Introduction: p-values, Bayes factors and e-values

Workshop Safe Statistics

Rianne de Heide, July 18, 2023

Menu

- p-values and why do we need a new theory for hypothesis testing?
- Are Bayes factors the solution?
- E-values

P-values and why do we need a new theory for hypothesis testing?

P-values

- History: Karl Pearson (1900) and Ronald Fisher (1925)

Why do we need a new theory for hypothesis testing?

- 100 years later: replicability crisis in social and medical science

- Medicine: J. Ioannidis, Why most published research findings are false , PLoS Medicine 2(8) (2005).

- Social Science: 270 authors, Estimating the reproducibility of psychological science, Science 349 (6251), 2015.

Why do we need a new theory for hypothesis testing?

Reproducibility crisis in social and medical science

Causes:

Why do we need a new theory for hypothesis testing?

Reproducibility crisis in social and medical science

Causes:

- publication bias

Why do we need a new theory for hypothesis testing?

Reproducibility crisis in social and medical science

Causes:

- publication bias
- fraud

Why do we need a new theory for hypothesis testing?

Reproducibility crisis in social and medical science

Causes:

- publication bias
- fraud
- lab environment vs. natural environment

Why do we need a new theory for hypothesis testing?

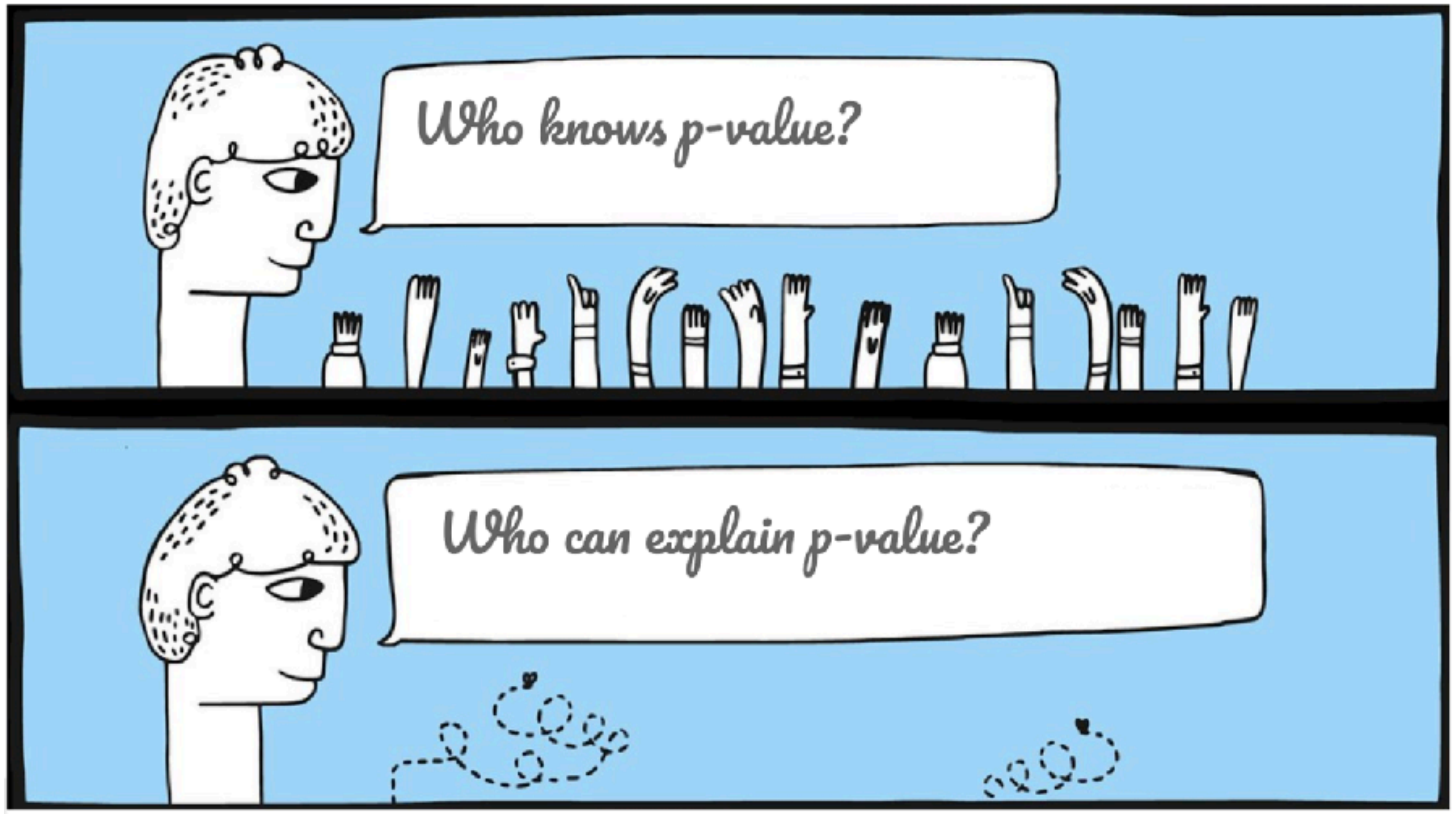
Reproducibility crisis in social and medical science

