

# Introduction to multiple Testing

Boris Hejblum

October 15<sup>th</sup>, 2019

# Introduction

# Multiple univariate statistical tests

- easy way to explore data: (many) **univariate** tests  
*e.g. between each explanatory variable and the outcome*
- can be a (bad) way to select variables for a small(er) multivariate model  
     $\Rightarrow$  *screening method*
- when multivariate models do not make sense or are not feasible

# One statistical test

Your decision/Reality	$H_0$ is true	$H_0$ is false
Do not reject $H_0$ (test non-significant)	Correct decision	Wrong decision
Reject $H_0$ (test significant)	Wrong decision	Correct decision

# 20 statistical tests

If we repeat the procedure 20 times, that is, *if we perform 20 univariate tests **without changing anything***:  $\implies$  great chance of having false positive detections...

How many tests do we expect will be false positive detection ?

# $m$ statistical tests

- How do we adapt to the fact that we have many ( $m$ ) tests ?
- Can we **adjust** the level of significance  $\alpha$  accordingly ?

# Russian roulette



Assume that a gun has 20 locations and contains one bullet.

- *pull the trigger*  
1 time: gun fires with **5%** probability
- *pull the trigger*  
10 times: gun fires with **40%** probability
- *pull the trigger*  
25 times: gun fires with **72%** probability

# Probability

The probability of no undesirable event (false positive or gun firing) is:

$$(1 - \alpha)^m$$



Type-I error for multiple tests

# Multiple testing notations

- $m$ : total number of tests
- $\mathcal{M}_0$ : the set of true null hypotheses
- $V$ : number of false positive (null hypotheses wrongly rejected)
- $R$ : number of null hypotheses rejected

Null hypotheses	True	False	Total
Non-rejected	$U$	$T$	$W$
Rejected	$V$	$S$	$R$
Total	$m_0$	$m - m_0$	$m$

# Family Wise Error Rate (FWER)

$$FWER = P(V > 0 | \mathcal{M}_0)$$

$\implies$  the probability to get at least one false positive, knowing the true null hypotheses.

- we want to stricly control the number of false positives
- $m$  is “not so large”

*Often for confirmatory analyses*

# False Discovery Rate (FDR)

$$FDR = E[V/R | \mathcal{M}_0]$$

$\implies$  the expected number of false positives on average among rejected null hypotheses, knowing the true null hypotheses.

- $m$  is “really large”
- the FWER is too conservative  
(we do not reject any null hypothesis)

*Often for exploratory/hypothesis generating analyses*

# Multiple testing correction

# Correction for multiple tests

When we take into account the number of tests, we can either:

- Correct p-values (called **adjusted p-values**) and keep the significance level  $\alpha$  fixed
- Keep raw p-values and correct significance level  $\alpha$

$\implies$  Both ways are equivalent *but software usually use adjusted p-values*

# Correction to control the FWER (1)

*Bonferroni correction:*

Compare p-values to  $\alpha/m$  instead of  $\alpha$

$\implies$  adjusted p-values:  $q = \min(1, mp)$

- **Controls the FWER**  
(playing *Russian roulette* 10 times with a gun with 200 slots is “safer”)
- Too conservative as soon as  $m$  get large

# Correction to control the FWER (2)

*Holm correction:*

1. Compare the smallest p-value to  $\alpha/m$ . If the associated null hypothesis is not rejected **stop**, otherwise **continue**
2. Compare the second smallest p-value to  $\alpha/(m - 1)$   
If the associated null hypothesis is not rejected **stop**, otherwise **continue**
3. ...
4. Compare the largest p-value to  $\alpha/1$  and **conclude**

⇒ also **controls the FWER** but a bit less conservative than Bonferonni  
(more null hypotheses could be rejected, if  $m$  is not too large)



