GISMA
BUSINESS SCHOOL

M506

Research Method and Scientific Work:

Analysing Quantitative Data

Week 8, Nov 2022

Prof. Dr. Tilmann Lindberg

Conceptual Model & Hypotheses

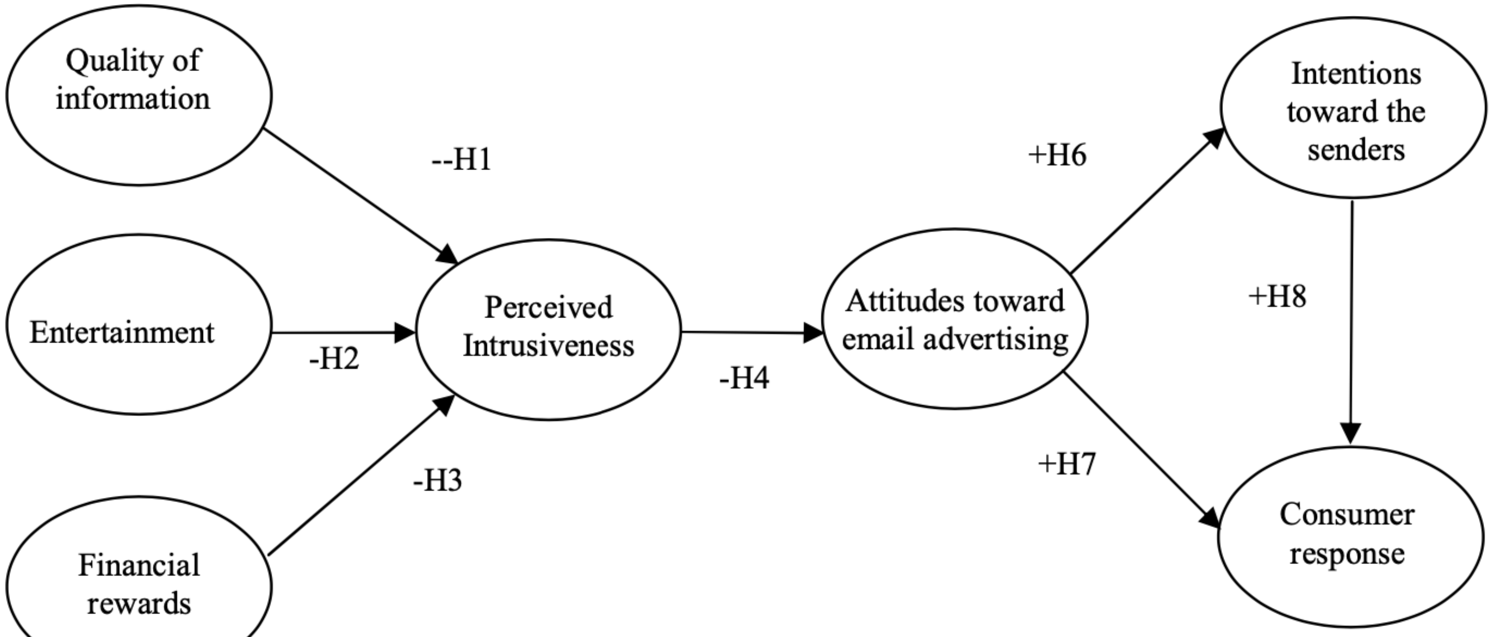


Figure 1. Conceptual framework

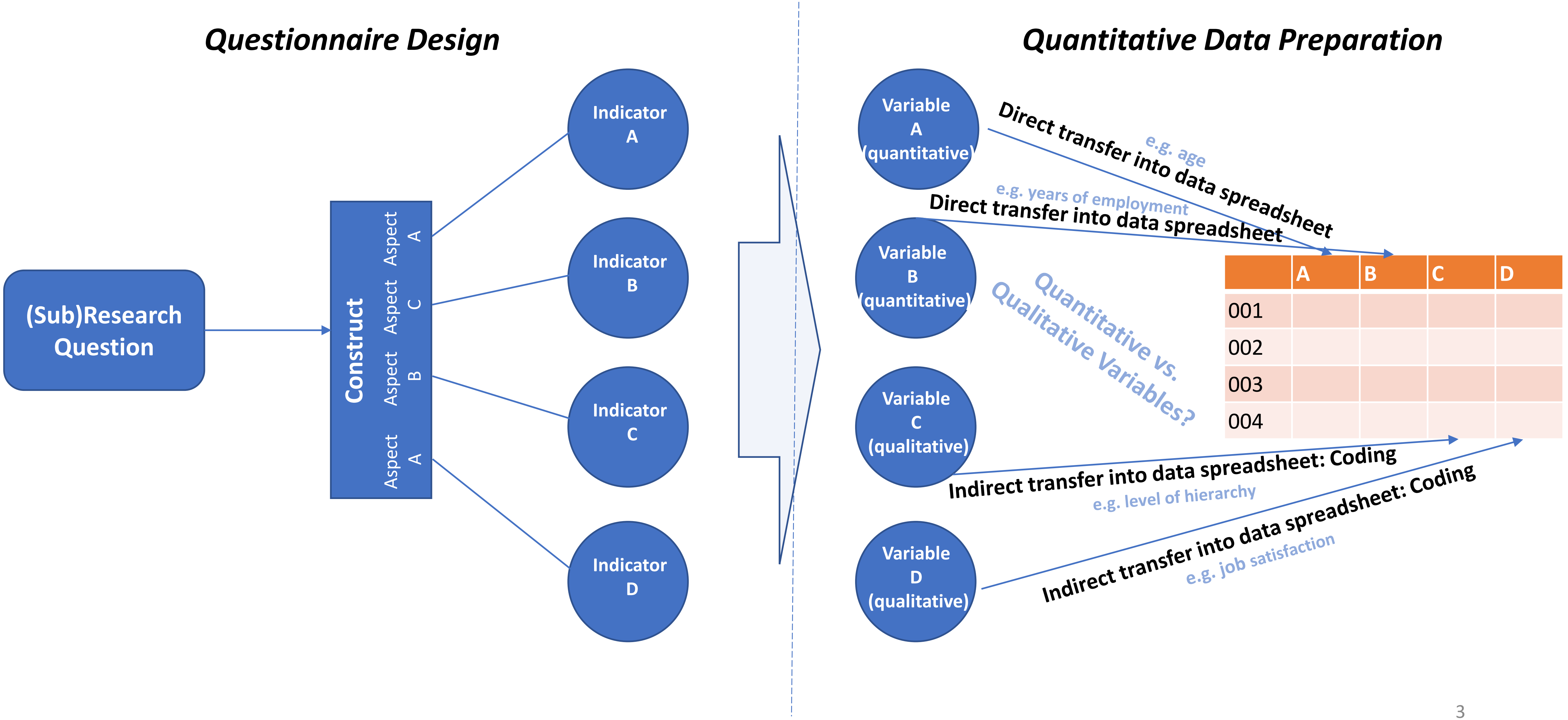
Source: Hamid et al., 2013

Question:
How do you read/interpret the graphic? What does it tell you in terms of the hypotheses?

Preparing your data: Creating a Data Spreadsheet

Questionnaire Design

Quantitative Data Preparation


(Sub)Research Question
Construct
Aspect A
Aspect C
Aspect B
Aspect A
Indicator A
Indicator B
Indicator C
Indicator D
Variable A (quantitative)
Variable B (quantitative)
Variable C (qualitative)
Variable D (qualitative)
Direct transfer into data spreadsheet
e.g. age
Direct transfer into data spreadsheet
e.g. years of employment
Quantitative vs. Qualitative Variables?
Indirect transfer into data spreadsheet: Coding
e.g. level of hierarchy
Indirect transfer into data spreadsheet: Coding
e.g. job satisfaction

	A	B	C	D
001				
002				
003				
004				

Preparing your data: Different Types of Scales

Nominal Scale

→ ***Describing categorical differences between variables without any numerical order***

- Gender
- Colour
- Political affiliations
- Preferences
- Levels of education
- Nationality
- Brands
- Etc.

Ordinal Scale

→ ***Describing categorical differences between variables with an inherent rank orders***

- Likert scale answers
- Rank order questions
- Semantic differential questions
- Etc.

Interval Scale

→ ***Describing numerical differences between variables with equal intervals between numbers without any zero point***

- Temperature in Celsius or Fahrenheit
- IQ Tests
- Test scores
- Time
- Voltage
- Etc.

Ratio Scale

→ ***Describing numerical differences between variables with equal intervals between numbers as well as a zero point***

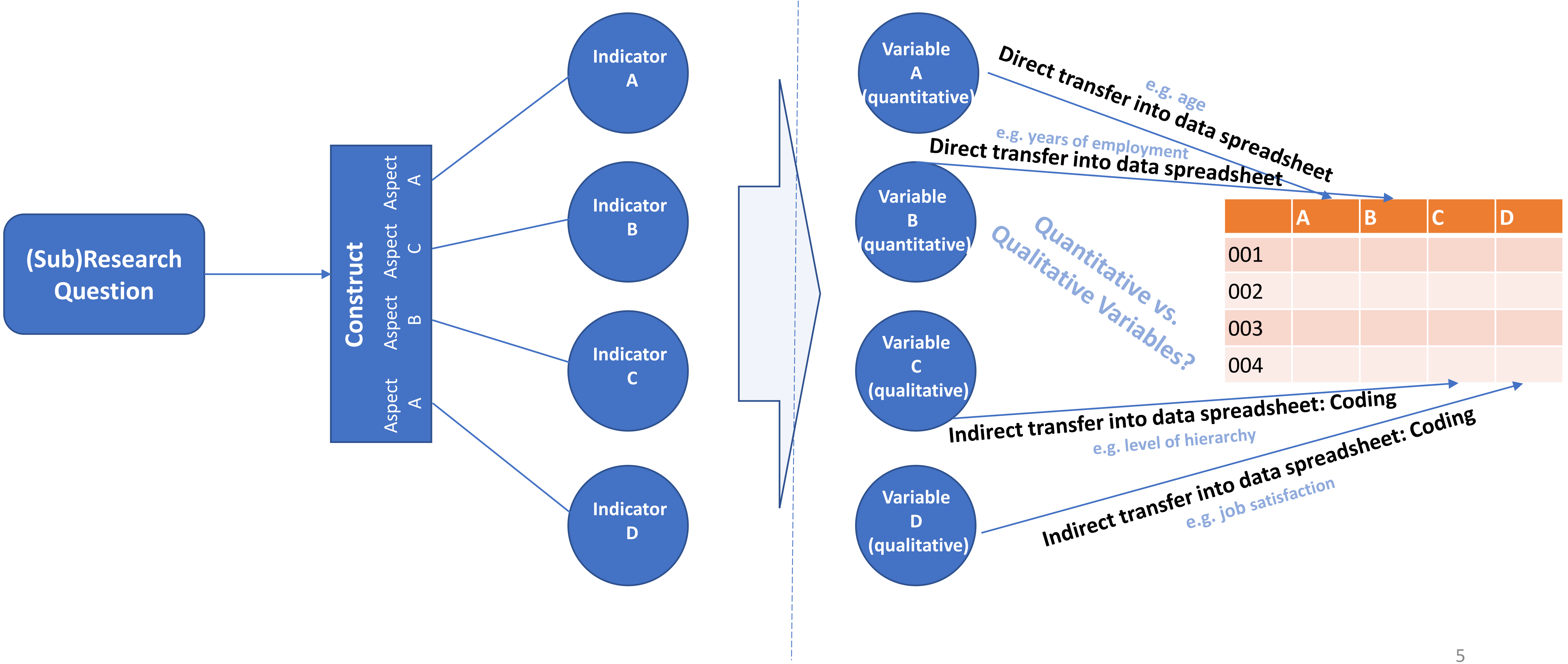
- Any monetary value (e.g. salary)
- Inventory amounts
- Height
- Votes (in elections)
- Etc.

