E-values

OBayes 2025 Tutorial

Dr. Rianne de Heide

Introduction

- @ University of Twente - The Netherlands
- Mathematical Statistics, Machine Learning Theory
- Co-developed new theory of hypothesis testing with e-values
- Recognition: PhD thesis prize, Cor Baayen Early Career Researcher Award, NWO VENI & M2 grants, Bernoulli Society New Researcher Award 2025

Foto: Patrick Siemons

Menu

- p-values and why do we need a new theory for hypothesis testing?
- Are Bayes factors the solution?
- e-values
- A trial
- Another e-value highlight: multiple testing

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?

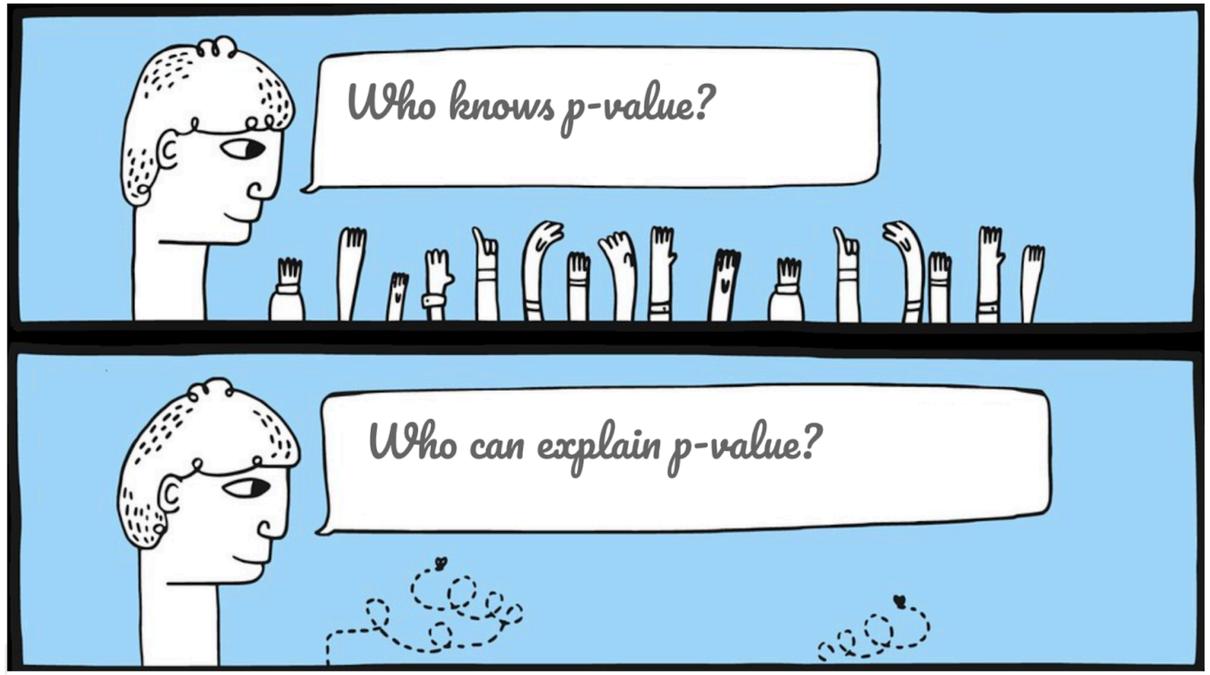
Reproducibility crisis in social and medical science

Causes:

- 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 3125 6009**

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%

Definition of the p-value

A p-value p is a random variable (i.e. a function) such that for every $P \in \mathscr{H}_0$, for $\alpha \in [0,1]$,

$$P(p \leq \alpha) \leq \alpha.$$

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): 55% 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): 55% 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

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, $P_A(\text{RAIN}) \geq P_B(\text{RAIN})$, 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?

Claims about optional stopping with Bayesian methods

- Lindley, 1957; Raiffa and Schlaifer, 1961, Edwards et al., 1963:

(with Bayesian methods) “it is entirely appropriate to collect data until a point has been proven or disproven, or until the data collector runs out of time, money, or patience.”

.

Claims about optional stopping with Bayesian methods

- Lindley, 1957; Raiffa and Schlaifer, 1961, Edwards et al., 1963:

  (with Bayesian methods) “it is entirely appropriate to collect data until a point has been proven or disproven, or until the data collector runs out of time, money, or patience.”

- Renewed interest: Wagenmakers 2007; Rouder 2014; Schönbrodt et al, 2017; Yu et al, 2014; Sanborn and Hills, 2014

“Bayes factors can handle optional stopping”

But what does that mean mathematically?

“Bayes factors can handle optional stopping”

But what does that mean mathematically?

Problems:

- Different authors mean different things by this claim
- Claims are often shown only in an informal sense, or restricted contexts

See the paper:

Optional Stopping with Bayes Factors: a categorization and extension of folklore results, with an application to invariant situations

Allard Hendriksen, Rianne de Heide, Peter Grünwald

Bayesian Analysis 16(3):961–989, 2021, doi:10.1214/20-BA1234.

“Bayes factors can handle optional stopping”

But what does that mean mathematically?

Problems:

- Different authors mean different things by this claim
- Claims are often shown only in an informal sense, or restricted contexts

Goal of the paper:

- systematic overview and formalization
- formal verification (proofs) and extension

Overview

- Identify 3 main mathematical senses in which Bayes factor methods can handle optional stopping

- Explain the practical notions of these notions

Conclusion

Whether Bayes factors can handle optional stopping is subtle, depending on the specifics of the given situation: what models are used, what priors, and what is the goal of the analysis.

Setting

- Hypothesis testing: H_0 versus H_1, sets of distributions, represented by unique probability distributions $\bar{P}_0$ and $\bar{P}_1$

Setting

- Hypothesis testing: H_0 versus H_1, sets of distributions, represented by unique probability distributions $\bar{P}_0$ and $\bar{P}_1$
- $\bar{P}_0$ and $\bar{P}_1$ are Bayes marginal distributions:

$$\bar{P}_0(A) = \int_{\Theta_0} P_{\theta|0}(A)\mathrm{d}\pi_0(\theta); \quad \bar{P}_1(A) = \int_{\Theta_1} P_{\theta|1}(A)\mathrm{d}\pi_1(\theta)$$

Setting

- Hypothesis testing: H_0 versus H_1, sets of distributions, represented by unique probability distributions $\bar{P}_0$ and $\bar{P}_1$
- $\bar{P}_0$ and $\bar{P}_1$ are Bayes marginal distributions:

$$\bar{P}_0(A) = \int_{\Theta_0} P_{\theta|0}(A)\mathrm{d}\pi_0(\theta); \quad \bar{P}_1(A) = \int_{\Theta_1} P_{\theta|1}(A)\mathrm{d}\pi_1(\theta)$$

- $$\frac{\pi(H_1 \mid A)}{\pi(H_0 \mid A)} = \frac{P(A \mid H_1)}{P(A \mid H_0)} \cdot \frac{\pi(H_1)}{\pi(H_0)}$$

