PhD in Economics and Statistics - University of Milano-Bicocca

Lecture 4: The post-hoc inference problem

May 8, 2019

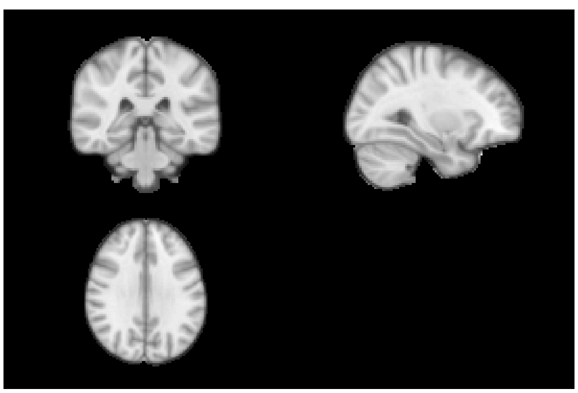
Lecturer: Aldo Solari

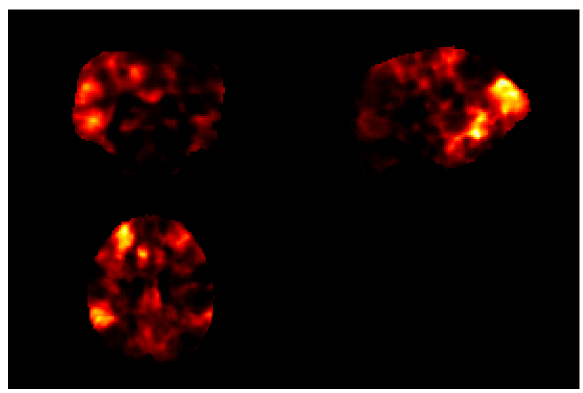
Closed testing (Marcus et al., 1976) is a fundamental principle of FWER control. [3] showed that closed testing can be used to obtain simultaneous confidence bounds for the false discovery proportion (FDP). Used in this way, closed testing allows a form of *post-selection inference*. Theoretical results for the special case of Simes local tests are discussed in [1]. Applications to genomics and fMRI data are discussed in [2] and [4], respectively.

1 fMRI data

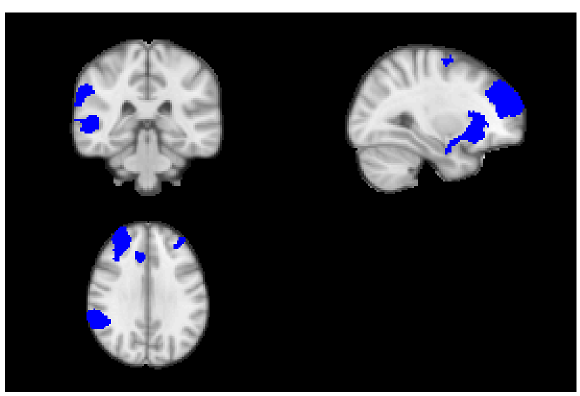
Analysis of fMRI data supplies an *activation map*: a p-value testing the null hypothesis of 'no activation' (e.g. $H_i : \mu_i = 0$) at each location (*voxel*).

The goal is to find brain regions of activations. Note that the data is at higher resolution than units of interest: we have p-values at the voxel level, but we wish to perform inference at the region level.





We can select interesting regions by looking at the data (activation map):



Now the problem is the following: *How to assess the significance of selected regions?* Regions are both selected and tested with the same data. We need to correct for the overoptimism in inference due to data-driven selection.

