Multiple Testing

FGCZ Protein Informatics Training

Witold Wolski wew@fgcz.ethz.ch

25 February, 2021

Multiplicity

Overview

- Traditional vs OMICS hypothesis testing
- Where does multiplicity arise
- What is Family Wise Error Rate (FWER)
- What is FDR

Traditional vs OMICS hypothesis testing

Testing hypothesis for a single observation

- Does diet impact weight of person
- Is treatment preventing infection with COVID?
- Does treatment changes the abundance of a specific protein?
- Does the proteome of the muscle tissue and the liver tissue differ?

OMICS experiments

- Which of the many proteins are differentially expressed because of treatment?
- Which of the many groups of proteins are differentially expressed because of treatment?

Weight loss example

Experiment:

- 250 subjects chosen "randomly".
- Diet for 1 week.
- Repeated Measurement (Data in kg.):
  - Weight at the start of the week
  - Weight at the end of week.

	n	Mean	StDev	SE Mean
Weight before	250	58.435	12.628	0.799
Weight after	250	58.309	12.636	0.799
Difference	250	0.126	1.081	0.068

Average weight loss is 0.13kg. Paired t-test for weight loss gives a t-statistic of $t = 0.126/0.068 = 1.84$, giving a p-value of 0.067 (using a two-sided test).
Not quite significant at the 5 level!

Example from Andersen (1990)

Weight loss example

```r
2*(1 - pt(1.84, df = 250 - 1))
```

```
## [1] 0.06695843
```

```r
2*(1 - pnorm(1.84, 0,1))
```

```
## [1] 0.06576824
```

```r
# Asymptotic test
(1 - pt(1.84, df = 250 - 1))
```

```
## [1] 0.03347921
```

```r
# one sided tests
```

	n	Mean	StDev	SE Mean
Weight before	250	58.435	12.628	0.799
Weight after	250	58.309	12.636	0.799
Difference	250	0.126	1.081	0.068

Assymptotic test - does not "help" (and is biased); smaller sample size larger bias.

Weight loss example

Why is the 1-sided test not acceptable?

You use a one-tailed test to improve the test's ability to learn whether the new diet is better.

However, that's unethical because the test cannot determine whether it is less effective. You risk missing valuable information by testing in only one direction.

Weight loss example

Can anything be done to get a significant result out of this study?

- Look at subgroups of the data by their sign of the zodiac. (additional factor)
- *12* instead of *1* test

- Conclusion: Those born under the sign of Aries are particularly suited to this new dietary control.

Mean Weight Loss by Sign of the Zodiac

Zodiac sign	n	mean weight loss	SE(mean)	t	p-value	
Aquarius	26	0.313	0.217	1.44	0.161	
Aries	15	0.543	0.205	2.65	0.019	**
Cancer	21	0.271	0.249	1.09	0.289	
Capricorn	27	-0.191	0.222	-0.86	0.397	
Gemini	18	0.068	0.266	0.26	0.801	
Leo	22	0.194	0.234	0.83	0.416	
Libra	26	0.108	0.217	0.50	0.623	
Pisces	19	0.362	0.232	1.56	0.136	
Sagittarius	12	0.403	0.294	1.37	0.197	
Scorpio	20	0.030	0.274	0.11	0.248	
Taurus	22	-0.315	0.183	-1.72	0.099	?
Virgo	22	0.044	0.238	0.18	0.955	

Weight loss example

What is the problem of this approach?

- Hypothesis that Arieans are good dieters was suggested by the fact that it gave an *apparently* significant result.
- By increasing the number of tests you increase the chance of false positive results (type I error).

Where does multiplicity arise

- **Multiple endpoints**
  - many outcome measures to asses an intervention.
    In mass spectrometry: *MS1 intensity and MS2 intensity (DIA).*
  - Solution : choose primary outcome, adjust p-values, multivariate analysis.
- **Interim Analysis**
  - analyse the data from a trial *periodically* as it becomes available
  - Solution: adjust p-values
- **Multiple Regression**
  - regression analysis involving many explanatory variables
  - Solution: Use background knowledge to suggest possible models, include only few interaction terms, adjust p-values.