Wilson 2014, 265f.; 10 Interval Data Examples: Interval Scale Definition & Meaning (intellspot.com) (04.11.2021)

Preparing your data: Creating a Data Spreadsheet

Questionnaire Design

Quantitative Data Preparation


(Sub)Research Question
Construct
Aspect A
Aspect B
Aspect C
Aspect A
Indicator A
Indicator B
Indicator C
Indicator D
Variable A (quantitative)
Variable B (quantitative)
Variable C (qualitative)
Variable D (qualitative)
Direct transfer into data spreadsheet
e.g. age
Direct transfer into data spreadsheet
e.g. years of employment
Quantitative vs. Qualitative Variables?
Indirect transfer into data spreadsheet: Coding
e.g. level of hierarchy
Indirect transfer into data spreadsheet: Coding
e.g. job satisfaction

	A	B	C	D
001				
002				
003				
004				

Preparing your data: Translating your Qualitative Data into Codes

Questionnaire 001

Statement	Strongly Agree	Agree	Neither agree nor disagree	Disagree	Strongly Disagree
Your employer offers attractive career opportunities.		X			
Climbing the career ladder at your employer is highly competitive.	X				
Your employer offers excellent employee incentive schemes.			X		
	1	*2*	*3*	*4*	*5*

Questionnaire 002

Statement	Strongly Agree	Agree	Neither agree nor disagree	Disagree	Strongly Disagree
Your employer offers attractive career opportunities.			X		
Climbing the career ladder at your employer is highly competitive.				X	
Your employer offers excellent employee incentive schemes.		X			
	1	*2*	*3*	*4*	*5*

Case Number	Your employer offers attractive career opportunities	Climbing the career ladder at your employer is highly competitive.	Your employer offers excellent employee incentive schemes.	…
001	2	1	3	…
002	3	4	2	…
003	3	1	4	…
…				

Wilson 2014, 234

Preparing your data: Different Types of Scales

Nominal Scale

→ ***Describing categorical differences between variables without any numerical order***

- Gender
- Colour
- Political affiliations
- Preferences
- Levels of education
- Nationality
- Brands
- Etc.

Ordinal Scale

→ ***Describing categorical differences between variables with an inherent rank orders***

- Likert scale answers
- Rank order questions
- Semantic differential questions
- Etc.

Interval Scale

→ ***Describing numerical differences between variables with equal intervals between numbers without any zero point***

- Temperature in Celsius or Fahrenheit
- IQ Tests
- Test scores
- Time
- Voltage
- Etc.

Ratio Scale

→ ***Describing numerical differences between variables with equal intervals between numbers as well as a zero point***

- Any monetary value (e.g. salary)
- Inventory amounts
- Height
- Votes (in elections)
- Etc.

Wilson 2014, 265f.; 10 Interval Data Examples: Interval Scale Definition & Meaning (intellspot.com) (04.11.2021)

Summarising Data: Frequency Tables

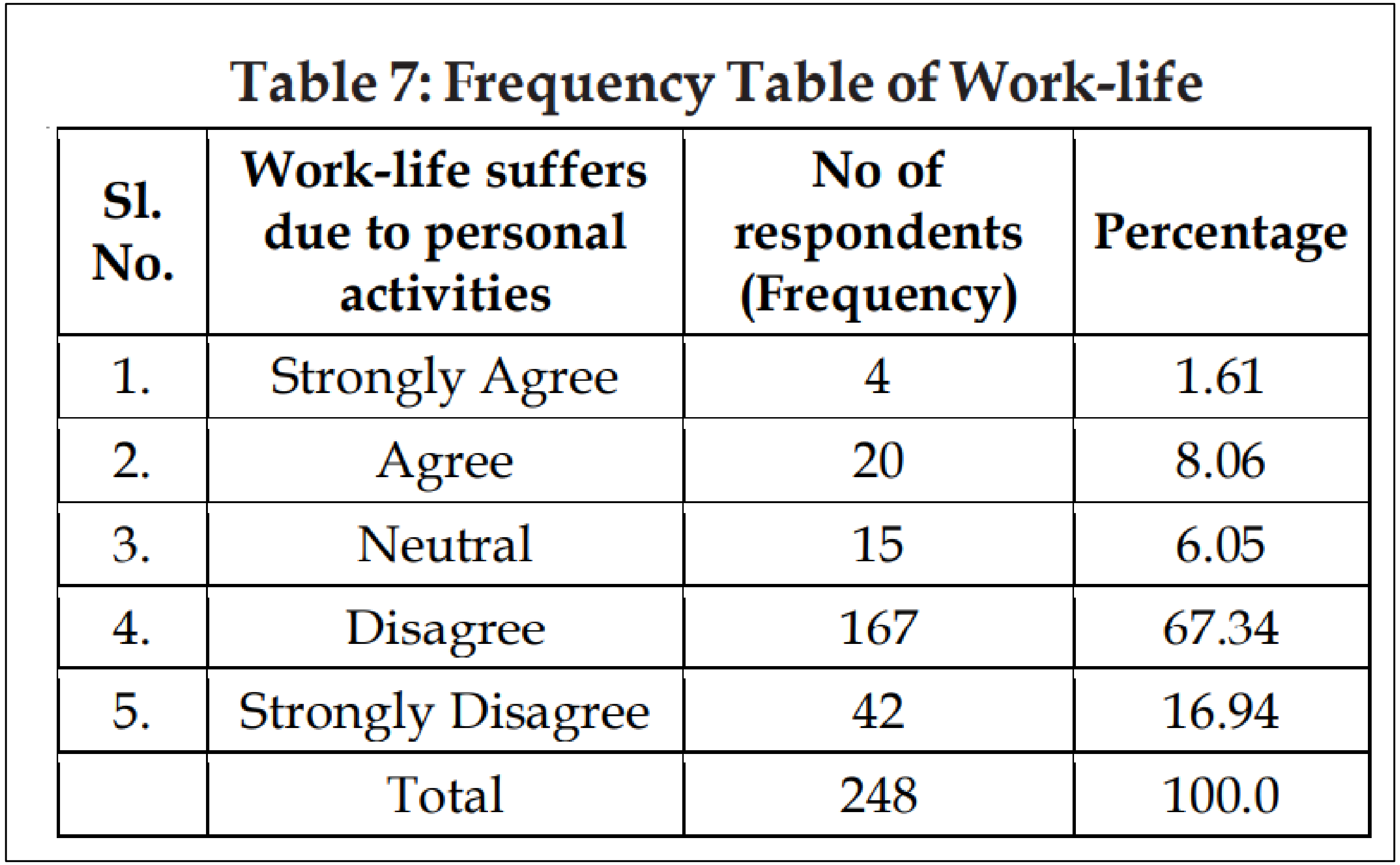


Table 7: Frequency Table of Work-life

Sl. No.	Work-life suffers due to personal activities	No of respondents (Frequency)	Percentage
1.	Strongly Agree	4	1.61
2.	Agree	20	8.06
3.	Neutral	15	6.05
4.	Disagree	167	67.34
5.	Strongly Disagree	42	16.94
	Total	248	100.0

Lombardi and Pastore 2014

Summarising Data: Frequency Tables

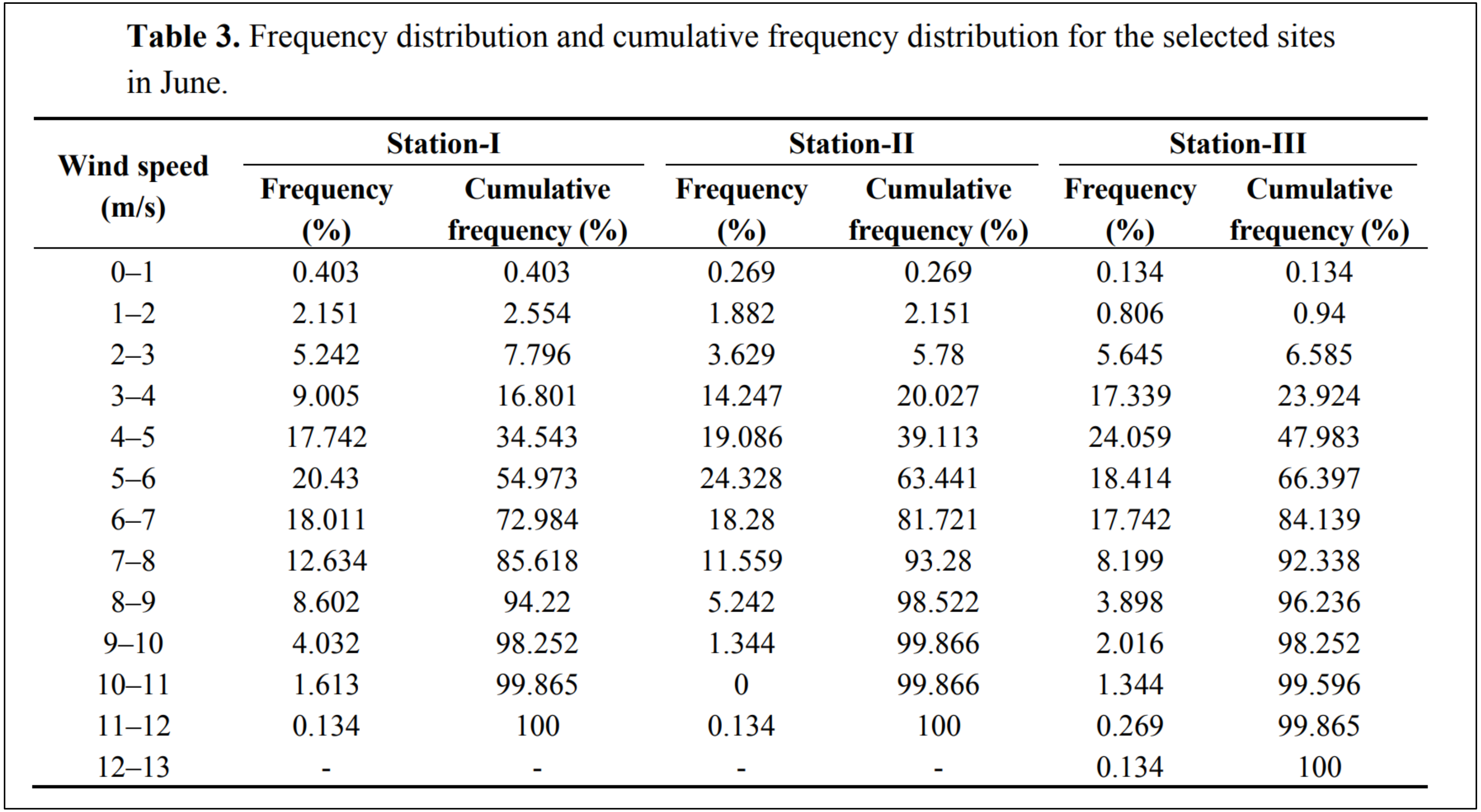


Table 3. Frequency distribution and cumulative frequency distribution for the selected sites in June.

Wind speed (m/s)	Station-I		Station-II		Station-III	
	Frequency (%)	**Cumulative frequency (%)**	**Frequency (%)**	**Cumulative frequency (%)**	**Frequency (%)**	**Cumulative frequency (%)**
0–1	0.403	0.403	0.269	0.269	0.134	0.134
1–2	2.151	2.554	1.882	2.151	0.806	0.94
2–3	5.242	7.796	3.629	5.78	5.645	6.585
3–4	9.005	16.801	14.247	20.027	17.339	23.924
4–5	17.742	34.543	19.086	39.113	24.059	47.983
5–6	20.43	54.973	24.328	63.441	18.414	66.397
6–7	18.011	72.984	18.28	81.721	17.742	84.139
7–8	12.634	85.618	11.559	93.28	8.199	92.338
8–9	8.602	94.22	5.242	98.522	3.898	96.236
9–10	4.032	98.252	1.344	99.866	2.016	98.252
10–11	1.613	99.865	0	99.866	1.344	99.596
11–12	0.134	100	0.134	100	0.269	99.865
12–13	-	-	-	-	0.134	100

Azad et al 2014, 3070)

