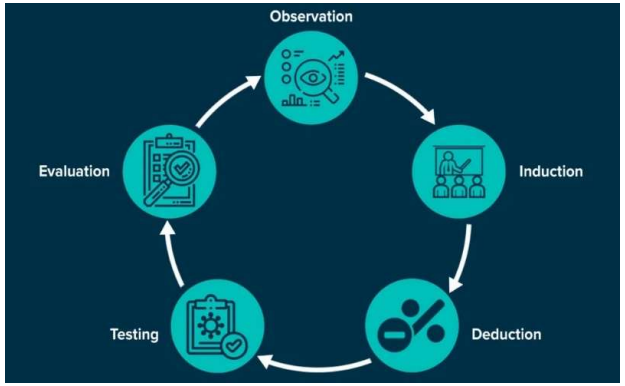
Empirical research in management and economics

Descriptive data analysis



Thorsten Pachur

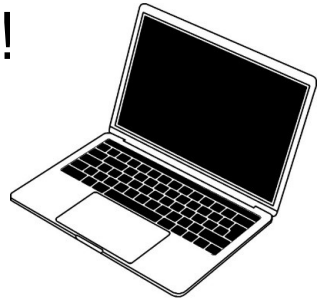
Technical University of Munich
School of Management
Chair of Behavioral Research Methods

For the Exercise

- Download & install JASP

  https://jasp-stats.org/download/

- Please bring a **laptop** to the exercise session!
- Download **Materials folder** from the module website!



Recap from last week

- Give four goals of empirical research
- What are the steps of the hypothetico-deductive model?
- What is the difference between deduction and induction?
- What are three criteria for establishing a causal relationship between two variables?
- Give three methodological challenges when working with empirical data
- What is the difference between correlational and experimental research?

Agenda for the semester

Session	Date	Topic
1	13 October	Introduction
2	**20 October**	**Descriptive data analysis**
3	27 October	Hypothesis development and research design
4	3 November	Inferential data analysis I
5	10 November	Inferential data analysis II
6	17 November	Simple regression
7	24 November	Multiple regression
8	1 December	Logistic regression
9	8 December	Factor analysis
10	15 December	Cluster analysis
11	12 January	Conjoint analysis
12	19 January	The replication crisis and open science
13	26 January	Summary and questions
	11 February	Exam

Learning goals for today

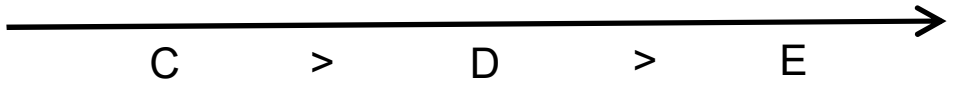
- Know how to distinguish different levels of measurement
- Know different plot types for data visualization
- Understand key statistics of descriptive data analysis
  - Characterizing the distribution of a single variable
    - Central tendency of a distribution
    - Variability of a distribution
    - Shape of a distribution
  - Characterizing the association between two variables

Levels of measurement

- *Nominal* scale

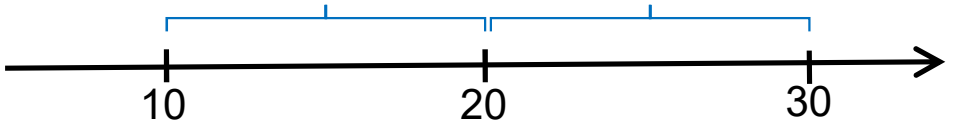
Examples: color labels, disease categories, marital status

- *Ordinal* scale



Examples: school grades, placement in a competition, product preferences

- *Interval* scale



Examples: temperature, opinion strength (e.g., on rating scale 1-8)

