A general principle for multiple testing

Statistics Seminar, 20 November 2025

Rianne de Heide, University of Twente and Centrum Wiskunde & Informatica Amsterdam

About

- Associate professor
- University of Twente (0.8)
- Centrum Wiskunde & Informatica Amsterdam (0.2)
- Research:
  e-values, multiple testing, bandits, trials, Bayesian learning, information-theoretic learning, foundations.
- Grants: NWO VENI, NWO M2

Group: Aurèle, Fabian, Rovanos, Yury

De e...
zebra-voo...
zebra.cls
File outline
Het toetse...
De hyp...
Type I ...
Type II...
Termin...
Hypothese...
Fisher'...
Interm...
De e-waarde
Hypothese...
Leuke ...
Wat is een ...
Een experime...
De toekomst v...

Editing

```
598     0.2 \text{ als } X=0 \\
599     0.6 \text{ als } X=1\\
600     0.2 \text{ als } X=2
601     \end{cases}
602     \]
603     Is $X$ dan een e-waarde? Daarvoor moeten we drie
        dingen checken:
604     \begin{enumerate}
605         \item Is $X$ een toevalsvariable? Ja: het is
            een functie van uitkomsten (data), en de
            output is een re\"eel getal, want dat kan
            alleen $0, 1$ of $2$ zijn.
606         \item Is $X$ niet-negatief? Ja, want de
            output kan alleen $0, 1$ of $2$ zijn.
607         \item Is de verwachting onder elke
            kansverdeling in de nulhypothese ten hoogste
            $1$? Er is maar $1$ kansverdeling in de
            nulhypothese: $P$. Laten we de verwachting
            van $X$ onder $P$ berekenen:
608         \[
609         \mathbb{E}_P [X] = \sum_{i \in \{1, 2, 3\}
            }x_i\cdot p_i = 0 \cdot 0.2 + 1 \cdot 0.6 + 2
            \cdot 0.2 = 1.
610         \]
611         Dat is niet groter dan $1$, dus $X$ is
            inderdaad een e-waarde voor deze nulhypothese!
612     \end{enumerate}
613 \end{voorbeeld}
614
615 \begin{opgave}
616     Stel we gooien een munt met kans $q$ op kop en
        kans $1-q$ op munt. Definieer de toevalsvariabele
        $X(\text{kop}) = 2$ en $X(\text{munt}) = 0$. Voor
        welke $q$ is $X$ een e-waarde?
617 \end{opgave}
618
619 \begin{opgave}
620     Stel we gooien een dobbelsteen met uitkomsten $X
        \in \{1, 2, 3, 4, 5, 6 \}$ met kans $p_1$ op $1$,
```

Hoofdstuk 3

De e-waarde

We willen een nulhypothese $\mathcal{H}_0$ testen, soms (maar niet noodzakelijk) ten opzichte van een alternatieve hypothese $\mathcal{H}_1$.

Definitie 3.1 (E-waarde)**.** Een *e-waarde* is een niet-negatieve toevalsvariabele E waarvoor geldt:

$$\text{voor alle } P \in \mathcal{H}_0 : \mathbb{E}_P[E] \leq 1. \tag{2}$$

Laten we kijken wat dit precies betekent. De e-waarde is een toevalsvariabele: zoals in hoofdstuk 2 beschreven is het dus een functie van uitkomsten, en in onze statistische toepassing is dat: de data. Een e-waarde is dus een functie van de data, en heeft als output een niet-negatief reëel getal, dus dat is 0 of een positief getal.

De verwachting onder P van de e-waarde is ten hoogste 1. En omdat de e-waarde niet-negatief is, moet de verwachting dus in het interval $[0, 1]$ liggen. Wat betekent nou 'de verwachting onder P'? Zoals beschreven in hoofdstuk 2, is een verwachting een *kansgewogen gemiddelde*. Welke kansen moeten we nemen om de uitkomsten te wegen? Dat wordt beschreven in de kansverdeling P. Een (nul)hypothese is niets meer dan een verzameling kansverdelingen, en als het gemiddelde van de toevalsvariable E ten hoogste 1 is, als het is gewogen met elke kansverdeling in de nulhypothese, dan is het een e-waarde.

