

Outline

Hypothesis testing

Multiple hypothesis testing

P-value correction

Multiple comparisons

Multiple testing

C.C. Günther

Outline

Hypothesis testing

Multiple testing

P-value correction

Multiple
comparisons

Outline

Hypothesis testing

Multiple hypothesis testing

P-value correction

Multiple comparisons

Multiple testing

C.C. Günther

Outline

Hypothesis testing

Multiple testing

P-value correction

Multiple
comparisons

The general idea

Define the null hypothesis H_0 and alternative hypothesis H_1 .

Perform experiment.

How likely is the outcome given that the null hypothesis is true?

Reject or accept null hypothesis.

The hypothesis test

Example: Is the coin
fair, or is either head or tail more likely?

H_0 : The coin is fair. H_1 : The coin is not fair.

1. Toss coin N times.
2. Count the number of heads and tails.
3. Compare to what
would be expected from a fair coin.



If the number of heads and tails is consistent with what could be expected from a fair coin, the *null-hypothesis* that the coin is fair should be accepted; if not, the null-hypothesis should be rejected.

The hypothesis test

Multiple testing

C.C. Günther

Outline

Hypothesis testing

Multiple testing

P-value correction

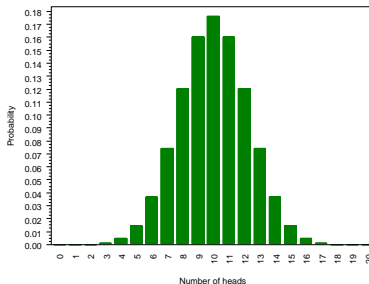
Multiple
comparisons

Example:

If we toss a fair coin 20 times, we can compute the probability of getting x heads ($x = 0, \dots, 20$).

The probability of getting at most 5 heads is appr. 2%; that of 15 more is also appr. 2%.

Our test: The number of heads should be between 6 and 14, otherwise we should reject the null-hypothesis (i.e. that the coin is fair).



The hypothesis test

Multiple testing

C.C. Günther

Outline

Hypothesis testing

Multiple testing

P-value correction

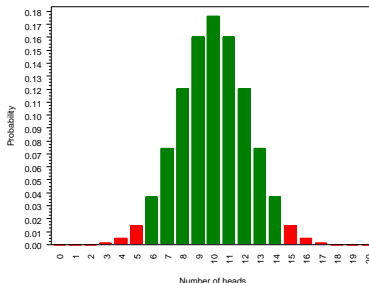
Multiple
comparisons

Example:

If we toss a fair coin 20 times, we can compute the probability of getting x heads ($x = 0, \dots, 20$).

The probability of getting at most 5 heads is appr. 2%; that of 15 more is also appr. 2%.

Our test: The number of heads should be between 6 and 14, otherwise we should reject the null-hypothesis (i.e. that the coin is fair).



Type I and type II errors

What if our decision is wrong?

There are two types of errors to make:

	H_0 is true	H_0 is false
Reject H_0	False positive Type I error	OK
Accept H_0	OK	False negative Type II error

Type I and Type II errors

Null-hypothesis:

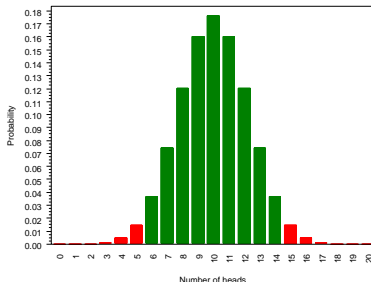
The coin is fair.

Our test: Toss 20 times.

Reject null-hypothesis
if number of heads is less
than 6 or greater than 14.

Type I error: Rejecting
the null hypothesis when
it is true. Even if the coin
is fair, we have 4% probability of rejecting the
null-hypothesis.

Type II error: Not rejecting the null hypothesis when it is
not true. Even if the coin is biased, we may end up
accepting the null-hypothesis.



