Repetition of statistics

Julius Fenn[1]

[1]University of Freiburg
PhD student at the Institute of Experimental Psychology Freiburg (Cognition, Action, and Sustainability)

14.12.2021

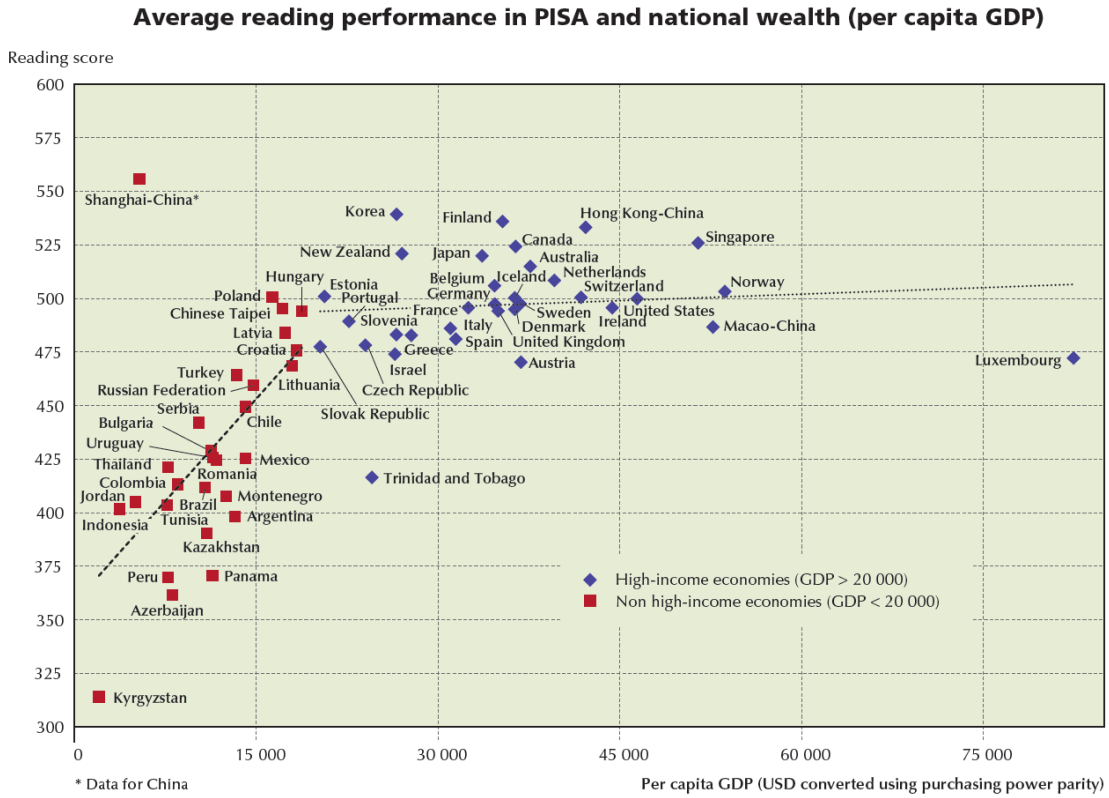
Any questions?

Motivation statistics

- Central areas of life science, among others, deal with numbers.
- Empirical phenomena are usually not deterministic, but stochastic
- an intuitive approach is not sufficient for the description of a variety of empirical phenomena, so that a mathematically sound treatment is necessary
- Statistics is a key qualification that is increasingly demanded by the labour market
- ...

Motivation statistics: Investment in the education system

data $\rightarrow$ statistic $\rightarrow$ political decisions



PISA in Focus 13
O'Connell (2019)

Motivation statistics: unstoppable (?) disruptive technologies

Table 1. Five successive technological revolutions, 1770s to 2000s

Technological revolution	*Popular name for the period*	*Core country or countries*	*Big-bang initiating the revolution*	*Year*
First	The "Industrial Revolution"	Britain	Arkwright's mill opens in Cromford	1771
Second	Age of steam and railways	Britain (spreading to Continent and USA)	Test of the "Rocket" steam engine for the Liverpool-Manchester railway	1829
Third	Age of steel, electricity, and heavy engineering	USA and Germany forging ahead and overtaking Britain	The Carnegie Bessemer steel plant opens in Pittsburgh, Pennsylvania	1875
Fourth	Age of oil, the automobile, and mass production	USA (with Germany at first vying for world leadership), later spreading to Europe	First Model-T comes out of the Ford plant in Detroit, Michigan	1908
Fifth	Age of information and telecommunications	USA (spreading to Europe and Asia)	The Intel microprocessor is announced in Santa Clara, California	1971

`Kaldor (2018)`
10 most expensive companies worldwide

Course structure

1. fundamentals: basic terms, data collection, excursus on measurement, scale levels
2. descriptive, explorative statistics
   - univariate: graphs, measures, symmetry of distributions
   - multivariate: graphs, correlation; excursus term theory
3. probabilities, randomness
4. distributions: discrete, continuous
5. (inferential statistical like hypothesis testing, linear regression, ...)

central learning platform: ILIAS

You can find all literature recommendations, slides, ... on ILIAS „**WS 21 - Fundamentals of Questionnaire Design and Analysis using lab.js and R**“

central terms: Describe, Search, Conclude

Statistics represents the theoretical basis of all empirical research and is a subfield of mathematics[1]

- **descriptive statistics:** descriptive and graphical processing and compression of data, as well as for the detection of errors in the data set (data validation).
- **explorative statistics:** search for structures and peculiarities in data in order to arrive at new questions and the choice of suitable statistical models
- **inductive statistics:** using probability theory / stochastics to reach more general conclusions beyond the surveyed sample

[1] Separation according to the method controversy into two camps using purely quantitative / qualitative methods is (from my point of view) extremely harmful, as the methods complement and overlap each other, see e.g. `Wolf (2008)`, `Kelle (2008)`

central terms

- **statistical units:** Objects over which quantities (characteristics) of interest are observed
- **characteristics:** Observed quantities of the statistical units, e.g.
  - characteristic X_{gender} with the (characteristic) expressions x_{female}, x_{male}, x_{diverse} (also observation).
- **population:** Set of all statistical units about which one wants to arrive at statements
- **sample:** subset of the population under investigation

  $\rightarrow$ Sample should be as representative as possible (see explanations from slide 12 onwards)

central terms

- **variable:** characteristics which can take at least two different values
- **constant:** characteristic with only one value

	ID	IDZP	GR	GES	OG	VG	CC	KU	IQ	KG	IQ.cat
1	1	1	1	2	1913.88	1005	6.08	54.7	96	57.607	unterdurchschnittlich
2	2	1	2	2	1684.89	963	5.73	54.2	89	58.968	unterdurchschnittlich
3	3	2	1	2	1902.36	1035	6.22	53	87	64.184	unterdurchschnittlich
4	4	2	2	2	1860.24	1027	5.8	52.9	87	58.514	unterdurchschnittlich
5	5	3	1	2	2264.25	1281	7.99	57.8	101	63.958	mittel
6	6	3	2	2	2216.4	1272	8.42	56.9	103	61.690	mittel
7	7	4	1	2	1866.99	1051	7.44	56.6	103	133.358	mittel
8	8	4	2	2	1850.64	1079	6.84	55.3	96	107.503	unterdurchschnittlich
9	9	5	1	2	1743.04	1034	6.48	53.1	127	62.143	überdurchschnittlich
10	10	5	2	2	1709.3	1070	6.43	54.8	126	83.009	überdurchschnittlich
11	11	6	2	1	1689.6	1173	7.99	57.2	101	61.236	mittel
12	12	6	1	1	1806.31	1079	8.76	57.2	96	61.236	unterdurchschnittlich
13	13	7	2	1	2136.37	1067	6.32	57.2	93	83.916	unterdurchschnittlich
14	14	7	1	1	2018.92	1104	6.32	57.2	88	79.380	unterdurchschnittlich
15	15	8	2	1	1966.81	1347	7.6	55.8	94	97.524	unterdurchschnittlich
16	16	8	1	1	2154.67	1439	7.62	57.2	85	99.792	unterdurchschnittlich
17	17	9	1	1	1767.56	1029	6.03	57.2	97	81.648	unterdurchschnittlich
18	18	9	2	1	1827.92	1100	6.59	56.5	114	88.452	überdurchschnittlich
19	19	10	2	1	1773.83	1204	7.52	59.2	113	79.380	mittel
20	20	10	1	1	1971.63	1160	7.67	58.5	124	72.576	überdurchschnittlich

Study: Tramo et al. 1998 Brain size, head size, and intelligence quotient in monozygotic twins

Study of identical twins

Sample size $N = 20$

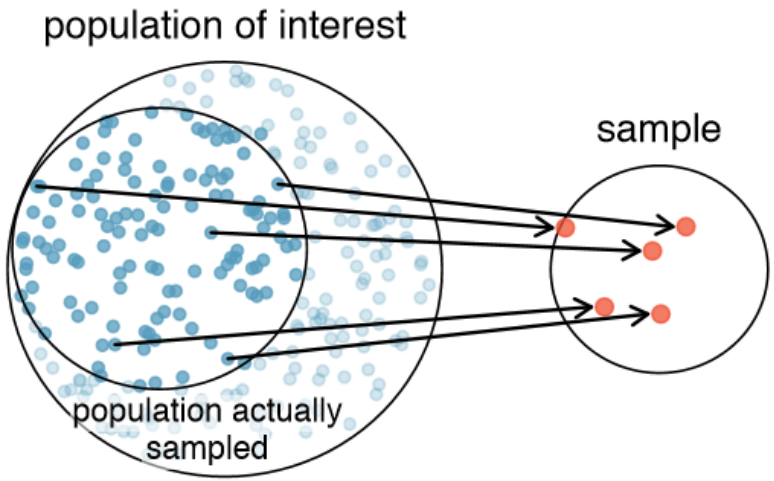
Variables:

- **ID:** Identifier of the individual person
- **IDZP:** Identifier of the twin pair
- **GR:** Birth order
- **GES:** Sex (1 = male, 2 = female)
- **OG:** Surface area of the cerebral cortex in cm^2
- **VG:** Volume of the forebrain in cm^3
- **CC:** Area of the corpus callosum in cm^2
- **KU:** Circumference of head in cm
- **IQ:** Intelligence quotient
- **KG:** Body weight in kg
- additionally **IQ.cat:** grouped intelligence quotient in 3 groups

Data collection: basics

Sampling theory (*sampling*) is a separate field of research with the aim of exploring methods of sampling without making systematic errors

Example of systematic error - causes e.g. *non-response bias*, incomplete *frame population* (see slide 17):

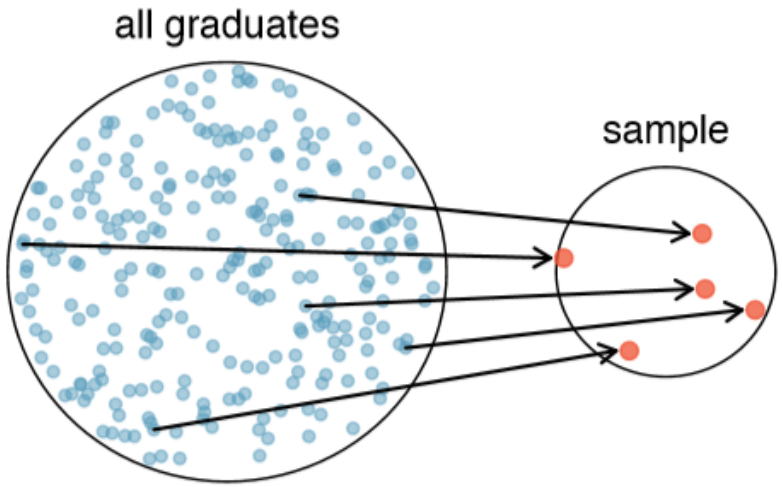


Graphics taken from: OpenIntro Statistics (Diez et al 2019); Zucchini et al. (2009), 27ff.

Data collection: sampling methods I

- **simple random sample:** random selection of statistical units; each object has the same probability of being included in the sample.
  - Advantage: systematic errors are avoided; can be representative

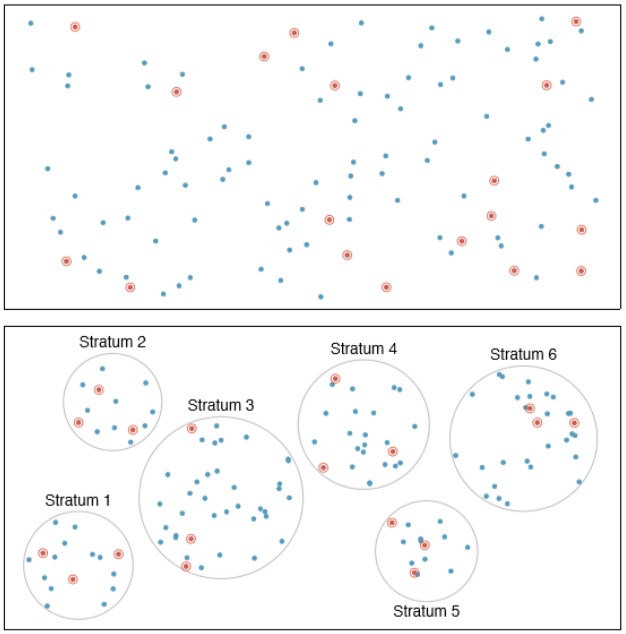
Example simple random sample:



Data collection: sampling methods II

- **simple random sample**
- **stratified random sample**. Population is divided into groups (*strata*) with the aim of finding strata that are similar in their characteristics - „divide-and-conquer sampling strategy“

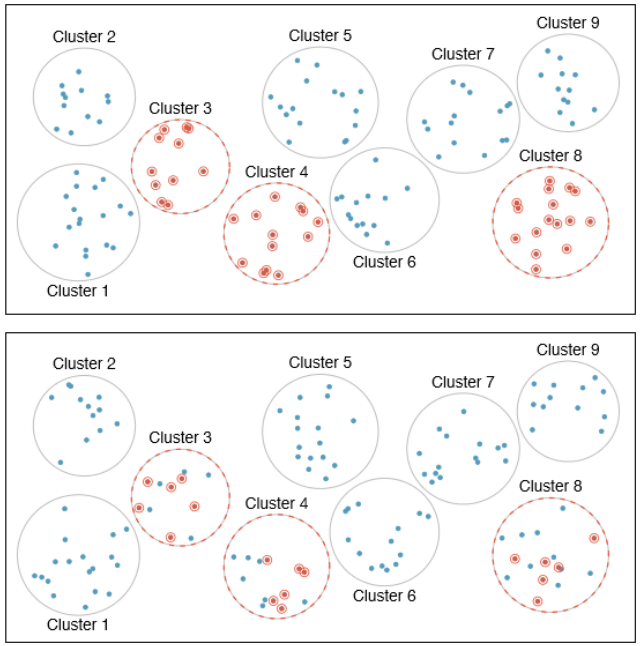
Example (1) simple random sample, (2) stratified random sample:



Data collection: sampling methods III

- **cluster sampling:** also groups (mostly according to natural concentrations of study units) formed; no systematic difference in their characteristics and thus small images of the population

Example cluster sampling (*multistage sampling*):



Data collection: sampling methods IV

other procedures (but these are methodologically criticised !):

- **quota selection:** certain quotas of samples are fixed from the outset (e.g. 90% of all education science students are male).
- **selection of typical cases:** statistical objects are selected as typical representatives of the population according to subjective criteria
- **selection according to convenience (*convenience sample*):** Statistical objects that are easy to reach have a higher probability of being included in the sample
- ...

Framework data collection

The population from which the sample is drawn determines the general validity of the results and conclusions!

$\rightarrow$ Diagram can be rewritten as $T \subseteq F \subseteq R \subseteq S$

- Presence of non-response: $S \subset R$
- List of statistical objects to be drawn (*frame population*) not complete, *target population*not covered: $F \subset T$

Leeuw, Hox und Dillman (2012), 318ff.

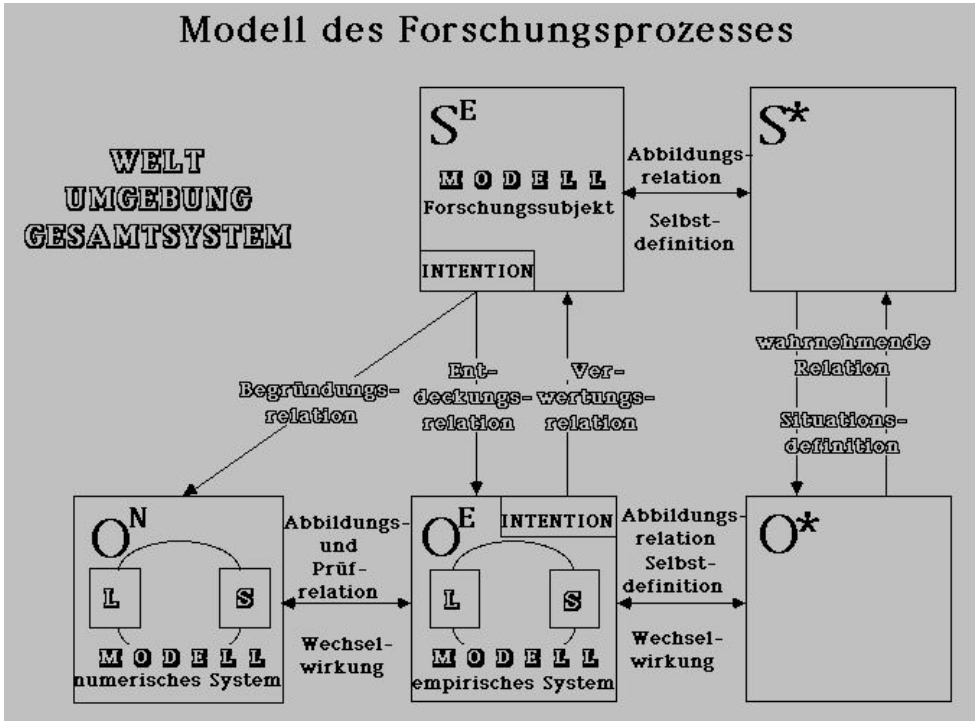
Take Home Messages: central terms, data collection

- Statistics can be divided into descriptive, exploratory and inductive statistics.
- Basic concepts allow the description of data sets (row are the statistical units and columns are the variables; the observed value of a variable represents a characteristic)
- Knowledge about methods of data collection allow (already within the experimental design) to critically question the generalisability of the results

Level of measurement: briefly measurement

Level of measurement indicates the information contained in a **measurement**

Definition of measurement: the entire process of observing empirical facts (research as a total system) up to the allocation of numbers to the observed phenomena (mapping relation).



Messung als Modellbildung (Stangl, 2019); Macho (2016), 1ff.

