Intelligent Data Analysis

Christian Braune, M.Sc.
Pascal Held, M.Sc., Alexander Dockhorn, B.Sc.

Computational Intelligence Group
Faculty of Computer Science
cbraune@iws.cs.ovgu.de

About me

2010 Bachelor of Science, Magdeburg, Germany

2012 Master of Science, Magdeburg, Germany

since then: Trying to get a PhD

Research: neuro-informatics, clustering, data analysis

Organisational

Lecture

Tuesday, 9.15-10.45, G29-E037

Christian Braune

Consultation: everytime, I'm in my office

Preferred way of contact: personal

Exercises

Tuesday 13.15-14.45 G29-K058

Tutor: Pascal Held

at the moment: Friday 9.15-10.45 G29-K059

Tutor: Alexander Dockhorn

Updated Information on the Course

`http://fuzzy.cs.uni-magdeburg.de/wiki/pmwiki.php?n=Lehre.IDA2015`

Acknowledgement

We thank Christian Borgelt for providing several slides for this course, that he produced during his time as a scientific researcher in our institute.

Organisational

Conditions for Certificates (Scheine)

ticked at least two thirds of the assignments,

presented at least twice a solution during the exercise, and

passed a small colloquium (approx. 10 min) or a written test (if there are more than 20 students) after the course.

Conditions for Exams

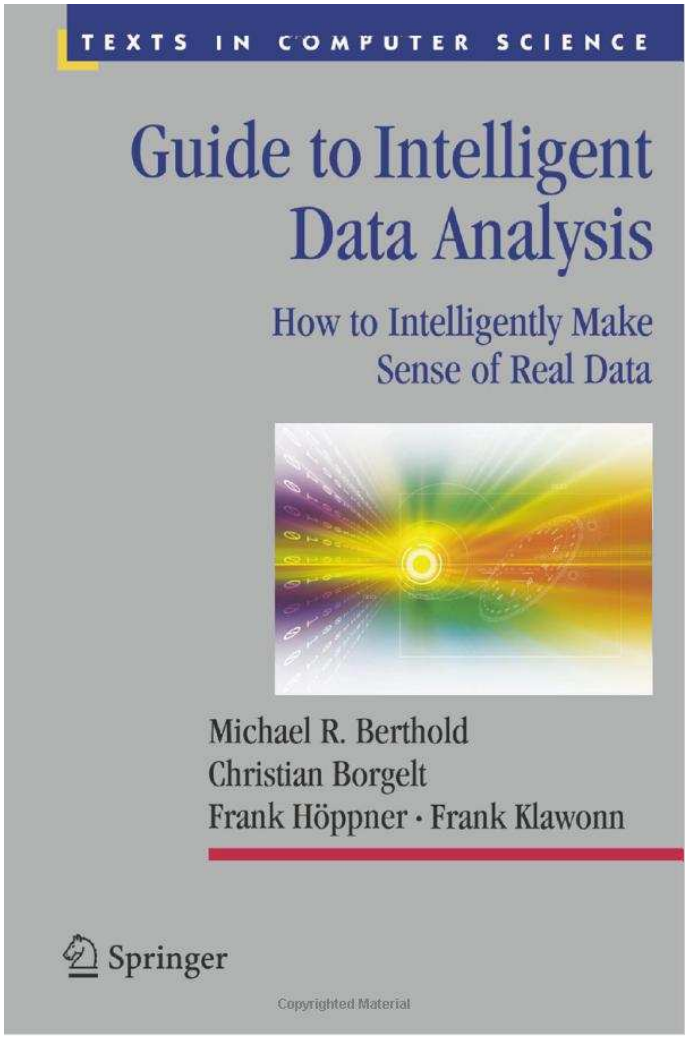
ticked at least two thirds of the assignments,

presented at least twice a solution during the exercise, and

passed a colloquium (approx. 20 min) or a written test (if there are more than 20 students) after the course.

Book

Berthold, Borgelt, Höppner, Klawonn: Guide to Intelligent Data Analysis, Springer 2011



Introduction

Data and Knowledge

- Characteristics and Differences of Data and Knowledge
- Quality Criteria for Knowledge
- Example: Tycho Brahe and Johannes Kepler

Knowledge Discovery and Data Mining

- How to Find Knowledge?
- The Knowledge Discovery Process (KDD Process)
- Data Analysis / Data Mining Tasks
- Data Analysis / Data Mining Methods

Summary

Introduction

Today every enterprise uses electronic information processing systems.

- Production and distribution planning
- Stock and supply management
- Customer and personnel management

Usually these systems are coupled with a database system
(e.g. databases of customers, suppliers, parts etc.).

Every possible individual piece of information can be retrieved.

However: **Data alone are not enough.**

- In a database one may "not see the wood for the trees".
- General patterns, structures, regularities go undetected.
- Often such patterns can be exploited to increase turnover
  (e.g. joint sales in a supermarket).

Data

Examples of Data

"Columbus discovered America in 1492."

"Mr Jones owns a Volkswagen Golf."

Characteristics of Data

refer to single instances
(single objects, persons, events, points in time etc.)

describe individual properties

are often available in huge amounts (databases, archives)

are usually easy to collect or to obtain
(e.g. cash registers with scanners in supermarkets, Internet)

do not allow us to make predictions

Examples of Knowledge

"All masses attract each other."

"Every day at 5 pm there runs a train from Magdeburg to Berlin."

Characteristic of Knowledge

refers to *classes* of instances
(*sets* of objects, persons, points in time etc.)

describes general patterns, structure, laws, principles etc.

consists of as few statements as possible (this is an objective!)

is usually difficult to find or to obtain
(e.g. natural laws, education)

allows us to make predictions

Criteria to Assess Knowledge

Not all statements are equally important, equally substantial, equally useful.

$\Rightarrow$ Knowledge must be assessed.

Assessment Criteria

Correctness (probability, success in tests)

Generality (range of validity, conditions of validity)

Usefulness (relevance, predictive power)

Comprehensibility (simplicity, clarity, parsimony)

Novelty (previously unknown, unexpected)

Priority

The priorities of criteria in science and economy are different

Optimality is not always necessary in economy

Economy often focuses more on usefulness, comprehensibility, novelty

Tycho Brahe (1546–1601)

Who was Tycho Brahe?

Danish nobleman and astronomer

In 1582 he built an observatory on the island of Ven (32 km NE of Copenhagen).

He determined the positions of the sun, the moon and the planets
(accuracy: one angle minute, without a telescope!).

He recorded the motions of the celestial bodies for several years.

Brahe's Problem

He could not summarize the data he had collected in a uniform and consistent scheme.

The planetary system he developed (the so-called Tychonic system) did not stand the test of time.

Johannes Kepler (1571–1630)

Who was Johannes Kepler?

German astronomer and assistant of Tycho Brahe

He advocated the Copernican planetary system.

He tried all his life to find the laws that govern the motion of the planets.

He started from the data that Tycho Brahe had collected.

Kepler's Laws

Each planet moves around the sun in an ellipse, with the sun at one focus.

The radius vector from the sun to the planet sweeps out equal areas in equal intervals of time.

The squares of the periods of any two planets are proportional to the cubes of the semi-major axes of their respective orbits: $T \sim a^{\frac{3}{2}}$.

How to find Knowledge?

We do not know any universal method to discover knowledge.

Problems

Today huge amounts of data are available in databases.

We are drowning in information,
but starving for knowledge. John Naisbett

Manual methods of analysis have long ceased to be feasible.

Simple aids (e.g. displaying data in charts) are too limited.

Attempts to Solve the Problems

Intelligent Data Analysis

Knowledge Discovery in Databases

Data Mining

Knowledge Discovery and Data Mining

As a response to the challenge raised by the growing volume of data a new research area has emerged, which is usually characterized by one of the following phrases:

Knowledge Discovery in Databases (KDD)

Usual characterization:

KDD is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. [Fayyad et al. 1996]

Data Mining

- Data mining is that step of the knowledge discovery process in which data analysis methods are applied to find interesting patterns.
- It can be characterized by a set of types of tasks that have to be solved.
- It uses methods from a variety of research areas.
  (statistics, databases, machine learning, artificial intelligence, soft computing etc.)

Data Mining Tasks 1

Classification

Predict outcome of an experiment with a finite number of possible results (e. g. yes/no, bird/plane/superman, good/neutral/bad)

Applicable for binary or categorical results

Prediction may be less expensive or easier to check

Examples

Is this customer creditworthy?

Will this customer respond to out mailing?

Will the quality of this product be acceptable?

Data Mining Tasks 2

Regression

Similar to classification

Prediction of a numerical value

Examples

What will be tomorrow's temperature?

How much will a customer spend?

How much will a machine's temperature increase in the next production cycle?

Data Mining Tasks 3

Cluster Analysis

Summarizing data. split data set into (mostly) disjunctive sub sets.

No need to examine data set as a whole but inspect clusters only

Gain insight in the structure of the data

Examples

Are there different groups of customers?

How many operating points does the machine have and how do they look like?

Data Mining Tasks 4

Association Analysis

Find correlations or interdependencies between items

Focus on relationships between all attributes

Examples

What optional equipments of a car often go together?

If a customer already bought A and B, what will they also buy?

Deviation Analysis

Find observations that do not follow a general trend

Outliers w.r.t. some concept

Examples

Under which circumstances does the system behave differently

What have customers in common that stand out of the crowd

CRISP-DM

Cross Industry Standard Process for Data Mining

Data Mining Process Model developed within an EU project

Several phases that are repeated until data mining project is finished

CRISP-Phases

Project understanding

Data understanding

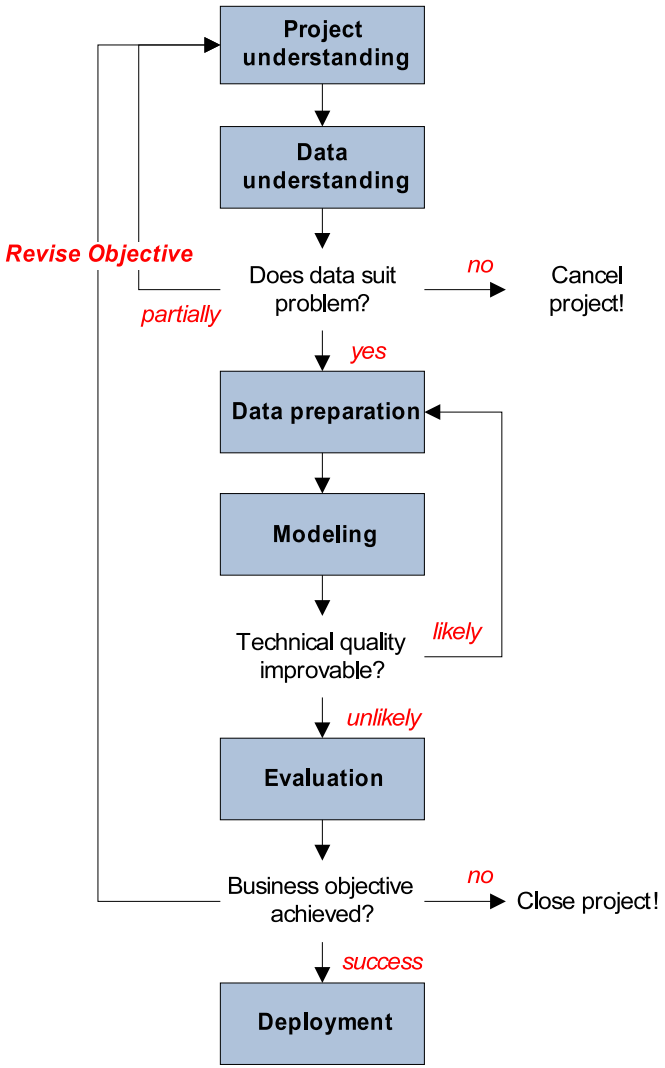
Data preparation

Modeling

Evaluation

Deployment

CRISP-DM Model



Project understanding

- What exactly is the problem, what the expected benefit?
- What should a solution look like?
- What is known about the domain?

Data understanding

- What (relevant) data is available?
- What about data quality/quantity/recency?

Data preparation

- Can data quality be increased?
- How can it be transformed for modeling?

Modeling

- What models is best suited to solve the problem?
- What is the best technique to get the model?
- How good does the model perform technically?

Evaluation

- How good is the model in terms of project requirements?
- What have we learned from the project?

Deployment

- How can the model be best deployed?
- Is there a way to know if the model is still valid?

Statistics

Descriptive Statistics

- Tabular and Graphical Representations
- Characteristic Measures
- Principal Component Analysis

Inductive Statistics

- Parameter Estimation
  (point and interval estimation, finding estimators)
- Hypothesis Testing
  (parameter test, goodness-of-fit test, dependence test)
- Model Selection
  (information criteria, minimum description length)

Summary

Statistics: Introduction

Statistics is the art to collect, to display, to analyze, and to interpret data in order to gain new knowledge.

[Sachs 1999]

[...] statistics, that is, the mathematical treatment of reality, [...]

Hannah Arendt

There are lies, damned lies, and statistics.

Benjamin Disraeli

Statistics, n. Exactly 76.4% of all statistics (including this one) are invented on the spot. However, in 83% of cases it is inappropriate to admit it.

The Devil's IT Dictionary

86.8748648% of all statistics pretend an accuracy that is not justified by the applied methods.

source unknown

Object, Case
Data describe objects, cases, persons etc.

(Random) Sample
The objects or cases described by a data set is called a *sample*,
their number the *sample size*.

Attribute
Objects and cases are described by *attributes*,
patients in a hospital, for example, by age, sex, blood pressure etc.

(Attribute) Value
Attributes have different possible *values*.
The age of a patient, for example, is a non-negative number.

Sample Value
The value an attribute has for an object in the sample is called *sample value*.

Scale Types / Attribute Types

Scale Type	Possible Operations	Examples
nominal (categorical, qualitative)	equality	sex blood group
ordinal (rank scale, comparative)	equality greater/less than	exam grade wind strength
metric (interval scale, quantitative)	equality greater/less than difference maybe ratio	length weight time temperature

Nominal scales are sometimes divided into *dichotomic* (two values) and *polytomic* (more than two values).

Metric scales may or may not allow us to form a ratio:
weight and length do, temperature does not.
time as duration does, time as calender time does not.

Descriptive Statistics

Tabular Representations: Frequency Table

Given data set: $x = (3, 4, 3, 2, 5, 3, 1, 2, 4, 3, 3, 4, 4, 1, 5, 2, 2, 3, 5, 3, 2, 4, 3, 2, 3)$

a_k	h_k	r_k	$\sum_{i=1}^{k} h_i$	$\sum_{i=1}^{k} r_i$
1	2	$\frac{2}{25} = 0.08$	2	$\frac{2}{25} = 0.08$
2	6	$\frac{6}{25} = 0.24$	8	$\frac{8}{25} = 0.32$
3	9	$\frac{9}{25} = 0.36$	17	$\frac{17}{25} = 0.68$
4	5	$\frac{5}{25} = 0.20$	22	$\frac{22}{25} = 0.88$
5	3	$\frac{3}{25} = 0.12$	25	$\frac{25}{25} = 1.00$

Absolute Frequency h_k (frequency of an attribute value a_k in the sample).

Relative Frequency $r_k = \frac{h_k}{n}$, where n is the sample size (here $n = 25$).

Cumulated Absolute/Relative Frequency $\sum_{i=1}^{k} h_i$ and $\sum_{i=1}^{k} r_i$.

Tabular Representations: Contingency Tables

Frequency tables for two or more attributes are called **contingency tables**.

They contain the absolute or relative frequency of **value combinations**.

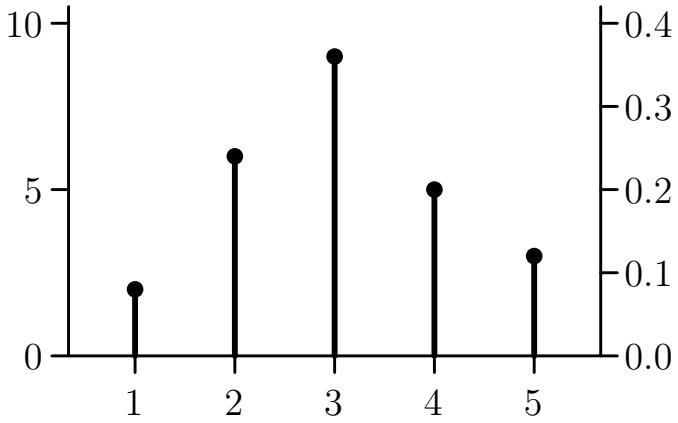
	a_1	a_2	a_3	a_4	$\sum$
b_1	8	3	5	2	18
b_2	2	6	1	3	12
b_3	4	1	2	7	14
$\sum$	14	10	8	12	44

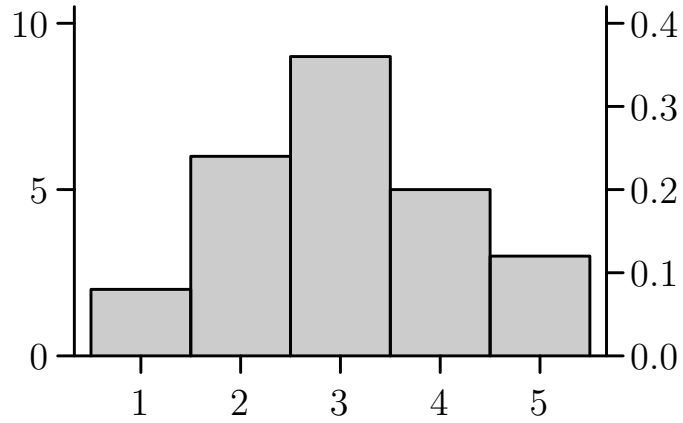
A contingency table may also contain the **marginal frequencies**,
i.e., the frequencies of the values of individual attributes.

Contingency tables for a higher number of dimensions (> 4) may be difficult to read.

Graphical Representations: Pole and Bar Chart

Numbers, which may be, for example, the frequencies of attribute values are represented by the lengths of poles (left) or bars (right).



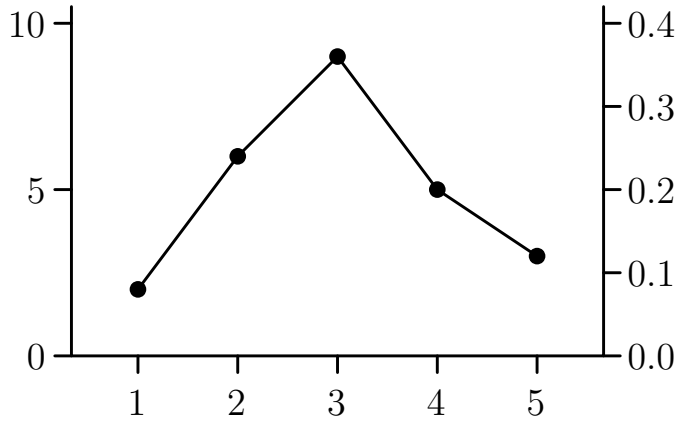


Bar charts are the most frequently used and most comprehensible way of displaying absolute frequencies.

A wrong impression can result if the vertical scale does not start at 0 (for frequencies or other absolute numbers).

Frequency Polygon and Line Chart

Frequency polygon: the ends of the poles of a pole chart are connected by lines. (Numbers are still represented by lengths.)



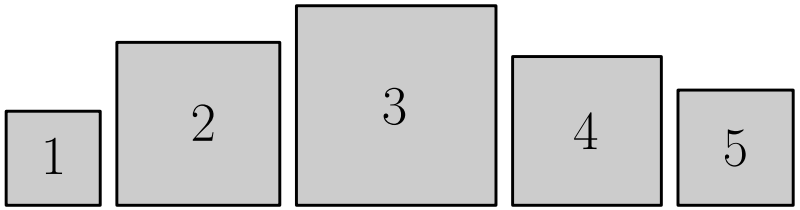
If the attribute values on the horizontal axis are not ordered, connecting the ends of the poles does not make sense.

Line charts are frequently used to display time series.

Area and Volume Charts

Numbers may also be represented by other geometric quantities than lengths, like areas or volumes.

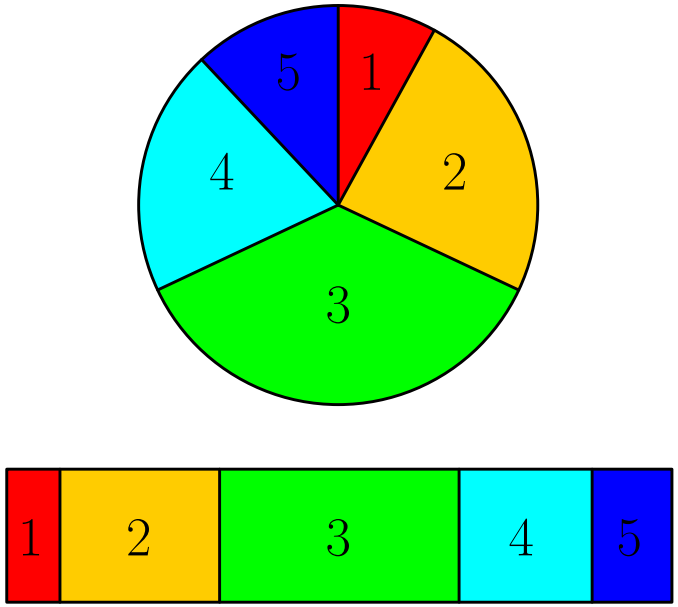
Area and volume charts are usually less comprehensible than bar charts, because humans have more difficulties to compare areas and especially volumes than lengths. (exception: the represented numbers are areas or volumes)

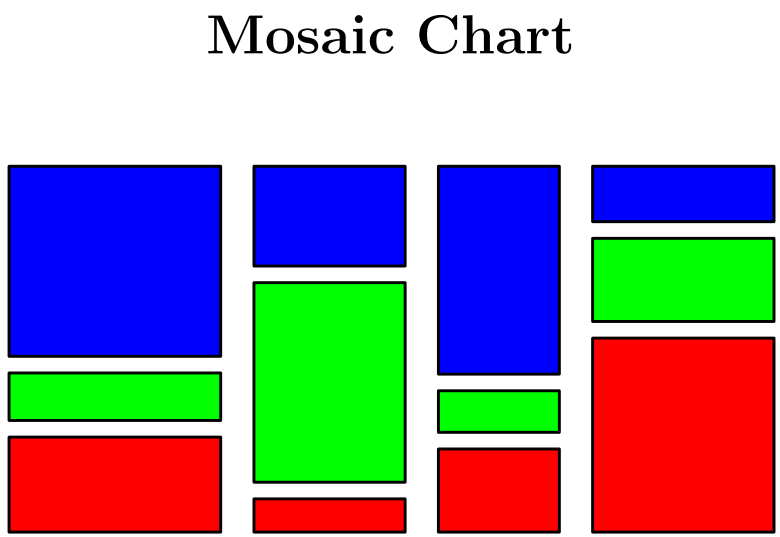


Sometimes the height of a two- or three-dimensional object is used to represent a number. The diagram then conveys a misleading impression.

Pie and Stripe Charts

Relative numbers may be represented by angles or sections of a stripe.





Mosaic charts can be used to display contingency tables.

More than two attributes are possible, but then separation distances and color must support the visualization to keep it comprehensible.

