Utrecht University

CENTRE FOR DIGITAL HUMANITIES 3 March 2022

Basics of Statistics Session One (1)

training for researchers and teachers in the Humanities

Hugo Quené

h.quene@uu.nl www.hugoquene.nl

Centre for Digital Humanities, Utrecht University
Utrecht inst of Linguistics OTS, Utrecht University

1

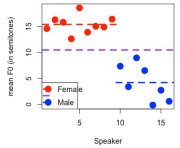
Introduction

teacher: Hugo Quené www.hugoquene.nl

- background in **speech** research
- speech is highly variable, hence **statistics**

today's tutorial

- quantitative (vs. qualitative)
- parametric (vs. nonparametric)
- frequentist (vs. Bayesian)



2

who are you? what do you want?

- on your mobile phone, go to

  https://www.wooclap.com/**QONDAR**

3

Principle 1
Data are sampled

- observed data are only a **sample** of larger population
  - population may be infinite and unknown (trees, humans, texts, sentences, responses)
- sample is ideally **random**, but may be **biased**:
  - e.g. selection bias, response bias …
- we try to find pattern in imperfectly sampled data, allowing for **uncertainty** from sampling

4

4

Principle 2
Observed data *vary*, randomly and systematically

variable: sth capable of varying

- **systematically** ("signal")
  observed effect, or pattern, often obscured
- **randomly** ("noise")
  due to sample variability, and measurement error, and unknown sources of variation
- pooled effects of random variation typically result in "normal" or "gaussian" distribution of random error
- errors tend to **cancel out** each other (on average)
  large sample: errors "disappear", patterns aggregate!

Why statistical analysis?

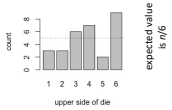
aims to discover **pattern** in data,
to discern meaningful **signal** from noise,
to **learn** from data,
to **make sense** of data

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

Example 1: fair die

```
> table(x)
x
1 2 3 4 5 6
3 3 6 7 2 9
```

- die is cube, six sides, each with probability of 1/6
- outcome is **discrete** or **categorical** variable
- outcomes of ***n*=30** throws:
  3 3 4 2 4 1 6 2 6 6 6 6 4 5 6 4 6 6 4 3 3 1 6 4 3 1 5 3 2 4
- left: frequencies (counts) in table form
- right: frequencies (counts) in "bar chart" figure form
  - **categorical**: spaces between discrete bars
- sampling variability: expected vs observed pattern

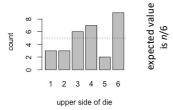


Example 1: fair die (continued)

```
> table(x)
x
1 2 3 4 5 6
3 3 6 7 2 9
```

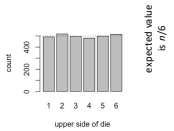
- categorical variable
- center:
  median (50% percentile) 4
  - between 15th and 16th ranked observation

  mode (most frequent value) 6
- dispersion:
  median absolute deviation (mad) 2.2



Example 1: fair die (continued)

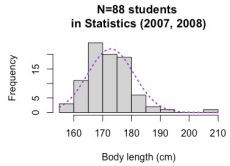
- as sample size *n* increases:
  clearer pattern, less noise
- **because independent sampling errors tend to cancel out each other**



9

9

Example 2: students' body length



- from N=88 students from two cohorts
- right: frequencies (counts) in histogram bins
  - **continuous**: no spaces between bars
- sampling variability: expected vs observed pattern

- centre: mean 173 median 172
- dispersion: std.dev. 8.0 mad 7.4

between 44th and 45th ranked observation

21 July 2021 10

10

Know your variables

- independent — grouping, factor, predictor
- dependent — outcome (depends on sample)

- categorical — e.g. die, gender
- continuous — e.g. body length, shoe size

- examples… — last vote (party), boosted, self-test outcome, T-shirt size, **age**…

predictor, or outcome ?
categorical, or continuous ?

12

Know your "levels of measurement"

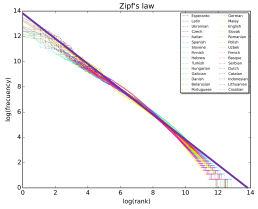
	name	properties	operations	example
categorical	nominal	no natural order	count	ice cream flavour
	ordinal	with natural order, no distances	count, order	education level, die
continuous	interval	equal intervals, no zero	add, subtract	temp'ture celsius
	ratio	equal intervals, with zero	multiply, divide	body length

13

Statistical model

model: Zipf's Law (straight purple line)
error: deviations from predicted values

- simplified version of reality (or rather: of data taken from reality)
- data = model + error
- simplest model: the **mean** (average)
  simplest error: **standard deviation**
- assuming interval or ratio level, assuming approximately normal distribution, assuming independent observations, …



https://upload.wikimedia.org/wikipedia/commons/a/ac/Zipf_30wiki_en_labels.png

14

variance and standard deviation

$$s = \sqrt{s^2} = \sqrt{\frac{\sum(x_i - \overline{x})^2}{n-1}}$$

- s^2 variance (in squared units)
  s standard deviation (sd, in orig units)

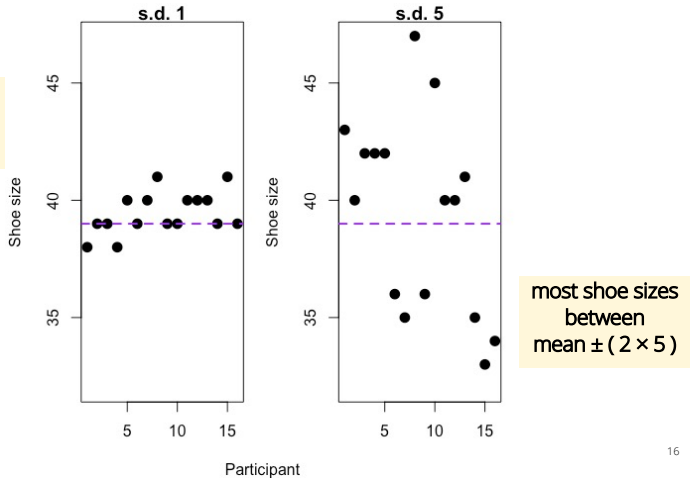
x: { 1, 2, 3 }
n=3, n-1=2
mean: (1+2+3)/n = 2
deviations: { -1, 0, +1 }
$(dev)^2$: { 1, 0, 1 }
SS dev: 1+0+1 = 2
variance = 2/2 = 1
std.dev. = 1

https://hugoquene.github.io/QMS-EN/ch-centre-and-dispersion.html

15

15

most shoe sizes between mean ± (2 × 1)



most shoe sizes between mean ± (2 × 5)

16

16

Questions ?

17

17