CAS Applied Data Science – Module 2 – Day 1

Statistical Inference for Data Science

Dr. Anja Mühlemann

29. August 2023

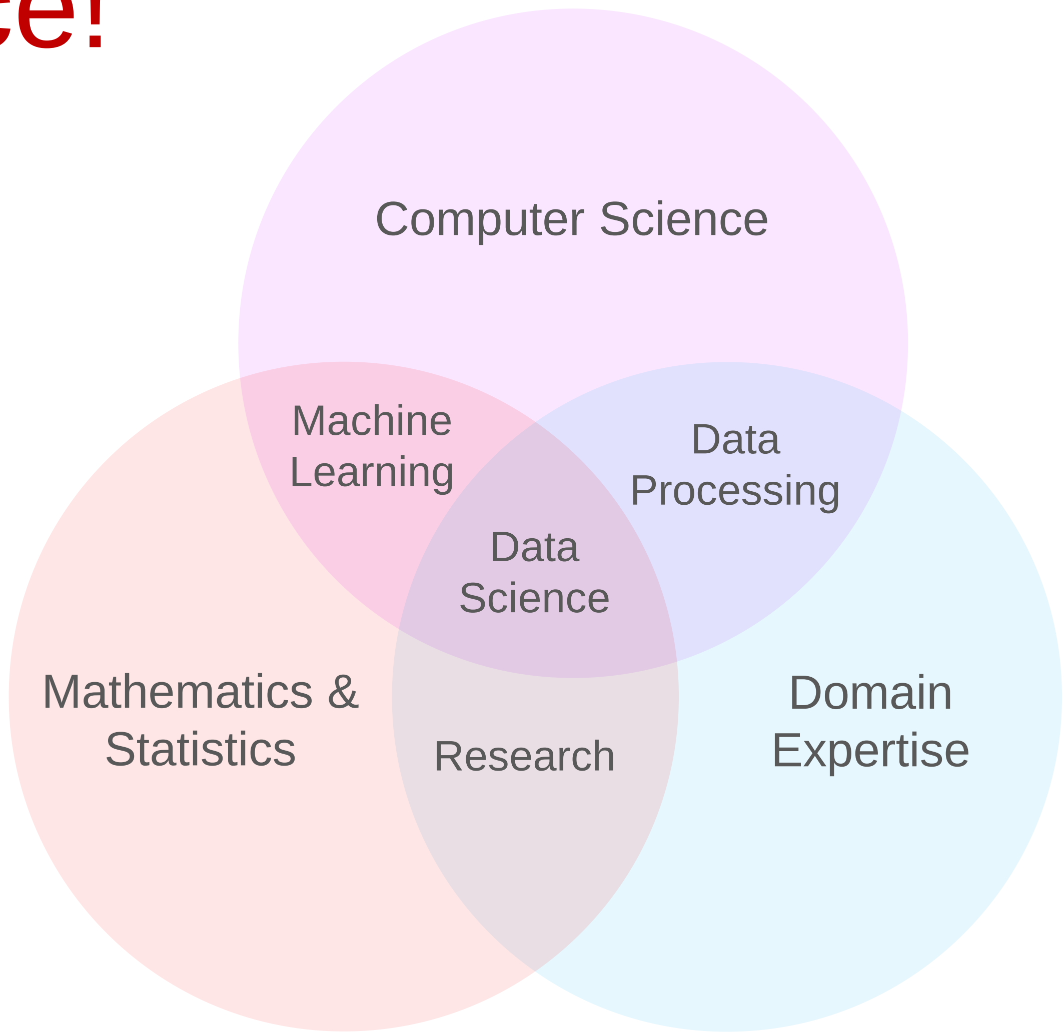
UNIVERSITÄT
BERN

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	Presentation session (date to be fixed!)

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

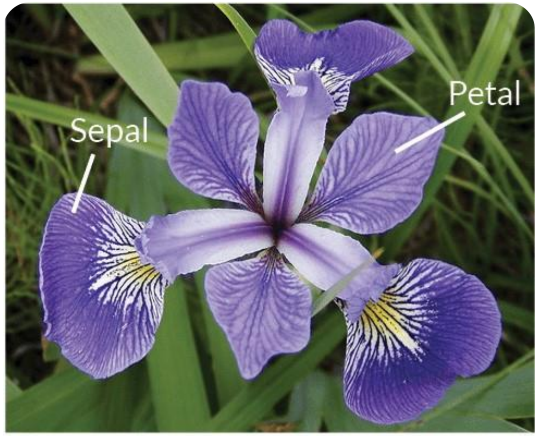
- *Due to time restrictions we use a single data set in this module*

- 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

What do I want to investigate?How can I investiga

Planning

Data Collection & Preprocessing

Helps to find some problems and act quickly.

Descriptive Statistics

Classification, Clustering,…

We want to test our hypothesis: after describing the data we use them to get a conclusion.

