E-values

Athens University of Economics and Business

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?
- E-values, e-processes and e-value based multiple testing
- A problem
- A proposition
- A solution

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?

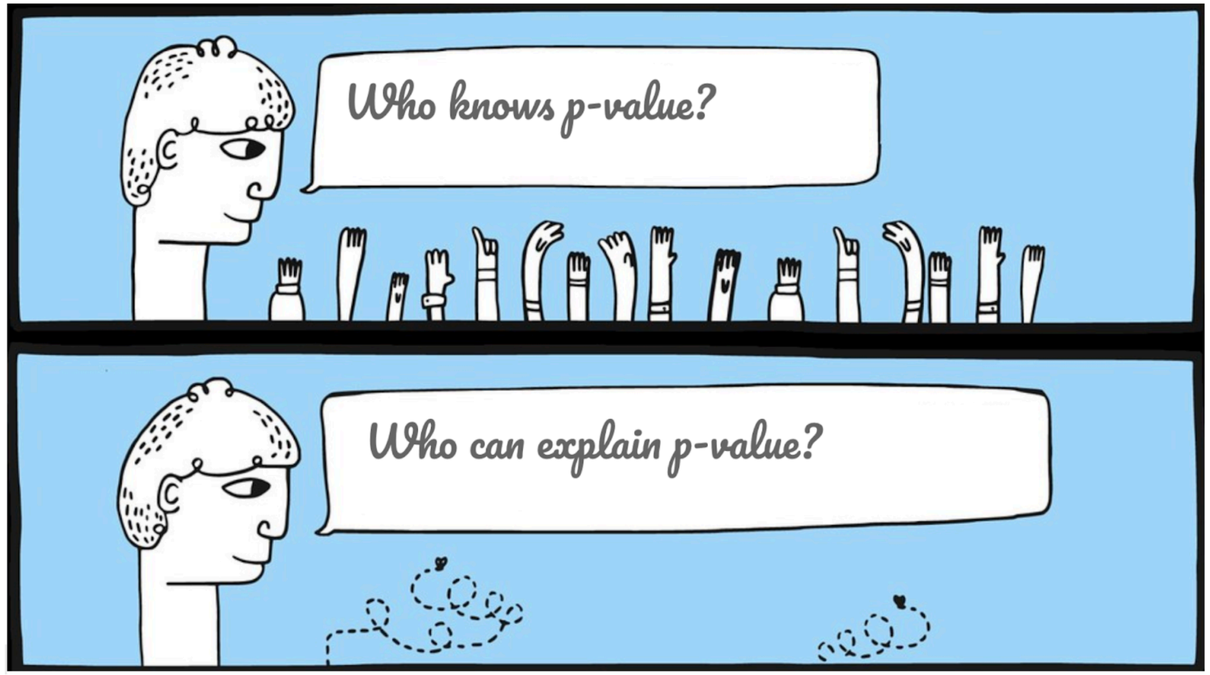
Reproducibility crisis in social and medical science

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 5550 2085**

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 $\alpha \in [0,1]$,

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

What is a p-value actually?

- The p-value gives the probability of observing this data, or more extreme data, given that the null hypothesis is true. Notation: $P(X^n \mid \mathscr{H}_0)$.

What is a p-value actually?

- The p-value gives the probability of observing this data, or more extreme data, given that the null hypothesis is true. Notation: $P(X^n \mid \mathscr{H}_0)$.

- Not: probability of the null hypothesis given the data: $P(\mathscr{H}_0 \mid X^n)$.

What is a p-value actually?

- The p-value gives the probability of observing this data, or more extreme data, given that the null hypothesis is true. Notation: $P(X^n \mid \mathscr{H}_0)$.

- Not: probability of the null hypothesis given the data: $P(\mathscr{H}_0 \mid X^n)$.

- But that *is* the quantity we are interested in, right?

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? menti.com 5550 2085

A) You add 10 subjects to the study, and you calculate a new p-value based on the total data, i.e. 80 subjects.

B) You calculate a new p-value for the 10 new subjects, and you multiply that p-value by 0.06, the p-value from the first 70 subjects.