Where does multiplicity arise

- **Repeated measures**
  - e.g. protein abundance at intervals of 1, 3, 6, 12 and 24 hours after ingestion of a drug.
  - Solution:
    - two-sample t-tests at each time point in sequence, e.g (3 vs 1, 6 vs 3 etc.) then adjust p.value
    - use summary measure (e.g. fit line and test line coefficients)
- **Subgroup comparison**
  - Samples are subdivided on baseline factors : gender, age-groups, sign of zodiac
  - Solution :
    - adjust p-values
    - ANOVA analysis to test factors

Example : 50 Control samples, 50 Treatment, no significant result. Split data into female and male, and young and old group. Four tests, instead of one and maybe one is significant.

Types of error when testing hypothesis

A **type I error** (false positive) occurs when
the null hypothesis (H0) is true, but is rejected.
The *type I error rate* or **significance level** (p-Value)
is the probability of rejecting the
null hypothesis given that it is true.

A **type II error** (false negative) occurs when
the null hypothesis is false,
but erroneously fails to be rejected.
The *the type II error rate* is denoted by the Greek letter β
and is related to the **power of a test** (which equals $1 - \beta$).

For a given test, the only way to reduce both error rates
is to **increase the sample size**, and this may not be feasible.

		reality	
		H_0 = true	H_0 = false
conclusion	H_0 is not rejected	OK	type II error
	H_0 is rejected	type I error	OK

Family-wise error rate (FWER)

In statistics, family-wise error rate (FWER) is the probability of making one or more false discoveries, or type I errors when performing multiple hypotheses tests.

If multiple hypotheses are tested, the chance of a rare event increases, and therefore, the likelihood of incorrectly rejecting a null hypothesis (i.e., making a type I error) increases.

P-value adjustment - Bonferroni correction

The Bonferroni correction compensates for that increase by testing each individual hypothesis at a significance level of $\epsilon = \alpha/k$, α is the desired overall size of test and k is the number of hypotheses.

$$k = 20;\ \alpha = 0.05; \qquad \epsilon = 0.05/20 = 0.0025$$

Bonferroni adjustments are typically very conservative (it assumes that the tests are independent - however they are frequently correlated) and more complex methods are usually used.

- R function `p.adjust` transforms the p-values (makes them larger) instead transforming the threshold.

wikipedia, Medical Statistics - Sheffield University 2018

P-value adjustment - Bonferroni correction

Family Wise Error Rate (FWER) - control the probability of at least one Type I error.

$$
\begin{aligned}
Pr(\text{at least one Type I error}|H_0) = \epsilon &= 1 - Pr(\text{no rejections}|H_0) \\
&= 1 - \prod^{k} Pr(p_i > \alpha) \\
&= 1 - \prod^{k} (1 - \alpha) \\
&= 1 - (1 - \alpha)^k
\end{aligned}
$$

