Lecture 1: Visualization of Data and Frequencies

Dr. L. Sanna Stephan

2023

RUG Groningen

Learning objectives

You will learn about

1. ... the **source** of data in economics, econometrics and OR and its **consequences** for data quality, subjectivity and ethical aspect and
2. ... ways to **characterize** a data-set.

LO1: Data Sources and Threats *(further reading: DAPEB Ch.1)*

- Sources of (Socio-economic) Data
- Things to Look Out for as a Literate Data Analyst
- Ethical Aspects

LO2: Characterising Data

- General Measures of the Data Set
- Characterising One Variable
  (further reading: DABEP 2.1, 3.1-3.8)
- Challenges
  (further reading: DABEP Ch. 2.8-2.12)
- Analysing Multiple Variables
  (further reading: DABEP Ch. 4.1, 4.2, 4.6)

LO1: Data Sources and Threats

(further reading: DAPEB Ch.1)

LO1: Data Sources and Threats

(further reading: DAPEB Ch.1)

Sources of (Socio-economic) Data

Data sources overview

- Administrative data
- Sensor data
- Data from/ for other research
- Web data
- Surveys
- Field or lab experiments

Administrative data (land/birth registry ...)

Cost: low (often open access)

Benefit: measured regularly, high quality

☠ **Threat:** sometimes not clear how a variable is defined or measured

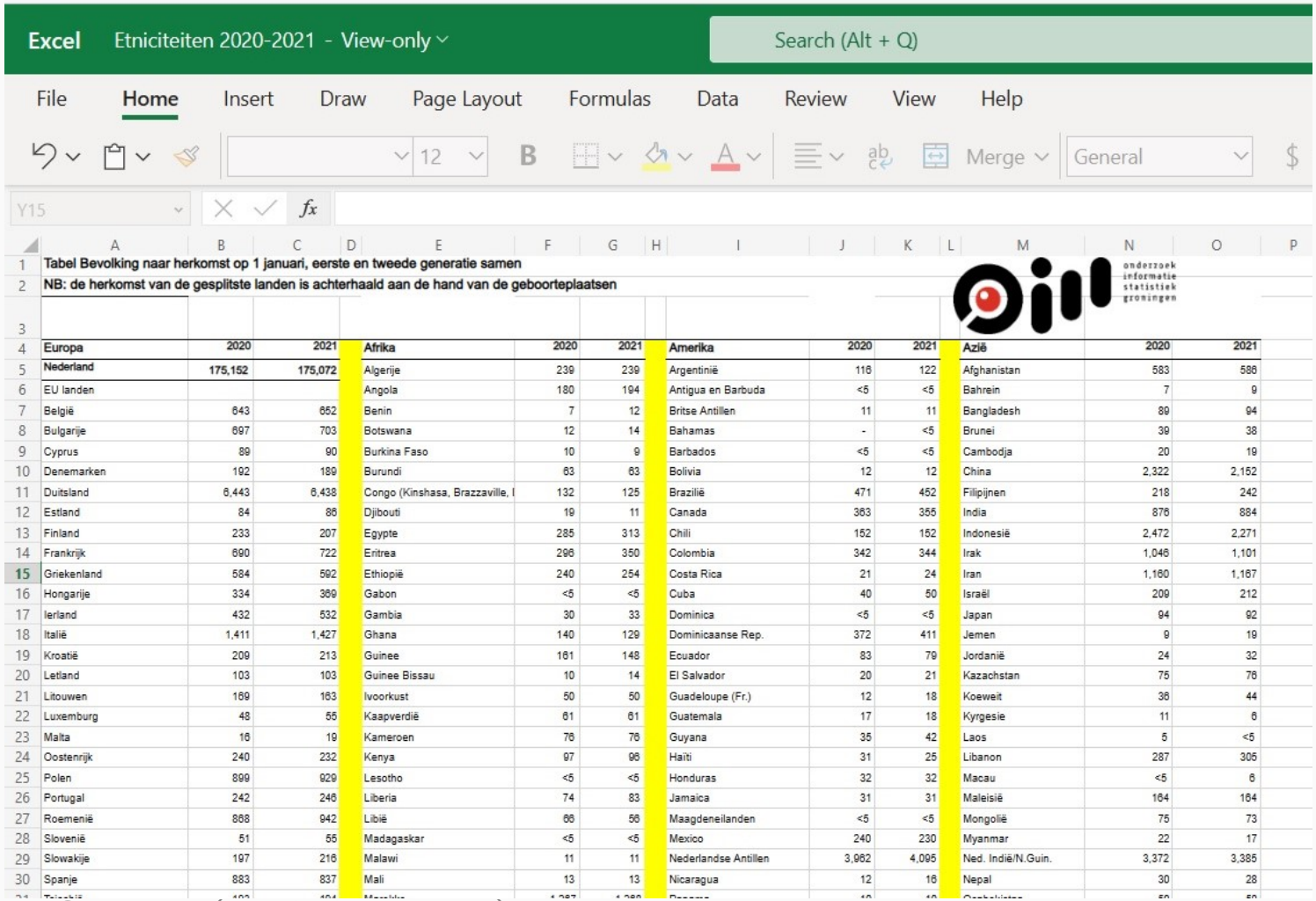
✓ **solution:** read documentation/ communicate with data owner

☠ **Threat:** not necessarily all variables needed are available

✓ **Solution:** merge data from different sources

Administrative data - example

Data from the municipality of Groningen: country of origin of registered inhabitants (https://data.groningen.nl)



Excel Etniciteiten 2020-2021 - View-only

Search (Alt + Q)

