Cherry-picking
Multiple Testing for Exploratory Research

Jelle Goeman Aldo Solari

Leiden University Medical Center

University of Milano-Bicocca

Journées de Statistique, 2012-05-24

A genomics data analysis result

Top 10 genes

Gene	p-value	multiplicity-corrected p-value
OCIAD2	5.5e-6	0.015
NEK3	6.7e-6	0.019
TAF5	7.1e-6	0.020
FOXD4L6	7.5e-6	0.021
ADIG	8.8e-6	0.025
ZNF19	1.3e-5	0.038
ERICH1	1.5e-5	0.044
SKP1	1.7e-5	0.050
GDF3	2.0e-5	0.059
CCDC25	2.0e-5	0.059
⋮	⋮	⋮

The empirical cycle

Confirmatory data analysis

- Limited number of research questions
- Research questions well-defined a priori
- Focus: strict error control
- Traditionally: (multiple) testing is important

Exploratory data analysis

- Many possible research questions
- Research questions not well-defined a priori
- Focus: finding promising research avenues
- Traditionally: (multiple) testing not so important

Microarray data analysis

More like exploratory than confirmatory research

- Probing many genes simultaneously
- Decision which questions are interesting taken a posteriori
- Findings are subject to follow up validation

Still: multiple testing performed

Reason: prevent unsuccessful validation experiments

Exploratory data analysis

Mild

It is not bad to select some true null hypotheses

Flexible

Procedures should not completely prescribe what to reject

Post hoc

Decide what/how much to follow up after seeing the data

Exploratory data analysis

Mild

It is not bad to select some true null hypotheses

Flexible

Procedures should not completely prescribe what to reject

Post hoc

Decide what/how much to follow up after seeing the data

Multiple testing in exploratory research

- Should sanction mild, flexible, post hoc inference
- Should advise, not prescribe

Set-up

Hypotheses

$H_1, \ldots, H_n$

True hypotheses

$T \subseteq \{1, \ldots, n\}$ indices of true hypotheses

Rejections

$R \subseteq \{1, \ldots, n\}$ set of rejected hypotheses (usually random)

Type I errors

$T \cap R \subseteq \{1, \ldots, n\}$

FWER, FDR, k-FWER

User role

Before seeing the data
Choose error rate to be controlled

$$\begin{aligned} \text{FWER:} &: \mathrm{P}(T \cap R = \emptyset) \\ \text{FDR} &: \mathrm{E}\left(\frac{\#(T \cap R)}{\#R \vee 1}\right) \\ \text{k-FWER} &: \mathrm{P}\big(\#(R \cap T) \geq k\big) \end{aligned}$$

Procedure

Chooses R that controls the chosen error rate

Alterative: exploratory inference

Role of the user

In complete freedom the user rejects collection of hypotheses R.

Alterative: exploratory inference

Role of the user

In complete freedom the user rejects collection of hypotheses R.

Role of the multiple testing procedure

Inform user of the number of false rejections incurred

Alterative: exploratory inference

Role of the user

In complete freedom the user rejects collection of hypotheses R.

Role of the multiple testing procedure

Inform user of the number of false rejections incurred

Number of false rejections

$= \#(T \cap R)$
$=$ function of the model parameters
$=$ something we can estimate or make a confidence interval for

Alterative: exploratory inference

Role of the user

In complete freedom the user rejects collection of hypotheses R.

Role of the multiple testing procedure

Inform user of the number of false rejections incurred

Number of false rejections

$= \#(T \cap R)$
$=$ function of the model parameters
$=$ something we can estimate or make a confidence interval for

Post hoc

If we make a simultaneous CI, post hoc choice of R is allowed

Closed Testing: ingredients

Marcus, Peritz and Gabriel (1976)