Solving for α gives

$$\epsilon \approx \alpha / k$$

or exact

$$\epsilon = 1 - \exp(1/k \log(1 - \alpha))$$

http://genomicsclass.github.io/book/pages/multiple_testing.html

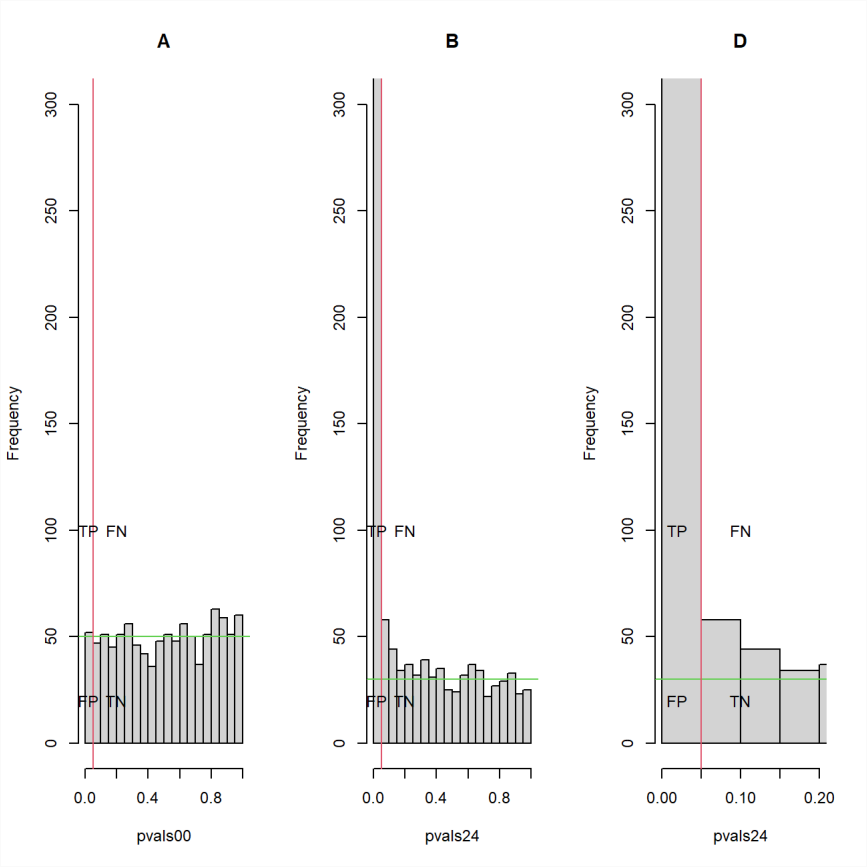
FWER - Conclusion

Controlling the FWER, will demand a *unrealistically small p-value*.

- Limit the number of tests.
- Summarize your measurements e.g. fit time courses
- Use package `multcomp` to correct p-values
  - takes correlation among observations into account
  - uses asymptotic properties, not suited if sample sizes are small.

FDR Motivation

```r
m ← 1000
simulate.p.values ← function(
  i, delta = 2, fraction = 0.1){
  control ← rnorm(6,0,1)
  treatment ← rnorm(6,0,1)
  if (runif(1) < fraction)
   treatment ← treatment + delta
  return(t.test(treatment,control)$p.value)
}
pvals00 ← sapply(1:m, simulate.p.values,
                 delta = 0,fraction = 0 )
pvals24 ← sapply(1:m, simulate.p.values,
                 delta = 2, fraction = 0.4 )
```



Simulating p-values Figure A) 1000 p-values where H0 true, B) 600 p-values where H0 true and 400 HA true. C) closup

False Discovery Rate (FDR)

- Figure A (previous slide) shows that even if only H0 true we have some p-values which are below the significance threshold. These are false positives (FP).
- In Figure D we have p-values less than significance threshold where H0 is true (FP) and a proportion of those where HA is true (TP).
- FDR-controlling procedures are designed to control the expected **proportion of "discoveries"** (rejected null hypotheses) **that are false** (incorrect rejections).

