Multiplicity Control in Clinical Trial

False Discovery Rate

Livio Finos

I thank Aldo Solari, Jelle Goeman and Florian Klinglmueller for the ideas and the materials we shared along these years. The material is the result of all this resoning together.

Outline

False Discovery Rate (FDR)

- Definition
- Methods

More about FDR

FWER or FDR?

A contingency table

Contingency table for multiple hypothesis testing

Rejection versus truth or falsehood of hypotheses

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

FDP, FWER and FDR

False Discovery Proportion

$$\mathrm{FDP} = \begin{cases} V/R & \text{if } R > 0 \\ 0 & \text{otherwise,} \end{cases}$$

Defined for every rejected set R

Familywise error rate

$$\mathrm{FWER} = \mathrm{P}(V > 0)$$

False discovery rate

$$\mathrm{FDR} = \mathrm{E}(\mathrm{FDP})$$

False Discovery Rate [1]

BH procedure

1. Sort the p-values: $p_{(1)}, \ldots, p_{(m)}$
2. Find j', the largest j such that $p_{(j)} \leq j\alpha/m$
3. Reject all hypotheses with p-values at most $p_{j'}$

Benjamini and Hochberg

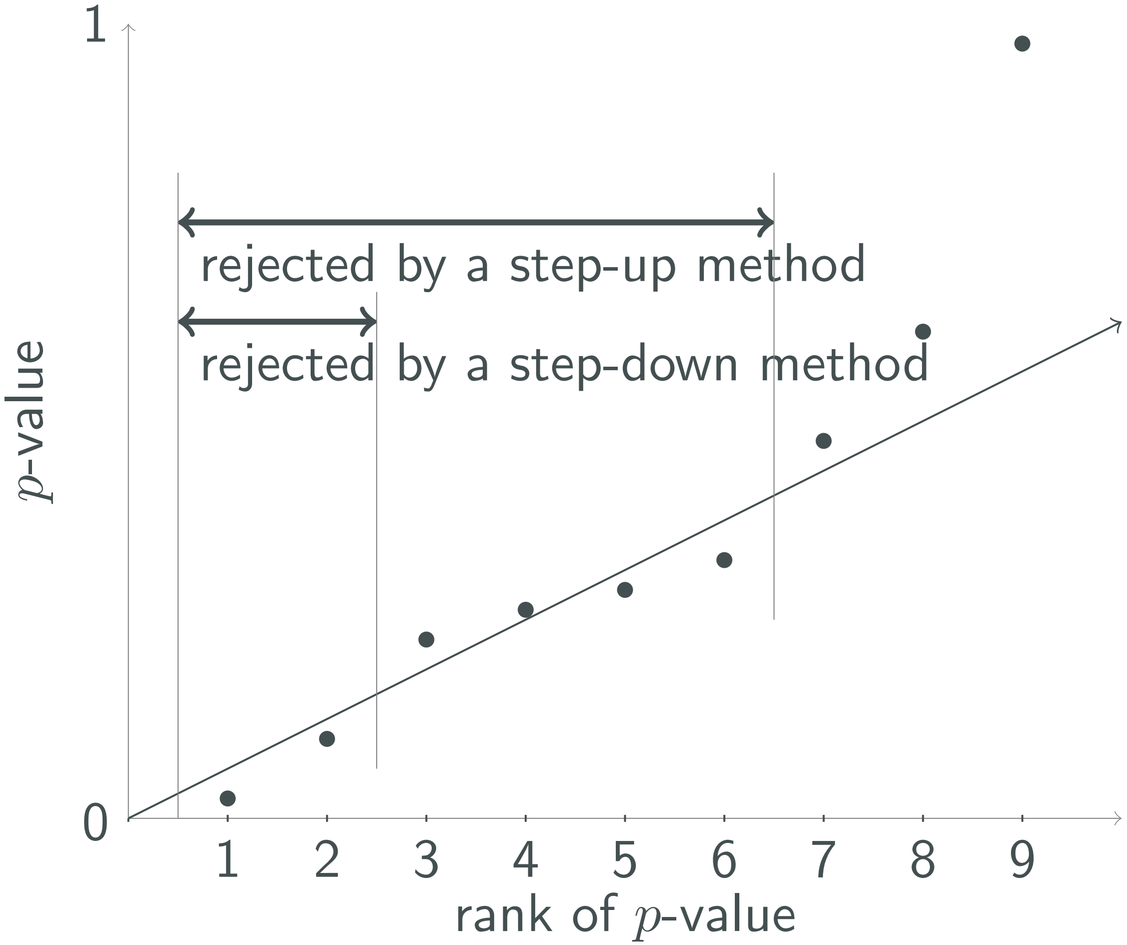
This procedure controls FDR under independence
Control is at $\pi_0\alpha$ (compare Bonferroni), with $\pi_0 = m_0/m$

