Multiple Hypothesis Testing in Genomics I

Basic Error Rates and Procedures

Jelle Goeman Aldo Solari

Radboud University Medical Center

University of Milano-Bicocca

International Society for Clinical Biostatistics
23 August 2015

Outline

1. Introduction
2. Error rates
3. Basic FWER methods
4. Basic FDR methods
5. Basic FDP estimation
6. Outlook

Outline

Aldo Solari

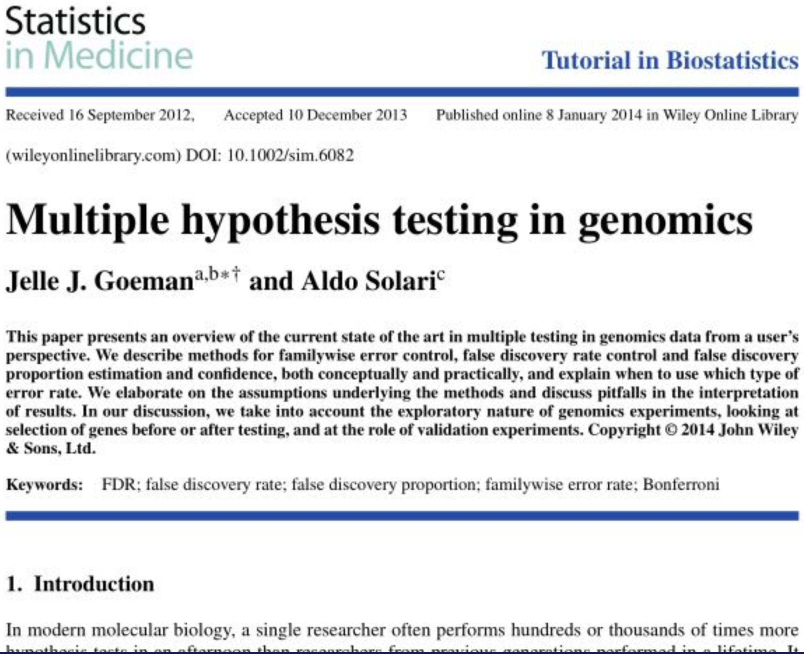
Associate Professor of Statistics at University of Milano-Bicocca

Jelle Goeman

Professor of Biostatistics at Radboud University Medical Center

A course in four parts

- Basic concepts and error rates (Jelle)
- Correlation and permutations (Aldo)
- Confidence for the False Discovery Proportion (Jelle)
- Structured problems (Aldo)



Statistics
in Medicine

Tutorial in Biostatistics

Received 16 September 2012, Accepted 10 December 2013 Published online 8 January 2014 in Wiley Online Library (wileyonlinelibrary.com) DOI: 10.1002/sim.6082

Multiple hypothesis testing in genomics

Jelle J. Goeman[a,b*†] and Aldo Solari[c]

This paper presents an overview of the current state of the art in multiple testing in genomics data from a user's perspective. We describe methods for familywise error control, false discovery rate control and false discovery proportion estimation and confidence, both conceptually and practically, and explain when to use which type of error rate. We elaborate on the assumptions underlying the methods and discuss pitfalls in the interpretation of results. In our discussion, we take into account the exploratory nature of genomics experiments, looking at selection of genes before or after testing, and at the role of validation experiments. Copyright © 2014 John Wiley & Sons, Ltd.

Keywords: FDR; false discovery rate; false discovery proportion; familywise error rate; Bonferroni

1. Introduction

In modern molecular biology, a single researcher often performs hundreds or thousands of times more hypothesis tests in an afternoon than researchers from previous generations performed in a lifetime. It

The programme today

Tutorial

Loosely followed. Read to support the course

Emphasis: what does it all mean?

- Concepts and understanding
- Which error rate to choose when
- Caveats

Also: practical execution of methods

Easy in R. Difficult in standard statistical software

Type I errors

Discovery

Rejection of a hypothesis = a scientific finding

Type I versus type II errors

A type II error is a failure to take a step forward.

