The Labyrinth of Multiple Testing: How to avoid the pitfall of false positives

FDR control

12th SISMEC National Congress 2023

Livio Finos and Angela Andreella

Outline

A contingency table

		Null hypothesis		
		False	True	Tot
Test	Rejected	S	V	R
	Not rejected	T	U	$m - R$
	Tot	m_1	m_0	m

False Discovery Proportion:

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

Defined for every rejected set $\mathcal{R}$ where $\#\mathcal{R} = R$.

FWER and FDR

Familywise error rate

$$\boxed{\text{FWER} = \mathbb{P}(V > 0)}$$

A procedure controls it if FWER $\leq \alpha$.

FWER focuses on the probability that the rejected set contains **any error**.

False discovery rate

$$\boxed{\text{FDR} = \mathbb{E}(\text{FDP})}$$

A procedure controls it if FDR $\leq \alpha$

FDR looks at the **expected proportion** of errors among the rejections.

Benjamini and Hochberg (BH) procedure [1]

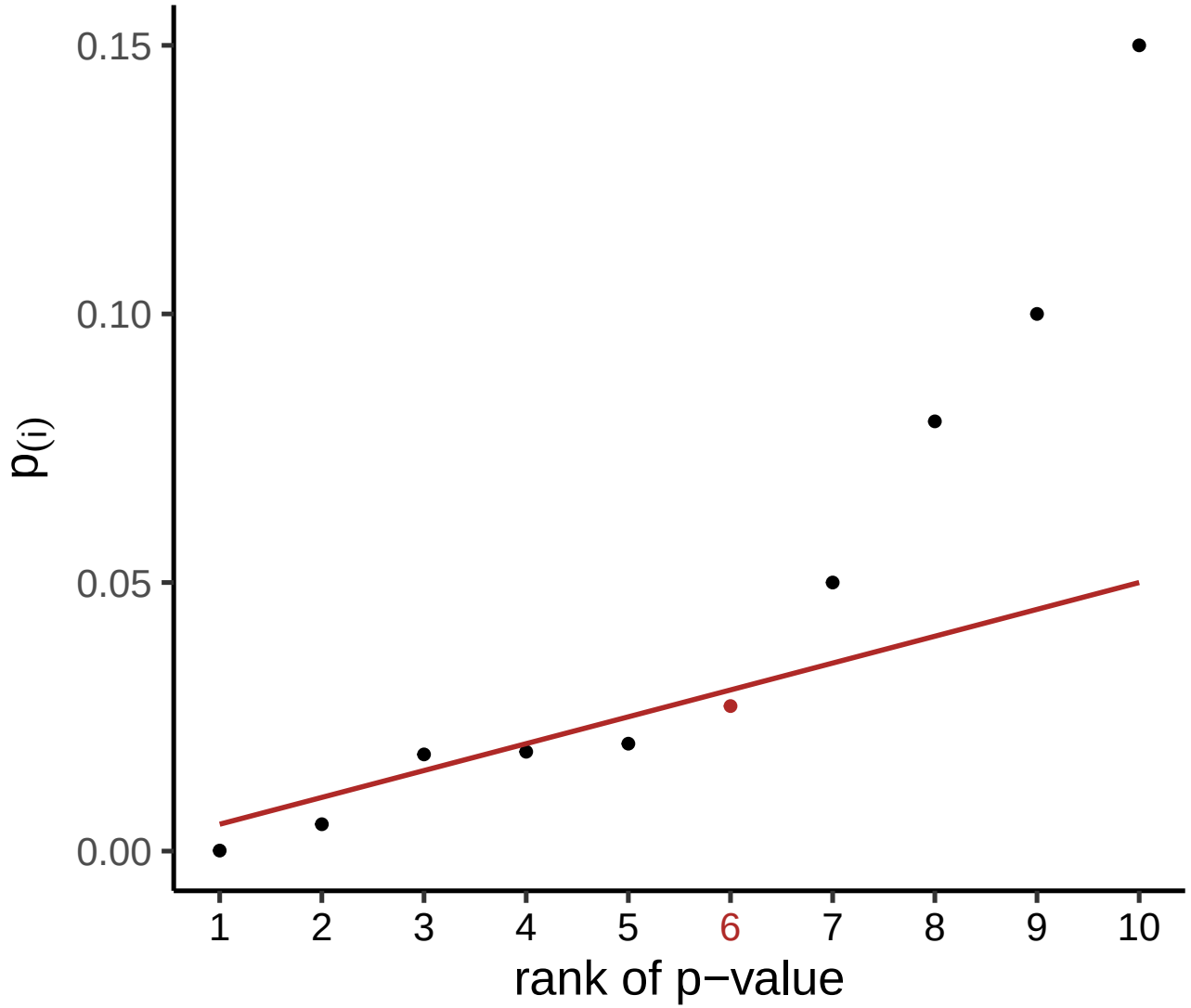
1. Sort the p-values: $p_{(1)}, \ldots, p_{(m)}$
2. Find j', the largest j such that

