Intelligent Data Analysis

Prof. Rudolf Kruse
Christian Braune

Computational Intelligence Group
Faculty of Computer Science
kruse@iws.cs.uni-magdeburg.de

About me

- 1979 diploma (Mathematics) degree from the University of Braunschweig, Germany
- 1980 PhD in Mathematics, 1984 the venia legendi in Mathematics from the same university
- 2 years at the Fraunhofer Gesellschaft
- 1986 joined the University of Braunschweig as a professor of computer science
- Since 1996 he is a full professor at the Department of Computer Science of the University of Magdeburg
- **Research:** statistics, artificial intelligence, expert systems, fuzzy control, fuzzy data analysis, computational intelligence, and information mining

Organisational

Lecture

- Thursday, 15.15-16.45, 22b-102
- Prof. Kruse
- Consultation: wednesday, 11 - 12 a.m., G29-008
- Preferred way of contact: `kruse@iws.cs.uni-magdeburg.de`

Exercises

- Monday 9.15-10.45 29-K058
- Tutor: Christian Braune
- G29-013, `christian.braune@st.ovgu.de`

Updated Information on the Course

- `http://fuzzy.cs.uni-magdeburg.de/wiki/pmwiki.php?n=Lehre.IDA2012`

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)

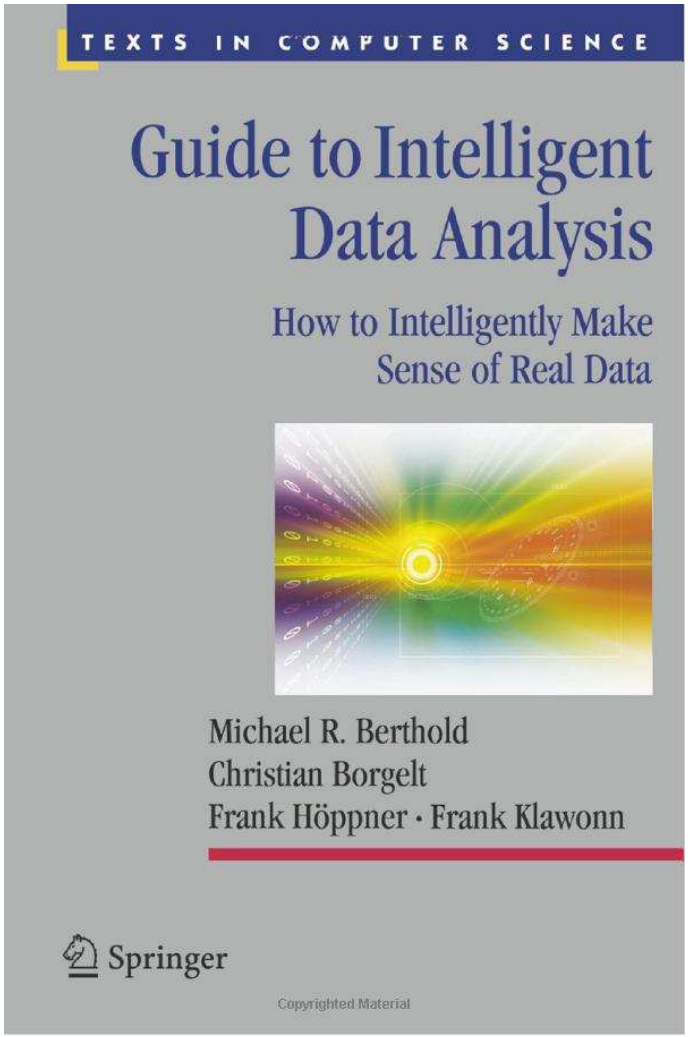
- ticked at least two thirds of the assignments,
- presented at least two times a solution during the exercise, and
- passed a small colloquium (approx. 10 min) or a written test (if there are more than 20 students) after the course.

Conditions for Exams

- no ticking required

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

Knowledge

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

- Science: correctness, generality, simplicity
- Economy: 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

1. Each planet moves around the sun in an ellipse, with the sun at one focus.
2. The radius vector from the sun to the planet sweeps out equal areas in equal intervals of time.
3. 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?*

Data Mining Tasks 5

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*

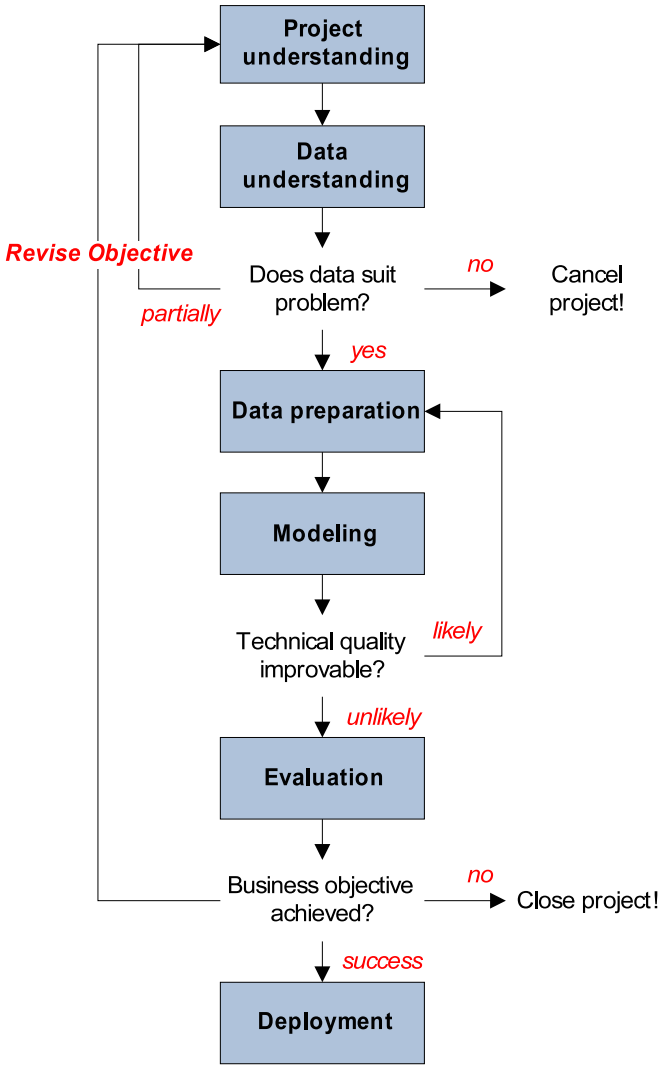
Cross Industry Standard Process for Data Mining

- Data Mining Process Model developed within an EU project
- Several phases that are repeated until data mining project is finished

CRISP-Phases

1. Project understanding
2. Data understanding
3. Data preparation
4. Modeling
5. Evaluation
6. Deployment

CRISP-DM Model



1. Project understanding
   - What exactly is the problem, what the expected benefit?
   - What should a solution look like?
   - What is known about the domain?
2. Data understanding
   - What (relevant) data is available?
   - What about data quality/quantity/recency?
3. Data preparation
   - Can data quality be increased?
   - How can it be transformed for modeling?
4. 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?
5. Evaluation
   - How good is the model in terms of project requirements?
   - What have we learned from the project?
6. Deployment
   - How can the model be best deployed?
   - Is there a way to know if the model is still valid?

Statistics

Statistics

- **Descriptive Statistics**
  - Tabular and Graphical Representations
  - Characteristic Measures
  - Principal Component Analysis
  - Data Visualization
  - Outliers
  - Missing Values
- **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

Basic Notions

- **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

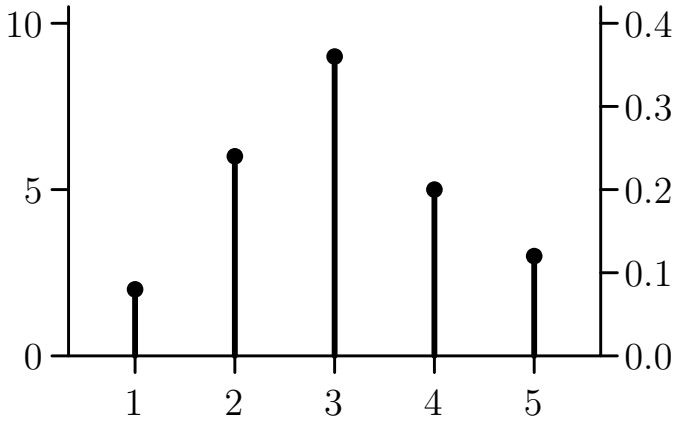
- Frequency tables for two or more attributes are called **contingency tables**.
- They contain the absolute or relative frequency of **value combinations**.

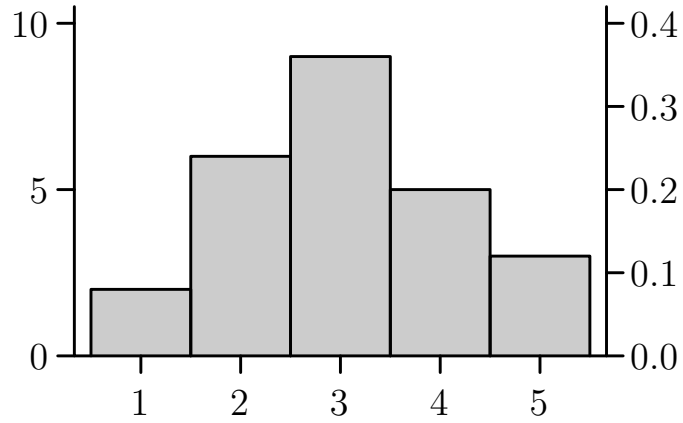
	a_1	a_2	a_3	a_4	$\sum$
b_1	8	3	5	2	18
b_2	2	6	1	3	12
b_3	4	1	2	7	14
$\sum$	14	10	8	12	44

- 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).

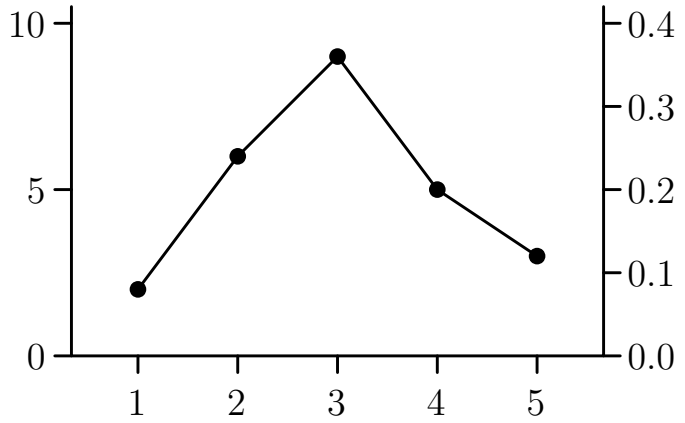




- Bar charts are the most frequently used and most comprehensible way of displaying absolute frequencies.
- A wrong impression can result if the vertical scale does not start at 0 (for frequencies or other absolute numbers).

