CAS Applied Data Science – Module 2 – Day 1

Statistical Inference for Data Science

Dr. Anja Mühlemann

27. August 2024

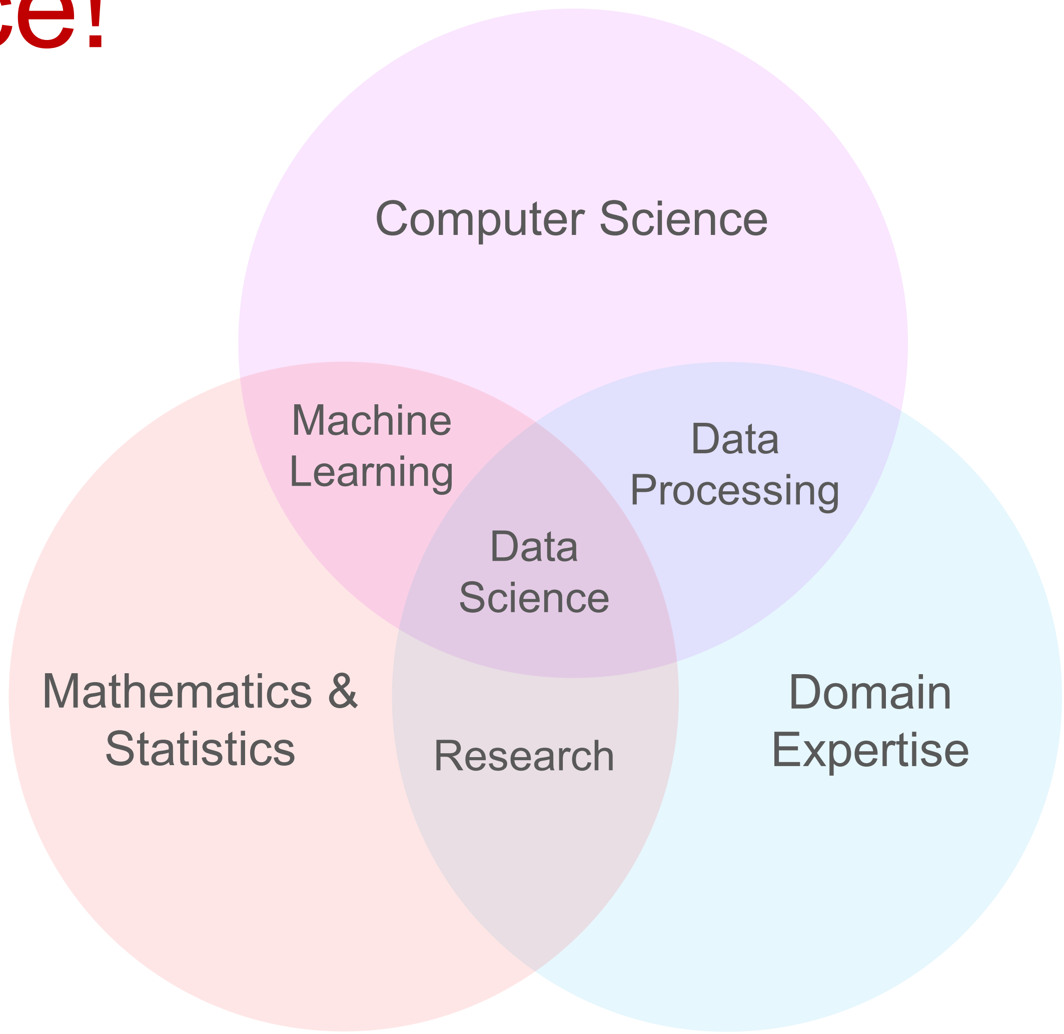
UNIVERSITÄT
BERN

Welcome to Data Science!

Data Science uses

- Mathematics and Statistics
- Computer Science
- Domain expertise

on data to build information and extract knowledge.



Module 2

Tuesday

09:00 - 12:30 Descriptive Statistics
13:30 - 17:00 Notebook 2 (self study)

Wednesday

09:00 - 12:30 Parameter estimation
13:30 - 17:00 Self study

Thursday

09:00 - 12:30 Hypothesis testing
13:30 - 17:00 Prepare test for presentation (self study)

Friday

09:00 - 10:30 Presentations
10:30 - 11:00 Coffee Break
11:00 - 12:30 Notebook 5

Caution

- This module aims to give a brief overview on basic statistics.
- That means in a short amout of time we'll see lot.
- While this may be repetition for some,
- For others there may be a lot of new things.
- I'll try my best to accomodate everyones needs.