Multiple testing

C.C. Günther

Outline

Hypothesis testing

Multiple testing

P-value correction

Multiple
comparisons

Significance level of a test

Multiple testing

C.C. Günther

Outline

Hypothesis testing

Multiple testing

P-value correction

Multiple
comparisons

Significance level: The probability of type I error (false positive) of a given test.

It is very common to perform tests at the 5% significance level: i.e. so that the false positive risk is *at most* 5%.

If the false positive risk is less than the selected significance level, the test is *conservative*.

If the false positive risk is larger than the selected significance level, the test is **wrong**!

The power of a test

Multiple testing

C.C. Günther

Outline

Hypothesis testing

Multiple testing

P-value correction

Multiple
comparisons

The probability that a false H_0 is rejected.

It is 1 minus the probability of a type II error.

A test with high power have a higher probability to draw the correct conclusion to reject the null hypothesis than a test with low power.

If the probability of a type I error decreases, the power also decreases.

How do we know when to reject H_0 ?

Multiple testing

C.C. Günther

Outline

Hypothesis testing

Multiple testing

P-value correction

Multiple
comparisons

Calculate the p -value and compare with the chosen significance level.

The p -value is the probability of observing what we have observed or something 'more extreme' when H_0 is true.

Small p -values \Rightarrow Reject H_0 .

Large p -values \Rightarrow Accept H_0 .

P-values

Our experiment:

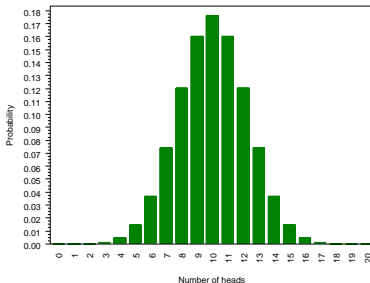
We toss the coin
20 times and get 7 heads.

P-value:

The probability of getting
this outcome or one
that deviates even more
from what is expected
under the null-hypothesis.

$$P = \Pr[X \leq 7 \text{ or } X \geq 13 \mid \text{null-hyp.}] = 0.263 \text{ (or 26.3\%).}$$

The deviation from the null-hypothesis is *statistically significant* at the *5% significance level* if $P \leq 0.05$.



Multiple testing

C.C. Günther

Outline

Hypothesis testing

Multiple testing

P-value correction

Multiple
comparisons

P-values

Our experiment:

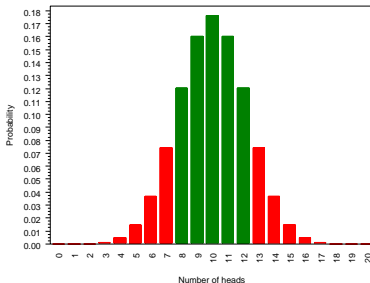
We toss the coin
20 times and get 7 heads.

P-value:

The probability of getting
this outcome or one
that deviates even more
from what is expected
under the null-hypothesis.

$$P = \Pr[X \leq 7 \text{ or } X \geq 13 \mid \text{null-hyp.}] = 0.263 \text{ (or 26.3\%).}$$

The deviation from the null-hypothesis is *statistically significant* at the *5% significance level* if $P \leq 0.05$.



Multiple testing

C.C. Günther

Outline

Hypothesis testing

Multiple testing

P-value correction

Multiple
comparisons

P-values

The P-values give a measure of the statistical strength of the evidence against the null-hypothesis.

$P > 0.05$ At the 5% significance level, this is considered to be what you could expect if the null-hypothesis is true.

P from 0.01 to 0.05 Considered statistically significant, but not strong evidence.

$P < 0.01$ Fairly strong evidence.

$P < 0.001$ Strong evidence.

The P-value does not tell if the deviation from the null-hypothesis is small or large, important or unimportant.

Confidence intervals

What if we don't assume that the coin is fair?

Assume the coin has probability p of head in each toss for some probability $p \in [0, 1]$.

