Closed testing

Aldo Solari
Statistical Inference II
PhD in Economics, Statistics and Data Science
University of Milano-Bicocca

XXXVII cycle

Outline

Comparing three groups

Simultaneous control of false discovery proportions

Main references

- Goeman, J.J. and Solari, A. (2022) Comparing three groups. The American Statistician, 76, 168-176.
- Goeman, J.J. and Solari, A. (2011) Multiple Testing for Exploratory Research. Statistical Science, 26, 584–597.
- Goeman, J.J., Meijer, R.J., Krebs, J.T.P. and Solari, A. (2019) Simultaneous Control of All False Discovery Proportions in Large-Scale Multiple Hypothesis Testing. Biometrika, 106, 841–856.

Table of Contents

Suppose that genetically similar seeds are randomly assigned to be raised either under standard conditions (control) or in two different nutritionally enriched environments (treatments I and II).

After a predetermined period all plants are harvested, dried and weighed. The results, expressed as dried weight in grams, for samples of $n = 10$ plants from each group are given in the following Table (data from Dobson, 1983, Table 7.1):

Control	4.17	5.58	5.18	6.11	4.50	4.61	5.17	4.53	5.33	5.14
Treatment I	4.81	4.17	4.41	3.59	5.87	3.83	6.03	4.89	4.32	4.69
Treatment II	6.31	5.12	5.54	5.50	5.37	5.29	4.92	6.15	5.80	5.26

Four hypotheses

We may formulate four null hypotheses to compare the group means μ_1, μ_2, and μ_3.

First, the so-called ‘global’ null hypotheses that all three group means are equal:

$$H_{123} \colon \mu_1 = \mu_2 = \mu_3.$$

Next, there are the three pairwise comparisons between groups:

$$H_{12} \colon \mu_1 = \mu_2; \qquad H_{13} \colon \mu_1 = \mu_3; \qquad H_{23} \colon \mu_2 = \mu_3.$$

Four scenarios

For the three-group comparison case we distinguish four scenarios for the choice of primary and secondary hypotheses.

1. *The global hypothesis H_{123} is primary*: This is natural when the presence of any difference between the means can directly be meaningfully interpreted, regardless of the location of such difference.
2. *All three pairwise hypotheses, H_{12}, H_{13} and H_{23}, are primary*: This is natural when the three groups represent categories of a nominal variable, and all three groups are equally important.
3. *Two of the pairwise hypotheses, say H_{12} and H_{13}, are primary*: This is natural when Group 1 represents a reference against which both other groups are compared.
4. *One of the pairwise hypotheses, say H_{12}, is primary*: This is natural when one of the groups (Group 3) is of secondary interest.

Scenario A is appropriate if we would first and foremost want to show that there is some effect of different growing conditions, regardless of which.

Scenario B would be chosen if we would be equally interested in showing a difference between any of the groups, but if only rejecting the global hypothesis would be unsatisfactory.

Scenario C would be appropriate if we would be primarily interested in finding at least one of the treatments is different from the control.

Scenario D prioritizing H_{12} would be most appropriate if demonstrating the effectiveness of treatment I with respect to the control would be of primary interest.

F tests

The standard tests for H_{123}, H_{12}, H_{13} and H_{23} in the one-way ANOVA model (with equal-size groups) are the (partial) *F*-tests based on the estimates $\hat{\mu}_1$, $\hat{\mu}_2$, $\hat{\mu}_3$, and pooled variance estimate $\hat{\sigma}^2$.

The partial *F*-test statistic for H_{12} is proportional to the standardized squared group difference

$$S_{12} = \frac{(\hat{\mu}_2 - \hat{\mu}_1)^2}{\hat{\sigma}^2}$$

analogous for H_{13} and H_{23}. The distributions of S_{12}, S_{13}, S_{23} are identical under the null hypotheses; let c_α be the $1 - \alpha$-quantile of that distribution.

For H_{123} the *F* test is proportional to the test statistic

$$S_{123} = S_{12} + S_{13} + S_{23}$$

Tukey HSD and Dunnett methods

Tukey's Honest Significant Difference (HSD) method rejects when $S_{ij} \geq \tilde{c}_\alpha$, where $\tilde{c}_\alpha$ is the $(1-\alpha)$-quantile of the distribution of

$$\tilde{S}_{123} = \max(S_{12}, S_{13}, S_{23}),$$

which is proportional to a studentized range distribution

Dunnett's procedure rejects H_{12} and/or H_{13} when the corresponding test statistics exceed $\tilde{c}_\alpha^1$, where $\tilde{c}_\alpha^1$ is the $(1-\alpha)$-quantile of the distribution of

$$\tilde{S}_1 = \max(S_{12}, S_{13}).$$

Note that Dunnett's critical value is less stringent than Tukey's one, i.e. $c_\alpha < \tilde{c}_\alpha^1 < \tilde{c}_\alpha$.

Restricted combinations

The four hypotheses H_{123}, H_{12}, H_{13}, and H_{23} are logically related to each other: if any two are true, then all must be true.

For example, if H_{12} and H_{13} are true, then $\mu_1 = \mu_2$ and $\mu_1 = \mu_3$, so that we have $\mu_1 = \mu_2 = \mu_3$, which implies that H_{123} and H_{23} are also true.

The number of true hypotheses among H_{123}, H_{12}, H_{13}, and H_{23} can therefore be either 0, 1, or 4, but never 2 or 3.