Voorbeeld 3.2. Stel we hebben een toevalsvariabele X die data als input neemt en de volgende waarden aanneemt: $\{0, 1, 2\}$. Stel we hebben 1 kansverdeling in de de nulhypothese, P, die de volgende kansen toekent aan X:

$$P(X) = \begin{cases} 0.2 \text{ als } X = 0 \\ 0.6 \text{ als } X = 1 \\ 0.2 \text{ als } X = 2 \end{cases}$$

Is X dan een e-waarde? Daarvoor moeten we drie dingen checken:

(1) Is X een toevalsvariable? Ja: het is een functie van uitkomsten (data), en de output is een reëel getal, want dat kan alleen $0, 1$ of 2 zijn.
(2) Is X niet-negatief? Ja, want de output kan alleen $0, 1$ of 2 zijn.

Other things about me

The e-value

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 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.

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 nonnegative 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?

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”.

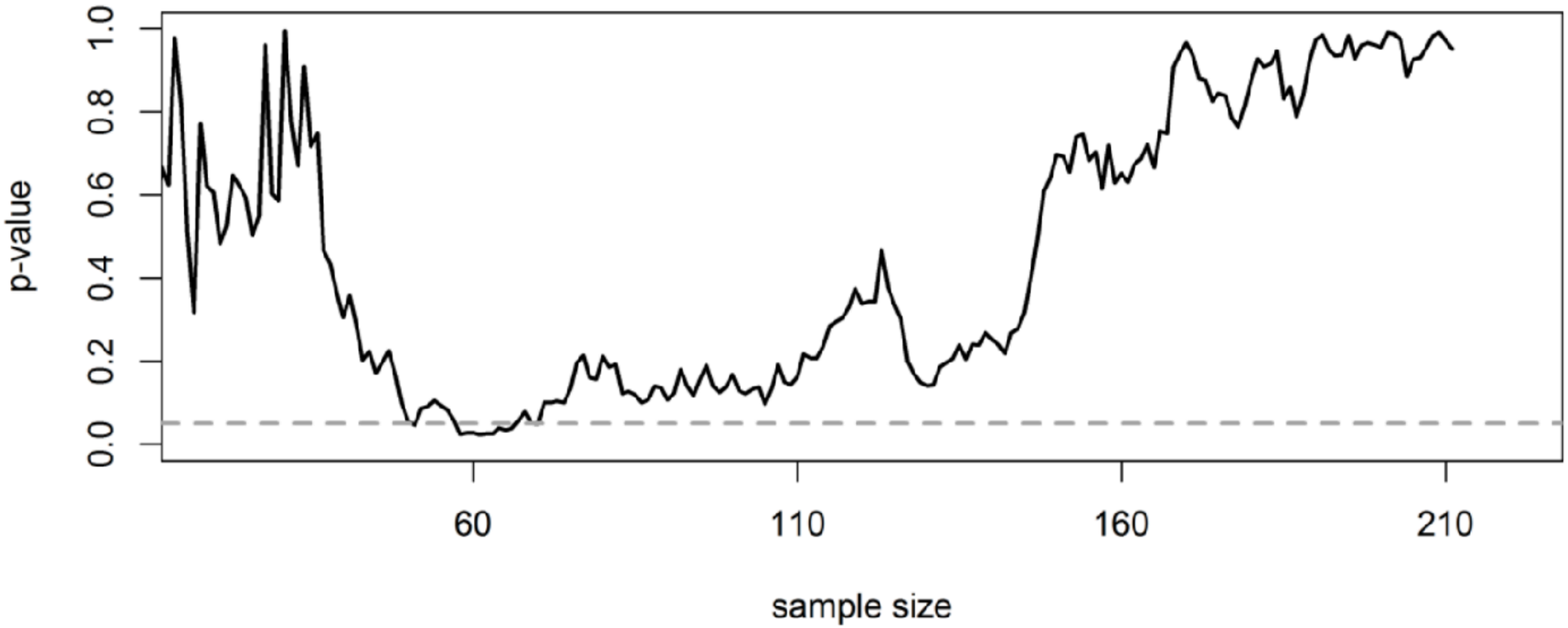
- This is called optional stopping, and invalidates p-values and their error guarantees

Type I error guarantee

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

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

Stopping rules and p-values



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

E-value history

- 2019: first two papers about e-values around the same time on ArXiv:

  Safe Testing - Grünwald, De Heide & Koolen

  Testing by Betting - Glenn Shafer

E-value history

- 2019: first two papers around the same time about e-values on ArXiv:

  Safe Testing - Grünwald, De Heide & Koolen

  Testing by Betting - Glenn Shafer

- Theory on hypothesis testing with e-values, and theory on “what is a good e-value?”

E-value history

- 2019: first two papers around the same time about e-values on ArXiv:

  Safe Testing - Grünwald, De Heide & Koolen

  Testing by Betting - Glenn Shafer

- Theory on hypothesis testing with e-values, and theory on “what is a good e-value?”
- Now: 100s of papers on e-values and many groups, mostly in mathematical statistics, around the world (Stanford, CMU, ETH Zurich, CWI, etc.) investigating e-values. Many grants, prizes, and international recognition :)

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 $P \in \mathscr{P}$.