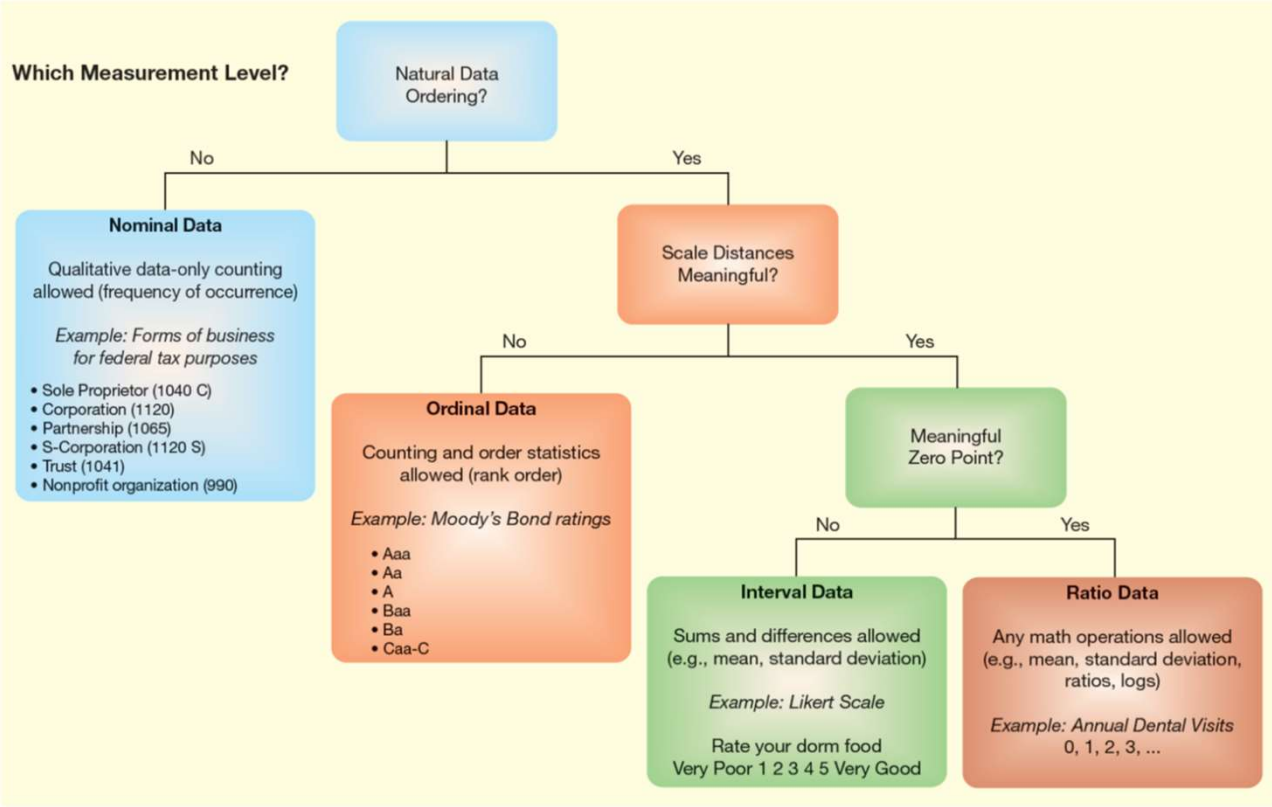
- *Ratio* scale

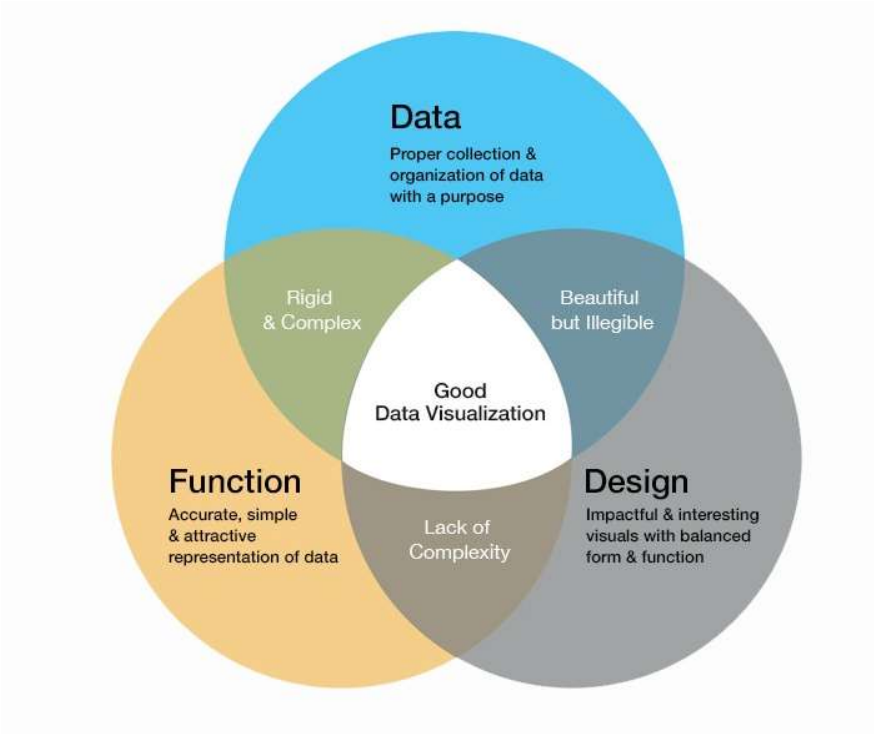
Examples: length, income, sales, frequency, weight

Levels of measurement

- ***Nominal* scale**
  Measurement that involves putting observations into qualitatively different categories
  Examples: color labels, disease categories, marital status
- ***Ordinal* scale**
  Measurement that involves ordering data according to the value on the variable
  Examples: school grades, placement in a competition, product preferences
- ***Interval* scale**
  Measurement that involves data on a number line for which any two adjacent values are the same distance from one another as any other pair of adjacent values
  → Size of intervals between values can be interpreted
  Examples: temperature, opinion strength (e.g., on rating scale 1-8)
- ***Ratio* scale**
  As interval scale, but with also having a natural zero point
  → The zero point denotes that the characteristic represented by the variable is absent
  Examples: length, income, sales, frequency, weight

Determining the measurement level of a variable





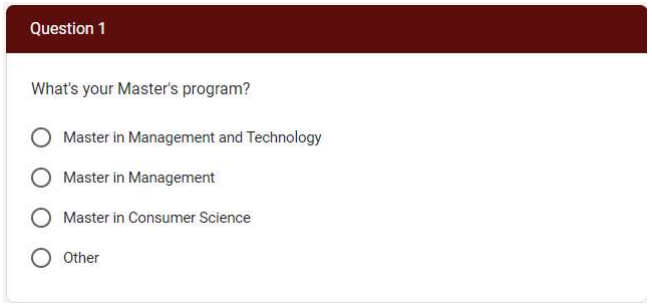
Data visualization

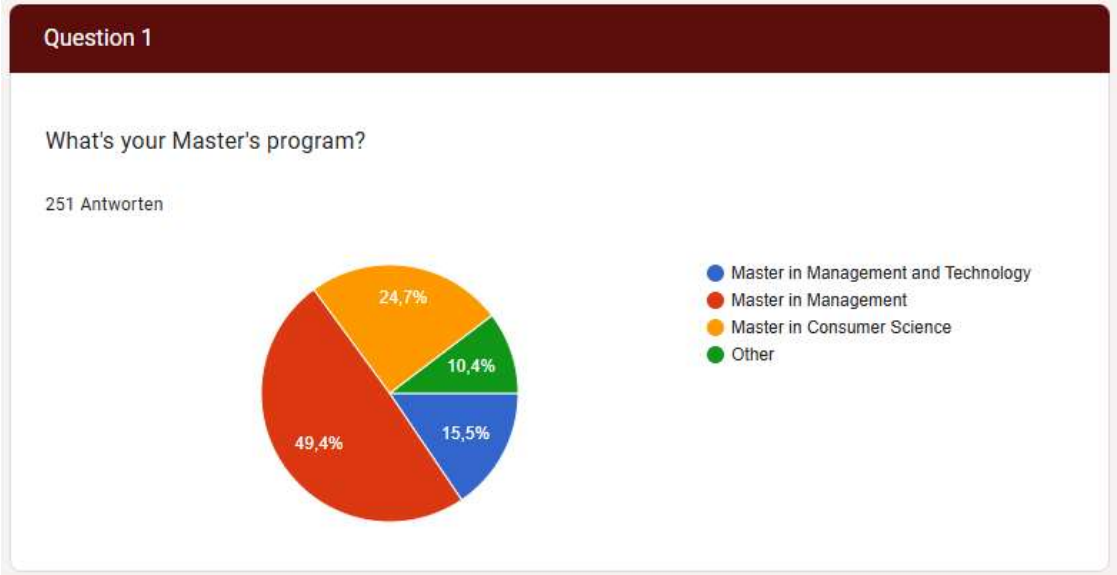
Data visualization

Pie chart

→ For frequency distribution of responses on a nominal-level variable

Nominal data, frequency of responses

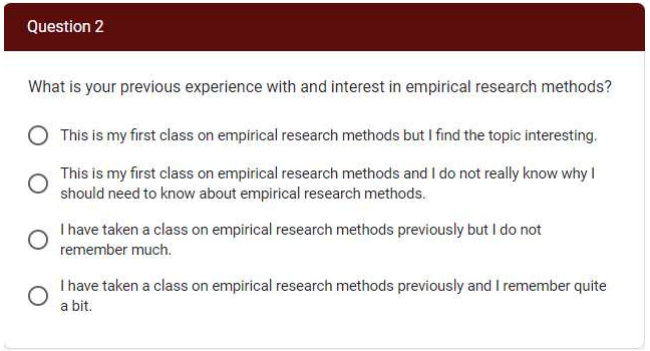


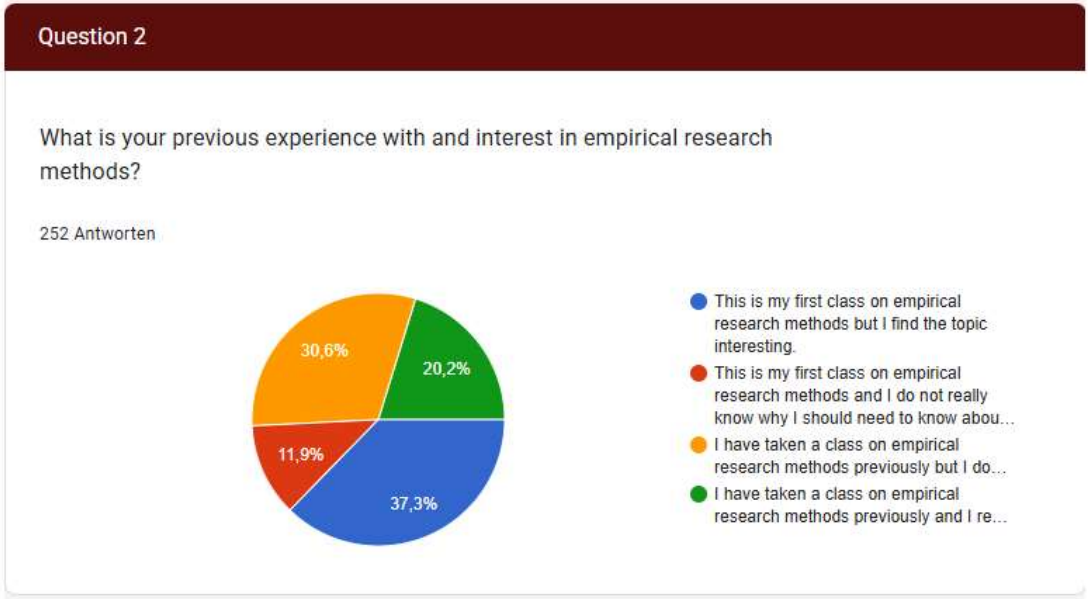


Data visualization

Pie chart

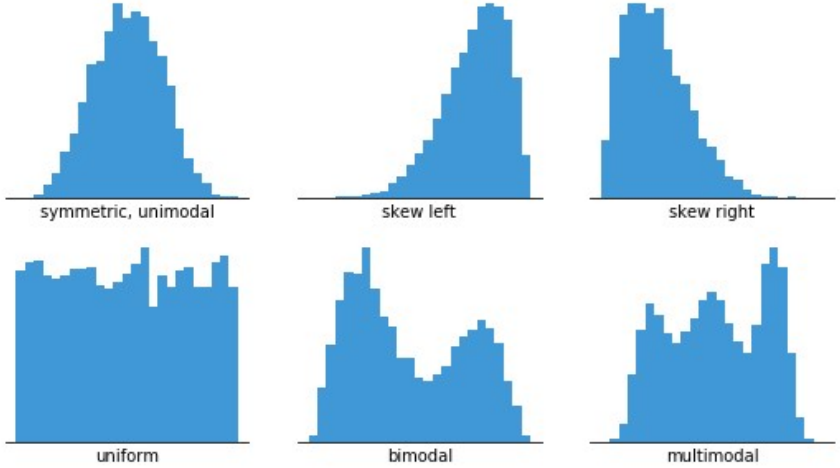
Nominal, ordinal?