Additionally, if only one hypothesis is true, this cannot be H_{123}.

These logical implications between hypotheses are also known as restricted combinations

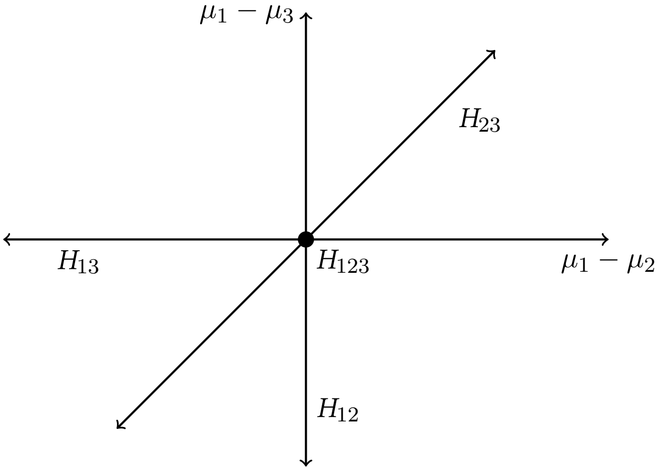


Figure: Visualization of the four hypotheses H_{12}, H_{13}, H_{23} and H_{123} in a parameter space with axes $\mu_1 - \mu_2$ and $\mu_1 - \mu_3$. Note that H_{23} is the diagonal line for which $\mu_1 - \mu_2 = \mu_1 - \mu_3$. In the origin all four hypotheses are true; elsewhere at most one.

A closed testing procedures in the three-group design take this general form:

1. Test H_{123} with a valid α-level test
2. If H_{123} was not rejected, stop; otherwise, test each of H_{12}, H_{13}, and H_{23} with a valid α-level test.

The four procedures have the following Step 1, with test statistics chosen so as to maximize power of the primary hypotheses:

1. *Classic closed testing*: H_{123} is tested with test statistic S_{123};
2. *Closed Tukey*: H_{123} is tested with test statistic $\tilde{S}_{123} = \max(S_{12}, S_{13}, S_{23})$;
3. *Closed Dunnett*: H_{123} is tested with test statistic $\tilde{S}_1 = \max(S_{12}, S_{13})$;
4. *Gatekeeping*: H_{123} is tested with test statistic S_{12}.

For H_{12}, H_{13}, and H_{23} we will always simply use the test that rejects when $S_{ij} \geq c_\alpha$

The adjusted p-value of H_{ij} in a closed testing procedure is therefore

$$\tilde{p}_{ij} = \max(p_{ij}, \tilde{p}_{123}),$$

where $\tilde{p}_{123}$ is the p-value for H_{123} in the procedure. These we can calculate for each of the four procedures as follows:

$$\begin{array}{rcl} \tilde{p}^{A}_{123} & = & p_{123}; \\ \tilde{p}^{B}_{123} & = & \min(\tilde{p}^{Tuk}_{12}, \tilde{p}^{Tuk}_{13}, \tilde{p}^{Tuk}_{23}); \\ \tilde{p}^{C}_{123} & = & \min(\tilde{p}^{Dun}_{12}, \tilde{p}^{Dun}_{13}); \\ \tilde{p}^{D}_{123} & = & p_{12}. \end{array}$$

Method	H_{12}	H_{13}	H_{23}	H_{123}
(A) Classic closed testing	0.194	0.088	0.016	0.016
(B) Closed Tukey	0.194	0.088	0.012	0.012
(C) Closed Dunnett	0.194	0.153	0.153	0.153
(D) Gatekeeping	0.194	0.194	0.194	0.194

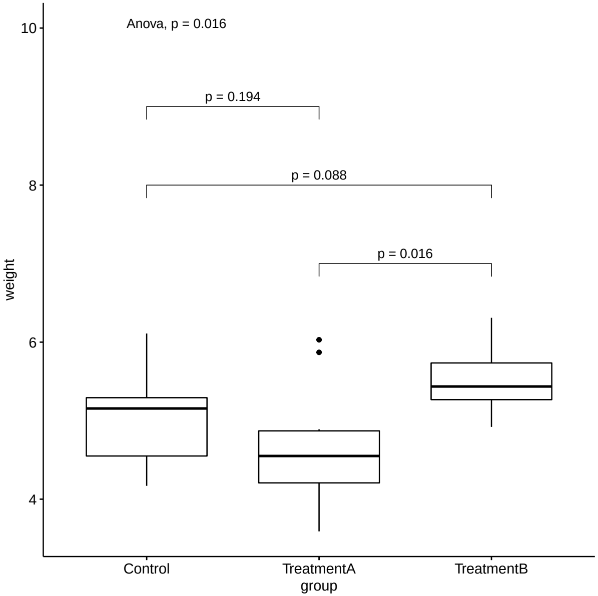

Anova, p = 0.016
p = 0.194
p = 0.088
p = 0.016
weight
10
8
6
4
Control
TreatmentA
TreatmentB
group

Table of Contents

fMRI experiment

Subjects perform mental tasks in MRI scanner
MRI measures oxygenated blood flow in brain (brain activity)

Brain activity map

Significance (*p*-value) for brain activity at each location (*voxel*)