- **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.

Safe Testing: 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?

Safe Testing: 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 with e-values: Main Theorem

- 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

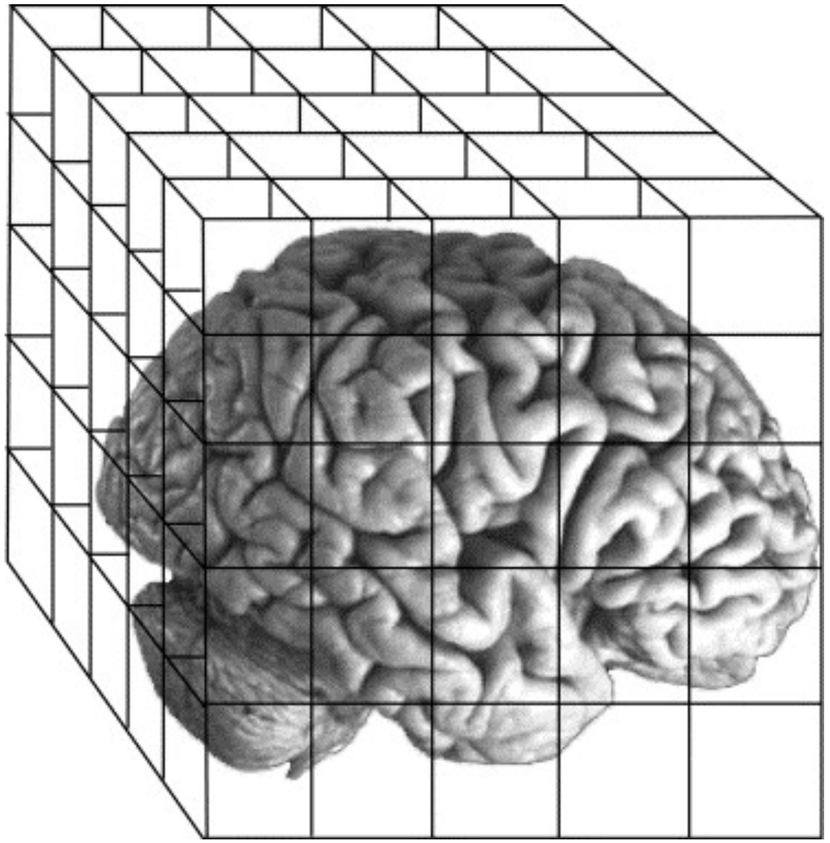
Bringing closure to FDR control: a general principle for multiple testing

