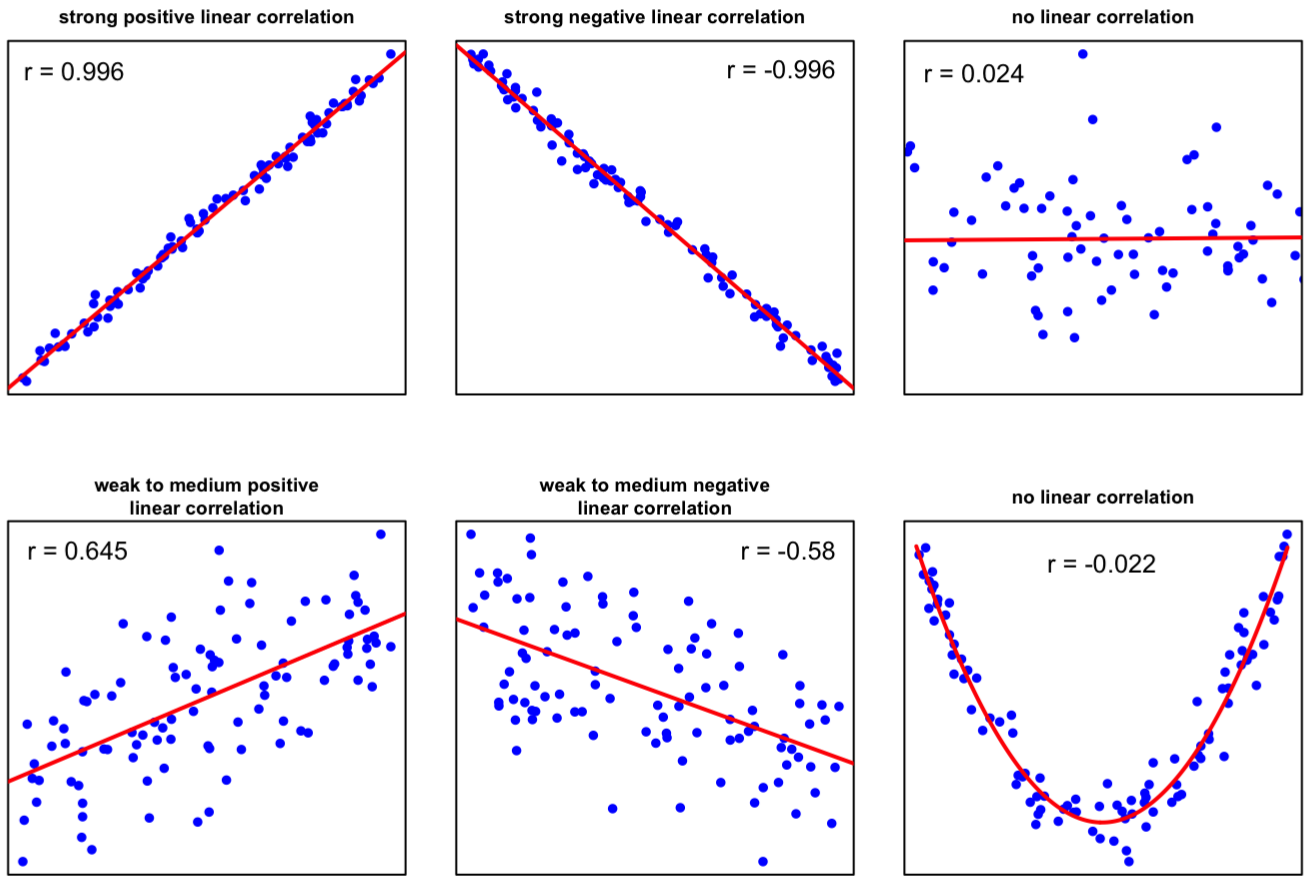
- Pearson Correlation (2 numerical features)

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Simultaneous description III

(of two features)

- Pearson Correlation (2 numerical features)



Probability I

Empirical data is usually regarded as random. Why?

