UNIVERSITÉ DU LUXEMBOURG

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE
Multiomics Data Science

BIOSTATISTICS

Lecture 9

Correction for Multiple Testing

Petr Nazarov

Email: petr.nazarov@lih.lu
Skype: pvn.public

16-04-2021

UNIVERSITÉ DU LUXEMBOURG

OUTLINE

Lecture 8

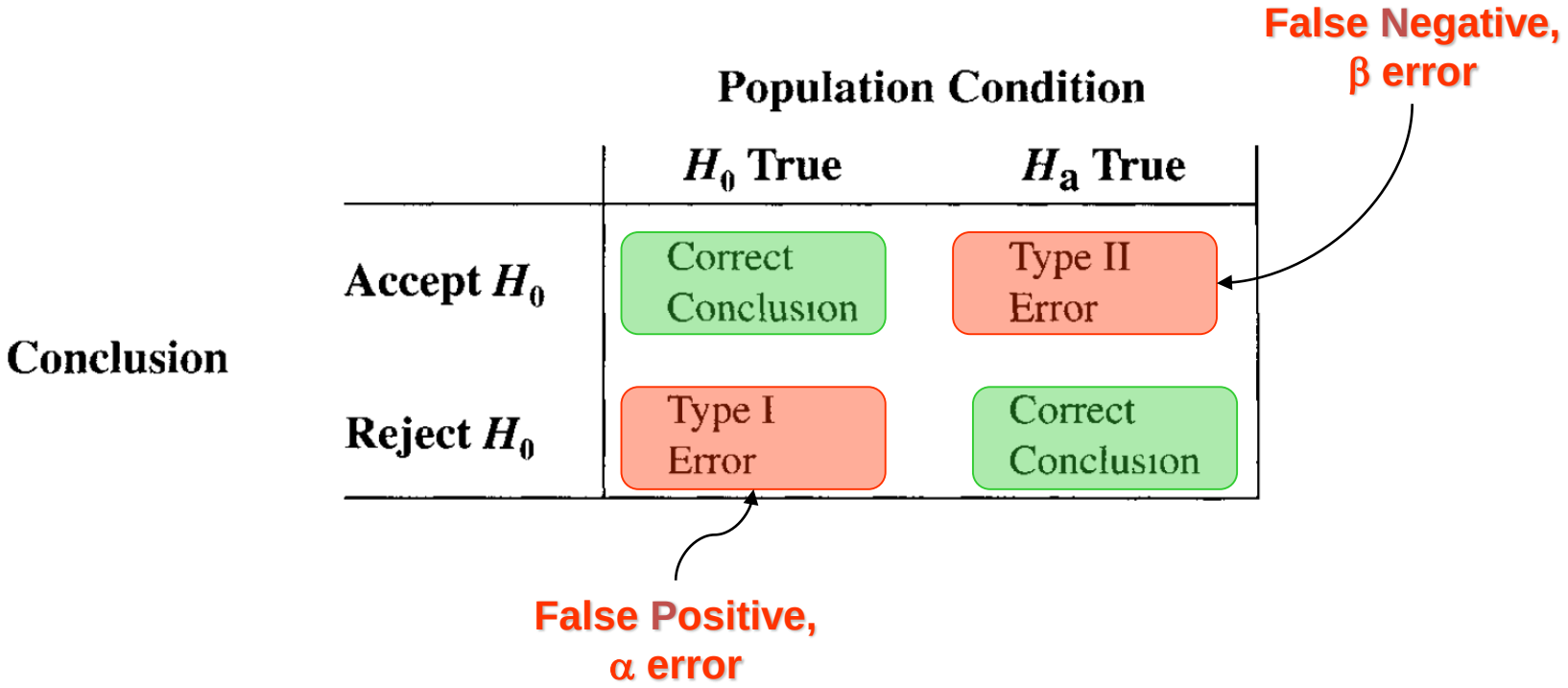
PART I → Test

PART II

- **Multiple testing problem**
- **False discovery rate**

MULTIPLE TESTING

Correct Results and Errors



Probability of an error in a multiple test:

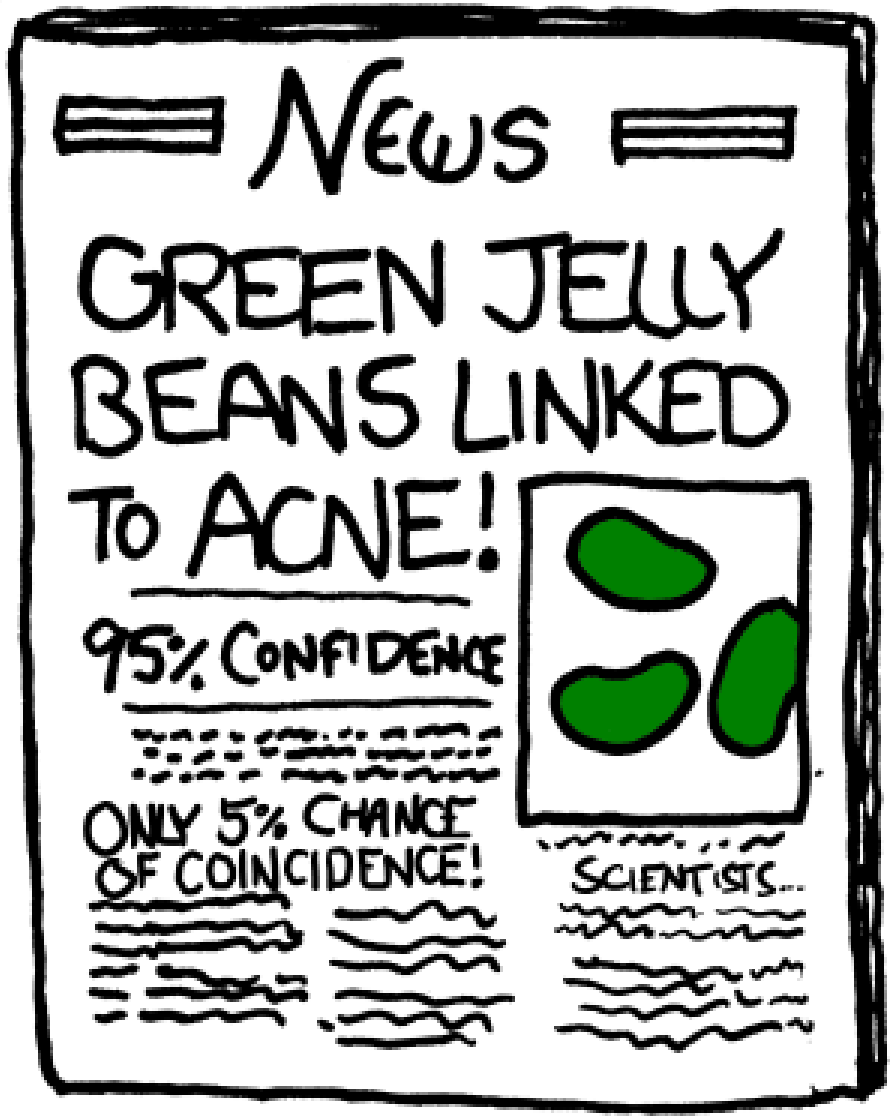
$1-(0.95)^{\text{number of comparisons}}$

n.o.c = 10 → p(error) = 0.4

n.o.c=100 → p(error) = 0.99

MULTIPLE TESTING

Example



http://www.xkcd.com/882/

MULTIPLE TESTING

False Discovery Rate

False discovery rate (FDR)

FDR control is a statistical method used in multiple hypothesis testing to correct for multiple comparisons. In a list of rejected hypotheses, FDR controls the expected proportion of incorrectly rejected null hypotheses (type I errors).