1.1 Classical multiple testing

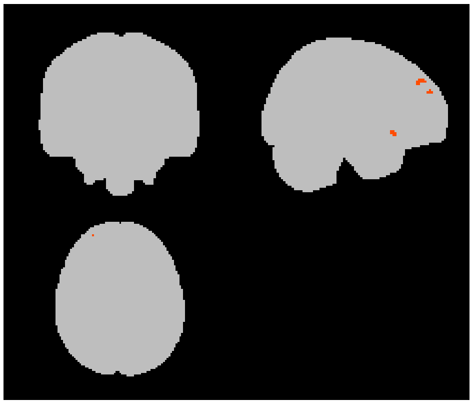
Suppose we have m hypotheses $H_1, \ldots, H_m$ with corresponding p-values $p_1, \ldots, p_m$. Denote by $T \subseteq \{1, \ldots, m\}$ and $F = \{1, \ldots, m\} \setminus T$ the index sets of true and false hypotheses, respectively. For a selection $S \subseteq \{1, \ldots, m\}$, let

$$\pi_1(S) = \frac{|F \cap S|}{|S|}$$

be the proportion of false hypotheses (true activations) in S, where $|\cdot|$ denotes the size of a set.

If we use Bonferroni method at level α, we reject all the hypotheses with indices in

$$R_\alpha = \{i : p_i \leq \alpha/m\}$$



Then the proportion of true activations in a selected region S can be estimated by

$$\hat{\pi}_1(S) = \frac{|R_\alpha \cap S|}{|S|}$$

with

$$\mathrm{P}(\pi_1(S) \leq \hat{\pi}(S)\ \forall S) \geq 1 - \alpha$$

FWER has the *subsetting property* that if a set R of hypotheses is rejected by an FWER-controlling procedure, then FWER control is also guaranteed for any subset $S \subset R$, i.e.

$$\mathrm{P}(R \cap T = \emptyset) \geq 1 - \alpha \Rightarrow \mathrm{P}(S \cap T = \emptyset) \geq 1 - \alpha \quad \forall S \subseteq R$$

This property does not hold for FDR control. If FDR control holds for the full set R only, this does not translate to any subset S, i.e.

$$\mathrm{E}\Big(\frac{|R \cap T|}{|R| \vee 1}\Big) \leq \alpha \not\Rightarrow \mathrm{E}\Big(\frac{|S \cap T|}{|S| \vee 1}\Big) \leq \alpha$$

2 Closed testing

Let $\theta \in \Theta$ be a parameter of interest. Suppose we have m hypotheses $H_1, \ldots, H_m$ with $H_i \subset \Theta$ true if and only if $\theta \in H_i$. Denote by $T \subseteq \{1, \ldots, m\}$ the index set of true hypotheses.

In the closed testing procedure, the collection of hypotheses is augmented with all possible intersection hypotheses

$$H_I = \bigcap_{i \in I} H_i$$

with $I \subseteq \{1, \ldots, m\}$. An intersection hypothesis H_I is true if and only if H_i is true for all $i \in I$. Note that $H_i = H_{\{i\}}$, so all original hypotheses, known as *elementary hypotheses*, are also intersection hypotheses.

The closed testing procedure starts by testing all intersection hypotheses with a *local test*, i.e. an α-level test for H_I:

$$\sup_{\theta \in H_I} \mathrm{P}_\theta(\phi_I = 1) \leq \alpha$$

Let the collection of all subsets of $\{1, \ldots, m\}$ be denoted by

$$\mathcal{I} = 2^{\{1, \ldots, m\}}$$

and the collection of index sets corresponding to true intersection hypotheses by

$$\mathcal{T} = \{I \in \mathcal{I} : I \subseteq T\}$$

We define $\mathcal{U}_\alpha$ as the collection of $I \in \mathcal{I}$ such that H_I is rejected by a local test at level α

$$\mathcal{U}_\alpha = \{I \in \mathcal{I} : \phi_I = 1\}$$

For each I, H_I is rejected by the closed testing procedure if and only if $I \in \mathcal{X}_\alpha$ where

$$\mathcal{X}_\alpha = \{I \in \mathcal{I} : J \in \mathcal{U}_\alpha \ \forall J \supseteq I\}$$

