ADVANCED BIOSTATISTICS
ABSTAT18
Multiple Testing Issues

Lisete Sousa

Department of Statistics and Operations Research | CEAUL
Faculty of Sciences of Lisbon University

Ciências ULisboa

CEAUL
Centro de Estatística e Aplicações
Universidade de Lisboa

IGC, April 3rd - 6th, 2018

Contents

- V: r.v. which represents the number of false positives among m (multiple hypotheses),
  $P(V \geq 1) = 1 - P(V = 0) = 1 - (1 - \alpha)^m$,
  where α is the probability of rejecting the null hypothesis when it is true: (*Type I Error*), $P(\text{Rej } H_0 | H_0 \text{ True})$.

Number of hypothesis tested (m)	False positives incidence ($m \times \alpha$)	Probability of 1 or more false positives by chance ($1 - (1 - 0.05)^m$)
1	$1/20 = 0.05$	0.050
2	$2 \times (1/20) = 0.1$	0.098
20	$20 \times (1/20) = 1$	0.642
100	$100 \times (1/20) = 5$	0.994

Problem: When many hypotheses are tested, the probability of a type I error (false positive) increases sharply with the number of hypotheses.

Types of error control

- Let $H_{01}, H_{02}, ..., H_{0m}$ denote the null hypotheses corresponding to the m tests.
- Suppose m_0 null hypotheses are true and m_1 null hypotheses are false.
- Let c denote a value between 0 and 1 that will serve as a cutoff for significance:
  - Reject H_{0i} if $p_i \leq c$ (declare significant difference in the expression levels)
  - Do not reject H_{0i} if $p_i > c$ (declare non-significant difference in the expression levels)

	Rejected H_0	Not Rejected H_0	
True H_0	V	U	m_0
False H_0	S	T	m_1
	R	$m - R$	m

V: false positives (type I error)

T: false negatives (type II error)

- Suppose one test of interest has been conducted for each of m genes in a microarray experiment.
- Let $p_1, p_2, \ldots, p_m$ denote the p-values corresponding to the m tests.

- **PCER:** Per-comparison error rate, the expected value of the number of Type I errors over the number of hypotheses, PCER$=\frac{\mathrm{E}(V)}{m}$.
- **PFER:** Per-family error rate, the expected number of Type I errors, PFER$=\mathrm{E}(V)$.
- **FWER:** Family-wise error rate: the probability of at least one type I error, FWER$=\mathrm{P}(V \geq 1)$.
- **FDR:** False discovery rate, is the expected proportion of incorrectly rejected null hypotheses , FDR$=\frac{\mathrm{E}(V)}{R}$ for $R > 0$ (number of rejected null hypotheses).

FWER – Family-Wise Error Rate

- Many procedures have been developed to control the Family-Wise Error Rate (the probability of at least one type I error): $P(V \geq 1)$
- Two general types of FWER corrections:
  1. Single Step: equivalent adjustments made to each p-value.
  2. Sequential: adaptive adjustment made to each p-value.

Single-step approach: Bonferroni

- The Bonferroni's Method is the simplest way to achieve control of the FWER at any desired level α.
- Simply choose $c = \alpha/m$.
- With this value of c, the FWER will be no larger than α for any family of m tests.

Sequential Adjustments: Holm's method

- Let $p_{(1)}, p_{(2)}, \ldots, p_{(m)}$ denote the m p-values ordered from smallest to largest.
- Find the **largest** integer k so that:

$$p_{(i)} \leq \frac{\alpha}{m-i+1} \text{ for all } i = 1, \ldots, k.$$

Some considerations

- FWER is appropriate when you want to guard against ANY false positives.
- FWER criteria may be too restrictive because control of false positives implies a considerable increase of false negatives.
- FWER is to conservative because it depends on the overall number of tests (m).
- Holm's method is less conservative than the Bonferroni's Method.

FDR – False Discovery Rate

In practice, however, many biologists seem willing to accept that some errors will occur, as long as this allows findings to be made. For example a researcher might consider acceptable a small proportion of errors (say 10%, 20%) between her findings. In this case, the researcher is expressing interest in controlling the **false discovery rate** (FDR).

	Rejected H_0	Not Rejected H_0	
True H_0	V	U	m_0
False H_0	S	T	m_1
	R	$m - R$	m

V: false positives (type I error)
T: false negatives (type II error)

- FDR is designed to control the proportion of false positives among the set of rejected hypothesis (R).
- $\text{FDR} = \frac{\text{E}(V)}{R}$

- FDR is the proportion of false positives among all the genes initially identified as being differentially expressed.
- If one obtains a list of differentially expressed genes where the FDR is controlled at, say, 20%, one will expect that a 20% of these genes will represent false positive results.

The Benjamini and Hochberg Procedure for Strongly Controlling FDR at Level α

- Let $p_{(1)}, p_{(2)}, \ldots, p_{(m)}$ denote the m p-values ordered from smallest to largest.
- Find the **largest integer** k so that $p_{(k)} \leq \frac{k \times \alpha}{m}$.

This correction is the least stringent of all previous options, and therefore tolerates more false positives.

There will be also less false negative genes.

Adjust p-values in R – p.adjust(p,method)

Possible methods:

```
"holm","hochberg","hommel","bonferroni","BH","BY","fdr","none"
```

Suppose we are testing 10 hypotheses resulting in the following p-values:

```
p<-c(0.001,0.002,0.006,0.013,0.024,0.168,0.231,0.254,0.319,0.56)
```

Apply p.adjust function according to Benforroni, Holm's and BH methods:

```
p.adjust(p,"bonferroni");p.adjust(p,"holm");p.adjust(p,"BH")
[1] 0.01 0.02 0.06 0.13 0.24 1.00 1.00 1.00 1.00 1.00
[1] 0.01 0.018 0.048 0.091 0.144 0.840 0.924 0.924 0.924 0.924
[1] 0.01 0.01 0.02 0.032 0.0480 0.280 0.317 0.317 0.354 0.560
```

Some final considerations

FWER vs FDR

- The decision of controlling FDR or FWER depends on the goals of the experiment.
- If the objective is *gene fishing*, allowing a certain number of false positives to be reasonable, then FDR is preferable.
- If instead one is working with a shorter number of hypotheses, in which we want to verify if some specific ones are significant, then FWER is the appropriate criteria.
- FDRs are more appropriate in large sets of hypotheses.

Remarks

- Which multiple tests correction should be used? As long as the conditions you have for the data meet with the assumptions in particular multiple tests corrections, use the one that gives the highest power. **Using an FDR method is common these days.**
- 5% (or 95% confidence) is a convention, not a magic number (same to 1% or 0.1%). If you do not have any particular reason to favour a particular threshold, use a convention.

Acknowledgements:

Antónia Turkman (DEIO – FCUL) and Carina Silva-Fortes (ESTeSL – IPL), for allowing the use of some material produced by us in previous courses.

Bibliography:

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, Vol. 57, 289–300.

Storey, J.D. and Tibshirani, R. (2001). Estimating false discovery rates under dependence, with applications to DNA microarrays. Preprint

Westfall, P.H. and Young, S.S. (1993). *Resampling-based multiple testing: examples and methods for p-value adjustment.* Wiley.