Teaching

- Introductory lectures
- In-depth self-study of the content with notebooks
- Discussion sessions based on your questions
  Please ask questions ☺
- I am open to modifications if wished for!

Project

Formal

- Group of 2-3 people
- 15min presentation, 15min discussion
- Half-day presence on presentation session

Content

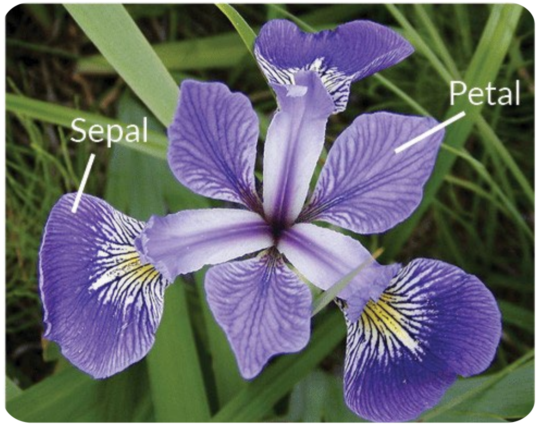
- Choose your own data set
- answer research questions using statistics

Iris data set

- *Due to time restrictions we use a single data set in this module*
- 3 classes: versicolor, setosa, virginica
- 4 characteristics
  petal: *length, width*
  sepal: *length, width*

Iris setosa

Iris virginica



Iris versicolor

Any questions so far?

General Procedure

Planning

↓

Data Collection & Preprocessing

↓

Descriptive Statistics

↓

Classification, Clustering,…

↓

Inferential Statistics

Descriptive Statisics

Why?

- Get an overview of the data
- Identify Patterns
- Identify possible problems eg. outliers
- Get a feeling for the quality of the data

➡ good description is the basis for good inference

Descriptive Statisics

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.

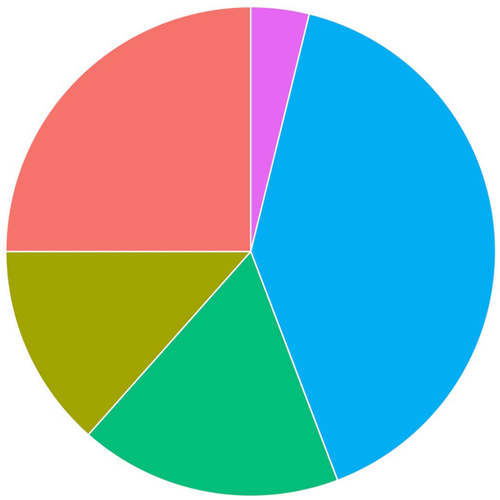
Categorical Variables

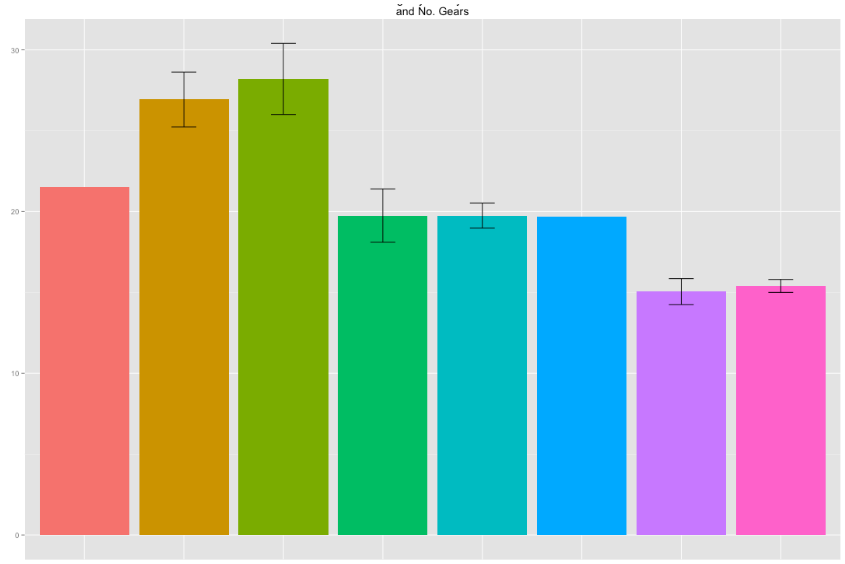
(quantitative)

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

Categorical Variables

(graphical)





