Controlling the False Discovery Rate (Benjamini & Hochberg, 1995)

Aditya Ghosh & Michael Salerno

Presenting for Stats 319 (Journal Club)

January 29, 2024

Acknowledgment: Our presentation is largely inspired by the 300C lectures by Prof. Emmanuel Candès

A review of the hypothesis testing framework

Data are sampled from some distribution paramaterized by θ:

- $X \sim \mathbb{P}_\theta$
- $\theta \in \Omega$

Furthermore, the parameter space Ω can be split into disjoint subclasses known as "hypotheses":

$$H_0 : \theta \in \Omega_0 \subset \Omega \qquad \text{(null hypothesis)}$$
$$H_1 : \theta \in \Omega_1 = \Omega \setminus \Omega_0 \qquad \text{(alternative hypothesis)}$$

Our goal is to infer which hypothesis is correct.

	Reject H_0	Retain H_0
$\theta \in \Omega_0$	Type I error	Good
$\theta \in \Omega_1$	Good	Type II error

- **Level of significance**: A level-α test guarantees that $\mathbb{P}_{H_0}(\text{Type I error}) \leq \alpha$.
- **power** $= 1 - \mathbb{P}_{H_1}(\text{Type II error})$

Under the Neyman-Pearson paradigm, a test procedure maximizes the power subject to the level of significance. **The only guarantee is a type I error rate less than** α.

Typically, θ is an unobservable state of the universe which interests us, and H_0 represents our default state of belief:

- H_0: Male and female births are equally likely.
- H_0: No difference in expected blood pressure after treatment.
- H_0: The true regression coefficient β_1 is zero.

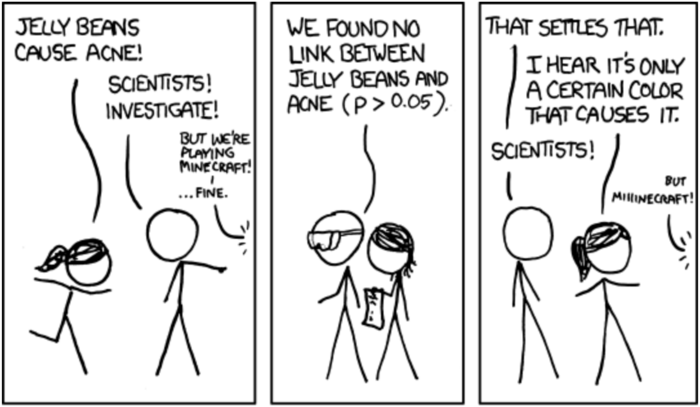
But what if we perform multiple tests?

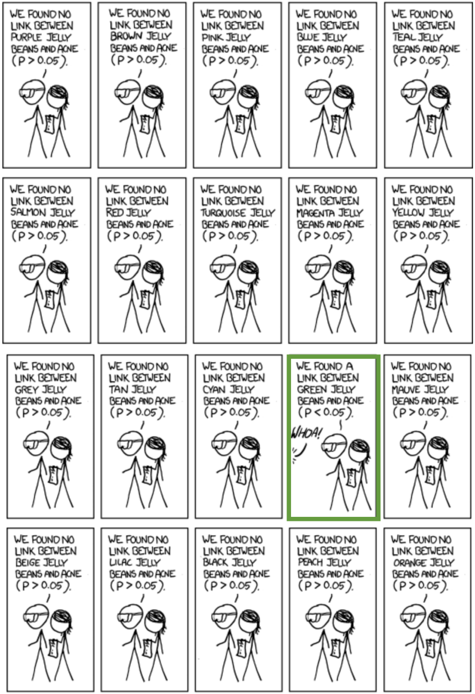
- H_{0j}: No difference in expected blood pressure after treatment j.
- H_{0j}: The true *jth* regression coefficient β_j is zero.