Theorem 2.1. *The closed testing procedure controls the FWER for all hypotheses H_I at level α, i.e.*

$$\mathrm{P}(\mathcal{T} \cap \mathcal{X}_\alpha = \emptyset) \geq 1 - \alpha$$

Proof. We have

$$\{T \notin \mathcal{U}_\alpha\} \subseteq \{I \notin \mathcal{X}_\alpha \ \forall I \subseteq T\} \subseteq \{\mathcal{T} \cap \mathcal{X}_\alpha = \emptyset\}$$

and because H_T is tested by an α level test

$$1 - \alpha \leq \mathrm{P}(T \notin \mathcal{U}_\alpha) \leq \mathrm{P}(\mathcal{T} \cap \mathcal{X}_\alpha = \emptyset)$$

$\square$

2.1 Closed testing for FWER control

Let

$$R_\alpha = \{i : \{i\} \in \mathcal{X}_\alpha\}$$

be the index set of elementary hypotheses rejected by the closed testing procedure. Then $\mathrm{P}(\mathcal{T} \cap \mathcal{X}_\alpha = \emptyset) \geq 1 - \alpha$ implies

$$\mathrm{P}(T \cap R_\alpha = \emptyset) \geq 1 - \alpha$$

Local tests tend to be easy to specify in most models, as each local test is a test of a single hypothesis, so that standard statistical test theory may be used. Below some examples of local tests based on p-values:

- Bonferroni local test: reject H_I if and only if

  $$\bigcup_{i \in S} \Big\{ p_i \leq \frac{\alpha}{|I|} \Big\}$$

  The CT procedure based on Bonferroni local tests gives Holm's method

- Simes local test: reject H_I if and only if

  $$\bigcup_{i \in S} \Big\{ p_{(i:I)} \leq \frac{i\alpha}{|I|} \Big\}$$

  where $p_{(i:I)}$ is the ith smallest p-value among the multiset $\{p_i : i \in I\}$. The Simes local test is level α if we assume the Simes' inequality, i.e.

  $$\mathrm{P}\Big(\bigcup_{i \in T} \Big\{ p_{(i:T)} \leq \frac{i\alpha}{m_0} \Big\} \Big) \leq \alpha$$

  The CT procedure based on Simes local tests gives the Hommel method: compute

  $$\hat{m}_0 = \max\{i \in \{0, \ldots, m\} : ip_{(m-i+j)} > j\alpha \text{ for } j = 1, \ldots, i\}$$

  and

  $$R_\alpha = \{i : p_i \leq \alpha / \hat{m}_0\}$$

  Hommel method always rejects at least as much as Hochberg's, and possibly more.

- Fisher local test: reject H_I if and only if

  $$-2 \sum_{i \in I} \log(p_i) \geq c_{|I|}$$

  where c_r is the $1 - \alpha$ quantile of a χ^2 distribution with $2r$ degrees of freedom. Fisher's combination test is level α if we assume that the null p-values are independent.

3 Simultaneous confidence bounds for the FDP

Let $S \subseteq \{1, \ldots, m\}$ be the index set of selected hypotheses, the discoveries. The number of false discoveries for S is

$$\tau(S) = |T \cap S|$$

[3] show how to calculate upper confidence bounds $\hat{\tau}_\alpha$ for $\tau(S)$ from the closed testing procedure:

$$\hat{\tau}_\alpha(S) = \max\{|I| : I \subseteq S, I \notin \mathcal{X}_\alpha\}$$

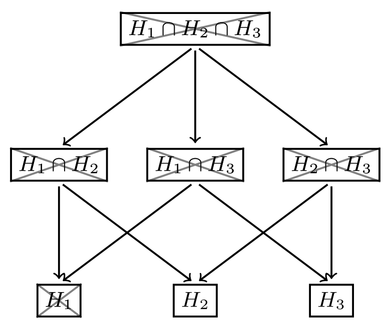
[3] prove that

$$\mathrm{P}(\tau(S) \leq \hat{\tau}_\alpha(S) \ \forall S \in \mathcal{I}) \geq 1 - \alpha$$