Histograms

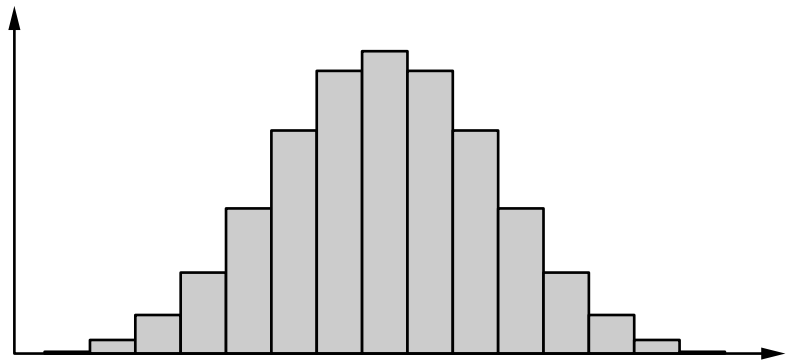
Intuitively: **Histograms are frequency bar charts for metric data.**

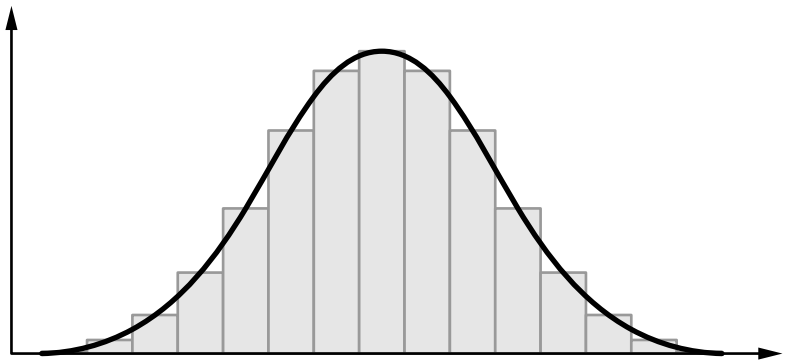
However: Since there are so many different values,
values have to be grouped in order to arrive a proper representation.

Most common approach: form equally sized intervals (so-called **bins**)
and count the frequency of sample values inside each interval.

Attention: Depending on the size and the position of the bins
the histogram may look considerably different.

In sketches often only a rough outline of a histogram is drawn:

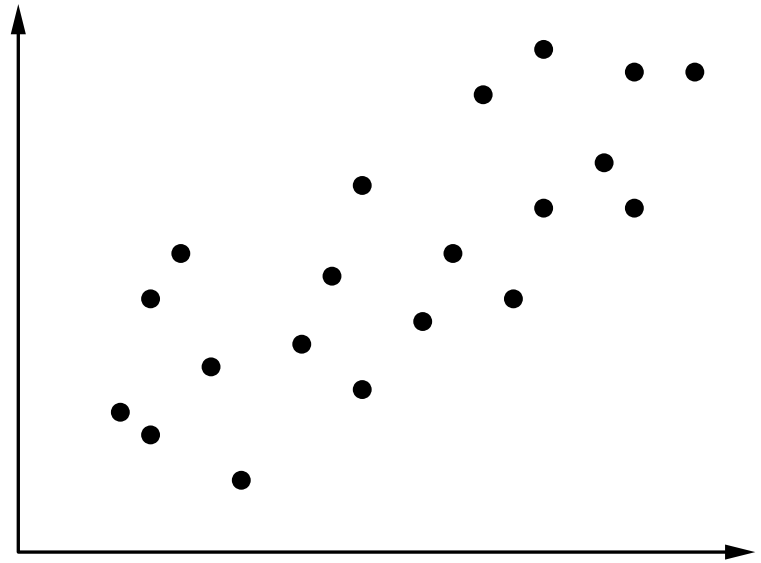




Scatter Plots

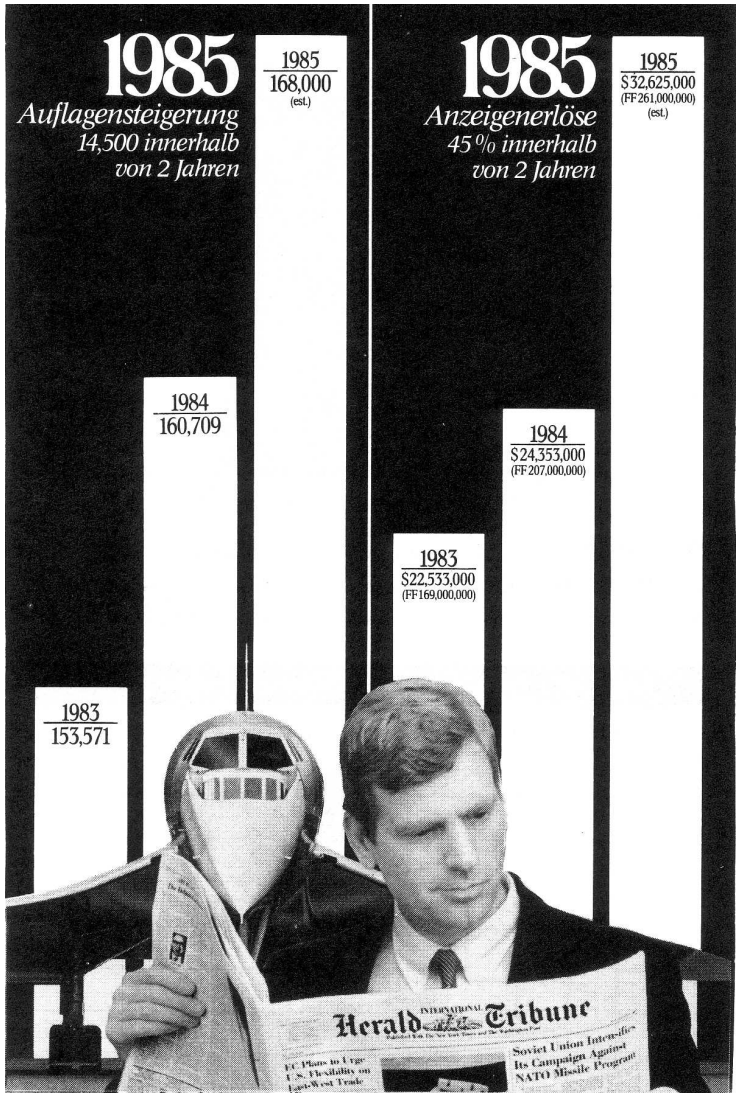
Scatter plots are used to display two-dimensional metric data sets.

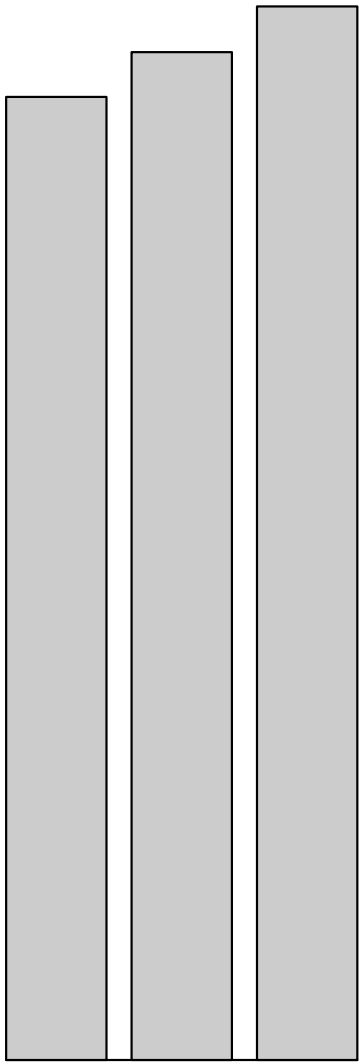
Sample values are the coordinates of a point.
(Numbers are represented by lengths.)



Scatter plots provide a simple means for checking for dependency.

How to Lie with Statistics





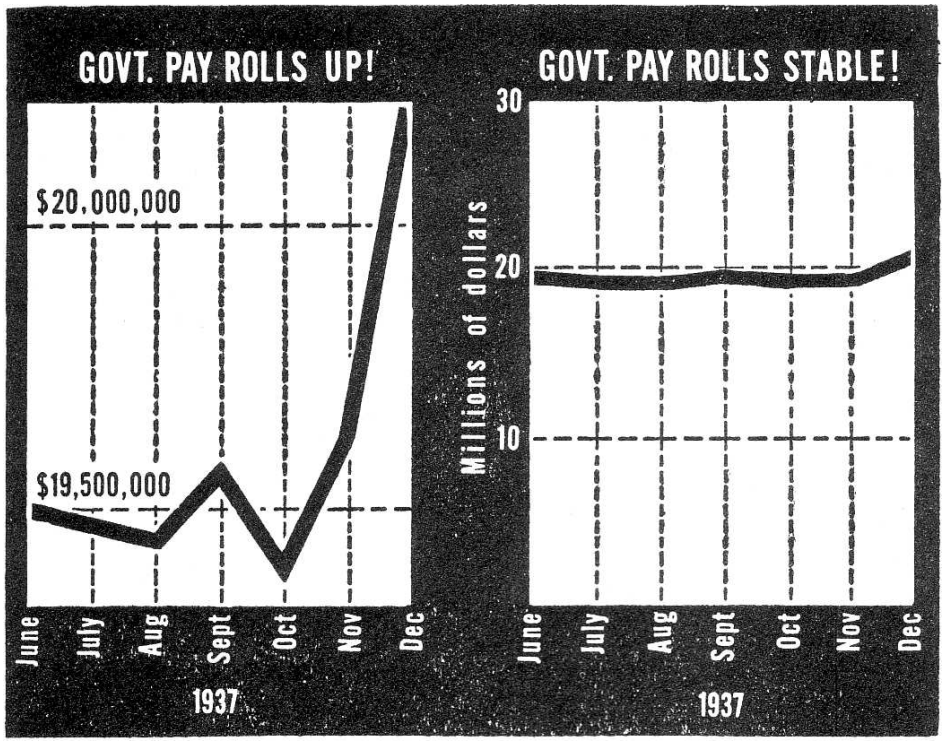
Often the vertical axis of a pole or bar chart does not start at zero, but at some higher value.

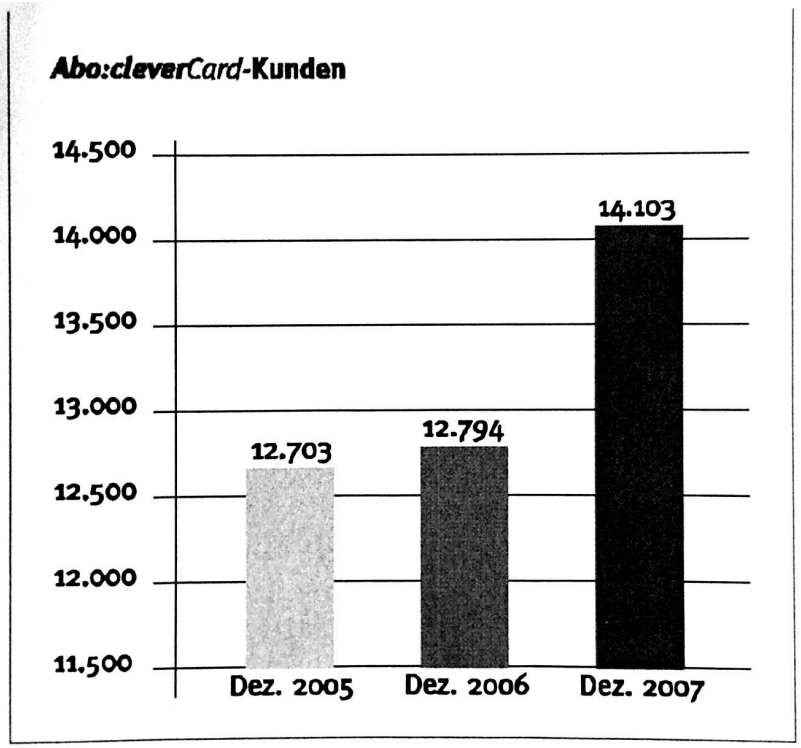
In such a case the conveyed impression of the ratio of the depicted values is completely wrong.

This effect is used to brag about increases in turnover, speed etc.

Sources of these diagrams and those on the following transparencies:
D. Huff: How to Lie with Statistics.
W. Krämer: So lügt man mit Statistik.

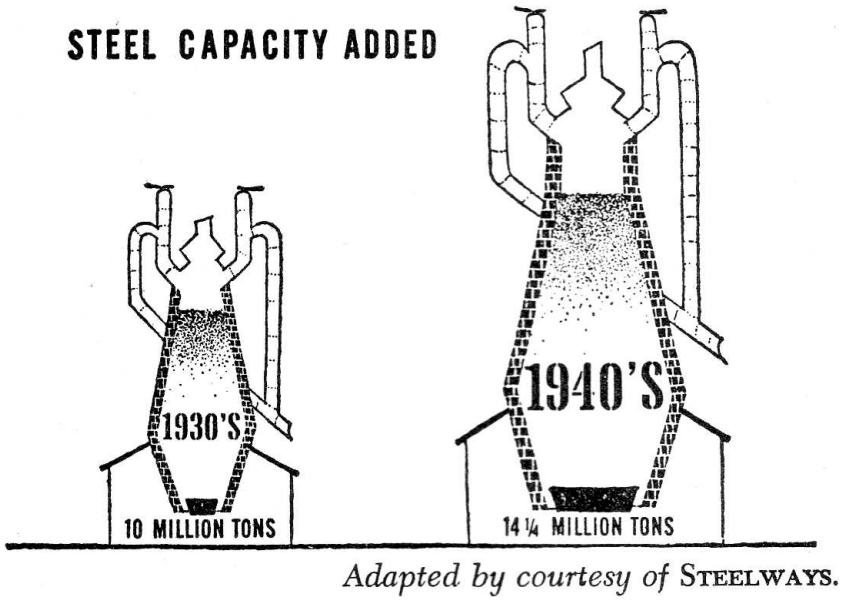
How to Lie with Statistics





Depending on the position of the zero line of a pole, bar, or line chart completely different impressions can be conveyed.

How to Lie with Statistics

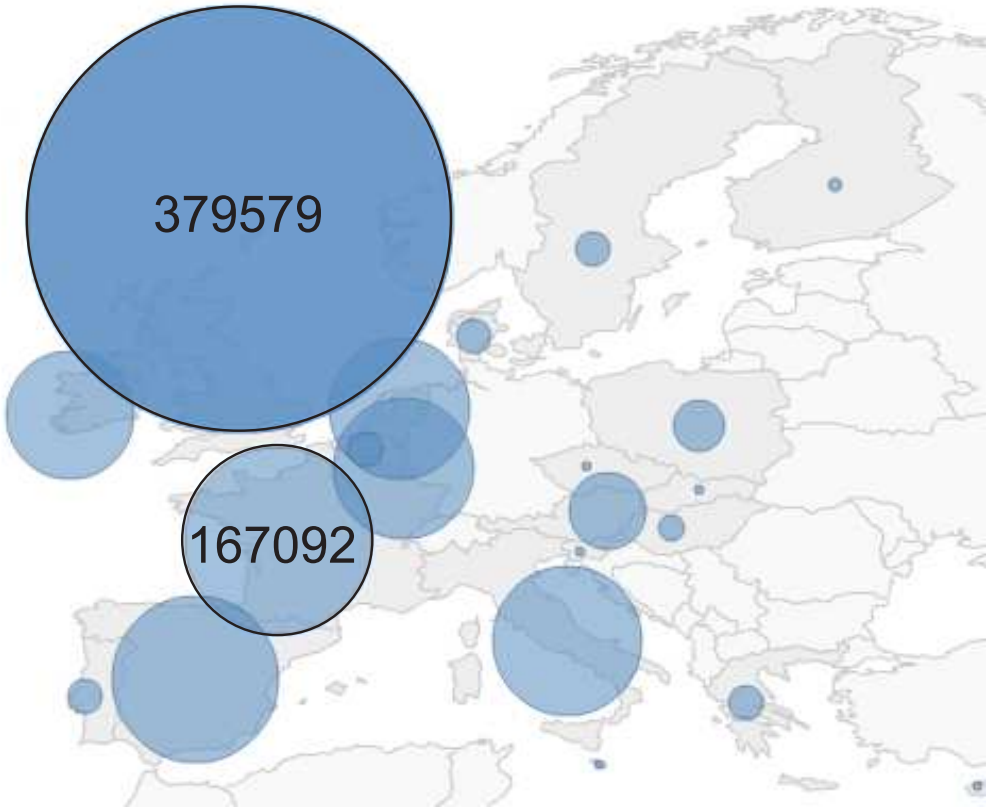


Poles and bars are frequently replaced by (sketches of) objects in order to make the diagram more aesthetically appealing.

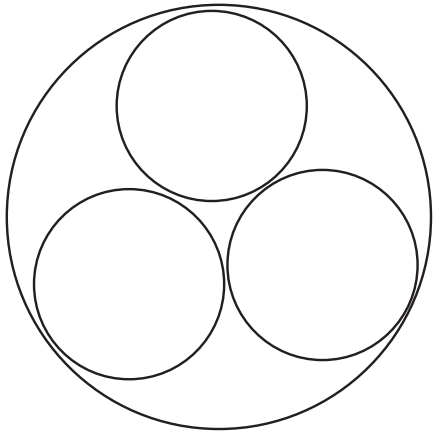
However, objects are perceived as 2- or even 3-dimensional and thus convey a completely different impression of the numerical ratios.

How to Lie with Statistics

Foreign outstanding debits of German banks in million Euros as of 2010:



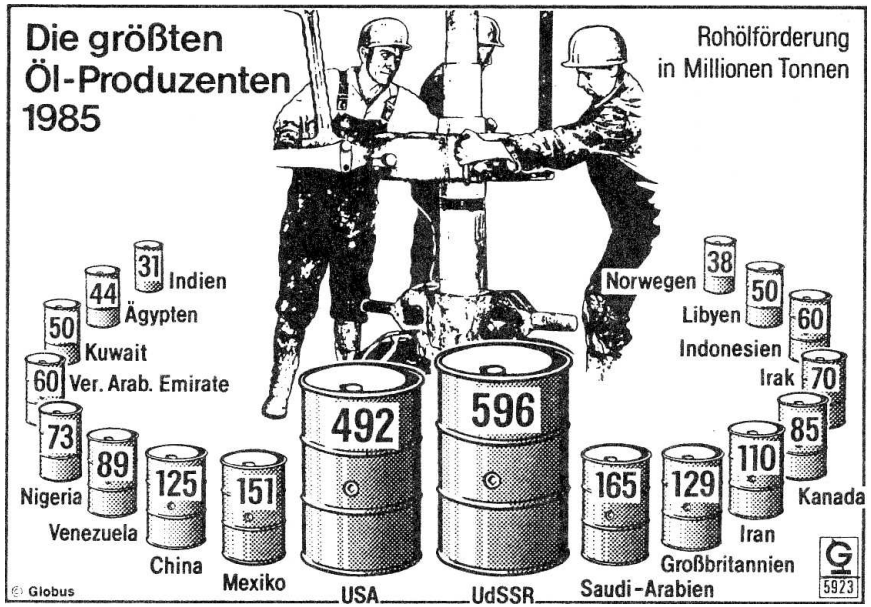
Source: Spiegel Online

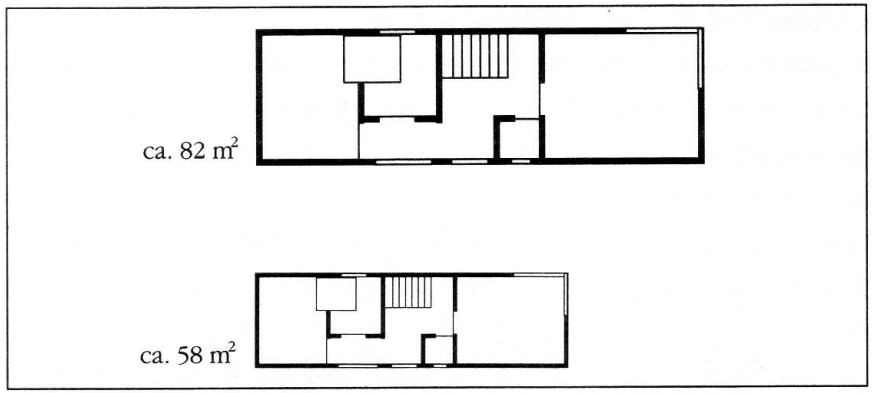


$$\frac{379579}{167092} \approx 2.2$$

$$\frac{A_{\mathrm{UK}}}{A_{\mathrm{F}}} \approx 5.0$$

How to Lie with Statistics





Quelle: "Zahlenspiegel" Bundesrepublik Deutschland - DDR – Ein Vergleich. 2. Auflage, Juli 1983, S. 63. Herausgeber: Bundesministerium für innerdeutsche Beziehungen.

In the left diagram the areas of the barrels represent the numerical value. However, since the barrels are drawn 3-dimensional, a wrong impression of the numerical ratios is conveyed.

The right diagram is particularly striking: an area measure is represented by the *side length* of a rectangle representing the apartment.

Descriptive Statistics: Characteristic Measures

Idea: Describe a given sample by few characteristic measures and thus summarize the data.

Localization Measures

Localization measures describe, usually by a single number, where the data points of a sample are located in the domain of an attribute.

Dispersion Measures

Dispersion measures describe how much the data points vary around a localization parameter and thus indicate how well this parameter captures the localization of the data.

Shape Measures

Shape measures describe the shape of the distribution of the data points relative to a reference distribution. The most common reference distribution is the normal distribution (Gaussian).

Location Measures for Metric Attributes: Mode and Median

Mode x^*

The mode is the attribute value that is most frequent in the sample.
It need not be unique, because several values can have the same frequency.
It is the most general measure, because it is applicable for all scale types.

Median $\tilde{x}$

The median minimizes the sum of absolute differences:

$$\sum_{i=1}^{n} |x_i - \tilde{x}| = \min . \qquad \text{and thus it is} \qquad \sum_{i=1}^{n} \operatorname{sgn}(x_i - \tilde{x}) = 0$$

If $x = (x_{(1)}, \ldots, x_{(n)})$ is a sorted data set, the median is defined as

$$\tilde{x} = \begin{cases} x_{(\frac{n+1}{2})}, & \text{if } n \text{ is odd,} \\ \frac{1}{2}\left(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}\right), & \text{if } n \text{ is even.} \end{cases}$$

The median is applicable to ordinal and metric attributes.

Arithmetic Mean $\bar{x}$

The arithmetic mean minimizes the sum of squared differences:

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \min. \qquad \text{and thus it is} \qquad \sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n} x_i - n\bar{x} = 0$$

The arithmetic mean is defined as

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

The arithmetic mean is only applicable to metric attributes.

Even though the arithmetic mean is the most common localization measure, the **median** is preferable if

- there are few sample cases,
- the distribution is asymmetric, and/or
- one expects that outliers are present.

Dispersion Measures: Range and Interquantile Range

A man with his head in the freezer and feet in the oven
is *on the average* quite comfortable.

old statistics joke

Range R
The range of a data set is the difference between the maximum and the minimum value.

$$R = x_{\max} - x_{\min} = \max{}_{i=1}^{n} x_i - \min{}_{i=1}^{n} x_i$$

Interquantile Range
The p-quantile of a data set is a value such that a fraction of p of all sample values are smaller than this value. (The median is the $\frac{1}{2}$-quantile.)