File Home Insert Draw Page Layout Formulas Data Review View Help

Tabel Bevolking naar herkomst op 1 januari, eerste en tweede generatie samen

NB: de herkomst van de gesplitste landen is achterhaald aan de hand van de geboorteplaatsen

onderzoek informatie statistiek groningen

Europa	2020	2021	Afrika	2020	2021	Amerika	2020	2021	Azië	2020	2021
Nederland	175,152	175,072	Algerije	239	239	Argentinië	116	122	Afghanistan	583	588
EU landen			Angola	180	194	Antigua en Barbuda	<5	<5	Bahrein	7	9
België	643	652	Benin	7	12	Britse Antillen	11	11	Bangladesh	89	94
Bulgarije	697	703	Botswana	12	14	Bahamas	-	<5	Brunei	39	38
Cyprus	89	90	Burkina Faso	10	9	Barbados	<5	<5	Cambodja	20	19
Denemarken	192	189	Burundi	63	63	Bolivia	12	12	China	2,322	2,152
Duitsland	6,443	6,438	Congo (Kinshasa, Brazzaville, I	132	125	Brazilië	471	452	Filipijnen	218	242
Estland	84	86	Djibouti	19	11	Canada	363	355	India	876	884
Finland	233	207	Egypte	285	313	Chili	152	152	Indonesië	2,472	2,271
Frankrijk	690	722	Eritrea	296	350	Colombia	342	344	Irak	1,046	1,101
Griekenland	584	592	Ethiopië	240	254	Costa Rica	21	24	Iran	1,160	1,167
Hongarije	334	369	Gabon	<5	<5	Cuba	40	50	Israël	209	212
Ierland	432	532	Gambia	30	33	Dominica	<5	<5	Japan	94	92
Italië	1,411	1,427	Ghana	140	129	Dominicaanse Rep.	372	411	Jemen	9	19
Kroatië	209	213	Guinee	161	148	Ecuador	83	79	Jordanië	24	32
Letland	103	103	Guinee Bissau	10	14	El Salvador	20	21	Kazachstan	75	76
Litouwen	169	163	Ivoorkust	50	50	Guadeloupe (Fr.)	12	18	Koeweit	36	44
Luxemburg	48	55	Kaapverdië	61	61	Guatemala	17	18	Kyrgesie	11	6
Malta	16	19	Kameroen	76	76	Guyana	35	42	Laos	5	<5
Oostenrijk	240	232	Kenya	97	96	Haiti	31	25	Libanon	287	305
Polen	899	929	Lesotho	<5	<5	Honduras	32	32	Macau	<5	6
Portugal	242	246	Liberia	74	83	Jamaica	31	31	Maleisië	164	164
Roemenië	868	942	Libië	66	56	Maagdeneilanden	<5	<5	Mongolië	75	73
Slovenië	51	55	Madagaskar	<5	<5	Mexico	240	230	Myanmar	22	17
Slowakije	197	216	Malawi	11	11	Nederlandse Antillen	3,962	4,095	Ned. Indië/N.Guin.	3,372	3,385
Spanje	883	837	Mali	13	13	Nicaragua	12	16	Nepal	30	28

Sensor data

Sensors can e.g. measure temperature, air quality or track movements of individuals or vehicles for operational planning in government and business.

Cost: low (normally freely available)

Benefit: data is big (frequency) and accurate

☠ **Threat:** only variables that can be measured by sensor available

✓ **Solution:** adjust research question, merge with other data

Sensor data - example

Sensor that counts movements on Cambridge Mill road to guide city development and public transport decisions (OGV=official government vehicle, LGV=local government vehicle).

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Local Time (Sens	Date	Time	countlineName	direction	Car	Pedestrian	Cyclist	Motorbike	Bus	OGV1	OGV2	LGV	
2	3/6/2019 1:00	3/6/2019	1:00:00	S1_MillRoad_CAM003	in	42	8	8	2	0	0	0	5	
3	3/6/2019 1:00	3/6/2019	1:00:00	S1_MillRoad_CAM003	out	21	4	1	0	0	0	0	1	
4	3/6/2019 2:00	3/6/2019	2:00:00	S1_MillRoad_CAM003	in	21	5	2	0	0	0	0	1	
5	3/6/2019 2:00	3/6/2019	2:00:00	S1_MillRoad_CAM003	out	32	4	0	0	0	0	0	0	
6	3/6/2019 3:00	3/6/2019	3:00:00	S1_MillRoad_CAM003	in	57	1	3	0	0	0	0	2	
7	3/6/2019 3:00	3/6/2019	3:00:00	S1_MillRoad_CAM003	out	18	2	0	1	0	0	0	0	
8	3/6/2019 4:00	3/6/2019	4:00:00	S1_MillRoad_CAM003	in	17	1	2	0	0	1	0	1	
9	3/6/2019 4:00	3/6/2019	4:00:00	S1_MillRoad_CAM003	out	23	4	2	0	0	0	0	0	
10	3/6/2019 5:00	3/6/2019	5:00:00	S1_MillRoad_CAM003	in	29	2	2	1	0	0	0	5	
11	3/6/2019 5:00	3/6/2019	5:00:00	S1_MillRoad_CAM003	out	32	9	17	1	0	1	0	4	
12	3/6/2019 6:00	3/6/2019	6:00:00	S1_MillRoad_CAM003	in	87	3	18	3	2	5	0	9	
13	3/6/2019 6:00	3/6/2019	6:00:00	S1_MillRoad_CAM003	out	93	21	45	5	1	1	1	11	
14	3/6/2019 7:00	3/6/2019	7:00:00	S1_MillRoad_CAM003	in	157	16	43	0	2	3	1	22	
15	3/6/2019 7:00	3/6/2019	7:00:00	S1_MillRoad_CAM003	out	180	41	71	4	3	7	2	53	
16	3/6/2019 8:00	3/6/2019	8:00:00	S1_MillRoad_CAM003	in	205	40	86	2	9	2	1	29	
17	3/6/2019 8:00	3/6/2019	8:00:00	S1_MillRoad_CAM003	out	210	80	145	1	9	3	0	52	
18	3/6/2019 9:00	3/6/2019	9:00:00	S1_MillRoad_CAM003	in	162	39	54	2	6	2	2	47	
19	3/6/2019 9:00	3/6/2019	9:00:00	S1_MillRoad_CAM003	out	197	72	82	4	8	5	0	45	
20	3/6/2019 10:00	3/6/2019	10:00:00	S1_MillRoad_CAM003	in	158	19	29	1	6	4	1	52	
21	3/6/2019 10:00	3/6/2019	10:00:00	S1_MillRoad_CAM003	out	169	68	49	3	9	11	0	42	
22	3/6/2019 11:00	3/6/2019	11:00:00	S1_MillRoad_CAM003	in	162	14	31	5	9	9	1	41	
23	3/6/2019 11:00	3/6/2019	11:00:00	S1_MillRoad_CAM003	out	186	47	26	0	8	4	0	39	
24	3/6/2019 12:00	3/6/2019	12:00:00	S1_MillRoad_CAM003	in	198	27	37	4	7	2	1	44	
25	3/6/2019 12:00	3/6/2019	12:00:00	S1_MillRoad_CAM003	out	195	63	54	5	6	2	2	45	
26	3/6/2019 13:00	3/6/2019	13:00:00	S1_MillRoad_CAM003	in	232	29	38	5	7	2	1	35	
27	3/6/2019 13:00	3/6/2019	13:00:00	S1_MillRoad_CAM003	out	223	64	46	7	10	1	1	34	