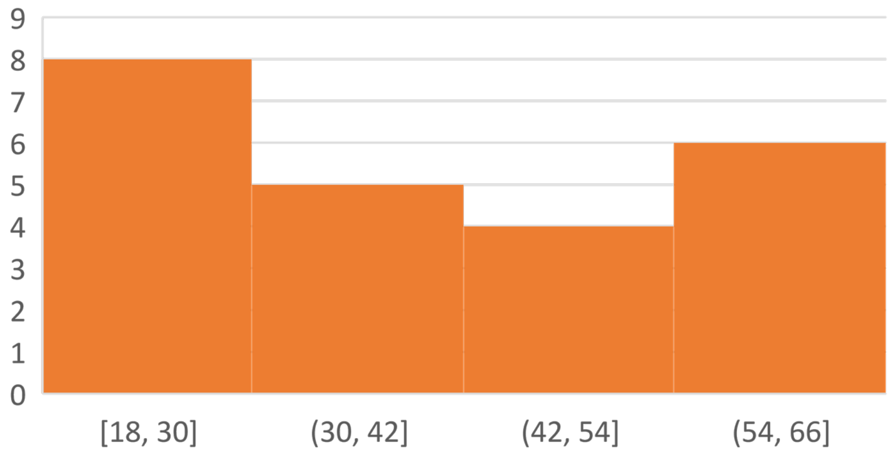
(Either absolute or relative frequencies can be displayed)

Numerical Variables
(categorization)

Summary tables

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

Histograms



Location

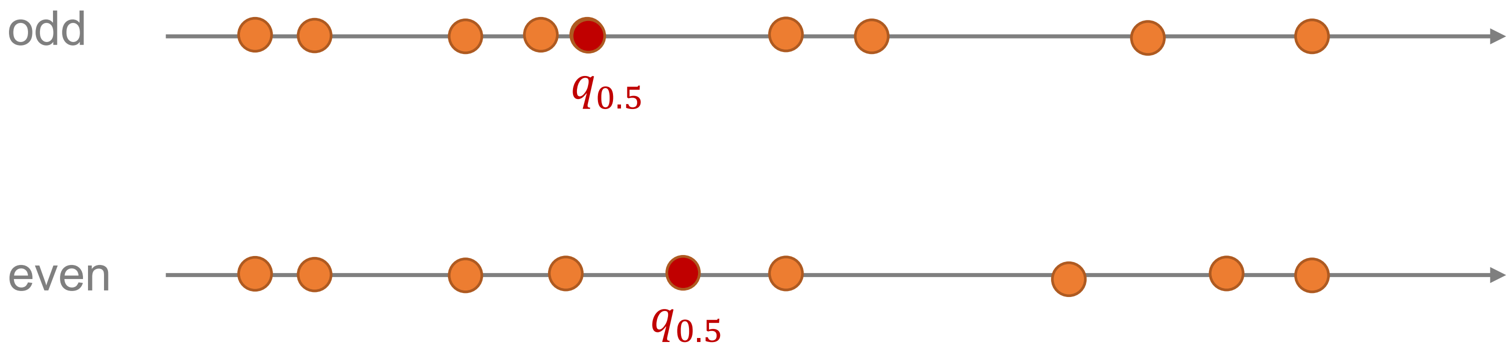
(Numerical Variables)