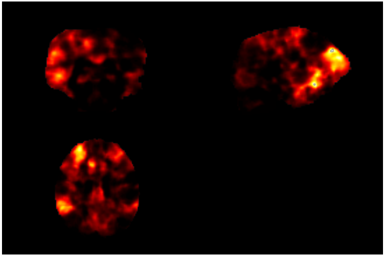
Goal

Identify emphregions of brain activity

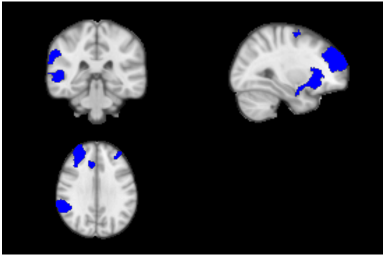
Aggregation

Micro-inferences (voxels) $\rightarrow$ larger-scale inferences (regions)

fMRI data



Brain activity map



Selection

The problem of post-selection inference

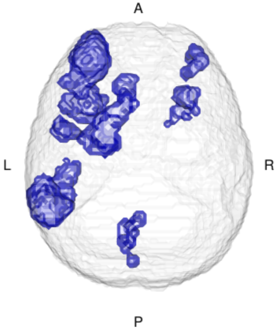
Examining the data to *select* interesting patterns,
then carrying out *inference* about the selection with the same data

Question

How to correct for overoptimism in inference due to data-driven selection?

Selected clusters

cluster ≡ contiguous voxels with $p < t = 0.0007$



9 clusters of size 2191, 1835, 1400, 698, 421, 304, 245, 232, 187

Simultaneous inference

For every selected region, return

$$\text{estimate} \quad \underbrace{[(1-\alpha)\ \text{confidence lower bound}, 100\%]}_{\text{one}-\text{sided confidence interval}}$$

for the *true discovery proportion* in the selection

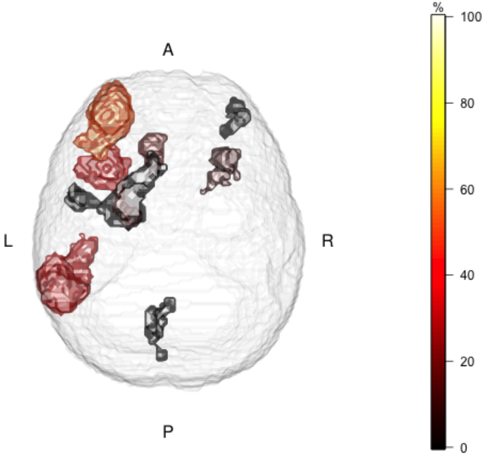
All lower bounds are *simultaneously* correct with probability $\geq 1-\alpha$

True discovery proportion

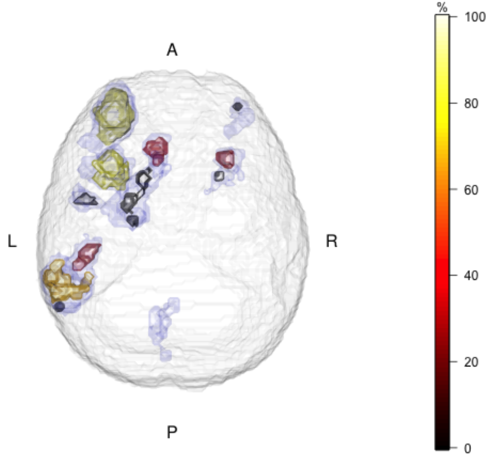
selection	*size*	$\widehat{\mathrm{TDP}}$	$[\,\underline{\mathrm{TDP}}\,,\ 100\%\,]$
S_1	2191	88%	[29% , 100%]
S_2	1835	86%	[46% , 100%]
S_3	1400	81%	[32% , 100%]
S_4	698	62%	[0% , 100%]
S_5	421	42%	[6% , 100%]
S_6	304	49%	[11% , 100%]
S_7	245	0%	[0% , 100%]
S_8	232	20%	[0% , 100%]
S_9	187	1%	[0% , 100%]

All lower bounds are correct with probability $\geq 95\%$

True discovery proportion



Zoom in



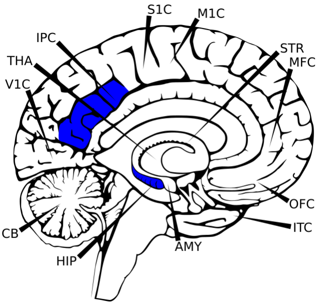
Sub-clusters

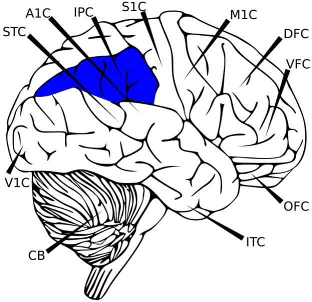
sub-cluster ≡ contiguous voxels with $p < t' = 0.00003$

selection	*threshold*	*size*	TDP
S_1	$p < t$	2191	29 %
S_1'	$p < t'$	405	66 %
S_1''	$p < t'$	133	23 %
S_1'''	$p < t'$	6	0 %
S_2	$p < t$	1835	46 %
S_2'	$p < t'$	963	86 %

$\vdots$

Domain-knowledge regions





$M = \{1, \ldots, m\}$	collection of $m = \|M\|$ voxels
$M_0 \subseteq M$	null voxels with $m_0 = \|M_0\|$ and $\pi_0 = m_0 / m$
$M_1 = M \setminus M_0$	non-null voxels with $m_1 = m - m_0$ and $\pi_1 = 1 - \pi_0$
$H_i : i \in M_0$	voxel null hypothesis with p-value $p_i, \quad i \in M$