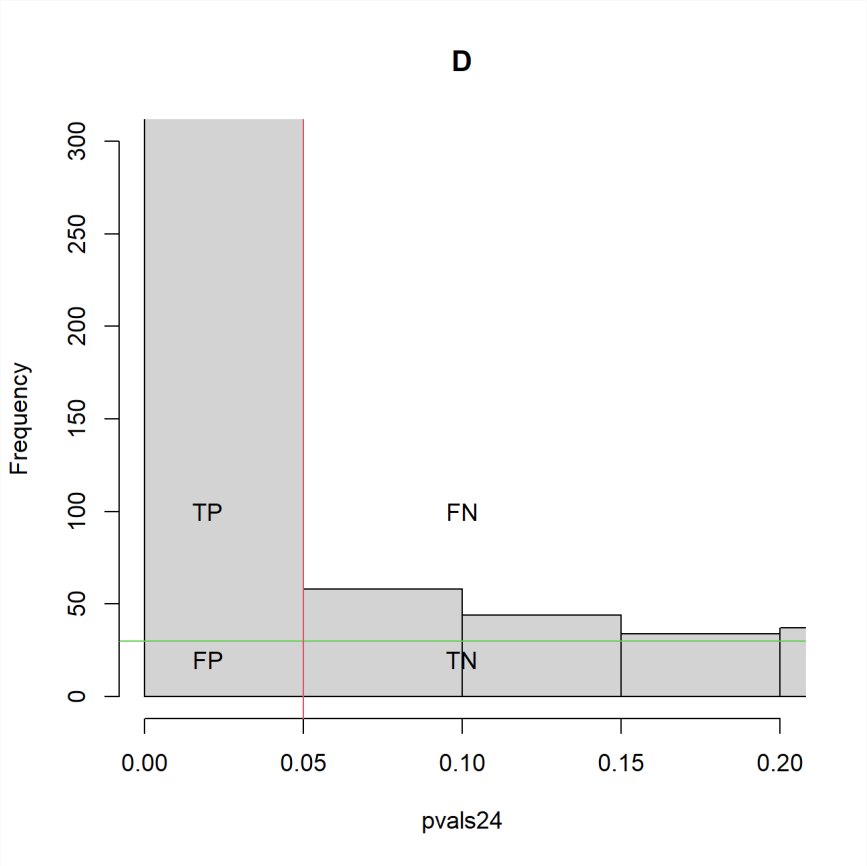
$$FDR = \frac{FP}{FP + TP}$$

- Particularly useful in the discovery fase where even FDR's of up to 50% are feasible.

FDR and p-value distribution

- TP true positives (H0 rejected if HA true)
- FP false positives (H0 rejected if H0 true)
- FN false negatives (H0 accepted if HA true)
- TN true negatives (H0 accepted if H0 true)

$$FDR = \frac{FP}{FP + TP}$$



John D.Storey 2002; Storey and Tibshirani 2003; Prummer 2012

FDR - Benjamini Hochberg - procedure

Definition of FDR as given in the Benjamini and Hochberg paper 1995.

R/C	H0 TRUE	HA	Total
Reject H0	V (FP)	S (TP)	R
Accept H0	U (TN)	T (FN)	m-R
Total	m_0	m-m_0	m

the proportion of false discoveries among the discoveries (rejections of the null hypothesis)

$Q = V/R = V/(V+S); \quad where \quad Q = 0 \ \ if \ \ R = 0$
$FDR = Q_e = E[Q]$ (expected value of Q).

https://en.wikipedia.org/wiki/False_discovery_rate, Benjamini-Hochberg (1995)

FDR - Benjamini Hochberg - procedure

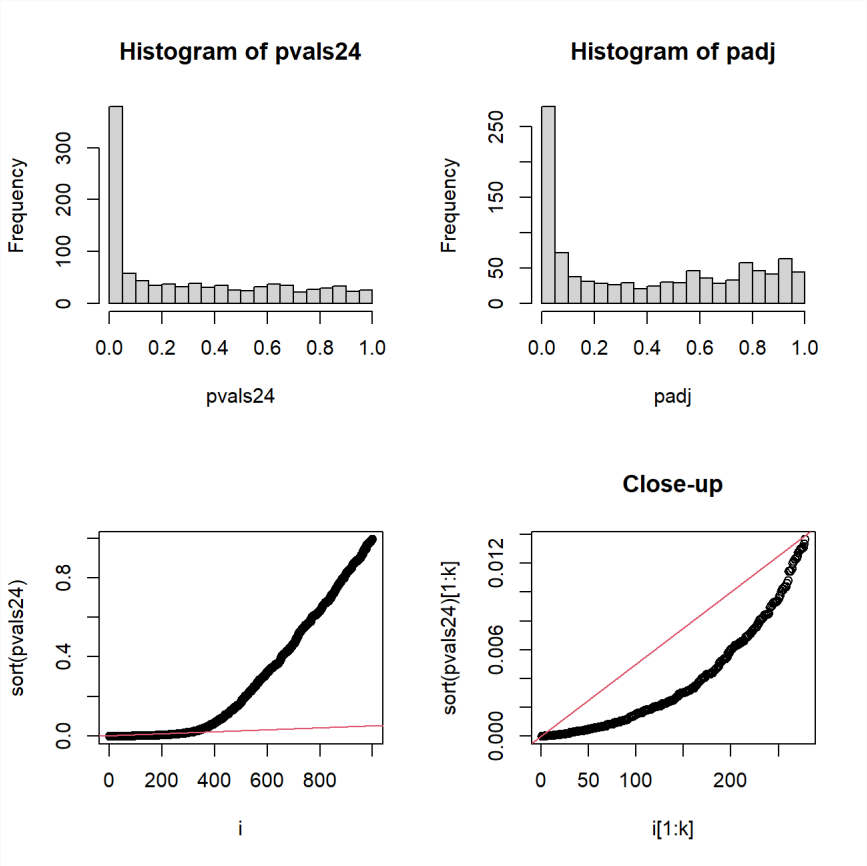
For any given FDR level α, the Benjamini-Hochberg (1995) procedure is very practical because it simply requires that we are able to compute p-values for each of the individual tests and this permits a procedure to be defined.