Data from/ for other research

Cost: low (often open access)

Benefit: often high quality (scientific standard)

☠ **Threat:** not necessarily all variables we need are available

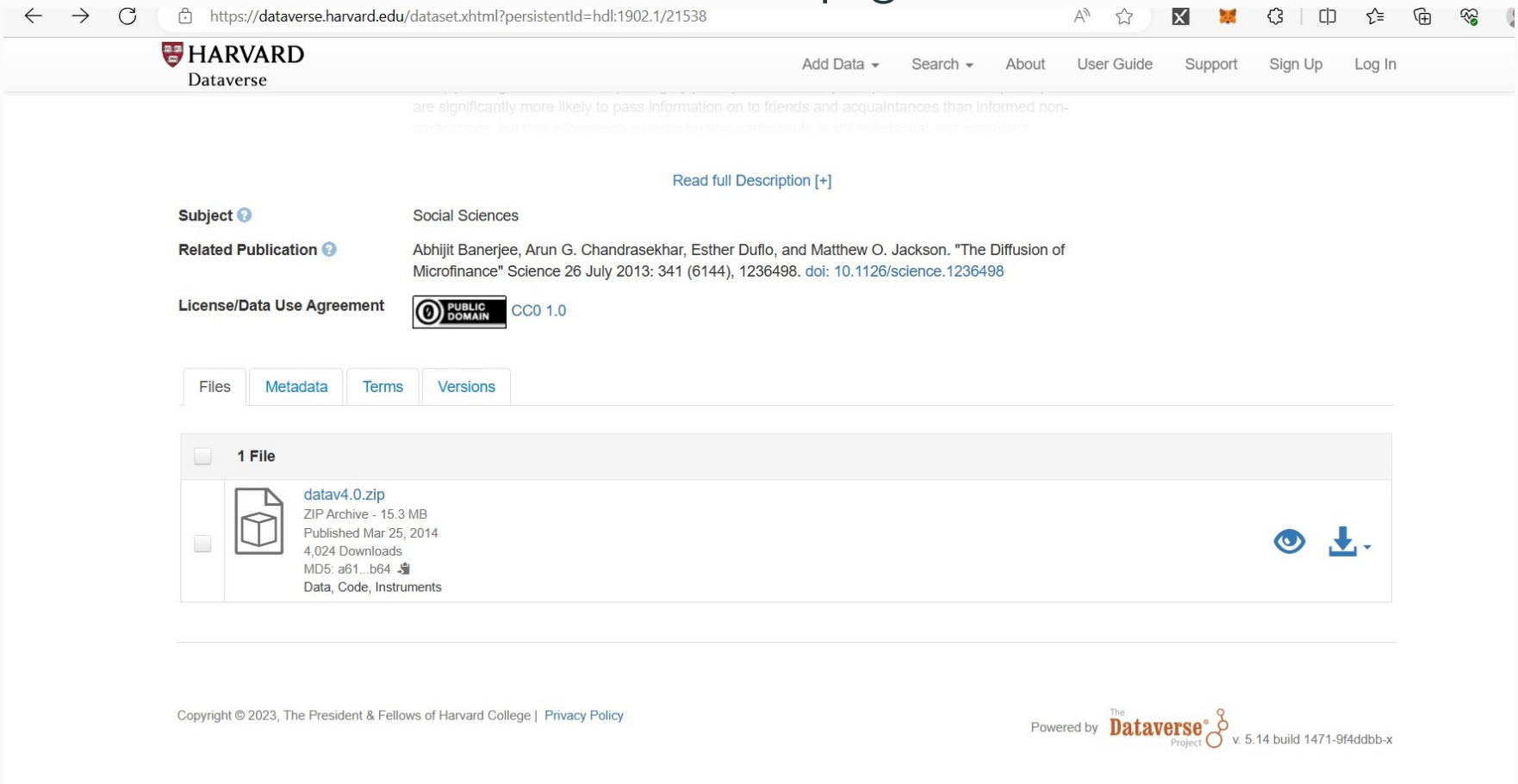
✓ **Solution:** merge data from different sources

☠ **Threat:** data may have already been studied extensively

✓ **Solution:** creativity (find new research question)

Data from/ for other research - example

You can download data from many papers on the University's or the author's page



Data from the web

Web scraping (FB, youtube, bitcoin prices ...) e.g. automated data collection from the web

Cost: low (software easy and free)

Benefit: often big (high frequency, many variables, many units of observation)