meaning that the set

$$\{0, \ldots, \hat{\tau}_\alpha(S)\}$$

is a $(1-\alpha)$-confidence set of the parameter $\tau(S)$. The quantity $\hat{\tau}_\alpha(S)$ is the size of the largest subset of S for which the corresponding intersection hypothesis is not rejected by the closed testing procedure. For, example, if the rejected hypotheses by the closed testing procedure are those marked with a cross



we obtain

S	Confidence set for $\tau(S)$
$\{1\}$	$\{0\}$
$\{2\}$	$\{0, 1\}$
$\{3\}$	$\{0, 1\}$
$\{1, 2\}$	$\{0, 1\}$
$\{1, 3\}$	$\{0, 1\}$
$\{2, 3\}$	$\{0, 1\}$
$\{1, 2, 3\}$	$\{0, 1\}$

To prove the coverage, remember that if the event $E = \{T \notin \mathcal{U}_\alpha\}$ has happened, then all rejections that the closed testing procedure has made are correct. Given that E has happened, the value of $\tau(S)$ cannot be greater than the value of $\hat{\tau}_\alpha(S)$, because otherwise a true intersection hypothesis would have been rejected, which is inconsistent with the definition of E.

The important thing to note about confidence sets is that they are simultaneous confidence sets, which all depend on exactly the same event E for their coverage. Because these confidence sets are simultaneous, the user can review all these confidence sets, and select the one that he or she likes best, while still keeping correct $1-\alpha$ coverage of the selected confidence set: under the event E, all confidence sets cover the true parameter simultaneously, and therefore, under the same event E, the selected confidence set covers the true parameter. Consequently, the selected confidence set has coverage $\mathrm{P}(E) \geq 1-\alpha$. The simultaneity of the sets makes their coverage robust against post hoc selection.

The bounds can be equivalently formulated in terms of the false discovery proportion $\pi_0(S) = |T \cap S|/|S|$ as

$$\pi_0(S) = \frac{\hat{\tau}_\alpha(S)}{|S|}$$

or in terms of the true discovery proportion $\pi_1(S) = |F \cap S|/|S|$ as

$$\pi_1(S) = \frac{|S| - \hat{\tau}_\alpha(S)}{|S|}$$

4 Closed testing with Simes local tests

Naive application of a closed testing procedure requires 2^m local tests to be performed, which severely limits the usefulness of the procedure in large problems. For this reason *shortcuts* have been developed, which are algorithms that limit the computation time of the closed testing procedure. [1] provided shortcuts for closed testing with Simes local tests.

Assume that Simes' inequality holds, i.e.

$$\mathrm{P}\Big(\bigcup_{i\in T}\Big\{p_{(i:T)} \leq \frac{i\alpha}{m_0}\Big\}\Big) \leq \alpha$$

The raw p-value of Simes test for testing H_I is

$$p_I = \min_{1\leq i\leq |I|}\Big\{\frac{|I|}{i} p_{(i:I)}\Big\}$$

where $p_{(i:I)}$ is the ith smallest p-value in $\{p_i : i \in I\}$.
Hommel (1988) defined the quantity

$$\hat{m}_0 = \max\{i \in \{0,\ldots,m\} : ip_{(m-i+j)} > j\alpha \text{ for } j = 1,\ldots,i\}$$

It turns out that $\hat{m}_0$ is an $1-\alpha$ upper bound for the number of true hypotheses m_0:

$$\mathrm{P}(m_0 \leq \hat{m}_0) \geq 1-\alpha$$

Furthermore, $\hat{m}_0$ can be used to determine whether H_I is rejected by the closed testing procedure:

Lemma 4.1. *$I \in \mathcal{X}_\alpha$ if and only if* $\min_{1 \le i \le |I|} \left\{ \frac{\hat{m}_0}{i} p_{(i:I)} \right\} \le \alpha$