C) You say to your boss: sorry, this is not possible. You are left with your p-value of 0.06, and you cannot conclude any significant result from your research.

D) You calculate a new p-value for the new 10 subjects, and you use a method they also use in meta-analyses to combine the p-values.

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”.

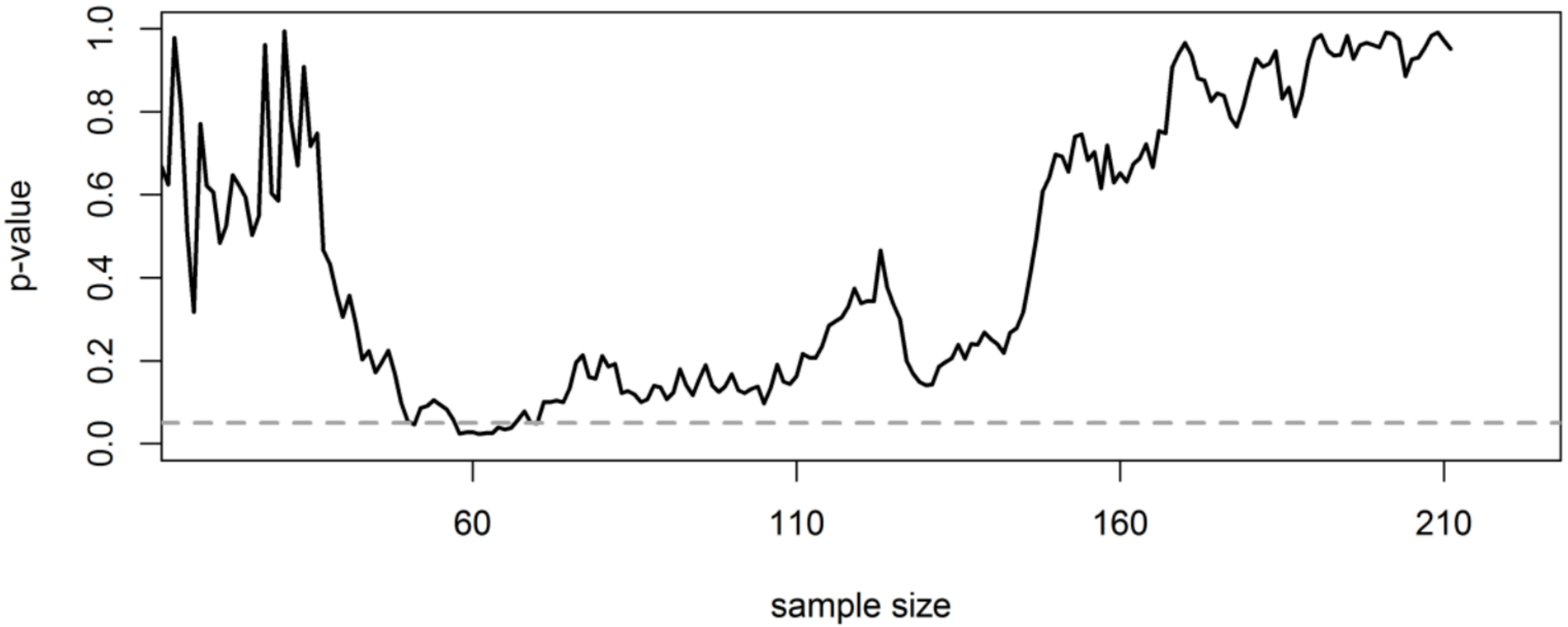
- This is called optional stopping, and invalidates p-values and their error guarantees

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

E-values, e-processes and e-value based multiple testing

Hypothesis testing

- We assume that there is some distribution P^* that is generating data $X := (X_1, X_2, \ldots)$.
- We wish to test a null hypothesis $H_0 : P^* \in \mathscr{P}$, where $\mathscr{P}$ is a set of probability distributions on some sample space.
- Often (but not always), we test H_0 in contrast with an alternative hypothesis $H_1 : P^* \in \mathscr{Q}$, where $\mathscr{Q}$ is again a set of probability distributions, on the same sample space as $\mathscr{P}$.