The p-interquantile range, $0 < p < \frac{1}{2}$, is the difference between
the $(1-p)$-quantile and the p-quantile.

The most common is the *interquartile range* ($p = \frac{1}{4}$)

Average Absolute Deviation

The average absolute deviation is the average of the absolute deviations of the sample values from the median or the arithmetic mean.

Average Absolute Deviation from the **Median**

$$d_{\tilde{x}} = \frac{1}{n} \sum_{i=1}^{n} |x_i - \tilde{x}|$$

Average Absolute Deviation from the **Arithmetic Mean**

$$d_{\bar{x}} = \frac{1}{n} \sum_{i=1}^{n} |x_i - \bar{x}|$$

It is always $d_{\tilde{x}} \leq d_{\bar{x}}$, since the median minimizes the sum of absolute deviations. (see the definition of the median)

Dispersion Measures: Variance and Standard Deviation

Variance s^2

It would be natural to define the variance as the average squared deviation:

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

However, inductive statistics suggests that it is better defined as

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

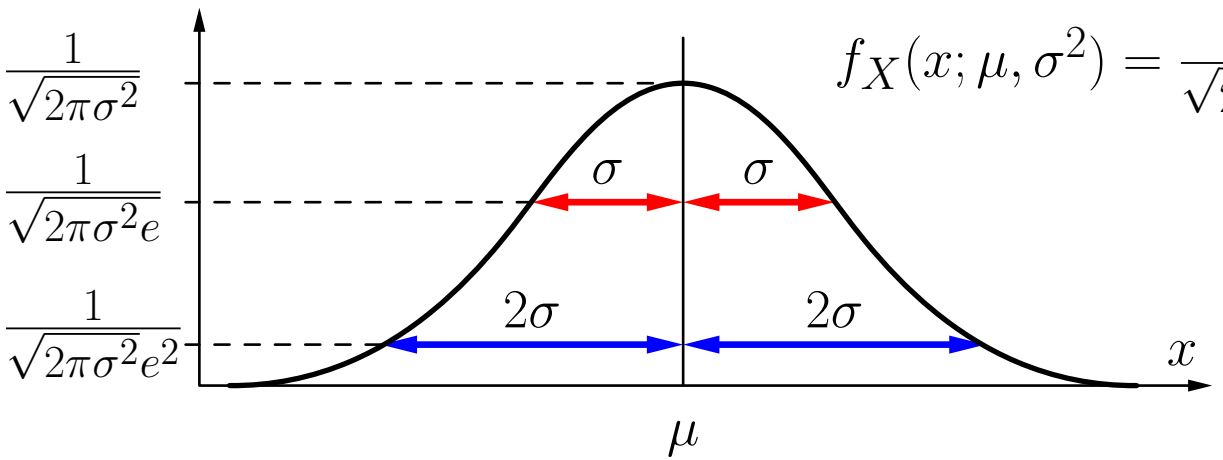
Standard Deviation s

The standard deviation is the square root of the variance, i.e.,

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

Special Case: Normal/Gaussian Distribution

The variance/standard deviation provides information about the height of the mode and the width of the curve.



$$f_X(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

μ:	expected value,	estimated by mean value $\bar{x}$
σ^2:	variance,	estimated by (empirical) variance s^2
σ:	standard deviation,	estimated by (empirical) standard deviation s

(Details about parameter estimation are studied later.)

Such a fit is only applicable to metric data!

Dispersion Measures: Variance and Standard Deviation

Note that it is often more convenient to compute the variance using the formula that results from the following transformation:

$$
\begin{aligned}
s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i^2-2x_i\bar{x}+\bar{x}^2\right)\\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n}x_i^2-2\bar{x}\sum_{i=1}^{n}x_i+\sum_{i=1}^{n}\bar{x}^2\right)\\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n}x_i^2-2n\bar{x}^2+n\bar{x}^2\right) = \frac{1}{n-1}\left(\sum_{i=1}^{n}x_i^2-n\bar{x}^2\right)\\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n}x_i^2-\frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)^2\right)
\end{aligned}
$$

Advantage: The sums $\sum_{i=1}^{n} x_i$ and $\sum_{i=1}^{n} x_i^2$ can both be computed in the same traversal of the data and from them both mean and variance are computable.

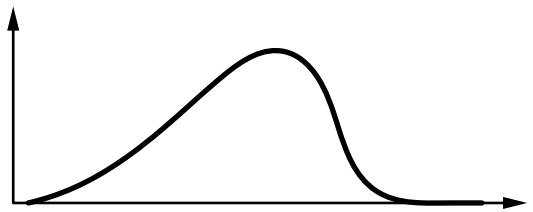
Shape Measures: Skewness

The **skewness** α_3 (or **skew** for short) measures whether, and if, how much, a distribution differs from a symmetric distribution.

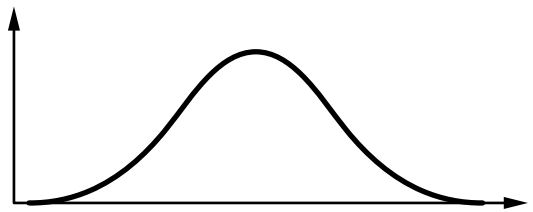
It is computed from the 3rd moment about the mean, which explains the index 3.

$$\alpha_3 = \frac{1}{n \cdot v^3} \sum_{i=1}^{n} (x_i - \bar{x})^3 = \frac{1}{n} \sum_{i=1}^{n} z_i^3$$

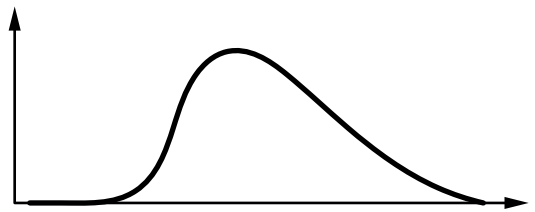
$$\text{where} \quad z_i = \frac{x_i - \bar{x}}{v} \quad \text{and} \quad v^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2.$$



$\alpha_3 < 0$: right steep



$\alpha_3 = 0$: symmetric



$\alpha_3 > 0$: left steep

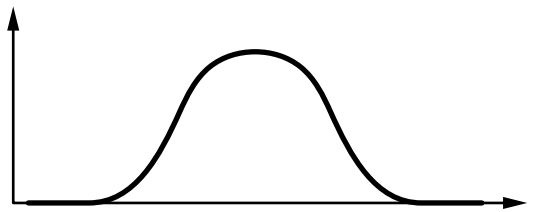
Shape Measures: Kurtosis

The **kurtosis** or **excess** α_4 measures how much a distribution is arched, usually compared to a Gaussian distribution.

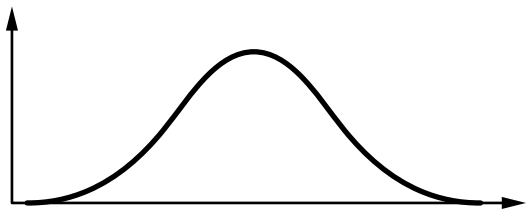
It is computed from the 4th moment about the mean, which explains the index 4.

$$\alpha_4 = \frac{1}{n \cdot v^4} \sum_{i=1}^{n} (x_i - \bar{x})^4 = \frac{1}{n} \sum_{i=1}^{n} z_i^4$$

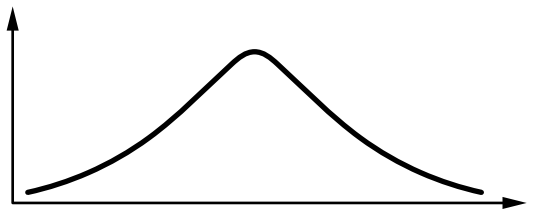
$$\text{where} \quad z_i = \frac{x_i - \bar{x}}{v} \quad \text{and} \quad v^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2.$$



$\alpha_4 < 3$: leptokurtic



$\alpha_4 = 3$: Gaussian



$\alpha_4 > 3$: platikurtic

Moments of Data Sets

The k-th **moment** of a dataset is defined as

$$m'_k = \frac{1}{n} \sum_{i=1}^{n} x_i^k.$$

The first moment is the **mean** $m'_1 = \bar{x}$ of the data set.
Using the moments of a data set the **variance** s^2 can also be written as

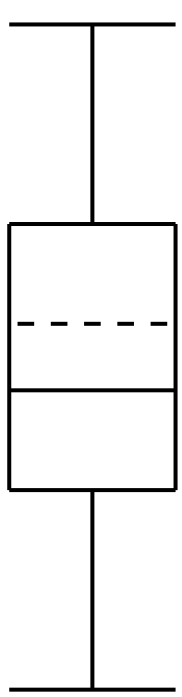
$$s^2 = \frac{1}{n-1} \left(m'_2 - \frac{1}{n} m'^2_1 \right) \qquad \text{and also} \qquad v^2 = \frac{1}{n} m'_2 - \frac{1}{n^2} m'^2_1.$$

The k-th **moment about the mean** is defined as

$$m_k = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^k.$$

It is $m_1 = 0$ and $m_2 = v^2$ (i.e., the **average squared deviation**).
The **skewness** is $\alpha_3 = \frac{m_3}{m_2^{3/2}}$ and the **kurtosis** is $\alpha_4 = \frac{m_4}{m_2^2}$.

Visualizing Characteristic Measures: Box Plots



$x_{\max}$ maximum

Q_3 3. quartile

$\bar{x}$ arithmetic mean

$\tilde{x} = Q_2$ median/2. quartile

Q_1 1. quartile

$x_{\min}$ minimum

A box plot is a common way to combine some important characteristic measures into a single graphic.

Often the central box is drawn laced ($\rangle\langle$) w.r.t. the arithmetic mean in order to emphasize its location.

Box plots are often used to get a quick impression of the distribution of the data by showing them side by side for several attributes.

Multidimensional Characteristic Measures

General Idea: Transfer the formulae to vectors.

Arithmetic Mean

The arithmetic mean for multidimensional data is the vector mean of the data points. For two dimensions it is

$$\overline{(x,y)} = \frac{1}{n}\sum_{i=1}^{n}(x_i, y_i) = (\bar{x}, \bar{y})$$

For the arithmetic mean the transition to several dimensions only combines the arithmetic means of the individual dimensions into one vector.

Other measures are transferred in a similar way.
However, sometimes the transfer leads to new quantities, as for the variance.

Excursion: Vector Products

For the variance, the square of the difference to the mean has to be generalized.

Inner Product
Scalar Product

$$\vec{v}^\top \vec{v} \qquad \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_m \end{pmatrix}$$

$$(v_1, v_2, \ldots, v_m) \quad \sum_{i=1}^{m} v_i^2$$

Outer Product
Matrix Product

$$\vec{v}\vec{v}^\top \qquad \begin{pmatrix} v_1, & v_2, & \ldots, & v_m \end{pmatrix}$$

$$\begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_m \end{pmatrix} \quad \begin{pmatrix} v_1^2 & v_1 v_2 & \cdots & v_1 v_m \\ v_1 v_2 & v_2^2 & \cdots & v_2 v_m \\ \vdots & & \ddots & \vdots \\ v_1 v_m & v_2 v_m & \cdots & v_m^2 \end{pmatrix}$$

In principle both vector products may be used for a generalization.

The second, however, yields more information about the distribution:

- a measure of the (linear) dependence of the attributes,
- a description of the direction dependence of the dispersion.

Covariance Matrix

Compute variance formula with vectors (square: outer product $\vec{v}\vec{v}^\top$):

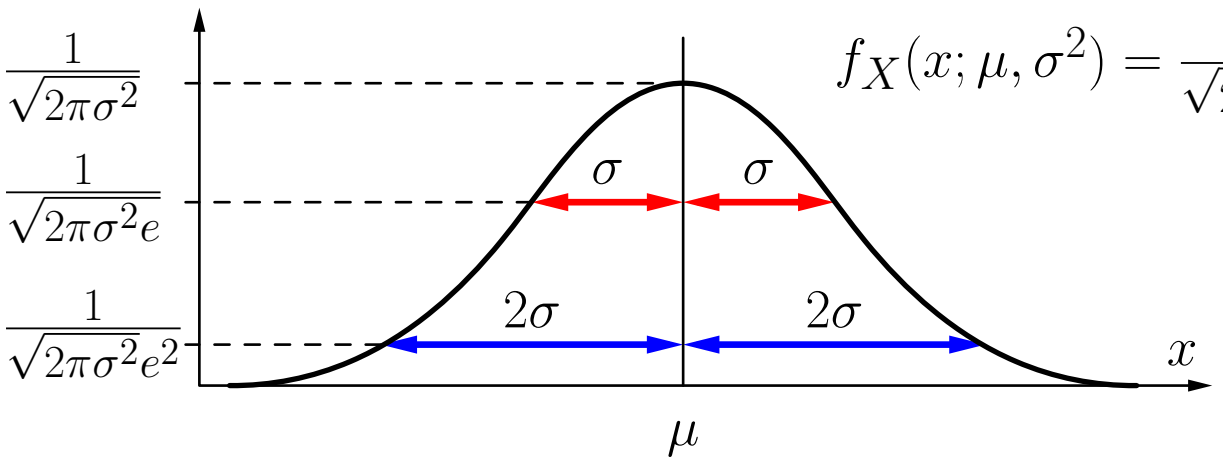
$$\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\begin{pmatrix} x_i \\ y_i \end{pmatrix} - \begin{pmatrix} \bar{x} \\ \bar{y} \end{pmatrix}\right)\left(\begin{pmatrix} x_i \\ y_i \end{pmatrix} - \begin{pmatrix} \bar{x} \\ \bar{y} \end{pmatrix}\right)^\top = \begin{pmatrix} s_x^2 & s_{xy} \\ s_{xy} & s_y^2 \end{pmatrix}$$

where

$$\begin{aligned} s_x^2 &= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) && \text{(variance of } x\text{)} \\ s_y^2 &= \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right) && \text{(variance of } y\text{)} \\ s_{xy} &= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}\right) && \text{(covariance of } x \text{ and } y\text{)} \end{aligned}$$

Special Case: Normal/Gaussian Distribution

The variance/standard deviation provides information about the height of the mode and the width of the curve.



$$f_X(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

μ:	expected value,	estimated by mean value $\bar{x}$,
σ^2:	variance,	estimated by (empirical) variance s^2,
σ:	standard deviation,	estimated by (empirical) standard deviation s.

Important: standard deviation has same unit as expected value.

Multivariate Normal Distribution

A **univariate normal distribution** has the density function

$$f_X(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

μ:	expected value,	estimated by mean value $\bar{x}$,
σ^2:	variance,	estimated by (empirical) variance s^2,
σ:	standard deviation,	estimated by (empirical) standard deviation s.

A **multivariate normal distribution** has the density function

$$f_{\vec{X}}(\vec{x}; \vec{\mu}, \mathbf{\Sigma}) = \frac{1}{\sqrt{(2\pi)^m |\mathbf{\Sigma}|}} \cdot \exp\left(-\frac{1}{2}(\vec{x}-\vec{\mu})^\top \mathbf{\Sigma}^{-1}(\vec{x}-\vec{\mu})\right)$$

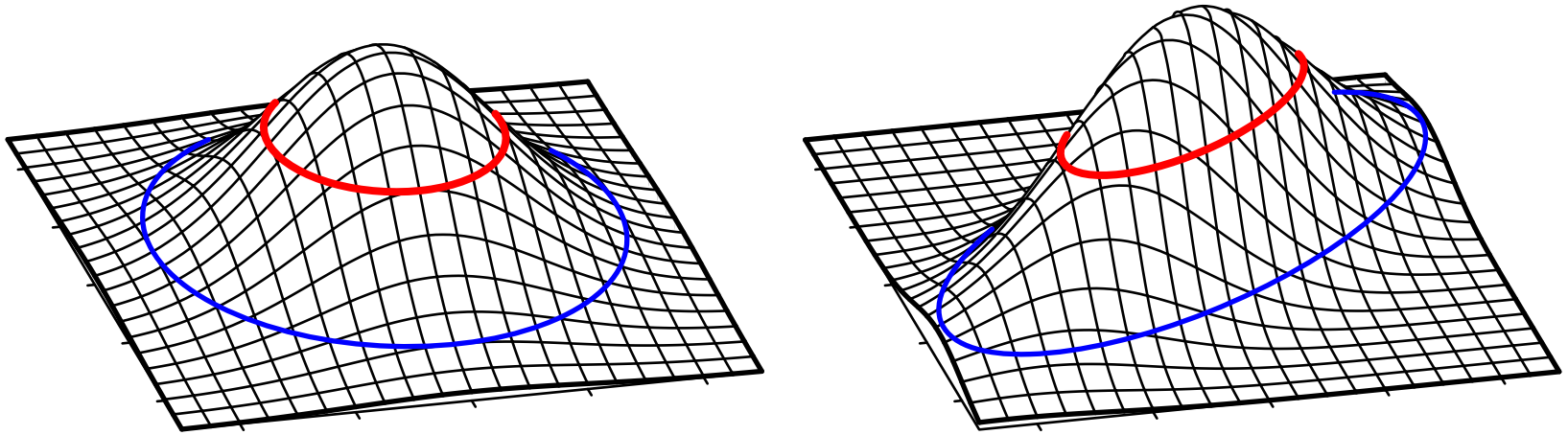
m:	size of the vector $\vec{x}$	(it is m-dimensional),
$\vec{\mu}$:	mean value vector,	estimated by (empirical) mean value vector $\bar{\vec{x}}$,
$\mathbf{\Sigma}$:	covariance matrix,	estimated by (empirical) covariance matrix $\mathbf{S}$,
$\lvert\mathbf{\Sigma}\rvert$:	determinant of the covariance matrix $\mathbf{\Sigma}$.	

Interpretation of a Covariance Matrix

The variance/standard deviation relates the spread of the distribution to the spread of a **standard normal distribution** ($\sigma^2 = \sigma = 1$).

The covariance matrix relates the spread of the distribution to the spread of a **multivariate standard normal distribution** ($\mathbf{\Sigma} = \mathbf{1}$).

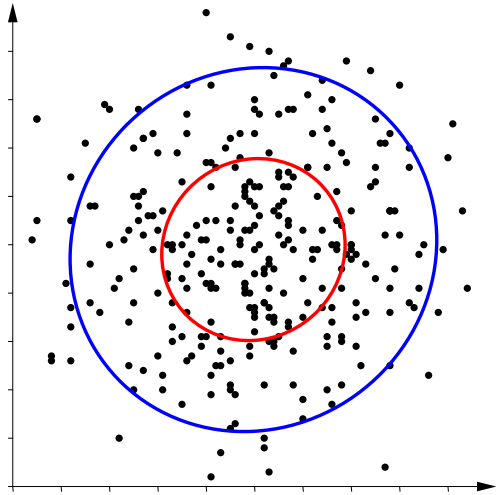
Example: bivariate normal distribution



standard general

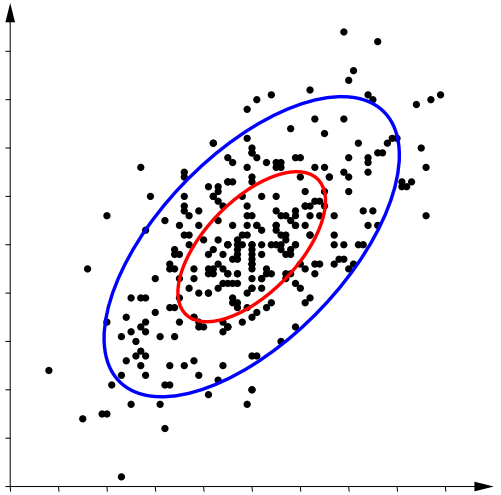
Question: Is there a multivariate analog of standard deviation?

Covariance Matrices of Example Data Sets



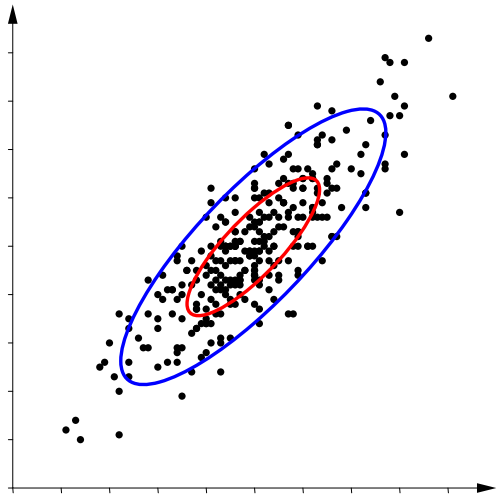
$$\boldsymbol{\Sigma} = \begin{pmatrix} 3.59 & 0.19 \\ 0.19 & 3.54 \end{pmatrix}$$

$$\mathbf{L} = \begin{pmatrix} 1.90 & 0 \\ 0.10 & 1.88 \end{pmatrix}$$



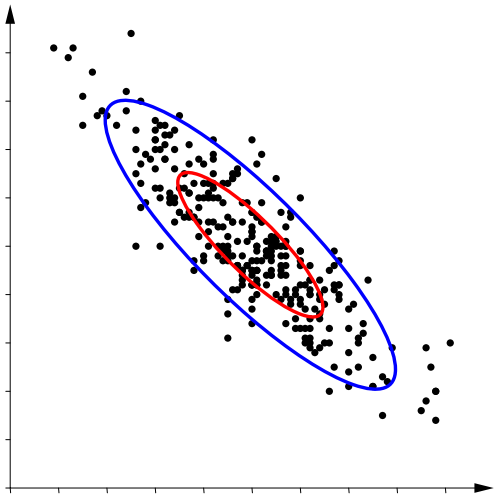
$$\boldsymbol{\Sigma} = \begin{pmatrix} 2.33 & 1.44 \\ 1.44 & 2.41 \end{pmatrix}$$

$$\mathbf{L} = \begin{pmatrix} 1.52 & 0 \\ 0.95 & 1.22 \end{pmatrix}$$



$$\boldsymbol{\Sigma} = \begin{pmatrix} 1.88 & 1.62 \\ 1.62 & 2.03 \end{pmatrix}$$

$$\mathbf{L} = \begin{pmatrix} 1.37 & 0 \\ 1.18 & 0.80 \end{pmatrix}$$



$$\boldsymbol{\Sigma} = \begin{pmatrix} 2.25 & -1.93 \\ -1.93 & 2.23 \end{pmatrix}$$

$$\mathbf{L} = \begin{pmatrix} 1.50 & 0 \\ -1.29 & 0.76 \end{pmatrix}$$

Correlation and Principal Component Analysis

Correlation Coefficient

The covariance is a measure of the strength of **linear dependence** of the two quantities.

However, its value depends on the variances of the individual dimensions.
$\Rightarrow$ Normalize to unit variance in the individual dimensions.