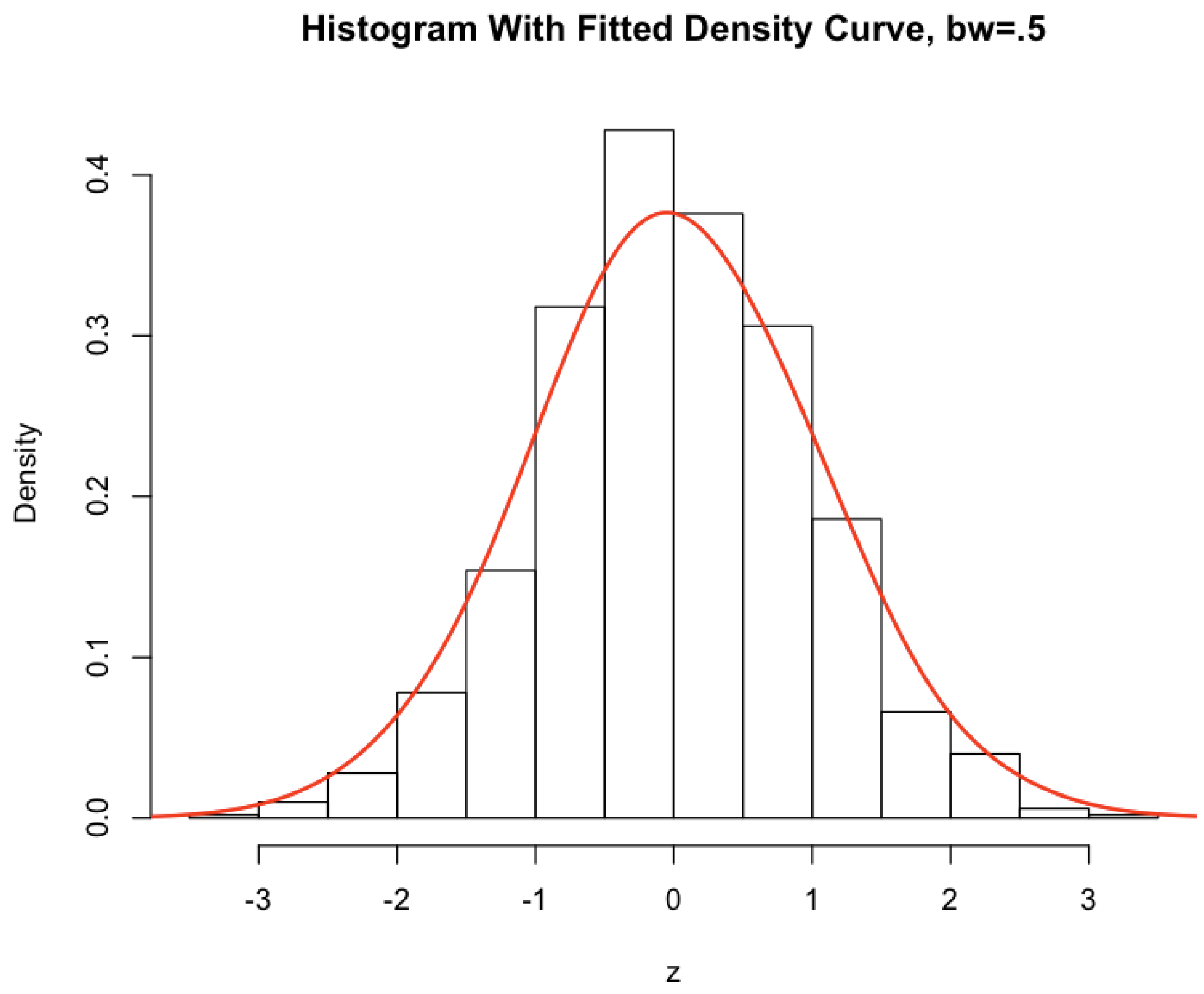
1. Objects under consideration (e.g. persons) as a random sample from a larger population
   ➡ conclusions about the population

2. physical or chemical measurements have random measurement errors
   ➡ must be taken into account to evaluate data reliably

- view observations of features as random draw from population
- feature has a pdf
- use knowledge on this pdf to quantify the uncertainty of our conclusions

Probability II

Probability density function (pdf)



Probability Interpretation

Frequentist

Bayesian

Iris data set

3 classes: versicolor, setosa, virginica

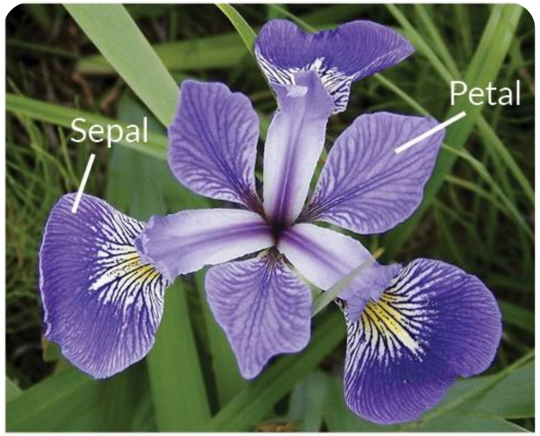
4 characteristics

- petal: *length, width*
- sepal: *length, width*

Let's analyse this data set descriptively

Iris Virginica

Iris Setosa



Iris Versicolor