The multi-tasking of multiple testing

Juggling significance and false discoveries

Anna Vesely

Psicostat - 28.04.2023

University of Bremen
vesely@uni-bremen.de

Motivation

Multiple hypothesis testing

In many fields, interest lies in making inference on a (potentially high) number m of features:

- medical data (1) - effect of different drugs on a symptom
- medical data (2) - effect of a drug on different symptoms
- genomics - (differential) expression of genes
- neuroimaging - brain activation in voxels
- ...

The goal is detecting signal while keeping the errors under control

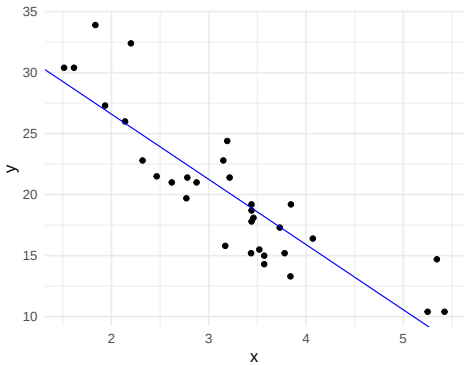
Multiple linear regression

$$y_j = \beta_0 + \sum_{i=1}^{m} \beta_i x_{ij}$$

We investigate which covariates have an effect on the outcome

Covariate i:

- null hypothesis $H_i : \beta_i = 0$
- p-value p_i



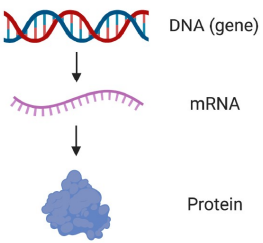
Differential gene expression

Gene expression is generally measured quantifying levels of the gene product (often a protein)

We look for differences between populations in the expression of $\approx$ 20,000 genes

Gene i:

- null hypothesis H_i : *no difference in gene i*
- p-value p_i from first-level analysis



Individual hypothesis testing

Test on a single feature

Consider a single null hypothesis H_0, e.g.,

$$H_0 : \textit{a drug is not effective}$$

The main goal is keeping under control the probability of

$$\text{type I error} \longleftrightarrow \text{false discovery} \longleftrightarrow \text{falsely reject } H_0 \text{ when it is true}$$

Standard methods allow to bound this probability of error by an 'acceptable' risk α (e.g., $\alpha = 0.05$)

Error control

		null hypothesis	
		false	true
		(drug is effective)	*(drug is not effective)*
test	rejected	true discovery	type I error
	not rejected	type II error	true negative

$$\mathbb{P}(\text{type I error}) = \mathbb{P}(\text{reject} \mid H_0) \leq \alpha$$

Simulations

- Simulate observations for a non-active feature: $X \sim \mathcal{N}(\mu, 1)$ with $\mu = 0$
- Test activation: $H_0 : \mu = 0$ (two-sided alternative)
- Using a one-sample t-test, obtain a p-value p

$$H_0 \text{ is true} \quad \longrightarrow \quad p \sim \text{Unif}[0, 1]$$

Over many simulations, the proportion of rejections is $\approx \alpha$

Multiple hypothesis testing

Tests on multiple features

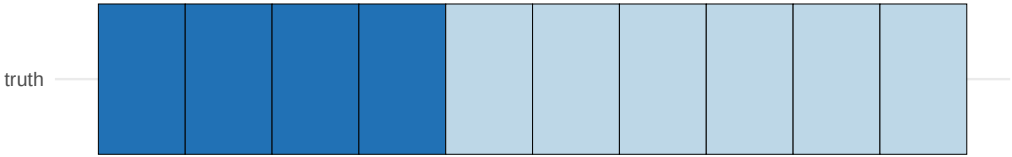
The goal is testing m hypotheses $H_1, \ldots, H_m$ simultaneously from the same data

This is a non-trivial extension of the individual case!