Summarising Data: Diagrams

Bar Diagram

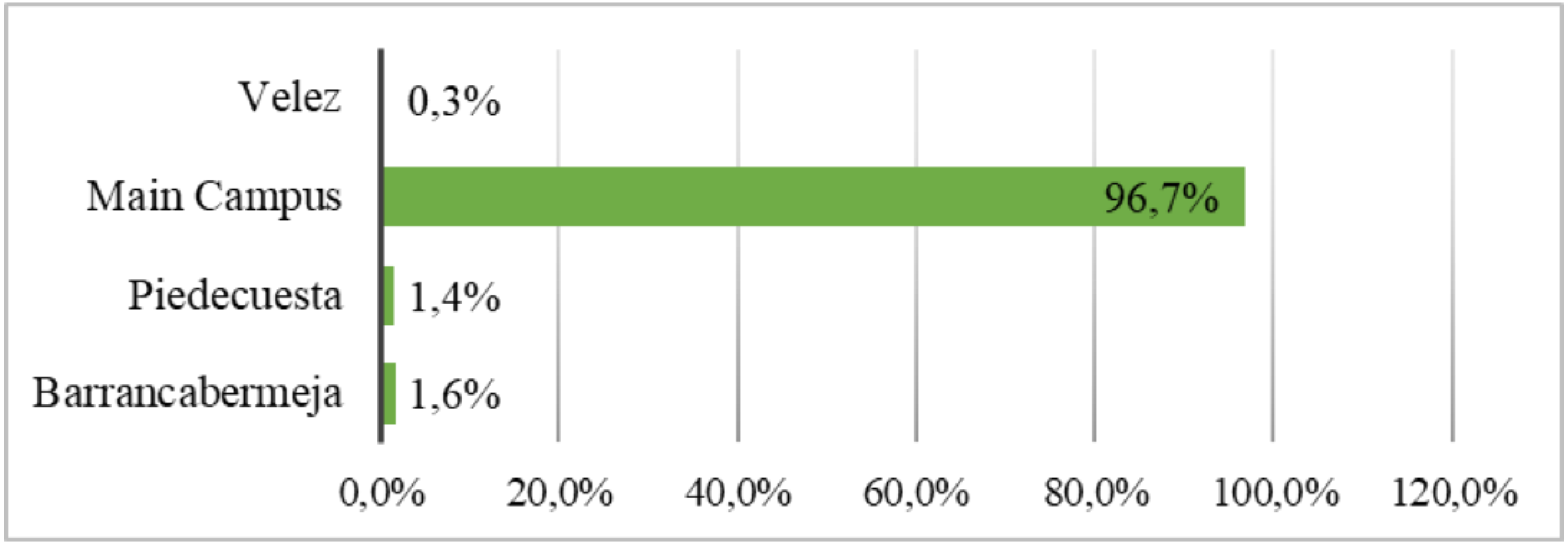
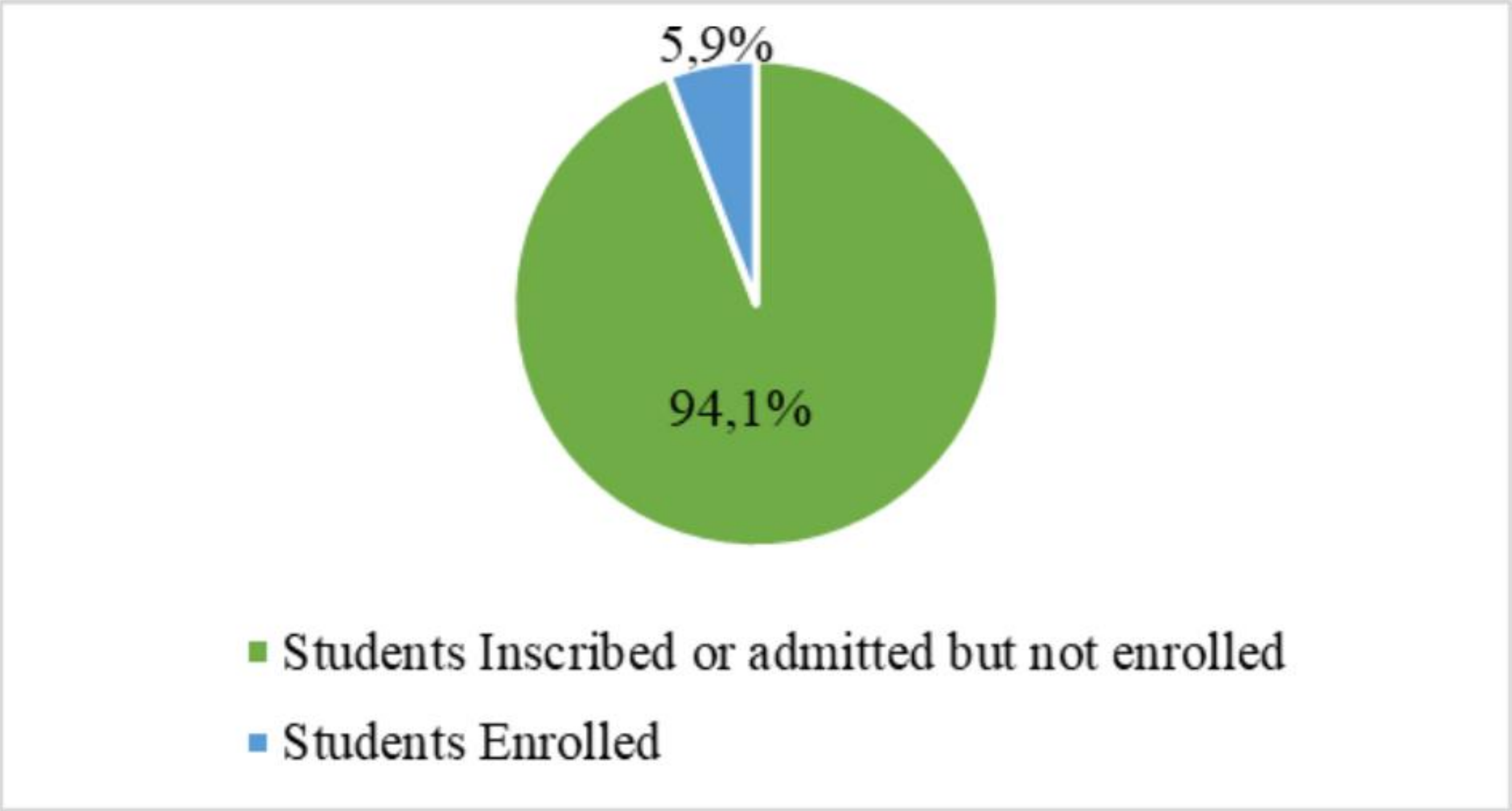


Figure 2. Bar graph of the percentage distribution of PREIN students by headquarter

Diaz and Rueda 2020

Summarising Data: Diagrams

Pie Charts



Diaz and Rueda 2020

Summarising Data: Diagrams

GISMA
BUSINESS SCHOOL

Line Graph

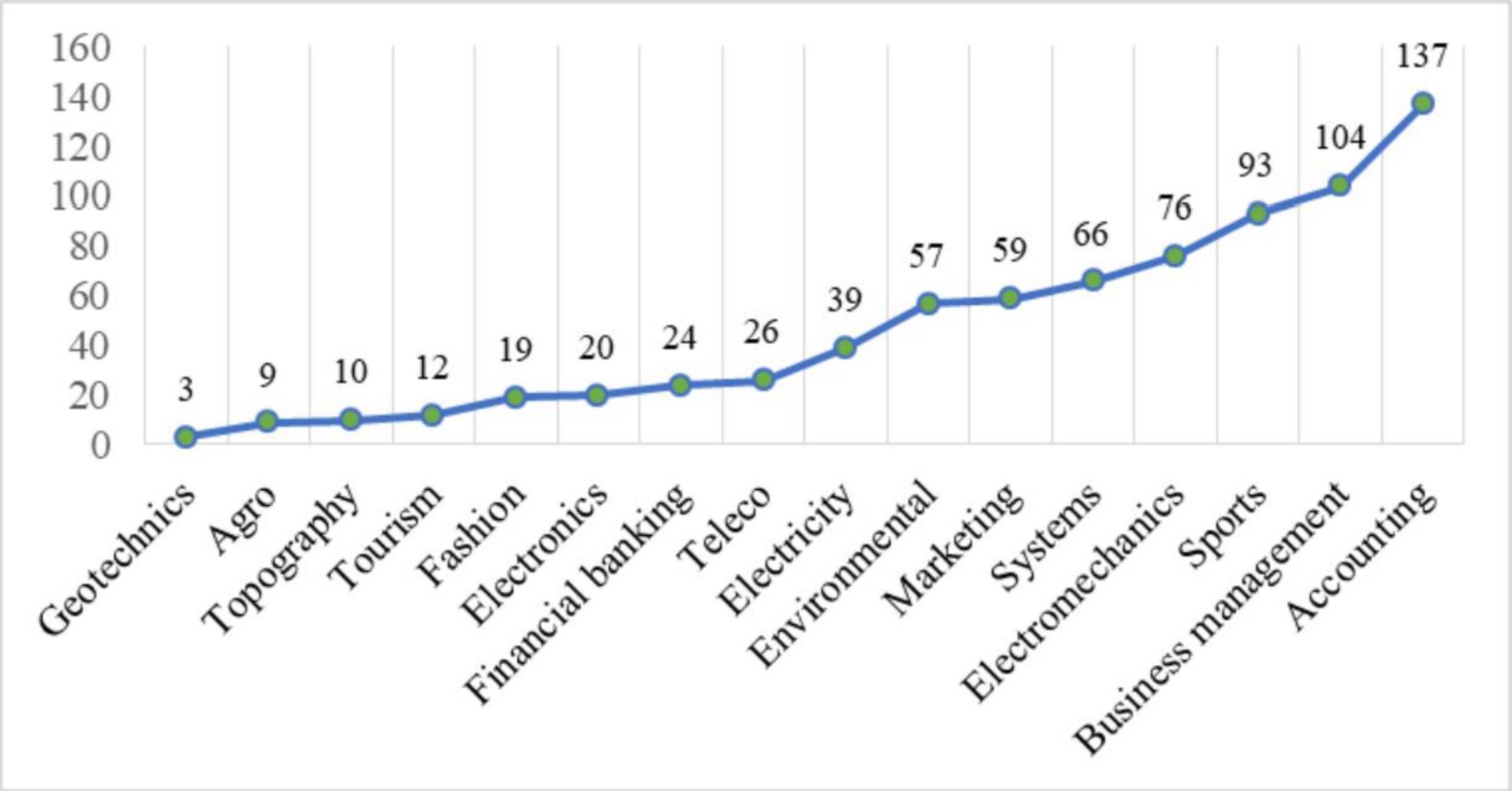


Figure 4. Line graph of the distribution of the number of PREIN students by academic programs

Diaz and Rueda 2020

Descriptive Analytics of Data: Central Tendency and Dispersion

Measuring central tendency

Mean

Arithmetical average
of a data distribution

$$\frac{\textit{Sum of all measured values}}{\textit{Total count of values}}$$

Median

Middle value of
a set of data

- Uneven amounts of values: middle number
- Even amounts of values: average of the two middle numbers

Mode

The value with the most frequent
occurrence in a set of data

What is value has
the highest frequency?

Measuring dispersion

Standard Deviation

Average deviation of values
from the mean – showing extent
of distribution

The average of all distances
between the values and the mean.

Range

Distance between the lowest and
the highest value

Overall spread of value

Interquartile range

Distance between the upper and
the lower quartile

Range without influence
of extreme values

Wilson 2014, 244f

Descriptive Analytics of Data: Describing Changes

GISMA BUSINESS SCHOOL

Describing Change

Index Numbers

Measure the change of quantity
over time of one homogenous item
(e.g. change of car prices over time)

$$\frac{\textit{Current year item price or costs}}{\textit{Base year item price or costs}}$$

Weighted Index Numbers

Measure the change of quantity
of time of group of heterogenous item
(e.g. retail price index)

Like normal index numbers, only weighted with item quantity in numerator and denominator

Describing frequency distribution

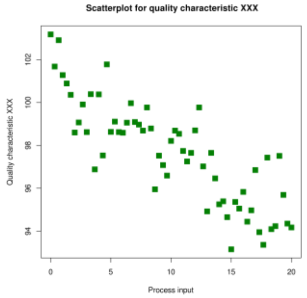
Cross-tabulations

Table showing the joint distribution of two variables

Nationality	Male	Female
British	35	37
German	29	23

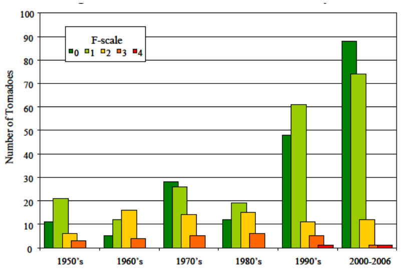
Scatter Diagrams

Graph showing relationship between two variables



Multiple Bar Chart

Bar chart comparing two or more variables for each year of comparison



By DanielPenfield - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=9402369; Wirlson 2014, 250

Inferential Analytics of Data

Method	Purpose	Example of Application
Hypothesis testing	Estimation	H0 – There is no difference in the mean exam marks between male and female manager H1 – There is no difference in the mean exam marks between male and female manager
Confidence testing	Estimation	Calculating a 95% confidence interval for the proportion of small firms in London that business with Europe.
Time series intervals	Forecasting	One-month changing averages of retail sales data
Pearson's product moment correlation coefficient (P)	Measuring association	Correlating gender with height
Spearman's rank correlation coefficient (NP)	Measuring association	Comparing two manager's ranked assessments of ten employees
Chi-squared test	Measuring difference	Do some manufactures produce more faulty goods than others?
Student's t-test	Measuring difference	Comparing the sample means of ages of female finance and marketing managers (independent t-test)
Simple regression	Assessing strength of relationship between variables	Strength of relationship between advertising and sales
Multiple regression	Assessing strength of relationship between variables	Strength of relationship between advertising spend and training spend on sales

