



Data Analytics Onboarding

Part (1): Basic Introduction

Deck I: What is Data?

Gabriel Malli, Sonja Radkohl



drahtwarenhandlung
film & animation
datenjournalismus





The SEVA Project

“Self-Explanatory Visual Analytics for Data-Driven Insight Discovery”

The aim of the project is to develop suitable **onboarding methods for data journalists**. Our vision is to have self-explanatory tools which actively support users in interpreting visualizations and analysis methods.

This slide deck is dedicated to **data analytics onboarding** to help aspiring data journalists take their **first steps into data analysis**.





Content of Deck I

- I.1 What is data?
- I.2 What is data journalism?
- I.3 Where do I get data from?



1.1 What is data?

The term data is defined in different ways. For our purposes, we use the following definition:

Data are numerical values...

- ...which are supposed to represent certain sections of reality
- ...which are generated or collected by certain methods (e.g. measurements, observations, surveys)
- ...which can be analyzed by statistical methods

(cf. Prietl/Houben 2018)





1.1 What is data?

Data is collected for a variety of purposes, including

→ *Scientific analyses*

→ *political considerations (e.g., as a basis for decision-making)*

→ *Commercial reasons*



Which sections of reality are converted into data form depends on human decisions and interests. Data are therefore never completely objective.



I.1 What is data?

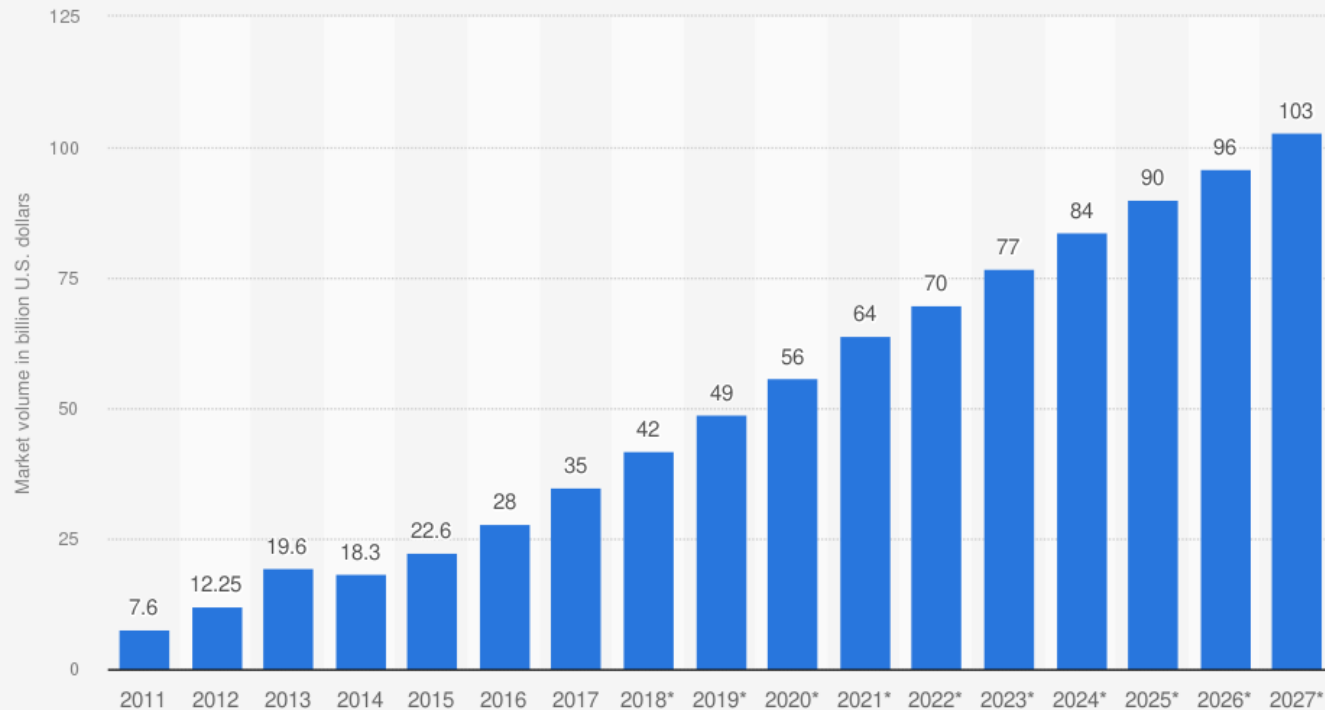
- Historically: Nation states with supremacy over data production (national statistics, censuses...)
- Present: Private companies as major players in data generation
- Internet as a central driver of a **datification** of society
 - Increase of number-based environments
 - Users generate data themselves
 - Automated data collection happens with less human intervention

(cf. Prietl/Houben 2018)



I.1 What is data?

Big data market size revenue forecast worldwide from 2011 to 2027 (in billion U.S. dollars)



Sources
Wikibon; SiliconANGLE
© Statista 2023

Additional Information:
Worldwide; Wikibon; 2014 to 2018

(SiliconANGLE 2018)

(Big) Data as a growing commercial factor:

→ *Product optimization*

→ *Targeted advertising*

(Cf. Zuboff 2019 for a critical discussion)





1.2 What is data journalism?

Data journalism is a journalist genre...

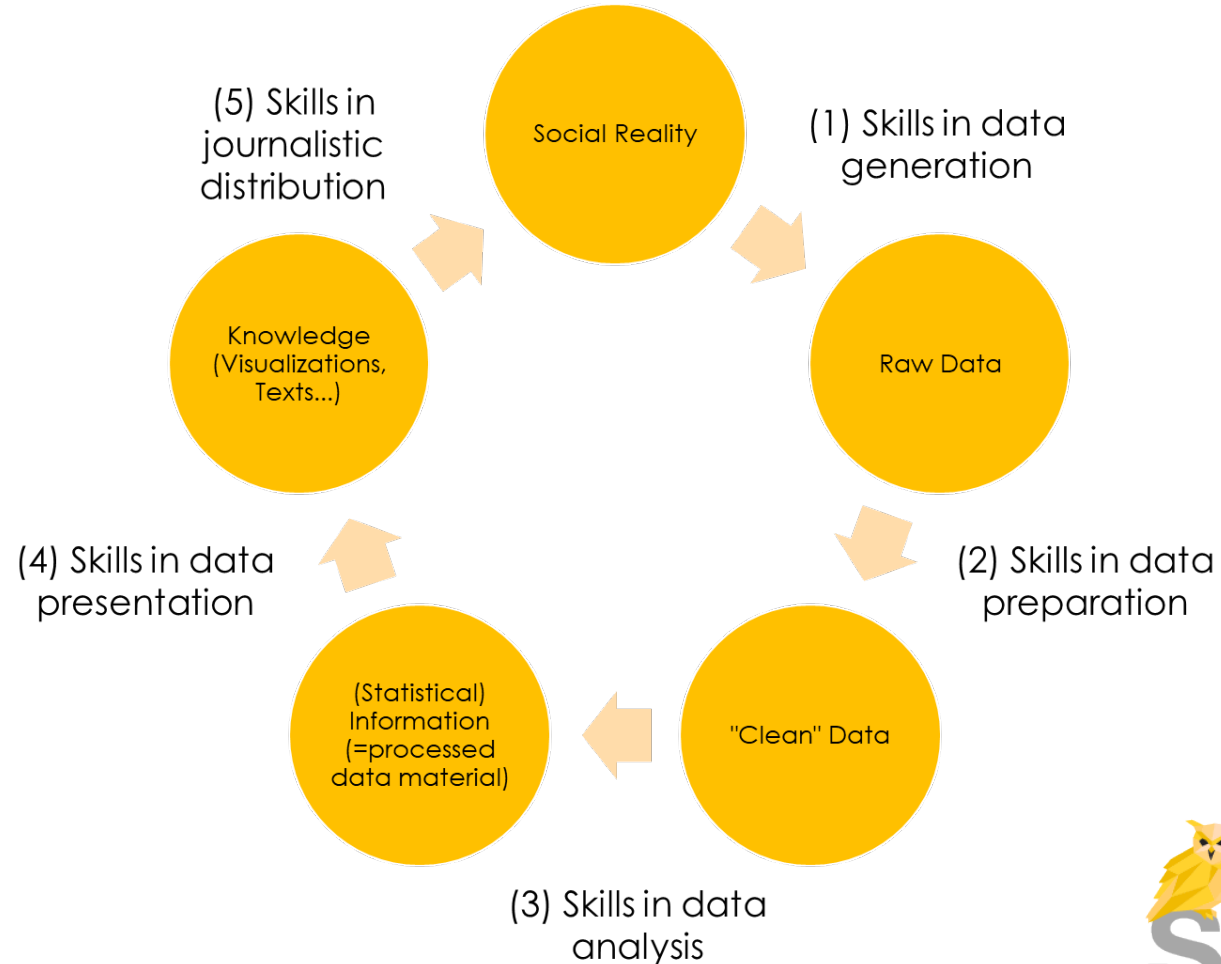
- ...that tells journalistic stories based on numerical data
- ...that makes data stories accessible to a broad audience through statistical processing and visualization

