ADVANCED BIOSTATISTICS ABSTAT17

Multiple Testing Issues

Lisete Sousa

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

IGC, April 10–13, 2017

Contents

MULTIPLE TESTING – P-VALUES CORRECTION

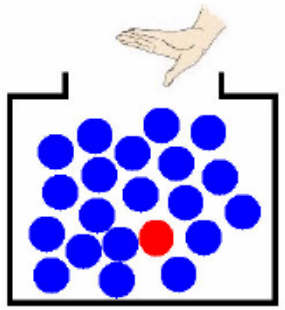
- Why should we do it?
- Committing an error (false positives).
- Two types of error control
- FWER
- Bonferroni's Method
- Holm's Method
- FDR
- Benjamini and Hochberg's Method

Short example to
1. understand the methods
2. learn how to use function `p.adjust()`

MULTIPLE TESTING – GENE EXPRESSION DATA

- Which type of error control?
- Example: R package (`RankProd`) for gene expression data
- Read the output

The multiple comparison problem



- Imagine a solution with 20 spheres: 19 are blue and 1 is red. What are de odds of randomly sampling the red sphere by chance? It is 1 out of 20.
- Now let's say that you get to sample a single sphere (and put it back into the solution) 20 times. Have a much higher chance to sample the red sphere. This is exactly what happens when testing several thousand tests at the same time.

- V: r.v. which represents the number of false positive genes, $P(V \geq 1) = 1 - P(V = 0) = 1 - (1 - \alpha)^n$, 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 (n)	False positives incidence ($n \times \alpha$)	Probability of 1 or more false positives by chance ($1 - (1 - 0.05)^n$)
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.

The multiple comparison problem

in gene expression data

- There is a serious consequence of performing statistical tests on many genes in parallel, which is known as multiple testing problem.
- How do we know that the genes that appear to be differentially expressed are truly differentially expressed and are not just an artefact introduced because we are analyzing a large number of genes?
- Is this gene truly differentially expressed, or could it be a false positive result?
- In such studies, the probability of at least one false positive result is near certain.

- We will often be interested not just in the probability of one error, but in the expected total number of errors. The expected number of false positives is simply α multiplied by the number of tests

- **E-value (expected value)**: For $m = 10\ 000$ independent t-tests with $\alpha = 0.1$, the expected number of false positives is $10\ 000 \times 0.10 = 1000$ false positives.

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)

- **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.

Example 1:

- Suppose we conduct 5 tests and obtain the following p-values for tests 1 through 5.

Test	1	2	3	4	5
p-value	0.042	0.001	0.031	0.014	0.007

- Which tests' null hypotheses will you reject if you wish to control the FWER at level 0.05?
- Use the Bonferroni's Method to answer this question.

Solution

Test	1	2	3	4	5
p-value	0.042	0.001	0.031	0.014	0.007

$p_1 = 0.042 > 0.01$
$p_2 = 0.001 \leq 0.01$
$p_3 = 0.031 > 0.01$
$p_4 = 0.014 > 0.01$
$p_5 = 0.007 \leq 0.01$

The cutoff for significance is $c = 0.05/5 = 0.01$ using the Bonferroni's Method. Thus we would reject the null hypothesis for tests 2 and 5 with the Bonferroni's Method.

How to do this in R?

There are several packages for multiple testing correction, as for example:

- `p.adjust` (the simplest – available at `stats` package)
- `multtest` (the most popular for gene expression data)
- `qvalue`
- `fdrtool`
- `structSSI` (for hypotheses with hierarchical or group structure)

In example 1, we will apply `p.adjust`.