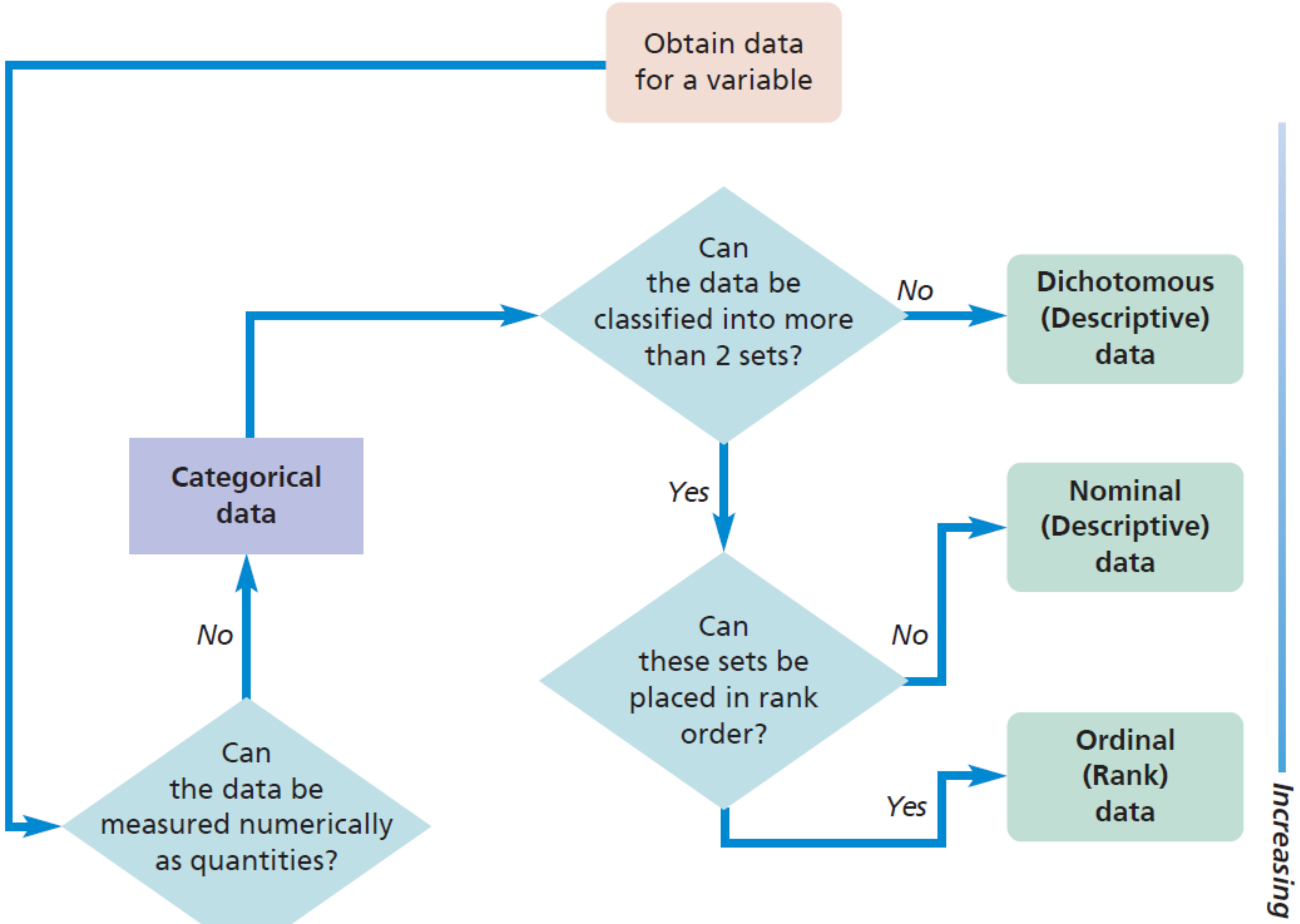
Wilson 2014, 255

Quantifying data

- Assign values to answers
- Coding
- Entered in a data file
- “Data matrix”

Defining the data type

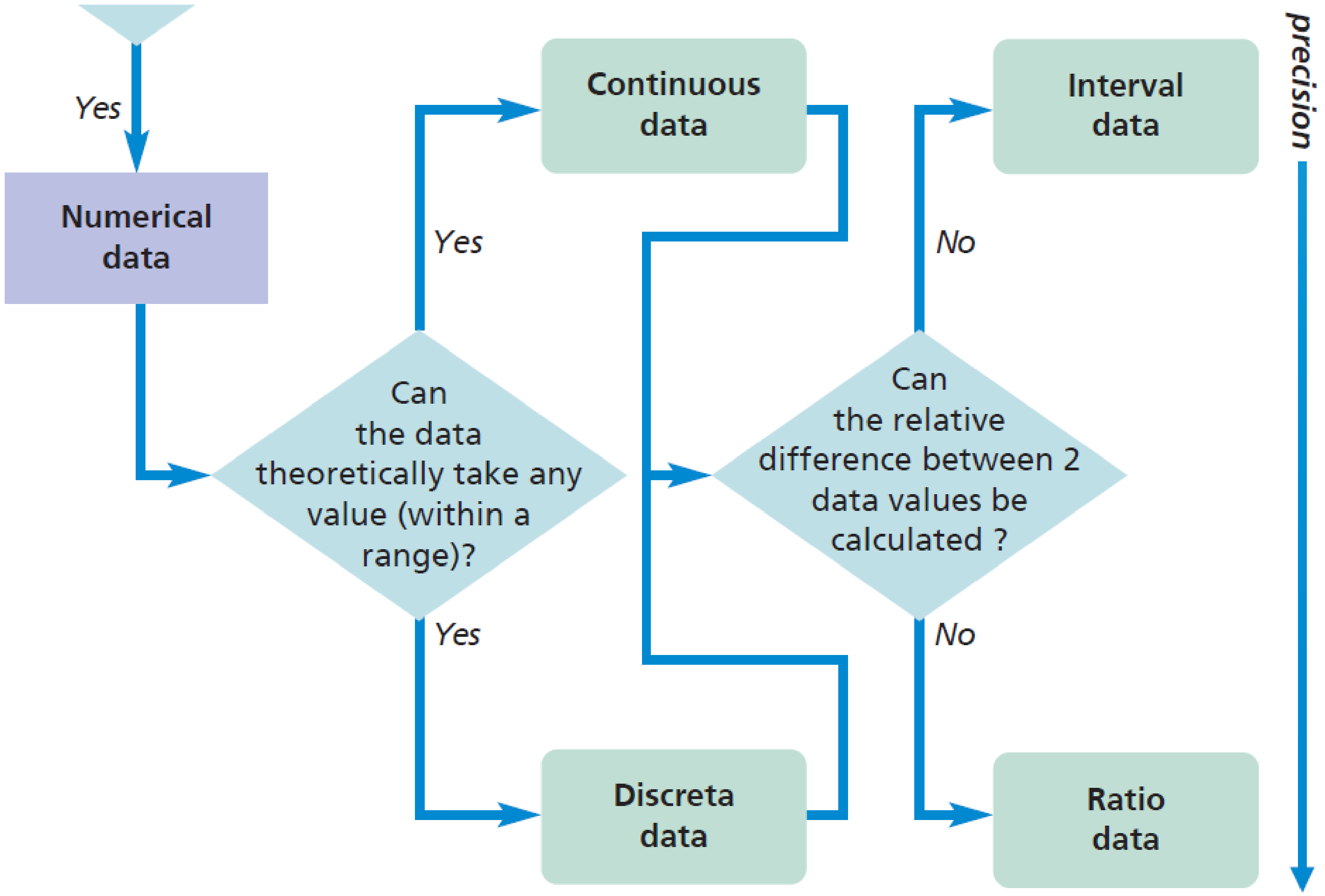
GISMA BUSINESS SCHOOL



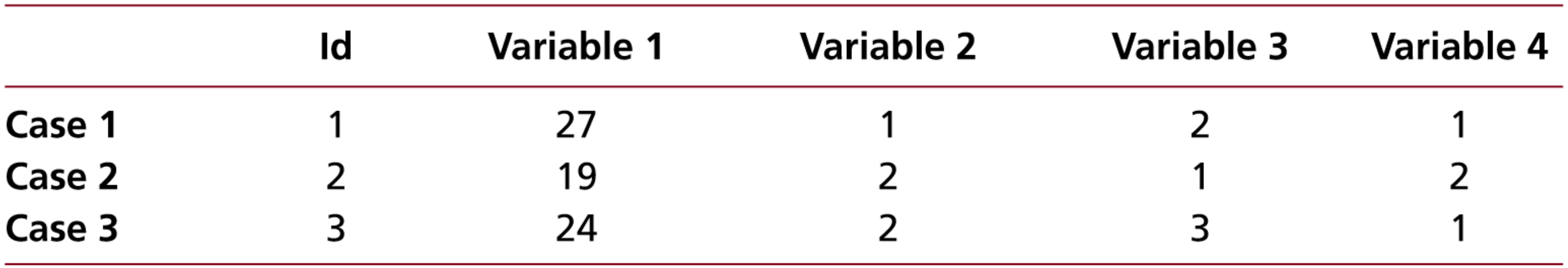
Obtain data for a variable
Can the data be classified into more than 2 sets?
No
Dichotomous (Descriptive) data
Categorical data
Yes
Nominal (Descriptive) data
No
No
Can these sets be placed in rank order?
Can the data be measured numerically as quantities?
Yes
Ordinal (Rank) data
Increasing

Source: © Mark Saunders, Philip Lewis and Adrian Thornhill 2018

Defining the data type


Yes
Numerical data
Can the data theoretically take any value (within a range)?
Yes
Continuous data
Yes
Discreta data
Can the relative difference between 2 data values be calculated ?
No
Interval data
No
Ratio data
precision

The data matrix



	Id	Variable 1	Variable 2	Variable 3	Variable 4
Case 1	1	27	1	2	1
Case 2	2	19	2	1	2
Case 3	3	24	2	3	1

What is in the matrix?

- Original data (coded from the questionnaire)
- Computed variables
  - Composite variables
    - Scale scores
  - Recoded variables

Codebook

- Variable name
- Related question(s)
- Values and labels
- Format

Quality control: data editing

- Value checks
  - 6 on a five point scale
- Data should be checked on logic
  - Certain combinations are illogical
    - 12 years and married
    - 6 years and college educated
    - Man and pregnant
    - 85-year-old granny, smokes 60 cigarettes per day, runs marathons

Relation variables and questions

- One multiple response question translates into multiple variables
- Which of the following brands do you use?
  - ❑ Nike
  - ❑ Adidas
  - ❑ Converse
  - ❑ Puma
  - ❑ Others, …

- ➢ Four variables coded 0/1
- ➢ The fifth category is an open question
  - ➢ Unlisted brands have to be coded afterward!
  - ➢ Leading to some more 0/1 coded questions!

