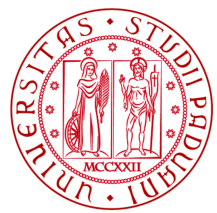
Permutation-based multiple testing with R - theory 1/2

Livio Finos
livio.finos@unipd.it

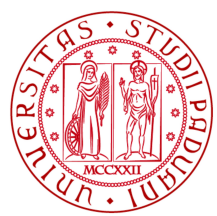
University of Padua

Ribno (Bled), Slovenia, 23/09/18



Outline

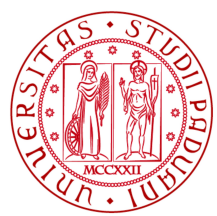
1. **Introduction**
   - Some examples
   - Some remarks
2. **FamilyWise Error Rate (FWER)**
   - Definition + Bonferroni
   - Holm (step-wise)
   - Closed Testing
3. **FWER control by Permutations tests**
   - Permutation Bonferroni
   - Westfall & Young: permutation Holm[1]
   - Closed Testing
4. **Conclusion**



[1] Westfall PH, Young SS (1993) Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment. Wiley.

Outline

1. **Introduction**
   - Some examples
   - Some remarks
2. **FamilyWise Error Rate (FWER)**
   - Definition + Bonferroni
   - Holm (step-wise)
   - Closed Testing
3. **FWER control by Permutations tests**
   - Permutation Bonferroni
   - Westfall & Young: permutation Holm[2]
   - Closed Testing
4. **Conclusion**



[2] Westfall PH, Young SS (1993) Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment. Wiley.

Microarray study

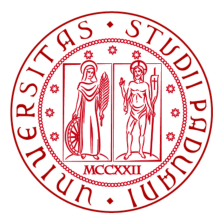
Compare 20,000 genes activity between disease and control group:

Top 5 genes out of 20,000

Gene	p-value
OCIAD2	5.5e-6
NEK3	6.7e-6
TAF5	7.1e-6
FOXD4L6	7.5e-6
ADIG	8.8e-6
⋮	⋮

Small p-value?

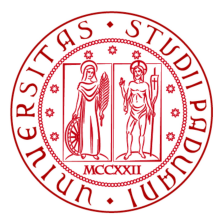
- a p=5.5e-6: very unlikelly (evidende in favour of H_1)
- it is unlikelly also on the light of 20,000 tested hypotheses?
- is OCIAD2 really differentially expressed?
- what about NEK3?



Surveys

e.g. self report inventory (web), psychological tests or social surveys

For each question we can get one or more tests
(e.g. comparing the opinions of two groups of people)

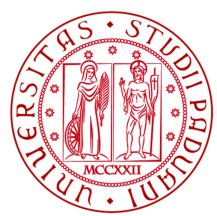


Surveys

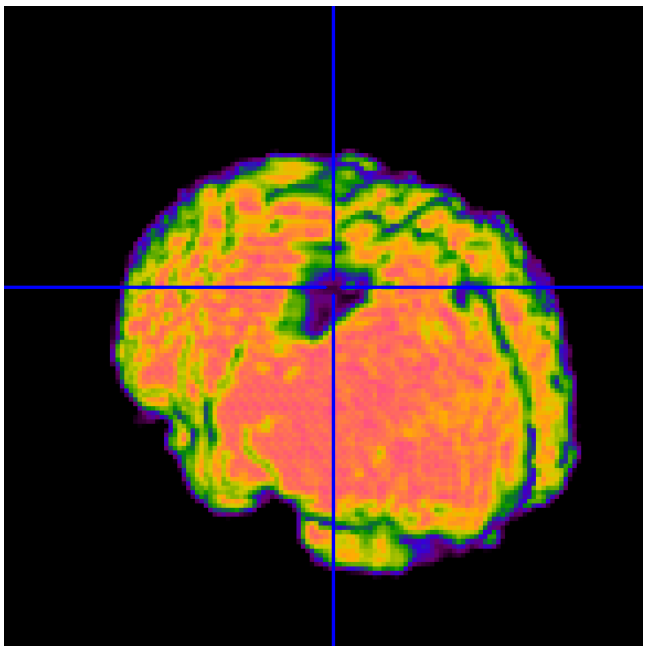
e.g. self report inventory (web), psychological tests or social surveys

For each question we can get one or more tests
(e.g. comparing the opinions of two groups of people)

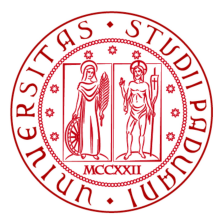
Question: Do we need to control for multiple type I errors?



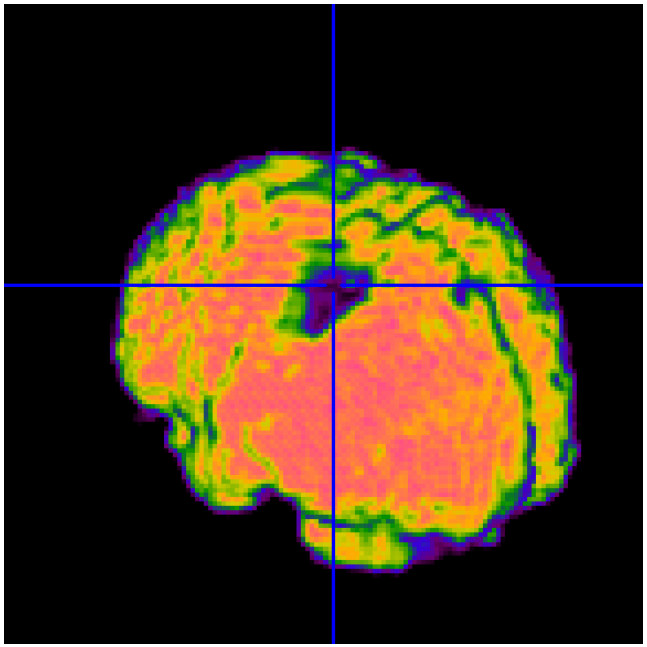
A further example: fMRI studies



A map of brain activity for each subject

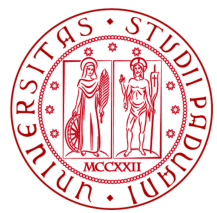


A further example: fMRI studies

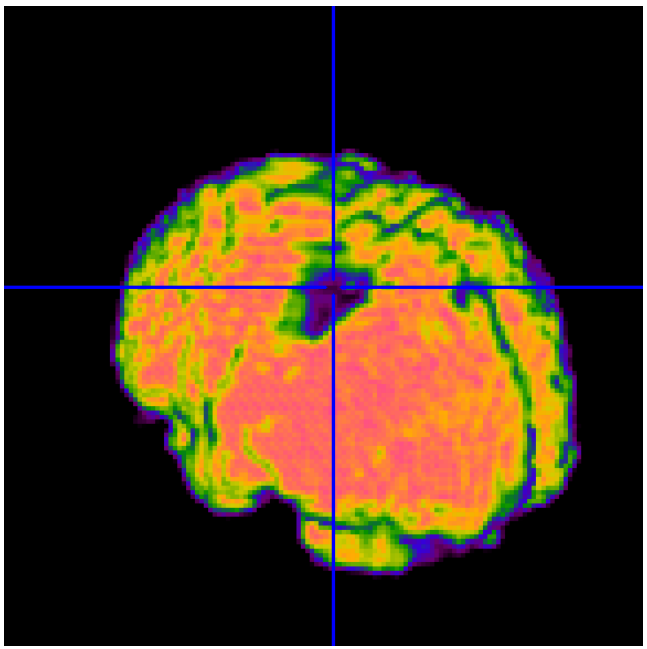


A map of brain activity for each subject

A p-value for each voxel (=pixel, point)



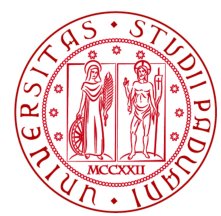
A further example: fMRI studies



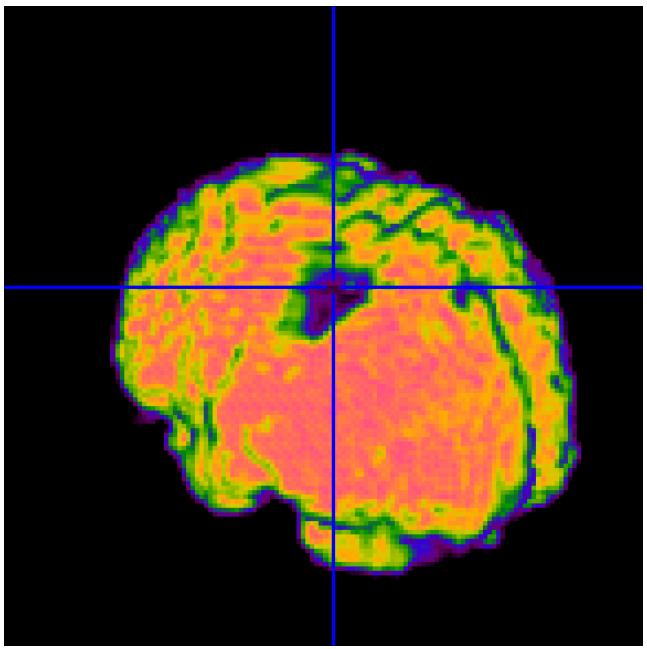
A map of brain activity for each subject

A p-value for each voxel (=pixel, point)

The output is a list of most significant voxels (among the thousands tested)



A further example: fMRI studies

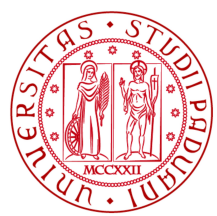


A map of brain activity for each subject

A p-value for each voxel (=pixel, point)

The output is a list of most significant voxels (among the thousands tested)

Question: Do we need to control for multiple type I errors?



Further examples

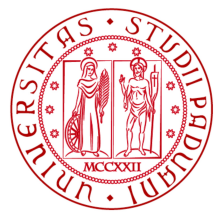
Kinematics of human movements
a test for each aspect of the movement

Regression models (e.g. LM e GLM)
a t-test for each parameter of the model

Anova
All pairs comparisons (post-hoc)

Every times we get more than one p-value ...

Question: Do we need to control for multiple type I errors?



Further examples

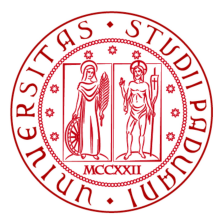
Kinematics of human movements
a test for each aspect of the movement

Regression models (e.g. LM e GLM)
a t-test for each parameter of the model

Anova
All pairs comparisons (post-hoc)

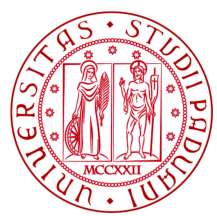
Every times we get more than one p-value ...

Question: Do we need to control for multiple type I errors?
Answer: Yes, of course!



Outline

1. **Introduction**
   - Some examples
   - Some remarks
2. **FamilyWise Error Rate (FWER)**
   - Definition + Bonferroni
   - Holm (step-wise)
   - Closed Testing
3. **FWER control by Permutations tests**
   - Permutation Bonferroni
   - Westfall & Young: permutation Holm[3]
   - Closed Testing
4. **Conclusion**



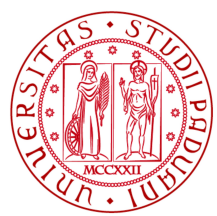
[3] Westfall PH, Young SS (1993) Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment. Wiley.