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

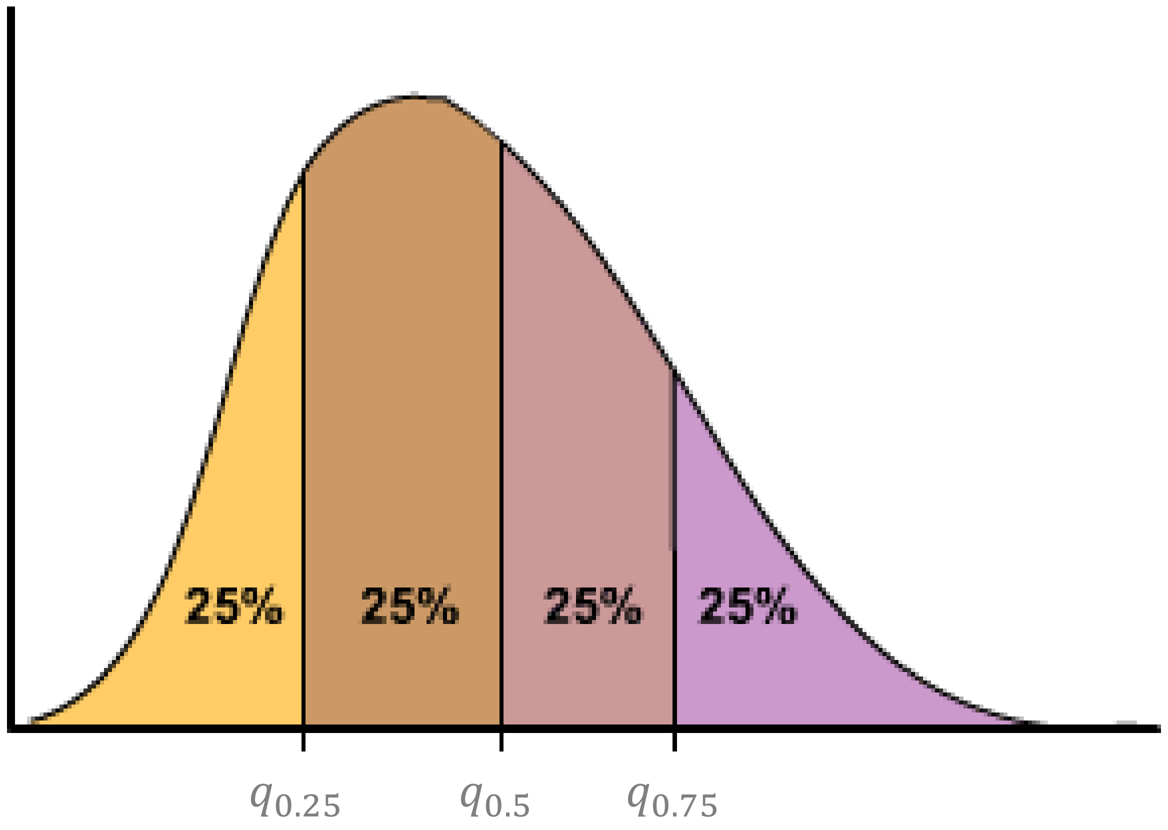
Quantiles

(Numerical Variables)

Generalizing the idea of the median to other fractions.

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

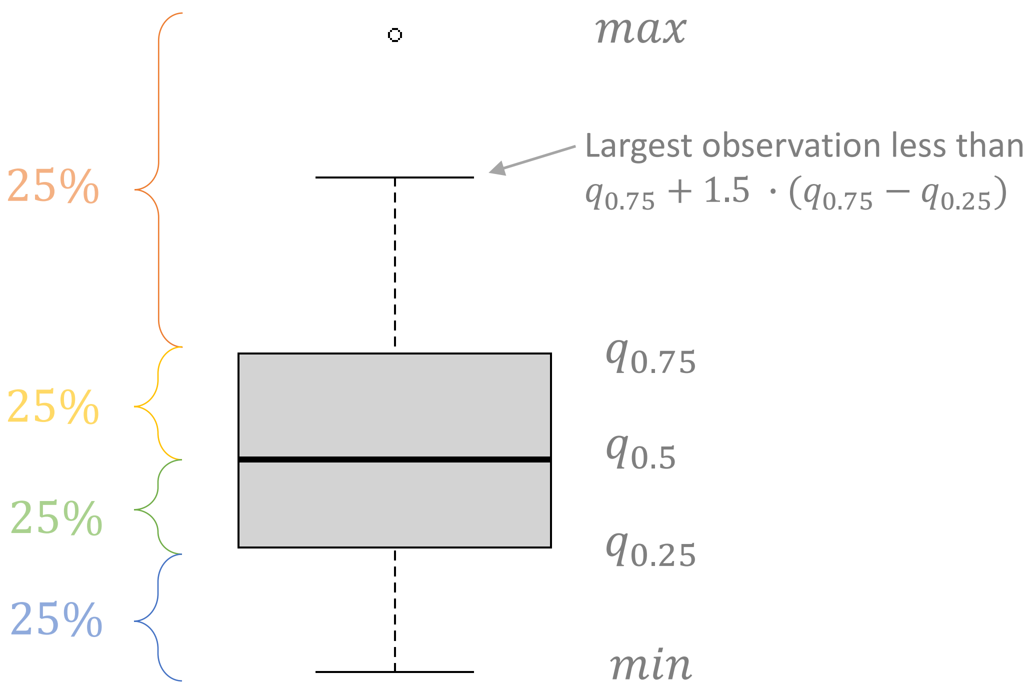
Typical for hypothesis testing: $q_{0.01}, q_{0.05}, q_{0.95}, q_{0.99}$



Boxplots

(Numerical Variables)

Graphical display of the quantiles



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):