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

Usually not more than 2D

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

There is two kind of variables:- Categorical: they represent the categ

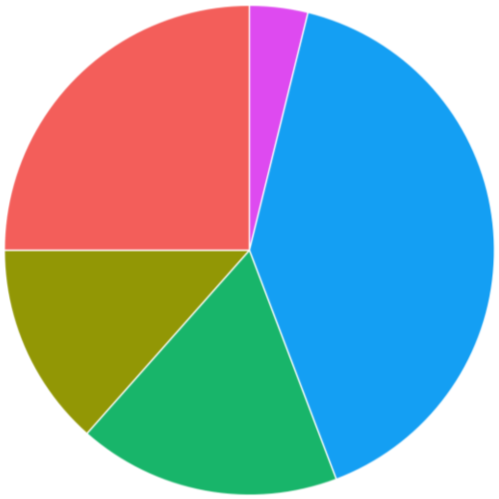
Categorical Variables

(quantitative)

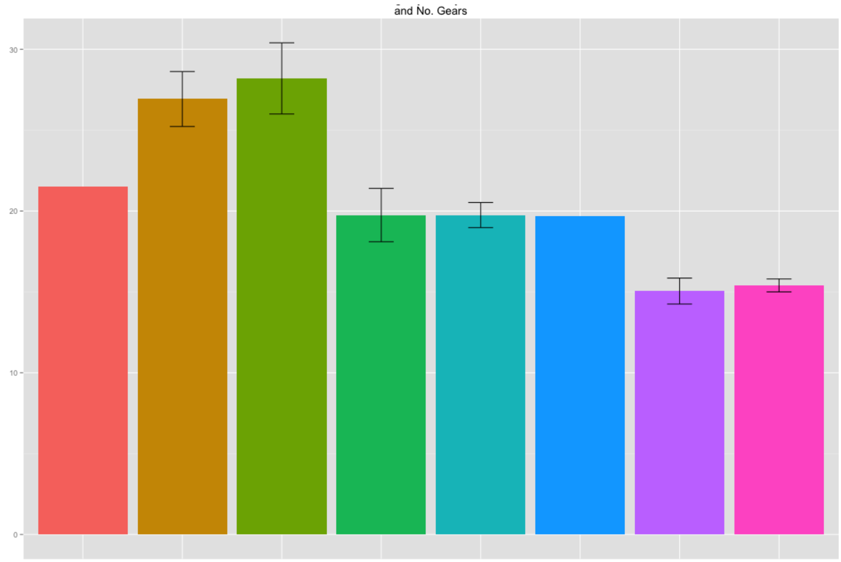
- Absolute frequency (eg. number of female participants)
- Relative frequency (eq. number of female participants divided by the sample size)

Categorical Variables
(graphical)

pie chart: nice but sometimes
may be difficult to grasp differences
between the data (see example in notebook)



With more than 2-3 slices, better to go with a bar chart



(Either absolute or relative frequencies can be displayed)

Numerical Variables

(categorization)

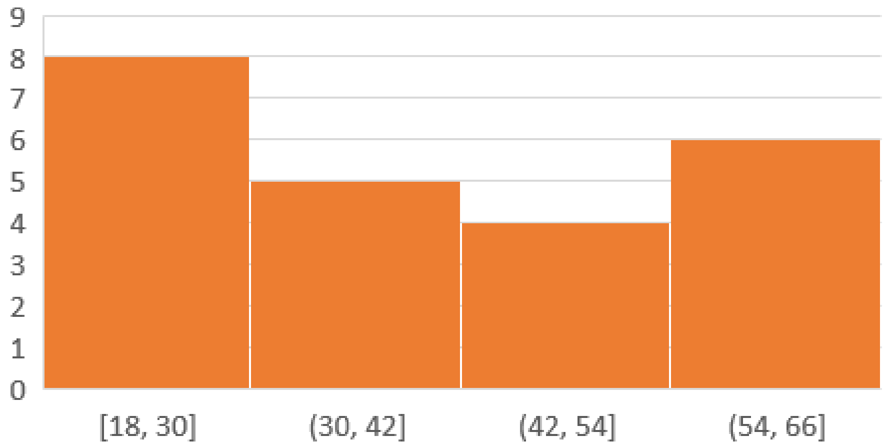
There is different measures depending on the system we are studying.

Summary tables

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

We could cluster and reduce a numerical variable in a categorical variable but we would lose some information.

Histograms



The result of a clustering is often a histogram.

Location

(Numerical Variables)

Where in the x line lies our variable

Location alone gives some information but not all the informations

What are typical values for the variable X?

- Sample Mean:

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

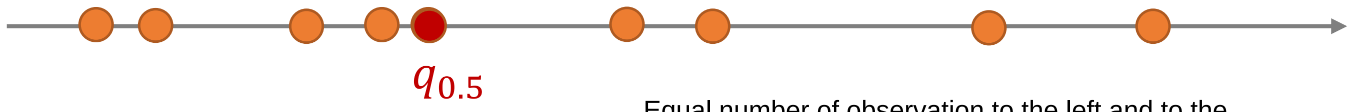
Mean describe location, the size of the variable, where my numbers are distributed around.

The mean though has a defect: it tends to be drag by the big and small numbers.