A type I error is a step in the wrong direction

Central tenet of multiple testing

Focus on type I errors which are worse than type II errors

But

Not true in every context

Today's problem

Hypotheses

$H_1, \ldots, H_m$

True hypotheses

H_i is true if $i \in \mathcal{T} \subseteq \{1, \ldots, m\}$. $\mathcal{T}$ is fixed and unknown.

Rejected hypotheses

We reject all H_j for $j \in \mathcal{R} \subseteq \{1, \ldots, m\}$. Two flavors:

- $\mathcal{R}$ is a predetermined function of the data
- $\mathcal{R}$ is chosen freely after seeing the data

Goal

Have a large set $\mathcal{R}$ with small $\mathcal{T} \cap \mathcal{R}$. Type I errors: $\#(\mathcal{T} \cap \mathcal{R})$.

Hypotheses

What is a hypothesis?

A submodel $H \subset \mathcal{M}$ of an encompassing model $\mathcal{M}$.

- Given by a full model with constraint, e.g. $\mu = 0$ in $\mathcal{N}(\mu, \sigma^2)$
- Direct formulation: $\mathcal{N}(0, \sigma^2)$

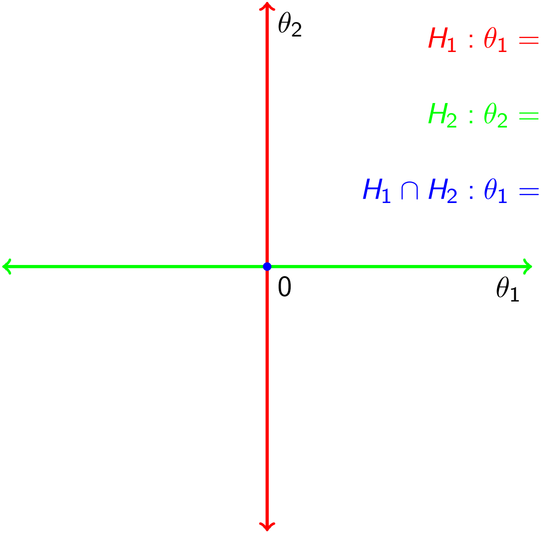
True distribution

We typically assume the true data generating distribution $t \in \mathcal{M}$

True hypotheses

H is true if and only if $t \in H$

Hypotheses as subspaces of the parameter space



$H_1 : \theta_1 = 0$

$H_2 : \theta_2 = 0$

$H_1 \cap H_2 : \theta_1 = \theta_2 = 0$

P-values

P-value based

Most basic methods discussed today start from *p*-values

Common definition

"Probability of observing a test statistic as extreme or more extreme than the observed test statistic"

Horrible definition

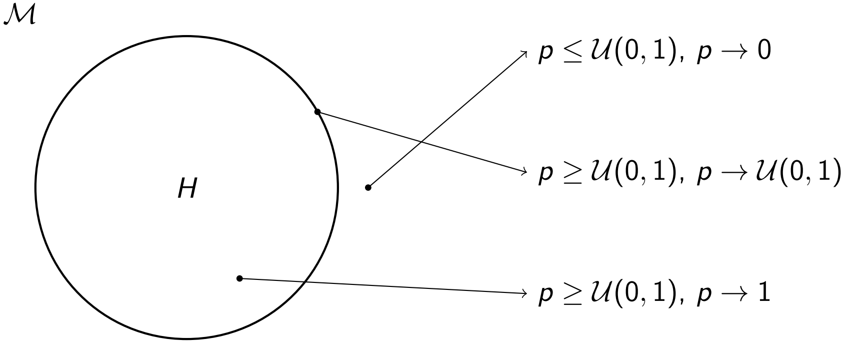
- Convoluted
- Suggests that a *p*-value is a probability
- Difficult to understand
- Does not capture the essence of *p*-values

A *p*-value is a standardized test statistic

A *p*-value is a random variable

A test statistic standardized to get a specific distribution

Distribution of the *p*-value


$\mathcal{M}$
H
$p \leq \mathcal{U}(0,1),\ p \to 0$
$p \geq \mathcal{U}(0,1),\ p \to \mathcal{U}(0,1)$
$p \geq \mathcal{U}(0,1),\ p \to 1$

Alternative (more fundamental) definition

α-level of a test

If we have a family of tests parameterized by type I error α

A p-value is

The maximal α-level at which the test rejects

Distributional properties

Follow from this definition

Generalizes to adjusted p-values

Maximal α-level at which the test procedure rejects

Joint distribution of *p*-values

Marginal distribution of *p*-values

Firmly under control

Joint distribution

May be anything. Typically *p*-values are correlated.