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

$S = 1.16, IQR = 1.34$

$S = 4.05, IQR = 5.93$

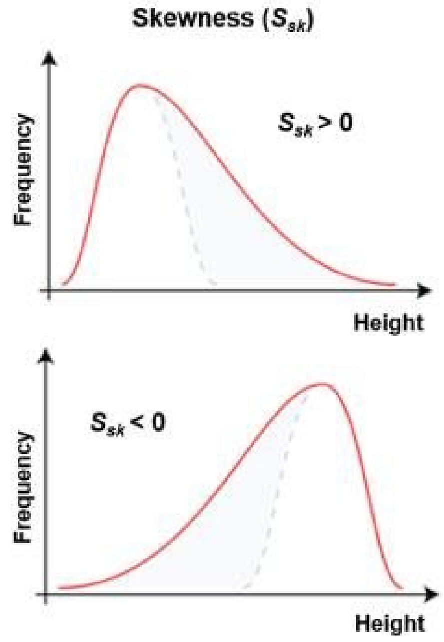
Shape

(Numerical Variables)

Is the distribution symmetric?

- Skewness:

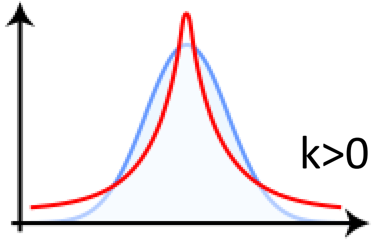
$$S_{sk} = \frac{1}{n}\sum_{i=1}^{n}\frac{(x_i - \bar{x})^3}{s^3}$$

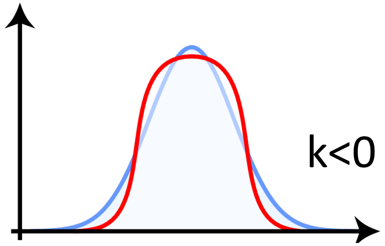


Does the distribution look like a bell curve?

- Kurtosis:

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



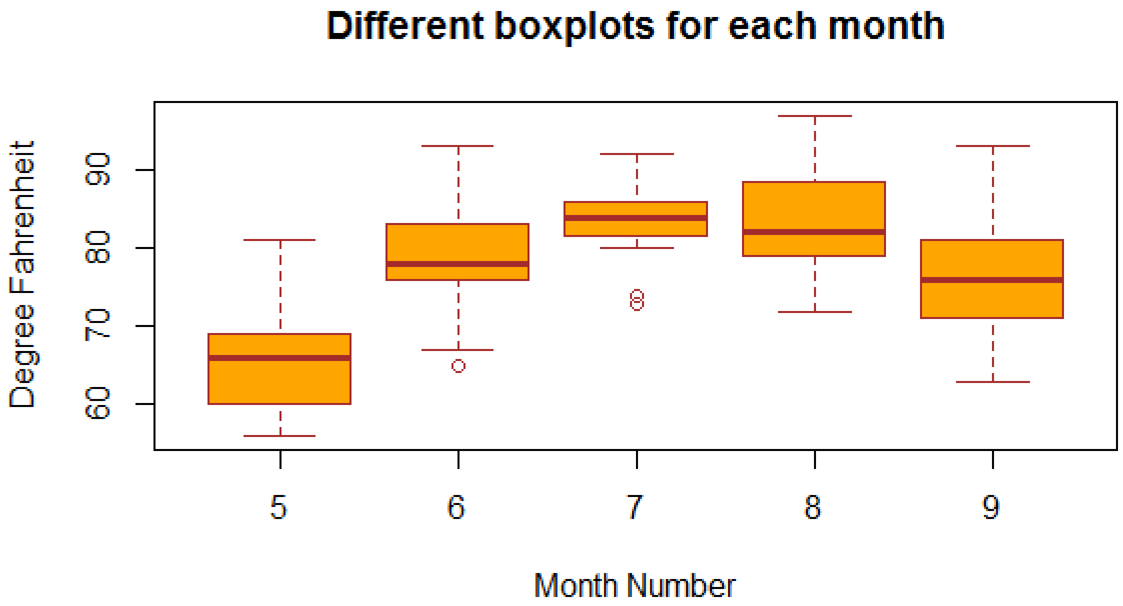


Simultaneous description
(of two features)

- Contigency table (2 categorical features)