Frequency Polygon and Line Chart

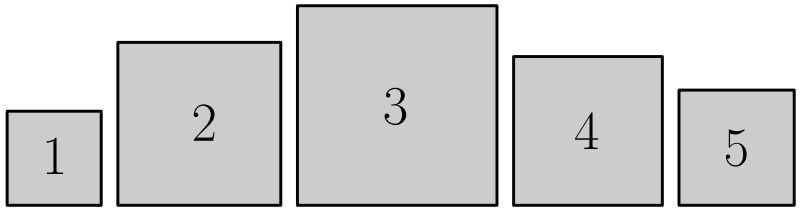
- Frequency polygon: the ends of the poles of a pole chart are connected by lines. (Numbers are still represented by lengths.)



- 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

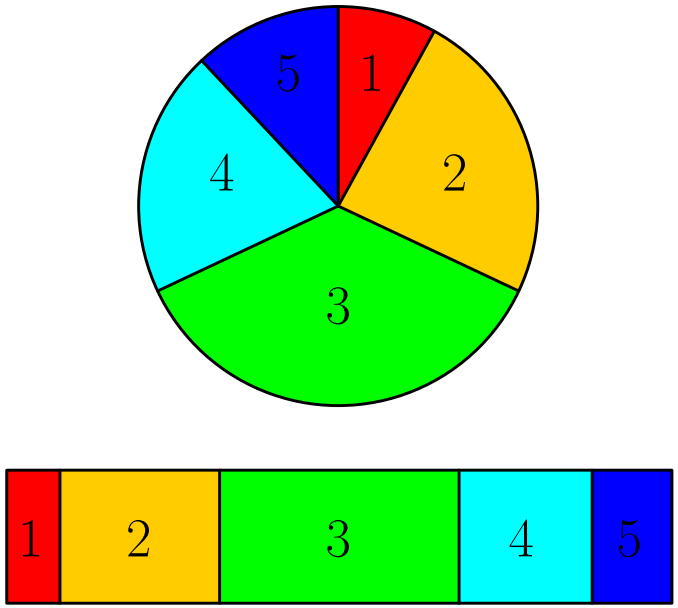
- Numbers may also be represented by other geometric quantities than lengths, like areas or volumes.
- Area and volume charts are usually less comprehensible than bar charts, because humans have more difficulties to compare areas and especially volumes than lengths. (exception: the represented numbers are areas or volumes)

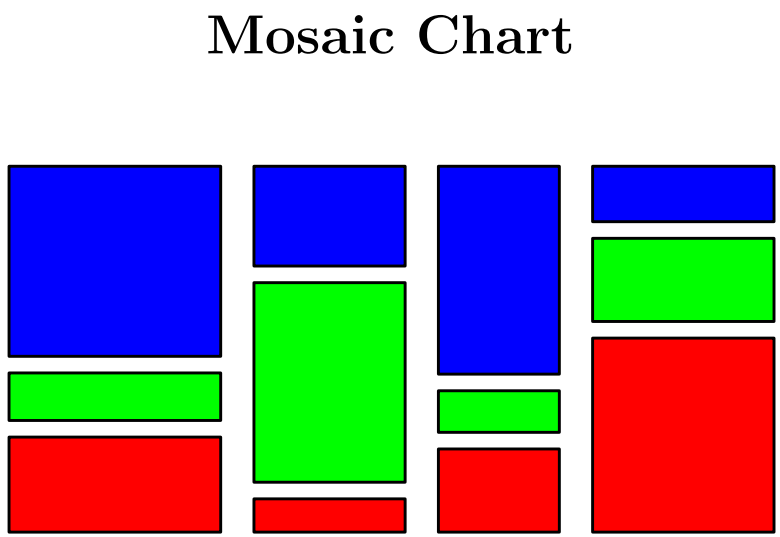


- Sometimes the height of a two- or three-dimensional object is used to represent a number. The diagram then conveys a misleading impression.

Pie and Stripe Charts

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



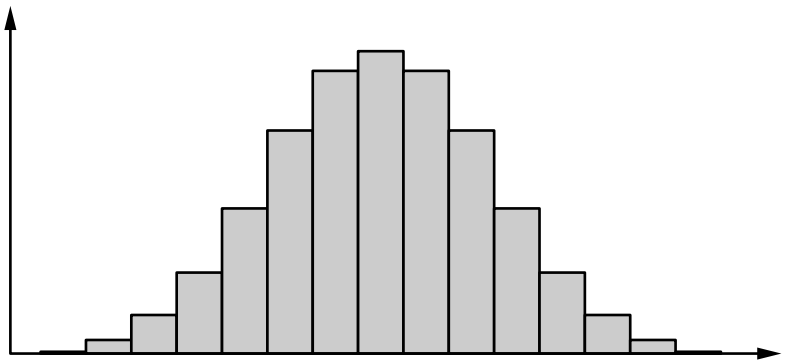


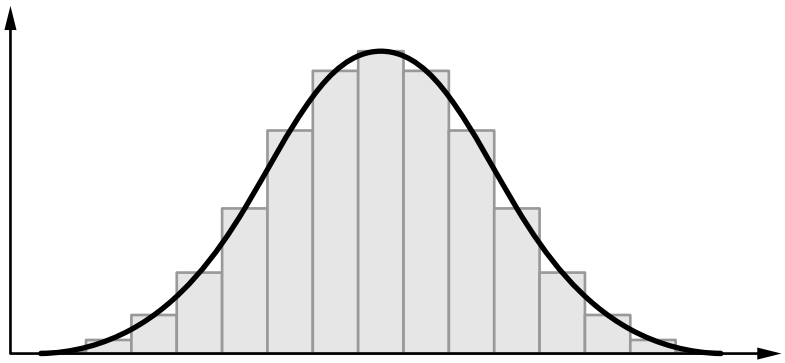
- Mosaic charts can be used to display contingency tables.
- More than two attributes are possible, but then separation distances and color must support the visualization to keep it comprehensible.

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.

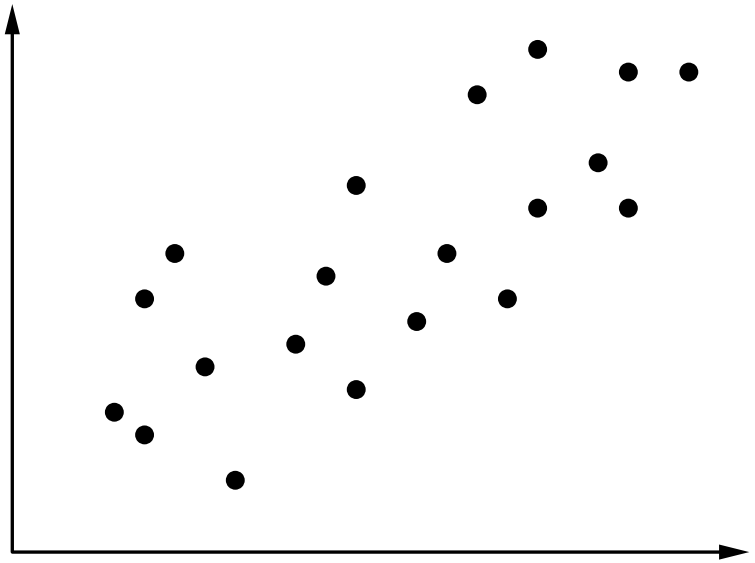
  Most common approach: form equally sized intervals (so-called **bins**)
  and count the frequency of sample values inside each interval.
- **Attention:** Depending on the size and the position of the bins
  the histogram may look considerably different.
- In sketches often only a rough outline of a histogram is drawn:





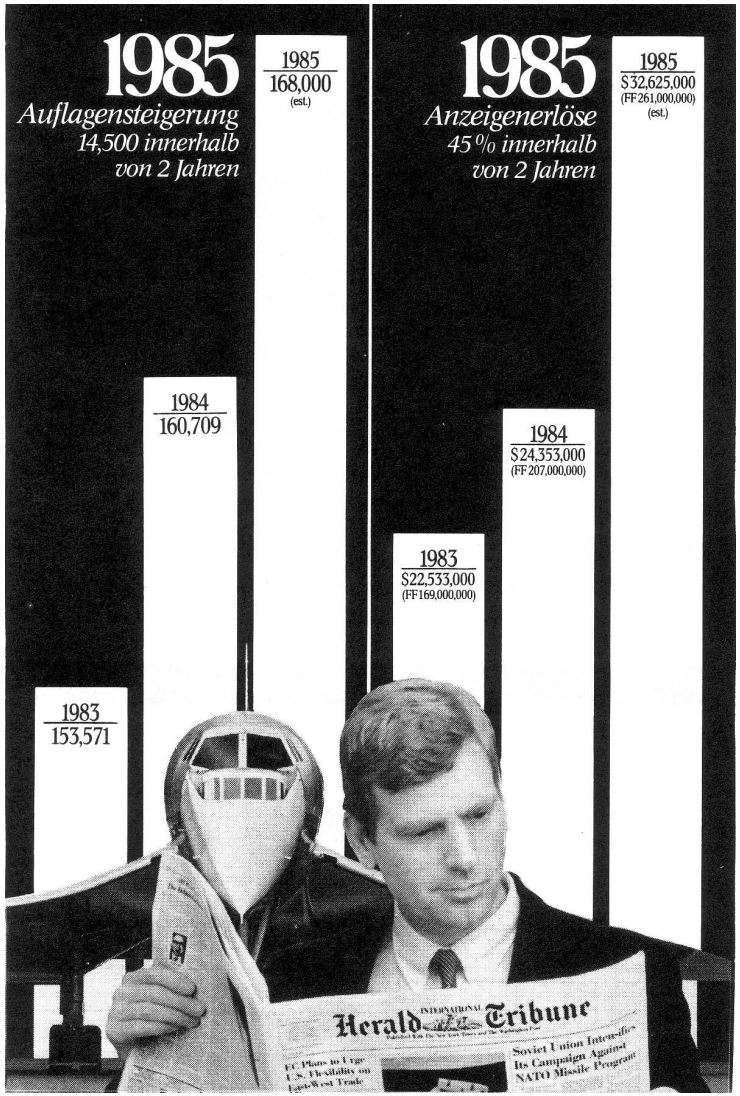
Scatter Plots

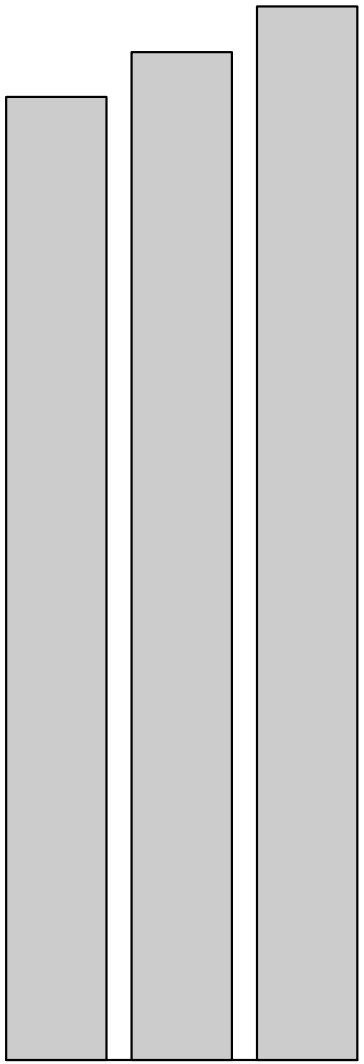
- Scatter plots are used to display two-dimensional metric data sets.
- Sample values are the coordinates of a point.
  (Numbers are represented by lengths.)



- Scatter plots provide a simple means for checking for dependency.

How to Lie with Statistics





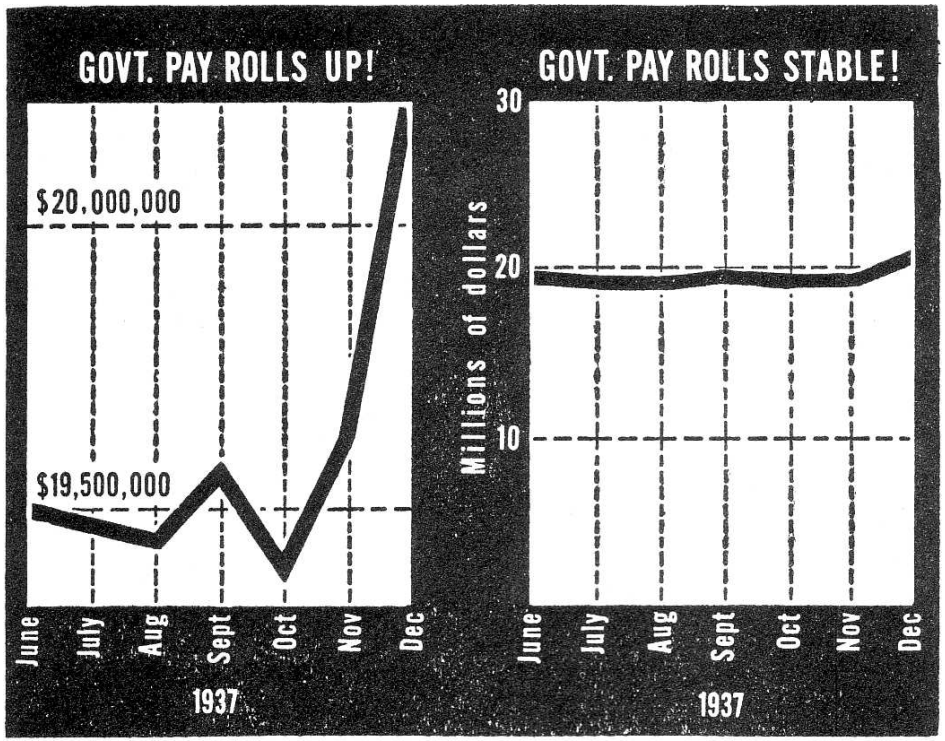
Often the vertical axis of a pole or bar chart does not start at zero, but at some higher value.

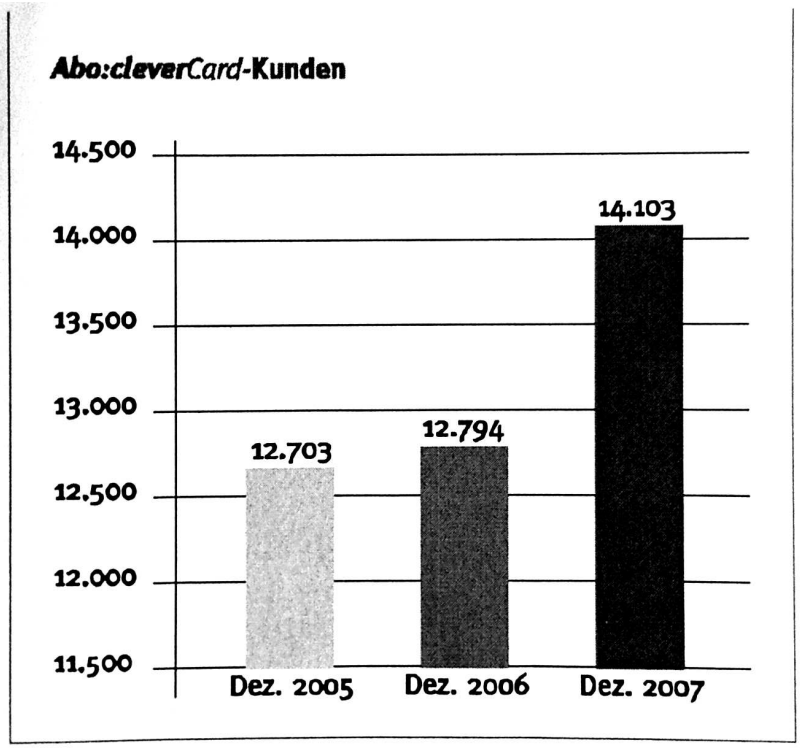
In such a case the conveyed impression of the ratio of the depicted values is completely wrong.

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.

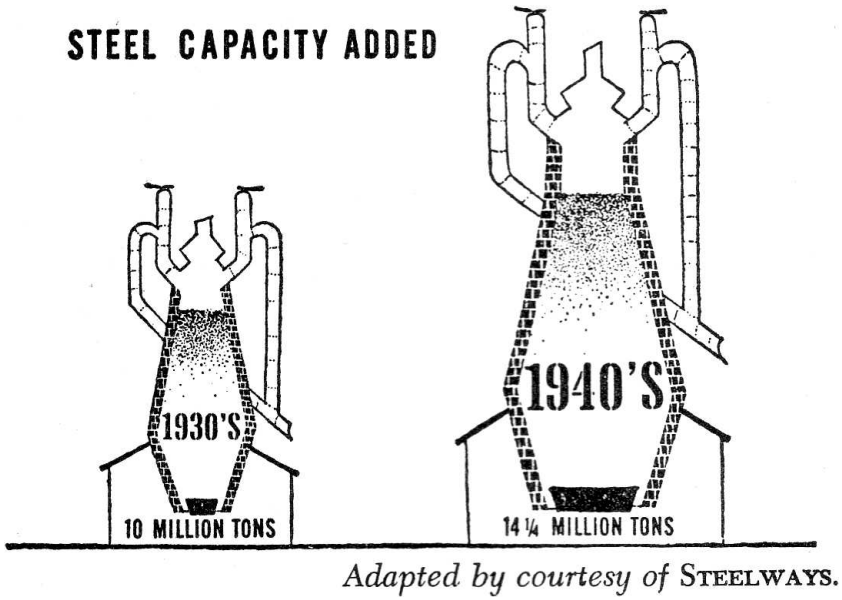
How to Lie with Statistics





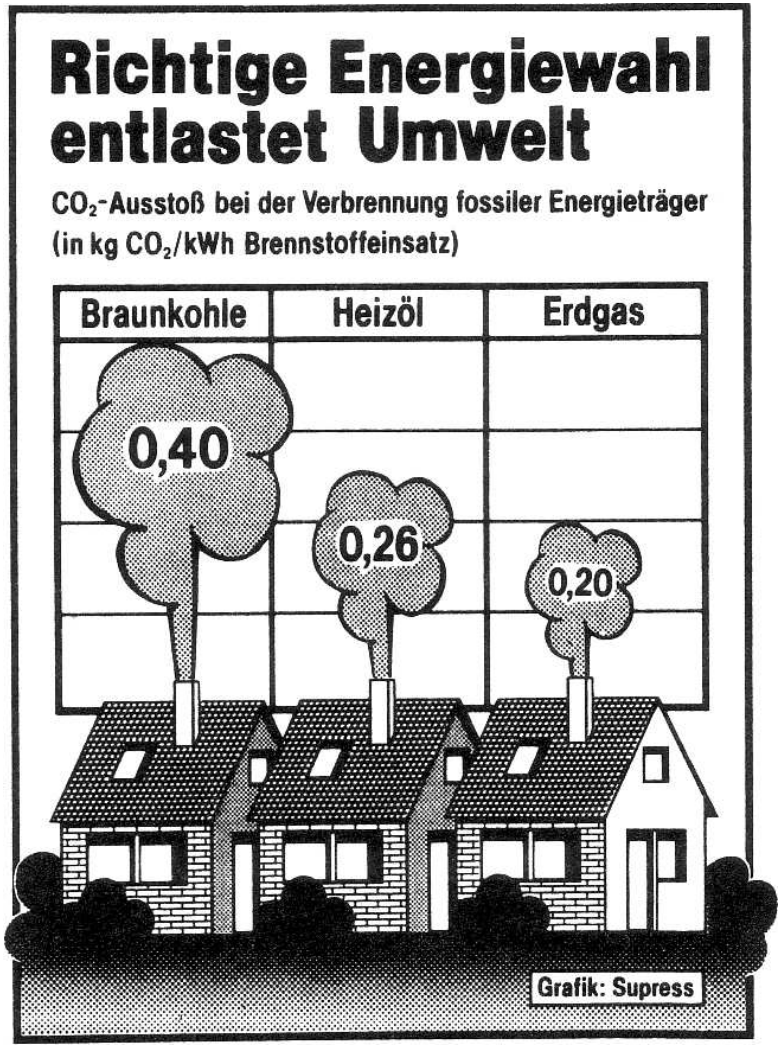
- Depending on the position of the zero line of a pole, bar, or line chart completely different impressions can be conveyed.