(cf. Bradshaw 2023)



1.2 What is data journalism?

Skills in the circuit of
Data Journalism (cf.
Malli et al. 2023)





1.2 What is data journalism?

Examples for excellent data-based stories from 2023

- [Detroit segregation wall still stands, a stark reminder of racial divisions \(nbcnews.com\)](#) [USA]
- [See where water is scarcest in the world — and why we need to conserve - Washington Post](#) [USA]
- [Muss der Bundestag diverser werden? | ZDFheute](#) [Germany]





1.3 Where do I get data from?

Raw Data from governments and national statistical agencies

- [Find open data - data.gov.uk](https://data.gov.uk) [UK]
- [GovData | Datenportal für Deutschland – GovData](https://www.govdata.de) [Germany]
- [open.data von Statistik Austria](https://open.data.gov.at) [Austria]

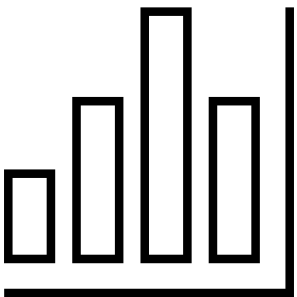




1.3 Where do I get data from?

Raw Data from scientific surveys

- [Home | European Social Survey](#)
- [Database for scientific data: DANS | Centre of expertise & repository for research data \(know.nl\)](#)





1.3 Where do I get data from?

Create your own surveys

- [LimeSurvey — Free Online Survey Tool](#)





1.3 Where do I get data from?

Data mining

- Automated extraction of meaningful information from large amounts of data
- Possible area of application: Data from social media

! Overview of free tools for mining social media data:
[Social Media Analysis — SAGE Ocean | Supporting Social Scientists Working with Big Data & Tech \(sagepub.com\)](#)





Literature

- Bradshaw, Paul. 2023. The online journalism handbook: skills to survive and thrive in the digital age. Third edition. Abingdon, Oxon ; New York, NY: Routledge.
- Malli, Gabriel, Radkohl, Sonja, Goldgruber, Eva. 2023. Pragmatic Data Craft – Conceptions of skillful data journalism between journalist values, scientific approaches, and economic boundaries. Proceedings of the 21th Annual STS Conference Graz 2023 „Critical Issues in Science, Technology and Society Studies“.
- Prietl, Bianca, and Daniel Houben. 2018. „Einführung. Soziologische Perspektiven auf die Datafizierung der Gesellschaft“. S. 7–32 in Digitale Gesellschaft. Bd. 17, herausgegeben von D. Houben und B. Prietl. Bielefeld, Germany: transcript Verlag.
- SiliconANGLE. (March 9, 2018). Big data market size revenue forecast worldwide from 2011 to 2027 (in billion U.S. dollars) [Graph]. In Statista. Retrieved December 11, 2023, from <https://www.statista.com/statistics/254266/global-big-data-market-forecast/>



Data Analytics Onboarding

Part (1): Basic Introduction
Deck II: What is Statistics?
Gabriel Malli, Sonja Radkohl



drahtwarenhandlung
film & animation
datenjournalismus





Content of Deck II

- II.1 What is statistics?
- II.2 Variables and scale levels of measurement
- II.3 Statistical Tools



II.1 What is statistics?

Statistics

- Set of quantitative mathematical methods for describing and analysing empirical findings on "mass phenomena"
- Originates in the description of the population of a state, but can be applied to a wide variety of areas

(Kamps 2018)

[Statistik • Definition | Gabler
Wirtschaftslexikon](#)





II.1 What is statistics?

Main areas
of statistics

Focus of this
course

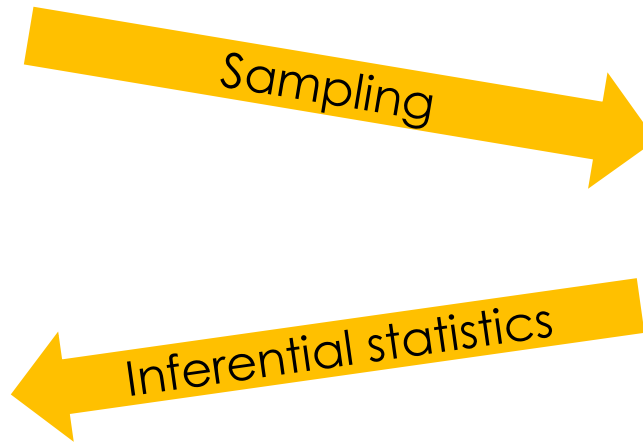
Descriptive
statistics

Inferential
statistics

II.1 What is statistics?



Population



Descriptive statistics

(cf. DATAtab Team 2023a)



II.1 What is statistics?

Descriptive statistics

- ...provides tools to describe and illustrate a sample
- ...helps to get an overview of the distribution of data

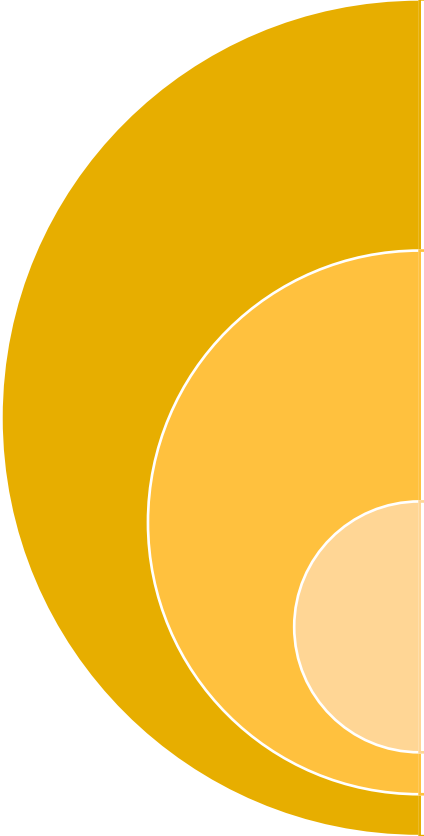
(cf. DATAtab Team 2023a)





II.1 What is statistics?

Tools of descriptive statistics



Location parameters
(Part (2) – Deck III)

- Mean
- Median
- Modal value

Dispersion parameters
(Part (2) – Deck IV)

- Standard deviation
- Variance
- Range

Visualization forms
(Part (2) – Deck V)

- Tables
- Charts



II.1 What is statistics?

Inferential statistics

- ...tests statements about the population based on the characteristics of a sample
- ...uses statistical instruments to test hypotheses

II.1 What is statistics?