The e-variable

- **Definition**: e-variable

  An e-variable E for $\mathscr{P}$ is a non-negative random variable satisfying $\mathbb{E}_P[E] \leq 1$ for all $P \in \mathscr{P}$.

The e-variable

- **Definition**: e-variable

  An e-variable E for $\mathscr{P}$ is a non-negative random variable satisfying $\mathbb{E}_P[E] \leq 1$ for all $P \in \mathscr{P}$.

- The value taken by the e-variable after observing the data is called the e-value. However, often, as also happens with the infamous p-value (p-variable), the random variable E itself is also often called e-value.

Tests and the type I error guarantee

- **Definition:** binary test

  A binary test ϕ is a $\{0,1\}$-valued random variable. The type-I error of a test ϕ for P is $\mathbb{E}_P[\phi]$. A test has level $\alpha \in [0,1]$ for $\mathscr{P}$ if its type-I error is at most α for every $P \in \mathscr{P}$.

Tests and the type I error guarantee

- **Definition:** binary test

  A binary test ϕ is a $\{0,1\}$-valued random variable. The type-I error of a test ϕ for P is $\mathbb{E}_P[\phi]$. A test has level $\alpha \in [0,1]$ for $\mathscr{P}$ if its type-I error is at most α for every $P \in \mathscr{P}$.

- **Markov's inequality for e-variables**

  Let E be an e-variable for $\mathscr{P}$. We have $P(E \geq 1/\alpha) \leq \alpha$ for all $P \in \mathscr{P}$ and $\alpha \in (0,1]$. Hence, $\mathbf{1}_{\{E \geq 1/\alpha\}}$ is a binary test of level α.

The e-process: any-time valid inference (1)

- Any-time valid inference is a new paradigm for sequential testing (or estimation) that allows the statistician to stop at any arbitrary stopping time, possibly not anticipated or specified in advance.

- E-processes are the central object for any-time valid inference

The e-process: any-time valid inference (2)

- E-processes are the central object for any-time valid inference
- Filtrations are important
- A filtration is an increasing nested sequence of sigma-algebra's, which we denote with $\mathfrak{F} := (\mathscr{F}_t)_{t \geq 0}$.
- A sequence of random variables $(Y_t)_{t \geq 0}$ is called a process if it is adapted to a filtration $\mathfrak{F}$, that means that every Y_t is measurable with respect to $\mathscr{F}_t$.
- A stopping time (or rule) τ is a nonnegative integer valued random variable such that $\{\tau \leq t\} \in \mathscr{F}_t$ for each $t \geq 0$. Let $\mathscr{T}$ be the set of all stopping times.

The e-process: any-time valid inference (3)

- **Definition:** e-process

  A nonnegative process E is called an e-process if it satisfies $\mathbb{E}_P[E_\tau] \leq 1$ for any stopping time $\tau \in \mathscr{T}$ and any $P \in \mathscr{P}$.

- (There is an equivalent definition that an e-process is a process upper bounded by a family of test martingales.)

Intermezzo: Why are filtrations important?

Pérez-Ortiz, Lardy, De Heide, Grünwald - AoS '24

- Suppose data $X_1, X_2, \ldots$ is sample from $\mathcal{N}(\mu, \sigma^2)$, and let $\delta = \mu/\sigma$. Consider testing $\mathcal{H}_0 : \delta = \delta_0$ vs. $\mathcal{H}_1 : \delta = \delta_1$
- Let $\mathbb{F}$ be the `full' data filtration, and let $\mathbb{G}$ denote the scale-invariant **coarsening** of $\mathbb{F}$:
  $$\mathcal{G}_t = \sigma\left(\frac{X_1}{|X_1|}, \ldots, \frac{X_t}{|X_1|}\right), \forall t$$
- In $\mathbb{G}$, an (in a certain sense optimal) e-process $(\mathbf{e}_t)_{t \geq 0}$ for this problem can be derived.
- This process is not an e-process w.r.t. $\mathbb{F}$:

  If $\tau^{\mathbb{F}} = 1 + \mathbf{1}\{|X_1| \in [0.44, 1.7]\}$, then $\mathbb{E}[\mathbf{e}_{\tau^{\mathbb{F}}}] \approx 1.19 > 1.$

The e-process: any-time valid inference (4)

- **Ville's inequality**

  If E is an e-process for $\mathscr{P}$, then

  $$P\left(\exists t \in \mathbb{N} : E_t \geq \frac{1}{\alpha}\right) \leq \alpha \text{ for every } P \in \mathscr{P} \text{ and } \alpha \in [0,1]$$

  $$\Leftrightarrow \sup_{P \in \mathscr{P}} \left(\sup_{t \in \mathbb{N}} E_t \geq \frac{1}{\alpha} \right) \leq \alpha \text{ for every } \alpha \in [0,1].$$

What are good e-variables?

- E-variable: 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-variable?

- 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) - JRSS-B 2024

- 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

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

Multiple testing: the problem

- If we test n true null hypotheses at level α, then on average we will (falsely) reject αn of them.

Multiple testing: the problem

- If we test n true null hypotheses at level α, then on average we will (falsely) reject αn of them.

- Examples:
  - testing whether some of 20.000 genes are linked to a disease
  - fMRI: 200.000 voxels
  - DNA methylation: 500.000 sites

Multiple testing: the problem

- If we test n true null hypotheses at level α, then on average we will (falsely) reject αn of them.
- Examples:
  - testing whether some of 20.000 genes are linked to a disease
  - fMRI: 200.000 voxels
  - DNA methylation: 500.000 sites
- We need other measures of acceptance/rejection errors.

Multiple testing: the problem

- If we test n true null hypotheses at level α, then on average we will (falsely) reject αn of them.
- Examples:
  - testing whether some of 20.000 genes are linked to a disease
  - fMRI: 100.000 voxels
  - DNA methylation: 500.000 sites
- We need other measures of acceptance/rejection errors.
- We need statistical procedures to control these measures of errors.

Notation

- Goal of a multiple testing procedure: choose a collection $\mathscr{R}$ of hypotheses to reject (often termed: discoveries)

- Errors that can be made:
  - Type I (false positives): rejected and true
  - Type II (false negatives): not rejected and false

	True	False	Total
Rejected	V	U	R
Not rejected	$m_0 - V$	$m_1 - U$	$m - R$
Total	m_0	m_1	m

Error rates

- False discovery proportion (FDP) $Q = \begin{cases} V/R & \text{if } R > 0 \\ 0 & \text{otherwise} \end{cases}$

  i.e. the proportion of false rejections among the rejections.

- In general $\mathscr{R}$ is random, so both V and Q are random variables, so we must focus on relevant aspects of their distribution.

- Family-wise error (FWER) $= P(V > 0) = P(Q > 0)$

- False Discovery Rate (FDR) = $\mathbb{E}(Q)$

e-BH (Wang & Ramdas, JRSS-B 2022)

- Associate each H_j with e-value e_j.
- 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

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

-

e-BH (Wang & Ramdas, JRSS-B 2022)

- Associate each H_j with e-value e_j.
- 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

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

- 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.

A problem

Testing a single hypothesis: no problem

- One e-process: $E_1, E_2, \ldots$
- Suppose we already gathered E_{t+1} before properly evaluating E_t.
- If $E_t \geq 1/\alpha$ but $E_{t+1} < 1/\alpha$, we can still go back to E_t and reject H_0, because Ville's inequality states

$$\sup_{P \in \mathcal{P}} \left(\sup_{t \in \mathbb{N}} E_t \geq \frac{1}{\alpha} \right) \leq \alpha$$

Testing multiple hypotheses: a problem (1)

- Let's focus on e-BH for FDR control.
- Consider K parallel e-processes $E_t^1, \ldots, E_t^K$
- e-BH guarantees: $\sup_t \text{FDR}(E_t^1, \ldots, E_t^K) \leq \alpha$

Testing multiple hypotheses: a problem (2)

- Consider $K = 2$, then e-BH rejects any hypothesis for which the e-value exceeds $2/\alpha$, and both hypotheses if both e-values exceed $1/\alpha$.