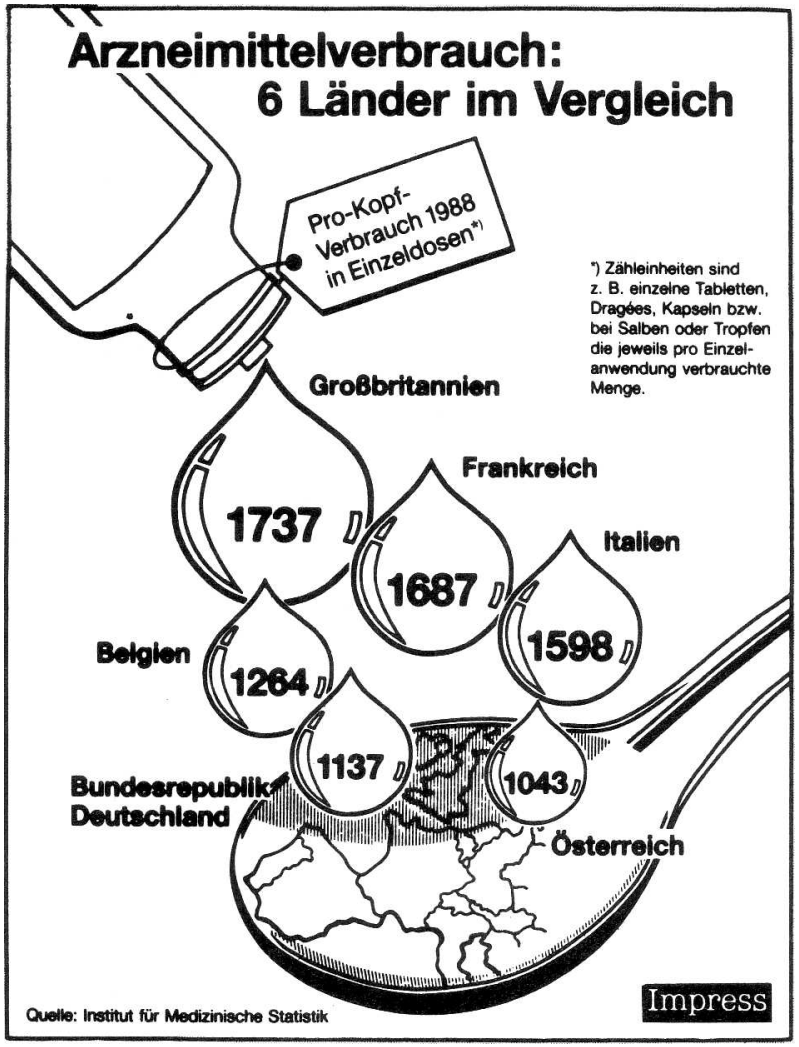
How to Lie with Statistics



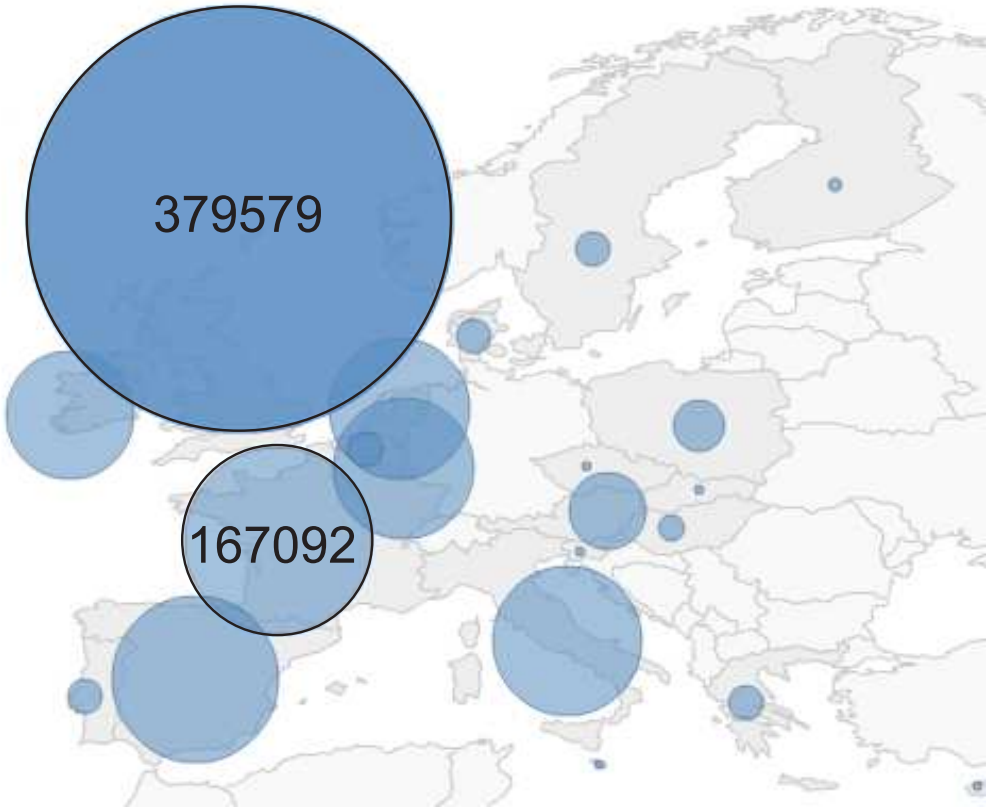
- Poles and bars are frequently replaced by (sketches of) objects in order to make the diagram more aesthetically appealing.
- However, objects are perceived as 2- or even 3-dimensional and thus convey a completely different impression of the numerical ratios.

How to Lie with Statistics

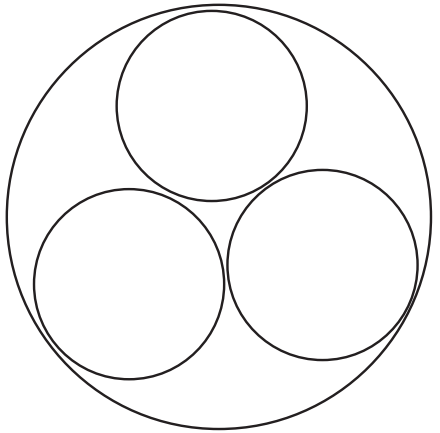




Foreign outstanding debits of German banks in million Euros as of 2010:



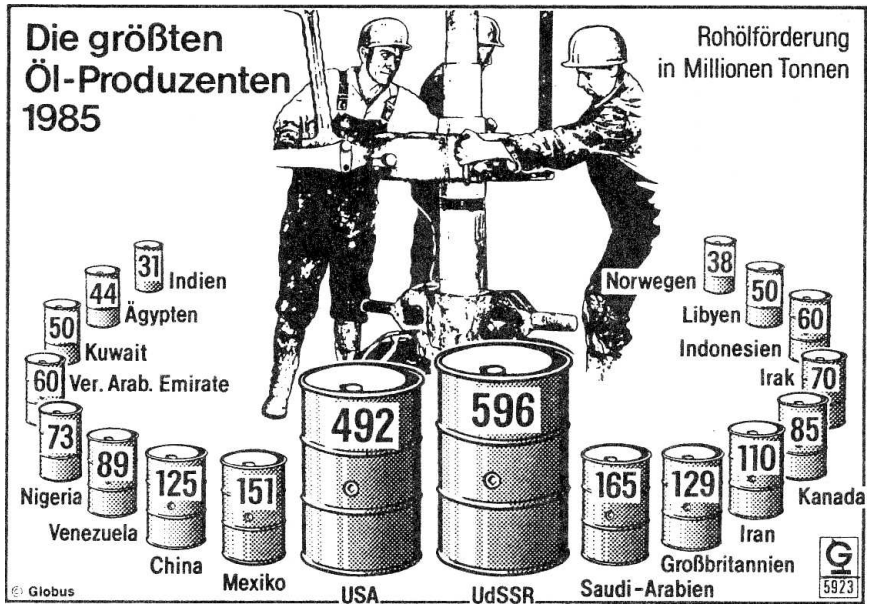
Source: Spiegel Online

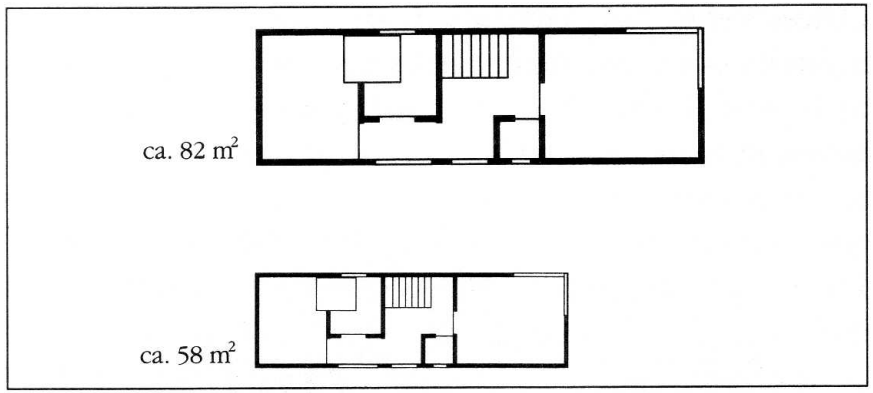


$$\frac{379579}{167092} \approx 2.2$$

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

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).

Localization Measures: 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.