Selection

$S \subseteq M$	selected voxels
$m_1(S) = \|M_1 \cap S\|$	number of true discoveries in the selection
$m_0(S) = \|S\| - m_1(S)$	number of false discoveries in S
$\pi_0(S) = m_0(S)/\|S\|$	false discovery proportion in S
$\pi_1(S) = 1 - \pi_0(S)$	true discovery proportion in S

Simultaneous confidence bound

$$\mathrm{P}\Big(\forall\, S \subseteq M :\ \underset{\text{lower bound}}{\underline{m}_1(S)} \ \le \underset{\text{parameter}}{m_1(S)}\Big) \ge 1-\alpha$$

Closed testing

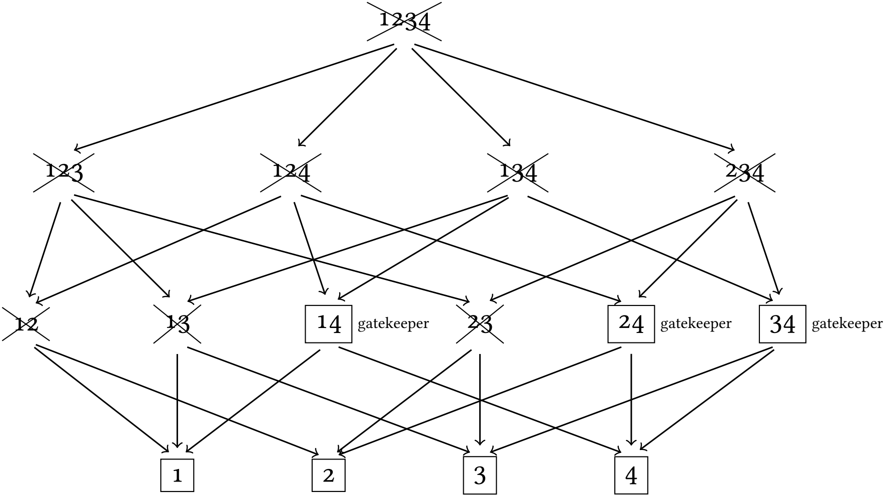
$H_1, \ldots, H_m$	elementary hypotheses
$H_S = \bigcap_{i \in S} H_i \qquad \forall\ S \subseteq M$	intersection hypotheses
$\phi_S = \mathbb{1}\{H_S \text{ rejected at level } \alpha\}$	local tests
$\tilde{\phi}_S = \min \left\{ \phi_K : S \subseteq K \subseteq M \right\}$	closed testing adjusted tests

Closed testing guarantees familywise error rate control at α over all intersection hypotheses

Four-pixel brain

1	2
3	4

Closed testing rejections



Confidence bound

$$\underline{m}_1(S) = |S| - \max_{K \subseteq S} \left\{ |K| : \tilde{\phi}_K = 0 \right\}$$

The size of S minus the size of the largest subset of S for which the corresponding intersection hypothesis is not rejected by closed testing

S	$\underline{m}_1(S)$	$\underline{\pi}_1(S)$
$\{1\}$	0	0%
$\{2\}$	0	0%
$\{3\}$	0	0%
$\{4\}$	0	0%
$\{1, 2\}$	1	50%
$\{1, 3\}$	1	50%
$\{1, 4\}$	0	0%
$\{2, 3\}$	1	50%
$\{2, 4\}$	0	0%
$\{3, 4\}$	0	0%
$\{1, 2, 3\}$	2	66.6%
$\{1, 2, 4\}$	1	33.3%
$\{1, 3, 4\}$	1	33.3%
$\{2, 3, 4\}$	1	33.3%
$\{1, 2, 3, 4\}$	2	50%

Closed testing bottleneck

The required number of tests is 2^m

Shortcut

Computation time can be reduced to polynomial time by specific choice of local tests

Simes test

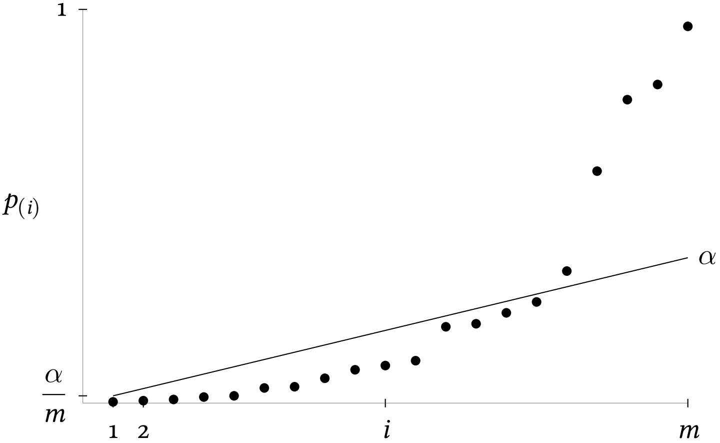
Simes test for H_S

$$\phi_S = \mathbb{1}\Big\{ \bigcup_{i \in S} \{ p_{(i\,:\,S)} \leq \frac{i\alpha}{|S|} \} \Big\}$$

where $p_{(i\,:\,S)}$ is the ith smallest p-value in $\{p_i : i \in S\}$

Assumption
Simes inequality holds for null p-values

$$\mathrm{P}\left(\bigcap_{i=1}^{m_0} \left\{ p_{(i\,:\,M_0)} > \frac{i\alpha}{m_0} \right\} \right) \geq 1 - \alpha$$



