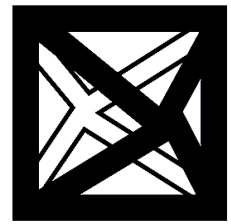
Multiple testing

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



XXXVIII cycle

Outline

Error rates

Methods for familywise error rate control

Methods for false discovery rate control

Table of Contents

In a single test, the probability of making a type I error is bounded by α, conventionally set at 5%.

Problems arise, however, when researchers do not perform a single hypothesis test but many of them. We will focus on two types of multiple testing methods:

1. those that control the *FamilyWise Error Rate*
2. those that control the *False Discovery Rate*

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

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

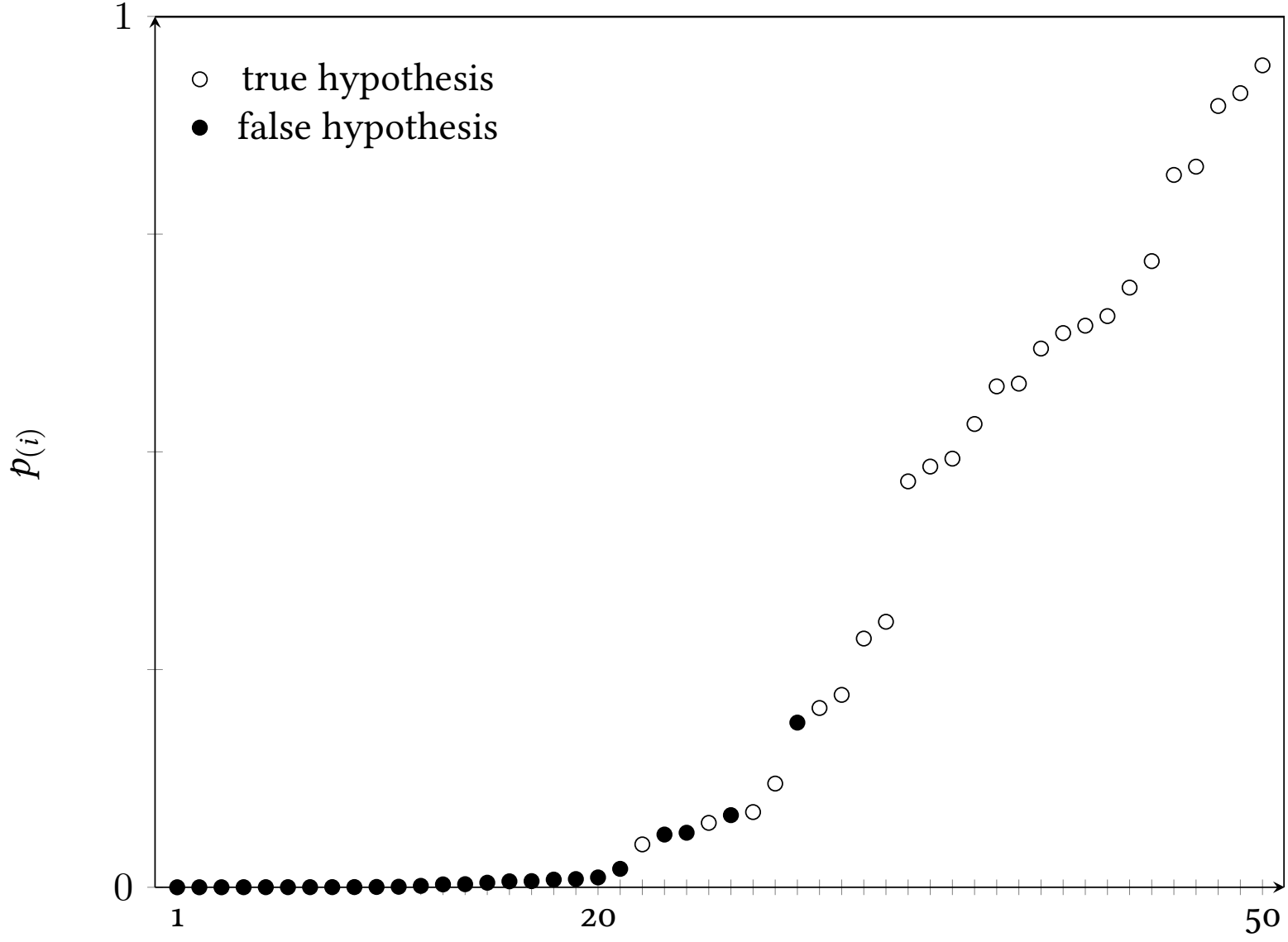
Both FDR and FWER are proper generalizations of the concept of Type I error to multiple hypotheses: if $m = 1$, the two error rates are identical and equal to the Type I error rate

