

Advanced Topics in Biostatistics

Multiple Testing

December 16, 2020

Tim Holland-Letz

DKFZ

Multiplicities are everywhere

My guarantee: There is only a 5 % chance that this bike will break down. Buy it!



Multiplicities are everywhere – what is the problem?



...well, there is a 5% chance that the tires could go flat, but also a 5 % chance that the spokes could break or the frame ...





With 100 single parts there is a 99.4% chance that at least one of these parts will break down.
You should not have bought it!

Long long ago...

... in Basic Biostatistics: **Single Hypothesis Testing**

Test Problem: Null hypothesis H_0 vs. alternative hypothesis H_1

	H_0 not rejected	H_0 rejected
H_0 true	o.k.	α (Type I error)
H_0 false	β (Type II error)	o.k.

Concept of Hypothesis Testing:

Control the Type I error at a fixed significance level α (usually this will be 0.05) and choose a test statistics that maximizes the power $1-\beta$.

Multiple Hypothesis Testing: Counting Errors

Assume we are testing M null hypotheses: H_{0g} , $g=1, \dots, M$

	# H_{0g} not rejected	# H_{0g} rejected	
# H_{0g} true	U	V	h_0
# H_{0g} false	T	S	h_1
	M - R	R	M

h_0 = # true null hypotheses

R = # rejected null hypotheses

V = # Type I errors [false positives]

T = # Type II errors [false negatives]

Introduction Multiple Testing

From Single to Multiple Hypothesis Testing:

Suppose that we perform **10 tests**, each with significance level $\alpha = 0.05$.

What is the probability that we will get at least one false positive decision?

Increasing the numbers of tests:

$$10 \text{ tests} \rightarrow 1 - (0.95)^{10} = \mathbf{0.401}$$

$$100 \text{ tests} \rightarrow 1 - (0.95)^{100} = \mathbf{0.994}$$

$$1000 \text{ tests} \rightarrow 1 - (0.95)^{1000} \approx \mathbf{1}$$

Bonferroni Correction

adjust the local significance level α_i of each test

$$\alpha_i = \frac{\alpha}{M}, \quad i = 1, \dots, M;$$

M is the number of tests

$\alpha :=$ global significance level

Increasing M , decreases the local significance level:


$$10 \text{ tests} \rightarrow 1 - (1 - 0.005)^{10} = \mathbf{0.049}$$

$$100 \text{ tests} \rightarrow 1 - (1 - 0.0005)^{100} = \mathbf{0.049}$$

$$1000 \text{ tests} \rightarrow 1 - (1 - 0.00005)^{1000} = \mathbf{0.049}$$

Problem: The Bonferroni correction leads to very small values of the local significance levels, i.e. the null hypotheses is often not rejected

Example: DNA Microarray Studies

- Aim: Identification of differentially expressed genes (expression levels are associated with a response or covariate of interest)
- The question of differential expression can be reformulated as a problem in multiple testing: **Simultaneous** test for each gene of the null hypothesis of no association between the expression levels and the response or covariates
- Microarray experiments measure thousands of gene expression levels **simultaneously**  large **multiplicity** problem

Example: DNA Microarray Studies

- In such microarray studies, the probability of at least one false positive decision is near certain.

➡ $P(\text{at least one false positive decision}) = P(V \geq 1) \approx 1$

- Interest not just in the probability of one Type I error, but in the expected number of false positive decisions $E(V)$.

Example:

For 10,000 tests with $\alpha = 0.05$, the expected number of false positives is $10,000 * 0.05 = 500 = E(V)$.

➡ **Multiple Testing Procedures** have been developed to protect against making a false positive conclusion.

Type I Error Rates

- **Family-Wise Error Rate (FWER):**
the probability of at least one Type I error
$$\text{FWER} = P(V \geq 1)$$

- **False Discovery Rate (FDR):**
the expected proportion of Type I errors among the rejected hypotheses
$$\text{FDR} = E(V/R; R > 0) = E(V/R \mid R > 0)P(R > 0)$$

Type I Error Rates

- A **Multiple Testing Procedure** controls the Type I error rate at nominal (global) level α , e.g. regarding FDR the aim is to restrict the proportion of Type I errors among the rejected null hypotheses below a certain level, i.e. α .
Typical values of the nominal level α are 0.05, 0.1, or 0.2.
- In general, for a given multiple testing procedure,

$$\mathbf{FDR \leq FWER}$$

- Under the complete null, i.e. all null hypotheses are true,

$$\mathbf{FDR = FWER}$$

Comparison FDR vs. FWER

The FWER

- is extremely conservative, only few genes are called significant.
- is used when one needs to be certain that all called genes are truly positive, i.e. a trustworthy list of differentially expressed genes is needed.
- is important when making decisions about the admittance of medical treatments.
- by controlling the FWER one can miss out on potentially important genes (false negatives).