Correlation Coefficient
(more precisely: Pearson's Product Moment Correlation Coefficient)

$$r = \frac{s_{xy}}{s_x s_y}, \qquad r \in [-1, +1].$$

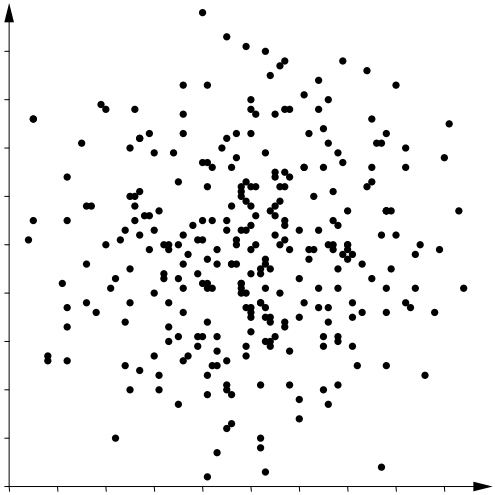
r measures the strength of linear dependence:

$r = -1$: the data points lie perfectly on a descending straight line.
$r = +1$: the data points lie perfectly on an ascending straight line.

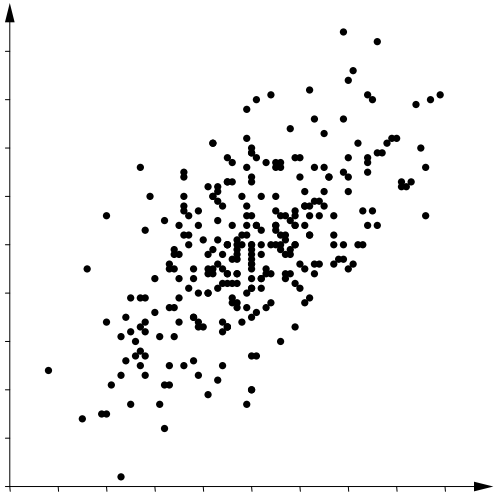
$r = 0$: there is no **linear** dependence between the two attributes (but there may be a non-linear dependence!).

Correlation Coefficients of Example Data Sets

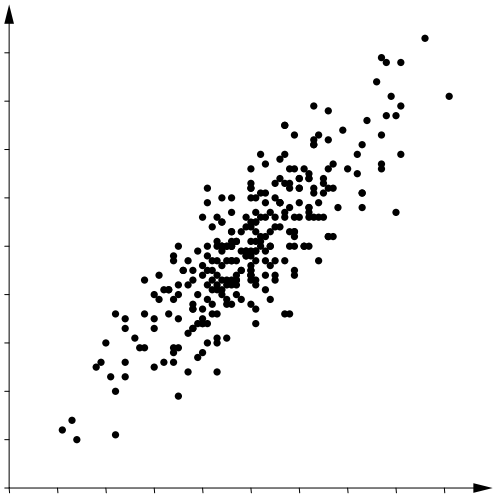
no correlation ($r \approx 0.05$)



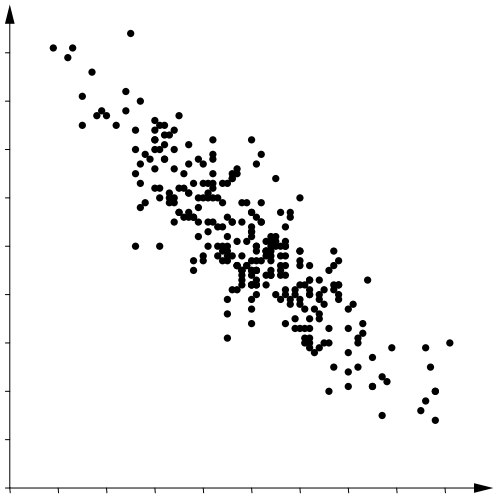
weak positive correlation ($r \approx 0.61$)



strong positive correlation ($r \approx 0.83$)



strong negative correlation ($r \approx -0.86$)



Normalize Data

Transform data to mean value 0 and variance/standard deviation 1:

$$\forall i; 1 \leq i \leq n: \qquad x'_i = \frac{x_i - \bar{x}}{s_x}, \qquad y'_i = \frac{y_i - \bar{y}}{s_y}.$$

Compute Covariance Matrix of Normalized Data

Sum outer products of transformed data vectors:

$$\mathbf{\Sigma'} = \frac{1}{n-1} \sum_{i=1}^{n} \begin{pmatrix} x'_i \\ y'_i \end{pmatrix} \begin{pmatrix} x'_i \\ y'_i \end{pmatrix}^\top = \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}$$

Subtraction of mean vector is not necessary (because it is $(0, 0)^\top$).
Diagonal elements are always 1 (because of unit variance in each dimension).

Normalizing the data and then computing the covariances or
computing the covariances and then normalizing them has the same effect.

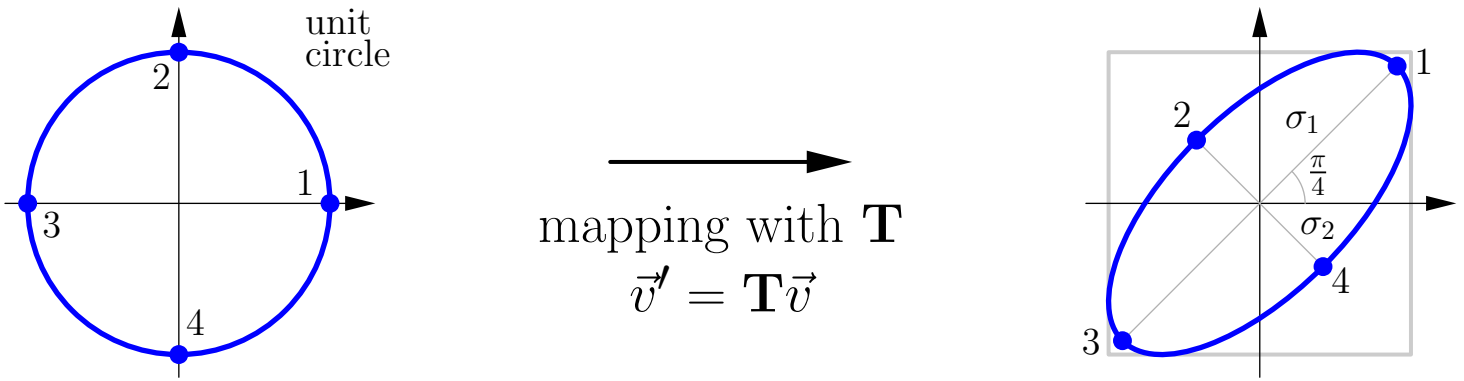
Special Case: Two Dimensions

Correlation matrix

$$\mathbf{\Sigma}' = \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}, \qquad \begin{aligned} \text{eigenvalues:} \quad & \sigma_1^2,\ \sigma_2^2 \\ \text{correlation:} \quad & r = \frac{\sigma_1^2 - \sigma_2^2}{\sigma_1^2 + \sigma_2^2} \end{aligned}$$

Eigenvalue decomposition

$$\mathbf{T} = \begin{pmatrix} c & -s \\ s & c \end{pmatrix} \begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{pmatrix}, \qquad \begin{aligned} s = \sin\tfrac{\pi}{4} = \frac{1}{\sqrt{2}}, \quad & \sigma_1 = \sqrt{1+r}, \\ c = \cos\tfrac{\pi}{4} = \frac{1}{\sqrt{2}}, \quad & \sigma_2 = \sqrt{1-r}. \end{aligned}$$



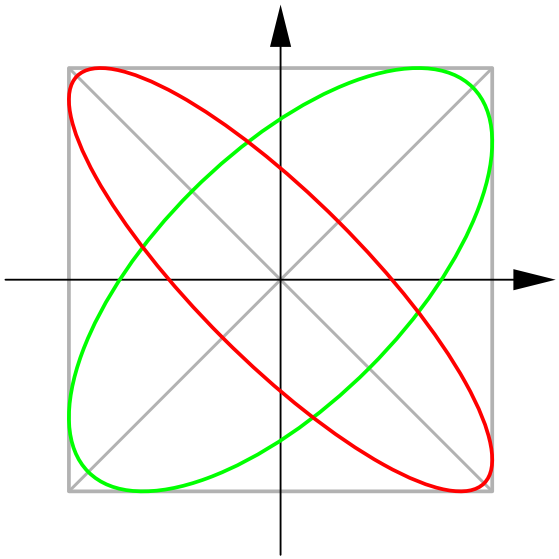
Correlation Matrix: Interpretation

For two dimensions the eigenvectors of a correlation matrix are always

$$\vec{v}_1 = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right) \quad \text{and} \quad \vec{v}_2 = \left(-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)$$

(or their opposites $-\vec{v}_1$ or $-\vec{v}_2$ or exchanged).

The reason is that the normalization transforms the data points in such a way, that the ellipse, the unit circle is mapped to by the "square root" of the covariance matrix of the normalized data, is always inscribed into the square $[-1, 1] \times [-1, 1]$. Hence the ellipse's major axes are the square's diagonals.

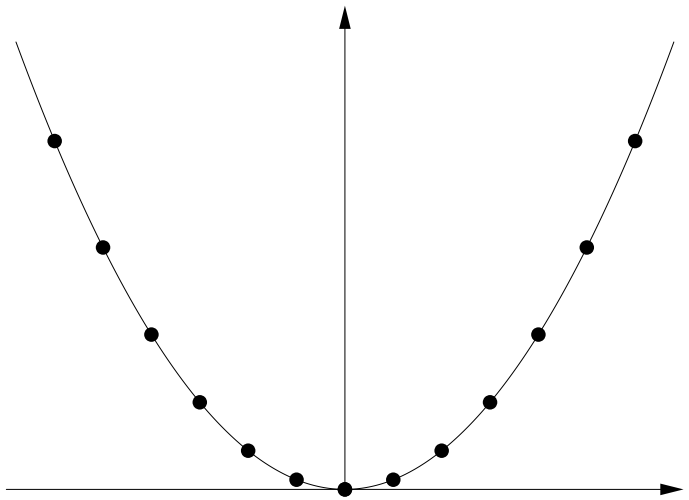


The situation is analogous in m-dimensional spaces: the eigenvectors are always m of the 2^{m-1} diagonals of the m-dimensional unit (hyper-)cube around the origin.

Correlation and Stochastic (In)Dependence

Note: stochastic independence $\Rightarrow r = 0$,
but: $r = 0 \not\Rightarrow$ stochastic independence.

Example: Suppose the data points lie symmetrically on a parabola.



The correlation coefficient of this data set is $r = 0$,
because there is **no linear** dependence between the two attributes.
However, there is a perfect **quadratic** dependence,
and thus the two attributes are **not** stochastically independent.

Regression Line

Since the covariance/correlation measures linear dependence,
it is not surprising that it can be used to define a **regression line**:

$$(y - \bar{y}) = \frac{s_{xy}}{s_x^2}(x - \bar{x}) \qquad \text{or} \qquad y = \frac{s_{xy}}{s_x^2}(x - \bar{x}) + \bar{y}.$$

The regression line can be seen as a conditional arithmetic mean:
there is one arithmetic mean for the y-dimensions for each x-value.

This interpretation is supported by the fact that the regression line minimizes the sum of squared differences in y-direction.
(Reminder: the arithmetic mean minimizes the sum of squared differences.)

More information on **regression** and the **method of least squares**
in the corresponding chapter.

Principal Component Analysis

Correlations between the attributes of a data set can be used to **reduce the number of dimensions**:

- Of two strongly correlated features only one needs to be considered.
- The other can be reconstructed approximately from the regression line.
- However, the feature selection can be difficult.

Better approach: **Principal Component Analysis** (PCA)

- Find the direction in the data space that has the highest variance.
- Find the direction in the data space that has the highest variance among those perpendicular to the first.
- Find the direction in the data space that has the highest variance among those perpendicular to the first and second and so on.
- Use first directions to describe the data.

Principal Component Analysis: Physical Analog

The rotation of a body around an axis through its center of gravity can be described by a so-called **inertia tensor**, which is a 3×3-matrix

$$\Theta = \begin{pmatrix} \Theta_{xx} & \Theta_{xy} & \Theta_{xz} \\ \Theta_{xy} & \Theta_{yy} & \Theta_{yz} \\ \Theta_{xz} & \Theta_{yz} & \Theta_{zz} \end{pmatrix}.$$

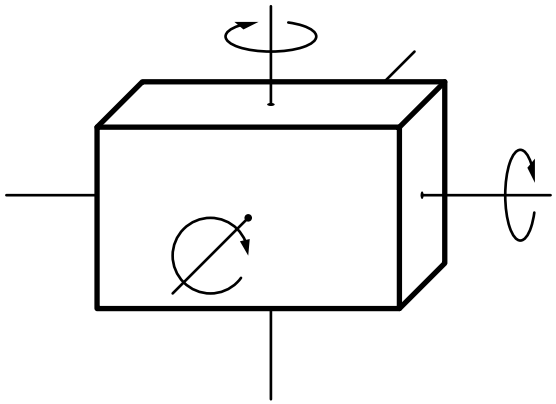
The diagonal elements of this tensor are called the **moments of inertia**.
They describe the "resistance" of the body against being rotated.

The off-diagonal elements are the so-called **deviation moments**.
They describe forces vertical to the rotation axis.

All bodies possess three perpendicular axes through their center of gravity, around which they can be rotated without forces perpendicular to the rotation axis. These axes are called **principal axes of inertia**.

There are bodies that possess more than 3 such axes (example: a homogeneous sphere), but all bodies have at least three such axes.

Principal Component Analysis: Physical Analog



The principal axes of inertia of a box.

The deviation moments cause "rattling" in the bearings of the rotation axis, which cause the bearings to wear out quickly.

A car mechanic who balances a wheel carries out, in a way, a principal axes transformation. However, instead of changing the orientation of the axes, he/she adds small weights to minimize the deviation moments.

A statistician who does a principal component analysis, finds, in a way, the axes through a weight distribution with unit weights at each data point, around which it can be rotated most easily.

Principal Component Analysis: Formal Approach

Normalize all attributes to arithmetic mean 0 and standard deviation 1:

$$x' = \frac{x - \bar{x}}{s_x}$$

Compute the **correlation matrix $\mathbf{\Sigma}$**
(i.e., the covariance matrix of the normalized data)

Carry out a **principal axes transformation** of the correlation matrix, that is, find a matrix $\mathbf{R}$, such that $\mathbf{R}^\top \mathbf{\Sigma} \mathbf{R}$ is a diagonal matrix.

Formal procedure:

- Find the **eigenvalues** and **eigenvectors** of the correlation matrix, i.e., find the values λ_i and vectors $\vec{v}_i$, such that $\mathbf{\Sigma}\vec{v}_i = \lambda_i \vec{v}_i$.
- The eigenvectors indicate the desired directions.
- The eigenvalues are the variances in these directions.

Principal Component Analysis: Formal Approach

Select dimensions using the **percentage of explained variance**.

- The eigenvalues λ_i are the variances σ_i^2 in the principal dimensions.
- It can be shown that the sum of the eigenvalues of an $m \times m$ correlation matrix is m. Therefore it is plausible to define $\frac{\lambda_i}{m}$ as the share the i-th principal axis has in the total variance.
- Sort the λ_i descendingly and find the smallest value k, such that

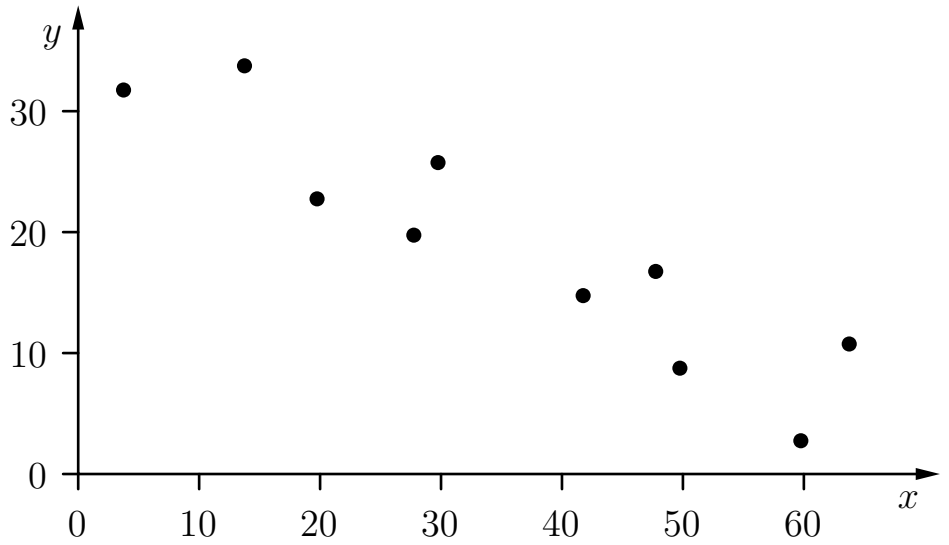
$$\sum_{i=1}^{k} \frac{\lambda_i}{m} \geq \alpha,$$

  where α is a user-defined parameter (e.g. $\alpha = 0.9$).
- Select the corresponding k directions (given by the eigenvectors).

Transform the data to the new data space by multiplying the data points with a matrix, the rows of which are the eigenvectors of the selected dimensions.

Principal Component Analysis: Example

x	5	15	21	29	31	43	49	51	61	65
y	33	35	24	21	27	16	18	10	4	12



Strongly correlated features $\Rightarrow$ Reduction to one dimension possible.

Principal Component Analysis: Example

Normalize to arithmetic mean 0 and standard deviation 1:

$$
\begin{aligned}
\bar{x} &= \frac{1}{10}\sum_{i=1}^{10} x_i &&= \frac{370}{10} &&= 37, \\
\bar{y} &= \frac{1}{10}\sum_{i=1}^{10} y_i &&= \frac{200}{10} &&= 20, \\
s_x^2 &= \frac{1}{9}\left(\sum_{i=1}^{10} x_i^2 - 10\bar{x}^2\right) &&= \frac{17290 - 13690}{9} = 400 && \Rightarrow s_x = 20, \\
s_y^2 &= \frac{1}{9}\left(\sum_{i=1}^{10} y_i^2 - 10\bar{y}^2\right) &&= \frac{4900 - \ \ 4000}{9} = 100 && \Rightarrow s_y = 10.
\end{aligned}
$$

x'	-1.6	-1.1	-0.8	-0.4	-0.3	0.3	0.6	0.7	1.2	1.4
y'	1.3	1.5	0.4	0.1	0.7	-0.4	-0.2	-1.0	-1.6	-0.8

Principal Component Analysis: Example

Compute the correlation matrix (covariance matrix of normalized data).

$$\mathbf{\Sigma} = \frac{1}{9}\begin{pmatrix} 9 & -8.28 \\ -8.28 & 9 \end{pmatrix} = \begin{pmatrix} 1 & -\frac{23}{25} \\ -\frac{23}{25} & 1 \end{pmatrix}.$$

Find the eigenvalues and eigenvectors, i.e., the values λ_i and vectors $\vec{v}_i$, $i = 1, 2$, such that

$$\mathbf{\Sigma}\vec{v}_i = \lambda_i \vec{v}_i \qquad \text{or} \qquad (\mathbf{\Sigma} - \lambda_i \mathbf{1})\vec{v}_i = \vec{0}.$$

where $\mathbf{1}$ is the unit matrix.

Here: Find the eigenvalues as the roots of the characteristic polynomial.

$$c(\lambda) = |\mathbf{\Sigma} - \lambda\mathbf{1}| = (1-\lambda)^2 - \frac{529}{625}.$$

For more than 3 dimensions, this method is numerically unstable and should be replaced by some other method (Jacobi-Transformation, Householder Transformation to tridiagonal form followed by the QR algorithm etc.).

Principal Component Analysis: Example

The roots of the characteristic polynomial $c(\lambda) = (1-\lambda)^2 - \frac{529}{625}$ are

$$\lambda_{1/2} = 1 \pm \sqrt{\frac{529}{625}} = 1 \pm \frac{23}{25}, \qquad \text{i.e.} \qquad \lambda_1 = \frac{48}{25} \quad \text{and} \quad \lambda_2 = \frac{2}{25}$$

The corresponding eigenvectors are determined by solving for $i = 1, 2$ the (underdetermined) linear equation system

$$(\boldsymbol{\Sigma} - \lambda_i \mathbf{1})\vec{v}_i = \vec{0}$$

The resulting eigenvectors (normalized to length 1) are

$$\vec{v}_1 = \left(\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}\right) \qquad \text{and} \qquad \vec{v}_2 = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right),$$

(Note that for two dimensions always these two vectors result.
Reminder: directions of the eigenvectors of a correlation matrix.)

Principal Component Analysis: Example

Therefore the transformation matrix for the principal axes transformation is

$$\mathbf{R} = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}, \qquad \text{for which it is} \qquad \mathbf{R}^\top \mathbf{\Sigma} \mathbf{R} = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$$

However, instead of $\mathbf{R}^\top$ we use $\sqrt{2}\mathbf{R}^\top$ to transform the data:

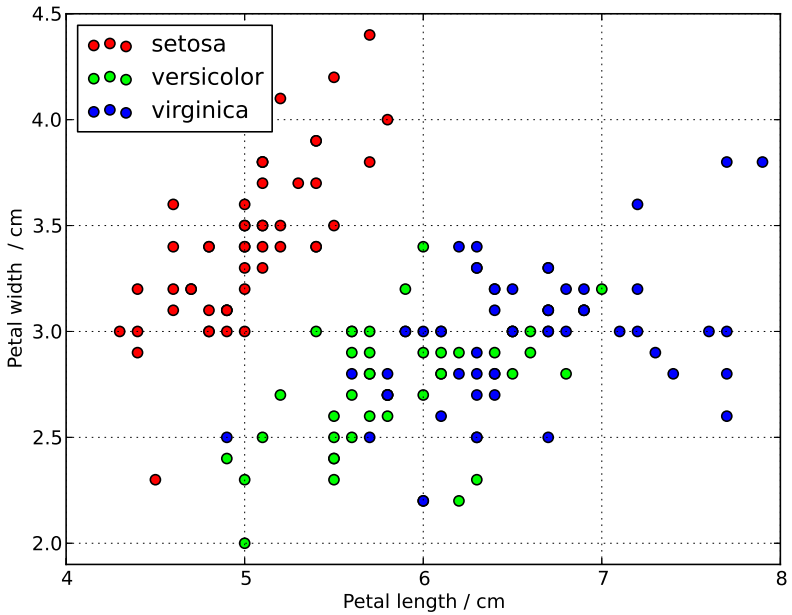
$$\begin{pmatrix} x'' \\ y'' \end{pmatrix} = \sqrt{2} \cdot \mathbf{R}^\top \cdot \begin{pmatrix} x' \\ y' \end{pmatrix}.$$

Resulting data set:

x''	-2.9	-2.6	-1.2	-0.5	-1.0	0.7	0.8	1.7	2.8	2.2
y''	-0.3	0.4	-0.4	-0.3	0.4	-0.1	0.4	-0.3	-0.4	0.6

y'' is discarded ($s^2_{y''} = 2\lambda_2 = \frac{4}{25}$) and only x'' is kept ($s^2_{x''} = 2\lambda_1 = \frac{96}{25}$).