Localization Measures: Arithmetic Mean

- **Arithmetic Mean** $\bar{x}$
  The arithmetic mean minimizes the sum of squared differences:

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

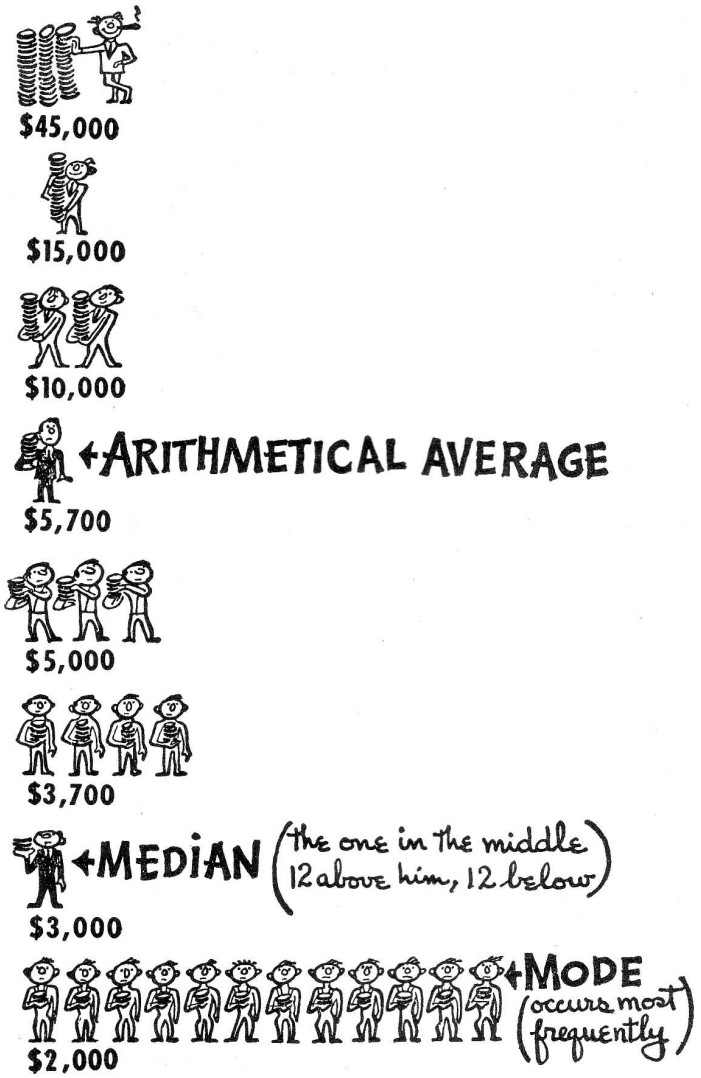
  The arithmetic mean is defined as

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

  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.

How to Lie with Statistics



»Sollen wir das arithmetische Mittel als durchschnittliche Körpergröße nehmen und den Gegner erschrecken, oder wollen wir ihn einlullen und nehmen den Median?«

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* ($p = \frac{1}{4}$)

Dispersion Measures: Average Absolute Deviation

- **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

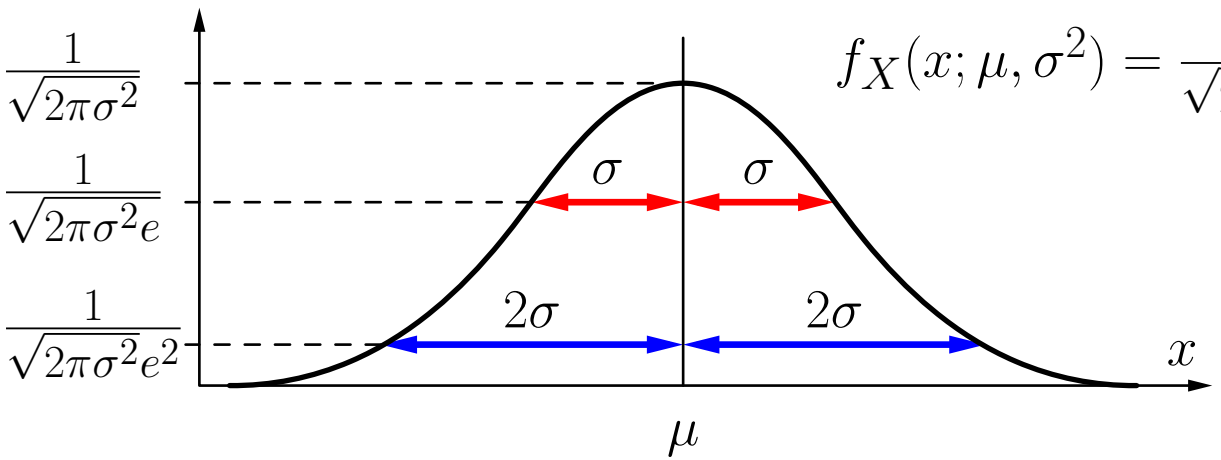
$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

- **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.)

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:

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

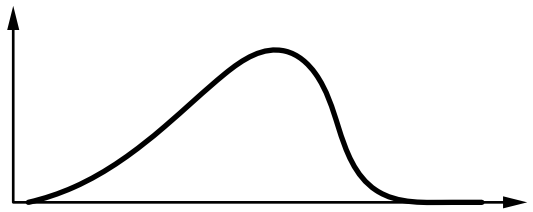
- Advantage: The sums $\sum_{i=1}^{n} x_i$ and $\sum_{i=1}^{n} x_i^2$ can both be computed in the same traversal of the data and from them both mean and variance are computable.

Shape Measures: Skewness

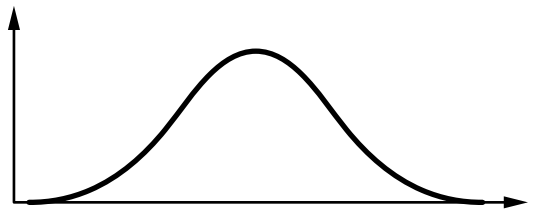
- The **skewness** α_3 (or **skew** for short) measures whether, and if, how much, a distribution differs from a symmetric distribution.
- It is computed from the 3rd moment about the mean, which explains the index 3.

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

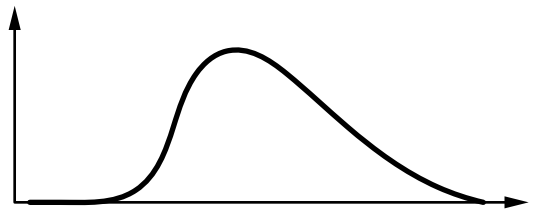
$$\text{where} \quad z_i = \frac{x_i - \bar{x}}{v} \quad \text{and} \quad v^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2.$$



$\alpha_3 < 0$: right steep



$\alpha_3 = 0$: symmetric



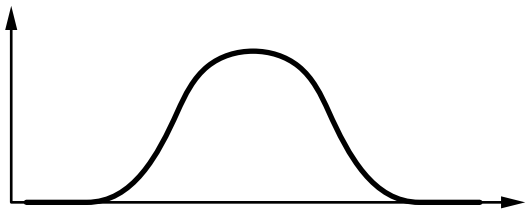
$\alpha_3 > 0$: left steep

Shape Measures: Kurtosis

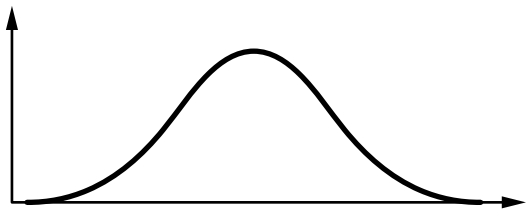
- The **kurtosis** or **excess** α_4 measures how much a distribution is arched, usually compared to a Gaussian distribution.
- It is computed from the 4th moment about the mean, which explains the index 4.

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

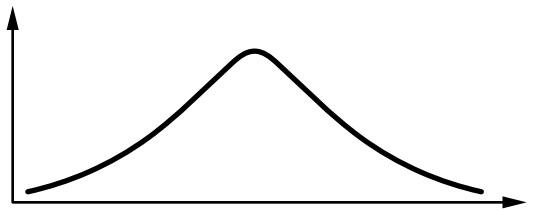
$$\text{where} \quad z_i = \frac{x_i - \bar{x}}{v} \quad \text{and} \quad v^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2.$$



$\alpha_4 < 3$: 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

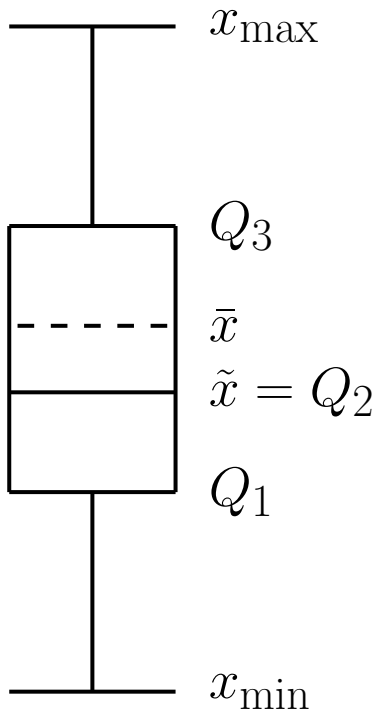
$$s^2 = \frac{1}{n-1} \left(m'_2 - \frac{1}{n} m'^2_1 \right) \qquad \text{and also} \qquad v^2 = \frac{1}{n} m'_2 - \frac{1}{n^2} m'^2_1.$$

- 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



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**		**Outer Product** **Matrix Product**	
$\vec{v}^\top \vec{v}$	$\begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_m \end{pmatrix}$	$\vec{v}\vec{v}^\top$	$\begin{pmatrix} v_1, & v_2, & \ldots, & v_m \end{pmatrix}$
$\begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_m \end{pmatrix}$	$(v_1, v_2, \ldots, v_m)$	$\begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_m \end{pmatrix}$	$\begin{pmatrix} v_1^2 & v_1 v_2 & \cdots & v_1 v_m \\ v_1 v_2 & v_2^2 & \cdots & v_2 v_m \\ \vdots & & \ddots & \vdots \\ v_1 v_m & v_2 v_m & \cdots & v_m^2 \end{pmatrix}$
$\sum_{i=1}^{m} v_i^2$			

- 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 $\vec{v}\vec{v}^\top$):

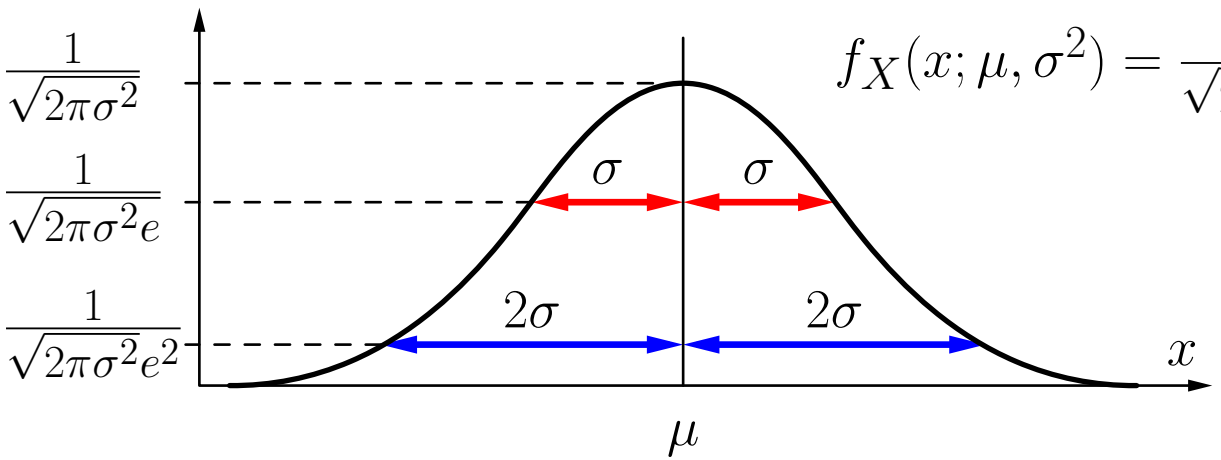
$$\mathbf{S} = \frac{1}{n-1} \sum_{i=1}^{n} \left(\begin{pmatrix} x_i \\ y_i \end{pmatrix} - \begin{pmatrix} \bar{x} \\ \bar{y} \end{pmatrix} \right) \left(\begin{pmatrix} x_i \\ y_i \end{pmatrix} - \begin{pmatrix} \bar{x} \\ \bar{y} \end{pmatrix} \right)^\top = \begin{pmatrix} s_x^2 & s_{xy} \\ s_{xy}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

  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.