Read the data:

```
> raw.p.values<-c(0.042,0.001,0.031,0.014,0.007)
> raw.p.values
```

```
[1] 0.042 0.001 0.031 0.014 0.007
```

Adjust the p-values:

```
> corr.p.values.Bonf<-p.adjust(p=raw.p.values,"bonferroni")
> corr.p.values.Bonf
```

```
[1] 0.210 0.005 0.155 0.070 0.035
```

Visualize tests corresponding to the rejected null hypotheses (position of the corrected p-value in the vector):

```
> global.alpha<-0.05
> which(corr.p.values.Bonf<global.alpha)
```

```
[1] 2 5
```

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:

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

- If no such k exists, set $c = 0$ (declare nothing significant).
- Otherwise set $c = p_{(k)}$ (reject the null hypotheses corresponding to the smallest k p-values).
- The point here is that we do not multiply every p_i by the same factor m.

Example 1 (cont.):

Test	1	2	3	4	5
p-value	0.042	0.001	0.031	0.014	0.007

$p_{(1)} = 0.001 \leq 0.05/(5-1+1) = 0.01$
$p_{(2)} = 0.007 \leq 0.05/(5-2+1) = 0.0125$
$p_{(3)} = 0.014 \leq 0.05/(5-3+1) = 0.0167$
$p_{(4)} = 0.031 > 0.05/(5-4+1) = 0.025$
$p_{(5)} = 0.042 \leq 0.05/(5-5+1) = 0.05$

These calculations indicate that **Holm's method** would reject null hypotheses for tests 2, 4 and 5. Here, $k = 3$ and $c = p_{(3)} = 0.014$.

How to do this in R?

Adjust the p-values:

```
> corr.p.values.Holm<-p.adjust(p=raw.p.values,"holm")
> corr.p.values.Holm
```

```
[1] 0.062 0.005 0.062 0.042 0.028
```

Visualize tests corresponding to the rejected null hypotheses (position of the corrected p-value in the vector):

```
> which(corr.p.values.Holm<global.alpha)
```

```
[1] 2 4 5
```

Some considerations

- FWER is appropriate when you want to guard against ANY false positives.
- However, in many cases (particularly in genomics) we can live with a certain number of 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.
- The methods will provide the same results for many data sets, but sometimes Holm's method will result in more rejected null hypotheses.

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).

- FDR which is the proportion of false positives among all the genes initially identified as being differentially expressed.
- Unlike a significance level which is determined before looking at the data, FDR is a post data measure of confidence.
- FDR uses information available in the data to estimate the proportion of false positive results that have occurred.
- 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.

	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).
- $\text{FDR} = \frac{\text{E}(V)}{R}$

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 $p_{(k)} \leq \frac{k \times \alpha}{m}$.
- If no such k exists, set $c = 0$ (declare nothing significant).
- Otherwise set $c = p_{(k)}$ (reject the null hypotheses corresponding to the smallest k p-values).

Example 1 (cont.):

Test	1	2	3	4	5
p-value	0.042	0.001	0.031	0.014	0.007

$$p_{(1)} = 0.001 \leq 1 \times 0.05/5 = 0.01$$
$$p_{(2)} = 0.007 \leq 2 \times 0.05/5 = 0.02$$
$$p_{(3)} = 0.014 \leq 3 \times 0.05/5 = 0.03$$
$$p_{(4)} = 0.031 \leq 4 \times 0.05/5 = 0.04$$
$$p_{(5)} = 0.042 \leq 5 \times 0.05/5 = 0.05$$

The **Benjamini and Hochberg's Method** would reject the null hypotheses for all 5 tests. Here, $k = 5$ and $c = p_{(5)} = 0.042$.

How to do this in R?

Adjust the p-values:

```
> corr.p.values.BH<-p.adjust(p=raw.p.values,"BH")
> corr.p.values.BH
```

```
[1] 0.04200000 0.00500000 0.03875000 0.02333333 0.01750000
```

Visualize tests corresponding to the rejected null hypotheses (position of the corrected p-value in the vector):

```
> which(corr.p.values.BH<global.alpha)
```

```
[1] 1 2 3 4 5
```

- This correction is the least stringent of all previous options, and therefore tolerates more false positives.
- There will be also less false negative genes.
- The correction becomes more stringent as the p-value decreases, similarly as the Bonferroni Step-down correction.