Unknown joint distribution

Greatest practical problem in multiple testing

Three basic approaches

No assumptions

Use general probability inequalities ('worst case')

Resulting methods conservative for most p-value distributions

Assume Simes' inequality

Generally but not universally valid probability inequality

Resulting methods conservative for some p-value distributions

Permutation-based methods

Only useable with some null hypotheses in some models

Resulting methods exact for all p-value distributions

Assumptions and error rates

Error rates, methods, assumptions

Assumptions	Error criterion		
	FWER control	FDR control	FDP confidence
None	Holm	Benjamini & Yekutieli	Goeman & Solari
Simes	Hommel	Benjamini & Hochberg	Goeman & Solari
Permutations	Westfall & Young	—	Meinshausen

Note: FDP estimation same for all assumptions

- Point estimates unaffected by correlation between *p*-values
- But accuracy of estimates highly affected!

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

with $R = \#\mathcal{R}$, $V = \#(\mathcal{R} \cap \mathcal{T})$, and $U = \#(\mathcal{R} \setminus \mathcal{T})$.

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

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

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

Assumptions and error rates

Error rates, methods, assumptions

Assumptions	Error criterion		
	FWER control	FDR control	FDP confidence
None	Holm	Benjamini & Yekutieli	Goeman & Solari
Simes	Hommel	Benjamini & Hochberg	Goeman & Solari
Permutations	Westfall & Young	—	Meinshausen

Note: FDP estimation same for all assumptions

- Point estimates unaffected by correlation between *p*-values
- But accuracy of estimates highly affected!

The family

How big is the multiple testing problem?

How many or which hypotheses to take together in one error rate?

Rules of thumb

- Focus on selection and selective emphasis
- More families = more errors
- Where is the theoretical controversy?

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*

Boole

Boole's inequality

For any events A and B:

$$\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B)$$

so

$$\mathrm{P}(A \cup B) \leq \mathrm{P}(A) + \mathrm{P}(B)$$

For any events $A_1, \ldots, A_m$:

$$\mathrm{P}(\bigcup_{i=1}^{m} A_i) \leq \sum_{i=1}^{m} \mathrm{P}(A_i)$$

Equality

Equality holds if events are disjoint

Bonferroni

Simple Bonferroni

Reject all hypotheses with p-value below α/m

Proof that Bonferroni works

$$\mathrm{P}\Big(\bigcup_{i=1}^{m_0}\{q_i \leq \alpha/m\}\Big) \leq \sum_{i=1}^{m_0}\mathrm{P}(q_i \leq \alpha/m) \leq m_0\frac{\alpha}{m} \leq \alpha$$

with $q_1, \ldots, q_{m_0}$ the p-values of true hypotheses.

Three inequalities

1. Uses Boole's inequality
2. Uses (super)uniformity of null p-values: $\mathrm{P}(q_i \leq t) \leq t$
3. Uses $m_0 \leq m$

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

“Effective number of tests”

Sometimes proposed

“Effective number of tests” if p-values are correlated

Example: genome-wide significance level

A p-value in GWAS is significant if $\leq 5 \times 10^{-8}$
“Effectively 10^6 independent tests in the genome”

Concept

Has no theoretical foundation: should depend on α and other factors

Instead

What is important is the distribution of $\min_i p_i$

Sequential rejection

Sequential rejection principle

A FWER procedure may always be designed as follows

- Reject a number of hypotheses controlling FWER
- Start over with the remaining hypotheses as if the rejected hypotheses never existed
- Even (Shaffer) may use the information that rejected hypotheses are certainly false
- Repeat until no new rejections occur

Why does this work?