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

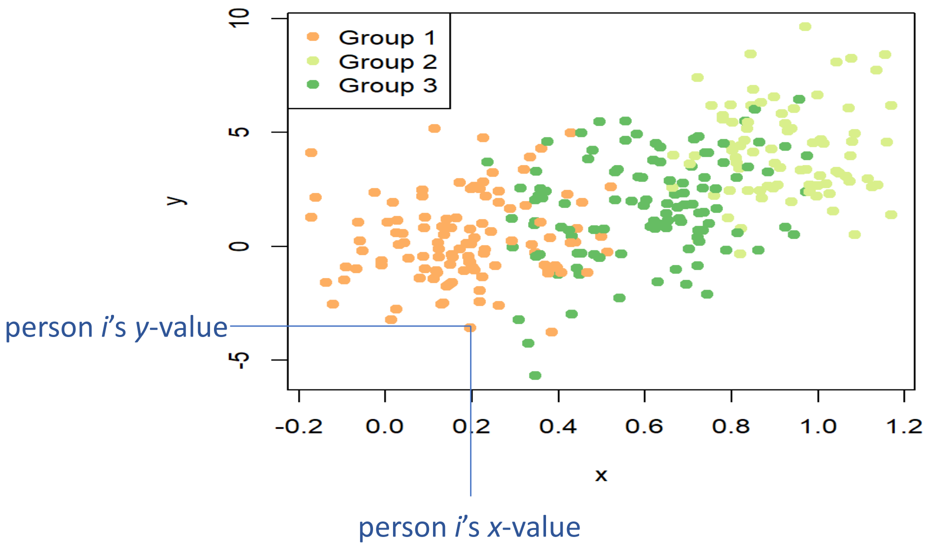
- Boxplots (1 categorical and 1 numerical feature)



Simultaneous description

(of two features)

- Scatterplot (2 numerical features)



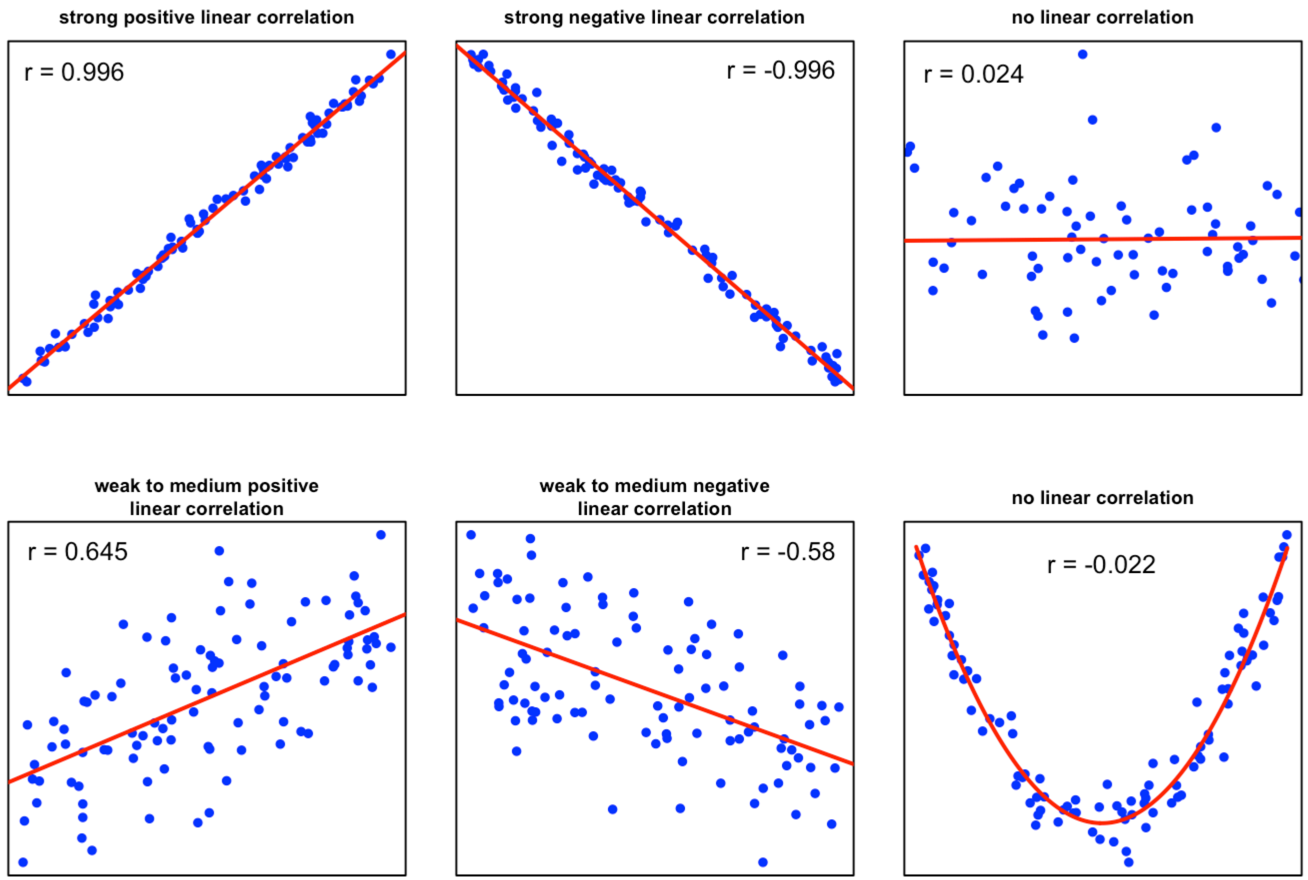
- Pearson Correlation (2 numerical features)

