Reporting and Assessing Statistical Evidence 2

PSM 2

Bennett Kleinberg

26 Feb 2019

Probability, Statistics & Modeling II

Lecture 7

Statistical reporting 2 + assessment details

Which question do you have?

Today

- recap week 1-5
- statistical reporting
- assessment details

Week 1

- marginal probability
- conditional probability
- Bayes' Rule
- Simpson's paradox

Week 2

- modelling data
- regression idea
- regression effects

Week 3

- GLM idea
- logistic regression
- model comparison

Week 4

- GLM for t-tests
- GLM for ANOVAs
- Relationship between t-test, ANOVA, lm

Week 5

- non-paramtric tests
- esp. ranking methods
- discrete data
- R by C (ChiSquare)
- X by Y by Z (loglinear)

Week 5

Statistical reporting

A tricky time

- Reproducibility crisis (2015)
- Statistics crisis (2016)

A tricky time

- QRPs widespread
- Some remedies in place
  - preregistration
  - data sharing

A tricky time

Problems:

- preregistration a long way from the norm
- data sharing is often a taboo

So what can we do without relying on (old) researchers improving their research practices?

Ways forward

- get the journals on board
- get the funding bodies on board
- tackle the NHST problem

My first “science” surprise

The Mind of a Con Man

New York Times article

Three of Jens Förster's papers have been retracted; more retractions may follow. HUMBOLDT-STIFTUNG/SVEN-MÜLLER

No tenure for German social psychologist accused of data manipulation

source, see also: Retraction Watch article

The NHST problem

Null hypothesis significance testing highly debated

- issue of “researcher’s degrees of freedom”
- analytical techniques
- exclusions
- no. of variables
- …

Leads to high prevalence of QRPs

Open Science to the rescue?

Remember:

- preregistration not common
- replication not common
- data sharing not common

Hypothetical scenario

Even if these practices become the norm …

… we're still stuck with arbitrary p-value thresholds!

And anyway:

How are we supposed to report anything now?

3 approaches to the p-value issue

1. Apply a new common threshold
2. Always justify your threshold
3. A completely different statistics framework

A new alpha threshold

However, we believe that a leading cause of non-reproducibility has not yet been adequately addressed: statistical standards of evidence for claiming new discoveries in many fields of science are simply too low.