- Sample Median: «center of the observations»

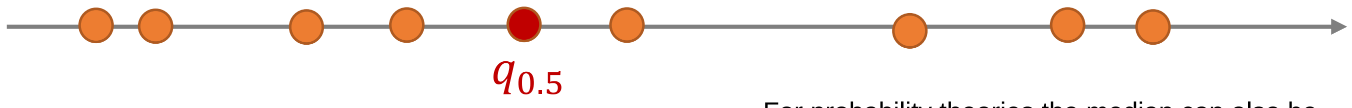
For variables that are skewed (lots of observation together and few large/small observation) it is better to use the median.

odd



Equal number of observation to the left and to the right of the median

even



For probability theories the median can also be a list of values.

➡ median ist more robust than the mean

Quantiles

(Numerical Variables)

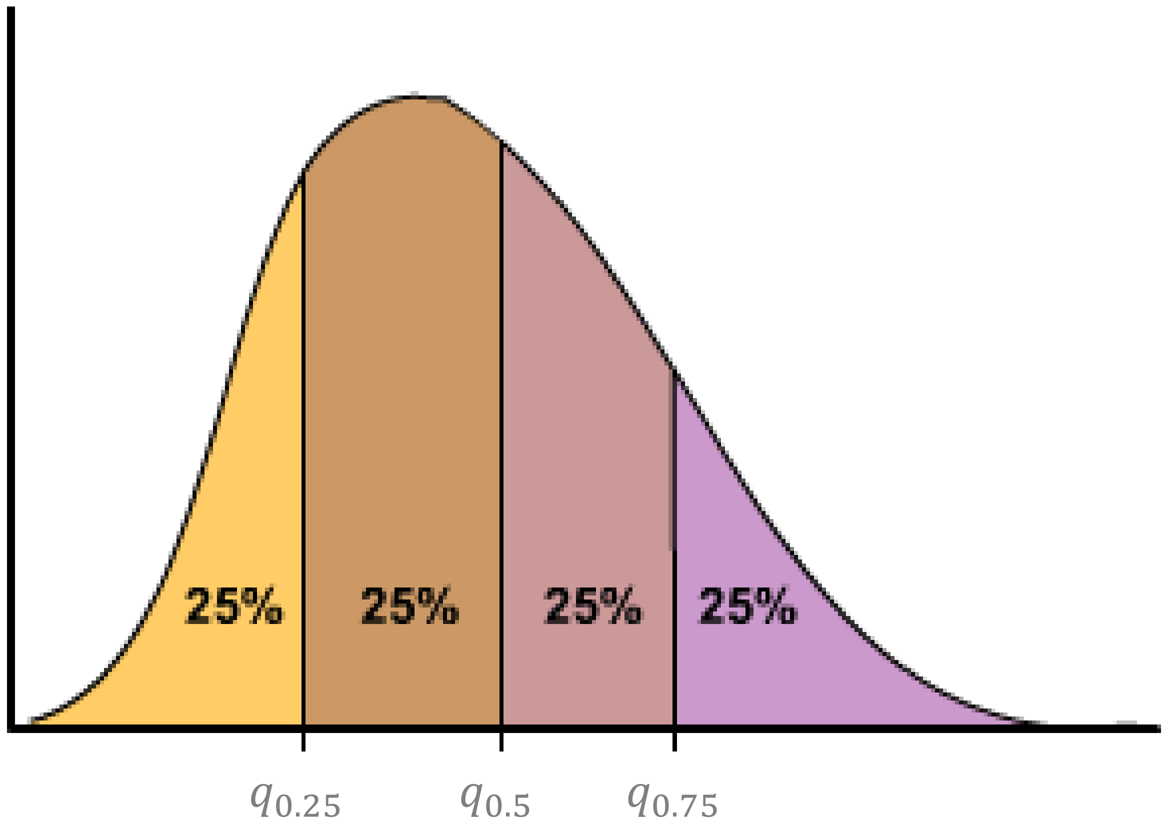
A quantile is a value where a certainqu

Generalizing the idea of the median to other fractions.

Every quantile is in principle possible but for some analysis some are more commonly used.

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

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



As a general rule we should not cancel out outliers unless we have a good reasoning behind.

Boxplots

(Numerical Variables)