1) τ -independence

- Given a stopping time τ, and a data sequence x^n compatible with τ, we have

$$\frac{\pi(H_1 \,|\, X^n = x^n, \tau = n)}{\pi(H_0 \,|\, X^n = x^n, \tau = n)} = \frac{P(\tau = n \,|\, X^n = x^n, H_1) \cdot \pi(H_1 \,|\, X^n = x^n)}{P(\tau = n \,|\, X^n = x^n, H_0) \cdot \pi(H_0 \,|\, X^n = x^n)}$$

$$= \frac{\pi(H_1 \,|\, X^n = x^n)}{\pi(H_0 \,|\, X^n = x^n)}$$

1) τ -independence

- Given a stopping time τ, and a data sequence x^n compatible with τ, we have

$$\frac{\pi(H_1 \,|\, X^n = x^n, \tau = n)}{\pi(H_0 \,|\, X^n = x^n, \tau = n)} = \frac{P(\tau = n \,|\, X^n = x^n, H_1) \cdot \pi(H_1 \,|\, X^n = x^n)}{P(\tau = n \,|\, X^n = x^n, H_0) \cdot \pi(H_0 \,|\, X^n = x^n)}$$

$$= \frac{\pi(H_1 \,|\, X^n = x^n)}{\pi(H_0 \,|\, X^n = x^n)}$$

- $$\frac{\overbrace{\pi(H_1 \,|\, X^n = x^n, \tau = n)}^{\gamma(x^n)}}{\pi(H_0 \,|\, X^n = x^n, \tau = n)} = \frac{\overbrace{\bar{P}_1(X^n = x^n)}^{\beta(x^n)}}{\bar{P}_0(X^n = x^n)} \cdot \frac{\pi(H_1)}{\pi(H_0)}$$

2) Calibration

Rouder (2014)

- Nominal posterior odds: $\gamma(x^n)$
- Observed posterior odds: $\dfrac{\pi(H_1 \,|\, \gamma(x^n) = c)}{\pi(H_0 \,|\, \gamma(x^n) = c)}$

2) Calibration

Rouder (2014)

- Nominal posterior odds: $\gamma(x^n)$
- Observed posterior odds: $\dfrac{\pi(H_1 \,|\, \gamma(x^n) = c)}{\pi(H_0 \,|\, \gamma(x^n) = c)}$
- Calibration under optional stopping: $c = \dfrac{P(\beta(x^\tau) = c \,|\, H_1)}{P(\beta(x^\tau) = c \,|\, H_0)}$

2) Calibration

Rouder (2014)

- Nominal posterior odds: $\gamma(x^n)$
- Observed posterior odds: $\dfrac{\pi(H_1 \,|\, \gamma(x^n) = c)}{\pi(H_0 \,|\, \gamma(x^n) = c)}$
- Calibration under optional stopping: $c = \dfrac{P(\beta(x^\tau) = c \,|\, H_1)}{P(\beta(x^\tau) = c \,|\, H_0)}$
- Note: result relies on priors not depending on the stopping time

3) (semi-)frequentist optional stopping

Def. A function $S : \cup_{i=m}^{T} \mathcal{X}^i \rightarrow \{0,1\}$ is said to be a frequentist sequential test with significance level α that is robust under optional stopping relative to H_0 if for all $P \in H_0$,

$$P(\exists n \leq T : S(X^n) = 1) \leq \alpha,$$

that is, the probability that there exists an n at which $S(X^n) = 1$ is bounded by α.

3) (semi-)frequentist optional stopping

Def. A function $S : \cup_{i=m}^{T} \mathcal{X}^i \to \{0,1\}$ is said to be a frequentist sequential test with significance level α that is robust under optional stopping relative to H_0 if for all $P \in H_0$,

$$P(\exists n \leq T : S(X^n) = 1) \leq \alpha,$$

that is, the probability that there exists an n at which $S(X^n) = 1$ is bounded by α.

Fact: $\bar{P}_0 \left(\exists n, 0 < n \leq T : \frac{1}{\beta(x^n)} \leq \alpha \right) \leq \alpha$

Why should we care?

- This all shows that Bayesian methods can deal with optional stopping, right? (Except for the case of fully frequentist OS with composite H_0)

Why should we care?

- This all shows that Bayesian methods can deal with optional stopping, right? (Except for the case of fully frequentist OS with composite H_0)

- Well, it's more subtle…

Why should we care?

- This all shows that Bayesian methods can deal with optional stopping, right? (Except for the case of fully frequentist OS with composite H_0)

- Well, it's more subtle…

- In many practical situations, results become non-intepretable or even undefined.

When problems arise: Subjective vs. Pragmatic and Default priors

- Bayesians view probabilities as degree of belief

When problems arise: Subjective vs. Pragmatic and Default priors

- Bayesians view probabilities as degree of belief, which is expressed as a prior

When problems arise: Subjective vs. Pragmatic and Default priors

- Bayesians view probabilities as degree of belief, which is expressed as a prior
- Then the prior is updated with data, and the posterior can be used to base decisions on

When problems arise: Subjective vs. Pragmatic and Default priors

- Bayesians view probabilities as degree of belief, which is expressed as a prior
- Then the prior is updated with data, and the posterior can be used to base decisions on

- For *subjectivists*, this is the full story

When problems arise: Subjective vs. Pragmatic and Default priors

- Bayesians view probabilities as degree of belief, which is expressed as a prior
- Then the prior is updated with data, and the posterior can be used to base decisions on

- For *subjectivists*, this is the full story
- *Objectivists*: indifference, a single, rational probability function

When problems arise: Subjective vs. Pragmatic and Default priors

- Bayesians view probabilities as degree of belief, which is expressed as a prior
- Then the prior is updated with data, and the posterior can be used to base decisions on

- For *subjectivists*, this is the full story
- *Objectivists*: indifference, a single, rational probability function
- Pragmatic Bayesians: *default* priors

When problems arise: Subjective vs. Pragmatic and Default priors

- Recent papers that advocate the use of Bayesian methods are based on such *default* priors (Rouder et al. 2009, 2012; Jamil et al. 2016)

When problems arise: Subjective vs. Pragmatic and Default priors

- Recent papers that advocate the use of Bayesian methods are based on *default* priors (Rouder et al. 2009, 2012; Jamil et al. 2016)
- Within the statistics community, a *pragmatic* stance is most common nowadays

When problems arise: Subjective vs. Pragmatic and Default priors

- Recent papers that advocate the use of Bayesian methods are based on *default* priors (Rouder et al. 2009, 2012; Jamil et al. 2016)
- Within the statistics community, a *pragmatic* stance is most common nowadays
- Pragmatic/default priors have some arbitrary aspects: sensitivity analyses become important

When problems arise: Subjective vs. Pragmatic and Default priors

- Recent papers that advocate the use of Bayesian methods are based on *default* priors (Rouder et al. 2009, 2012; Jamil et al. 2016)
- Within the statistics community, a *pragmatic* stance is most common nowadays
- Pragmatic/default priors have some arbitrary aspects: sensitivity analyses become important

- τ-independence and calibration are fully subjective definitions of OS!