Conclusion	Population Condition: H_0 is TRUE	H_0 is FALSE	Total
Accept H_0 (non-significant)	U	T	$m - R$
Reject H_0 (significant)	V	S	R
Total	m_0	$m - m_0$	m

$$FDR = E\left(\frac{V}{V+S}\right)$$

MULTIPLE EXPERIMENTS

False Discovery Rate

Assume we need to perform *m* = 100 comparisons, and select maximum **FDR = α = 0.05**

Independent tests

The Simes procedure ensures that its expected value $\mathrm{E}\left[\frac{V}{V+S}\right]$ is less than a given α (Benjamini and Hochberg 1995). This procedure is valid when the m tests are independent. Let $H_1 \ldots H_m$ be the null hypotheses and $P_1 \ldots P_m$ their corresponding p-values. Order these values in increasing order and denote them by $P_{(1)} \ldots P_{(m)}$. For a given α, find the largest k such that $P_{(k)} \leq \frac{k}{m}\alpha.$

Then reject (i.e. declare positive) all $H_{(i)}$ for $i = 1, \ldots, k$.

Note that the mean α for these m tests is $\frac{\alpha(m+1)}{2m}$ which could be used as a rough FDR, or RFDR, "α adjusted for m indep. tests." The RFDR calculation shown here provides a useful approximation and is not part of the Benjamini and Hochberg method; see AFDR below.

MULTIPLE EXPERIMENTS

False Discovery Rate

Assume we need to perform m = 100 comparisons, and select maximum **FDR = α = 0.05**

k – is rank of p-value (order #)

$$FDR = E\left(\frac{V}{V+S}\right)$$

Expected value for FDR < α if $P_{(k)} \le \frac{k}{m}\alpha$

Benjamini-Hochberg (FDR)

Benjamini-Hochberg's FDR →

$$\frac{mP_{(k)}}{k} \le \alpha$$

Other Methods

Bonferroni – simple, but too stringent, not recommended $mP_{(k)} \le \alpha$

Holm – a more powerful and less stringent version of Bonferroni (ok)

$$(m-k+1)P_{(k)} \le \alpha$$

MULTIPLE EXPERIMENTS

Example: Random Data

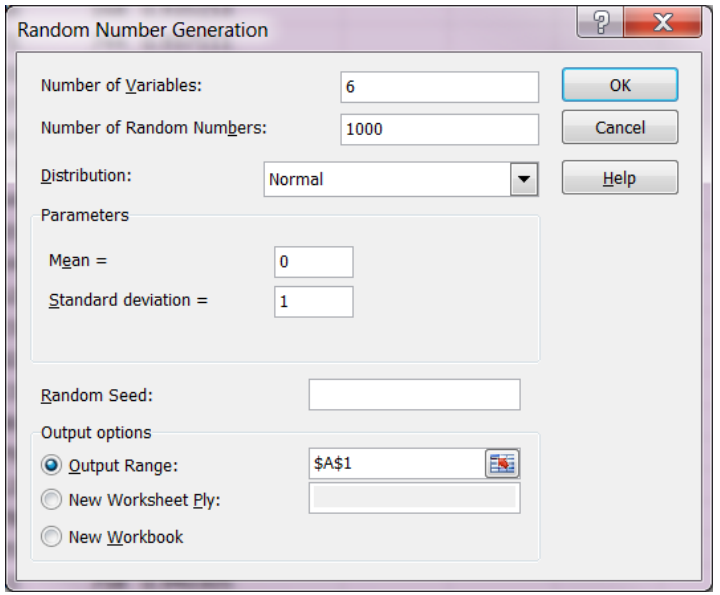
- Generate 6 columns of normal random variables (1000 points/candidates in each).
- Consider the first 3 columns as “treatment”, and the next 3 columns as “control”.
- Using t-test calculate p-values b/w “treatment” and “control” group. How many candidates have p-value<0.05 ?
- Calculate FDR. How many candidates you have now?

Candidates. 5% are false

Same candidates. Just sorted

Top 5% selected

???