Hypotesis testing, a single test

compare the two hypos:

- H_0: two groups are Equal, Y doesn't depend on X, nothing to publish :(
- H_1: two groups are Different, Y does depend on X, publish it! :)

Each test produces a p-value p,
if $p \leq .05$ **(**$\alpha = .05$**) reject** H_0 **(and incline to** H_1**)**



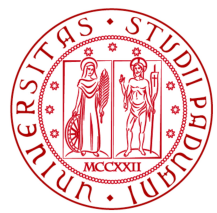
Errors

- **Type I** (false positive): Reject H_0 when it is TRUE
  $P(\textit{Type I error}) = P(p \leq .05|H_0) = .05$
- **Type II** (false negative): Don't reject H_0 when it is FALSE
  $P(\textit{Type II error}) = P(p > .05|H_1)$
  Power:

$$\begin{aligned} P(p \leq .05|H_1) &= 1 - P(p > .05|H_1) \\ &= 1 - P(\textit{Type II error}) \end{aligned}$$

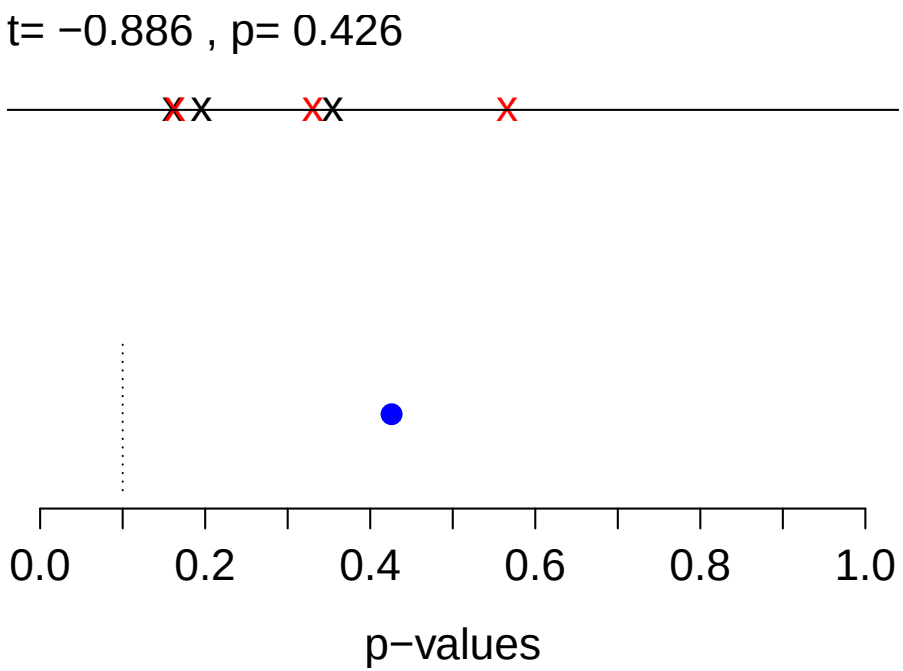
Power is nothing without control
(Asymmetric relevance of errors)

We control the $P(\textit{type I Error})$ (ie $\leq .05$)
and (try) to maximize the power (i.e. minimize *type II error*)



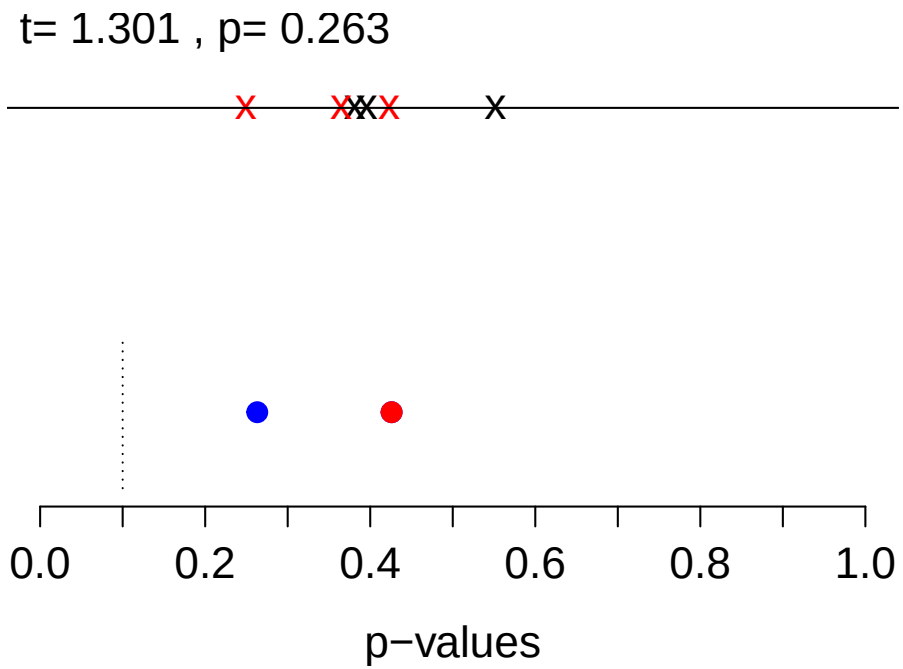
Type I Error:

$$\mathbf{P}(\mathbf{p} \leq .\mathbf{05} | \mathbf{H_0} = \text{ groups are equal}) = ?$$



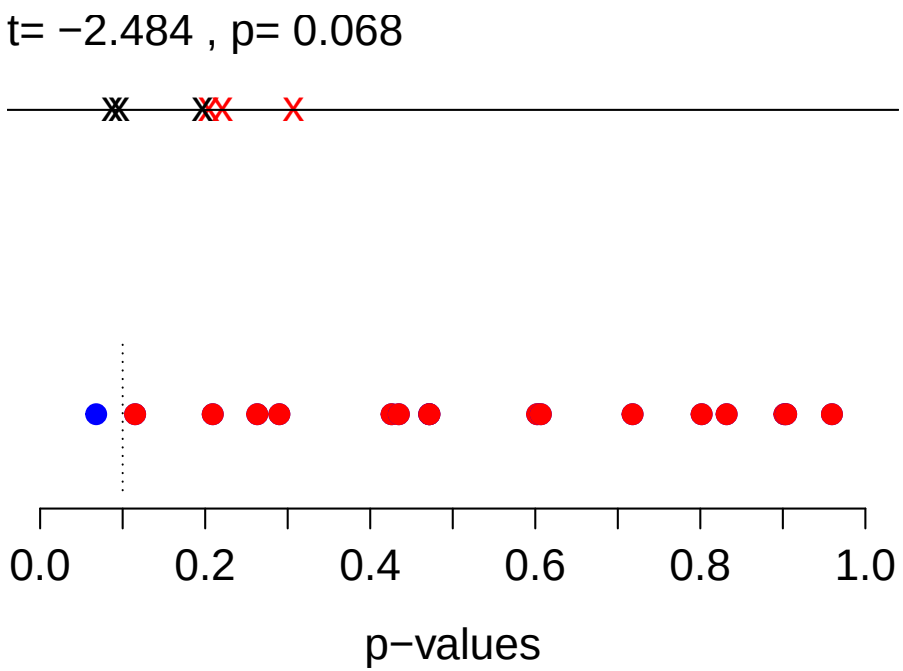
Type I Error:

$$\mathbf{P}(\mathbf{p} \leq .\mathbf{05} | \mathbf{H_0} = \text{ groups are equal}) = ?$$



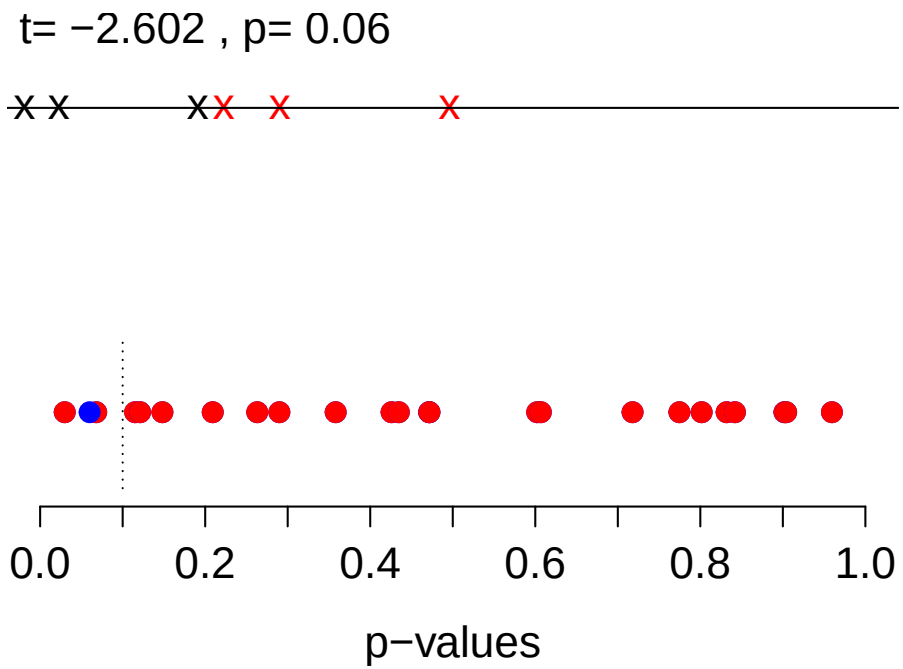
Type I Error:

$$\mathbf{P}(\mathbf{p} \leq .\mathbf{05} | \mathbf{H_0} = \text{ groups are equal}) = ?$$



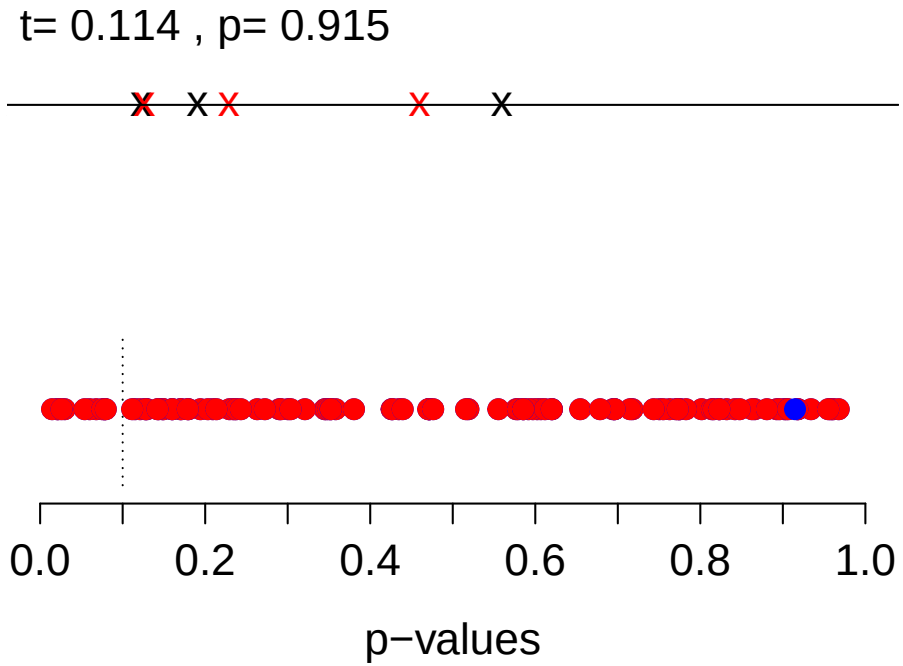
Type I Error:

$$\mathbf{P}(\mathbf{p} \leq .\mathbf{05} | \mathbf{H_0} = \text{ groups are equal}) = ?$$



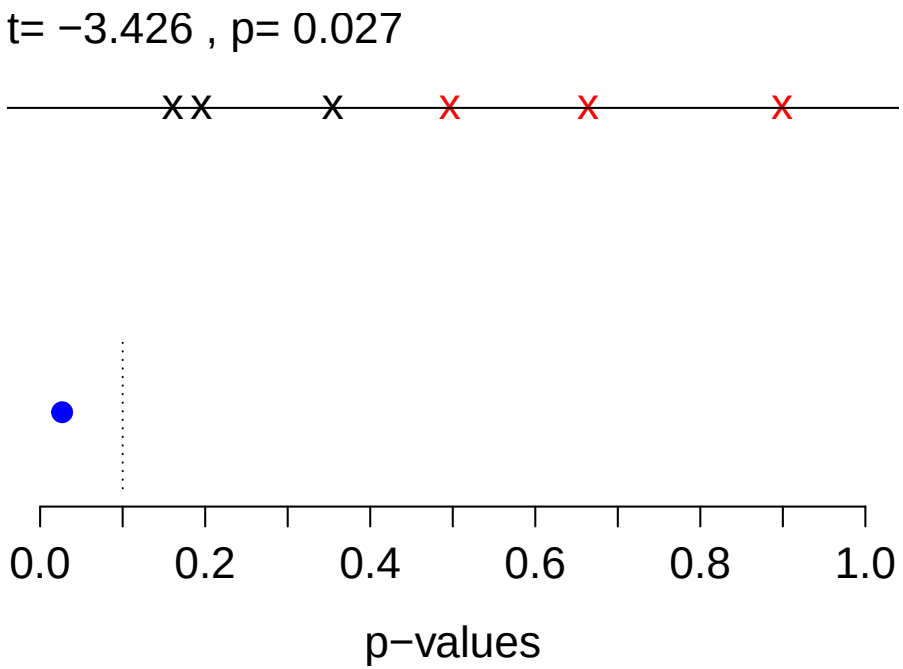
Type I Error:

$$\mathbf{P(p \leq .05 | H_0 = } \text{ groups are equal}) = \mathbf{0.05}$$



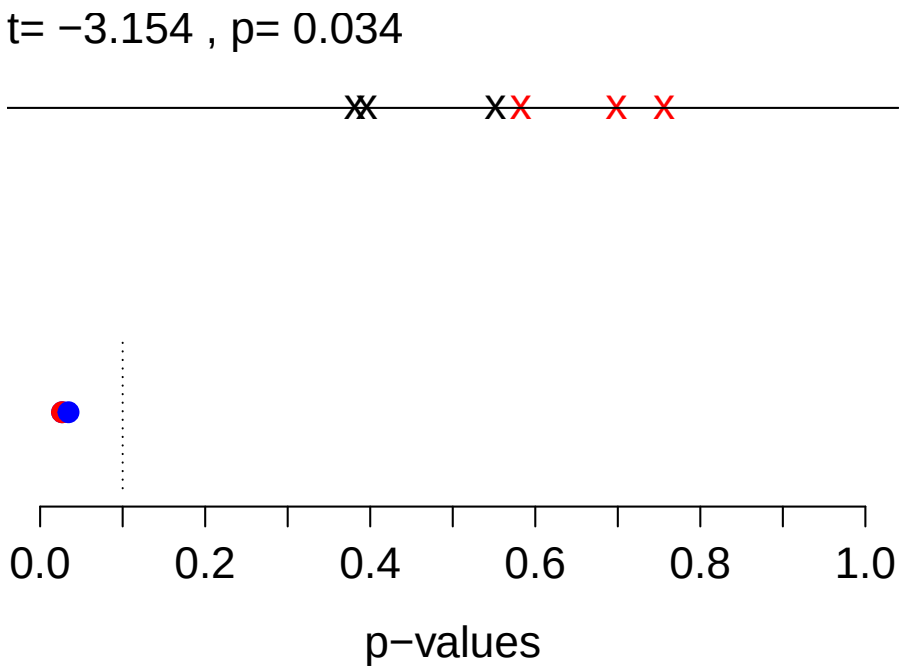
Power:

$$\mathbf{P}(\mathbf{p} \leq .\mathbf{05} | \mathbf{H_1} = \text{ groups are NOT equal})$$



