Fachhochschule
Dortmund
University of Applied Sciences and Arts

Data Science

05: Central tendencies

Klaus Kaiser
WiSe 2024 / 2035

Data Science

Recap: 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

Recap Anonymization

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.

Data Science

Recap Types of statistics

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 probability theory.

Data Science

Recap Frequencies

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

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

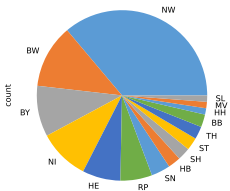
Furthermore, we call

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

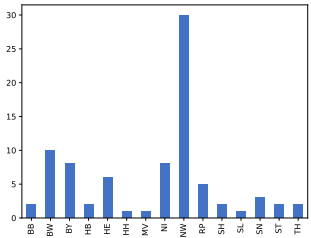
Data Science

Recap Frequencies

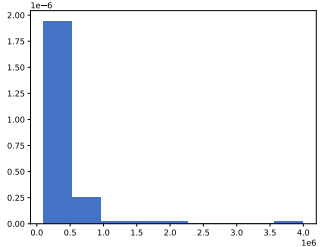
There are several ways to present frequencies with the help of plots.



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

Today

Histogram and **empiric distribution function** + Computing and plotting of **characteristic values** of a set of data!

Frequencies & Histogram

Frequencies & Histogram

Plotting frequencies

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:

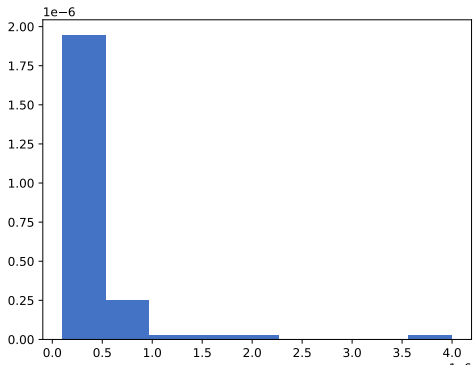
$$d_j \cdot g_j = d_j \frac{f_j}{d_j} = f_j$$

- 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 & Histogram Frequencies

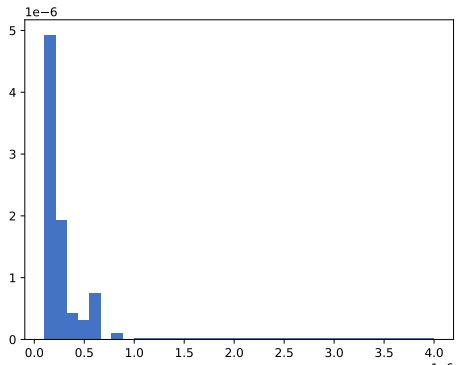
j	c_{j-1}	c_j	h_j	f_j	g_j
1	100000	533333	70	0.84	$1.95 \cdot 10^{-06}$
2	533333	966666	9	0.11	$2.50 \cdot 10^{-07}$
3	966666	1400000	1	0.01	$2.78 \cdot 10^{-08}$
4	1400000	1833333	1	0.01	$2.78 \cdot 10^{-08}$
5	1833333	2266666	1	0.01	$2.78 \cdot 10^{-08}$
6	2266666	2700000	0	0.0	0.0
7	2700000	3133333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