Tools of inferential statistics (among others)

Simple test procedures

- t-test
- Chi-square-test
- ...

Correlation analysis

- Pearson correlation analysis
- Spearman rank correlation
- ...

Regression analysis

- Simple linear regression
- Multiple regression

Detailed further information on these and other methods can be found at [First steps with DATAtab](#)





11.2 Variables and scale levels

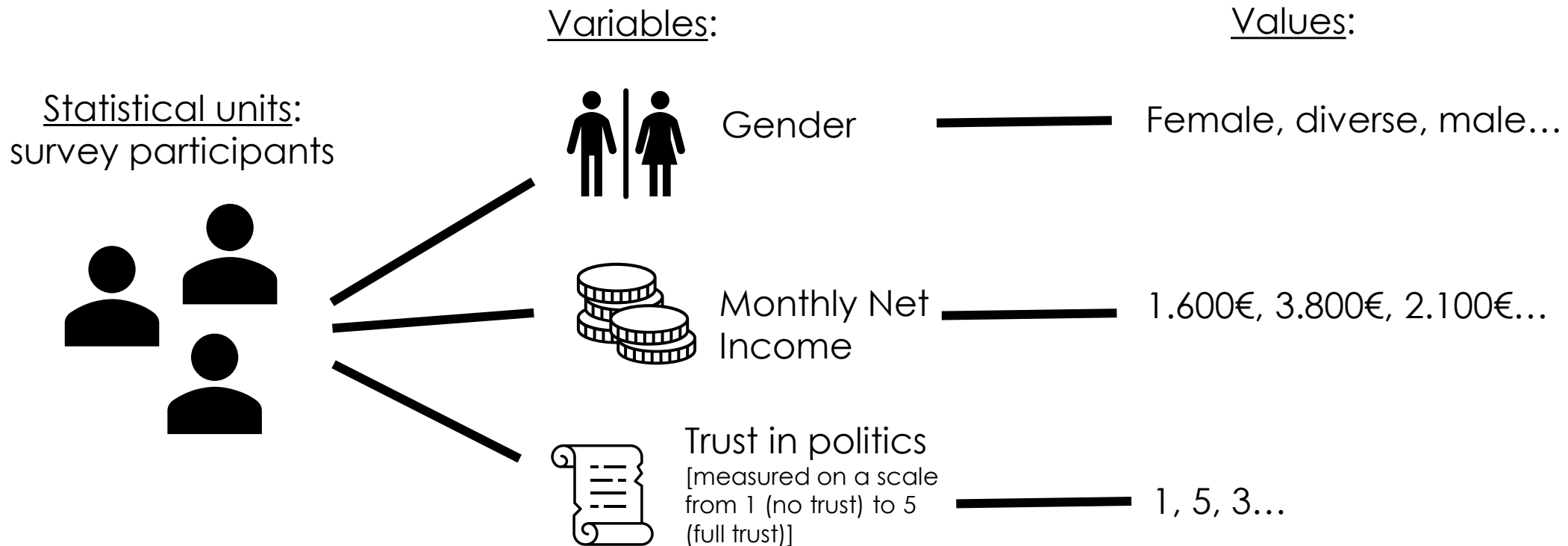
Variables

- ...are the "stuff" from which statistical statements are made
- ...are characteristics of a statistical unit from the sample that are analysed statistically
- ...can take on different values

(cf. Statista n.d.)

11.2 Variables and scale levels

Example: For a data journalism story, a journalist wants to analyse the connections between gender, income and trust in politics of people in Austria. For this purpose, he works with a survey dataset, which has collected these data for a sample.





11.2 Variables and scale levels

Scale levels of measurement

- Properties of variables that specify which statistical operations can be applied



nominal

ordinal

metric

(cf. DATAtab Team 2023b)





11.2 Variables and scale levels

Nominal level of measurement

- No logical ranking between values possible
- Only relations of “equal” or “inequal” possible



Example: Gender

1 = male
2 = female
3 = inter
4 = gender fluid

1 ≠ 3; 2 ≠ 4

11.2 Variables and scale levels

Ordinal level of measurement

- Logical ranking of values possible
- Meaningful relations of “greater” and “smaller” are possible



Example: Trust in politics

1 = no trust
2 = low trust
3 = moderate trust
4 = high trust
5 = full trust

$1 < 4 < 5$



11.2 Variables and scale levels

Metric level of measurement

- Logical ranking of values possible
- Meaningful differences and sums can be calculated from values



Example: Monthly net income

3.200€

1.600€

1.200€

2.300€

4.100€

1.200

4.100





11.3 Statistical Tools

- Microsoft Excel
- IBM SPSS Statistics: [IBM SPSS Statistics](#)
- PSPP: [PSPP - GNU Project - Free Software Foundation](#)
[free replacement]
- R project: [R: The R Project for Statistical Computing \(r-project.org\)](#) [free]
- DATAtab: [Online Statistik Rechner: t-Test, Chi-Quadrat, Regression, Korrelation, Varianzanalyse \(datatab.de\)](#)
[free web-based environment]





Literature

- DATAtab Team (2023a): Descriptive and inferential statistics, [Descriptive and Inference Statistics • Simply explained – DATAtab](#)
- DATAtab Team (2023b): Level of measurement, [Level of measurement • Simply explained - DATAtab](#)
- Kamps, U. (2018): Statistik. In Gabler Wirtschaftslexikon Online, <https://wirtschaftslexikon.gabler.de/definition/statistik-45267/version-268564>
- Statista (n.d.): Definition Statistik für Anfänger - die Variable, [Statistik für Anfänger - die Variable | Statista](#)



Data Analytics Onboarding

Part (2): Descriptive Statistics

Deck III: Location parameters

Gabriel Malli, Sonja Radkohl



FH | JOANNEUM
University of Applied Sciences

datavisyn



drahtwarenhandlung
film & animation
datenjournalismus





Content of Deck III

- III.1 What are location parameters?
- III.2 Mode
- III.3 Median
- III.4 Arithmetic Mean



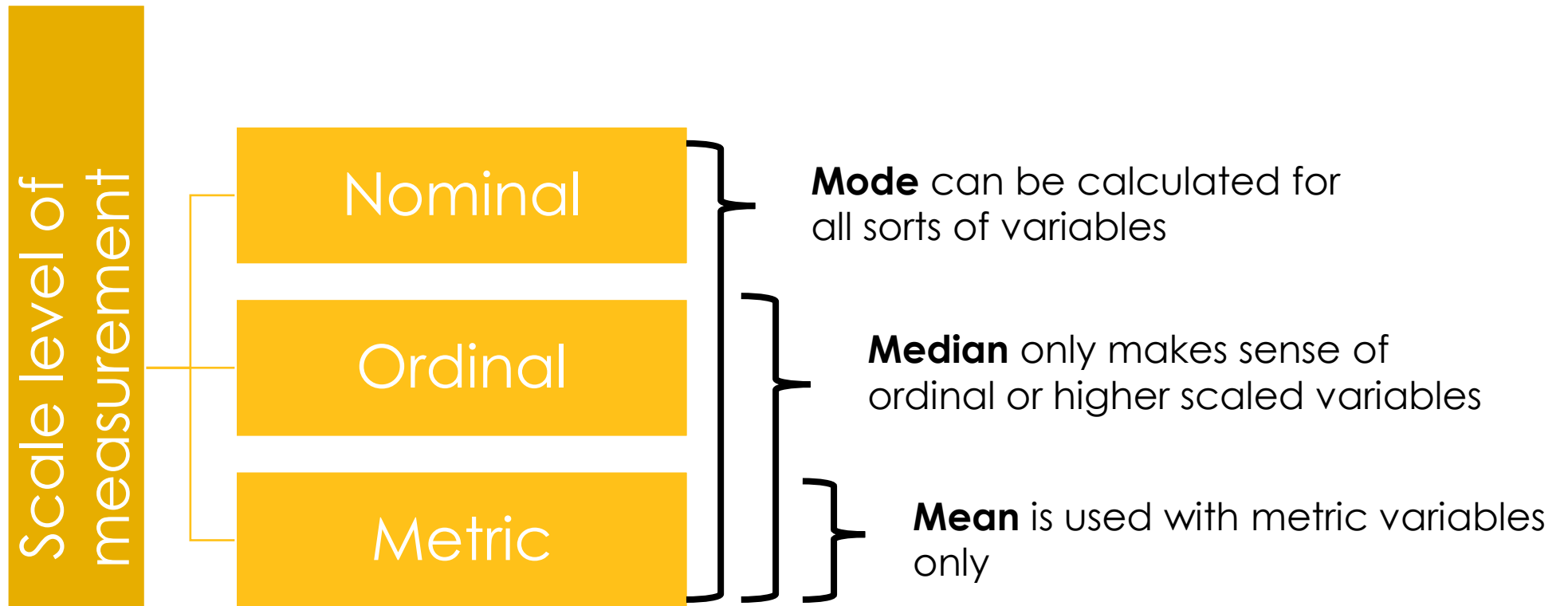
III.1 What are location parameters?