Causes:

- publication bias
- fraud
- lab environment vs. natural environment
- use of p-values

What is a p-value actually?

We wish to test a null hypothesis $\mathscr{H}_0$, often in contrast with an alternative hypothesis $\mathscr{H}_1$.



What is a p-value actually?

We wish to test a null hypothesis $\mathscr{H}_0$, often in contrast with an alternative hypothesis $\mathscr{H}_1$.

P-value:

- "Probability under the null hypothesis of obtaining a real-valued test statistic at least as extreme as the one obtained"
- "The *P*-value is the smallest level of significance that would lead to rejection of the null hypothesis *H0* with the given data."
- "P-value is the level of marginal significance within a statistical hypothesis test, representing the probability of the occurrence of a given event."
- "A p-value, or probability value, is a number describing how likely it is that your data would have occurred by random chance."

What do doctors know about statistics?

A controlled trial of a new treatment led to the conclusion that it is significantly better than placebo: $p < 0.05$. Which of the following statements do you prefer? **menti.com 5150 7926**

A. It has been proved that the treatment is better than placebo.

B. If the treatment is not effective, there is less than 5 percent chance of obtaining such results.

C. The observed effect of the treatment is so large that there is less than 5 percent chance that the treatment is no better than placebo.

D. I do not really know what a p-value is and do not want to guess.

What do doctors know about statistics?

A controlled trial of a new treatment led to the conclusion that it is significantly better than placebo: $p < 0.05$. Which of the following statements do you prefer?

A. It has been proved that the treatment is better than placebo. 20%

B. If the treatment is not effective, there is less than 5 percent chance of obtaining such results. 13%

C. The observed effect of the treatment is so large that there is less than 5 percent chance that the treatment is no better than placebo. 51%

D. I do not really know what a p-value is and do not want to guess. 16%

Stopping rules and p-values

- Suppose you are doing a trial on 70 subjects. The p-value is promising but just not significant ($p = 0.06$). Your boss says there is some more money for adding 10 more subjects to the the trial. What do you do?

Stopping rules and p-values

- Suppose you are doing a trial on 70 subjects. The p-value is promising but just not significant ($p = 0.06$). Your boss says there is some more money for adding 10 more subjects to the the trial. What do you do?

- John et al (2012): 58% of psychologists admits to “Deciding whether to collect more data after looking to see whether the results were significant”.

Stopping rules and p-values

- Suppose you are doing a trial on 70 subjects. The p-value is promising but just not significant ($p = 0.06$). Your boss says there is some more money for adding 10 more subjects to the the trial. What do you do?

- John et al (2012): 58% of psychologists admits to “Deciding whether to collect more data after looking to see whether the results were significant”.