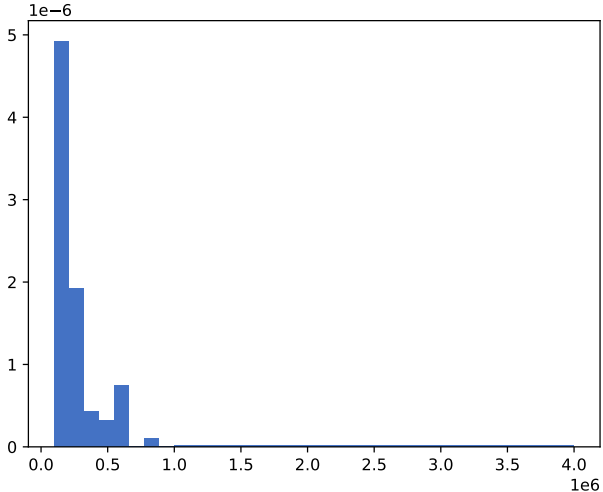
Frequencies & Histogram Frequencies

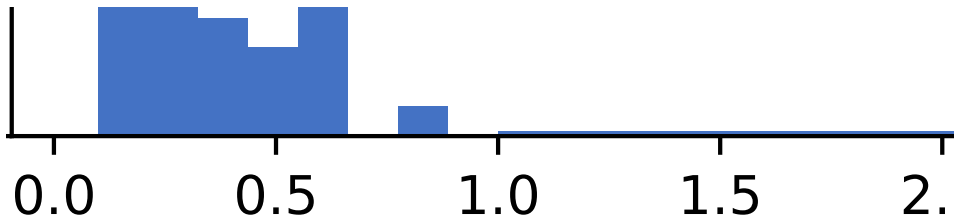
j	c_{j-1}	c_j	h_j	f_j	g_j
1	100000	212500	46	0.55	$4.93 \cdot 10^{-06}$
2	212500	325000	18	0.22	$1.92 \cdot 10^{-06}$
3	325000	437500	4	0.05	$4.28 \cdot 10^{-07}$
4	437500	550000	3	0.04	$3.21 \cdot 10^{-07}$
5	550000	662500	7	0.08	$7.50 \cdot 10^{-07}$
6	662500	775000	0	0.0	0.0
7	775000	887500	1	0.01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



Data Science

Frequencies & Histogram Frequencies




0.0
0.5
1.0
1.5
2.

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 & Histogram

Empirical Distribution Function

Data Science

Frequencies & Histogram 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[1].

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

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

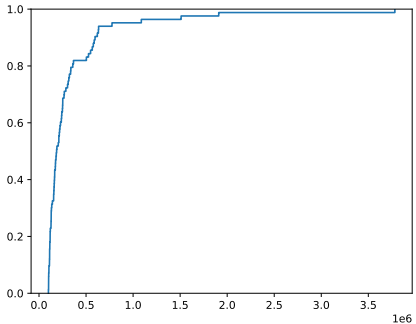
and the **empirical distribution function** $F_n : \mathbb{R} \to [0, 1]$ as

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

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

Data Science

Frequencies & Histogram Motivation



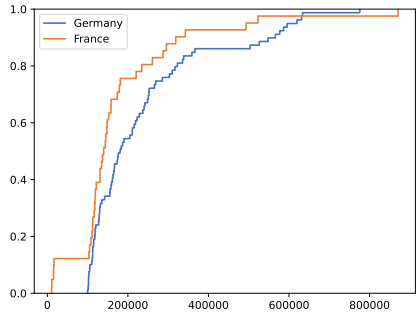


Figure: Empirical distribution function of large cities in Germany (left) and large cities which are no metropolis in Germany and France (right)

Central tendencies

To describe a set of observations it could be useful to reduce in on one or few characteristic values - or **central tendencies**.

- Which value is the most common?
- Which value is the one in the middle?
- Which value is the averaged one?

Depending on the observations and question different values could be interesting.

Data Science

Central tendencies Motivation

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

In the case of an ordinal or metric variable, we define $x_{(i)}$ to be the i^{th} element of the observations ordered by value, i.e.,

$$x_{(1)} \leq \cdots \leq x_{(i)} \leq x_{(i+1)} \leq \cdots \leq x_{(n)}$$

- $x_{(i)}$ is also referred as the i^{th} rank value

Central tendencies

Mode, Median and Mean

Data Science

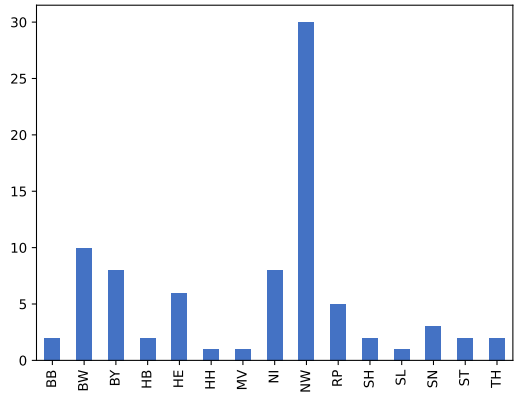
Mode, Median and Mean Motivation

	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
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Task: Which state stands out and why?