Using packages which include multiple testing correction

in gene expression data

As we have seen previously in this lecture, multiple comparison problem arises when performing statistical tests on many genes simultaneously.

The aim of this section, is to show how to read and interpret the results when applying statistical packages from Bioconductor to identify differentially expressed genes.

We chose the package `RankProd`, which is one of the most used packages in this area. The data considered is the swirl zebrafish data, which was already used previously in order to exemplify the EM algorithm.

Short introduction to `RankProd` package

- Breitling et al. (2004) present a technique for identifying differentially expressed genes that originates from an analysis of biological reasoning.
- The technique is based on calculating rank products (RP) from replicate experiments.
- At the same time, it provides a statistical way to determine the significance level for each gene and allows for the flexible control of the false-detection rate (FDR).

Theoretical aspects

The assumptions made for RP method are relatively weak. It is assumed that

(1) relevant expression changes affect only a minority of genes,

(2) measurements are independent between replicate arrays,

(3) most changes are independent of each other,

(4) measurement variance is about equal for all genes.

- *RP* is a non-parametric statistic used to detect genes that are consistently highly ranked (strongly up-regulated/ down-regulated) in a number of replicate experiments.
- It is assumed that, under the null hypothesis that the order of all genes is random, the probability of finding a specific gene among the top r of n genes in a replicate is

$$p = r/n.$$

- Multiplying these probabilities allows the calculation of the corresponding combined probability as a rank product

$$RP = \prod_i r_i / n_i,$$

where r_i is the position of a specific gene in the i-th replicate and n_i is the total number of genes in the i-th replicate sorted by increasing/decreasing (up-regulated/down-regulated) values.

- The smaller the *RP* value, the smaller the probability that the observed position of the gene at the top of the lists is due to chance.
- A simple permutation-based estimation procedure provides a very convenient way to determine how likely it is to observe a given *RP* value, or better, in a random experiment.
- If there is high variability in gene-specific variances, RP tends to give overly optimistic p-values. The average ranks may constitute an alternative (Breitling and Herzyk, 2006).

Advantages and disadvantages

Rank Products method has several advantages:

- Fast and simple;
- Results are reliable in highly noisy data;
- Result in an increased power and accuracy at small number of replicates;
- Able to combine data sets from different laboratories into one analysis to increase the power of the identification;
- Analyzes both Affymetrix and cDNA microarrays (designed with a common reference or a direct two-color design).

The main disadvantage of RP method is that there is a significant loss of performance, when the equal-variance assumption is seriously violated and the number of replicates is higher than three.

Application to swirl zebrafish dataset

Log ratios (wt/swirl) matrix for each gene and slide are set into `data`.

For replicates 2 and 4 the ratio corresponds to (swirl/wt), therefore you need to make a correction previously:

```
> MA$M[,2] <- (-1)*MA$M[,2]
> MA$M[,4] <- (-1)*MA$M[,4]
> data<-MA$M
```

The number of samples (k) is printed, in order to fix the dimension of vector `c1`.

This vector, c1, contains only 1's as each sample (column) corresponds to expression ratios of two channels (paired samples).

```
> k <- dim(data)[2]
> k
[1] 4

> c1 <- c(rep(1,k))
> c1
[1] 1 1 1 1
```

The analysis is done using,

```
> RP.out <- RP(data,c1,num.perm=1000,logged=TRUE)
```

In the output, the channel used as the numerator is called as class 1 (wt) and the channel used as denominator as class 2 (swirl).

The list of selected up- and down-regulated genes is based on the estimated percentage of false positive predictions (pfp), which is also known as false discovery rate (FDR).