$$\boxed{p_{(j)} \leq j\alpha/m = c_j^{BH}}$$

3. Reject all hypotheses with p-values at most $p_{j'}$

[1] Benjamini, Y., Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B (Methodological), 57(1), 289-300.

Benjamini and Hochberg (BH) procedure



Benjamini and Hochberg (BH) procedure

In this procedure

- FDR control is at $\pi_0\alpha$ (compare Bonferroni), with $\pi_0 = m_0/m$
- controls FDR is valid under **independence** and **positive dependence through stochastic ordering** (i.e., non-negatively associated p-values):
  - One-sided tests: as long as test statistics not negatively correlated
  - Two-sided tests: If test statistics are (asymptotically) Normal
- we gain in power with respect to FWER-based methods when m_0 is large
- $\tilde{p}_{(i)} = \min\left(\frac{p^{\text{raw}}_{(i)} m}{i}, 1\right)$

Benjamini & Yekutieli (BY)[2]

Variant of BH valid for any distribution of p-values

How does it work?

Same as BH, but

$$\boxed{p_{(j)} \leq \frac{j\alpha}{mL} = c_j^{BY}}$$

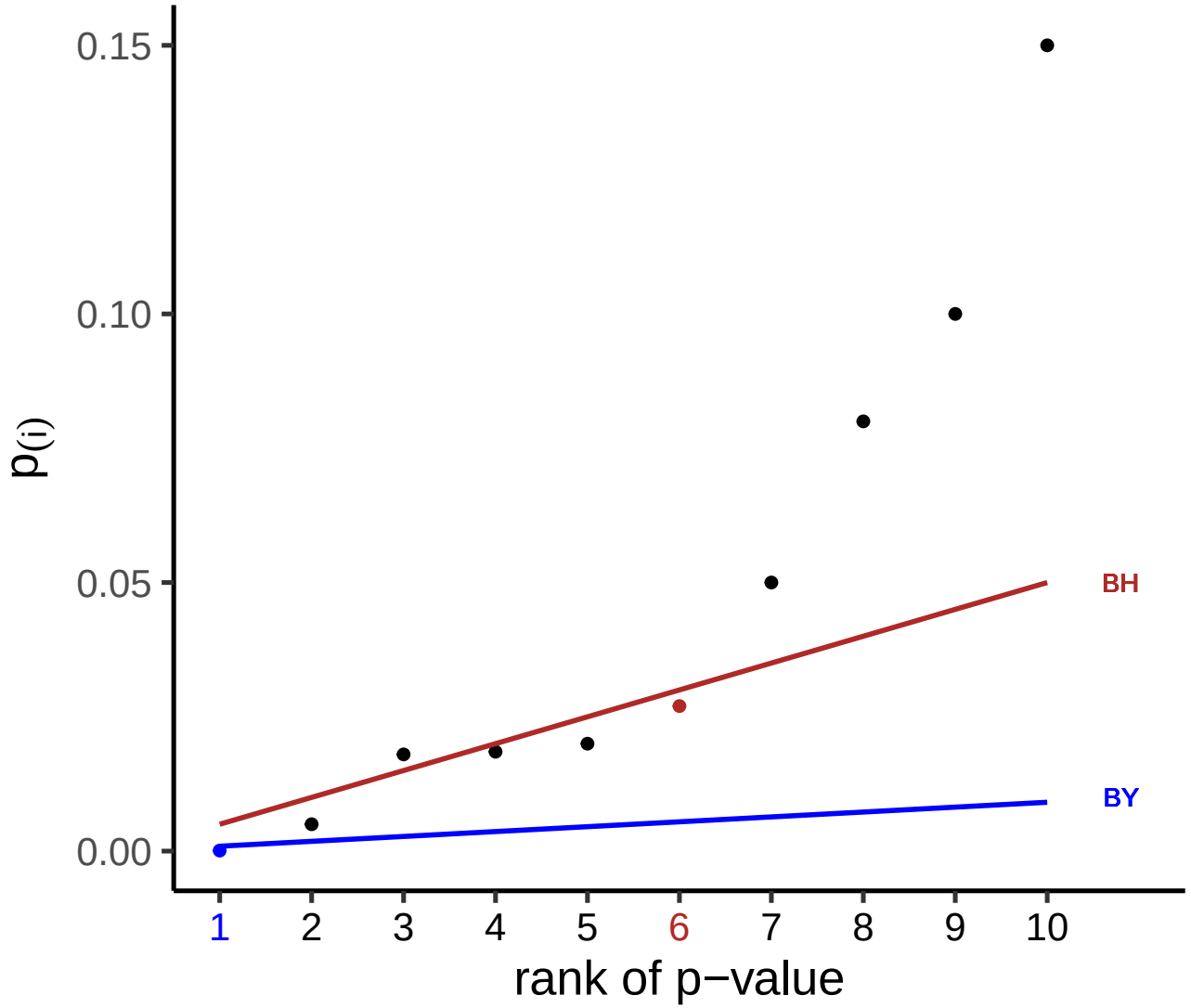
where $L = \sum_{j=1}^{m} 1/j$ (es $m = 3$: $L = 1/1 + 1/2 + 1/3$)