Test which values of p may be rejected, and which must be accepted as possible values. If tests are at the 5% significance level, the accepted values of p form the *95% confidence interval*.

The null-hypothesis that the coin is fair ($p = 1/2$) is accepted if $p = 1/2$ is contained in the confidence interval.

For 7 heads in 20 tosses, the 95% confidence interval for the probability of heads is $[0.15, 0.59]$, which contains $1/2$.

Outline

Hypothesis testing

Multiple hypothesis testing

P-value correction

Multiple comparisons

Multiple testing

C.C. Günther

Outline

Hypothesis testing

Multiple testing

P-value correction

Multiple
comparisons

Testing multiple hypotheses at one time

Multiple testing

C.C. Günther

Outline

Hypothesis testing

Multiple testing

P-value correction

Multiple
comparisons

Example:

Let's test five coins to see if they are fair.

Toss each coin 20 times, and use our test.

If the coins are fair, for each we have 4% probability of a type I error.

What is the probability of making at least one type I error?



Testing multiple hypotheses at one time

Multiple testing

C.C. Günther

Outline

Hypothesis testing

Multiple testing

P-value correction

Multiple
comparisons

What is the probability
of making at least one type I error?



$$\begin{aligned}P(\text{at least one type I error}) &= 1 - P(\text{no type I errors}) \\&= 1 - P(\text{no type I error coin 1}) \cdot \\&\quad \dots \cdot P(\text{no type I error coin 5}) \\&= 1 - (1 - 0.04)^5 = 0.18\end{aligned}$$

The risk of making at least one type I error is 18%.

Example: 10 000 genes

H_0^i : gene i is not differentially expressed, $i = 1, \dots, 10000$

Assume: No differentially expressed genes, H_0^i true for all i .

Significance level $\alpha = 0.01$.

Expect $10000 \cdot \alpha = 10000 \cdot 0.01 = 100$ genes to have a p-value smaller than 0.01 by chance.

We expect to find 100 differentially expressed genes when in fact none of them are!

Many tests \rightarrow many false positives \rightarrow not good!

The problem of multiple hypothesis testing

Multiple testing

C.C. Günther

Outline

Hypothesis testing

Multiple testing

P-value correction

Multiple
comparisons

When performing several tests, the chance of getting one or more false positives increases.

Multiple testing problem: Need to control the risk of false positives (type I error) when performing a large number of tests.

Bad solution to the multiple testing problem

Multiple testing

C.C. Günther

The big DON'T: It is **not** permissible to perform several tests and only present those that gave the desired outcome.

Outline

Hypothesis testing

Multiple testing

P-value correction

Multiple comparisons



Bad solution to the multiple testing problem

Multiple testing

C.C. Günther

The big DON'T: It is **not** permissible to perform several tests and only present those that gave the desired outcome.

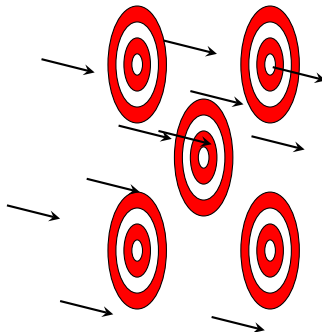
Outline

Hypothesis testing

Multiple testing

P-value correction

Multiple comparisons



All-against-all correlations

Example data: Large B-cell lymphoma data.

Correlation between gene expression signatures.

Pearson correlation P-value	sign_ germB	sign_ lymph	sign_ prolif	BHP6	MHC
sign_germB Germinal center B cell sign.	1.00000	0.16336 0.0113	-0.05530 0.3938	-0.08362 0.1967	0.17837 0.0056
sign_lymph Lymph node signature	0.16336 0.0113	1.00000	-0.31586 <.0001	-0.02660 0.6818	0.15082 0.0194
sign_prolif Proliferation signature	-0.05530 0.3938	-0.31586 <.0001	1.00000	0.14079 0.0292	-0.13411 0.0379
BHP6 BMP6	-0.08362 0.1967	-0.02660 0.6818	0.14079 0.0292	1.00000	0.08650 0.1817
MHC MHC class II signature	0.17837 0.0056	0.15082 0.0194	-0.13411 0.0379	0.08650 0.1817	1.00000

Computing all pairwise correlations and then presenting only those that are statistically significant, is not acceptable!