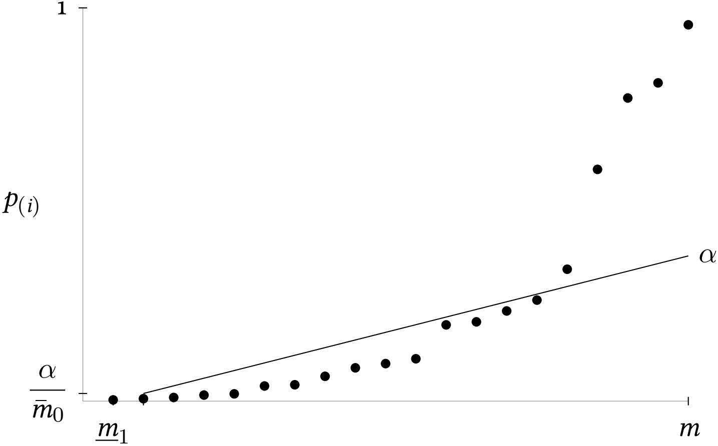
$m_0 = m$? No. Then $m_0 \leq \bar{m}_0 = m - 1$

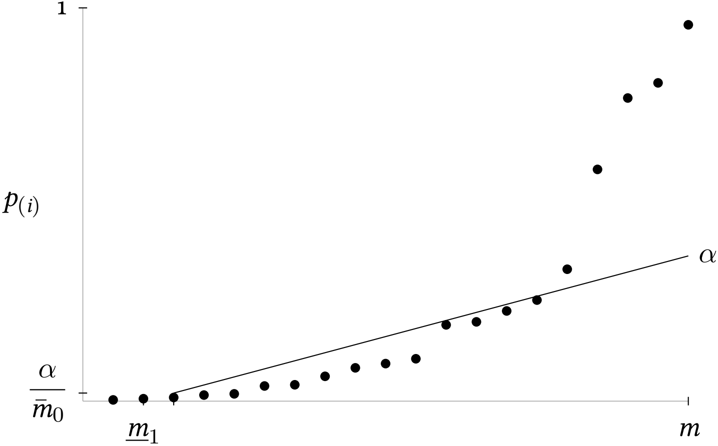
Upper bound for m_0

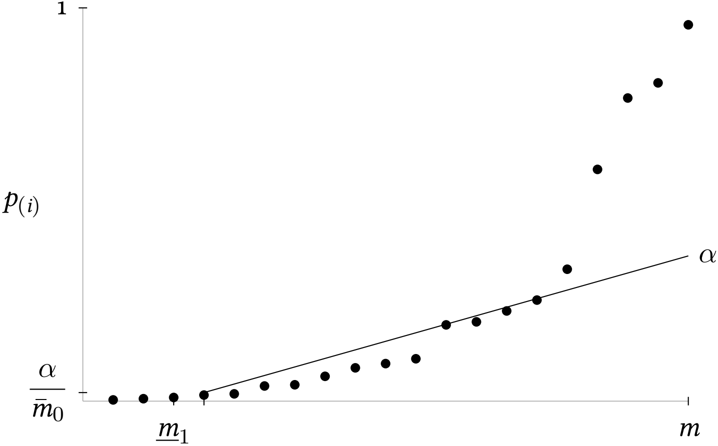
$$\bar{m}_0 = \max\left\{0 \leq k \leq m : \bigcap_{i=1}^{k} \left\{p_{(m-k+i)} > \frac{i\alpha}{k}\right\}\right\}$$