Testing multiple hypotheses: a problem (2)

- Consider $K = 2$, then e-BH rejects any hypothesis for which the e-value exceeds $2/\alpha$, and both hypotheses if both e-values exceed $1/\alpha$.
- Suppose that at time t we have $1/\alpha \leq E_t^1 < 2/\alpha$ and $E_t^2 < 1/\alpha$.

Testing multiple hypotheses: a problem (2)

- Consider $K = 2$, then e-BH rejects any hypothesis for which the e-value exceeds $2/\alpha$, and both hypotheses if both e-values exceed $1/\alpha$.
- Suppose that at time t we have $1/\alpha \leq E_t^1 < 2/\alpha$ and $E_t^2 < 1/\alpha$.
- Gathering more data for both e-processes, we could arrive at

  $E_{t+1}^1 < 1/\alpha$ and $1/\alpha \leq E_{t+1}^2 < 2/\alpha$.

Testing multiple hypotheses: a problem (2)

- Consider $K = 2$, then e-BH rejects any hypothesis for which the e-value exceeds $2/\alpha$, and both hypotheses if both e-values exceed $1/\alpha$.
- Suppose that at time t we have $1/\alpha \leq E_t^1 < 2/\alpha$ and $E_t^2 < 1/\alpha$.
- Gathering more data for both e-processes, we could arrive at

  $E_{t+1}^1 < 1/\alpha$ and $1/\alpha \leq E_{t+1}^2 < 2/\alpha$.

- Now, if we would not have gathered one more data point for E^1, then we would have rejected both hypotheses.

FDR control for e-processes

- e-BH guarantees: $\sup_t \text{FDR}\left(E_t^1, \ldots, E_t^K\right) \leq \alpha$
- But we would like to control:

  $$\text{FDR}\left(\sup_t E_t^1, \ldots, \sup_t E_t^K \right) \leq \alpha$$

- We would like to argue that this is the proper FDR criterion for e-processes.
- *(Note: in this talk I focus on FDR with e-BH, but we can show the same for other methods and FWER.)*

A proposition

Working with the running maxima of e-processes

- Clearly, if we work with $M_t^k := \max_{1 \le s \le t} E_s^k$, the running maximum processes,

  we obtain a non-decreasing sequence of sets of rejected hypotheses.

Working with the running maxima of e-processes

- Clearly, if we work with $M_t^k := \max_{1 \le s \le t} E_s^k$, the running maximum processes,

  we obtain a non-decreasing sequence of sets of rejected hypotheses.

- The running maximum process is not an e-process itself, therefore we cannot straightforwardly prove FDR control.

Working with the running maxima of e-processes

- Clearly, if we work with $M_t^k := \max_{1 \le s \le t} E_s^k$, the running maximum processes,

  we obtain a non-decreasing sequence of sets of rejected hypotheses.

- The running maximum process is not an e-process itself, therefore we cannot straightforwardly prove FDR control.

- In case the e-processes are independent, $1/M_t^k$ is a valid p-value (p-process), thus, we obtain FDR control via classical BH.

Proposition

(Tavyrikov, Goeman, De Heide, ’25)

- **Proposition**

  The cumulative maximum e-BH procedure applied to arbitrarily dependent e-processes does not control the FDR at level α.

- Proof: example on the next slide.

Proof

$$E_t^1 = X_0^1 \prod_{s=1}^{t} e_s^1, \quad E_t^2 = X_0^2 \prod_{s=1}^{t} e_s^2$$

$$(X_0^1, X_0^2) \sim \begin{cases} (0,0), & \text{with probability } 1-2\alpha, \\ \left(\frac{1}{2\alpha}, \frac{1}{2\alpha}\right), & \text{with probability } 2\alpha, \end{cases}$$

and the increments (e_s^1, e_s^2) are distributed as

$$\begin{cases} \left(\frac{1}{2}, \frac{1}{2}\right), & \text{with probability } \frac{1}{3}, \\ \left(2, \frac{1}{2}\right), & \text{with probability } \frac{1}{3}, \\ \left(\frac{1}{2}, 2\right), & \text{with probability } \frac{1}{3}. \end{cases}$$