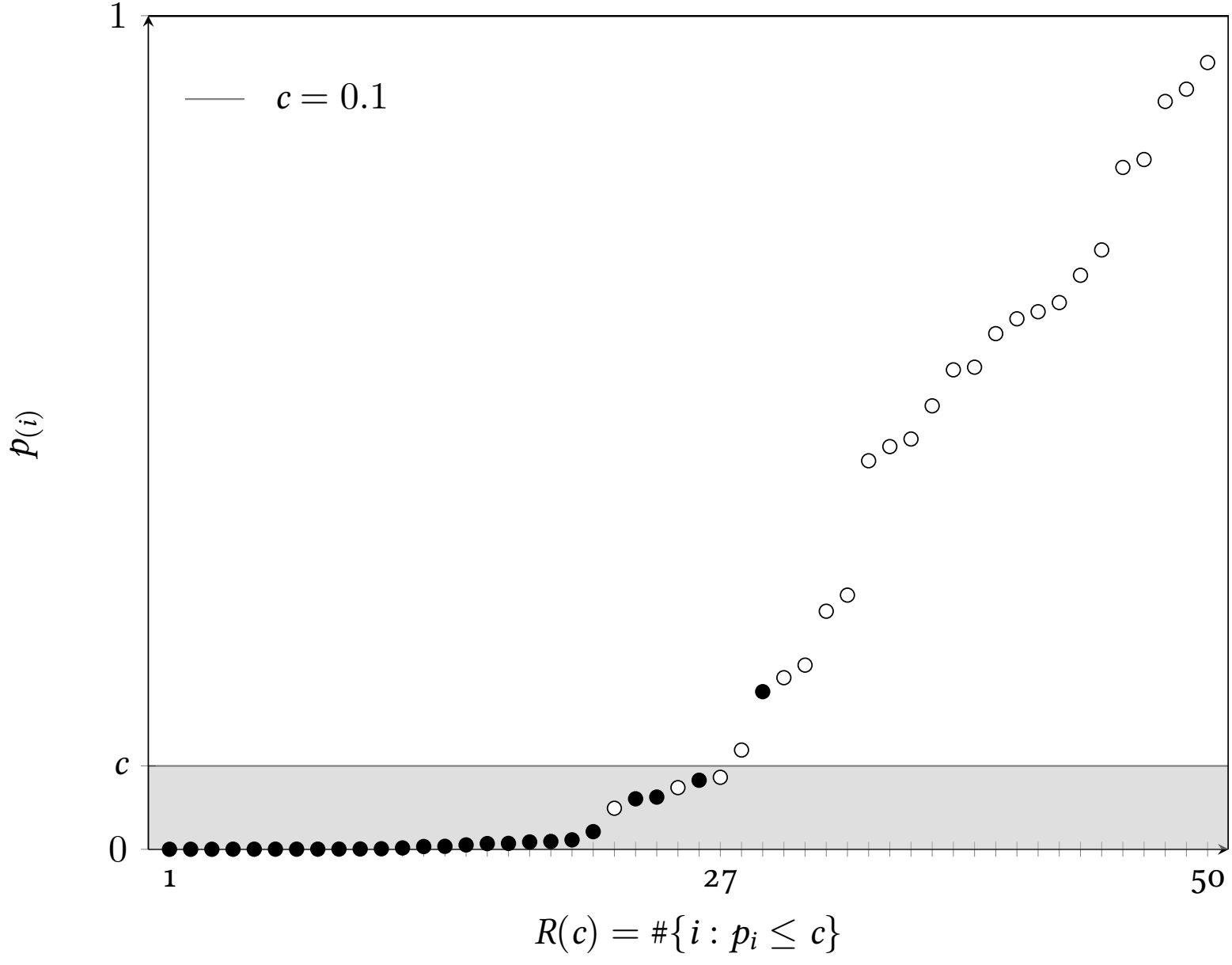
p.adjust example

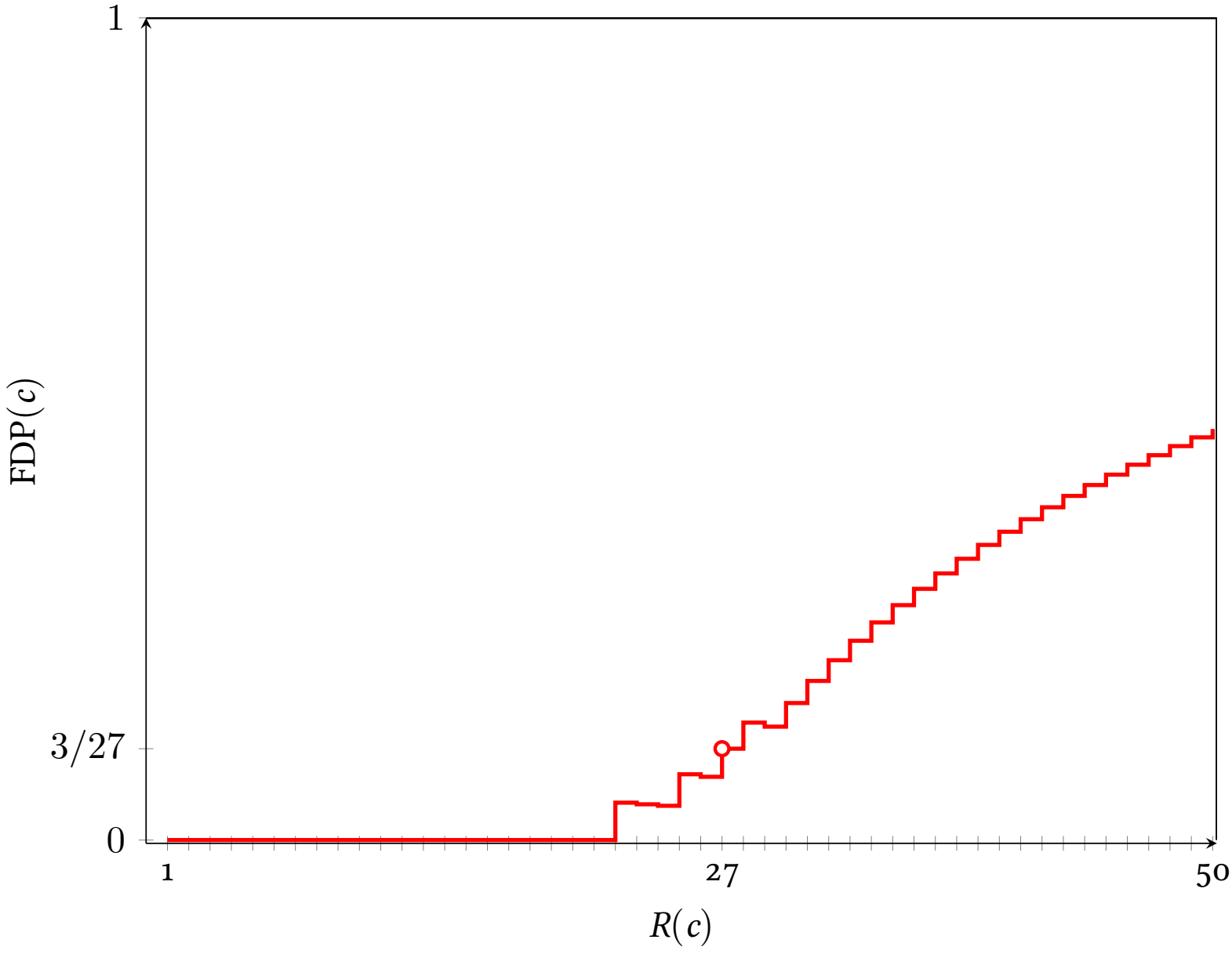
```
?p.adjust

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)

$m = 20$, $m_0 = 15$, $Y_i \sim N(\mu_i, 1)$, $\mu_i = 0$ if H_i true, $\mu_i = 2$ otherwise

	1	2	3	4	5	6	7	8	9	10
R	4	4	5	5	3	2	4	4	5	2