so the lower bound for the overall number of true discoveries is

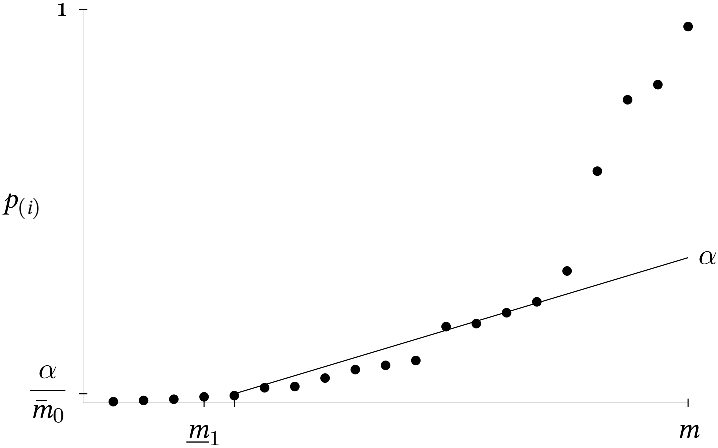
$$\underline{m}_1 = m - \bar{m}_0$$

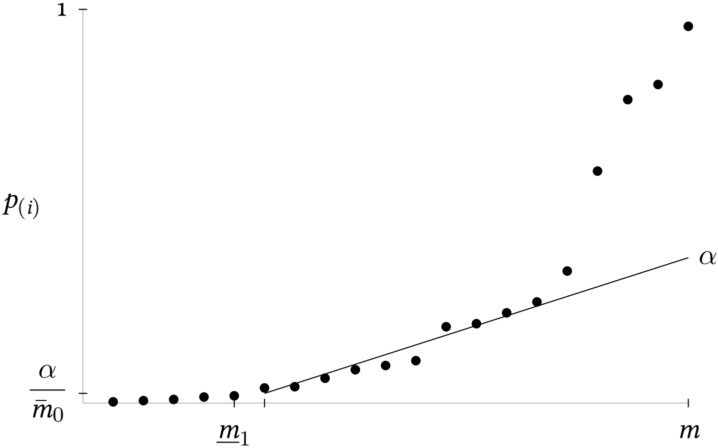

1
$p_{(i)}$
α
$\frac{\alpha}{\bar{m}_0}$
$\underline{m}_1$
m

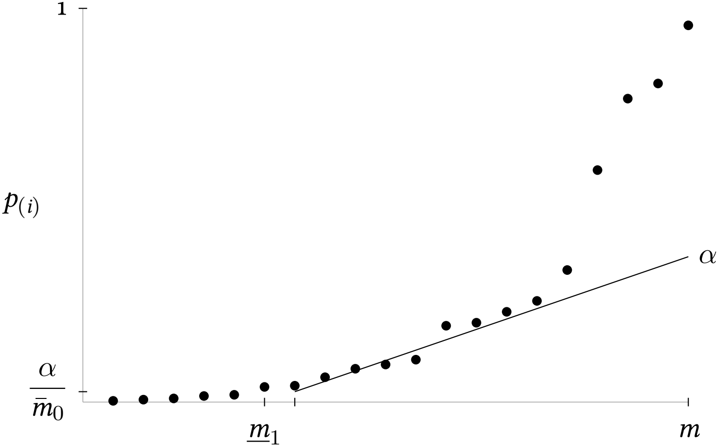

1
$p_{(i)}$
α
$\frac{\alpha}{\bar{m}_0}$
$\underline{m}_1$
m



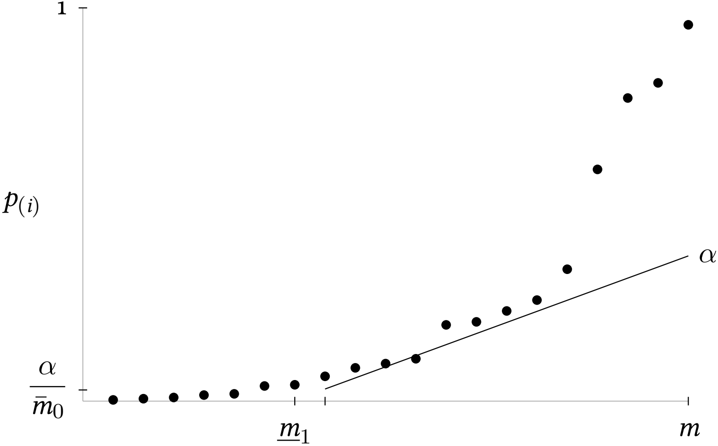
1
$p_{(i)}$
α
$\frac{\alpha}{\bar{m}_0}$
$\underline{m}_1$
m