Problems with different types of priors

- Type 0: Right-Haar priors on group invariant nuisance parameters

Problems with different types of priors

- Type 0: Right-Haar priors on group invariant nuisance parameters

- Type I: Default/pragmatic priors that do *not* depend on any aspects of the experimental set-up

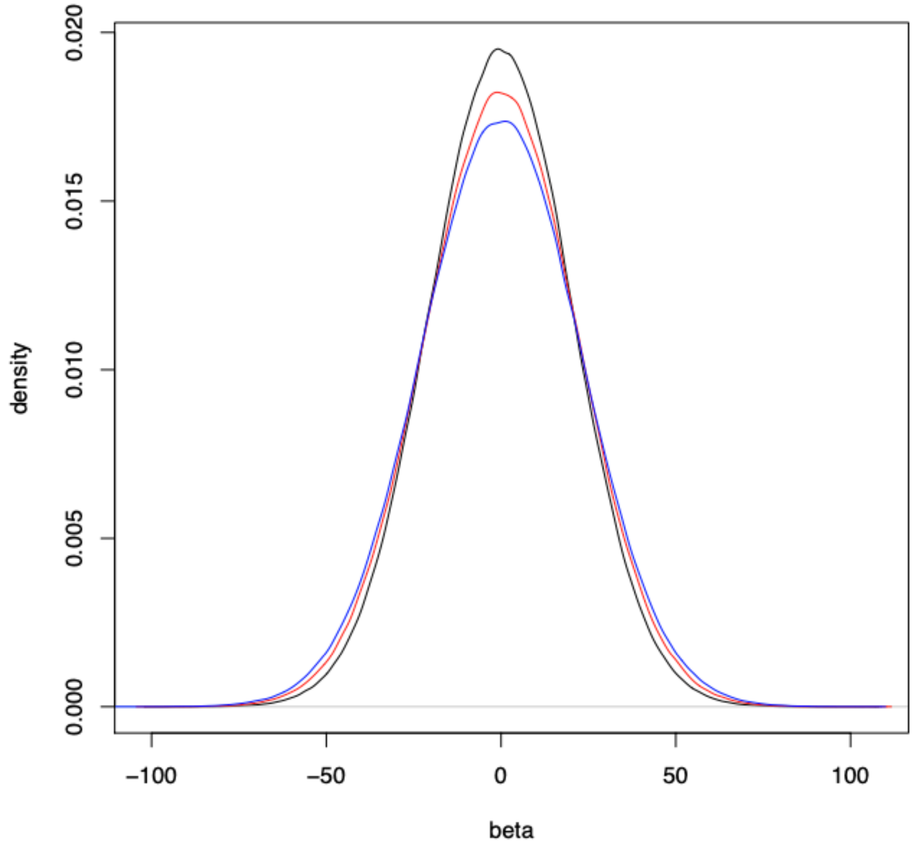
Problems with different types of priors

- Type 0: Right-Haar priors on group invariant nuisance parameters

- Type I: Default/pragmatic priors that do *not* depend on any aspects of the experimental set-up

- Type II: Default/pragmatic priors not of type 0 or I

The problem with type II priors



$$y \sim \mathrm{N}\left(\mu + X\beta, \sigma^2\right),$$
$$\beta \sim \mathrm{N}\left(0, g\sigma^2 n(X'X)^{-1}\right),$$
$$g \sim \mathrm{IG}\left(\frac{1}{2}, \frac{\sqrt{2}}{8}\right).$$

- Not defined under optional stopping

The problem with type I priors

- Example: Jeffreys' Bayesian t-tests: Cauchy prior (type I) on the effect size

The problem with type I priors

- Example: Jeffreys' Bayesian t-tests: Cauchy prior (type I) on the effect size

- The Issue: do we really believe that a Cauchy prior accurately reflects our prior beliefs? Example: test of fertilizer on wheat growth.

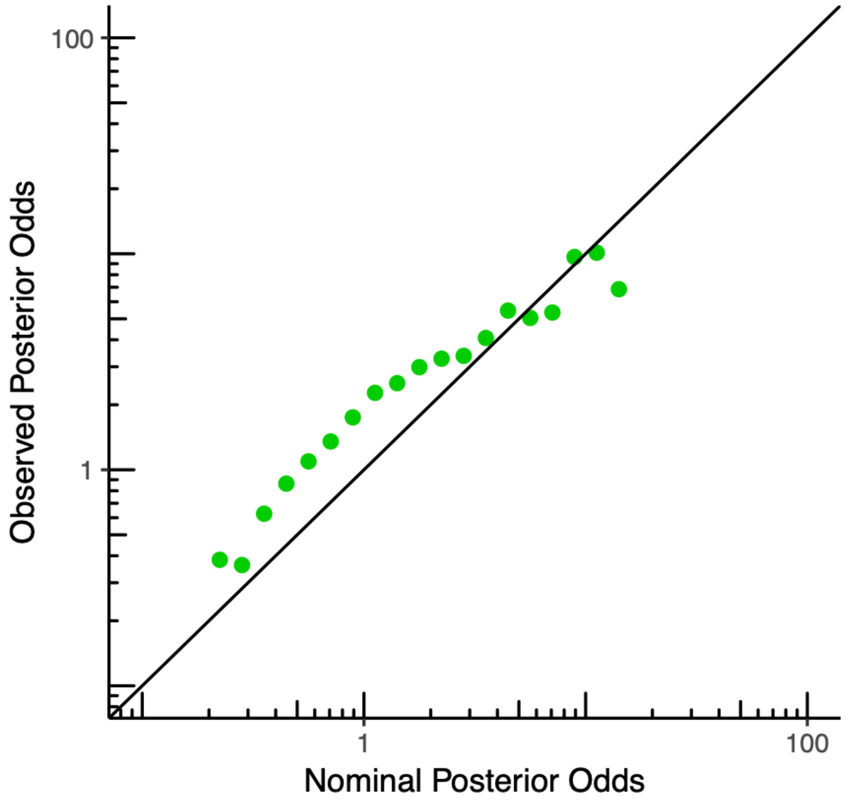
The problem with type I priors

- Example: Jeffreys' Bayesian t-tests: Cauchy prior (type I) on the effect size

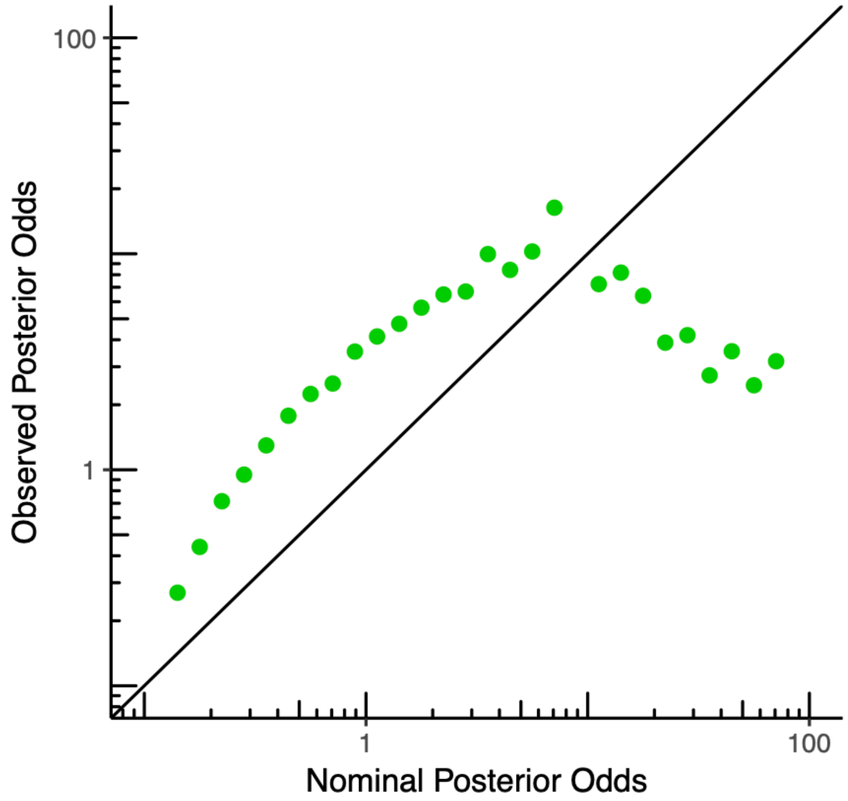
- The Issue: do we really believe that a Cauchy prior accurately reflects our prior beliefs? Example: test of fertilizer on wheat growth.
- Objective Bayesians change their priors depending on the inference task
- The prior is used as a tool in inferring likely parameters or hypotheses, and not to be thought of as something that prescribes how actual data will arise or tend to look like