Level of measurement: nominal

- Expressions are names or categories assigned to statistical units
- Examples:
  - characteristic X_{gender} with the expressions x_{female}, x_{male}, x_{diverse}
  - Characteristic $Y_{\text{Hair colour}}$ with the expressions y_{brown}, y_{blond}, y_{other}
- lowest level of information
- nominal scaled variables can be assigned any labels and symbols / numbers in the dataset

allowed are injective transformations:
a, b, x, y Values of a nominal scale where $a = b$ and $x \neq y$
$\rightarrow f(a) = f(b)$ und $f(x) \neq f(y)$

Erklärvideos: studyflix Skalenniveau

Level of measurement: ordinal

- Expressions can be ordered, but intervals are not interpretable
- Example:
  - school grades[2] 1 better 2 but the grade interval between 1.2 and 4.5 cannot be compared / interpreted
- characteristic values are often assigned to ranks, these fulfil the ordering property of numbers (numbers can be put in an order with regard to their height)

allowed are strictly monotonic transformations:
the less-than relation is maintained if the following applies $x < y \rightarrow f(x) < f(y)$

[2] School grades are often assumed to be interval-scaled, but their measurement-theoretical goodness is critical, see e.g.. Neumann (2007), Lintorf (2012)

Level of measurement: interval

- distances between numbers can be interpreted in a meaningful way
- there is no natural zero (not useful to form quotients)
- Examples:
  - Temperature in Celsius (set 0° at freezing point, 100° at boiling point water).
  - Results of an intelligence test (often set at IQ: 100 ± 15).

allowed are positive affine transformations:
$f(x) = a + b * x$ mit $b > 0$

Note: Quadratic transformations are not allowed

Level of measurement: ratio

- distances between numbers can be interpreted meaningfully and statements about quotients (e.g. person 1 is twice as tall as person 2) are possible
- the ratio scale has a natural zero point and a defined unit of measurement.
- Examples:
  - temperature in Fahrenheit
  - Body height in mm, cm, m
- highest level of information

allowed are formal transformations:
$f(x) = b * x$ mit $b > 0$

Level of measurement: summary

Skalenart	count	order	form differences	form ratios
nominal	yes	no	no	no
ordinal	yes	yes	no	no
interval	yes	yes	yes	no
ratio	yes	yes	yes	yes

Tabelle: Meaningful calculations for the different scale levels

→ Nominal scaled variables are also called **qualitative variables**
→ Interval scaled and ratio scaled variables are also called **quantitative variables**

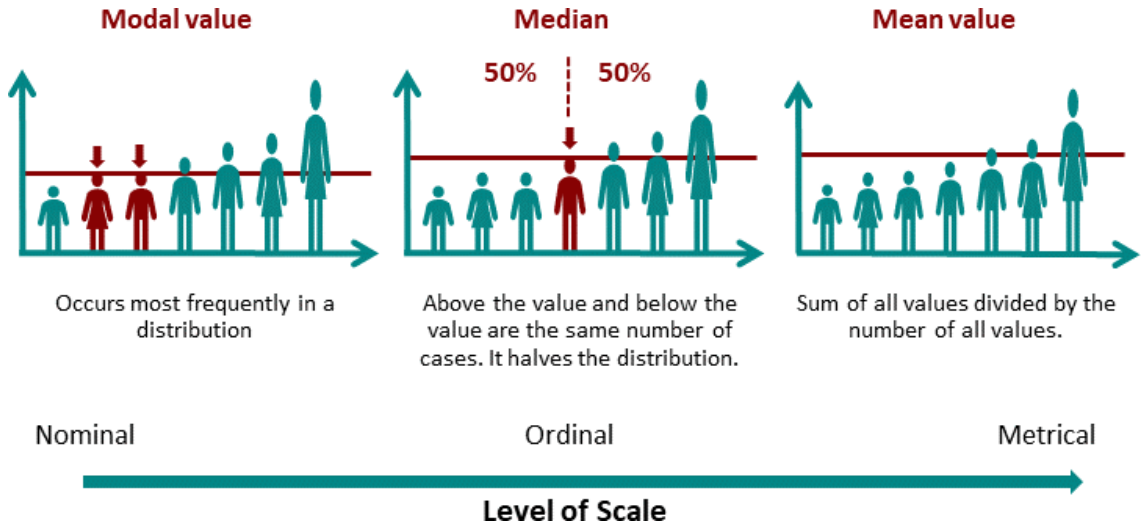
→ Nominal scaled and ordinal scaled variables represent **discrete variables**.
→ Interval-scaled and ratio-scaled variables represent **metric variables**

Level of measurement: classification of variables

- **qualitative:** Variables can reflect different categories of a characteristic
- **quantitative:** Variables can reflect a different extent or intensity of a characteristic
  - ordinal scaled variables are in a grey area here, because values can be ordered, but distances between the characteristic expressions (quantitative aspect) cannot be interpreted.
- **discrete:** set of expressions of a variable can be represented by the first n natural numbers ($\mathbb{N} = \{1, 2, ..., n\}$) and is thus countable
- **continuous:** between two values there can always be (theoretically) infinitely many other values, range of values are all real numbers ($\mathbb{R} = \{-\infty, \infty\}$)

Take Home Messages: level of measurement

- Scale level of a variable indicates the content of information of a measurement.
- the information content of the scales is as follows: ratio > interval > ordinal > nominal
- scale levels have a **downward compatibility** any procedure that we may apply to a low scale level we can also apply to a higher one



R-commands section level of measurement

R is an object-oriented language and it must override the „object“ - i.e. the variable of the record - with the following syntax structure:
`variable <- command(variable)`

- nominal scale: `factor(x = daten$variable)`
- ordinal scale:
  `factor(x = daten$variable, ordered = TRUE)`
- metric scales:
  `as.numeric(as.character(x = daten$variable))`
  → the distinction between interval and ratio scales is not relevant for **R** (as well as e.g. SPSS)

note: furthermore, **R** distinguishes between integer ($\mathbb{N}$) and continuous ($\mathbb{R}$) variables by the commands: `as.integer(x = )` resp. `as.numeric(x = )`

Data generation process

Data generation process: „When [...] data set has been created, a good model has to be constructed for the data generation process (DGP). This is the stage at which the actual [...] analysis begins... When an [...] model has been constructed for the DGP, it should **only be used for the analysis if it reflects the ongoings in the system of interest properly**“ - Lütkepohl und Krätzig (2004)

$\rightarrow$ a model should make a parsimonious, plausible and substantially meaningful contribution to the data generation process - Hoyle (2012)

$\rightarrow$ in statistics, we assume that we can represent the DGP by discrete and continuous distributions

Distributions I

a random variable is defined by assigning numbers to an event of a random process (realisation), for this purpose **distributions** are necessary:

- Random processes can only be described by functional relations $f(X)$, types of representation:
  - for discrete random variables $\rightarrow$ probability function
  - for continuous random variables $\rightarrow$ density function

$\rightarrow$ distributions are determined by functions and their parameters

A random variable $f(X)$ is a function that assigns a real number to each possible outcome of a random experiment / process.

furter information see: Wikipedia random variables; Fahrmeir et al. (2016); Zucchini et al. (2009)

Distributions II: example I

Random variables can only take on finitely many values or countably infinitely many values
$\rightarrow$ each possible event can be assigned a natural number

Example: rolling 6 or 30 fair dices several times at once and we want to know $A = \{\text{result is } 1\} = \{1\} \rightarrow P(\{1\}) = 1/6$ (binomial distributed):

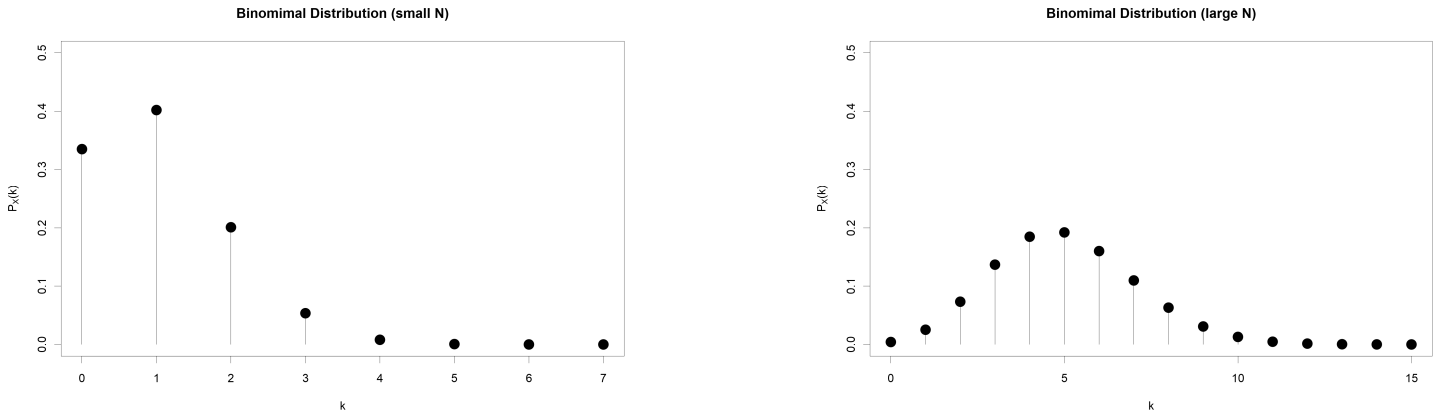


Abbildung: left rolling 6 dices, right 30 dices

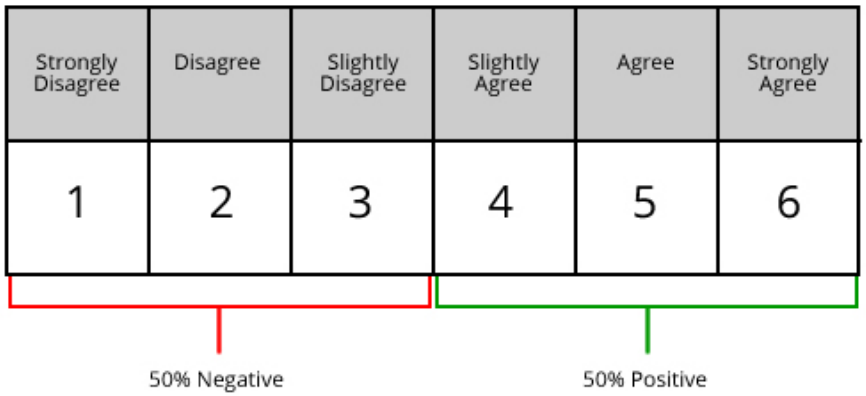
Is rolling a dice a Gauss distribution?

Distributions II: example II

single items can be binomial distributed (! however the real distribution is never known)

Imagine that participants should answer the following two items on a 6 point Likert Scale:

- Should something be done about climate change for future generations?
- Would you be willing to accept personal restrictions for the fight against climate change?



→ What would you expect?

Take-Home-Message: surveys

- individual items of a Likert scale result from a binomially distributed process ($Y_1 \sim Bin(n, p)$, ...)
  - $\rightarrow$ we want to achieve a high variance in the items (all response options are used)
  - $\rightarrow$ variance or covariance (if multiple items) is the most important building block for the statistical models
  - $\rightarrow$ the answering of items should be independent from other people or the ordering of the items
- our statistical models should reflect the ongoings in the system of interest properly / sufficiently and should be theoretically derived

> Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful
>
> *George Box (1987)*

Descriptive statistics: Motivation

- to be able to analyse data, one needs an impression of the available information
- data sets are so large that direct viewing of the data is not useful
- goals include:
  - description and distribution of individual variables
  - detection of data errors (implausible values, outliers,..)
  - investigation of correlations between variables

Descriptive statistics: Fundamentals

In general, statistical procedures can be divided into how many variables are considered at once:

- univariate: Representation of values one variable (also: denotes dependence on only one variable, one-dimensional)
- bivariate: Representation of the values of two variables
- multivariate: Representation of the values of several variables

$\rightarrow$ the statistical procedures and the assumed distributions behind them (see from slide **??**) become increasingly complex

Descriptive statistics: univariate

The aim is the representation, description of data consisting of observations of a single characteristic (univariate, one-dimensional data)
$\rightarrow$ Comparison of several univariate results enables impressions about possible dependencies that can be further analysed using multivariate procedures

univariate representations / statistical methods:

- frequency distributions
  - frequency table
  - bar chart / histogram
- central measures (from slide 48)
- symmetry of distributions (from slide 82)

Frequency table I

Initial situation: Observed characteristics $x_1, ..., x_n$ of a feature X
$\rightarrow$ Raw data is summarised according to different occurring counts / characteristics $a_1, ..., a_k$, where $k \leq n$ (for categorical characteristics k is equal to the number of categories)

Representation of a part of the original list for the variable IQ.cat (grouped intelligence quotient in 3 groups):

1	2	3 - 18	19	20
below average	below average	...	medium	above average

1	2	3 - 18	19	20
1	1	...	2	3

$\rightarrow$ here is an ordinal scaled variable, thus $3 > 2 > 1$ (this would not be the case with a nominal scaled variable)

Frequency table II

Frequency table for the variable IQ.cat:

Characteristic	class	absolute frequency h	relative frequency f
below averag	$85 < .. \leq 95$	11	$11/20 = .55$
medium	$96 < .. \leq 113$	5	$5/20 = .25$
above average	$114 < .. \leq 127$	4	$4/20 = .20$

- absolute frequency h: $h(a_j) = h_j$, number of x_i from $x_1, .., x_n$ with $x_i = a_j$.
- relative frequency f: $f(a_j) = f_j = h_j/n$

Bar chart

- Row chart: a bar perpendicular to the x-axis with absolute heights $h_1, ..., h_k$ or relative heights $f_1, ..., f_k$ is plotted over different occurring characteristics $a_1, ..., a_k$.
- Bar chart: like bar chart but with x-axis placed vertically instead of horizontally

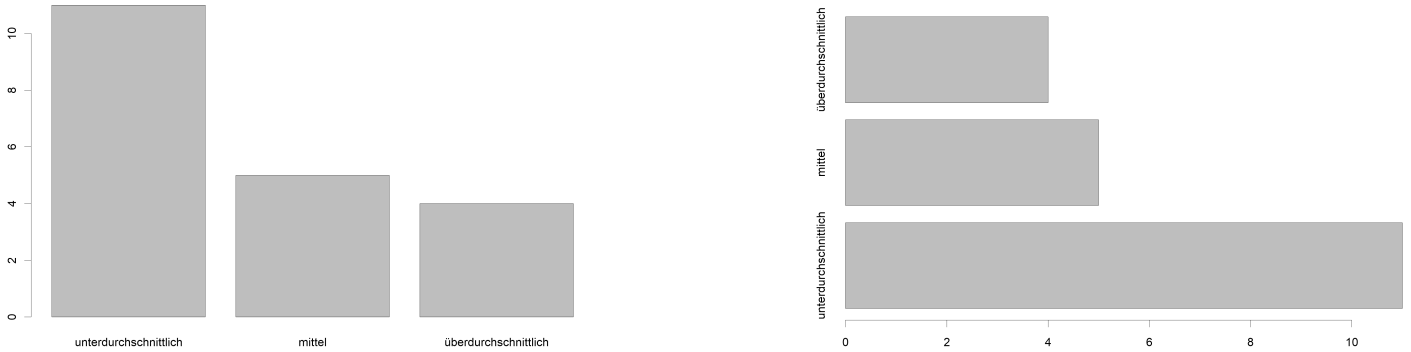


Abbildung: left bar chart, right row chart

Histogram I: construction

The construction of a histogram is done under the principle of area fidelity, so that the areas are proportional to the absolute h_j or relative frequencies f_j:

1. Data are grouped into classes $c_o, .., c_k$, where these have adjacent intervals: $[c_o, c_1), .., [c_{k-1}, c_k)$
   - Rectangles in the histogram have the width $d_j = c_j - c_{j-1}$
2. Height of the rectangles in the histogram are defined by $frequency / classwidth$
   - absolute frequencies: h_j / dj
   - relative frequencies: f_j / dj
3. Area over the intervals results from $Area = Width \, x \, Height$

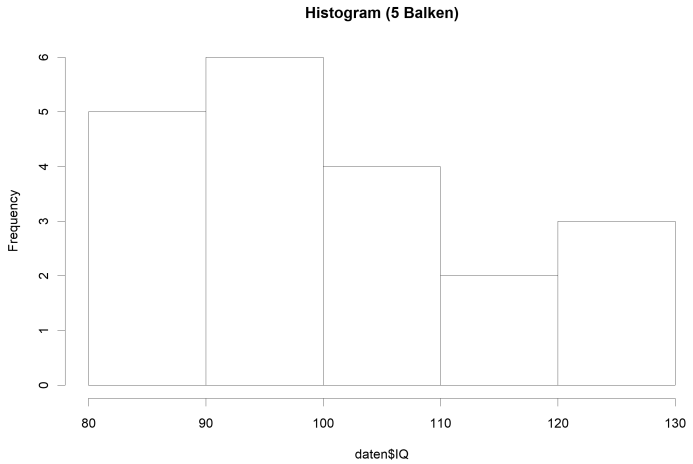
$\rightarrow$ if class widths d_j are equal, the absolute h_j or relative frequencies f_j are simply deducted as height

Histogram II: example

Representation of a histogram for the variable IQ:

$(c_{j-1}, c_j]$	h_j	b_j	h_j / b_j
$80 \leq x \leq 90$	5	10	5
$90 \leq x \leq 100$	6	10	6
$100 \leq x \leq 110$	4	10	4
$110 \leq x \leq 120$	2	10	2
$120 \leq x \leq 130$	3	10	3

$\rightarrow$ all class widths d_j are the same size



Histogram III: number of classes

Formation of classes always represents a certain loss of information (*data aggregation*) $\rightarrow$ Number of classes has a decisive influence on the visual impression of a distribution

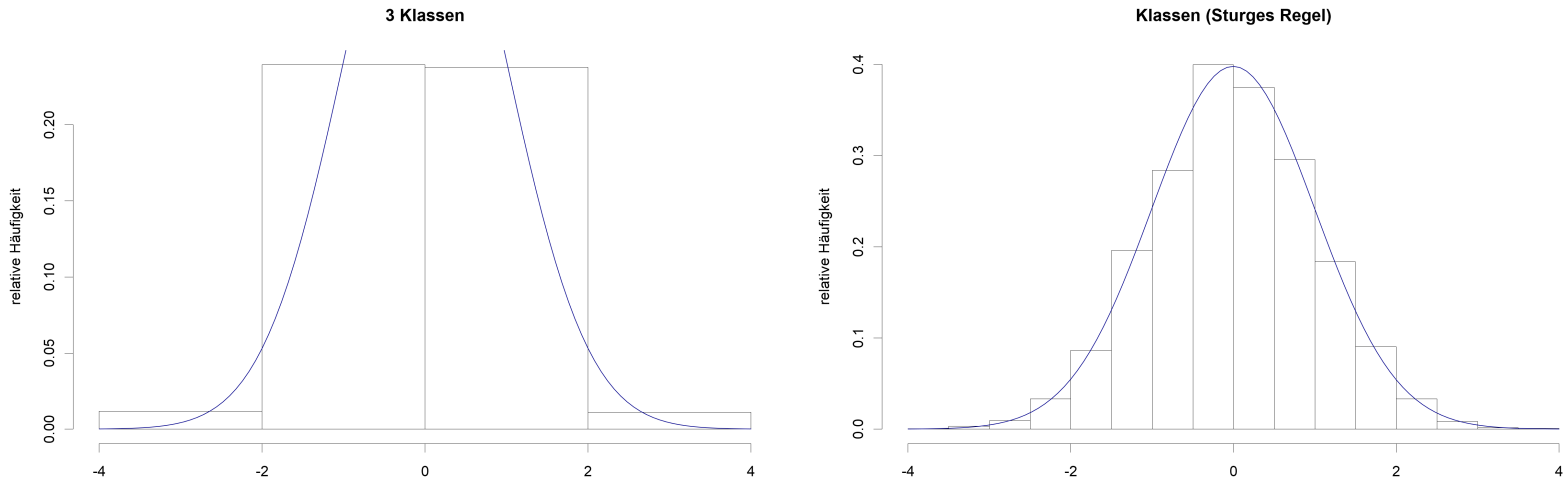


Abbildung: left 3 classes, right Sturges rule*

$\rightarrow$ blue line represents the true distribution, this is a standard normal distribution $X \sim \mathcal{N}(0, 1)$ (see later from slide 153)

