Fachhochschule
Dortmund
University of Applied Sciences and Arts

Data Science

04: Statistics basics

Klaus Kaiser
WiSe 2024 / 2035

Data Science

Recap: Primary vs. secondary data

Primary vs. secondary data

It is common to distinguish between **primary data** and **secondary data**. While primary data is generated for the investigated purpose, secondary data is obtained from different sources and created for different purposes.

Data Science

Recap: Motivating example

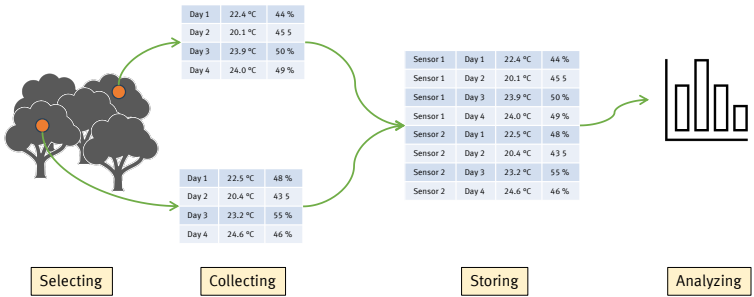


Figure: Example of how a data capturing process could work: Analyzing the climate of a forest by measuring temperature and humidity.

Data Science

Recap: NoSQL databases

NoSQL

NoSQL databases do not follow a relational approach. Therefore, they do not require a structured table scheme and try to avoid relational techniques like joins.

- Databases which break with ”traditional” SQL databases
- Relational databases: good all-rounder! NoSQL databases: experts for their special purpose
- Often: focus on horizontal scaling and special types of data

Data Science

Recap: FAIR data

The **FAIR** principles are defined to make data usage sustainable [1]

- **Findability:** Metadata and data should be easy to find for both humans and computers.
- **Accessibility:** It should be clear how data can be accessed, possibly including authentication.
- **Interoperability:** Metadata and data use names and vocabulary, which follow common principles.
- **Reusability:** Metadata and data should be well-described so that they can be replicated and/or combined in different settings

With **FAIR/O data,** FAIR data is meant which are also open available.

Data Science

Today

scraping, **data protection** and **first statistical basics**.

Collecting data

Data Science

Collecting data: Motivating example

Sometimes, data is already available, but it cannot be retrieved directly.

Examples

- Needed data could be given on a website, but the website does not offer an API to retrieve the data.
- Needed data is hidden in log files or other output of programs.

Data Science

Collecting data: Boundaries of relational databases

Task: Which data could be crawled from websites?

Data Science

Collecting data: Data scraping

Data scraping refers to the technique of extracting data from the outputs of other programs.

- Input for scraping is output which is intended to be presented to an end user
- Output is mostly structured, meaning the same information is always given at the same position

Scraping should only be used if no other technique to obtain the data, e.g. APIs, are available.

- Structure of the output could change, thus information is located at a different position
- Scraping could be restricted by the source, e.g. limiting page calls etc.

Data Science

Collecting data: Alert

Attention!

In some cases, scraping a website could be illegal - consider data protection and copyright laws, especially, if the scraped data is intended to be published or used commercially!

Data Science

Collecting data: Example

Largest cities of Germany

For a project we want to extract the *Katasterfläche* from the largest cities in Germany.

Task: Which steps would you perform to obtain the information?

Data Science

Collecting data: Example

Largest cities of Germany

1. Information is available in the Wikipedia
2. Collecting a list of all relevant Wikipedia pages - by hand or automatic:

Die größten deutschen Städte 2020

Die Einwohnerzahlen und der Gebietsstand beziehen sich auf den 31. Dezember 2020.

Rang	Stadt	Einwohnerzahl	Land
1.	Berlin	3.664.088	Berlin
2.	Hamburg	1.852.478	Hamburg
3.	München	1.488.202	Bayern
4.	Köln	1.083.498	Nordrhein-Westfalen
5.	Frankfurt am Main	764.104	Hessen

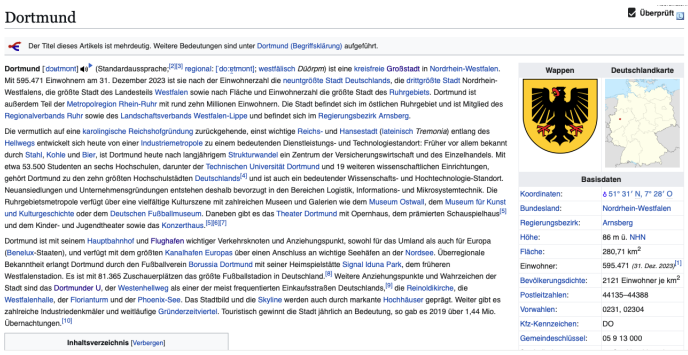
Figure: List of the largest cities in Germany, taken from [2]

Data Science

Collecting data: Example

Largest cities of Germany

3. Pages of the cities have a similar structure, e.g. on the right side is a box with some facts of the city:



Dortmund

Überprüft

Der Titel dieses Artikels ist mehrdeutig. Weitere Bedeutungen sind unter Dortmund (Begriffsklärung) aufgeführt.