Because FWER does not care about second errors

Holm

Holm = sequential Bonferroni

Repeatedly apply Bonferroni until no new rejections occur

Start with $c = \alpha/m$

Repeat

1. Reject all hypotheses with p-value $\leq c$
2. Recalculate $c = \alpha/(m - r)$
   with r number of so far rejected hypothesis

Improvement of Bonferroni

Uniformly more powerful, but usually only a little bit

Holm: example

Example: p-values

$\boxed{0.011}$ $\boxed{0.15}$ $\boxed{0.001}$ $\boxed{0.003}$ $\boxed{0.005}$ $\boxed{0.009}$ $\boxed{0.87}$ $\boxed{0.64}$ $\boxed{0.002}$

Critical value

1. 9 hypotheses, $\alpha = 0.05$, so $c = 0.05/9 = 0.0056$
2. 5 remaining hypotheses, $\alpha = 0.05$, so $c = 0.05/5 = 0.01$
3. 4 remaining hypotheses, $\alpha = 0.05$, so $c = 0.05/4 = 0.0125$
4. 3 remaining hypotheses, $\alpha = 0.05$, so $c = 0.05/3 = 0.0167$

Logically related hypotheses

Example

Anova model. Three subgroups.

Hypotheses: pairwise comparisons between subgroups.

$$\begin{aligned} H_{12} &: \mu_1 = \mu_2 \\ H_{23} &: \mu_2 = \mu_3 \\ H_{13} &: \mu_1 = \mu_3 \end{aligned}$$

Relationships

If H_{12} is false, H_{23} and H_{13} cannot be both true.

Restricted combinations

Not all combinations of truth/falsehood of hypotheses are viable

Shaffer's method

Variant of Holm's method with restricted combinations

Start with $c = \alpha/m$

Repeat

1. Reject all hypotheses with p-value $\leq c$
2. Recalculate $c = \alpha/s$
   with s the maximum number of hypotheses that can still be true given that all the rejections made so far are correct

Compare to Holm

Method is valid under the same general assumptions as Holm

Less conservative than Holm in case of restricted combinations

Shaffer: example

Hypotheses and data

$$\begin{aligned} H_{12} &: \mu_1 = \mu_2 & p_{12} = 0.01 \\ H_{23} &: \mu_2 = \mu_3 & p_{23} = 0.04 \\ H_{13} &: \mu_1 = \mu_3 & p_{13} = 0.53 \end{aligned}$$

Shaffer's procedure

1. Reject all hypotheses with p-value $\leq \alpha/3 \rightarrow$ reject H_{12}
2. If H_{12} is false, at most one of H_{23} and H_{13} can be simultaneously true
3. Reject all hypotheses with p-value $\leq \alpha/1 \rightarrow$ reject H_{23}
4. Continue:. . . No further rejections possible

Adjusted *p*-values

General

Multiplicity-adjusted p-value is the smallest FWER α at which the hypothesis would be rejected in a multiple testing procedure

Compare

The definition of the regular p-value

Example: Bonferroni

Adjusted p is $\min(mp_i, 1)$

Analogous in other FWER-procedures

Calculation can be more complicated

Adjusted *p*-values for Holm's procedure

Start with *p*-values for m hypotheses

1. Sort the *p*-values $p_{(1)}, \ldots, p_{(m)}$.
2. Multiply each $p_{(i)}$ by its adjustment factor $a_i = m - i + 1$, $i = 1, \ldots, m$
3. If the multiplication in step 2 violates the original ordering, repair this: increase the smallest *p*-value in all violating pairs:

$$\tilde{p}_{(i)} = \max_{j=1,\ldots,i} a_j p_{(j)}$$

4. Set $\tilde{p}_{(i)} = \min(\tilde{p}_{(i)}, 1)$ for all i.

Subsetting property of FWER

FWER guarantees

With 95% probability the rejected set $\mathcal{R}$ contains no type I errors

Individual hypotheses within $\mathcal{R}$

Are also 95% confidently no type I errors

Subsetting property

FWER control translates to FWER control in a subset

Why is this useful?