- List these p-values in ascending order and denote them by $P_{(1)} \dots P_{(m)}$.
- For a given FDR level α, find the largest k such that $P_{(k)} \leq \frac{k}{m}\alpha$.
- Reject the null hypothesis (i.e., declare discoveries) for all $H_{(i)}$ for $i = 1, \dots, k$.

FDR - Benjamini Hochberg - procedure

```r
alpha ← 0.05
i = seq(along=pvals24)
k ← max(which(sort(pvals24) < i/m*alpha))
padj ← p.adjust(pvals24,method="BH")

par(mfrow=c(2,2))
hist(pvals24, breaks=20)
hist(padj , breaks = 20)
plot(i,sort(pvals24))
abline(0,i/m*alpha, col=2)
plot(i[1:k],sort(pvals24)[1:k],type="b",
     main="Close-up")
abline(0,i/m*alpha, col=2)
```



Highlighted code illustrates the Benjmini Hochberge procedure (top line) and how you would compute the FDR in R using the method `p.adjust`.

FDR - Conclusion

- $FDR \leq 0.05$ is a much more lenient requirement then $FWER \leq 0.05$.

Although we will end up with more false positives, FDR gives us much more power. This makes it particularly appropriate for discovery phase experiments where we may accept FDR levels much higher than 0.05.

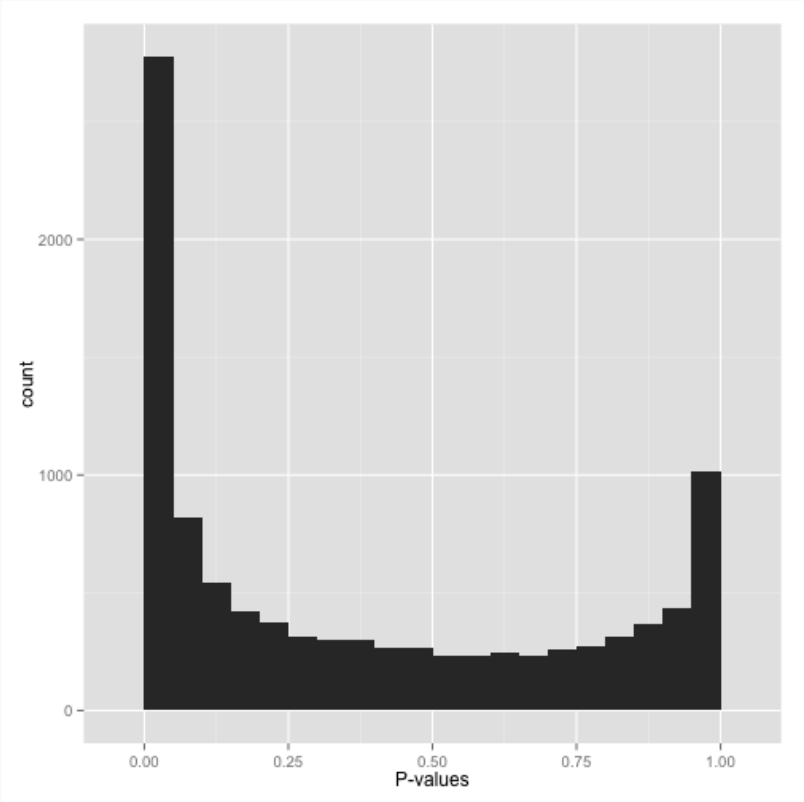
The BH procedure is valid when the m tests are independent, and also in various scenarios of dependence, but is **not universally valid**. (e.g. gene sets.)