Location parameters

- Statistical indicators that provide information on the "central tendency" of a variable
 - Which value occurs most frequently? → Mode
 - Which value is in the centre of the distribution? → Median
 - What is the average value of a variable? → Mean

(Kuckartz et al. 2013)

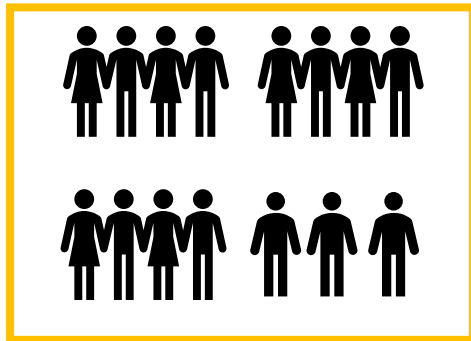
III.1 What are location parameters?



III.2 Mode

The **mode** is the value that occurs most frequently in a data set

Sample



Variable: Gender

Value	Frequency
M	9
F	6

Mode: Male

III.3 Median

The **median** divides a set of data exactly in the middle, so that 50% of the values are above and 50% of the values are below the median

Survey respondent	Monthly net Income
1	2.000€
2	1.200€
3	2.200€
4	3.400€
5	1.900€
6	2.000€
7	2.100€
8	10.200€
9	1.800€

Sort the values in ascending order

1.200, 1.800, 1.900, 2.000, 2.000, 2.100, 2.200, 3.400, 10.200

2.000 is exactly in the middle and therefore represents the median

III.4 Arithmetic Mean

The **arithmetic mean** represents the average of all available values

Survey respondent	Monthly net Income
1	2.000€
2	1.200€
3	2.200€
4	3.400€
5	1.900€
6	2.000€
7	2.100€
8	10.200€
9	1.800€

Add up all values and divide by the number:

$$\frac{(1.200 + 1.800 + 1.900 + 2.000 + 2.000 + 2.100 + 2.200 + 3.400 + 10.200)}{9} = 2977,8$$

Arithmetic mean





III.4 Arithmetic mean

Attention: As we have seen, the arithmetic mean is not particularly robust against outliers in the data. A single very high value (€10,200) has massively increased the average. Without this value, the mean value would only be €2,075.

This (German-language) blog entry explains why it can sometimes make sense to choose the median

[Warum der Durchschnitt manchmal nicht hilft – datengeschichten](#)



LITERATURE

- Kuckartz, Udo, Stefan Rädiker, Thomas Ebert, und Julia Schehl. 2013. *Statistik: Eine verständliche Einführung*. Wiesbaden: VS Verlag für Sozialwissenschaften.



Data Analytics Onboarding

Part (2): Descriptive Statistics

Deck IV: Dispersion parameters

Gabriel Malli, Sonja Radkohl



drahtwarenhandlung
film & animation
datenjournalismus





Content of Deck IV

- IV.1 What are dispersion parameters?
- IV.2 Range
- IV.3 Interquartile range
- IV.4 Standard deviation and variance



IV.1 What are dispersion parameters?

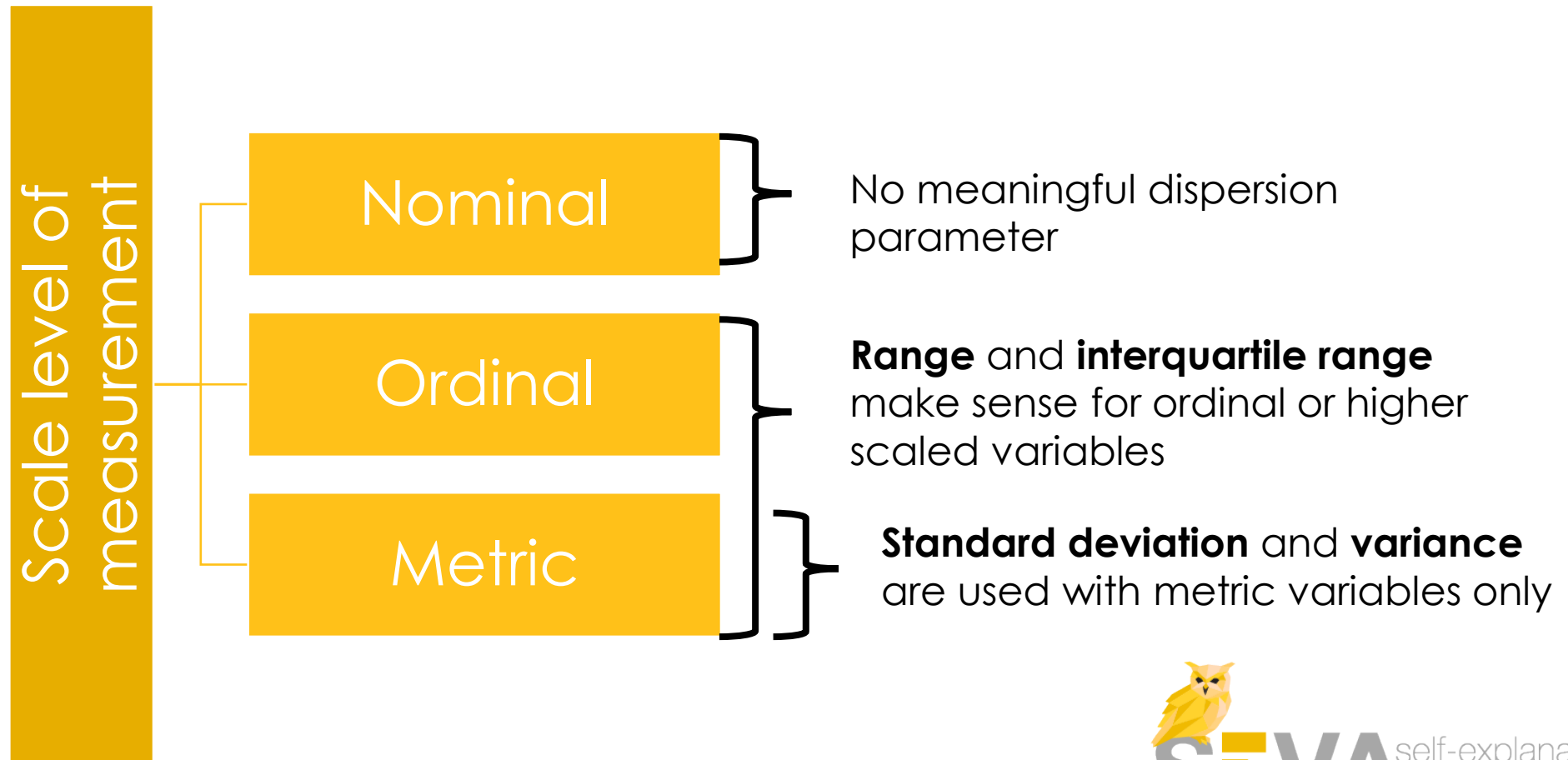
Dispersion parameters

- Statistical parameters that indicate how data in a sample is scattered around its "centre"
- Position parameters do not say anything about the distribution of the data; this requires dispersion parameters
 - How large is the distance between the largest and the smallest value? → Range
 - In which range do 50% of the values lie? → Interquartile range
 - How far away are the values of a distribution from the arithmetic mean? → Standard deviation and variance

(Kuckartz et al. 2013)



IV.1 What are dispersion parameters?



