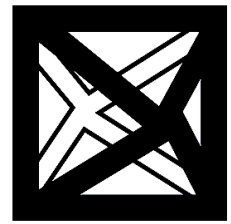
Multiple testing

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



XXXVII cycle

Outline

Global testing

Error rates

Methods for familywise error rate control

Methods for false discovery rate control

Main references

- Candés (2022) Stats 300C - Theory of Statistics. Lectures 1-7 https://candes.su.domains/teaching/stats300c/index.html
- Goeman and Solari (2014) Multiple Hypothesis Testing in Genomics. Statistics in Medicine, 33, 1946–78.

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_m \end{pmatrix} \sim N_m \left(\begin{pmatrix} \mu_1 \\ \vdots \\ \mu_m \end{pmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdot & \sigma_{1m} \\ \sigma_{21} & \sigma_2^2 & \cdot & \sigma_{2m} \\ \cdot & \cdot & \cdot & \cdot \\ \sigma_{m1} & \cdot & \cdot & \sigma_m^2 \end{bmatrix} \right)$$

The parameter of interest is $\mathrm{E}(Y) = \mu$, where $\mu_j = 0$ means "no effect" and $\mu_j \neq 0$ means "effect" in the jth component

The nuisance parameter is the variance/covariance matrix $\mathrm{Var}(Y) = \Sigma$

Three questions

1. *Detecting effects*: There is at least one μ_j different from 0?
2. *Counting effects*: How many μ_j are different from 0?
3. *Identifying effects*: Which μ_j are different from 0?

Table of Contents

Global null hypothesis

$$H_0 : \mu = 0, \text{ i.e. } \bigcap_{j=1}^{m} \{\mu_j = 0\} \text{ vs } H_1 : \mu \neq 0, \text{ i.e. } \bigcup_{j=1}^{m} \{\mu_j \neq 0\}$$

For simplicity, consider $\Sigma = I_m$ and the one-sided alternative

$$H_0 : \bigcap_{j=1}^{m} \{\mu_j = 0\} \text{ vs } H_1 : \bigcup_{j=1}^{m} \{\mu_j > 0\}$$

MaxT test

$(Y_1, \ldots, Y_m)^t \overset{H_0}{\sim} N_m(0, I_m)$

$T_{\max} = \max(Y_1, \ldots, Y_m)$

The critical value $t_{1-\alpha}$ of $T_{\max}$ is

$$\mathrm{pr}_0(T_{\max} \geq t_{1-\alpha}) = \alpha$$

where $t_{1-\alpha}$ is the $1-\alpha$ quantile of the distribution of the maximum of m independent standard normal variables

$$\int_{t_{1-\alpha}}^{\infty} m\phi(y)\Phi(y)^{m-1} dy = \alpha$$

where ϕ and Φ are the density and cdf of $N(0,1)$

Bonferroni approximation

We can replace $t_{1-\alpha}$ by $z_{1-\frac{\alpha}{m}}$

$$\begin{aligned}
\mathrm{pr}_0(T_{\max} \geq z_{1-\frac{\alpha}{m}}) &= \mathrm{pr}_0\Big(\bigcup_{j=1}^{m}\{Y_j \geq z_{1-\frac{\alpha}{m}}\}\Big) \\
&\leq \sum_{j=1}^{m}\mathrm{pr}_0(Y_j \geq z_{1-\frac{\alpha}{m}}) = m\frac{\alpha}{m} = \alpha
\end{aligned}$$

The union bound might seem crude, but with independent Y_js the size of the test is very near α

$$\begin{aligned}
\mathrm{pr}_0(T_{\max} \geq z_{1-\frac{\alpha}{m}}) &= 1 - \prod_{j=1}^{m}\mathrm{pr}_0(Y_j < z_{1-\frac{\alpha}{m}}) \\
&= 1 - \left(1 - \frac{\alpha}{m}\right)^m \overset{m\to\infty}{\to} 1 - e^{-\alpha}
\end{aligned}$$

For $\alpha = 0.05$, $1 - e^{-\alpha} = 0.0487$

Magnitude of Bonferroni critical value

How large is the threshold $z_{1-\frac{\alpha}{m}}$? For large m

$$\begin{aligned} z_{1-\frac{\alpha}{m}} &\approx \sqrt{2\log m} - \frac{\log(2\log m) + \log 2\pi}{2\sqrt{2\log m}} \\ &\approx \sqrt{2\log m} \end{aligned}$$

with no dependence on α

$$\frac{\phi(t)}{t}(\frac{t^2}{t^2+1}) \leq \text{pr}(N(0,1) > t) \leq \frac{\phi(t)}{t}$$

where $\phi(t)$ is the probability density function of $N(0,1)$. This result implies that for large t, $\frac{\phi(t)}{t}$ is a good approximation to the normal tail probability. Let $z^* = z_{1-\frac{\alpha}{m}}$. We have
$\frac{\alpha}{m} = \text{pr}(N(0,1) > z_{1-\frac{\alpha}{m}}) \approx \frac{\phi(z^*)}{z^*}$, which implies
$\alpha/m \approx \frac{1}{z^*\sqrt{2\pi}} e^{-\frac{(z^*)^2}{2}}$. Taking the logarithm

$$\log m \approx \frac{1}{2}\log(2\pi) + \frac{1}{2}(z^*)^2 + \log(z^*) + \log(\alpha)$$

Note that z^* is increasing in m, i.e. $m \to \infty$ induces $z^* \to \infty$. As $\frac{1}{2}\log(2\pi) + \log(z^*) + \log(\alpha)$ is negligible compared to $(z^*)^2$ when m goes to ∞, it gives

$$z_{1-\frac{\alpha}{m}} \approx \sqrt{2\log m}$$

Needle in a haystack problem

$$H_0 : \mu_j = 0 \text{ for all } j = 1, \ldots, m$$
$$H_1 : \mu_j = c_m > 0, \mu_k = 0 \text{ for } k \neq j$$

What is the limiting power of Bonferroni test?

$$\lim_{m \to \infty} \text{pr}_1\left(T_{\max} > z_{1-\frac{\alpha}{m}}\right)$$

Assume without loss of generality that $\mu_1 = c_m$ and let $\epsilon > 0$ small.

Suppose $c_m > (1+\epsilon)\sqrt{2\log m}$. Then, for $m \to \infty$

$$\mathrm{pr}_1(T_{\max} > z_{1-\frac{\alpha}{m}}) \geq \mathrm{pr}_1(Y_1 > z_{1-\frac{\alpha}{m}}) = \mathrm{pr}(N(0,1) > z_{1-\frac{\alpha}{m}} - c_m) \to 1$$

Suppose $c_m < (1-\epsilon)\sqrt{2\log m}$. Then for $m \to \infty$

$$\begin{aligned}
\mathrm{pr}_1(T_{\max} > z_{1-\frac{\alpha}{m}}) &\leq \mathrm{pr}(Y_1 > z_{1-\frac{\alpha}{m}}) + \mathrm{pr}(\max_{j>1} Y_j > z_{1-\frac{\alpha}{m}}) \\
&= \mathrm{pr}(N(0,1) > z_{1-\frac{\alpha}{m}} - c_m) + \mathrm{pr}(\max_{j>1} Y_j > z_{1-\frac{\alpha}{m}}) \\
&\to 0 + (1 - e^{-\alpha})
\end{aligned}$$

and Bonferroni test has no power

Can we do better than this test? The optimal test given by Neyman-Pearson lemma for the simple hypotheses

$$H_0 : \mu_j = 0 \text{ for all } j$$
$$H_1 : \{\mu_j\} \sim \pi$$

where π selects a coordinate $\mathcal{J}$ uniformly and sets $\mu_{\mathcal{J}} = c_m$ with all other $\mu_j = 0$.

However, even the optimal likelihood ratio test fails when $c_m = (1 - \epsilon)\sqrt{2 \log m}$:

$$\text{pr}_1(\text{type II error}) \to 1 - \alpha$$

In summary, there is no test that is asymptotically able to distinguish between the null and alternative hypotheses when the mean of the needle in the haystack, c_m, is smaller than the $\sqrt{2 \log m}$ threshold

MinP test

Let $p_j = 1 - \Phi(Y_j)$ be the jth p-value, $j = 1, \ldots m$

Assume $p_1, \ldots, p_m$ i.i.d. Uniform$(0, 1)$ under H_0

The MinP test is based on the minimum p-value

$$p_{\min} = \min(p_1, \ldots, p_m) \overset{H_0}{\sim} \text{Beta}(1, m)$$

The MinP test rejects H_0 if $p_{\min} \leq 1 - (1 - \alpha)^{\frac{1}{m}}$ and has size α:

$$\begin{aligned} \text{pr}_0(p_{\min} \leq 1 - (1 - \alpha)^{\frac{1}{m}}) &= 1 - \text{pr}_0\left(\bigcap_{j=1}^{m} \{p_j > 1 - (1 - \alpha)^{\frac{1}{m}}\} \right) \\ &= 1 - [(1 - \alpha)^{\frac{1}{m}}]^m = \alpha \end{aligned}$$

Bonferroni method

Assume that p_j is a valid p-value under H_0, i.e.

$\text{pr}(P_j \leq u; H_0) \leq u$ for all $u \in (0, 1)$

The Bonferroni method (Bonferroni, 1936) rejects H_0 if $p_{\min} \leq \alpha/m$:

$$\begin{aligned} \text{pr}_0(p_{\min} \leq \alpha/m) &= \text{pr}_0\Big(\bigcup_{j=1}^{m}\{p_j \leq \alpha/m\}\Big) \\ &\leq \sum_{j=1}^{m}\text{pr}_0\Big(p_j \leq \alpha/m\Big) \\ &= m\frac{\alpha}{m} = \alpha \end{aligned}$$

An appealing property of Bonferroni's method is that it controls the Type I error rate even when the p-values $p_1, \ldots, p_m$ are arbitrarily dependent.

Simes test

Assume $p_1, \ldots, p_m$ i.i.d. Uniform$(0, 1)$ under H_0

Sort the p-values

$$p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$$

The null distribution of jth ordered p-value is

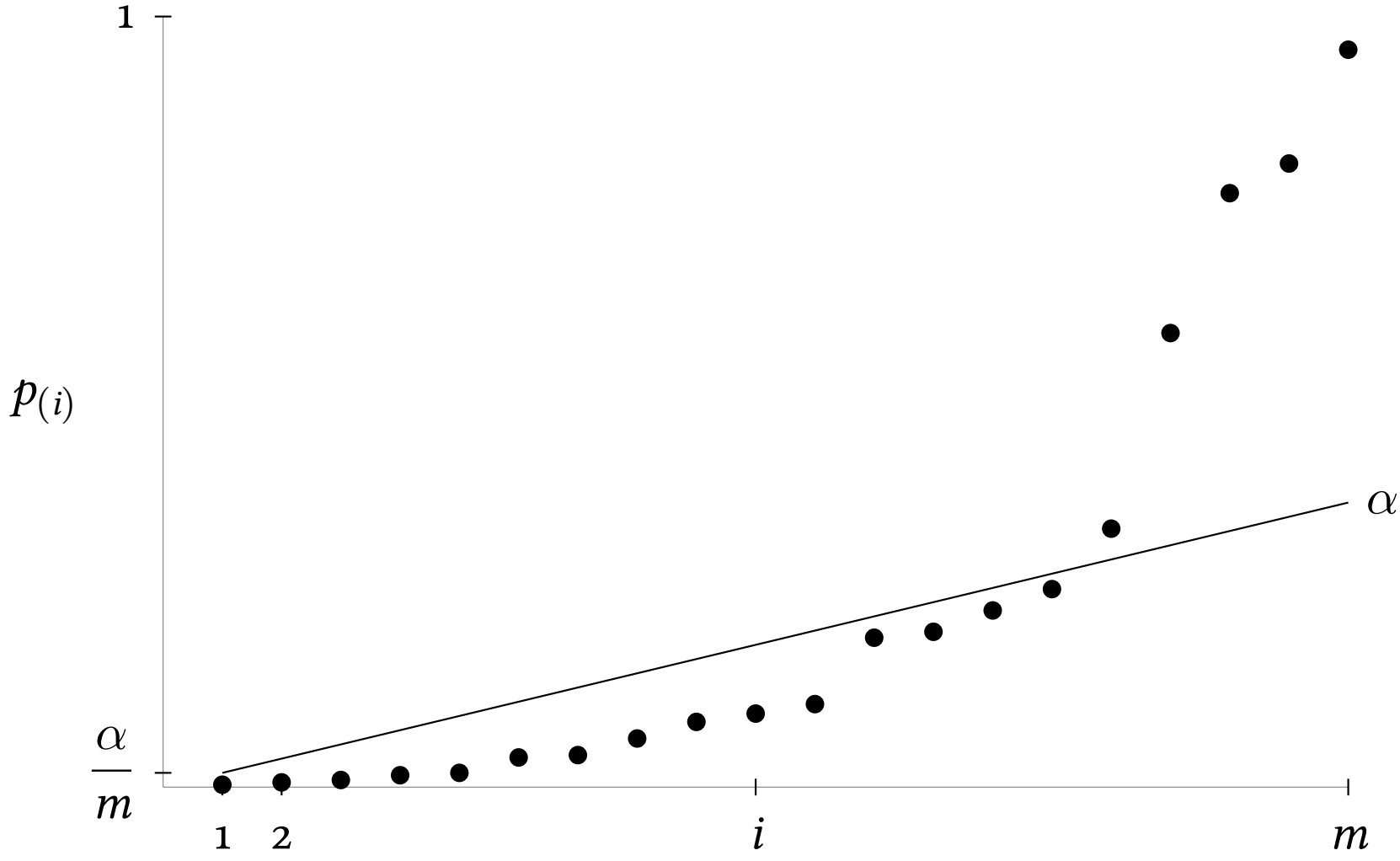
$$p_{(i)} \overset{H_0}{\sim} \text{Beta}(i, m - i + 1)$$