Data Science

Mode, Median and Mean Motivation



Task: Which state stands out and why?

We call a value a_j **mode** or **modal,** if it fulfills

$$h(a_j) \geq h(a_i) \text{ for all } i = 1, \ldots, l$$

- The mode is the value with the largest frequency
- There could be multiple modes in a set of observations
- The mode is especially interesting for nominal data

Task: What is the modal value in the following examples?

Blood group of patients in a therapy (nominal)

A A B 0 AB 0 0 A B A

Clothing sizes in a group (ordinal)

XL L S S M M M L XS L

Daily temperature in °C at 10 o'clock (metric)

11.2 13.3 14.1 13.7 12.2 11.3 9.9

Data Science

Mode, Median and Mean Motivation

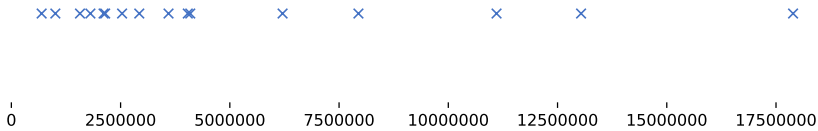
state	population
Baden-Württemberg	11104040
Bayern	13038724
Berlin[1]	3596999
Brandenburg	2534075
Bremen	693204
Hamburg	1808846
Hessen	6207278
Mecklenburg-Vorpommern	1570817
Niedersachsen	7943265
Nordrhein-Westfalen	17890489
Rheinland-Pfalz	4094169
Saarland	1006864
Sachsen	4038131
Sachsen-Anhalt	2146443
Schleswig-Holstein	2927542
Thüringen	2110396

state	population
Nordrhein-Westfalen	17890489
Bayern	13038724
Baden-Württemberg	11104040
Niedersachsen	7943265
Hessen	6207278
Rheinland-Pfalz	4094169
Sachsen	4038131
Berlin[1]	3596999
Schleswig-Holstein	2927542
Brandenburg	2534075
Sachsen-Anhalt	2146443
Thüringen	2110396
Hamburg	1808846
Mecklenburg-Vorpommern	1570817
Saarland	1006864
Bremen	693204

Table: Population of all federal states of Germany in 2022[2]

Data Science

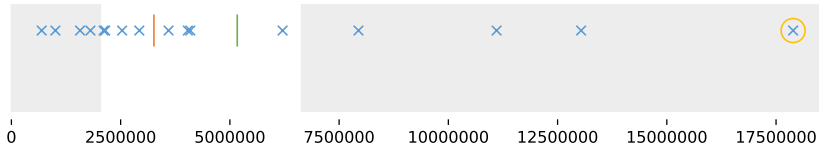
Mode, Median and Mean Motivation



Task: Which values and areas might be interesting - how can we describe the data?

Data Science

Mode, Median and Mean Motivation

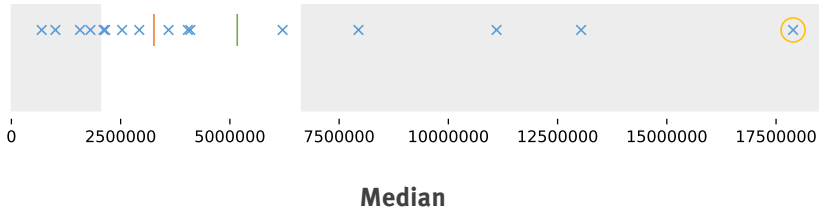


Task: Which values and areas might be interesting - how can we describe the data?

- **Mean-value:** What is the mean population of the states (green)
- **Median-value:** Which population splits the data in 50 percent less and 50 percent more inhabitants (orange)
- **Quantile:** which are the lower and upper 25 percent of the values (gray areas)?
- **Outlier:** Which values are extreme (yellow circle)?