.

Strong calibration



Fixed sample size



Optional stopping

Conclusion

- Can we do optional stopping with Bayes factors?

  Whether Bayes factors can handle optional stopping is subtle, depending on the specifics of the given situation: what models are used, what priors, and what is the goal of the analysis.

- For most practical Bayesian hypothesis testing problems, one should be careful with optional stopping

Bayes factors and full frequentist optional stopping

- When H_0 is simple, we have the bound

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

  (we will later see that BF here is an *e-value*)

Bayes factors and optional stopping

- When H_0 is simple, we have the bound

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

- 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, though it becomes subtle as to which filtration the process is then adapted to)

e-values

A fair coin?

$$L_0 = 1$$

A fair coin?

$L_0 = 1$

$\lambda_1 = 0.2$ (on heads)

A fair coin?

$L_0 = 1$

$\lambda_1 = 0.2$ (on heads)

A fair coin?

$L_0 = 1$

$B_1 = -1$

$\lambda_1 = 0.2$ (on heads)

A fair coin?

$L_0 = 1$

$B_1 = -1$

$\lambda_1 = 0.2$ (on heads)

$L_1 = L_0 \cdot (1 + \lambda_1 B_1) = 0.8$

A fair coin?

$L_0 = 1$

$B_1 = -1$

$\lambda_1 = 0.2$ (on heads)

$L_1 = L_0 \cdot (1 + \lambda_1 B_1) = 0.8$

$\lambda_2 = 0.4$ (on heads)

A fair coin?

$$L_0 = 1$$

$$B_1 = -1$$

$$\lambda_1 = 0.2 \text{ (on heads)}$$

$$L_1 = L_0 \cdot (1 + \lambda_1 B_1) = 0.8$$

$$B_2 = +1$$

$$\lambda_2 = 0.4 \text{ (on heads)}$$

$$L_2 = L_1 \cdot (1 + \lambda_2 B_2) = 1.12$$

A fair coin?

$$L_0 = 1$$

$$B_1 = -1$$

$$\lambda_1 = 0.2 \text{ (on heads)}$$

$$L_1 = L_0 \cdot (1 + \lambda_1 B_1) = 0.8$$

$$B_2 = +1$$

$$\lambda_2 = 0.4 \text{ (on heads)}$$

$$L_2 = L_1 \cdot (1 + \lambda_2 B_2) = 1.12$$

$$L_t := \prod_{s=1}^{t} (1 + \lambda_s B_s)$$

A fair coin?

$L_0 = 1$

$B_1 = -1$

$\lambda_1 = 0.2$ (on heads)

$L_1 = L_0 \cdot (1 + \lambda_1 B_1) = 0.8$

$B_2 = +1$

$\lambda_2 = 0.4$ (on heads)

$L_2 = L_1 \cdot (1 + \lambda_2 B_2) = 1.12$

$L_t := \prod_{s=1}^{t} (1 + \lambda_s B_s)$; Under $\mathscr{H}_0$, $(L_t)_{t \in \mathbb{N}}$ is a non-negative martingale.

A fair coin?

$$L_t := \prod_{s=1}^{t} (1 + \lambda_s B_s) \; ; \qquad \text{Under } \mathscr{H}_0, \; (L_t)_{t \in \mathbb{N}} \text{ is a non-negative martingale.}$$

At any stopping time τ, we have $\mathbb{E}_{\mathscr{H}_0}[L_\tau] = 1$ (optional stopping theorem).

A fair coin?

$$L_t := \prod_{s=1}^{t} (1 + \lambda_s B_s) \; ; \qquad \text{Under } \mathscr{H}_0, \; (L_t)_{t \in \mathbb{N}} \text{ is a non-negative martingale.}$$

At any stopping time τ, we have $\mathbb{E}_{\mathscr{H}_0}[L_\tau] = 1$ (optional stopping theorem).

Ville's inequality:

$$\mathbb{P}(\exists t \in \mathbb{N} : L_t > 1/\alpha) \leq \alpha$$

p-value equivalent:

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

A fair coin?

$$L_t := \prod_{s=1}^{t} (1 + \lambda_s B_s) \; ; \qquad \text{Under } \mathscr{H}_0, \; (L_t)_{t \in \mathbb{N}} \text{ is a non-negative martingale.}$$

At any stopping time τ, we have $\mathbb{E}_{\mathscr{H}_0}[L_\tau] = 1$ (optional stopping theorem).

Ville's inequality:

$$\mathbb{P}(\exists t \in \mathbb{N} : L_t > 1/\alpha) \leq \alpha$$

p-value equivalent:

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

L_t is called an e-value

L_t measures evidence against $\mathscr{H}_0$

E-values

- e-value: non-negative random variable E satisfying

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

E-values

- e-value: non-negative random variable E satisfying

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

- We can define hypothesis tests based on e-values.

E-values

- e-value: non-negative random variable E satisfying

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

- But what is a good e-value?

E-values

- e-value: non-negative random variable E satisfying

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

- But what is a good e-value?

- GROW: Growth-Rate Optimal in Worst case: the e-value E^* that achieves

$$\max_{E : E \text{ is an e-value}} \min_{P \in \mathcal{H}_1} \mathbb{E}_P[\log E]$$

Safe Testing (Grünwald, De Heide, Koolen)

- The GROW e-value $E^*_{W_1}$ exists (for composite $\mathscr{H}_0$), and satisfies