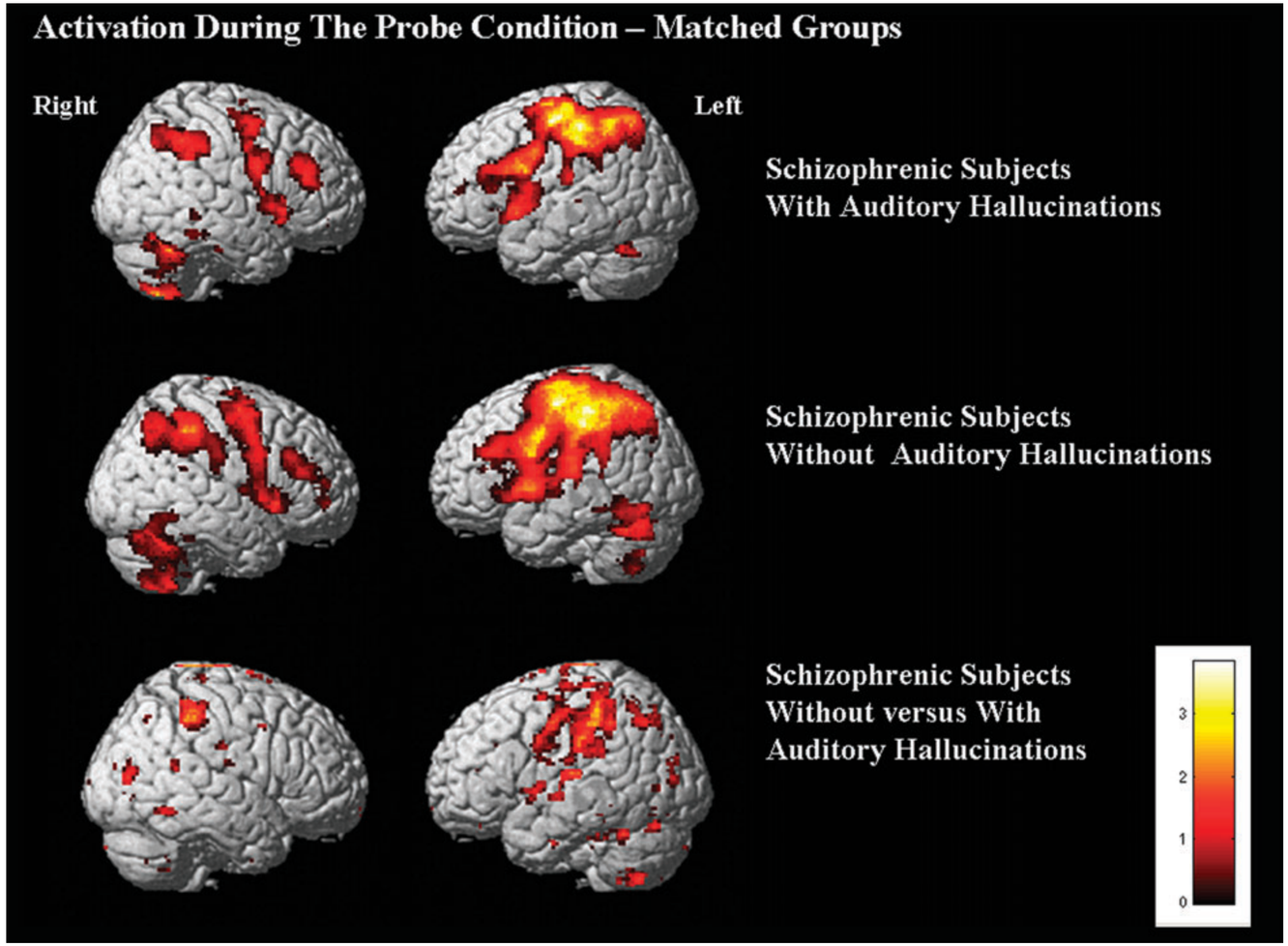
Ziyu Xu, Aldo Solari, Lasse Fischer, Rianne de Heide, Aaditya Ramdas and Jelle Goeman

https://arxiv.org/pdf/2509.02517

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 .

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.

Error rates

$N \subseteq [m]$ hypotheses are true null; the rest are potential discoveries

Famous error rates:

- Familywise error rate (FWER): $P(|R \cap N| > 0)$
- Per-family error rate: $\mathbb{E}(|R \cap N|)$
- False Discovery rate (FDR): $\mathbb{E}\left(\frac{|R \cap N|}{R}\right)$

General form

Control some expected loss: $\mathbb{E}(f_N(R))$

FWER history

- Closure principle: necessary and sufficient for the construction of valid methods (Sonnemann, 1982, 2008)

FWER history

- Closure principle: necessary and sufficient for the construction of valid methods (Sonnemann, 1982, 2008)
- Any method that controls FWER is a special case of a closed testing procedure

FWER history

- Closure principle: necessary and sufficient for the construction of valid methods (Sonnemann, 1982, 2008)
- Any method that controls FWER is a special case of a closed testing procedure
- Challenged by the Partitioning Principle (Finner and Strassburger, 2002)

FWER history

- Closure principle: necessary and sufficient for the construction of valid methods (Sonnemann, 1982, 2008)
- Any method that controls FWER is a special case of a closed testing procedure
- Challenged by the Partitioning Principle (Finner and Strassburger, 2002)
- They are equivalent: Goeman et al. (2021)

False Discovery Proportion (FDP) history

- Genovese and Wasserman (2006) and Goeman and Solari (2011) have extended closed testing to control of false discovery proportions (FDPs)

False Discovery Proportion (FDP) history

- Genovese and Wasserman (2006) and Goeman and Solari (2011) have extended closed testing to control of false discovery proportions (FDPs)

- Goeman et al. (2021) showed that all methods controlling a quantile of the distribution of FDP are either equivalent to a closed testing procedure or are dominated by one, extending the Closure Principle to all methods controlling FDP.

Why is the closure principle nice?

- reduces the complex task of constructing a multiple testing method to the simpler task of choosing hypothesis tests for intersection hypotheses

Why is the closure principle nice?

- reduces the complex task of constructing a multiple testing method to the simpler task of choosing hypothesis tests for intersection hypotheses
- helps to handle complex situations such as restricted combinations

Why is the closure principle nice?