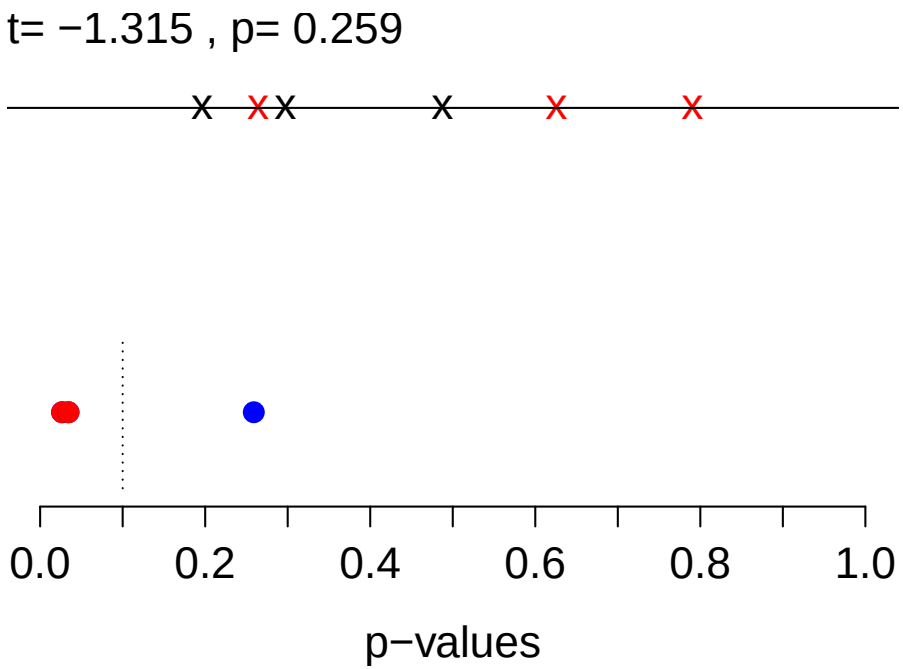
Power:

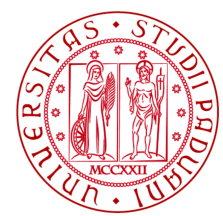
$$\mathbf{P}(\mathbf{p} \leq .\mathbf{05} | \mathbf{H_1} = \text{ groups are NOT equal})$$



Power:

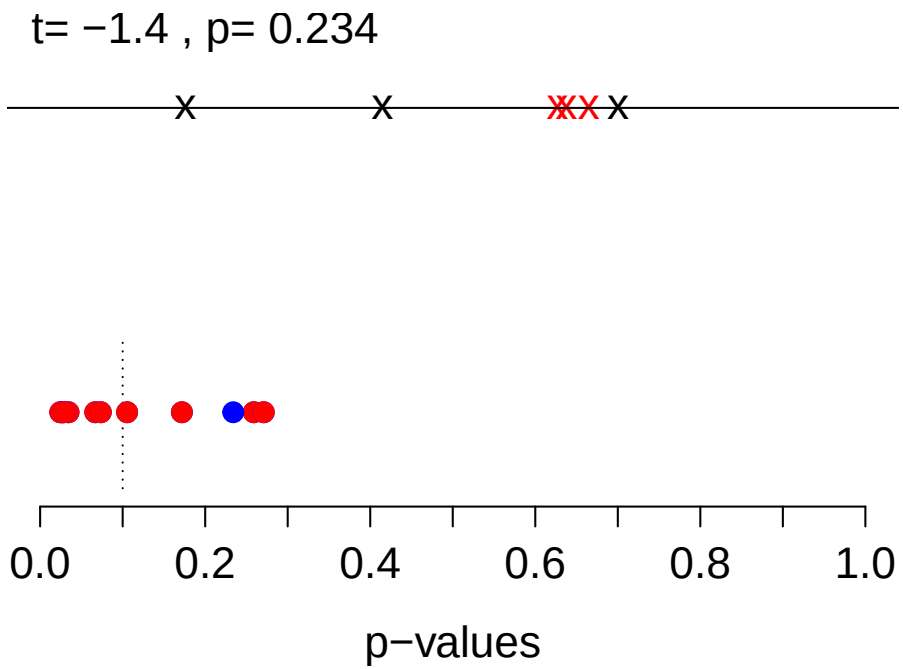
$$\mathbf{P}(\mathbf{p} \leq .\mathbf{05} | \mathbf{H_1} = \text{ groups are NOT equal})$$

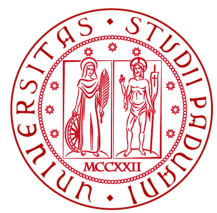




Power:

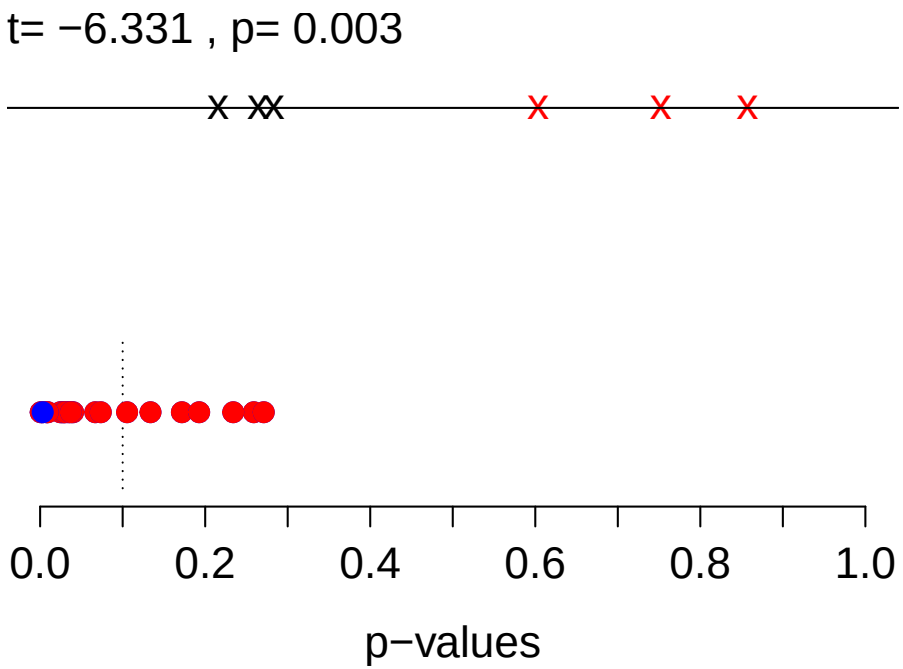
$$\mathbf{P}(\mathbf{p} \leq .\mathbf{05} | \mathbf{H_1} = \text{ groups are NOT equal})$$

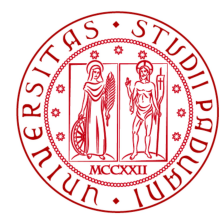




Power:

$$\mathbf{P}(\mathbf{p} \leq .\mathbf{05} | \mathbf{H_1} = \text{ groups are NOT equal})$$

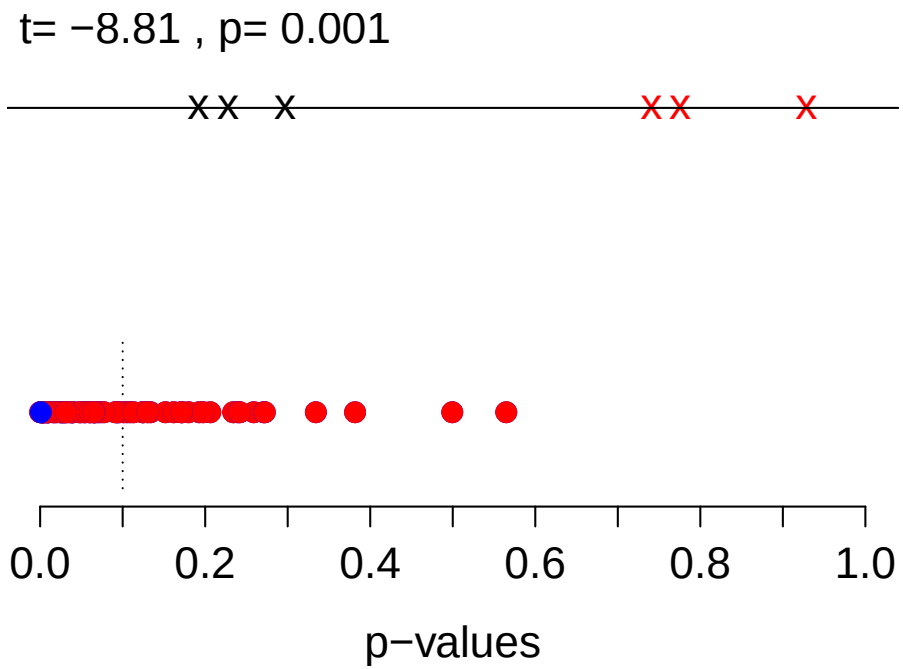


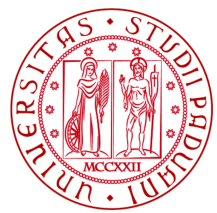


Power:

$\mathbf{P(p \leq .05 | H_1 =}$ groups are NOT equal)

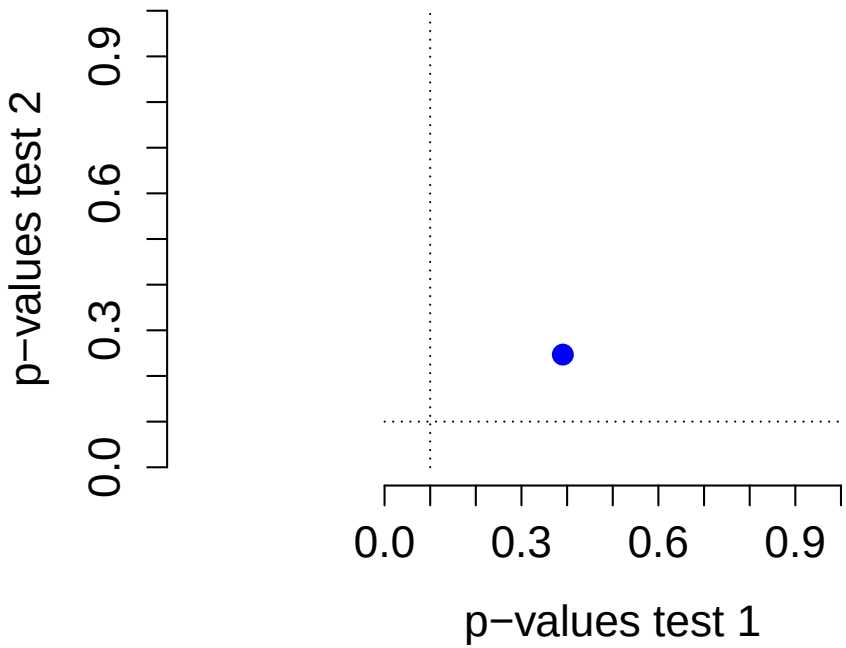
e.g.: *Power* : $P(p \leq 0.05 | H_1) = 0.75$

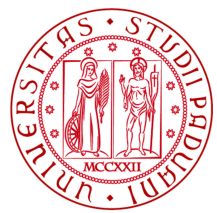




Type I Error, Two Tests

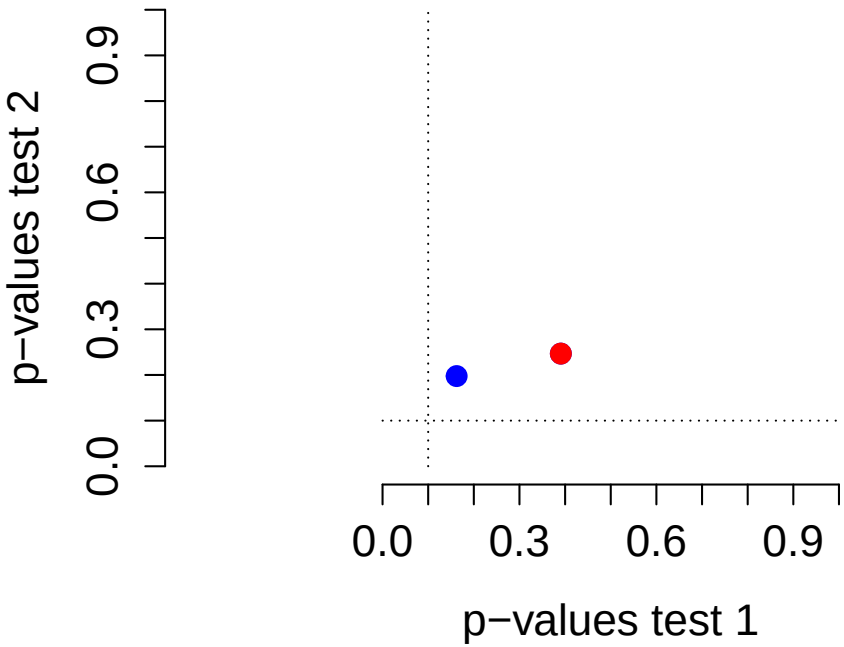
Probability of at LEAST one (false) rejection?