Considering a cutoff of 0.001 for `pfp`, 56 genes are selected to be DE:

```
> length(RP.out$pfp[RP.out$pfp<0.001])
[1] 56
```

This means that among the selected top 56 regulated genes, we expect to have 0.56 false positives!

Note that, since this method is based on permutations, different runs may generate slightly different results.

Top 56 genes selected to be differentially expressed:

```
> table <- topGene(RP.out,cutoff=0.001,logged=TRUE,
+ logbase=2,method="pfp")

Table1: Genes called significant under class1 < class2
Table2: Genes called significant under class1 > class2

> table

$Table1
      gene.index RP/Rsum FC:(class1/class2)   pfp P.value
 [1,]       7036  9.3246              0.3931 0e+00       0
 [2,]       7491 11.7725              0.3990 0e+00       0
 [3,]       4546 12.0935              0.4257 0e+00       0
 [4,]        683 12.1175              0.4040 0e+00       0
 [5,]       5075 14.1421              0.4130 0e+00       0
 [6,]       4032 14.1640              0.4138 0e+00       0
 [7,]       8295 14.2456              0.4143 0e+00       0
 [8,]        515 15.6508              0.4158 0e+00       0
 [9,]       4380 19.3273              0.4274 0e+00       0
```

[10,]	2276	19.9724	0.4303	0e+00	0
[11,]	7307	23.1055	0.4367	0e+00	0
[12,]	7602	26.3533	0.4463	1e-04	0
[13,]	6457	26.4745	0.4778	1e-04	0
[14,]	4647	26.7552	0.4600	1e-04	0
[15,]	3790	26.8945	0.4487	1e-04	0
[16,]	4623	28.3417	0.4618	1e-04	0
[17,]	7542	31.7084	0.4611	1e-04	0
[18,]	1697	34.3205	0.4711	1e-04	0
[19,]	6449	34.3649	0.4767	1e-04	0
[20,]	2945	36.5476	0.4829	1e-04	0
[21,]	1146	36.7377	0.4814	1e-04	0
[22,]	315	36.7559	0.4781	1e-04	0
[23,]	2715	37.2771	0.4760	1e-04	0
[24,]	59	37.6705	0.4545	1e-04	0
[25,]	3674	41.1883	0.3959	2e-04	0
[26,]	3695	45.7635	0.4972	2e-04	0
[27,]	988	48.1066	0.5006	3e-04	0
[28,]	6023	49.1878	0.5073	2e-04	0
[29,]	736	57.4403	0.5205	5e-04	0
[30,]	4865	59.5970	0.5267	6e-04	0

```
[31,]        4988 62.8162                0.5227 6e-04        0
[32,]         229 67.5730                0.5445 7e-04        0
[33,]        1643 68.9837                0.5444 7e-04        0
[34,]        6117 70.1293                0.5369 7e-04        0
[35,]         125 72.6411                0.5489 8e-04        0
[36,]        4996 72.8878                0.5494 8e-04        0

$Table2
      gene.index RP/Rsum FC:(class1/class2)   pfp P.value
 [1,]       2961  4.3004                6.4590 0e+00       0
 [2,]       1609  7.9922                4.9111 0e+00       0
 [3,]       1611  9.3977                4.5330 0e+00       0
 [4,]       3723 10.2761                4.5470 0e+00       0
 [5,]       3721 10.6168                4.6117 0e+00       0
 [6,]        157 18.4893                3.5398 0e+00       0
 [7,]       7649 28.4849                3.1126 0e+00       0
 [8,]       4263 37.2009                2.6502 0e+00       0
 [9,]       2151 40.1657                2.6314 0e+00       0
[10,]       6375 44.2341                2.5848 2e-04       0
[11,]       3726 44.6339                2.4984 4e-04       0
[12,]       4188 50.6936                2.3837 4e-04       0
```

[13,]	319	52.3108	2.4404	5e-04	0
[14,]	6903	53.3336	2.3912	4e-04	0
[15,]	6728	54.5806	2.3761	4e-04	0
[16,]	2679	55.2085	2.3799	4e-04	0
[17,]	5265	62.9521	2.3157	9e-04	0
[18,]	4454	64.0852	2.3282	9e-04	0
[19,]	3200	65.1399	2.2839	9e-04	0
[20,]	1782	66.1038	2.3614	9e-04	0

From both tables, we can see that the selected top 56 regulated genes, contains 36 down-regulated (`Table1`) genes and 20 up-regulated genes (`Table2`), for swirl mutation.