Fundamental principle of FWER control

Intersection hypothesis

$H_C = \bigcap_{i \in C} H_i$, for $C \subseteq \{1, \ldots, n\}$

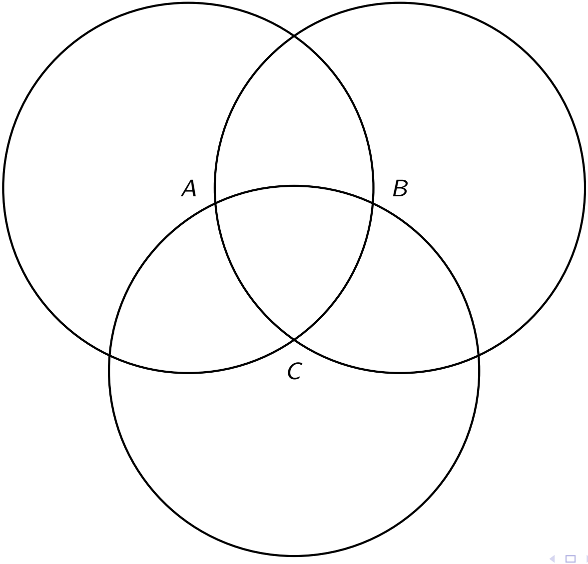
Closure

Collection of all intersection hypotheses
$\mathcal{C} = \{H_C : C \subseteq \{1, \ldots, n\}\}$

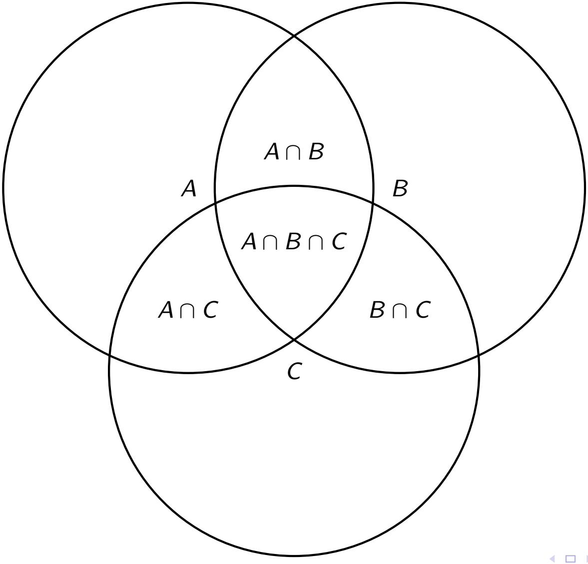
Local test

Valid α-level test for every intersection hypothesis

Closed testing (graphically)



Closed testing (graphically)



Closed testing: procedure

Raw rejections

Hypotheses $\mathcal{U} \subseteq \mathcal{C}$ rejected by the local test

Multiplicity-rejected rejections

Reject $H \in \mathcal{C}$ if $J \in \mathcal{U}$ for every $J \subseteq H$

Closed testing: procedure

Raw rejections

Hypotheses $\mathcal{U} \subseteq \mathcal{C}$ rejected by the local test

Multiplicity-rejected rejections

Reject $H \in \mathcal{C}$ if $J \in \mathcal{U}$ for every $J \subseteq H$

Statement

$$\mathrm{P}(\mathcal{R} \cap \mathcal{T} = \emptyset) \geq 1 - \alpha$$

with $\mathcal{R} = \{C \in \mathcal{C} : C \text{ rejected}\}$ and $\mathcal{T} = \{C \in \mathcal{C} : C \text{ true}\}$

Closed testing: procedure

Raw rejections

Hypotheses $\mathcal{U} \subseteq \mathcal{C}$ rejected by the local test

Multiplicity-rejected rejections

Reject $H \in \mathcal{C}$ if $J \in \mathcal{U}$ for every $J \subseteq H$

Statement

$$\mathrm{P}(\mathcal{R} \cap \mathcal{T} = \emptyset) \geq 1 - \alpha$$

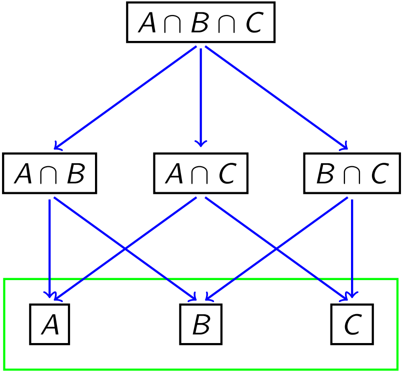
with $\mathcal{R} = \{C \in \mathcal{C} : C \text{ rejected}\}$ and $\mathcal{T} = \{C \in \mathcal{C} : C \text{ true}\}$

Proof

$$\{\mathcal{R} \cap \mathcal{T} = \emptyset\} \supseteq \{H_T \notin \mathcal{U}\}$$

Consonance

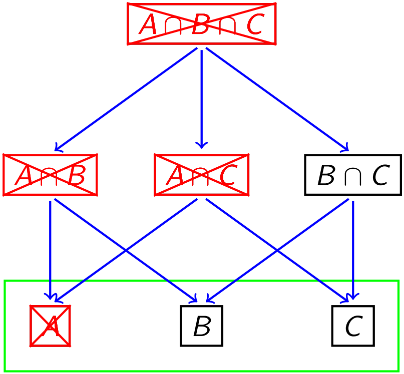
Traditionally, only rejection of elementary hypotheses is of interest