Histogram IV: number of classes

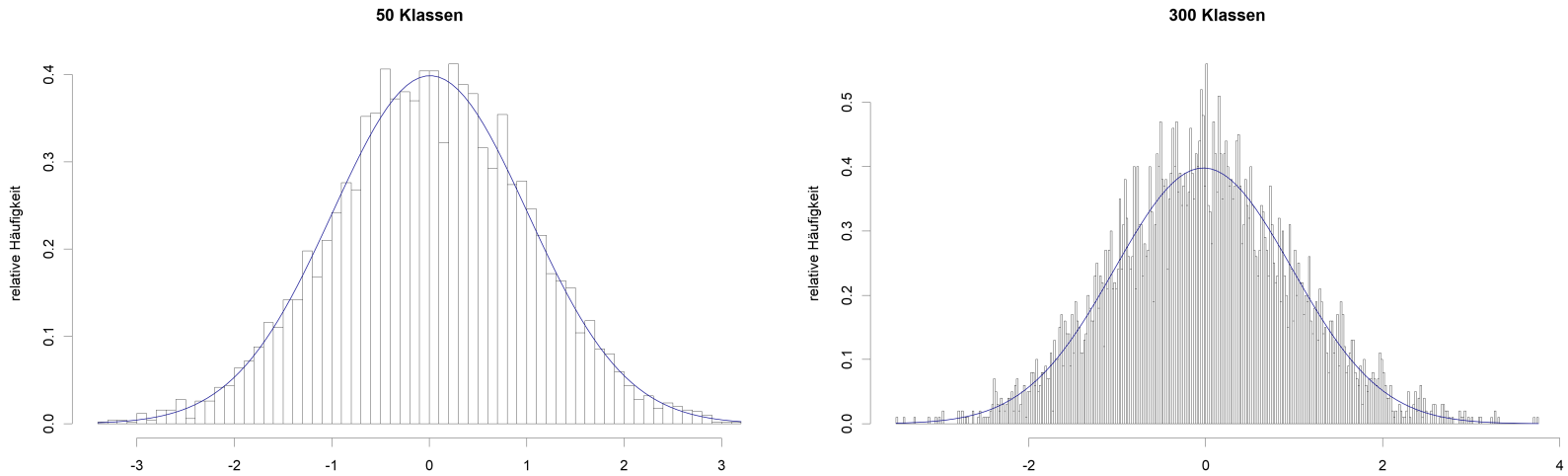


Abbildung: left 50 classes, right 300 classes

Histogram V: number of classes

Number of classes significantly influences the shape of the histogram

- too low number give only a smooth estimate $\rightarrow$ low variability
- too high number give a good approximation of the data, but a very rough estimate $\rightarrow$ large variability

this is a classic **bias-variance trade off**, i.e.

- J small: large bias but low variability
- J large: small bias but large variability

Number of classes is often estimated by the statistics programme using rules of thumb, e.g. using the *Sturges rule: $\lfloor (1 + \log_2 n)$

Cumulative frequency distribution: construction

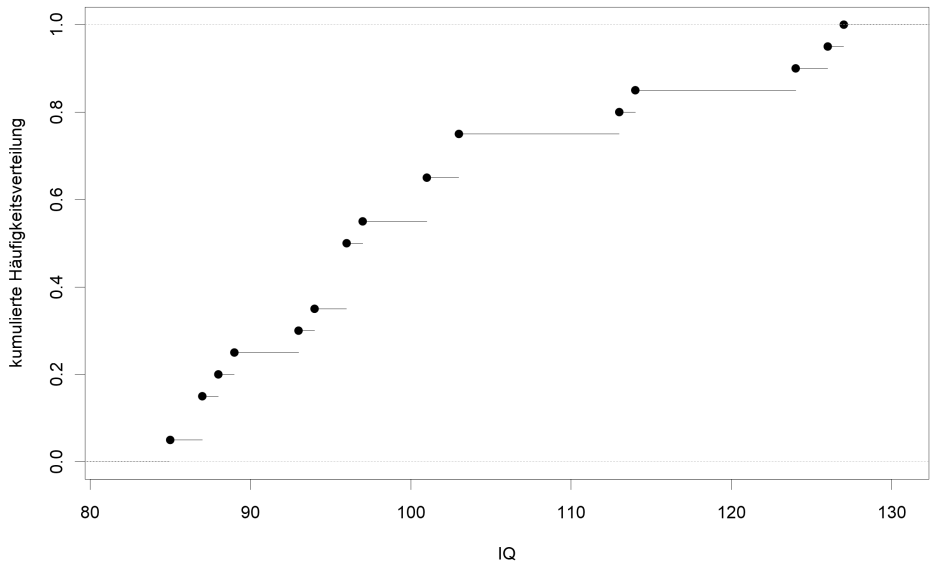
In contrast to the histogram, allows the most exact representation of the measured values (! no loss of information) with metric variables

- absolute cumulative frequency distribution: determine the number of observations (characteristic values) that are less than or equal to x: $H(x) =$ Anzahl der Werte x_i mit $x_i \leq x$
  - these absolute frequencies can be written as $H(x) = h(a_1) + ... + h(a_j)$ (sum of the steps)
- relative cumulative frequency distribution (also empirical distribution function): division of the absolute frequencies by the number of observations: $F(x) = H(x)/n$

$\rightarrow$ these are monotonically increasing staircase functions which jump upwards at the values $a_1, ..., a_k$.

Cumulative frequency distribution: example

What proportion of the data is less than or equal to a value of interest x?



Representation of the variable IQ in the data set of Trame et al. 1998 (see slide 10)

Take Home Messages: Choice of presentation

- discrete variables with few expressions: pie chart and bar chart
- metric variables, but also discrete variables with many values: Histograms (! certain loss of information)
- metric variables with many expressions: cumulative frequencies (! no loss of information, but the shape of the distribution is not visible - see from slide 82)

$\rightarrow$ Linking level of measurement with forms of representation

R-commands section univariate statistics using graphs

- Create a frequency table: `table(x)`
- Bar chart: `barplot(table(x), horiz = TRUE)`, Säulendiagramm: `barplot(table(x), horiz = FALSE)`; here, the argument `horiz` only specifies whether the bars are horizontal or vertical
- Histogram with absolute frequencies: `hist(x)`, Histogram with relative frequencies: `hist(x, freq = FALSE)`
- Cumulative frequency distribution: `plot(x = ecdf(x))`

Measures: Basics I

- Measures (*also parameters*) summarise the essential properties of a frequency distribution, these are divided into:
  - Location parameters / measures of central tendency: represent central tendency, centre of gravity of data.
  - dispersion parameter / dispersion: represent distributions / dispersions of data around the centre of gravity
  - shape parameter / skewness: represent the shape of the distribution of data (statements about symmetry, see from slide 82)
- Note: some of the measures correspond to moments of random variables (see from slide 143)

Measures: Basics II

Which measure is appropriate depends on the scale level (and context, as well as the question):

Beschreibung der	Statistik	Skalenniveau
zentralen Tendenz	Mittelwert	mind. Intervallskala
	Median	mind. Ordinalskala
	Modus	mind. Nominalskala
Position in der Verteilung	Quantile (Quartile, Dezile)	mind. Ordinalskala
Streuung	Varianz	mind. Intervallskala
	Standardabweichung	mind. Intervallskala
	Interquartilabstand	mind. Intervallskala
	Spannweite	mind. Intervallskala
Schiefe	Schiefekoeffizient	mind. Intervallskala
	Quartilskoeffizient	mind. Ordinalskala
Wölbung	Kurtosiskoeffizient	mind. Intervallskala

Measures: Basics II

Which measure is appropriate depends on the scale level (and context, as well as the question):

Scale type	Empirical properties	Permissible statistics	Examples
Nominal	Equivalence	Mode, chi square	Eye colour, place of birth, etc…
Ordinal	Equivalence, order (greater or less)	Median, percentile	Surface hardness, military rank, etc…
Interval	Equality, order, distance (addition or subtraction)	Mean, standard deviation, correlation, regression, analysis of variance	Temperature in °C, serial numbers, etc…
Ratio	Equality, order, distance, ratio (multiplication or division)	All statistics permitted for interval scales plus the following: geometric mean, harmonic mean, coefficient of variation, logarithms	Temperature in K, weight, age, number of children, etc…

Measures of central tendency: Mean - formula

Prerequisite: metric variables

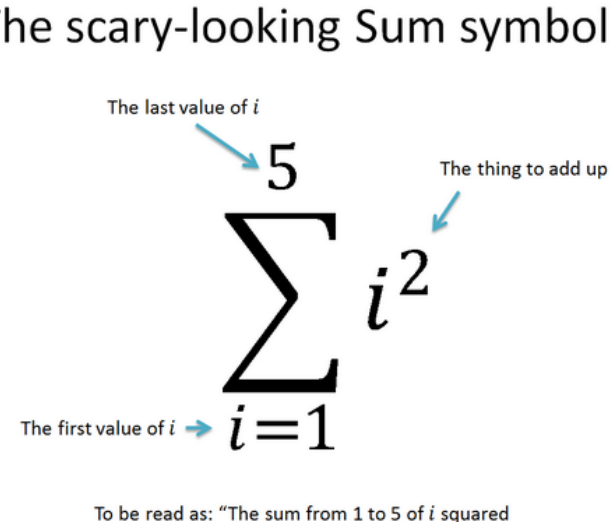
Formula:

$$\overline{x} = \frac{1}{n}(x_1 + .. + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Formula for stratification for r group averages:

$$\overline{x} = \frac{1}{n}(n_1\overline{x}_1 + .. + n_1\overline{x}_1) = \frac{1}{n}\sum_{j=1}^{r} n_j x_j$$

briefly sum sign



- Sum sign consists of the capital greek letter Σ followed by a function
- running index (also count variable), this appears again in the function
- the running index starts from its start value and ends at the given end value (*here 5 steps*)

Wikipedia Summation

Measures of central tendency: Mean - properties I

the mean value has a so-called **center of gravity property**, sum of deviations between x_i and $\overline{x}$ disappears:

$$\sum_{i=1}^{n}(x_i - \overline{x}) = 0$$

$\rightarrow$ no statement about the dispersion of the data possible, therefore choice of squared differences - the smaller this value, the better the mean value represents the data:

$$\overline{x} = \min_{c \in \mathbb{R}} \left(\sum_{i=1}^{n}(x_i - c)^2 \right) \tag{1}$$

Note: Term states that no value c from the real numbers $\mathbb{R}$ other than the mean $\overline{x}$ leads to a smaller sum of the squares of the deviations

Measures of central tendency: Mean - properties II

How representative are the group means (green) for the available data (*example: test performance of 5 schools*)?

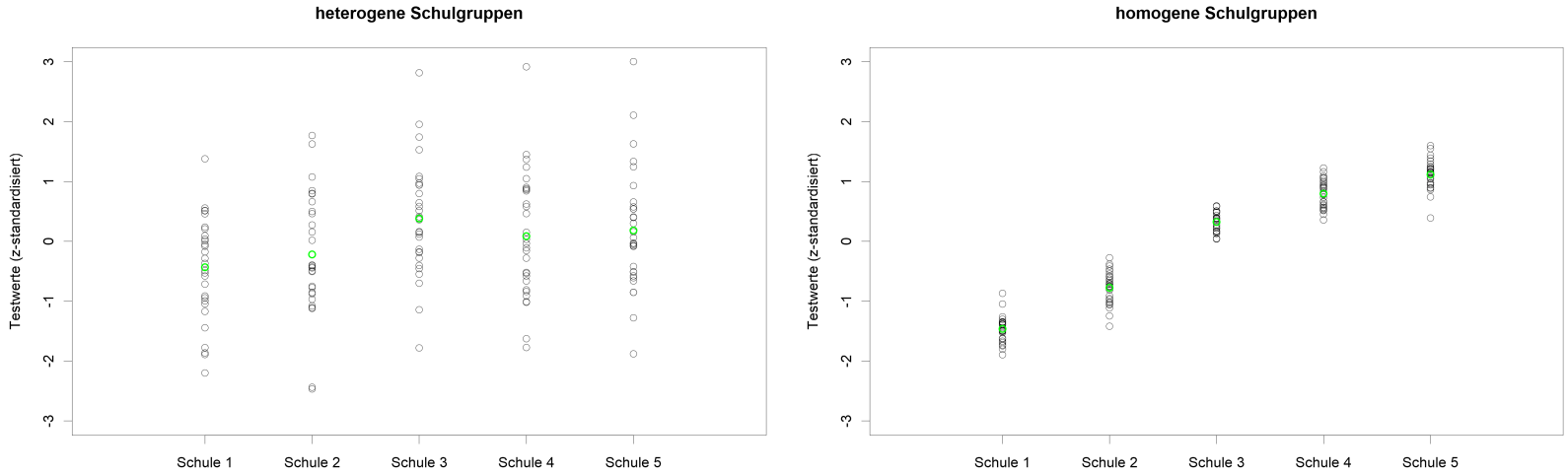


Abbildung: Heterogeneous school groups on the left, homogeneous school groups on the right (group mean representative)

Measures of central tendency: Median

Prerequisite: at least ordinal scaled variables
$\rightarrow$ the median is **robust / resistant** to extreme values (also outliers)[3]

„calculation“: Median is placed in the middle of the data so that at least 50% of the values are less than / equal to and at the same time at least 50% of the values are greater than / equal to the median.

$$x_{med} \begin{cases} = x_{(\frac{n+1}{2})} & \text{if n odd} \\ \in [x_{(\frac{n}{2})}, x_{(\frac{n+1}{2})}] & \text{if n even} \end{cases}$$

$\rightarrow$ for even n, the following rule is often used:

$$x_{med} = \frac{1}{2}(x_{(\frac{n}{2})} + x_{(\frac{n+1}{2})})$$

[3] extreme values: measured value that does not meet expectations (! subjective), often values (arbitrarily) are called outliers that lie outside 1.5 times the quartile distance (see from slide 87, as well as from slide **??**)

Measures of central tendency: Median - properties I

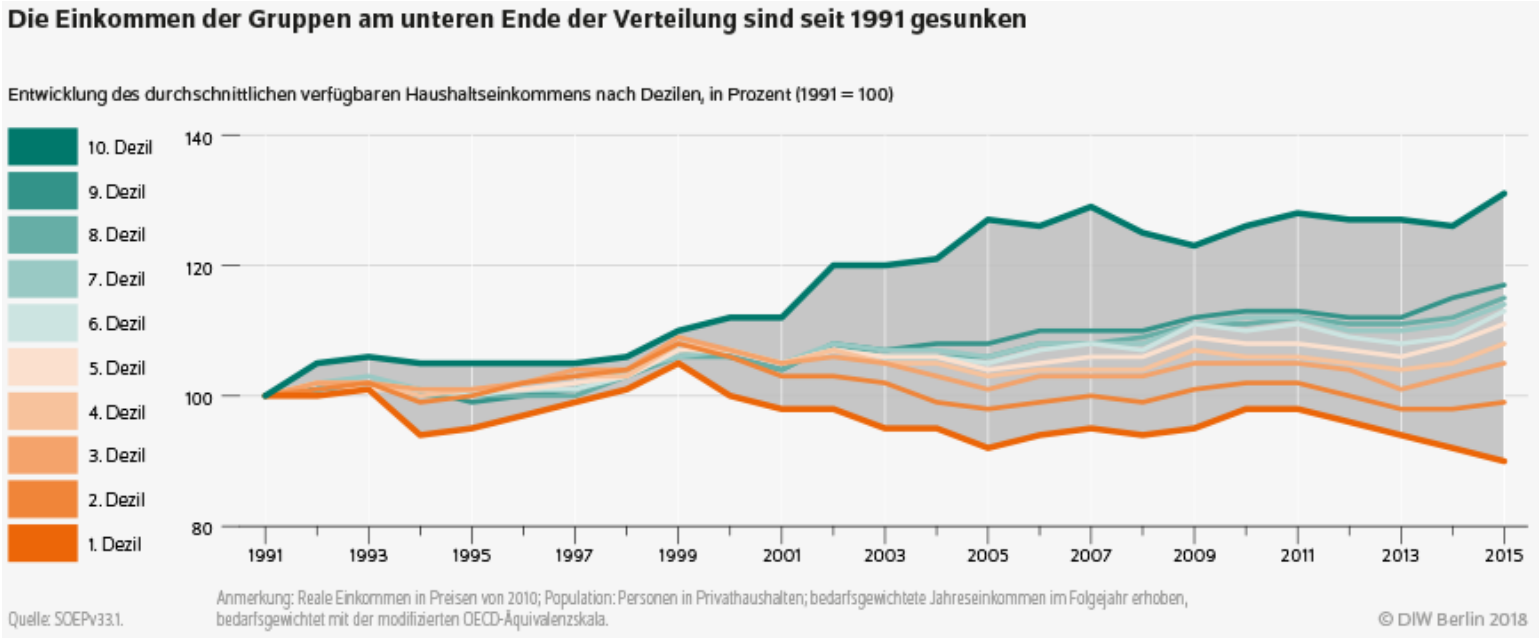
the median is exactly in the centre of the data

$$x_{med} = \min_{c \in \mathbb{R}} \left(\sum_{i=1}^{n} |x_i - c| \right) \tag{2}$$

Note: Term states that no value c from the real numbers $\mathbb{R}$ other than the median leads to a smaller sum of the deviation

Measures of central tendency: Median - properties II

Is the mean income in Germany determined by the mean or median?