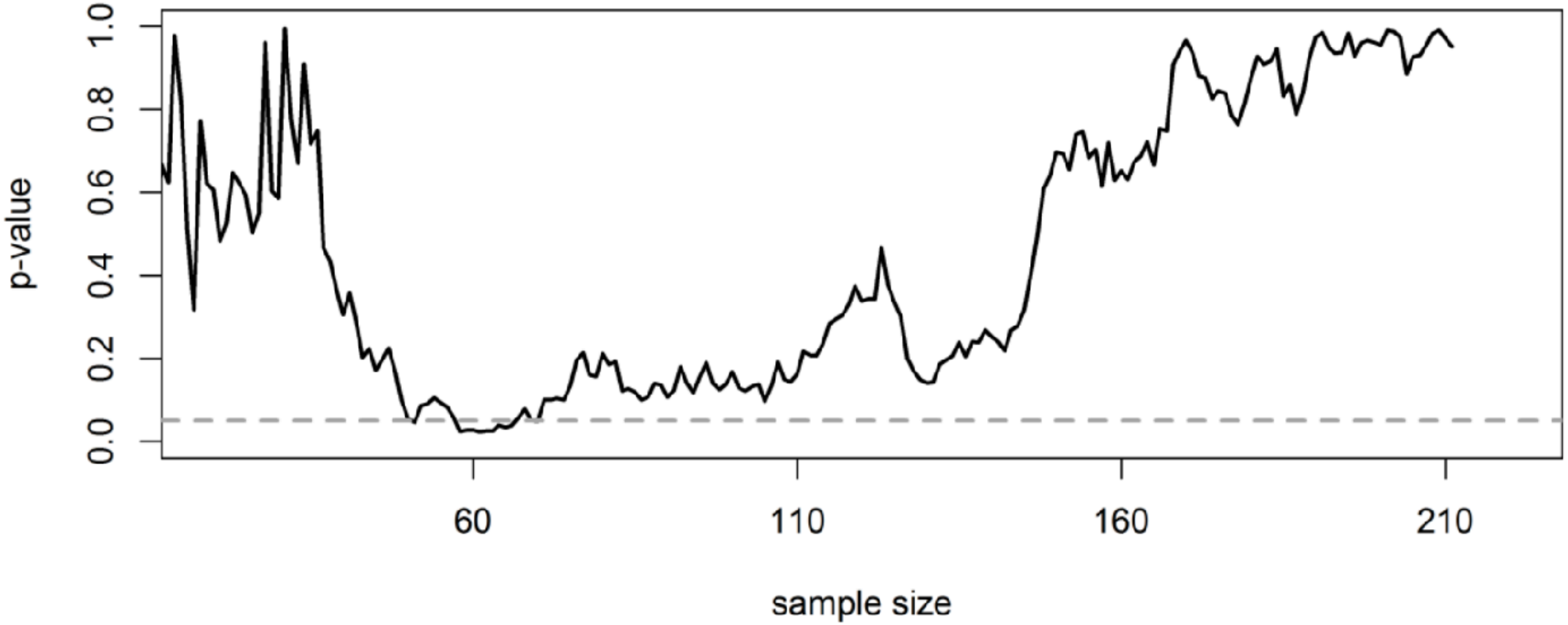
- This is called optional stopping, and invalidates p-values and their error guarantees (more about peeking in Peters talk)

Type I error guarantee

Fix $\alpha \in (0,1)$, then

$$\mathbb{P}(\text{reject } \mathscr{H}_0) \leq \alpha$$

Stopping rules and p-values



$$\mathbb{P}(\exists t \in \mathbb{N} : p_t < \alpha) = 1$$

Other disadvantages with p-values

- Combining evidence from different (possibly dependent) studies

  Hospitals A and B perform similar trials, and they report p-values p_A and p_B. How to combine the evidence?

  A meta-analysis is done. However, the subsequent studies were only done because the previous studies were promising, so there is a complicated (and unknown) dependency. How to combine the evidence?