Benjamin et al., 2018

A new alpha threshold

Redefine statistical significance

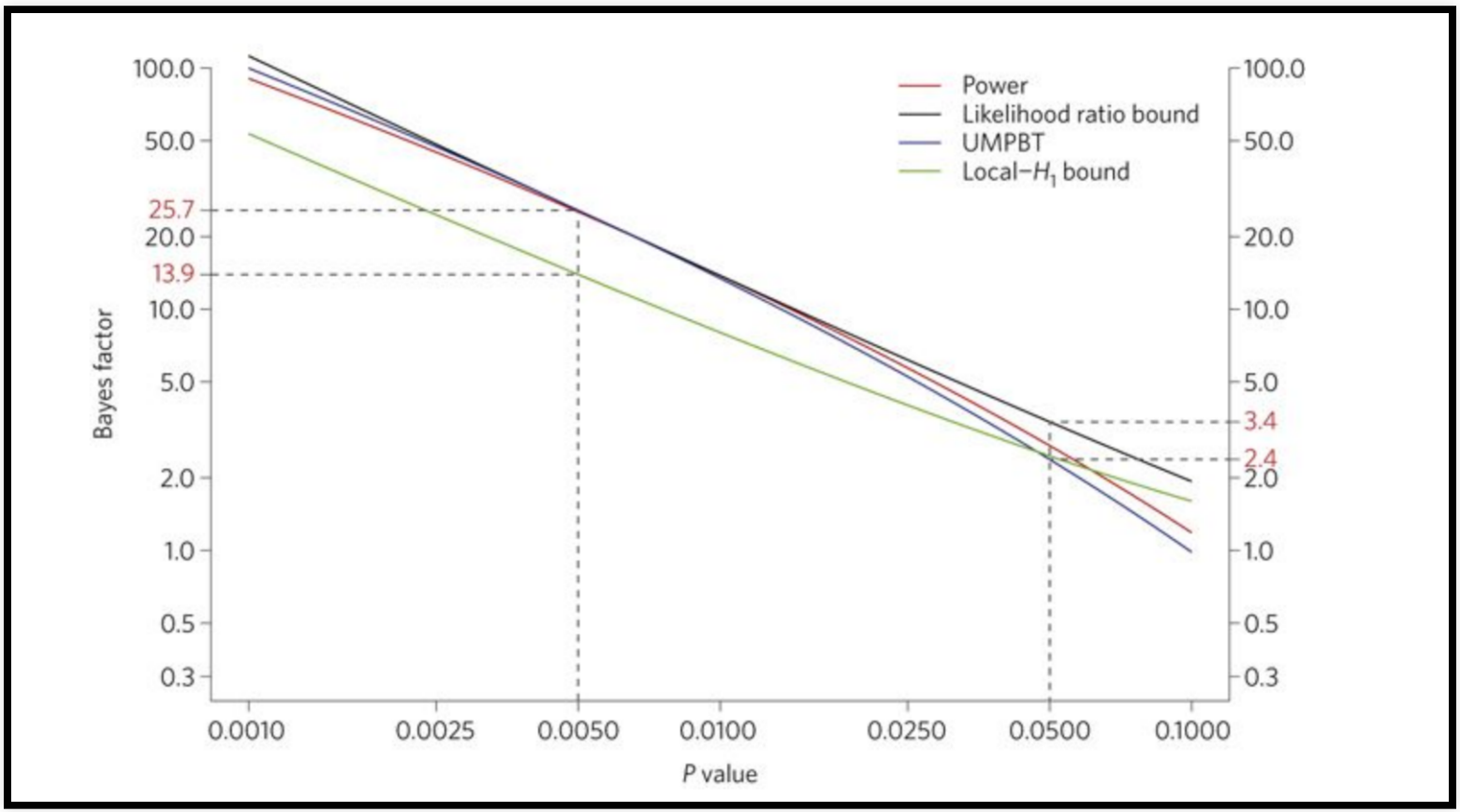
Daniel J. Benjamin ✉, James O. Berger, […] Valen E. Johnson ✉

Nature Human Behaviour **2**, 6–10 (2018) | Download Citation ↓

We propose to change the default *P*-value threshold for statistical significance from 0.05 to 0.005 for claims of new discoveries.