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

Simultaneous description

(of two features)

- Pearson Correlation (2 numerical features)

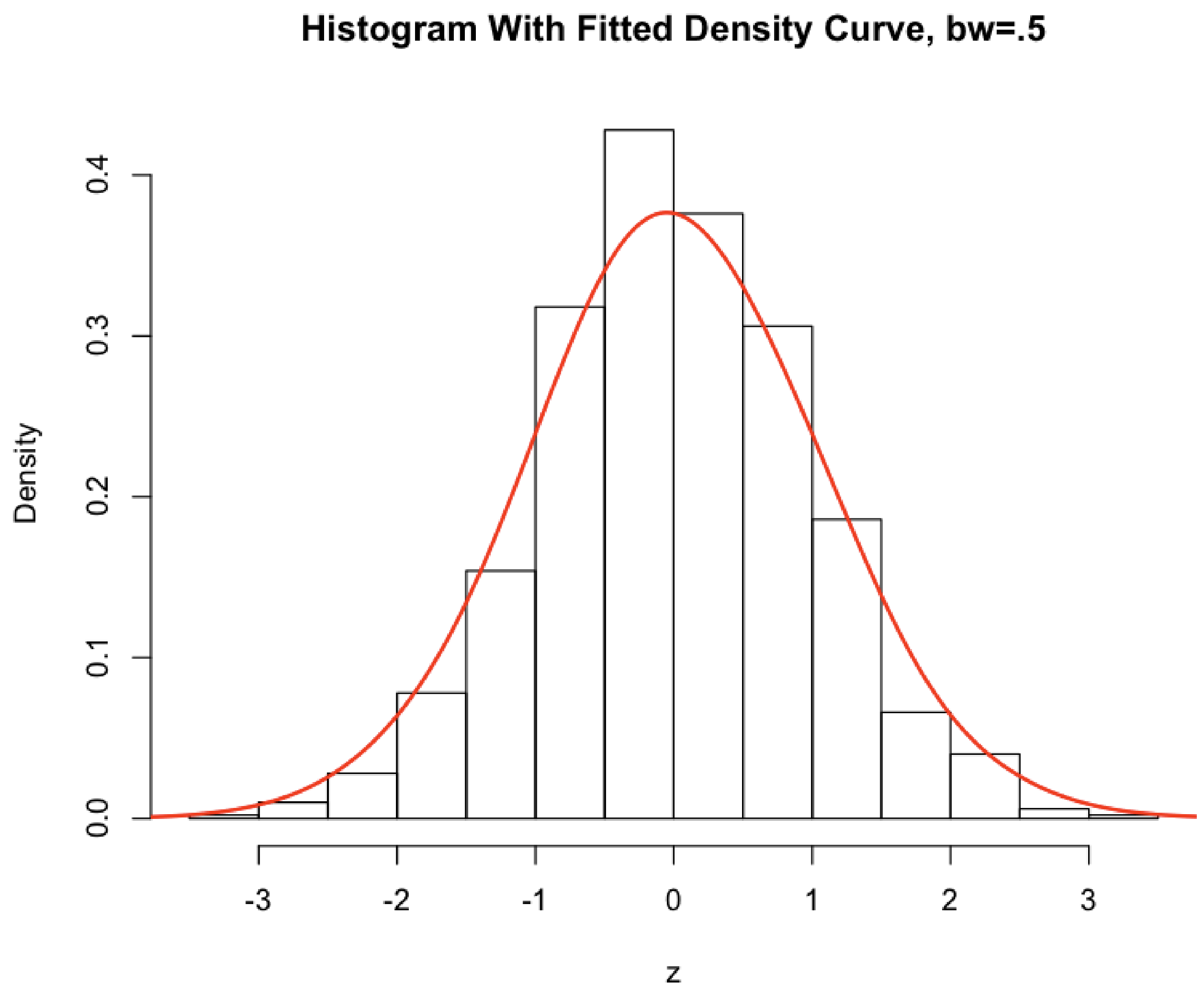


Probability

- Descriptive statistic is an important first step but does not provide us with the means we aim for eventually.
- In general, we want confirm a hypothesis on a population based on sample of said population.
- To this end, we need a mathematical framework for dealing this uncertainty.
- To quantify the uncertainty one often works with probability distributions.