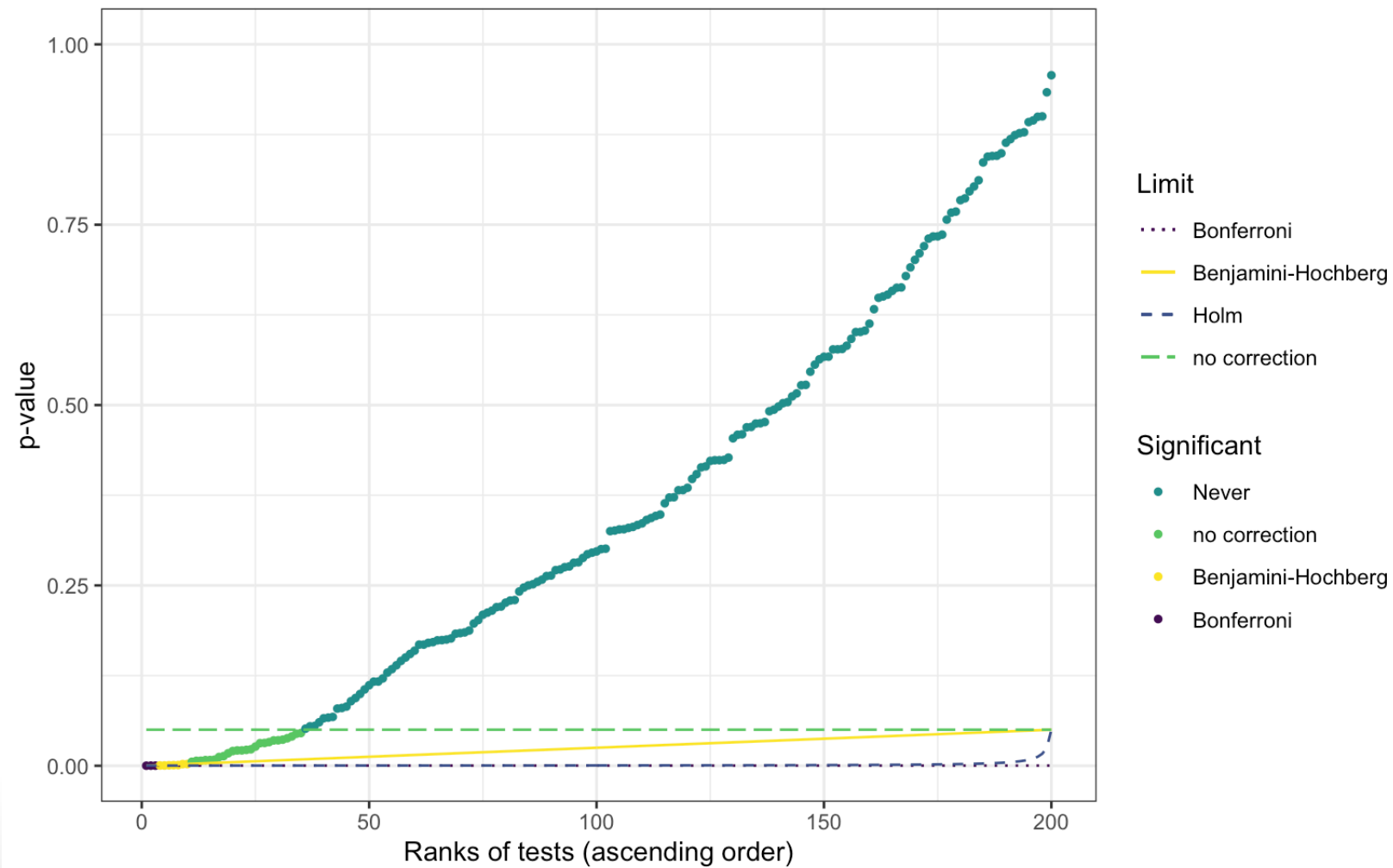
# Correction to control the FDR

1. Compare the smallest p-value to  $\alpha/m$ , the second smallest p-value to  $2\alpha/m$ , ..., and the largest p-value to  $\alpha$ .
2. Find the largest p-value that is strictly less than its associated threshold. We note this p-value  $p^*$ .
3. Reject all null hypotheses associated to p-values smaller than  $p^*$ .