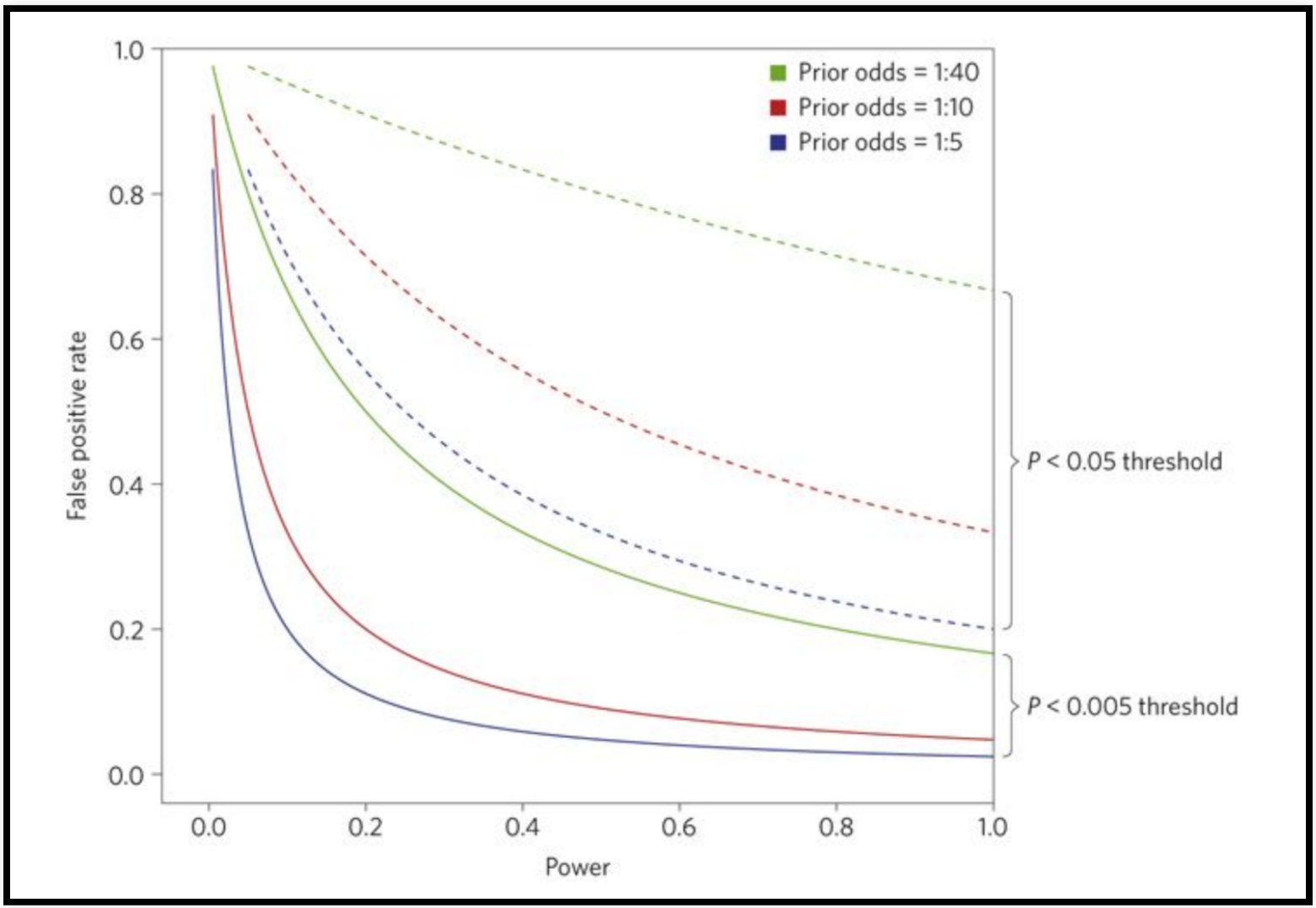
Benjamin et al., 2018

A new alpha threshold



Benjamin et al., 2018

A new alpha threshold



Benjamin et al., 2018

A new alpha threshold

For research communities that continue to rely on null hypothesis significance testing, reducing the P value threshold for claims of new discoveries to 0.005 is an actionable step that will immediately improve reproducibility.

Benjamin et al., 2018

Justify your alpha

Justify your alpha

Daniel Lakens ✉, Federico G. Adolfi, […] Rolf A. Zwaan

Nature Human Behaviour **2**, 168–171 (2018) | Download Citation ⬇

In response to recommendations to redefine statistical significance to $P \leq 0.005$, we propose that researchers should transparently report and justify all choices they make when designing a study, including the alpha level.

Justify your alpha

- “Weak justifications for the α = .005 threshold”
- “How a threshold of p ≤.005 might harm scientific practice”
- “No one alpha to rule them all”

Justify your alpha

We call for a broader mandate beyond p-value thresholds whereby all justifications of key choices in research design and statistical practice are transparently evaluated, fully accessible, and preregistered whenever feasible.

Lakens et al., 2018

Morale of the story

- single studies overrated
- challenge of single summary statistic remains
- more meaningful
  - replications
  - “many labs” projects
  - RRRs
  - “many analysts” projects

About many analyst projects...

Silberzahn et al. (2018):

> *"The primary research question tested [...] was whether soccer players with dark skin tone are more likely than those with light skin tone to receive red cards from referees"*

- 146,028 dyads of players and referees

Many analysts

- 29 teams, 61 researchers
- Whole range of academic experience

What do you think?