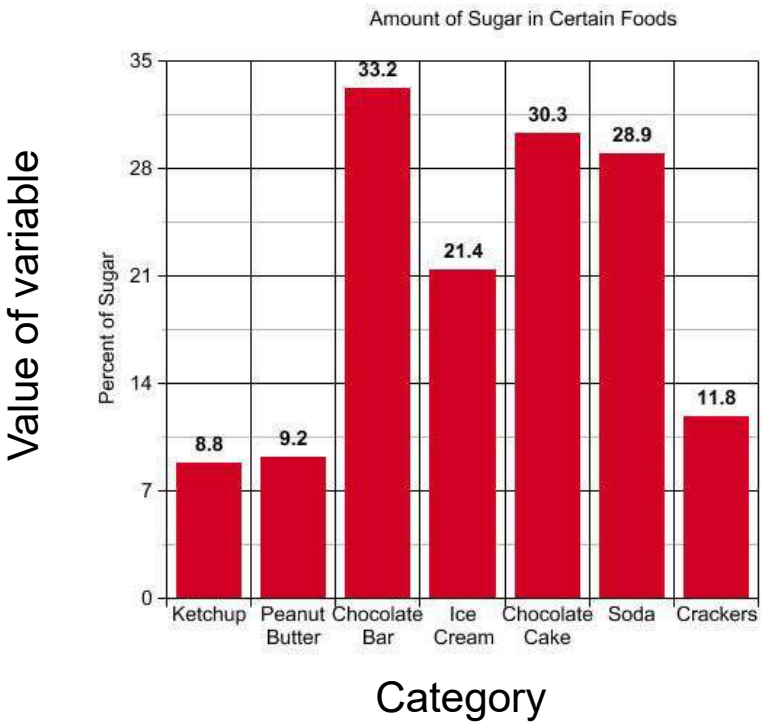
Data visualization

Histogram → For **frequency distribution of values** on a variable with at least ordinal level

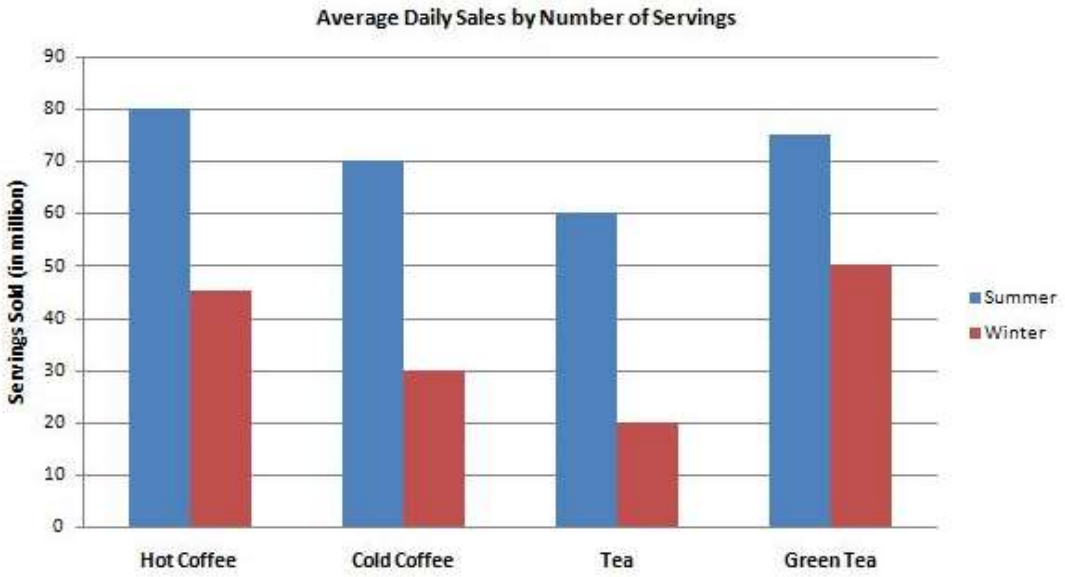


Data visualization

Bar plot → For continuous variables of objects in different (nominal) categories



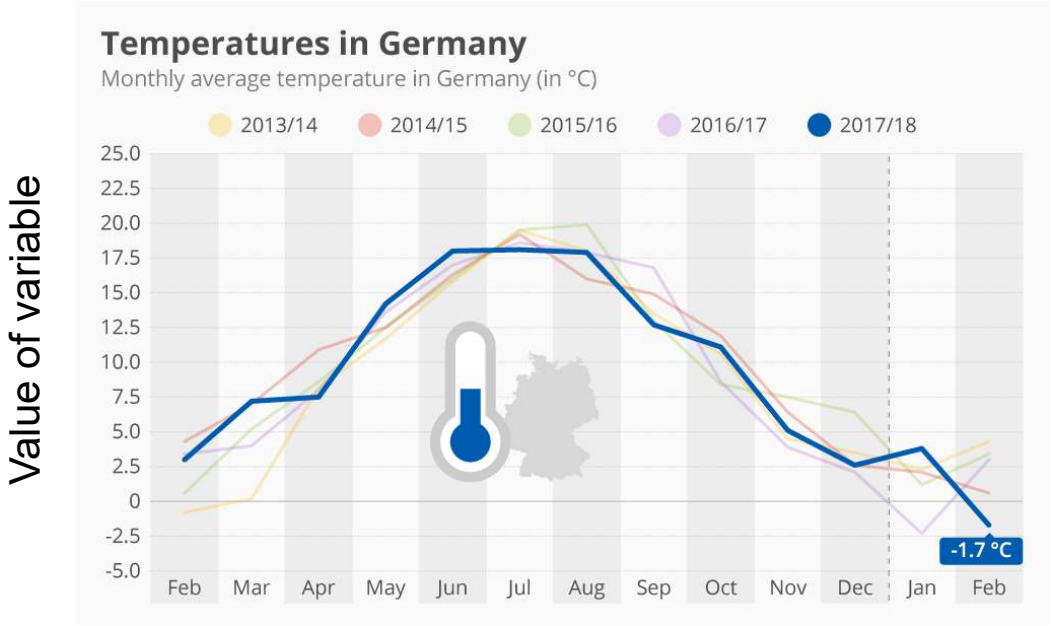
"Grouped" bar plot



Data visualization

Line plot

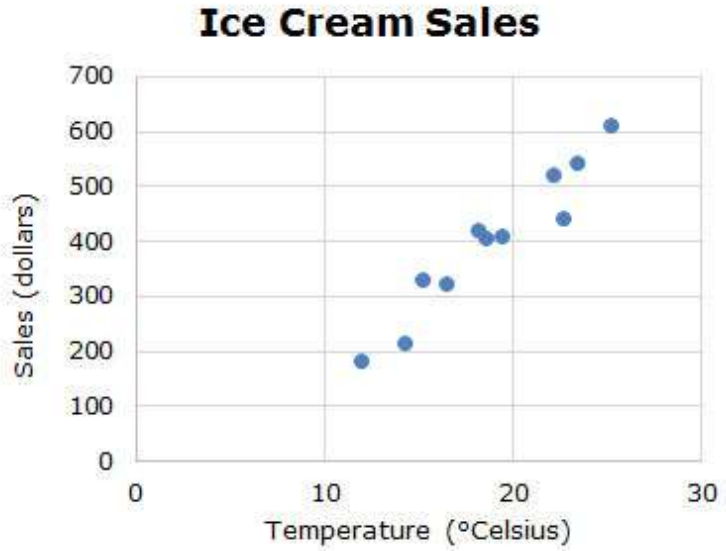
→ For continuous values of objects in different categories that can be ***ordered meaningfully***