Scatter Plots 1

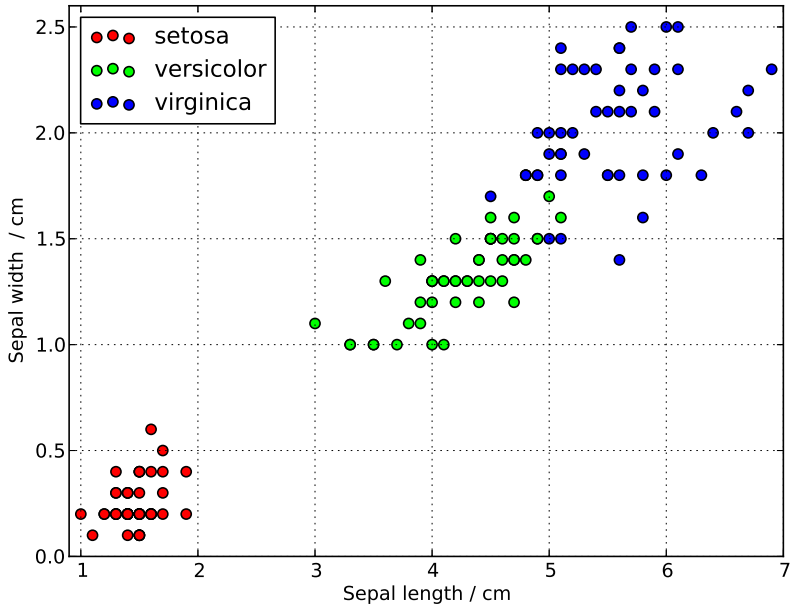


Iris data set (150 samples, 3 classes, 4 numerical attributes)

Only *sepal length* and *sepal width* used to plot data

Different colors have been used for the different classes

Scatter Plots 2

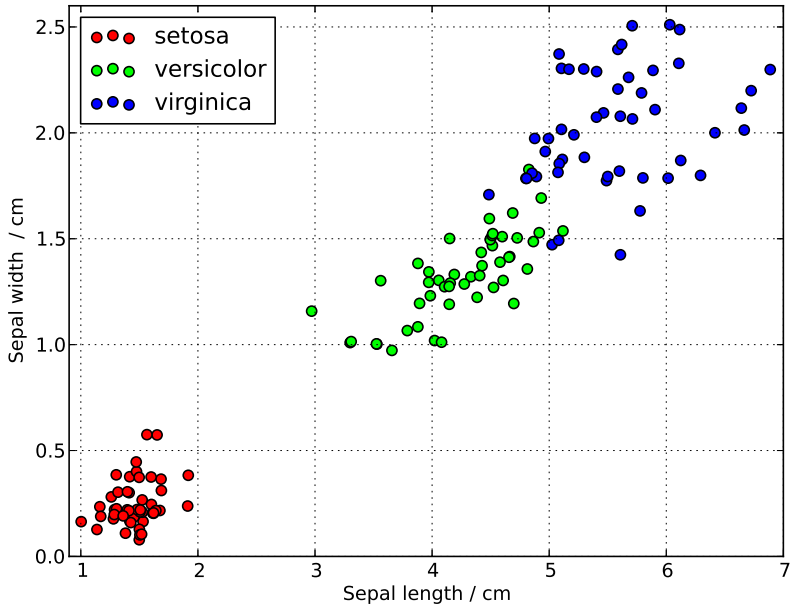


Different combinations of attributes may reveal correlations otherwise unseen

petal length and *petal width* provide a better separation of the classes

Iris setosa can already be clearly identified

Scatter Plots 3



Jitter can be added to the data points to make points visible that are otherwise invisible

Small random numbers are added to the coordinates before plotting

Categorical attributes *need* to be jittered

Methods for higher-dimensional data 1

A display or plot is by definition two-dimensional, so that only two axes (attributes) can be incorporated.

3D-techniques can be used to incorporate three axes (attributes)

The number of possible scatter plots grows in a quadratic fashin with the number of attributes. For m attributes there are $\binom{m}{2} = m(m-1)$ possible scatter plots. For 50 attributes there are different 2450 scatter plots.

Methods for higher-dimensional data 2

Principal approach for incorporating all attributes in a plot:

Try to preserve as much of the *structure* of the high-dimensional data set when plotting data in two or three dimensions.

Define a measure that avaluates lower-dimensional representations of the data in terms of how well a representation preserves the original *structure* of the high-dimensional data set.

Find the representation that gives the best value for the defined measure.

There is no unique measure for structure preservation.

Multidimensional Scaling (MDS)

Multidimensional scaling (MDS) is not restricted to mappings in the form of simple projections. In contrast to PCA, MDS does not even construct an explicit mapping from the high-dimensional space to the low-dimensional space. It only positions the data points in the low-dimensional space.

The representation of the data in the low-dimensional space constructed by MDS aims at preserving the distances between the data points and not like PCA the variance in the data set.

Multidimensional Scaling

MDS requires a distance matrix $D \in \mathbb{R}^{n \times n}$ where each $d_{ij}, 1 \leq i, j, \leq n$ is the distance between data object i and data object j.

The distance should be non-negative: $d_{ij} \geq 0, \forall i, j$.

The distance should be symmetric: $d_{ij} = d_{ji}, \forall i, j$.

The entries on the principal diagonal should be zero: $d_{ii} = 0, \forall i$.
(Each data object has zero distance to itself.)

Usually, the distances are the Euclidean distances of the data objects (*after normalization*) in the high-dimensional space.

Multidimensional Scaling

MDS must define a point $p_i \in \mathbb{R}^q$ (usually $q = 2$, sometimes also $q = 3$) for each data object x_i.

The distances d^*_{ij} between the points p_i and p_j should be roughly the same as the distances d_{ij} between the original data objects x_i and x_j.

Usually $d^*_{ij} = \|p_i - p_j\|$.

Multidimensional Scaling: Objective Functions

$$E_0 = \sum_{i=1}^{n} \sum_{j=i+1}^{n} \left(d_{ij}^* - d_{ij} \right)^2 \text{ (absolute squared error)}$$

$$E_1 = \frac{1}{\sum_{i=1}^{n} \sum_{j=i+1}^{n} \left(d_{ij} \right)^2} \sum_{i=1}^{n} \sum_{j=i+1}^{n} \left(d_{ij}^* - d_{ij} \right)^2 \text{ (normalised absolute squared error)}$$

The normalisation factor $\frac{1}{\sum_{i=1}^{n} \sum_{j=i+1}^{n} \left(d_{ij} \right)^2}$ does not have an influence on the location of the minimum of the objective function.

In contrast to E_0, the value of E_1 does neither depend in the number of data objects nor on the magnitude of the original distances.

Multidimensional Scaling: Objective Functions

$$E_2 = \sum_{i=1}^{n} \sum_{j=i+1}^{n} \left(\frac{d_{ij}^* - d_{ij}}{d_{ij}} \right)^2 \text{ (relative squared error)}$$

$$E_3 = \frac{1}{\sum_{i=1}^{n} \sum_{j=i+1}^{n} (d_{ij})^2} \sum_{i=1}^{n} \sum_{j=i+1}^{n} \left(\frac{d_{ij}^* - d_{ij}}{d_{ij}} \right)^2$$

(mixture between relative and absolute squared error)

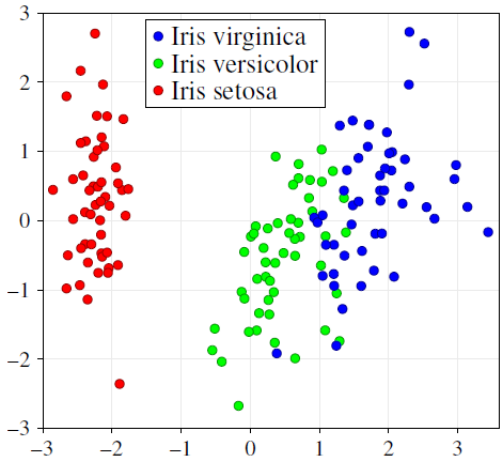
MDS based on E_3 is called **Sammon mapping**. The value of E_3 is called *stress*.

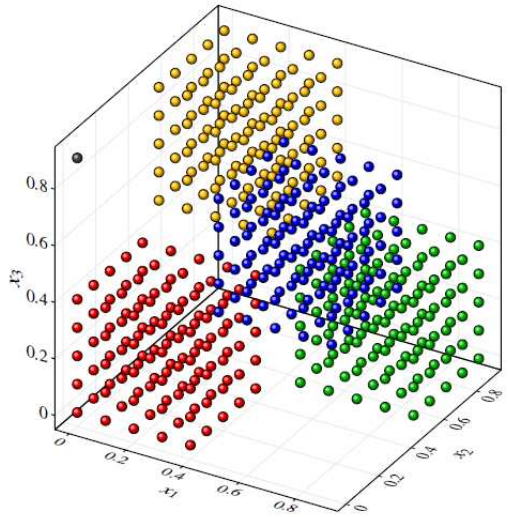
Multidimensional Scaling

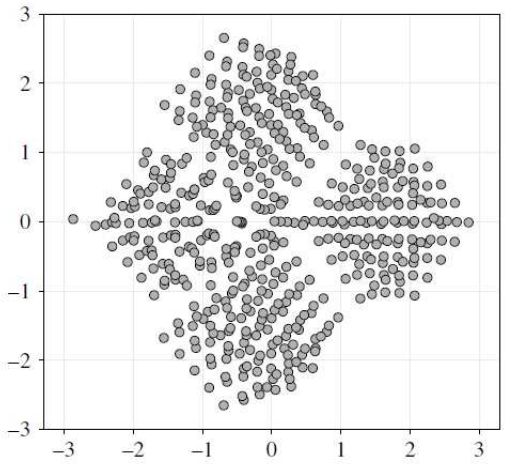
MDS represents a non-linear optimisation problem with $q \cdot n$ ($2n$ for $q = 2$) parameters to be optimised.
Even for a small data set like the Iris data set, a two-dimensional MDS representation requires the optimisation of 300 parameters.

Since the problem is non-linear, a gradient descent method is used to minimise the objective function for MDS.

Multidimensional Scaling







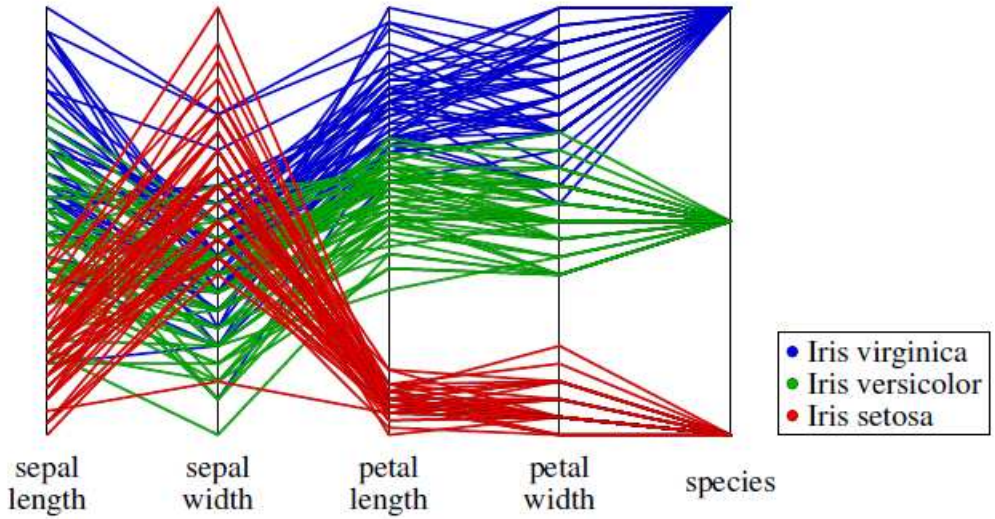
MDS (Sammon mapping) for the Iris data set, the Cube data, and MDS for the Cube data set.

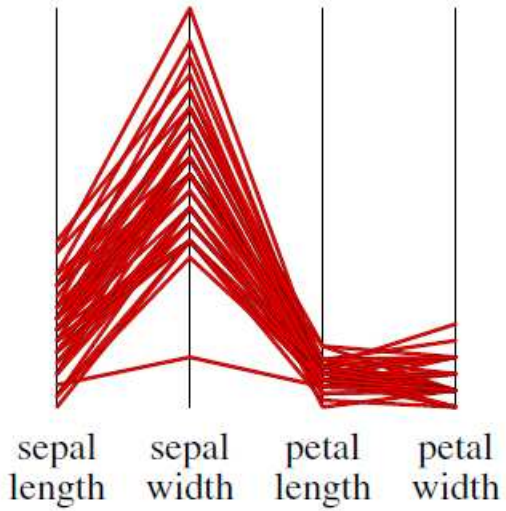
Parallel coordinates

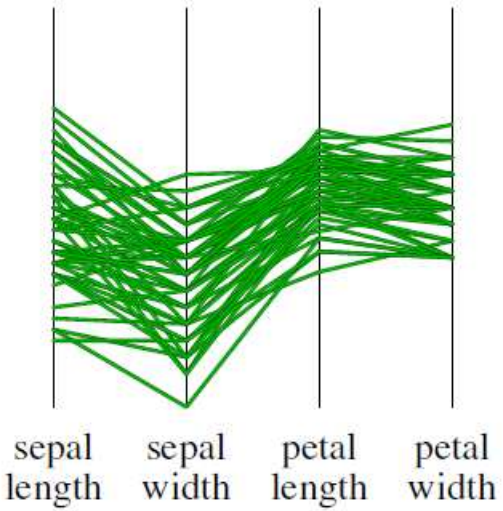
Parallel coordinates draw the coordinate axes parallel to each other, so that there is no limitation for the number of attributes to be shown simultaneously.

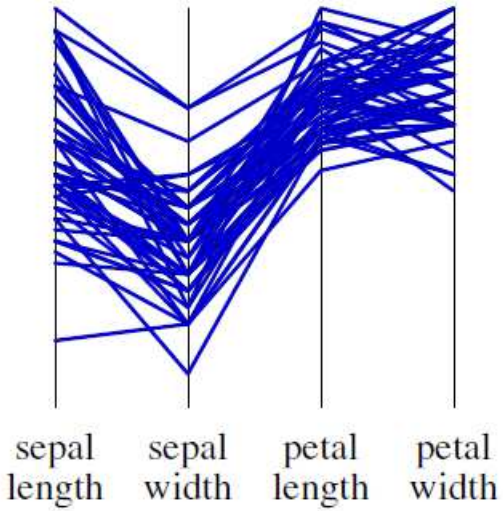
For a data object a polyline is drawn connecting the values of the data attribute on the corresponding axes.

Parallel coordinates









Outlier Detection

An *outlier* is a value or data object that is far away or very different from all or most of the other data.
Causes for outliers:

Data quality problems (erroneous data coming from wrong measurements or typing mistakes)

Exceptional or unusual situations/data objects.

Outliers coming from erroneous data should be excluded from the analysis.

Even if the outliers are correct (exceptional data), it is sometime useful to exclude them from the analysis. For example, a single extremely large outlier can lead to completely misleading values for the mean value.

Outlier Detection: Single Attributes

Categorical attributes: An outlier is a value that occurs with a frequency extremely lower than the frequency of all other values.

In some cases, the outliers can even be the target objects of the analysis.

Example: Automatic quality control system

Goal: Train a classifier, classifying the parts as correct or with failures based on measurements of the produced parts. The frequency of the correct parts will be so high that the parts with failure might be considered as outliers.

Outlier Detection: Single Attributes

Numerical attributes:

Outliers in boxplots.

Problems: Asymmetric distribution, large data sets

Statistical tests, for example *Grubb's test*:

Define the statistic

$G = \frac{\max\{\|x_i - \overline{x}\| : 1 \leq i \leq n}{s}$, where $x_1, \ldots, x_n$ is the sample, $\overline{x}$ its mean value and s its empirical standard deviation. For a given significance level α, the null hypothesis that the sample coming from a *normal distribution* does not contain outliers is rejected if

$$G > \frac{n-1}{\sqrt{n}} \sqrt{\frac{t^2_{1-\alpha/(2n),n-2}}{n-2+t^2_{1-\alpha/(2n),n-2}}}$$

where $t_{1-\alpha/(2n),n-2}$ denotes the $(1-\alpha/(2n))$-quantile of the t-distribution with $n-$ degrees of freedom.

Outlier Detection: Single Attributes

Grubb's test applied to the Iris data set:

attribute	p-value
sepal length	0.92
sepal width	0.13
petal length	1.0
petal width	1.0

The p-values do not indicate any outliers.

Note that the assumption of normal distrbuted values is not correct. The attributes from one species might follow a normal distribution, but not the values from all species together.

Outlier Detection for Multidimensional Data

Scatter plots for (visually detecting) outliers w.r.t. two attributes.

PCA or MDS plots for (visually detecting) outliers.

Cluster analysis techniques: Outliers are those points which cannot be assigned to any cluster.

Missing Values

For some instances values of single attributes might be missing.
Causes for missing values:

- Broken sensors.
- Refusal to answer a question.
- Irrelevant attribute for the corresponding object (e.g. *Pregnant (yes/no)?* for men).

Missing values might not necessarily be indicated as missing (instead: zero or default values).

Types of Missing Values

Consider the attribute X_{obs}. A missing value is denoted by ?.
X is the true value of the considered attribute, i.e. we have

$$X_{obs} = X, \text{ if } X_{obs} \neq ?$$

Let Y be the (multivariate) (random) variable denoting the other attributes apart from X.

Types of Missing Values

Missing completely at random (MCAR): The probability that a value for X is missing does neither depend on the true value of X nor on other variables.

$$P(X_{obs} = ?) = P(X_{obs} = ? \,|X, Y)$$

Example: The maintenance staff sometimes forgets to change the batteries of a sensor, so that the sensor sometimes does not provide any measurements.

MCAR is also called **Observed at random (OAR).**

Types of Missing Values

Missing at random (MAR): The probability that a value for X is missing does not depend on the true value of X.

$$P(X_{obs} = ?|Y) = P(X_{obs} = ?|X)$$

Example: The maintenance staff does not change the batteries of a sensor when it is raining, so that the sensor does not always provide measurements when it is raining.

Types of Missing Values

Nonignorable: The probability that a value for X is missing depends on the true value of X.

Example: A sensor for the temperature will not work when there is frost.
In the cases of MCAR and MAR, the missing values can be estimated - at least in principle, when the data set is large enough - based on the values of the other attributes. (The cause for the missing values is *ignorable*.) In the extreme case of the sensor for the temperature, it is impossible to provide any statement concerning temperatures below $0°C$.

Types of Missing Values

In the case of MCAR, it can be assumed that the missing values follow the same distribution as the observed values of X.

In the case of MAR, the missing values might not follow the distribution of X. But by taking the other attributes into account, it is possible to derive reasonable imputations for the missing values.

In the case of nonignorable missing values it is impossible to provide sensible estimations for the missing values.

Types of Missing Values

If it is not known based on domain knowledge which kind of missing values can be expected, the following strategy can be applied.

Turn the considered attribute X into a binary attribute, replacing all measured values by the values *yes* and all missing values by the value *no*.

Build a classifier with now binary attribute X as the target attribute and use all other attributes for the prediction of the class values *yes* and *no*.

Determine the misclassification rate. The misclassification rate is the proportion of data objects that are not it assigned to the correct class by the classifier.

Types of Missing Values

In the case of **OAR**, the other attributes should not provide any information, whether X has a missing value or not. Therefore, the misclassification rate of the classifier should not differ significantly from pure guessing, i.e. if there 10% missing values for the attribute X, the misclassification rate of the classifier should not be much smaller than 10%.

If, however, the misclassification rate of the classifier is significantly better than pure guessing, this is an indicator that there is a correlation between missing values for X and the values of the other attributes. The missing values are not **OAR**.

MAR and **nonignorable** cannot be distinguished in this way.

Checklist for Data Understanding

Get an idea of the data quality. Standard problems like syntactic accuracy can be easily checked.

Outliers can be a problem for data analysis. There are various methods for finding outliers. Visualisation methods like boxplots, scatter plots, projections based on PCA or MDS may be useful.

Simple correlations between attributes can be easily detected by scatter plots as well.

Specific assumptions made by some methods (e.g. normal distribution) should be checked during data understanding.

Checklist for Data Understanding

Missing values can be a problem. Depending on the reason why they are missing (OAR, MAR, nonignorable) the missing values can be estimated. OAR can be detected by checking the misclassification rate of a classifier that tries to predict whether a value is missing or not.

Missing values might not be explicitly marked as missing! Be aware of default values. (E.g. *DATE* in mySQL databases has a default of January, 1^{st} 1970.)

Probability Foundations

Reminder: Probability Theory

Goal: Make statements and/or predictions about results of physical processes.

Even processes that seem to be simple at first sight may reveal considerable difficulties when trying to predict.

Describing real-world physical processes always calls for a simplifying mathematical model.

Although everybody will have some intuitive notion about probability, we have to formally define the underlying mathematical structure.

Randomness or chance enters as the incapability of precisely modelling a process or the inability of measuring the initial conditions.

- *Example*: Predicting the trajectory of a billard ball over more than 9 banks requires more detailed measurement of the initial conditions (ball location, applied momentum etc.) than physically possible according to Heisenberg's uncertainty principle.

Reality vs. Model

Producing a result of a physical process is referred to as an **observed outcome**.

Assessing or predicting the probability of every possible outcome is not straightforward but often implicitly assumed to be clear.

We will study this "non-straightforwardness" with three real-world examples:

- Rolling a die.
- Arrivals of inquiries at a call center.
- The weight of a bread roll purchased from a bakery.
  (Inspired by a broadcast of Quarks & Co. from WDR.)

Obviously, all examples differ in the nature of the space of possible observable outcomes.

Example 1: Rolling a Die

Physical Process
Shaking a six-sided die in a dice cup.
Then cast it and read off the number of pips.

Possible Outcomes

Sources of Randomness

- Inaccurate knowledge about locations, momenta.
- Inellastic collisions inside the dice cup.
- Inhomogeneous material distribution of the die.
- Uneven table surface.
- Unknown frictions, airflow etc.

Model
Outcomes have equal probability.

Example 2: Phone Calls at a Call Center

Physical Process
Counting the number of phone calls that arrive
at a call center within a predefined time window.

Possible Outcomes
The events (if any) happening in time and space.

Sources of Randomness

- Calls are initiated by human beings: no predictability.
- Misdialed calls.
- Technical problems resulting in lost calls.

Model
Poisson distribution of number of calls.

Example 3: Bread Rolls at a Bakery

Physical Process
Baking a bread roll from a piece of dough.
Measuring its weight (with arbitrary precision).

Possible Outcomes
Bread rolls.

Sources of Randomness

- Amount of dough put on the baking sheet.
- Baking process (ingredients, temperature, time).

Model
Gaussian distribution of the weight.

Formal Approach on the Model Side

We conduct an experiment that has a set Ω of possible outcomes.
E. g.:

- Rolling a die ($\Omega = \{1, 2, 3, 4, 5, 6\}$)
- Arrivals of phone calls ($\Omega = \mathbb{N}_0$)
- Bread roll weights ($\Omega = \mathbb{R}_+$)

Such an outcome is called an **elementary event**.

All possible elementary events are called the **frame of discernment** Ω (or sometimes **universe of discourse**).

The set representation stresses the following facts:

- All possible outcomes are covered by the elements of Ω. (**collectively exhaustive**).
- Every possible outcome is represented by exactly one element of Ω. (**mutual disjoint**).