Data Science

Mode, Median and Mean Motivation



$$x_1 \leq x_2 \leq x_3 \leq x_4 \leq x_5 \leq x_6 \leq x_7 \leq x_8 \leq x_9 \leq x_{10} \leq x_{11} \leq x_{12} \leq x_{13} \leq x_{14} \leq x_{15} \leq x_{16} \leq x_{17}$$

$$\underbrace{x_1 \leq x_2 \leq x_3 \leq x_4 \leq x_5 \leq x_6 \leq x_7 \leq x_8}_{\approx 50\%} \leq x_9 \leq \underbrace{x_{10} \leq x_{11} \leq x_{12} \leq x_{13} \leq x_{14} \leq x_{15} \leq x_{16} \leq x_{17}}_{\approx 50\%}$$

Median

Value in the Middle of the data - approximately half of the data is less or equal to this value and approximately half of the data is larger or equal to this value.

We define x_{med} as the **Median** given by

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

- X must be at least ordinal, otherwise the observations could not be ordered.
- If X is ordinal, the case n even could not be computed, then one could choose $x_{\left(\frac{n}{2}\right)}$ or $x_{\left(\frac{n}{2}+1\right)}$ as the value of x_{med}

Median is invariant concerning linear transformations. This means that for any $a, b \in \mathbb{R}$ with $a > 0$ and a sample $x_1, \ldots x_n$ of the variable X there holds

$$\text{if } y_i = ax_i + b \text{ then } y_{med} = ax_{med} + b$$

Task: What is the median value in the following examples?

Blood group of patients in a therapy (nominal)

A A B 0 AB 0 0 A B A

Clothing sizes in a group (ordinal)

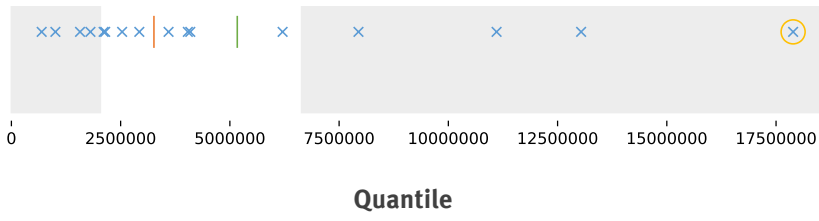
XL L S S M M M L XS L

Daily temperature in °C at 10 o'clock (metric)

11.2 13.3 14.1 13.7 12.2 11.3 9.9

Data Science

Mode, Median and Mean Quantile



$$x_1 \leq x_2 \leq x_3 \leq x_4 \leq x_5 \leq x_6 \leq x_7 \leq x_8 \leq x_9 \leq x_{10} \leq x_{11} \leq x_{12} \leq x_{13} \leq x_{14} \leq x_{15} \leq x_{16} \leq x_{17}$$

$$\underbrace{x_1 \leq x_2 \leq x_3 \leq x_4}_{\approx 25\%} \leq x_5 \leq \underbrace{x_6 \leq x_7 \leq x_8 \leq s_9 \leq x_{10} \leq x_{11} \leq x_{12}}_{\approx 50\%} \leq x_{13} \leq \underbrace{x_{14} \leq x_{15} \leq x_{16} \leq x_{17}}_{\approx 25\%}$$

p-Quantile

Value which divides the dataset into two parts, where p portion of the data is less or equal to this value and $1 - p$ portion of the data is larger or equal to this value.

We define x_p with $0 < p < 1$ as the ***p*-quartile** given by

$$\begin{cases} x_{(\lfloor n \cdot p \rfloor + 1)} & \text{if } n \cdot p \notin \mathbb{N} \\ \frac{1}{2}\left(x_{(n \cdot p)} + x_{(n \cdot p + 1)}\right) & \text{if } n \cdot p \in \mathbb{N} \end{cases}$$

where $\lfloor a \rfloor$ denotes the largest natural number which is smaller than a.

- The median equals to the 0.5-quartile: $x_{med} = x_{0.5}$
- A p-quantile divides the data into a p and $1 - p$ portion.

Special p-quantiles

- $x_{0.25}$ is denoted as **lower quartile**
- $x_{0.75}$ is denoted as **upper quartile**
- $x_{0.1}$ is denoted as **lower decile**
- $x_{0.9}$ is denoted as **upper decile**

Furthermore, a p-quantile can be used to compute additional central tendencies

- $\frac{1}{2}(x_p + x_{1-p})$ is denoted as **p-quantile middle**
- $\frac{1}{2}(x_{0.25} + x_{0.75})$ is denoted as **quartile middle**
- $\frac{1}{2}(x_{0.1} + x_{0.9})$ is denoted as **decile middle**

Data Science

Mode, Median and Mean p-Quantil

Task: What is the 0.25-quantil value in the following examples?