Software

- Excel (Basic)

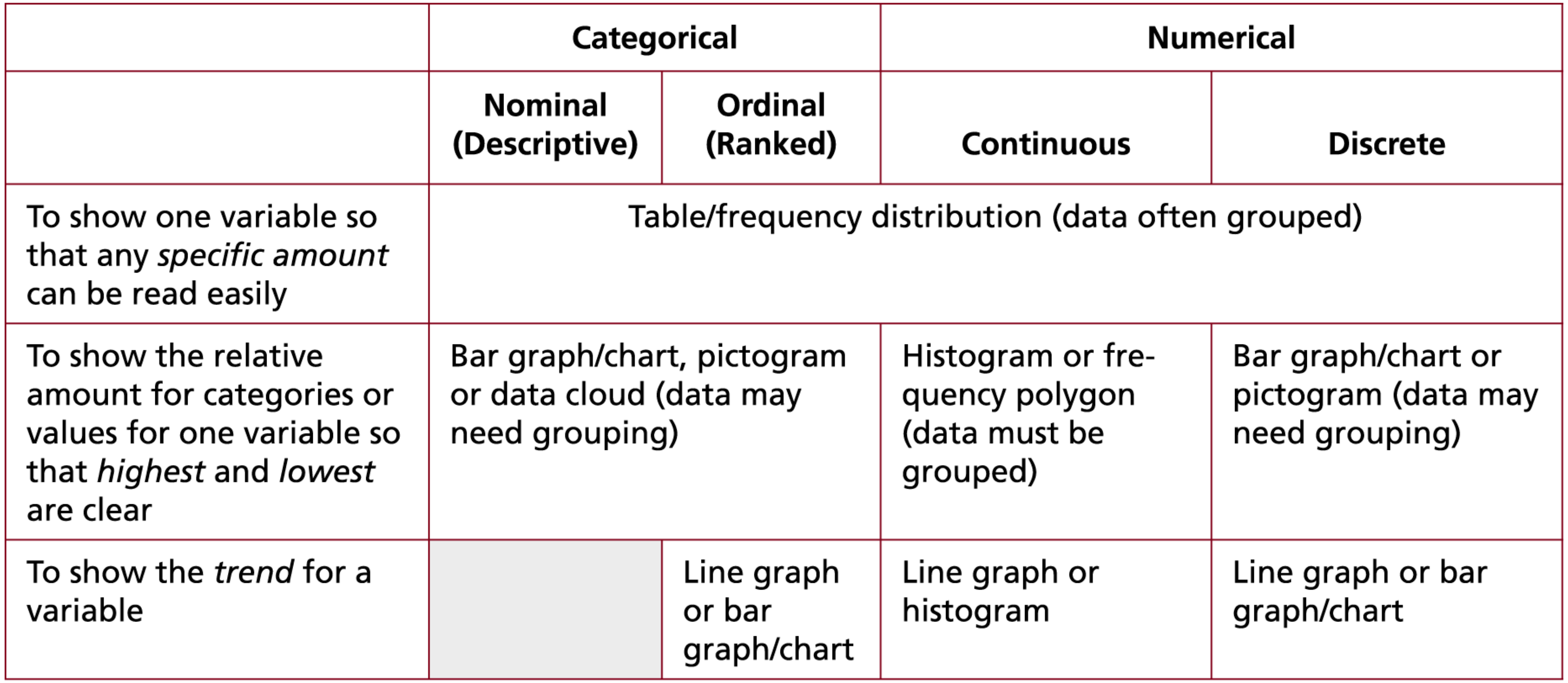
Advanced:

- R (advanced); freely downloadable
- Python (advanced); freely downloadable
- SPSS (expensive)
- STATA (available in quarterly licenses)
- Advantages:
  - Codebook is included
  - Powerful statistical techniques are available
  - Syntax: all the data manipulations and procedures are stored in a file (log file; or script). Everything can be repeated with one press on the button!
  - Replicable; repeatable

Compatibility:

- Most formats can be imported and exported
- E.g. data entered in Excel can be imported in SPSS, STATA or in R

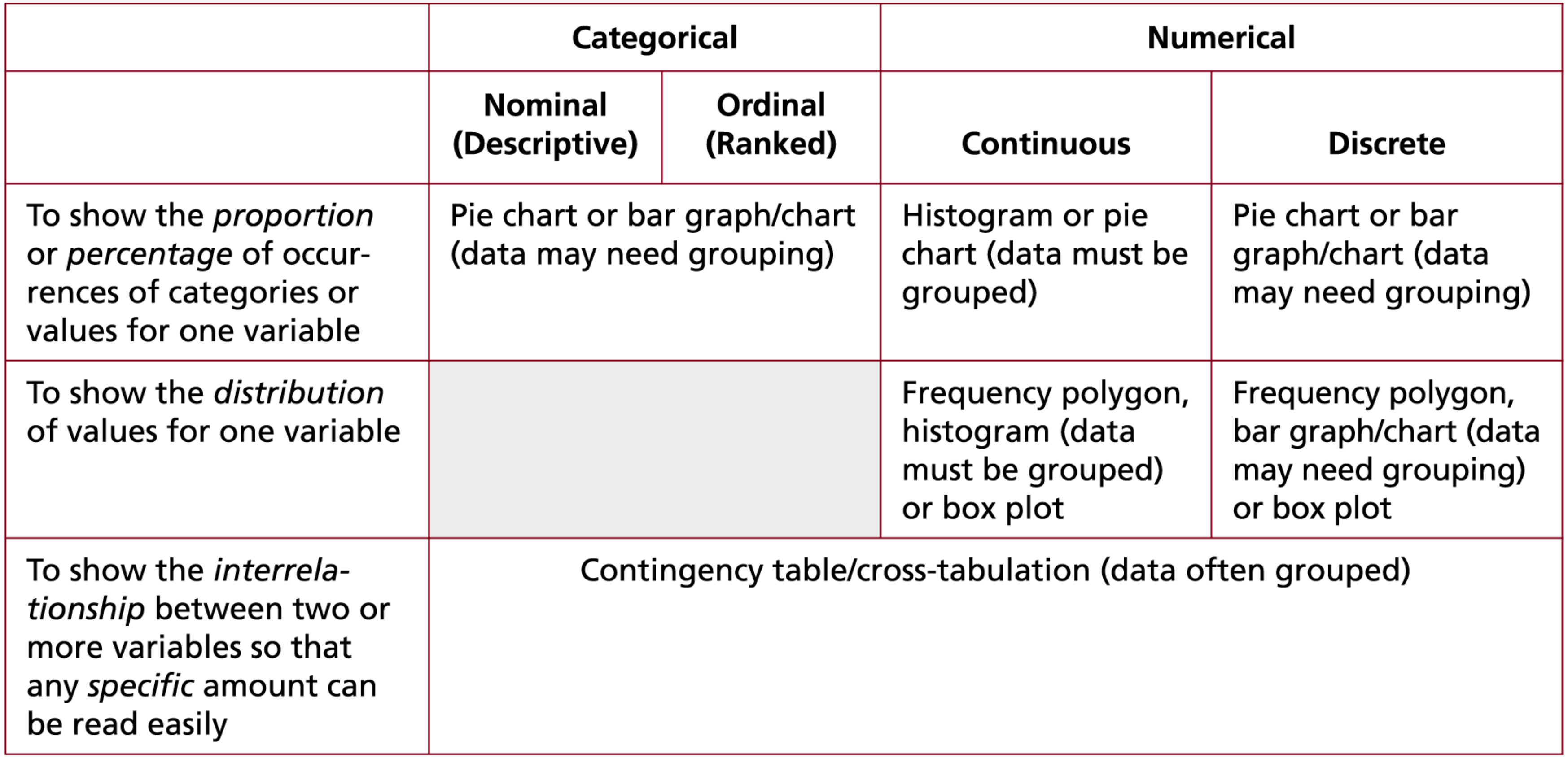
Data presentation by data type: A summary



	Categorical		Numerical	
	Nominal (Descriptive)	**Ordinal (Ranked)**	**Continuous**	**Discrete**
To show one variable so that any *specific amount* can be read easily	Table/frequency distribution (data often grouped)			
To show the relative amount for categories or values for one variable so that *highest* and *lowest* are clear	Bar graph/chart, pictogram or data cloud (data may need grouping)		Histogram or frequency polygon (data must be grouped)	Bar graph/chart or pictogram (data may need grouping)
To show the *trend* for a variable		Line graph or bar graph/chart	Line graph or histogram	Line graph or bar graph/chart

Source: © Mark Saunders, Philip Lewis and Adrian Thornhill 2018

Data presentation by data type: A summary



	Categorical		Numerical	
	Nominal (Descriptive)	**Ordinal (Ranked)**	**Continuous**	**Discrete**
To show the *proportion* or *percentage* of occurrences of categories or values for one variable	Pie chart or bar graph/chart (data may need grouping)		Histogram or pie chart (data must be grouped)	Pie chart or bar graph/chart (data may need grouping)
To show the *distribution* of values for one variable			Frequency polygon, histogram (data must be grouped) or box plot	Frequency polygon, bar graph/chart (data may need grouping) or box plot
To show the *interrelationship* between two or more variables so that any *specific* amount can be read easily	Contingency table/cross-tabulation (data often grouped)			

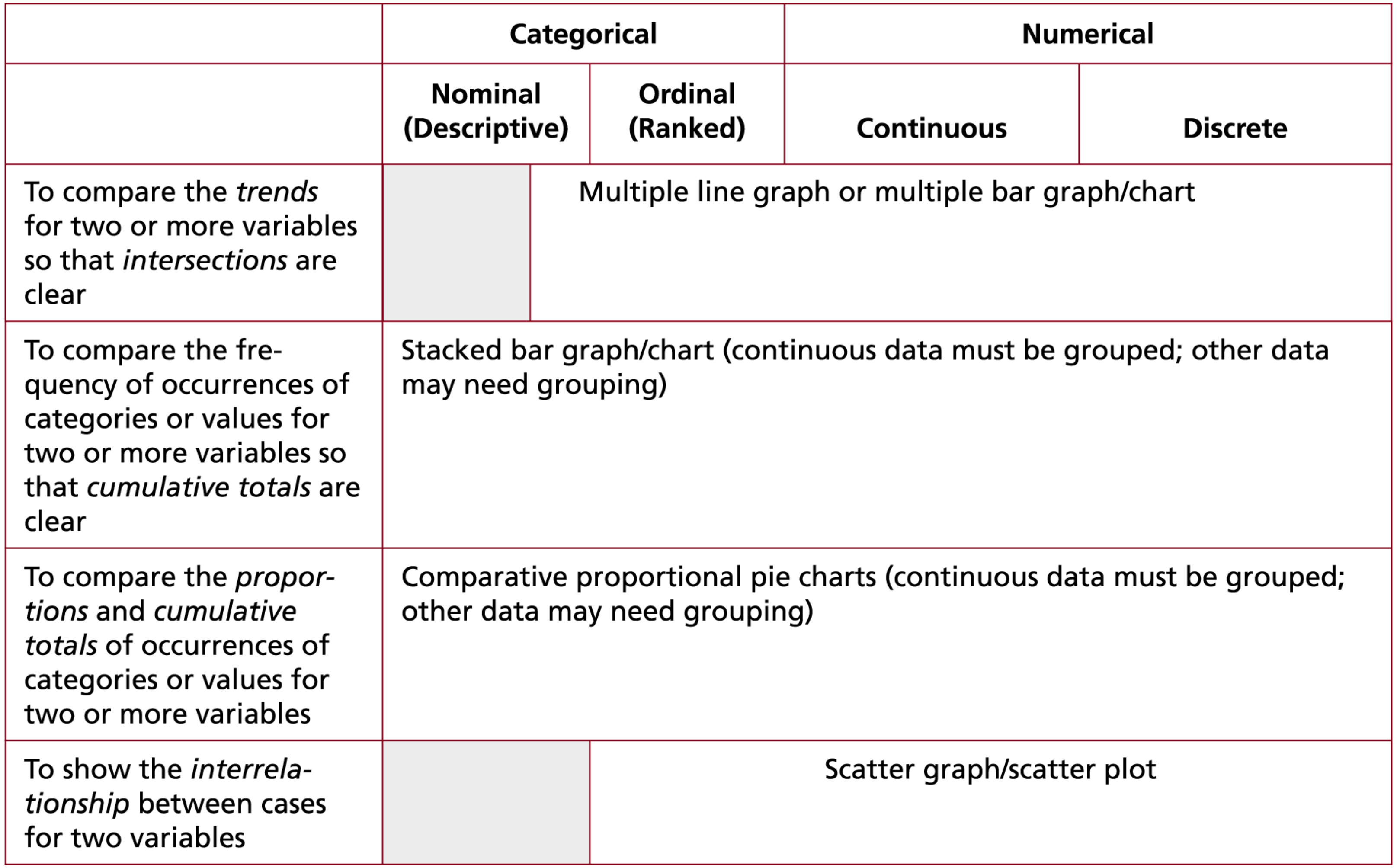
Source: © Mark Saunders, Philip Lewis and Adrian Thornhill 2018

Data presentation by data type: A summary

	Categorical		Numerical	
	Nominal (Descriptive)	**Ordinal (Ranked)**	**Continuous**	**Discrete**
To compare the relative amount for categories or values for two or more variables so that *highest* and *lowest* are clear	Multiple bar graph/chart (continuous data must be grouped; other data may need grouping)			
To compare the *proportions* or *percentages* of occurrences of categories or values for two or more variables	Comparative pie charts or percentage component bar graph/chart (continuous data must be grouped; other data may need grouping)			
To compare the *distribution* of values for two or more variables			Multiple box plot	

Source: © Mark Saunders, Philip Lewis and Adrian Thornhill 2018

Data presentation by data type: A summary



	Categorical		Numerical	
	Nominal (Descriptive)	Ordinal (Ranked)	Continuous	Discrete
To compare the *trends* for two or more variables so that *intersections* are clear		Multiple line graph or multiple bar graph/chart		
To compare the frequency of occurrences of categories or values for two or more variables so that *cumulative totals* are clear	Stacked bar graph/chart (continuous data must be grouped; other data may need grouping)			
To compare the *proportions* and *cumulative totals* of occurrences of categories or values for two or more variables	Comparative proportional pie charts (continuous data must be grouped; other data may need grouping)			
To show the *interrelationship* between cases for two variables		Scatter graph/scatter plot		

Source: © Mark Saunders, Philip Lewis and Adrian Thornhill 2018

Data preparation (2)

- Missing values
  - Partial non-response
  - Coding them
    - Blank (dangerous!)
    - Preferred: 8 or 9 (or 88 or 99)
    - Tell software to treat those values as missing
  - Analyzing them
    - Delete list-wise or pair-wise

Another group assignment

- What effect does social media have on people’s minds?
- What effect does daily use of Twitter have on the attention span of under-16s?