Dortmund [ˈdɔʁtmʊnt] (Standardaussprache:[2][3] regional: [ˈdoːɐ̯tmʊnt]; westfälisch *Düörpm*) ist eine kreisfreie **Großstadt** in Nordrhein-Westfalen. Mit 595.471 Einwohnern am 31. Dezember 2023 ist sie nach der Einwohnerzahl die neuntgrößte Stadt Deutschlands, die drittgrößte Stadt Nordrhein-Westfalens, die größte Stadt des Landesteils Westfalen sowie nach Fläche und Einwohnerzahl die größte Stadt des Ruhrgebiets. Dortmund ist außerdem Teil der Metropolregion Rhein-Ruhr mit rund zehn Millionen Einwohnern. Die Stadt befindet sich im östlichen Ruhrgebiet und ist Mitglied des Regionalverbands Ruhr sowie des Landschaftsverbands Westfalen-Lippe und befindet sich im Regierungsbezirk Arnsberg.

Die vermutlich auf eine karolingische Reichshofgründung zurückgehende, einst wichtige Reichs- und Hansestadt (lateinisch *Tremonia*) entlang des Hellwegs entwickelt sich heute von einer Industriemetropole zu einem bedeutenden Dienstleistungs- und Technologiestandort: Früher vor allem bekannt durch Stahl, Kohle und Bier, ist Dortmund heute nach langjährigem Strukturwandel ein Zentrum der Versicherungswirtschaft und des Einzelhandels. Mit etwa 53.500 Studenten an sechs Hochschulen, darunter der Technischen Universität Dortmund und 19 weiteren wissenschaftlichen Einrichtungen, gehört Dortmund zu den zehn größten Hochschulstädten Deutschlands[4] und ist auch ein bedeutender Wissenschafts- und Hochtechnologie-Standort. Neuansiedlungen und Unternehmensgründungen entstehen deshalb bevorzugt in den Bereichen Logistik, Informations- und Mikrosystemtechnik. Die Ruhrgebietsmetropole verfügt über eine vielfältige Kulturszene mit zahlreichen Museen und Galerien wie dem Museum Ostwall, dem Museum für Kunst und Kulturgeschichte oder dem Deutschen Fußballmuseum. Daneben gibt es das Theater Dortmund mit Opernhaus, dem prämierten Schauspielhaus[5] und dem Kinder- und Jugendtheater sowie das Konzerthaus.[5][6][7]

Dortmund ist mit seinem Hauptbahnhof und Flughafen wichtiger Verkehrsknoten und Anziehungspunkt, sowohl für das Umland als auch für Europa (Benelux-Staaten), und verfügt mit dem größten Kanalhafen Europas über einen Anschluss an wichtige Seehäfen an der Nordsee. Überregionale Bekanntheit erlangt Dortmund durch den Fußballverein Borussia Dortmund mit seiner Heimspielstätte Signal Iduna Park, dem früheren Westfalenstadion. Es ist mit 81.365 Zuschauerplätzen das größte Fußballstadion in Deutschland.[8] Weitere Anziehungspunkte und Wahrzeichen der Stadt sind das Dortmunder U, der Westenhellweg als einer der meist frequentierten Einkaufsstraßen Deutschlands,[9] die Reinoldikirche, die Westfalenhalle, der Florianturm und der Phoenix-See. Das Stadtbild und die Skyline werden auch durch markante Hochhäuser geprägt. Weiter gibt es zahlreiche Industriedenkmäler und weitläufige Gründerzeitviertel. Touristisch gewinnt die Stadt jährlich an Bedeutung, so gab es 2019 über 1,44 Mio. Übernachtungen.[10]

Inhaltsverzeichnis [Verbergen]

Wappen	Deutschlandkarte

Basisdaten

Koordinaten:	δ 51° 31′ N, 7° 28′ O
Bundesland:	Nordrhein-Westfalen
Regierungsbezirk:	Arnsberg
Höhe:	86 m ü. NHN
Fläche:	280,71 km²
Einwohner:	595.471 (31. Dez. 2023)[1]
Bevölkerungsdichte:	2121 Einwohner je km²
Postleitzahlen:	44135–44388
Vorwahlen:	0231, 02304
Kfz-Kennzeichen:	DO
Gemeindeschlüssel:	05 9 13 000

Figure: Screenshot of the Wikipedia article of Dortmund [3]

Largest cities of Germany

4. Underlying, the source of the website is HTML, which can be parsed to extract the information needed

```
<tr class="hintergrundfarbe-basis">
    <td>
        <a href="/wiki/Katasterfl" title="Katasterfläche">
    Fläche</a>:
    </td>
    <td>
        280,71 km <sup>2</sup>
    </td>
</tr>
```

Data Science

Collecting data: Data Scraping

- Alternatives to web scraping: listening to data feeds of web servers
- Log scraping: Extracting information from log files
- Screen scraping: Extracting information from screen (e.g. from screenshots)

Data protection

Data Science

Data protection: Motivation

We have seen: Data is everywhere, but can we always use data? Are there limits or restrictions?