Other disadvantages with p-values

- Combining evidence from different (possibly dependent) studies (e.g. two different populations; meta-analysis)

- Limited applicability: unknown probabilities (counterfactuals)

  Consider two weather forecasters A and B. On sunny days, P_A(RAIN) $\geq P_B$(RAIN), and on rainy days their accuracy is approximately the same. Is B better than A? We can’t do this with p-values

Other disadvantages with p-values

- Combining evidence from different (possibly dependent) studies (e.g. two different populations; meta-analysis)

- Limited applicability: unknown probabilities (counterfactuals)

  Consider two weather forecasters A and B. On sunny days, P_A(RAIN) $\geq P_B$(RAIN). Is B better than A?

- Interpretational problems: misunderstanding (hence misuse) of p-values

Are Bayes factors the solution?

The Bayes factor

- Prior odds $\dfrac{P(H_1)}{P(H_0)}$

The Bayes factor

- Prior odds $\dfrac{P(H_1)}{P(H_0)}$

- Bayes marginal / model evidence / evidence $P(X_1, \ldots, X_n \mid H_i), i = 0,1$

The Bayes factor

- Prior odds $\dfrac{P(H_1)}{P(H_0)}$
- Bayes marginal / model evidence / evidence $P(X_1, \ldots, X_n \,|\, H_i), i = 0,1$
- Posterior odds $\dfrac{P(H_1 \,|\, X_1, \ldots, X_n)}{P(H_0 \,|\, X_1, \ldots, X_n)} = \dfrac{P(H_1)}{P(H_0)} \dfrac{P(X_1, \ldots, X_n \,|\, H_1)}{P(X_1, \ldots, X_n \,|\, H_0)}$

The Bayes factor

- Prior odds $\dfrac{P(H_1)}{P(H_0)}$
- Bayes marginal / model evidence / evidence $P(X_1, \ldots, X_n \mid H_i), i = 0{,}1$
- Posterior odds $\dfrac{P(H_1 \mid X_1, \ldots, X_n)}{P(H_0 \mid X_1, \ldots, X_n)} = \dfrac{P(H_1)}{P(H_0)} \dfrac{P(X_1, \ldots, X_n \mid H_1)}{P(X_1, \ldots, X_n \mid H_0)}$
- Bayes factor: $= \dfrac{p(X_1, \ldots, X_n \mid H_1)}{p(X_1, \ldots, X_n \mid H_0)} = \dfrac{\int_{\Theta_1} p_{\theta_1}(X_1, \ldots, X_n \mid \theta_1) w(\theta_1) d\theta_1}{\int_{\Theta_0} p_{\theta_0}(X_1, \ldots, X_n \mid \theta_0) w(\theta_0) d\theta_0}$

Bayes factors and optional stopping

- When H_0 is simple, we have the bound

  $P(\exists t \in \mathbb{N}, \text{BF} > 1/\alpha) \leq \alpha$

Bayes factors and optional stopping

- When H_0 is simple, we have the bound

  $$P(\exists t \in \mathbb{N}, \text{BF} > 1/\alpha) \leq \alpha$$

- When H_0 is composite, this does not hold, i.e., the type I error guarantee is not preserved under optional stopping, just as with p-values (exception: group-invariant Bayes factors, s.a. the Bayesian t-test)

Bayes factors and optional stopping

- When H_0 is simple, we have the bound

  $$P(\exists t \in \mathbb{N}, \text{BF} > 1/\alpha) \leq \alpha$$

- When H_0 is composite, this does not hold, i.e., the type I error guarantee is not preserved under optional stopping, just as with p-values.

- Other notions of (Bayesian) optional stopping: see De Heide and Grünwald (2021) and Hendriksen, De Heide and Grünwald (2021).

E-values

A lady tasting tea

A lady tasting tea

Null hypothesis: the lady has no ability to distinguish the teas.