In practice

- Quite conservative (especially if m_0 is large):
  - $c_j^{BY} < c_j^{BH}$
  - $\tilde{p}_{(i)} = \min\left(\frac{p_{(i)}^{\text{raw}} Lm}{i}, 1\right)$
- Not often needed, not often used

[2] Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. Annals of Statistics, 1165-1188.

Benjamini & Yekutieli (BY)



False Discovery Rate Control

BH and BY methods are implemented in `R` by `p.adjust`:

- `p.adjust(p, "BH")`
- `p.adjust(p, "BY")`

Outline

FDP estimation

Difference between FDR control and FDP estimation

- **FDR control**: starts with the choice of α to be controlled and the procedure finds $\mathcal{R}$
- **FDP estimation**: starts with $\mathcal{R}$ (not necessarily the hypotheses with top p-values) and finds an estimate (or confidence interval) for FDP of that set.

To formulate the point estimation approach:

- $V(t) = \#\{\text{true null } H_i : p_i \leq t\}$
- $\mathcal{R}(t) = \{H_i : p_i \leq t\}$

$\rightarrow FDR(t) = \mathbb{E}(V(t)/\#\mathcal{R}(t))$ with $t \in (0, 1]$

Storey's FDP estimate

Intuition

By uniformity of p-values under the null

$$\mathrm{FDP}(t) = V(t)/R(t) \approx m_0 t/R(t)$$

Estimate of m_0 (again by uniformity)

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

where $0 < \lambda < 1$ constant (e.g., $\lambda = 1/2$, $\lambda = \alpha$).

Resulting estimate of FDP

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

Storey's FDP estimate

Storey's estimate is sometimes used as a way to control FDR, rather than as a way to estimate FDP: selecting the highest value of t such that the estimate $\widehat{\text{FDP}} \leq \alpha$.

Close relationship with BH

An alternative way of constructing BH rejected set

1. Estimate $\hat{m}_0 = m$ instead of Storey's estimate $\rightarrow$ $\hat{FDP} = mt/(\#\{p_i \leq t\})$
2. Take t the largest value such that $\widehat{\text{FDP}} \leq \alpha$

Alternative look at Storey

Storey's method = adaptive BH FDR control

Alternative look at BH

Conservative estimates of FDP

Storey's FDP estimate

Method of moments estimate

- Only dependent on means $\rightarrow$ unaffected by correlation structure
- Standard errors available for independent p-values only
- Variability of estimate can be large if p-values correlated $\rightarrow$ FDP can be (widely) underestimated.

Outline

Bonferroni-bashing

Often heard

"Never use Bonferroni: it is too conservative"

Is this true?

- Is $m_0 \ll m$?
- Are p-values highly superuniform (conservative, i.e., distribution around 1)?
- Are p-values highly positively correlated?