Finally, it can be proved that

Theorem 4.2. *The lower bound $\hat{\pi}_1(S)$ for $\pi_1(S)$ is given by*

$$\hat{\pi}_1(S) = \min \left\{ 0 \le k \le |S| : \min_{1 \le i \le |S|-k} \left\{ \frac{\hat{m}_0}{i} p_{(i+k:S)} \right\} > \alpha \right\} (|S|)^{-1}$$

such that

$$\mathrm{P}(\hat{\pi}_1(S) \le \pi_1(S) \text{ for all } S) \ge 1 - \alpha$$

4.1 Algorithm's complexity

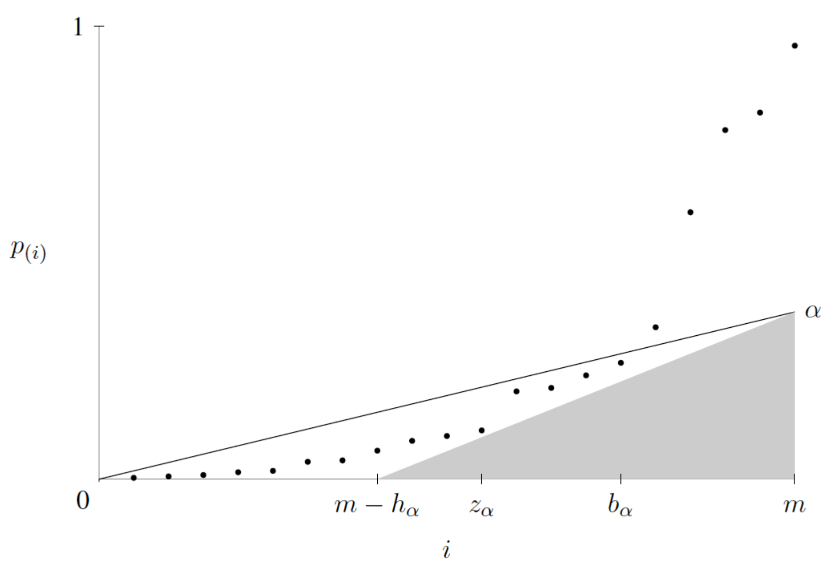
- Sort the p-values (once): linearitmic, i.e. $O(m \log m)$
- Compute h_α (once): linearitmic, i.e. $O(m \log m)$
- Compute $\hat{\pi}_1(S)$: linear in the size of S, i.e. $O(|S|)$

4.2 Relationship with Benjamini-Hochberg

Define the rejection set of Benjamini-Hochberg method by

$$B = \{ i : p_i \le (\alpha b)/m \}$$

where $b = \max(i : p_{(i)} \le i\alpha/m)$. A graphical illustration of $\hat{m}_0$ and b is given in the Figure below:



The value of $\hat{m}_0$ (h_α in the Figure) is the largest such that the gray triangle does not contain any of the points; the value of b (b_α in the Figure) is the index of the largest point below the line.

It can be proved that

$$\frac{|R|}{m} \leq \hat{\pi}_1 \leq \frac{|B|}{m}$$

where $R = \{i : p_i \leq \alpha/\hat{m}_0\}$ is the rejected set of Hommel's method for FWER control.

The probability of FDP exceeding a certain threshold γ can be bounded by Markov's inequality:

$$\mathrm{P}(Q > t) \leq \frac{\mathrm{E}(Q)}{t}$$

If B_γ is the rejection set of BH at level γ, then $Q = \pi_0(B_\gamma) = \dfrac{|B_\gamma \cap T|}{|B_\gamma| \vee 1}$ and for $t = \gamma/\alpha$ then

$$\mathrm{P}\Big(\pi_0(B_\gamma) \leq \frac{\gamma}{\alpha}\Big) \geq 1 - \alpha$$

For example, for $\alpha = 0.1$ and $\gamma = 0.01$, then for $B_{0.01}$ we have $\mathrm{E}(\mathrm{FDP}) \leq 0.01$ and

$$\mathrm{P}(\mathrm{FDP} < 0.1) \geq 0.9$$

Lemma 4.3. *For any $\gamma \in (0, 1)$, we have $\hat{\pi}_0(B_\gamma) \leq \hat{\pi}_0 \dfrac{\gamma}{\alpha}$ and*

$$\mathrm{P}\Big(\pi_0(B_\gamma) \leq \hat{\pi}_0 \frac{\gamma}{\alpha}\Big) \geq 1 - \alpha$$