V	1	0	1	1	0	0	1	1	2	0
V>0	1	0	1	1	0	0	1	1	1	0
V/R	0.25	0.00	0.20	0.20	0.00	0.00	0.25	0.25	0.40	0.00

Reject H_i if $p_i \leq 0.05$ gives FWER = 0.536 and FDR = 0.168

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

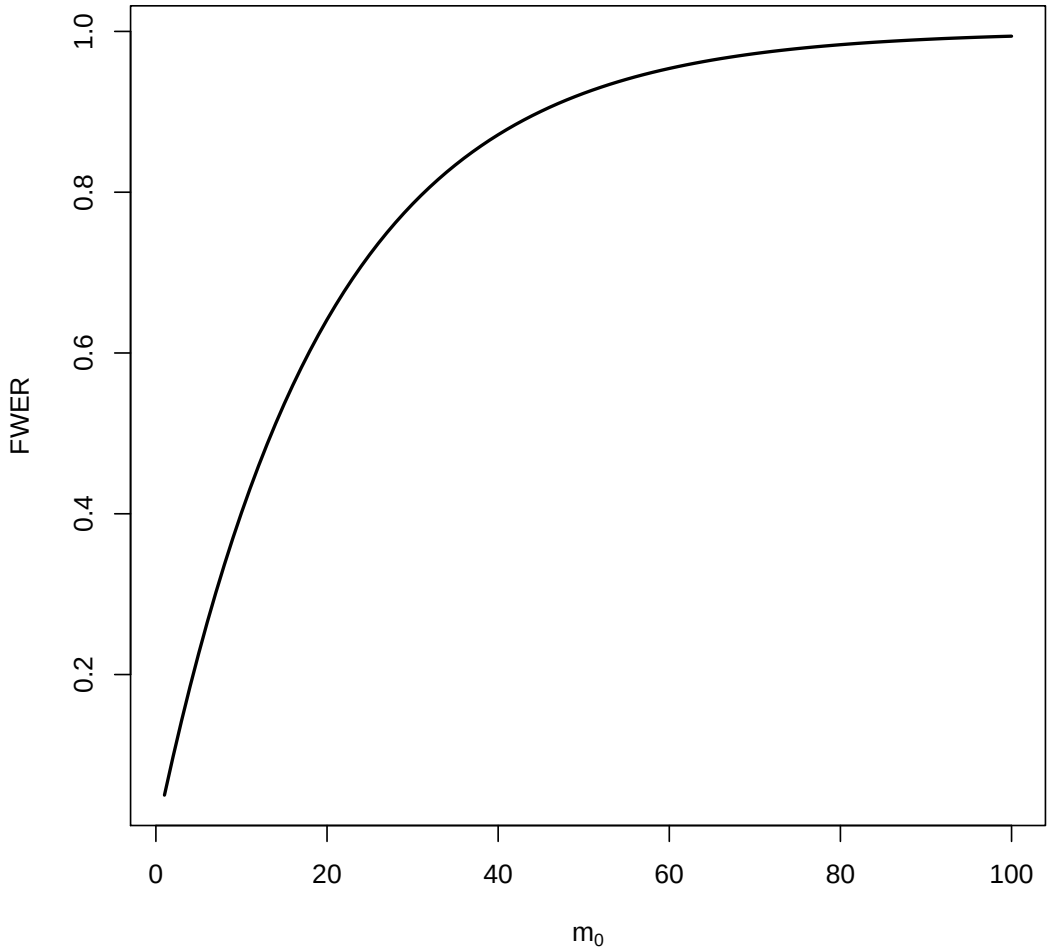
Assumptions on the p-values often involve only the