Multiple testing

C.C. Günther

Outline

Hypothesis testing

Multiple testing

P-value correction

Multiple
comparisons

Large scale T-testing

Multiple testing

C.C. Günther

Example data: Expression from 100 genes, outcome is survival. Perform t-test for each gene.

H_0^i : gene i is not differentially expressed, $i = 1, \dots, 100$.

Rank	Gene	P-value	Rank	Gene	P-value	...
1	GENE84X	0.00037	13	GENE6X	0.02083	
2	GENE73X	0.00431	14	GENE71X	0.02401	
3	GENE48X	0.00544	15	GENE49X	0.02463	
4	GENE1X	0.00725	16	GENE38X	0.02751	
5	GENE81X	0.00769	17	GENE46X	0.02804	
6	GENE91X	0.00793	18	GENE75X	0.02892	
7	GENE96X	0.00803	19	GENE36X	0.04072	
8	GENE22X	0.00907	20	GENE83X	0.04519	
9	GENE95X	0.00977	21	GENE8X	0.04608	
10	GENE58X	0.01734	22	GENE21X	0.05213	
11	GENE77X	0.01911	23	GENE78X	0.06940	
12	GENE33X	0.01974	24	GENE16X	0.07046	

Presenting only those with small P-value is inappropriate when we have done 100 tests!

Outline

Hypothesis testing

Multiple testing

P-value correction

Multiple comparisons

Other cases where multiple testing occurs

Multiple testing

C.C. Günther

Outline

Hypothesis testing

Multiple testing

P-value correction

Multiple
comparisons

Example: A researcher wants to compare incidence of disease between rural and urban populations. He finds a difference for two out of ten common diseases ($P=0.02$ and 0.03 resp.).

Example: A researcher wants to check if health depends on social status. Both health and social status can be measured in many different, although similar, ways. He checks all combinations.

Example: A researcher cannot decide which is more appropriate to use: Pearson correlation or Spearman. He picks the one that gives the lowest P-value.

Outline

Hypothesis testing

Multiple hypothesis testing

P-value correction

Multiple comparisons

Multiple testing

C.C. Günther

Outline

Hypothesis testing

Multiple testing

P-value correction

Multiple
comparisons

Corrected p -values

Multiple testing

C.C. Günther

Outline

Hypothesis testing

Multiple testing

P-value correction

Multiple
comparisons

The original p -values do not tell the full story.

Instead of using the original p -values for decision making, we should use corrected ones.

False positive rate under multiple tests

Multiple testing

C.C. Günther

Outline

Hypothesis testing

Multiple testing

P-value correction

Multiple
comparisons

Result: If you perform N tests at a significance level α , then the probability of having at least one false positive is at most $N \times \alpha$.

In many cases, the risk will be less, but this result is true even in the worst of cases.

It is also correct if some of the null-hypotheses are actually wrong.

May use this to formulate a *multiple test* that controls the over-all risk of having a false positive.

Bonferroni's p -value correction

Multiple testing

C.C. Günther

Outline

Hypothesis testing

Multiple testing

P-value correction

Multiple
comparisons

Bonferroni: If you perform N tests at a significance level α/N , then the probability of having at least one false positive is at most α .

Bonferroni p -value: If you run N tests, multiply all the p -values by N to get the Bonferroni corrected p -values.

Result: The probability of getting a Bonferroni corrected p -value less than α for a true null-hypothesis is at most α .