☠ **Threat:** only suitable for web-based interaction

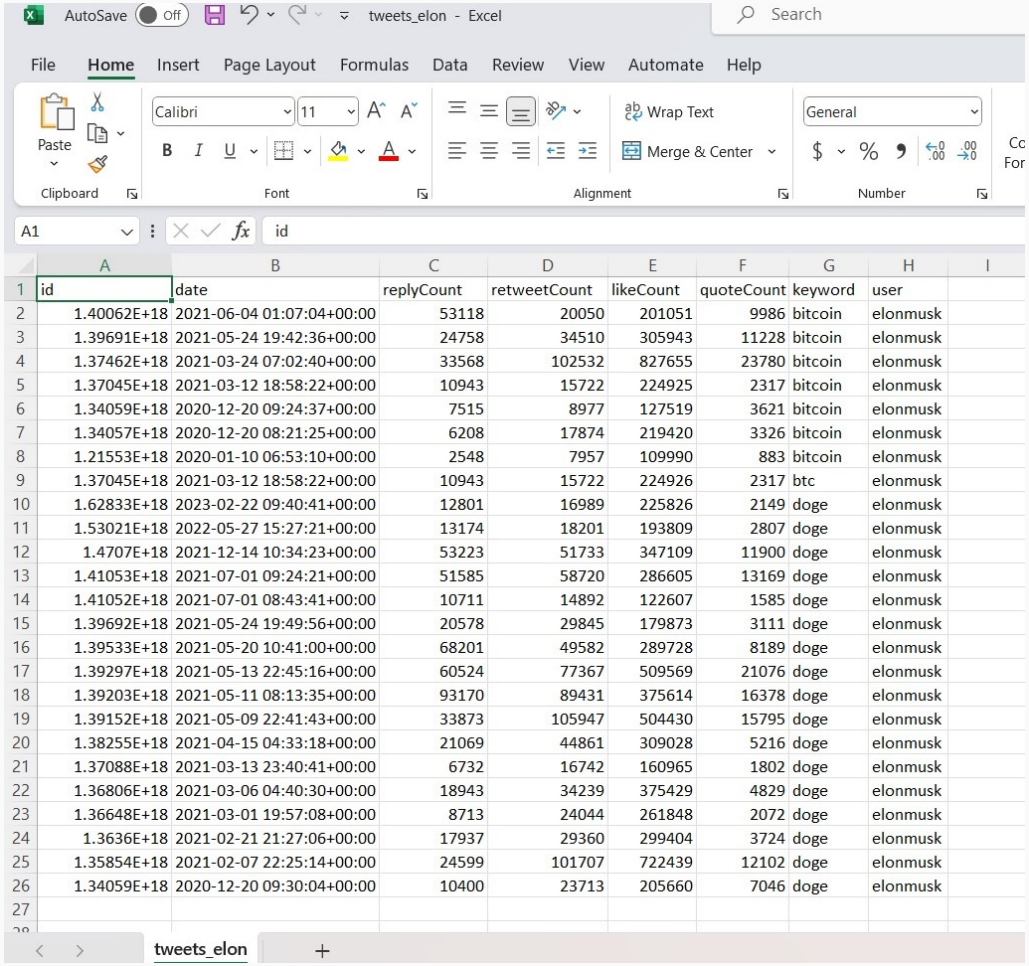
✓ **Solution:** adjust research question

☠ **Threat:** blindly scraping can entail error (e.g. sarcastic negative tweet recorded as positive)

✓ **Solution:** more sophisticated machine learning (e.g. sarcasm detection)

Data from the web - example

Tweets from Elon Musk scraped by a Bsc student



	A	B	C	D	E	F	G	H
1	id	date	replyCount	retweetCount	likeCount	quoteCount	keyword	user
2	1.40062E+18	2021-06-04 01:07:04+00:00	53118	20050	201051	9986	bitcoin	elonmusk
3	1.39691E+18	2021-05-24 19:42:36+00:00	24758	34510	305943	11228	bitcoin	elonmusk
4	1.37462E+18	2021-03-24 07:02:40+00:00	33568	102532	827655	23780	bitcoin	elonmusk
5	1.37045E+18	2021-03-12 18:58:22+00:00	10943	15722	224925	2317	bitcoin	elonmusk
6	1.34059E+18	2020-12-20 09:24:37+00:00	7515	8977	127519	3621	bitcoin	elonmusk
7	1.34057E+18	2020-12-20 08:21:25+00:00	6208	17874	219420	3326	bitcoin	elonmusk
8	1.21553E+18	2020-01-10 06:53:10+00:00	2548	7957	109990	883	bitcoin	elonmusk
9	1.37045E+18	2021-03-12 18:58:22+00:00	10943	15722	224926	2317	btc	elonmusk
10	1.62833E+18	2023-02-22 09:40:41+00:00	12801	16989	225826	2149	doge	elonmusk
11	1.53021E+18	2022-05-27 15:27:21+00:00	13174	18201	193809	2807	doge	elonmusk
12	1.4707E+18	2021-12-14 10:34:23+00:00	53223	51733	347109	11900	doge	elonmusk
13	1.41053E+18	2021-07-01 09:24:21+00:00	51585	58720	286605	13169	doge	elonmusk
14	1.41052E+18	2021-07-01 08:43:41+00:00	10711	14892	122607	1585	doge	elonmusk
15	1.39692E+18	2021-05-24 19:49:56+00:00	20578	29845	179873	3111	doge	elonmusk
16	1.39533E+18	2021-05-20 10:41:00+00:00	68201	49582	289728	8189	doge	elonmusk
17	1.39297E+18	2021-05-13 22:45:16+00:00	60524	77367	509569	21076	doge	elonmusk
18	1.39203E+18	2021-05-11 08:13:35+00:00	93170	89431	375614	16378	doge	elonmusk
19	1.39152E+18	2021-05-09 22:41:43+00:00	33873	105947	504430	15795	doge	elonmusk
20	1.38255E+18	2021-04-15 04:33:18+00:00	21069	44861	309028	5216	doge	elonmusk
21	1.37088E+18	2021-03-13 23:40:41+00:00	6732	16742	160965	1802	doge	elonmusk
22	1.36806E+18	2021-03-06 04:40:30+00:00	18943	34239	375429	4829	doge	elonmusk
23	1.36648E+18	2021-03-01 19:57:08+00:00	8713	24044	261848	2072	doge	elonmusk
24	1.3636E+18	2021-02-21 21:27:06+00:00	17937	29360	299404	3724	doge	elonmusk
25	1.35854E+18	2021-02-07 22:25:14+00:00	24599	101707	722439	12102	doge	elonmusk
26	1.34059E+18	2020-12-20 09:30:04+00:00	10400	23713	205660	7046	doge	elonmusk
27								

Surveys (customer satisfaction survey, voter survey ...)

Cost: high (design, implement, enter data requires personal)

Benefit: variables needed often not available elsewhere

☠ **Threat:** relies on the respondent's ability to understand/interpret the question

✓ **Solution:** careful design, supervise interviewers, small pilot survey first

☠ **Threat:** often voluntary participation

✓ **Solution:** careful design, stratified sampling

☠ **Threat:** may be influenced by the interviewer, the time, the survey instruments, the setting

✓ **Solution:** careful design, cross checking with other data

Survey - examples

Survey questionnaire to measure individuals food choices (www.examples.com)

Food Questionnaire

Eating Out	Daily	4-5 times a week	2-3 times a week	Once a week	Rarely
How often do you eat out for breakfast?					
How often do you eat out for lunch?					
How often do you eat out for dinner?					
If you answered at least once per week for any of the questions above, then please answer the following questions					
How often do you eat at buffets?					
How often do you eat at "fast food chains"?					
How often do you eat at a "sit down" restaurant?					