Blood group of patients in a therapy (nominal)

A A B 0 AB 0 0 A B A

Clothing sizes in a group (ordinal)

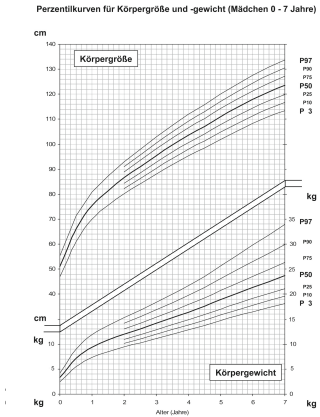
XL L S S M M M L XS L

Daily temperature in °C at 10 o'clock (metric)

11.2 13.3 14.1 13.7 12.2 11.3 9.9

Data Science

Mode, Median and Mean p-Quantil Example



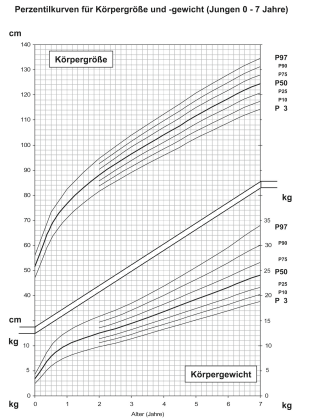
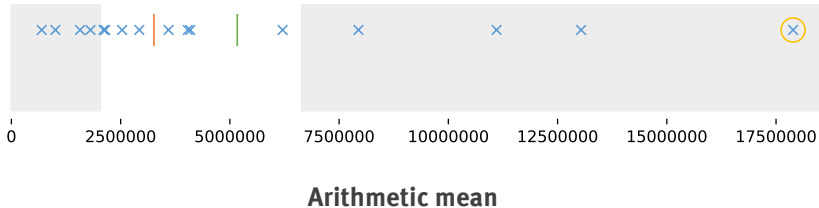


Figure: Growth curves given in the ”U-Heft” called examination booklet of children[3].

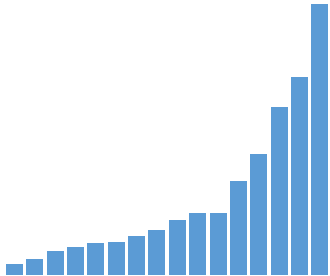
Data Science

Mode, Median and Mean Arithmetic Mean



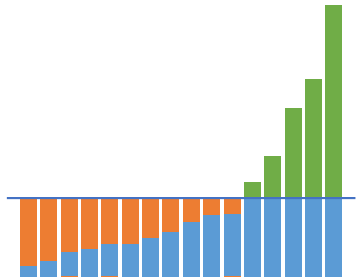
Data Science

Mode, Median and Mean Arithmetic Mean



Data Science

Mode, Median and Mean Arithmetic Mean



Arithmetic Mean

Value averaging the data - value which an observation take if every observation has the same value while keeping the sum.

We define $\bar{x}$ as the **arithmetic mean** given by

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

- The arithmetic mean is only useful for metric variables

Let $x_1, \ldots, x_n$ and $y_1, \ldots, y_m$ be observations of two metric variables X and Y. Then

- gives the arithmetic mean the centroid of the observations $\sum_{i=1}^{n} (x_i - \bar{x}) = 0$
- takes the function $f(z) = \sum_{i=1}^{n} (x_i - z)^2$ its minimum in $z_{min} = \bar{x}$
- for a linear transformation $z_i = ax_i + b$ with $i = 1, \ldots, n$, $a, b \in \mathbb{R}$ there holds $\bar{z} = a\bar{x} + b$
- for $z_i = x_i + y_i$ with $i = 1, \ldots, n$ there holds $\bar{z} = \bar{x} + \bar{y}$

Data Science

Mode, Median and Mean Arithmetic mean

Task: What is the arithmetic mean value in the following examples?

