Utrecht University CENTRE FOR DIGITAL HUMANITIES 7 Oct 2021

Basics of Statistics Session Three

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

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)

2

(hypotheses about) relations between variables

- **H0: data** = constant+**error** — **no** pattern (no effect of IV on DV)
  H1: data = constant+**pattern+error** — some effect
- falsification principle (Popper):
  reject H0 if data provide **significant** evidence against H0
  i.e., if P(data|H0) is very low (we know what data to expect if H0 were true)
- decision is based on imperfect (sampled) data, containing errors,
  hence decision may be incorrect!

3

3

	PCR test result: neg	PCR test result: pos	
healthy	true neg	**false pos,** *quarantaine*	specificity estim 98%
COVID19	**false neg,** *infectuous*	true pos	sensitivity, recall, **88%** (N=3818 pat.)
prevalence, unknown	NPV estim 96%	precision, PPV estim 94%	

Jarrom D, Elston L, Washington J, et al
Effectiveness of tests to detect the presence of SARS-CoV-2 virus, and antibodies to SARS-CoV-2, to inform COVID-19 diagnosis: a rapid systematic review
BMJ Evidence-Based Medicine Published Online First: 01 October 2020. doi: 10.1136/bmjebm-2020-111511

4

	test result: neg keep H0	test result: pos reject H0	
H0 true	true neg	***false pos, spurious*** *(Type I error)*	
H0 false	***false neg, miss*** *(Type II error)*	true pos	
prevalence, *unknown proportion of false H0's*			

5

P for significance

Note: significant effect has low P

- significance = risk of Type I error (false positive outcome)
- P(data|H0)
  - frequentist: in large number of repeated samples
- not P(H1|data)
- not 1-P(H0|data)
- significance = **effect size × size of study**
  (Rosnow & Rosenthal, 2008)

6

6

ES for effect size

- amount of difference
  standardized to amount of dispersion
- many different measures of effect size
- e.g. $d = (M_1 - M_0) / s$
- similar to Z score: difference divided by pooled sd
- not sensitive to N
  - while significance is sensitive to N
- example: gender effect in adult voice pitch,
  d = 9 semitones / 5 semitones = 1.7

7

7

power

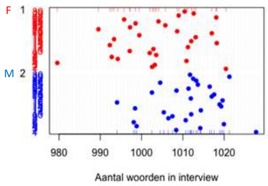
- 1 – P(Type II error)
- P (reject H0 | H0 false)
- H0 is false and H0 is rejected: right!
- should be determined a priori
- depends ...
  on effect size,
  on sample size N,
  on chosen level of significance

power increases with...
larger effect
larger sample
larger P(Type I error)

8

8

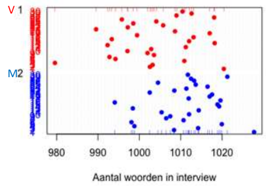
H0: women and men are equally talkative



H1: F ≠ M
H0: F = M

how certain or justified
is the decision
to reject H0 (i.e. to claim
that there is a difference)?

9



if we would repeat the study many times, then in 95% of replications the 95%CI would contain the true difference

mean(F)=1004
mean(M)=1011
difference: 7 words (<1%)

n=30 per group (N=60)
s (pooled) = 9
ES = 7 / 9 ≈ 0.8

$t(58) = 3.2$

$p = .002$

95% CI: (3, 12)

Reject H0

t test value combines differences between groups and s.d. within groups

p is probability of finding this *t* test value, or larger, if H0 is true

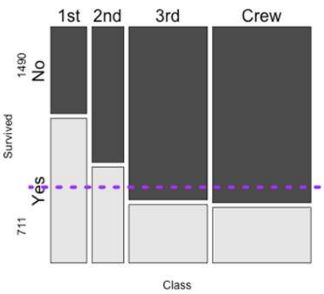
10

Example: surviving the *Titanic* disaster (1912)

random variation in individual outcomes, due to many circumstances

H0: no systematic effect of class on survival (no pattern)

chi² test: reject H0
1st and 2nd class had better odds of surviving than 3rd class and crew
[$\chi^2(3) = 190$, $p < .001$, Cramer's V = 0.29]