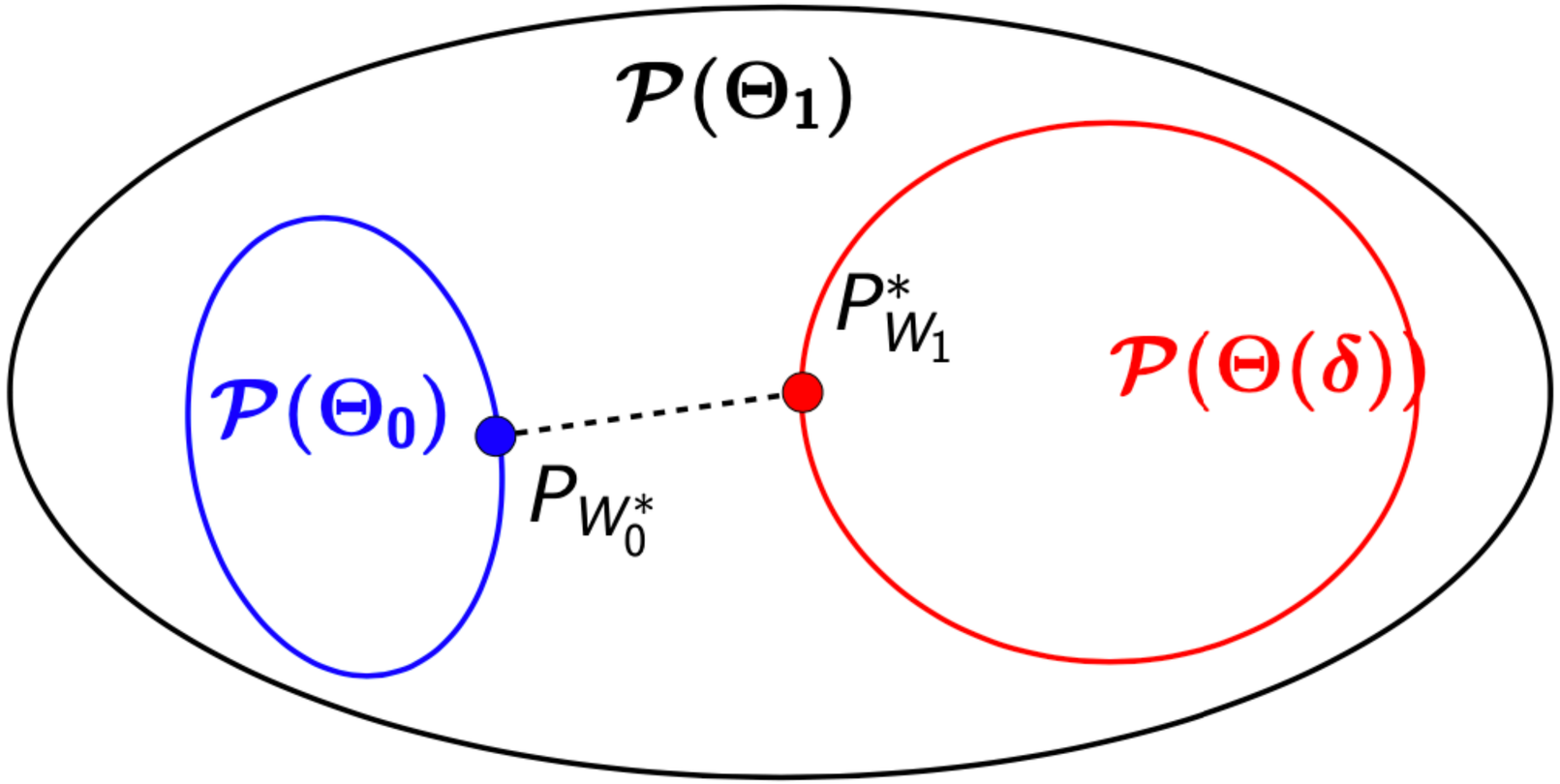
$$\mathbb{E}_{Z\sim P_{W_1}}[\log E^*_{W_1}] = \sup_{E\in\mathscr{E}} \mathbb{E}_{Z\sim P_{W_1}}[\log E] = \inf_{W_0\in\mathscr{W}_0} D(P_{W_1} \,\|\, P_{W_0})$$

- if the inf is achieved by some W_0°, the GROW e-value takes a simple form: $E^*_{W_1} = p_{W_1}(Z)/p_{W_0^\circ}(Z)$

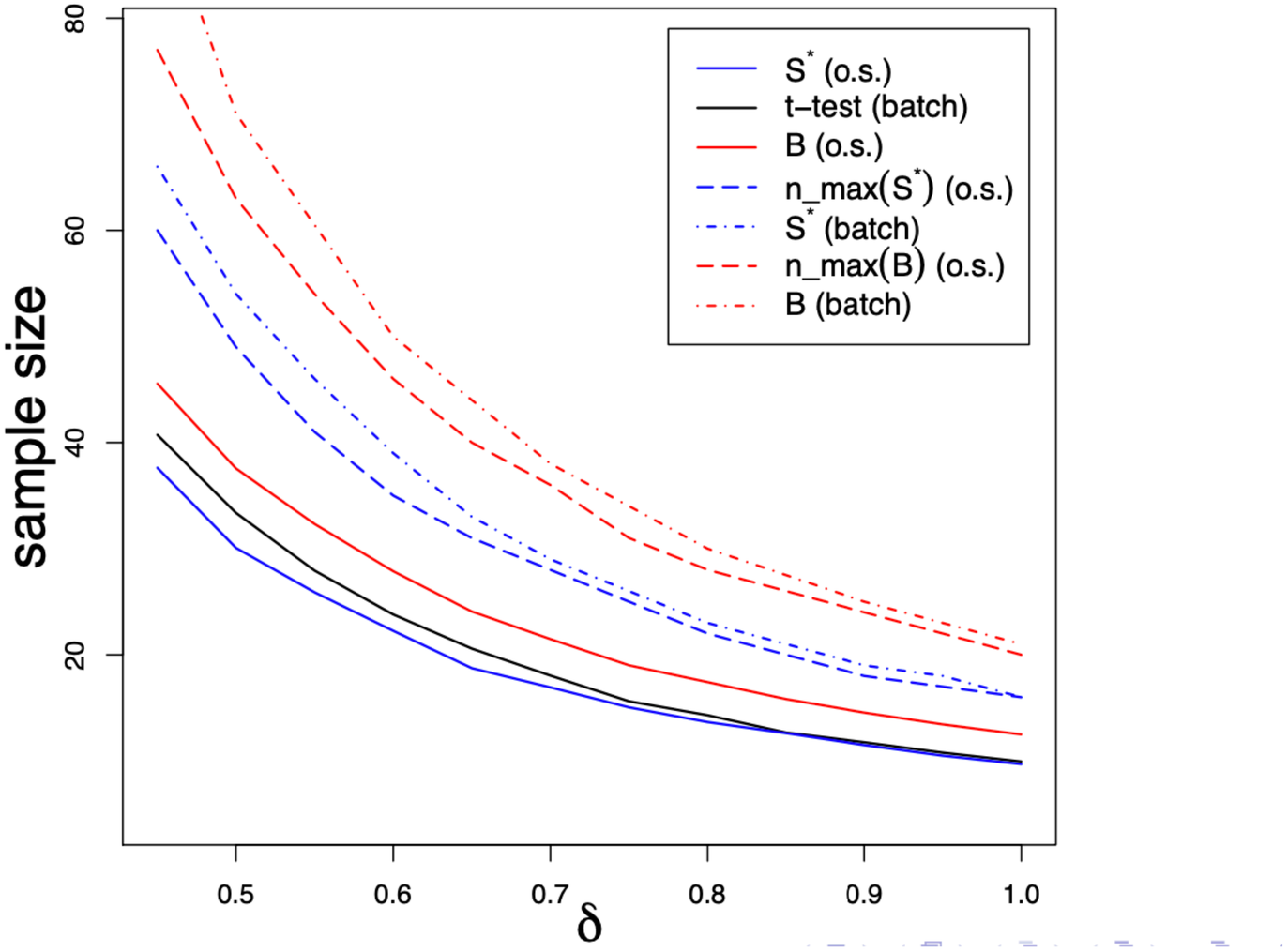
- GROW e-values $E^*_{\mathscr{W}_1} = p_{W_1^*}(Z)/p_{W_0^*}(Z)$ can be found by a double KL-minimization problem $\min_{W_1\in\mathscr{W}_1} \min_{W_0\in\mathscr{W}_0} D(P_{W_1} \,\|\, P_{W_0})$ and they satisfy

$$\inf_{W\in\mathscr{W}_1} \mathbb{E}_{Z\sim P_W}[\log E^*_{\mathscr{W}_1}] = \sup_{E\in\mathscr{E}} \inf_{W\in\mathscr{W}_1} \mathbb{E}_{Z\sim P_W}[\log E] = D(P_{W_1^*} \,\|\, P_{W_0^*})$$