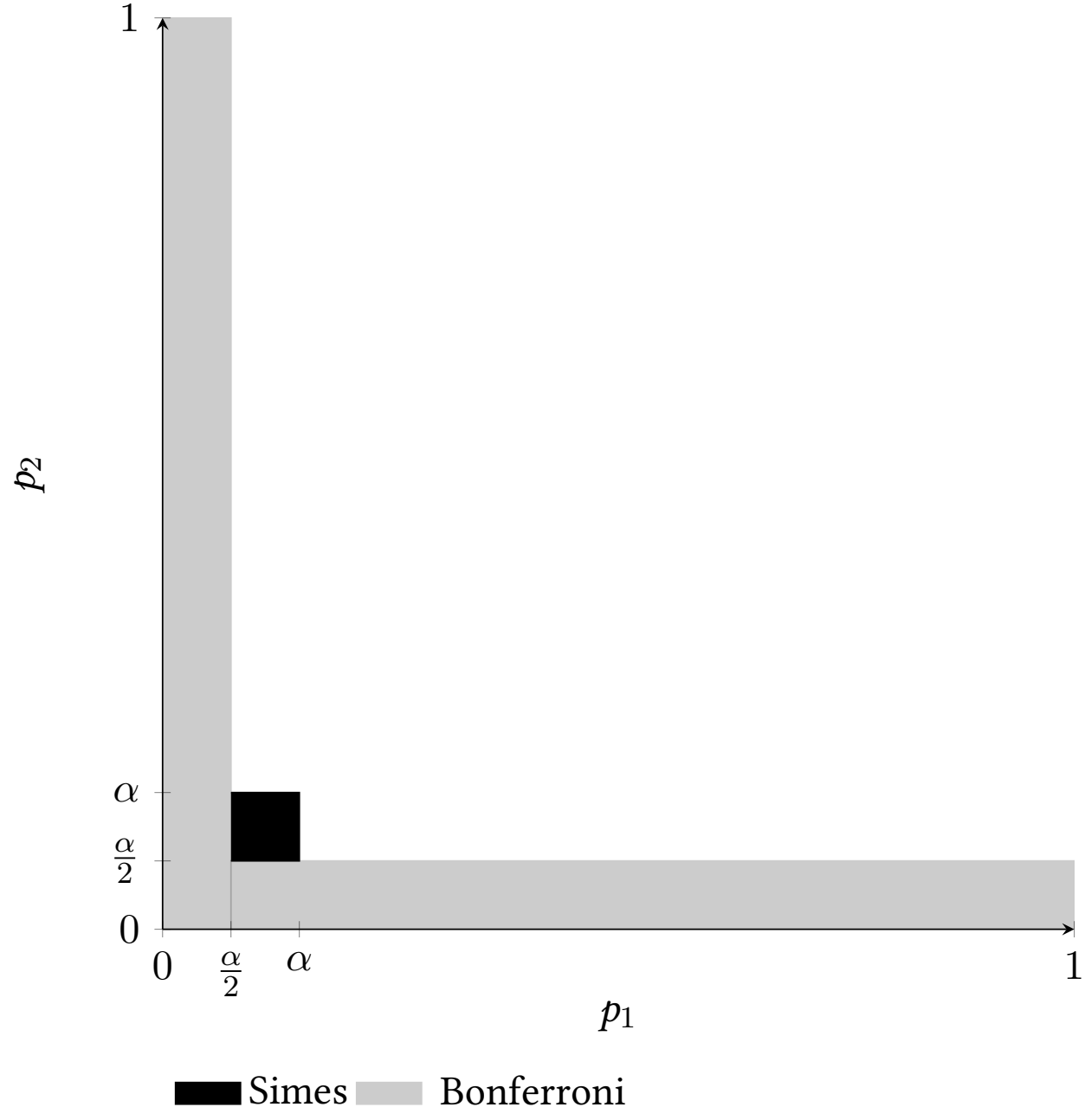
Simes test (Simes, 1986) rejects H_0 if

$$\exists\, j : p_{(i)} \leq \frac{\alpha i}{m}$$

with p-value

$$p_{\text{Simes}} = \min_{i=1,\ldots,m} \left\{ p_{(i)} \frac{m}{i} \right\} \overset{H_0}{\sim} \text{Uniform}(0, 1)$$


1
$p_{(i)}$
α
$\frac{\alpha}{m}$
1
2
i
m


1
p_2
α
$\frac{\alpha}{2}$
0
0
$\frac{\alpha}{2}$
α
1
p_1
Simes
Bonferroni

Kolmogorov-Smirnov test

Empirical cdf $\hat{F}_m(t) = \dfrac{\sum_{j=1}^{m} \mathbb{1}\{p_j \leq t\}}{m}$ for $t \in [0,1]$.

Assume $p_1, \ldots, p_m$ i.i.d. Uniform$(0,1)$ under H_0. Then

$$m\hat{F}_m(t) \overset{H_0}{\sim} \text{Binomial}(m, t)$$

Kolmogorov (one-sided) test statistic (Kolmogorov, 1933) is

$$T_{\text{KS}} = \sup_{t \in (0,1)} \{\hat{F}_m(t) - t\}$$

A useful inequality developed by Massart (1990) shows that

$$\text{pr}_0(T_{\text{KS}} \geq u) \leq e^{-2mu^2}$$

for $u \geq \sqrt{\log 2/2m}$

Tukey's higher criticism

Number of significant tests at level α

$$\frac{\text{observed} - \text{expected}}{\text{standard deviation}} = \frac{m\hat{F}_m(\alpha) - m\alpha}{\sqrt{m\alpha(1-\alpha)}}$$

Tukey's higher criticism statistic (Tukey, 1976; Donoho and Jin, 2004) is

$$T_{\text{hc}} = \max_{\alpha \leq \alpha_0} \frac{\hat{F}_m(\alpha) - \alpha}{\sqrt{\alpha(1-\alpha)/m}}$$

Sparse mixture

Assume that

$$\begin{aligned} H_0 &: Y_j \overset{i.i.d}{\sim} N(0,1) \\ H_1 &: Y_j \overset{i.i.d}{\sim} \pi_0 N(0,1) + \pi_1 N(\mu,1) \end{aligned}$$

where $\pi_0 + \pi_1 = 1$

Asymptotic analysis with

$$\begin{aligned} \pi_1(m) &= m^{-\beta} & \frac{1}{2} < \beta < 1 \\ \mu(m) &= \sqrt{2r \log m} & 0 < r < 1 \end{aligned}$$

Needle in a haystack problem: $\beta = 1$ and $r = 1$

If π_1 and μ were known, then the optimal test would be the likelihood ratio test

Detection boundary

$$\rho(\beta) = \begin{cases} \beta - \frac{1}{2} & \text{if } \frac{1}{2} < \beta \leq \frac{3}{4} \\ (1 - \sqrt{1-\beta})^2 & \text{if } \frac{3}{4} \leq \beta < 1 \end{cases}$$

If $r > \rho(\beta)$, then the Neyman-Pearson optimal test has full power. Higher criticism also has full power, i.e.

$$\text{pr}_1(\text{reject } H_0) \to 1 \quad m \to \infty$$

without knowledge of π_1 and/or μ

If $r < \rho(\beta)$, then the Neyman-Pearson optimal test has no power.

Bonferroni method has suboptimal threshold if $\beta \in (1/2, 3/4)$

$$\rho_{\text{Bonferroni}}(\beta) = \quad (1 - \sqrt{1-\beta})^2 \quad \text{if } \frac{1}{2} < \beta < 1$$

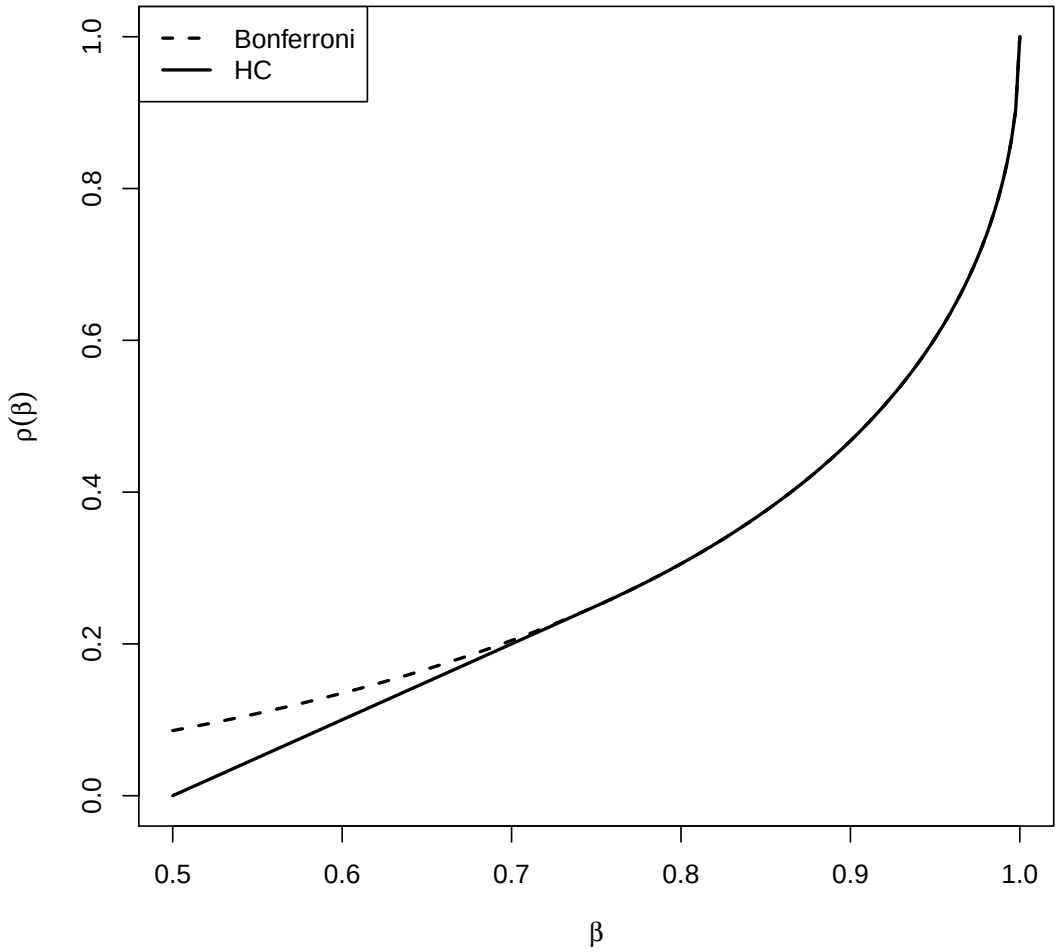

Bonferroni
HC
ρ(β)
β
0.0
0.2
0.4
0.6
0.8
1.0
0.5
0.6
0.7
0.8
0.9
1.0

Table of Contents

In a single test, the probability of making a type I error is bounded by α, conventionally set at 0.05

Problems arise, however, when researchers do not perform a single hypothesis test but many of them

There are many ways of dealing with type I errors. We will focus on three types of multiple testing methods:

1. those that control the *FamilyWise Error Rate*
2. those that control the *False Discovery Rate*
3. those that estimate the *False Discovery Proportion* or make confidence intervals for it

Rejections

Suppose we have a collection $\mathcal{H} = \{H_1, \ldots, H_m\}$ of m null hypotheses.

An unknown number m_0 of these hypotheses is true, whereas the other $m_1 = m - m_0$ is false. The proportion of true hypotheses is $\pi_0 = m_0/m$

The collection of true hypotheses is $\mathcal{T} \subseteq \mathcal{H}$ and of false hypotheses is $\mathcal{F} = \mathcal{H} \setminus \mathcal{T}$

The goal of a multiple testing procedure is to choose a collection $\mathcal{R} \subseteq \{H_1, \ldots, H_m\}$ of hypotheses to reject. If we have p-values $p_1, \ldots, p_m$ for $H_1, \ldots, H_m$, a natural choice is

$$\mathcal{R} = \{H_i : p_i \leq c\}$$

rejecting all hypotheses with a p-value below a critical value c

Errors

Ideally, the set of rejected hypotheses $\mathcal{R}$ should coincide with the set $\mathcal{F}$ of false hypotheses as much as possible. However, two types of error can be made:

Type I errors: true hypotheses that we rejected, i.e. $\mathcal{R} \cap \mathcal{T}$

Type II errors: false hypotheses that we failed to reject, i.e. $\mathcal{F} \setminus \mathcal{R}$

Rejected hypotheses are sometimes called *discoveries*, hence the terms *true discovery* and *false discovery* are sometimes used for correct and incorrect rejections

Type I errors

Type I errors are traditionally considered more problematic than type II errors

If a rejected hypothesis allows publication of a scientific finding, a type I error brings a false discovery, and the risk of publication of a potentially misleading scientific result

Type II errors, on the other hand, mean missing out on a scientific result. Although unfortunate for the individual researcher, the latter is, in comparison, less harmful to scientific research as a whole

2×2 table

We can summarize the numbers of errors in a contingency table:

	true	false	total
rejected	V	U	R
not rejected	$m_0 - V$	$m_1 - U$	$m - R$
total	m_0	m_1	m

We can observe m and $R = |\mathcal{R}|$, but all quantities in the first two columns of the table are unobservable

False Discovery Proportion

The False Discovery Proportion (FDP) Q is defined as

$$Q = \frac{V}{\max(R, 1)} = \begin{cases} V/R & \text{if } R > 0 \\ 0 & \text{otherwise,} \end{cases}$$

the proportion of false rejections among all rejections, defined as o if no rejections are made

FamilyWise Error Rate and False Discovery Rate

$$\text{FWER} = \text{pr}(V > 0) = \text{pr}(Q > 0)$$

the probability that the rejections contains any Type I error

$$\text{FDR} = \text{E}(Q)$$

the expected proportion of Type I errors among the rejections

We say that FWER or FDR is *controlled* at level α when the set $\mathcal{R}$ is chosen in such a way that the corresponding aspect of the distribution of Q is guaranteed to be at most α, i.e.