German Institute for Economic Research: Real Income Germany between 1991 and 2015
vs.
Wikipedia Income distribution in Germany
current report: WSI-Verteilungsbericht 2019

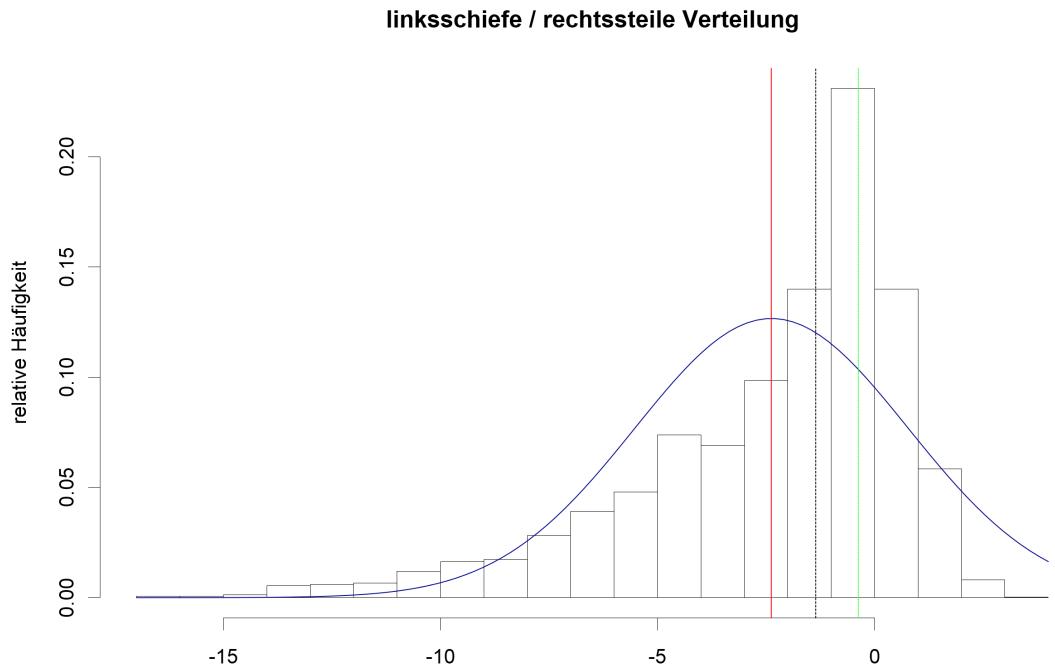
Measures of central tendency: Mode

Prerequisite: at least nominally scaled variables, most important positional measure for categorical variables
the mode, like the median, is **robust / resistant** to extreme values

„Calculation“: x_{mod} is the expression with the highest frequency, it is only unique if the frequency distribution has only one maximum

Measures of central tendency: Summary

- mean indicates centre of gravity of the data, cf. equation (1)
- Median indicates the midpoint (50%) of the data, cf. equation (2)
- Mode indicates the most frequently occurring value



vertical lines from left to right: mean, median, mode

Measures of the central tendency: prerequisites

- mean
  - metric features
  - also useful for binary 0 / 1 coded characteristics
  - partly used for ordinal scaled variables (like grades) (! interpret with care)
- median
  - ordinal scaled variables
- mode
  - Applicable to any scale level, but rarely interpretable for metric variables

Measures of central tendency: allowed transformations

- mean allows positive affine transformations (= interval scale, see from slide 20):

$$f(x) = a + b * x \;\rightarrow\; \overline{x} = a + b * \overline{x} \quad mit\ b > 0$$

- Median allows strictly monotonic transformations (= ordinal scale):

$$x < y \;\rightarrow\; f(x) < f(y)$$

- mode allows injective transformations (= nominal scale): a, b, x, y values of a nominal scale where $a = b$ and $x \neq y$ and $f(a) = f(b)$ and $f(x) \neq f(y)$

Position in the distribution: quantiles

Prerequisite: at least ordinal scaled variables
Measure is a generalisation of the median (median divides the data into the 50% quantile)

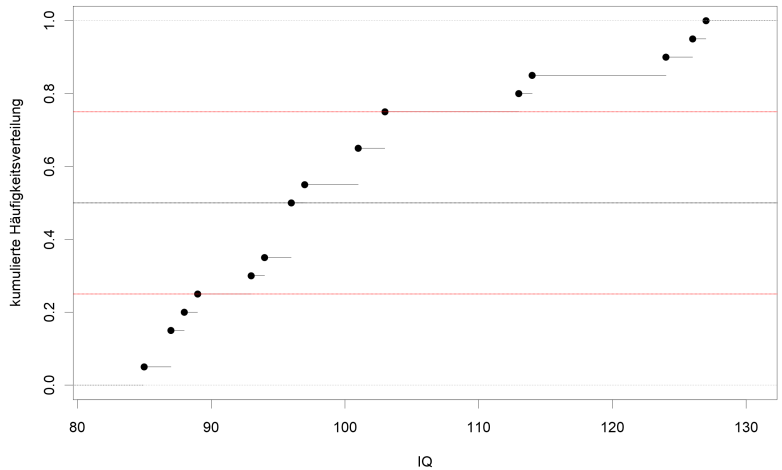
„Calculation“: at least $p * 100\%$ of the data are less than or equal to and $(1 - p) * 100\%$ are greater than or equal to the p-quantile x_p, where $0 \leq p \leq 1$

$$x_p \begin{cases} = x_{(np+1)} & \text{, if n*p is not integer} \\ \in [x_{(np)}, x_{(np+1)}] & \text{, if n*p is integer} \end{cases}$$

Position in the distribution: quantiles - example

quartiles (25% p-quantiles) of the variable IQ in the data set of Trame et al. 1998 (see slide 10):

```
p*100 = %:    0%    25%    50%    75%   100%
x:          85.0   92.0   96.5  105.5  127.0
```



Position in the distribution: quantiles - types

- **Quartile:** $x_{.00}, x_{.25}, x_{.50}, x_{.75}, x_{1.00}$
- **Decile:** $x_{.10}, x_{.20}, .., x_{.90}$
- **Percentile:** $x_{.01}, x_{.02}, .., x_{.99}$
- $x_{.25}$: lower quantile
- $x_{.50} = y_{med}$: Median
- $x_{.75}$: upper quantile
- as well as *Inter Quartile Range, IQR* $d_Q = x_{.75} - x_{.25}$
  - in this range are the 50% of the data; this is a measure of dispersion; answers the question: *How far apart are the data?*
- as well as **span** $d_Q = x_{1.00} - x_{0.00}$ (maximum - minimum)

$\rightarrow$ relevant for the 5-point summary (see slide 87)

Dispersion: Variance - Formula

Definition: Measure of the dispersion of the probability density (values) around its centre of gravity;
Prerequisite: metric variables

Formulas, compare equation (1):
empirical variance:

$$\widetilde{s}^2 = \frac{1}{n}\left((x_1 - \bar{x})^2 + .. + (x_n - \bar{x})^2)\right) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Sampling variance:

$$s^2 = \frac{1}{n-1}\left((x_1 - \bar{x})^2 + .. + (x_n - \bar{x})^2)\right) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$\rightarrow$ mostly the sample variance is calculated, since the empirical variance underestimates the population variance (practical reason); as well as only $n-1$ variances vary freely and thus one averages by the number of degrees of freedom (theoretical reason)

Dispersion: Variance - properties I

The idea of defining variance

1. Deviations between $x_i - \bar{x}$ measure how much the data scatter around mean $\bar{x}$.
2. by squaring, all deviations receive a positive sign: $(x_i - \bar{x})^2$
   - sensitive to outliers: large x-values enter the formula strongly
   - ! the variance s^2 no longer has the same unit of measurement due to squaring, e.g. body height measured in cm, so the unit of variance is cm^2
3. variance is the mean of the squared deviations and is small if the values scatter closely around the mean value

Dispersion: Variance - properties II: calculation

1. $x_i - \bar{x}$
2. $(x_i - \bar{x})^2$
3. Interpretation, *Is there an outlier?*

Example for the variable IQ: $\bar{x} = 101$, $s^2 = 174.53$, $s = 13.21$

IQ	$IQ_i - \bar{IQ}$	$(IQ_i - \bar{IQ})^2$
96	-5	25
89	-12	144
87	-14	196
...	...	...
113	12	144
124	23	529
$\sum_{i=1}^{n}$	0	3316

$$s^2 = \frac{1}{20-1} * 3316 = 174.53, \quad \sqrt{s^2} = 13.21$$

Dispersion: Standard deviation

Objective: Variance of a variable is always given in square units, standard deviation has the same unit as the variable itself
Prerequisite: metric variables

Formula: Root of the sample variance

$$s = \sqrt{s^2}$$

$\rightarrow$ the units of the standard deviation s now correspond to the original unit of measurement

Dispersion: variance, standard deviation - properties

- sensitive to outliers / extreme values: due to squaring, strong deviations have a high weight.
  - *e.g. last value of variable IQ is replaced by 220, change variance:* $s^2 = 174.53 \rightarrow s^2 = 867.75$ *(also mean:* $\bar{x} = 101 \rightarrow \bar{x} = 105.8$*)*
- variance can be linearly transformed: $fy_i = a + b * x_i$, so variance is unchanged for a and for b: $s_y^2 = b^2 s^2$
- by means of the standard deviation and the arithmetic mean under the assumption that the characteristic is normally distributed (see from slide 153), classic confidence intervals can be formed (see from slide **??**)

Skewness: basics

! link to symmetry of one-dimensional distributions (see from slide 82)

the dimensions characterise the following:

- skewness: statistical ratio describing the type and strength of the asymmetry of a probability distribution.
  - $g_m < 0$: distribution skewed to the right (right skew, left skew, negative skew).
  - $g_m > 0$ distribution skewed to the left (left-skewed, right-skewed, positive skewness)
- kurtosis: measure of the steepness of a (single-peaked) probability function (also density function, frequency distribution).
  - $\gamma > 0$: steep-peaked (also leptokurtic), more peaked distribution compared to normal distribution
  - $\gamma < 0$: flat-peaked (also platykurtic), flattened distribution compared to normal distribution

Skewness: coefficient I

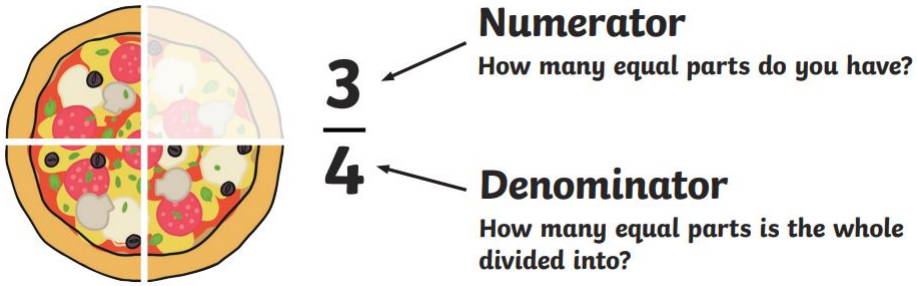
Prerequisite: metric variables
Skewness of a random variable X is the central moment[4] 3rd order, normalised to the standard deviation s formula:

$$g_m = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}^3} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^3 = \frac{m_3}{\widetilde{s}^3}$$

$\rightarrow$ Note second term: this is nothing more than a z-standardisation raised to the power of 3 (see slide 153)

[4] central moments describe the distribution of the probability mass around the expected value $\mu = \mathbb{E}(X)$ (mean) of the random variable X, which are defined as $\mu_k := \mathbb{E}\left((X - \mu)^k\right)$

briefly importance to understand fractions



→ by subtracting the mean in the numerator and dividing the denominator (often standard deviation), a variety of measures are **normalised** (also standardised) in statistics to make their range of values independent of the metric of the variable (e.g. body weight measured in g vs. kg), examples are:

- Coefficient of variation: $v = \frac{s}{\bar{x}}, \bar{x} > 0$ - Fahrmeir et al. (2016), 68
  - → normalised to: $0 \leq v \leq 1$
- Correlation coefficient
  - → normalised to: $-1 \leq r_{XY} \leq 1$
- ...

Skewness: coefficient II

the skewness coefficient g_m is constructed as follows:

- in the numerator, the deviations from the mean are calculated and raised to the power of 3
- in the denominator the standard deviation is raised to the power of 3
  - denominator is always positive in contrast to the numerator (*e.g.* $-2^3 = -8$)

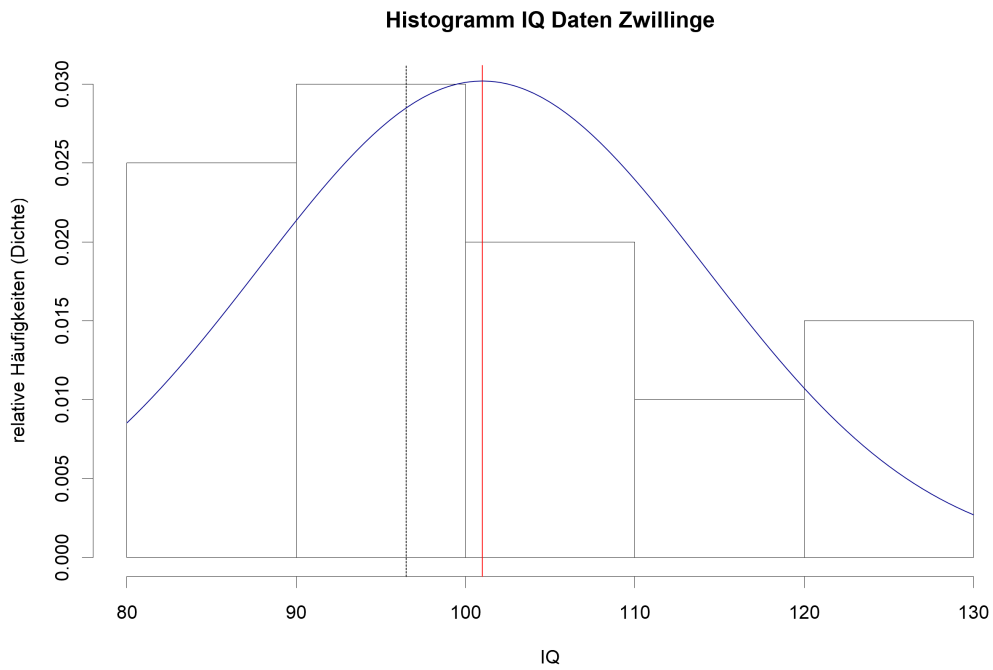
  $\rightarrow$ symmetric distribution: sum of negative and positive deviation terms in the numerator is approximately equal

  $\rightarrow$ left skewed distribution: sum of the deviation terms in the numerator is larger for the observations below the mean; vice versa for right skewed distributions

$\rightarrow$ the further the skewness coefficient g_m deviates from 0, the more left-skewed or right-skewed is the distribution

Skewness: coefficient III: example

Example of the variable IQ - graphical representation:
positive skewness: Schiefekoeffizient = .81 (also right-skewed, right-tailed, left-leaning curve)



vertical lines from left to right: median, mean

Skewness: quartile coefficient I

Prerequisite: at least ordinal scaled variables
$\rightarrow$ Advantage: in contrast to the skewness coefficient, the quartile coefficient is not influenced by outliers

formula:

$$g_p = \frac{(x_{1-p} - x_{med}) - (x_{med} - x_p)}{x_{1-p} - x_p}$$

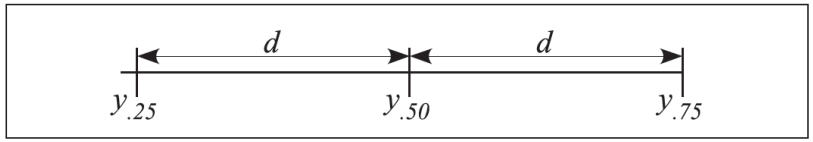
often $p = .25$ is set to obtain the classical quartile coefficient:

$$g_p = \frac{(x_{.75} - x_{.50}) - (x_{.50} - x_{.25})}{x_{.75} - x_{.25}}$$

Skewness: quartile coefficient II

the quartile coefficient g_p is constructed as follows:

- this calculates the difference between the upper quantile and the median, as well as the difference between the median and lower quantile, example symmetrical distribution:



- $g_{.25} < 0$: Distribution skewed to the right (negative skewness).
- $g_{.25} > 0$ distribution tilted to the left (positive skewness)

Skewness: Example of skewness and quartile coefficient

Representation of two boxplots for the variable IQ (with plotted values):

Abbildung: links Variable IQ, rechts Variable IQ mit Datenfehler

Result:

- $g_m = .81$: right-skewed distribution (positive skewness) respectively $g_m = -1.67$ left-skewed distribution (negative skewness)
- $g_{.25} = .33$: right-skewed distribution (positive skewness) respectively $g_{.25} = .07$: symmetrical distribution

Kurtosis (curvature): coefficient I

Prerequisite: metric variables
kurtosis of a random variable X is the 4th order central moment normalised to the standard deviation s formula:

$$g_m = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}^4} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^4 = \frac{m_4}{\widetilde{s}^4} := \underbrace{\frac{m_4}{\widetilde{s}^4} - 3}_{\text{comparision to N()}}$$

$\rightarrow$ Note second term: this is nothing more than a z-standardisation raised to the power of 4 (see from slide 153).
$\rightarrow$ to better assess kurtosis, it is compared with the kurtosis of a normal distribution (last term) - *e.g. implemented like this in SPSS*.

Kurtosis : coefficient II

- $\gamma > 0$: steep peaked (also leptokurtic), distribution more peaked compared to the normal distribution.
- $\gamma < 0$: flat-peaked (also platykurtic), distribution flattened compared to normal distribution
- $\gamma \approx 0$: approximately normally distributed (also mesokurtic)

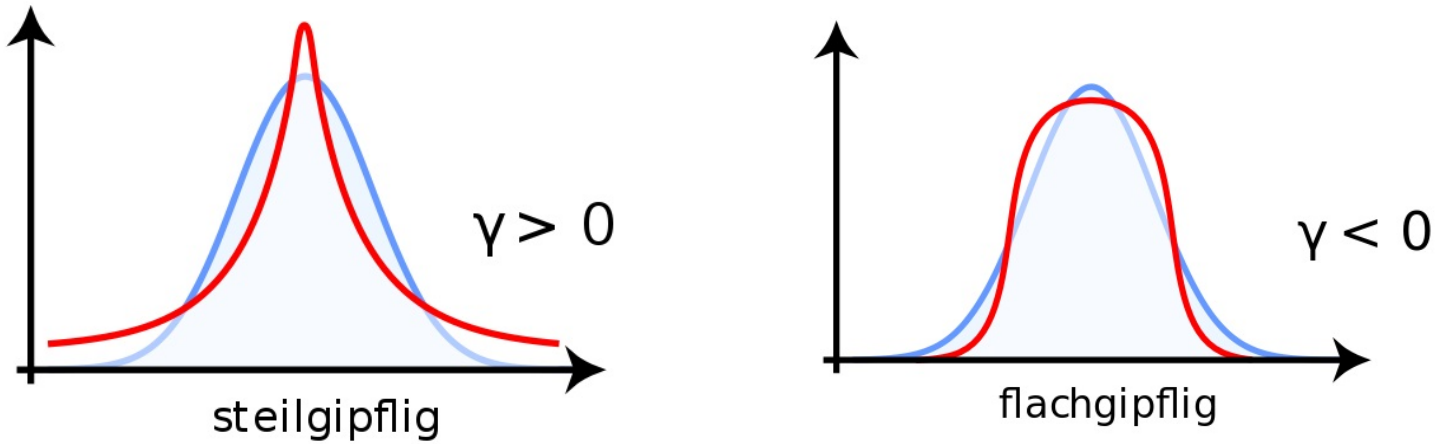


Abbildung: left steep peaked, right flat-peaked

Wikipedia Kurtosis

Take Home Messages: measures

- in dependence on the scale level, different measures can be calculated for individual variables, which are subdivided into:
  - Measures of central tendency (centre of gravity of the data)
  - measures of dispersion (variation in the data from the mean)
  - Measures of skewness / kurtosis (statements about symmetry of a distribution compared to the normal distribution)
- often underestimated is the possibility to calculate quantiles of data (*example income*)

$\rightarrow$ linking level of measurement with measures

R-commands section measures

- computation mean: `mean(x)`
- median: `median(x)`
- mode: `getmode(x)` $\rightarrow$ is not implemented in **R**, so it is necessary to write own function
- quantiles: `quantile(x, probs = ..)`
  - using the argument`probs` you can determine which quantiles are calculated (quartiles by default)
- variance: `var(x)`
- standard deviation: `sd(x)`
- quartile coefficient: `getquartilecoef(x)` $\rightarrow$ not implemented in **R**
- skewness: `moments::skewness(x)`
- kurtosis: `moments::kurtosis(x)`

Symmetry: Examples

- often one encounters certain (partly undesirable) forms of distribution in reality / in surveys:
  - income distributions are almost always **right-skewed** (high incomes run out to the right).
  - questions that are answered in a socially desirable way are often **right-skewed** (question is strongly rejected) or **left-skewed** (question is strongly favoured)
  - standardised tests are often said to have a **symmetrical** distribution, but can be right-skewed if questions are too easy for example
  - ...

