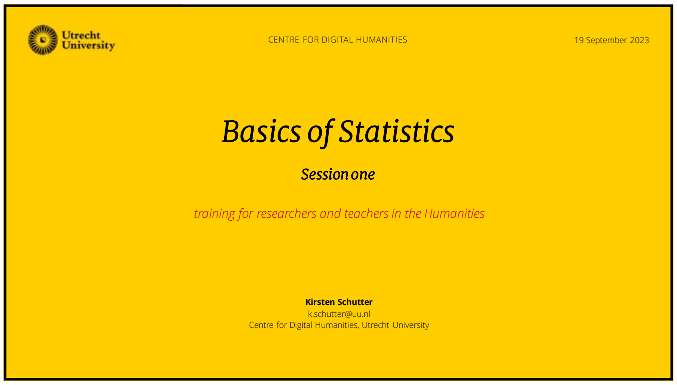


CENTRE FOR DIGITAL HUMANITIES

19 September 2023

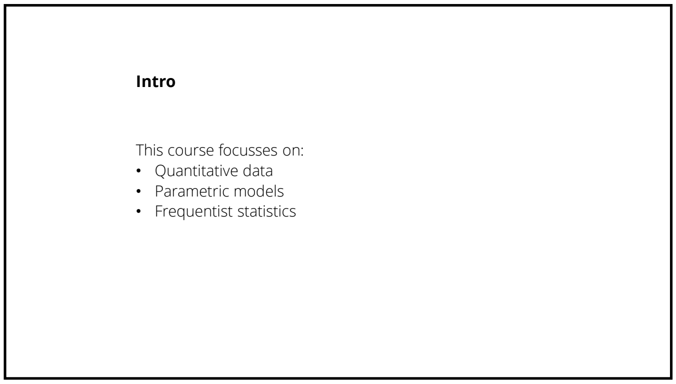
Basics of Statistics

Session one

training for researchers and teachers in the Humanities

Kirsten Schutter
k.schutter@uu.nl
Centre for Digital Humanities, Utrecht University

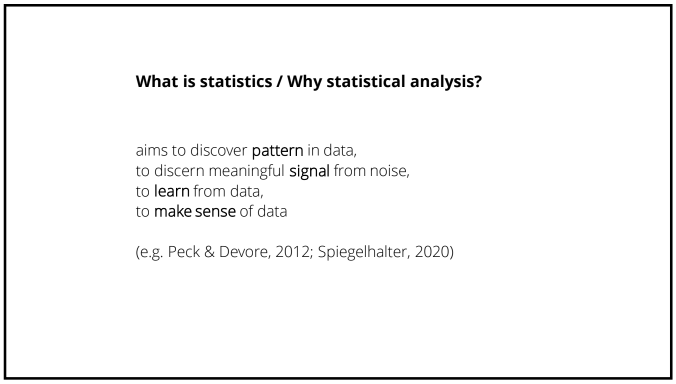
1



Intro

This course focusses on:
- Quantitative data
- Parametric models
- Frequentist statistics

2

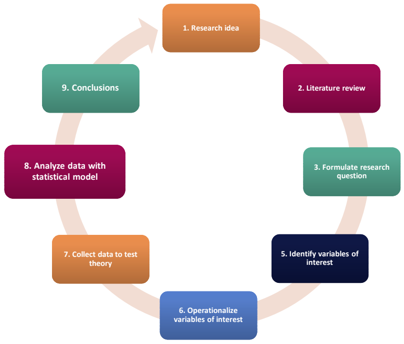


What is statistics / Why statistical analysis?

aims to discover pattern in data,
to discern meaningful signal from noise,
to learn from data,
to make sense of data

(e.g. Peck & Devore, 2012; Spiegelhalter, 2020)

3


1. Research idea
2. Literature review
3. Formulate research question
5. Identify variables of interest
6. Operationalize variables of interest
7. Collect data to test theory
8. Analyze data with statistical model
9. Conclusions

4

Research question

Specifying a research question is the methodological point of departure of scientific research

- Every scientific study starts with a research question
- The goal of a study is to answer this research question
- Thus, the study needs to be designed to answer the research question
- This is the research methodology / research design
- The research question defines the focus of your study
- A research question should have clear methodological implications for data collection and analysis

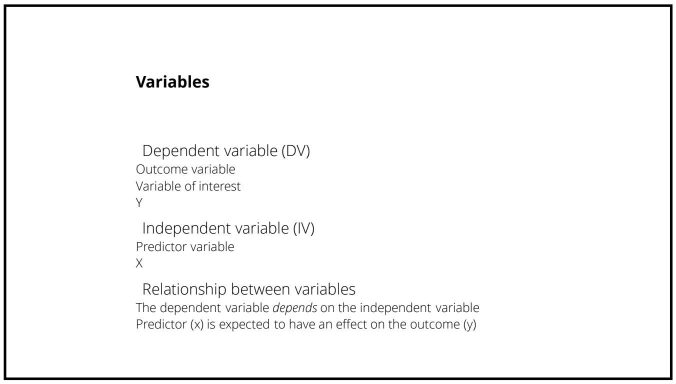
5

Variables

A variable is something you can measure (quantify) that varies across units

Participant	Sex	Height (cm)	Shoe size (EU)
1	Female	166	37
2	Female	170	39
3	Male	182	42
4	Male	173	41
5	Female	186	38