That single extraordinary finding is reliable

The adjusted p-value is meaningful for individual hypotheses

When to use FWER in genomics?

Using FDP or FDR-based methods

- Intermediate stage analyses
- When the individual findings are less important
- When type II errors are an issue
- When power is low

Using FWER-based methods

- Final-stage analyses
- When the individual findings are to be vouched for
- When type I errors matter most
- When power is good

Benjamini and Hochberg

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

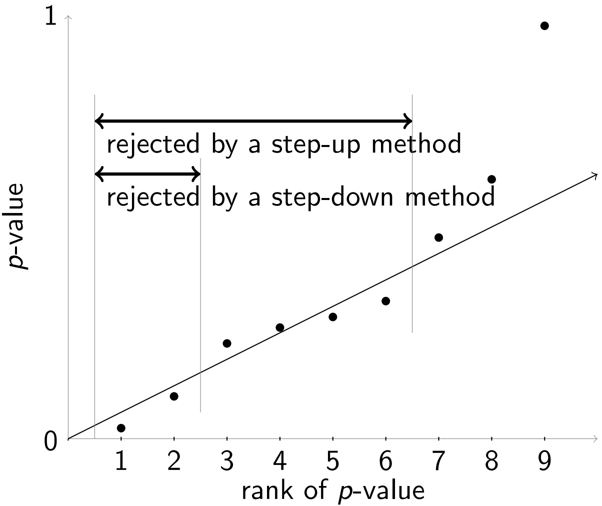
This procedure controls FDR under independence

Control is at $\pi_0\alpha$ (compare Bonferroni), with $\pi_0 = m_0/m$

Later

Conditions relaxed

Step-down and step-up



Assumptions

One-sided tests

As long as test statistics not negatively correlated

Two-sided tests

If test statistics are (asymptotically) normal

Exact limits of validity of BH procedure

Subject to much ongoing research

Related

Simes inequality (more later)

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

Subsetting property

Meaning of FDR control

If we generate $\mathcal{R}$ and randomly pick a hypothesis from it this is a type I error with probability $\leq \alpha$

Property

Of $\mathcal{R}$ (or procedure leading to $\mathcal{R}$)

Subsetting property

FDR control on $\mathcal{R}$ does not translate to subsets

In particular not to individual hypotheses

Exception

The subset with the lowest *p*-values has FDR control

Leniency scaling

FDR: small proportion of errors

Consequence: large sets treated differently from small sets

Small sets

Few errors allowed → FDR behaves like FWER

Large sets

Many errors allowed → large probability of errors present

Consequence

'Tails' of large sets $\mathcal{R}$ are likely type I errors

FDR-adjusted *p*-values

Adjusted *p*-value

Highest α-level at which procedure rejects a hypothesis

Lack of subsetting property

FDR is about the set $\mathcal{R}$, not about individual hypotheses

Meaningless

To report an individual hypotheses from $\mathcal{R}$ with its adjusted *p*-value

Meaning of an adjusted *p*-value

Same FDR-adjusted *p*-value indicates a higher chance of a type I error if part of a large set $\mathcal{R}$ than if part of a small set

Calculating FDR-adjusted p-values

Start with p-values for m hypotheses

1. Sort the p-values $p_{(1)}, \ldots, p_{(m)}$.
2. Multiply each $p_{(i)}$ by its adjustment factor $a_i = m/i$
3. If the multiplication in step 2 violates the original ordering, repair this: Decrease the highest p-value in all violating pairs:

$$\tilde{p}_{(i)} = \min_{j=i,\ldots,m} a_j p_{(j)}$$

4. Set $\tilde{p}_{(i)} = \min(\tilde{p}_{(i)}, 1)$ for all i.

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

Benjamini & Yekutieli

Assumptions of Benjamini and Hochberg

Non-negatively associated p-values

Benjamini and Yekutieli

Variant valid for any distribution of p-values

How does it work?