Proof

$$E_t^1 = X_0^1 \prod_{s=1}^{t} e_s^1, \quad E_t^2 = X_0^2 \prod_{s=1}^{t} e_s^2$$

$$(X_0^1, X_0^2) \sim \begin{cases} (0,0), & \text{with probability } 1-2\alpha, \\ \left(\frac{1}{2\alpha}, \frac{1}{2\alpha}\right), & \text{with probability } 2\alpha, \end{cases}$$

and the increments (e_s^1, e_s^2) are distributed as

$$\begin{cases} \left(\frac{1}{2}, \frac{1}{2}\right), & \text{with probability } \frac{1}{3}, \\ \left(2, \frac{1}{2}\right), & \text{with probability } \frac{1}{3}, \\ \left(\frac{1}{2}, 2\right), & \text{with probability } \frac{1}{3}. \end{cases}$$

These are valid e-processes, i.e. $\mathbb{E}[E_t^1] = \mathbb{E}[E_t^2] = 1$ at all times.

They are not independent.

Rejection happens for the event

$$E_t^1 \geq \frac{2}{\alpha}, \quad E_t^2 \geq \frac{2}{\alpha}, \quad \text{or both } E_t^1, E_t^2 \geq \frac{1}{\alpha}$$

Simulations show that the FDR of the running maxima of these processes is 1.08α.

A solution

Adjusters

- An (admissible) adjuster is an increasing function $A : [1,\infty] \to [1,\infty]$ that is right-continuous, $A(\infty) = \lim_{e\to\infty} A(e) = \infty$ and $\int_1^\infty A(e)/e^2 \mathrm{d}e = 1$.

Adjusters

- An (admissible) adjuster is an increasing function $A : [1,\infty] \to [1,\infty]$ that is right-continuous, $A(\infty) = \lim_{e\to\infty} A(e) = \infty$ and $\int_1^\infty A(e)/e^2 \mathrm{d}e = 1$.
- If $M_t^k := \sup_{1\le s\le t} E_s^k$ is the running maximum of an e-process, then

  $(A(M_t^k))_{t\ge 0}$ is also an e-process.

Adjusters

- An (admissible) adjuster is an increasing function $A : [1,\infty] \to [1,\infty]$ that is right-continuous, $A(\infty) = \lim_{e\to\infty} A(e) = \infty$ and $\int_1^\infty A(e)/e^2 \mathrm{d}e = 1$.
- If $M_t^k := \sup_{1\le s\le t} E_s^k$ is the running maximum of an e-process, then

  $(A(M_t^k))_{t\ge 0}$ is also an e-process.
- Since the adjusted process is non-decreasing, it will never result in regretting having collected more evidence.

The adjusted running maximum process has FDR-sup $\leq \pi_0 \alpha$ (Tavyrikov, Goeman, De Heide, ’25)

Let $\mathscr{R}^{\text{e-BH}}(\mathbf{e})$ denote the e-BH rejection set based on a set of e-values $\mathbf{e} = e^1_{t_1}, \ldots, e^K_{t_K}$, which can be obtained from different time points $t_1, \ldots, t_K$.

I.e., $j \in \mathscr{R}^{\text{e-BH}}(\mathbf{e})$ if and only if $e^j_{t_j} > m/(\alpha \,|\, \mathscr{R}^{\text{e-BH}}(\mathbf{e}) \,|\,)$. Let A be an adjuster, so that $A\left(\sup_{t \leq s}(e^j_t)\right)$ is an adjusted running maximum process.