$\rightarrow$ in addition to the shape of a distribution, extreme values or outliers can also be detected

Symmetry: basics

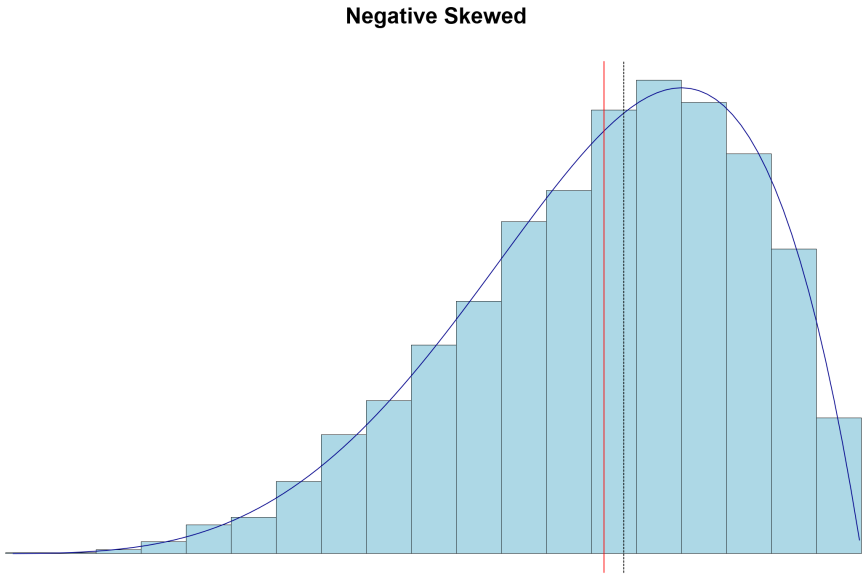
Symmetry or skewness can be assessed via so-called (Fechner's) position rules:

- Symmetrical distribution: $\overline{x} \approx x_{med} \approx x_{mod}$
  - only approximate, since empirical distributions are rarely exactly symmetric
- Left-skewed distribution: $\overline{x} > x_{med} > x_{mod}$
- Right-skewed distribution: $\overline{x} < x_{med} < x_{mod}$

$\rightarrow$ mainly relevant for **unimodal** distributions (see from slide 88)

Symmetry: negative skew (left-skewed)

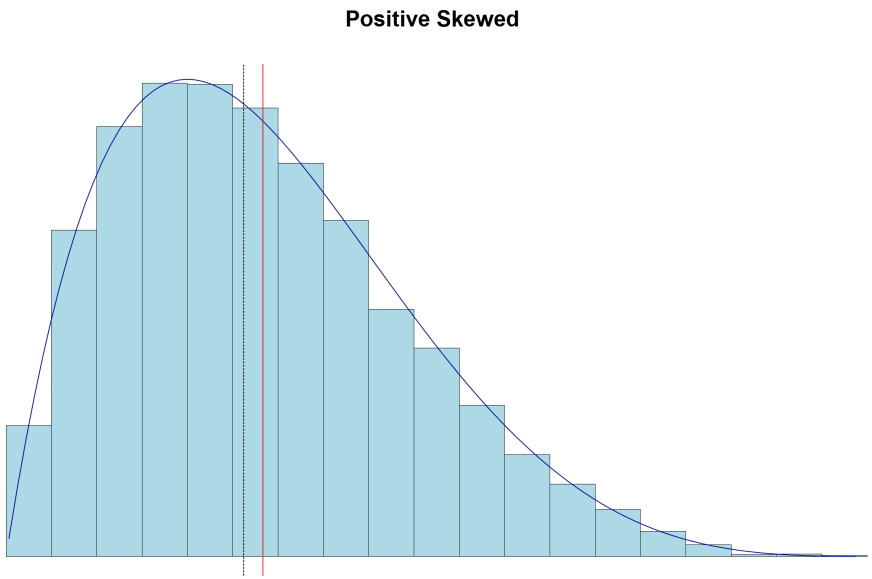
negative skew: skewness coefficient < 0
$\rightarrow$ falls more flat on the left side than on the right side



vertical lines from left to right: mean, median

Symmetry: positive skew (right-skewed)

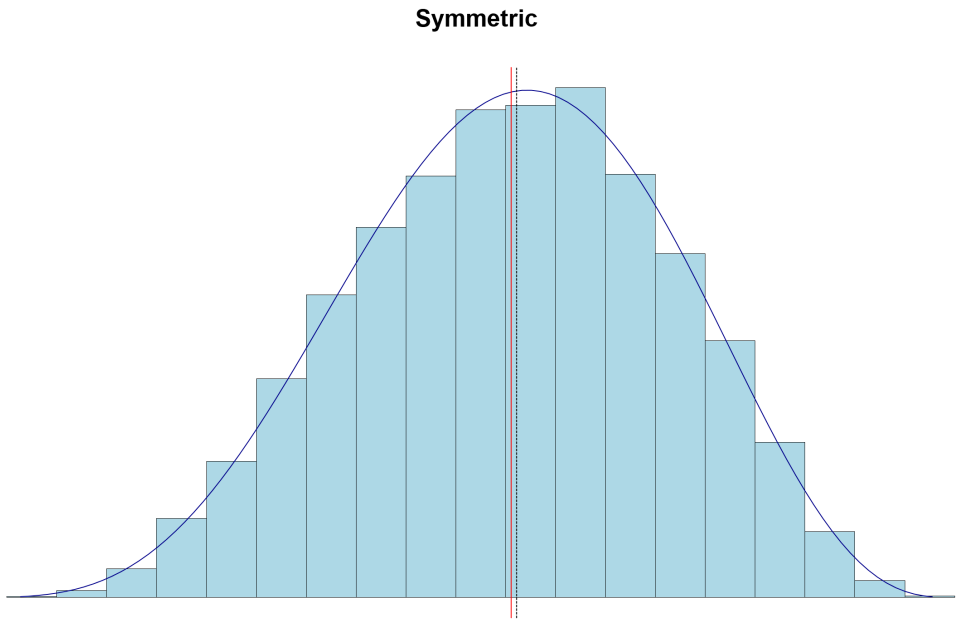
positive skew: skewness coefficient > 0
$\rightarrow$ falls more flat on the right side than on the left side



vertical lines from left to right: median, median

Symmetry: symmetrical distribution

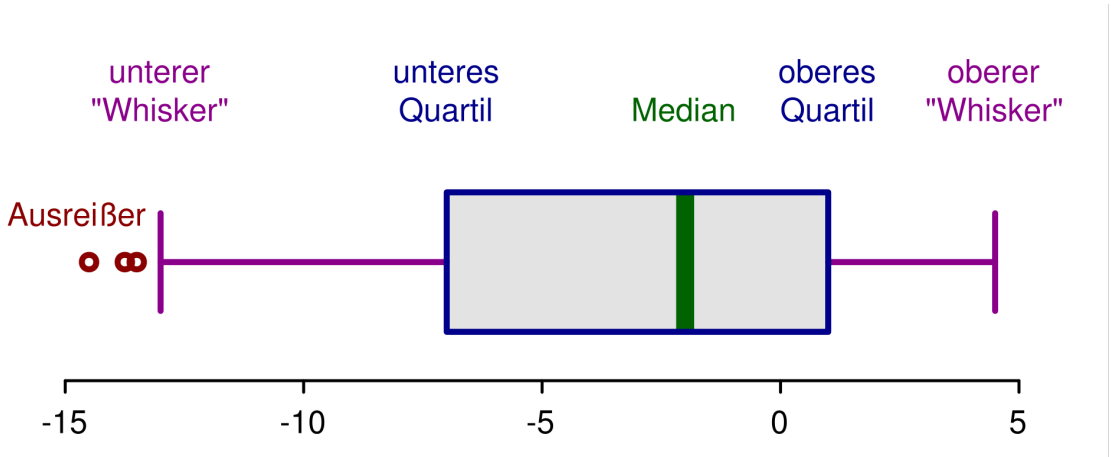
positive skew: skewness coefficient ≈ 0
$\rightarrow$ falls evenly on both sides



vertical lines: median $\approx$ median

Boxplot: 5-point summary

a boxplot is derived from quantiles and visually represents the following measures:: $x_{\text{min}}, x_{0.25}, x_{\text{med}}, x_{0.75}, x_{\text{max}}$
outlier defined as $[x_{0.25} - 1.5 * \text{IQR}, x_{0.75} + 1.5 * \text{IQR}]$



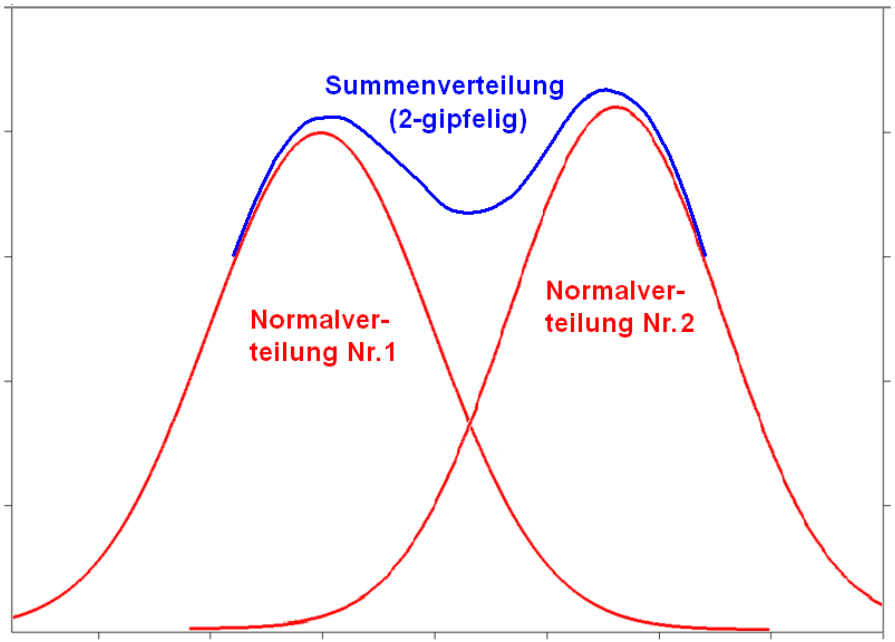
Interpretation:
Median within the box slightly shifted to the right $\rightarrow$ right-sided distribution
several outliers $\rightarrow$ no normal distribution

Wikipedia Box-Plot

Symmetry: peak of a distribution

a distribution may have the following peaks:

- unimodal (single-peaked): distribution with one peak / maximum
- bimodal (two-peaked): has several maxima; as well as multimodal (multi-peaked) distributions
  - → data come from different subpopulations, but are mixed in the representation; subpopulations can be separated by the minimum between the two maxima



Take Home Messages: symmetry

unimodal distributions can have the following possible forms:

- Symmetric distribution
- Left-skewed distribution
- Right-skew distribution

multimodal distributions, on the other hand, are an indication that the observed phenomenon comes from several subpopulations that can (possibly) be separated into unimodal distributions

$\rightarrow$ it is important to theoretically explain properties of symmetry of distributions by possible causes (such as response tendencies, intrinsically non-symmetrical phenomenon, ...)

R-commands section symmetry

- Calculation of the skewness: `moments::skewness(x)`
- Boxplot: `boxplot(x)`; several boxplots next to each other: `boxplot(x ~ y)`, where y is a categorical (discrete) variable
- 5-point summary (several descriptive measures): `summary(x)`

Multivariate descriptive statistics: Basics I

Looking at two or more variables at the same time can answer questions about **relationships**:

- Do the values of one variable tend to go hand in hand with certain values of another variable
  - Example: *People who are taller have higher body weight*
- Measures of correlation are also to be interpreted as measures of difference, example:
  - *Average weight of taller people is higher than average weight of people with medium or short height*

Multivariate descriptive statistics: Basics II

The aim is the representation, description of data consisting of observations of two or more characteristics (variables) (bivariate, two-dimensional data or multivariate, multidimensional data).

$\rightarrow$ also allows comparison of several univariate results to gain impressions about possible dependencies, which can be further analysed with multivariate procedures

bivariate and multivariate representations:

- contingency table
- two-dimensional bar chart
- two-dimensional histogram
- scatter plot
  - pairwise scatterplots (scatterplot matrix)

Contingency table I

Prerequisite: discrete variables with relatively few characteristics (also possible to categorise metric characteristics)
Contingency tables use only the nominal scale level, thus the lowest information content

Example: Are there differences between the sexes on the categorised IQ variable?

Variable IQ.cat	male (1)	female (2)
below average	6	5
medium	2	3
above average	2	2

Contingency table II

		Outcome: lung cancer		
		Yes	No	Total
Exposure: smoked	Yes	75	25	100
	No	10	90	100
	Total	85	115	N

measure of risk; Basic idea frequency table from slide 36

Contingency table III

Contingency tables represent the marginal distributions of two variables X, Y, where the

- marginal frequency of feature X is given by: $h_{i.} = h_{i1} + ... + h_{im}$, where $i = 1, .., k$ *is the marginal frequency of feature* Y *given by:* $h_{.j} = h_{1j} + ... + h_{kj}$, *where* $j = 1, .., m$

$\rightarrow$ Dot notation: $h_{i.} = \sum_j h_{ij}$ (*sum over all columns*)

Example:

Variable IQ.cat	male (1)	female (2)	$h_{i.}$
below average	6	5	11
medium	2	3	5
above average	2	2	4
$h_{.j}$	10	10	20

Contingency table III - conditional frequency

Groups are (almost always) of different sizes, so it requires conditional frequencies:
conditional frequency distribution of Y under the condition $X = a_i$:

$$f_Y(b_1|a_i) = \frac{h_{i1}}{h_{i.}}, .., f_Y(b_m|a_i) = \frac{h_{im}}{h_{i.}}$$

conditional frequency distribution of X under the condition $Y = b_j$:

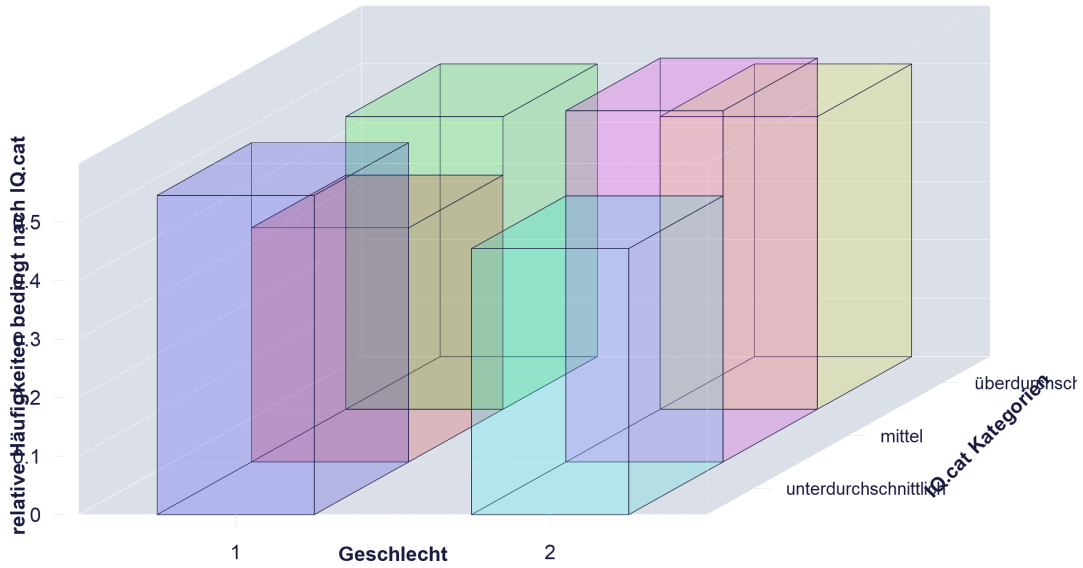
$$f_X(a_1|b_j) = \frac{h_{1j}}{h_{.j}}, .., f_X(a_k|b_j) = \frac{h_{kj}}{h_{.j}}$$

Example of fixed level of variable IQ.cat ($X = a_i$):

Variable IQ.cat	male (1)	female (2)	
below average	.55	.45	1
medium	.4	.6	1
above average	.5	.5	1

Three-dimensional bar chart: example

Example of the calculated conditional frequencies from the previous slide:



Three-dimensional histogram I: construction

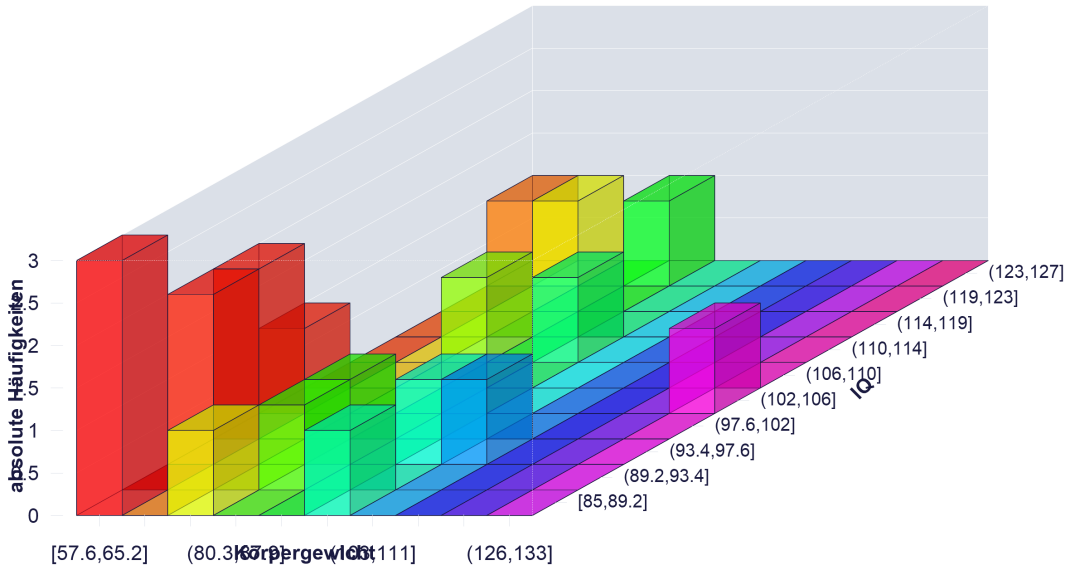
Prerequisite: discrete variables with relatively few expressions
Analogous to the construction of a one-dimensional histogram:

1. data are grouped into rectangle classes $c_o, ..., c_k$, as well as $e_o, ..., e_j$, where these have adjacent intervals: $[c_o, c_1)x[e_o, e_1), ..., [c_{i-1}, c_i)x[e_{j-1}, e_j)$
   - Blocks in the histogram have the base edge $[c_{i-1}, c_i)$ in the x coordinate and $[e_{j-1}, c_j)$ in the y coordinate, respectively.
2. absolute height of the blocks: $\frac{h_{ij}}{(c_i - c_{i-1})(e_j - e_{j-1})}$
   - Note: Term divides the frequencies within the intervals (h_{ij}) by the base area.
   - is done under the principle of area fidelity (volume is proportional to height and thus to absolute frequencies)

$\rightarrow$ same disadvantages as with a one-dimensional histogram (see from slide 39)

Three-dimensional histogram II: example

Representation of the variables IQ x body weight:

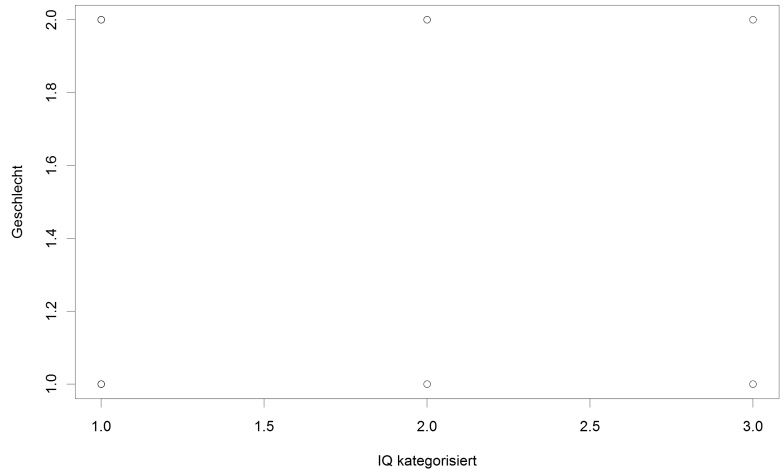


→ interrelationships would be shown by „mountains and valleys“ in the data (data centroids)

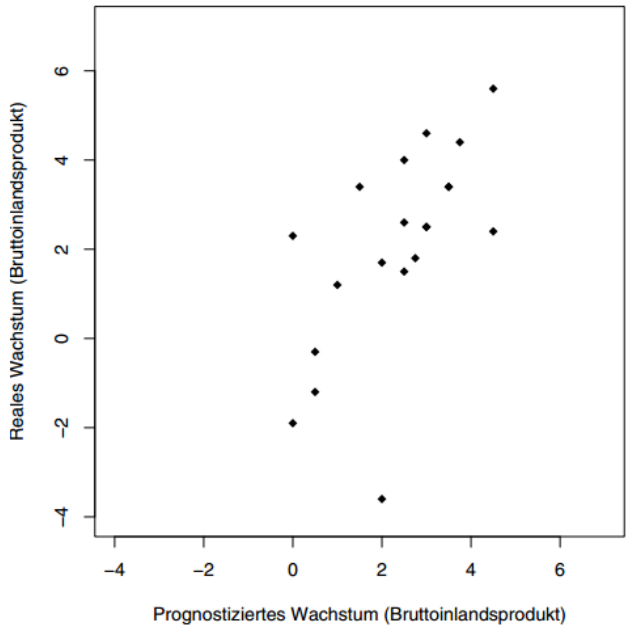
Scatter plot I: construction, example

Prerequisite: metric variables with relatively few identical values. pairs of measurements $(x_1, y_1), ..., (x_n, y_n)$ are represented in an x-y-coordinate system (same index n: need equal number of observations for both variables)

Visualisation scatter plot with multiple identical values:



Scatter plot II: example



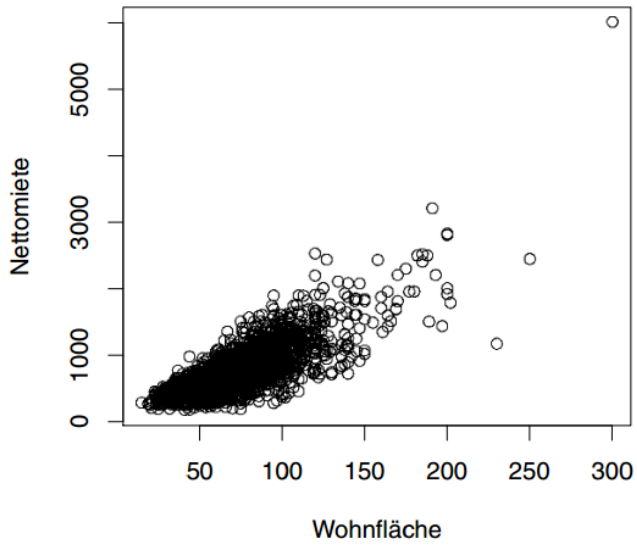


Abbildung: left projected and real growth rate for the Council of Economic Experts' (Sachverständigenrat) forecast in 1975-94, right living space times net rent in Munich

Fahrmeir et al. (2016), 122

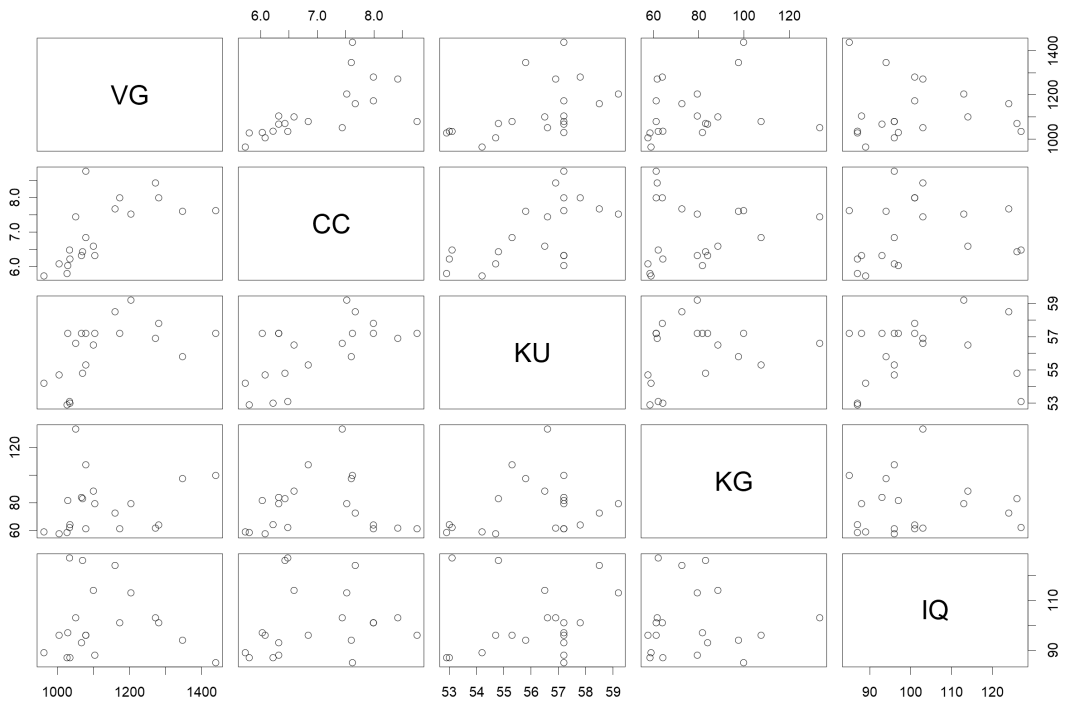
pairwise scatterplots (scatterplot matrix) I: notes

Prerequisite: pairwise metric variables with relatively few identical values

- scatter plots of two variables are formed in pairs
- on the diagonal is the variable name from the data set
- ! Relationships are shown without taking the other variables into account.

pairwise scatterplots (scatterplot matrix) II: example

Plot scatterplot matrix for the variables VG (volume of the forebrain in cm^3), CC (area of the corpus callosum in cm^2), KU (head circumference in cm), KG (body weight in kg), IQ



pairwise scatterplots II: correlation matrix

Correlation matrix (empirical correlation coefficient according to Bravais-Pearson) for the variables VG (volume of the forebrain in cm^3), CC (area of the corpus callosum in cm^2), KU (head circumference in cm), KG (body weight in kg), IQ:

	VG	CC	KU	KG	IQ
VG	1.00	0.66	0.51	0.21	-0.06
CC	0.66	1.00	0.61	0.08	0.16
KU	0.51	0.61	1.00	0.24	0.14
KG	0.21	0.08	0.24	1.00	0.00
IQ	-0.06	0.16	0.14	0.00	1.00

$\rightarrow$ VG, CC, KU are highly correlated with each other, but no (low) correlation with KG and with IQ

Multivariate descriptive statistics: measures

if individual variables show a (linear) correlation, this can be quantified using statistical measures

bivariate (and multivariate) statistical measures:

- for two variables with the same scale level:
  - Empirical correlation coefficient according to Bravais-Pearson
  - Spearman's correlation coefficient
  - Alternative rank correlation measures
- *not mentioned:* there are statistical procedures if two variables have a different scale level:
  - η^2 (Eta-squared)
  - r_{pbis} (point biserial correlation)

Fahrmeir et al. (2016), 126ff.; MatheGuru Korrelation, Korrelationskoeffizient

Empirical correlation coefficient I - formula

Prerequisite: two metric variables

Formula Empirical correlation coefficient according to Bravais-Pearson:

$$r = r_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{\widetilde{s}_{XY}}{\widetilde{s}_X \widetilde{s}_Y}$$

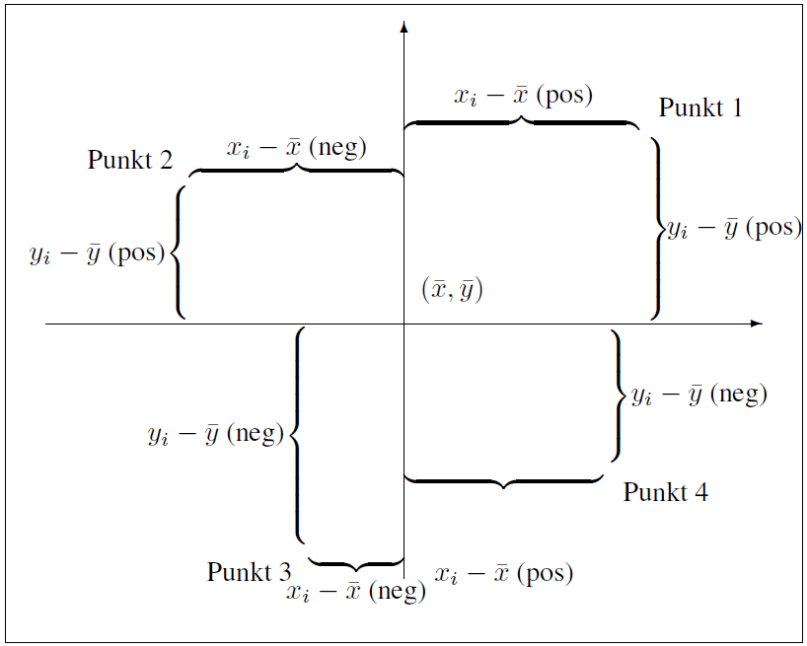
$\rightarrow$ the denominator is used for normalisation[5], range of values between perfectly positive correlation ($r_{XY} = 1$), no correlation ($r_{XY} = 0$) and perfectly negative correlation ($r_{XY} = -1$)

[5] the common variation of the values can be at most as large as the variation of the individual values

Empirical correlation coefficient II - properties

equidistant **linear** connection
Empirical covariance:

$$\widetilde{s}_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$



Fahrmeir et al. (2016), 129

briefly properties covariance

the following properties of the covariance Cov also apply to the correlation:

- symmetry: $Cov(X, Y) = Cov(Y, X)$
- sensitive to outliers
- independence: if two random variables X and Y are independent, then the covariance is $Cov(X, Y) = 0$.
- constant parameter: if the data consists of constant values a in addition to a random variable X, the covariance $Cov(X, a) = 0$.

Relationship to variance:

- $Cov(X, X) = Var(X)$
- with the covariance, in contrast to the variance, the unit of measurement remain
    - , however, the correlation is dimensionless (regardless of the unit of measurement, it takes the values $[-1 \leq r_{SP} \leq 1]$)

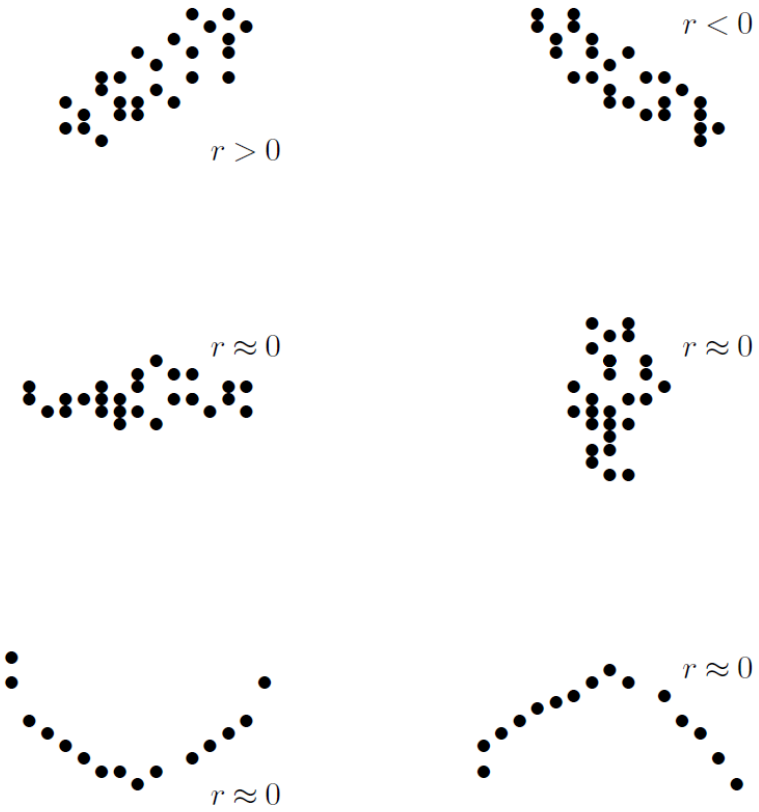
MatheGuru Kovarianz

Empirical correlation coefficient III - properties

equidistant **linear** connection

- clouds of points in the 1st, 3rd quadrant $r_{XY} > 0$, same-sense correlation
  - extreme case $r_{XY} = 1$: all points lie on a straight line with positive slope.
- clouds of points in 2nd, 4th quadrant $r_{XY} < 0$, opposite relation
- point clouds in all quadrants $r_{XY} \approx 0$ similarly strongly distributed

Empirical correlation coefficient IV - examples



$\rightarrow$ Empirical correlation coefficient only maps **linear** correlations

Fahrmeir et al. (2016), 130

Empirischer Korrelationskoeffizient V - Interpretation Stärke

Cohen (1992) suggests the following classification to determine the strength of correlations:

- $r_{XY} = .1$: corresponds to a weak effect
- $r_{XY} = .3$: corresponds to a medium effect
- $r_{XY} = .5$: corresponds to a strong effect

$\rightarrow$ for a linear regression, the coefficient of determination is mathematically equal to the squared empirical correlation coefficient
$R^2 = r_{XY}^2$
$r_{XY}^2 * 100$ indicates how much variance can be explained by the included variables

Cohen (1992)

Spearman's correlation coefficient I - formula

Prerequisite: two ordinal scaled variables

Spearman's correlation coefficient formula:

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

$\rightarrow$ the denominator is used to normalise the range of values: $[-1 \leq r_{SP} \leq 1]$

Example ranks:

x_i	2.17	8	1.09	2.16
$rg(x_i)$	3	4	1	2

Spearman's correlation coefficient II - properties

gleichsinniger **monotoner** Zusammenhang

- $r_{SP} > 0$, if in tendency x large $\rightarrow$ also y large and vice versa, monotonic relationship in the same direction
- $r_{SP} < 0$, if in tendency x large $\rightarrow$ y small and vice versa, opposite monotonic relationship
- $r_{SP} \approx 0$, if no tendency is visible, no monotonic relationship

Spearman's correlation coefficient III - examples

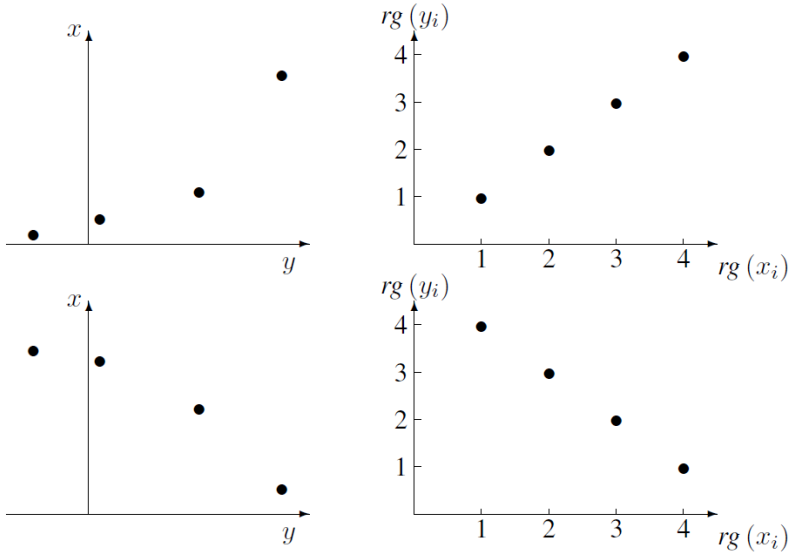


Abbildung: extreme cases for Spearman's correlation coefficients: top $r_{SP} = 1$, bottom $r_{SP} = -1$

$\rightarrow$ Empirical correlation coefficient only maps **monotone** correlations , e.g. for upper two graphs always holds if $x_i < x_j$, then $y_i < y_j$ for any $i \neq j$

Fahrmeir et al. (2016), 135

Alternative rank correlation measures: basics I

in contrast to Spearman's correlation coefficients, only the discrepancy and not the difference between the ranks is considered

- compared to Spearman's correlation coefficients, the values of **concordance measures** are consistently smaller
- to be applied especially when the data are not normally distributed, the scales have unequal answering options, for very small sample sizes

Alternative rank correlation measures: basics II

Concordance measures build on the comparison of observation tuples:

- concordant (C): if $x_i < x_j$ then $y_i < y_j$, values behave in the same way, these are summed up in N_c.
- discordant (D): if $x_i < x_j$ then $y_i > y_j$, values behave in opposite senses, these are summed up in N_d
- bound pairs (T, *ties*) exist in 3 different forms:
  - T_{XY} if $x_i = x_j$ then $y_i = y_j$
  - T_X if $x_i = x_j$ then $y_i \neq y_j$
  - T_Y if $x_i \neq x_j$ then $y_i = y_j$

Concordance measures I - formula

Prerequisite: two ordinal scaled variables

Kendall's τ_α:

$$\tau_\alpha = \frac{N_c - N_d}{n(n-1)/2}$$

Goodman and Kruskal γ:

$$\gamma = \frac{N_c - N_d}{N_c + N_d}$$

$\rightarrow$ the denominator is used to normalise the range of values: $[-1 \leq r_{SP} \leq 1]$

Concordance measures II - example

1. in the first step, the first person (statistical object) is compared with all others that are below it.
2. in the second step the second person is compared with all others that are below it
3. ...