We are not allowed to use every data, we could get!

Data Science

Data protection Disclaimer

Attention!

This is only a brief overview of potential risks. They are intended to sensitize you! I'm not a lawyer! If there are questions - consult your data protection officer!

Lectures with further information

- Informatik und Gesellschaft
- IT-Recht

Data Science

Data protection

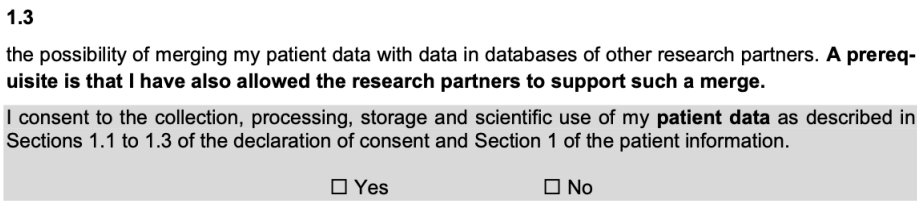
- **Copyright:** Some data might be under copyright, thus using it might be restricted.
- **Company secrets:** Data can contain the secrets of a company! Thus, using this data might be restricted by the company.
- **Ethics:** Depending on what should be done with data, this might be forbidden!

Many companies store personal data of their customers. Due to laws (e.g., GDPR), this data is protected!

- Personal data is only allowed to be stored as long as it is needed for the purpose
- The usage of personal data is bound to the purpose
- The person must give its consent to store and use personal data

A new patient

If a new patient comes to a doctor, it is common for the patient to sign a data protection form. With this form the patient gives the consent that their personal data to be shared, e.g., to be sent to a laboratory or another doctor.
In a clinic, might also be a question regarding whether the personal data is permitted to be used for scientific purposes.



1.3

the possibility of merging my patient data with data in databases of other research partners. **A prerequisite is that I have also allowed the research partners to support such a merge.**

I consent to the collection, processing, storage and scientific use of my **patient data** as described in Sections 1.1 to 1.3 of the declaration of consent and Section 1 of the patient information.

☐ Yes ☐ No

Figure: Example of a consent for scientific usage of personal data [4]

Data protection

Anonymization

The GDPR (in Germany DSGO) gives the possibility to use data for a different purpose.

Art. 6 GDPR Lawfulness of processing - item 4

Where the processing for a purpose other than that for which the personal data have been collected is not based on the data subject's consent or on a Union or Member State law which constitutes a necessary and proportionate measure in a democratic society to safeguard the objectives referred to in Article 23(1), the controller shall, in order to ascertain whether processing for another purpose is compatible with the purpose for which the personal data are initially collected, take into account, inter alia:

[...]

(e) the existence of appropriate safeguards, which may include encryption or pseudonymisation.

Anonymization

Data is anonymized by removing or editing all personal information, such that the person cannot be identified anymore.

- Sometimes, some personal information are needed to make the dataset usable.
- If anonymization is not possible, pseudonymization could be possible
  - **Example:** It is needed that the dataset distinguish between persons. Thus, some kind of ID is needed.

Pseudonymization

Data is pseudonymized by editing all personal information, such that the person can only be identified if additional (not generally available) information is needed.

Depending on the question, which should be answered with the help of a dataset, anonymization may be possible or pseudonymization may be needed.

name	first name	zip-code	product	count	price
Hans	Doe	44227	Device A	2	434
Hans	Doe	44227	Device B	4	234
Alice	Doe	44227	Device C	1	120

Table: Example for a dataset which could be used for analysis.

- **Task A:** Give the volume of sales for every product
- **Task B:** Give the volume of sales for every customer

Exercise

Give the volume of sales for every product

Task: How to anonymize the customer?

Exercise

Give the volume of sales for every customer

Task: How to pseudonymize the customer?

Exercise

Give the volume of sales for every product

Task: How to anonymize the customer?

product	count	price
Device A	2	434
Device B	4	234
Device C	1	120

Exercise

Give the volume of sales for every customer

Task: How to pseudonymize the customer?

customer	product	count	price
43123	Device A	2	434
43123	Device B	4	234
23342	Device C	1	120

Hash-function

A hash-function is a function which maps arbitrary data to a value of fixed size, i.e. $h : K \to S$ with $|K| > |S|$.

- The hash function should be surjective: Every hash-value should be possible
- The hash function should be efficient to be calculated
- The hash values should be equaly distributed on the expected input values

For a common hash-function like **sha-256** it is very rare that two strings have the same hash-value, i.e. the probability is equal to

$$\left(\frac{1}{2}\right)^{256} \approx 8.64 \cdot 10^{-78}$$

Task: Imagine you daily life, where do you find hash-functions or hash-values?