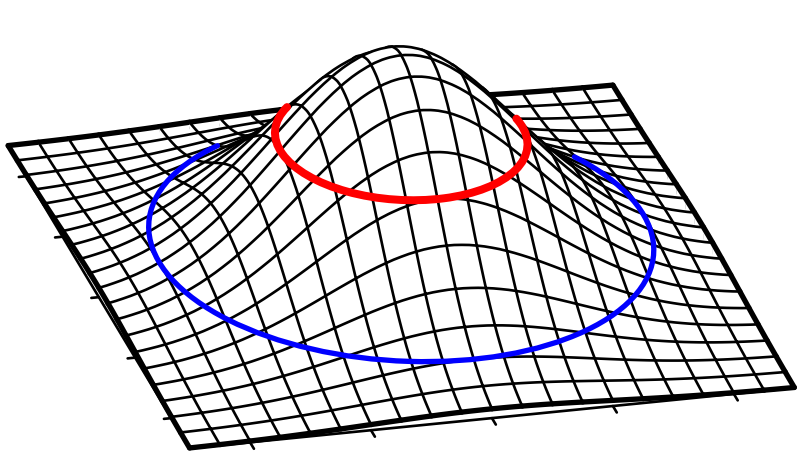
- A **multivariate normal distribution** has the density function

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

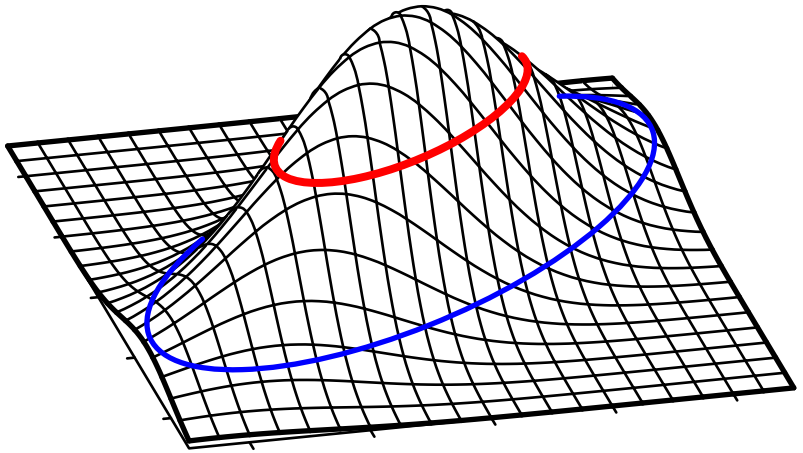
m:	size of the vector $\vec{x}$	(it is m-dimensional),
$\vec{\mu}$:	mean value vector,	estimated by (empirical) mean value vector $\bar{\vec{x}}$,
$\mathbf{\Sigma}$:	covariance matrix,	estimated by (empirical) covariance matrix $\mathbf{S}$,
$\lvert\mathbf{\Sigma}\rvert$:	determinant of the covariance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



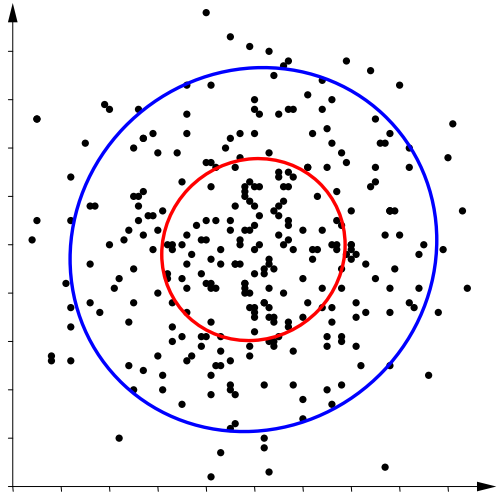
standard



general

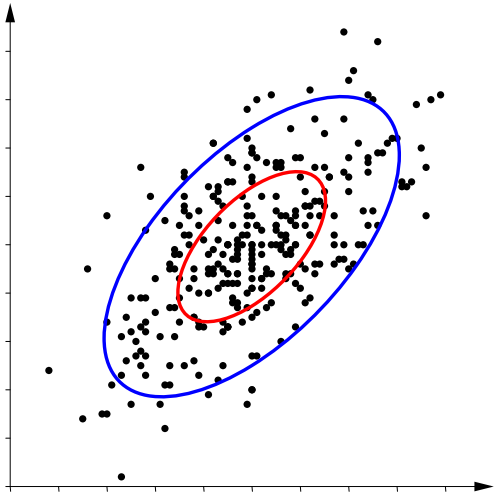
- **Question:** Is there a multivariate analog of standard deviation?

Covariance Matrices of Example Data Sets



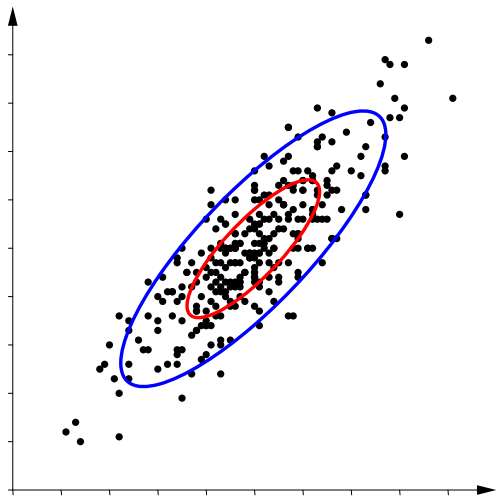
$$\boldsymbol{\Sigma} = \begin{pmatrix} 3.59 & 0.19 \\ 0.19 & 3.54 \end{pmatrix}$$

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



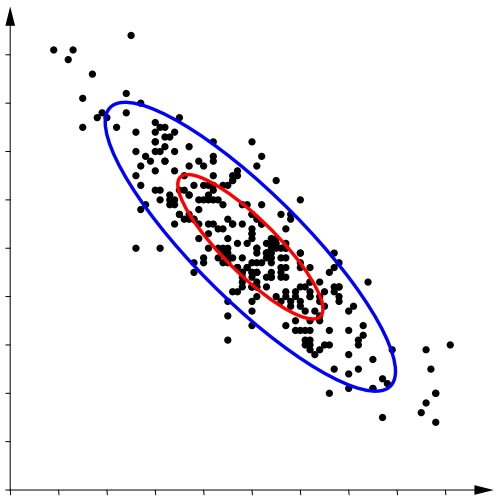
$$\boldsymbol{\Sigma} = \begin{pmatrix} 2.33 & 1.44 \\ 1.44 & 2.41 \end{pmatrix}$$

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



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

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



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

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

Correlation and Principal Component Analysis

Correlation Coefficient

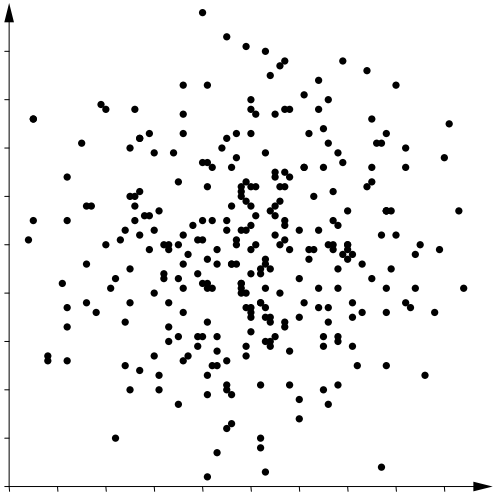
- The covariance is a measure of the strength of **linear dependence** of the two quantities.
- However, its value depends on the variances of the individual dimensions.
  $\Rightarrow$ Normalize to unit variance in the individual dimensions.
- **Correlation Coefficient**
  (more precisely: Pearson's Product Moment Correlation Coefficient)

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

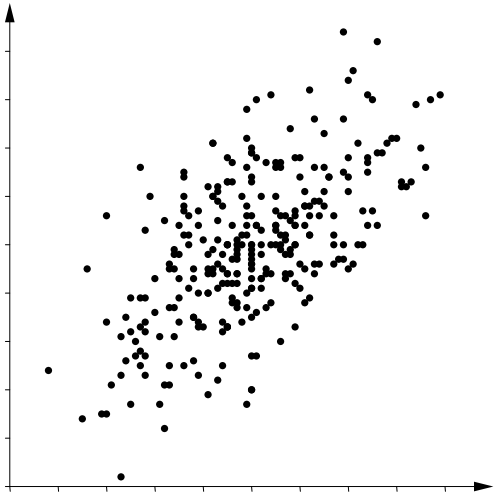
- r measures the strength of linear dependence:
  $r = -1$: the data points lie perfectly on a descending straight line.
  $r = +1$: the data points lie perfectly on an ascending straight line.
- $r = 0$: there is no **linear** dependence between the two attributes (but there may be a non-linear dependence!).