Blood group of patients in a therapy (nominal)

A A B 0 AB 0 0 A B A

Clothing sizes in a group (ordinal)

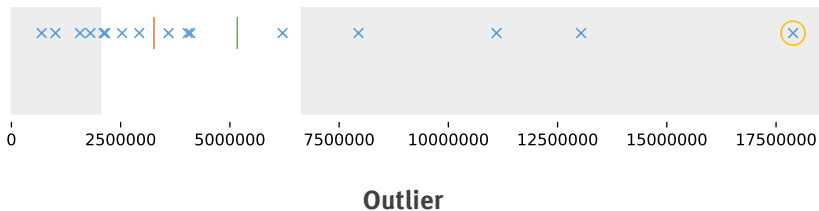
XL L S S M M M L XS L

Daily temperature in °C at 10 o'clock (metric)

11.2 13.3 14.1 13.7 12.2 11.3 9.9

Data Science

Mode, Median and Mean Outlier



A value larger than

$$z_{upper} = x_{0.75} + \frac{3}{2}(x_{0.75} - x_{0.25})$$

or lower than

$$z_{lower} = x_{0.25} - \frac{3}{2}(x_{0.75} - x_{0.25})$$

is often referred as an **outlier**.

- Outlier could have different reasons, e.g. by chance, measurement errors, novel data, ...

Arithmetic mean

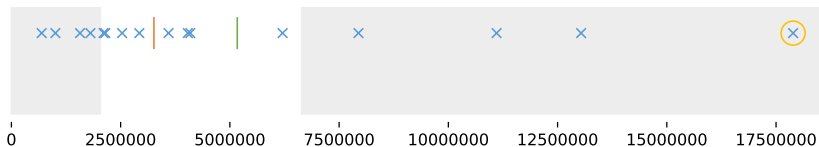
The arithmetic mean is very sensitive concerning outlier - every observation has the same weight $1/n$

Median

The median is robust concerning ”outlier” in an observation, i.e. an outlier has only a small influence on the Median value.

Data Science

Mode, Median and Mean Outlier



Overview

Types of data

	nominal	ordinal	metric
Mode	✓	✓	(✓)
Median / quartile	✗	✓	✓
Arithmetic mean	✗	✗	✓

- Using the mode value for metric data is not recommended

Outlier

	Mode	Median / quartile	Arithmetic mean
Robustness	✓	✓	✗

Data Science

Mode, Median and Mean Example

data	**mean**	**median**	0.25**-quartile**	0.75**-quantile**
Federal states	5169455.13	3262270.5	2035008.5	6641274.75
Large cities	330394.33	187119	128619	313420.5

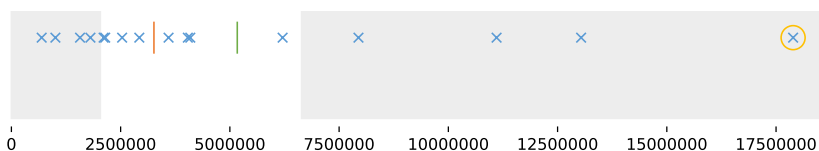
Central tendencies

Box plots

Data Science

Box plots Motivation

We can compute different describing central tendencies of data. Is there a way to visualize the data with the help of these tendencies?



The simple box plot (whisker plot) consists of

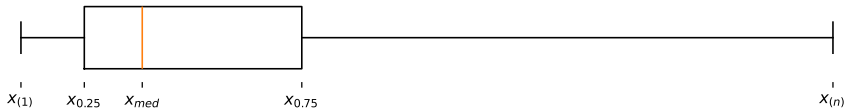
- a box reaching from the upper ($x_{0.25}$) to the lower quartile ($x_{0.75}$)
- horizontal lines at the lowest ($x_{(1)}$) and highest ($x_{(n)}$) values
- whisker lines connecting the lowest value with the lower quartile and the highest value with the upper quantile
- a horizontal line in the box at the median value

A box plot displays the five number summary of a set of data: minimum, first quartile, median, third quartile and maximum.