Events

Often, we are interested in *higher-level* events
(e. g. casting an odd number, arrival of at least 5 phone calls or
purchasing a bread roll heavier than 80 grams)

Any subset $A \subseteq \Omega$ is called an **event** which **occurs**, if the outcome $\omega_0 \in \Omega$ of the random experiment lies in A:

$$\text{Event } A \subseteq \Omega \text{ occurs} \quad \Leftrightarrow \quad \bigvee_{\omega \in A} (\omega = \omega_0) \;=\; \mathsf{true} \quad \Leftrightarrow \quad \omega_0 \in A$$

Since events are sets, we can define for two events A and B:

- $A \cup B$ occurs if A or B occurs; $A \cap B$ occurs if A and B occurs.
- $\overline{A}$ occurs if A does not occur (i. e., if $\Omega \backslash A$ occurs).
- A and B are *mutually exclusive*, iff $A \cap B = \emptyset$.

Event Algebra

A family of sets $\mathcal{E} = \{E_1, \ldots, E_n\}$ is called an **event algebra**, if the following conditions hold:

- The **certain event** Ω lies in $\mathcal{E}$.
- If $E \in \mathcal{E}$, then $\overline{E} = \Omega \backslash E \in \mathcal{E}$.
- If E_1 and E_2 lie in $\mathcal{E}$, then $E_1 \cup E_2 \in \mathcal{E}$ and $E_1 \cap E_2 \in \mathcal{E}$.

If Ω is uncountable, we require the additional property:

For a series of events $E_i \in \mathcal{E}, i \in \mathbb{N}$, the events $\bigcup_{i=1}^{\infty} E_i$ and $\bigcap_{i=1}^{\infty} E_i$ are also in $\mathcal{E}$.

$\mathcal{E}$ is then called a σ**-algebra**.

Side remarks:

Smallest event algebra: $\mathcal{E} = \{\emptyset, \Omega\}$

Largest event algebra (for finite or countable Ω): $\mathcal{E} = 2^{\Omega} = \{A \subseteq \Omega \mid \mathsf{true}\}$

Probability Function

Given an event algebra $\mathcal{E}$, we would like to assign every event $E \in \mathcal{E}$ its probability with a **probability function** $P: \mathcal{E} \to [0, 1]$.

We require P to satisfy the so-called **Kolmogorov Axioms**:

- $\forall E \in \mathcal{E}: \; 0 \leq P(E) \leq 1$
- $P(\Omega) = 1$
- For pairwise disjoint events $E_1, E_2, \ldots \in \mathcal{E}$ holds:

$$P(\bigcup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} P(E_i)$$

  Note that for $|\Omega| < \infty$ the union and sum are finite also.

From these axioms one can conclude the following (incomplete) list of properties:

- $\forall E \in \mathcal{E}: \; P(\overline{E}) = 1 - P(E)$
- $P(\emptyset) = 0$
- If $E_1, E_2 \in \mathcal{E}$ are mutually exclusive, then $P(E_1 \cup E_2) = P(E_1) + P(E_2)$.

Question 1: How to calculate P?

Question 2: Are there "default" event algebras?

Idea for question 1: We have to find a way of distributing (thus the notion *distribution*) the unit mass of probability over all elements $\omega \in \Omega$.

- If Ω is finite or countable a **probability mass function** p is used:

$$p: \Omega \to [0,1] \quad \text{and} \quad \sum_{\omega \in \Omega} p(\omega) = 1$$

- If Ω is uncountable (i. e., continuous) a **probability density function** f is used:

$$f: \Omega \to \mathbb{R} \quad \text{and} \quad \int_{\Omega} f(\omega)\, \mathrm{d}\omega = 1$$

“Default” Event Algebras

Idea for question 2 (“default” event algebras) we have to distinguish again between the cardinalities of Ω:

- Ω finite or countable: $\mathcal{E} = 2^{\Omega}$
- Ω uncountable, e. g. $\Omega = \mathbb{R}$: $\mathcal{E} = \mathcal{B}(\mathbb{R})$

$\mathcal{B}(\mathbb{R})$ is the **Borel Algebra**, i. e., the smallest σ-algebra that contains all closed intervals $[a, b] \subset \mathbb{R}$ with $a < b$.

$\mathcal{B}(\mathbb{R})$ also contains all open intervals and single-item sets.

It is sufficient to note here, that all intervals are contained

$$\{[a, b] \,,]a, b] \,,]a, b[\,, [a, b[\subset \mathbb{R} \mid a < b\} \subset \mathcal{B}(\mathbb{R})$$

because the event of a bread roll having a weight between 80 g and 90 g is represented by the interval $[80, 90]$.

Random Variable

A function $X : D \to M$ is called a **random variable** if and only if the preimage of any value of M is an event (in some probability space).

If X is numeric, we call $F(x)$ with

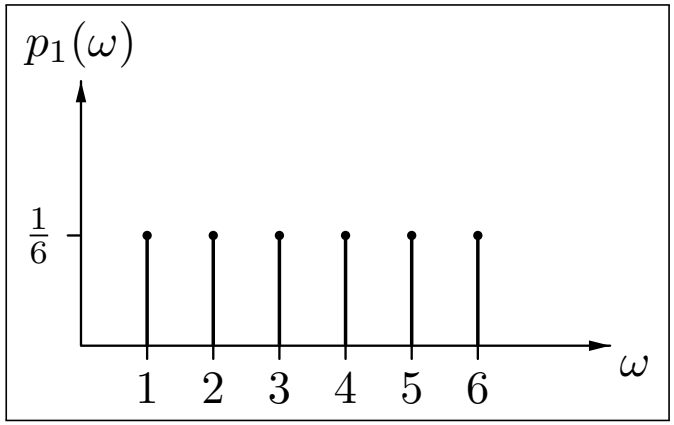
$$F(x) = P(X \leq x)$$

the **distribution function** of X.

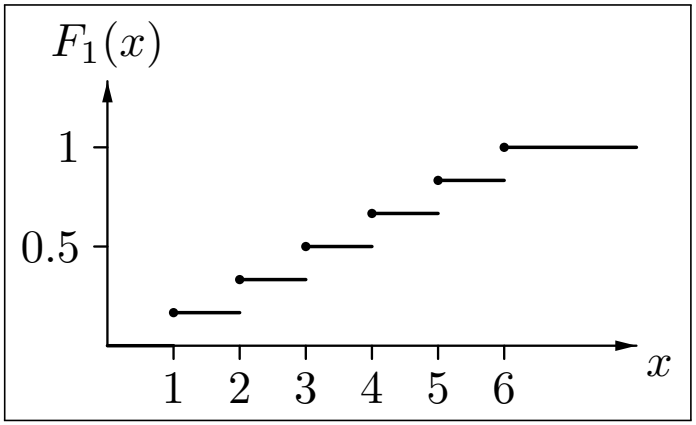
Example: Rolling a Die

$\Omega = \{1, 2, 3, 4, 5, 6\} \quad X = \mathrm{id}$

$$p_1(\omega) = \tfrac{1}{6}$$



$$F_1(x) = P(X \leq x)$$



$$\sum_{\omega \in \Omega} p_1(\omega) = \sum_{i=1}^{6} p_1(\omega_i) = \sum_{i=1}^{6} \frac{1}{6} = 1$$

$$P(X \leq x) = \sum_{x' \leq x} P(X = x')$$

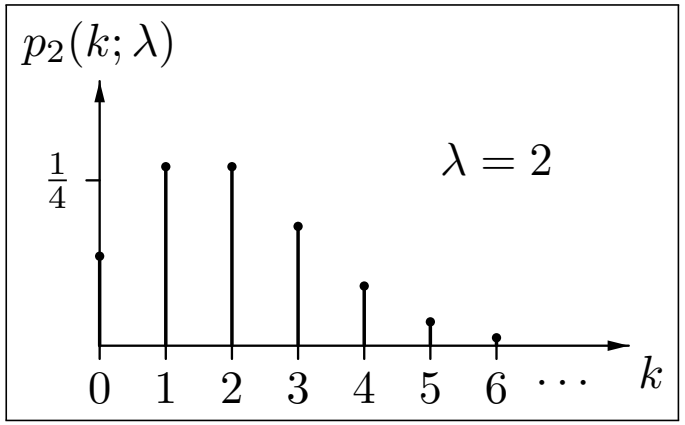
$$P(a < X \leq b) = F_1(b) - F_1(a)$$

$$P(X = x) = P(\{X = x\}) = P(X^{-1}(x)) = P(\{\omega \in \Omega \mid X(\omega) = x\})$$

Example: Arriving Phone Calls

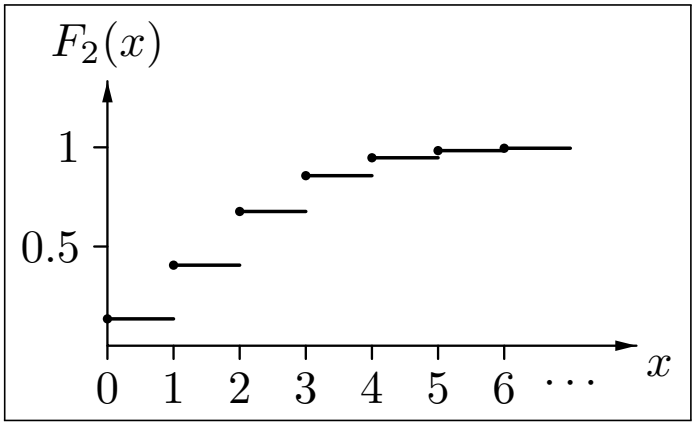
$\Omega = \mathbb{N}_0 \quad X = \mathrm{id}$

$$p_2(k;\lambda) = e^{-\lambda} \cdot \frac{\lambda^k}{k!}$$



$$\sum_{k\in\mathbb{N}_0} p_2(k;\lambda) = \sum_{k=0}^{\infty} e^{-\lambda} \cdot \frac{\lambda^k}{k!}$$
$$= e^{-\lambda} \cdot \underbrace{\sum_{k=0}^{\infty} \frac{\lambda^k}{k!}}_{=e^{\lambda}}$$
$$= e^{-\lambda} \cdot e^{\lambda} = 1$$

$$F_2(k;\lambda) = \sum_{i=0}^{k} e^{-\lambda} \cdot \frac{\lambda^i}{i!}$$



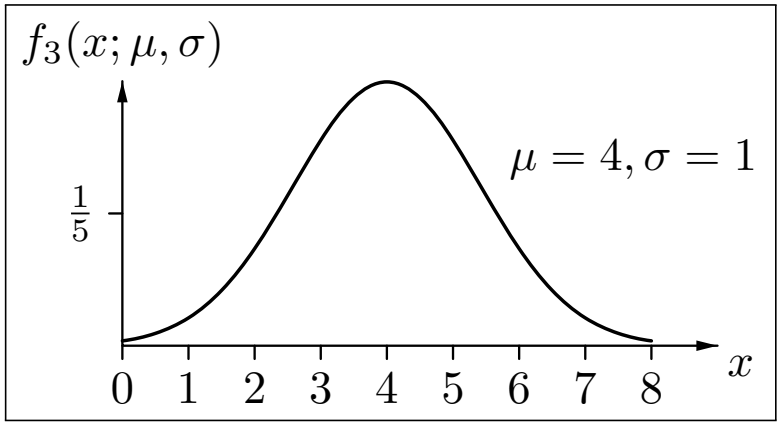
$$P(X \leq x) = \sum_{x' \leq x} P(X = x')$$

$$P(a < X \leq b) = F_2(b) - F_2(a)$$

Example: Weight of a Bread Roll

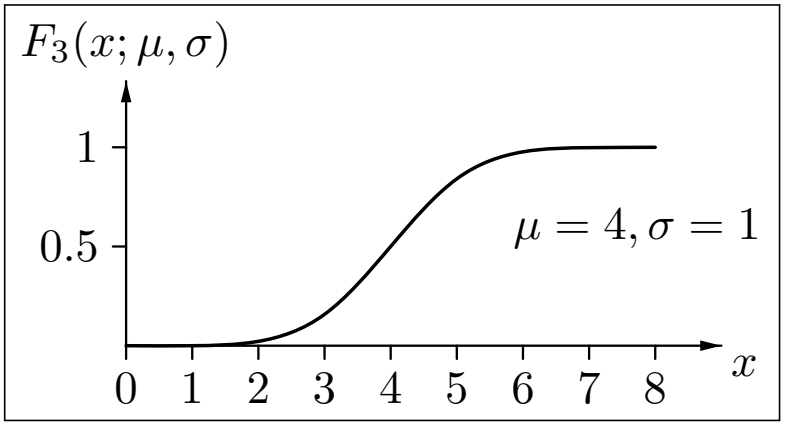
$\Omega = \mathbb{R} \quad X = \mathrm{id}$

$$f_3(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$



$$\int_{-\infty}^{+\infty} f_3(x)\,\mathrm{d}x = 1$$

$$F_3(x) = \int_{-\infty}^{x} f_3(x)\,\mathrm{d}x$$



$$\begin{aligned} P(X \leq x) &= P(]-\infty, x]) \\ &= \int_{-\infty}^{x} f_3(x)\mathrm{d}x \end{aligned}$$

$$\begin{aligned} P(a < X \leq b) &= P(]a, b]) \\ &= \int_{a}^{b} f_3(x)\mathrm{d}x \\ &= F_3(b) - F_3(a) \end{aligned}$$

Poisson Distribution

Limit case of the Binomial distribution:

$$\lim_{n\to\infty} b_X(k; n, p) \;=\; \lim_{n\to\infty} \binom{n}{k} p^k (1-p)^{n-k} \;=\; e^{-\lambda} \cdot \frac{\lambda^k}{k!}$$

with $k = 0, 1, 2, \ldots$ and $\lambda = n \cdot p$.

Expected Value: $E(X) = \lambda$

Variance: $V(X) = \lambda$

Models, e. g.

- Number of cars that pass a gate.
- Number of customers at a register.
- Number of calls at a call center.

λ is the rate parameter (i. e., occurrences per unit time)

Exponential Distribution

A continuous random variable with density function

$$f_X(x;\lambda) = \begin{cases} \lambda \cdot e^{-\lambda x} & \text{if } x \geq 0, \lambda > 0 \\ 0 & \text{otherwise} \end{cases}$$

is **exponentially distributed**.

Expected Value: $E(X) = \frac{1}{\lambda}$ $\quad$ $F_X(x;\lambda) = \begin{cases} 1 - e^{-\lambda x} & \text{if } x \geq 0, \lambda > 0 \\ 0 & \text{otherwise} \end{cases}$

Variance: $V(X) = \frac{1}{\lambda^2}$

Models, e. g.

- Lifetime of electrical devices.
- Waiting times in a queue.
- Time between failures of a system.

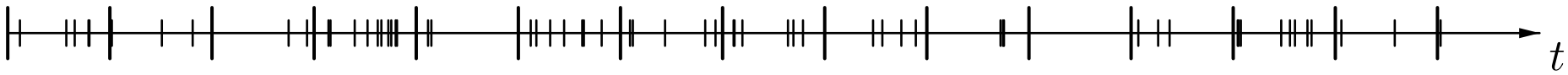
Relation between Poisson and Exponential Distributions

Assume an arrival process with λ arrivals (per unit time, say 1h)

The random variable that describes the **number of arrivals** within the next unit time interval is **Poisson distributed** with parameter λ.

The random variable that describes the probability of the **waiting times between two arrivals** is **exponentially distributed** with (the same!) λ.

Example:



Small ticks denote arrivals, large ticks mark unit time windows.

60 arrivals, 15 unit time windows.

Poisson sample $\vec{x}_P = (4, 3, 2, 10, 2, 7, 5, 6, 4, 3, 0, 3, 8, 2, 1)$

Exponential sample $\vec{x}_E = (0.1192, 0.4544, 0.0821, 0.1352, \ldots)$

$\lambda = 4$

Inductive Statistics

Parameter Estimation
Given an assumption about the type of distribution of the underlying random variable the parameter(s) of the distribution function is estimated.

Hypothesis Testing
A hypothesis about the data generating process is tested by means of the data.

- *Parameter Test*
  Test whether a parameter can have certain values.
- *Goodness-of-Fit Test*
  Test whether a distribution assumption fits the data.
- *Dependence Test*
  Test whether two attributes are dependent.

Model Selection
Among different models that can be used to explain the data the best fitting is selected, taking the complexity of the model into account.

In inductive statistics probability theory is applied to make inferences about the process that generated the data. This presupposes that the sample is the result of a random experiment, a so-called **random sample**.

The random variable yielding the sample value x_i is denoted X_i.
x_i is called a **instantiation** of the random variable X_i.

A random sample $x = (x_1, \ldots, x_n)$ is an instantiation of the **random vector** $X = (X_1, \ldots, X_n)$.

A random sample is called **independent** if the random variables $X_1, \ldots, X_n$ are (stochastically) independent, i. e. if

$$\forall c_1, \ldots, c_n \in \mathbb{R}: \quad P\left(\bigwedge_{i=1}^{n} X_i \leq c_i\right) = \prod_{i=1}^{n} P(X_i \leq c_i).$$

An independent random sample is called **simple** if the random variables $X_1, \ldots, X_n$ have the same distribution function.

Given:

A data set and

a family of parameterized distributions functions of the same type, e.g.

- the family of binomial distributions $b_X(x; p, n)$ with the parameters p, $0 \leq p \leq 1$, and $n \in \mathbb{N}$, where n is the sample size,
- the family of normal distributions $N_X(x; \mu, \sigma^2)$ with the parameters μ (expected value) and σ^2 (variance).

Assumption:

The process that generated the data can be described well by an element of the given family of distribution functions.

Desired:

The element of the given family of distribution functions (determined by its parameters) that is the best model for the data.

Parameter Estimation

Methods that yield an estimate for a parameter are called **estimators**.

Estimators are **statistics**, i.e. functions of the values in a sample.

As a consequence they are functions of (instantiations of) random variables and thus (instantiations of) random variables themselves.

Therefore we can use all of probability theory to analyze estimators.

There are two types of parameter estimation:

- **Point Estimators**
  Point estimators determine the best value of a parameter
  w.r.t. the data and certain quality criteria.
- **Interval Estimators**
  Interval estimators yield a region, a so-called **confidence interval**,
  in which the true value of the parameter lies with high certainty.

Not all statistics, that is, not all functions of the sample values are reasonable and useful estimator. Desirable properties are:

Consistency
With growing data volume the estimated value should get closer and closer to the true value, at least with higher and higher probability.
Formally: If T_n is an estimator for the parameter θ, it should be

$$\forall \varepsilon > 0 : \quad \lim_{n \to \infty} P(|T_n(X) - \theta| < \varepsilon) = 1,$$

where $T_n(X)$ calculates the estimate for θ from the independent and simple random sample $X = \{x_1, \ldots, x_n\}$.

Unbiasedness
An estimator should not tend to over- or underestimate the parameter.
Rather it should yield, on average, the correct value.
Formally this means

$$E(T) = \theta.$$

Efficiency
The estimation should be as precise as possible, that is, the deviation from the true value should be as small as possible. Formally: If T and U are two estimators for the same parameter θ, then T is called *more efficient* than U if

$$D^2(T) < D^2(U).$$

Sufficiency
An estimator should exploit all information about the parameter contained in the data. More precisely: two samples that yield the same estimate should have the same probability (otherwise there is unused information).
Formally: an estimator T for a parameter θ is called sufficient iff for all samples $x = (x_1, \ldots, x_n)$ with $T(x) = t$ the expression

$$\frac{f_{X_1}(x_1; \theta) \cdots f_{X_n}(x_n; \theta)}{f_T(t; \theta)}$$

is independent of θ.

Point Estimation: Example

Given: a family of **uniform distributions** on the interval $[0, \theta]$, i. e.

$$f_X(x; \theta) = \begin{cases} \frac{1}{\theta}, & \text{if } 0 \leq x \leq \theta, \\ 0, & \text{otherwise.} \end{cases}$$

Desired: an estimate for the unknown parameter θ.

We will now consider two estimators for the parameter θ
and compare their properties.

- $T = \max\{X_1, \ldots, X_n\}$ (if it is clear from the context, we will write T instead of T_n)
- $U = \frac{n+1}{n} \max\{X_1, \ldots, X_n\}$

General approach:

- Find the probability density function of the estimator.
- Check the desirable properties by exploiting this density function.

Point Estimation: Example

To analyze the estimator $T = \max\{X_1, \ldots, X_n\}$, we compute its density function:

$$\begin{aligned} f_T(t;\theta) &= \frac{\mathrm{d}}{\mathrm{d}t} F_T(t;\theta) = \frac{\mathrm{d}}{\mathrm{d}t} P(T \leq t) \\ &= \frac{\mathrm{d}}{\mathrm{d}t} P(\max\{X_1, \ldots, X_n\} \leq t) \\ &= \frac{\mathrm{d}}{\mathrm{d}t} P\Big(\bigwedge_{i=1}^{n} X_i \leq t\Big) = \frac{\mathrm{d}}{\mathrm{d}t} \prod_{i=1}^{n} P(X_i \leq t) \\ &= \frac{\mathrm{d}}{\mathrm{d}t} (F_X(t;\theta))^n = n \cdot (F_X(t;\theta))^{n-1} f_X(t,\theta) \end{aligned}$$

where

$$F_X(x;\theta) = \int_{-\infty}^{x} f_X(x;\theta)\mathrm{d}x = \begin{cases} 0, & \text{if } x \leq 0, \\ \frac{x}{\theta}, & \text{if } 0 \leq x \leq \theta, \\ 1, & \text{if } x \geq \theta. \end{cases}$$

Therefore it is

$$f_T(t;\theta) = \frac{n \cdot t^{n-1}}{\theta^n} \quad \text{for} \quad 0 \leq t \leq \theta, \quad \text{and 0 otherwise.}$$

Point Estimation: Example

The estimator $T_n = \max\{X_1, \ldots, X_n\}$ is **consistent**:

$$\begin{aligned}
\lim_{n\to\infty} P(|T_n - \theta| < \epsilon) &= \lim_{n\to\infty} P(T_n > \theta - \epsilon) \\
&= \lim_{n\to\infty} \int_{\theta-\epsilon}^{\theta} \frac{n \cdot t^{n-1}}{\theta^n} \mathrm{d}t \;=\; \lim_{n\to\infty} \left[\frac{t^n}{\theta^n}\right]_{\theta-\epsilon}^{\theta} \\
&= \lim_{n\to\infty} \left(\frac{\theta^n}{\theta^n} - \frac{(\theta-\epsilon)^n}{\theta^n}\right) \\
&= \lim_{n\to\infty} \left(1 - \left(\frac{\theta-\epsilon}{\theta}\right)^n\right) \;=\; 1
\end{aligned}$$

It is **not unbiased**:

$$\begin{aligned}
E(T) &= \int_{-\infty}^{\infty} t \cdot f_T(t;\theta) \mathrm{d}t \;=\; \int_0^{\theta} t \cdot \frac{n \cdot t^{n-1}}{\theta^n} \mathrm{d}t \\
&= \left[\frac{n \cdot t^{n+1}}{(n+1)\theta^n}\right]_0^{\theta} \;=\; \frac{n}{n+1}\theta \;<\; \theta \quad \text{for } n < \infty.
\end{aligned}$$

Point Estimation: Example

The estimator $U = \frac{n+1}{n} \max\{X_1, \ldots, X_n\}$ has the density function

$$f_U(u;\theta) = \frac{n^{n+1}}{(n+1)^n} \frac{u^{n-1}}{\theta^n} \qquad \text{for } 0 \leq t \leq \frac{n+1}{n}\theta, \text{ and 0 otherwise.}$$

The estimator U is **consistent** (without formal proof).