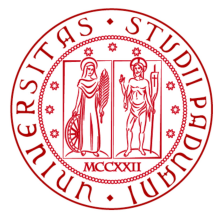




Type I Error, Two Tests

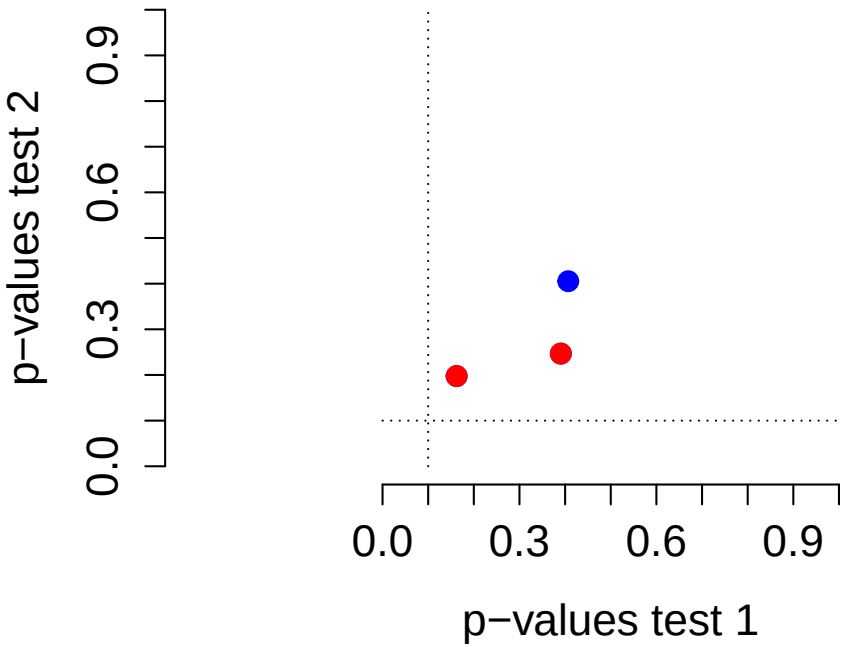
Probability of at LEAST one (false) rejection?

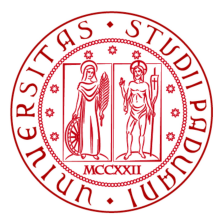




Type I Error, Two Tests

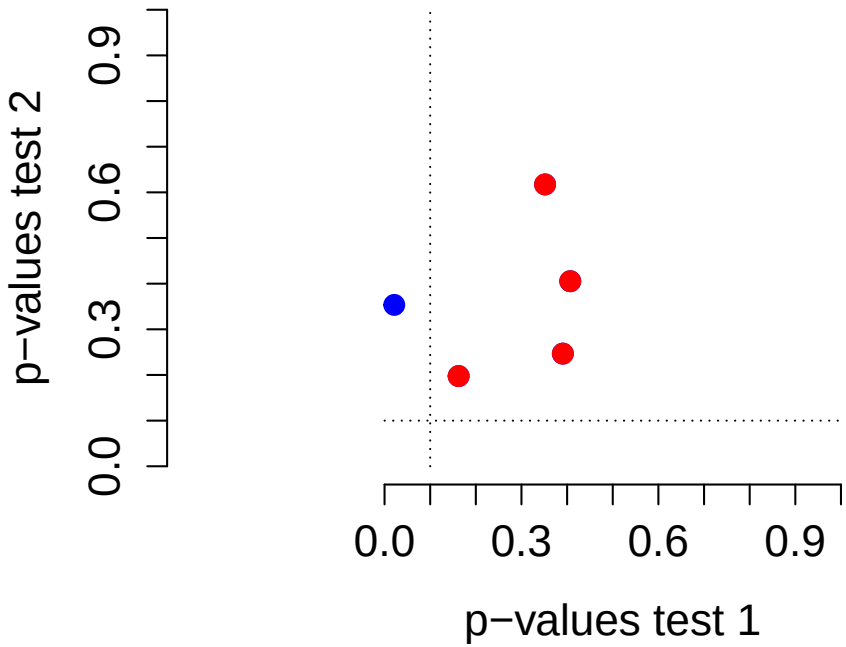
Probability of at LEAST one (false) rejection?

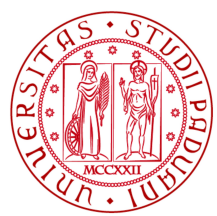




Type I Error, Two Tests

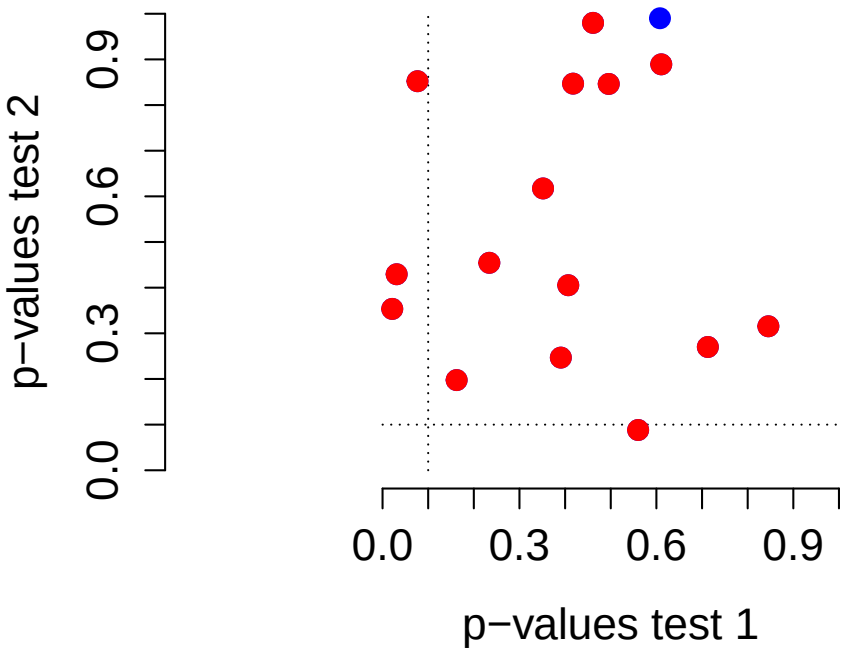
Probability of at LEAST one (false) rejection?

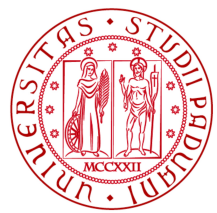




Type I Error, Two Tests

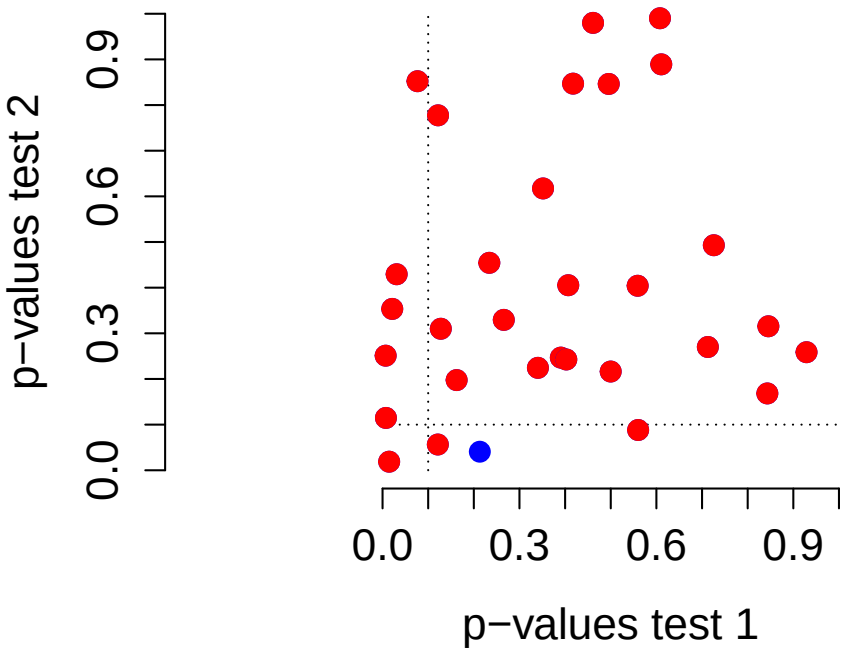
Probability of at LEAST one (false) rejection?

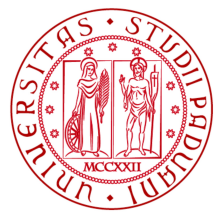




Type I Error, Two Tests

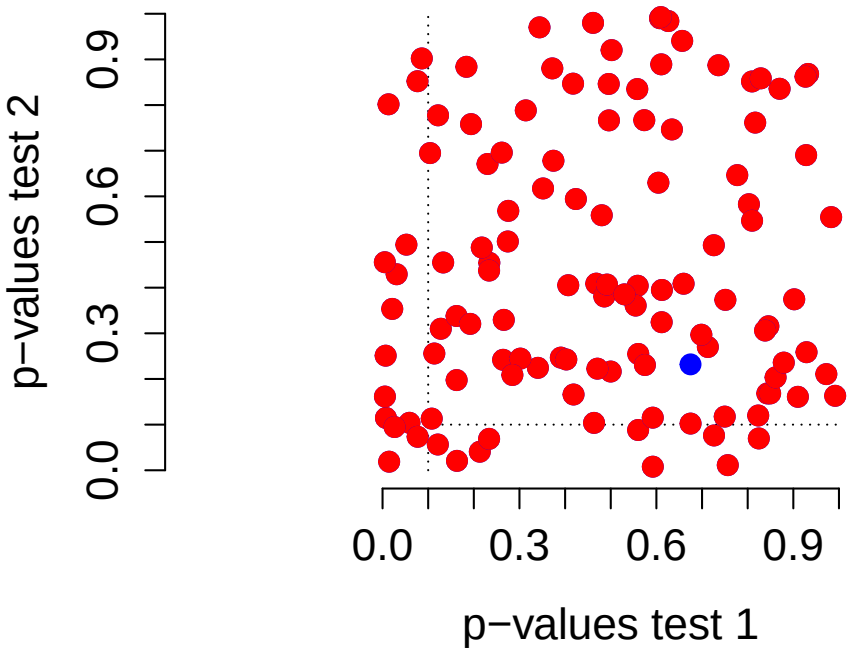
Probability of at LEAST one (false) rejection?

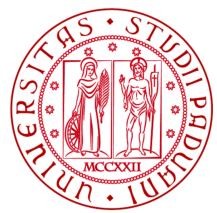




Type I Error, Two Tests

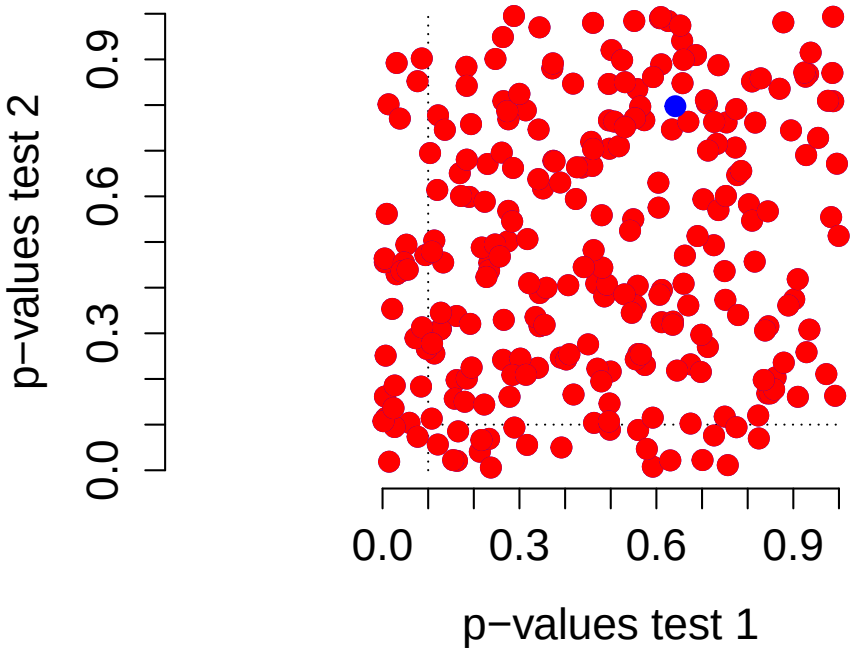
Probability of at LEAST one (false) rejection?