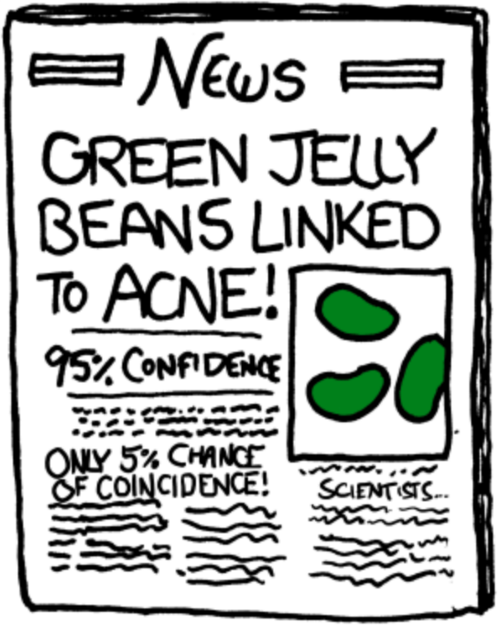
Recall that we only control the type I error rate, typically at level $\alpha = 0.05$.

What does this mean for the state of science?

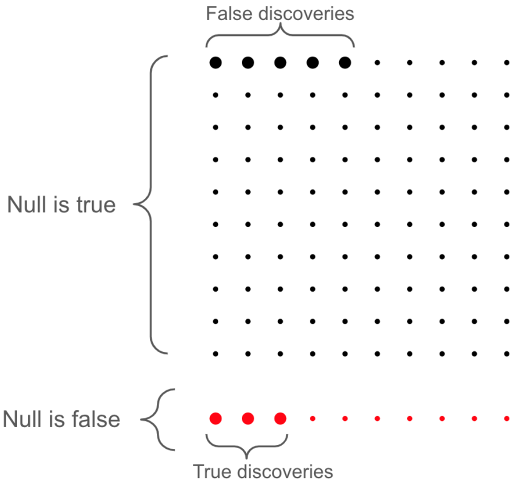
Testing multiple hypotheses


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
...FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!


WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).

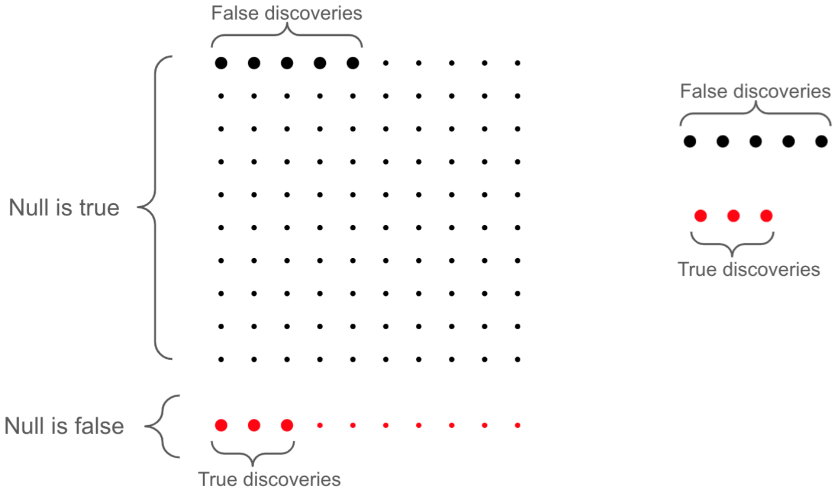

NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

Traditional type-I error control

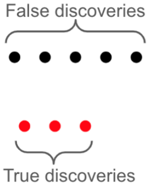


- Each dot represents a hypothesis being tested. A bold dot represents rejecting the null hypothesis (declaring a discovery).
- We imagine an army of scientists all around the world, all testing their own hypotheses.
- Type-I error control says: out of all the dots, all the hypotheses tested around the world, at most 5% are *bold black dots* (false discoveries).
- But if only the discoveries are published, we don't get to see all the dots!