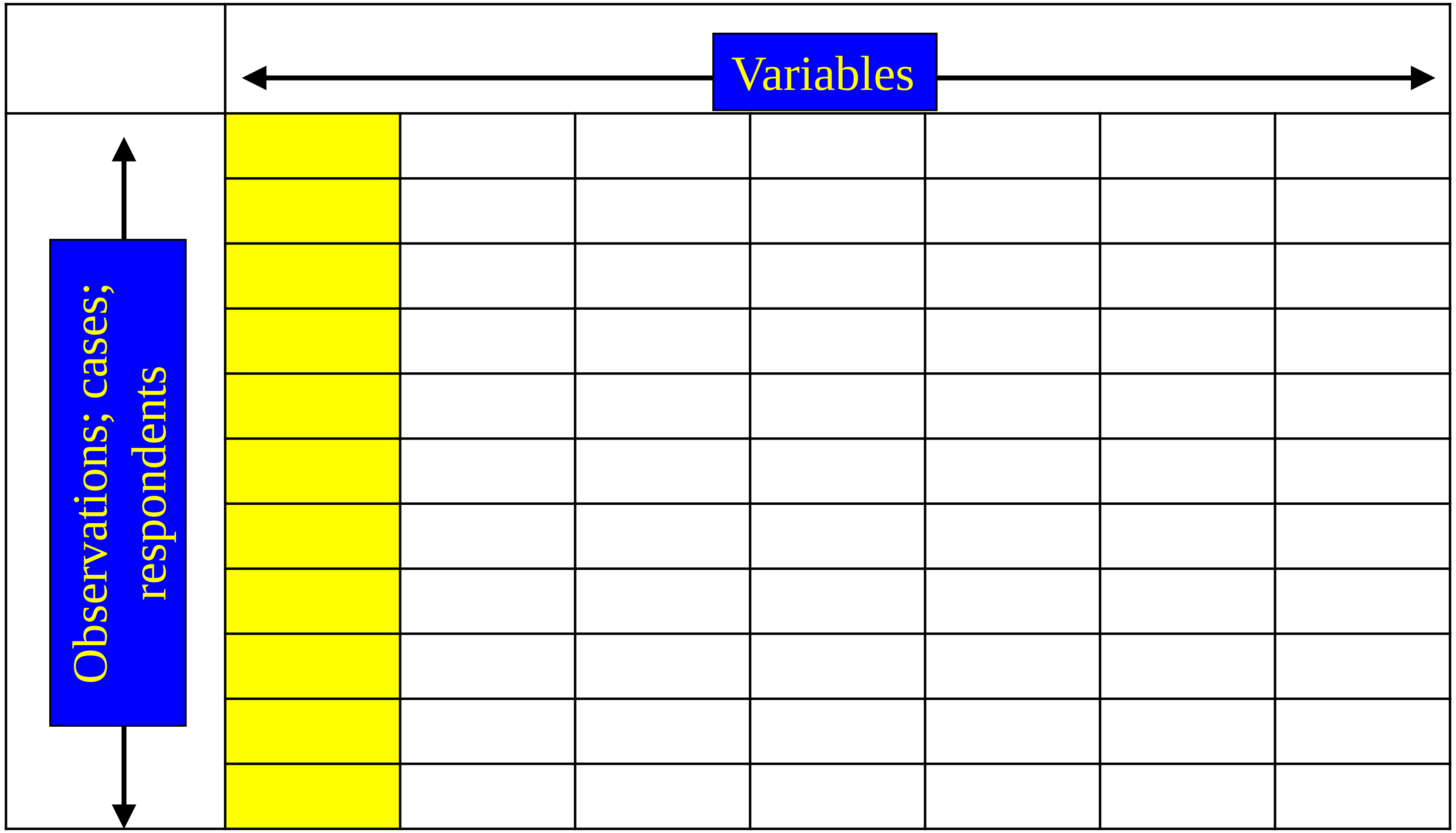
Methodology

- Describe how you have treated the data in the methodological section
  - After data collection
  - Before findings and analyses

Depending on the RQ …

1. Describing variables (*descriptives*)
   - Frequencies
   - Central tendency: mode, median, mean
2. Differences between groups
   - Chi-square
   - T-test, for difference between 2 groups
   - ANOVA (analysis of variance), for differences between 3 or more groups
3. Relationships between phenomena
   - Regression analysis (log-linear; ordinal)