Later

Conditions relaxed

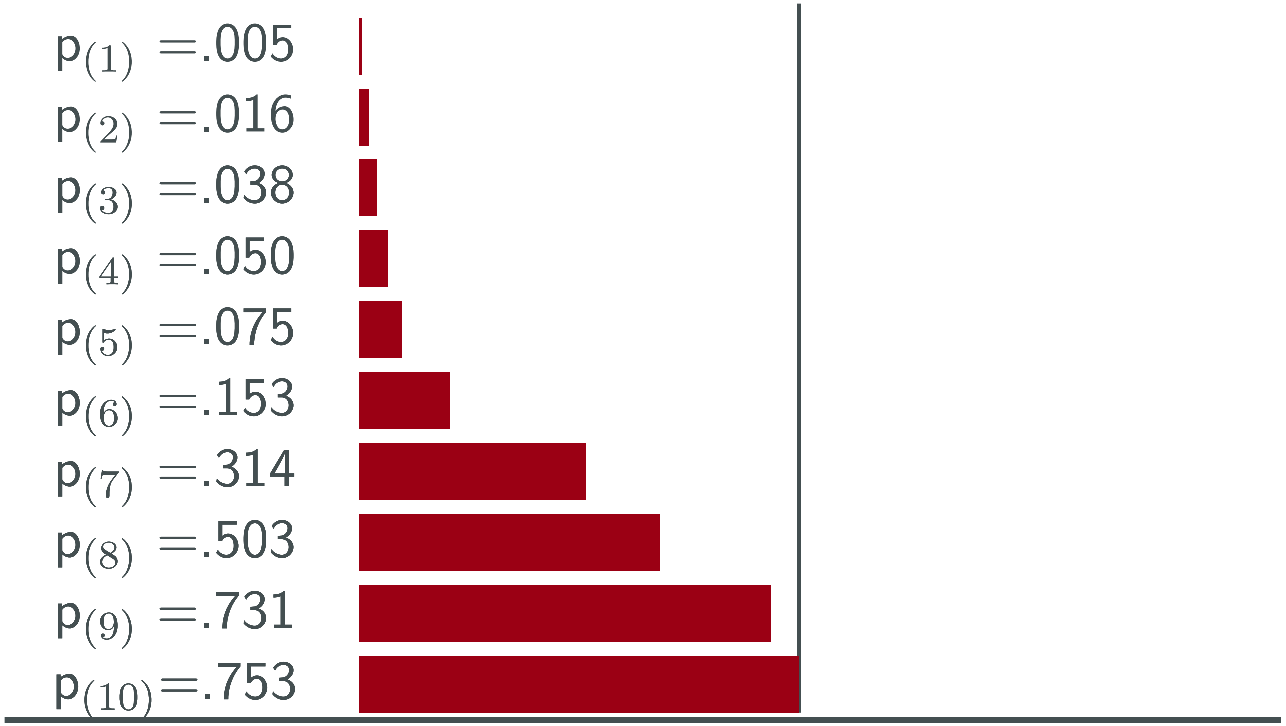
[1] Benjamini and Hochberg (1995). Journal of the Royal Statistical Society, Series B (Methodological) 57 (1): 289–300.

Step-down and step-up



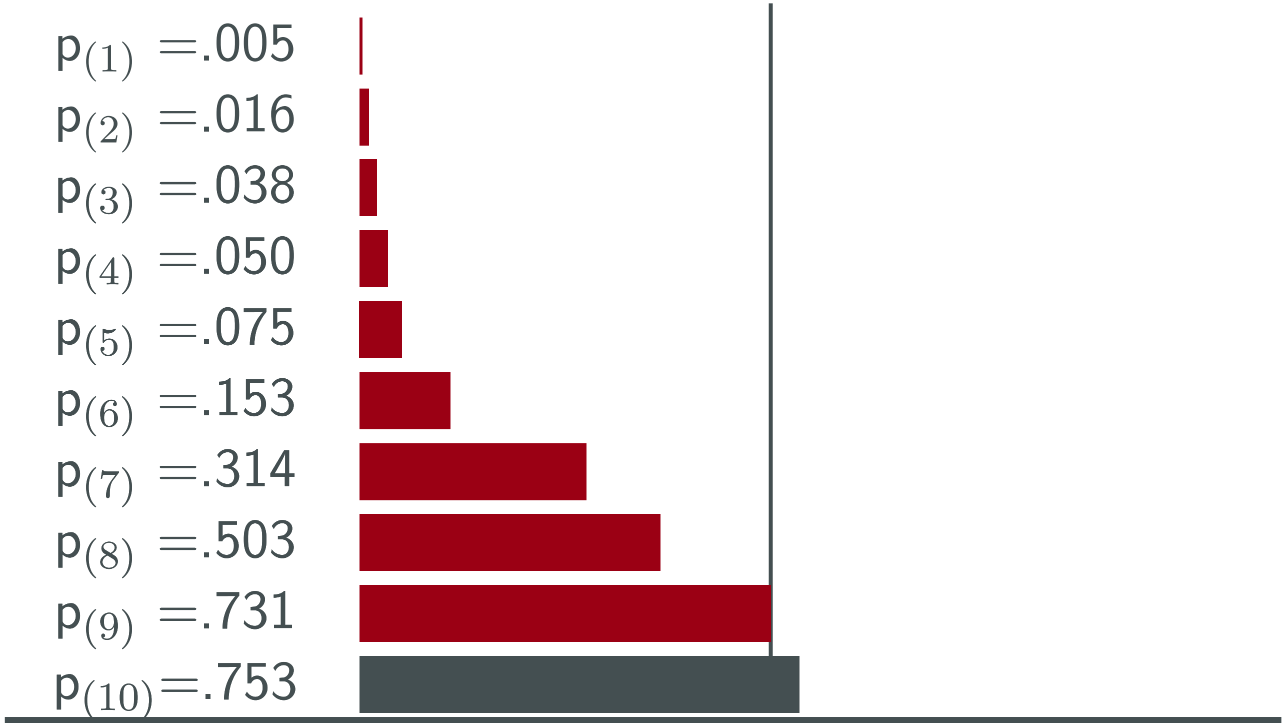
Benjamini and Hochberg (BH)

$$\frac{p_{(10)}\ \ m}{10} = \frac{0.753\ \ 10}{10} = 0.753 \overset{?}{\leq} \mathbf{q} = \mathbf{.10}\ :\ \text{No}$$



Benjamini and Hochberg (BH)

$$\frac{p_{(9)}\ m}{9} = \frac{0.731\ \ 10}{9} = 0.812 \overset{?}{\leq} \mathbf{q} = \mathbf{.10}\ :\ \text{No}$$