IV.2 Range

The **range** indicates the distance between the minimum and the maximum, i.e. the highest and the lowest value of a distribution

Survey respondent	Monthly net Income
1	2.000€
2	1.200€
3	2.200€
4	3.400€
5	1.900€
6	2.000€
7	2.100€
8	10.200€
9	1.800€

Maximum (10.200)

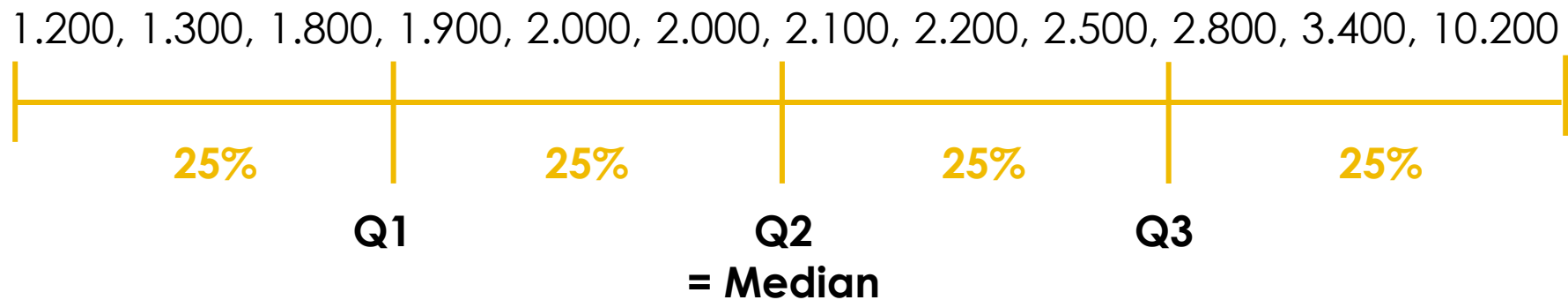
- minimum (1.200)

range (9.000)

IV.3 Interquartile range

Quartiles divide the sample into four equal parts.

[For reasons of clarity, the sample was extended by three values.]



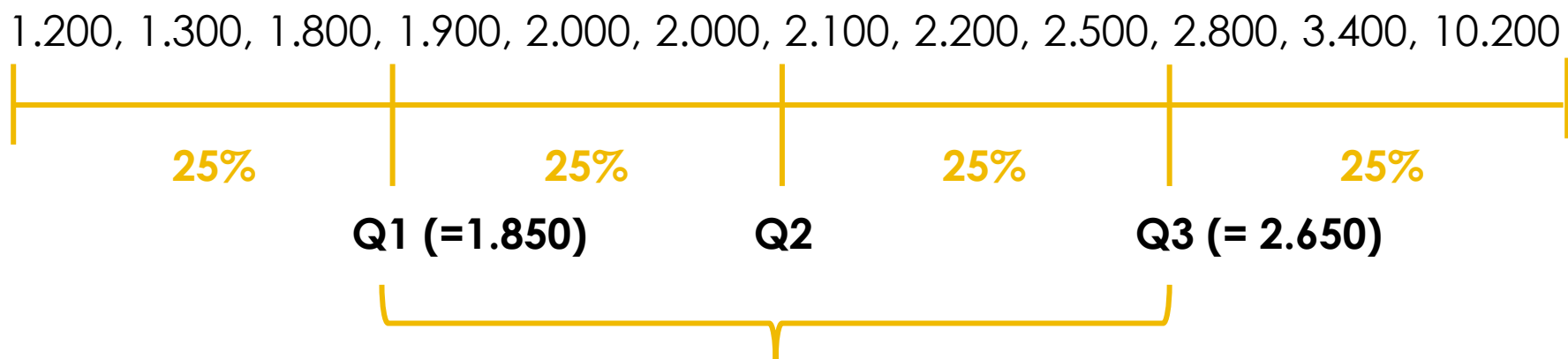
25% of the values are below the lowest quartile, 25% of the values are above the highest. 50% are between Q1 and Q3. Q2 divides the sample into equal parts and therefore corresponds to the median.

(cf. DATAtab Team 2023)



IV.3 Interquartile range

The **interquartile range** corresponds to the range between Q1 and Q3, i.e. the range of the middle 50% of the values.



When Q1 and Q3 lie between two values, they correspond to the mean of the two values

$$Q3 - Q1 = \text{Interquartile range} (= 800)$$

IV.4 Standard deviation and variance

Standard deviation (s) and **variance (var)** provide information on how much the values of a distribution scatter around the arithmetic mean.

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

Number of values

Size of individual value

Arithmetic mean

(Kuckartz et al. 2013: 71-73)



IV.4 Standard deviation and variance

How to calculate the variance:

- 1) Calculate the mean (\bar{X}):** Find the average of all the data points in the set.
- 2) Find the difference from the mean for each value:** Subtract the mean from each individual value.
- 3) Square Each of the Differences:** Square the result obtained for each data point.
- 4) Calculate the Average of the Squared Differences:** Sum up all squared results and divide them by the number of values



IV.4 Standard deviation and variance

Example: Calculating variance and standard deviation for this set of values

1.200, 1.300, 1.800, 1.900, 2.000, 2.000, 2.100, 2.200, 2.500, 2.800, 3.400, 10.200

1) Calculate the mean (\bar{X}):

$$(1.200 + 1.300 + 1.800 + 1.900 + 2.000 + 2.000 + 2.100 + 2.200 + 2.500 + 2.800 + 3.400 + 10.200) / 12 = 2783,33$$

2) Find the difference from the mean for each value:

$$1.200 - 2.783,33 = (-1.583,33) \quad 1.300 - 2.783,33 = (-1.483,33) \quad \dots \quad 10.200 - 2.783,33 = 7.416,67$$

3) Square Each of the Differences:

$$(-1.583,33)^2 = 2.506.933,89 \quad (-1.483,33)^2 = 2.200.267,89 \quad \dots \quad (7.416,67)^2 = 55.006.993,89$$

4) Calculate the Average of the Squared Differences:

$$(2.506.933,89 + 2.200.267,89 + \dots + 55.006.993,89) / 12 = 5.329.722,22 = \text{var}(x) = s^2$$

$$s = 2.308,62$$



IV.4 Standard deviation and variance

$$\bar{X} = 2783,33 \quad s = 2.308,62$$

How can these values be interpreted?

- The standard deviation of 2308,62€ indicates the average amount by which individual data points in the dataset deviate from the mean of €2783,33.
- The combination of a mean of 2783,33€ and a standard deviation of 2308,62€ suggests that the data points are spread out over a wide range.
- Income distribution appears to be unequal





IV.4 Standard deviation and variance

Free online calculator for standard deviation and variance of a dataset:

[Descriptive Statistics Calculator - DATAtab](#)





LITERATURE

- DataTAB Team. 2023. Dispersion parameter. [Dispersion parameter: Variance, standard deviation, range \(datatab.net\)](https://datatab.net)
- Kuckartz, Udo, Stefan Rädiker, Thomas Ebert, und Julia Schehl. 2013. *Statistik: Eine verständliche Einführung*. Wiesbaden: VS Verlag für Sozialwissenschaften.