Bonferroni's P-value correction

Multiple testing

C.C. Günther

Outline

Hypothesis testing

Multiple testing

P-value correction

Multiple
comparisons

Pearson correlation / P-value

sign_germB					-
Germinal center B cell sign.					
sign_lymph	0.16336				-
Lymph node signature	0.0113				
sign_prolif	-0.05530	-0.31586			-
Proliferation signature	0.3938	<.0001			
BHP6	-0.08362	-0.02660	0.14079		-
BMP6	0.1967	0.6818	0.0292		
MHC	0.17837	0.15082	-0.13411	0.08650	-
MHC class II signature	0.0056	0.0194	0.0379	0.1817	

Multiply each p -value by 10 to get the Bonferroni corrected P-value.

Bonferroni's P-value correction

Multiple testing

C.C. Günther

Outline

Hypothesis testing

Multiple testing

P-value correction

Multiple
comparisons

Pearson correlation / P-value

sign_germB	-				
Germinal center B cell sign.					
sign_lymph	0.16336	-			
Lymph node signature	0.0113				
sign_prolif	-0.05530	-0.31586	-		
Proliferation signature	0.3938	<.0001			
BHP6	-0.08362	-0.02660	0.14079	-	
BMP6	0.1967	0.6818	0.0292		
MHC	0.17837	0.15082	-0.13411	0.08650	-
MHC class II signature	0.0056	0.0194	0.0379	0.1817	

Multiply each p -value by 10 to get the Bonferroni corrected P-value.

Large scale T-testing

T-tests done for 100 genes. Bonferroni correction requires us to multiply all P-values with 100.

Rank	Gene	P-value	Rank	Gene	P-value	...
1	GENE84X	0.00037	13	GENE6X	0.02083	
2	GENE73X	0.00431	14	GENE71X	0.02401	
3	GENE48X	0.00544	15	GENE49X	0.02463	
4	GENE1X	0.00725	16	GENE38X	0.02751	
5	GENE81X	0.00769	17	GENE46X	0.02804	
6	GENE91X	0.00793	18	GENE75X	0.02892	
7	GENE96X	0.00803	19	GENE36X	0.04072	
8	GENE22X	0.00907	20	GENE83X	0.04519	
9	GENE95X	0.00977	21	GENE8X	0.04608	
10	GENE58X	0.01734	22	GENE21X	0.05213	
11	GENE77X	0.01911	23	GENE78X	0.06940	
12	GENE33X	0.01974	24	GENE16X	0.07046	

Only the smallest P-value survives Bonferroni correction.

Large scale T-testing

T-tests done for 100 genes. Bonferroni correction requires us to multiply all P-values with 100.

Rank	Gene	P-value	Rank	Gene	P-value	...
1	GENE84X	0.00037	13	GENE6X	0.02083	
2	GENE73X	0.00431	14	GENE71X	0.02401	
3	GENE48X	0.00544	15	GENE49X	0.02463	
4	GENE1X	0.00725	16	GENE38X	0.02751	
5	GENE81X	0.00769	17	GENE46X	0.02804	
6	GENE91X	0.00793	18	GENE75X	0.02892	
7	GENE96X	0.00803	19	GENE36X	0.04072	
8	GENE22X	0.00907	20	GENE83X	0.04519	
9	GENE95X	0.00977	21	GENE8X	0.04608	
10	GENE58X	0.01734	22	GENE21X	0.05213	
11	GENE77X	0.01911	23	GENE78X	0.06940	
12	GENE33X	0.01974	24	GENE16X	0.07046	

Only the smallest P-value survives Bonferroni correction.

Bonferroni's p -value correction

Bonferroni correction is the most well-known multiple testing correction:

- ▶ Very simple.
- ▶ Always correct: no model assumptions, no assumption of independence.
- ▶ Gives one new p -value for each test.
- ▶ Useable even if some hypotheses are false.
- ▶ If some tests produce false positives even after correction, it will still be reliable on other tests (unless correlated).

However, Bonferroni-correction is often conservative!