Otherwise

Bonferroni is not conservative, but FWER is strict

Meaning of FDR control

Recall that $\mathbb{E}(FDP) = \mathbb{E}(V/R) \leq \pi_0\alpha$

Therefore, FDR control is affected by FDP **variability** $\rightarrow R$ is **random**.

- Variability can be high if p-values correlated
- Users of FDR must be aware that control of FDR at α only controls FDP in **expectation** and that the actual FDP can often be $\gg$ than α.
- FDR control is a property of the procedure leading to a rejected set, not of the rejected set itself.

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 uninteresting hypotheses to be rejected so that more false rejections are allowed.

[3] Finner, H., & Roters, M. (2001). On the false discovery rate and expected type I errors. Biometrical Journal, 43(8), 985-1005.

Cheating

Adding uninteresting hypotheses to be rejected so that more false rejections are allowed.

Subsets

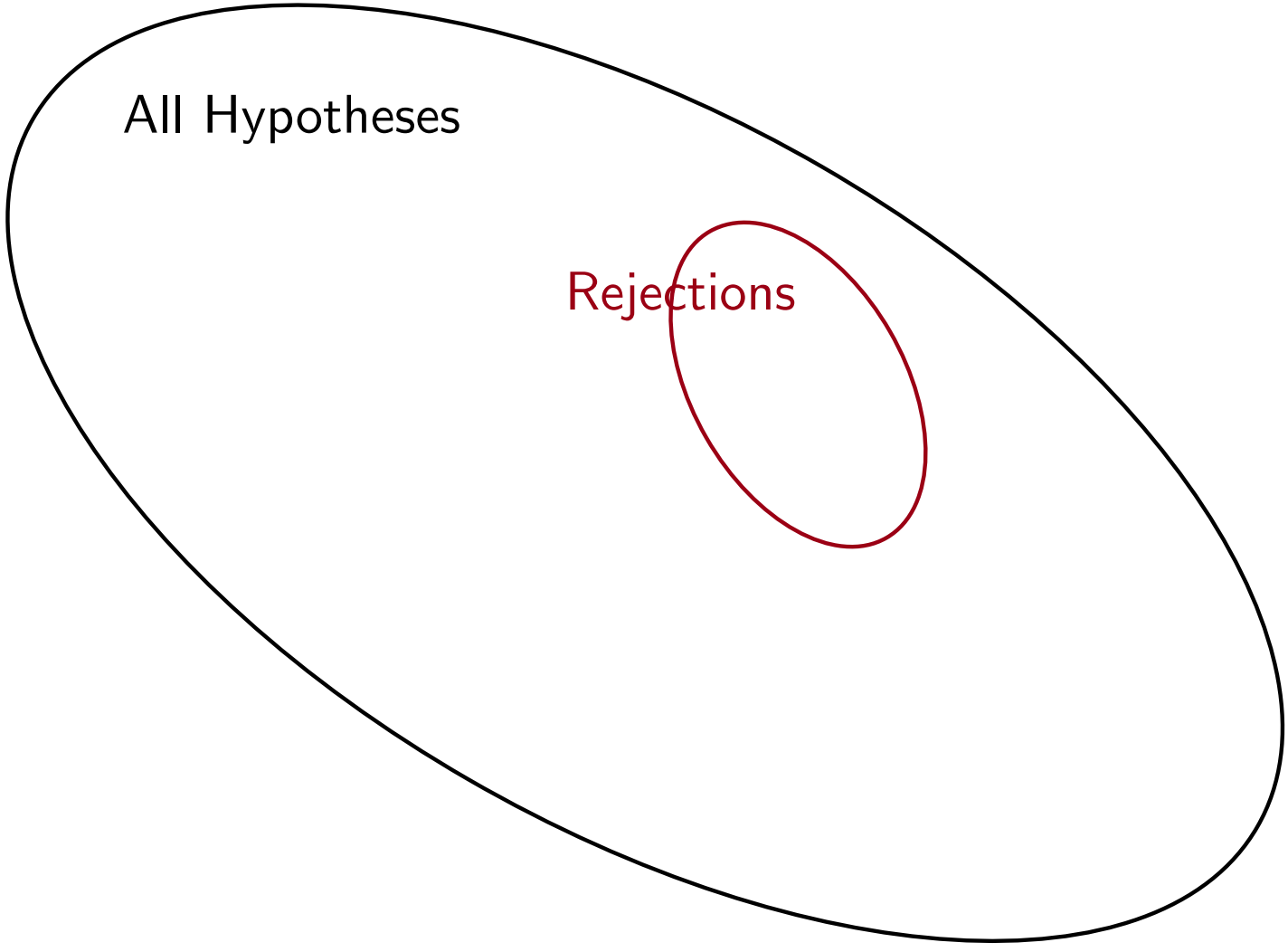
FDR is about the set R , not about individual hypotheses: Control of FDR in R does NOT imply control of the FDR in all subsets