The file-drawer effect: what do we actually see?



FDR lets us control the proportion of false discoveries out of all *discoveries*, not out of all *hypotheses tested*



FDR control puts an upper bound on

$$E\left(\frac{\text{false discoveries}}{\text{false discoveries} + \text{true discoveries}}\right)$$

Formal introduction to FDR control

	declared non-signif.	declared significant	Total
H_0 true	U	V	n_0
H_0 false	T	S	$n - n_0$
	$n - R$	R	n

- Familywise error rate (FWER) $= \mathbb{P}(V \geq 1)$
- False discovery proportion (FDP):

$$\mathrm{FDP} = \frac{V}{\max(R, 1)} = \begin{cases} V/R & \text{if } R \geq 1 \\ 0 & \text{if } R = 0 \end{cases}$$

- False discovery rate (FDR) $= \mathbb{E}[\mathrm{FDP}]$

Connections

- If all the hypotheses are true, then

$$\text{FDR control} \equiv \text{FWER control}$$

Connections

- If all the hypotheses are true, then

$$\text{FDR control} \equiv \text{FWER control}$$

- Any procedure that controls the FWER must also control the FDR (since $FDP = 0$ when $R = 0$ and $FDP \leq 1$ when $R \geq 1$)

Connections

- If all the hypotheses are true, then

$$\text{FDR control} \equiv \text{FWER control}$$

- Any procedure that controls the FWER must also control the FDR (since $FDP = 0$ when $R = 0$ and $FDP \leq 1$ when $R \geq 1$)

Control FDR instead of controlling FWER?

FWER vs. FDR (contd.)

- Small # hypotheses $\rightarrow$ FWER control ✓ (but, may lack power)
- Large-scale studies $\rightarrow$ FWER control may miss important findings

FWER vs. FDR (contd.)

- Small # hypotheses $\rightarrow$ FWER control $\checkmark$ (but, may lack power)
- Large-scale studies $\rightarrow$ FWER control may miss important findings
- FDR control sacrifices some stringency to permit exploration with a few false positives

FWER vs. FDR (contd.)

- Small # hypotheses $\rightarrow$ FWER control ✓ (but, may lack power)
- Large-scale studies $\rightarrow$ FWER control may miss important findings
- FDR control sacrifices some stringency to permit exploration with a few false positives
- FDR control does not assure a specific study, but ensures that science as a whole will be alright!

The BH procedure

The BH procedure

- Say we want to control the FDR at level α

The BH procedure

- Say we want to control the FDR at level α
- Compute p-values $p_1, \ldots, p_n$ for the n hypotheses $H_1, \ldots, H_n$

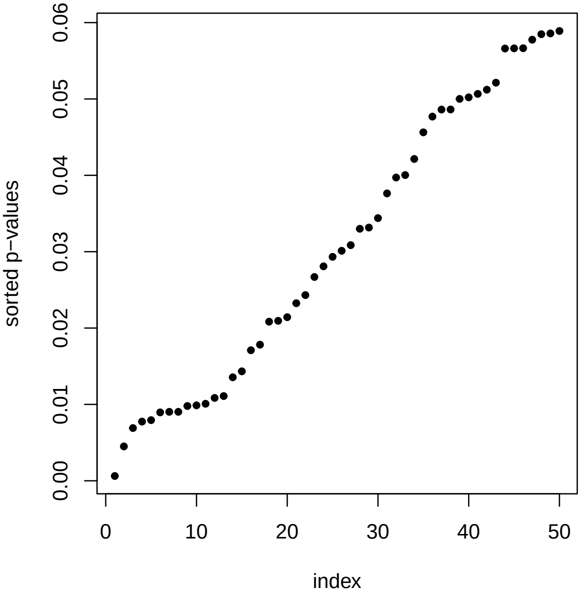
The BH procedure

- Say we want to control the FDR at level α
- Compute p-values $p_1, \ldots, p_n$ for the n hypotheses $H_1, \ldots, H_n$
- Sort the p-values: $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(n)}$