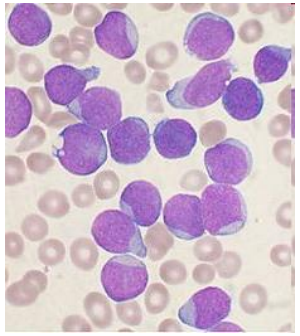


MULTIPLE EXPERIMENTS

Example: Acute Lymphoblastic Leukemia

http://edu.modas.lu/data

`all_data.xls`



Acute lymphoblastic leukemia (ALL), is a form of leukemia, or cancer of the white blood cells characterized by excess lymphoblasts.

all_data.xls contains the results of full-trancript profiling for ALL patients and healthy donors using Affymetrix microarrays. The data were downloaded from ArrayExpress repository and normalized. The expression values in the table are in $\log_2$ scale.

Let us analyze these data:

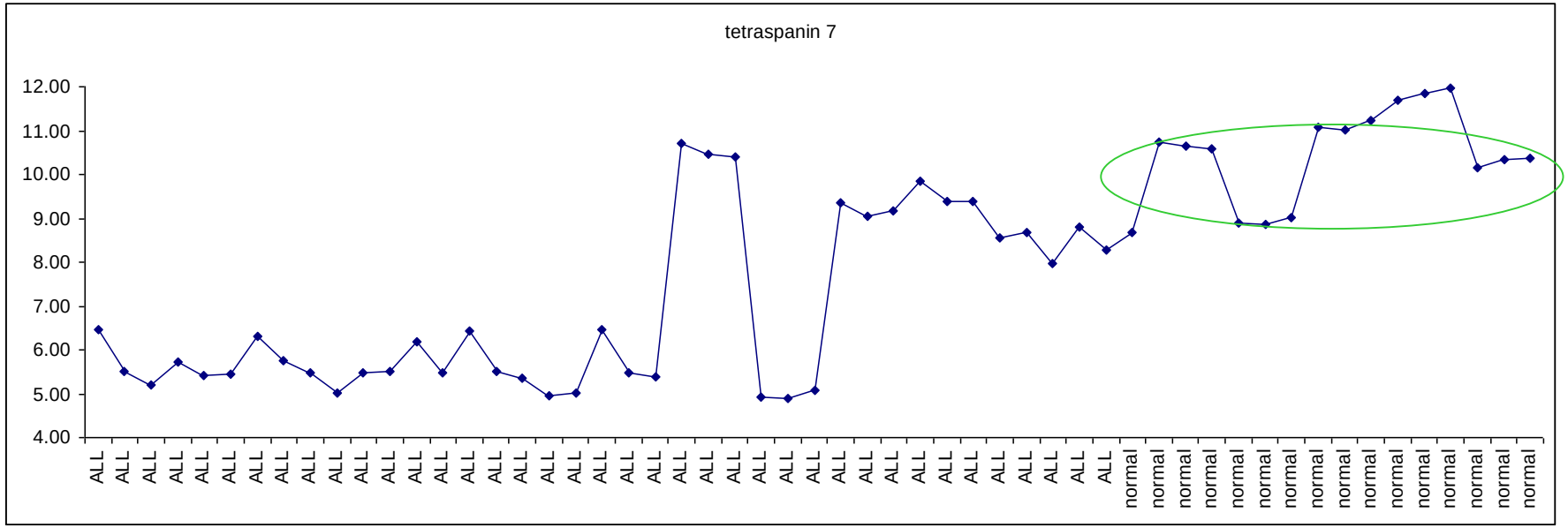
- Calculate log-ratio (logFC) for each gene
- Calculate the p-value based on t-test for each gene
- Perform the FDR-based adjustment of the p-value.

  Calculate the number of up and down regulated genes with FDR<0.01
- How would you take into account logFC?

Example score: $score = -\log(adj.p.value) \cdot |logFC|$

FDR (adj. p-value) is a main measure. Other only help…

MULTIPLE EXPERIMENTS



look for "tetraspanin 7" + leukemia in google ☺

Results are never perfect… Deeper investigation of the ALL subgroup should be recommended.

QUESTIONS ?

Thank you for your attention