$$\text{FWER} \leq \alpha \quad \text{or} \quad \text{FDR} \leq \alpha$$

FWER $\geq$ FDR

The two error rates FDR and FWER are related. Because $0 \leq Q \leq 1$, we have $Q \leq \mathbb{1}\{Q > 0\}$ and

$$\mathrm{E}(Q) \leq \mathrm{P}(Q > 0)$$

which means that FWER control implies FDR control

If all hypotheses are true, FDR and FWER are identical; because $R = V$ in this case, Q is a Bernoulli variable, and

$$\mathrm{E}(Q) = \mathrm{P}(Q > 0)$$

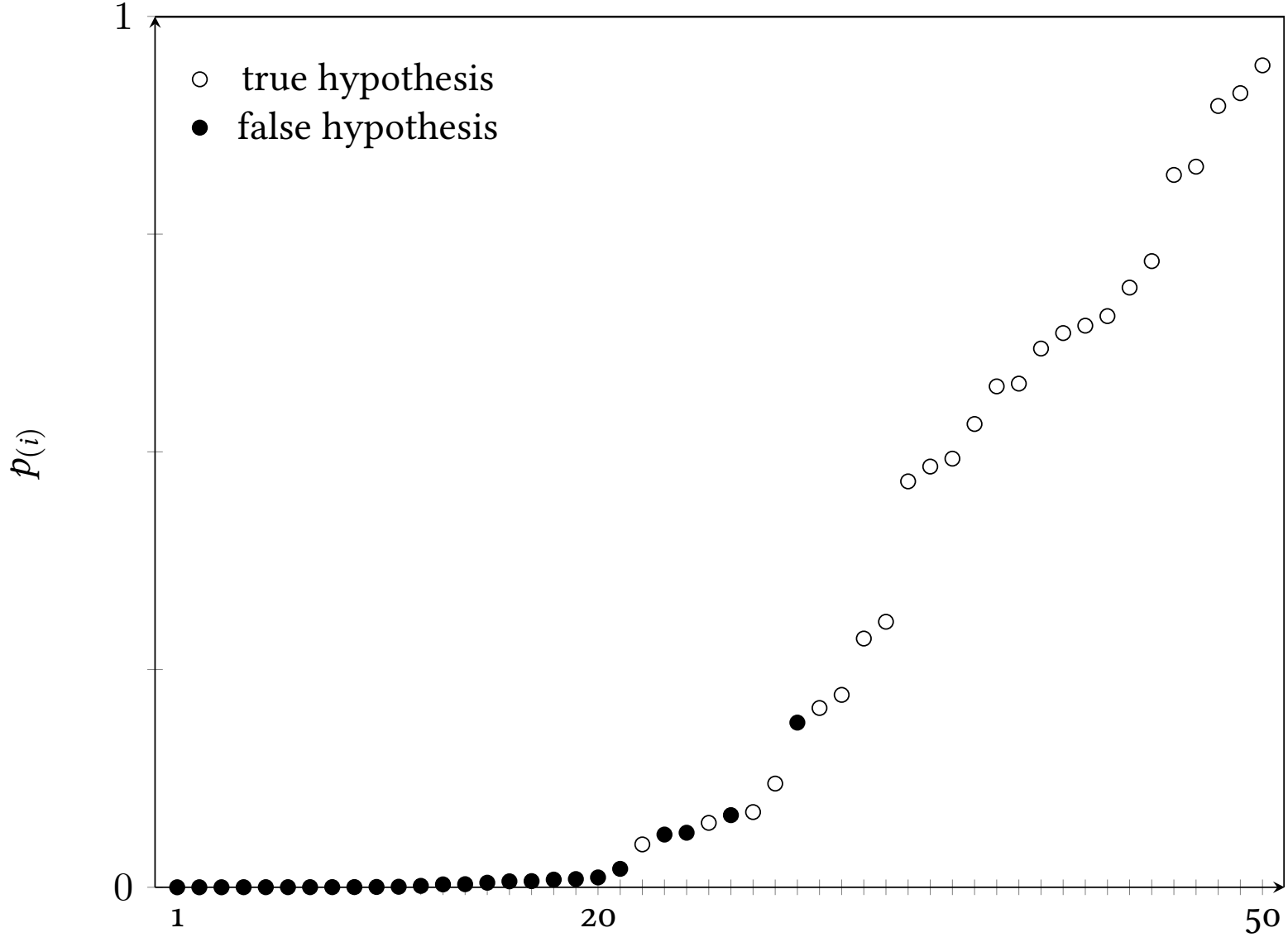
Both FDR and FWER are proper generalizations of the concept of Type I error to multiple hypotheses: if $m = 1$, the two error rates are identical and equal to the Type I error rate

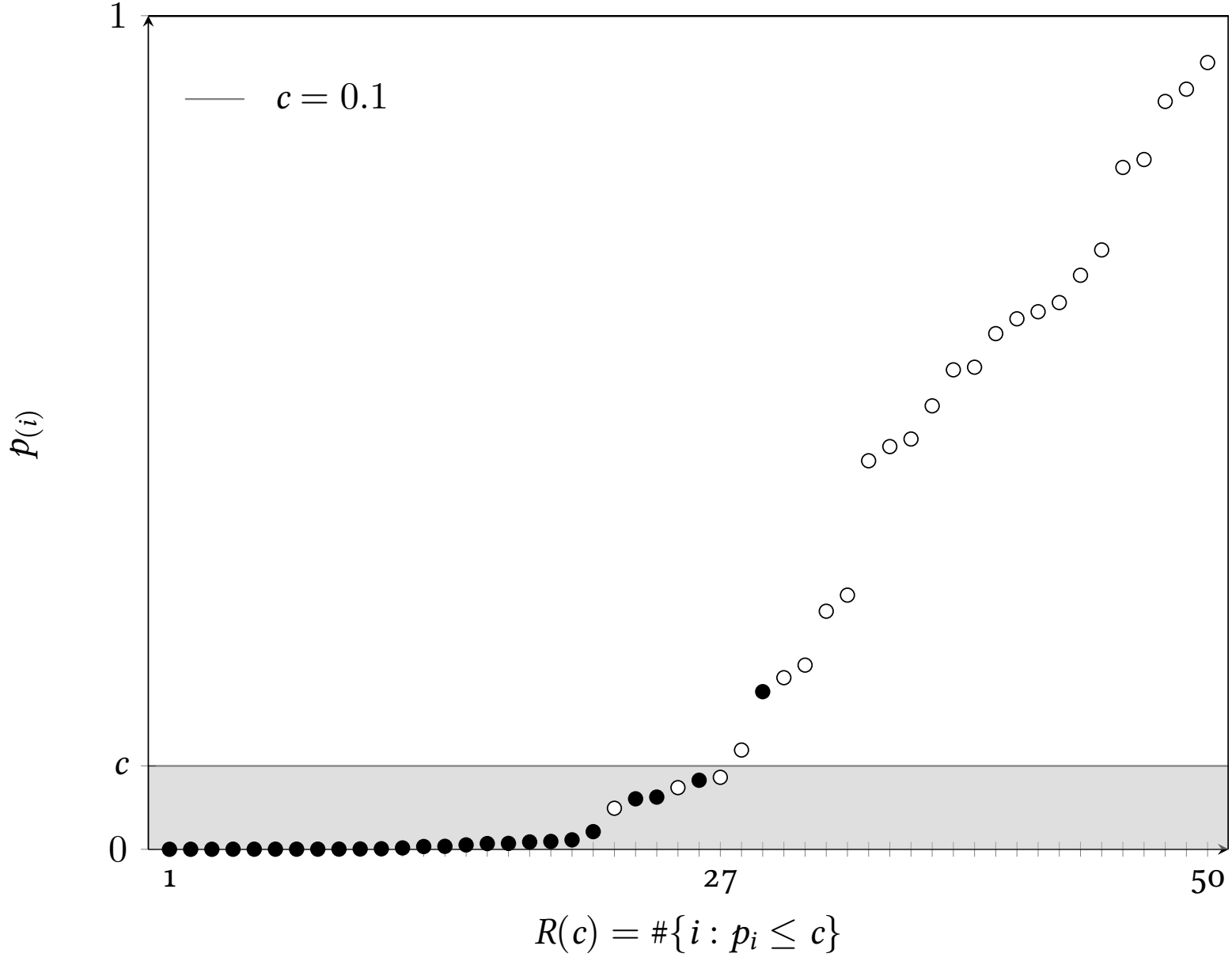
p.adjust example

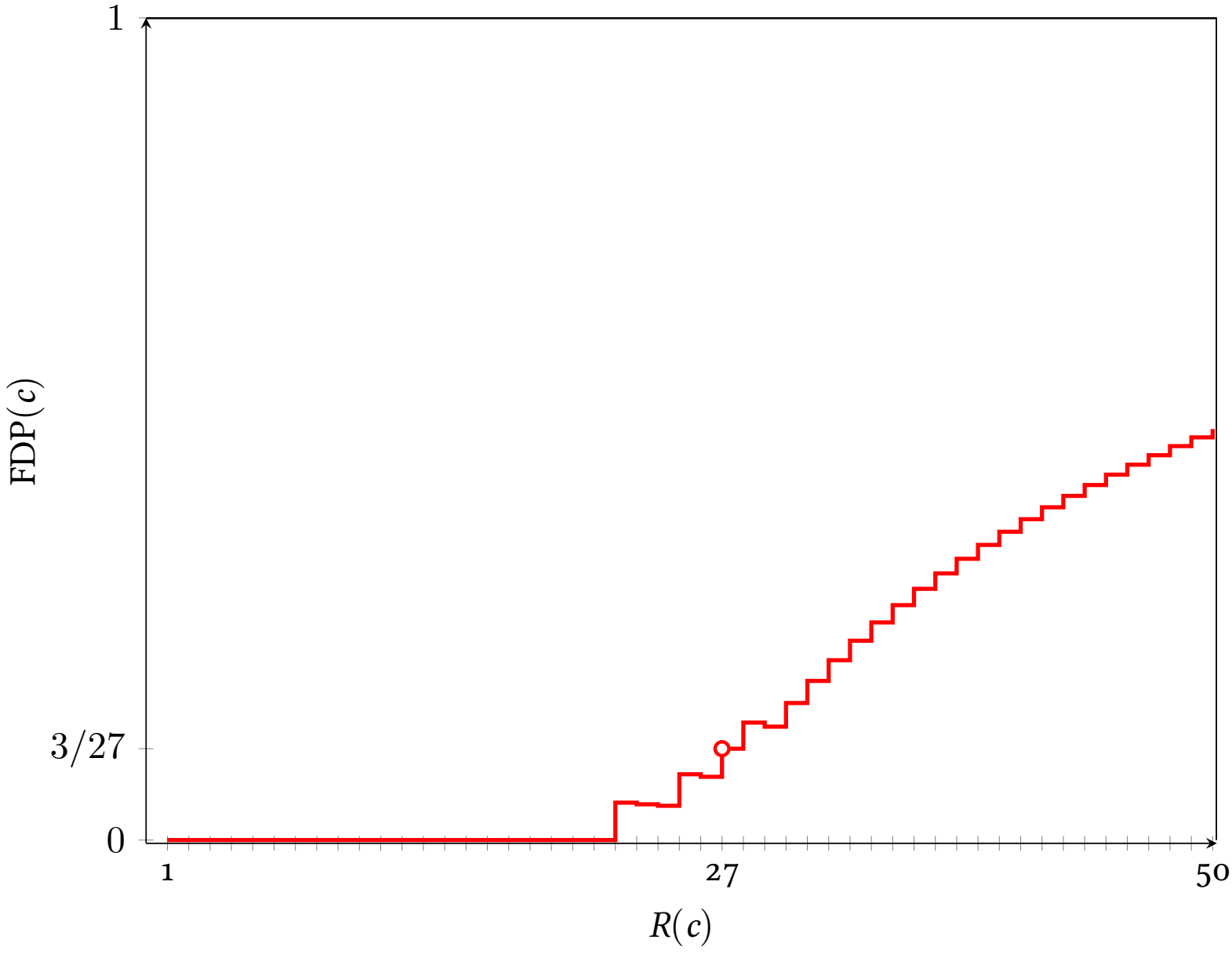
```
?p.adjust

set.seed(123)
x <- rnorm(50, mean = c(rep(0, 25), rep(3, 25)))
p <- 2*pnorm(sort(-abs(x)))

round(p, 3)
```


1
0
$p_{(i)}$
true hypothesis
false hypothesis
1
20
50


1
$c = 0.1$
$p_{(i)}$
c
0
1
27
50
$R(c) = \#\{i : p_i \leq c\}$


1
FDP(c)
3/27
0
1
27
50
R(c)

$m = 10$, $m_0 = 80$, $Y_i \sim N(\mu_i, 1)$, $\mu_i = 0$ if H_i true, $\mu_i = 2$ otherwise

	1	2	3	4	5	6	7	8	9	10
R	20	17	23	16	20	16	15	17	20	17
V	4	5	6	5	5	3	3	5	7	4
$\mathbb{1}\{V > 0\}$	1	1	1	1	1	1	1	1	1	1
V/R	0.20	0.29	0.26	0.31	0.25	0.19	0.20	0.29	0.35	0.24

Reject H_i if $p_i \leq 0.05$ gives FWER = 0.983 and FDR = 0.232

Null p-values

All methods we will consider start from a collection of p-values $p_1, \ldots, p_m$, one for each hypothesis tested.