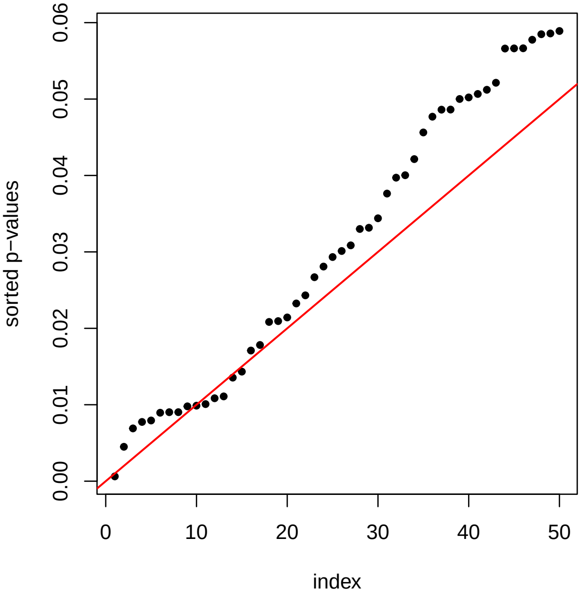
- Say we want to control the FDR at level α
- Compute p-values $p_1, \ldots, p_n$ for the n hypotheses $H_1, \ldots, H_n$
- Sort the p-values: $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(n)}$
- BH_α procedure: Reject $H_{(1)}, \ldots, H_{(i_0)}$ where