The closed graph of hypotheses A, B and C

Consonance

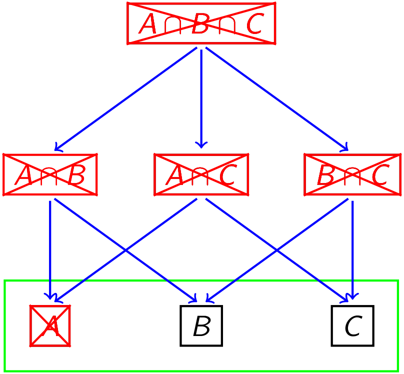
Traditionally, only rejection of elementary hypotheses is of interest



Consonant rejections

Consonance

Traditionally, only rejection of elementary hypotheses is of interest



Non-consonant rejection of $B \cap C$

Parameter, confidence bound and coverage

Parameter

$\tau(R) = \#(T \cap R)$ for a fixed set R

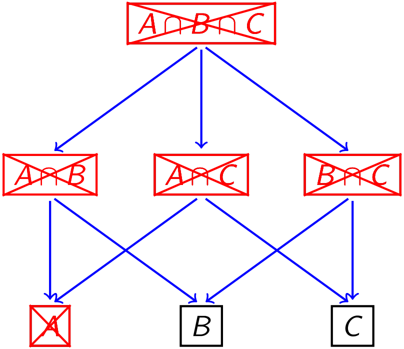
Closed testing

Let $\mathcal{X}$ be the collection of hypotheses rejected

Confidence bound

$t_\alpha(R) = \max(\#C : C \subseteq R, H_C \notin \mathcal{X}\}$

In the example



$t_\alpha(\{B, C\}) = 1$

Coverage

Coverage statement

$$\mathrm{P}(\tau(R) \leq t_\alpha(R)) \geq 1 - \alpha$$

Coverage

Coverage statement

$$P(\tau(R) \leq t_\alpha(R)) \geq 1 - \alpha$$

Proof

$$\{\tau(R) \leq t_\alpha(R)\} \subseteq \{H_T \notin \mathcal{U}\}$$

Coverage

Coverage statement

$$\mathrm{P}(\tau(R) \leq t_\alpha(R)) \geq 1 - \alpha$$

Proof

$$\{\tau(R) \leq t_\alpha(R)\} \subseteq \{H_T \notin \mathcal{U}\}$$

Confidence set

- Trivial lower bound $\tau(R) \geq 0$
- Confidence set $\{0, \ldots, t_\alpha(R)\}$
- Confidence set for $\phi(R) = \#R - \tau(R)$ immediate

Coverage

Coverage statement

$\mathrm{P}(\tau(R) \leq t_\alpha(R)) \geq 1 - \alpha$

Proof

$\{\tau(R) \leq t_\alpha(R)\} \subseteq \{H_T \notin \mathcal{U}\}$

Confidence set

- Trivial lower bound $\tau(R) \geq 0$
- Confidence set $\{0, \ldots, t_\alpha(R)\}$
- Confidence set for $\phi(R) = \#R - \tau(R)$ immediate

Simultaneous

- Simultaneous control over all R
- Consequence: coverage robust against post hoc selection of R

Reject hypotheses

R	confidence set for $\tau(R)$	confidence set for $\phi(R)$
$\{A\}$	{0}	{1}
$\{B\}$	{0,1}	{0,1}
$\{C\}$	{0,1}	{0,1}
$\{A, B\}$	{0,1}	{1,2}
$\{A, C\}$	{0,1}	{1,2}
$\{B, C\}$	{0,1}	{1,2}
$\{A, B, C\}$	{0,1}	{2,3}

Mild, post hoc, flexible

Mild

Allows a many or few false rejections as the researcher wants

Flexible

The researcher is completely free to choose the rejections

Post hoc

Allowed because confidence statements are simultaneous

Mild, post hoc, flexible

Mild

Allows a many or few false rejections as the researcher wants

Flexible

The researcher is completely free to choose the rejections

Post hoc

Allowed because confidence statements are simultaneous

Role of the multiple testing procedure

Advises rather than dictates

Selecting covariates

Which covariates are associated with a response?