	Vergleiche								
Nr.	1	2	3	4	5	6	7	8	9
1									
2	D								
3	C	C							
4	C	C	C						
5	C	C	C	T_{XY}					
6	C	C	C	C	C				
7	C	C	C	C	C	T_Y			
8	C	C	C	C	C	C	D		
9	C	C	C	C	C	C	C	C	
10	C	C	C	C	C	C	C	C	C

	Summe									
D	1						1			**2**
C	8	8	7	5	5	3	2	2	1	**41**
T_X										**0**
T_Y						1				**1**
T_{XY}				1						**1**

Daten			
X		***Y***	
x_1	2	y_1	1
x_2	1	y_2	2
x_3	3	y_3	3
x_4	4.5	y_4	4.5
x_5	4.5	y_5	4.5
x_6	6	y_6	6.5
x_7	8	y_7	6.5
x_8	7	y_8	8
x_9	9	y_9	9
x_{10}	10	y_{10}	10

Take Home Messages: multivariate descriptive statistics

a large number of questions can be described with the following methods:

Graphs:

- grouped or discrete variables: Contingency tables, three-dimensional bar chart.
- quantitative variables: Scatterplot, three-dimensional histogram, scatterplot matrix (pairwise scatterplots).

measures:

- intervallscaled variables: Empirical correlation coefficient according to Bravais-Pearson.
- ordinal scaled variables: Spearman's correlation coefficient; deviation normal distribution: concordance measures such as Kendall's τ_α

R-commands section multivariate descriptive statistics I

Graphs:

- contingency table: `table(x, y)`
- contingency table conditional frequencies: `prop.table(x, margin = )` , depending on whether `margin = 1` or `margin = 2` is conditioned on variable x or y.
- two-dimensional bar chart: `epade::bar3d.ade(x)`
- two-dimensional histogram: `epade::bar3d.ade(x = gplots::hist2d(x, y)$counts)`
- scatter plot: `plot(x, y)`
- scatter plot matrix (pairwise scatterplots): `pairs(x)`

R-commands section multivariate descriptive statistics II

statistical correlation measures:

- Empirical correlation coefficient according to Bravais-Pearson: `cor(x, y, method = "pearson")`.
- Spearman's correlation coefficient: `cor(x, y, method = "spearman")`
- Alternative rank correlation measures: `cor(x, y, method = "kendall")`

further correlation measures such as r_{pbis} (point biserial correlation) can be found in chapter 6 in Dormann (2013)

Correlation measures - no causality

> Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful
>
> *George Box (1987)*

Correlation measures (generally statistical methods) never (!) indicate causality unless a correlation meets very strong assumptions

→ Important contributions on the topic can be found, for example, in Shadish, Cook und Campbell (2002) or in contributions to complex thinking Elbow (1993); Tsoukas und Hatch (2001)

Causality assumptions

Necessary assumptions to attribute causality to a relationship between two variables X and Y:

- if the effect Y is assumed to be caused by the cause X, the cause must precede the effect ($X \overset{\text{cause}}{\longrightarrow} Y$).
- there is a correlation or observable covariation between the occurrence of X and Y.
- the correlation between X and Y is **isolated**. Thus, there are no other possible causes that cause the correlation of X and Y. The relationship between the two variables is also evident when other variables that may cause Y are statistically controlled for.
- the estimation procedure chosen for the estimation of the relationship between the two variables is correct. The distribution of the observed data is consistent with the distributional assumptions of the estimation procedure.
- The direction of the causal relationship is correctly specified.

Hoyle (2012), 113

stay critical!

Relationship structures between statements on integration and on political culture at an overall level (*correlation matrix)*:

Tab. 3 Einstellungen gegenüber Integration und Haltung zur Demokratie in Europa. (Quelle: Eigene Berechnungen; European Social Surveys 2012–2016)

	Soziales Vertrauen	Wichtigkeit Leben in einer Demokratie[a]	Wie demokratisch ist das eigene Land[a]	Zufriedenheit wie Demokratie im Land arbeitet
Erlaubnis zur Immigration von Mitgliedern der eigenen „Ethnie"/ Mehrheitsgesellschaft	+0,19	+0,16	+0,14	+0,16
Erlaubnis zur Immigration von Mitgliedern anderer „Ethnien" Minderheiten	+0,24	+0,14	+0,14	+0,19
Erlaubnis zur Immigration von Muslimen[b]	+0,23	–	–	+0,24
Einschätzung von Immigration als gut	+0,26	+0,17	+0,18	+0,27
Immigration macht Land weniger lebenswert	–0,27	–0,18	–0,18	–0,28
Immigration unterwandert Kultur des Landes	–0,27	–0,20	–0,24	–0,28

Pearsons Produkt-Moment-Korrelationen über alle Länder der Datensätze (kontrolliert auf Länder)
Alle Skalen 11-Punkt-Antwortvorgaben; $n = 32000–37000$
[a]2012
[b]2014

→ not test for measurement invariance (correlations have the same patterns between countries)

G. Pickel und S. Pickel (2018), 313

stay critical! own calculations

Relationship structures (*correlation matrix)* between statements on integration and on political culture separately for the countries Hungary and Germany:

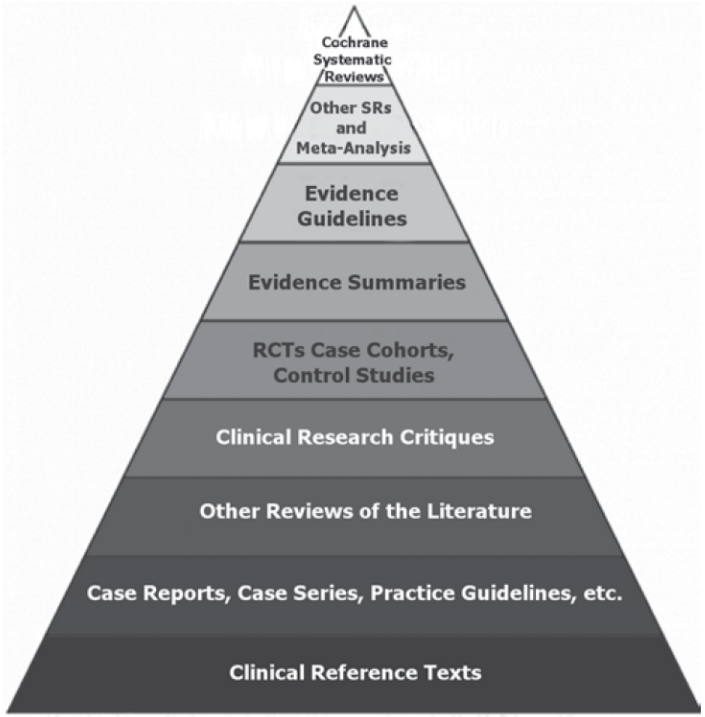
	St H.	Swd H.	St G.	Swd G.
Migration eigener Ethnien	0.13	0.06	0.27	0.28
Migration fremder Ethnien	0.15	-0.03	0.27	0.31
Besser Ort zu leben	0.27	0.14	0.34	0.42
Bereicherung Kultur	0.23	0.06	0.30	0.37

Tabelle: St = Social trust, Swd = Satisfaction with democracy, H. = Hungary, G. = Germany

$\rightarrow$ Relationship is **not stable** between countries

Standards of evidence

Hierarchy of evidence in the context of „evidence-based practice“ in medicine::



→ from the 6th level of the pyramid onwards, studies often have a much lower relevance

Bromme, Prenzel und Jäger (2014), 11ff.

Probability?

Fragen:

- How likely is it that the sun will rise tomorrow?
- What is the probability that it will rain tomorrow?
- How likely is it to roll the number 6 on a fair die?
- Bet on coin tosses: If tails comes up you win 2€, if heads comes up you lose 1€. How much money did you win after 100 rolls?

$\rightarrow$ Probabilities are used to quantify the degree of uncertainty of the occurrence of events

Zucchini et al. (2009), 73ff.; Fahrmeir et al. (2016), 165ff.

Probability concepts I

There are three approaches in everyday life and in science for determining **probabilities**:

- subjective probability: assignment of probabilities based on subjective beliefs
- logical probability (also classical, Laplace probability): logically seen / on the basis of plausible considerations there is nothing to be said against the assignment of probabilities
- frequentist (objective) probability: relative frequency with which a certain **event** occurs in an independent repetition of a **random experiment**

Probability concepts II

- **random experiment**: thought or actual experiment whose outcome (event) cannot be predicted with certainty
- **(outcome) set**: Set of all possible outcomes of a random experiment, denoted by symbol Ω (*Omega*)
  - normal coin: $\Omega = \text{head, number}\}$
  - normal die: $\Omega = \{1, 2, 3, 4, 5, 6\}$
  - flight from Hamburg to Rome: $\Omega = \{0, 1\}$ (0 do not land in Rome, 1 land in Rome)
  - Burning time of a light bulb which is not countable[6] has many outputs: $\Omega = x|x \geq 0$

[6] only possible probabilities with the following formula „$Px_1 \leq \text{result} \leq x_2$“ for any subintervals of the result set, where x_1, x_2 are real numbers

Probability concepts III

- **subsets** of the result set: are denoted with capital Latin letters
  - normal dice: $A = \{\text{Augenzahl ist gerade}\} = \{2, 4, 6\}$
  - burning time of a light bulb:
    $B = \{\text{Brenndauer ist länger als 500 Stunden}\} = \{x | x > 500\}$
- an **empty set** $\emptyset$ is an impossible event
  - normal dice: $\Omega = \{1, 2, 3, 4, 5, 6\}$ has event
    $C = \{\text{number of dice is a 7}\} = \{7\} \rightarrow C \notin \Omega$
- the complete result set Ω is always a **safe event**
  - normal dice: $\Omega = \{1, 2, 3, 4, 5, 6\}$ hat Ereignis
    $D = \{1, 2, 3, 4, 5, 6\} \rightarrow D \in \Omega$, also: $D \subseteq \Omega$

Definition of probability I

probability function is a function P which assigns a real number to all events from Ω and which must satisfy the free following axioms of Kolmogoroff:

A1 $0 \leq P(A) \leq 1$

A2 $P(\Omega) = 1$

A3 wenn $A \cap B = \emptyset$ (disjunkt), gilt $P(A \cup B) = P(A) + P(B)$

$\rightarrow$ what conditions must the function P fulfil in order to represent a probability (also applies to subjective probabilities)

Definition of probability II - example

Are these functions probabilities?

Ereignis	$\emptyset$	{Kopf}	{Zahl}	Ω
Wahrscheinlichkeit	$P(\emptyset)$	$P(\{Kopf\})$	$P(\{Zahl\})$	$P(\Omega)$
Fall 1	0.0	0.5	0.5	1.0
Fall 2	0.0	0.6	0.4	1.0
Fall 3	0.0	0.0	1.0	1.0
Fall 4	0.0	0.5	0.6	1.0
Fall 5	0.1	0.4	0.5	1.0

Verletzung folgender Axiome:

- case 4 violates A2: assuming that $A \cap B = \emptyset$, $P(\text{head}) + P(\text{number}) = 1.0 \neq 1.1$ holds
- case 5 violates A2: an impossible event never has a probability and it should hold $P(\Omega) = P(\Omega) + P(\emptyset)$ since $P(\emptyset) = 0$.

Zucchini et al. (2009), 77

Frequentist (objective) probability

10,000 times the toss was simulated with a coin, this has the result set $\Omega = \{\text{head}, \text{tail}\}$, here we are only interested in the event $A = \{\text{coin shows head}\} = \{\text{head}\}$.

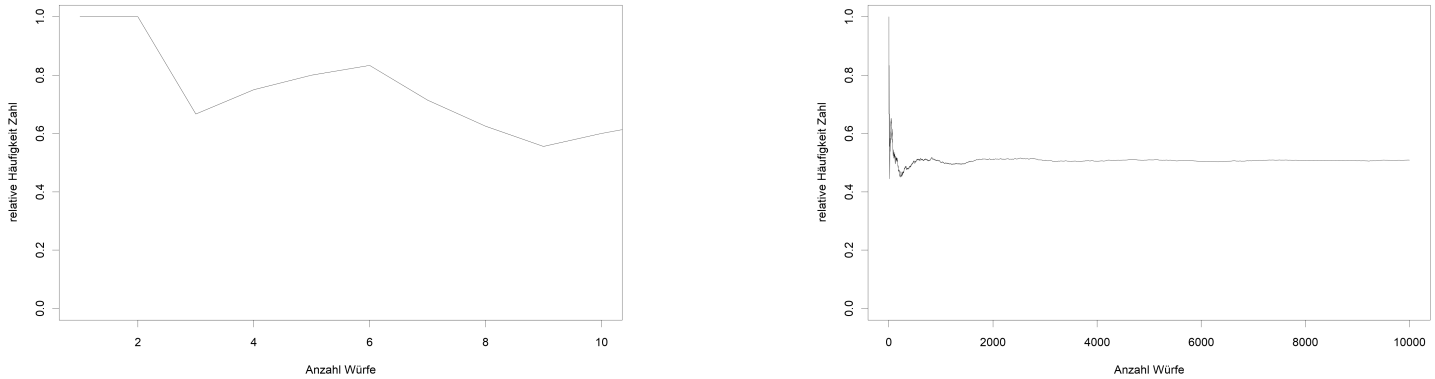


Abbildung: left relative frequency number to throw head at 10 throws, right at 10,000 throws

$\rightarrow$ relative frequency with which the event occurs in 10,000 independent repetitions of a random experiment is as expected at $P(A) = 0.5$

Frequentist (objective) probability: law of large numbers

law of large numbers: the final value of the relative frequency of an event A can be interpreted as the probability of A (relative frequency interpretation of probability)

$$f_n(A) \overset{n\to\infty}{\longrightarrow} P(A)$$

Problem: real random experiments are not symmetrical, i.e. all elementary events do not always have the same probability (one cannot repeat experiments under always the same conditions) $\rightarrow$ no Laplace probability

Random variable I: definition

random variable X is not a single value but has a whole range of possible values

→ behaviour of a random variable X can only be described by probabilities; justification:

- in reality, most situations are **stochastic**, i.e. only probabilities and no certainties can be stated, since correlations are subject to random fluctuations
  - *example: flight duration of the flight connection between Hamburg and Rome*
- if an event is **deterministic**, for each value of variable X, an exact value of variable Y can be determined.
  - *example: period of oscillation of a pendulum in a vacuum*

Zucchini et al. (2009), 35ff., 73ff.

Random variable II: definition

if the desired event (*flight duration of flight connection A between Hamburg and Rome*) occurred, then a **realisation** of the random variable X took place

- before the aircraft has landed, the actual flight duration of flight connection A is unknown.
- after the aircraft has landed safely, the flight duration is a simple number

$\rightarrow$ Probabilities allow the degree of uncertainty to be predicted (more or less)

Random variable III: conditional probability - basic idea

A variety of factors influence the likelihood of loading safely in Rome:

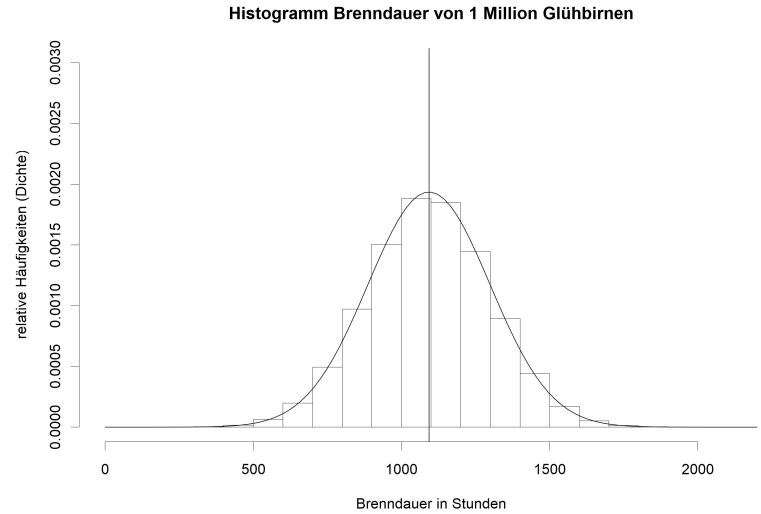
- Random experiment: Flight to Rome.
- Event of interest: I arrive in Rome in one piece.
- Initial assessment: My chance of arriving safely is good (high probability).
- Additional information:
  - One engine appears to be defective.
  - The pilot staggers through the cabin shouting pop songs

$\rightarrow$ new assessment: my chances have become worse (lower conditional probability)

conditional probability: $P(A|B)$, where A is event „land in Rome“ and B is additional information

Random variable: example light bulbs I

You are the operator of a small ghost train at the funfair / kermis and you have to buy 30 light bulbs. The light bulbs of a certain type have the following burning time:



→ Sadly, nothing compared to the centennial light at 4550 East Avenue, Livermore in California, which has been burning since 1901 (see Wikipedia Centennial Light)

Random variable: example light bulbs II

You have now been to two shops and your 30 randomly selected light bulbs have the following burning time:

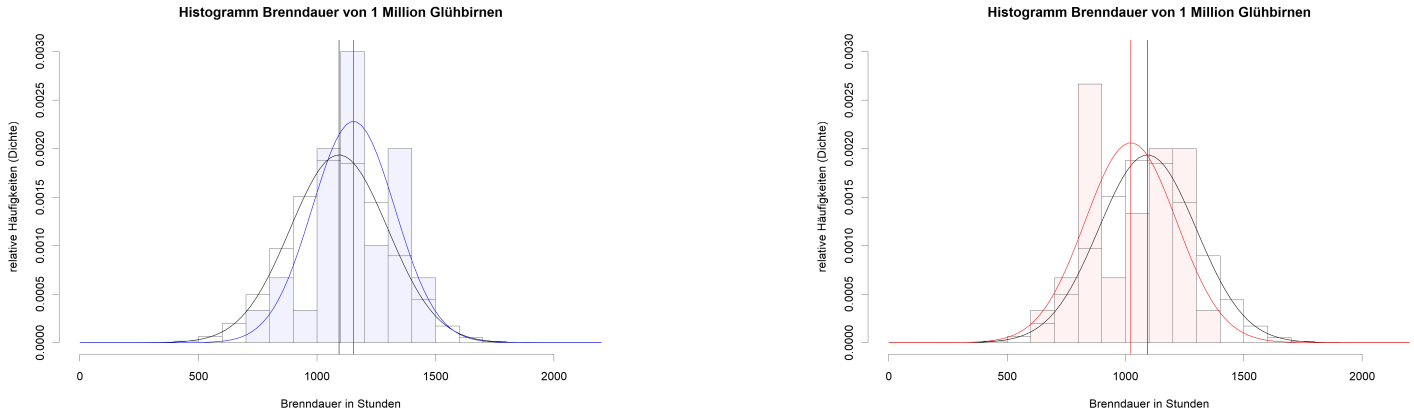
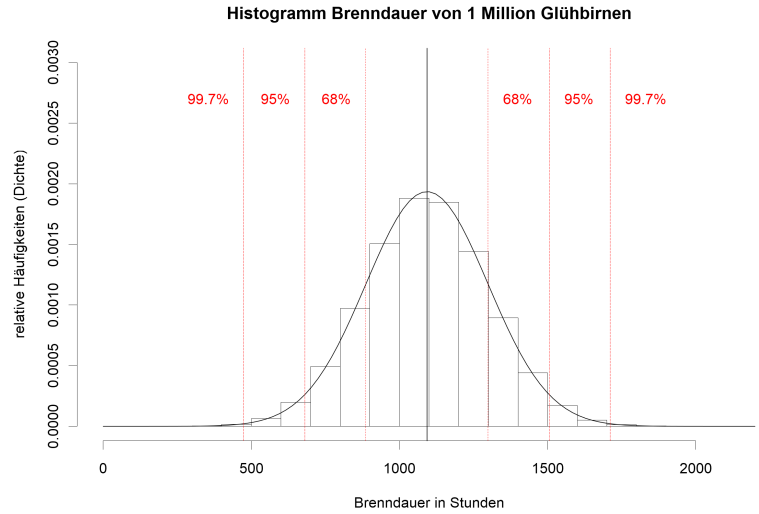


Abbildung: on the left 30 light bulbs in blue, on the right 30 light bulbs in red in comparison to the theoretical purchase of 1 million light bulbs

Where would you buy light bulbs in the future?