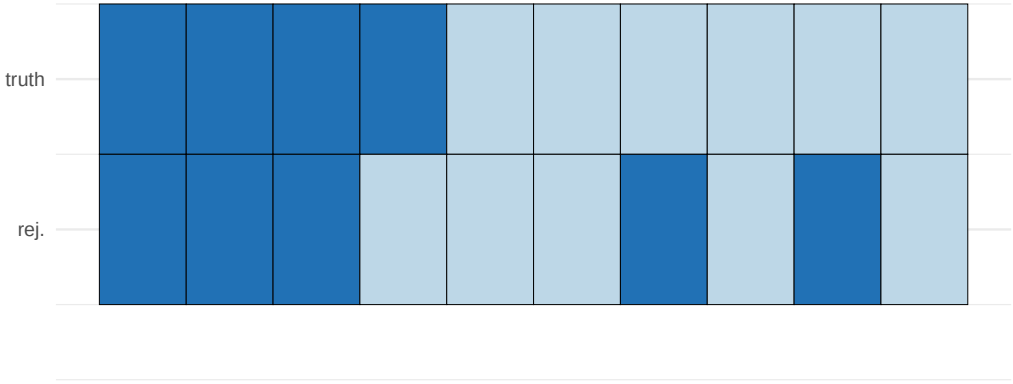
Each test carries the risk of making a type I error
$\longrightarrow$ the risk of having at least one may become unmanageable

How do we generalize the concept of type I error and control it?

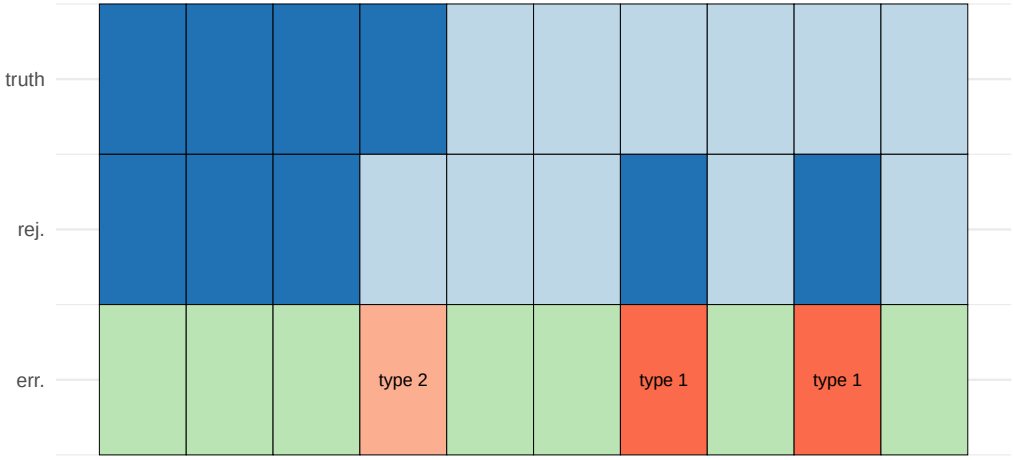
Example: ground truth



Example: tests



Example: errors



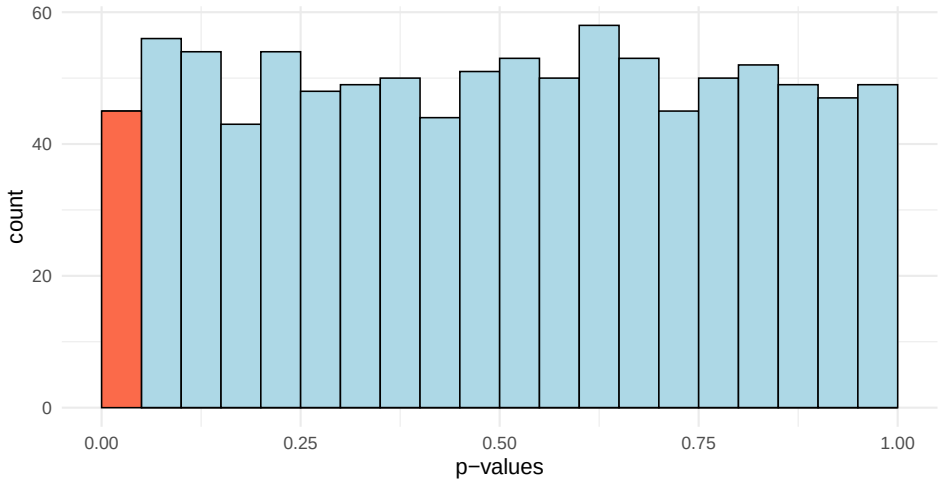
Simulations

- Repeat the previous simulations for m independent features: $X_i \sim \mathcal{N}(\mu_i, 1)$ with $\mu_i = 0$
- Test activation for each: $H_i : \mu_i = 0$ (two-sided alternative)
- Obtain m p-values