$p_{(1)}$ =.005

$p_{(2)}$ =.016

$p_{(3)}$ =.038

$p_{(4)}$ =.050

$p_{(5)}$ =.075

$p_{(6)}$ =.153

$p_{(7)}$ =.314

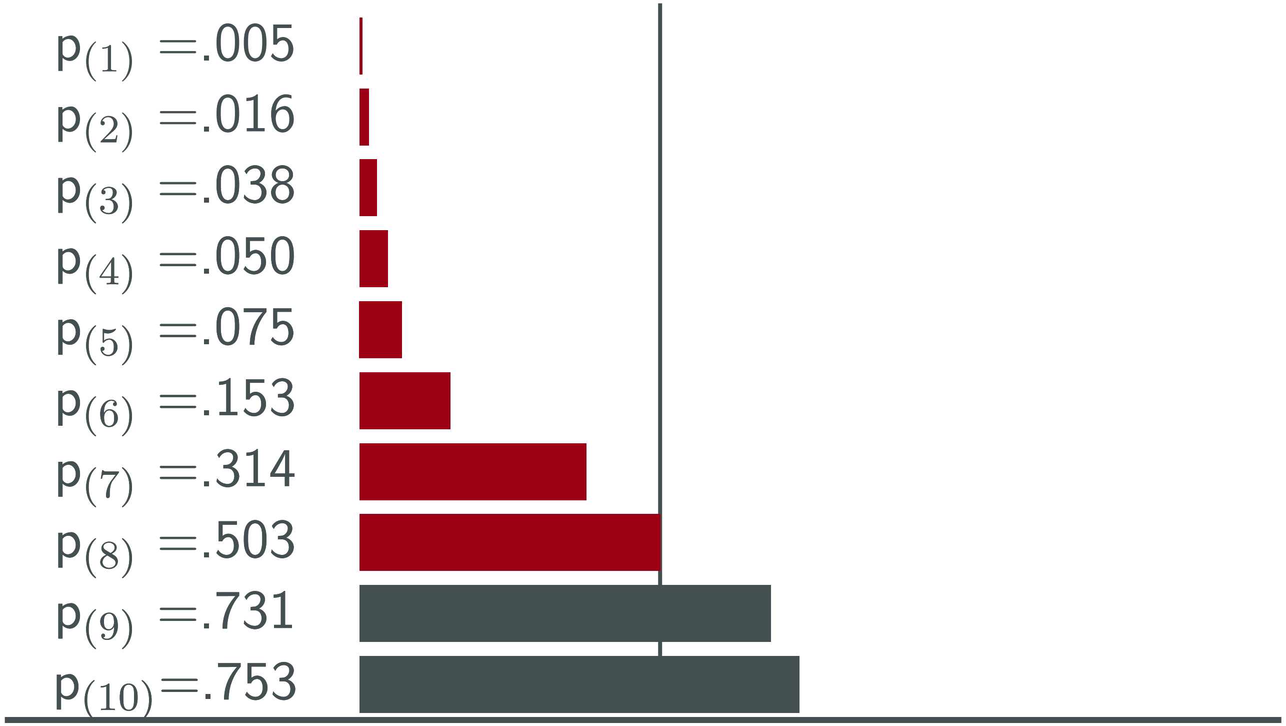
$p_{(8)}$ =.503

$p_{(9)}$ =.731

$p_{(10)}$=.753

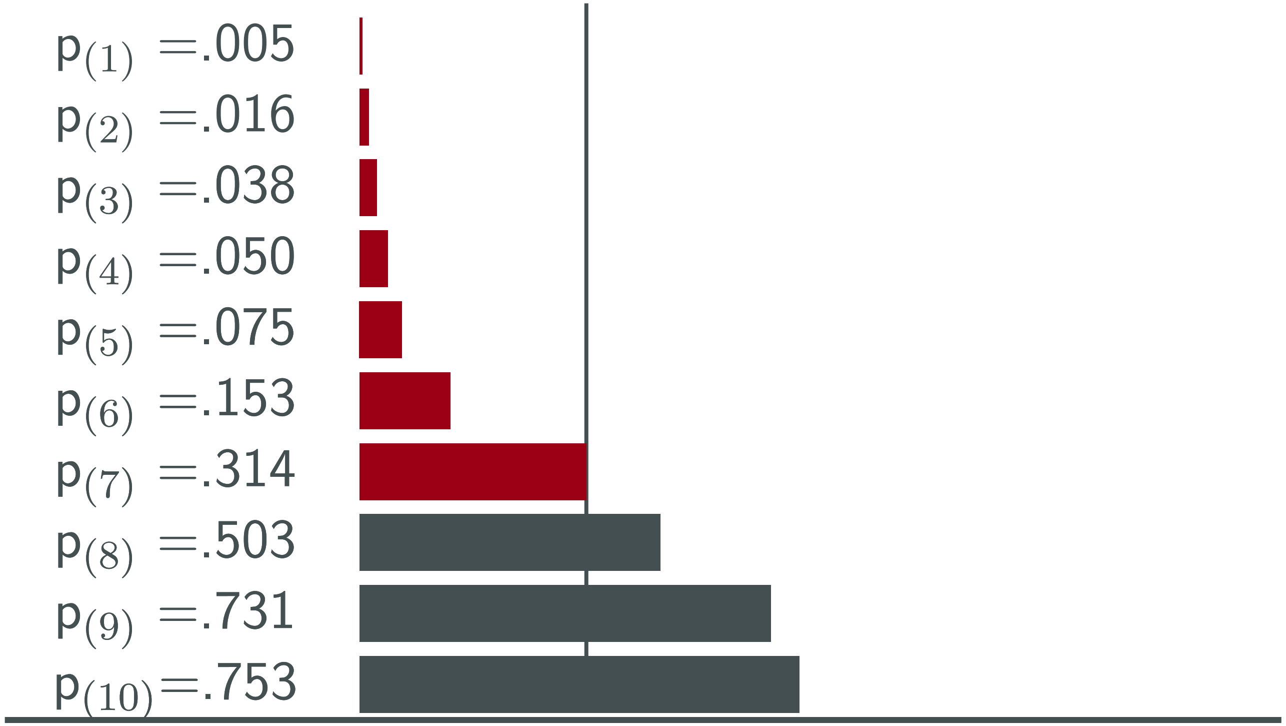
Benjamini and Hochberg (BH)

$$\frac{p_{(8)}\ m}{8} = \frac{0.503\ \ 10}{8} = 0.629 \overset{?}{\leq} \mathbf{q} = \mathbf{.10}\ :\ \text{No}$$



Benjamini and Hochberg (BH)

$$\frac{p_{(7)}\ m}{7} = \frac{0.314\ \ 10}{7} = 0.449 \overset{?}{\leq} \mathbf{q} = \mathbf{.10}\ :\ \text{No}$$



Benjamini and Hochberg (BH)

$$\frac{p_{(6)}\ m}{6} = \frac{0.153\ \ 10}{6} = 0.255 \overset{?}{\leq} \mathbf{q} = \mathbf{.10}$$: No