We call these p-values *raw* as they have not been corrected for multiple testing yet

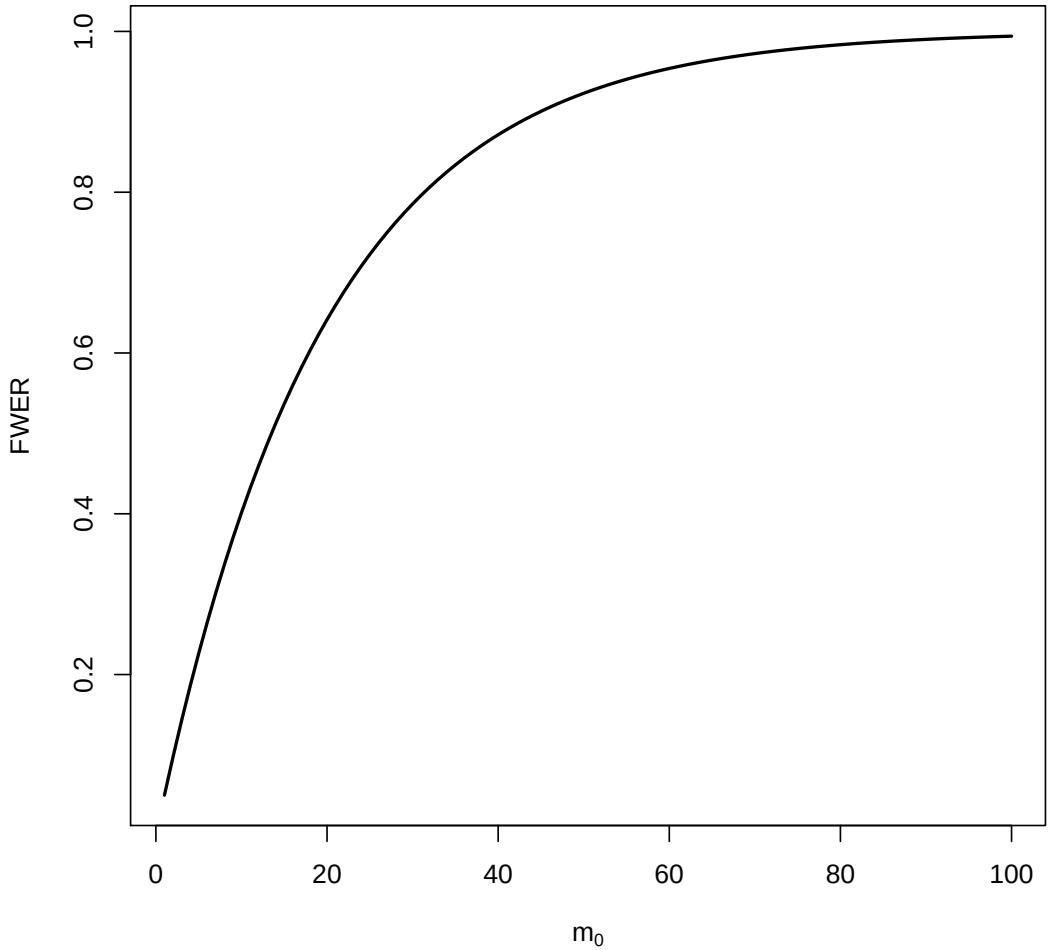
Assumptions on the p-values often involve only the p-values of true hypotheses. We denote these *null* p-values by

$$q_1, \ldots, q_{m_0}$$

Null p-values are assumed to be *valid* in the sense

$$\mathrm{P}(q_i \leq u) \leq u$$

with equality when $q_i \sim \mathrm{Uniform}(0, 1)$



Assume $q_1, \ldots, q_{m_0}$ i.i.d. Uniform$(0, 1)$.

Then $\mathcal{R} = \{H_i : p_i \leq 0.05\}$ has FWER $= 1 - (1 - 0.05)^{m_0}$

Table of Contents

Bonferroni method

$$\mathcal{R}_{\text{Bonferroni}} = \left\{ H_i : p_i \leq \frac{\alpha}{m} \right\}$$

Consider the expected number of type I errors $\mathrm{E}(V)$ (also called Per Family Error Rate, PFER). By Markov's inequality

$$\mathrm{pr}(V > 0) \leq \mathrm{E}(V)$$

Assume that null p-values are valid. Bonferroni method controls the PFER at level α:

$$\mathrm{E}(V) = \mathrm{E}\Big(\sum_{i=1}^{m_0} \mathbb{1}\{q_i \leq \frac{\alpha}{m}\} \Big) = \sum_{i=1}^{m_0} \mathrm{pr}\Big(q_i \leq \frac{\alpha}{m} \Big) \leq m_0 \frac{\alpha}{m}$$

Bonferroni conservativeness

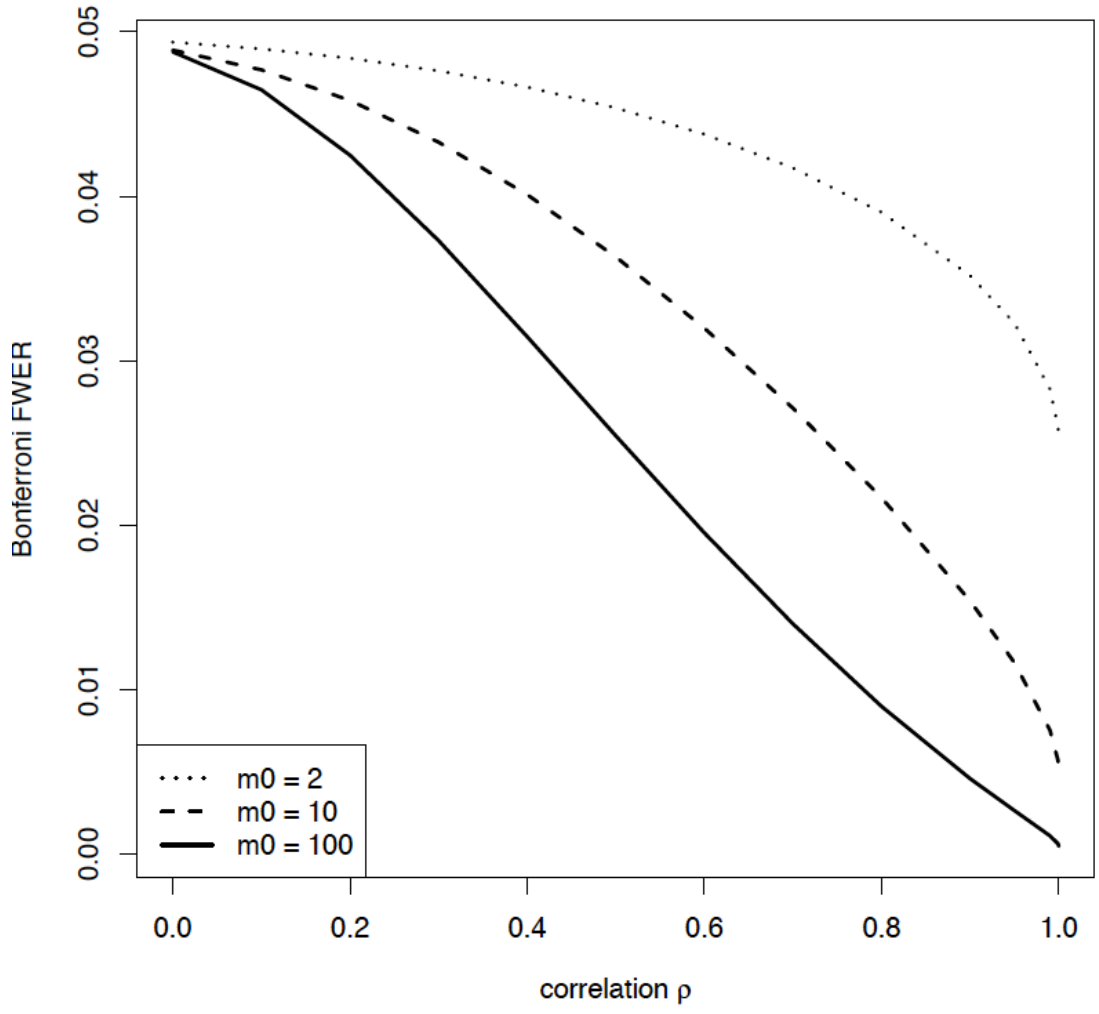
$$\text{pr}\Big(\bigcup_{i=1}^{m_0}\{q_i \leq \frac{\alpha}{m}\}\Big) \leq \sum_{i=1}^{m_0}\text{pr}\Big(q_i \leq \frac{\alpha}{m}\Big) \leq m_0\frac{\alpha}{m}$$

The two inequalities indicate in which cases the Bonferroni method can be *conservative*, i.e. FWER $< \alpha$

The right-hand one shows that Bonferroni controls the FWER at level $\pi_0\alpha$, where $\pi_0 = m_0/m$. If there are many false null hypotheses, Bonferroni will be conservative

The left-hand inequality is due to Boole's inequality, i.e. for any collection of events $E_1, \ldots, E_k$, we have $\text{P}\big(\bigcup_{i=1}^{k} E_i\big) \leq \sum_{i=1}^{k}\text{P}(E_i)$. This inequality is a strict one in all situations except the one in which all events $\{q_i \leq \alpha/m\}$ are disjoint. With independent p-values, the conservativeness is present but very minor

Much more serious conservativeness can occur if p-values are positively correlated. Suppose that the correlation matrix is such that $\{\Sigma\}_{ij} = \rho$ for $i \neq j$



Adjusted p-values

When testing a single hypothesis, we often do not only report whether a hypothesis was rejected, but also the corresponding p-value

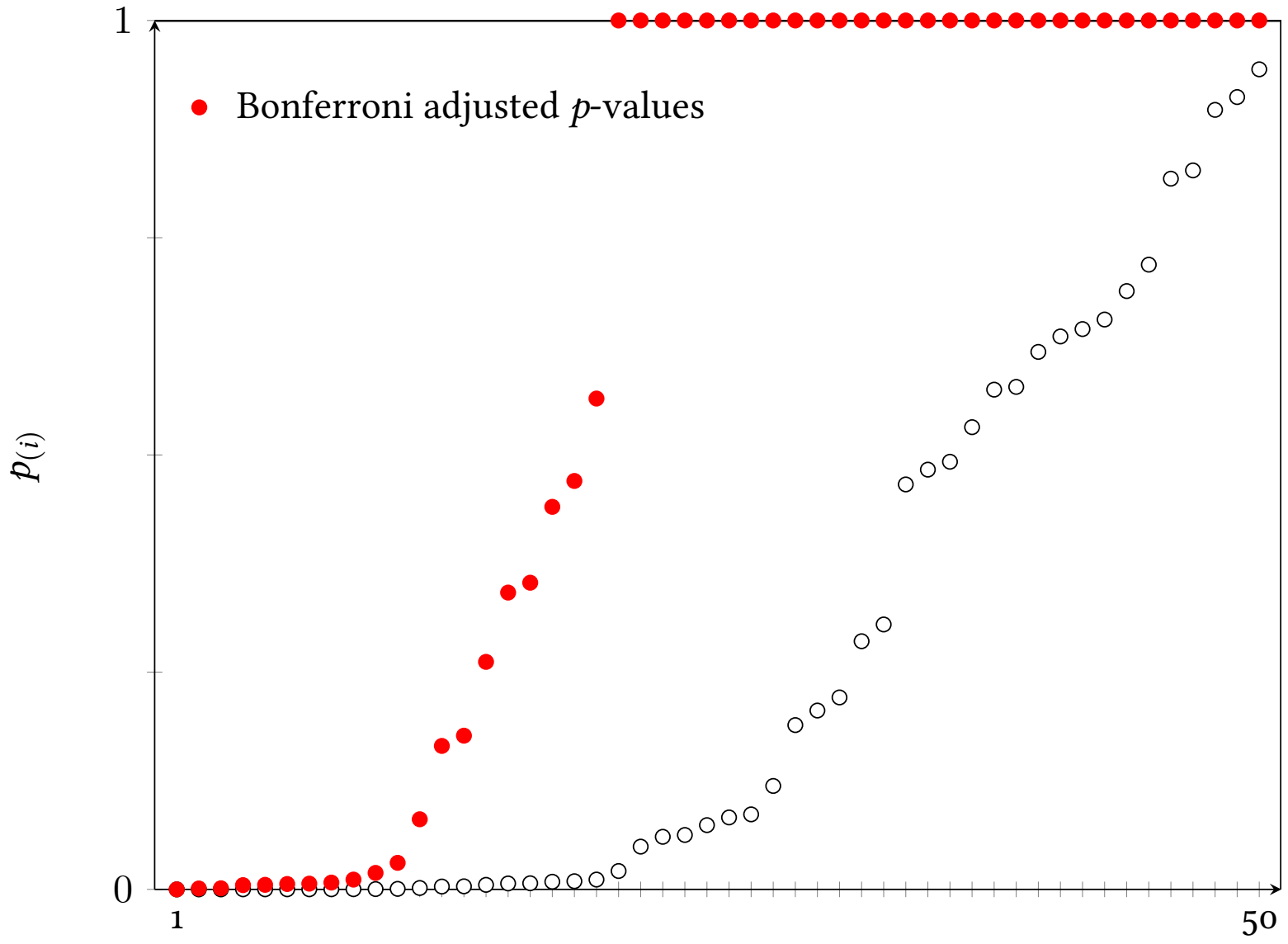
By definition, the p-value is the smallest chosen α-level of the test at which the hypothesis would have been rejected

The direct analogue of this in the context of multiple testing is the *adjusted* p-value, defined as the smallest α level at which the multiple testing method would reject the hypothesis.

For the Bonferroni method, this adjusted p-value is given by

$$\tilde{p}_i = \min(mp_i, 1)$$

where p_i is the raw p-value


1
0
$p_{(i)}$
Bonferroni adjusted p-values
1
50

Sidak method

$$\mathcal{R}_{\text{Sidak}} = \{H_i \in \mathcal{H} : p_i \leq 1 - (1-\alpha)^{1/m}\}$$

Assume that null p-values are i.i.d. Uniform$(0, 1)$. Sidak method controls the FWER at level α.