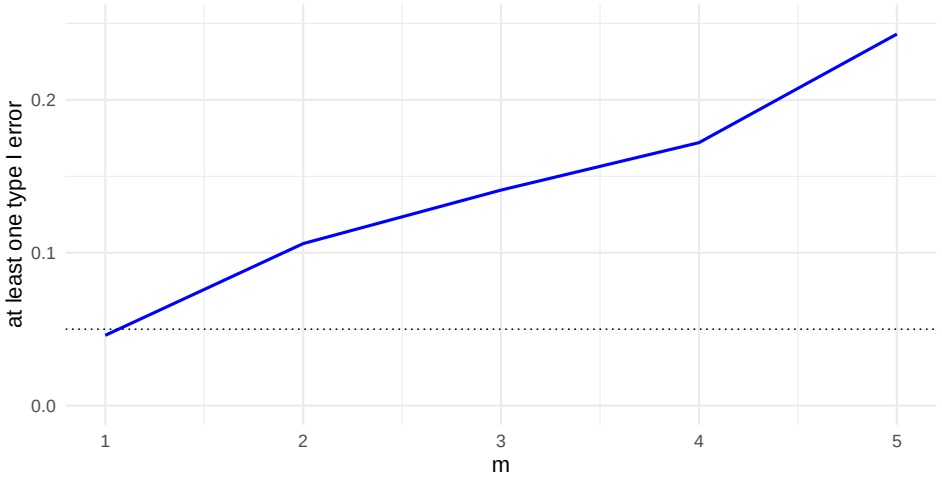
$$\text{All hypotheses are true} \quad \longrightarrow \quad \text{each } p_i \sim \text{Unif}[0, 1]$$

Without further adjustments, some of these p-values will be $\leq \alpha$!

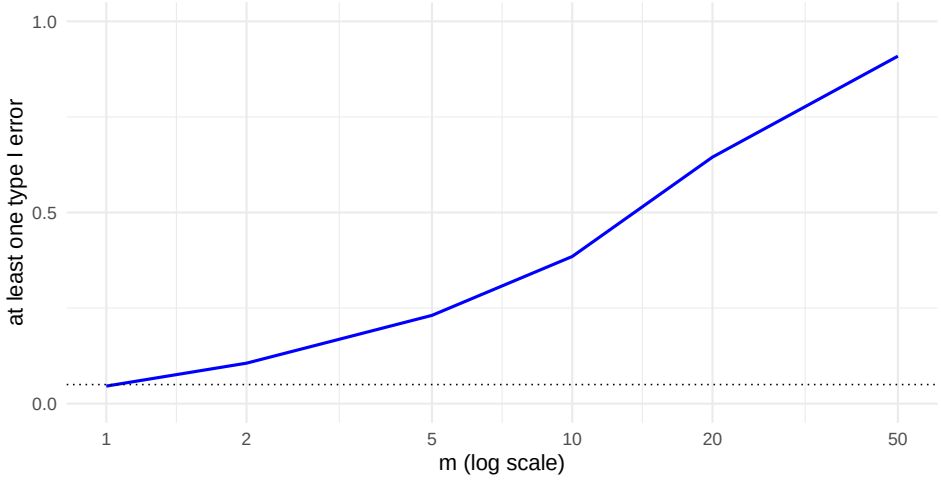
$m = 1000$ tests, 1 simulation



m tests, 1000 simulations



m tests, 1000 simulations



Error control

		null hypothesis		
		false	true	tot.
test	rejected	S	V	R
	not rejected	T	U	$m - R$
	tot.	m_1	m_0	m

We work on the false discoveries (rejections of true null hypotheses):