$\mathsf{p}_{(1)}$ =.005

$\mathsf{p}_{(2)}$ =.016

$\mathsf{p}_{(3)}$ =.038

$\mathsf{p}_{(4)}$ =.050

$\mathsf{p}_{(5)}$ =.075

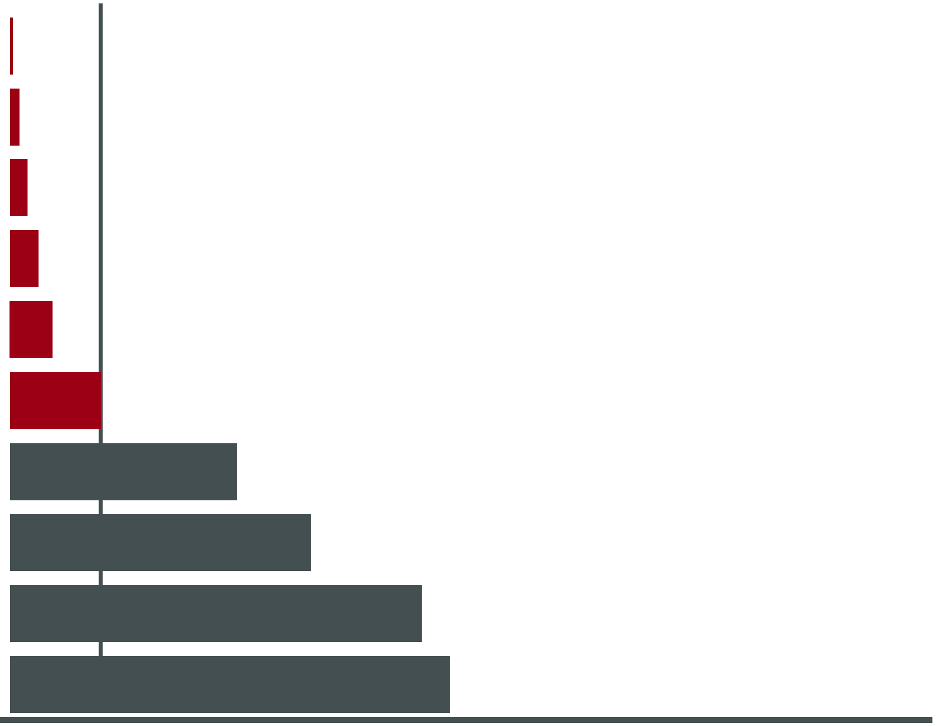
$\mathsf{p}_{(6)}$ =.153

$\mathsf{p}_{(7)}$ =.314

$\mathsf{p}_{(8)}$ =.503

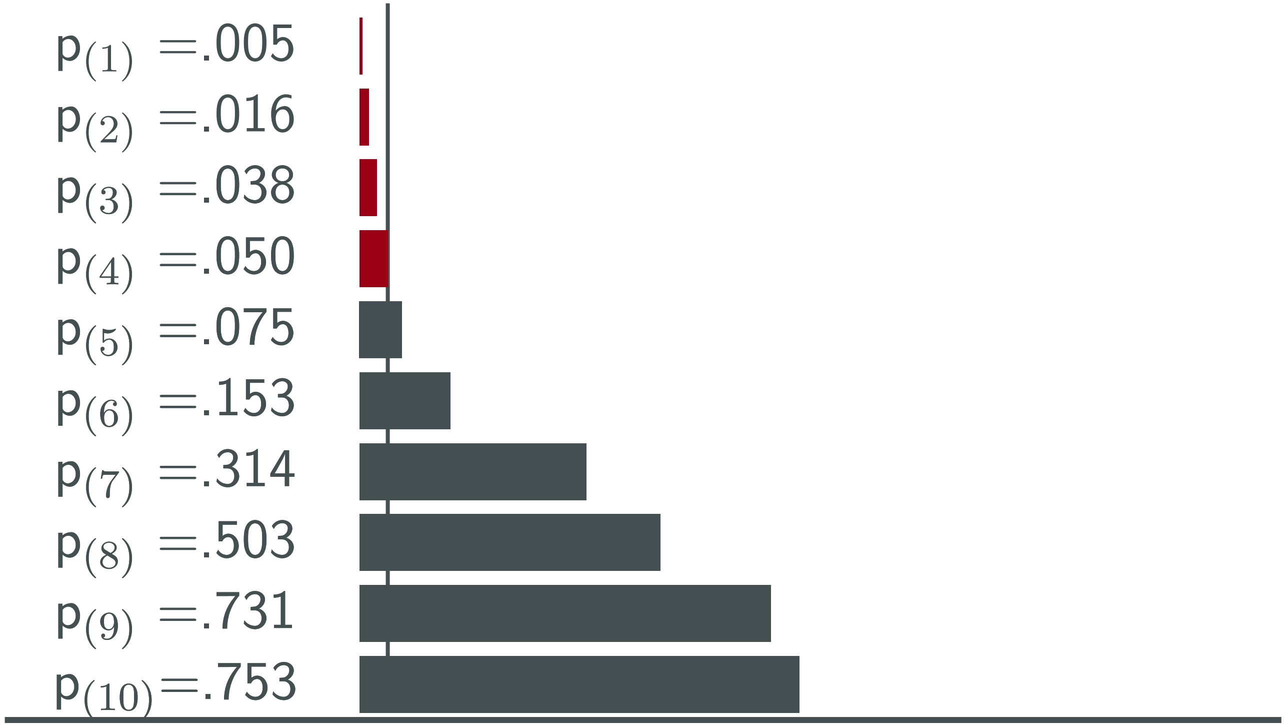
$\mathsf{p}_{(9)}$ =.731

$\mathsf{p}_{(10)}$=.753



Benjamini and Hochberg (BH)

etc.