Graphical display of the picture before

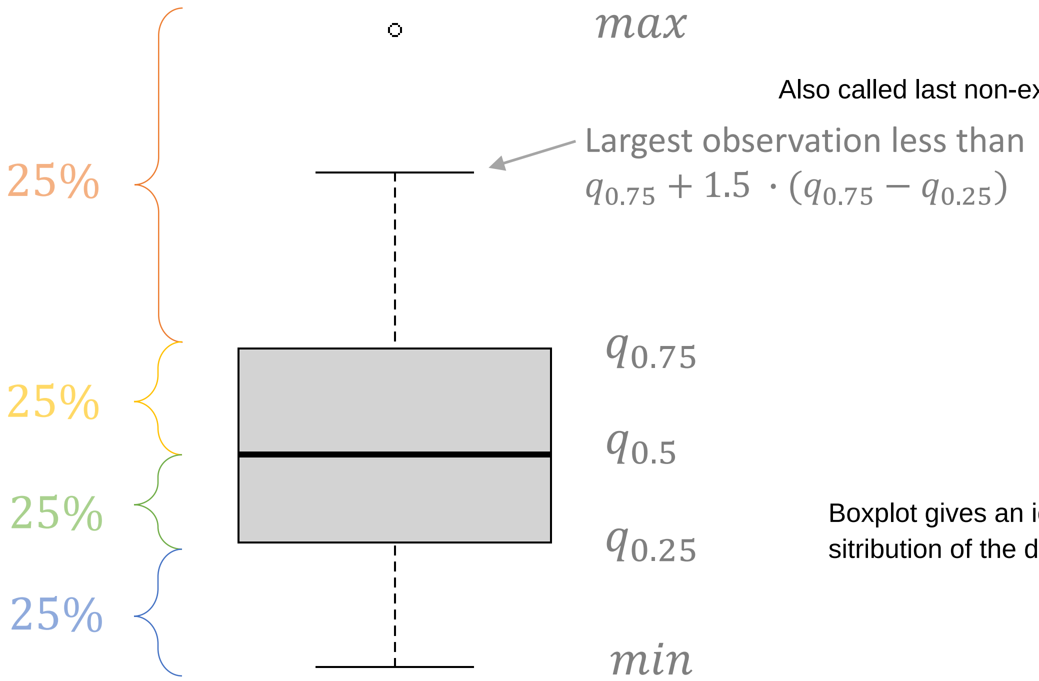
Boxplots are often used for the quarters

Graphical display of the quantiles

When the points lies away from q0.75+1.5(q0.75-q0.25) and same actually from the bottom, the data are considered by Python outliers and will be drown as individual points.

in some whiskers the outliers are drown separately

Also called last non-extreme observation



% of observations

Boxplot gives an idea of the stiribution of the data

Spread

(Numerical Variables)

How strong is the deviation from the center?

- Sample standard deviation:

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

Without the square root it is called thevariant.B

- IQR (inter quartile range):

Quartile is a quantile that is one of the quarters.

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

$S = 1.16, IQR = 1.34$

$S = 4.05, IQR = 5.93$

Shape

(Numerical Variables)

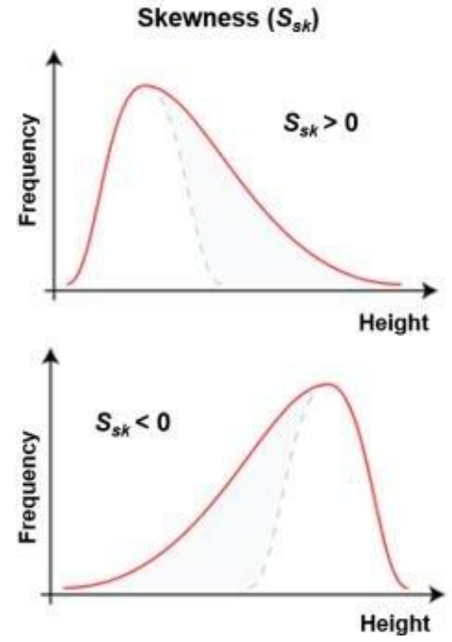
Is the distribution symmetric?

- Skewness:

Third power

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

normalizing helps
to analyze with the data



Suggestion is not to check only the skewness but to the histogram too.
Skewness useful especially for unimodal distribution

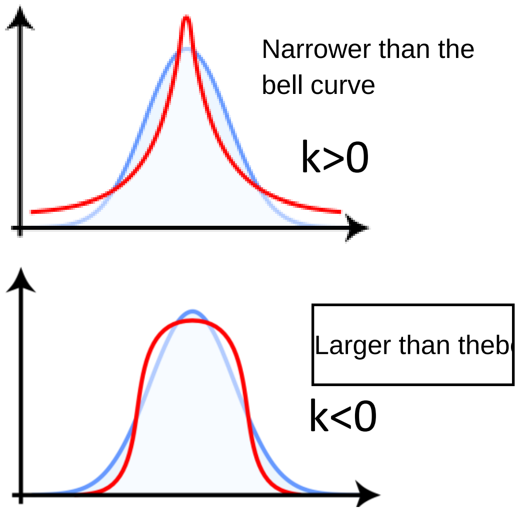
Does the distribution look like a bell curve?

- Kurtosis:

Fourth power

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

Bell curve is the normal distribution.