Finner and Roters[3]

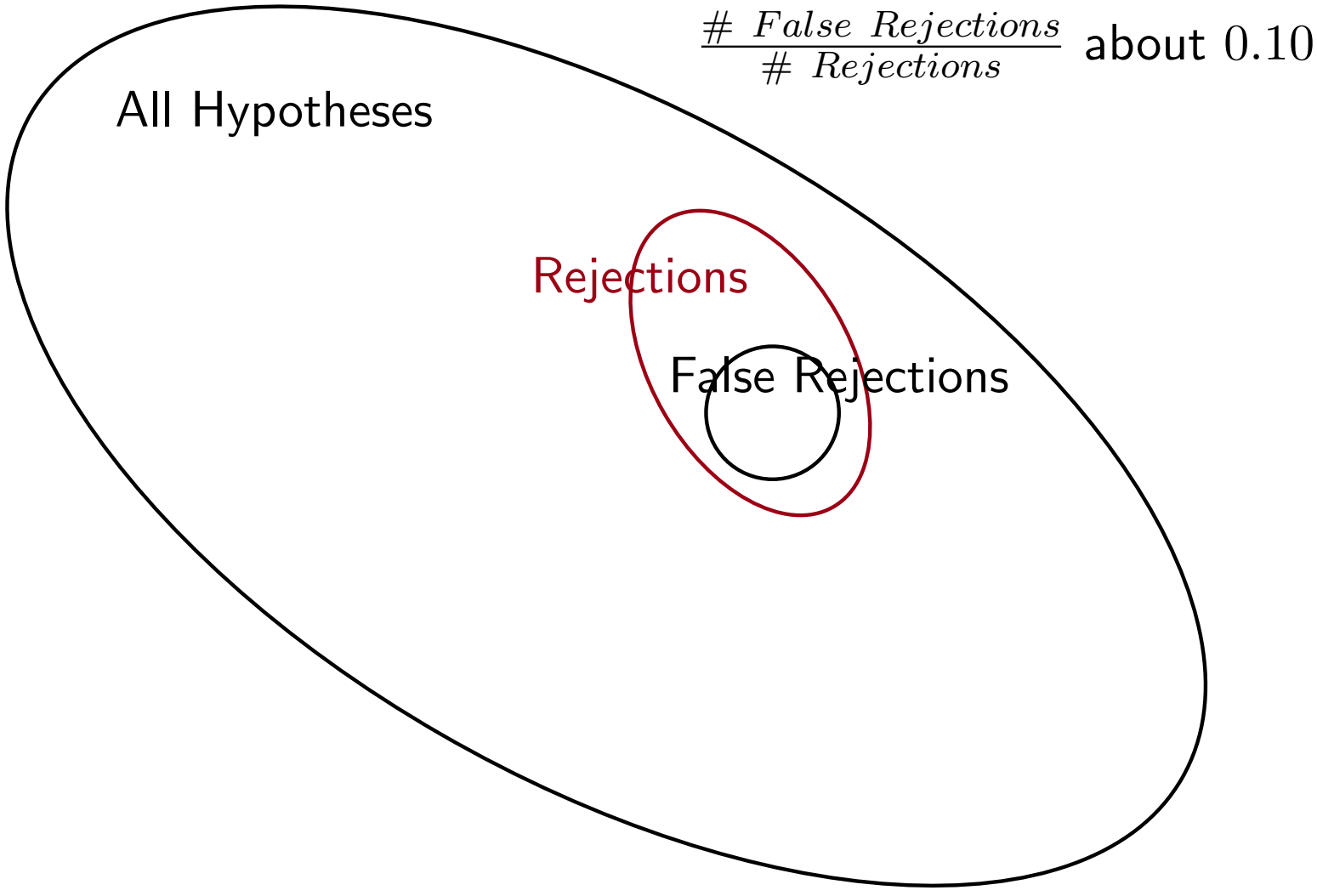
- FDR control on all subsets = FWER control
- FWER control on all subsets = FWER control

[3] Finner, H., & Roters, M. (2001). On the false discovery rate and expected type I errors. Biometrical Journal, 43(8), 985-1005.