Joint information projection



Simulation example: t-test



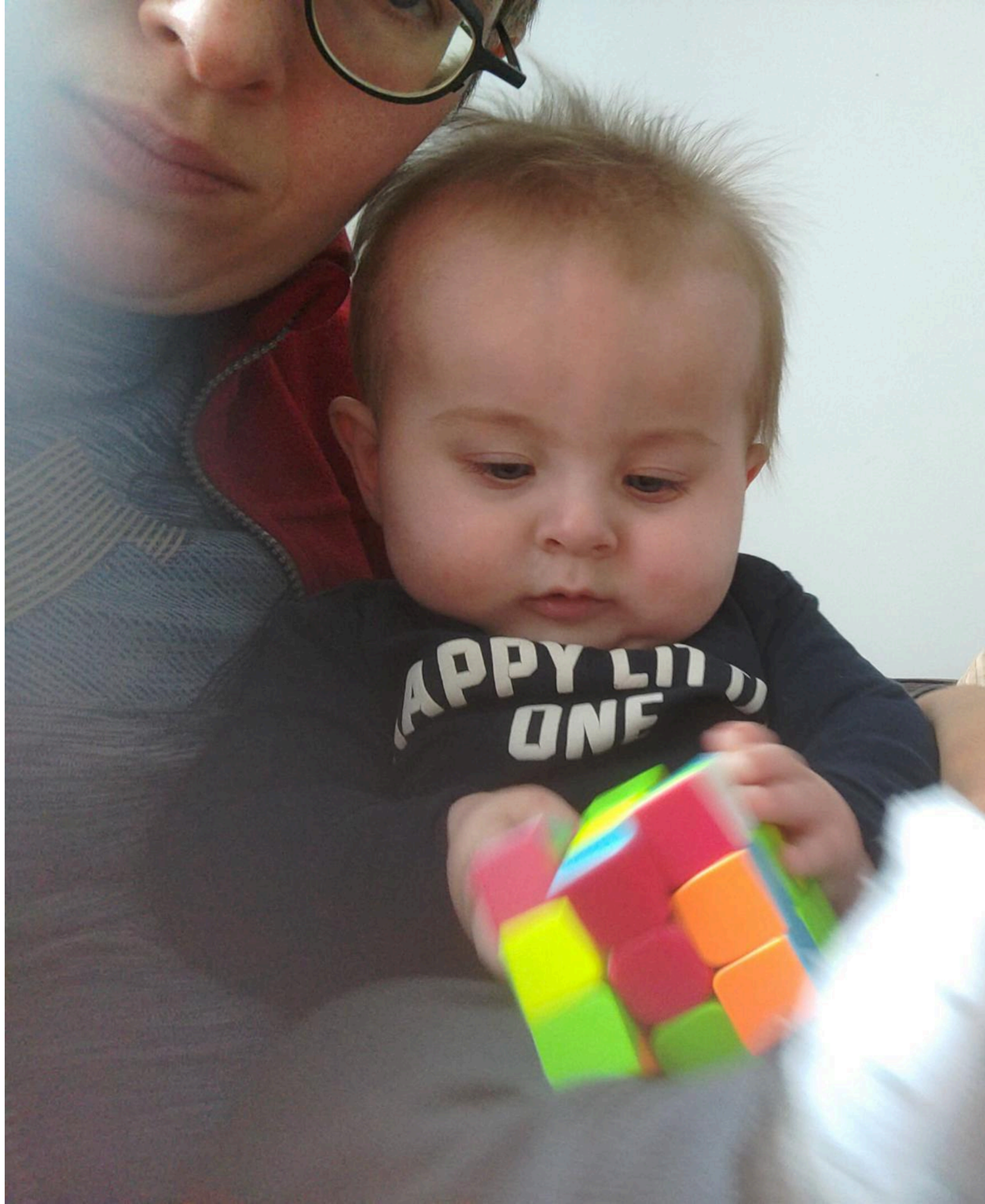
Advantages of e-values

- Any-time valid 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
- Many interesting properties, e.g. in multiple testing allowing for general dependence in FDR methods, derandomization of knock-offs, etc.

A trial

A (real) trial

- Group A: standard boosters
- Group B: new boosters
- Outcome: no leakage (0) or leakage (1)
- Assumption: data is i.i.d. Bernoulli with parameter θ determining the probability of leakage.