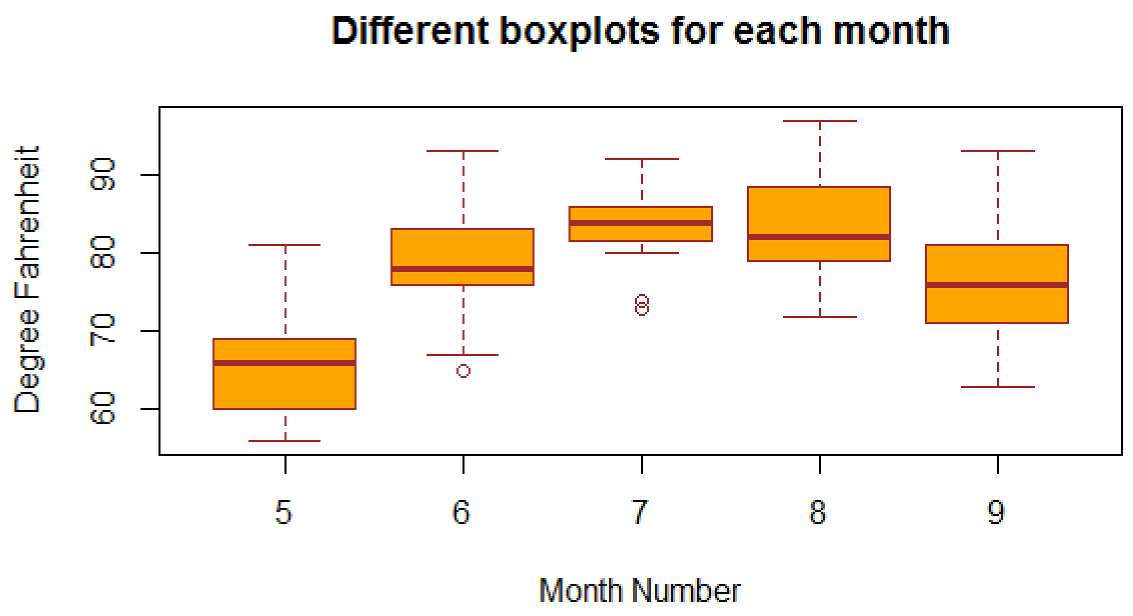
Simultaneous description

(of two features)

- Contigency table (2 categorical features)

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

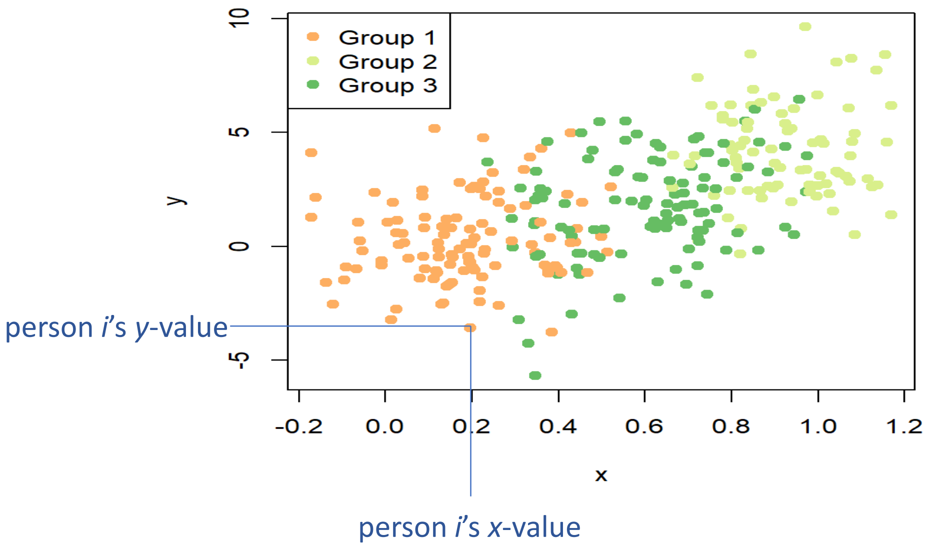
- Boxplots (1 categorical and 1 numerical feature)



Simultaneous description

(of two features)

- Scatterplot (2 numerical features)



- Pearson Correlation (2 numerical features)

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

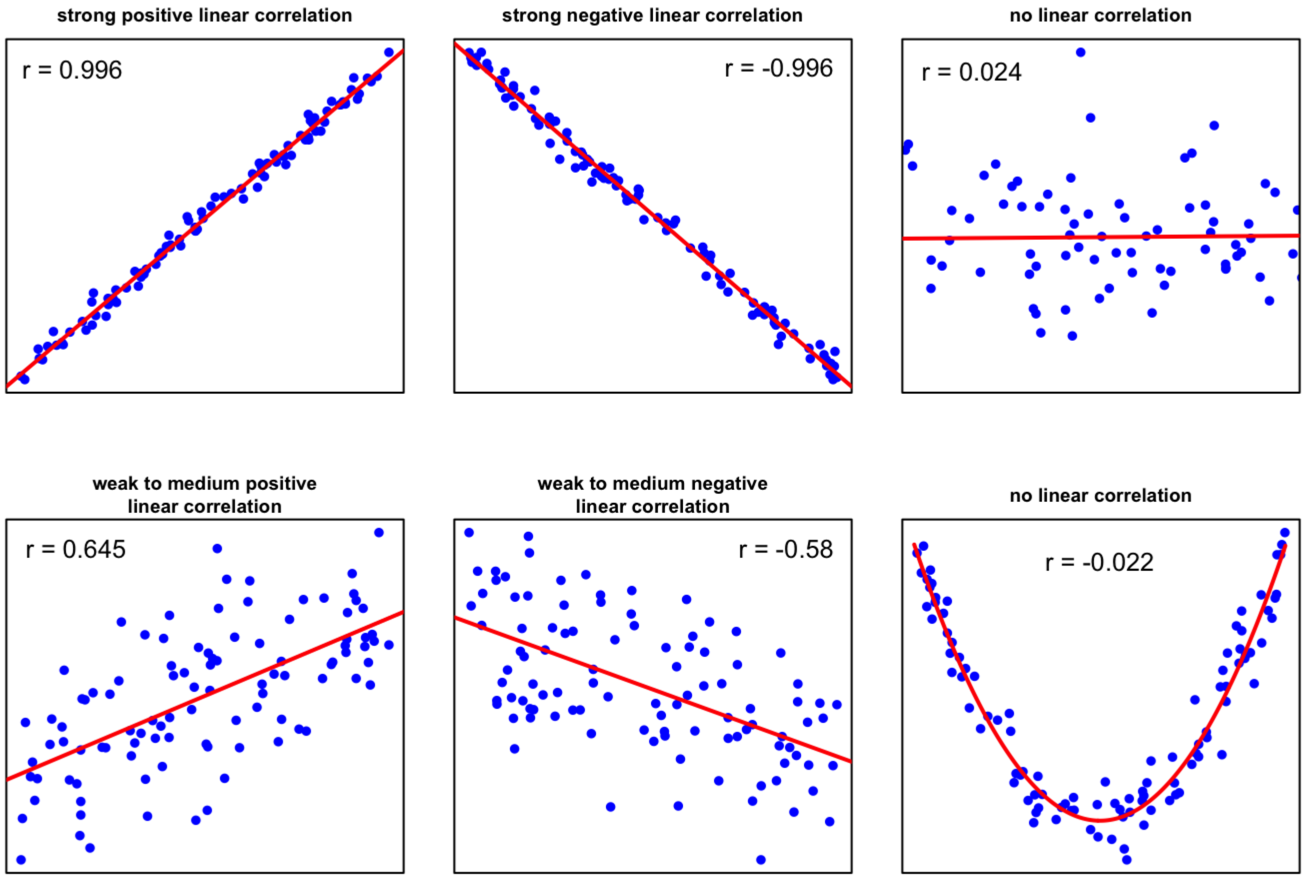
Often done to show that there is some correlation between the two variables.