$p_{(1)} = .005$

$p_{(2)} = .016$

$p_{(3)} = .038$

$p_{(4)} = .050$

$p_{(5)} = .075$

$p_{(6)} = .153$

$p_{(7)} = .314$

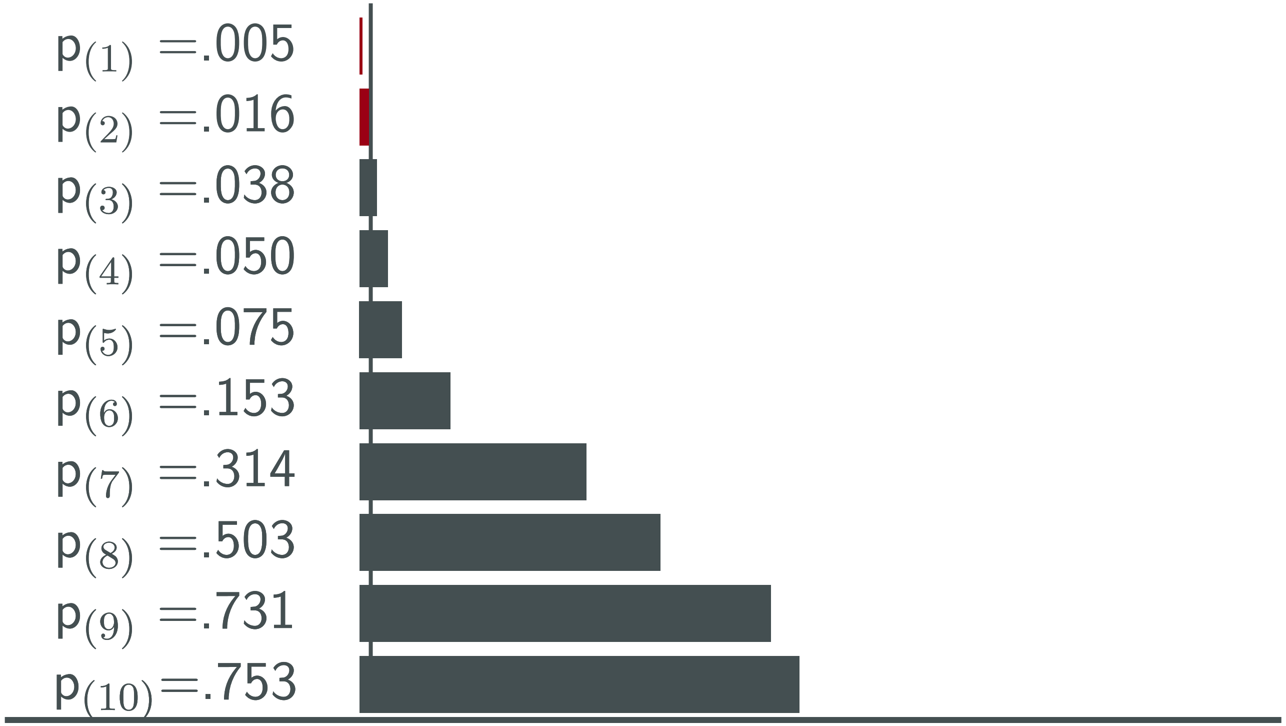
$p_{(8)} = .503$

$p_{(9)} = .731$

$p_{(10)} = .753$

Benjamini and Hochberg (BH)

$$\frac{p_{(2)}\ m}{2} = \frac{0.016\ \ 10}{2} = 0.080 \overset{?}{\leq} \mathbf{q} = \mathbf{.10}$$: Yes!, STOP



Meaning of FDR control

FDR control

On average, the $\mathcal{R}$ returned by BH has FDP $\leq \pi_0\alpha$

Variability of FDP

Due to variability of $\mathcal{R}$

Realized FDP

Varies around $\pi_0\alpha$.
Variability can be high if p-values correlated

Benjamini & Yekutieli (BY)[2]

Assumptions of Benjamini and Hochberg

Non-negatively associated p-values (*Positive Dependence through Stochastic ordering*)

i.e.

One-sided tests

As long as test statistics not negatively correlated

Two-sided tests

If test statistics are (asymptotically) normal

[2] Benjamini Y, Yekutieli D. (2001) The control of the false discovery rate in multiple testing under dependency. *Annals of statistics* 29(4):1165–1188

Benjamini & Yekutieli (BY)[3]

Benjamini and Yekutieli

Variant valid for any distribution of p-values

How does it work?

Same as BH but $\frac{p_{(i)}\ m}{i}\ L \overset{?}{\leq} q = .10$
with $L = \sum_{j=1}^{i} 1/j$ (es $i = 3$: $L = 1/1 + 1/2 + 1/3$)

In practice

- Quite conservative, especially if m_0 is large
- Not often needed, not often used

Sotware

BH e BY: `library(stats); p.adjust()`

[3] Benjamini Y, Yekutieli D. (2001) The control of the false discovery rate in multiple testing under dependency. *Annals of statistics* 29(4):1165–1188

Outline

False Discovery Rate (FDR)

- Definition
- Methods

More about FDR

FWER or FDR?

Adaptive FDR control

BH controls FDR at $\pi_0\alpha$

If π_0 were known, use $\tilde{\alpha} = \alpha/\pi_0$ instead

Adaptive FDR control idea

Estimate π_0 by $\hat{\pi}_0$ and use $\tilde{\alpha} = \alpha/\hat{\pi}_0$

Various methods available

- Higher power if π_0 low, lower power if $\pi_0 \approx 1$
- May reject hypotheses with p-values $> \alpha$
- FDR control under dependence not guaranteed

Storey's FDP estimate

Rejected set

Suppose we reject hypotheses $\mathcal{R} = \{H_i : p_i \leq t\}$

Intuition

By uniformity of p-values under the null $\mathrm{FDP} \approx m_0 t / \#\mathcal{R}$