Examples

- Checksum (e.g. IBAN)
- $h(s) = s \mod p$, where p is a prime number
- MD5, SHA256 etc.

Data Science

Anonymization Anonymization

First name	Family name	Birthdate	Data_1	Data_2	Data_3
Max	Mustermann	2008-04-02	12331	Device_A	4432
Max	Mustermann	2008-04-02	52345	Device_B	6543
Erika	Mustermann	2010-01-01	52342	Device_B	1231

Hashing

Values are pseudonymized by computing a hash value of all personal information:

`sha('MaxMustermann2024-04-02')` = `'756769e032c43f1d645cd561250be78b'`

`sha('ErikaMustermann2010-01-01')` = `'b87bc342c0f2a876a65ecc5860db879e'`

Hash	Data_1	Data_2	Data_3
756769e032c43f1d645cd561250be78b	12331	Device_A	4432
756769e032c43f1d645cd561250be78b	52345	Device_B	6543
b87bc342c0f2a876a65ecc5860db879e	52342	Device_B	1231

When the name and birthday of a person are known, the hash value can be computed.
Thus, the entries corresponding to a person could be identified.

Hashing+Salting

Values are pseudonymized by creating a salt value (random value) which is added to the string of personal information. The salt only known to the data owner:

`sha('MaxMustermann2024-04-0255234523') = 'd8d384f24db0006fd33fcf86b2ee33bf'`

`sha('ErikaMustermann2010-01-0155234523') = 'f4bcce2e15f17f72dcd1f00660a32f1e'`

Hash	Data_1	Data_2	Data_3
d8d384f24db0006fd33fcf86b2ee33bf	12331	Device_A	4432
d8d384f24db0006fd33fcf86b2ee33bf	52345	Device_B	6543
f4bcce2e15f17f72dcd1f00660a32f1e	52342	Device_B	1231

When performing an anonymization to release a dataset, it is important to consider all values given in the dataset to check if a person can also be identified even if all personal information is removed.

Example

AOL released a large dataset with search queries in 2006[5]. Personal information like IDs was deleted!

Task: What was the problem with the dataset?

Example

AOL released a large dataset with search queries in 2006[6]. Personal information like IDs was deleted!

Task: What was the problem with the dataset?

The dataset contained search queries, which gave information about the user. Thus, with the help of the dataset, many persons could be identified.

Statistics Introduction

Up to now

We discussed several ways to obtain data - we know how to get it**, but what next?**

	Name	population	category	area	population / area	state
1	Berlin	3782202	metropolis	891.12	4244	BE
2	Hamburg	1910160	metropolis	755.09	2530	HH
⋮	⋮	⋮	⋮	⋮	⋮	⋮
8	Leipzig	619879	big city	297.8	2082	SN
9	Dortmund	595471	big city	280.71	2121	NW
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table: List of the largest cities (83 cities with more than 100000 inhabitants) of Germany adapted from[7].

Task: Which questions could be answered with the help of this dataset?

We need to analyze the data to gain knowledge from it! Thus, we need statistics!

”Statistics is the art of stating in precise terms that which one does not know.”

William Kruskal[8]

or

Statistics is the discipline that concerns the collection, organization, analysis, interpretation, and presentation of data.[9]

Descriptive statistics

Description of data by computing frequencies, characteristic numbers and presenting data graphically.

Explorative Statistik

Search for structures and special features in the data. Generating of new questions and hypotheses.

inductive statistics

Conclusion on data-generating mechanism with propability theory.

Frequencies and histograms

We consider a variable X with $n \in \mathbb{N}$ observations $x_1, \ldots x_n$. Every entry equals to one of $l \in \mathbb{N}$ with $l \leq n$ different values $a_1, \ldots, a_l$.

observation	x_1	x_2	x_3	x_4	x_5	$\ldots$	x_{n-1}	x_n
value	a_1	a_2	a_2	a_1	a_2	$\ldots$	a_{l-1}	a_l

For a categorical variable (nominal or ordinal) l is equal to the number of categories and is in general much smaller than n. For a metric scale, l is close to n.

We define

- **absolute frequency** as

$$h(a_j) = h_j \text{ with } h_j = \sum_{i=1}^{n} (a_j = x_i) \text{ and } 1 \leq j \leq l$$

- **relative frequency** as

$$f(a_j) = \frac{h_j}{n}$$

Furthermore, we call

- the values $h_1, \ldots, h_k$ **absolute frequency distribution**
- the values $f_1 \ldots f_l$ **relative frequency distribution**.