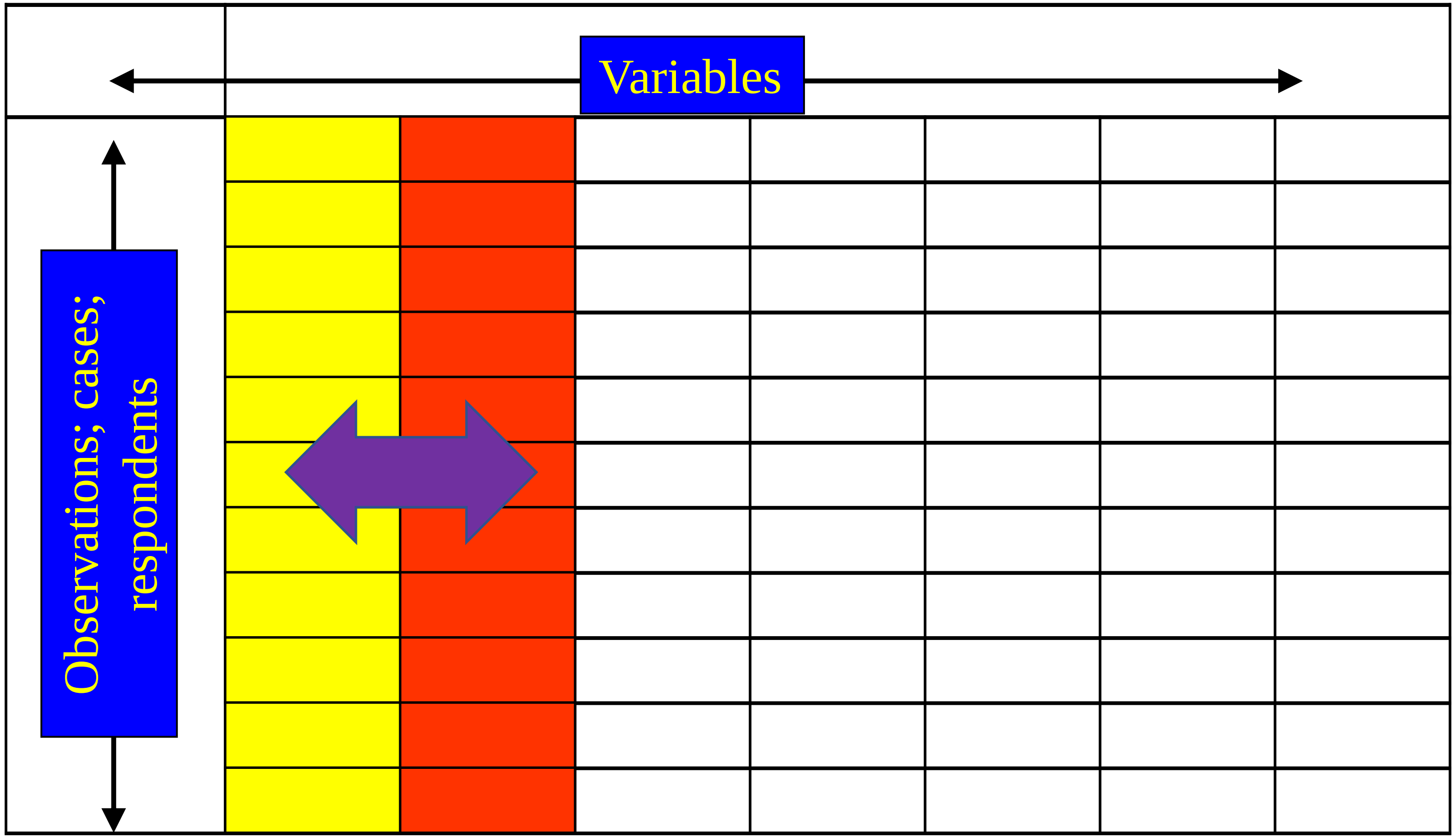
Data analysis (1): frequencies



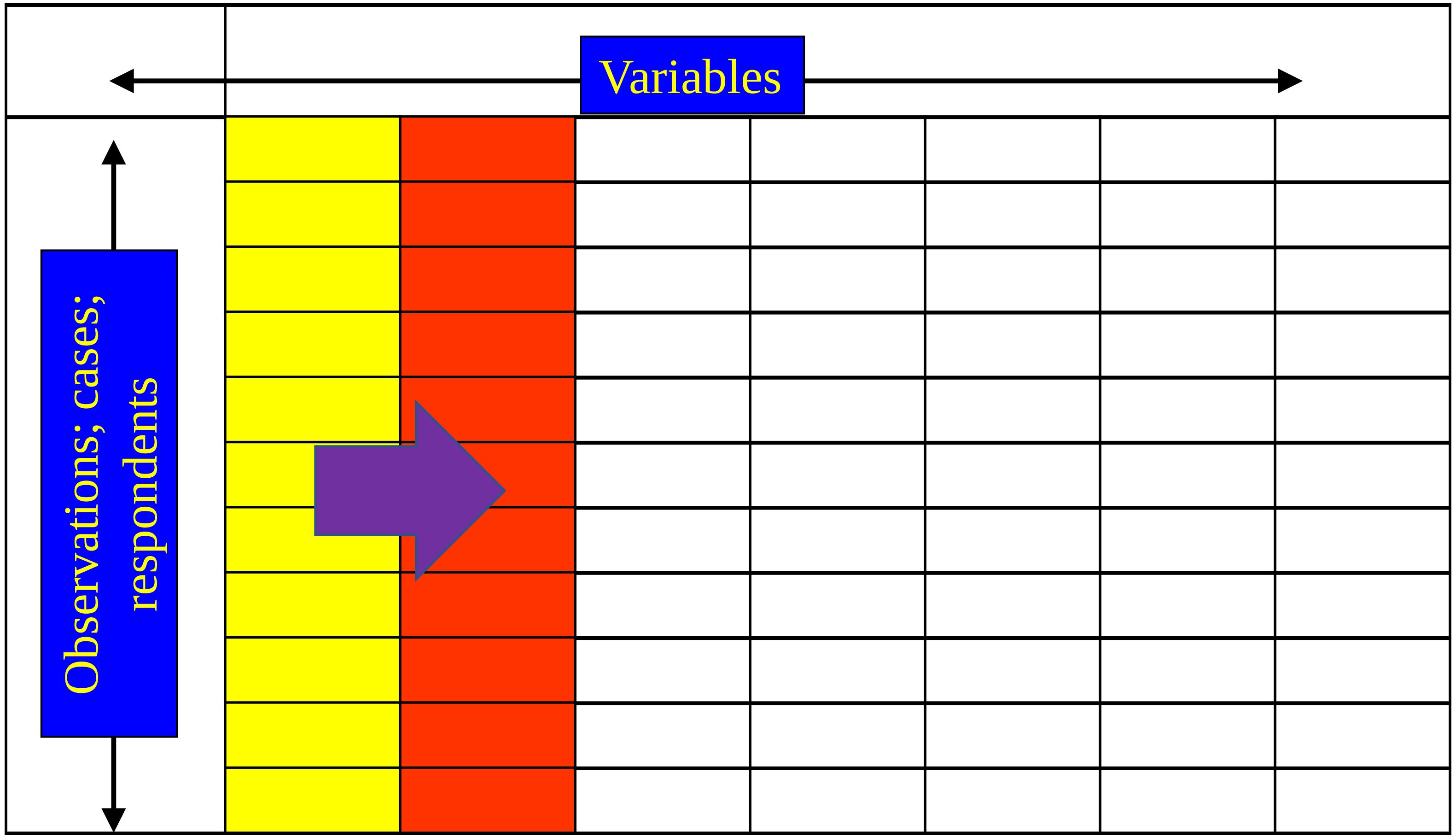
Variables

Observations; cases; respondents

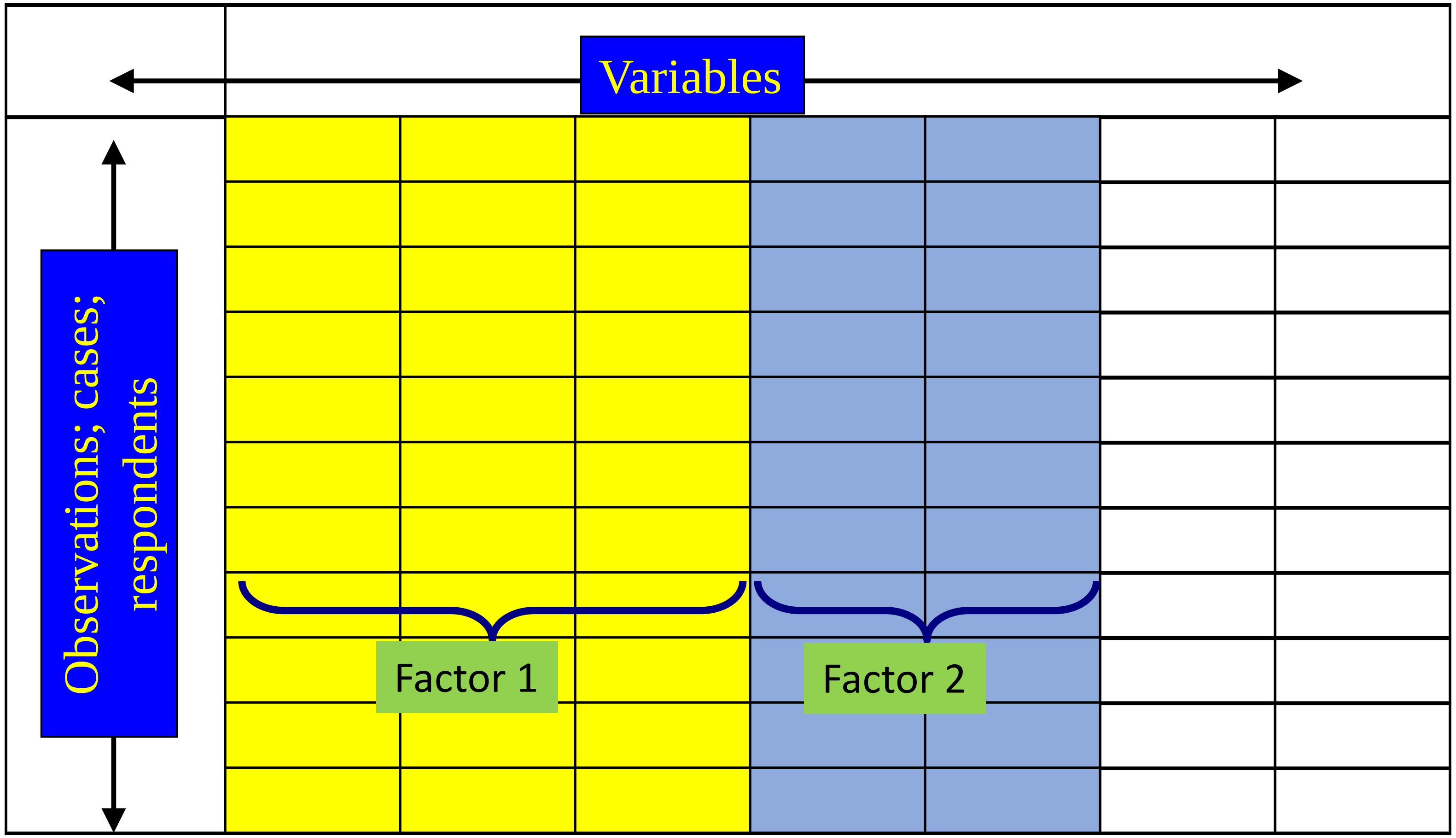
Data analysis (2): correlations



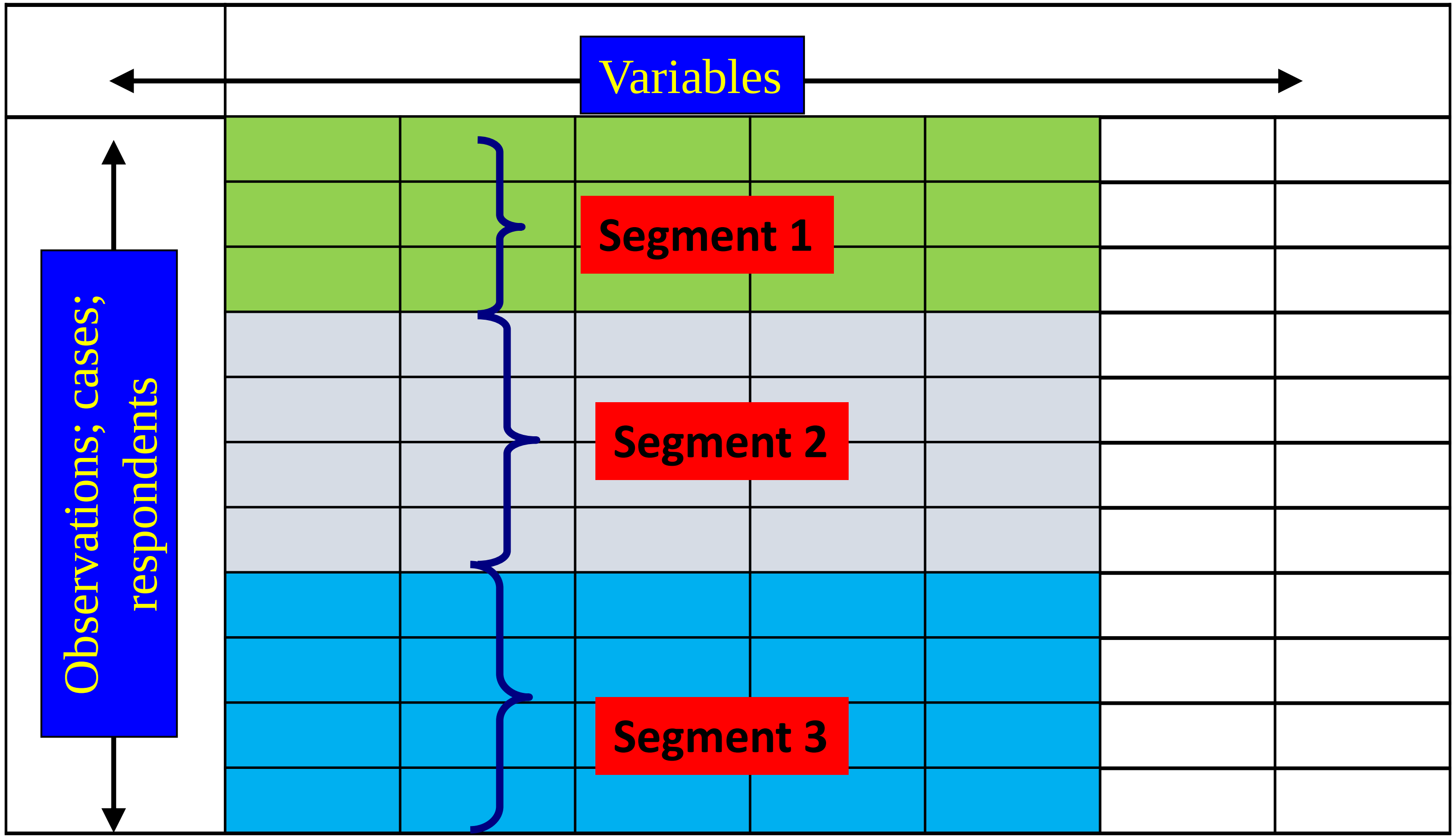
Data analysis (3): causal relations



Data analysis (4): data reduction



Data analysis (5): grouping cases



Describing variables

- Frequencies

- Descriptive
  - Minimum
  - Maximum
  - Percentiles
  - Mode, mean, median
  - Outliers

Percentiles

- The percentile or percentile rank is the percentage of values in a distribution that is <= a certain value.
- This can be used to answer the question: What percentage is smaller, lighter, worse ... ? - or the reverse question: How many percent are bigger, heavier, better ... ? ?

Percentiles - example

Of the 25 students in a class, 5 have written a grade A, 5 a grade B, 5 a grade C, 5 a grade D and 5 a grade E in an exam, where A is the best grade and E the worst grade.

If a student has, for example, a grade of B, he does not necessarily know yet whether he has really done well (perhaps all the others have a grade of A).

However, if the teacher tells the student with the grade B that he is in the 80% percentile, he knows that 80% of the students have a grade B or below ("worse") and only 20% above (just the 5 students who have a grade A).So the percentile can be used as a benchmark to rank a certain value.

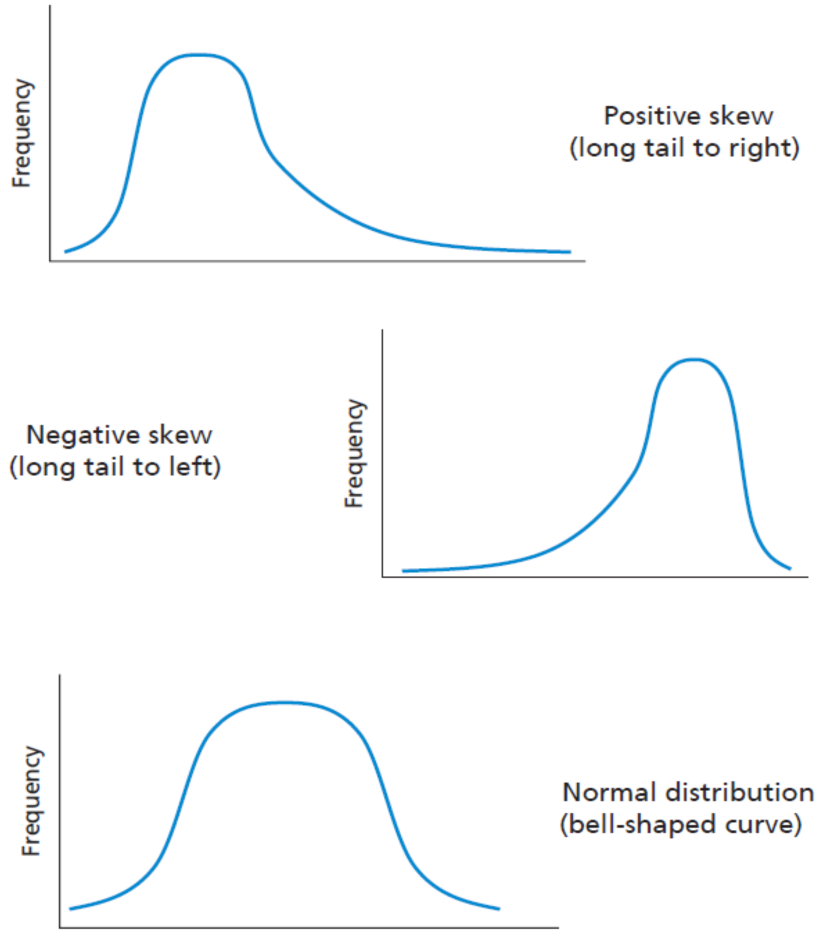
Relationships (Central tendency)

- For interval and ratio variables
- If the distribution of values is normal or symmetric:
  - Mode = Median = Mean
- If skewed to the right (outliers on upper end)
  - Mean > Median > Mode

Example

- The mean income of the population is 120
- The median income is 100
- Evidently there are outliers: persons with high incomes

Frequency polygons showing distributions of values



Positive skew
(long tail to right)

Negative skew
(long tail to left)

Normal distribution
(bell-shaped curve)

Source: © Mark Saunders, Philip Lewis and Adrian Thornhill 2018

Dispersion (Central tendency)

- Mean = average, calculated by adding all values, then dividing by number of items
- Example: age of employees in a team –
  - 22, 34, 46, 42, 43, 74, 42, 15, 42, 46, 37, 29, 51, 28, 21, 30
  - N = 16
  - Mean = 602/16 = 37,6
- Median = the middle value of a series of individual results when we have outliers
  - 15, 21, 22, 28, 29, 30, 34, **37**, **42**, 42, 42, 43, 36, 46, 51, 74
  - Median = the sum of two middle values/2 = (37+42)/2 = 39,5
- Mode = rarely used, indicates the most frequently occurring value, i.e. 42
- Range = the difference between the highest and lowest value i.e., 59 (15 – 74 years)
- Interquartile range = the dispersion of the inner 50%
  - Omit the highest and lowest quarter of the measures and measure the range of the inner 50%
  - Range 29 – 43 = 14 years

Difference between groups

T-test

- Independent samples
  - Two groups in cross-sectional design
- Dependent samples
  - One group, two measurements in longitudinal design
- Requirement: normally distributed variables
- Otherwise: non-parametric test

More than two groups

- Analysis of Variance (ANOVA)
- Dependent variable is interval/ratio
- Independent is nominal
  - Groups
  - Experimental variable (instrument)