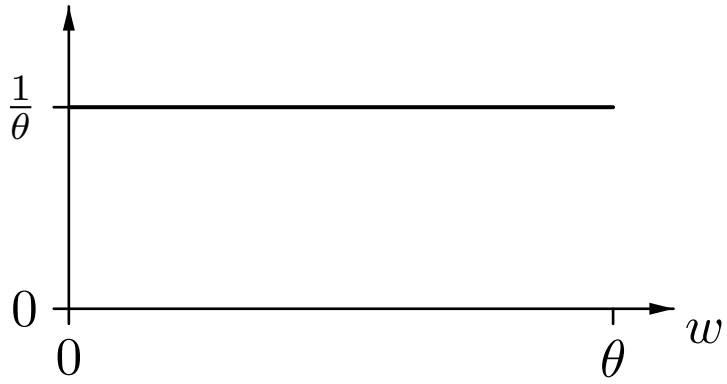
It is **unbiased**:

$$\begin{aligned}
E(U) &= \int_{-\infty}^{\infty} u \cdot f_U(u;\theta) \mathrm{d}u \\
&= \int_0^{\frac{n+1}{n}\theta} u \cdot \frac{n^{n+1}}{(n+1)^n} \frac{u^{n-1}}{\theta^n} \mathrm{d}u \\
&= \frac{n^{n+1}}{(n+1)^n \theta^n} \left[\frac{u^{n+1}}{n+1} \right]_0^{\frac{n+1}{n}\theta} \\
&= \frac{n^{n+1}}{(n+1)^n \theta^n} \cdot \frac{1}{n+1} \left(\frac{n+1}{n}\theta \right)^{n+1} = \theta
\end{aligned}$$

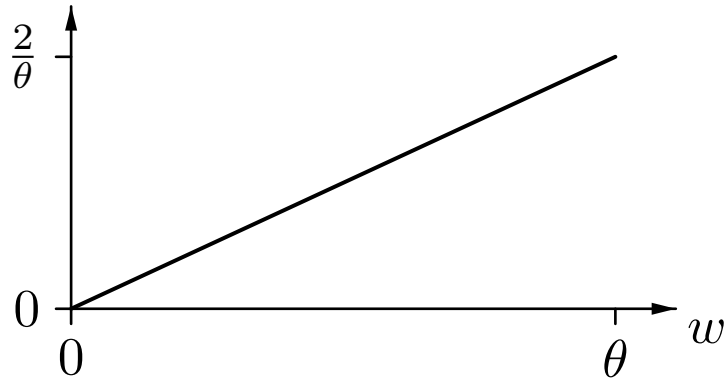
Densities of Estimators

What does the density of the estimator $W = \max\{X_1, \ldots, X_n\}$ look like? (w. r. t. n)

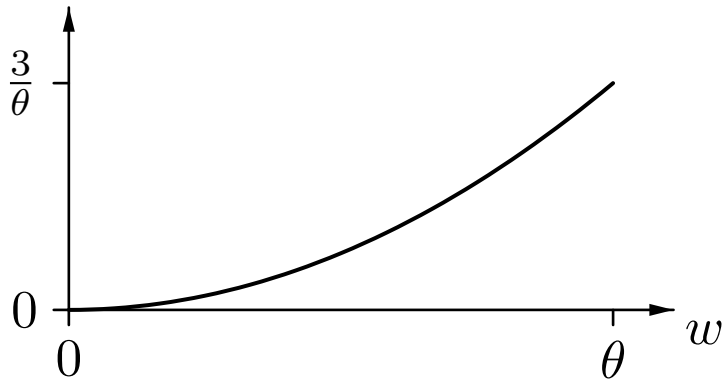
$f_W(w; \theta, 1)$



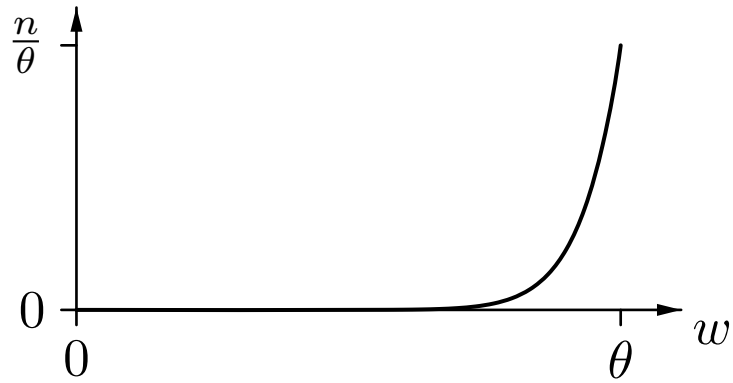
$f_W(w; \theta, 2)$



$f_W(w; \theta, 3)$



$f_W(w; \theta, n)$



Note the different scales for the y-axes!

Point Estimation: Example

Given: a family of **normal distributions** $N_X(x; \mu, \sigma^2)$

$$f_X(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Desired: estimates for the unknown parameters μ and σ^2.

The median and the arithmetic mean of the sample
are both consistent and unbiased estimators for the parameter μ.

The median is less efficient than the arithmetic mean.

The function $V^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ is a consistent, but **biased** estimator for the parameter σ^2 (it tends to underestimate the variance).

The function $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$, however, is a consistent and **unbiased** estimator for σ^2 (this explains the definition of the empirical variance).

Point Estimation: Example

Given: a family of **polynomial distributions**
(synonym: multinomial distribution)

$$f_{X_1,\ldots,X_k}(x_1,\ldots,x_k;\theta_1,\ldots,\theta_k,n) = \frac{n!}{\prod_{i=1}^{k} x_i!} \prod_{i=1}^{k} \theta_i^{x_i},$$

(n is the sample size, the x_i are the frequencies of the different values a_i, $i = 1,\ldots,k$, and the θ_i are the probabilities with which the values a_i occur.)

Desired: estimates for the unknown parameters $\theta_1,\ldots,\theta_k$

The relative frequencies $R_i = \frac{X_i}{n}$ of the different values a_i, $i = 1,\ldots,k$, are

- consistent,
- unbiased,
- most efficient, and
- sufficient estimators for the θ_i.

Polynomial Distribution: Example

Consider the random experiment of picking a person out of a population (with replacement, technically) and determine her eye color.

$k = 3$: $\quad a_1 \mathrel{\widehat{=}} \mathsf{blue}, \quad a_2 \mathrel{\widehat{=}} \mathsf{green}, \quad a_3 \mathrel{\widehat{=}} \mathsf{brown}$

$\theta_1 = 0.3, \quad \theta_2 = 0.3, \quad \theta_3 = 0.4$

The probability of finding 2 persons with blue eyes, 4 persons with green eyes and 4 persons with brown eyes (in a sample of size 10) is:

$$f_{X_1,X_2,X_3}(2, 4, 4;\, \theta_1, \theta_2, \theta_3, 10) = \frac{10!}{2!4!4!} \cdot 0.3^2 \cdot 0.3^4 \cdot 0.4^4 \approx 0.0588$$

Note:

$$\sum_{i=1}^{k} x_i = n \quad \text{and} \quad \sum_{i=1}^{k} \theta_i = 1$$

How Can We Find Estimators?

Up to now we analyzed given estimators,
now we consider the question how to find them.

There are three main approaches to find estimators:

- **Method of Moments**
  Derive an estimator for a parameter from the moments of a distribution and its generator function.
  (We do not consider this method here.)
- **Maximum Likelihood Estimation**
  Choose the (set of) parameter value(s) that makes the sample most likely.
- **Maximum A-posteriori Estimation**
  Choose a prior distribution on the range of parameter values, apply Bayes' rule to compute the posterior probability from the sample, and choose the (set of) parameter value(s) that maximizes this probability.

Maximum Likelihood Estimation

General idea: **Choose the (set of) parameter value(s) that makes the sample most likely.**

If the parameter value(s) were known, it would be possible to compute the probability of the sample. With unknown parameter value(s), however, it is still possible to state this probability as a function of the parameter(s).

Formally this can be described as choosing the value θ that maximizes

$$L(D; \theta) = f(D \mid \theta),$$

where D are the sample data and L is called the **Likelihood Function**.

Technically the estimator is determined by

- setting up the likelihood function,
- forming its partial derivative(s) w.r.t. the parameter(s), and
- setting these derivatives equal to zero (necessary condition for a maximum).

Brief Excursion: Function Optimization

Task: Find values $\vec{x} = (x_1, \ldots, x_m)$ such that $f(\vec{x}) = f(x_1, \ldots, x_m)$ is optimal.

Often feasible approach:

A necessary condition for a (local) optimum (maximum or minimum) is that the partial derivatives w.r.t. the parameters vanish (Pierre Fermat).

Therefore: (Try to) solve the equation system that results from setting all partial derivatives w.r.t. the parameters equal to zero.

Example task: Minimize $f(x, y) = x^2 + y^2 + xy - 4x - 5y$.

Solution procedure:

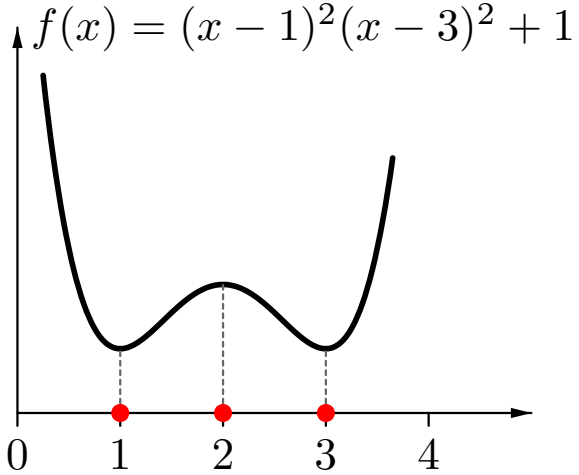
Take the partial derivatives of the objective function and set them to zero:

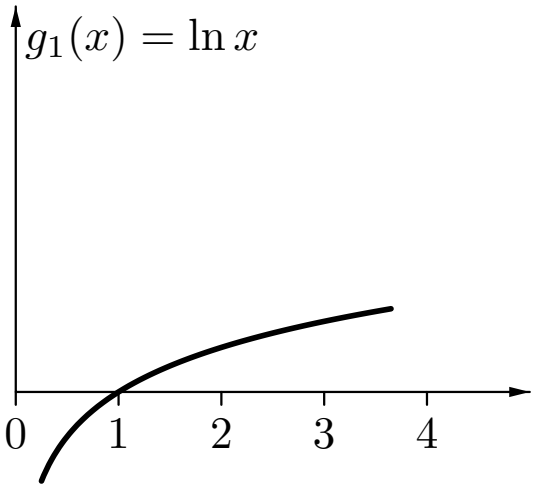
$$\frac{\partial f}{\partial x} = 2x + y - 4 = 0, \qquad \frac{\partial f}{\partial y} = 2y + x - 5 = 0.$$

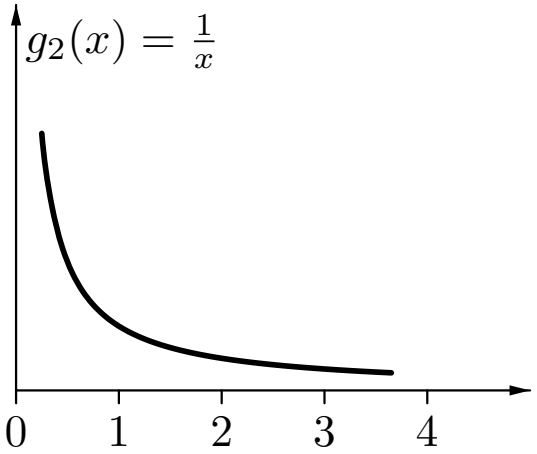
Solve the resulting (here: linear) equation system: $x = 1, \quad y = 2.$

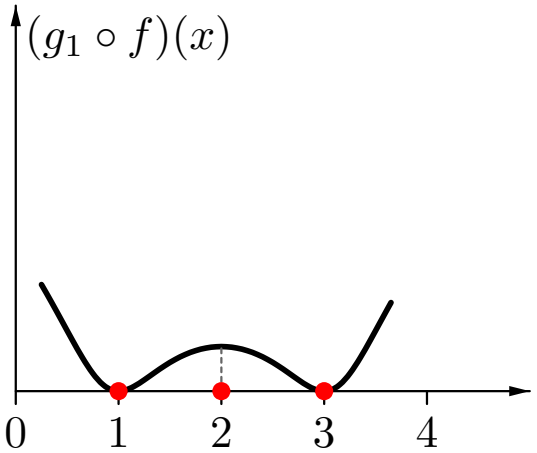
Optima of a Function

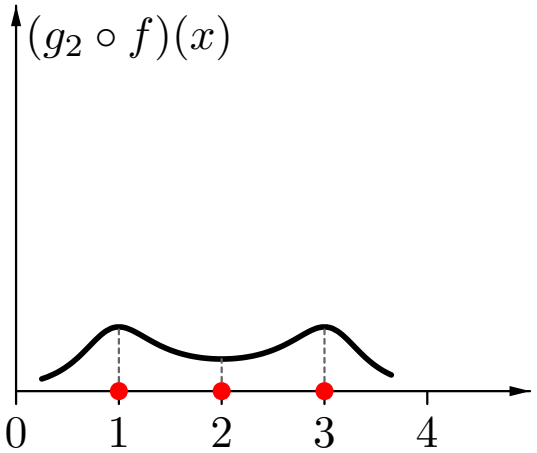
The **locations** of the optima of a function f do not change if f is composed with a **strictly monotonic** (increasing or decreasing) function g.











Maximum Likelihood Estimation: Example

Given: a family of **normal distributions** $N_X(x; \mu, \sigma^2)$

$$f_X(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Desired: estimators for the unknown parameters μ and σ^2.

The **Likelihood Function**, which describes the probability of the data, is

$$L(x_1, \ldots, x_n; \mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right).$$

To simplify the technical task of forming the partial derivatives,
we consider the natural logarithm of the likelihood function, i. e.

$$\ln L(x_1, \ldots, x_n; \mu, \sigma^2) = -n \ln\left(\sqrt{2\pi\sigma^2}\right) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2.$$

Maximum Likelihood Estimation: Example

Estimator for the **expected value** μ:

$$\frac{\partial}{\partial \mu} \ln L(x_1, \ldots, x_n; \mu, \sigma^2) = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) \stackrel{!}{=} 0$$

$$\Rightarrow \qquad \sum_{i=1}^{n} (x_i - \mu) = \left(\sum_{i=1}^{n} x_i \right) - n\mu \stackrel{!}{=} 0 \qquad \Rightarrow \qquad \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Estimator for the **variance** σ^2:

$$\frac{\partial}{\partial \sigma^2} \ln L(x_1, \ldots, x_n; \mu, \sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (x_i - \mu)^2 \stackrel{!}{=} 0$$

$$\Rightarrow \qquad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})^2 = \frac{1}{n} \sum_{i=1}^{n} x_i^2 - \frac{1}{n^2} \left(\sum_{i=1}^{n} x_i \right)^2 \qquad \text{(biased!)}$$

Maximum A-posteriori Estimation: Motivation

Consider the following three situations:

A drunkard claims to be able to predict the side on which a thrown coin will land (head or tails). On ten trials he always states the correct side beforehand.

A tea lover claims that she is able to taste whether the tea or the milk was poured into the cup first. On ten trials she always identifies the correct order.

An expert of classical music claims to be able to recognize from a single sheet of music whether the composer was Mozart or somebody else. On ten trials he is indeed correct every time.

Maximum likelihood estimation treats all situations alike, because formally the samples are the same. However, this is implausible:

We do not believe the drunkard at all, despite the sample data.

We highly doubt the tea drinker, but tend to consider the data as evidence.

We tend to believe the music expert easily.

Maximum A-posteriori Estimation

Background knowledge about the plausible values can be incorporated by

- using a **prior distribution** on the domain of the parameter and
- adapting this distribution with **Bayes' rule** and the data.

Formally maximum a-posteriori estimation is defined as follows:
find the parameter value θ that maximizes

$$f(\theta \mid D) = \frac{f(D \mid \theta) f(\theta)}{f(D)} = \frac{f(D \mid \theta) f(\theta)}{\int_{-\infty}^{\infty} f(D \mid \theta') f(\theta') \mathrm{d}\theta'}$$

As a comparison: maximum likelihood estimation maximizes

$$f(D \mid \theta)$$

Note that $f(D)$ need not be computed: It is the same for all parameter values and since we are only interested in the value θ that maximizes $f(\theta \mid D)$ and not the *value of* $f(\theta \mid D)$, we can treat it as a normalization constant.

Maximum A-posteriori Estimation: Example

Given: a family of **binomial distributions**

$$f_X(x; \theta, n) = \binom{n}{x} \theta^x (1-\theta)^{n-x}.$$

Desired: an estimator for the unknown parameter θ.

a) **Uniform prior**: $f(\theta) = 1, \qquad 0 \leq \theta \leq 1.$

$$f(\theta \mid D) = \gamma \binom{n}{x} \theta^x (1-\theta)^{n-x} \cdot 1 \qquad \Rightarrow \qquad \hat{\theta} = \frac{x}{n}$$

b) **Tendency towards** $\frac{1}{2}$: $f(\theta) = 6\theta(1-\theta), \qquad 0 \leq \theta \leq 1.$

$$f(\theta \mid D) = \gamma \binom{n}{x} \theta^x (1-\theta)^{n-x} \cdot \theta(1-\theta) = \gamma \binom{n}{x} \theta^{x+1} (1-\theta)^{n-x+1}$$

$$\Rightarrow \qquad \hat{\theta} = \frac{x+1}{n+2}$$

Excursion: Dirichlet's Integral

For computing the normalization factors of the probability density functions that occur with polynomial distributions, **Dirichlet's Integral** is helpful:

$$\int_{\theta_1} \dots \int_{\theta_k} \prod_{i=1}^{k} \theta_i^{x_i} \, \mathrm{d}\theta_1 \dots \mathrm{d}\theta_k \;=\; \frac{\prod_{i=1}^{k} \Gamma(x_i + 1)}{\Gamma(n + k)}, \qquad \text{where} \quad n = \sum_{i=1}^{k} x_i$$

and the Γ-function is the so-called **generalized factorial**:

$$\Gamma(x) = \int_0^\infty e^{-t} t^{x-1} \mathrm{d}t, \qquad x > 0,$$

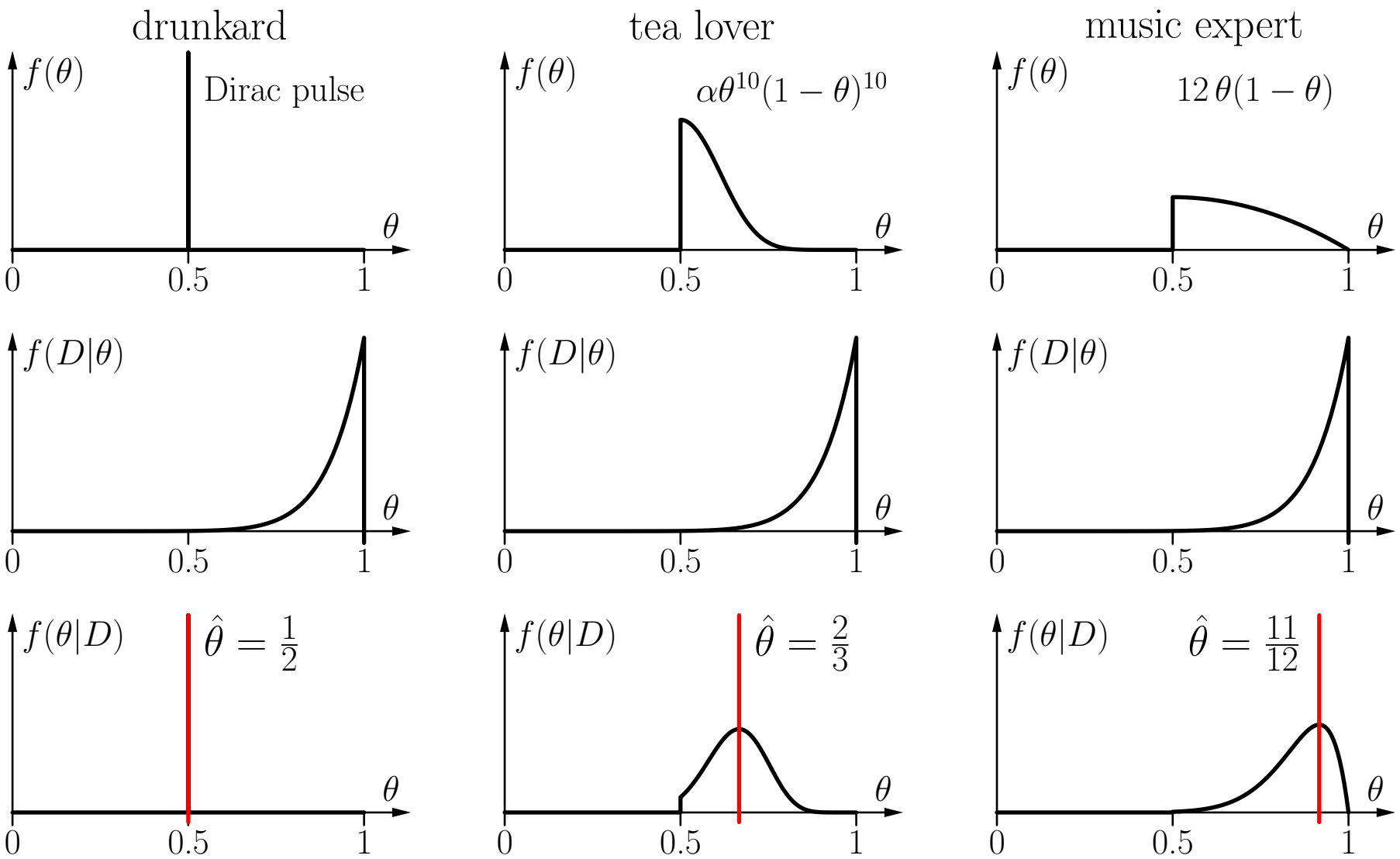
which satisfies

$$\Gamma(x+1) = x \cdot \Gamma(x), \qquad \Gamma(\tfrac{1}{2}) = \sqrt{\pi}, \qquad \Gamma(1) = 1.$$

Example: the normalization factor α for the binomial distribution prior $f(\theta) = \alpha\, \theta^2 (1-\theta)^3$ is

$$\alpha \;=\; \frac{1}{\int_\theta \theta^2 (1-\theta)^3 \mathrm{d}\theta} \;=\; \frac{\Gamma(5+2)}{\Gamma(2+1)\,\Gamma(3+1)} \;=\; \frac{6!}{2!\,3!} \;=\; \frac{720}{12} \;=\; 60.$$

Maximum A-posteriori Estimation: Example



Interval Estimation

In general the estimated value of a parameter will differ from the true value.

It is desirable to be able to make an assertion about the possible deviations.