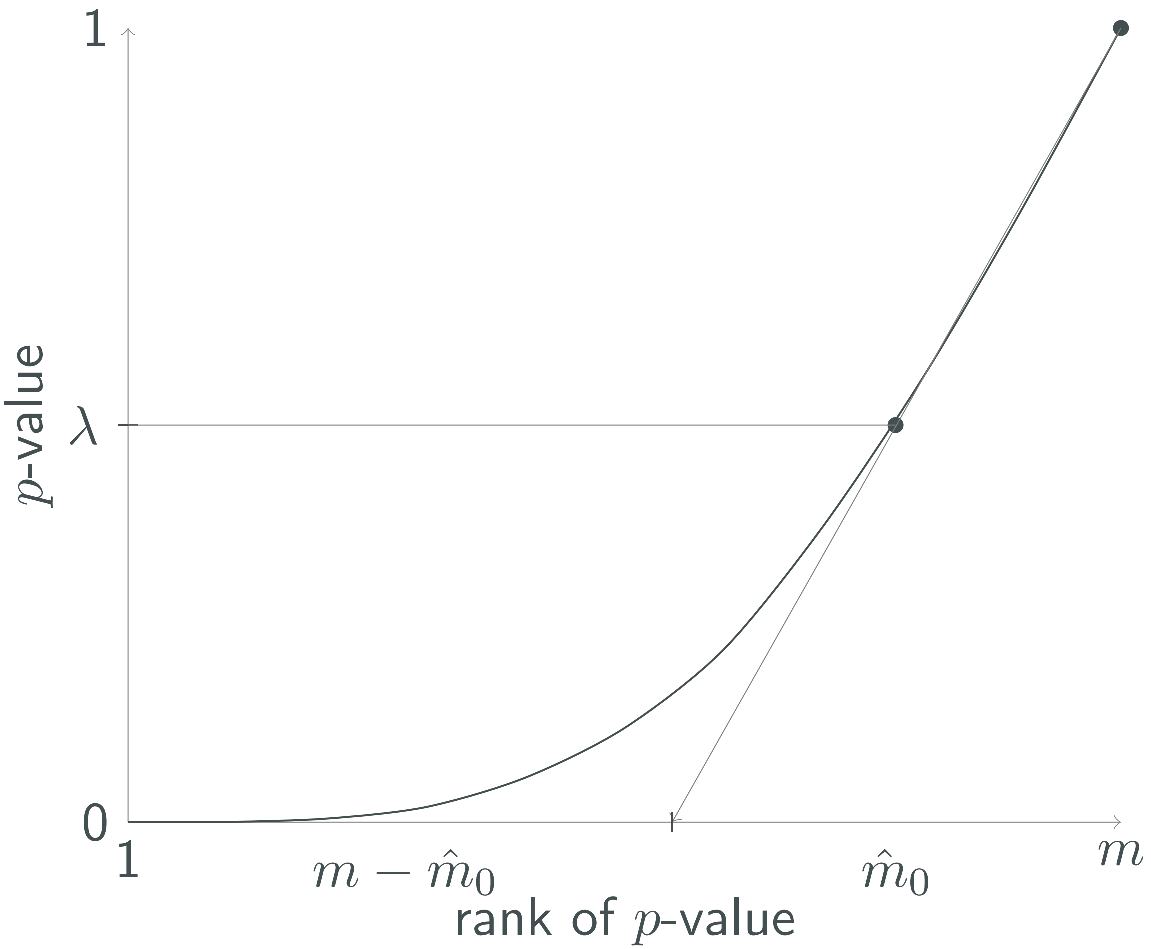
Estimate of m_0 (again by uniformity)

$$\hat{m}_0 = \frac{\#\{p_i > \lambda\} + 1}{1 - \lambda}$$

Resulting estimate of FDP ("q-value")

$$\widehat{\mathrm{FDP}} = \frac{\hat{m}_0 t}{\#\mathcal{R}} = \frac{t}{1 - \lambda} \frac{\#\{p_i > \lambda\} + 1}{\#\{p_i \leq t\}}$$

Storey's π_0 estimation



Storey and Benjamini & Hochberg

Close relationship

Alternative way of constructing BH rejected set

1. Estimate $\hat{m}_0 = 1$ instead of Storey's estimate
2. Take t the largest value such that $\widehat{\mathrm{FDP}} \leq \alpha$

Alternative look at Storey

Storey's method $=$ adaptive FDR control

Alternative look at BH

Conservative estimates of FDP

Storey and dependence

Method of moments estimate

Only dependent on means → unaffected by correlation structure

However

Variability of estimate can be large if p-values correlated

Standard errors unavailable

Available for independent p-values only

Use of FDP estimation

Point estimation

No standard errors

For the rest

Very similar to adaptive FDR methods

- Remember that FDP estimate is for the R set: No subsetting property
- FDP can be (widely) underestimated

Outline

False Discovery Rate (FDR)

- Definition
- Methods

More about FDR

FWER or FDR?

Bonferroni-bashing

Often heard

“Never use Bonferroni: it is too conservative”

Is this true?

- Is $m_0 \ll m$?
- Are p-values highly superuniform?
- Are p-values highly positively correlated?

Otherwise

Bonferroni is not conservative, but FWER is strict

Four flavors of multiple testing

FWER control at 5%

95% of experiments give no type I errors

FDR control at 5%

On average, experiments give no more than 5% FDP

FDP estimation

Get a (conservative) point estimate of FDP in every experiment

FDP confidence 95%

Overstate the FDP at most 5% of the time

FWER or FDR?

Implicit Assumptions in FDR

The hypotheses are exchangeable:
False Rejections compensate True Rejections

FWER or FDR?

Implicit Assumptions in FDR

The hypotheses are exchangeable:
False Rejections compensate True Rejections

Problems

- Cheating
- Subsets

Cheating

Adding un-interesting hypoteses to be rejected, so that more false rejections are allowed.

[4] Finner H, Roters M. (2001) On the false discovery rate and expected type I errors. *Biometrical Journal*; 43(8):985–1005

Cheating

Adding un-interesting hypoteses to be rejected, so that more false rejections are allowed.