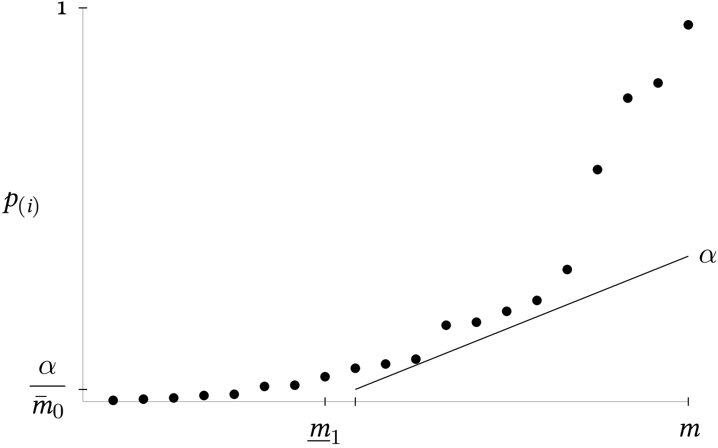

1
$p_{(i)}$
α
$\frac{\alpha}{\bar{m}_0}$
$\underline{m}_1$
m


1
$p_{(i)}$
α
$\frac{\alpha}{\bar{m}_0}$
$\underline{m}_1$
m

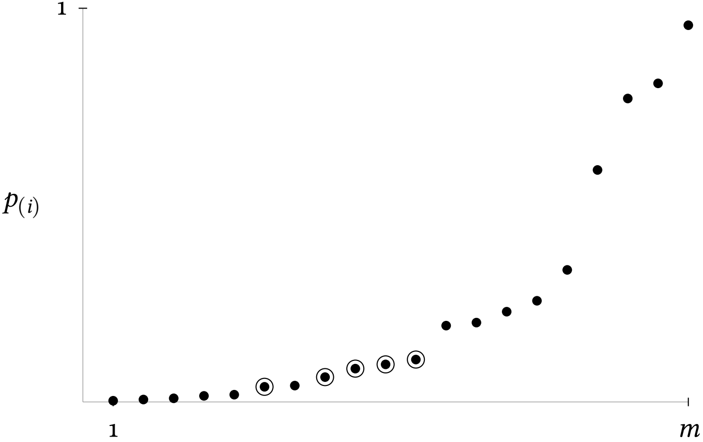


1
$p_{(i)}$
$\frac{\alpha}{\bar{m}_0}$
α
$\underline{m}_1$
m


1
$p_{(i)}$
$\frac{\alpha}{\bar{m}_0}$
$\underline{m}_1$
m
α


1
$p_{(i)}$
α
$\frac{\alpha}{\bar{m}_0}$
$\underline{m}_1$
m

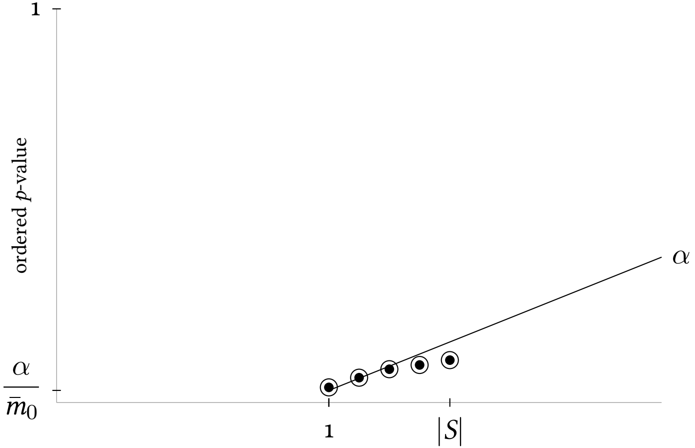
Arbitrary selection

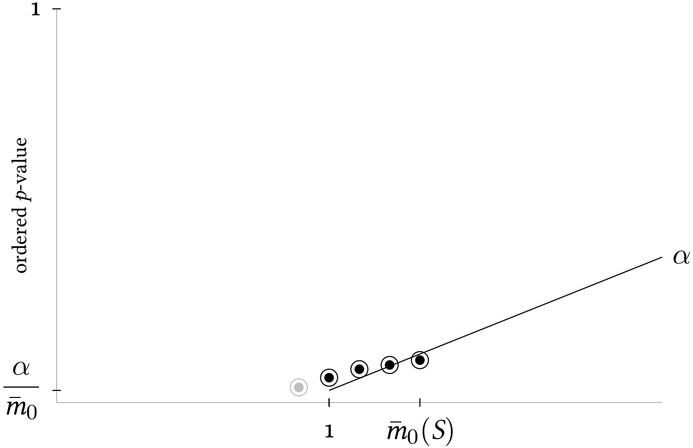


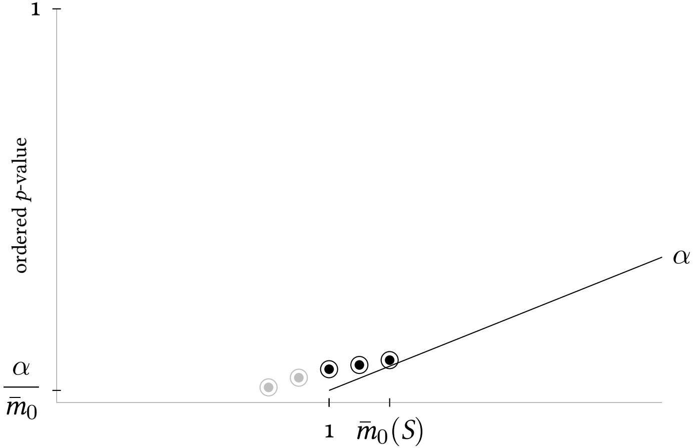
$S \subseteq M$

Confidence bound

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


1
ordered p-value
$\frac{\alpha}{\bar{m}_0}$
α
1
$|S|$


1
ordered p-value
α
$\frac{\alpha}{\bar{m}_0}$
1
$\bar{m}_0(S)$


1
ordered p-value
$\frac{\alpha}{\bar{m}_0}$
α
1
$\bar{m}_0(S)$

Algorithm

	Operation	*Complexity*
1	Sort the p-values	$O(m \log m)$
2	Compute $\bar{m}_0$	$O(m)$
3	Compute $\underline{m}_1(S)$	$O(\|S\|)$

- $\bar{m}_0$ in linear time

  Meijer, Krebs, Goeman (2019)

- Implemented in the R package `hommel`

Relationship to Hommel (FWER)

- Reject the hypotheses with indexes in

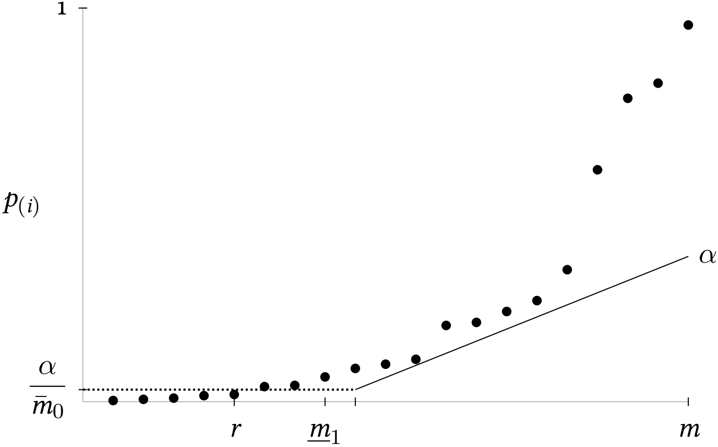
$$R = \left\{ i \in M : p_i \leq \frac{\alpha}{\bar{m}_0} \right\}$$