The FDR

- is used if the FWER is too stringent, one is more interested in having more true positives. The false positives can be sorted out in subsequent experiments (expensive).
- by controlling the FDR one can choose how many of the subsequent experiments one is willing to be in vain.

Adjusted p-Values

Unadjusted / adjusted p-Values:

If interest is in controlling, e.g., the FWER, the adjusted p-value for null hypothesis H_{0g} , $g=1, \dots, M$, is:

$$p_g^*$$

The null hypothesis H_{0g} is rejected at FWER α if $p_g^* \leq \alpha$.

Unadjusted p-values are called p_g .

Some Notations

Single Step Procedures:

Procedures which will take M unadjusted p-values and modify them separately.

Step-Wise Procedures:

More powerful procedures which adjust unadjusted p-values sequentially, from the smallest to the largest, or vice versa.

Ordering of observed p_g such that $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(M)}$.

Example Data

Notterman et al. (2001): Transcriptional Gene Expression Profiles of Colorectal Adenoma, Adenocarcinoma, and Normal Tissue Examined by Oligonucleotide Arrays, Cancer Research 61, 3124-3130.

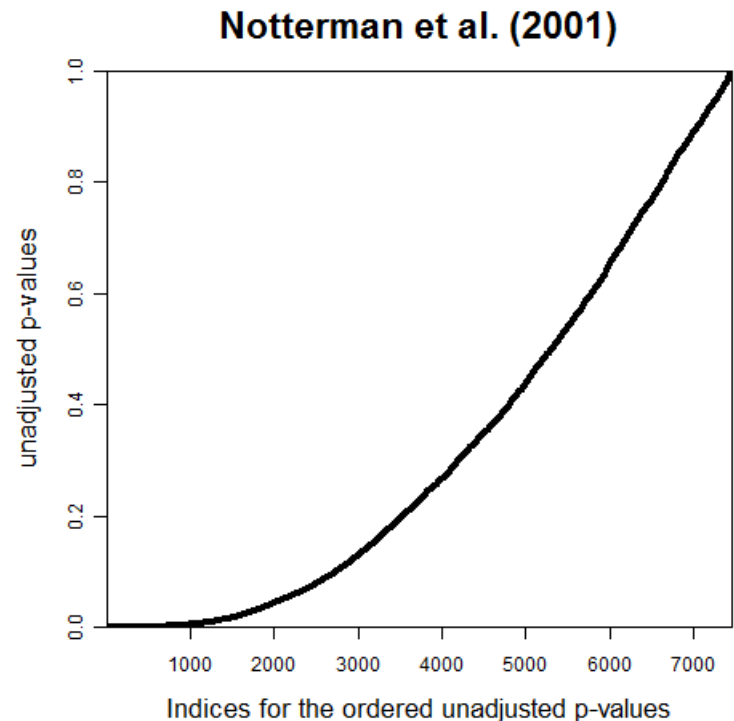
The Data:

Measurements on $M=7464$ oligonucleotides of 18 patients. For each patient tumor tissue and paired normal tissue has been analysed.

Hypotheses Testing:

H_{0g} : no difference between the expression of tumor and normal tissue.

Perform $M=7464$ paired t-tests, resulting in $M=7464$ unadjusted p-values.



Adjustment Methods: FWER (1)

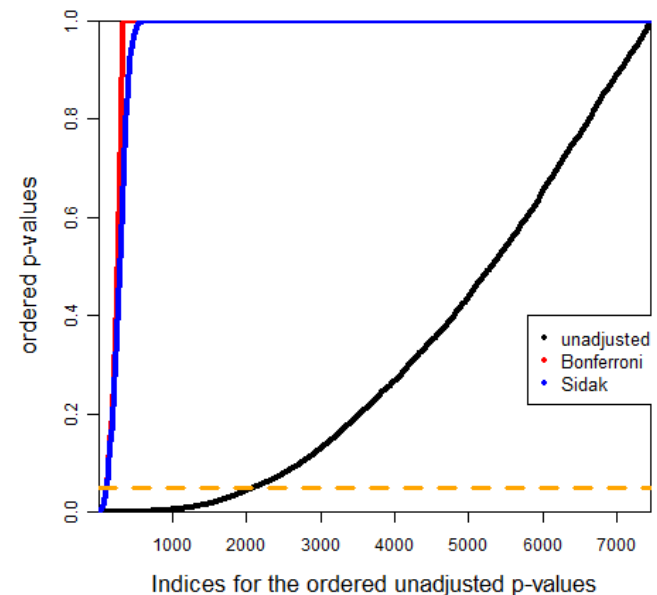
Single-Step Procedures

- Bonferroni correction:
 - single-step procedure
 - arbitrary dependency structure

$$p_g^* = \min(M \cdot p_g, 1)$$

- Sidak correction:
 - single-step procedure
 - **independence assumption**

$$p_g^* = 1 - (1 - p_g)^M$$



Example

5 p-values:

0.001, 0.021, 0.34, 0.88, 0.011

Ordered list	Bonferroni		
0.001	0.005 *		
0.011	0.055		
0.021	0.105		
0.34	1		
0.88	1		

*significant at 0.05 level

Adjustment Methods: FWER (2)

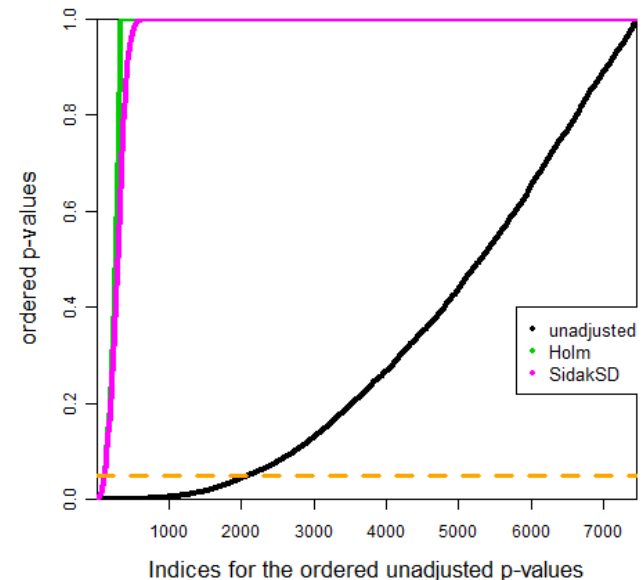
Step-Wise Procedures:

- (Bonferroni-)Holm (1979):
 - step-down procedure
 - arbitrary dependency

$$p_g^* = \max_{k=1,\dots,g} [\min((M-k+1) \cdot p_{(k)}, 1)]$$

- Sidak (1987):
 - step-down procedure
 - improvement over Holm (1979)
 - **independence assumption**

$$p_g^* = \max_{k=1,\dots,g} [1 - (1 - p_{(k)})^{M-k+1}]$$



Example

5 p-values:

0.001, 0.021, 0.34, 0.88, 0.011

Ordered list	Bonferroni	Bonferroni-Holm	
0.001	0.005 *	0.005*	
0.011	0.055	0.044*	
0.021	0.105	0.063	
0.34	1	n.s.	
0.88	1	n.s.	

*significant at 0.05 level

Adjustment Methods: FDR (1)

Procedures under independence assumption:

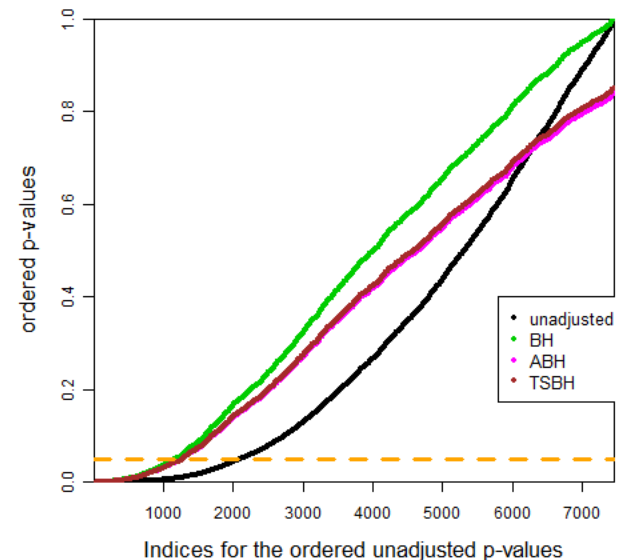
- Benjamini & Hochberg (1995):
step-up procedure

$$p_g^* = \min_{k=g, \dots, M} \left[\min \left(\frac{M}{k} \cdot p_{(k)}, 1 \right) \right]$$

- Others:

Adaptive BH (2006)

Two-Stage BH (2006)



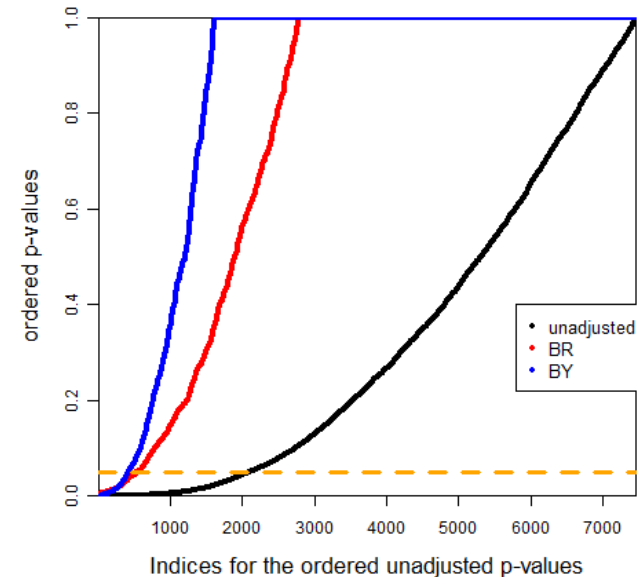
Adjustment Methods: FDR (2)