$\text{pr}\Big(\bigcup_{i=1}^{m_0}\big\{q_i \leq c\big\}\Big) = 1 - \prod_{i=1}^{m_0}\text{P}\Big(q_i > c\Big) = 1 - (1-c)^{m_0}$ which equals α for $c = 1 - (1-\alpha)^{1/m_0}$. Since we don't know m_0, we can use

$$1 - (1-\alpha)^{1/m} \leq 1 - (1-\alpha)^{1/m_0}$$

The ratio between the Bonferroni and Sidak critical values

$$\frac{\alpha/m}{1-(1-\alpha)^{1/m}} \overset{m\to\infty}{\to} \frac{-\log(1-\alpha)}{\alpha}$$

which evaluates to only 1.026 for $\alpha = 0.05$

Holm method

Holm's method always rejects at least as much as Bonferroni's method, and often a bit more, but still has valid FWER control under the same assumptions

Holm's method is a sequential variant of the Bonferroni method that always rejects at least as much as Bonferroni's method, and often a bit more, but still has valid FWER control under the same assumptions

In the first step, all hypotheses with p-values at most α/h_0 are rejected, with $h_0 = m$ just like in the Bonferroni method. Suppose this leaves h_1 hypotheses unrejected. Then, in the next step, all hypotheses with p-values at most α/h_1 are rejected, which leaves h_2 hypotheses unrejected, which are subsequently tested at level α/h_2. This process is repeated until either all hypotheses are rejected, or until a step fails to result in any additional rejections

Sort p-values $p_{(1)} \leq \ldots \leq p_{(m)}$ and corresponding hypotheses $H_{(1)} \leq \ldots \leq H_{(m)}$

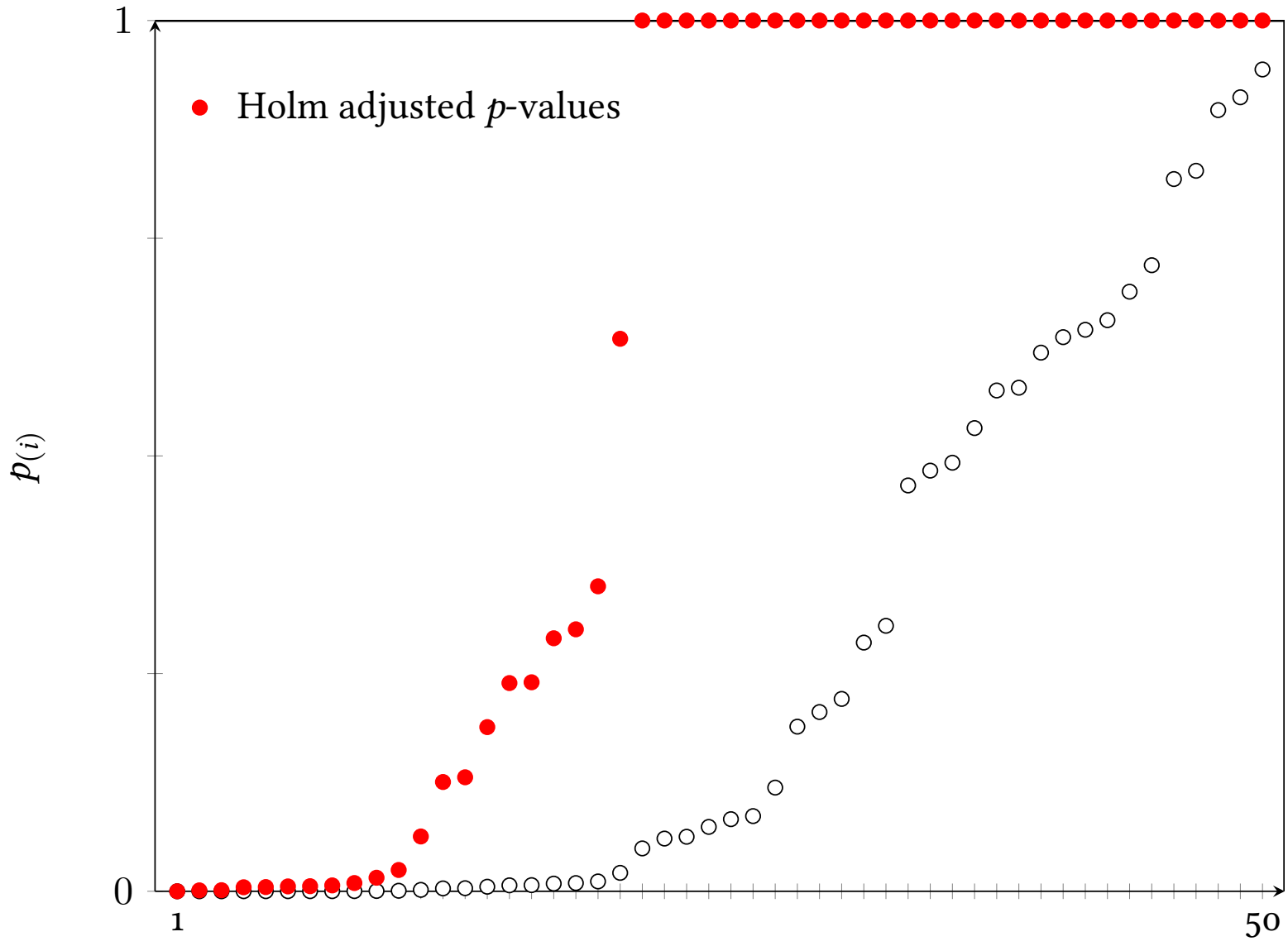
Algorithm 1 Holm

1: $i \leftarrow 0$
2: **while** $p_{(i+1)} \leq \frac{\alpha}{m-i}$ **do**
3: $\quad i \leftarrow i + 1$
4: **end while**
5: Reject $H_{(1)}, \ldots, H_{(i)}$

The Holm adjusted p-values for the hypotheses $H_{(1)}, \ldots, H_{(m)}$ are defined sequentially in the following way:

$$\tilde{p}_{(1)} = \min(1, mp_{(1)})$$

$$\tilde{p}_{(i)} = \max(\tilde{p}_{(i-1)}, (m - i + 1)p_{(i)}) \quad \text{for } i = 2, \ldots, m$$


1
0
$p_{(i)}$
Holm adjusted p-values
1
50

Closed testing

Suppose we have m hypotheses $H_1, \ldots, H_m$.

Denote by $T \subseteq \{1, \ldots, m\}$ the index set of true hypotheses.

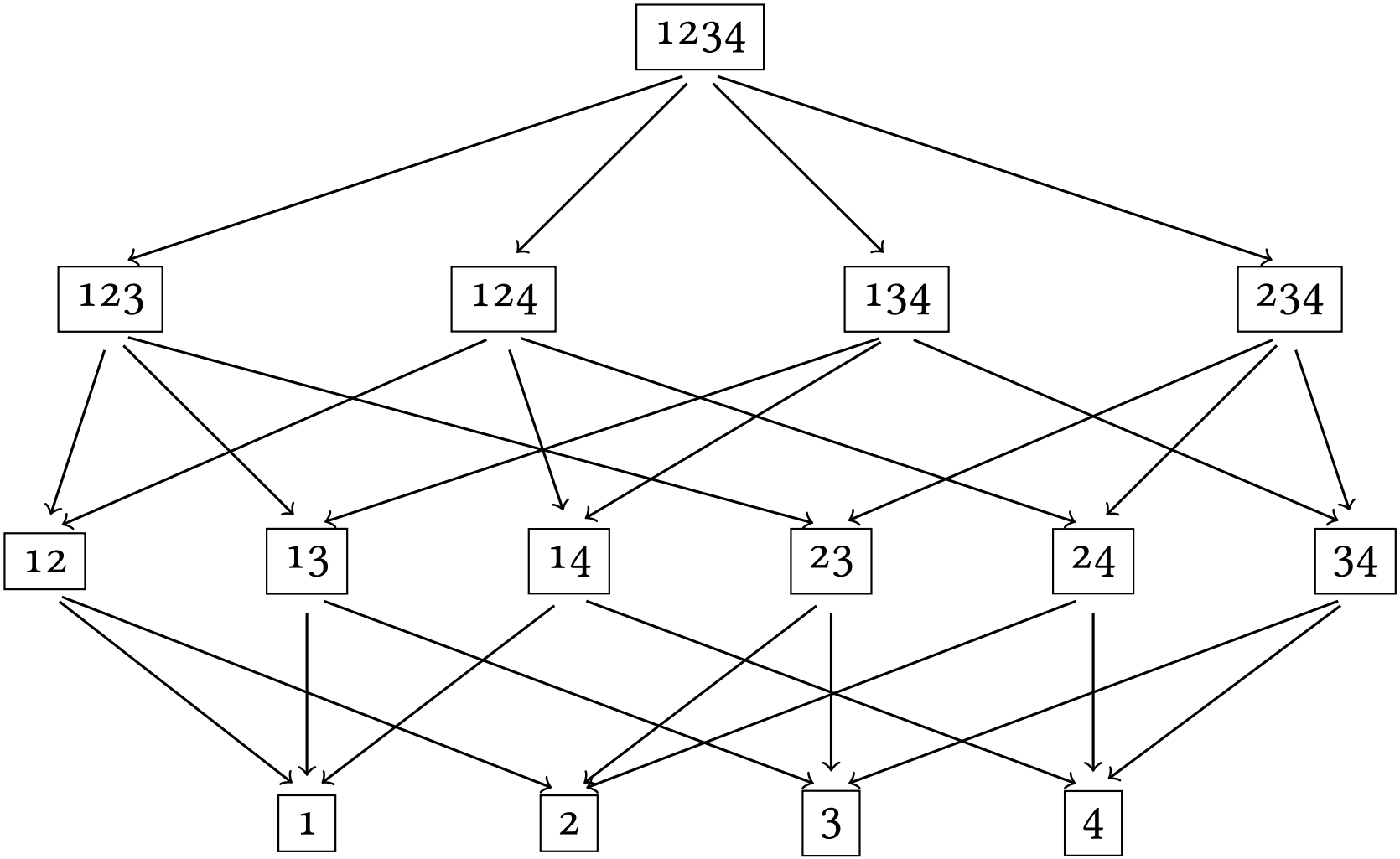
In the closed testing procedure of Marcus et al. (1976), the collection of hypotheses is augmented with all possible *intersection hypotheses*

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

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 that all original hypotheses, known as *elementary hypotheses*, are also intersection hypotheses.

Intersection hypotheses

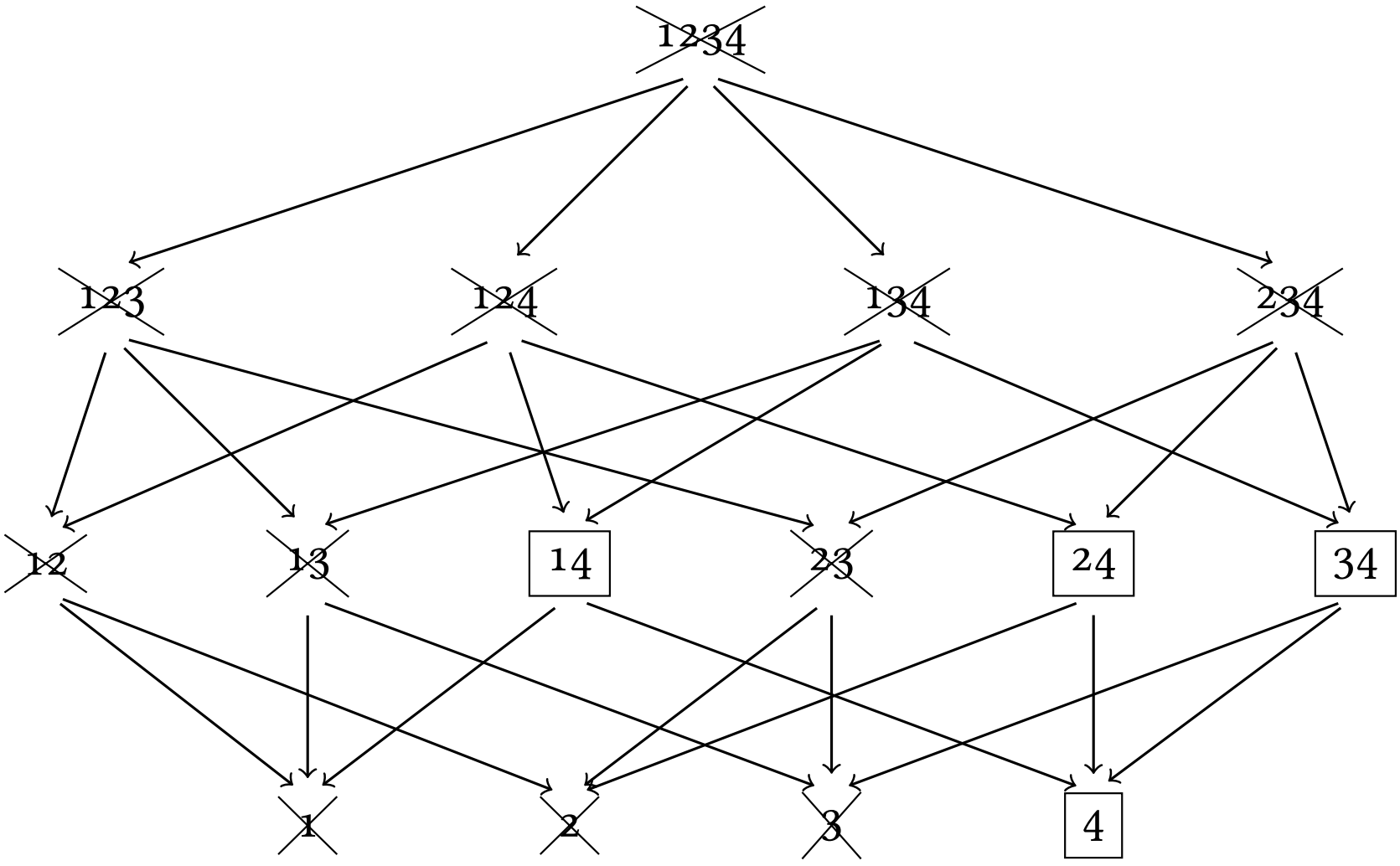


The closed testing procedure starts by testing all intersection hypotheses with a *local test*.

It will be helpful to use the notation $\mathcal{I} = 2^{\{1,\dots,m\}}$ for the collection of all subsets of $\{1, \dots, m\}$ and $\mathcal{T} = \{I \in \mathcal{I} \colon I \subseteq T\}$ for the collection of index sets corresponding to true intersection hypotheses.