- reduces the complex task of constructing a multiple testing method to the simpler task of choosing hypothesis tests for intersection hypotheses
- helps to handle complex situations such as restricted combinations
- methods constructed using closed testing often allow for some user flexibility, permitting researchers to modify some aspects of the multiple testing procedure post hoc without compromising error control

FDR history

- Blanchard and Roquain (2008) formulated two quite general sufficient conditions, self-consistency and dependence control, under which, if both hold, FDR control is guaranteed.

FDR history

- Blanchard and Roquain (2008) formulated two quite general sufficient conditions, self-consistency and dependence control, under which, if both hold, FDR control is guaranteed.

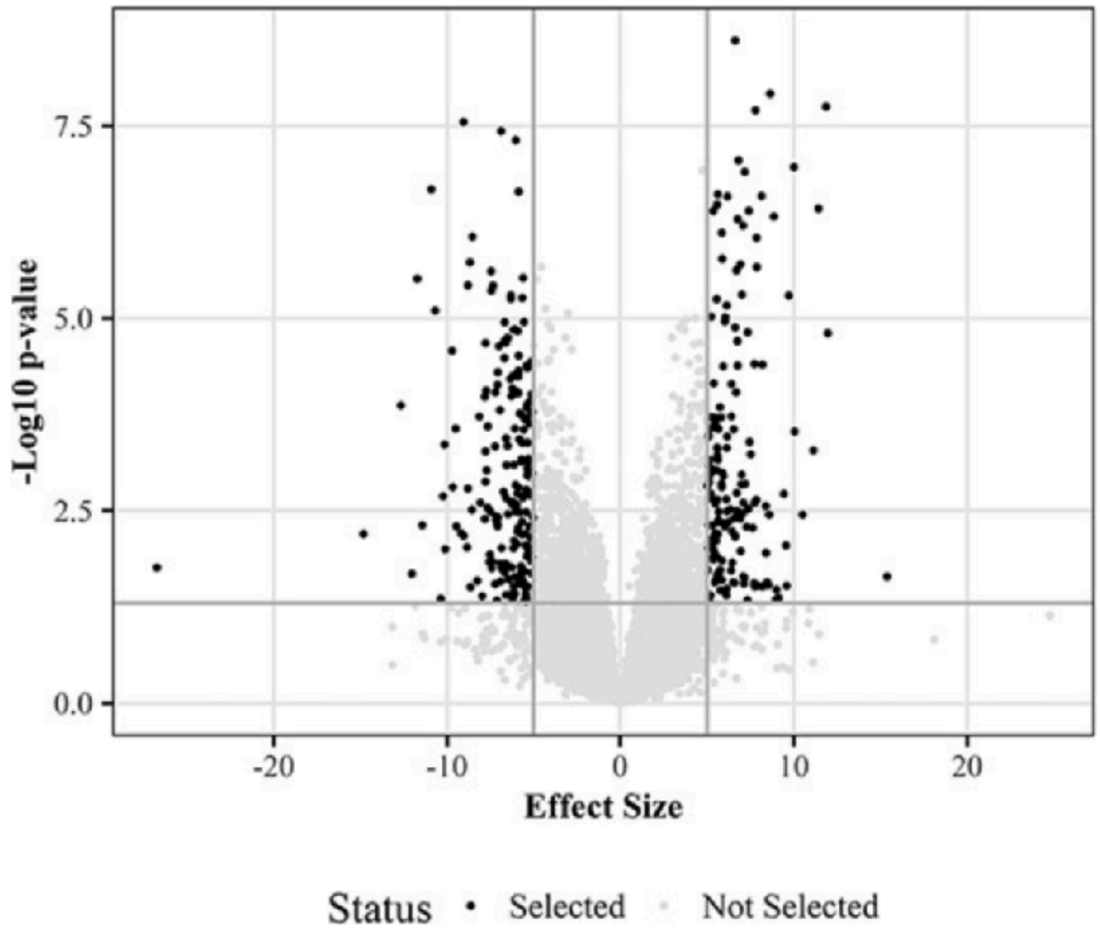
- However, self-consistency is sufficient but not necessary for FDR control: Solari and Goeman (2017) show uniform improvements of self-consistent methods by a non-self-consistent method.

FDR history

- Blanchard and Roquain (2008) formulated two quite general sufficient conditions, self-consistency and dependence control, under which, if both hold, FDR control is guaranteed.
- However, self-consistency is sufficient but not necessary for FDR control: Solari and Goeman (2017) show uniform improvements of self-consistent methods by a non-self-consistent method.
- No post-hoc user flexibility. Why is that problematic? Example on the next slide.

FDR control and volcano plots: no guarantees!