A lady tasting tea

Null hypothesis: the lady has no ability to distinguish the teas.

$$\binom{8}{4} = \frac{8!}{4!(8-4)!} = 70$$

Safe Testing

e-values in stead of p-values

- intuitive interpretation: betting
- sequential testing possible
- easy combination of several studies: by multiplication

Safe Testing - a lady tasting coffee

Safe Testing - a lady tasting coffee

Safe Testing - a lady tasting coffee

M C

Safe Testing - a lady tasting coffee

$B_1 = -1$

M C

Safe Testing - a lady tasting coffee

$B_1 = -1$

M C

Safe Testing - a lady tasting coffee

$B_1 = -1$

M C

Safe Testing - a lady tasting coffee

$B_1 = -1$

M C

M C

Safe Testing - a lady tasting coffee

$B_1 = -1$

M C

$B_2 = +1$

M C

A lady tasting coffee: guessing

$B_1 = -1$

M C

$B_2 = +1$

M C

A lady tasting coffee: guessing

$B_1 = -1$

M C

$B_2 = +1$

M C

$S_t = \exp\{u \sum_{s=1}^{t} B_s - vt\}$ is an e-value for certain choices of u, v > 0

A lady tasting coffee: guessing

$B_1 = -1$

M C

$B_2 = +1$

M C

$S_t = \exp\{u \sum_{s=1}^{t} B_s - vt\}$ is an e-value for certain choices of u, v > 0

$\mathscr{H}_0$: There is no difference between MC and CM.

A lady tasting coffee: guessing

$B_1 = -1$

M C

$B_2 = +1$

M C

$S_t = \exp\{u \sum_{s=1}^{t} B_s - vt\}$ is an e-value for certain choices of u, v > 0 ,

$\mathscr{H}_0$: There is no difference between MC and CM.

If we reject when S_t is large, we preserve Type I error guarantees under optional stopping.

E-value

- Simplified version (for fixed n): non-negative random variable E satisfying

$$\text{for all } P \in \mathcal{H}_0 : \quad \mathbb{E}_P[E] \leq 1.$$

E-value

- Simplified version (for fixed n): non-negative random variable E satisfying

$$\text{for all } P \in \mathscr{H}_0 : \quad \mathbb{E}_P[E] \leq 1.$$

- Bayes factors with special priors are e-values

$$\text{BF} = \frac{p(X_1, \ldots, X_n \mid H_1)}{p(X_1, \ldots, X_n \mid H_0)} = \frac{\int_{\Theta_1} p_{\theta_1}(X_1, \ldots, X_n \mid \theta_1) w(\theta_1) d\theta_1}{\int_{\Theta_0} p_{\theta_0}(X_1, . . ., X_n \mid \theta_0) w(\theta_0) d\theta_0}$$

Advantages of e-values

- Sequential testing, validity under optional stopping
- Easy combination (several studies/meta analysis)
- Easy interpretation: betting. High e-value is more evidence against H_0
- E-values can be constructed from different paradigms: frequentist, objective Bayesian, subjective Bayesian, strict Neyman-Pearsonian, and others
- Work on making optimal e-values (that grow fastest when H_0 is not true, see e.g. Grünwald, De Heide & Koolen 2024)

References

- Pearson, K. "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling". Philosophical Magazine. Series 5. 50 (302): 157–175. (1900).
- Fisher, R. Statistical Methods For Research Workers, Cosmo study guides. (1925).
- Ioannidis, J. Why most published research findings are false, PLoS Medicine 2(8) (2005).
- 270 authors, Estimating the reproducibility of psychological science, Science 349 (6251), 2015.
- Wulff HR, Andersen B, Brandenhoff P, Guttler F. What do doctors know about statistics?. Statistics in medicine. 1987 Jan;6(1):3-10.
- John LK, Loewenstein G, Prelec D. Measuring the prevalence of questionable research practices with incentives for truth telling. Psychological science. 2012 May;23(5):524-32.
- Hendriksen A, de Heide R, Grünwald P. Optional stopping with Bayes factors: a categorization and extension of folklore results, with an application to invariant situations. Bayesian Analysis. 2021 Sep;16(3):961-89.
- De Heide R, Grünwald PD. Why optional stopping can be a problem for Bayesians. Psychonomic Bulletin & Review. 2021 Jun;28:795-812.
- Grünwald, P., De Heide, R., Koolen, W., Safe Testing. JRSS-B (2024)
- Fisher, R. "Statistical Methods For Research Workers, Cosmo study guides." (1925).
- A. Ramdas - Lecture: http://stat.cmu.edu/~aramdas/betting/Feb11-class.pdf

Extra slides