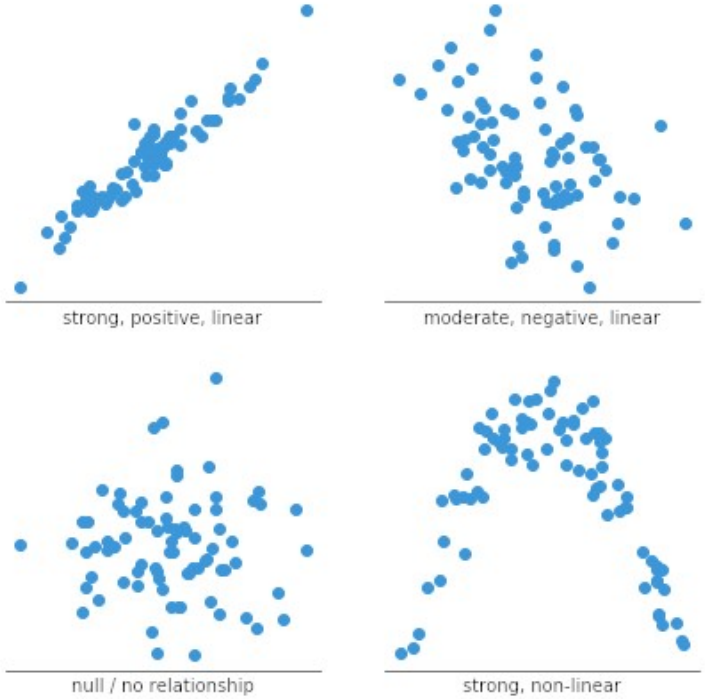
Data visualization

Scatter plot → To visualize the association between two variables (that are at least on ordinal level)