Data Analytics Onboarding

Part (2): Descriptive Statistics
Deck V: Visualization
Štefan Emrich, Gabriel Malli, Sonja Radkohl





Content of Deck V

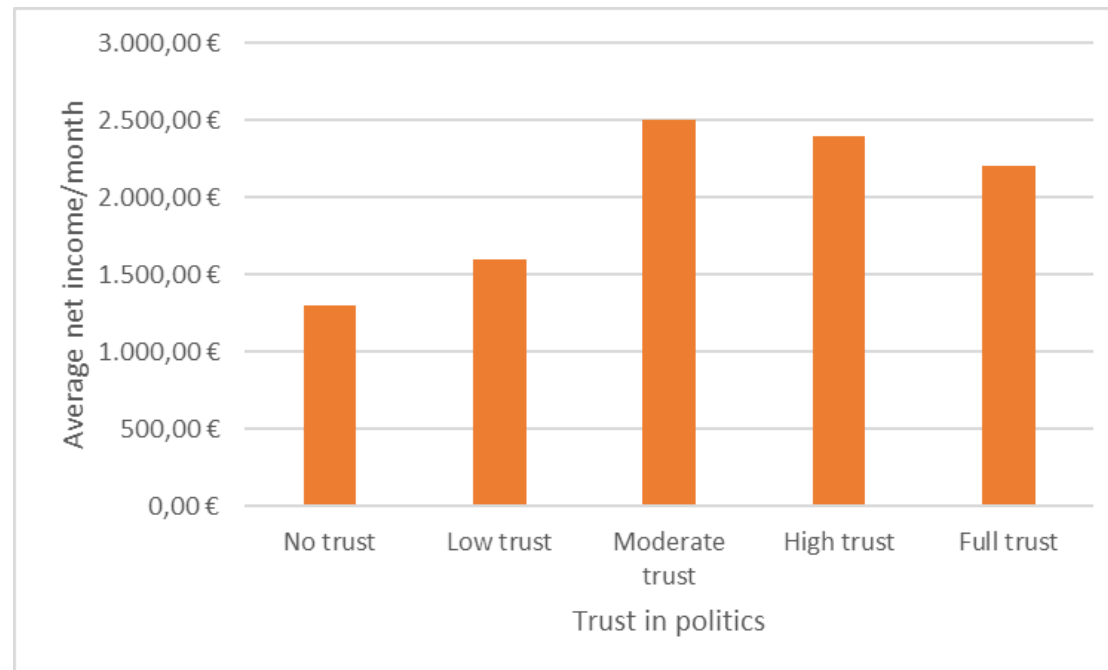
- V.1 Bar charts
- V.2 Line charts
- V.3 Area charts
- V.4 Pie charts
- V.5 Histograms
- V.6 Scatter plots
- V.7 Box plots
- V.8 Tools for data visualization

This deck is based on the *Wiki Visual & Data Analytics* by Štefan Emrich [start \[Visual & Data Analytics\] \(datengeschichten.at\)](#)



V.1 Bar charts

The **bar chart** is a diagram that shows the frequency distribution of a characteristic by means of columns/rectangles that are perpendicular to the x-axis and not adjacent to each other



(Fictitious values for visualization)



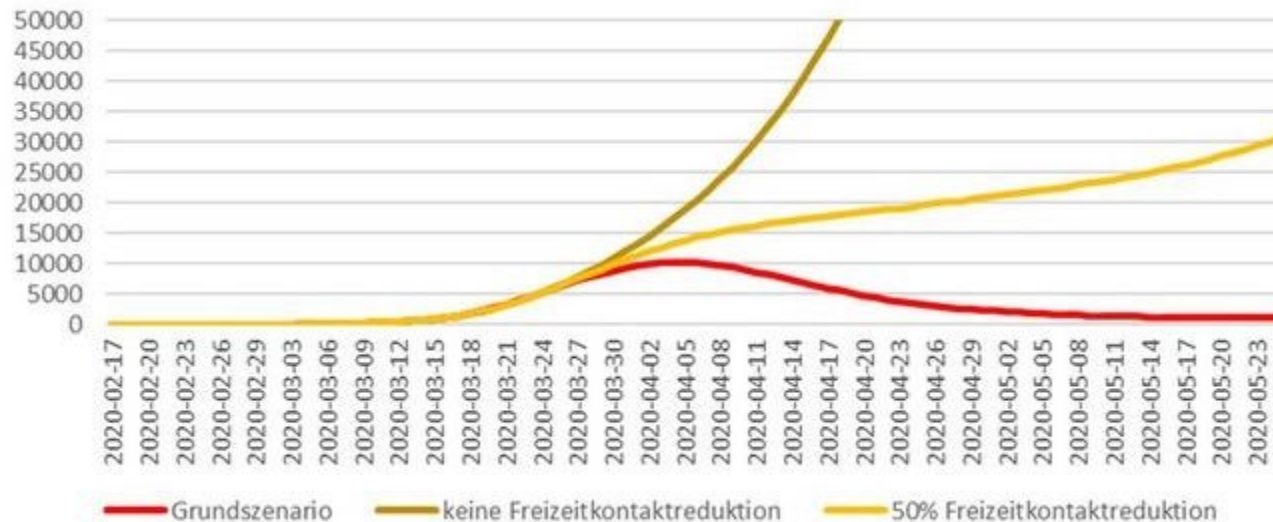
V.1 Bar charts

How to interpret	In a vertical bar chart, the characteristic values of different categories are displayed next to each other along the x-axis. The height of the bars corresponds to the value of the characteristic value. In contrast to the line representation, a bar visually includes all values from the origin to the final value.
Suitable for which data?	Vertical bar charts are suitable for unsorted data or distributions (categorical data), but also if the units can be clearly sorted. Bar charts are also a good option for discrete-time data (years, months). Overall, the variables must clearly be comparable.
How to construct	For the construction, it is sufficient to have the data of one variable.

V.1 Line charts

A **line chart** (also known as a curve chart) is the graphical representation of the functional relationship between two (in 2D representation) or three (in 3D representation) variables in line form.

The following example shows a line chart for three COVID-19 scenarios over time.





V.2 Line charts

Suitable for which data?

- Line charts are only suitable for categorical data if categories have a clear order.
- They are usually the first choice for visualising data and trends over time, especially with multiple data sets.
- They are well suited to showing correlating trends.