  with familywise error rate control at α

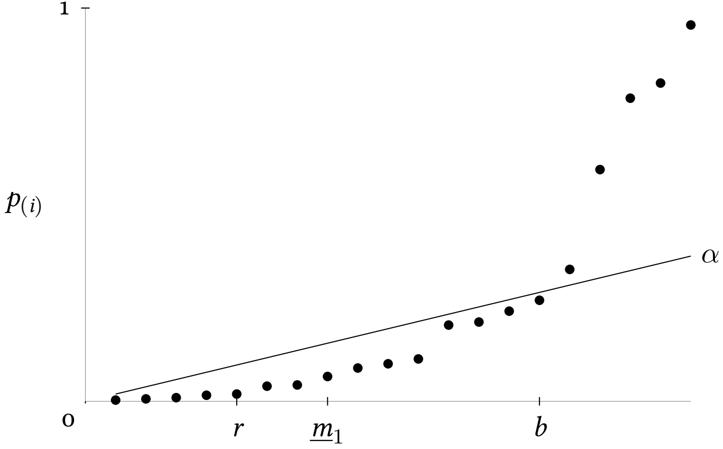
- Voxels in R represent *localized true discoveries*

$$\underline{m}_1(R) = |R| = r$$

Hommel rejections



Relationship to Benjamini-Hochberg (FDR)



Large-scale testing

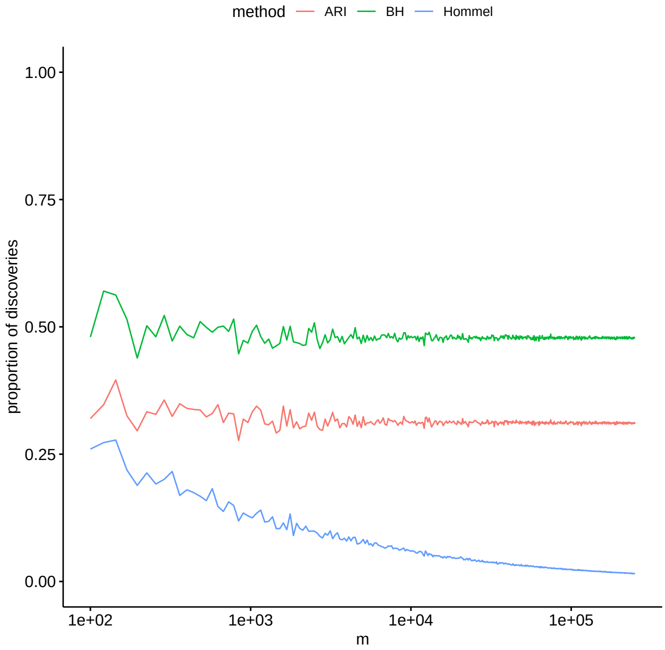
Assume $p_1, \ldots, p_m \overset{i.i.d.}{\sim} F$
with a mixture distribution $F(u) = \pi_0 u + \pi_1 F_1(u)$

Fix $\alpha \in (0,1)$. As the number of hypotheses $m \to \infty$

$$\operatorname*{plim}_{m\to\infty} \frac{r}{m} = 0 \qquad \operatorname*{plim}_{m\to\infty} \frac{m_1}{m} = k > 0 \qquad \operatorname*{plim}_{m\to\infty} \frac{b}{m} = k' > 0$$

if a minimal level of signal is present[1]

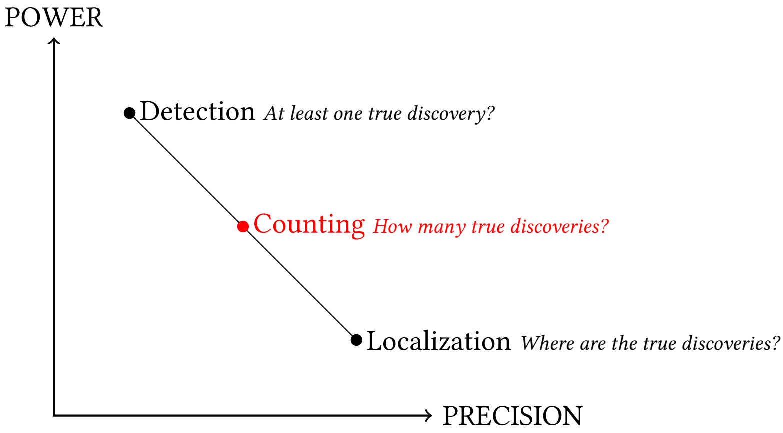
[1] "criticality" of Chi (2007), i.e. if $F(u\alpha) > u$ for at least one $0 \leq u < 1$


method
ARI
BH
Hommel
1.00
0.75
0.50
0.25
0.00
proportion of discoveries
1e+02
1e+03
1e+04
1e+05
m

Localized true discoveries

S	$\lvert S\rvert$	$\underline{\pi_1}(S)$	$\lvert S \cap R\rvert / \lvert S\rvert$
S_1	2191	29%	0.3%
S_2	1835	46%	4%
S_3	1400	32%	6%
S_4	698	0%	0%
S_5	421	6%	0%
S_6	304	11%	0%
S_7	245	0%	0%
S_8	232	0%	0%
S_9	187	0%	0%

Trade-off


POWER
Detection *At least one true discovery?*
Counting *How many true discoveries?*
Localization *Where are the true discoveries?*
PRECISION

The less specific the question is, the more power to answer it