Value of variable B



Value of variable A

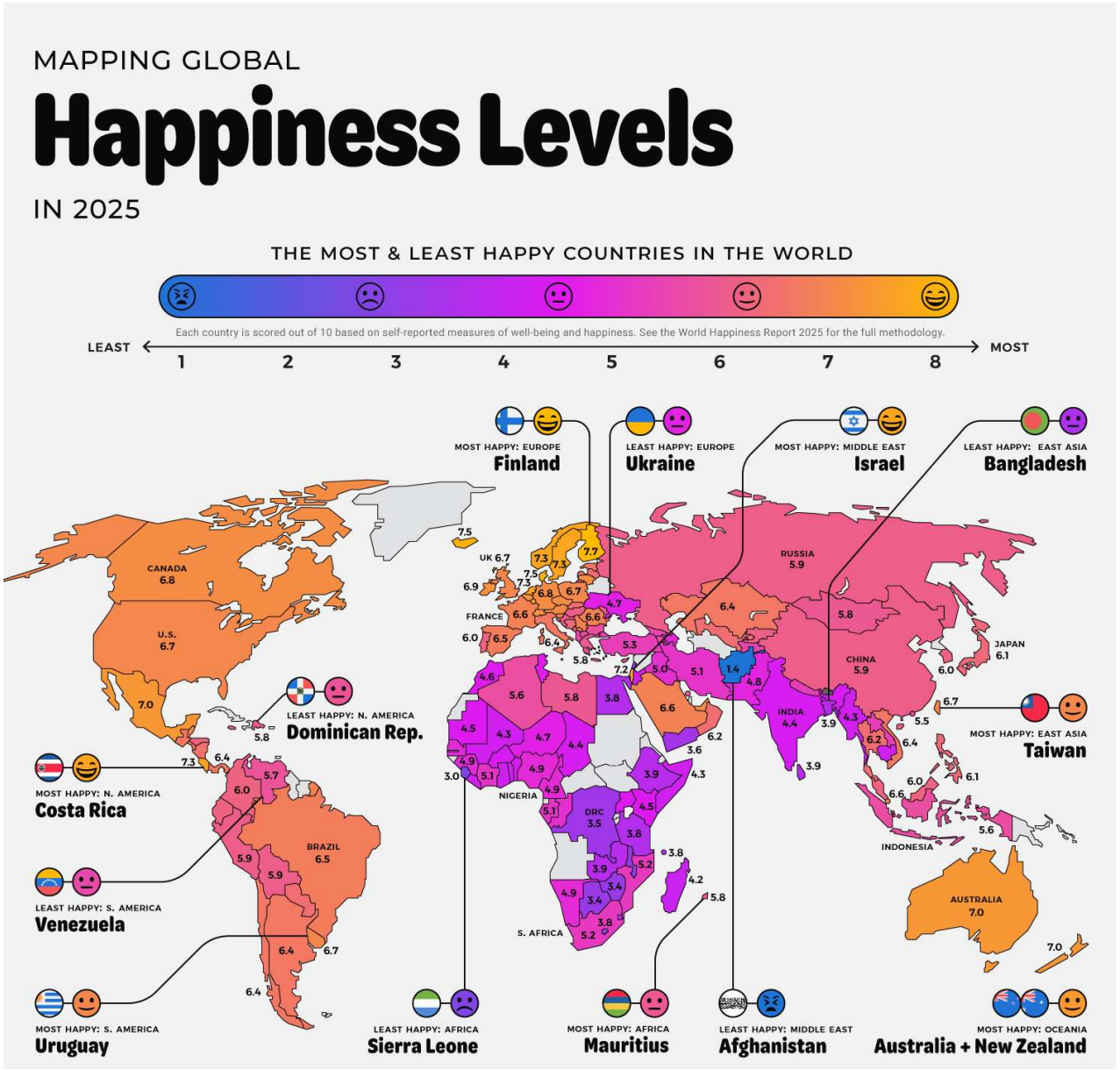


Descriptive data analysis

MAPPING GLOBAL

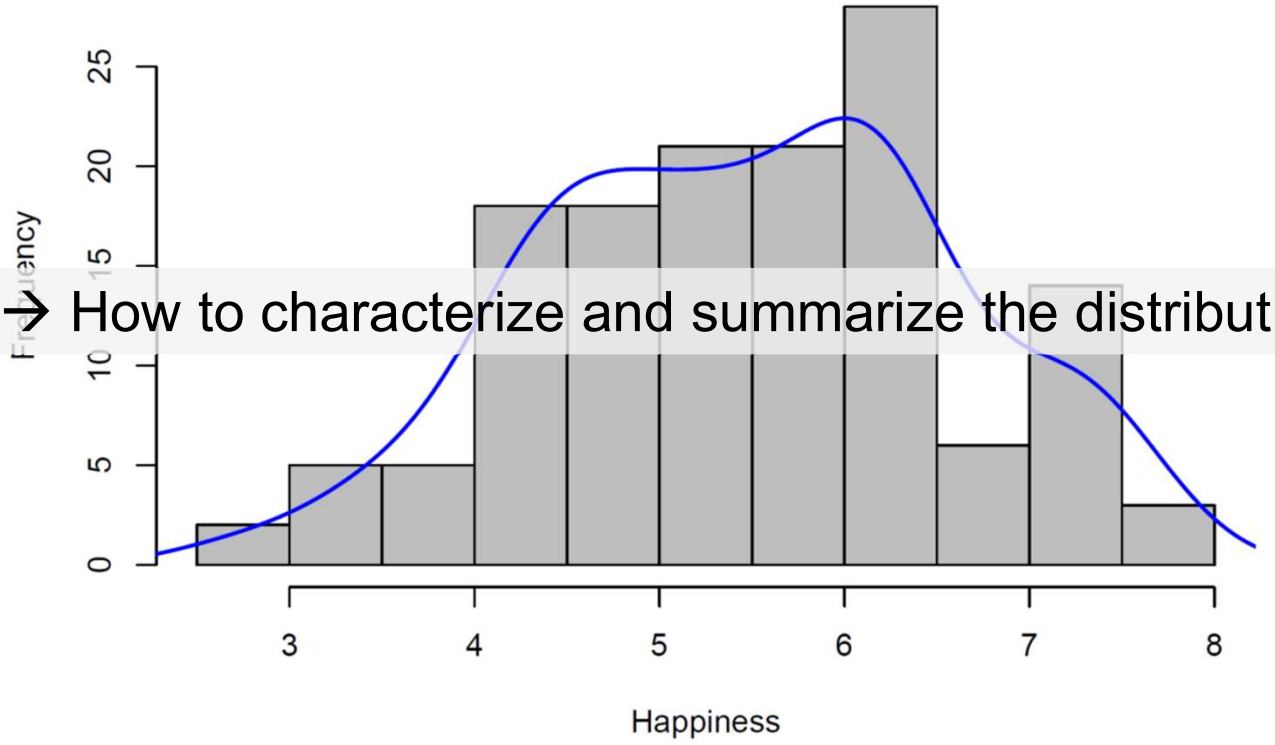
Happiness Levels

IN 2025



Distribution of happiness around the globe

N = 141 countries



→ How to characterize and summarize the distribution?

Measures of central tendency

- Mode (for all levels of measurement)
  → The most frequent value(s)

- Median (*Md*) (at least ordinal scale)
  → Value where 50% of the data are smaller (50% percentile, $Q_{50\%}$)

- (Arithmetic) Mean (*M*) (at least interval scale)

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

Distribution of happiness around the globe

N = 141 countries

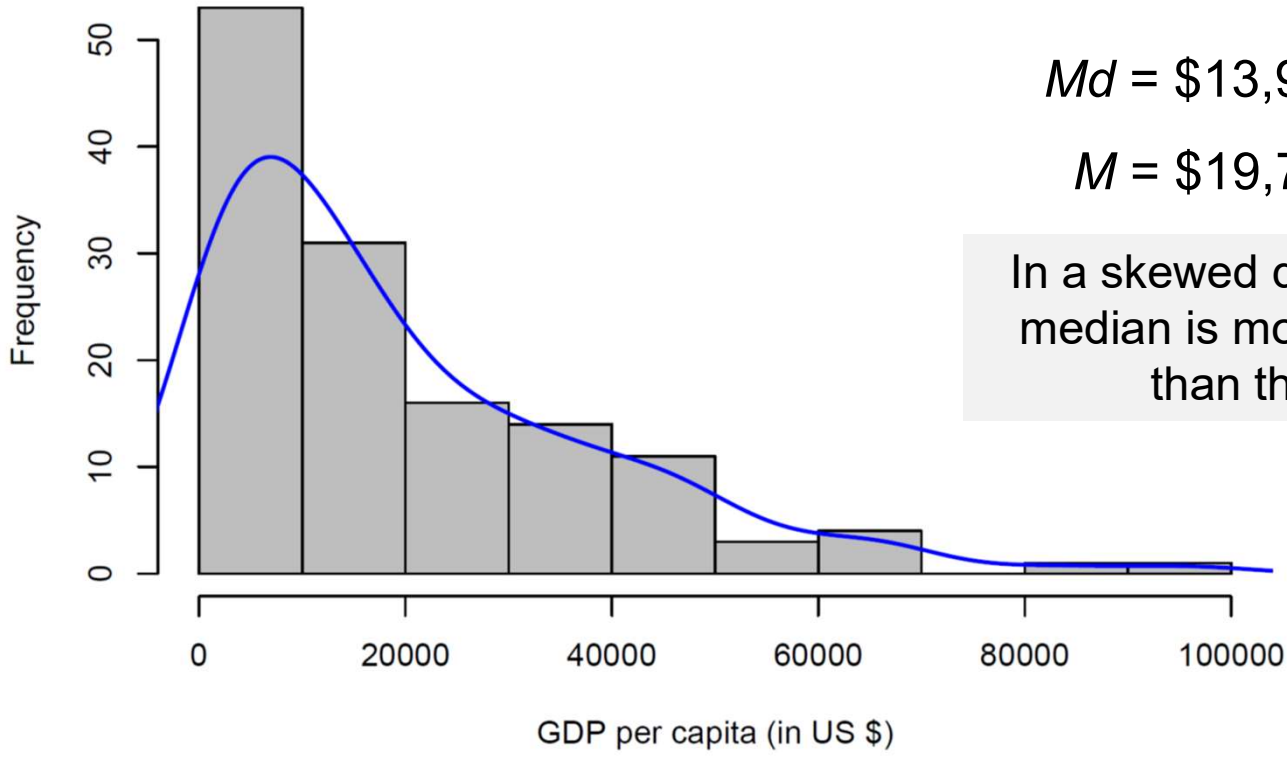
Mode = [4.31, 5.10, 5.15, 5.61]

Md = 5.58

M = 5.49

Central tendency in a skewed distribution

N = 141 countries



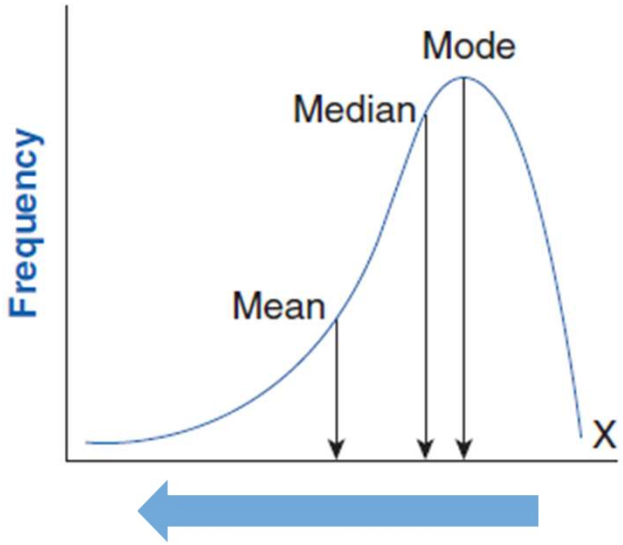
Md = $13,974

M = $19,777

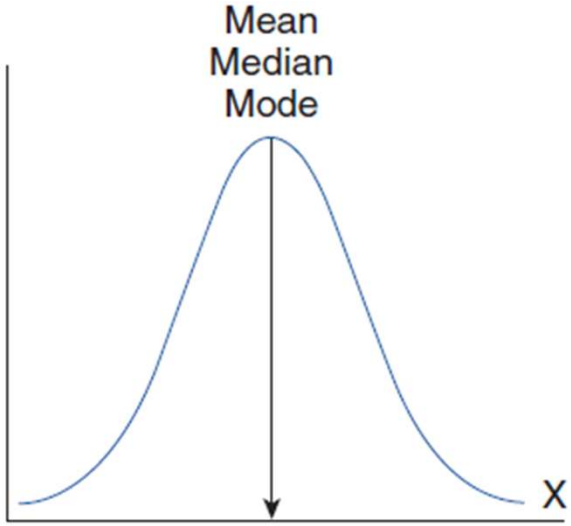
In a skewed distribution, the median is more informative than the mean !