How long does it take for you to eat a typical meal? _____ minutes

Do you pay attention or monitor your portion sizes? Yes No

Do you snack? Yes No

Have you ever been told you needed to lose weight by your physician? Yes No

Do you currently want to try to lose weight? Yes No

Lab or field experiments

Lab experiment: participants play game that mimics an economic situation.

Field experiment: a treatment variable is randomly assigned (e.g. drug/technology/scholarship...)

Cost: high (hire and compensate participants)

Benefit: enables design and randomization

☠ **Threat:** can not mimic all economic interactions in lab

✓ **Solution:** careful choice of research question

☠ **Threat:** field experiment can be difficult to justify on ethical grounds

✓ **Solution:** careful consideration of necessity

Online lab experiments - example

Amazon mechanical turk offers the possibility to run experiments online

LO1: Data Sources and Threats

(further reading: DAPEB Ch.1)

Things to Look Out for as a Literate Data Analyst

Threats

1. (Unintentional) data misinterpretation and misuse,
2. Intentional selective data usage and presentation,
3. Measurement error.

1. (Unintentional) data misinterpretation and misuse

Causes:

- insufficient feedback between survey design/ implementation and data analysis and
- insufficient understanding of data and context.

1. (Unintentional) data misinterpretation and misuse

Causes:

- insufficient feedback between survey design/ implementation and data analysis and
- insufficient understanding of data and context.

Data analyst should

... be involved in survey implementation,

... communicate with data owner and

... do background research on survey setting/variable definition.

1. (Unintentional) data misinterpretation and misuse

Example 1: field study to assess child well-being in rural Senegal.




1. (Unintentional) data misinterpretation and misuse

Example 1: field study to assess child well-being in rural Senegal.

Intended: how many adults take care of how many children?

1. (Unintentional) data misinterpretation and misuse

Example 1: field study to assess child well-being in rural Senegal.

Intended: how many adults take care of how many children?

Actual: "How many children are in your household?"

Senegalese reality: polygamy, multiple generations cohabit.

1. (Unintentional) data misinterpretation and misuse

Example 1: field study to assess child well-being in rural Senegal.

Intended: how many adults take care of how many children?

Actual: "How many children are in your household?"

Senegalese reality: polygamy, multiple generations cohabit.

Context and background matter!

1. (Unintentional) data misinterpretation and misuse

Example 2: marketing survey to assess how micro credit is used

Example 2: marketing survey to assess how micro credit is used

"Did you use the loan to (i) extend your current or (ii) start a new business?"

Example 2: marketing survey to assess how micro credit is used

"Did you use the loan to (i) extend your current or (ii) start a new business?"

Ambiguity: is starting shop/factory/branch/restaurant at a different location selling slightly different products/services (i) or (ii)?

Example 2: marketing survey to assess how micro credit is used

"Did you use the loan to (i) extend your current or (ii) start a new business?"

Ambiguity: is starting shop/factory/branch/restaurant at a different location selling slightly different products/services (i) or (ii)?

Choice matters for subsequent interpretation.

2. Intentional selective data usage

"Let the data speak..."

creative intervention in **data collection**, **data cleaning** and **data mining** in order to generate desired "evidence".

2. Intentional selective data usage: data collection

Vary survey **time or location.**

2. Intentional selective data usage: data collection

Vary survey **time or location.**

- **Time:** course satisfaction before or after exam, street survey during peak or non-peak hours, survey intentions at beginning or middle of the year
- **Location:** survey consumers at Lidl versus Coop, university library versus university sports centre

2. Intentional selective data usage: data collection

Vary the **survey instruments (questions)**

- Leading question: "How do you experience working with this highly distinguished scholar?"
- Positive versus double negative: "Was the facility not unclean?"
- Defining categories: "How many hours do you work?"
  1. less than 40, more than 40
  2. 30-35,35-40,40-45, 45-50,50-55,55-60, more than 60
- And many more

2. Intentional selective data usage: data wrangling

With many variables and/or many observations, you can ...

... **select** the variables (e.g. products included in consumption basket to evaluate inflation?) or the time frame,

... **omit** variables to establish a false causality (e.g. lighters provoke cancer (omitted: light cigarette)).

Data cleaning can involve

... deleting outliers or other "erroneous observations",

... changing the scale or normalising.

2. Intentional selective data usage: data mining

Data mining: extracting/ discovering patterns in large data sets
You can choose

1. the statistical method used and
2. the complexity of the model.

Simple example: statements is correct, but looses power in context.

- "In Africa, in every generation, there are more smokers" (... "but the population is also growing").
- "Teenagers spend less and less time on YouTube" (... "but more time on TikTok").
- "Chinese are rich, because they have a large GDP" (... "but they are also numerous").

2. Intentional selective data usage

- "Fishing for results" is unfortunately widespread.
- It is at best bad practice and at worst dangerous misinformation.
- There are many ways to do so
- The world is complex
- Data analysis is a difficult task requiring **context**, **theory** and careful (tedious) **cross checking** and **documentation/justification** of data manipulation

3. Measurement error

Even with careful design, errors in data accumulation occur.
Typical sources of inaccuracies:

- measurement error (human or technical),
- misunderstanding/ mis-interpretation by respondent.

LO1: Data Sources and Threats

(further reading: DAPEB Ch.1)

Ethical Aspects

Ethics

Data accumulation:

- Privacy/ personal data ownership/ sensitive data
- Who has the right to use it and for what purpose?