Possible p-value distributions

In practice we can observe various shapes of p-value distributions.

How to interpret a p-value histogram

This blog post discusses what types of p-value distrubutions you might encounter when analysing data and how to treat them.



Conclusions

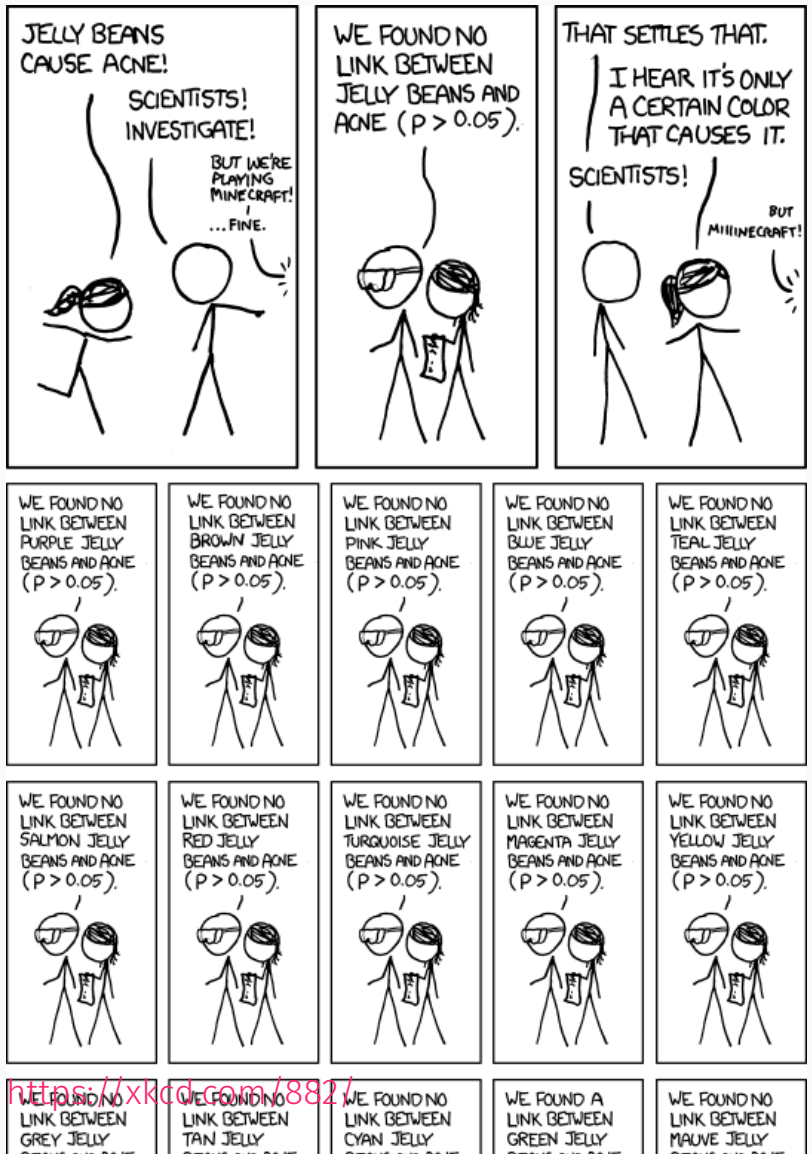
- In case of multiplicity do not report unadjusted p-value.
- *Family Wise Error Rates (FWER)*
  - use to adjust for number of tests for single protein
  - Typical threshold for FWER are 0.05 or 0.01
- *False Discovery Rate (FDR)*
  - controls error rates when selecting proteins for follow up (OMICS experiments)
  - FDR's of 0.1 or even 0.5 are acceptable.
- If you need a FDR estimate Limit the number of traditional hypothesis you test.

Conclusions

If you do subgroup analysis use
exploratory or **descriptive** data analysis:

- tabulating (e.g. venn diagrams)
- dimensionality reduction (e.g, PCA)
- clustering of samples and proteins (e.g, time series clustering)
- Use GSEA or ORA analysis to contrast subgroups.
- Do not over-interpret your findings
  by report FDR's (they are biased).

Thank you



https://xkcd.com/882/