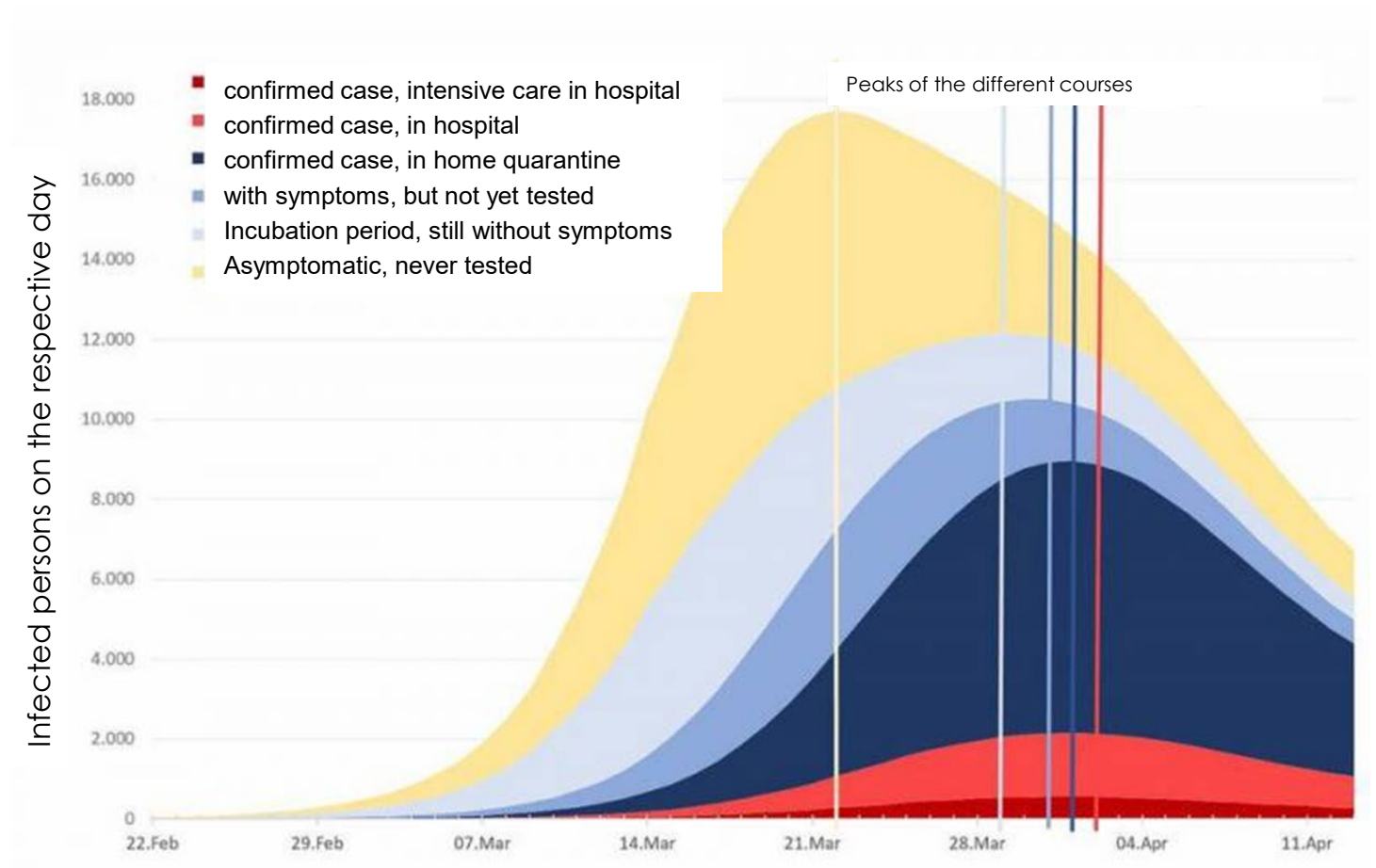


V.3 Area charts

An **area chart** graphically depicts the development of quantities. The form of representation is based on the line chart. The areas between the axis and the lines are highlighted with different colours, patterns or hatching. Two or more quantities are often compared. In this case, one speaks of a stacked or subdivided area chart.

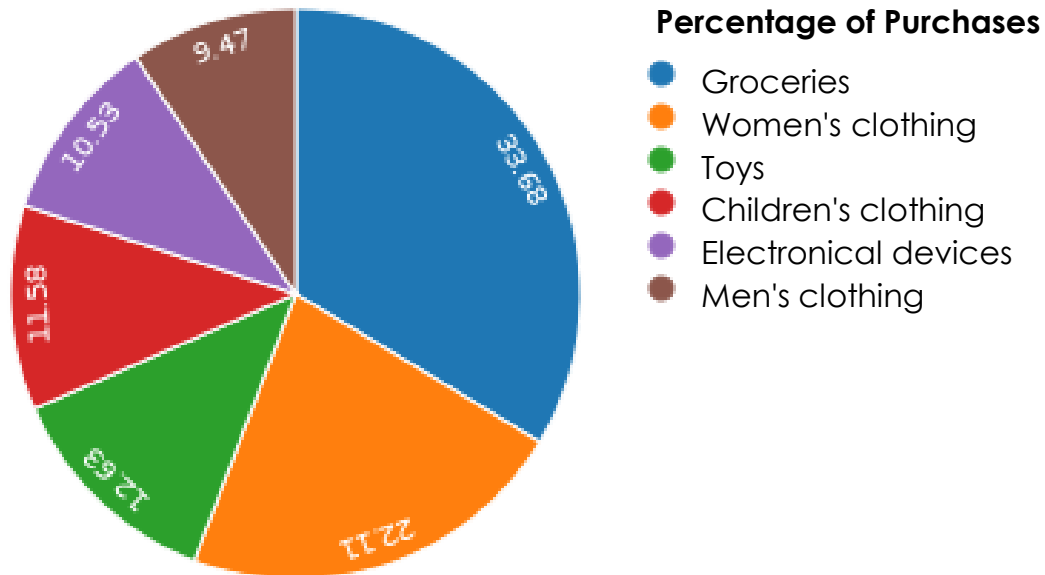
V.3 Area charts

The following example shows a stacked area diagram with six subpopulations of COVID-19 infected people.



V.4 Pie charts

The **pie chart** is a form of representation for partial values of a whole as parts of a circle. The pie chart is circular and divided into several circle sectors, with each circle sector representing a partial value and the circle thus representing the sum of the partial values (the whole).





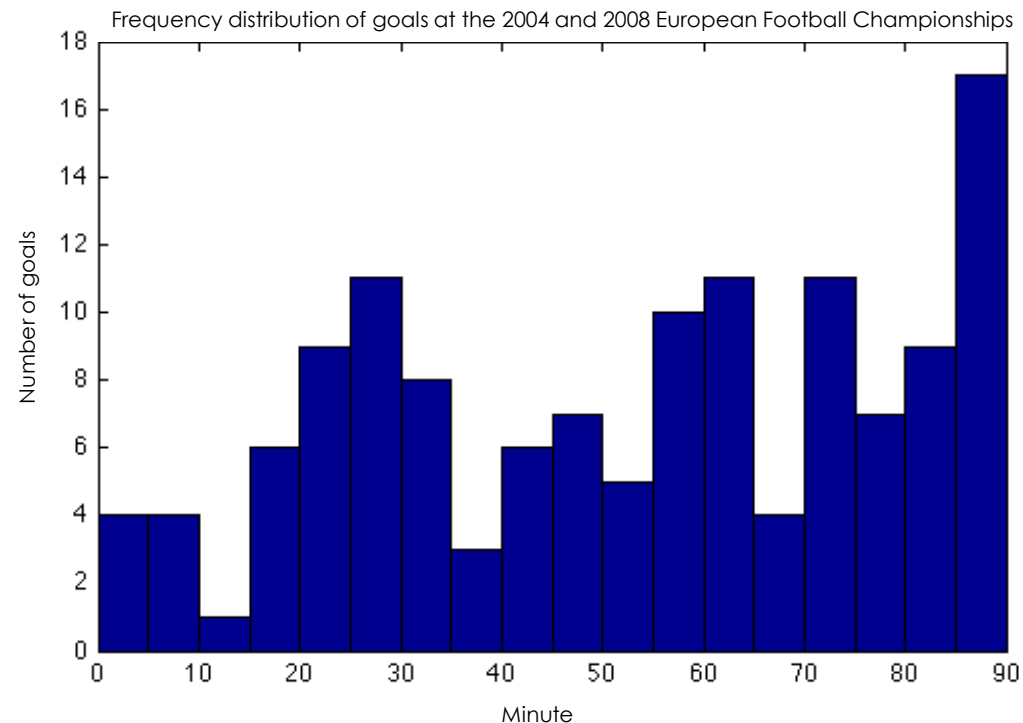
V.4 Pie charts

Suitable for which data?

Pie charts can only be used to a very limited extent. For categorical data, when no exact comparison of the segments is necessary and these add up to the total. They are suitable for size comparisons when proportions (percentage or absolute) of a whole are shown. In general, no more than a handful of partial values (approx. 5-7) should be shown, as readability quickly suffers.