Subsets

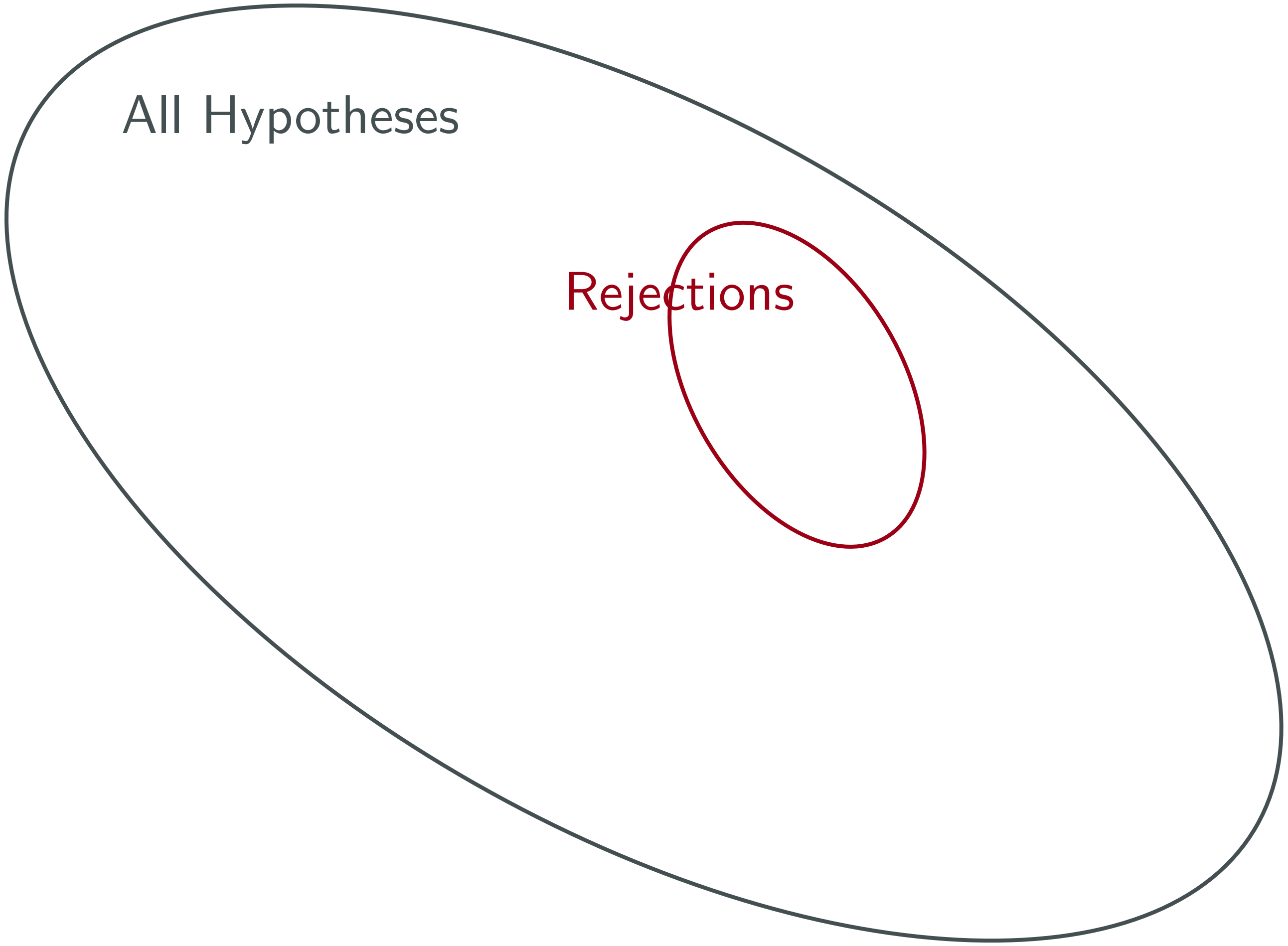
FDR is about the set R, not about individual hypotheses: Control of FDR in R does NOT imply control of the FDR in all subsets

Finner and Roters[4]

- FDR control on all subsets = FWER control
- FWER control on all subsets = FWER control

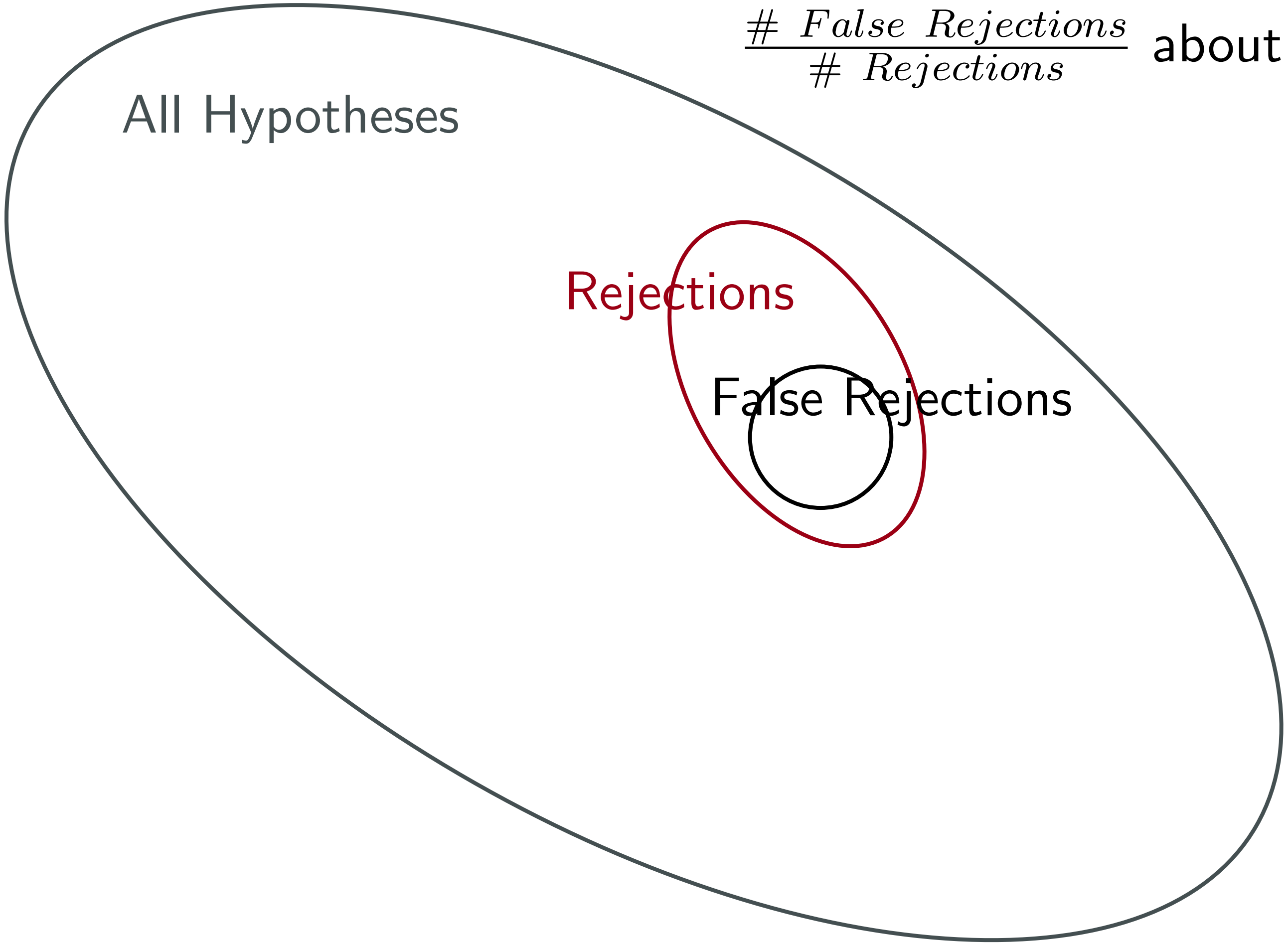
[4] Finner H, Roters M. (2001) On the false discovery rate and expected type I errors. *Biometrical Journal*; 43(8):985–1005