$$i_0 = \max\{i : p_{(i)} \leq i\alpha/n\}$$

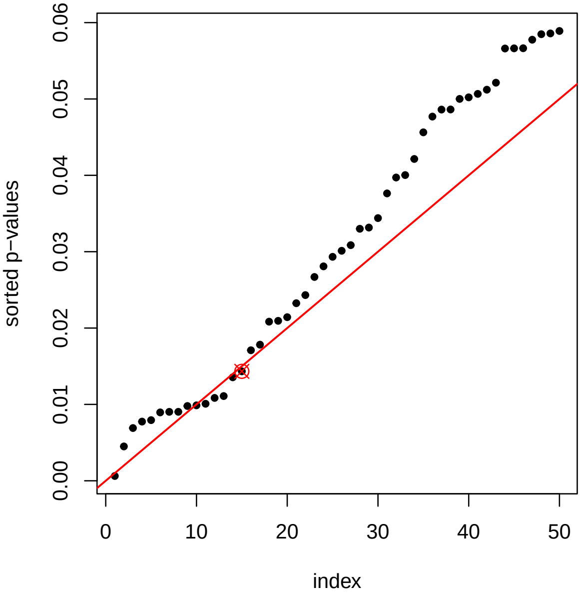
The BH procedure



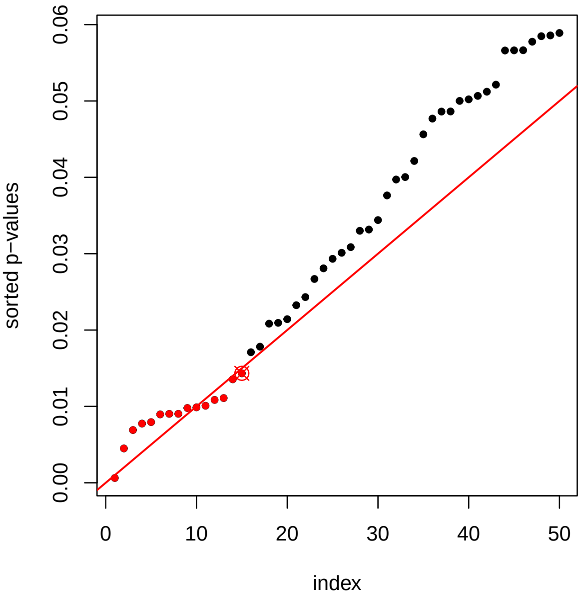
The BH procedure



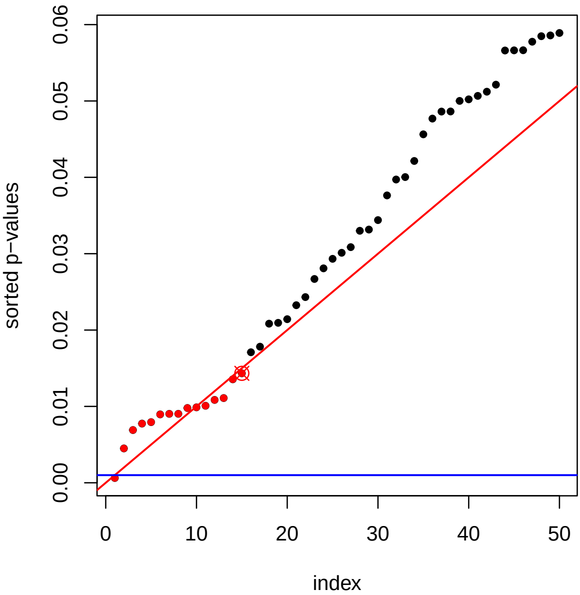
The BH procedure



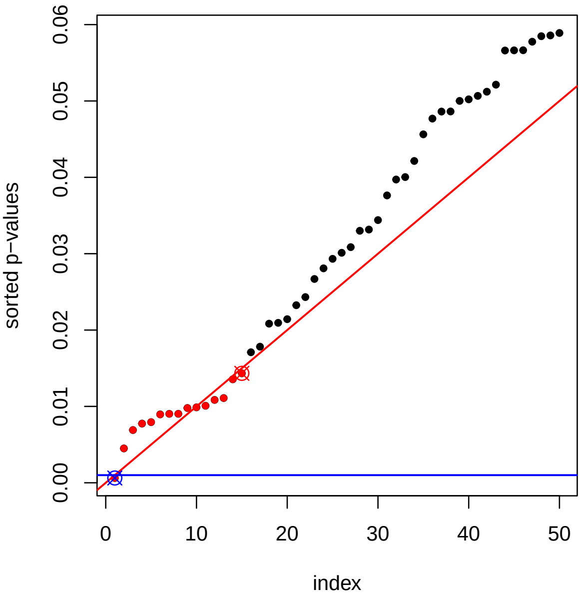
The BH procedure



BH allows more discoveries than Bonferroni



BH allows more discoveries than Bonferroni



Historical notes

- Simes (1986) mentioned BH procedure for weak FWER control

  controls FWER when all the hypotheses are true

Historical notes

- Simes (1986) mentioned BH procedure for weak FWER control
- Hommel (1988): it does not control FWER in the strong sense

  for some config of non-nulls, P(false discovery) can be more than α

- Simes (1986) mentioned BH procedure for weak FWER control
- Hommel (1988): it does not control FWER in the strong sense
- Hochberg (1988) gives a procedure for strong FWER control

$$i_0 = \max\left\{ i : p_{(i)} \leq \frac{\alpha}{n+1-i} \right\} \quad \text{vs} \quad i_0 = \max\left\{ i : p_{(i)} \leq \frac{i\alpha}{n} \right\}$$

Historical notes

- Simes (1986) mentioned BH procedure for weak FWER control
- Hommel (1988): it does not control FWER in the strong sense
- Hochberg (1988) gives a procedure for strong FWER control

$$i_0 = \max\left\{ i : p_{(i)} \leq \frac{\alpha}{n+1-i} \right\} \quad \text{vs} \quad i_0 = \max\left\{ i : p_{(i)} \leq \frac{i\alpha}{n} \right\}$$