p-values of true hypotheses. We denote these *null* p-values by

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

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

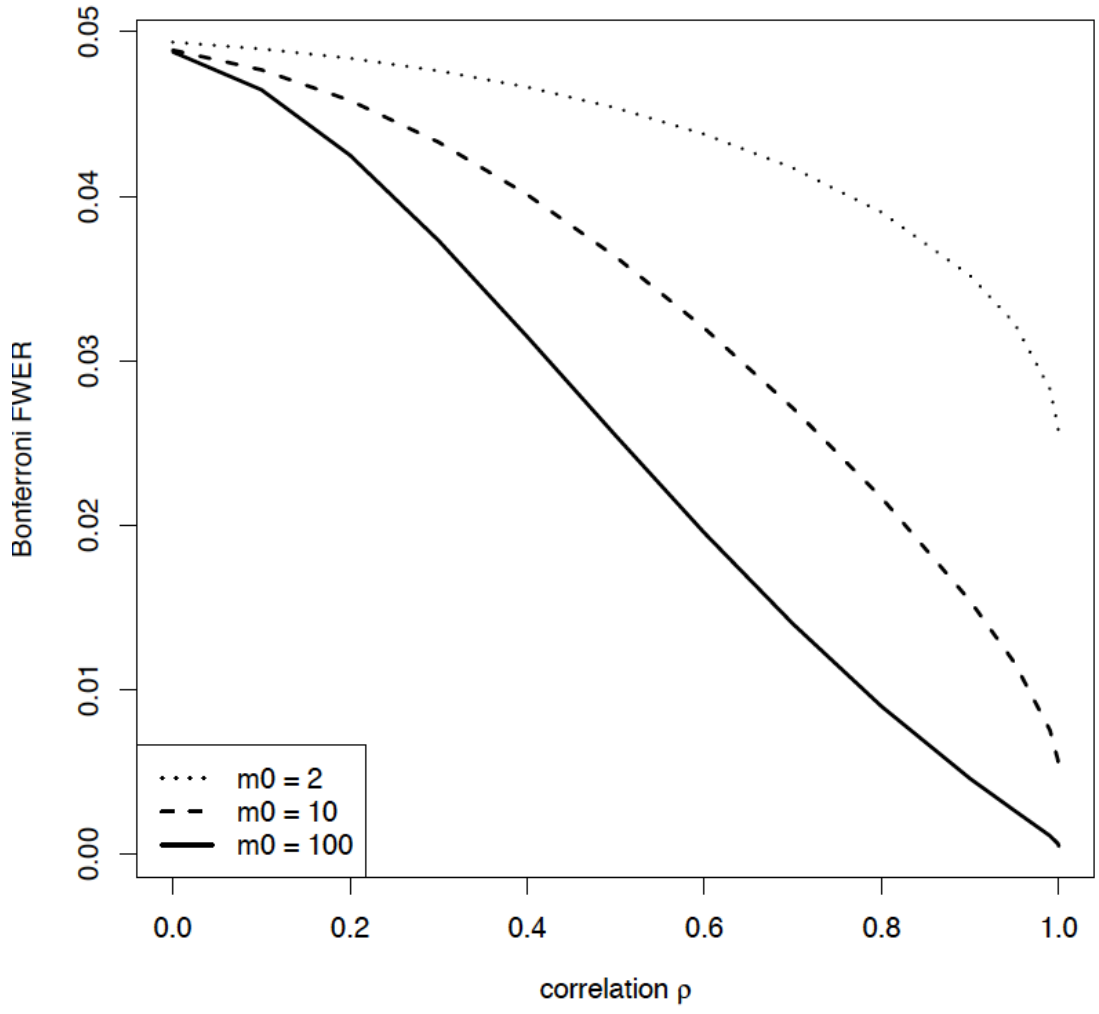
$$\text{pr}\Big(\bigcup_{i=1}^{m_0}\{q_i \leq \frac{\alpha}{m}\}\Big) \leq \sum_{i=1}^{m_0}\text{pr}\Big(q_i \leq \frac{\alpha}{m}\Big) \leq m_0\frac{\alpha}{m}$$

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

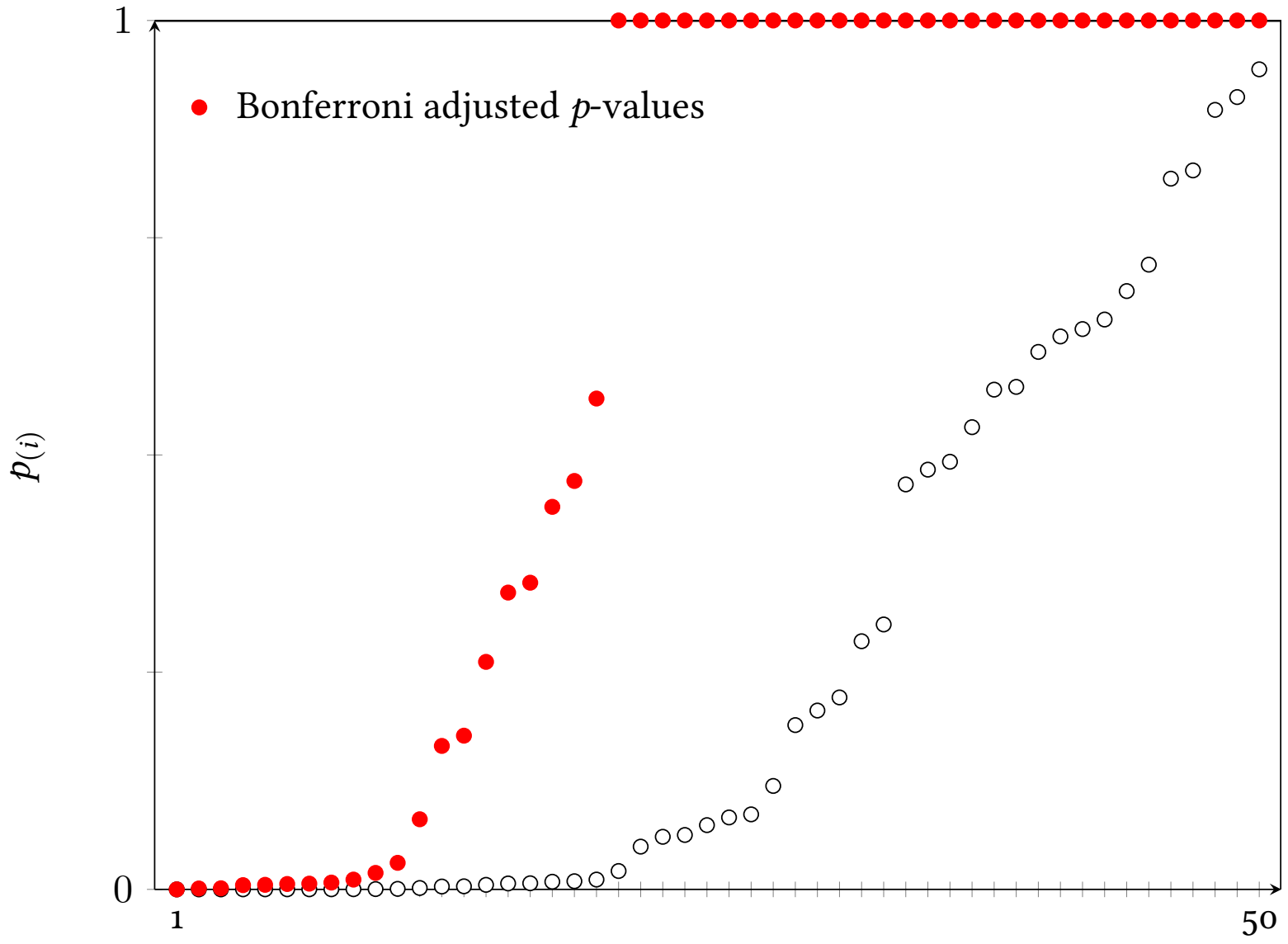
By definition, the p-value is the smallest chosen α-level of the test at which the hypothesis would have been rejected