Simultaneous description

(of two features)

- ## Pearson Correlation (2 numerical features)

When the points are closed to the line the r tends to 1. r can be from -1 to 1, the sign tells us if is an increasing or a decreasing relationship. The correlation numerically is only from 0 to 1. If the No. is close to 0 there is no correlation.



Pearson correlation is not good for curves like this. It s done for linear correlation only. In this case the Pearson correlation coefficient would be close to 0 but there is a clear correlation.

In statistic the results obtained from a populationdriv

Probability

Probability theory gives the theoretical background to move from the sample to the entire population.

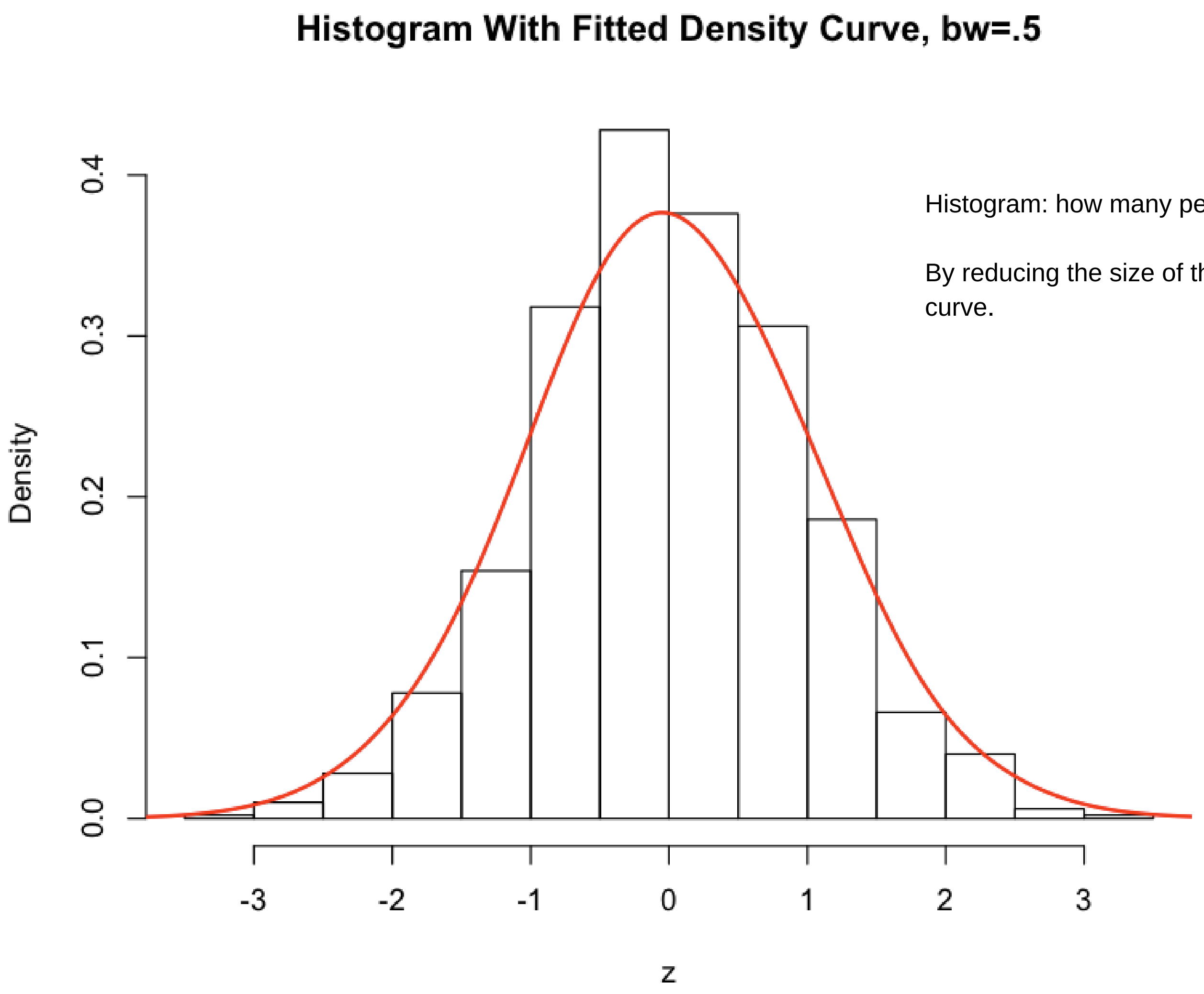
Probability works with the distributions. This allows also to evaluate how representative was the sample.

