Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Multiple Testing

Applied Multivariate Statistics – Spring 2012

Overview

- Problem of multiple testing
- Controlling the FWER:
  - Bonferroni
  - Bonferroni-Holm
- Controlling the FDR:
  - Benjamini-Hochberg
- Case study

Package repositories in R

- Comprehensive R Archive network (CRAN):
  - packages from diverse backgrounds
  - install packages using function “install.packages”
  - homepage: http://cran.r-project.org/

- Bioconductor:
  - biology context
  - download package manually, unzip, load into R using “library(…, lib.loc = ‘path where you saved the folder of the package’)”
  - homepage: http://www.bioconductor.org

- We are going to use the package “multtest” from Bioconductor

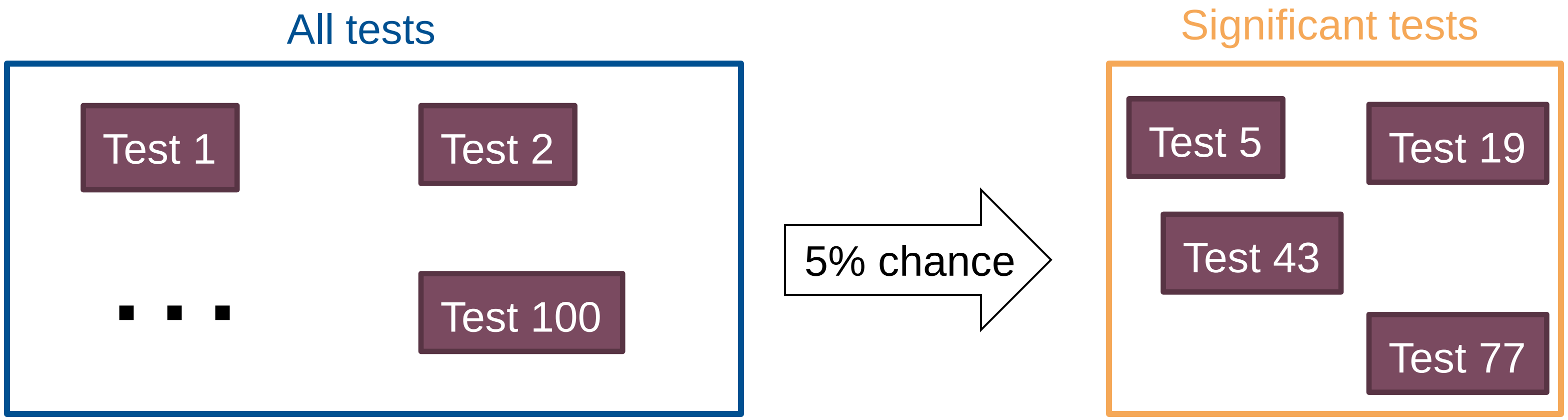
Example: Effect of “wonder-pill”

- Claim: Wonder pill has an effect!
- Random group of people
- Measure 100 variables before and after taking the pill: Weight, blood pressure, heart rate, blood parameters, etc.
- Compare before and after using a paired t-test for each variable on the 5% significance level

- Breaking news: 5 out of 100 variables indeed showed a significant effect !!

The problem of Multiple Testing

- Single test on 5% significance level:
  By definition, type 1 error is (at most) 5%
- Type 1 error: Reject H_0 if H_0 is actually true
  In example: Declare that wonder-pill changes variable, if in reality there is no change
- Let's assume, that wonder-pill has no effect at all.
  Then: Every variable has a 5% chance of being "significantly changed by the drug"
- Like a lottery: Nmb. Sign. Tests ~ Bin(100, 0.05)



All tests

Test 1

Test 2

...

Test 100

5% chance

Significant tests

Test 5

Test 19

Test 43

Test 77

Family Wise Error Rate (FWER)

- Family: Group of tests that is done
- FWER = Probability of getting at least one wrong significance (= one false positive test)
- $FWER = P(V \geq 1) \approx {}^{V}/_{M_0}$

	Declared non-sign.	Declared sign.	Total
True H_0	U	V	M_0
False H_0	T	S	M_1
Total	M-R	R	M

- Clinical trials: Food and Drug Administration (FDA) typically requires FWER to be less than 5%

FWER in example

- V: Number of incorrectly significant tests
- V ~ Bin(100, 0.05)
- $FWER = P(V \geq 1) = 1 - P(V = 0) = 1 \ - 0.95^{100} = 0.99$ (assuming independence among variables)
- We will most certainly have at least one false positive test!

Controlling FWER: Bonferroni Method

- “Corrects” p-values; only count a test as significant, if corrected p-value is less than significance level

- If you do M tests, reject each H_{0i} only if for the corresponding p-value P_i holds:

$$\mathrm{M} * P_i < \alpha$$

- FWER of this procedure is less or equal to α

- In example: Reject H_0 only if 100*p-value is less than 0.05

- Very conservative: Power to detect H_A gets very small

Example: Bonferroni

- P-values (sorted):
  $H_{0(1)}$: 0.005, $H_{0(2)}$: 0.011, $H_{0(3)}$: 0.02, $H_{0(4)}$: 0.04, $H_{0(5)}$: 0.13
- M = 5 tests; Significance level: 0.05
- Corrected p-value: 0.005*5 = 0.025 < 0.05: Reject $H_{0(1)}$
- Corrected p-value: 0.011*5 = 0.055: Don’t reject $H_{0(2)}$
- Corrected p-value: 0.02*5 = 0.1: Don’t reject $H_{0(3)}$
- Corrected p-value: 0.04*5 = 0.2: Don’t reject $H_{0(4)}$
- Corrected p-value: 0.13*5 = 0.65: Don’t reject $H_{0(5)}$