Bonferroni's p -value correction

Multiple testing

C.C. Günther

Outline

Hypothesis testing

Multiple testing

P-value correction

Multiple
comparisons

Pearson correlation / P-value

sign_germB Germinal center B cell sign.	-				
sign_lymph Lymph node signature	0.16336 0.0113	-			
sign_prolif Proliferation signature	-0.05530 0.3938	-0.31586 <.0001	-		
BHP6 BMP6	-0.08362 0.1967	-0.02660 0.6818	0.14079 0.0292	-	
MHC MHC class II signature	0.17837 0.0056	0.15082 0.0194	-0.13411 0.0379	0.08650 0.1817	-

Only one p -value would survive Bonferroni correction.

However, getting $P < 0.05$ for 5 of the remaining 9 correlations seems unlikely to happen by chance.

In this case, Bonferroni correction is quite conservative.

Bonferroni's p -value correction

Pearson correlation / P-value

sign_germB Germinal center B cell sign.	-				
sign_lymph Lymph node signature	0.16336 0.0113	-			
sign_prolif Proliferation signature	-0.05530 0.3938	-0.31586 <.0001	-		
BHP6 BMP6	-0.08362 0.1967	-0.02660 0.6818	0.14079 0.0292	-	
MHC MHC class II signature	0.17837 0.0056	0.15082 0.0194	-0.13411 0.0379	0.08650 0.1817	-

Only one p -value would survive Bonferroni correction.

However, getting $P < 0.05$ for 5 of the remaining 9 correlations seems unlikely to happen by chance.

In this case, Bonferroni correction is quite conservative.

Large scale T-testing

Multiple testing

C.C. Günther

Outline

Hypothesis testing

Multiple testing

P-value correction

Multiple
comparisons

Microarrays now contain more than 40.000 probes: Too many to test them one by one and hope that they can survive Bonferroni correction.

Assume $\alpha = 0.05$, $N = 40000$

H_0^i : gene i is not differentially expressed, $i = 1, \dots, 40000$.

Reject H_0^i if $p_i \cdot 40000 \leq 0.05$

i.e. if $p_i \leq 0.00000025$.

The original p-values must be very small in order to reject.

The problem of conservative corrections

Multiple testing

C.C. Günther

Outline

Hypothesis testing

Multiple testing

P-value correction

Multiple
comparisons

There are two problems with conservative correction:

1. Need very small p -value to reject H_0 .
2. The power of the test is low.

Alternative p -value corrections

Multiple testing

C.C. Günther

Outline

Hypothesis testing

Multiple testing

P-value correction

Multiple
comparisons

Several (less conservative) methods exist.

Two groups of methods:

- ▶ Methods that control the family-wise error rate (FWER).
- ▶ Methods that control the false discovery rate (FDR).

Alternative p -value corrections

Possible outcomes from m hypothesis tests:

	No. true	No. false	Total
No. accepted	U	T	$m - R$
No. rejected	V	S	R
Total	m_0	$m - m_0$	m

V : no. of type I errors (false positives)

T : no. of type II errors (false negatives)

Family-wise error rate (FWER)

- ▶ The probability of at least one type I error
 - ▶ $\text{FWER} = P(V \geq 1)$
- ▶ Control FWER at a level α .
 - ▶ Procedures that adjust the p-values separately.
 - ▶ Single step procedures.
 - ▶ More powerful procedures adjust sequentially, from the smallest to the largest, or vice versa.
 - ▶ Step-down and step-up methods
- ▶ The Bonferroni correction controls the FWER.

Methods that control the FWER

Multiple testing

C.C. Günther

Outline

Hypothesis testing

Multiple testing

P-value correction

Multiple
comparisons

- ▶ Bonferroni
- ▶ Sidak
- ▶ Bonferroni–Holm
- ▶ Westfall & Young

Sidak correction

Multiple testing

C.C. Günther

Outline

Hypothesis testing

Multiple testing

P-value correction

Multiple
comparisons

Assumes independent tests.