Reduce all critical values by a factor $\sum_{i=1}^{m} 1/i \approx \log(m)$

In practice

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

When to use FDR

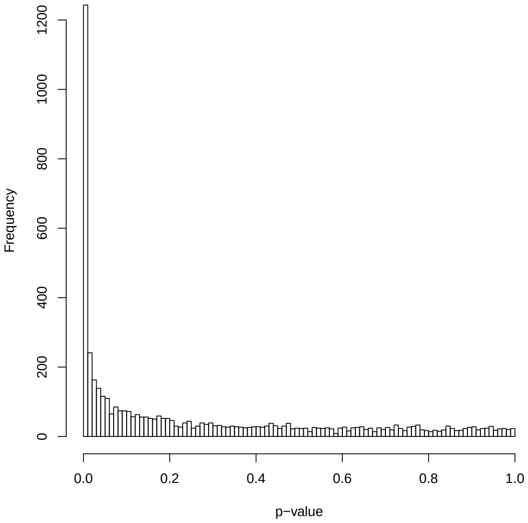
FDR is the norm

In most genomics literature (exception GWAS)

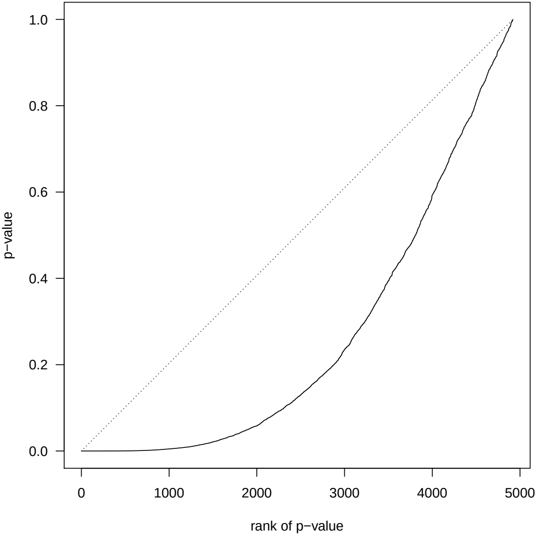
Use FDR especially

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

P-value histogram



Sorted *p*-value plot



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 $\mathrm{FDP} \approx m_0 t / \#\mathcal{R}$

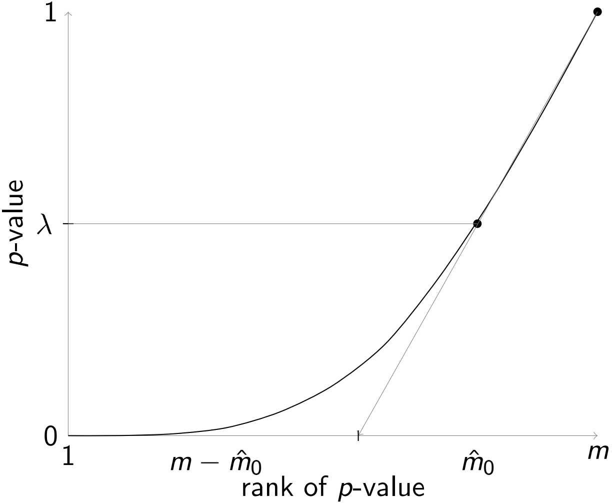
Estimate of m_0 (again by uniformity)

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

Resulting estimate of FDP ("q-value")

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

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 $\mathrm{F\hat{D}P} \leq \alpha$

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

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

Doing all this in R

Trivial calculations

Once you have the *p*-values

R

`p.adjust`

Other statistical software

Difficult. . .

Excel

Easy

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

Three ways of dealing with dependence

No assumptions

Boole's or Hommel's probability inequality

Assumptions underlying Simes' inequality

Allows Simes-based procedures (such as BH)

Can be assumed OK for two-sided asymptotically normal tests

Use permutations

If the null hypotheses and model allows

Assumptions and error rates

Error rates, methods, assumptions

Assumptions	Error criterion		
	FWER control	FDR control	FDP confidence
None	Holm	Benjamini & Yekutieli	Goeman & Solari
Simes	Hommel	Benjamini & Hochberg	Goeman & Solari
Permutations	Westfall & Young	—	Meinshausen

Note: FDP estimation same for all assumptions

- Point estimates unaffected by correlation between p-values
- But accuracy of estimates highly affected!