Skewed distributions

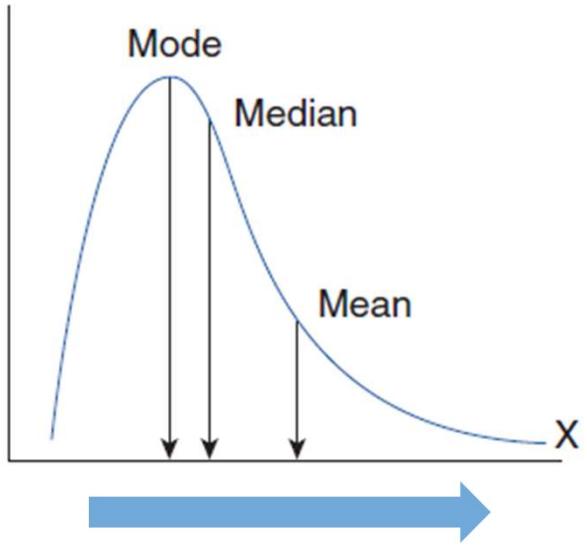
Left/negatively skewed



Symmetric



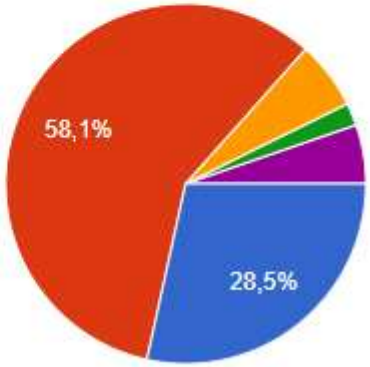
Right/positively skewed



Question 4

Which of the following statements is correct?

246 Antworten



YES!

- In a right-skewed distribution the arithmetic mean is typically lower than the median.
- In a right-skewed distribution the arithmetic mean is typically higher tha…
- In a right-skewed distribution the arithmetic mean is typically the same…
- In a right-skewed distribution the arithmetic mean is not defined.
- None of the above statements is correct.

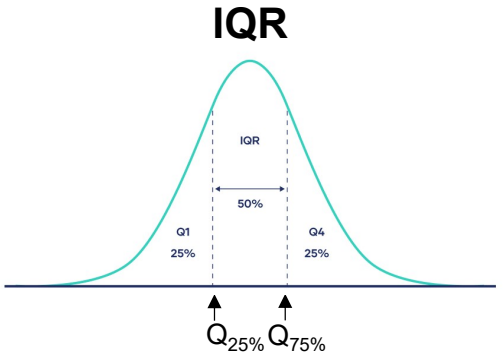
Distribution of happiness around the globe

N = 141 countries

Measures of variability

- Range (at least ordinal scale)
  → $x_{max} - x_{min}$

- Interquartile range (IQR) (at least ordinal scale)
  → $Q_{75\%} - Q_{25\%}$ (middle 50% of the data)



- Variance (at least interval scale)
  → Mean squared deviation from the average value $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$

  Standard deviation (*SD*): $SD = \sqrt{s^2}$

Distribution of happiness around the globe

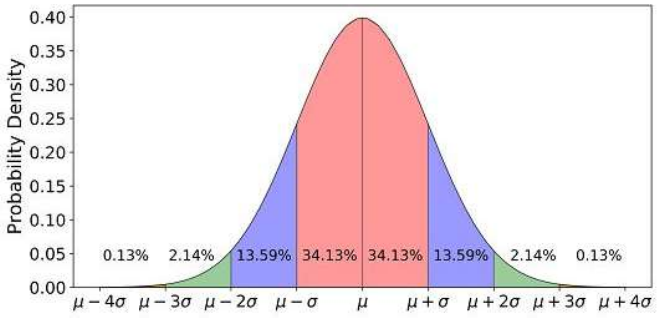
Range = 5.13

IQR = 1.64

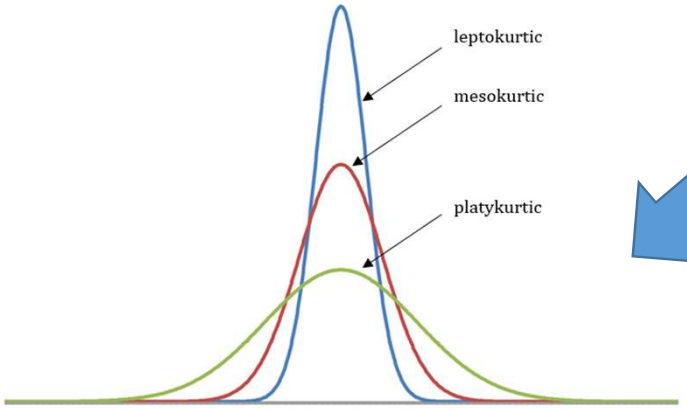
SD = 1.12

Shape of a distribution

Normal distribution (“Bell curve“)

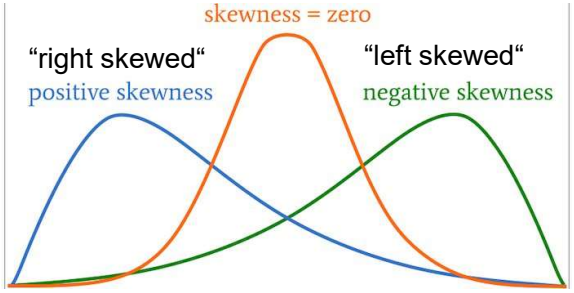


Carl-Friedrich Gauss



Peakness

Skewness

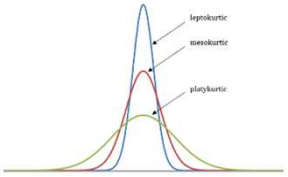


Measures of the shape of a distribution

- Kurtosis (peakedness)

$$\gamma = \frac{m_4}{s^4} - 3 \quad \text{with} \quad m_4 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4$$

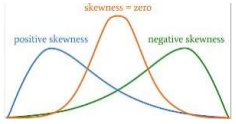
γ = 0: normal distribution
γ > 0: peaked distributions
γ < 0: flattened distributions



- Skewness

$$g_m = \frac{m_3}{s^3} \quad \text{with} \quad m_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3$$

g_m = 0: symmetric distributions
g_m > 0: right-skewed distributions
g_m < 0: left-skewed distributions