Example

List of the states of the largest cities of Germany:

BE	HH	BY	NW	HE	BW	NW	SN	NW	NW	HB	SN	NI	BY
NW	NW	NW	NW	NW	NW	BW	BW	BY	HE	NW	NW	NW	NI
SN	SH	ST	ST	BW	NW	RP	SH	TH	NW	MV	HE	NW	BB
SL	NW	RP	NI	NW	NI	NW	HE	BW	NW	BY	NW	NW	NW
BY	HE	BY	BW	BW	BW	BY	NI	NI	NW	BW	BY	RP	HB
NW	RP	NW	NW	TH	NW	NI	HE	NW	NI	NW	RP	BB	

Data Science

Frequencies and histograms Frequencies

a_j	NW	BW	BY	NI	HE	RP	SN	HB
$h(a_j)$	30	9	8	8	6	5	3	2
$f(a_j)$	0.36	0.11	0.10	0.10	0.07	0.06	0.04	0.02

a_j	SH	ST	TH	BB	BE	HH	MV	SL
$h(a_j)$	2	2	2	2	1	1	1	1
$f(a_j)$	0.02	0.02	0.02	0.02	0.01	0.01	0.01	0.01

The above table is also called **frequency table**.

For metric data, there are often many different values, i.e. one would consider approximately as many classes as there are entries ($l \approx n$).

Example

List of the sizes of the population of the largest cities in Germany. No entry occurs twice.

3782202	1910160	1510378	1087353	775790	633484	631217	619879	595471	586608
577026	566222	548186	526091	503707	366385	358938	338410	335789	322904
316877	309964	303150	285522	268943	265885	252769	252066	250681	248873
242172	240114	237244	228550	222889	219044	215675	211099	210795	204687
190490	187119	183509	180761	176110	174629	173255	166960	166414	164792
162960	161545	159465	157896	155749	155163	142308	135490	132032	130093
129942	128992	128246	127256	120261	118705	118528	117806	115298	114677
112970	112737	112660	111693	110791	105606	105039	103184	102464	102325
102114	101486	100010							

If there are many values, it could be useful to define classes, for grouping the values in.

1. Define the number of classes
2. Define the class-boundaries
3. Compute the frequencies for every class by counting the values which belongs to one class.

In this setting, especially for histograms, the classes are often referred to as **bins**.

The number $c \in \mathbb{N}$ of classes, can be computed by

- $\lfloor 1 + 3.3 \log_{10}(n) \rfloor$ (rule of Sturges)
- $\lfloor \sqrt{n} \rfloor$ or $2\lfloor \sqrt{n} \rfloor$

where n gives the number of values and $\lfloor a \rfloor$ the largest natural number which is smaller than a.

Example

The number of cities we consider is 83, thus we could use

- $\lfloor 1 + 3.3 \log_{10}(83) \rfloor = 15$
- $\lfloor \sqrt{83} \rfloor = 9$
- $2\lfloor \sqrt{83} \rfloor = 18$

categories.

When the number of classes is defined, the next step is choosing the class boundaries $c_0, \ldots, c_k$. The classes are then given by $[c_{j-1}, c_j)$ for $j = 1, \ldots, l$. Furthermore, they should fulfill:

- Every value a_i should be present in one class $[c_{j-1}, c_j)$ for $j = 1, \ldots, l$.
- The classes should be disjoint, i.e. no possible value could occur in two classes - $c_{j-1} < c_j$ for all $j = 1, \ldots, l$.
- For c_0 and c_k one can choose the value $-\infty$ and ∞

The distance between two class boundaries, $d_j = c_j - c_{j-1}$, is called **class width**.

- **Ideal:** All classes have the same class width
- Sometimes, the same class width is not useful, e.g.
  - if this results into empty classes
  - if one uses $\pm\infty$ as class boundary

Data Science

Frequencies and histograms Building classes

Example

3782202	1910160	1510378	1087353	775790	633484	631217	619879	595471	586608
577026	566222	548186	526091	503707	366385	358938	338410	335789	322904
316877	309964	303150	285522	268943	265885	252769	252066	250681	248873
242172	240114	237244	228550	222889	219044	215675	211099	210795	204687
190490	187119	183509	180761	176110	174629	173255	166960	166414	164792
162960	161545	159465	157896	155749	155163	142308	135490	132032	130093
129942	128992	128246	127256	120261	118705	118528	117806	115298	114677
112970	112737	112660	111693	110791	105606	105039	103184	102464	102325
102114	101486	100010							

Task: Let us consider $k = 9$ classes, how would you choose the class boundaries?

Example

There are different ways to choose the class boundaries, two possible examples:

1. Choosing all classes with the same size and choosing the left and right boundary in such a way that all values are in the classes (e.g. by rounding the values):

$$c_0 = 100000,\ c_l = 4000000 \text{ and } d_j \approx 443333 \text{ for all } j = 1, \ldots, l$$

2. Choosing $c_{l-1} = 1000000$ and $c_l = \infty$ to avoid many empty classes, the rest is chosen similar to the first suggestion:

$$c_0 = 100000,\ c_{l-1} = 1000000 \text{ and } d_j \approx 123750 \text{ for all } j = 1, \ldots, l$$