Give it a try: https://osf.io/gvm2z/

Many analysts

- 20 teams found sign. positive relationship
- 9 teams found no relationship
- 29 different analyses
- 21 unique combinations of covariates (e.g. position, country, …)

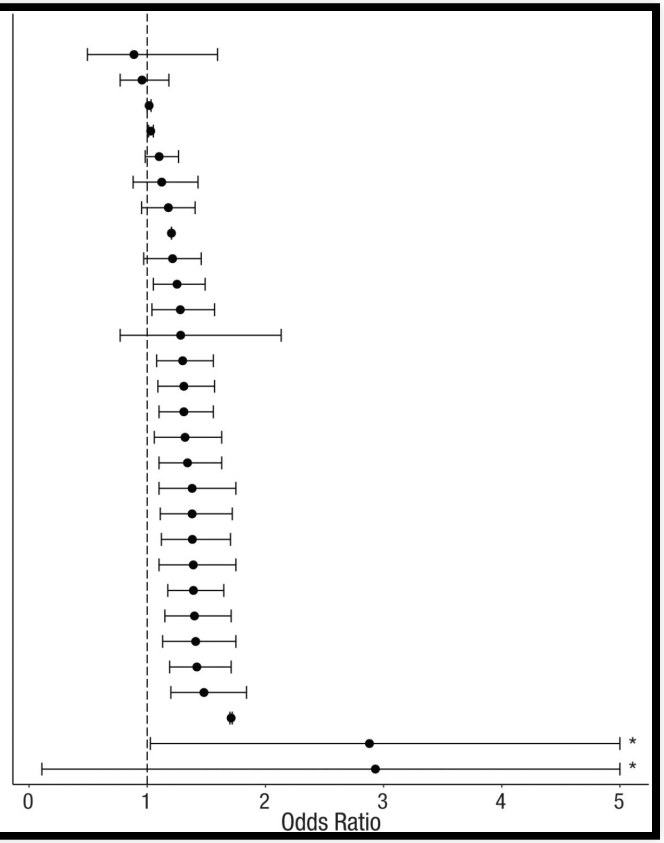
Many analysts

Table 3. Analytic Approaches and Results for Each Team

Team	Distribution	Treatment of nonindependence	Number of covariates	Analytic approach	OR
1	Linear	Clustered standard errors	7	Ordinary least squares regression with robust standard errors, logistic regression	1.18 [0.95, 1.41]
6	Linear	Clustered standard errors	6	Linear probability model	1.28 [0.77, 2.13]
14	Linear	Clustered standard errors	6	Weighted least squares regression with clustered standard errors	1.21 [0.97, 1.46]
4	Linear	None	3	Spearman correlation	1.21 [1.20, 1.21]
11	Linear	None	4	Multiple linear regression	1.25 [1.05, 1.49]
10	Linear	Variance component	3	Multilevel regression and logistic regression	1.03 [1.01, 1.05]
2	Logistic	Clustered standard errors	6	Linear probability model, logistic regression	1.34 [1.10, 1.63]
30	Logistic	Clustered standard errors	3	Clustered robust binomial logistic regression	1.28 [1.04, 1.57]
31	Logistic	Clustered standard errors	6	Logistic regression	1.12 [0.88, 1.43]
32	Logistic	Clustered standard errors	1	Generalized linear models for binary data	1.39 [1.10, 1.75]
8	Logistic	None	0	Negative binomial regression with a log link	1.39 [1.17, 1.65]
15	Logistic	None	1	Hierarchical log-linear modeling	1.02 [1.00, 1.03]
3	Logistic	Variance component	2	Multilevel logistic regression using Bayesian inference	1.31 [1.09, 1.57]
5	Logistic	Variance component	0	Generalized linear mixed models	1.38 [1.10, 1.75]
9	Logistic	Variance component	2	Generalized linear mixed-effects models with a logit link	1.48 [1.20, 1.84]
17	Logistic	Variance component	2	Bayesian logistic regression	0.96 [0.77, 1.18]
18	Logistic	Variance component	2	Hierarchical Bayes model	1.10 [0.98, 1.27]
23	Logistic	Variance component	2	Mixed-model logistic regression	1.31 [1.10, 1.56]
24	Logistic	Variance component	3	Multilevel logistic regression	1.38 [1.11, 1.72]
25	Logistic	Variance component	4	Multilevel logistic binomial regression	1.42 [1.19, 1.71]
28	Logistic	Variance component	2	Mixed-effects logistic regression	1.38 [1.12, 1.71]
21	Miscellaneous	Clustered standard errors	3	Tobit regression	2.88 [1.03, 11.47]
7	Miscellaneous	None	0	Dirichlet process Bayesian clustering	1.71 [1.70, 1.72]