For all the top 56 genes, the p-value (`P.value`) is approximately zero, meaning that each of these genes is significantly differentially expressed.

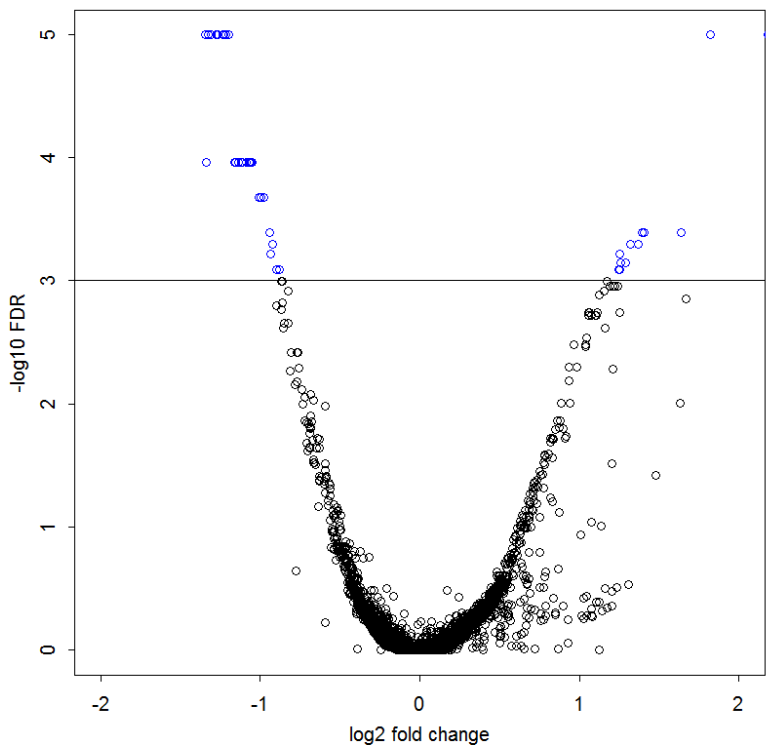
Graphical representation - VOLCANO Plot

If one chooses to compute the significance values (p-values) of the genes, it is interesting to compare the size of the fold change to the statistical significance level.

The **volcano plot** arrange genes along dimensions of biological and statistical significance. For details on how to implement this graphic check the R script at
`Script_Example_09 Multiple Testing.R`.

- The horizontal dimension is the fold change between the two groups (on a log scale, so that up and down regulation appear symmetric). Indicates biological impact of the change;

- The vertical axis represents the adjusted p-values (or the FDR) from the statistical test on a negative log scale, so smaller p-values (or FDR) appear higher up. Indicates the statistical evidence, or reliability of the change.



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.

Breitling, R., Armengaud, P., Amtmann, A., and Herzyk, P. (2004). Rank Products: A simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Letter*, 57383–57392.

Breitling, R. and Herzyk, P. (2006). Rank based methods as a non-parametric alternative of the t-statistic for the analysis of biological microarray data. *J. of Bioinformatics and Computational Biology* 3(5): 1171–1189.

Dudoit et al. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, Vol. 12, 111–139.

Hong, F. and Wittner, B. (2010). Bioconductor RankProd Package Vignette. Bioconductor.

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

Tusher, V.G. et al. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *PNAS*, Vol. 98, 5116–21.

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