- A box plot can be horizontal or vertical represented
- The box contains the middle 50% of the data

Data Science

Box plots Box plots



Data Science

Box plots Box plots

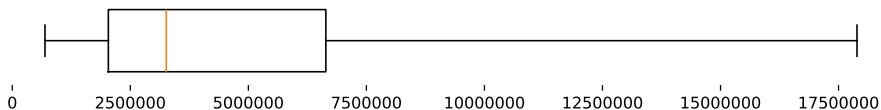
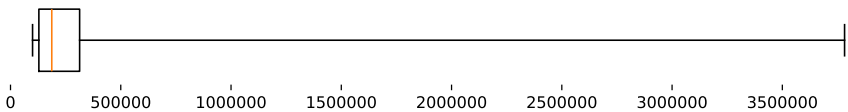


Figure: Box plot of the population of the federal states of Germany

Data Science

Box plots Box-plots

	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
⋮	⋮	⋮	⋮	⋮	⋮	⋮



Alternatively, the box-plot can also take outliers into account. Then, the horizontal lines at the largest values is replaced by the largest value which is smaller than

$$z_{upper} = x_{0.75} + \frac{3}{2}\left(x_{0.75} - x_{0.25}\right)$$

and the smallest value is replaced by the smallest value which is larger than

$$z_{lower} = x_{0.25} - \frac{3}{2}\left(x_{0.75} - x_{0.25}\right)$$

Every observation x_i which fulfills $x_i \notin [z_{lower}, z_{upper}]$ is marked, e.g. with a star or circle.

Data Science

Box plots Box plots

Data Science

Box plots Box plots

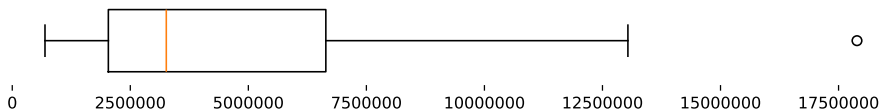
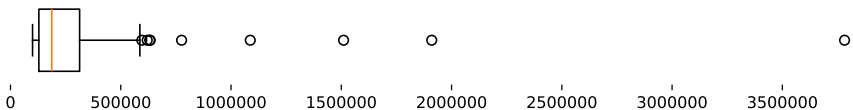


Figure: Box plot with outlier of the population of the federal states of Germany

Data Science

Box plots Box-plots

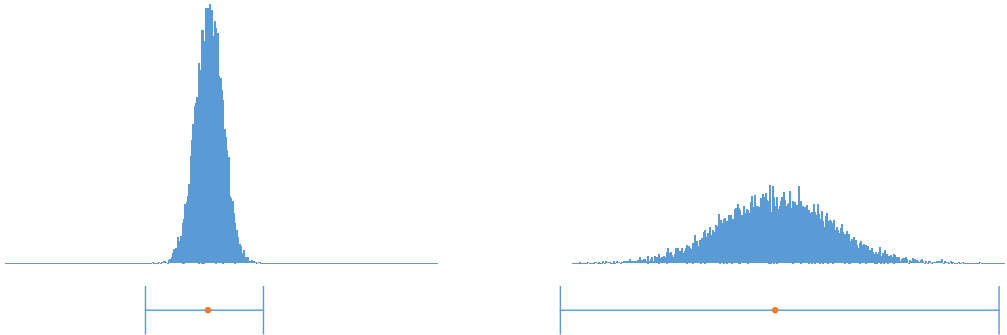
	Name	population	category	area	population / area	state
1	Berlin	3782202	metropolis	891.12	4244	BE
2	Hamburg	1910160	metropolis	755.09	2530	HH
⋮	⋮	⋮	⋮	⋮	⋮	⋮
8	Leipzig	619879	big city	297.8	2082	SN
9	Dortmund	595471	big city	280.71	2121	NW
⋮	⋮	⋮	⋮	⋮	⋮	⋮



Statistical dispersion

Data Science

Statistical dispersion Motivation



- Central tendencies, reduce the set of observations to one value
- Both sets of observations have the same mean value, but spread differently
- What is the distance between observations - how far do the observations spread around a defined center?

There are different kinds of statistical dispersion, which differ in which values are taken into account:

- Based on the difference between two central tendencies
- Based on the difference between observations and one central tendency
- relation between statistical dispersion and central tendency

The statistical dispersion helps to interpret the data:

- The larger the value, the more spread the values
- If the value is small, then the observations are more concentrated to one point

Let $x_1, \ldots, x_n$ be observations of a metric variable X. Then, we define the range r as the difference between the minimal $x_{(1)}$ and the maximal $x_{(n)}$ value:

$$r = x_{(n)} - x_{(1)}$$

- r only depends on the minimal and maximal value. All further information on $x_1, \ldots, x_n$ are lost.
- r is not robust concerning outlier!