6



Variables

Dependent variable (DV)

Outcome variable

Variable of interest

Y

Independent variable (IV)

Predictor variable

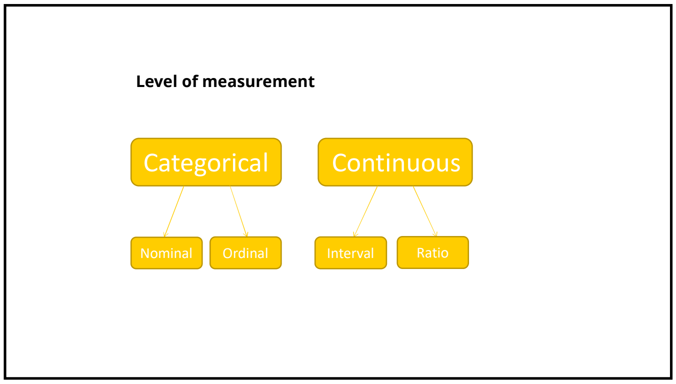
X

Relationship between variables

The dependent variable *depends* on the independent variable

Predictor (x) is expected to have an effect on the outcome (y)

7



Level of measurement

Categorical

Continuous

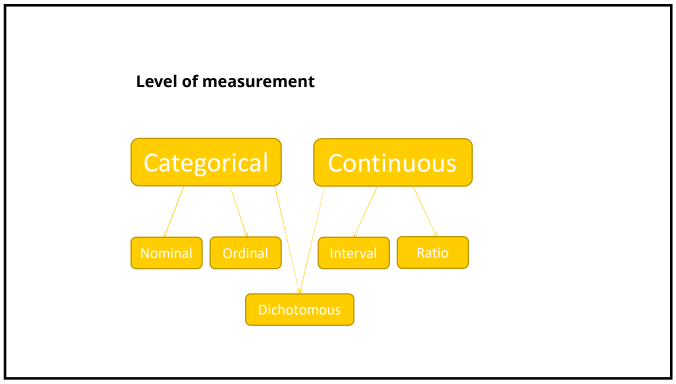
Nominal

Ordinal

Interval

Ratio

8



Level of measurement

Categorical

Continuous

Nominal

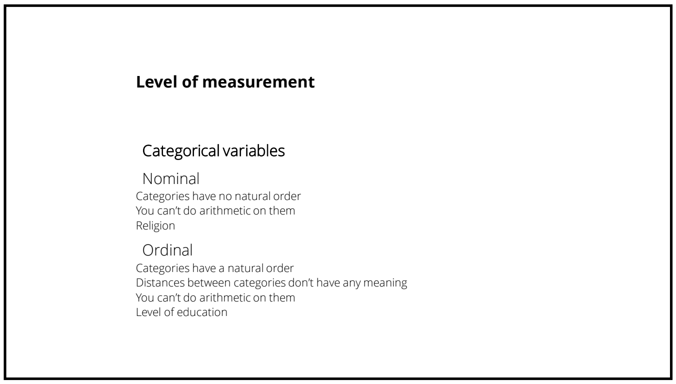
Ordinal

Interval

Ratio

Dichotomous

9



Level of measurement

Categorical variables

Nominal

Categories have no natural order

You can't do arithmetic on them

Religion

Ordinal

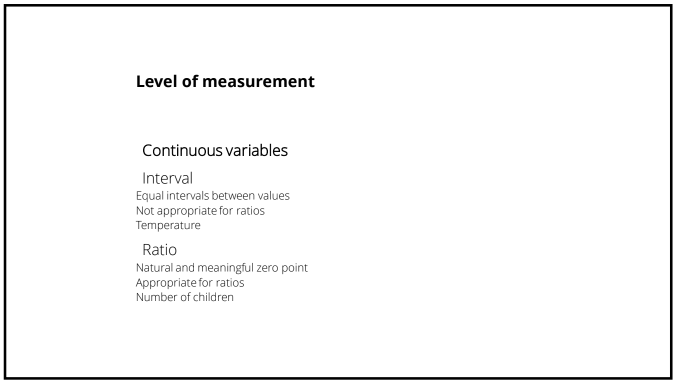
Categories have a natural order

Distances between categories don't have any meaning

You can't do arithmetic on them

Level of education

10



Level of measurement

Continuous variables

Interval

Equal intervals between values

Not appropriate for ratios

Temperature

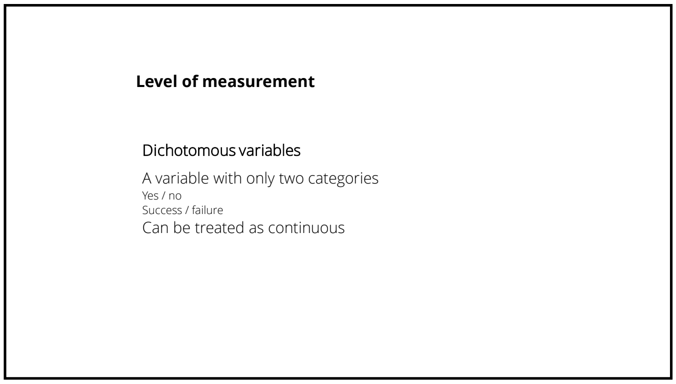
Ratio

Natural and meaningful zero point

Appropriate for ratios

Number of children

11