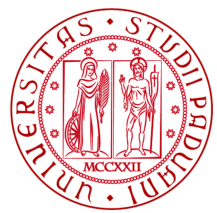




Type I Error, Two Tests

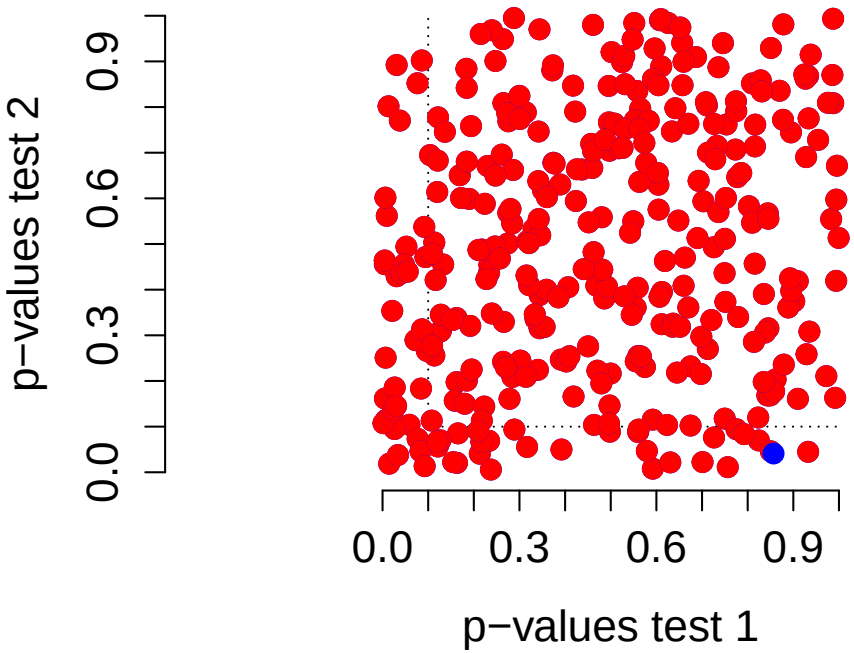
Probability of at LEAST one (false) rejection?

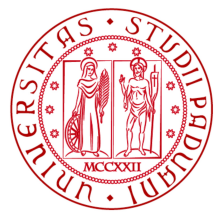




Type I Error, Two Tests

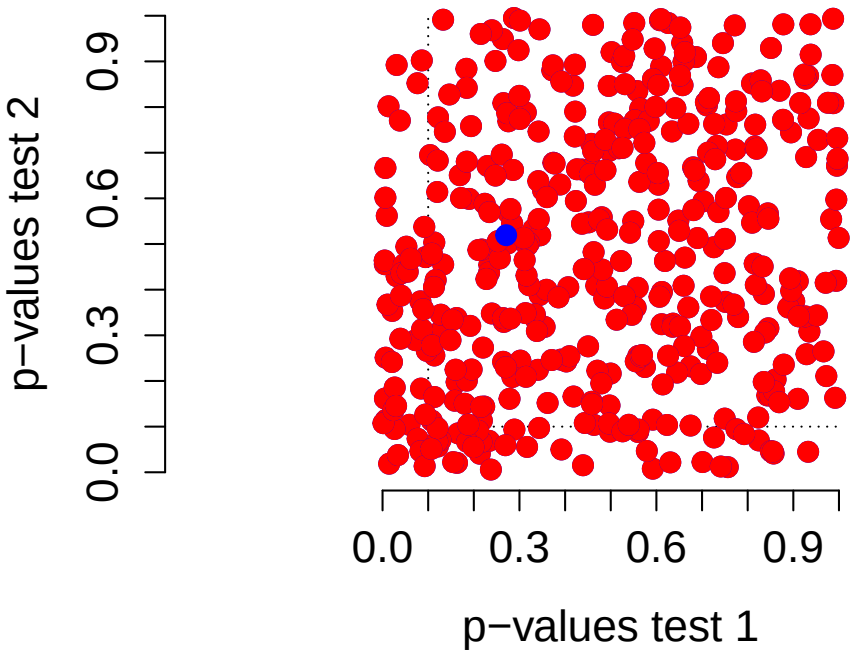
Probability of at LEAST one (false) rejection?

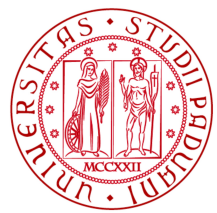




Type I Error, Two Tests

Probability of at LEAST one (false) rejection?

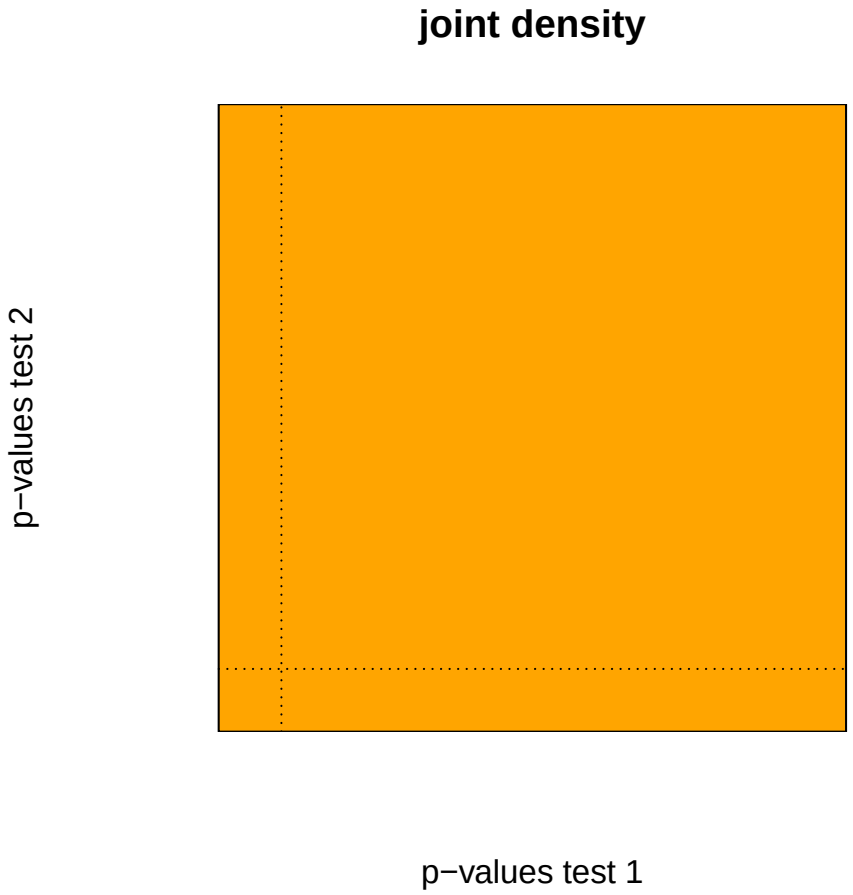


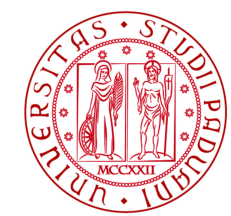


Type I Error, Two Tests

Probability of at LEAST one (false) rejection?

$= .05 + .05 - (.05 * .05) = 1 - (1 - .05)^2 = .0975 = 1 - (1 - \alpha)^2$





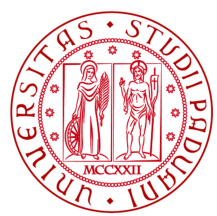
Probability of false rejections

m independent tests (p-values)

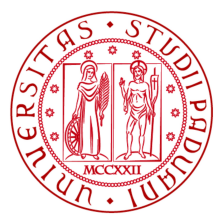
Reject H_0 if $p \leq \alpha$

Probability of AT LEAST a rejection

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

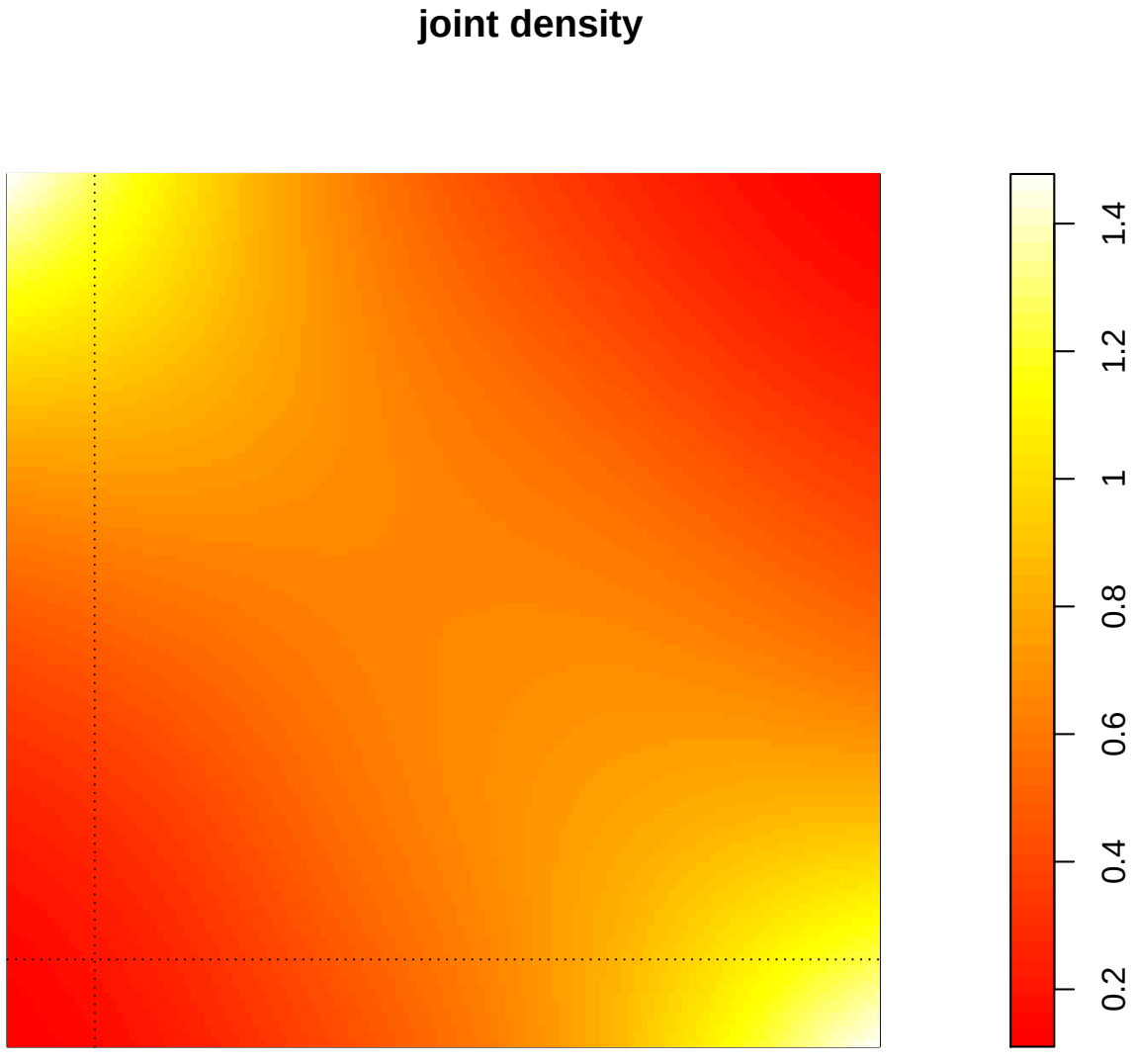


Probability of AT LEAST one Type I error as a function of number of tests



Dependent tests

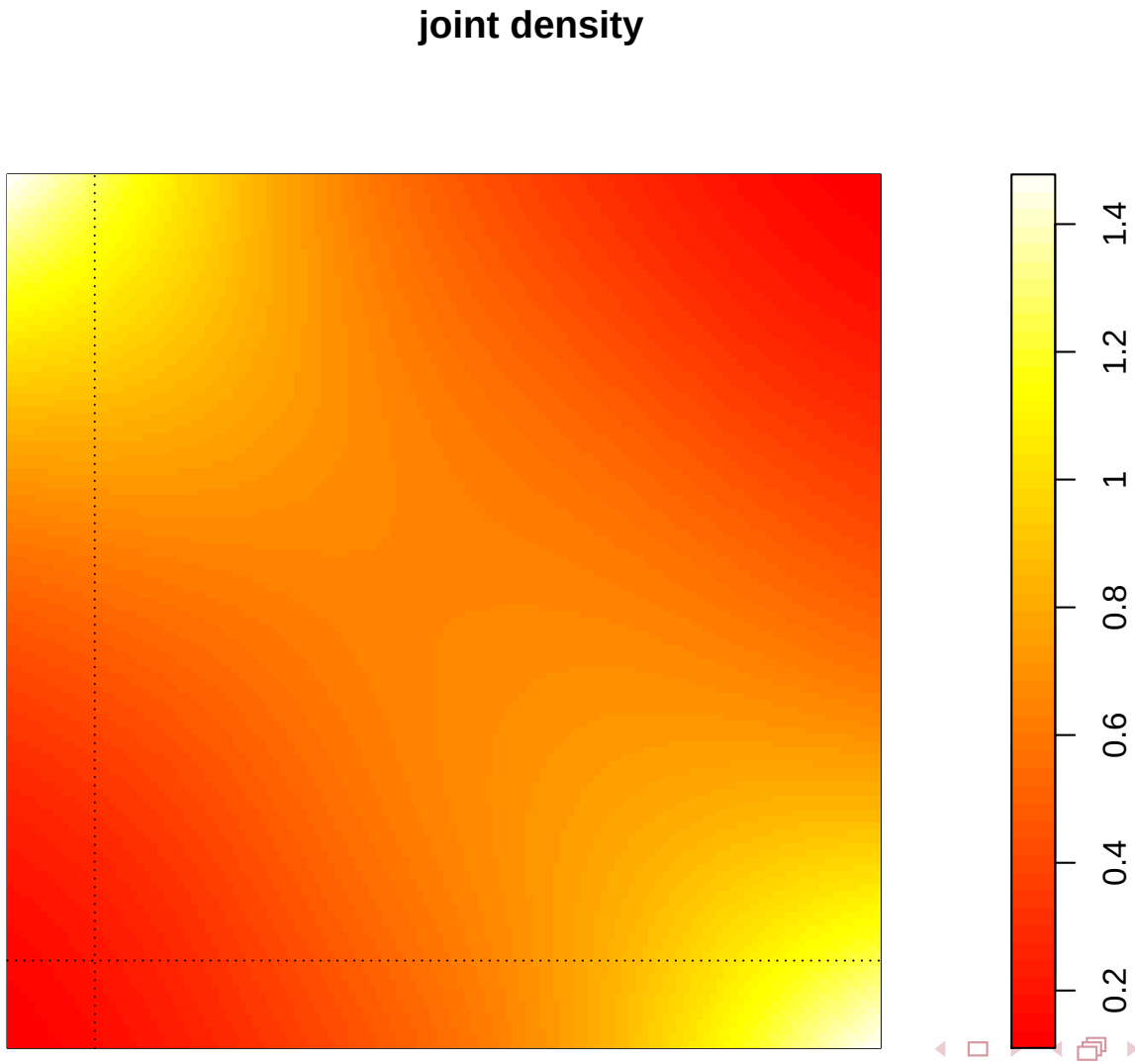
In most of the real cases:



Dependent tests

In most of the real cases:

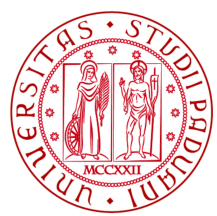
$P(\text{At least one false rejection}) > (!) 1 - (1 - \alpha)^2$



Type I errors

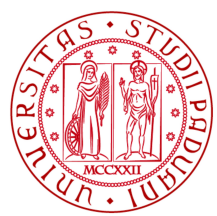
How to define the type I error when there are many hypotheses?

Which procedures control this error?



Outline

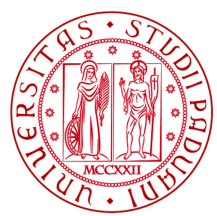
1. **Introduction**
   - Some examples
   - Some remarks
2. **FamilyWise Error Rate (FWER)**
   - Definition + Bonferroni
   - Holm (step-wise)
   - Closed Testing
3. **FWER control by Permutations tests**
   - Permutation Bonferroni
   - Westfall & Young: permutation Holm[4]
   - Closed Testing
4. **Conclusion**



[4] Westfall PH, Young SS (1993) Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment. Wiley.

Outline

1. **Introduction**
   - Some examples
   - Some remarks
2. **FamilyWise Error Rate (FWER)**
   - Definition + Bonferroni
   - Holm (step-wise)
   - Closed Testing
3. **FWER control by Permutations tests**
   - Permutation Bonferroni
   - Westfall & Young: permutation Holm[5]
   - Closed Testing
4. **Conclusion**



[5] Westfall PH, Young SS (1993) Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment. Wiley.

FamilyWise Error Rate (FWER)

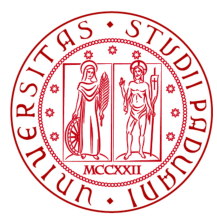
Probability of AT LEAST one false rejection

Bonferroni Inequality

Reject H_i if $p_i \leq \alpha/m$ ($m =$ number of hypotheses)

control FWER

$$
\begin{aligned}
\text{FWER} &= P\big(p_i \leq \alpha/m \text{ for at least one True null hypo}\big) \\
&= P\Big(\bigcup_{i \in \{\text{true null hypotheses}\}} \{p_i \leq \alpha/m\}\Big) \\
&\leq \sum_{i \in \{\text{true null hypotheses}\}} \mathrm{P}(p_i \leq \alpha/m) \\
&\leq \#\{\text{true null hypotheses}\}\frac{\alpha}{m} \leq \alpha
\end{aligned}
$$



Bonferroni procedure

Multiplicity-adjusted p-value

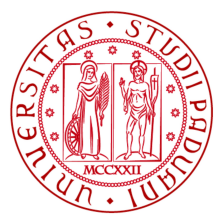
$\tilde{p}_i = mp_i$ $i = 1, \ldots, m$ and rejects it if $\leq \alpha$

Pros

- Very easy
- Control the FWER under any dependence

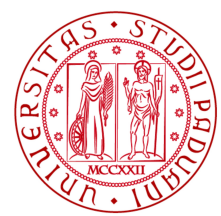
Cons

Conservative (Adj. P-values very high, few rejections)



Outline

1. **Introduction**
   - Some examples
   - Some remarks
2. **FamilyWise Error Rate (FWER)**
   - Definition + Bonferroni
   - Holm (step-wise)
   - Closed Testing
3. **FWER control by Permutations tests**
   - Permutation Bonferroni
   - Westfall & Young: permutation Holm[6]
   - Closed Testing
4. **Conclusion**



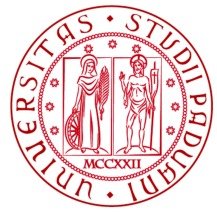
[6] Westfall PH, Young SS (1993) Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment. Wiley

Holm's procedures [7]

1. First step: adjusted p-value: $p \cdot m$; reject if $\leq \alpha$
2. After r rejections, adjusted p-value: $p \cdot (m - r)$
3. just do not reject anything, Stop.

Bonferroni

Adj. p-value:	$p_A 5$	$p_B 5$	$p_C 5$	$p_D 5$	$p_E 5$	$\leq ? \alpha$
$\mathcal{H} \setminus \mathcal{R}$:	A	B	C	D	E	
$\mathcal{R}$:						



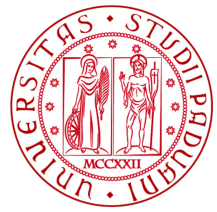
[7] Holm S. (1979) A simple sequentially rejective multiple test procedure.*Scandinavian Journal of Statistics*; 6 (2): 65-70.

Holm's procedures [7]

1. First step: adjusted p-value: $p \cdot m$; reject if $\leq \alpha$
2. After r rejections, adjusted p-value: $p \cdot (m - r)$
3. just do not reject anything, Stop.

Suppose p_A and p_C significant

Adj. p-value:	$p_A 5$	$p_B 5$	$p_C 5$	$p_D 5$	$p_E 5$	$\leq ? \alpha$
$\mathcal{H} \setminus \mathcal{R}$:	A	B	C	D	E	
$\mathcal{R}$:						



[7] Holm S. (1979) A simple sequentially rejective multiple test procedure.*Scandinavian Journal of Statistics*; 6 (2): 65-70.

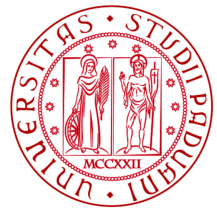
Holm's procedures [7]

1. First step: adjusted p-value: $p \cdot m$; reject if $\leq \alpha$
2. After r rejections, adjusted p-value: $p \cdot (m - r)$
3. just do not reject anything, Stop.

Adjusted p-value: $p \cdot 3$

Adj. p-value:	-	$p_B 3$	-	$p_D 3$	$p_E 3$	$\leq ?\alpha$
$\mathcal{H} \setminus \mathcal{R}$:		B		D	E	
$\mathcal{R}$:	A		C			

[7] Holm S. (1979) A simple sequentially rejective multiple test procedure.*Scandinavian Journal of Statistics*; 6 (2): 65-70.

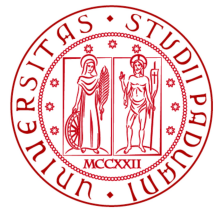


Holm's procedures [7]

1. First step: adjusted p-value: $p \cdot m$; reject if $\leq \alpha$
2. After r rejections, adjusted p-value: $p \cdot (m - r)$
3. just do not reject anything, Stop.

Suppose p_D significant

Adj. p-value:	-	$p_B 3$	-	$p_D 3$	$p_E 3$	$\leq ?\alpha$
$\mathcal{H} \setminus \mathcal{R}$:		B		D	E	
$\mathcal{R}$:	A		C			



[7] Holm S. (1979) A simple sequentially rejective multiple test procedure.*Scandinavian Journal of Statistics*; 6 (2): 65-70.

Holm's procedures [7]

1. First step: adjusted p-value: $p \cdot m$; reject if $\leq \alpha$
2. After r rejections, adjusted p-value: $p \cdot (m - r)$
3. just do not reject anything, Stop.

Adjusted p-value: $p \cdot 2$

Adj. p-value:	-	$p_B 2$	-	-	$p_E 2$	$\leq ? \alpha$
$\mathcal{H} \setminus \mathcal{R}$:		B			E	
$\mathcal{R}$:	A		C	D		

[7] Holm S. (1979) A simple sequentially rejective multiple test procedure.*Scandinavian Journal of Statistics*; 6 (2): 65-70.

Holm's procedures [7]

1. First step: adjusted p-value: $p \cdot m$; reject if $\leq \alpha$
2. After r rejections, adjusted p-value: $p \cdot (m - r)$
3. just do not reject anything, Stop.

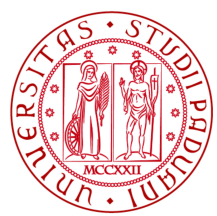
No rejections: Stop

Adj. p-value:	-	$p_B 2$	-	-	$p_E 2$ $\leq ? \alpha$
$\mathcal{H} \setminus \mathcal{R}$:		B			E
$\mathcal{R}$:	A		C	D	

[7] Holm S. (1979) A simple sequentially rejective multiple test procedure.*Scandinavian Journal of Statistics*; 6 (2): 65-70.

Outline

1. **Introduction**
   - Some examples
   - Some remarks
2. **FamilyWise Error Rate (FWER)**
   - Definition + Bonferroni
   - Holm (step-wise)
   - Closed Testing
3. **FWER control by Permutations tests**
   - Permutation Bonferroni
   - Westfall & Young: permutation Holm[8]
   - Closed Testing
4. **Conclusion**



[8] Westfall PH, Young SS (1993) Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment. Wiley.

Closed Testing[9]

Closure of hypos (all possible intersections)

hypos to test

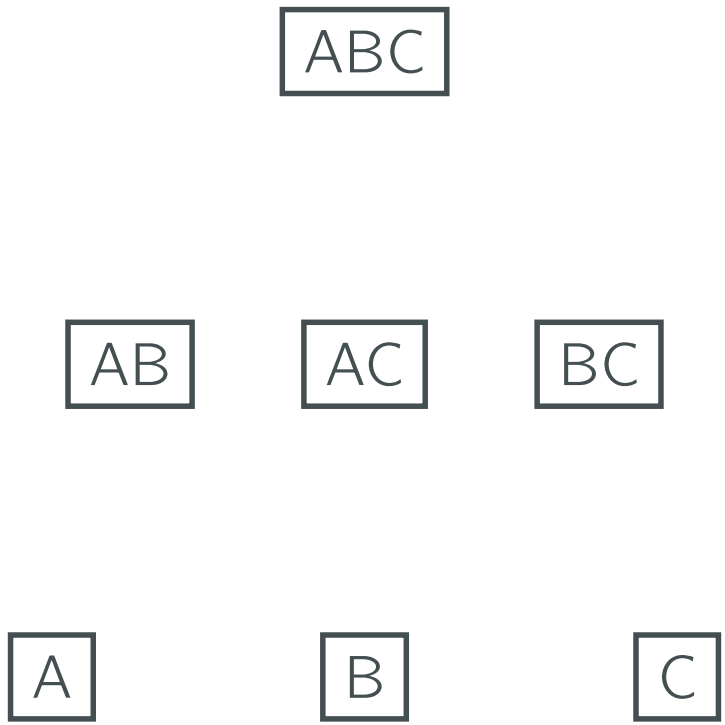
A B C

[9] R Marcus, E Peritz, KR Gabriel (1976). On closed testing procedures with special reference to ordered analysis of variance. Biometrika 63: 655-660.

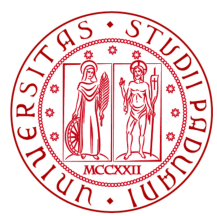
Closed Testing[9]

Test the overall null hypo (eg MANOVA)

Closure



[9] R Marcus, E Peritz, KR Gabriel (1976). On closed testing procedures with special reference to ordered analysis of variance. Biometrika 63: 655-660.



Closed Testing[9]

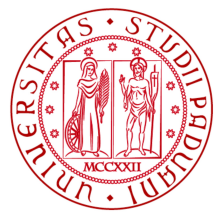
test at level α

ABC α

AB AC BC

A B C

[9] R Marcus, E Peritz, KR Gabriel (1976). On closed testing procedures with special reference to ordered analysis of variance. Biometrika 63: 655-660.