- number V
- proportion V/R

FWER control

Familywise error rate

		null hypothesis		
		false	true	tot.
test	rejected	S	V	R
	not rejected	T	U	$m - R$
	tot.	m_1	m_0	m

$$\text{FWER} = \mathbb{P}(\text{at least one type I error}) = \mathbb{P}(V > 0)$$

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

Bonferroni correction

Instead of rejecting all $p_i \leq \alpha$:

- obtain adjusted p-values $\tilde{p}_i = p_i \cdot m$
- reject all $\tilde{p}_i \leq \alpha$

The method:

- controls the FWER under any dependence structure of the data
- may be very conservative and lead to many false negatives

FWER controlling methods

- Bonferroni - always valid
- Holm-Bonferroni - improves Bonferroni and remains always valid
- Hochberg - valid under independence or positive dependence
- Hommel - as Hochberg, slightly more powerful but slower
- ...

The main methods are implemented in the `R` function `p.adjust`

Example: linear regression

```
data(mtcars)
fit = lm(mpg ~ disp + drat + wt, data = mtcars)
p = summary(fit)$coefficients[, 4][-1]
p_adj = p.adjust(p, method = "holm")
```

control	p-value	disp	drat	wt
no	raw	0.098	0.567	0.014
FWER	adjusted (Holm)	0.196	0.567	0.043

FWER controlling methods

FWER control may be very stringent, especially when m is large
$\longrightarrow$ it can lead to many false negatives, potentially missing important discoveries

This is not the only generalization of the type I error!

- If the goal is to minimize the risk of false discoveries $\longrightarrow$ stick to FWER
- If we may allow some false discoveries to occur, as long as the overall proportion is controlled $\longrightarrow$ ...

FDR control

False discovery rate

		null hypothesis		
		false	true	tot.
test	rejected	S	V	R
	not rejected	T	U	$m - R$
	tot.	m_1	m_0	m

$$\text{FDP} = \frac{\text{false rejections}}{\text{rejections}} = \frac{V}{R}, \qquad \text{FDR} = \mathbb{E}(\text{FDP})$$

A procedure controls it if $\text{FDR} \leq \alpha$

FDR controlling methods

- Benjamini-Hochberg - valid under independence, positive dependenceand many other settings (not always!)
- Benjamini-Yekutieli - always valid, may be more conservative
- ...

These methods are implemented in the same function `p.adjust`

Example: linear regression

```
data(mtcars)
fit = lm(mpg ~ disp + drat + wt, data = mtcars)
p = summary(fit)$coefficients[, 4][-1]
p_adj = p.adjust(p, method = "BH")
```

control	p-value	disp	drat	wt
no	raw	0.098	0.567	0.014
FWER	adjusted (Holm)	0.196	0.567	0.043
FDR	adjusted (BH)	0.147	0.567	0.043

Other methods

Other types of error control

- k-FWER - generalized FWER
- FDX - false discovery exceedance
- JER - joint error rate
- FDP - false discovery proportion
- ...

False discovery proportion

$$\text{FDP} = \frac{\text{false rejections}}{\text{rejections}} = \frac{V}{R}$$

A procedure controls it if it gives an upper $(1-\alpha)$-confidence bound B for it:

$$\mathbb{P}(\text{FDP} \leq B) \geq 1 - \alpha$$

It is desirable to control the FDP of all possible subsets simultaneously

Error control

Familywise error rate

$$\text{FWER} = \mathbb{P}(\text{at least one false discovery}) \quad \longrightarrow \quad \text{FWER} \leq \alpha$$

False discovery proportion

$$\text{FDP} = \frac{\text{false rejections}}{\text{rejections}} \quad \longrightarrow \quad \text{upper confidence bound}$$

False discovery rate

$$\text{FDR} = \mathbb{E}(\text{FDP}) \quad \longrightarrow \quad \text{FDR} \leq \alpha$$

Error control

- FWER - minimizes the risk of false discoveries
- FDR - allows some false discoveries, controls the overall proportion
- ...

Always state clearly which error is taken into account!

An overview

Goeman and Solari (2014). Multiple hypothesis testing in genomics. *Statistics in Medicine* 33