The simplest possibility is to state not only a point estimate,
but also the standard deviation of the estimator:

$$t \pm D(T) = t \pm \sqrt{D^2(T)}.$$

A better possibility is to find intervals that contain the true value with high probability. Formally they can be defined as follows:
Let $A = g_A(X_1, \ldots, X_n)$ and $B = g_B(X_1, \ldots, X_n)$ be two statistics, such that

$$P(A < \theta < B) = 1 - \alpha, \qquad P(\theta \leq A) = \frac{\alpha}{2}, \qquad P(\theta \geq B) = \frac{\alpha}{2}.$$

Then the random interval $[A, B]$ (or an instantiation $[a, b]$ of this interval) is called $(1-\alpha) \cdot 100\%$ **confidence interval** for θ. The value $1-\alpha$ is called **confidence level**.

This definition of a confidence interval is not specific enough:
A and B are not uniquely determined.

Common solution: Start from a point estimator T for the unknown parameter θ and define A and B as functions of T:

$$A = h_A(T) \quad \text{and} \quad B = h_B(T).$$

Instead of $A \leq \theta \leq B$ consider the corresponding event w.r.t. the estimator T, that is, $A^* \leq T \leq B^*$.

Determine $A = h_A(T)$ and $B = h_B(T)$ from the inverse functions $A^* = h_A^{-1}(\theta)$ and $B^* = h_B^{-1}(\theta)$.

$$\begin{array}{rlcl} \text{Procedure:} & P(A^* < T < B^*) & = & 1 - \alpha \\ \Rightarrow & P(h_A^{-1}(\theta) < T < h_B^{-1}(\theta)) & = & 1 - \alpha \\ \Rightarrow & P(h_A(T) < \theta < h_B(T)) & = & 1 - \alpha \\ \Rightarrow & P(A < \theta < B) & = & 1 - \alpha. \end{array}$$

Interval Estimation: Example

Given: a family of **uniform distributions** on the interval $[0, \theta]$, i.e.

$$f_X(x;\theta) = \begin{cases} \frac{1}{\theta}, & \text{if } 0 \leq x \leq \theta, \\ 0, & \text{otherwise.} \end{cases}$$

Desired: a confidence interval for the unknown parameter θ.

Start from the unbiased point estimator $U = \frac{n+1}{n} \max\{X_1, \ldots, X_n\}$:

$$P(U \leq B^*) = \int_0^{B^*} f_U(u;\theta)\mathrm{d}u = \frac{\alpha}{2}$$

$$P(U \geq A^*) = \int_{A^*}^{\frac{n+1}{n}\theta} f_U(u;\theta)\mathrm{d}u = \frac{\alpha}{2}$$

From the study of point estimators we know

$$f_U(u;\theta) = \frac{n^{n+1}}{(n+1)^n}\frac{u^{n-1}}{\theta^n}.$$

Interval Estimation: Example

Solving the integrals gives us

$$B^* = \sqrt[n]{\frac{\alpha}{2}}\,\frac{n+1}{n}\,\theta \qquad \text{and} \qquad A^* = \sqrt[n]{1-\frac{\alpha}{2}}\,\frac{n+1}{n}\,\theta,$$

that is,

$$P\left(\sqrt[n]{\frac{\alpha}{2}}\,\frac{n+1}{n}\,\theta < U < \sqrt[n]{1-\frac{\alpha}{2}}\,\frac{n+1}{n}\,\theta\right) = 1-\alpha.$$

Computing the inverse functions leads to

$$P\left(\frac{U}{\sqrt[n]{1-\frac{\alpha}{2}}\frac{n+1}{n}} < \theta < \frac{U}{\sqrt[n]{\frac{\alpha}{2}}\frac{n+1}{n}}\right) = 1-\alpha,$$

that is,

$$A = \frac{U}{\sqrt[n]{1-\frac{\alpha}{2}}\frac{n+1}{n}} \qquad \text{and} \qquad B = \frac{U}{\sqrt[n]{\frac{\alpha}{2}}\frac{n+1}{n}}.$$

Interval Estimation: Common Misconceptions

Since A and B are functions of random variables $\vec{X} = (X_1, \ldots, X_n)$ (modelling the underlying random sampling), they are random variables themselves and thus a statement like

$$P(A(\vec{X} < \theta < B(\vec{X})) = 1 - \alpha$$

makes sense.

If, however, applied to a specific random sample $\vec{x}$ the interval borders $a = A(\vec{x})$ and $b = B(\vec{x})$ become fixed and are not random anymore.

A probability statement about $a < \theta < b$ would be nonsensical because either $\theta \in [a, b]$ or $\theta \notin [a, b]$.

Therefore it is incorrect to say:
"The true parameter θ lies with $(1 - \alpha) \cdot 100\%$ probability within the confidence interval."

Correct: **"This confidence interval has been generated by a procedure which returns for $(1 - \alpha) \cdot 100\%$ of all possible samples $\vec{x}$ an interval that contains the true parameter θ."**

Interval Estimation: Common Misconceptions

Relation to sample size n and confidence level α.

Width of a confidence interval can be considered a measure of imprecision or inaccuracy, i. e., the smaller the interval the more accurate the estimation.
(although the real parameter may not be within the interval at all, of course).

Increasing n yields a smaller interval.

Increasing α yields a smaller interval. (Often misunderstood!)

Example: random variable X with binomial distribution: $b_X(x; p, n)$

Let x be the number of positive outcomes in the sample of size n.
The $(1 - \alpha) \cdot 100\%$ confidence interval for p reads:

$$\left[r - z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{r(1-r)}{n-1}}, \quad r + z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{r(1-r)}{n-1}} \right] \quad \text{with} \quad r = \frac{x}{n}$$

(with $z_a = \Phi^{-1}(a)$ and Φ being the standard normal distribution function)

Hypothesis Testing

A **hypothesis test** is a statistical procedure with which a decision is made between two contrary hypothesis about the process that generated the data.

The two hypotheses can refer to

- the value of a parameter **(Parameter Test)**,
- a distribution assumption **(Goodness-of-Fit Test)**,
- the dependence of two attributes **(Dependence Test)**.

One of the two hypothesis is preferred, that is, in case of doubt the decision is made in its favor. (One says that it gets the "benefit of the doubt".)

The preferred hypothesis is called the **Null Hypothesis** H_0,
the other hypothesis is called the **Alternative Hypothesis** H_a.

Intuitively: the null hypothesis H_0 is put on trial.
Only if the evidence is strong enough, it is convicted (i.e. rejected).
If there is doubt, however, it is acquitted (i.e. accepted).

Hypothesis Testing

The test decision is based on a **test statistic**,
that is, a function of the sample values.

The null hypothesis is rejected if the value of the test statistic
lies inside the so-called **critical region** C.

Developing a hypothesis test consists in finding the critical region
for a given test statistic and significance level (see below).

The test decision may be wrong. There are two possible types of errors:

> **Type 1:** The null hypothesis H_0 is rejected, even though it is correct.
> **Type 2:** The null hypothesis H_0 is accepted, even though it is false.

Type 1 errors are considered to be more severe,
since the null hypothesis gets "the benefit of the doubt".

Therefore it is tried to restrict the probability of a type 1 error to a certain maximum α. This maximum value α is called **significance level**.

Example: Outlier Detection - Single Attributes

Numerical attributes:

Outliers in boxplots.

Problems: Asymmetric distribution, large data sets

Statistical tests, for example *Grubb's test*:

Define the statistic

$G = \frac{\max\{\|x_i - \overline{x}\| : 1 \leq i \leq n}{s}$, where $x_1, \ldots, x_n$ is the sample, $\overline{x}$ its mean value and s its empirical standard deviation. For a given significance level α, the null hypothesis that the sample coming from a *normal distribution* does not contain outliers is rejected if

$$G > \frac{n-1}{\sqrt{n}} \sqrt{\frac{t^2_{1-\alpha/(2n),n-2}}{n-2+t^2_{1-\alpha/(2n),n-2}}}$$

where $t_{1-\alpha/(2n),n-2}$ denotes the $(1-\alpha/(2n))$-quantile of the t-distribution with $n-$ degrees of freedom.

Parameter Test

In a parameter test the contrary hypotheses refer to the value of a parameter, for example (one-sided test):

$$H_0: \quad \theta \geq \theta_0, \qquad\qquad H_a: \quad \theta < \theta_0.$$

For such a test usually a point estimator T is chosen as the test statistic.

The null hypothesis H_0 is rejected if the value t of the point estimator does not exceed a certain value c, the so-called **critical value** (i.e. $C = (-\infty, c]$).

Formally the critical value c is determined as follows: We consider

$$\beta(\theta) = P_\theta(H_0 \text{ is rejected}) = P_\theta(T \in C),$$

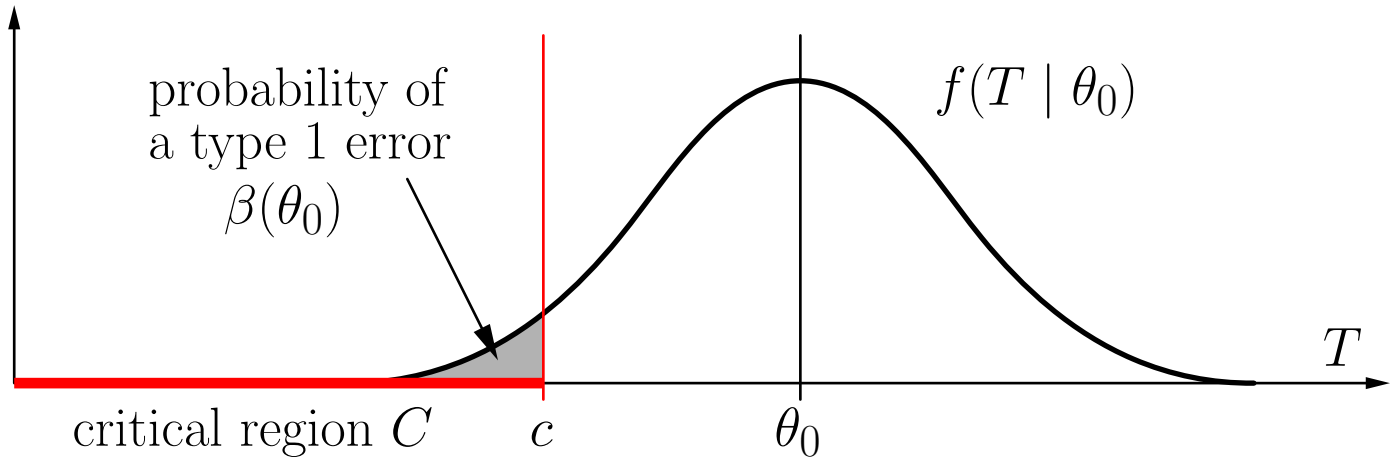
the so-called **power** β of the test.

The power must not exceed the significance level α for values θ satisfying H_0:

$$\max_{\theta:\theta \text{ satisfies } H_0} \beta(\theta) \leq \alpha. \qquad \text{(here:} \qquad \beta(\theta_0) \leq \alpha)$$

Parameter Test: Intuition

The probability of a type 1 error is the area under the estimator's probability density function $f(T \mid \theta_0)$ to the left of the critical value c. (Note: This example illustrates $H_0 : \theta \geq \theta_0$ and $H_a : \theta < \theta_0$.)

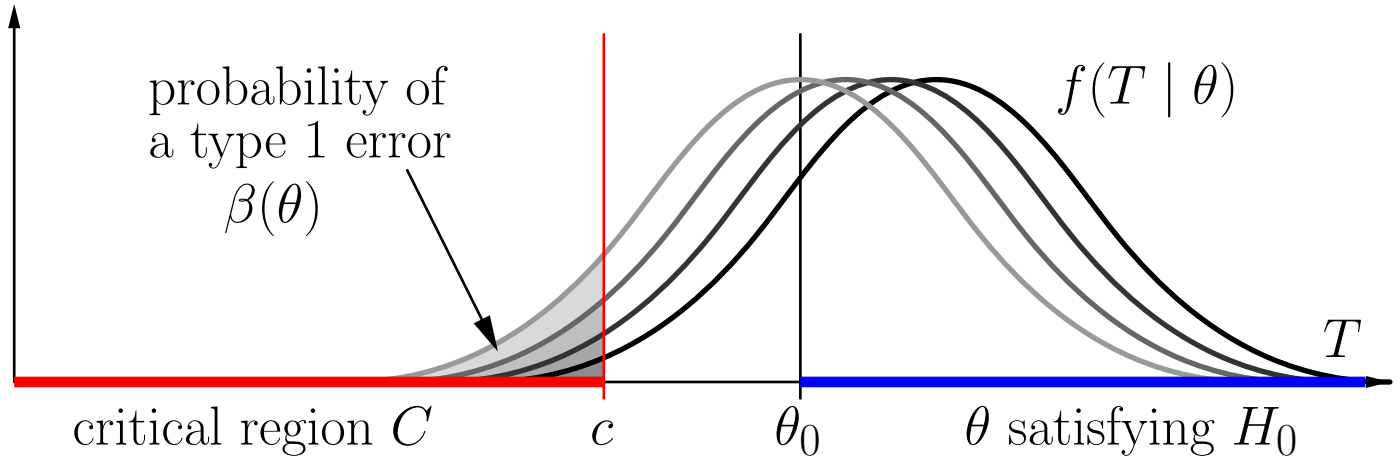


Obviously the probability of a type 1 error depends on the location of the critical value c: higher values mean a higher error probability.

Idea: Choose the location of the cricital value so that the maximal probability of a type 1 error equals α, the chosen significance level.

Parameter Test: Intuition

What is so special about θ_0 that we use $f(T \mid \theta_0)$?



In principle, all θ satisfying H_0 have to be considered,
that is, all density functions $f(T \mid \theta)$ with $\theta \geq \theta_0$.

Among these values θ, the one with the highest probability of a type 1 error
(i.e., the one with the highest power $\beta(\theta)$) determines the critical value.

Intuitively: we consider the **worst possible case**.

Parameter Test: Example

We consider a one-sided test of the expected value μ of a normal distribution $N(\mu, \sigma^2)$ with known variance σ^2, i.e., we consider the hypotheses

$$H_0: \quad \mu \geq \mu_0, \qquad\qquad H_a: \quad \mu < \mu_0.$$

As a test statistic we use the standard point estimator for the expected value

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

This point estimator has the probability density

$$f_{\bar{X}}(x) = N\left(x; \mu, \frac{\sigma^2}{n}\right).$$

Therefore it is (with the $N(0,1)$-distributed random variable Z)

$$\alpha = \beta(\mu_0) = P_{\mu_0}(\bar{X} \leq c) = P\left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \leq \frac{c - \mu_0}{\sigma/\sqrt{n}}\right) = P\left(Z \leq \frac{c - \mu_0}{\sigma/\sqrt{n}}\right).$$

Parameter Test: Example

We have as a result that

$$\alpha = \Phi\left(\frac{c - \mu_0}{\sigma/\sqrt{n}}\right),$$

where Φ is the distribution function of the standard normal distribution.

The distribution function Φ is tabulated, because it cannot be represented in closed form. From such a table we retrieve the value z_α satisfying $\alpha = \Phi(z_\alpha)$.

Then the critical value is

$$c = \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}.$$

(Note that the value of z_α is negative due to the usually small value of α. Typical values are $\alpha = 0.1$, $\alpha = 0.05$ or $\alpha = 0.01$.)

H_0 is rejected if the value $\bar{x}$ of the point estimator $\bar{X}$ does not exceed c, otherwise it is accepted.

Parameter Test: Example

Let $\sigma = 5.4$, $n = 25$ and $\bar{x} = 128$. We choose $\mu_0 = 130$ and $\alpha = 0.05$.

From a standard normal distribution table we retrieve $z_{0.05} \approx -1.645$ and get

$$c_{0.05} \approx 130 - 1.645\frac{5.4}{\sqrt{25}} \approx 128.22.$$

Since $\bar{x} = 128 < 128.22 = c$, we reject the null hypothesis H_0.

If, however, we had chosen $\alpha = 0.01$, it would have been (with $z_{0.01} \approx -2.326$):

$$c_{0.01} \approx 130 - 2.326\frac{5.4}{\sqrt{25}} \approx 127.49$$

Since $\bar{x} = 128 > 127.49 = c$, we would have accepted the null hypothesis H_0.

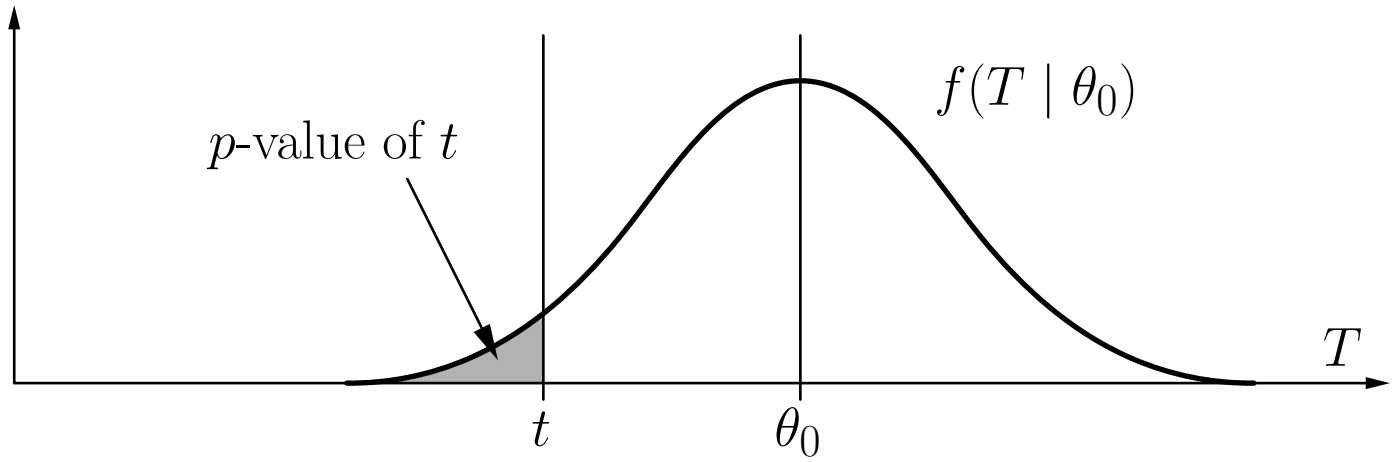
Instead of fixing a significance level α one may state the so-called **p-value**

$$p = \Phi\left(\frac{128 - 130}{5.4/\sqrt{25}}\right) \approx 0.032.$$

For $\alpha \geq p = 0.032$ the null hypothesis is rejected, for $\alpha < p = 0.032$ accepted.

Parameter Test: p-value

Let t be the value of the test statistic T
that has been computed from a given data set.
(Note: This example illustrates $H_0 : \theta \geq \theta_0$ and $H_a : \theta < \theta_0$.)



The **p-value** is the probability that a value of t or less
can be observed for the chosen test statistic T.

The p-value is a **lower limit for the significance level** α
that may have been chosen if we wanted to reject the null hypothesis H_0.

Attention: p-values are often misused or misinterpreted!

A low p-value does **not** mean that the result is very reliable!

All that matters for the test is whether the computed p-value
is **below the chosen significance level or not**.

(A low p-value could just be a chance event, an accident!)

The significance level may **not** be chosen **after** computing the p-value,
since we tend to choose lower significance levels if we know that they are met.

Doing so would undermine the reliability of the procedure!

Stating p-values is only a convenient way of avoiding a fixed significance level.
(Since significance levels are a matter of choice and thus user-dependent.)

However: A significance level must still be chosen
before a reported p-value is looked at.

Relevance of the Type-2 Error

Reminder: There are two possible types of errors:

Type 1: The null hypothesis H_0 is rejected, even though it is correct.

Type 2: The null hypothesis H_0 is accepted, even though it is false.

Type-1 errors are considered to be more severe,
since the null hypothesis gets "the benefit of the doubt".

However, **type-2 errors should not be neglected** completely:

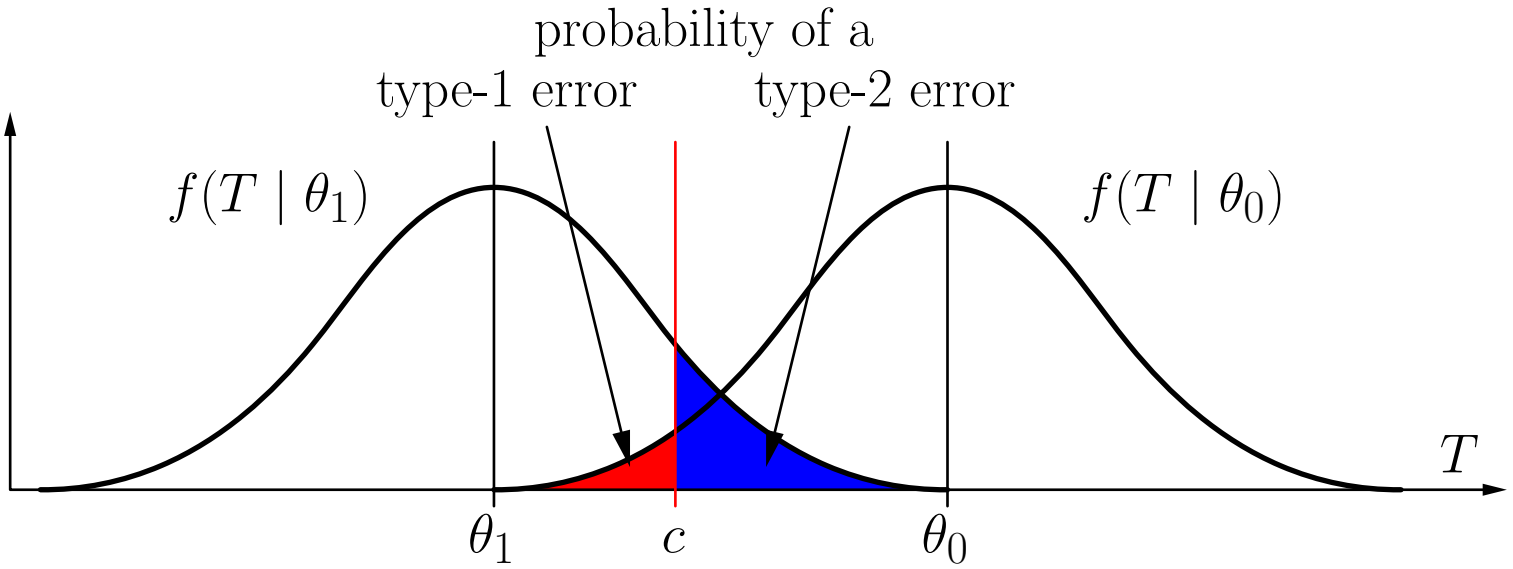
- It is always possible to achieve a vanishing probability of a type-1 error:
  Simply accept the null hypothesis in all instances, regardless of the data.
- Unfortunately such an approach maximizes the type-2 error.

Generally, **type-1 and type-2 errors are complementary quantities**:

The lower we require the type-1 error to be (the lower the significance level),
the higher will be the probability of a type-2 error.

Relationship between Type-1 and Type-2 Error

Suppose there are only two possible parameter values θ_0 and θ_1 with $\theta_1 < \theta_0$. (That is, we have $H_0 : \theta = \theta_0$ and $H_a : \theta = \theta_1$.)



Lowering the significance level α moves the critical value c to the left: lower type-1 error (red), but higher type-2 error (blue).

Increasing the significance level α moves the critical value c to the right: higher type-1 error (red), but lower type-2 error (blue).