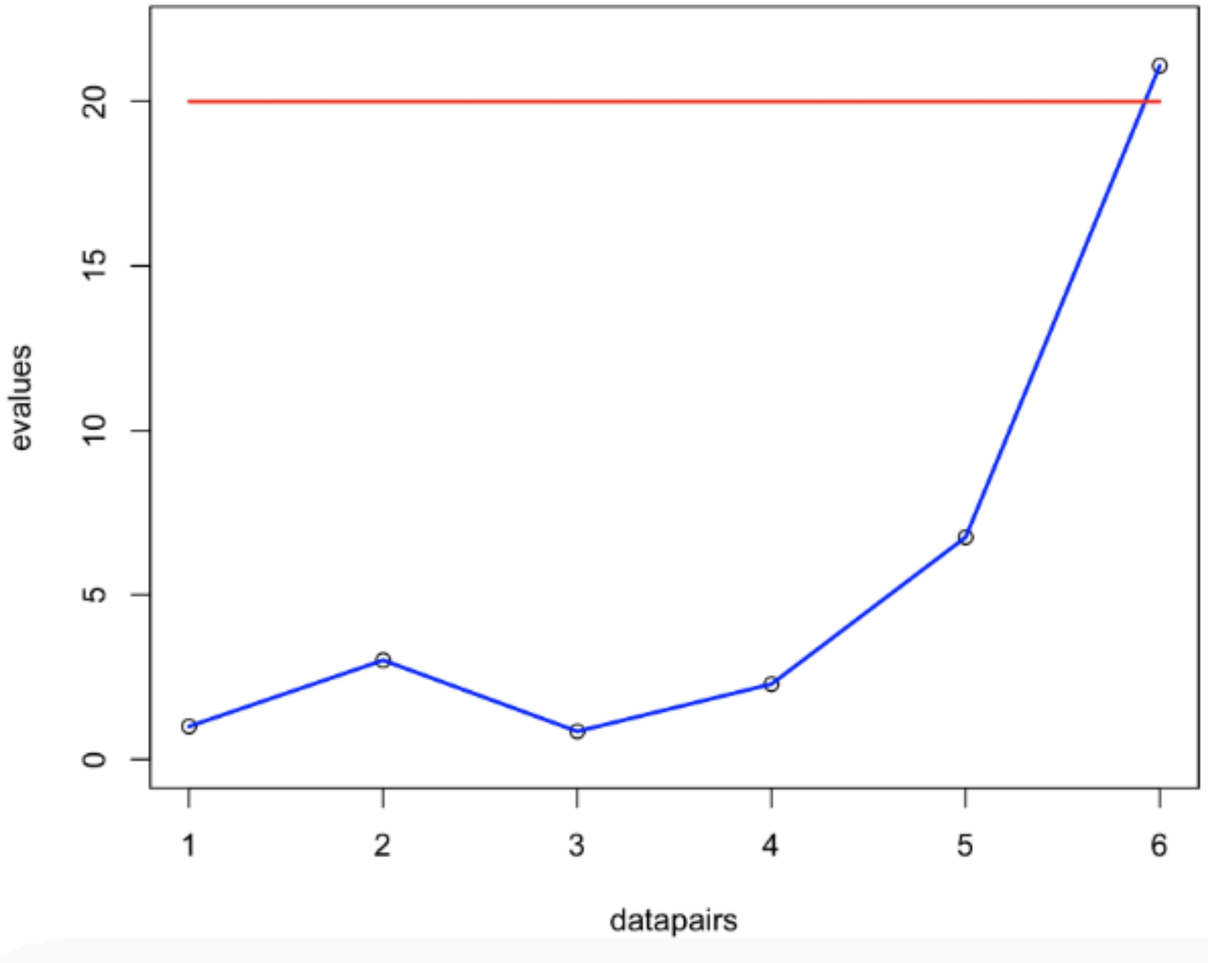
A (real) trial

- Data streams $Y_{1,A}, Y_{2,A}, \ldots \overset{i.i.d.}{\sim} P_{\theta_A}$ and $Y_{1,B}, Y_{2,B}, \ldots \overset{i.i.d.}{\sim} P_{\theta_B}$,
- $\mathscr{H}_0 : \theta_A = \theta_B$
- $\mathscr{H}_1 : \theta_A \neq \theta_B$
- Data is gathered in pairs. After each pair we calculate the e-value.
- We have a Type I error guarantee if we do this. We can stop whenever we like, in particular, if the e-value exceeds 20.

Analysis

- safe.prop.test(ya=ya, yb=yb, pilot=T)

	1	2	3	4	5	6
normale boosters	0	0	1	0	0	0
nieuwe boosters	1	1	1	1	1	1



How to do this with p-values?

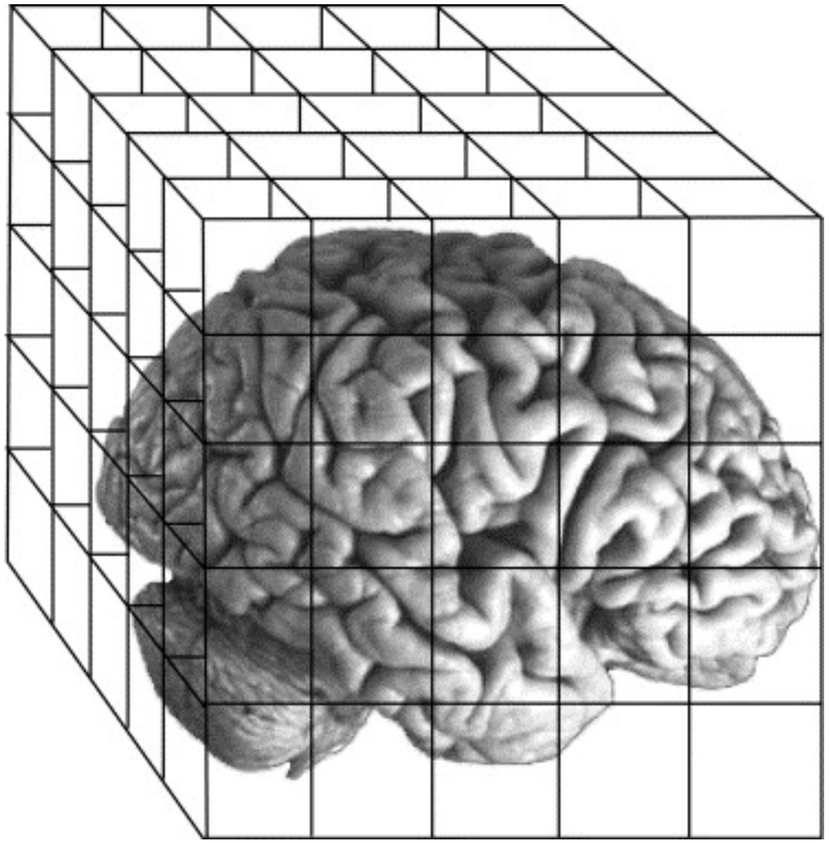
How to do this with p-values?

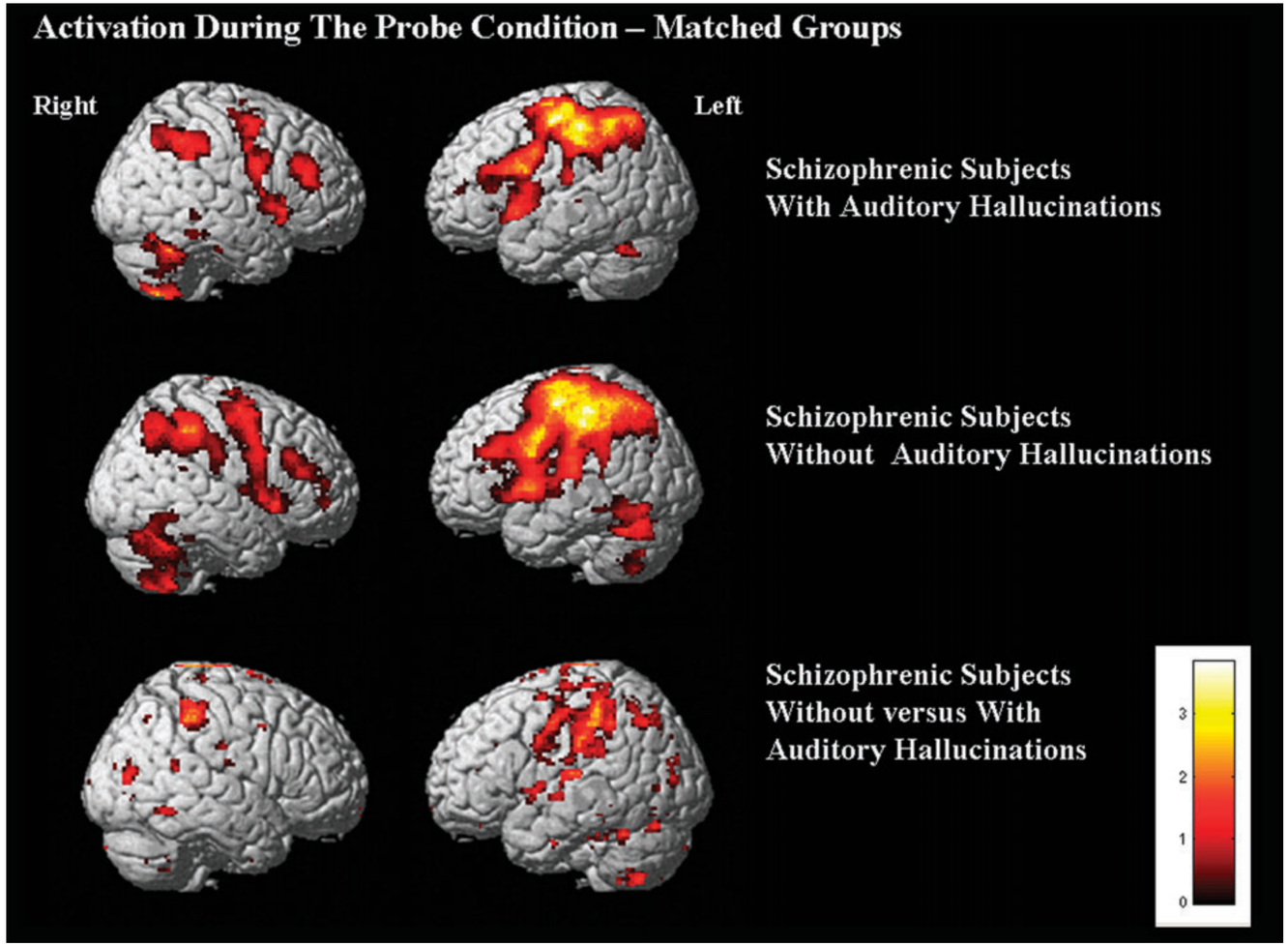
- No idea about the effect size, not even in which direction.
- Pilot study with 12 trials in either group.
- Then estimate the effect size.
- Then calculate the sample size needed.
- Then do the experiment.
- Suppose the (second) experiment would also take 12 nights: at least 18 nights with leakage: stop early because of ethical reasons. Not even possible to report a p-value.