- Conclusion:
  Reject $H_{0(1)}$, don’t reject $H_{0(2)}$, $H_{0(3)}$, $H_{0(4)}$, $H_{0(5)}$

Improving Bonferroni: Holm-Bonferroni Method

- “Corrects” p-values; only count a test as significant, if corrected p-value is less than significance level

- Sort all M p-values in increasing order: $P_{(1)}$, …, $P_{(M)}$
  $H_{0(i)}$ denotes the null hypothesis for p-value $P_{(i)}$
- Multiply $P_{(1)}$ with M, $P_{(2)}$ with M-1, etc.
- If $P_{(i)}$ smaller than the cutoff 0.05, reject $H_{0(i)}$ and carry on
  If at some point $H_{0(j)}$ can not be rejected, stop and don’t reject $H_{0(j)}$, $H_{0(j+1)}$, …, $H_{0(M)}$

- FWER of this procedure is less or equal to α
- Method “Holm” has never worse power than “Bonferroni” and is often better; still conservative

Example: Holm-Bonferroni

- P-values:
  $H_{0(1)}$: 0.005, $H_{0(2)}$: 0.011, $H_{0(3)}$: 0.02, $H_{0(4)}$: 0.04, $H_{0(5)}$: 0.13
- M = 5 tests; Significance level: 0.05
- Corrected p-value: 0.005*5 = 0.025 < 0.05: Reject $H_{0(1)}$
- Corrected p-value: 0.011*4 = 0.044 : Reject $H_{0(2)}$
- Corrected p-value: 0.02*3 = 0.06: Don't reject $H_{0(3)}$ and stop

- Conclusion:
  Reject $H_{0(1)}$ and $H_{0(2)}$, don't reject $H_{0(3)}$, $H_{0(4)}$, $H_{0(5)}$

False Discovery Rate (FDR)

- Controlling FWER is extremely conservative
  We might be willing to accept A FEW false positives
- FDR = Fraction of “false significant results” among the significant results you found
- $FDR = {}^{V}/_{R}$

	Declared non-sign.	Declared sign.	Total
True H_0	U	V	M_0
False H_0	T	S	M_1
Total	M-R	R	M

- FDR = 0.1 oftentimes acceptable for screening

Controlling FDR: Benjamini-Hochberg

- “Corrects” p-values; only count a test as significant, if corrected p-value is less than significance level

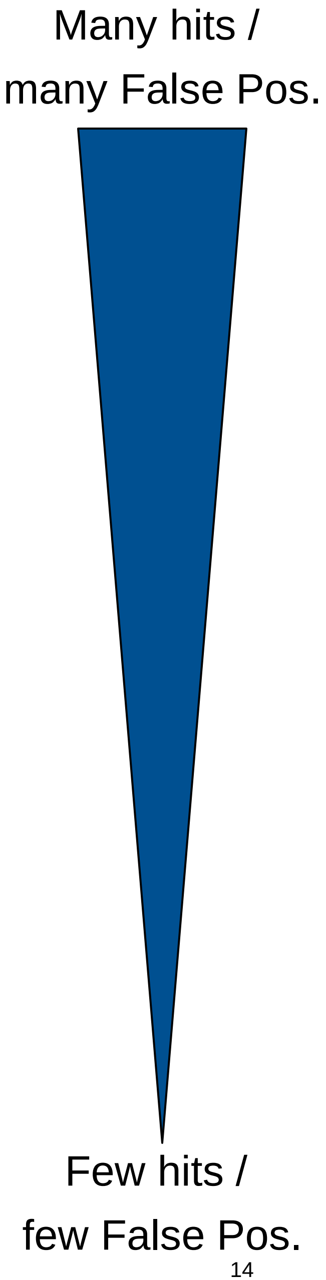
- Method a bit more involved; sequential as Holm-Bonferroni

Correcting for Multiple Testing in R

- Function “mt.rawp2adjp” in package “multtest” from Bioconductor
- Use option “proc”:
  - Bonferroni: “Bonferroni”
  - Holm-Bonferroni: “Holm”
  - Benjamini-Hochberg: “BH”

When to correct for multiple testing?

- **Don’t correct:**
  Exploratory analysis; when generating hypothesis
  Report the number of tests you do
  (e.g.: “We investigated 40 features, but only report on 10; 7 of those show a significant difference.”)
- **Control FDR (typically FDR < 10%):**
  Exploratory analysis; Screening: Select some features for further, more expensive investigation
  Balance between high power and low number of false positives
- **Control FWER (typically FWER < 5%):**
  Confirmatory analysis; use if you really don’t want any false positives



Case study: Detecting Leukemia types

- 38 tumor mRNA samples from one patient each:
  27 acute lymphoblastic leukemia (ALL) cases (code 0)
  11 acute myeloid leukemia (AML) cases (code 1)
- Expression of 3051 genes for each sample

- Which genes are associated with the different tumor types?

Concepts to know

- When to control FWER, FDR
- Bonferroni, Holm-Bonferroni, Benjamini-Hochberg

R functions to know

- “mt.rawp2adjp” in Bioconductor package “multtest”

Online Resources

- http://www.bioconductor.org/packages/release/bioc/html/multtest.html
- There: Section “Documentation”
- “multtest.pdf”: Practical introduction to multtest-package
- “MTP.pdf”: Theoretical introduction to multiple testing