Closed Testing[9]

Suppose significant

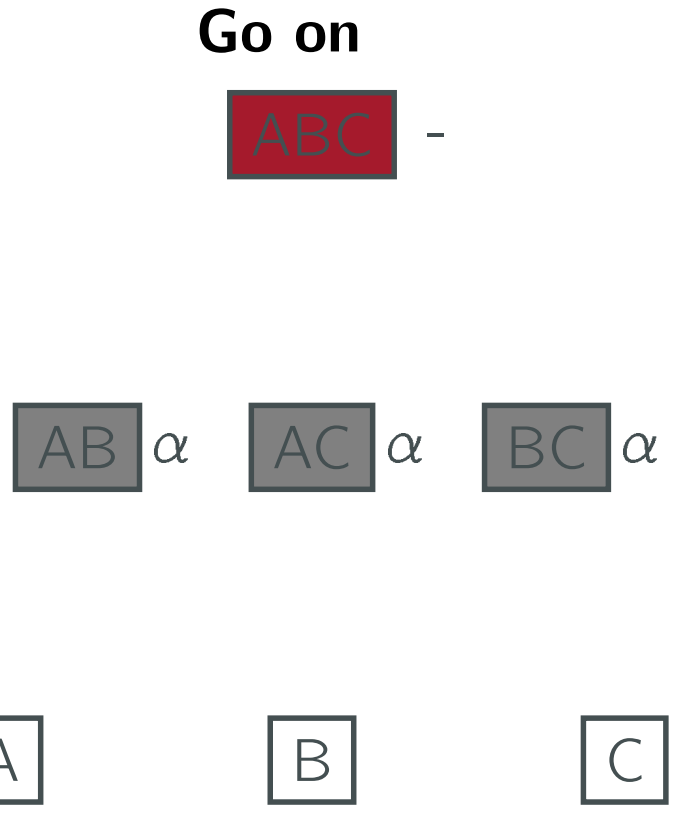
ABC -

AB AC BC

A B C

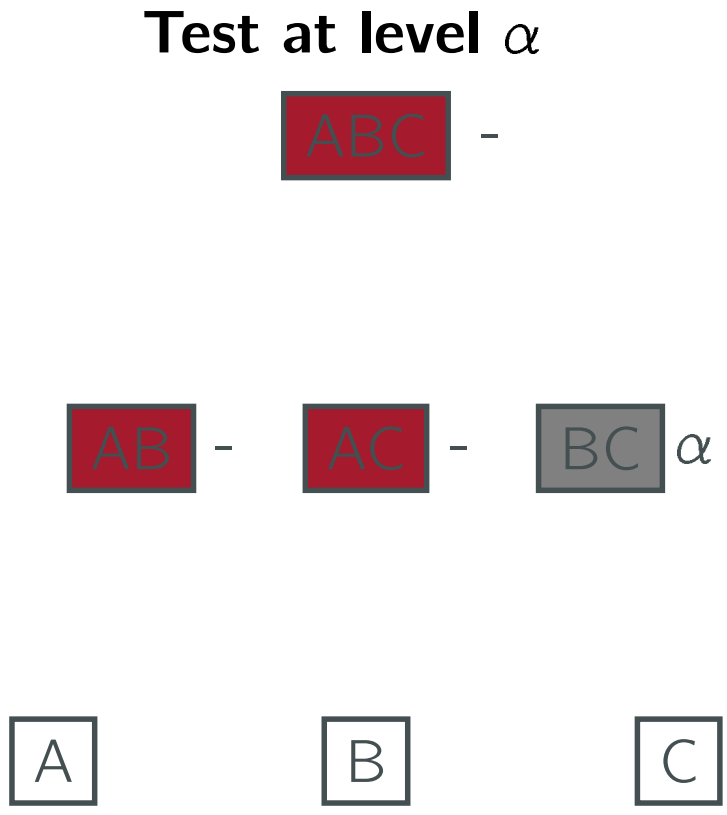
[9] R Marcus, E Peritz, KR Gabriel (1976). On closed testing procedures with special reference to ordered analysis of variance. Biometrika 63: 655-660.

Closed Testing[9]

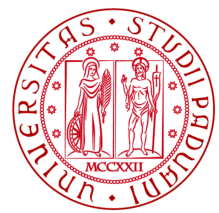


[9] R Marcus, E Peritz, KR Gabriel (1976). On closed testing procedures with special reference to ordered analysis of variance. Biometrika 63: 655-660.

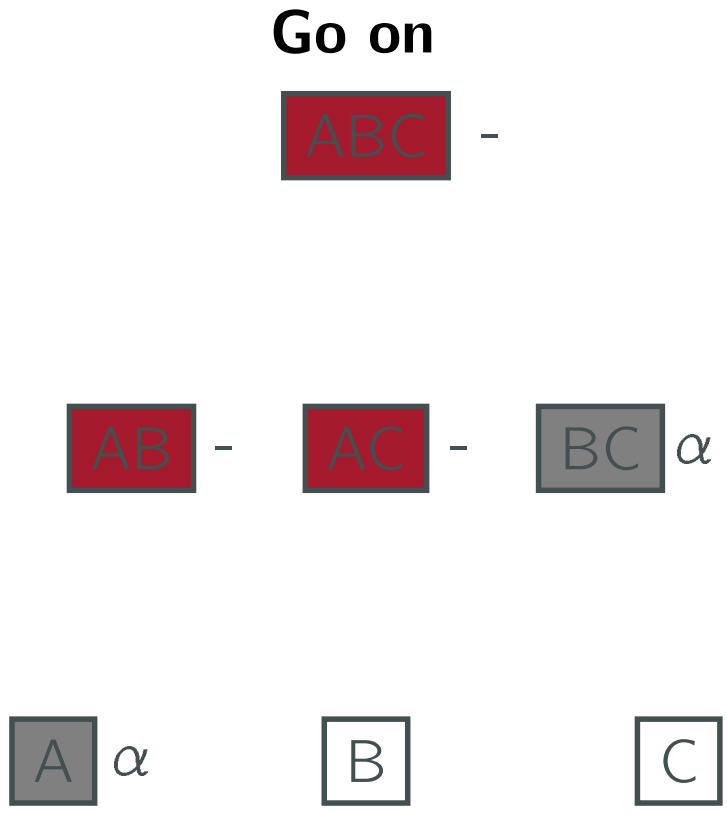
Closed Testing[9]



[9] R Marcus, E Peritz, KR Gabriel (1976). On closed testing procedures with special reference to ordered analysis of variance. Biometrika 63: 655-660.



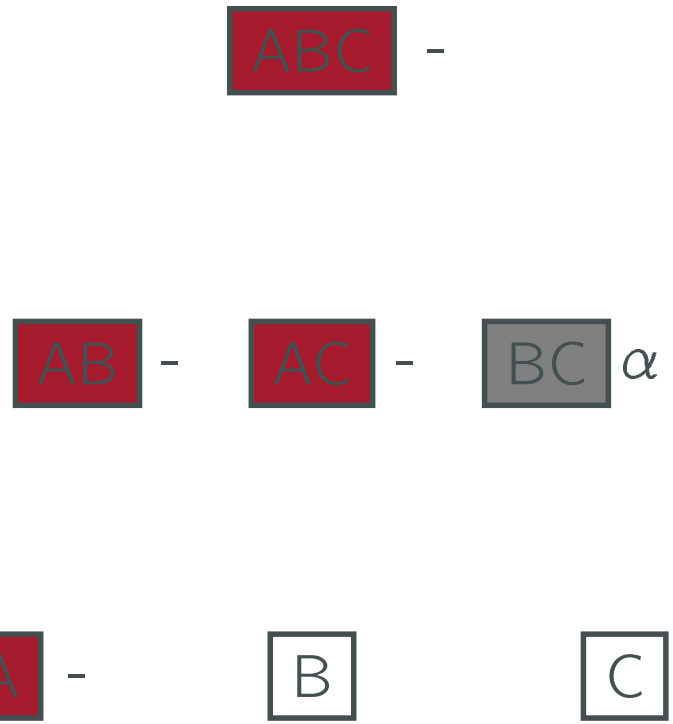
Closed Testing[9]



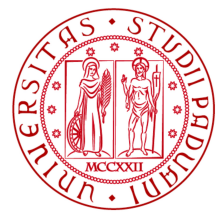
[9] R Marcus, E Peritz, KR Gabriel (1976). On closed testing procedures with special reference to ordered analysis of variance. Biometrika 63: 655-660.

Closed Testing[9]

A **is signif. if all above are signif.**



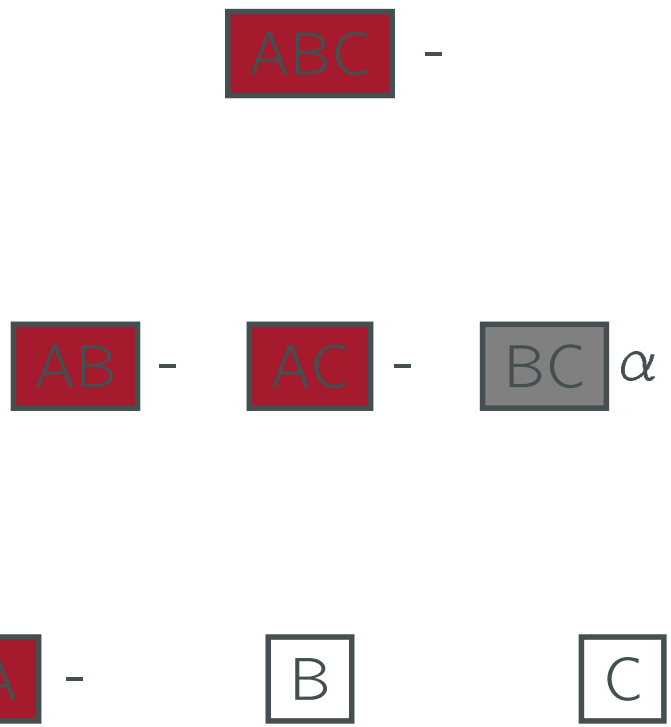
[9] R Marcus, E Peritz, KR Gabriel (1976). On closed testing procedures with special reference to ordered analysis of variance. Biometrika 63: 655-660.



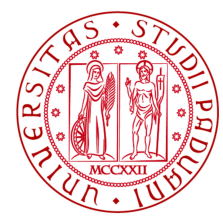
Closed Testing[9]

Cons: too much tests very quickly: $= 2^{\#hypos} - 1$

A is signif. if all above are signif.

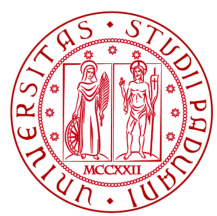


[9] R Marcus, E Peritz, KR Gabriel (1976). On closed testing procedures with special reference to ordered analysis of variance. Biometrika 63: 655-660.



Outline

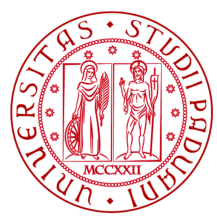
1. **Introduction**
   - Some examples
   - Some remarks
2. **FamilyWise Error Rate (FWER)**
   - Definition + Bonferroni
   - Holm (step-wise)
   - Closed Testing
3. **FWER control by Permutations tests**
   - Permutation Bonferroni
   - Westfall & Young: permutation Holm[10]
   - Closed Testing
4. **Conclusion**



[10] Westfall PH, Young SS (1993) Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment. Wiley

Outline

1. **Introduction**
   - Some examples
   - Some remarks
2. **FamilyWise Error Rate (FWER)**
   - Definition + Bonferroni
   - Holm (step-wise)
   - Closed Testing
3. **FWER control by Permutations tests**
   - Permutation Bonferroni
   - Westfall & Young: permutation Holm[11]
   - Closed Testing
4. **Conclusion**



[11] Westfall PH, Young SS (1993) Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment. Wiley.

Bonferroni is conservative

Bonferroni bound

Reject for p-values at most α/m

By Boole's inequality

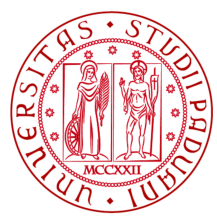
Guaranteed: FWER $\leq \alpha$, but often FWER $< \alpha$

Can we improve?

Reject for p-values at most $\tilde{\alpha} > \alpha/m$, while keeping FWER control

Yes we can

By permutations



Improved Bonferroni

Reduced α

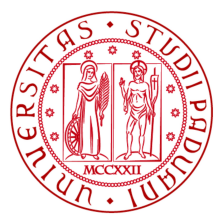
Reject H_i if $p_i \leq \tilde{\alpha}$

Control of FWER?

$$\begin{aligned} \text{FWER} &= \mathrm{P}\big(p_i \leq \tilde{\alpha} \text{ for at least one } i \text{ with } H_i \text{ true}\big) \\ &= \mathrm{P}\Big(\bigcup_{i \in T} \{p_i \leq \tilde{\alpha}\}\Big) \\ &= \mathrm{P}\Big(\min_{i \in T} p_i \leq \tilde{\alpha}\Big) \leq \alpha \end{aligned}$$

How can we determine the value of $\tilde{\alpha}$?