Practice: typically exploratory and post hoc

Example

- Physical data set
- Response: mass of male subjects
- Covariates: length and circumference of body parts
- 10 covariates
- 22 subjects

Classical Forward/Backward (p-values)

covariate	full model	selected model
(Intercept)	0.036	0.000
Forearm	0.061	0.000
Biceps	0.755	—
Chest	0.420	—
Neck	0.518	—
Shoulder	0.905	—
Waist	0.000	0.000
Height	0.033	0.005
Calf	0.303	—
Thigh	0.351	0.036
Head	0.105	—

Closed testing applied

Model

Linear model, 10+1 regression coefficients

Local test

Test H_C with an F-test of $\beta_i = 0$ for $i \in C$ against saturated

Closed testing applied

Model

Linear model, 10+1 regression coefficients

Local test

Test H_C with an F-test of $\beta_i = 0$ for $i \in C$ against saturated

Result

626 out of 1023 intersection hypotheses rejected

Number of false nulls (95% conf)

- F/B set: at least one false null hypothesis
- All 10: at least two false null hypotheses

Defining sets

Defining sets

- Rejected intersection hypotheses without rejected supersets
- Each contains at least one false null

$\{waist\}$
$\{forearm, neck, shoulder, height\}$
$\{forearm, biceps, shoulder, calf\}$
$\{forearm, shoulder, height, calf\}$
$\{forearm, biceps, chest, neck, shoulder, thigh\}$
$\{forearm, shoulder, height, thigh\}$
$\{forearm, calf, thigh\}$

Shortlist

Shortlist (idea suggested by Nicolai Meinshausen)

- Rewrite defining sets to intersection of unions
- At least one of these sets contains only false nulls

$\{waist, forearm\}$
$\{waist, shoulder, calf\}$
$\{waist, shoulder, thigh\}$
$\{waist, neck, height, calf\}$
$\{waist, neck, calf, thigh\}$
$\{waist, height, biceps, calf\}$
$\{waist, height, calf, chest\}$
$\{waist, height, biceps, thigh\}$
$\{waist, height, calf, thigh\}$

Shortcuts

General

Procedure can be used for any local test

Number of intersection hypotheses

2^n: computationally hard above 20 hypotheses

Concept: shortcut

Smart choice of local test to save calculations

Extend shortcut concept

Should also yield rejected non-consonant intersections

Shortcuts based on Simes' inequality

Simes' inequality

If $p_1, \ldots, p_k$ from true hypotheses, then simultaneously $p_{(i)} > \frac{i\alpha}{k}$.

Use Simes as local test

Reject if any $p_{(i)} \leq \frac{i\alpha}{k}$
Allows shortcut with n^2 rather than 2^n calculations

Link with Benjamini & Hochberg FDR control

Same assumptions, same weak FWER control

Link with Hochberg and Hommel procedures

Rejects exactly the same as Hommel for FWER

The Rosenwald data

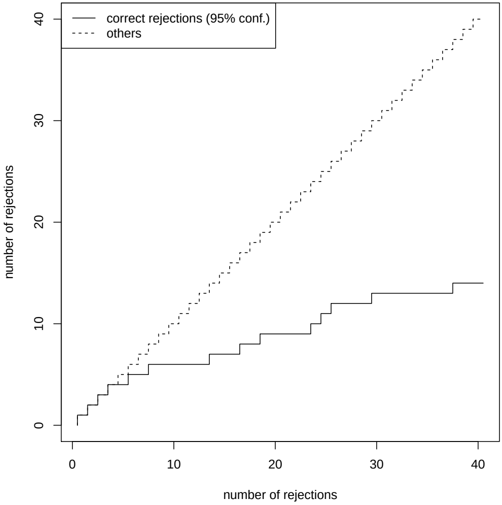
Rosenwald (2002)

- Diffuse Large B-cell Lymphoma
- Gene expression of 7399 probes
- Sample size of 240 patients
- Survival follow-up of median 8 years

Hypothesis tests

7399 likelihood ratio tests in the Cox model

Rejection curve (first part)



Package “cherry” on CRAN

General closed testing

- Specify any local test
- Up to 31 hypotheses

Special functions based on shortcuts

- Simes inequality
- Fisher combinations

Conclusion

New method

Between weak and strong FWER control

Nothing new

Just closed testing and simultaneous confidence sets

Suitable for exploratory research

Mild, post hoc, flexible

Read more?

Goeman and Solari (2011).
Multiple testing for exploratory research, with discussion
Statistical Science, **26** (4), 584-597.