Data Science

Frequencies and histograms Building classes

class	h
$[100000, 533333)$	70
$[533333, 966666)$	9
$[966666, 1400000)$	1
$[1400000, 1833333)$	1
$[1833333, 2266666)$	1
$[2266666, 2700000)$	0
$[2700000, 3133333)$	0
$[3133333, 3566666)$	0
$[3566666, 4000000)$	1

class	h
$[100000, 212500)$	46
$[212500, 325000)$	18
$[325000, 437500)$	4
$[437500, 550000)$	3
$[550000, 662500)$	7
$[662500, 775000)$	0
$[775000, 887500)$	1
$[887500, 1000000)$	0
$[1000000, \infty)$	4

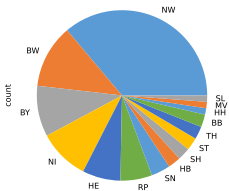
Frequencies and histograms

Plotting frequencies

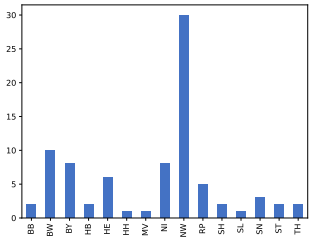
Data Science

Frequencies and histograms Frequencies

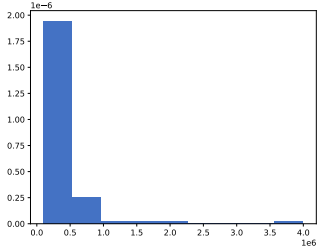
There are several ways to present frequencies with the help of plots.



Pie chart



Bar chart

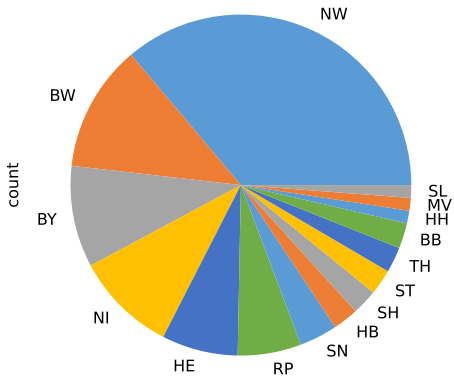


Histogram

While pie and bar charts are most likely to be used for nominal data or classes with the same distance, a histogram can be used for metric data.

Data Science

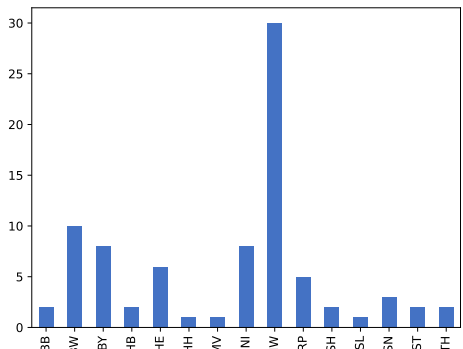
Frequencies and histograms Frequencies



In a **pie chart** the size and angle of a circular segment is proportional to the frequency of the quantity it represents.
The angle for the j^{th} segment is given by

$$2\pi \cdot f(a_j)$$

(In °: $360° \cdot f(a_j)$).



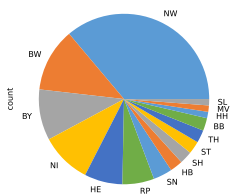
In a **bar chart** the values are given on the x-axis, the absolute or relative frequency are given on the y-axis. The frequency of a value is represented by a horizontal bar:

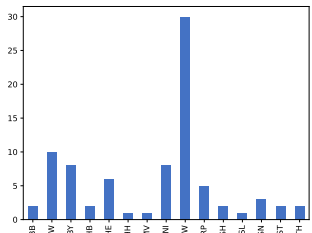
- position corresponds to the value
- length to the frequency

The choice of how to present frequencies depends on, what the goal of the presentation is! For example, it depends on, how the viewer should compare one value with others.

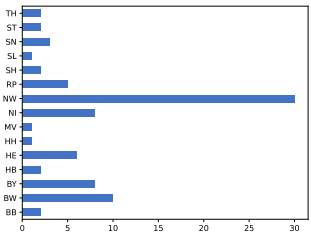
Task: How good can the following questions be answered with a pie or bar chart?

- How does the frequency of one value compare to blocks of other frequencies?
- How does the frequency of one value compare to another frequency?

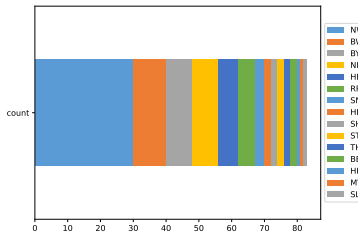




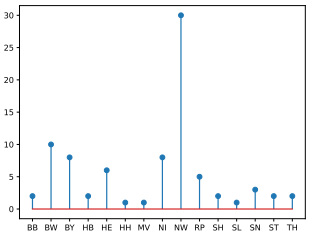
There are also some alternatives to representing frequencies of categorical variables with the help of pie or bar charts.