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

Rejections

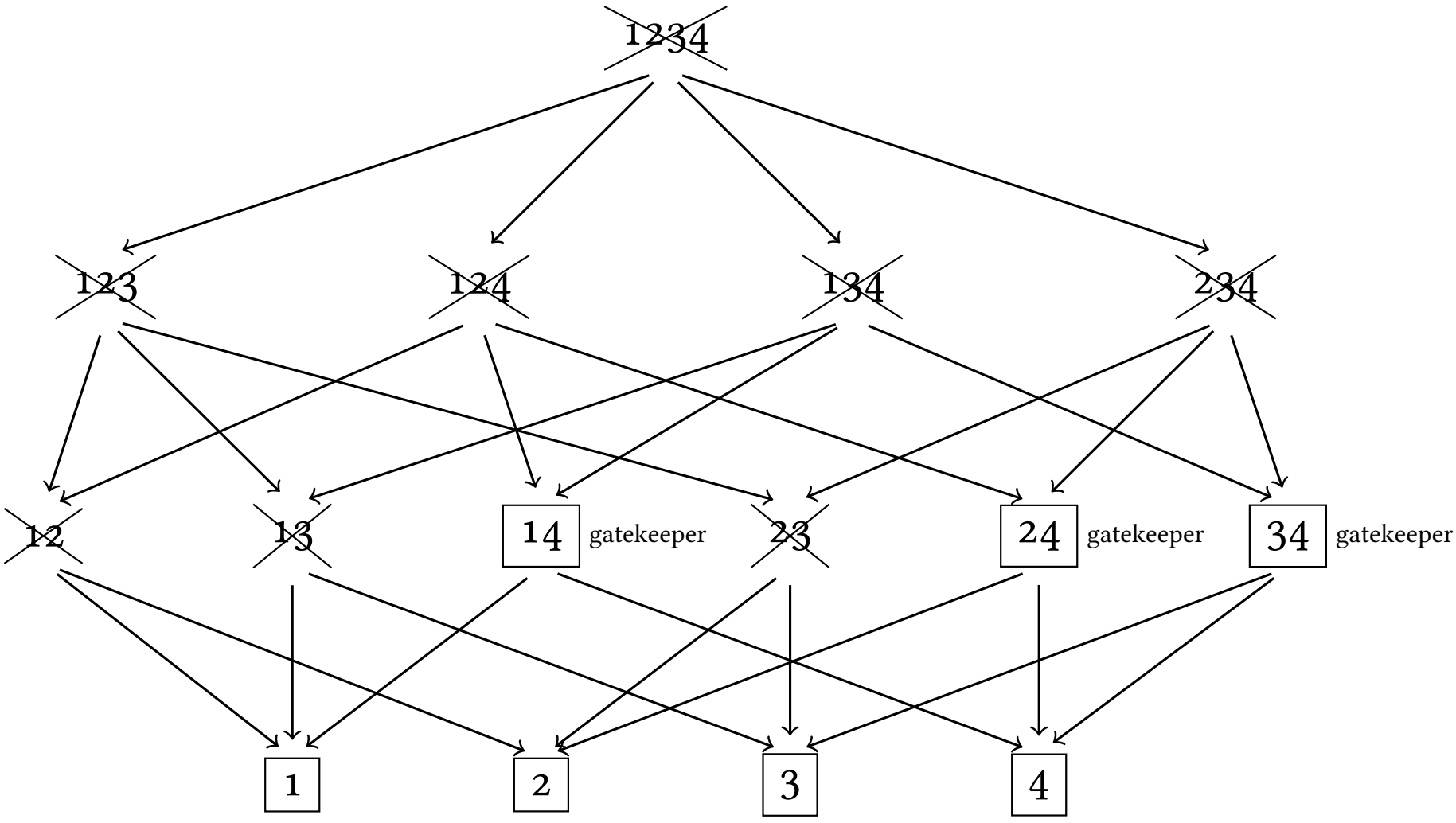


An intersection hypothesis H_I is subsequently rejected by the closed testing procedure if and only if all intersection hypotheses $H_\mathcal{J}$ with $I \subseteq \mathcal{J}$ are rejected by the local test.

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

$$\mathcal{X}_\alpha = \{I \in \mathcal{I} \colon \mathcal{J} \in \mathcal{U}_\alpha \text{ for all } \mathcal{J} \supseteq I\}.$$

Closed testing rejections



Under the assumption that

$$\text{pr}(T \notin \mathcal{U}_\alpha) \geq 1 - \alpha,$$

i.e. that the local test rejects H_T, the intersection of all true hypotheses, with probability at most α, the closed testing procedure controls the familywise error rate for all intersection hypotheses H_I, i.e.

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

This follows because $T \notin \mathcal{U}_\alpha$ implies that $I \notin \mathcal{X}_\alpha$ for every $I \subseteq T$, which in turn implies $\mathcal{T} \cap \mathcal{X}_\alpha = \emptyset$.

Consider the closed testing with Bonferroni local tests, i.e. with p-value for testing H_I equal to

$$p_I = \min\{p_i, i \in I\}|I|$$

(truncated at 1 if it exceeds 1)

Then the closed testing procedure gives Holm procedure.

Consider the closed testing with Simes local tests, i.e. with p-value for testing H_I equal to

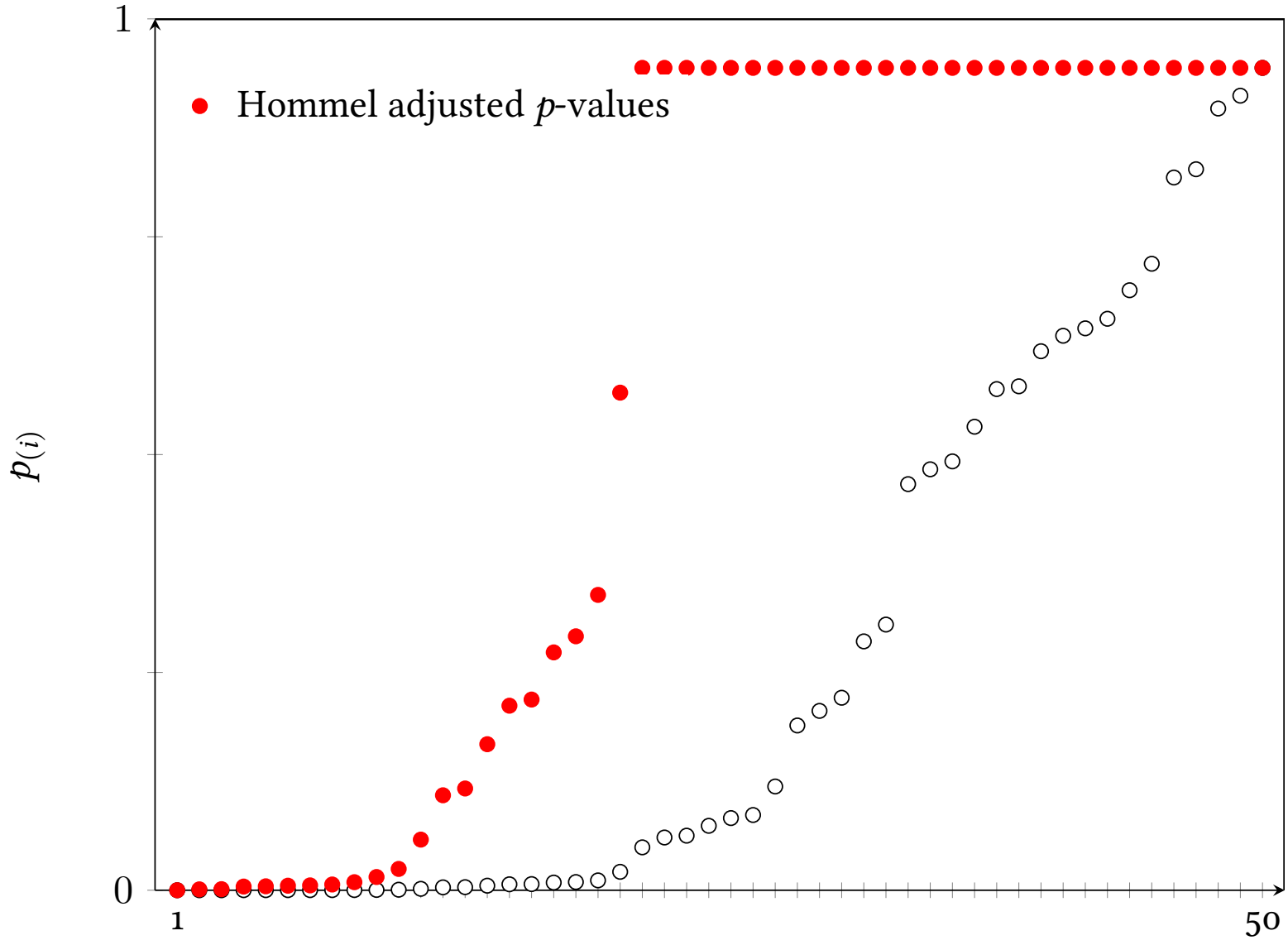
$$p_I = \min_{i=1,\ldots,|I|} \left\{ p_{(i:I)} \frac{|I|}{i} \right\}$$

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

Then the closed testing procedure gives Hommel procedure.

Assumption: Simes inequality holds for null p-values

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


1
0
$p_{(i)}$
Hommel adjusted p-values
1
50

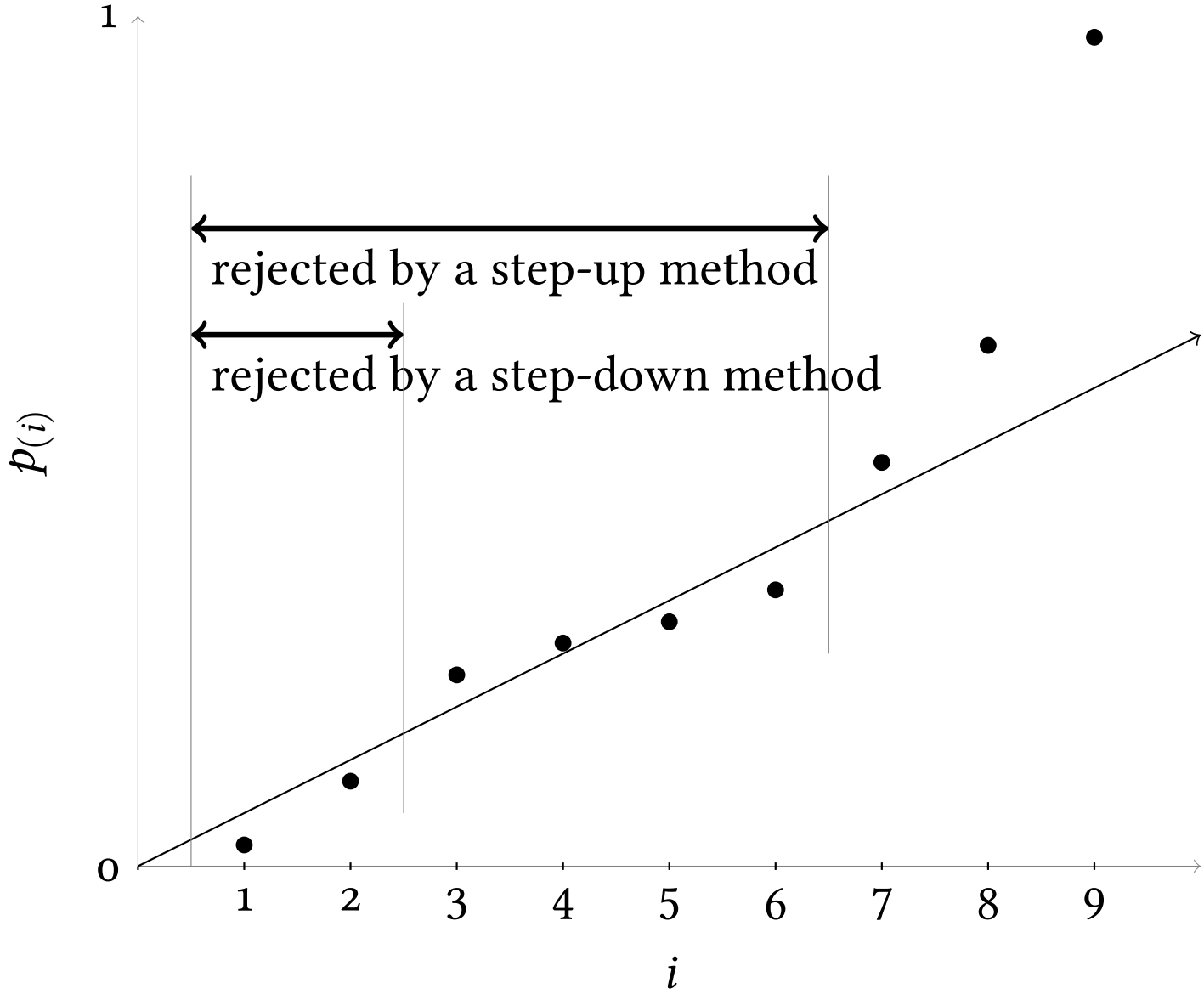
Hochberg method

While more conservative than Hommel's procedure, Hochberg's procedure is simpler to describe and still more powerful than Holm's.

Algorithm 2 Hochberg

1: $i \leftarrow m$
2: **while** $p_{(i)} > \frac{\alpha}{m-i+1}$ **do**
3: $\quad i \leftarrow i - 1$
4: **end while**
5: Reject $H_{(1)}, \ldots, H_{(i)}$

Step-down vs step-up



Sequential testing

Suppose that, prior to running our experiments, we have ordered our hypotheses $H_1, \ldots, H_m$ where H_1 is the most promising hypothesis.

The fixed sequence test first compares p_1 to α. If $p_1 > \alpha$ then the procedure terminates. If $p_1 \leq \alpha$, then we reject H_1 and compare p_2 to α, and so on.

Algorithm 3 Fixed Sequence

1: Ordered hypotheses $H_1, \ldots, H_m$
2: $i \leftarrow 1$
3: **while** $p_i \leq \alpha$ **do**
4: $\quad i \leftarrow i + 1$
5: **end while**
6: Reject $H_{(1)}, \ldots, H_{(i)}$

Let H_{i^*} be the first true null hypothesis. That is i^* is the smallest index i such that H_i is true.

$$\begin{aligned} \text{pr}(V \geq 1) &= \text{pr}(\text{reject } H_{i^*}) \\ &= \text{pr}(p_1 \leq \alpha, \ldots, p_{i^*} \leq \alpha) \\ &\leq \text{pr}(p_{i^*} \leq \alpha) \\ &\leq \alpha \end{aligned}$$

In the above proof it was important that the order of the hypotheses was chosen in advance and independently of the p-values.

Indeed, if the order depended on the p-values, then i^* would be a random variable and we may no longer have $\text{pr}(p_{i^*} \leq \alpha) \leq \alpha$

Fallback procedure

Suppose that, prior to running our experiments, we have ordered our hypotheses $H_1, \ldots, H_m$.