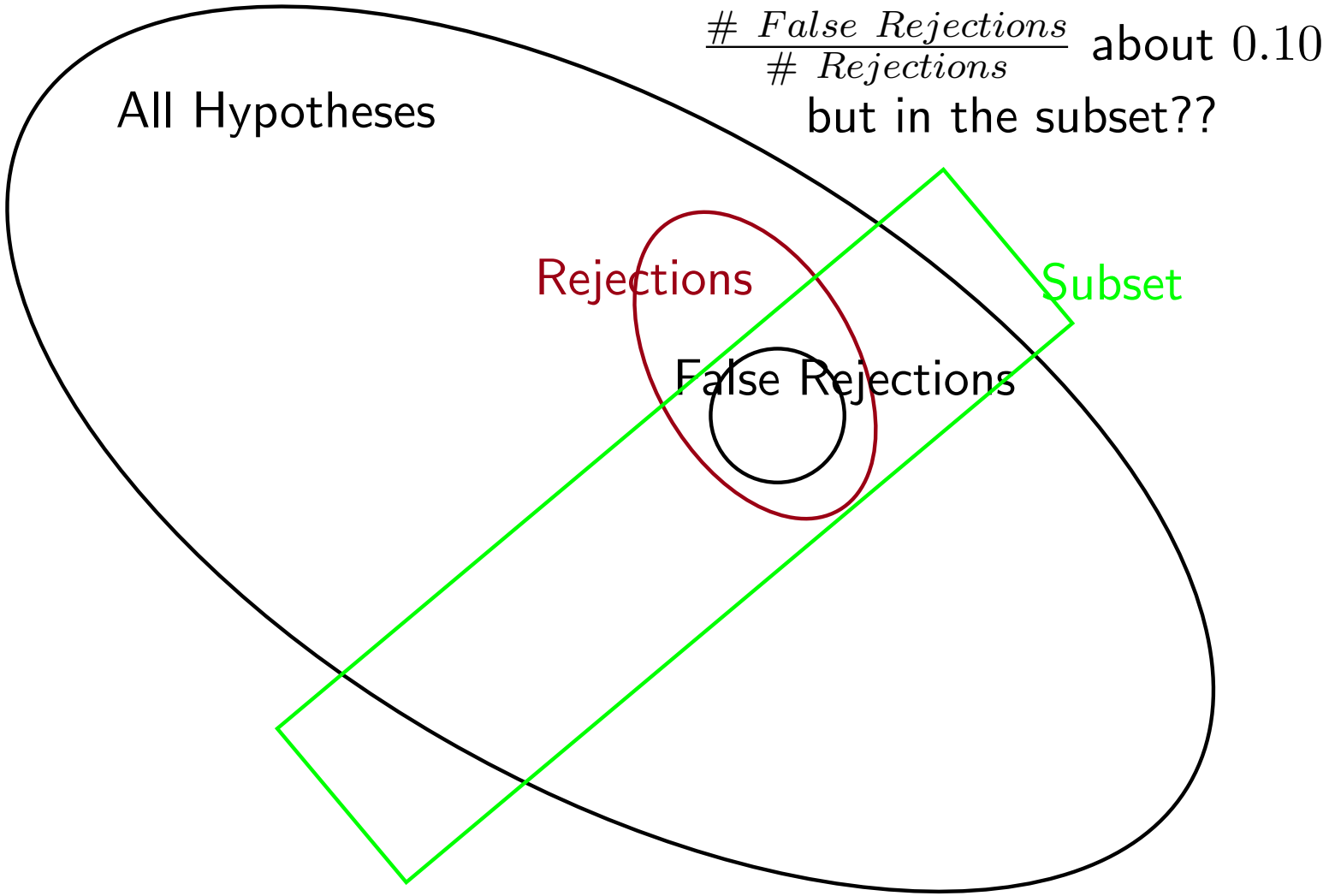
Subsets of Rejected hypotheses



Subsets of Rejected hypotheses



Subsets of Rejected hypotheses



Relationships between FWER and FDR

Dominance

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

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 preferable in Clinical Trials
  - more flexible
- FWER and FDR easy in `R`
- excellent tutorial: Goeman & Solari (2014) [4]

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