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:

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

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

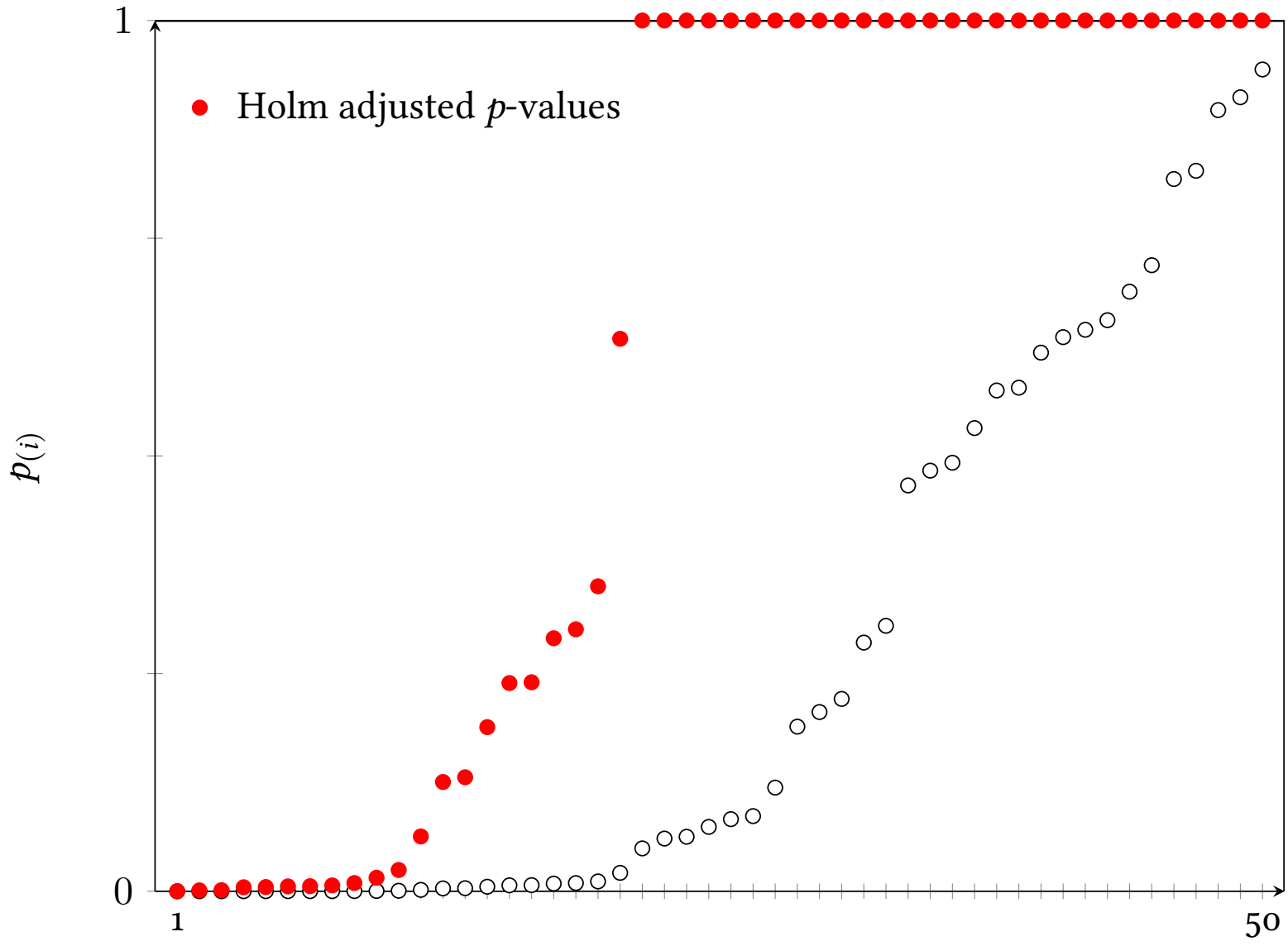

1
0
$p_{(i)}$
Holm adjusted p-values
1
50

Table of Contents

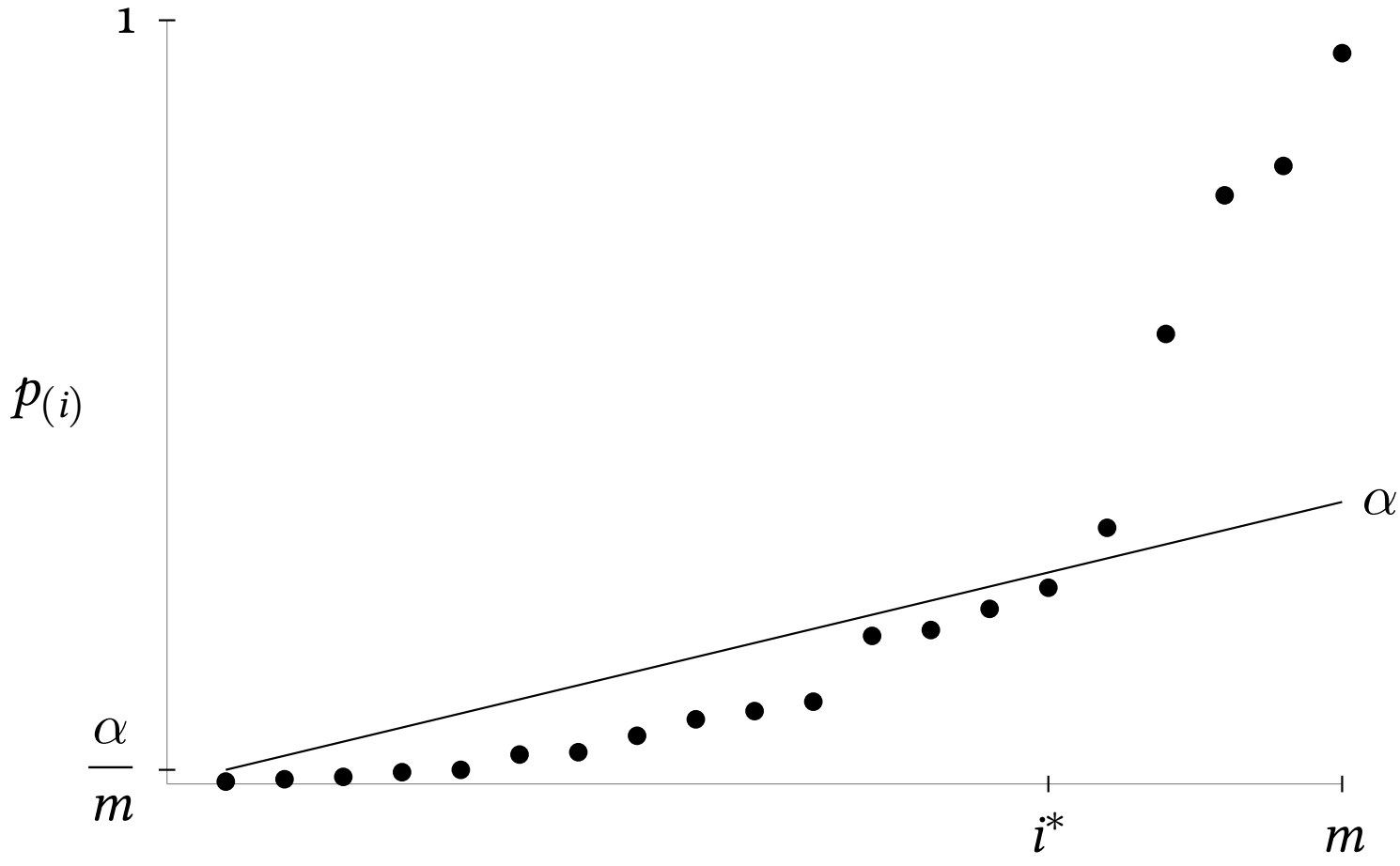