The adjusted p -value is found from the formula

$$\tilde{p}_i = 1 - (1 - p_i)^{1/n}$$

where p_i is the unadjusted p -value and n is the number of tests.

Very similar to the Bonferroni correction, very conservative.

Step-down procedure, adjust p -values sequentially.

Order the k p -values, let $p_{(1)}$ be the smallest, $p_{(2)}$ the second smallest and so on.

If $p_{(1)} < \alpha/k$, reject $H_{0,1}$ and continue...

If $p_{(2)} < \alpha/(k + 1 - 2) = \alpha/(k - 1)$, reject $H_{0,2}$

and so on...

until the hypothesis cannot be rejected.

The Bonferroni-Holm adjusted p -values \tilde{p} are then given by

$$\begin{aligned}\tilde{p}_1 &= k \cdot p_1 \\ \tilde{p}_j &= \max((k - j + 1) \cdot p_j, \tilde{p}_{j-1}), \quad 2 \leq j \leq k\end{aligned}$$

Adjusted p -values greater than 1 are set to 1.

Example: Bonferroni-Holm

Rank	P-value	Corrected P-value
1	0.00082	* 19 = 0.01558 *
2	0.00143	* 18 = 0.02574 *
3	0.00171	* 17 = 0.02907 *
4	0.00242	* 16 = 0.03872 *
5	0.00538	* 15 = 0.08070
6	0.00905	* 14 = 0.12670
7	0.01241	* 13 = 0.16133
8	0.03512	* 12 = 0.42144
9	0.04366	* 11 = 0.48026
10	0.07431	* 10 = 0.74311
11	0.14253	* 9 = 1.00000
12	0.15675	* 8 = 1.00000
13	0.21415	* 7 = 1.00000
14	0.25134	* 6 = 1.00000
15	0.41526	* 5 = 1.00000
16	0.46761	* 4 = 1.00000
17	0.57738	* 3 = 1.00000
18	0.75464	* 2 = 1.00000
19	0.89514	* 1 = 1.00000

Bonferroni-Holm p -value corresponds to removing tests as they are found to be significant and perform Bonferroni correction on the remaining.

Example: Bonferroni-Holm

Rank	P-value	Corrected P-value
1	0.00082	* 19 = 0.01558 *
2	0.00143	* 18 = 0.02574 *
3	0.00171	* 17 = 0.02907 *
4	0.00242	* 16 = 0.03872 *
5	0.00538	* 15 = 0.08070
6	0.00905	* 14 = 0.12670
7	0.01241	* 13 = 0.16133
8	0.03512	* 12 = 0.42144
9	0.04366	* 11 = 0.48026
10	0.07431	* 10 = 0.74311
11	0.14253	* 9 = 1.00000
12	0.15675	* 8 = 1.00000
13	0.21415	* 7 = 1.00000
14	0.25134	* 6 = 1.00000
15	0.41526	* 5 = 1.00000
16	0.46761	* 4 = 1.00000
17	0.57738	* 3 = 1.00000
18	0.75464	* 2 = 1.00000
19	0.89514	* 1 = 1.00000

Bonferroni-Holm p -value corresponds to removing tests as they are found to be significant and perform Bonferroni correction on the remaining.

Permutation tests

Statistical technique to use when distribution is unknown.

Example: Gene set measurements for patient and control group.

For each gene $i = 1, \dots, n$, a test statistic t_i is calculated.

Assume $|t_1| \geq |t_2| \geq \dots \geq |t_n|$.

Permute the 'patient' and 'control' labels \Rightarrow new dataset.

Calculate new $t_{i,b}^*$ for the permuted sample.

Repeat B times, B is large number.

The $t_{i,b}^*$, $b = 1, \dots, B$ now constitute a distribution for t_i under the null hypothesis.