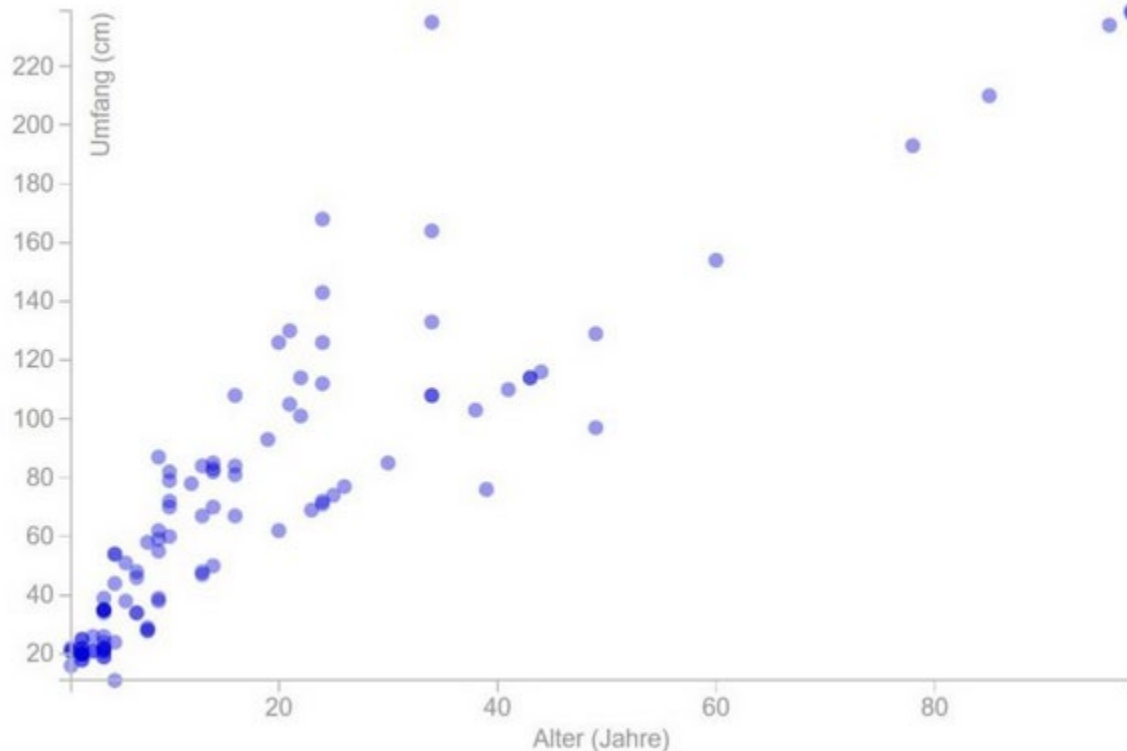
V.5 Histograms

The **histogram** is a graphical representation of the frequency distribution of metrically scaled variables. It requires the data to be divided into bins, which can have a constant or variable width. Rectangles of the width of the respective class are drawn directly next to each other, the areas of which represent the (relative or absolute) frequencies. The height of each rectangle then represents the (relative or absolute) frequency density, i.e. the (relative or absolute) frequency divided by the width of the corresponding bin.



V.6 Scatter plots

A **scatter plot**, also known as scatter cloud, is the graphical representation of pairs of values of two statistical characteristics. These pairs of values are entered into a Cartesian coordinate system, resulting in a scatterplot. The points can be visualised using various symbols. Scatterplots are very well suited to show a correlation of characteristics, as can be seen in the graphic below.



This scatterplot shows a random selection of 111 trees growing in the city of Vienna. These trees are shown with their circumference (y-axis) and age (x-axis). The example shows nicely that there is a (relatively linear) relationship between the two variables under consideration (age and circumference).

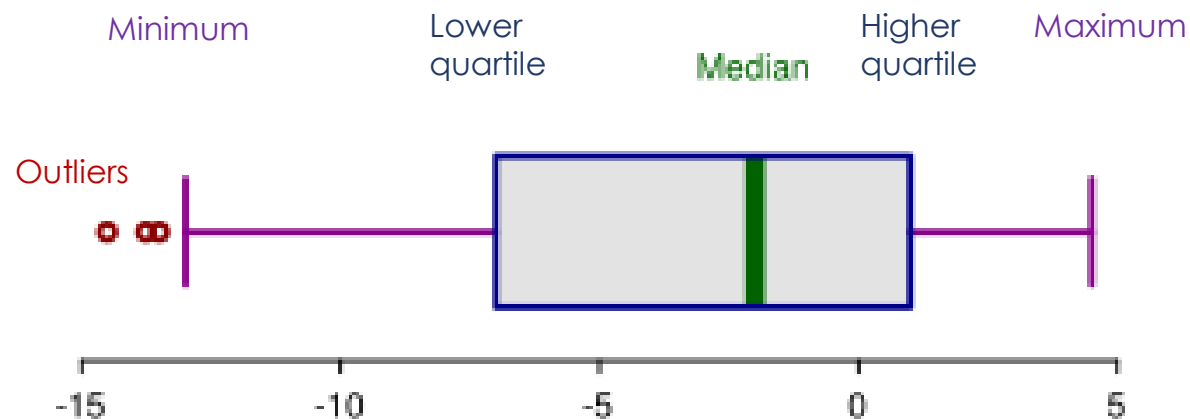


V.6 Scatter plots

How to interpret	<p>This interactive (German language) illustration explains how a scatterplot is created and how the data is coded.</p> <p><u>Der Scatterplot, das unbekannte Wesen – datengeschichten</u></p>
Suitable for which data?	<p>Scatterplots are suitable for discrete and continuous time data (years, months) and when 2 to (maximum) 4 different (ordinal or metric) characteristics are to be displayed.</p>
How to construct	<p>For a scatterplot, a data set must have at least two characteristics (per result unit) that are at least ordinal or metric in nature.</p>

V.7 Box plots

The **box plot**, also known as a whisker plot, whisker box plot or box graph, is a type of diagram that visualises the statistical variables median, quantiles (see Deck IV.3) and sometimes also outliers of a data set. A box plot is intended to give a quick impression of the range in which the data lies and how it is distributed over this range. There are vertical and horizontal variants of the box plot. This description is based on the horizontal box plot. The vertical variant is merely a 90° rotation of it.





V.7 Box plots

How to interpret

The box plot consists of a rectangle in the centre, the eponymous box, which is divided again by a line. This line indicates the position of the median of the data set. The left and right ends of the box/rectangle mark the position of the lower and upper quartiles. In other words, the threshold values below (lower quartile) and above (upper quartile) which 25% of all data points are located. This means that 50% of all values are located within the box; this distance is also called the interquartile range (IQA or IQR). There are different definitions for the sensors, whiskers or antennas, of the box. In the first variant, the ends of the antennas mark the last (or first) data point. This display form does not show outliers. In the second variant, the antennas are up to $1.5 \cdot \text{IQR}$ long (away from the box), but only extend to the last data point within this distance. In this variant, all points outside the whiskers are labelled as outliers. The shape of the box therefore immediately indicates the dispersion of the data. If the box is compact, the median is in the centre and the antennae are short, the data points are very close together.



V.7 Box plots

Suitable for which data?

Boxplots are suitable for visualising data sets with one characteristic or for comparing several data sets or samples with regard to the same characteristic.

Possible examples are, for example

- Prices for a product in different shops for different countries
- Alcohol content for different types of wine
- Income for a set of occupations for men and women



V.8 Tools for data visualization

- R, R-Studio, ggplot2
- Excel, Google Spreadsheets, freie Office-Suiten
- Python (mit entsprechenden Librarys)
- Tableau
- D3.js
- Datawrapper
- Graphistry
- MS Power BI
- Flourish