- BH argue how their procedure rejects more than the above one

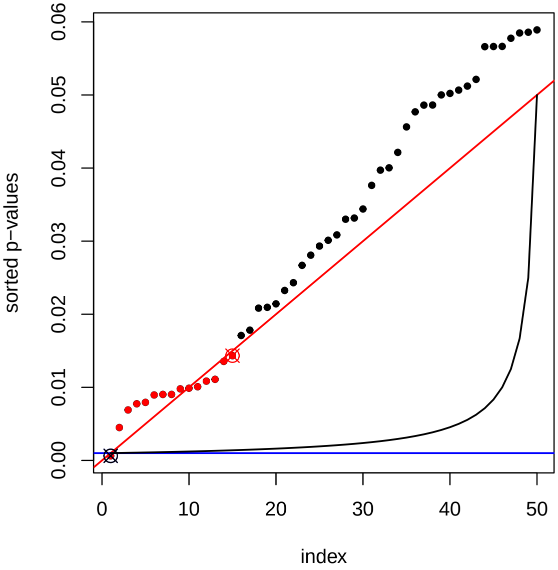


Figure 1: FDR control makes more rejections (and has more power) than FWER control

Theoretical guarantees

Theorem (Benjamini & Hochberg, 1995). The BH_α procedure controls the FDR at level α if the p-values are independent:

$$\mathrm{FDR} = \frac{n_0}{n}\alpha \leq \alpha.$$

FDR control

Theorem (Benjamini & Hochberg, 1995). The BH_α procedure controls the FDR at level α if the p-values are independent:

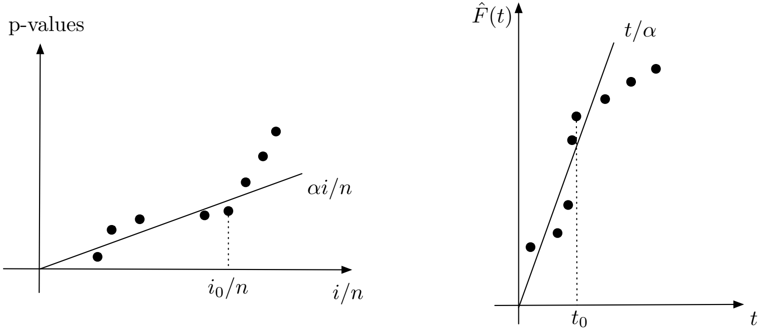
$$\text{FDR} = \frac{n_0}{n}\alpha \leq \alpha.$$

Numerous proofs, see our Stats 300C lecture notes for a couple of them

BH has more power

Theorem (BH, 1995). The BH procedure is a solution of the problem: choose t that maximizes the number of rejections at this level, $R(t)$, subject to the constraint $R(t)/n \geq t/\alpha$.

Theorem (BH, 1995). The BH procedure is a solution of the problem: choose t that maximizes the number of rejections at this level, $R(t)$, subject to the constraint $R(t)/n \geq t/\alpha$.



(a) p-values on the y axis, indices on x

(b) p-values on the x axis, indices on y

Image courtesy: Emmanuel Candès's 300C lecture notes

Proof of FDR control by Martingale theory (Storey et al., 2004)

Consider rejecting all H_i with p-values $p_i \leq t$, where $t \in (0, 1)$

	H_0 not rejected	H_0 rejected	Total
H_0 true	$U(t)$	$V(t)$	n_0
H_0 false	$T(t)$	$S(t)$	$n - n_0$
	$n - R(t)$	$R(t)$	n

Consider rejecting all H_i with p-values $p_i \leq t$, where $t \in (0,1)$

	H_0 not rejected	H_0 rejected	Total
H_0 true	$U(t)$	$V(t)$	n_0
H_0 false	$T(t)$	$S(t)$	$n - n_0$
	$n - R(t)$	$R(t)$	n

$V(t)/t$ is a backwards martingale $\mathbb{E}[\frac{V(s)}{s} \mid \mathcal{F}_{\geq t}] = \frac{1}{s}\frac{s}{t}V(t)$ for $s \leq t$