In particular, for $B = B_\alpha$ we have

$$\mathrm{E}(\pi_0(B)) \leq \pi_0 \alpha \quad \text{and} \quad \mathrm{P}\Big(\pi_0(B) \leq 2\hat{\pi}_0 \alpha\Big) \geq \frac{1}{2}$$

4.3 Scalability of power

Assume that p-values are drawn from a mixture distribution

$$\begin{aligned} P(u) &= \pi_0 P_0(u) + \pi_1 P_1(u) \\ &= \pi_0 u + \pi_1 P_1(u) \end{aligned}$$

where $P_0(u) = u$ is the distribution of the p-values for true null hypotheses and $P_1(u) \geq u$ is the distribution of the p-values for false null hypotheses.

As the number of hypotheses $m \to \infty$, the number of FWER rejections (Hommel) divided by m vanishes, i.e.

$$\operatorname*{plim}_{m \to \infty} \frac{|R^m|}{m} = 0$$

where $\operatorname{plim}_{m \to \infty} X_n = X$ is short for $\lim_{m \to \infty} \mathrm{P}(|X_m - X| > \epsilon) = 0$ However, if enough signal is present, i.e.

$$\exists u \in [0, 1) : P(u\alpha) > u$$

then the number of FDR rejections (BH) divided by m does not vanish, i.e.

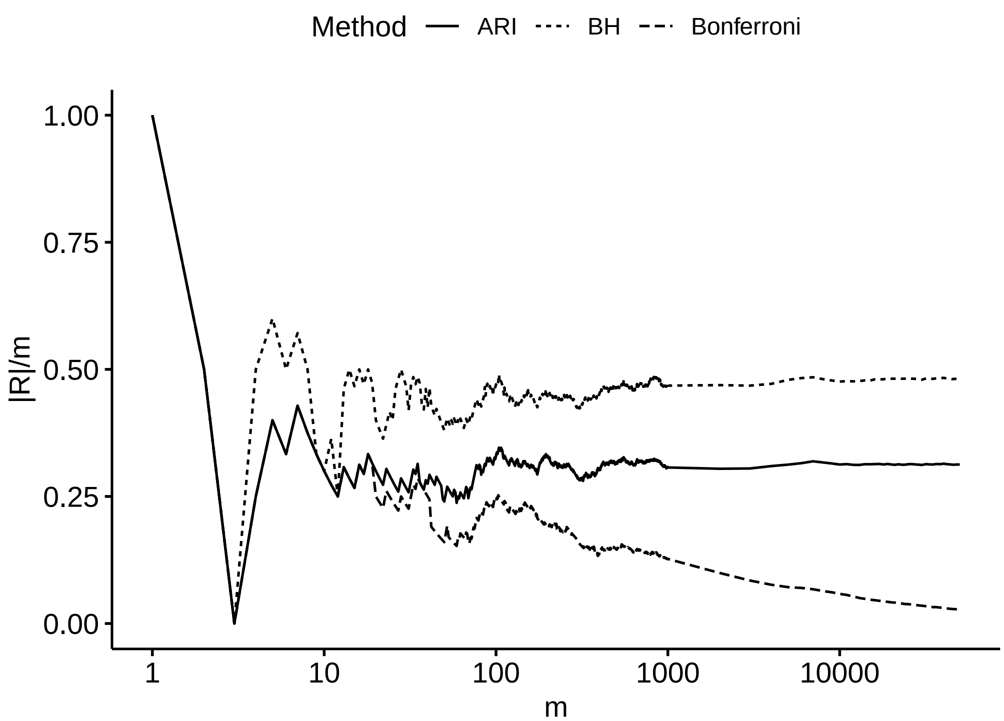
$$\operatorname*{plim}_{m\to\infty} \frac{|B^m|}{m} = k > 0$$

with $k = \dfrac{\sup\{u : P(u) \geq u/\alpha\}}{\alpha}$. What about $\hat{\pi}_1$?