Correlation Coefficients of Example Data Sets

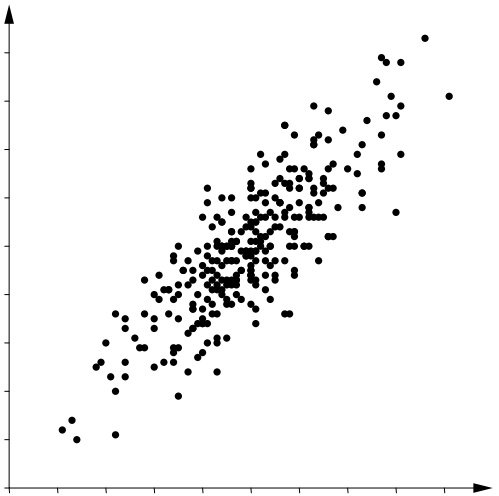
no correlation ($r \approx 0.05$)



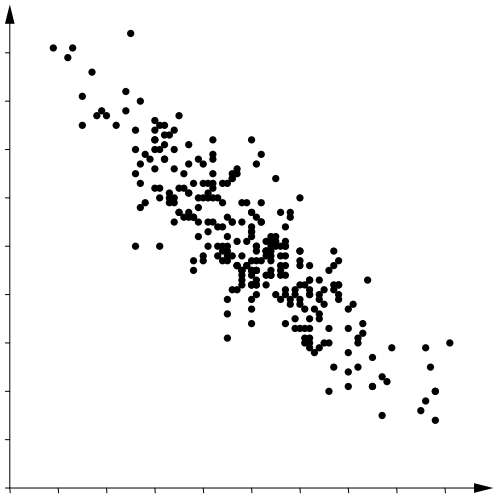
weak positive correlation ($r \approx 0.61$)



strong positive correlation ($r \approx 0.83$)



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



Correlation Matrix

- **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:

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

  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.

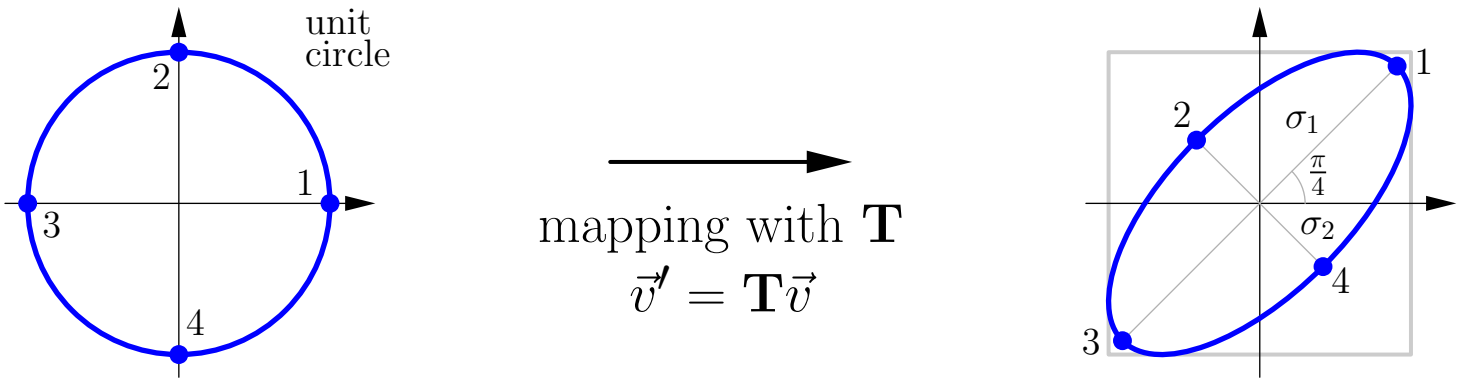
Special Case: Two Dimensions

- Correlation matrix

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

- Eigenvalue decomposition

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



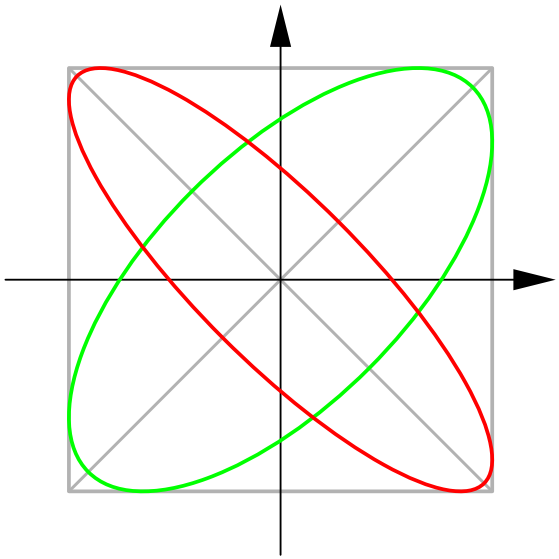
Correlation Matrix: Interpretation

- 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).

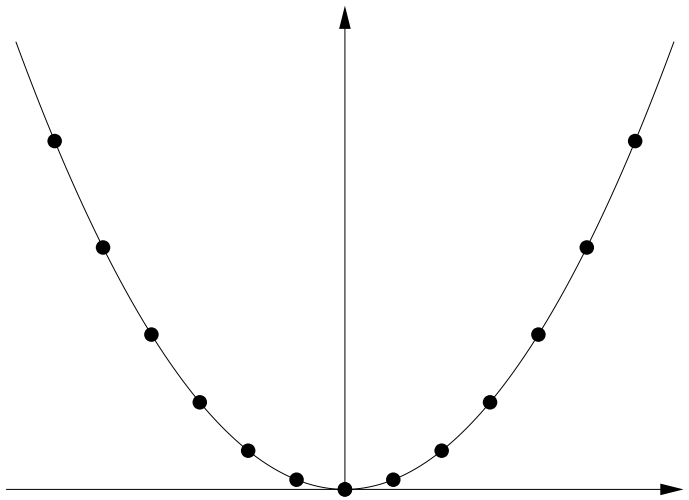
  The reason is that the normalization transforms the data points in such a way, that the ellipse, the unit circle is mapped to by the "square root" of the covariance matrix of the normalized data, is always inscribed into the square $[-1, 1] \times [-1, 1]$. Hence the ellipse's major axes are the square's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

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

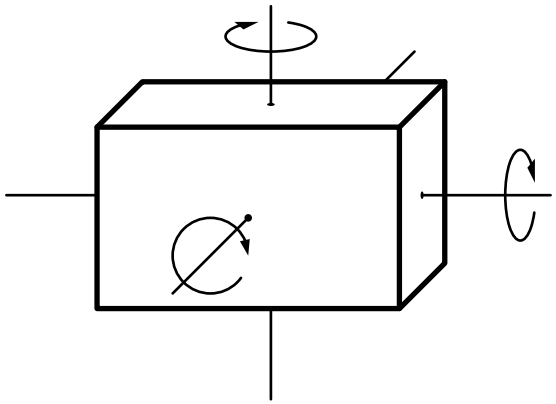
- The diagonal elements of this tensor are called the **moments of inertia**.
  They describe the "resistance" of the body against being rotated.

- 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

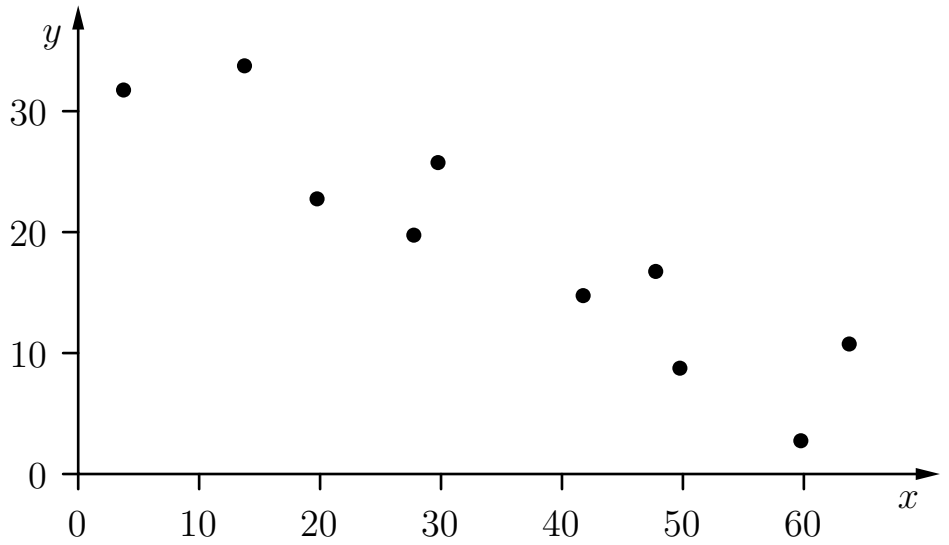
$$\sum_{i=1}^{k} \frac{\lambda_i}{m} \geq \alpha,$$

    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 ⇒ 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:

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

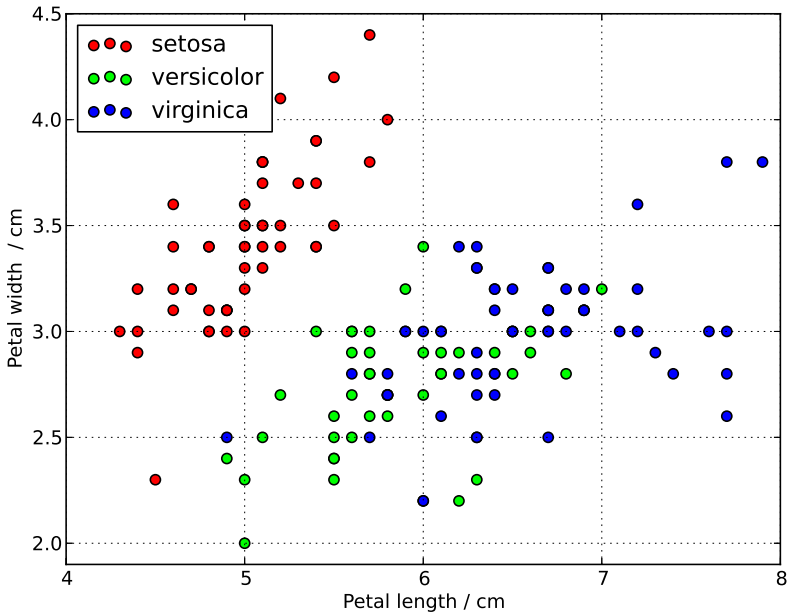
Resulting data set:

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 ($s^2_{x''} = 2\lambda_1 = \frac{96}{25}$).