Data Science

Statistical dispersion Range

Task: What is the range in the following example?

Daily temperature in °C at 10 o'clock (metric)

11.2 13.3 14.1 13.7 12.2 11.3 9.9

Data Science

Statistical dispersion Quartile range

Let $x_1, \ldots, x_n$ be observations of a metric variable X. Then, we define the quartile range (also interquartile range) qd as the difference between the upper quartile $x_{0.75}$ and lower quartile $x_{0.25}$ value:

$$qd = x_{0.75} - x_{0.25}$$

- qd is robust concerning outlier
- 50% of the "central" observations are given between $x_{0.25}$ and $x_{0.75}$

Data Science

Statistical dispersion Quartile range

Task: What is the quartile range in the following example?

Daily temperature in °C at 10 o'clock (metric)

11.2 13.3 14.1 13.7 12.2 11.3 9.9

Hint: $x_{0.25} = 11.25$ and $x_{0.75} = 13.5$.

Let $x_1, \ldots, x_n$ be observations of a metric variable X and $\bar{x}$ the corresponding arithmetic mean. Then we define the **empirical variance** of $x_1, \ldots, x_n$ as the mean squared deviation given by

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

Furthermore, we define

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

as the **empirical standard deviation**.

- The suffix *empirical* is used to differentiate it from the variance of a random variable (later!). The word shows, that the value was computed on concrete data.
- Often (especially in software products), the sampling variance, given by

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

  is preferred. Note that the difference is small for large n.
- The empirical variance can also be computed by

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

Data Science

Statistical dispersion Quartile range

Task: What is the empirical variance in the following example?

Daily temperature in °C at 10 o'clock (metric)

11.2 13.3 14.1 13.7 12.2 11.3 9.9

step	result
$\bar{x}$	12.24
$x_i - \bar{x}$	-1.04, 1.06, 1.86, 1.46, -0.04, -0.94, -2.34
$(x_i - \bar{x})^2$	1.09, 1.12, 3.45, 2.12, 0.00, 0.89, 5.49
$\sum_{i=1}^{n} (x_i - \bar{x})^2$	14.16
$\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$	2.02

Steiners theorem

Let $x_1, \ldots, x_n \in \mathbb{R}$, $a \in \mathbb{R}$. Then, there holds

$$\sum_{i=1}^{n} (x_i - a)^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 + n(\bar{x} - a)^2$$

Let $x_1, \ldots, x_n$ be observations of a metric variable X and $\tilde{s}_X^2$ the corresponding empirical variance. Furthermore, let $\tilde{s}_Y^2$ be the empirical variance of the variable Y which is given by the linear transformation

$$y_i = ax_i + b \text{ for } i \in \{1, \ldots, n\}$$

with $a, b \in \mathbb{R}$. Then, there holds:

$$\tilde{s}_Y^2 = a^2 \tilde{s}_X^2.$$

Data Science

Statistical dispersion Example

data	**mean**	**median**	0.25**-quartile**	0.75**-quantile**
Federal states	5169455.13	3262270.5	2035008.5	6641274.75
Large cities	330394.33	187119	128619	313420.5

data	**range**	**quartile range**	**empirical variance**	**empirical standard deviation**
Federal states	17197285	4606266.25	22795310650229.98	4774443.49
Large cities	3682192	184801.5	229138903806.70	478684.56

Summary & Outlook

Data Science

Summary & Outlook: Summary

- You know the definition of a histogram and can create one.
- You understand what central tendencies are and what they are used for
- You compute central tendencies to analyze a set of observations
- You can explain and create a box-plot of a set of observations
- You can compute different statistical deviations

Description of the correlation of two variables

References

Data Science

Summary & Outlook: Endnotes

[1] https://de.wikipedia.org/wiki/Liste_der_Großstädte_in_Deutschland
[2] https://de.wikipedia.org/wiki/Liste_der_deutschen_Bundesländer_nach_Bevölkerungsentwicklung
[3] https://www.g-ba.de/downloads/83-691-421/2016-08-04_Kinderuntersuchungsheft_WEB.pdf

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.