Lemma 4.4. *If enough signal is present, we have*

$$\operatorname*{plim}_{m\to\infty} \hat{\pi}_1 = c > 0$$

with $c = \sup\limits_{0\leq u<1} \dfrac{P(u\alpha) - u}{1-u}$.



4.4 Consistency

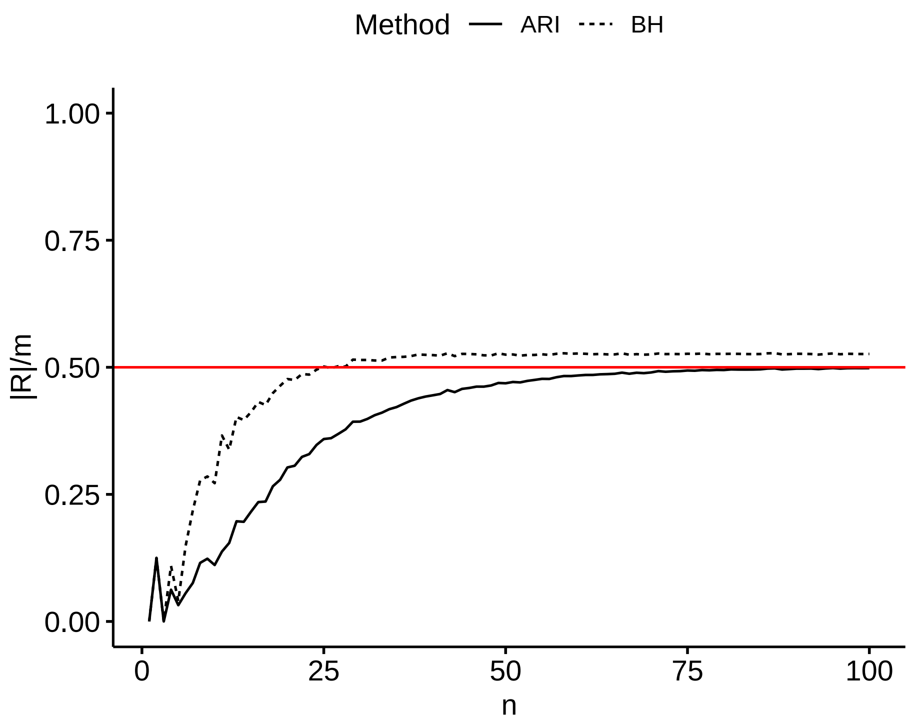
Is $\hat{\pi}_\alpha$ a consistent estimator of π?

Theorem 4.5. *Letting* $n \to \infty$ *and* $m \to \infty$*, we have, for fixed* $0 < \alpha < 1$*:*

$$\operatorname*{plim}_{(m,n)\to\infty} \hat{\pi}_1^{m,n} = \pi_1$$

Compare to Benjamini-Hochberg:

$$\operatorname*{plim}_{(m,n)\to\infty} \frac{|B^{m,n}|}{m} = \frac{\pi_1}{1-\alpha(1-\pi_1)} > \pi_1$$



References

[1] J. Goeman, R. Meijer, T. Krebs, and A. Solari. Simultaneous control of all false discovery proportions in large-scale multiple hypothesis testing. *Biometrika (to appear)*, 2019.

[2] J. J. Goeman and A. Solari. Multiple hypothesis testing in genomics. *Statistics in medicine*, 33(11):1946–1978, 2014.

[3] J. J. Goeman, A. Solari, et al. Multiple testing for exploratory research. *Statistical Science*, 26(4):584–597, 2011.

[4] J. D. Rosenblatt, L. Finos, W. D. Weeda, A. Solari, and J. J. Goeman. All-resolutions inference for brain imaging. *NeuroImage*, 181:786–796, 2018.