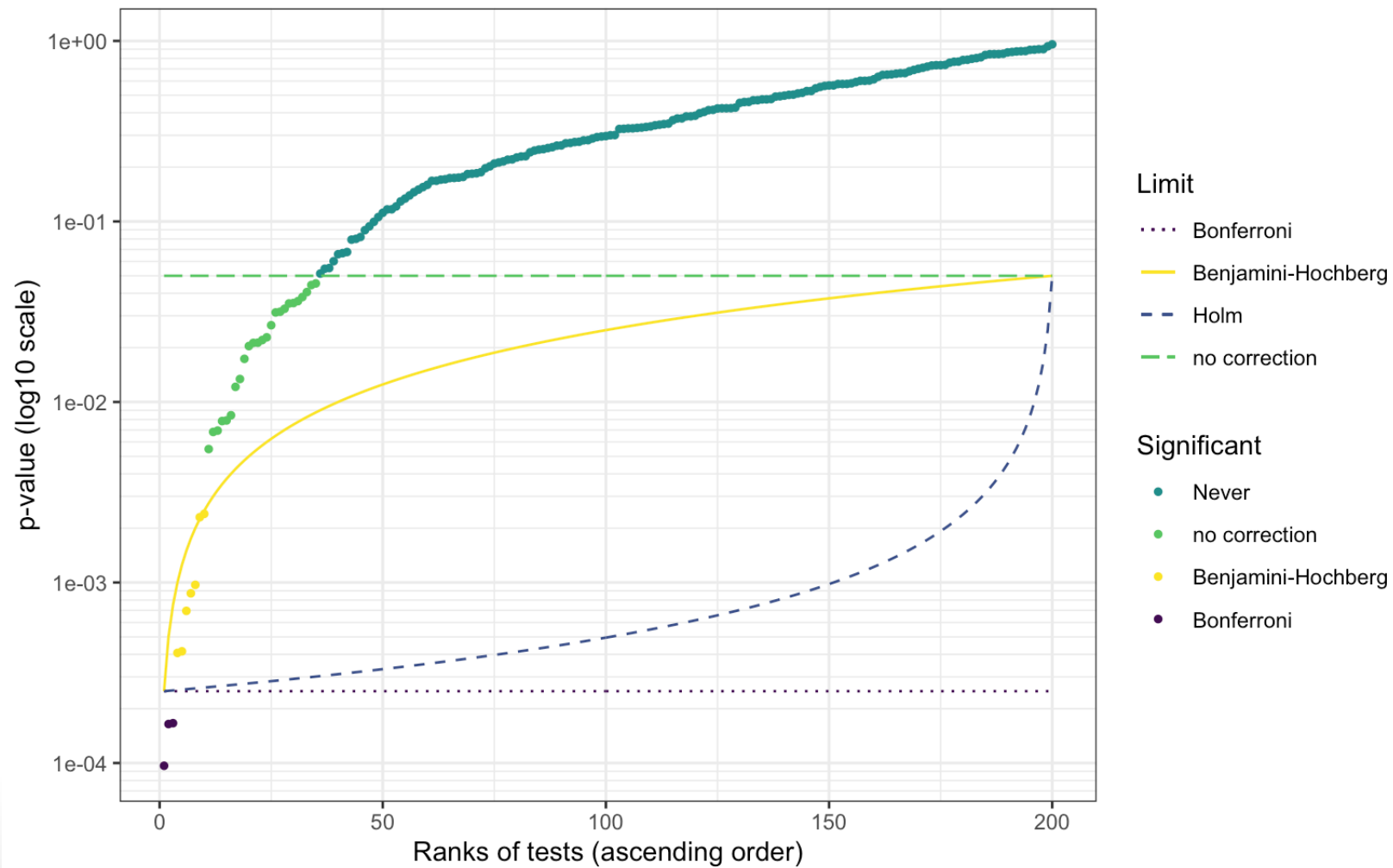
$\implies$  **controls the FDR** (but not the FWER)

Hence, even more flexible than Holm correction and usually used when  $m$  is very large

# Vizualizing an example



# Vizualizing an example – log-scale



# Conclusion

# Take home message

Multiple testing must be taken into account:

- $\Rightarrow$  p-values must be adjusted !
- choose the correction method according to your scientific objectives
- **NEVER** play at Russian roulette (seriously) !

# Further reading

- Dudoit, S & van der Laan, J. *Multiple Testing Procedures with Applications to Genomics*. Springer Series in Statistics (2008).
- Foulkes, AS. *Applied statistical genetics with R: for population-based association studies*. Springer Verlag (2009).
- Benjamini, Y & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* (1995).
- Holm, S. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*. (1979).
- Phipson B & Smyth GK. Permutation p-values should never be zero: calculating exact p-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology*. 31;9(1):1544–6115 (2010).