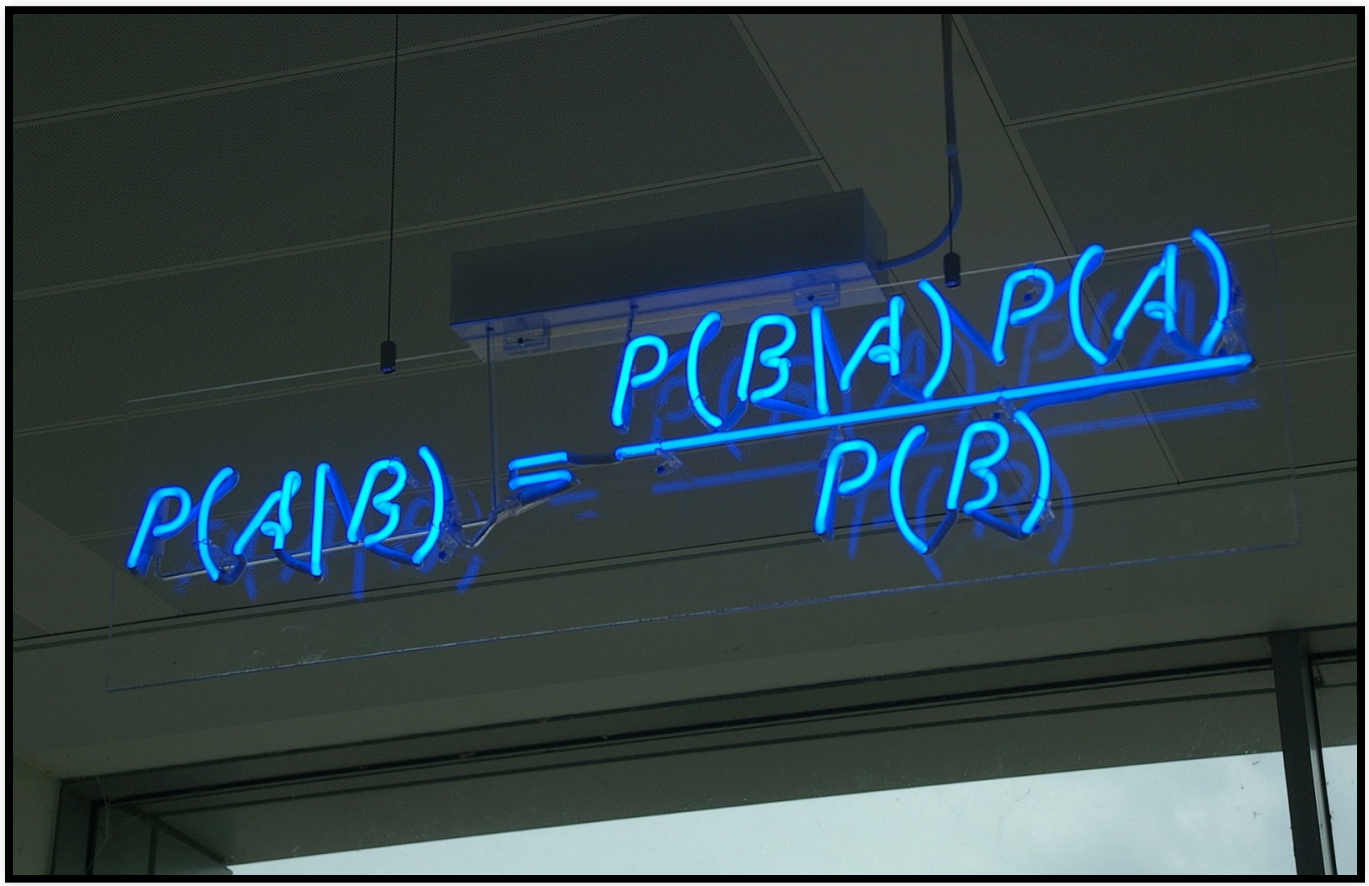
Many analysts

Team	Analytic Approach	Odds Ratio
12	Zero-Inflated Poisson Regression	0.89
17	Bayesian Logistic Regression	0.96
15	Hierarchical Log-Linear Modeling	1.02
10	Multilevel Regression and Logistic Regression	1.03
18	Hierarchical Bayes Model	1.10
31	Logistic Regression	1.12
1	OLS Regression With Robust Standard Errors, Logistic Regression	1.18
4	Spearman Correlation	1.21
14	WLS Regression With Clustered Standard Errors	1.21
11	Multiple Linear Regression	1.25
30	Clustered Robust Binomial Logistic Regression	1.28
6	Linear Probability Model	1.28
26	Hierarchical Generalized Linear Modeling With Poisson Sampling	1.30
3	Multilevel Logistic Regression Using Bayesian Inference	1.31
23	Mixed-Model Logistic Regression	1.31
16	Hierarchical Poisson Regression	1.32
2	Linear Probability Model, Logistic Regression	1.34
5	Generalized Linear Mixed Models	1.38
24	Multilevel Logistic Regression	1.38
28	Mixed-Effects Logistic Regression	1.38
32	Generalized Linear Models for Binary Data	1.39
8	Negative Binomial Regression With a Log Link	1.39
20	Cross-Classified Multilevel Negative Binomial Model	1.40
13	Poisson Multilevel Modeling	1.41
25	Multilevel Logistic Binomial Regression	1.42
9	Generalized Linear Mixed-Effects Models With a Logit Link	1.48
7	Dirichlet-Process Bayesian Clustering	1.71
21	Tobit Regression	2.88
27	Poisson Regression	2.93



“The best defense against subjectivity in science is to expose it.”

A completely different statistics framework



Next week

For you?

“The best defense against subjectivity in science is to expose it.”

Make yourself and the findings the least vulnerable.

The Applied Analysis Project

General info

- 50% of final grade
- preregistration + independent analysis
- demonstrate your understanding of statistical techniques
- conduct analyes in R

Details

- we provide a dataset