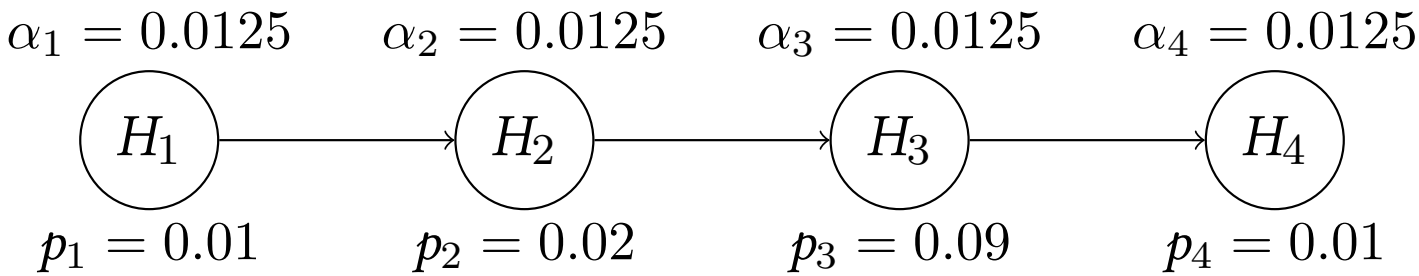
A set of threshold $\alpha_1, \ldots, \alpha_m \geq 0$ such that $\sum_{i=1}^{m} \alpha_i = 1$ must also be fixed in advance.

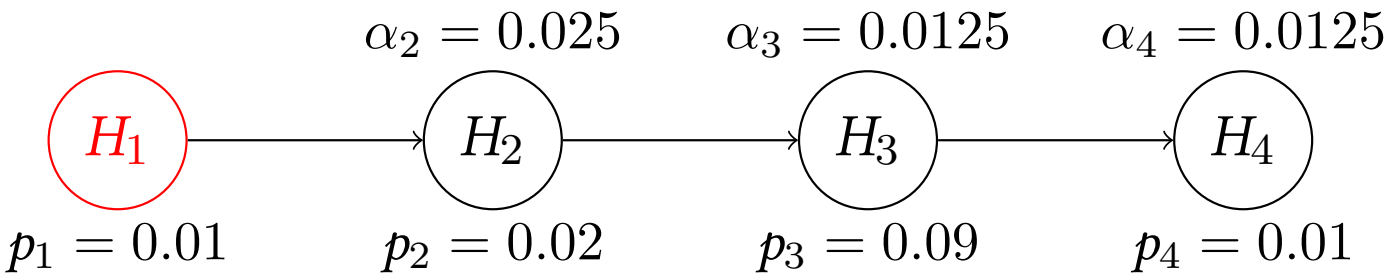
The fallback procedure does the following:

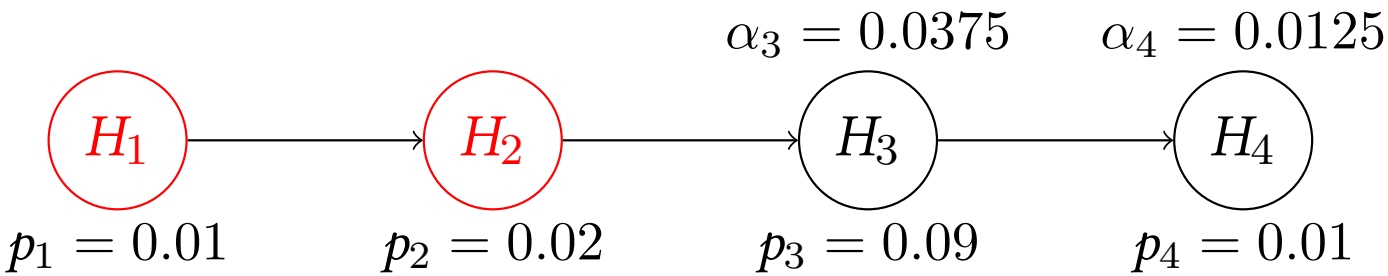
- Compare p_1 to α_1. If $p_1 \leq \alpha_1$, then we reject H_1 and change α_2 to $\alpha_2 + \alpha_1$. If $p_1 > \alpha_1$, then we fail to reject H_1 and we leave α_2 unchanged.
- We the reject H_2 iff $p_2 \leq \alpha_2$. If we reject H_2, then we change α_3 to $\alpha_3 + \alpha_2$
- etc.

Algorithm 4 Fallback procedure

Ordered hypotheses $H_1, \ldots, H_m$, threshold $\alpha_1, \ldots, \alpha_m \geq 0$, $\sum_{i=1}^{m} \alpha_i = 1$
$\mathcal{R} \leftarrow \emptyset$
for i=1,...,m **do**
    if $p_i \leq \alpha_i$ **then**
        $\mathcal{R} \leftarrow \mathcal{R} \cup \{i\}$
        $\alpha_{i+1} \leftarrow \alpha_{i+1} + \alpha_i$
    end if
end for
Reject $\mathcal{R}$


$\alpha_1 = 0.0125$
$\alpha_2 = 0.0125$
$\alpha_3 = 0.0125$
$\alpha_4 = 0.0125$
H_1
H_2
H_3
H_4
$p_1 = 0.01$
$p_2 = 0.02$
$p_3 = 0.09$
$p_4 = 0.01$


$\alpha_2 = 0.025$
$\alpha_3 = 0.0125$
$\alpha_4 = 0.0125$
H_1
H_2
H_3
H_4
$p_1 = 0.01$
$p_2 = 0.02$
$p_3 = 0.09$
$p_4 = 0.01$


$\alpha_3 = 0.0375$
$\alpha_4 = 0.0125$
H_1
H_2
H_3
H_4
$p_1 = 0.01$
$p_2 = 0.02$
$p_3 = 0.09$
$p_4 = 0.01$

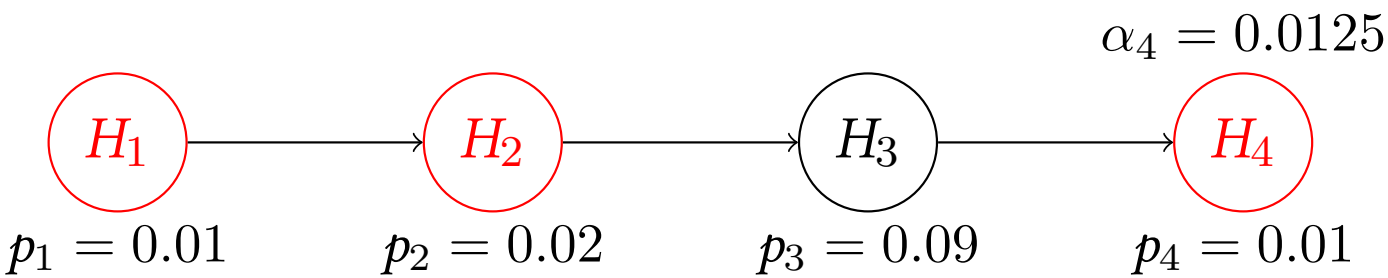

$\alpha_4 = 0.0125$
H_1
H_2
H_3
H_4
$p_1 = 0.01$
$p_2 = 0.02$
$p_3 = 0.09$
$p_4 = 0.01$

Table of Contents

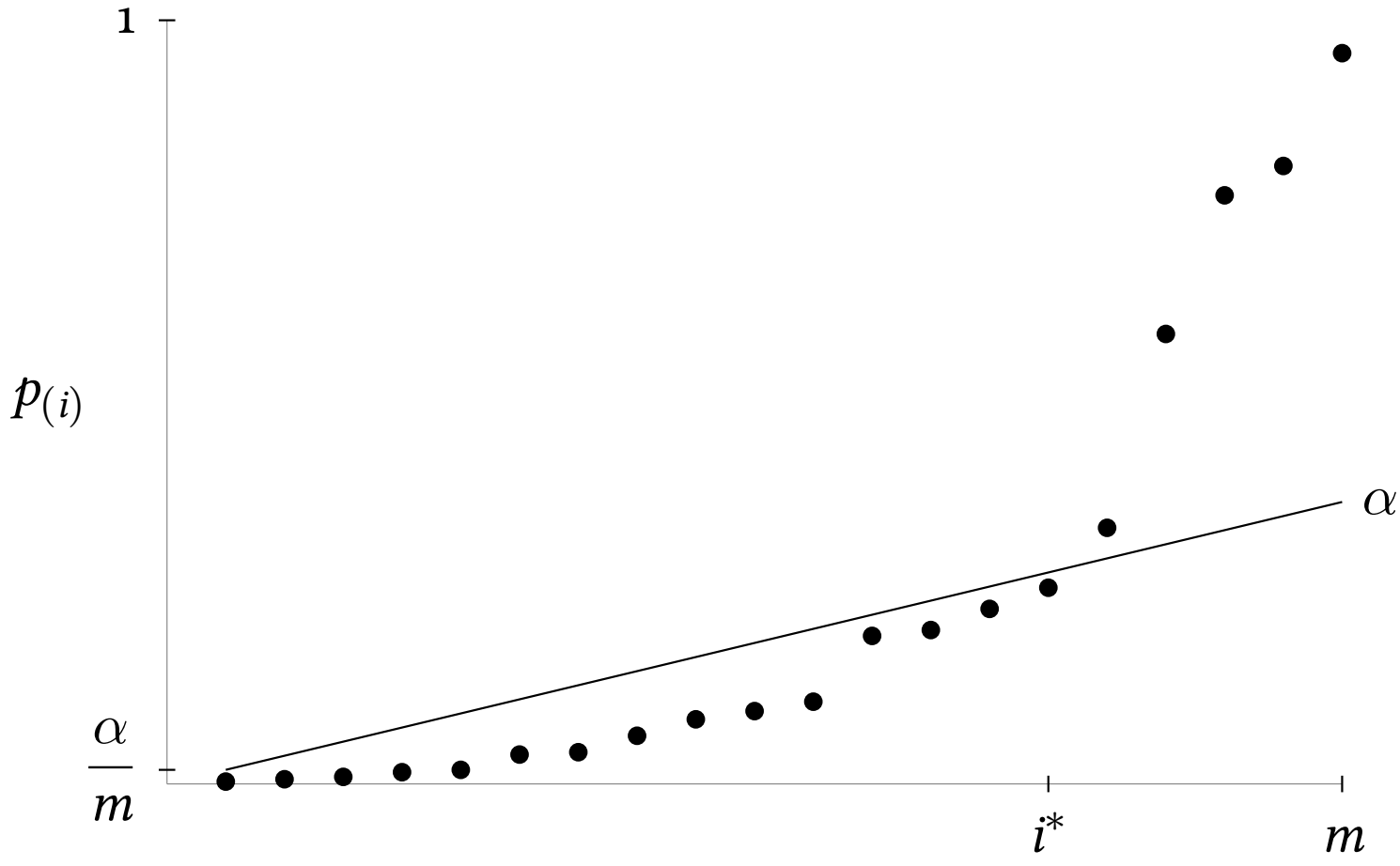
If we are testing millions of hypotheses at once, and making few false discoveries is not the end of the world

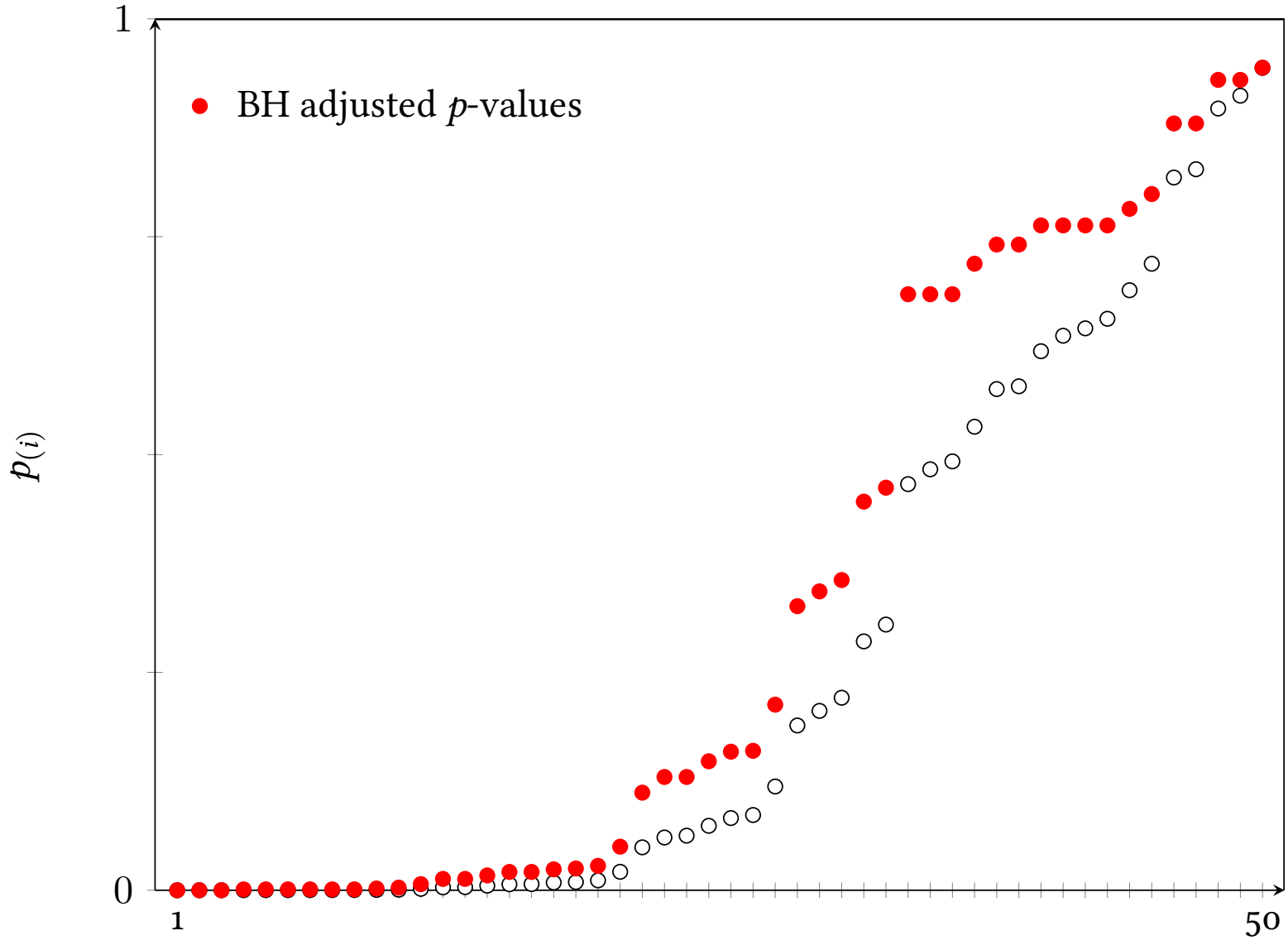
The concept of False Discovery Rate (FDR) has changed thinking about multiple testing quite radically, showing that FWER control is not only way to do of multiple testing, and stimulating the field of multiple testing enormously

FDR was introduced by Benjamini and Hochberg in 1995, and currently has 85K citations. It is one of the most-cited research of all time

Benjamini & Hochberg method

1. Sort the p-values
$$p_{(1)} \leq \cdots \leq p_{(m)}$$
2. If $p_{(i)} > \dfrac{i\alpha}{m}$ for all i, reject nothing, i.e. $\mathcal{R} = \emptyset$
3. Otherwise, let
$$i^* = \max\left\{ i \in \{1, \ldots, m\} : p_{(i)} \leq \frac{i\alpha}{m} \right\}$$
be the largest i for which $p_{(i)} \leq \dfrac{i\alpha}{m}$
4. Reject all $H_{(i)}$ with $i \leq i^*$, i.e. $\mathcal{R} = \left\{ H_i : p_i \leq \dfrac{i^*\alpha}{m} \right\}$


1
$p_{(i)}$
α
$\frac{\alpha}{m}$
i^*
m


1
0
$p_{(i)}$
BH adjusted p-values
1
50

For independent p-values $p_1, \ldots, p_m$ and null p-values $q_1, \ldots, q_{m_0}$ i.i.d. Uniform$(0, 1)$, the FDR of the Benjamini-Hochberg method is exactly $\pi_0 \alpha$.