Drawing conclusions and presenting them:

- Using the lack of statistical literacy of the audience
- Deliberate sloppy or selective reporting

LO2: Characterising Data

LO2: Characterising Data

General Measures of the Data Set

General measures of the data-set

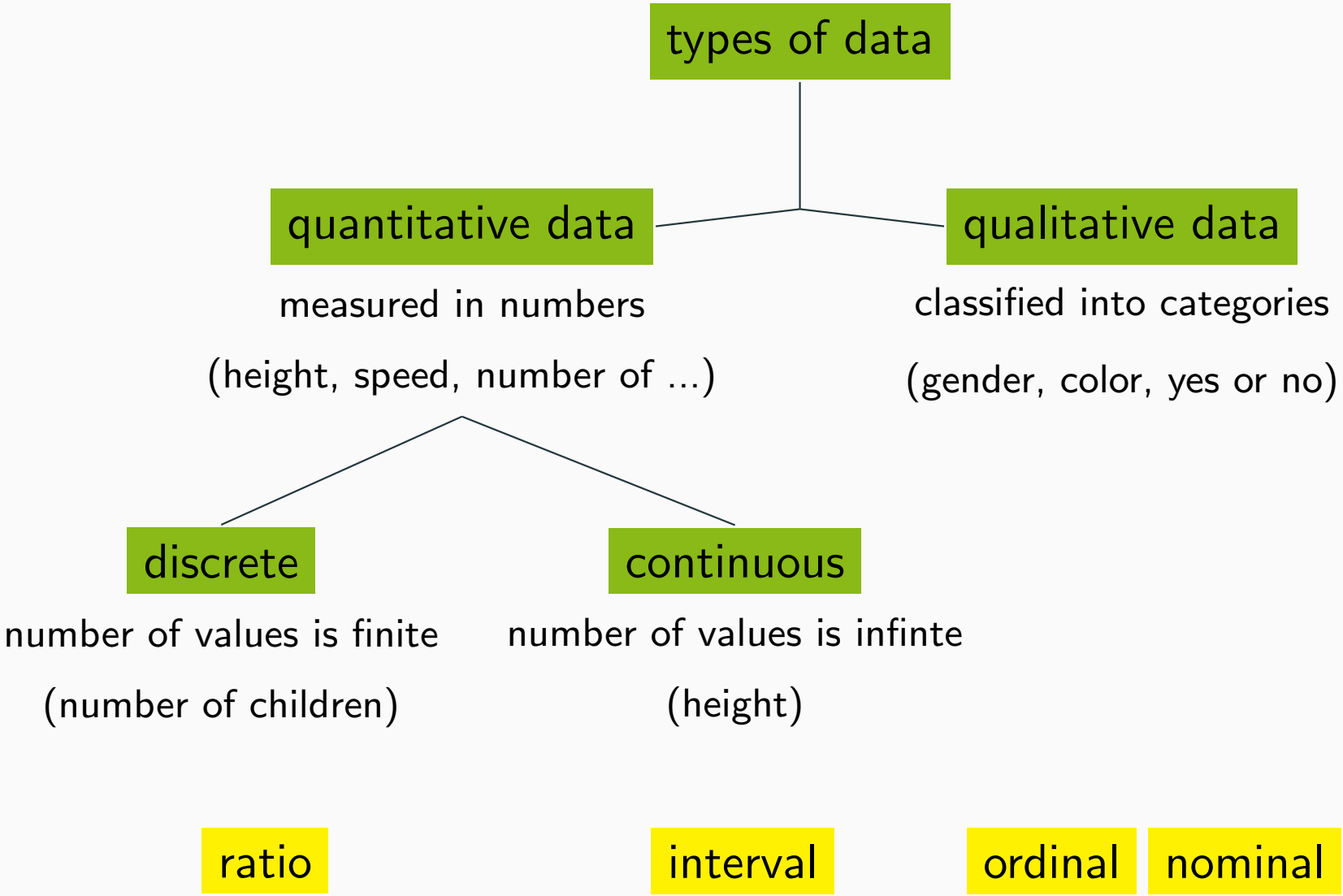
- Size (number of observations)
  - If measurements are taken at multiple points in time, we distinguish the unit of observation (individual, country...) and the points in time.
  - Then there are two sizes: the individual dimension (N) and the time dimension (T)
- Number of variables (measurements for each observation)

LO2: Characterising Data

Characterising One Variable

(further reading: DABEP 2.1, 3.1-3.8)

Types and scales of data



The use of mathematical formulas

- Use standard measures to characterise data sets $\Rightarrow$ easy comparison
- Mathematics: precise, concise language to communicate data set features

You are now learning about...

- Measures of central tendency:
  1. mean (average),
  2. median (separates the lower half from the upper half),
  3. mode (most frequent observation).
- Measures of dispersion:
  1. variance,
  2. inter-quartile range.
- Ways to illustrate the distribution of the data:
  1. frequencies,
  2. histograms.
- Challenges when analysing the above and
- Data transformations and their effects.

Measures of central tendency

For quantitative variables, we can analyse different measures of central tendency (aim: describe the center of the data).

The **mean**:

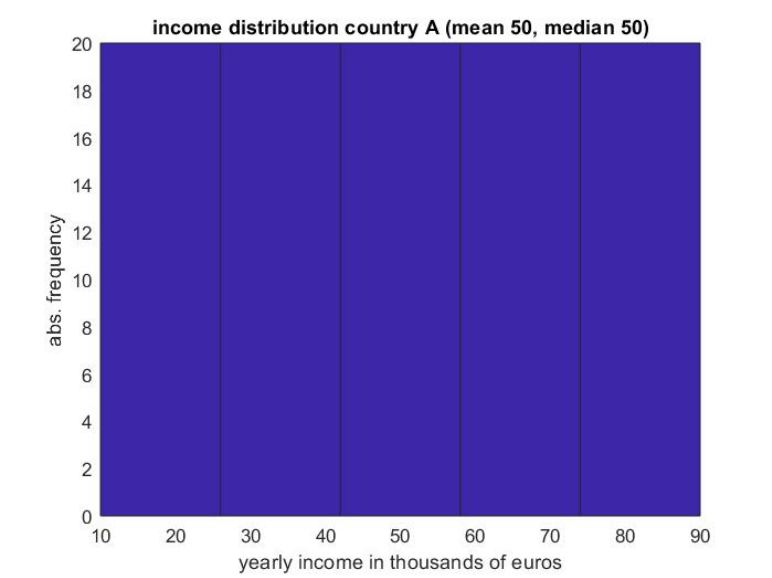
$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}$$