Level of measurement

Dichotomous variables

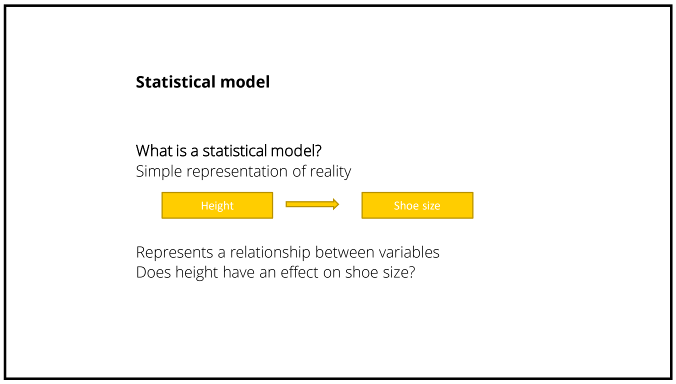
A variable with only two categories

Yes / no

Success / failure

Can be treated as continuous

12



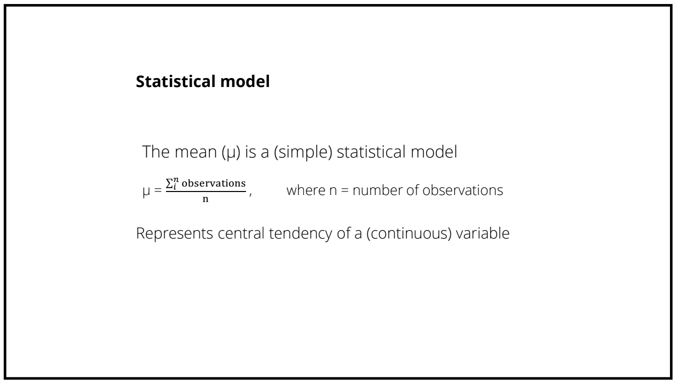
Statistical model

What is a statistical model?
Simple representation of reality

Height → Shoe size

Represents a relationship between variables
Does height have an effect on shoe size?

13



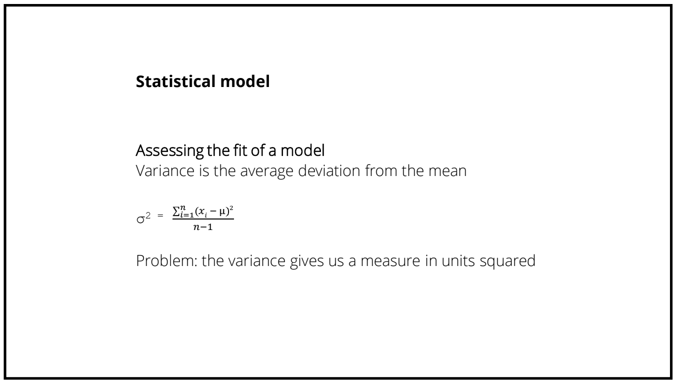
Statistical model

The mean (μ) is a (simple) statistical model

$\mu = \frac{\sum_i^n \text{observations}}{\text{n}}$, where n = number of observations

Represents central tendency of a (continuous) variable

14



Statistical model

Assessing the fit of a model
Variance is the average deviation from the mean

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n-1}$$

Problem: the variance gives us a measure in units squared

15

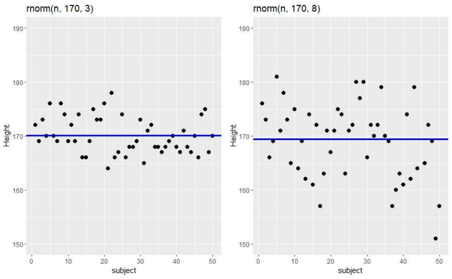
Statistical model

Standard deviation
Solution: we take the square root, this is called the standard deviation (s)

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

The smaller the deviance, the more accurate the mean represents the sample

16



17