The conclusion is obvious when $m_0 = 0$: assume $m_0 \geq 1$

Define $V_i = \mathbb{1}\{H_i \text{ rejected}\}$ for each $i \in T$ where $T = \{i : H_i \in \mathcal{T}\}$. We can express the FDP as

$$Q = \sum_{i \in T} \frac{V_i}{R \vee 1}$$

We claim that

$$\mathbb{E}\left(\frac{V_i}{R \vee 1}\right) = \frac{\alpha}{m}, \quad i \in T$$

based on which we have

$$\text{FDR} = \mathbb{E}(Q) = \sum_{i \in T} \mathbb{E}\left(\frac{V_i}{R \vee 1}\right) = \sum_{i \in T} \frac{\alpha}{m} = \pi_0 \alpha$$

What remains for the proof is to show that the claim is true

When there are $R = k$ rejections, then H_i is rejected if and only if $p_i \leq (\alpha k)/m$, and therefore, we have

$$V_i = \mathbb{1}\{p_i \leq (\alpha k)/m\}$$

Suppose $p_i \leq (\alpha k)/m$ (i.e. H_i is rejected). Let us take p_i and set its value to o, and denote the new number of rejections by $R(p_i \downarrow 0)$. This new number of rejections is exactly R, because we have only reordering the first k p-values, all of which remain below the threshold $(\alpha k)/m$. On the other hand, if $p_i > (\alpha k)/m$, then we do not reject H_i, and so $V_i = 0$. Therefore we have

$$V_i \mathbb{1}\{R = k\} = V_i \mathbb{1}\{R(p_i \downarrow 0) = k\}$$

Combining the observations above and taking the expectation conditional on all p-values except for p_i, i.e. $\mathcal{F}_i = \{p_1, \ldots, p_{i-1}, p_{i+1}, \ldots, p_m\}$, we have

$$\begin{aligned}
\mathbb{E}\Big(\frac{V_i}{R \vee 1}|\mathcal{F}_i\Big) &= \sum_{k=1}^{m} \frac{\mathbb{E}(\mathbb{1}\{p_i \leq (\alpha k)/m\}\mathbb{1}\{R(p_i \downarrow 0) = k\}|\mathcal{F}_i)}{k} \\
&= \sum_{k=1}^{m} \frac{\mathbb{1}\{R(p_i \downarrow 0) = k\}(\alpha k)/m}{k}
\end{aligned}$$

where the second equality holds because knowing $\mathcal{F}_i$ and $p_i = 0$ makes $\mathbb{1}\{R(p_i \downarrow 0)\}$ deterministic, and the fact that $p_i \sim U(0,1)$ and the p-values $p_1, \ldots, p_m$ are independent

Next, we have

$$\mathbb{E}\Big(\frac{V_i}{R \vee 1}|\mathcal{F}_i\Big) = \frac{\alpha}{m}\sum_{k=1}^{m} \mathbb{1}\{R(p_i \downarrow 0) = k\} = \frac{\alpha}{m}$$

after noticing that $\sum_{k=1}^{m} \mathbb{1}\{R(p_i \downarrow 0) = k\} = 1$

Since we have set p_i to 0, we must make at least one rejection - we will always reject H_i. Therefore $R(p_i \downarrow 0) \geq 1$, and $R(p_i \downarrow 0)$ must take a value between 1 and m

The tower property verifies that

$$\text{FDR} = \sum_{i \in T} \mathbb{E}\Big(\frac{V_i}{R \vee 1}\Big) = \sum_{i \in T} \mathbb{E}\Big[\mathbb{E}\Big(\frac{V_i}{R \vee 1}|\mathcal{F}_i\Big)\Big] = \sum_{i \in T} \frac{\alpha}{m} = \pi_0 \alpha$$

□

BH is valid under the more general assumption of *positive regression dependence on the subset of nulls* (PRDS).

One case under which the PRDS condition holds is one-sided test statistics that are jointly normally distributed, if all correlations between test statistics are positive.

For p-values satisfying the PRDS assumption, the Benjamini-Hochberg procedure controls the FDR at level $\pi_0 \alpha$.

Adaptive Benjamini-Hochberg

The Benjamini & Hochberg method, like Bonferroni, controls its error rate at level $\pi_0\alpha$, rather than at α. This suggests the possibility of an alternative, more powerful Benjamini & Hochberg procedure that uses critical values

$$\frac{i\alpha}{\hat{\pi}_0 m}$$

rather than $(i\alpha)/m$ if a good estimate $\hat{\pi}_0$ of the proportion of true hypotheses π_0 would be available

Such procedures are called *adaptive* procedures, and many have been proposed on the basis of various estimates of π_0

A problem with the adaptive approach, however, is that estimates of π_0 can have high variance, especially if p-values are strongly correlated. Naive plug-in procedures, in which this variance is not taken into account, will therefore generally not have FDR control

π_0 estimator

$$\hat{m}_0(\lambda) = \frac{\sum_{i=1}^{m} \mathbb{1}\{p_i > \lambda\}}{1 - \lambda}$$

If null p-values have marginal Uniform$(0, 1)$ distribution, a proportion $1 - \lambda$ is expected to be above λ:

$$\mathrm{E}(\sum_{i=1}^{m} \mathbb{1}\{p_i > \lambda\}) \geq \mathrm{E}(\sum_{i \in T} \mathbb{1}\{q_i > \lambda\}) = m_0(1 - \lambda)$$

thus $\mathrm{E}(\hat{m}_0) \geq m_0$.

$\hat{\pi}_0 = \frac{\hat{m}_0}{m}$ is a conservative estimator of π_0, i.e. $\mathrm{E}(\hat{\pi}_0) \geq \pi_0$

Storey method

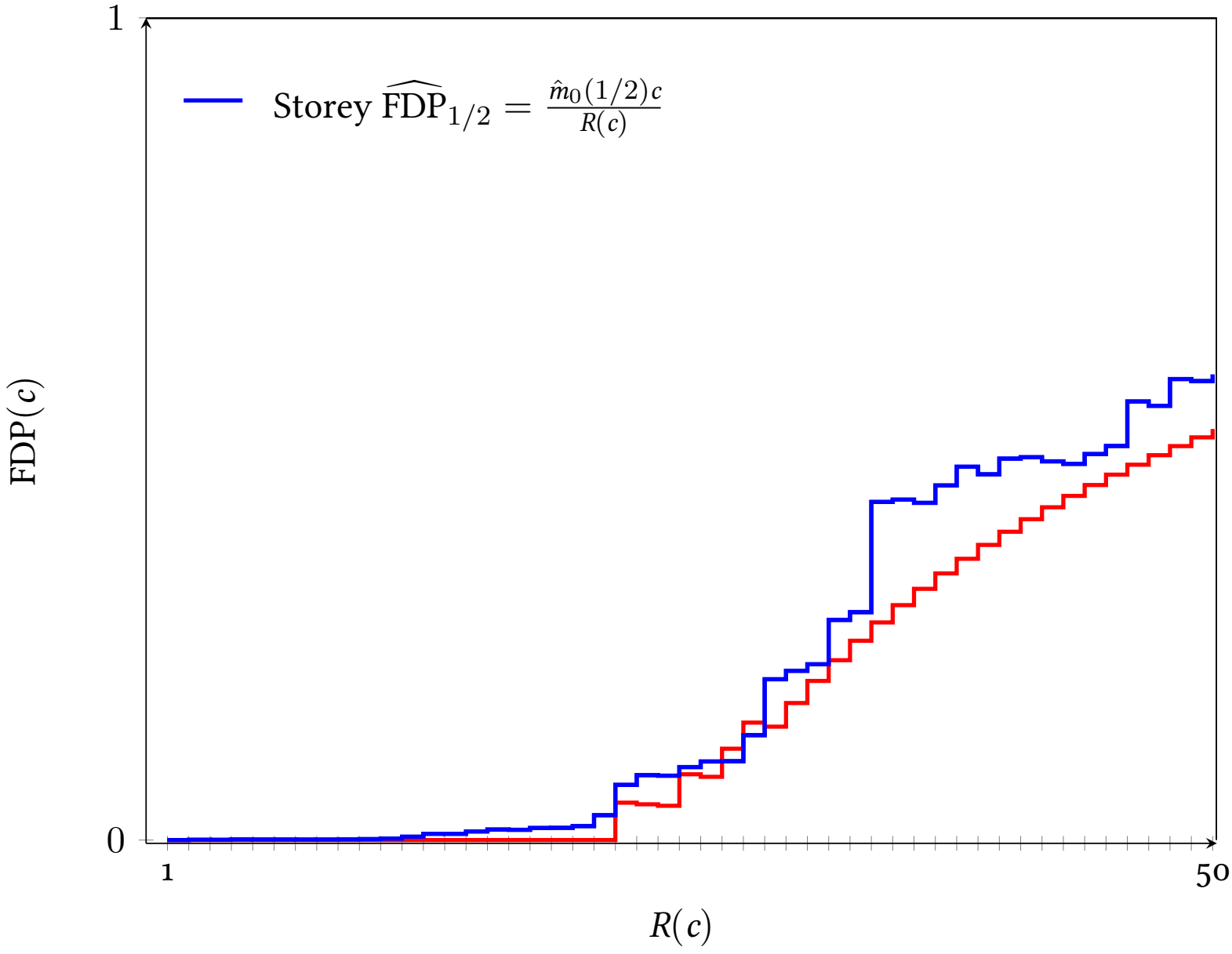
1. Choose $\lambda \in (0, 1)$. Estimate π_0 by

$$\hat{\pi}_0(\lambda) = \frac{\sum_{i=1}^{m} \mathbb{1}\{p_i > \lambda\} + 1}{(1 - \lambda)m}$$

2. Perform Benjamini-Hochberg procedure at level

$$\frac{\alpha}{\hat{\pi}_0(\lambda)}$$

- The addition of 1 to the numerator makes sure that $1/\hat{\pi}_0$ is always well-defined (but it may happen $\hat{\pi}_0 > 1$)
- The value of λ is typically $1/2$, although $\lambda = \alpha$ has also been advocated
- Storey method controls FDR under independence of p-values but generally not under positive dependence

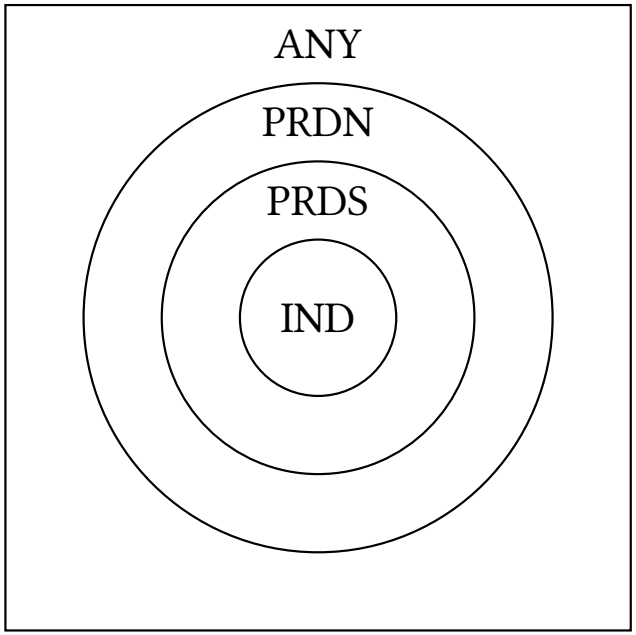

1
Storey $\widehat{\mathrm{FDP}}_{1/2} = \frac{\hat{m}_0(1/2)c}{R(c)}$
FDP(c)
0
1
50
$R(c)$

Arbitrary dependence

If the PRDS assumption is not valid, an alternative is the procedure of Benjamini & Yekutieli (2001), which is valid under general dependence

Benjamini & Yekutieli showed $\exists$ a worst-case joint distribution of p-values such that FDR[BH(α)] $= \alpha H_m$ with $H_m = \sum_{j=1}^{m} \frac{1}{j}$ harmonic number

Provable FDR control of BH



- IND : nulls i.i.d. $U(0,1)$ and indep. non-nulls (Benjamini, Hochberg, 1995) FDR $= \pi_0 \alpha$
- PRDS : pos. regr. dep. on the subset of nulls (Benjamini, Yekutieli, 2001) FDR $\leq \pi_0 \alpha$
- PRDN : pos. regr. dep. within nulls (Su, 2018) FDR $\leq \pi_0 \alpha + \pi_0 \alpha \log \frac{1}{\pi_0 \alpha}$
- ANY : arbitrary dependence (Benjamini, Yekutieli, 2001) FDR $\leq \pi_0 \alpha \sum_{i=1}^{m} (1/i)$