The adjusted running maximum process has FDR-sup $\leq \pi_0 \alpha$ (Tavyrikov, Goeman, De Heide, '25)

Then the FDR-sup at time s is

$$\text{FDR-sup}_s = \sum_{j \in H_0} \frac{\mathbf{1}\left\{ A\left(\sup_{t \leq s}(e_t^j)\right) \geq \frac{K}{\alpha\, |\mathscr{R}^{\text{e-BH}(\mathbf{e})}|} \right\}}{|\mathscr{R}^{\text{e-BH}(\mathbf{e})}| \vee 1}$$

$$\leq \sum_{j \in H_0} \mathbb{E}\left[A\left(\sup_{t \leq s}(e_t^j)\right) \frac{\alpha \frac{|\mathscr{R}^{\text{e-BH}(\mathbf{e})}|}{K}}{|\mathscr{R}^{\text{e-BH}(\mathbf{e})}| \vee 1} \right] \leq \frac{\alpha}{K} \sum_{j \in H_0} \mathbb{E}\left[A\left(\sup_{t \leq s}(e_t^j)\right) \right] \leq \frac{K_0}{K}\alpha\,.$$

Do we *need* adjusters?

- **Theorem (informal)**: The function is *necessarily* an adjuster, as long as it is an *increasing* function that maps the running maximum process to an e-process.

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.

Exciting new result: bringing closure to FDR

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

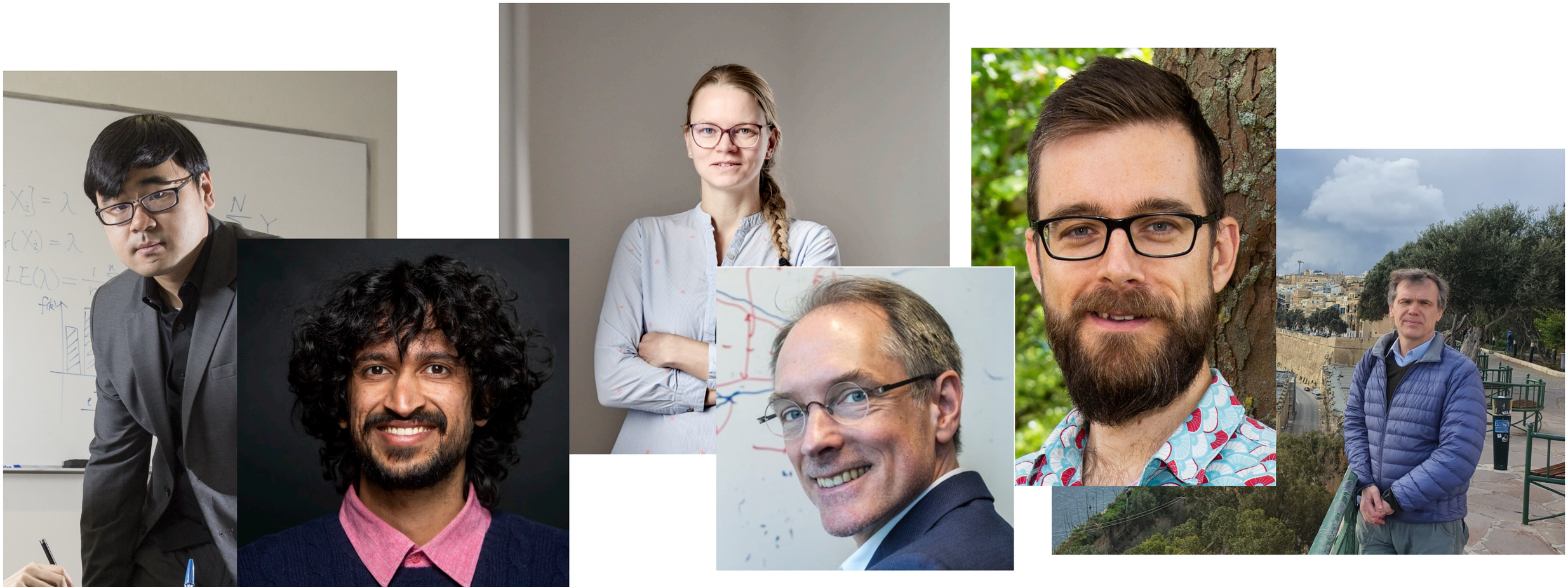
- Necessary and sufficient principle for multiple testing methods controlling an expected loss (think of FDR)
- Every such multiple testing method is a special case of a general closed testing procedure based on e-values.
- Uniform improvements of these methods
- Simultaneous error control
- Post-hoc 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