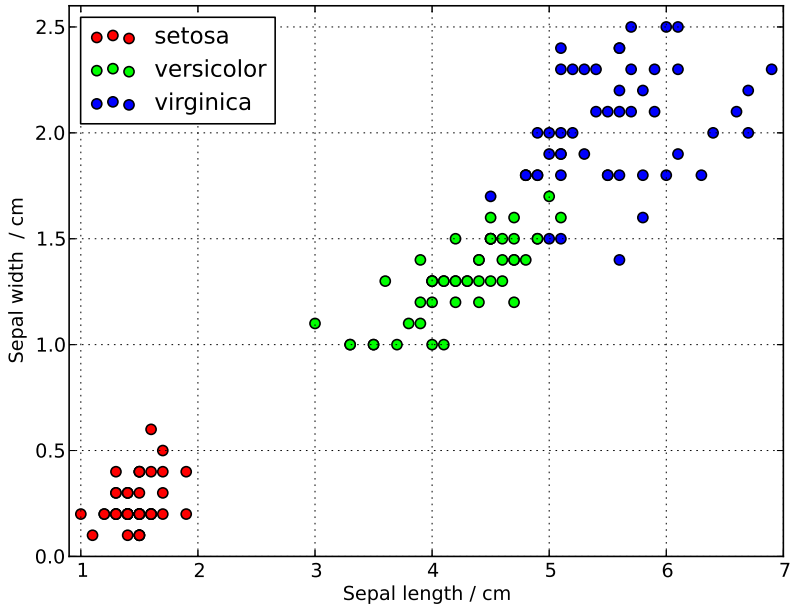
Data Visualization

Scatter Plots 1



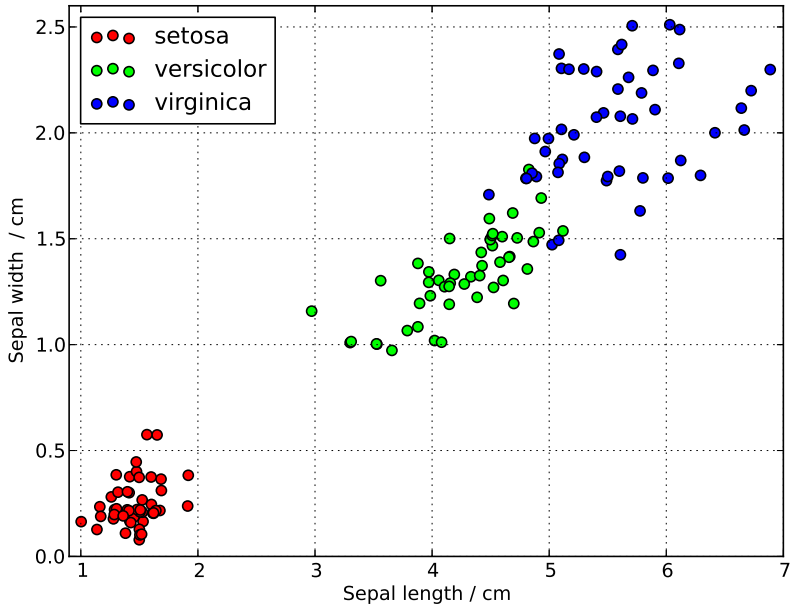
- Iris data set (150 samples, 3 classes, 4 numerical attributes)
- Only *sepal length* and *sepal width* used to plot data
- Different colors have been used for the different classes

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

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

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

- The normalisation factor $\frac{1}{\sum_{i=1}^{n} \sum_{j=i+1}^{n} (d_{ij})^2}$ 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

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

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

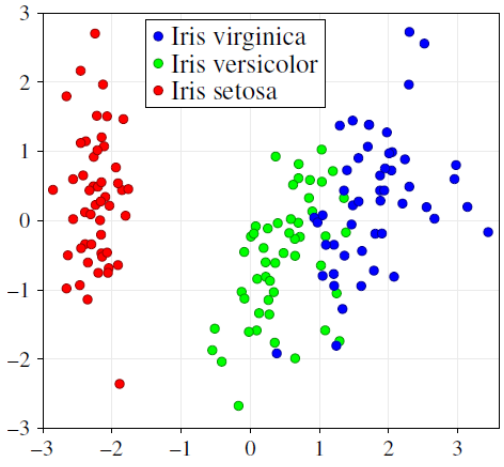
(mixture between relative and absolute squared error)

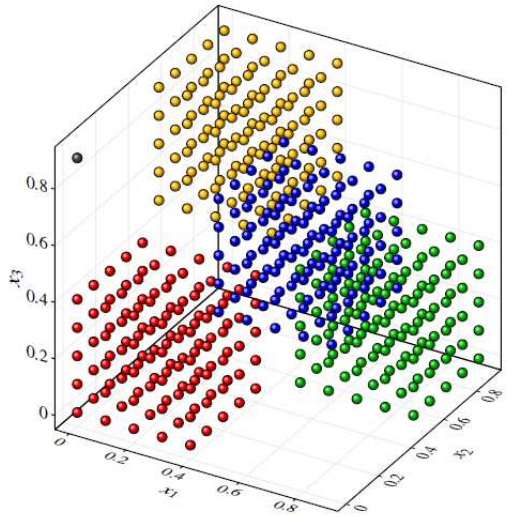
MDS based on E_3 is called **Sammon mapping**. The value of E_3 is called *stress*.

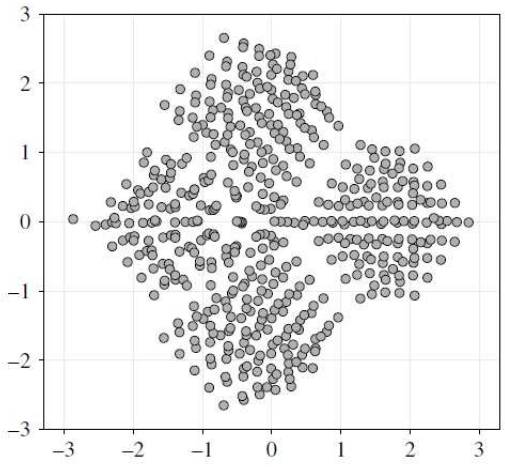
Multidimensional Scaling

- 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







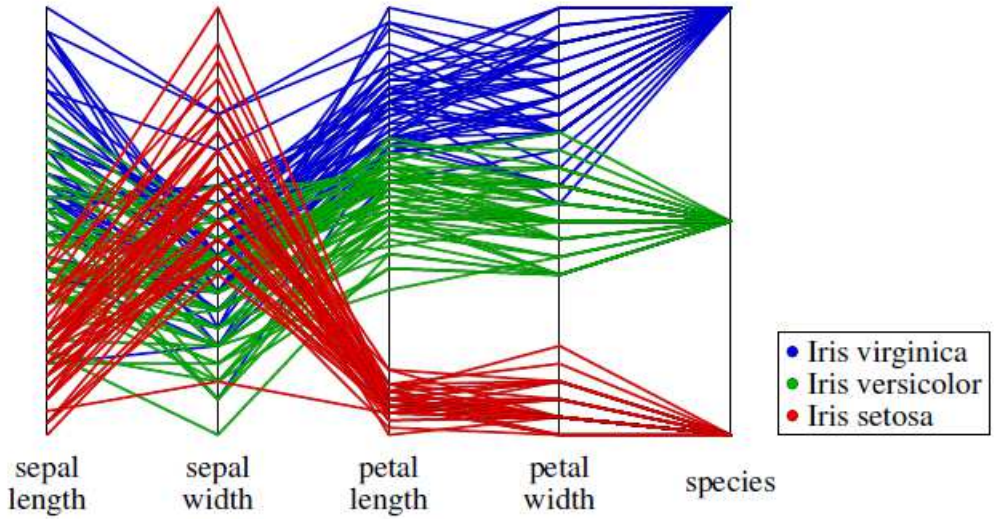
MDS (Sammon mapping) for the Iris data set, the Cube data, and MDS for the Cube data set.

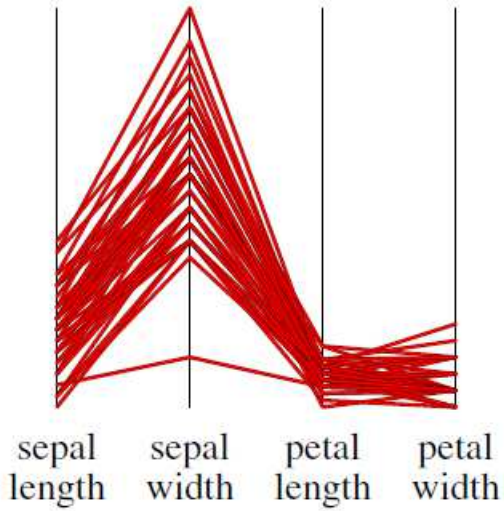
Parallel coordinates

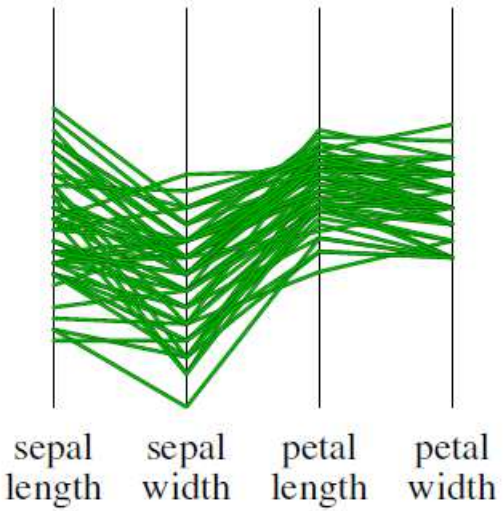
Parallel coordinates draw the coordinate axes parallel to each other, so that there is no limitation for the number of attributes to be shown simultaneously.

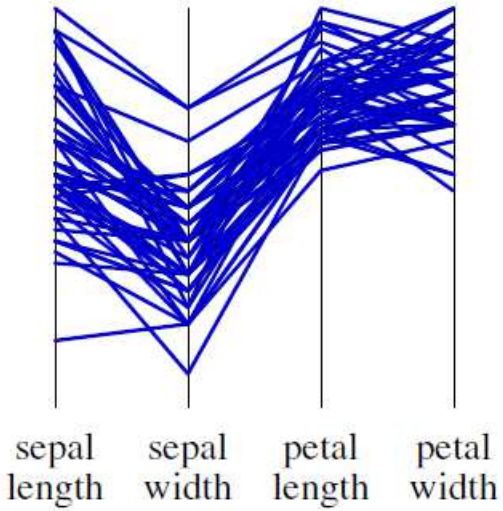
For a data object a polyline is drawn connecting the values of the data attribute on the corresponding axes.

Parallel coordinates









Outlier Detection

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* (later)

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

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 neither depend on the true value of X.

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

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 $0°$.

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.

1. 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*.
2. 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*.
3. 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.)