Procedures under arbitrary dependency:

- Benjamini & Yekutieli (2001):
conservative step-up procedure

$$p_g^* = \min_{k=g, \dots, M} \left[\min \left(\frac{M}{k} \sum_{j=1}^M \frac{1}{j} \cdot p_{(k)}, 1 \right) \right]$$

- Blanchard & Roquain (2008)



Example

5 p-values:

0.001, 0.021, 0.34, 0.88, 0.011

Ordered list	Bonferroni	Bonferroni-Holm	Benjamini-Hochberg
0.001	0.005 *	0.005*	0.005*
0.011	0.055	0.044*	0.0275*
0.021	0.105	0.063	0.035*
0.34	1	n.s.	0.425
0.88	1	n.s.	0.88

*significant at 0.05 level

How to Choose a Multiple Testing Procedure?

Problem:

There is a large variety of multiple testing procedures. How can we choose which to use?

Selection Criteria:

- Interpretation: does the chosen Type I error rate answer a relevant question for you?
- Validity: are the assumptions under which the procedure applies clear and definitely or plausibly true, or are they unclear and most probably not true?
- Computability: are the procedure's calculations straightforward to calculate accurately, or is there possibly numerical or simulation uncertainty, or discreteness?

How to Choose a Multiple Testing Procedure?

Rules of Thumb:

- If you want no false positives at all:
Bonferroni-Holm
- If you accept some false positives, but want to limit them to a fraction of your positive results:
Benjamini- Hochberg
- In the same situation, but you expect strong correlation between your tests:
Benjamini & Yekutieli
- False positives are of little consequence:
No adjustment

- **R function `p.adjust()`**

Example:

```
pvalues<-c(0.02,0.003,0.54,0.01)
p.adjust(pvalues,method='holm')
[1] 0.040 0.012 0.540 0.030
```

Other methods: 'bonferroni', 'BH', 'BY', ...

- **Chipster**
- **Mutoss (Graphical user interface for R)**

Alternative: Hierarchical testing

- Mainly used in clinical trials.
- Conduct all desired tests in a prespecified sequence.
- Only conduct a later test if the previous test was significant.
- Otherwise, testing is stopped.

Example:

A new treatment is supposed to both show at most 10% more side effects, and a superior effectiveness compared to standard.

Solution:

- First test, whether rate of side effects is $< \text{rate}(\text{standard}) + 10\%$.
- If this is significant, test whether effectiveness $> \text{effectiveness}(\text{standard})$.
- FWER will be kept at the desired level.

Multiple Testing: General comments

Any kind of repeated analysis of the same data is multiple testing:

- Interim analyses in clinical trials
- Trying different statistical tests for the same data („p-value hacking“)
- Even looking at the raw data to preselect interesting variables for further analysis

Increased chance of alpha errors can be the consequence in every case!

The „reproducibility crisis“ is in large parts caused by multiple testing.

Next Lecture

**January 13, 2021:
Linear mixed models**

Dr. Nicholas Schreck

Selected References

- Benjamini, Y and Hochberg, Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing JRSS B 57: 289-300
- Benjamini Y, Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. Annals of Statistics 29: 1165-1188
- Benjamini Y, Krieger AM, Yekutieli D (2006) Adaptive linear step-up procedures that control the false discovery rate. Biometrika 93: 491-507
- Blanchard G, Roquain E (2008) Two simple sufficient conditions for FDR control. Electronic Journal of Statistics 2: 963-992
- Dudoit S, van der Laan MJ (2007) Multiple Testing Procedures with Applications to Genomics. Springer
- Efron B, Tibshirani R, Storey JD, Tusher V (2001) Empirical Bayes analysis of a microarray experiment. Journal of the American Statistical Association 96: 1151-1160
- Holm S (1979) A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics 6: 65-70
- Hochberg Y (1988) A sharper Bonferroni procedure for multiple tests of significance. Biometrika 75: 800-802
- Sidak Z (1967) Rectangular confidence regions for the means of multivariate normal distributions. Journal of the American Statistical Association 62: 626-633
- Storey JD (2002) A direct approach to false discovery rates. Journal of the Royal Statistical Society: Series B 64: 479-498
- Westfall, PH and Young, SS (1993) Resampling-based multiple testing: Examples and methods for p-value adjustment, John Wiley & Sons, Inc