If we are testing millions of hypotheses at once, and making few false discoveries is not the end of the world

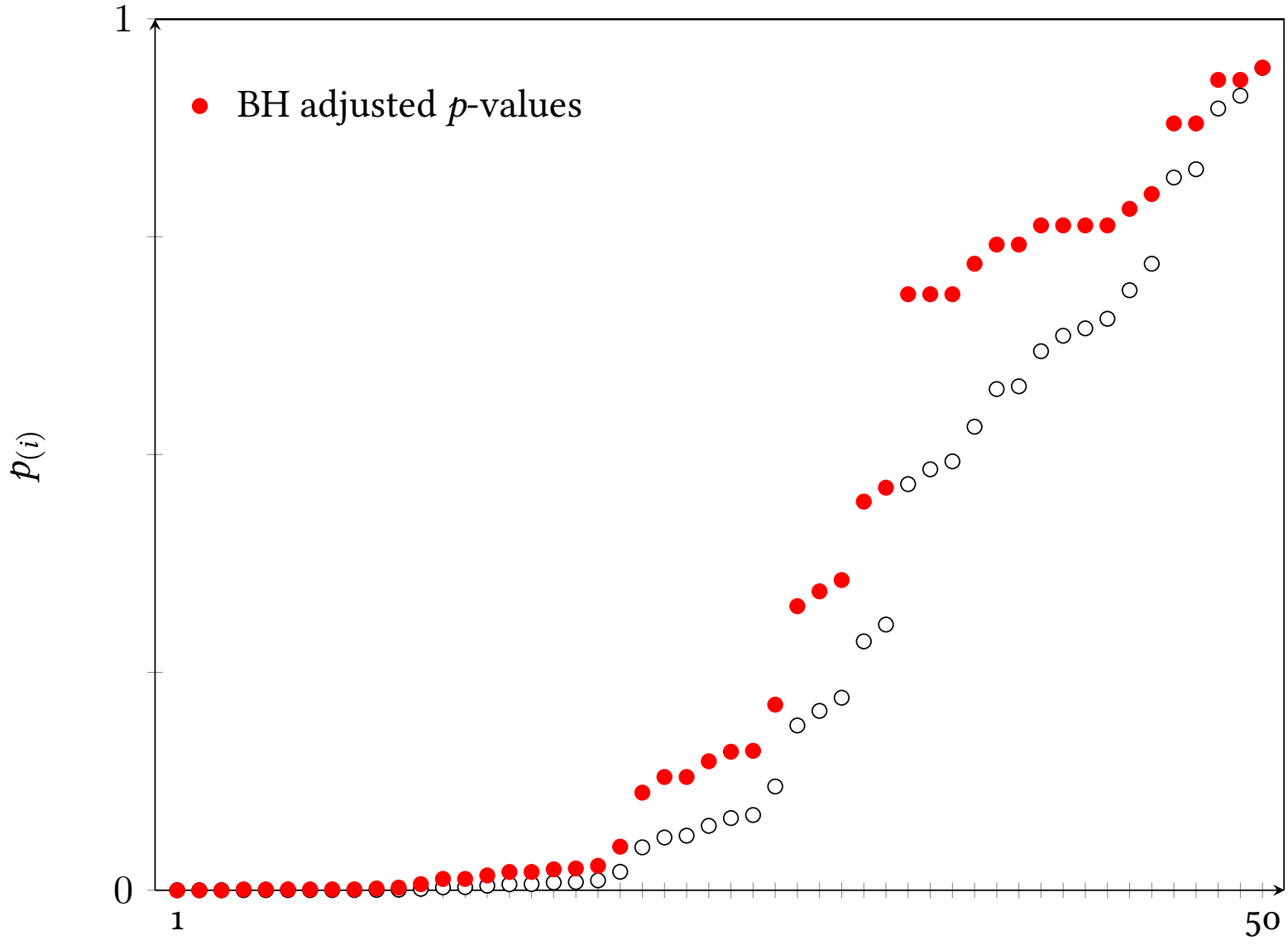
The concept of False Discovery Rate (FDR) has changed thinking about multiple testing quite radically, showing that FWER control is not only way to do of multiple testing, and stimulating the field of multiple testing enormously

FDR was introduced by Benjamini and Hochberg in 1995, and currently has 95K 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

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

□

PRDS

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 $\pi_0\alpha$.