- 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

Careful: histogram can hide a lot, by different binning some data can be shown or hidden, our target is to show the most that we can to have an accurate hypothesis.

Probability density function (pdf)



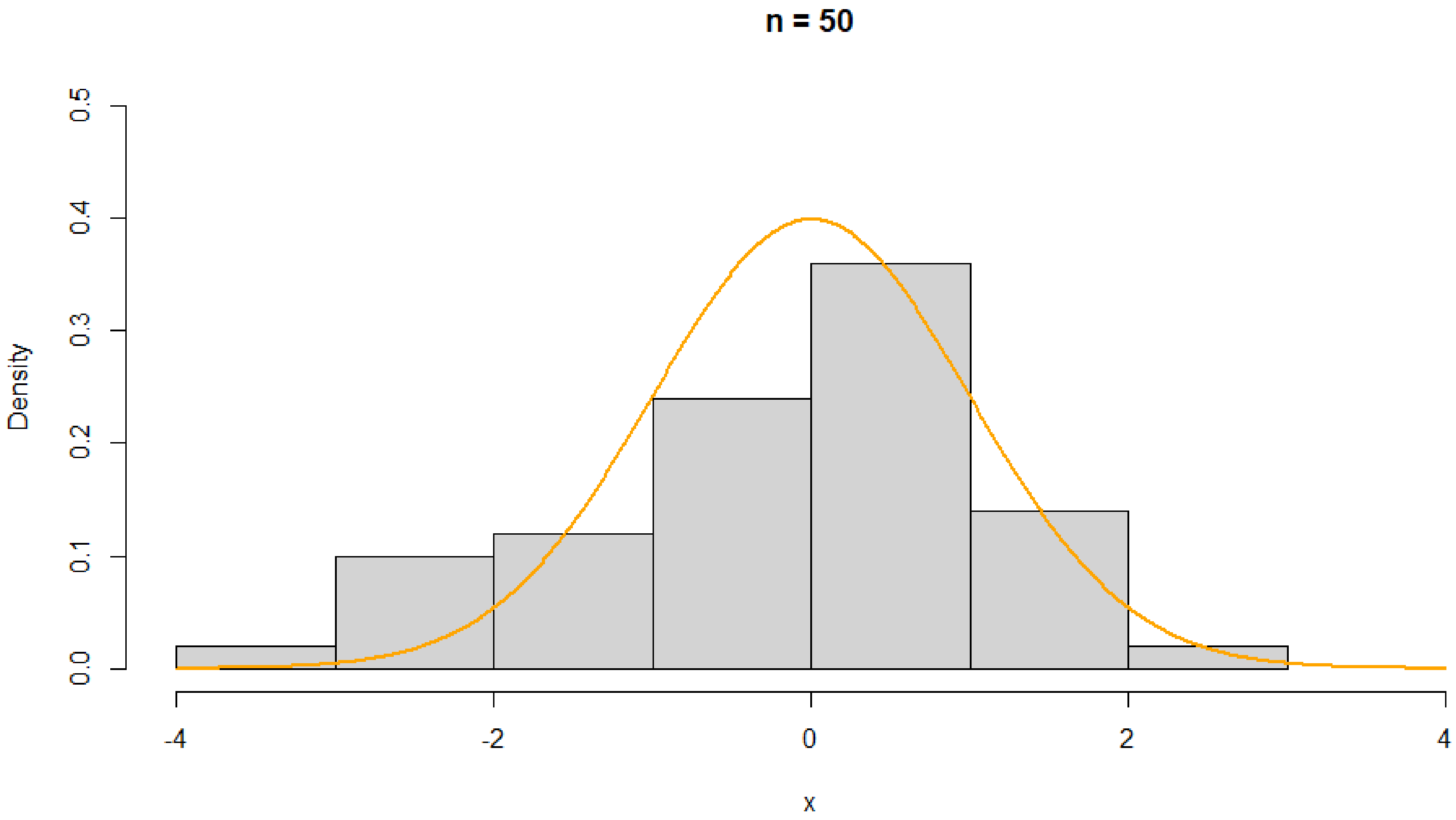
Histogram: how many people are in each bin.