Ebrahimpoor & Goeman (2021)



The e-closure principle

How to design a multiple testing procedure

- e-Closure
  A general recipe for making multiple testing methods

- Building blocks
  Intersection hypotheses and e-values

- Contributions

  - Recovers the Closure Principle for FWER
  - Extends to FDR
  - Uniformly improves a.o. eBH and BY
  - Introduces unprecedented flexibility in multiple testing

Intersection hypotheses

Intersection hypothesis

For $S \subseteq [m], H_S = \bigcap_{i \in S} H_i$, which is true iff all $H_i, i \in S$ true

The e-collection

$E = (e_S)_{S \subseteq [m]}$: local e-values such that $\mathsf{E}(e_N) \leq 1$

Sufficient

Each e_S is an e-value for $H_S, S \subseteq [m]$

The e-Closure Principle

- The e-Closed Procedure

$$\mathscr{R}_\alpha(E) = \left\{R \subseteq [m] : \alpha e_S \geq \mathrm{f}_S(R) \quad \forall S \subseteq [m]\right\}$$

- The e-Closure Principle

R controls $\mathrm{E}(\mathrm{f}_N(R)) \leq \alpha$ iff $R \in \mathscr{R}_\alpha(E)$ for e-collection E

- Simultaneous control

$\mathrm{E}(\mathrm{f}_N(R)) \leq \alpha$ simultaneously over $R \in \mathscr{R}_\alpha(E)$: $\quad \mathrm{E}\left(\max_{R \in \mathscr{R}_\alpha(E)} \mathrm{f}_N(R) \right) \leq \alpha$

Post hoc error rate

- All error rates

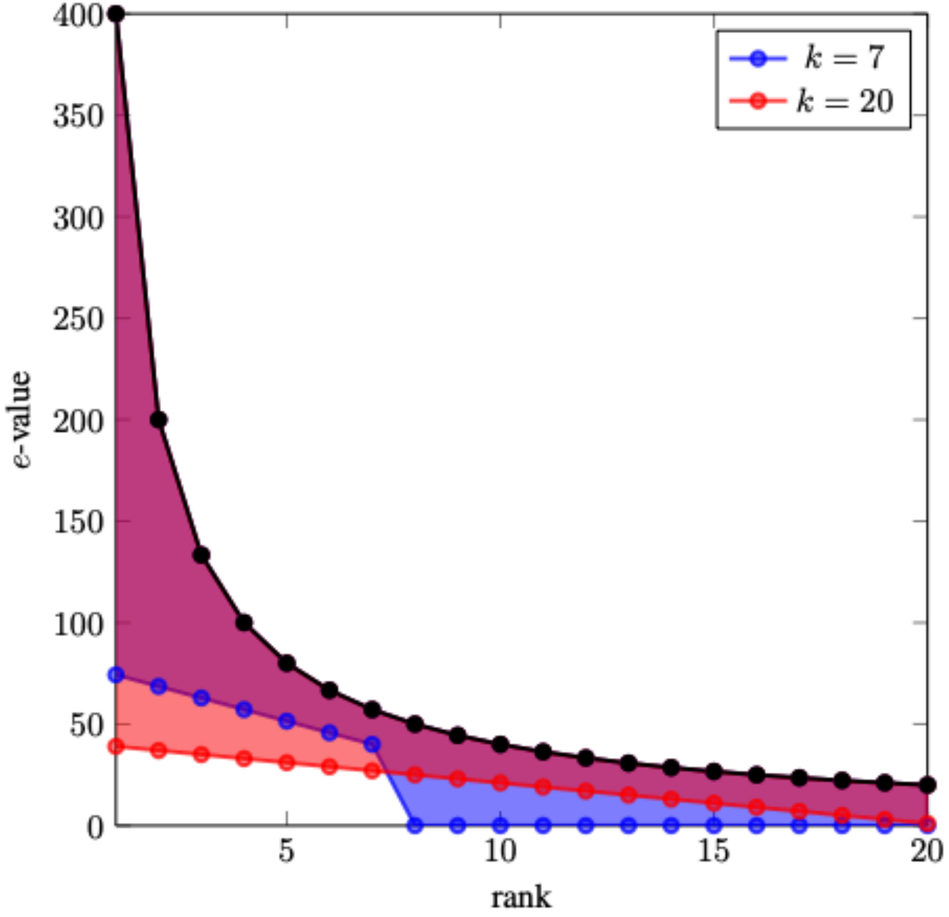
  $\mathscr{F} = \{\text{all functions } \mathrm{f}_N(R)\}$