Proof of FDR control by Martingale theory (Storey et al., 2004)

Consider rejecting all H_i with p-values $p_i \leq t$, where $t \in (0,1)$

	H_0 not rejected	H_0 rejected	Total
H_0 true	$U(t)$	$V(t)$	n_0
H_0 false	$T(t)$	$S(t)$	$n - n_0$
	$n - R(t)$	$R(t)$	n

$V(t)/t$ is a backwards martingale $\mathbb{E}[\frac{V(s)}{s} \mid \mathcal{F}_{\geq t}] = \frac{1}{s}\frac{s}{t}V(t)$ for $s \leq t$

BH rejects all H_i with $p_i \leq \tau \Rightarrow \tau$ is a stopping time

Proof of FDR control by Martingale theory (Storey et al., 2004)

Consider rejecting all H_i with p-values $p_i \leq t$, where $t \in (0,1)$

	H_0 not rejected	H_0 rejected	Total
H_0 true	$U(t)$	$V(t)$	n_0
H_0 false	$T(t)$	$S(t)$	$n - n_0$
	$n - R(t)$	$R(t)$	n

$V(t)/t$ is a backwards martingale $\mathbb{E}[\frac{V(s)}{s} \mid \mathcal{F}_{\geq t}] = \frac{1}{s}\frac{s}{t}V(t)$ for $s \leq t$

BH rejects all H_i with $p_i \leq \tau \Rightarrow \tau$ is a stopping time

$$\mathrm{FDR}(\tau) = \mathbb{E}\left[\frac{V(\tau)}{R(\tau) \vee 1}\right] \overset{pic}{\leq} \frac{\alpha}{n}\mathbb{E}\left[\frac{V(\tau)}{\tau}\right] \overset{\text{OST}}{=} \frac{\alpha}{n}\mathbb{E}\left[\frac{V(1)}{1}\right] \overset{def}{=} \alpha\frac{n_0}{n} \leq \alpha$$

Proof of FDR control by Martingale theory (Storey et al., 2004)

Consider rejecting all H_i with p-values $p_i \leq t$, where $t \in (0,1)$

	H_0 not rejected	H_0 rejected	Total
H_0 true	$U(t)$	$V(t)$	n_0
H_0 false	$T(t)$	$S(t)$	$n - n_0$
	$n - R(t)$	$R(t)$	n

$V(t)/t$ is a backwards martingale $\mathbb{E}[\frac{V(s)}{s} \mid \mathcal{F}_{\geq t}] = \frac{1}{s}\frac{s}{t}V(t)$ for $s \leq t$

BH rejects all H_i with $p_i \leq \tau \Rightarrow \tau$ is a stopping time

$$\mathrm{FDR}(\tau) = \mathbb{E}\left[\frac{V(\tau)}{R(\tau) \vee 1}\right] \overset{pic}{\leq} \frac{\alpha}{n}\mathbb{E}\left[\frac{V(\tau)}{\tau}\right] \overset{\text{OST}}{=} \frac{\alpha}{n}\mathbb{E}\left[\frac{V(1)}{1}\right] \overset{def}{=} \alpha\frac{n_0}{n} \leq \alpha$$

Storey's procedure improves upon BH, by doing better than $\dfrac{n_0}{n} \leq 1$

Theorem (Benjamini & Yekutieli, 2001). Under arbitrary dependence of the p-values, the BH_α procedure has the following guarantee

$$\mathrm{FDR} = \frac{n_0}{n}\alpha H(n) \leq \alpha H(n)$$

where $H(n) = 1 + \frac{1}{2} + \cdots + \frac{1}{n} \approx \log n + 0.577$.

Theorem (Benjamini & Yekutieli, 2001). Under arbitrary dependence of the p-values, the BH_α procedure has the following guarantee

$$\mathrm{FDR} = \frac{n_0}{n}\alpha H(n) \le \alpha H(n)$$

where $H(n) = 1 + \frac{1}{2} + \cdots + \frac{1}{n} \approx \log n + 0.577$.