By reducing the size of the bin we compute the curve.

Probability

In probability usually i is not proof that the hypothesis is correct but it is proven that the opposite is wrong.

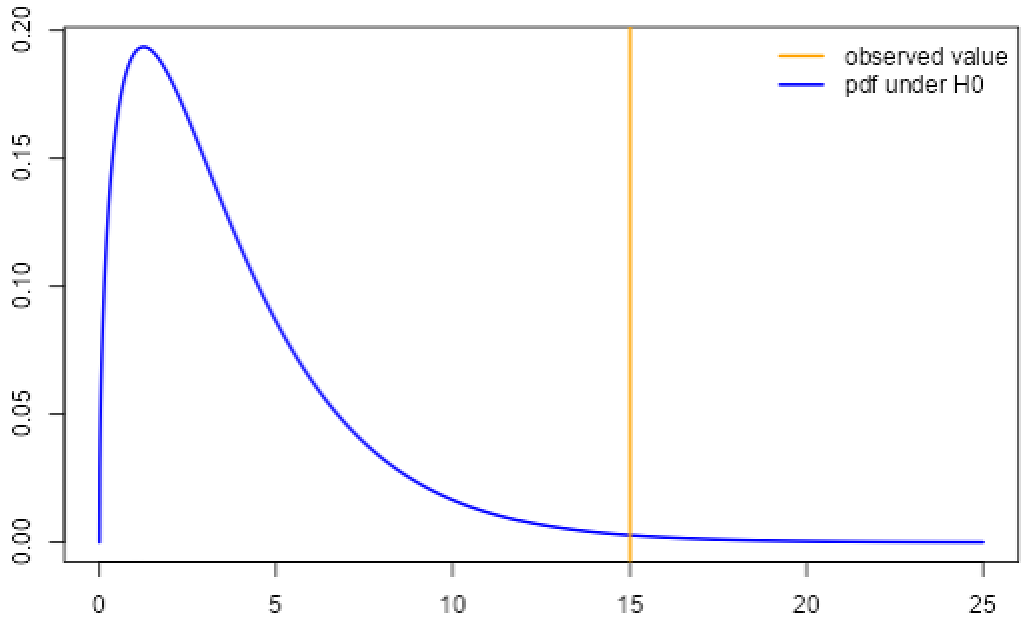
Probability density function (pdf)

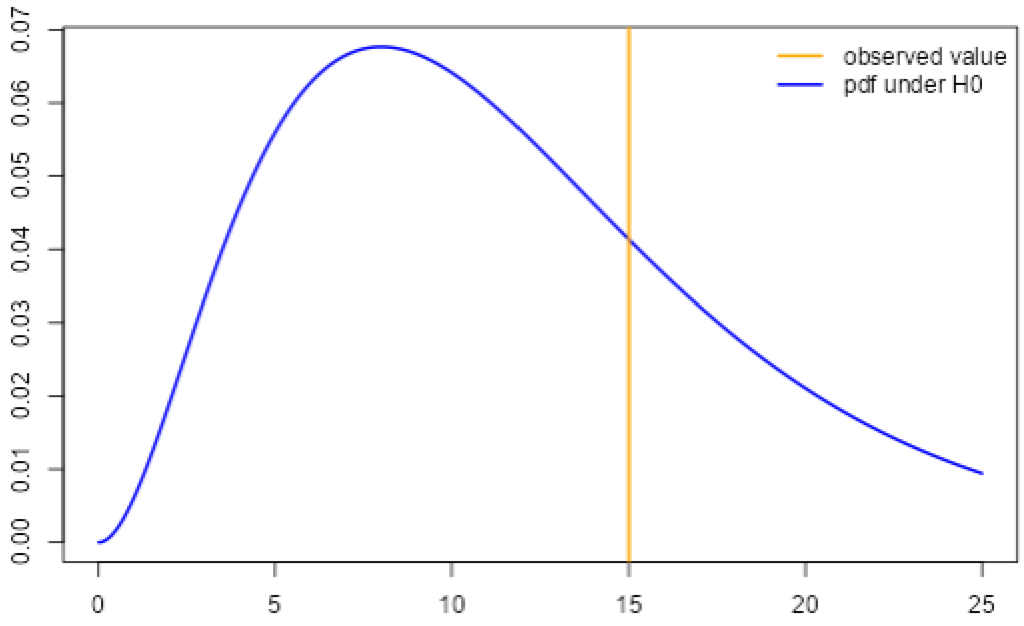


Tests are developed basing on assumptions built with the data knowledge.

In the example at the side we hypothise that the distribution will look like the one below. We test if our value with another distribution (like the one above) would give the planned result, if this fails means our theory holds.
We assign to this probability a value.

Sketch of idea





Summarizing: statistic is used indirectly as a tool to build a probability hypothesis to be then verified by the probability theory.
How good an hypothesis is, depends by how well the data were statistically analyzed.

NOTE: we are speaking here of frequency probability.