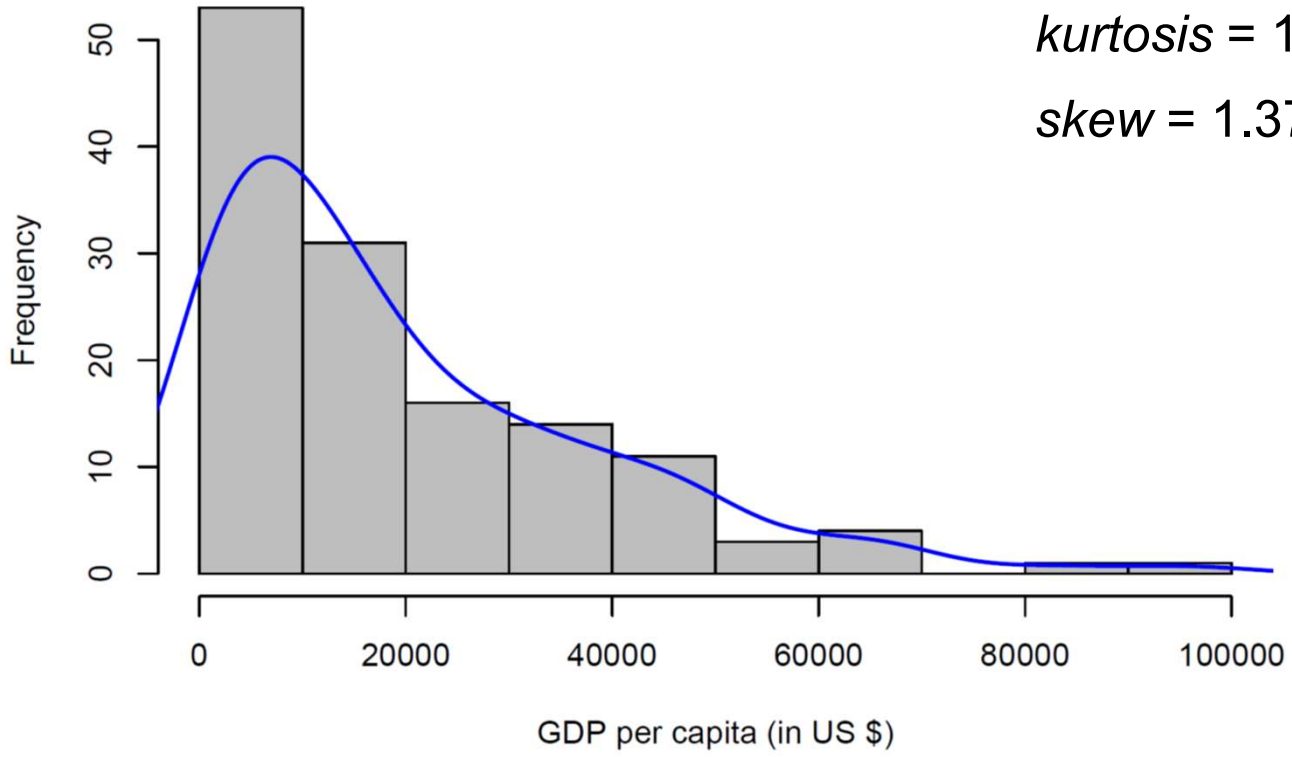


Distribution of happiness around the globe

kurtosis = -0.61 → flattened

skew = -0.08 → left-skewed

Distribution of GDP around the globe

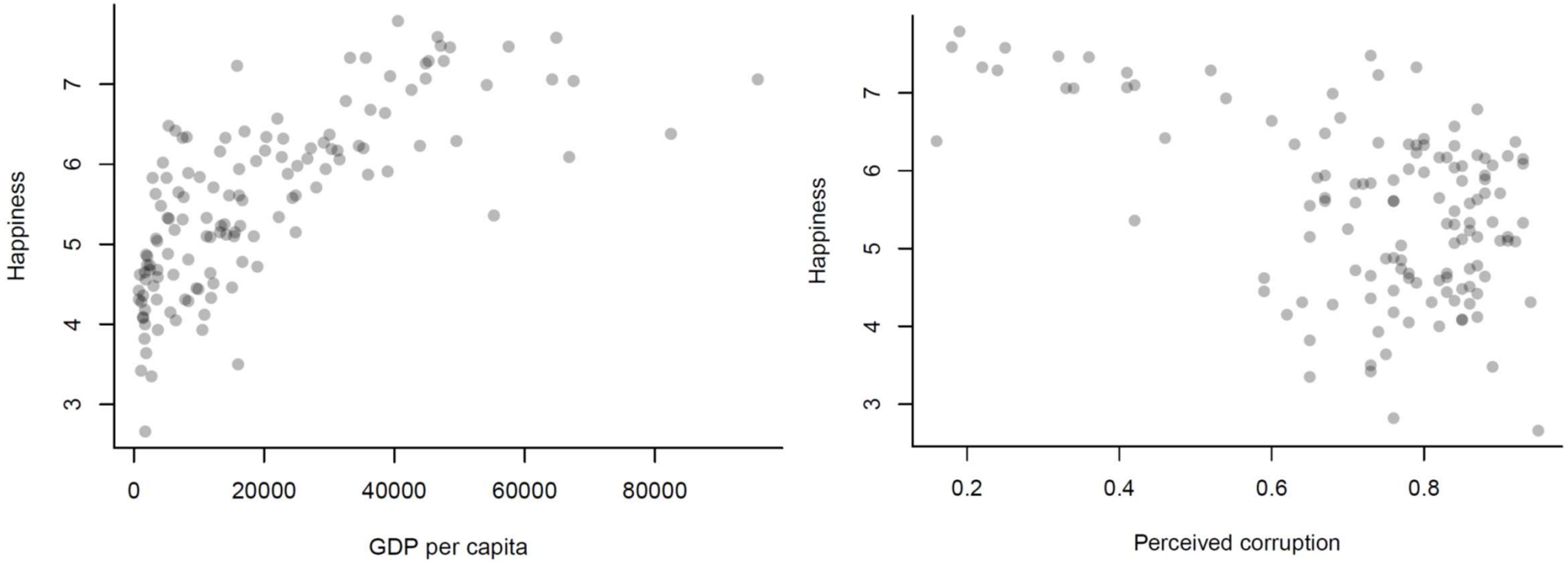


kurtosis = 1.77 → peaked

skew = 1.37 → right-skewed

Association between two variables

Variables with at least ordinal level of measurement



Association between two variables

Variables with nominal level of measurement → Frequency of co-occurrence of the variables‘ categories

		Previous experience with statistical software				
		I have no previous experience with statistical software.	I have previously used JASP.	I have previously used R.	I have previously used both R and JASP.	I have previously used other statistical software (e.g., SPSS, STATA).
Program	Master in Consumer Science	19	2	27	2	12
	Master in Management	67	2	38	2	15
	Master in Management and Technology	10	0	22	0	7
	Other	5	0	18	0	3

Association between two variables

- Product-moment correlation (Pearson) (at least **interval** scale)

  Ranges from -1 (=perfect *negative* association) to 0 (=no association) to 1 (=perfect *positive* association)

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

n is the number of cases

- Rank correlation (Spearman) (at least **ordinal** scale)

  Ranges from -1 to 0 to 1

$$r_s = 1 - \left(\frac{6 \times \sum D^2}{n^3 - n}\right)$$

n number of pairs of items in the sample

D difference between each pair of ranked measurements

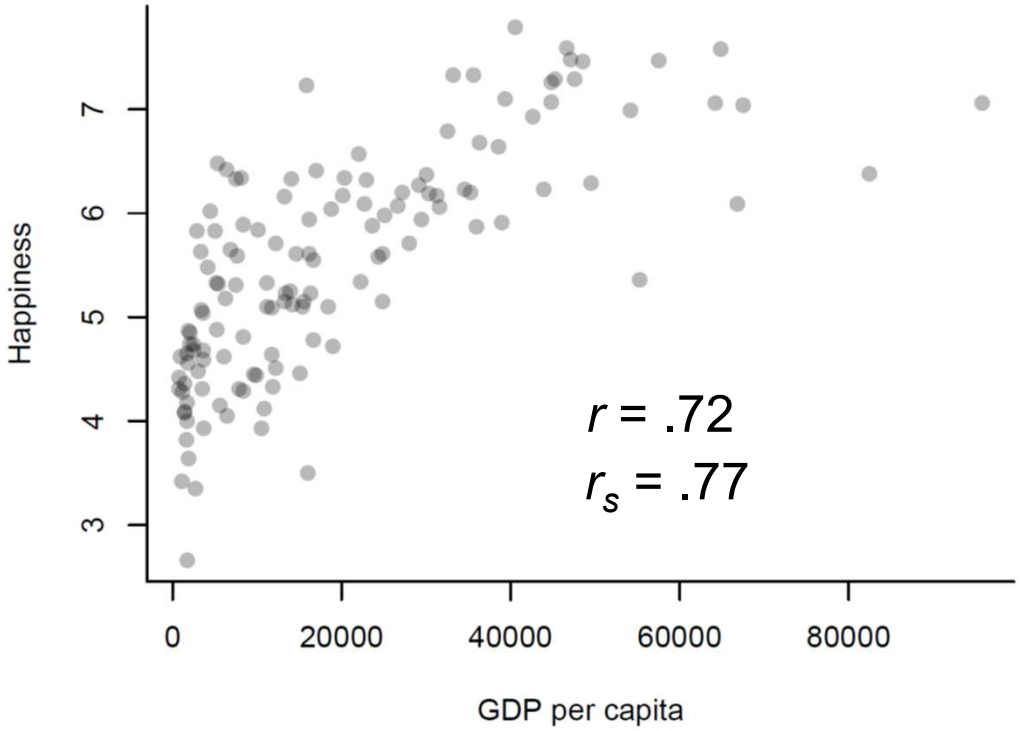
- Cramer's *V* (both variables are on **nominal** scale)