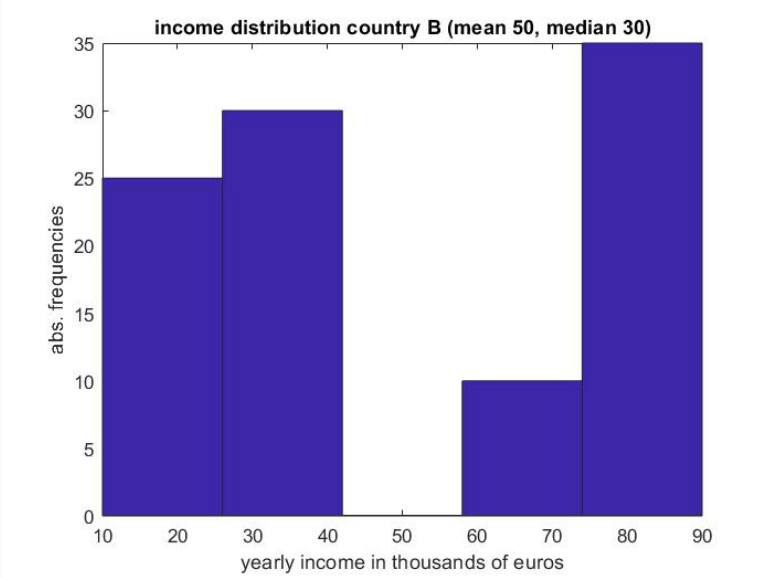
The **median** separates the data observations such that half are below and half are above

The **mode** is the value that occurs most often

Measures of central tendency

We use the following example to highlight why looking at the median can provide additional insights.





Measures of dispersion

For quantitative variables, we can analyse different measures of central tendency (aim: describe how much the data is spread out).

$$Var(x) = \frac{\sum_i (x_i - \bar{x})^2}{N-1}$$

(We divide by $N-1$ because $\bar{x}$ comes from our data).
The **variance** is the average of the squared deviations from the mean.

$$std(x) = \sqrt{Var(x)}$$

Standard deviation: square root of the variance
Range: difference between the largest and the smallest observation.

Measures of dispersion - Quartiles

Quartiles split sorted data into four parts, each with an equal number of observations.

- First quartile: 25% of the data is below, 75% of the data is above
- Second quartile (median): 50% of the data is below, 50% of the data is above
- Third quartile: 75% of the data is below, 25% of the data is above

Quartiles measure dispersion of the data.
Inter-quartile range (IQR): distance between the third and the first quartile.

A tool: indicator function

...is a function that we can apply to a data entry.
It takes the value one if the condition is true and the value zero otherwise.

$$\mathbf{I}(gender_i = female)$$

$$\mathbf{I}(x_i \leq 100)$$

An indicator function creates a **binary** or **dummy** variable

Frequencies

Let X be one of the variables in our data-set
Assume a discrete variable on an interval scale (number of children) with a range of $[0, 20]$
Observations $i = 1, \ldots, N \quad \rightarrow x_1, \ldots, x_N$

$$\text{Absolute frequency: } af_j = \sum_i \mathbf{I}(x_i = j) \quad j = 1, \ldots, 20$$

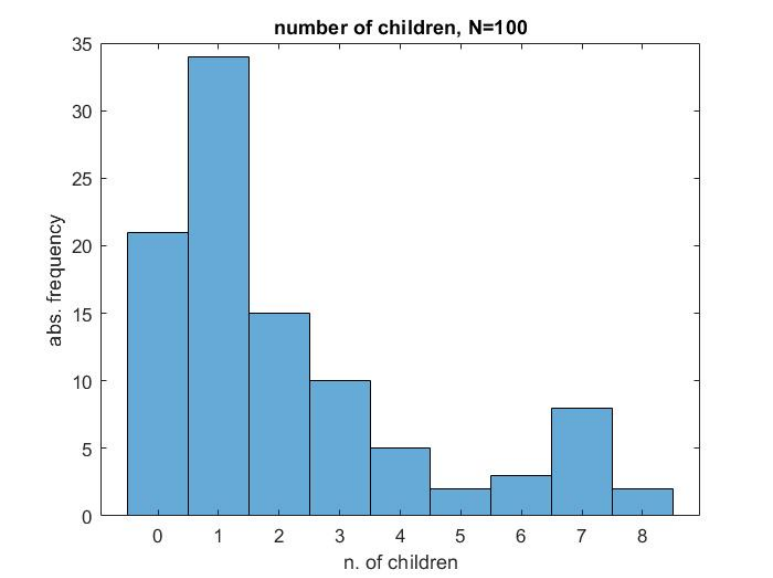
$$\text{Relative frequency: } rf_j = \frac{\sum_i \mathbf{I}(x_i = j)}{N} \quad j = 1, \ldots, 20$$

A **histogram** plots the frequencies for either each value or each class of values,

Histogram example

We will use this example to highlight why looking at the histogram can provide important additional insights

n. of children	0	1	2	3	4	5	6	7	8	9	10
n. of occurrences	21	34	15	10	5	2	3	8	2	0	0

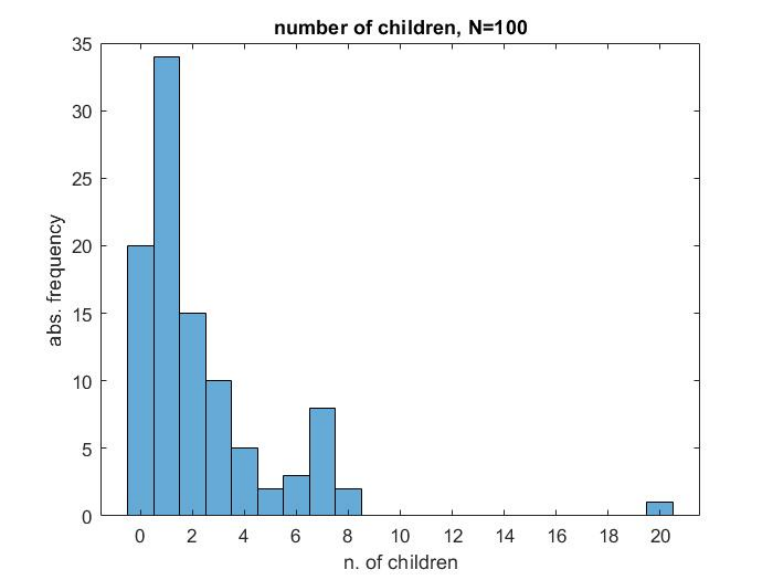


$$\bar{x} = 2.14, \quad Var(x) = 4.8893$$

histogram example

Let us change just one data entry

n. of children	0	1	2	3	4	5	6	7	8	9	10	20
n. of occurrences	20	34	15	10	5	2	3	8	2	0	0	1



$$\bar{x} = 2.3400, \quad Var(x) = 8.0246$$

Mean and variance are severely affected by outliers, because each observation enters the calculation (unlike the median).