Veni project:
Multiple testing with e-values

Example: multiple testing in neuroimaging

130.000 voxels





Wible, Cynthia G. et al. “fMRI activity correlated with auditory hallucinations during performance of a working memory task: data from the FBIRN consortium study.” Schizophrenia bulletin 35 1 (2009): 47-57 .

Bringing flexibility to multiple testing

- Researchers want to work interactively with the data, which is not possible with current methods

- How can this be achieved? New theory of hypothesis testing with e-values

Bringing flexibility to multiple testing

- Researchers want to work interactively with the data, which is not possible with current methods

- How can this be achieved? New theory of hypothesis testing with e-values

- Current research aim: rigorous mathematical theory for multiple testing with e-values and e-processes

e-BH (Wang & Ramdas, 2021)

- Let $e_{[k]}$ be the kth order statistic of $e_1, \ldots, e_K$, from the largest to the smallest.
- Define the test procedure which rejects hypotheses with the largest $k_e^\star$ e-values, where

$$k_e^\star = \max \left\{ k \in \mathcal{K} : \frac{k e_{[k]}}{K} \geq \frac{1}{\alpha} \right\}.$$

- This procedure controls the FDR at level α even under unknown arbitrary dependence between the e-values.
- BH and BY are special cases of e-BH.

Exciting new result: bringing closure to FDR

With Jelle Goeman, Aldo Solari, Aaditya Ramdas, Neil Xu, Lasse Fisher

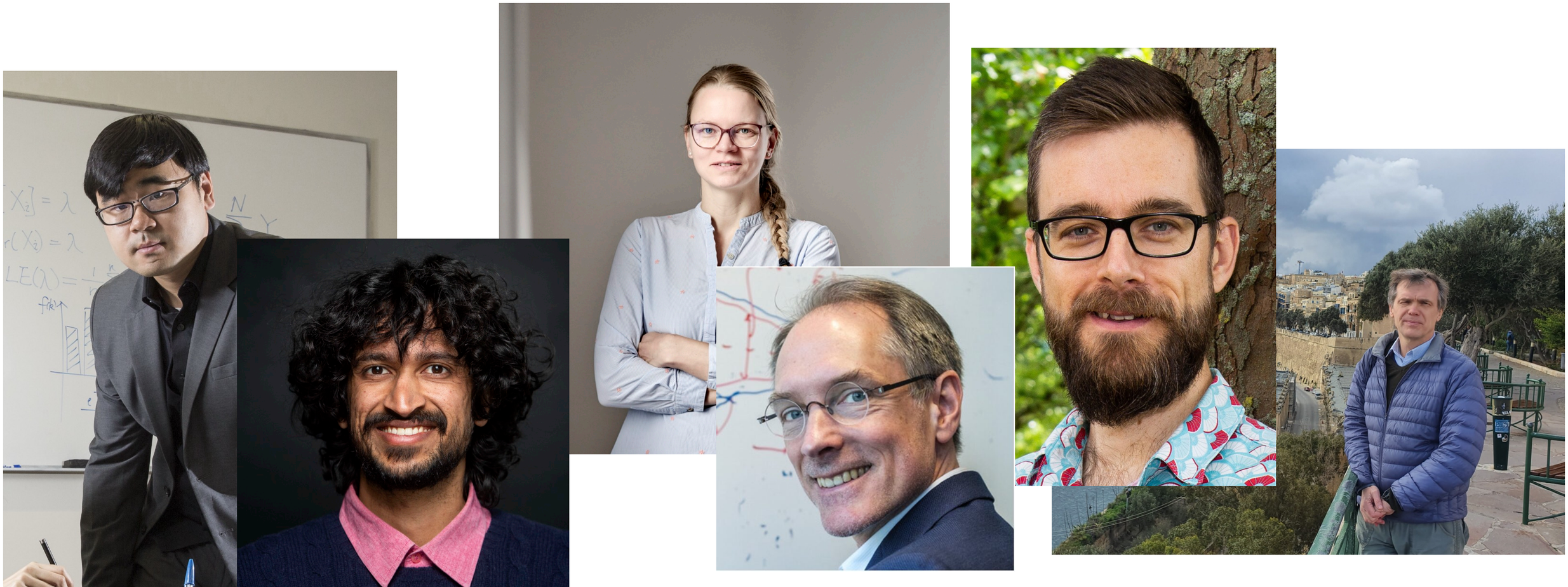
- Necessary and sufficient principle for multiple testing methods controlling an expected loss (think of FDR)
- Every such multiple testing method is a special case of a general closed testing procedure based on e-values.
- Uniform improvements of these methods
- Simultaneous error control
- Post-hoc flexibility for the user choice of alpha, target error rate, and sometimes even nominal error rate
- Restricted combinations possible - exploiting logical relationships between hypotheses

The e-Partitioning Principle of False Discovery Rate Control
J Goeman, R de Heide, A Solari - arXiv preprint arXiv:2504.15946, 2025

Bringing closure to FDR control: beating the e-Benjamini-Hochberg procedure
Z Xu, L Fischer, A Ramdas - arXiv preprint arXiv:2504.11759, 2025

The future of e-values

- Many groups studying e-values now (in mathematical statistics, probability theory): e.g. CWI, CMU, ETH, Waterloo, London, Stanford, Twente…



Questions?

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