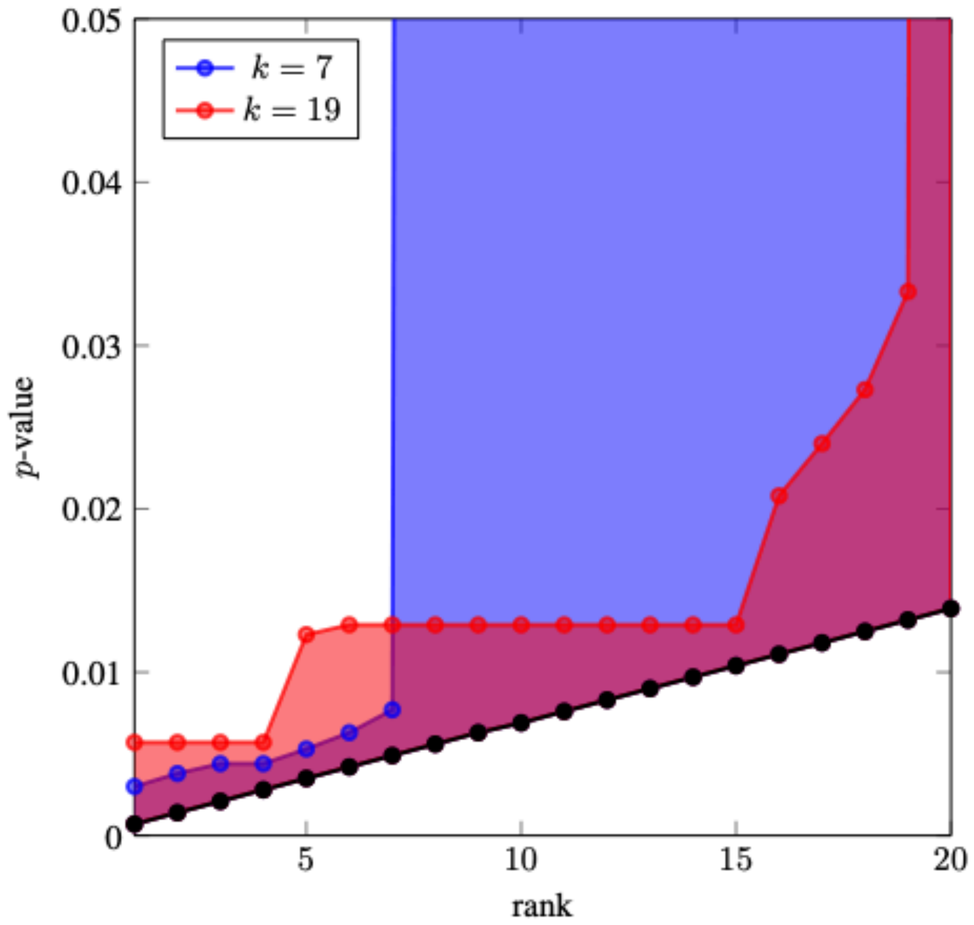
- Simultaneous (= post hoc) choice of error

  $$\mathrm{E}\left(\sup_{\mathrm{f}\in\mathscr{F}} \max_{R\in\mathscr{R}_\alpha^{\mathrm{f}}(E)} \mathrm{f}_N(R) \right) \leq \alpha$$

- So: possible to switch from FWER to FDR if not much signal present

Improving existing procedures: eBH and BY





BY vs closed BY in standard data sets

Dataset	m	BY / $\overline{BY}$ rejections		source
		$\alpha = 5\%$	$\alpha = 10\%$	
APSAC	15	3 / 3	3 / 5	BH '95
NAEP	34	6 / 8	8 / 11	BH '00
PADJUST	50	12 / 15	17 / 20	`p.adjust`
PVALUES	4289	129 / 145	225 / 275	`fdrtool`
VANDEVIJVER	4919	614 / 677	779 / 866	Goeman Solari '14
GOLUB	7128	617 / 648	743 / 799	Efron Hastie '16

More properties: post hoc α

- Choose rejected set post hoc
- Choose error loss post hoc
- One step further: choose α post hoc (Koning 2023)
- Requires: e-value does not depend on α. Then we have:

$$\mathrm{E}\left(\sup_{\alpha \in (0,1)} \sup_{\mathrm{f} \in \mathscr{F}} \max_{R \in \mathscr{R}_\alpha^{\mathrm{f}}(E)} \frac{\mathrm{f}_N(R)}{\alpha} \right) \leq 1$$