11

11

Statistical tests may produce misleading results

10.1126/science.aac4716

- replication crisis:
  n=100 replications of high-impact psych studies,
  - only 39 replications show similar effect
  - effect size about half of original study
- problems due to
  insufficient power (probability of rejecting H0)
  - due to small effect size and/or small sample size
- and due to base rate fallacy (cf breast cancer analogy):
  low prevalence of true hypotheses

10.1371/journal.pmed.0020124

12

12

Intermezzo: know your symbols

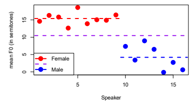
symbol	*English*	*Dutch*
P, p	probability	kans, waarschijnlijkheid
N, n	number	aantal
S, s	spread	spreiding (st.dev)
D, d	difference	verschil
M, m	mean	gemiddelde
R, r	cor-relation	cor-relatie
L	likelihood	?

Roman symbols
for known properties
of **sample**
(M, s)

Greek symbols
for unknown properties
of **population**
(μ, σ)

13

data = model + error



isMale (dummy):
0 for Female,
1 for Male

error: defined as
s.d. of difference
from dashed line
(ffrom prediction)

- f0: voice pitch, in semitones relative to 110 Hz (piano keys re A2)
- M0: f0 = b_0 + error (baseline model, purple)
  b_0 = 10.5 ST RMSE = 6.1
  predicted pitch: 10.5 ST for all speakers
- M1: F0 = b_0 + *isMale**b_1 + error
  b_0 = 15.4 ST
  b_1 = -11.2 ST RMSE = 2.5
  predicted pitch: for females 15.4 ST, for males 4.2 ST
- M1 has lower error, better fit to data (p<.0001), prefer M1

14

14

data = model + error

constant:
value 1 for all obs

isMale (dummy):
0 for female,
1 for male

- M0: height = (*constant**β_0) + error (null model)
  β_0 = 177.6 cm RMSE = 12.8
- predicted height: 178 cm for all students
- M1: height = (*constant**β_0 + *isMale**β_1) + error
  β_0 = 166.6 cm
  β_1 = 22.1 cm RMSE = 6.0

- M1 has lower error, better fit to data (p<.001), prefer M1
- predicted height: for females 167 cm, for males 189 cm

21 July 2021 15

15

data = model + error

- also applies to...
  χ^2 test, *t* test, ANOVA (for categorical predictor/s),
  regression, GLM (for continuous predictor/s)

- BUT only under several assumptions and conditions

16

16

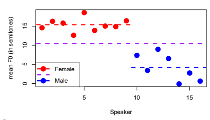
key assumptions

- independence:
  each observation is independently drawn from population
  - otherwise: use hierarchical models
- robustness:
  model has only few parameters (e.g. *N*/20)
  - otherwise: overspecification, poor generalizability
- multicollinearity:
  predictors should not be mutually correlated

17

17

comparing two models



the chance under H0 that M1 fits data better than M0 does, just by **accident**, is p<.0001

M1 is significantly better (p<.0001)

based on **residuals** (errors) of two models
M0: RMSE 6.1 (SD relative to purple line)
M1: RMSE 2.5 (SD relative to red/blue lines)
$F(1,14) = 71.7$, p<.0001
- probability of this reduction of resid, or larger reduction, if H0 is true

M1 has smaller residuals, prefer M1

if there is really **no** effect of gender on voice pitch,
and if we **repeat** the same study (resampling speakers from the same population) 10000 times,
then one of the replications will **accidentally** show a gender effect of this size or larger

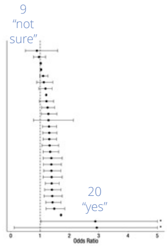
18

18

19

One data set, many analyses, different outcomes

- RQ: "whether soccer referees are more likely to give red cards to dark-skin-toned players than to light-skin-toned players."
- one data set, 29 teams, 61 data analysts
- "Uncertainty in interpreting research results is … a function of the many reasonable decisions that researchers must make in order to conduct the research. (…) [M]any subjective decisions are part of the research process and can affect the outcomes. "



doi:10.1177/2515245917747646

20

Skilled interpretation is required

- how was sample drawn? possible biases?
- which "noisy" variables have been considered? how?
  e.g. player position, league, previous encounters…
- was analysis appropriate and adequate for these data? for this RQ? for this design of study?
  https://www.hugoquene.nl/qm/CheatSheetQuantRes.pdf
- how robust is analysis? how generalizable are results?

21

questions?

- questions?
- next: hands-on practical session
- build and explore your own statistical models!

22