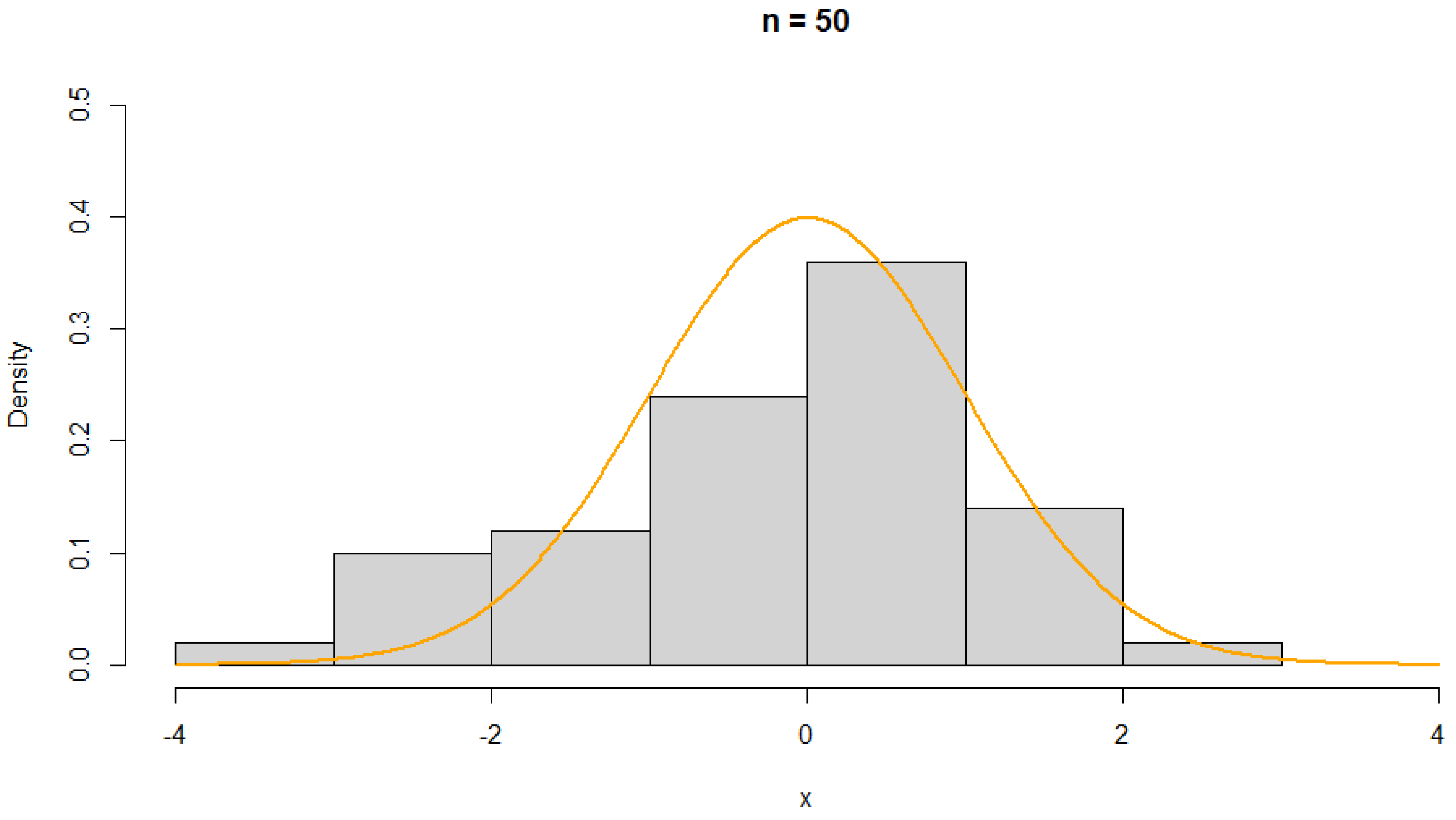
Probability

Probability density function (pdf)



Probability

Probability density function (pdf)



Sketch of idea