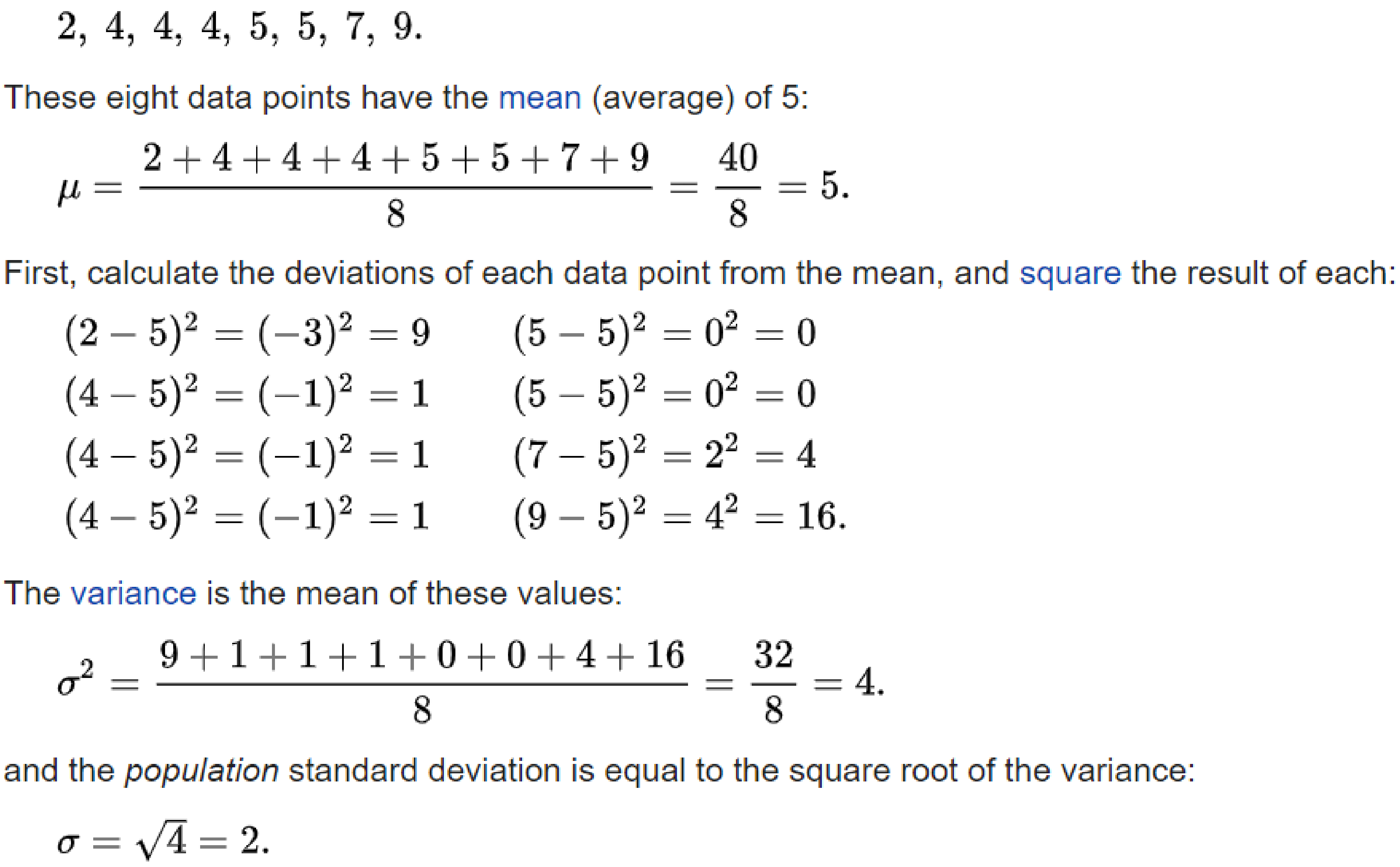
T-test

- For example, you have to treatments i.e., one group that has undergone some treatment (the experimental group) and one group that has not undergone a treatment (the baseline group)
- You find different means and what to find out whether the difference is significant i.e., an effect of the treatment (or random)
- The *t*-test can be used, for example, to determine if the means of two sets of data are significantly different from each other.

Standard Deviation

- a measure of the amount of variation or dispersion of a set of values
- low standard deviation indicates that the values tend to be close to the mean (also called the expected value) of the set, while a high standard deviation indicates that the values are spread out over a wider range
- A standard deviation is a statistic that measures the dispersion of a dataset relative to its mean. The standard deviation is calculated as the square root of variance by determining each data point's deviation relative to the mean. If the data points are further from the mean, there is a higher deviation within the data set; thus, the more spread out the data, the higher the standard deviation

Standard Deviation



$$2,\ 4,\ 4,\ 4,\ 5,\ 5,\ 7,\ 9.$$

These eight data points have the mean (average) of 5:

$$\mu = \frac{2+4+4+4+5+5+7+9}{8} = \frac{40}{8} = 5.$$

First, calculate the deviations of each data point from the mean, and square the result of each:

$$\begin{array}{ll} (2-5)^2 = (-3)^2 = 9 & (5-5)^2 = 0^2 = 0 \\ (4-5)^2 = (-1)^2 = 1 & (5-5)^2 = 0^2 = 0 \\ (4-5)^2 = (-1)^2 = 1 & (7-5)^2 = 2^2 = 4 \\ (4-5)^2 = (-1)^2 = 1 & (9-5)^2 = 4^2 = 16. \end{array}$$

The variance is the mean of these values:

$$\sigma^2 = \frac{9+1+1+1+0+0+4+16}{8} = \frac{32}{8} = 4.$$

and the *population* standard deviation is equal to the square root of the variance:

$$\sigma = \sqrt{4} = 2.$$

Output

- What to present?
- Computer output is
  - Complicated
  - Detailed
  - Full of redundancies in which the reader is not interested!
- Therefore:
  - NEVER use computer output directly in the text!!
  - PREFERABLY do not use it even in the appendices!!
  - Stick to key statistics, and use your own preferred format
- Key statistics: significance
  - Concept of significance
  - Understand what a significance test is all about, in the statistical procedure you have used!!

Statistics to examine relationships, differences and trends **by data type:** A **summary**

	Categorical		Numerical	
	Nominal (Descriptive)	Ordinal (Ranked)	Continuous	Discrete
To test *normality* of distribution			Kolmogorov-Smirnov test, Shapiro-Wilk test	
To test whether two variables are *independent*	Chi square (data may need grouping)		Chi square if variable grouped into discrete classes	
To test whether two variables are *associated*	Cramer's V and Phi (both variables must be dichotomous)			

Source: © Mark Saunders, Philip Lewis and Adrian Thornhill 2018

Statistics to examine relationships, differences and trends **by data type: A summary**

	Categorical		Numerical	
	Nominal (Descriptive)	**Ordinal (Ranked)**	**Continuous**	**Discrete**
To test whether two groups (categories) are *different*		Kolmogorov-Smirnov (data may need grouping) or Man-Whitney *U* test	Independent *t*-test or paired *t*-test (often used to test for changes over time) or Mann-Whitney *U* test (where data skewed or a small sample)	
To test whether three or more groups (categories) are *different*			Analysis of variance (ANOVA)	

Source: © Mark Saunders, Philip Lewis and Adrian Thornhill 2018

Statistics to examine relationships, differences and trends by data type: A summary

	Categorical		Numerical	
	Nominal (Descriptive)	**Ordinal (Ranked)**	**Continuous**	**Discrete**
To assess the *strength of relationship* between two variables		Spearman's rank correlation coefficient or Kendall's rank order correlation coefficient	Pearson's product moment correlation coefficient (PMCC)	
To assess the strength of a relationship between one dependent and one independent variable			Coefficient of determination	

Source: © Mark Saunders, Philip Lewis and Adrian Thornhill 2018

Statistics to examine relationships, differences and trends **by data type:** A **summary**

	Categorical		Numerical	
	Nominal (Descriptive)	Ordinal (Ranked)	Continuous	Discrete
To assess the strength of a relationship between one dependent and two or more independent variables			Coefficient of multiple determination	
To *predict* the value of a dependent variable from one or more independent variables			Regression equation	
To explore *relative change* over time			Index numbers	

Source: © Mark Saunders, Philip Lewis and Adrian Thornhill 2018

Statistics to examine relationships, differences and trends **by data type: A summary**

	Categorical		Numerical	
	Nominal (Descriptive)	Ordinal (Ranked)	Continuous	Discrete
To compare *relative changes* over time			Index numbers	
To determine the trend over time of a series of data			Time series, moving averages or regression equation (regression analysis)	

Source: © Mark Saunders, Philip Lewis and Adrian Thornhill 2018

Significance

- Is there really an effect; are groups really different?
- Or are the results due to chance?
- Statistical significance
  - Tests and techniques always indicate significance
  - High significance is indicated, mostly, by low P-values
    - P-value represents the likelihood that a test statistic (χ^2; T; F; ..) is larger than the value found

Hypothesis testing

- H_0 is 'null hypothesis'
  - Is the hypothesis to be tested
  - For example: there is no difference between groups
- H_1 is 'alternative hypothesis'
  - What you expect as a researcher!
- The null hypothesis is
  - Tested;
    - If p<5% then we reject the null hypothesis
  - That is not the same as: "our" alternatively is proven!!
    - We just found support for our theory
  - "All crows are black": it's highly likely but you cannot prove it; you can reject it if you find a white one!

Example

You expect that absenteeism of sales staff differs from personnel department

- H_0: Absenteeism between the departments does not differ
- By rejecting H_0, you point out that personnel department differs from sales staff (although you have not proven it!)

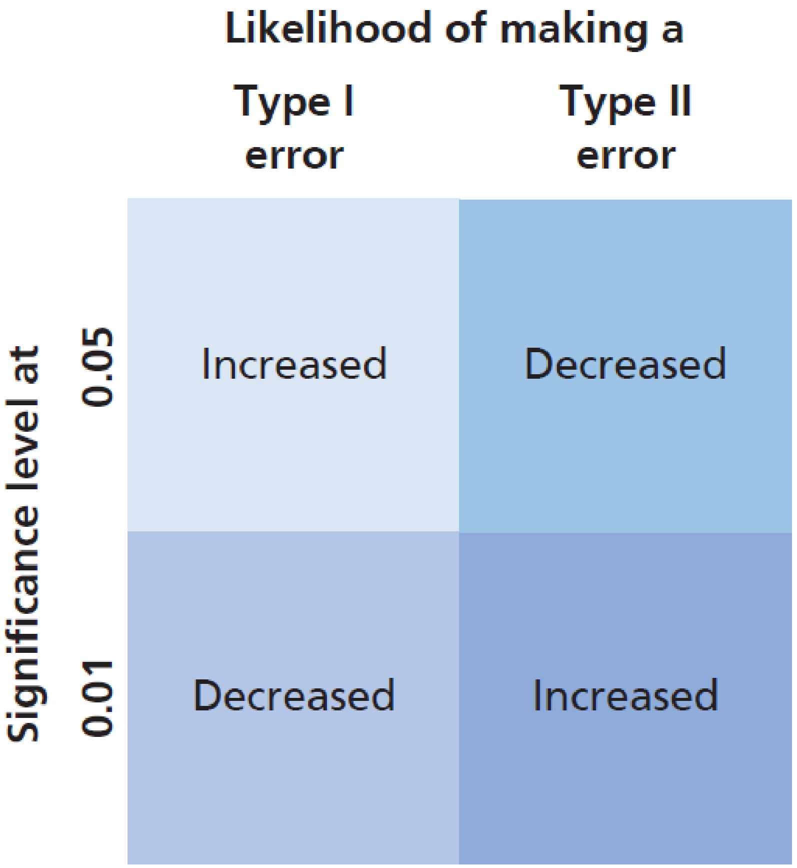
Types of errors

- Type I error: α
  - Null hypothesis is incorrectly rejected
- Type II error: β
  - Null hypothesis is incorrectly accepted
  - Power of the test: 1- β
- Decreasing one, increases the other!

<u>How high should α and β be?</u>

- In social and economic research for historical reasons, α is set at .05 (5%)
  - The confidence level is 95%
  - We reject the null hypothesis if the probability of a certain test statistic based on the null hypothesis to be true is less than 5%
- The power of the test, 1- β, is normally set at 80%
  - The chances that we find an effect if there is one, should be 80%
- Setting the α and β determines the size of the sample; in procedures that we call power analysis
- Setting α and β strongly depends on the context!

Type I and Type II errors



	Likelihood of making a	
Significance level at	Type I error	Type II error
0.05	Increased	Decreased
0.01	Decreased	Increased

Source: © Mark Saunders, Philip Lewis and Adrian Thornhill 2018