Horizontal bar chart



Stacked bar chart



Lollipop chart

A histogram is one way to present frequencies of continuous / metric data. For this, the data is classified into classes (see one of the previous slides) with class boundaries $c_0, \ldots, c_k$. Then, the j^{th} class, for $j = 1, \ldots, k$, is represented by a box starting from c_{j-1} to c_j (width equals to $d_j = c_j - c_{j-1}$) and height

$$g_j := \frac{f_j}{d_j} = \frac{f_j}{c_j - c_{j-1}}.$$

- The area of the box equals to the frequency of the class:

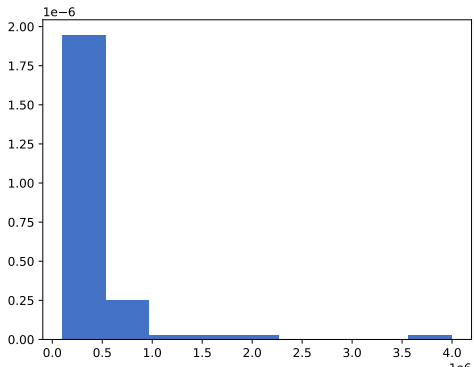
$$d_j \cdot g_j = d_j \frac{f_j}{d_j} = f_j$$

- The complete area of all boxes of the histogram equals to 1.
- If every class width has the same size, the height g_j is proportional to the relative frequency f_j.

Data Science

Frequencies and histograms Frequencies

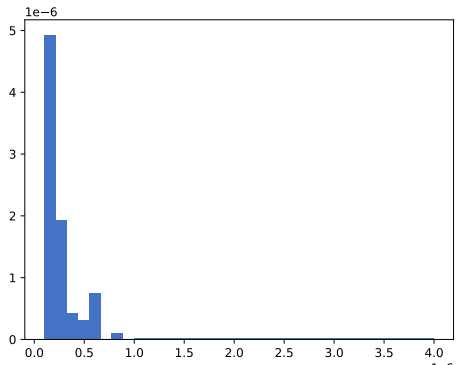
j	c_{j-1}	c_j	h_j	f_j	g_j
1	100000	533333	70	0.84	$1.95 \cdot 10^{-06}$
2	533333	966666	9	0.11	$2.50 \cdot 10^{-07}$
3	966666	1400000	1	0.01	$2.78 \cdot 10^{-08}$
4	1400000	1833333	1	0.01	$2.78 \cdot 10^{-08}$
5	1833333	2266666	1	0.01	$2.78 \cdot 10^{-08}$
6	2266666	2700000	0	0.0	0.0
7	2700000	3133333	0	0.0	0.0
8	3133333	3566666	0	0.0	0.0
9	3566666	4000000	1	0.01	$2.78 \cdot 10^{-08}$



Data Science

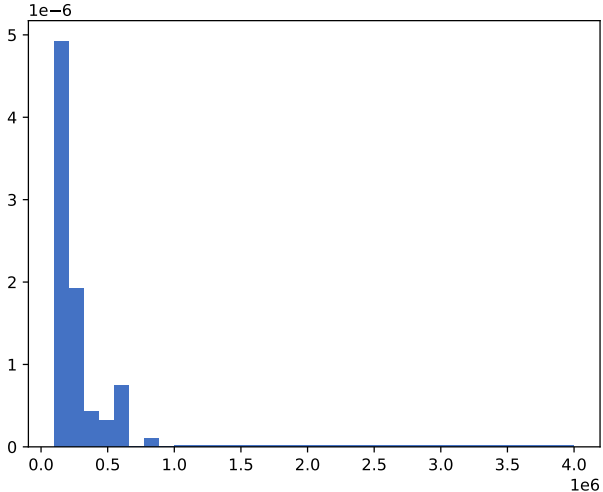
Frequencies and histograms Frequencies

j	c_{j-1}	c_j	h_j	f_j	g_j
1	100000	212500	46	0.55	$4.93 \cdot 10^{-06}$
2	212500	325000	18	0.22	$1.92 \cdot 10^{-06}$
3	325000	437500	4	0.05	$4.28 \cdot 10^{-07}$
4	437500	550000	3	0.04	$3.21 \cdot 10^{-07}$
5	550000	662500	7	0.08	$7.50 \cdot 10^{-07}$
6	662500	775000	0	0.0	0.0
7	775000	887500	1	0.01	$1.07 \cdot 10^{-07}$
8	887500	1000000	0	0.0	0.0
9	1000000	4000000	4	0.05	$4.28 \cdot 10^{-07}$



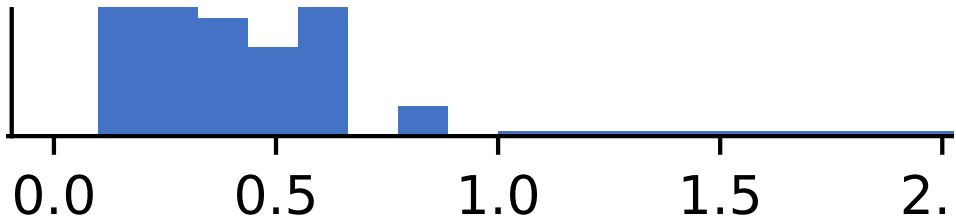
Data Science

Frequencies and histograms Frequencies



Data Science

Frequencies and histograms Frequencies



A histogram can also represent the absolute frequencies, in this case, the size of one box is equal to the absolute frequency and the sum of all sizes is equal to the total number of observations.