Using permutations to find the distribution of the minimum p-value



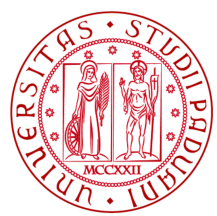
Multiple testing using permutations

The single step min-P method

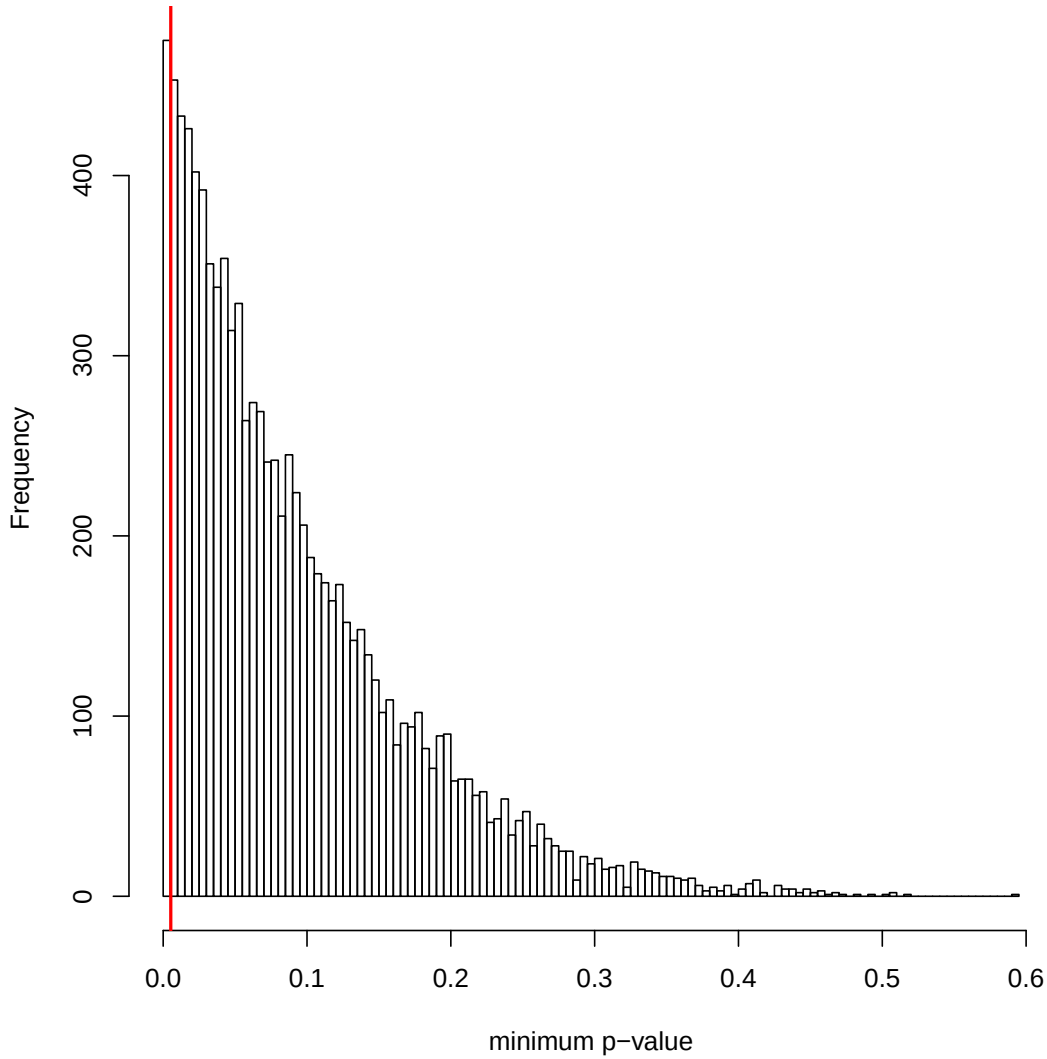
1. Calculate the smallest p-value m for the real data
2. Randomly permute the data
3. Calculate new p-values for all tests based on permuted data
4. Calculate the smallest p-value m^π for permuted data
5. Repeat permutation many (say k=1000) times: $m_1^\pi, \ldots, m_k^\pi$
6. Calculate $\tilde{\alpha}$ as the α-quantile of $m_1^\pi, \ldots, m_k^\pi$

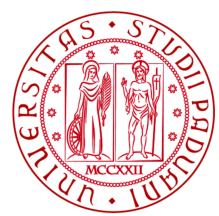
Multiple testing result

Reject all hypotheses with (non-permuted) p-values at most $\tilde{\alpha}$



Histogram of the smallest p-value of 10 independent tests





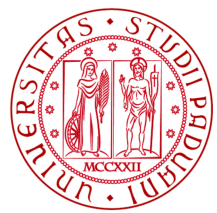
Correlation structure of p-values

Permutation

- Destroys correlation between covariates and response
- Retains correlation among covariates

Consequence

- P-values of correlated genes remain correlated in permutations
- Distribution of minimum p-value correctly takes correlations into account



Gain relative to Bonferroni

Min-P permutation critical value

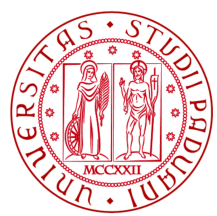
In the example data: $\tilde{\alpha} = 0.00527$ vs $\alpha/m = 0.005$

When is the gain large?

Negatively correlated p-values: typically no gain
Independent p-values: minimal gain
Positively correlated p-values: gain can be large

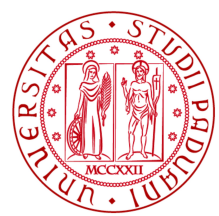
Genomic data

Genes typically strongly positively correlated



Outline

1. **Introduction**
   - Some examples
   - Some remarks
2. **FamilyWise Error Rate (FWER)**
   - Definition + Bonferroni
   - Holm (step-wise)
   - Closed Testing
3. **FWER control by Permutations tests**
   - Permutation Bonferroni
   - Westfall & Young: permutation Holm[12]
   - Closed Testing
4. **Conclusion**



[12] Westfall PH, Young SS (1993) Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment. Wiley

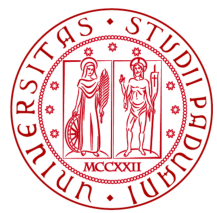
Sequential permutation multiple testing

Single step

Single step min-P is permutation equivalent of Bonferroni

What about Holm?

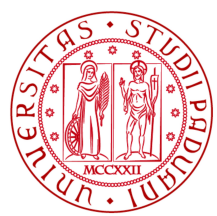
Permutation equivalent of Holm's method: Westfall & Young



Westfall & Young min-P

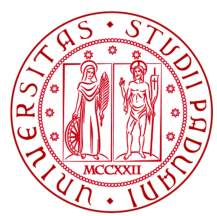
The min-P algorithm

- Start with all hypotheses
- Repeat
  - Do single step min-P to calculate $\tilde{\alpha}$
  - Reject hypotheses with p-value $\leq \tilde{\alpha}$
  - Remove rejected hypotheses
- Until no new rejections occur



Westfall & Young min-p (and max-t)

Westfall & Young min-p: shortcut using min-p combining function (m tests).

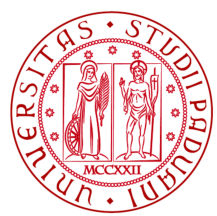


Westfall & Young min-p (and max-t)

Westfall & Young min-p: shortcut using min-p combining function (m tests).
Suppose three hypotheses tested and $p_A \leq p_B \leq p_C$

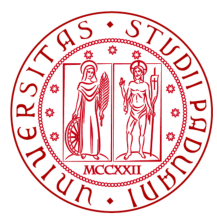
- Test H_A, H_B and H_C using min-p: p_{ABC}
  - if $p_{ABC} \leq \alpha$ reject H_A and go on
  - if $p_{ABC} > \alpha$ STOP
- Test H_B and H_C using min-p: p_{BC}
  - if $p_{BC} \leq \alpha$ reject H_B and go on
  - if $p_{BC} > \alpha$ STOP
- Test H_C: p_C
  - if $p_C \leq \alpha$ reject H_C and STOP
  - if $p_C > \alpha$ STOP

For max-t procedure substitute $p_{ABC} \leq \alpha$ with $t_{ABC} \geq t_\alpha$



Outline

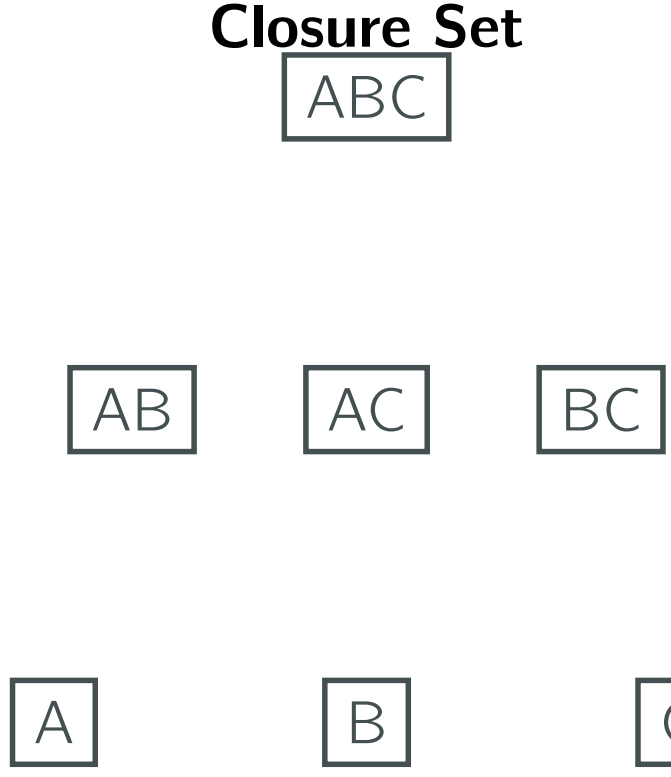
1. **Introduction**
   - Some examples
   - Some remarks
2. **FamilyWise Error Rate (FWER)**
   - Definition + Bonferroni
   - Holm (step-wise)
   - Closed Testing
3. **FWER control by Permutations tests**
   - Permutation Bonferroni
   - Westfall & Young: permutation Holm[13]
   - Closed Testing
4. **Conclusion**



[13] Westfall PH, Young SS (1993) Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment. Wiley.

Closed Testing[14]

Test in each node: any multivariate permutation test (eg alternative to Manova)

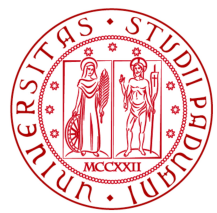


Adjusted $\tilde{p}_A = \max(p_A, p_{AB}, p_{AC}, p_{ABC})$

[14] R Marcus, E Peritz, KR Gabriel (1976). On closed testing procedures with special reference to ordered analysis of variance. Biometrika 63: 655-660.

Outline

1. **Introduction**
   - Some examples
   - Some remarks
2. **FamilyWise Error Rate (FWER)**
   - Definition + Bonferroni
   - Holm (step-wise)
   - Closed Testing
3. **FWER control by Permutations tests**
   - Permutation Bonferroni
   - Westfall & Young: permutation Holm[15]
   - Closed Testing
4. **Conclusion**



[15] Westfall PH, Young SS (1993) Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment. Wiley.

Accounting for dependences: adjusted p-value lower (i.e. more rejections)

When?

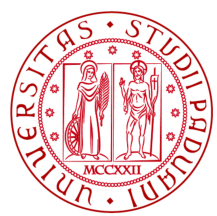
Negative correlation: generally no gain
p-value Independents: little or no gain
Positive correlation: big gain, usually
(NB: a test with bi-directional alternative and with negative correlation produce p-value positively correlated)

How?

in R: `library(flip); flip(); flip.adjust()`

Real data

The variables of real datasets are often correlated
then ...



Accounting for dependences: adjusted p-value lower (i.e. more rejections)

When?

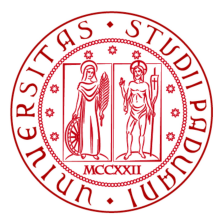
Negative correlation: generally no gain
p-value Independents: little or no gain
Positive correlation: big gain, usually
(NB: a test with bi-directional alternative and with negative correlation produce p-value positively correlated)

How?

in R: `library(flip); flip(); flip.adjust()`

Real data

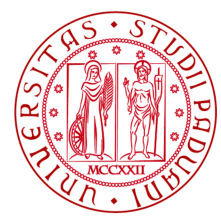
The variables of real datasets are often correlated
then . . . permutations are (often) convenient



Summary

Error FamilyWise

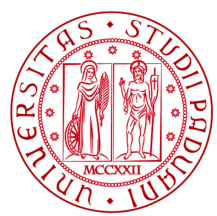
- Generalize the Type I error to the case of multiple hypotheses



Summary

Error FamilyWise

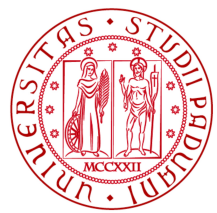
- Generalize the Type I error to the case of multiple hypotheses
- Control the probability of at LEAST a false rejections



Summary

Error FamilyWise

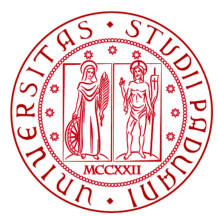
- Generalize the Type I error to the case of multiple hypotheses
- Control the probability of at LEAST a false rejections
- corrects the p-value (adjusted p-value always equal to or greather than the unadjusted p-value)



Summary

Error FamilyWise

- Generalize the Type I error to the case of multiple hypotheses
- Control the probability of at LEAST a false rejections
- corrects the p-value (adjusted p-value always equal to or greather than the unadjusted p-value)

R Software

- Bonferroni and Holm `library(stats); p.adjust()`
- Testing Closed `library(cherry); closed()`
- Permutations - Westfall & Young `library(flip); flip.adjust()`