More properties: restricted combinations

- Logically related hypotheses, for example pairwise combinations

$$H_{1=2} : \mu_1 = \mu_2; \quad H_{1=3} : \mu_1 = \mu_3; \quad H_{2=3} : \mu_2 = \mu_3$$

- Logical relationships = gain in power

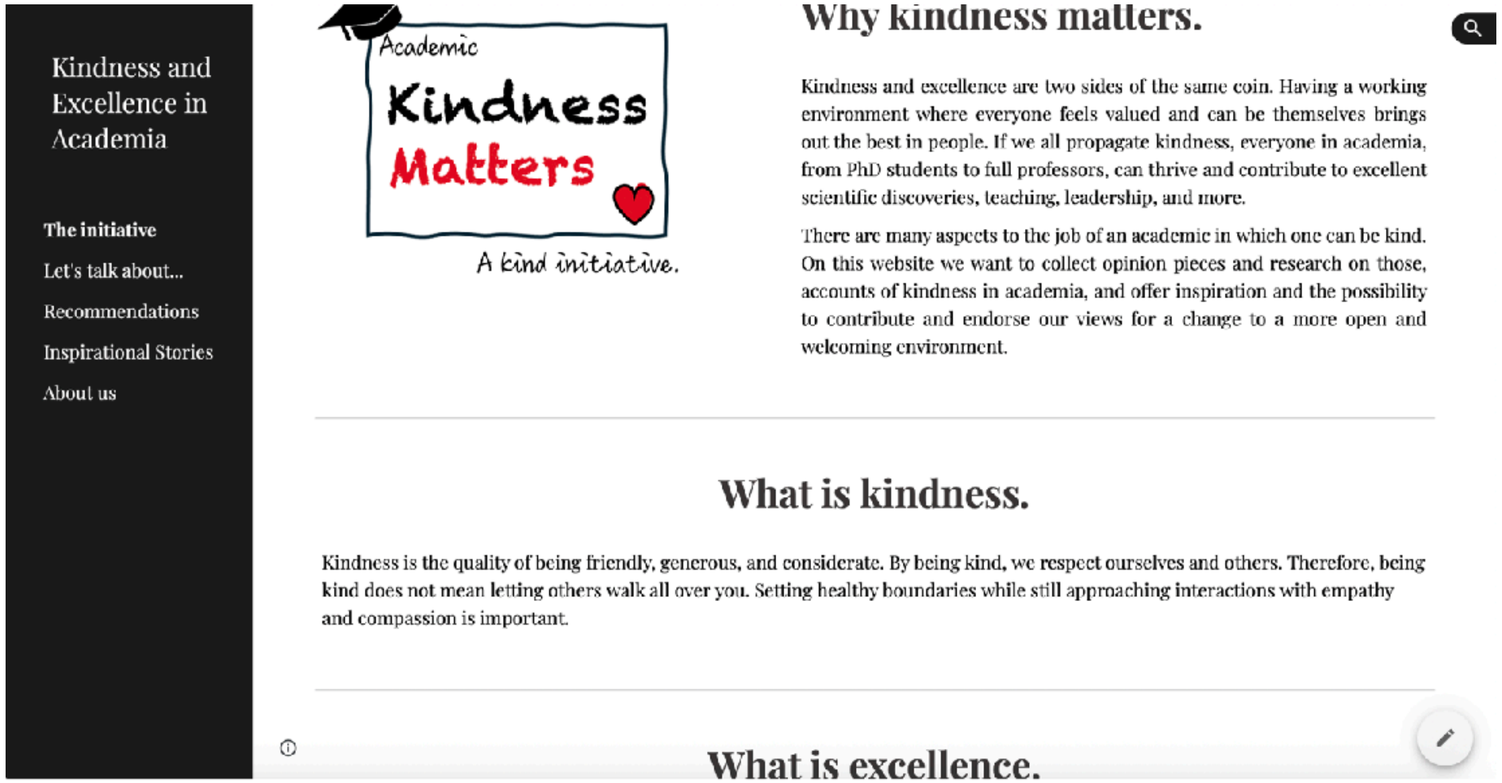
  Up to now only known for FWER, unknown for FDR

Summary: e-Closure

- General Necessary and Sufficient Principle: unites all multiple testing methods
- Simplifies multiple testing: Choose how to summarise evidence against H_S; rest is computation
- Flexibility: Simultaneous over rejected sets, error rates, α
- Power: Uniformly improves known methods

A general recipe for making multiple testing methods

Kindness and excellence in academia



- Research shows that having a working environment where everyone feels valued and can be themselves leads not only to happy employees, but also to higher quality and quantity of production.
- Join us! https://sites.google.com/view/kindness-and-excellence/the-initiative