Subsets of Rejected hypotheses

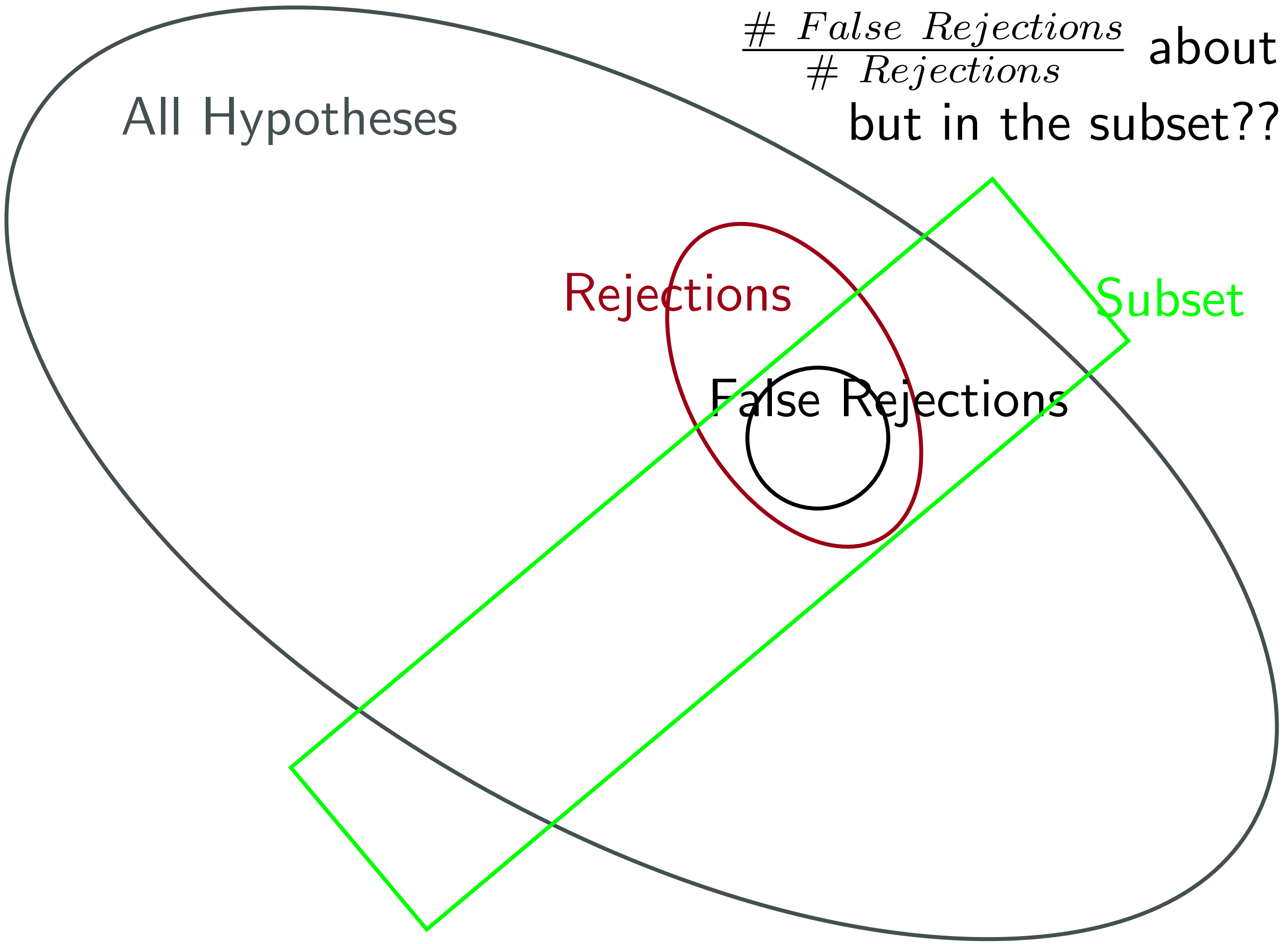


Subsets of Rejected hypotheses

$\frac{\# \ False\ \ Rejections}{\# \ Rejections}$ about 0.10



Subsets of Rejected hypotheses



$\frac{\#\ False\ Rejections}{\#\ Rejections}$ about 0.10

but in the subset??

Relationships between FWER and FDR

Dominance

$$\mathrm{P}(V > 0) = \mathrm{E}(\mathbf{1}\{V > 0\}) \geq \mathrm{E}(\mathrm{FDP})$$

Consequence: Control of FWER implies control of FDR

Complete null hypothesis

If all hypotheses true, $\mathrm{FDP} = \mathbf{1}\{V > 0\}$
Consequence: If all hypotheses true, FDR $=$ FWER

Single hypothesis

If only one hypothesis, $\mathrm{FDP} = \mathbf{1}\{V > 0\}$
Consequence: If only one hypothesis, FDR $=$ FWER $=$ Type I error

FWER vs. FDR: scaling

Scaling

As the size m of the problem grows
(complete null not true)

FWER

- Number of rejections remains limited
- Number of errors remains limited

FDR

- Number of rejections grows with m
- Number of errors grows with m

When to use FDR

- If collection of rejections important
- If validation experiments follow
- If hypotheses are exchangeable
- If power is an issue

Take-home message

- molteplicity control is mandatory in Clinical Trials
- FWER: controlling the probability of at least one error
- FDR: controlling the proportion of false rejection (on average)
- FWER is
  - a stronger control
  - usually preferible in Clinical Trials
  - more flexible
- FWER and FDR easy in R
- excellent tutorial: Goeman & Solari (2014) [5]

[5] JJ Goeman, A Solari (2014) Tutorial in biostatistics: multiple hypothesis testing in genomics. Statistics in medicine, Volume 33, Issue 11