Utrecht University

CENTRE FOR DIGITAL HUMANITIES

7 Oct 2021

Basics of Statistics

training for researchers and teachers in the Humanities

Hugo Quené and Kirsten Schutter
h.quene@uu.nl k.schutter@uu.nl
Centre for Digital Humanities, Utrecht University
Utrecht inst of Linguistics OTS, Utrecht University

1

Intro

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

2

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

For more on the empirical cycle see https://hugoquene.github.io/QMS-EN/

4

Variables

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

Subject	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

5

Variables

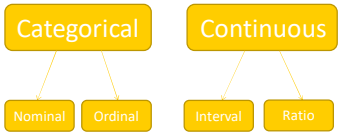
Dependent variable (DV)
Outcome variable
Y

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)

6

Level of measurement



7

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

8

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

9

Level of measurement

Dichotomous variables

A variable with only two categories
Also known as a binomial variable
Yes / no
Success / failure
Can be treated as continuous

10

Statistical model

What is a statistical model?
Simple representation of reality

Prediction

For example, the mean is a simple model

11

Statistical model

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

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

Represents central tendency of a (continuous) variable

12

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

13

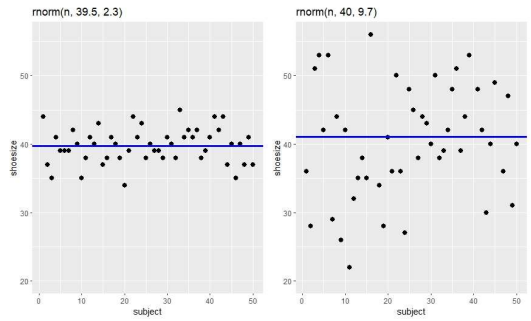
Statistical model

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

14



15

Statistical model

A relationship between variables

Does height have an effect on shoe size?

16