LO2: Characterising Data

Challenges

(further reading: DABEP Ch. 2.8-2.12)

Outliers

Outliers are observations that are extreme relative to the rest of the sample
Keep or delete?

- Keep: they are meaningful since they occur
- Delete: they have a somewhat extreme impact on mean and variance
- Delete: we could claim that they are not representative for the data-set

Outliers

Outliers are observations that are extreme relative to the rest of the sample
Keep or delete?

- Keep: they are meaningful since they occur
- Delete: they have a somewhat extreme impact on mean and variance
- Delete: we could claim that they are not representative for the data-set

Best practice:

- Show both, results with and without outliers and
- be transparent in your choice.

Missing values

Investigate reason!

- Systematic missing or
- missing at random ?

Decision:

- Delete or
- approximate/ impute value ?

Missing values

Investigate reason!

- Systematic missing or
- missing at random ?

Decision:

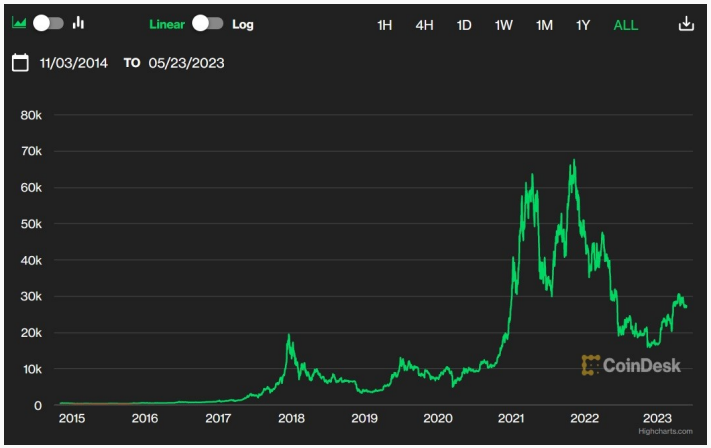
- Delete or
- approximate/ impute value ?

Best Practice:

- Show both, results with and without imputed missing values and
- be transparent in your choice.

Transforming data

- Data can be measured and illustrated in different ways (e.g. units of measurement).
- Common data transformation: take the natural logarithm.



Looks very different! left: historical prices, right: logarithm of prices

LO2: Characterising Data

Analysing Multiple Variables

(further reading: DABEP Ch. 4.1, 4.2, 4.6)

Analysing multiple variables

You will learn about

1. correlations
2. sub-samples and conditioning.

Analysing multiple variable

Positive (negative) **correlation** between two variables: a deviation from the mean is more likely to occur for both variables in the same (different) direction.
Else we say that the two variables are uncorrelated

$$Cov(x, y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{N - 1}$$

Only measures a **linear** relationship!

Spurious correlation

Careful when looking at correlations!

- It may be that x and y correlate, because both of them correlate with a third variable z.
- In this case, we call the correlation between x and y "spurious".

Example:

- Positive correlation between "being in the hospital" and "dying".
- Conclusion: shut down hospitals, because this is where people die.
- Omitted variable: "being seriously sick" correlated positively with both "hospital" and "dying".

Spurious correlation example

Does a vegan diet lead to better health ?

Let $v_i = 1$ for vegans, $v_i = 0$ for non vegans,

let s_i be a health score provided by a gp (the higher the better)

Question: expected sign of $cov(v, s)$?

What conclusion can you (not) draw from this?

Spurious correlation example

Does a vegan diet lead to better health ?

Let $v_i = 1$ for vegans, $v_i = 0$ for non vegans,

let s_i be a health score provided by a gp (the higher the better)

Question: expected sign of $cov(v, s)$?

What conclusion can you (not) draw from this?

Answer: expect a positive correlation BUT the correlation may be partly spurious: vegans are also likely to have a generally better lifestyle (exercise, meditation, mental health awareness...)

Sub-samples and conditioning

When we have multiple variables, we can also split the sample along one variable.

Characteristics persons	Smoking behaviour, 12 years or older Smoking status Smokers	Length and weight Under- and overweight, 4 years or older Overweight
	%	
Total	19.4	44.5
Sex: Male	23.1	46.4
Sex: Female	15.8	42.7
Level of educ.: pre-prim.and prim. educ.	24.4	62.2
Level of educ.: lower sec. education	23.3	60.0
Level of educ.:upper secondary	22.8	57.7
Level of educ: bachelor degrees	15.6	47.3
Level of educ: master degrees and doc.	11.7	38.1

Source: CBS

Here we have two outcome variables (smoking, overweight) and two ways to split the sample (gender, education).

Sub-samples and conditioning

- Sub-samples do not overlap (no observation occurs twice).
- Sub-samples cover the entire data-set (every observation occurs once).

Sub-samples and conditioning

- Sub-samples do not overlap (no observation occurs twice).
- Sub-samples cover the entire data-set (every observation occurs once).

- Use full sample for **general** statements about everybody (e.g. smoking).
- Use sub sample for **specific** statements about a group (e.g. smoking among young adults).

Sub-samples and conditioning

Characteristics persons	Smoking behaviour, 12 years or older Smoking status Smokers	Length and weight Under- and overweight, 4 years or older Overweight
	%	
Total	19.4	44.5
Sex: Male	23.1	46.4
Sex: Female	15.8	42.7

The average number of smokers per 100 individuals is 19.4.
Sub-sample averages are **conditional** averages.

- Among females, the average number of smokers per 100 individuals is 15.8.
- The average number of smokers per 100 individuals, conditional on being female, is 15.8.

Wrap up

Now we have learned about

... where our data comes from and why this can create specific challenges and

... different ways to characterise one or several variables