Frequencies and histograms

Empirical Distribution Function

Data Science

Frequencies and histograms Motivation

	Name	population	category	area	population / area	state
1	Berlin	3782202	metropolis	891.12	4244	BE
2	Hamburg	1910160	metropolis	755.09	2530	HH
⋮	⋮	⋮	⋮	⋮	⋮	⋮
8	Leipzig	619879	big city	297.8	2082	SN
9	Dortmund	595471	big city	280.71	2121	NW
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table: List of the largest cities (83 cities with more than 100000 inhabitants) of Germany adapted from[10].

What is the proportion of large cities with less than or equals to 250000 inhabitants?

3782202	1910160	1510378	1087353	775790	633484	631217	619879	595471	586608
577026	566222	548186	526091	503707	366385	358938	338410	335789	322904
316877	309964	303150	285522	268943	265885	252769	252066	250681	248873
242172	240114	237244	228550	222889	219044	215675	211099	210795	204687
190490	187119	183509	180761	176110	174629	173255	166960	166414	164792
162960	161545	159465	157896	155749	155163	142308	135490	132032	130093
129942	128992	128246	127256	120261	118705	118528	117806	115298	114677
112970	112737	112660	111693	110791	105606	105039	103184	102464	102325
102114	101486	100010							

The number of cities with less than 250000 inhabitants is given by:

$$\sum_{j=1}^{n} \begin{cases} 1 & x_j \leq 250000 \\ 0 & \textit{otherwise} \end{cases} = \sum_{a_j \leq 250000} h(a_j) = 54.$$

Thus, the proportion equals to 0.65.

For an ordinal or continuous variable X with observations $x_1 \dots x_n$ with different values $a_1, \dots, a_l$. We define the **absolute cumulative frequency** as

$$H(x) = \sum_{a_j \leq x} h(a_j)$$

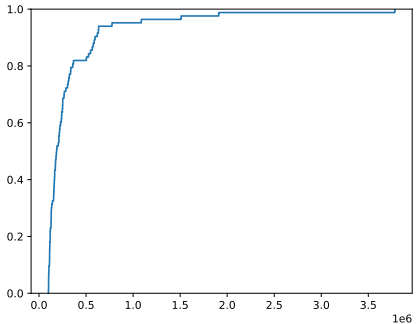
and the **empirical distribution function** $F_n : \mathbb{R} \to [0, 1]$ as

$$F_n(x) = \sum_{a_j \leq x} f(a_j) = \frac{H(x)}{n}.$$

For variables with interpretable distances, we can plot the empirical distribution function as a piecewise constant, or staircase / step, function.

Data Science

Frequencies and histograms Motivation



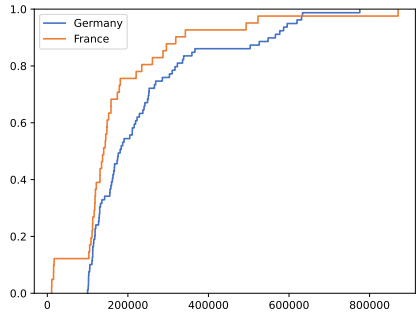


Figure: Empirical distribution function of large cities in Germany (left) and large cities which are no metropolis in Germany and France (right)

Summary & Outlook

Data Science

Summary & Outlook: Summary

- You are aware on problems concerning data protection
- You know ways to anonymize or pseudonymize data
- You have a first idea what statistics is
- You are able to compute frequencies of different data
- You can distinguish different plotting methods and are able to choose a proper one

Data Science

Summary & Outlook: Outlook

More about statistics: characteristic values

References

Data Science

Summary & Outlook: Endnotes

[1] http://www.go-fair.org/fair-principles/.
[2] https://de.wikipedia.org/wiki/Liste_der_größten_deutschen_Städte
[3] https://de.wikipedia.org/wiki/Dortmund
[4] https://www.medizininformatik-initiative.de/sites/default/files/2020-11/MII_WG-Consent_Patient-Consent-Form_v1.6d_engl-version.pdf
[5] https://en.wikipedia.org/wiki/AOL_search_log_release
[6] https://en.wikipedia.org/wiki/AOL_search_log_release
[7] https://de.wikipedia.org/wiki/Liste_der_Großstädte_in_Deutschland
[8] William Kruskal (1965). "STATISTICS, MOLIERE AND HENRY ADAMS": 80.
[9] https://en.wikipedia.org/wiki/Statistics
[10] https://de.wikipedia.org/wiki/Liste_der_Großstädte_in_Deutschland

Data Science

Summary & Outlook: Acknowledgement

Parts of the lecture base on the lecture ”Statistics” (FH Dortmund)
by
Prof. Dr. Sonja Kuhnt and Prof. Dr. Nadja Bauer.