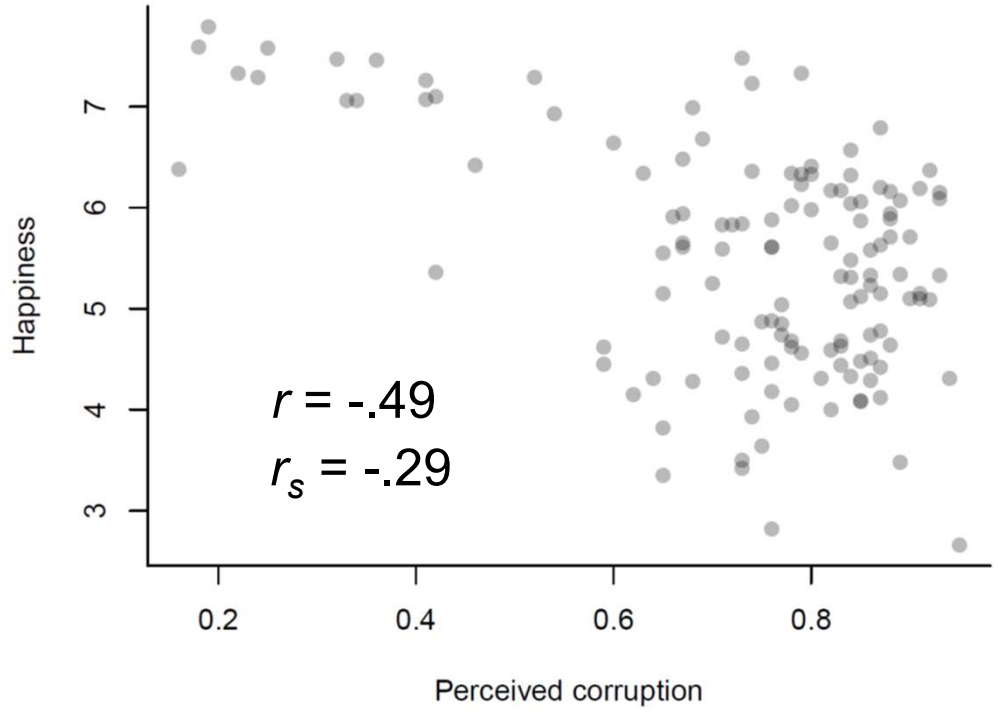
  Ranges from 0 (=no association) to 1 (=perfect association)

$$\text{Cramer's V} = \sqrt{\frac{\chi^2}{n \cdot \min(r-1, c-1)}}$$

r and c are the number of categories that the two variables have
n is the number of cases

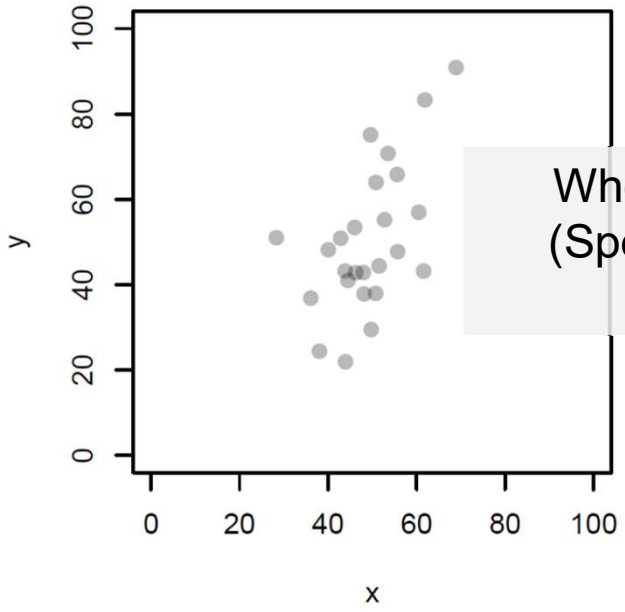
Association between two variables



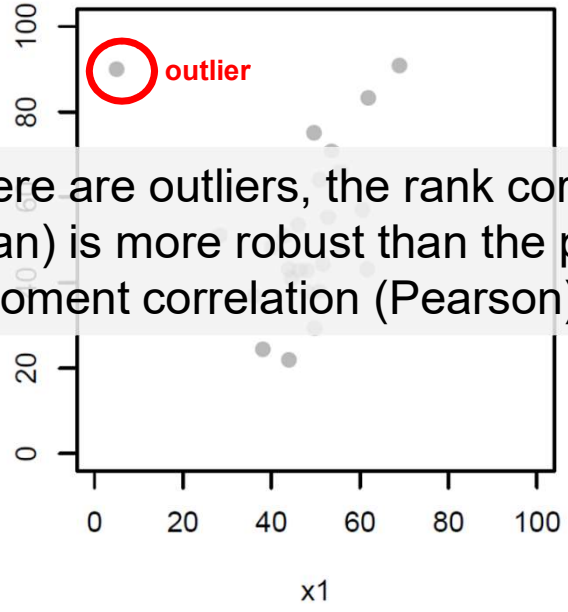


Beware of outliers

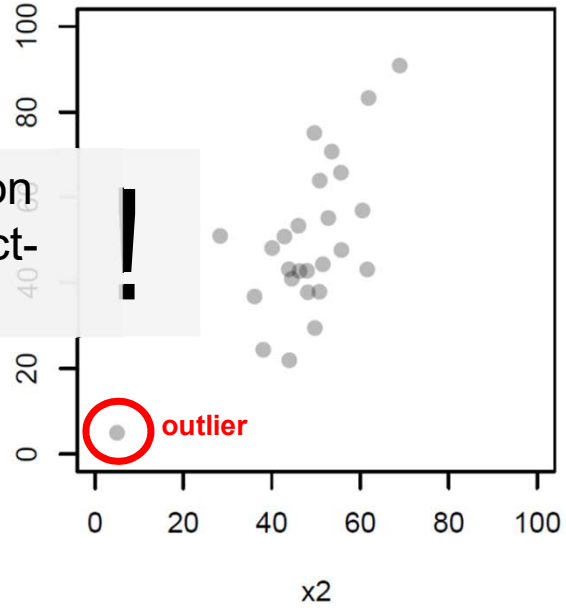
r = .37
r_s = .43



r = -.09
r_s = .27



r = .63
r_s = .49



When there are outliers, the rank correlation (Spearman) is more robust than the product-moment correlation (Pearson)

!

Association between two variables

Variables with **nominal level** of measurement → frequency of co-occurrence of the variables‘ categories

		Previous experience with statistical software				
		I have no previous experience with statistical software.	I have previously used JASP.	I have previously used R.	I have previously used both R and JASP.	I have previously used other statistical software (e.g., SPSS, STATA).
Program	Master in Consumer Science	19	2	27	2	12
	Master in Management	67	2	38	2	15
	Master in Management and Technology	10	0	22	0	7
	Other	5	0	18	0	3

Cramer‘s V = .20

Self-quiz questions

- What is the difference between a nominal, ordinal, interval, and ratio level of measurement?
- Describe for which type of data the following types of plots are indicated: Pie chart, histogram, bar plot, line plot, scatter plot
- Give indices for characterizing the distribution of a variable in terms of a) central tendency, b) variability, and c) shape
- Describe indices for expressing the strength (and direction) of the association between two variables and when each index is indicated
  - Pearson and Spearman correlations: How do they differ?
  - Cramer‘s *V*: Why can it not be negative?

Background readings for next week

Beins, B. C. (2019). Measurement and sampling. In: B. C. Beins, *Research methods* (p. 127–159). Cambridge University Press.

Dienes, Z. (2008). Karl Popper and demarcation. In: Z. Dienes, *Understanding psychology as a science: An introduction to scientific and statistical inference* (p. 1–32). Palgrave Macmillan.