Is *BC* the new *BH*?

A tale of two multiple hypothesis testing procedures

Tim Barry

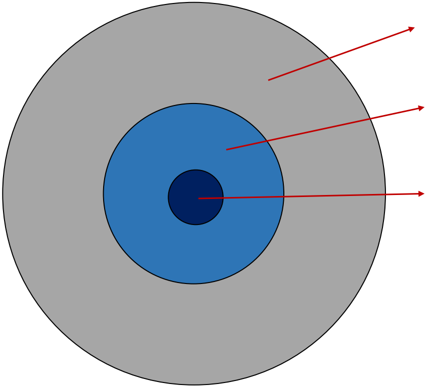
January 28, 2021

Shifting my statistical focus

- **Before**
  - exponential families
  - measurement error models
  - latent variable models
- **Now**
  - multiple testing and false discovery rates
  - conditional independence testing
  - negative control methods
  - robust inference

Multiple testing review

Consider m hypothesis $H_1, \ldots, H_m$. Suppose that we test these hypotheses and make R discoveries. Of these discoveries, suppose that V are *false* discoveries (i.e., true nulls) and that $V - R$ are *true discoveries* (i.e., true alternatives).



All *m* hypotheses

Discoveries *R*

False discoveries *V*

Multiple testing review: FDR and FWER

The family-wise error rate (FWER) is the probability of making even 1 false discovery:

$$FWER = \mathbb{P}(V \geq 1).$$

The false discovery rate (FDR) is the expected fraction of false discoveries:

$$FDR = \mathbb{E}\left(\frac{V}{\max\{R,1\}}\right) := \mathbb{E}\left(FDP\right).$$

If $R = 0$ (i.e., no discoveries), then

$$FDR = 0.$$

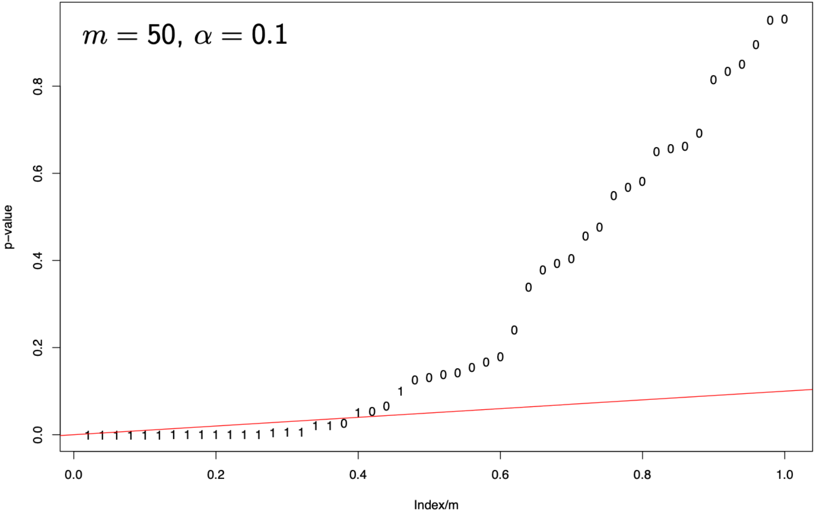
Benjamini-Hochberg (BH)

BH Procedure: Suppose that we calculate p-values $p_1, \ldots, p_m$ for the hypotheses $H_1, \ldots, H_m$. Suppose that the p-values corresponding to the true null hypotheses are $U(0,1)$, and suppose that the p_is are independent. Let $p_{(1)} \leq \ldots, \leq p_{(m)}$ be the ordered p-values and $H_{(1)}, \ldots, H_{(m)}$ the corresponding ordered hypotheses. Let $\alpha \in (0,1)$ be the user-chosen FDR level (typically, $\alpha \in \{0.05, 0.1, 0.2\}$). Define

$$\hat{k} = \mathrm{argmax}_k \{p_{(k)} \leq \alpha k/m\}.$$

Reject $H_{(1)}, \ldots, H_{(\hat{k})}$.
Theorem: The BH procedure controls the FDR at level α, i.e. $FDR \leq \alpha$.

Visual interpretation of BH



(source: Chris Genovese)

The Barber-Candés (BC) procedure

BC procedure (Barber and Candés 2015) : Let $X_1, \ldots X_m$ be test statistics corresponding to the hypotheses $H_1, \ldots, H_m$, with large values indicating evidence against the null hypothesis. Assume that the density ψ of the null test statistics is symmetric about 0, i.e. $\psi(x) = \psi(-x)$ for all $x \geq 0$. Also, assume that the X_is are independent. Let $|X| := \{|X_i| : i = 1, \ldots, n\}$ be the set of sample absolute values, and let

$$\widehat{FPD}(t) := \frac{1 + \#\{i : X_i \leq -t\}}{\max\left(1, \#\{i : X_i \geq t\}\right)}$$

be the empirical false discovery proportion for given $t \in |X|$. Finally, let $\alpha \in (0, 1)$ be the FDR target. The BC threshold is $\tau_{\mathrm{BC}} = \min\left\{t \in |X| : \widehat{FDP}(t) \leq \alpha\right\}$. Reject all $X_i \geq \tau_{\mathrm{BC}}$.

Theorem: The BC procedure controls FDR at level α, i.e. $FDR \leq \alpha$.

BC procedure note

Note: The null X_is are **not** p-values. Instead, the null X_is are test statistics with a shared, symmetric density. For example:

- Gaussian
- Double exponential (AKA Laplace)
- Student's t
- Rademacher
- Unknown symmetric distribution
- etc.