- your task is to act like an analyst
- answer 5 specific questions
- plus 1 question of your own

Dataset details

- data on offenders charged with gang-related crimes
- variable codebook online and on Moodle

Dataset release

- Phase 1: pseudo-dataset
  - identical in structure
  - different in values and length
  - serves as input for your preregistration
  - dataset available **now** on Moodle
- Phase 2: full release
  - individualised for each student
  - released via email on 30 March 2019

Preregistration

- Word limit: 300 words
- detail the hypotheses
  - for the 5 given questions
  - plus for your own question
- specify all exclusions, transformations, selection criteria, etc.
- use the template online on Moodle (based on the OSF template)
- deadline for preregistration: **29 March 2019**
  - submission via TII on Moodle

Full analysis

- datasets released on **30 March 2019**
- base your analysis on the preregistration
- specify and justify all deviations from the prereg.

Full analysis report

- no introduction section
- Advised structure:
  - hypotheses
  - method
  - data
  - analytical plan
  - results
  - discussion

Full analysis report

- Word limit: 1,500 words
- Submit as anonymised PDF on TII on Moodle
- Additionally: submit your code supplement
  - R Notebook
  - fully reproducible code
  - submit as an anonymised html on Moodle
- deadline: **16 April 2019**

Feedback session

- 1-on-1 feedback on 26 March
- 10 min slot for each student
- Template for feedback submission online and on Moodle

All dates/deadlines

- dataset release 1: *now*
- deadline feedback submission tenplate: *22 March* on TII
- deadline preregistration: *29 March* on TII
- dataset release 2: *30 March*
- deadline final report + code: *16 April*

Questions about the assessment?

Outlook

Next week: Bayesian hypothesis testing

Homework: Revise week 1-5

END