Theorem (Guo & Rao, 2008). There are joint distributions of p-values for which FDR of the BH procedure is at least $\min\{\alpha H(n), 1\}$.

- **e-value**: f is an e-value if $\mathbb{E}(f) \leq 1$ (under null)

The e-BH procedure (Wang & Ramdas, 2020)

- **e-value**: f is an e-value if $\mathbb{E}(f) \leq 1$ (under null)
- $1/(e\text{-value})$ is a valid p-value $\quad \mathbb{P}(f^{-1} \leq \alpha) = \mathbb{P}(f \geq \frac{1}{\alpha}) \leq \alpha \mathbb{E} f \leq \alpha$

The e-BH procedure (Wang & Ramdas, 2020)

- **e-value**: f is an e-value if $\mathbb{E}(f) \leq 1$ (under null)
- $1/(e\text{-value})$ is a valid p-value $\quad \mathbb{P}(f^{-1} \leq \alpha) = \mathbb{P}(f \geq \frac{1}{\alpha}) \leq \alpha \mathbb{E} f \leq \alpha$
- **e-BH** procedure: apply BH to a bunch of $(e\text{-values})^{-1}$

The e-BH procedure (Wang & Ramdas, 2020)

- **e-value**: f is an e-value if $\mathbb{E}(f) \leq 1$ (under null)
- $1/(e\text{-value})$ is a valid p-value $\qquad \mathbb{P}(f^{-1} \leq \alpha) = \mathbb{P}(f \geq \frac{1}{\alpha}) \leq \alpha \mathbb{E} f \leq \alpha$
- **e-BH** procedure: apply BH to a bunch of $(e\text{-values})^{-1}$
- **Theorem** (Wang & Ramdas, 2020). The e-BH procedure has FDR at most $\alpha n_0/n \leq \alpha$ (same guarantee as for the usual BH procedure with independent p-values)

Editorializing

- Traditional type-I error control fails when you test multiple hypotheses but suppress null findings.
- FDR is a *statistical* fix. But we also need *sociological* or *cultural* fixes: change the incentives in science so we can see more of the null findings.
  - Preregistration
  - Journals for null results
  - Evaluation criteria for job candidates, tenure, prestigious awards: do we value shocking results, or careful study design?

What are Open Science Badges?

- Badges to acknowledge open science practices are incentives for researchers to share data, materials, or to preregister
- Badges signal to the reader that the content has been made available and certify its accessibility in a persistent location.
- Currently, over **100** journals offer Open Science Badges to signal and reward when underlying data, materials, or preregistrations are available, see below.

Journal of Articles in Support of the Null Hypothesis

INDEX ABOUT MANUSCRIPT SUBMISSION REVIEWER SUBMISSION EDITORIAL BOARD CONTACT

Welcome to the *Journal of Articles in Support of the Null Hypothesis*. In the past other journals and reviewers have exhibited a bias against articles that did not reject the null hypothesis. We seek to change that by offering an outlet for experiments that do not reach the traditional significance levels ($p < .05$). Thus, reducing the file drawer problem, and reducing the bias in psychological literature. Without such a resource researchers could be wasting their time examining empirical questions that have already been examined. We collect these articles and provide them to the scientific community free of cost.

Journal of Negative Results in Biomedicine

Article Talk

From Wikipedia, the free encyclopedia

The ***Journal of Negative Results in Biomedicine*** was a peer-reviewed open access medical journal. It published papers that promote a discussion of unexpected, controversial, provocative and/or negative results in the context of current research. The journal was established in 2002 and ceased publishing in September 2017. It was abstracted and indexed in the Emerging Sources Citation Index,[1] Index Medicus/MEDLINE/PubMed,[2] and Scopus.[3]

Thank You!

Questions?