Random variable: example light bulbs III

Probabilities can be calculated via the density function*:



this corresponds to the 68-95-99.7 rule of a **normal distribution** ($\mu \hat{=} \overline{x}, \sigma \hat{=} s$:)

- $P(\mu - \sigma \leq x \leq \mu + \sigma) \approx 0.6828$
- $P(\mu - 2\sigma \leq x \leq \mu + 2\sigma) \approx 0.9545$
- $P(\mu - 3\sigma \leq x \leq \mu + 3\sigma) \approx 0.9973$

MatheGuru Normalverteilung

Random variable: example light bulbs VI

Calculation of the 68-rule (further explanation see slide 153):
Normal distribution:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \quad [-\infty < x < \infty]$$

Integral to be calculated for the 68 rule:

$$P(886.804 < X < 1298.961) = \int_{886.804}^{1298.961} \frac{1}{206.1\sqrt{2\pi}} e^{\frac{-(x-1092.9)^2}{2206.1^2}} dx = 0.6828$$

basic step-by-step calculation:

1. calculate root function
2. determine individual integrals and subtract them from each other

→ Calculation for Normal Distribution youtube: Integrating Normal Density Function

Take Home Messages: Probabilities

- many processes in reality are stochastic, thus they are (more or less) random processes.
- random processes are described by probabilities
- probabilities serve to quantify the degree of uncertainty of the occurrence of events and thus prevent surprises
- conditional probabilities describe the change of probabilities when prior knowledge is known

Distributions: describe random processes I

the aim is the description of results (events) of random processes by numbers

- numbers are here the characteristic value of a variable
  - Variable X whose characteristic x is the result of a random process is called random variable X, $x \in \mathbb{R}$ is the realization of X

$\rightarrow$ behaviour of X can only be described by probabilities

example of discrete random variables Y, Z:

$$Y = \text{Sum of two dice}$$

$$Z = \begin{cases} 1 & \text{if sum of two dice} = 7 \\ 0 & \text{else} \end{cases}$$

$\rightarrow$ in the example, the underlying random process is not of primary interest or cannot always be described, otherwise the value range of the variable is directly considered as the result space

Distributions: describe random processes II

Random variable is defined by assigning numbers to an event of a random process (realisation), for which **distributions** are necessary:

- it is not desirable or feasible to assign a probability to every possible event
- Random processes can only be described by functional relations, types of representation:
  - for discrete random variables $\rightarrow$ probability function
  - for continuous random variables $\rightarrow$ density function

$\rightarrow$ Distributions are determined by functions and their parameters

A random variable $f(X)$ is a function that assigns a real number to each possible outcome of a random experiment / process.

Example discrete random variables

Random variables can only take on finitely many values or countably infinitely many values
$\rightarrow$ every possible event can be assigned a natural number

Example: rolling 6 or 30 fair dices several times at once and we want to know $A = \{\text{result is } 1\} = \{1\} \rightarrow P(\{1\}) = 1/6$ (binomial distributed):

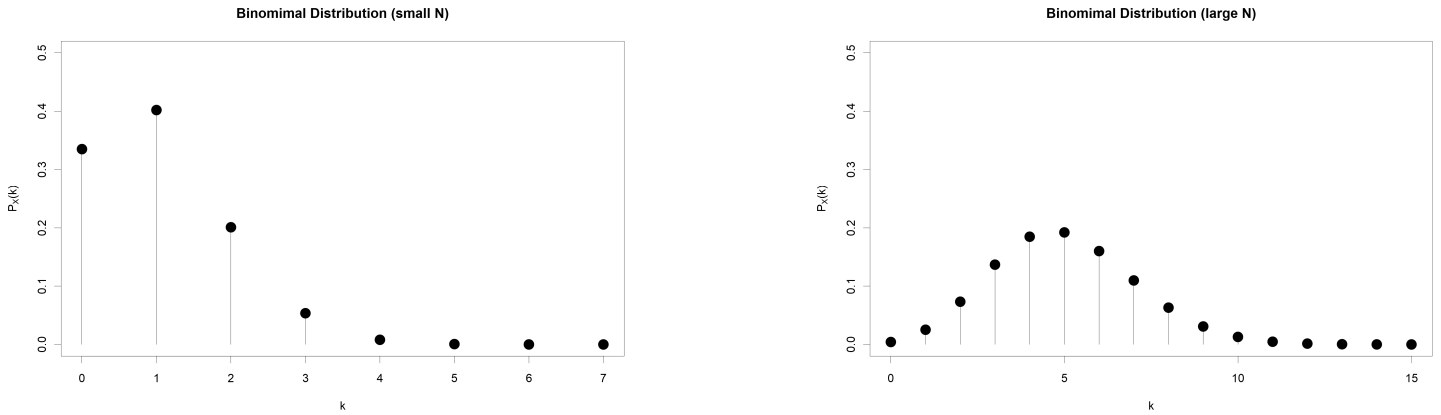


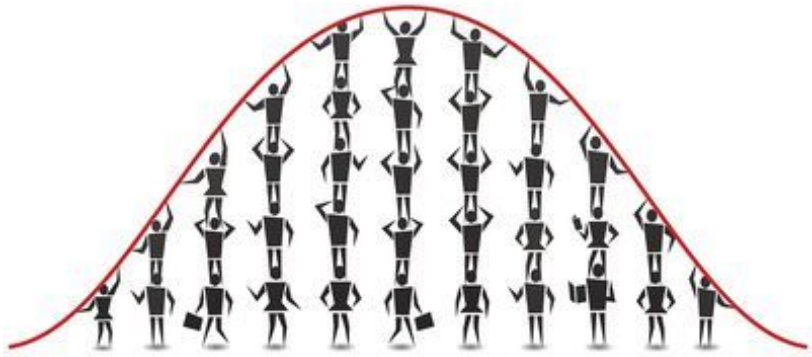
Abbildung: left rolling 6 dices, right 30 dices

Is rolling a dice a Gauss distribution?

Example continuous random variables

Random variables can take on an infinite number of values
$\rightarrow$ no probability can be assigned to an possible single event, e.g.
Probability that a machine will fail after 7.483428383... days is zero
(real numbers)

Example body size of a specific population (normal distribution):



9 Real Life Examples Of Normal Distribution

Which distributions exist?

Probability distributions that are significant in theory or in practical application have specific names, (non-exhaustive) list of distributions:

- List of probability distributions

→ List of convolutions of probability distributions

→ Flowchart of relationships between probability distributions (last page)

Distribution of discrete random variables

Random variable X is called discrete if it can only take on finitely or countably infinitely many values $x_1, x_2, ..., x_k$.

determined by probabilities:

$$P(X = x_i) = p_i, \quad i = 1, 2, .., k$$

Probability function:

$$f(x) = \begin{cases} P(X = x_i) = p_i & x = x_i \in \{x_1, x_2, .., x_k, ..\} \\ 0 & \text{else} \end{cases}$$

Example: 2 dice were thrown and the event of interest was „The sum of the eyes of both dice is at most equal to 4“:

$$P(X \leq 4) = \{(1,1), (1,2), (1,3), (2,1), (2,2), (3,1)\}$$
$$= P(1,1) + .. + P(3,1) = \frac{6}{36} = \frac{1}{6}$$

Binomial distribution I

$$P(X = x) = \binom{n}{x} \pi^x (1-\pi)^{n-x}, \quad x = 0, 1, ..., n$$

- parameter π constant probability of success of each run n.
- number n: independent repetitions of a Bernoulli experiment

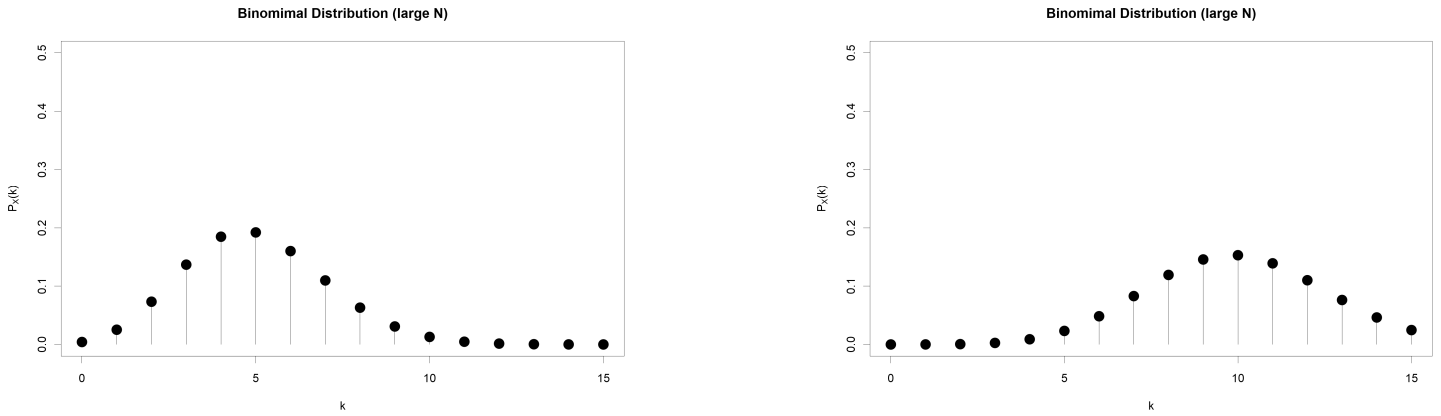


Abbildung: left $\pi = 1/6$, right $\pi = 1/3$ with 30 dice each

Binomial distribution II

Discrete distributions have formally very similar properties to empirical distributions:

Binomial distribution: Expected value:

$$E(X) = E(X_1) + .. + E(X_n) = n\pi$$

Variance:

$$Var(X) = Var(X_1) + .. + Var(X_n) = n\pi(1 - \pi)$$

$\rightarrow$ expected value $E(X)$ is analogous to the arithmetic mean $\bar{x}$
Interpretation: if a random experiment is repeated very often, the arithmetic mean approaches the expected value

Distribution of continuous random variables

Random variable X is called continuous if it can take on an infinite number of values $x_1, x_2, ..., x_k$ (is continuous)

Probabilities are determined by the area $[a, b]$ and the overlying function $f(x)$:

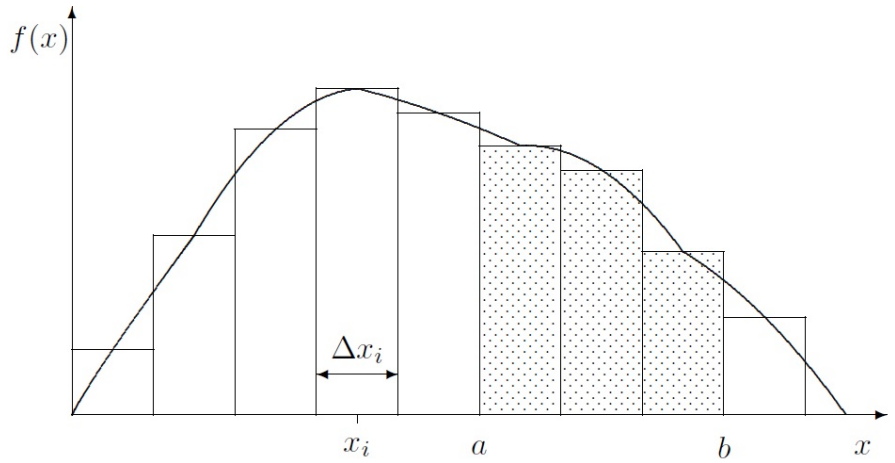
$$P(a < X < b) = \int_a^b f(x)dx$$

Function $f(x)$ is called (probability) density of X

$\rightarrow$ for continuous variables, the x-values in $[a, b]$ are no longer countable and a single value x is $P(X = x) = 0$, since the limits $a = b$ would be

briefly density

following is an approximating probability histogram, where the rectangular areas over $\triangle x_i$ can be calculated using $P(X_d = x_i) = f(x_i)\triangle x_i$.
$\rightarrow$ If the class widths $\triangle x_i$ approach zero, X_d becomes a continuous random variable



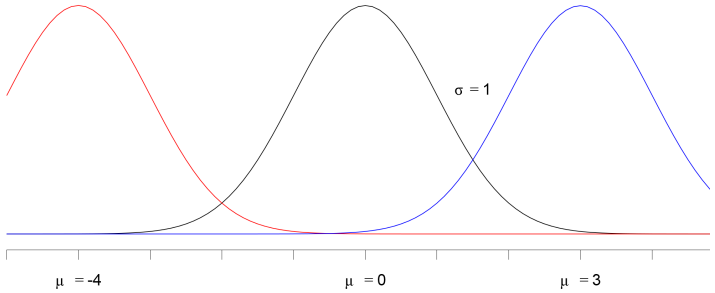
Fahrmeir et al. (2016), 253

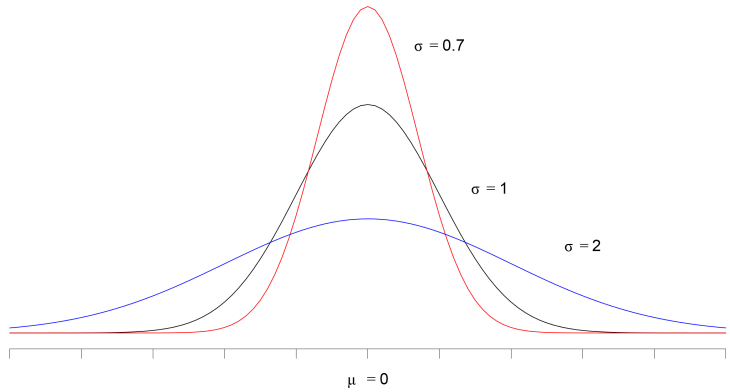
Normal distribution I

Normal distribution (also Gaussian bell curve):

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \quad [-\infty < x < \infty]$$

- parameter μ (left figure): maximum of the normal distribution.
- parameter σ (right figure): dispersion of the normal distribution





Normal distribution II

a normally distributed random variable X can be standardised as follows:

$$Z = \frac{X - \mu}{\sigma}$$

Standard normal distribution $X \sim N(\mu = 0, \sigma = 1)$:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} \quad [-\infty < x < \infty]$$

Standardisation of normally distributed random variables has multiple advantages: Comparison of variables of different metrics, $Cov(X, Y) = r_{xy}$, in a regression coefficients directly correspond to effect sizes, ...

Take Home Message: Distributions

- Random variables describe the outcome of a random experiment by numbers
- the most important continuous distribution is the normal distribution
- Note: Binomial distribution can be approximated by the normal distribution under the following conditions (*limiting form*)
  - number of independent trials goes towards infinity $n \to \infty$
  - probability of success π does not go towards 0 or 1

Literature

Bromme, Rainer, Manfred Prenzel und Michael Jäger (2014). „Empirische Bildungsforschung und evidenzbasierte Bildungspolitik“. In: *Zeitschrift für Erziehungswissenschaft* 17.4, S. 3–54.

Cohen, Jacob (1992). „A Power Primer.“. In: *Psychological Bulletin* 112.1, S. 155–159.

Dormann, Carsten (2013). *Parametrische Statistik Verteilungen, maximum likelihood und GLM in R*. Springer Spektrum.

Elbow, Peter (1993). „The uses of binary thinking“. In: *Journal of Advanced Composition*, S. 51–78.

Fahrmeir, Ludwig et al. (2016). *Statistik: Der Weg zur Datenanalyse*. Springer-Verlag.

Hoyle, Rick (2012). *Handbook of structural equation modeling*. Guilford Press.

Kaldor, Mary (2018). „Cycles in world politics“. In: *International Studies Review* 20.2, S. 214–222.

Kelle, Udo (2008). *Die Integration qualitativer und quantitativer Methoden in der empirischen Sozialforschung*. Springer.

Leeuw, Edith, Joop Hox und Don Dillman (2012). *International handbook of survey methodology*. Routledge.

Lintorf, Katrin (2012). *Wie vorhersagbar sind Grundschulnoten? Prädiktionskraft individueller und kontextspezifischer Merkmale*. Springer-Verlag.

Lütkepohl, Helmut und Markus Krätzig (2004). *Applied time series econometrics*. Cambridge University Press.

Macho, Siegfried (2016). „Moderne Testtheorie“. Université de Fribourg.

Neumann, Astrid (2007). *Briefe schreiben in Klasse 9 und 11. Beurteilungskriterien, Messungen, Textstrukturen und Schülerleistungen*. Waxmann Verlag.

O'Connell, Michael (2019). „Is the impact of SES on educational performance overestimated? Evidence from the PISA survey“. In: *Intelligence* 75, S. 41–47.

Pickel, Gert und Susanne Pickel (2018). „Migration als Gefahr für die politische Kultur?“ In: *Zeitschrift für Vergleichende Politikwissenschaft* 12.1, S. 297–320.

Shadish, William R, Thomas D Cook und Donald Thomas Campbell (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Wadsworth Cengage Learning.

Tsoukas, Haridimos und Mary Hatch (2001). „Complex thinking, complex practice: The case for a narrative approach to organizational complexity“. In: *Human relations* 54.8, S. 979–1013.

Wolf, Sabrina (2008). „Der Methodenstreit quantitativer und qualitativer Sozialforschung unter besonderer Berücksichtigung der grundlegenden Unterschiede beider Forschungstraditionen“. Bachelor Thesis. Universität Augsburg.

Zucchini, Walter et al. (2009). *Statistik für Bachelor-und Masterstudenten: eine Einführung für Wirtschafts-und Sozialwissenschaftler*. Springer-Verlag.