Tim's thought: We can run BC on transformed p-values. Let $p_i \sim \mathrm{U}(0, 1)$ be a null p-value. For $c > 0$, let $X_i := c(-p_i + 1/2)$. Then $X_i \sim \mathrm{U}(-c, c)$, with large values indicating evidence against the null. **Therefore, BC is more flexible than BH.**

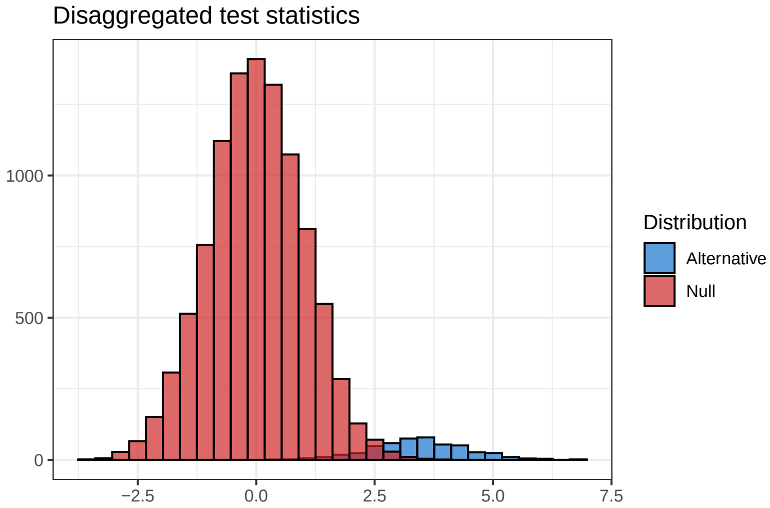
Visual interpretation of BC

- Let $X_1, \ldots, X_{10,000} \sim N(0, 1)$ be the test statistics under the null hypothesis. Let $Y_1, \ldots, Y_{500} \sim N(3.5, 1)$ be the test statistics under the alternative hypothesis. Let

  $$Z := [X_1, \ldots, X_{10,000}, Y_1, \ldots, Y_{500}]$$

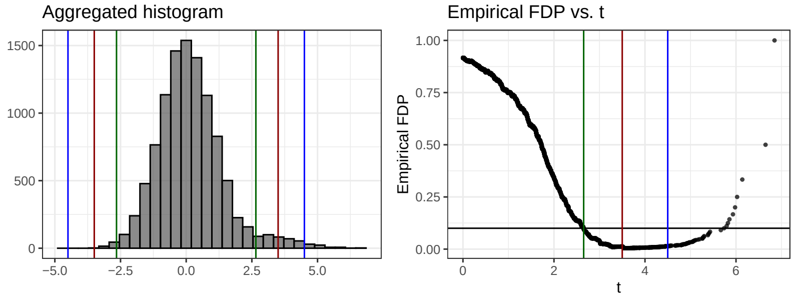
  be the full vector of test statistics.
- We apply BC to Z with the goal of discovering the Y_is with FDR control at level 0.1.

Visual interpretation of BC



Visual interpretation of BC

The blue, red, and green vertical lines represent different candidate thresholds (at 4.5, 3.5, and 2.65, respectively).



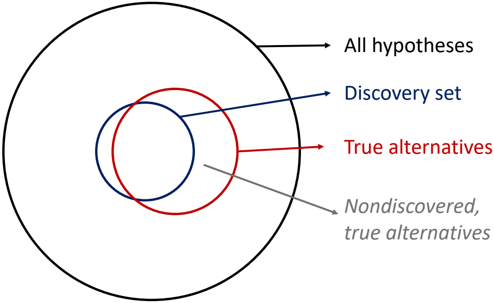
The BC threshold is at 2.65; thus, we reject all test statistics greater than this value.

Power

- Let S be the number of *true alternatives*, and let Q be the number of *non-discovered* true alternatives. The *false non-discovery rate* (FNR) is

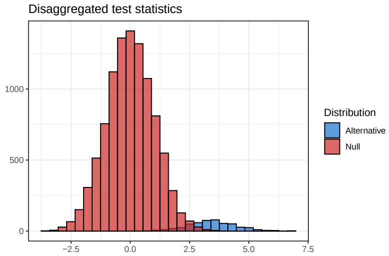
$$FNR := \mathbb{E}\left[\frac{Q}{\max\{S, 1\}}\right].$$

- FDR is analogous to type I error; FNR is analogous to type II error.



Power: numerical experiment

I ran a small simulation experiment ($B = 50$ replicates) to compare the FDR (target: $< 10\%$) and FNR (smaller is better) of BC and BH in the example above. The results were similar across methods.



	BH	BC
FDR	9.7 %	9.8 %
FNR	19.1 %	19.2 %

Power: theoretical result

- Arias-Castro and Chen (2017) showed that BC and BH have asymptotically identical power when the test statistics are Gaussian (as above).
- The empirical experiments of Arias-Castro and Chen (2017) confirm their theoretical results. However, BC empirically seems lose power in “ultrasparse” (i.e., $< 1/1000$ hypotheses true) settings. More investigation is required.

BC opens the door to new strategies for high-dimensional, robust, and/or nonparametric FDR control.

1. High-dimensional, nonparametric two-sample testing (Ge et al. 2021).
2. Signal recovery in the (possibly high-dimensional) linear model (Barber and Candés 2015).
2. Doubly-robust, finite-sample calibration with negative controls (us 2022+?).

Arias-Castro, Ery and Shiyun Chen (2017). “Distribution-free multiple testing”. In: *Electronic Journal of Statistics* 11.1, pp. 1983–2001.

Barber, Rina Foygel and Emmanuel J. Candés (2015). “Controlling the false discovery rate via knockoffs”. In: *Annals of Statistics* 43.5, pp. 2055–2085.

Ge, Xinzhou et al. (2021). “Clipper: p-value-free FDR control on high-throughput data from two conditions”. In: *Genome Biology* 22.1, pp. 1–29.