The p -value of t_i can be calculated as

$$p_i = \frac{\text{number of permutations with } |t_{i,b}^*| \geq |t_i|}{\text{number of permutations } B}$$

Permutation tests

Multiple testing

C.C. Günther

Outline

Hypothesis testing

Multiple testing

P-value correction

Multiple
comparisons

The Westfall and Young step-down correction calculates adjusted p -values directly through permutation.

These p -values take correlations between the tests into account.

$$\tilde{p}_i = \frac{\text{number of permutations with } u_{i,b} \geq |t_i|}{\text{number of permutations}}$$

where $u_{n,b} = |t_{n,b}^*|$

$u_{i,b} = \max_{l=i,\dots,n}(u_{l+1,b}, |t_{l,b}^*|)$, $i = n-1, \dots, 1$

Disadvantage: Computer intensive method.

Alternative p -value corrections

Possible outcomes from m hypotheses tests:

	No. true	No. false	Total
No. accepted	U	T	$m - R$
No. rejected	V	S	R
Total	m_0	$m - m_0$	m

V : no. of type I errors (false positives)

T : no. of type II errors (false negatives)

False discovery rate (FDR)

- ▶ The expected proportion of false positives among the rejected hypotheses.
 - ▶ $FDR = E[V/R | R > 0] \cdot P(R > 0)$
- ▶ Example: If 100 null hypotheses are rejected, with an FDR of 5%, 5 of them will be false positives.
- ▶ Various procedures
 - ▶ The Benjamini-Hochberg procedure
 - ▶ Other versions

Controlling the false discovery rate

Multiple testing

C.C. Günther

Outline

Hypothesis testing

Multiple testing

P-value correction

Multiple
comparisons

The Benjamini-Hochberg procedure

Assumes independent p -values.

Let $p_{(1)}, \dots, p_{(m)}$ be the ordered p -values p_1, \dots, p_m .

Start with $p_{(m)}$. Reject $H_{0,m}$ if $p_{(m)} \leq \alpha$.

For the remaining p -values:

Reject $H_{0,i}$ if $\tilde{p}_{(i)} \leq \alpha$

where $\tilde{p}_{(i)} = \min_{k \in \{i, \dots, n\}} \frac{m \cdot p_{(k)}}{k}$.

Other variations exist.

Simple example

The Benjamini-Hochberg procedure

Assume the unadjusted p -values are 0.007, 0.02, 0.4, 0.5.

The adjusted p -values are then $\tilde{p}_{(i)} = \min_{k \in \{i, \dots, n\}} \frac{m \cdot p_{(k)}}{k}$:

$$\tilde{p}_{(4)} = 0.50$$

$$\tilde{p}_{(3)} = 4 \cdot 0.4/3 = 0.53 > \tilde{p}_{(4)} \Rightarrow \tilde{p}_{(3)} = 4 \cdot 0.5/4 = 0.50$$

$$\tilde{p}_{(2)} = 4 \cdot 0.02/2 = 0.04$$

$$\tilde{p}_{(1)} = 4 \cdot 0.007/1 = 0.028$$

Example: Adjusting to control the FDR

Multiple testing

C.C. Günther

Outline

Hypothesis testing

Multiple testing

P-value correction

Multiple
comparisons

Rank	P-value	FDR (5%)
1	0.00082	* $19 / 3 = 0.01083$
2	0.00143	* $19 / 3 = 0.01083$
3	0.00171	* $19 / 3 = 0.01083$
4	0.00242	* $19 / 4 = 0.01150$
5	0.00538	* $19 / 5 = 0.02044$
6	0.00905	* $19 / 6 = 0.02867$
7	0.01241	* $19 / 7 = 0.03368$
8	0.03512	* $19 / 8 = 0.08341$
9	0.04366	* $19 / 9 = 0.09217$
10	0.07431	* $19 / 10 = 0.014119$
11	0.14253	* $19 / 11 = 0.024619$
12	0.15675	* $19 / 12 = 0.24819$
13	0.21415	* $19 / 13 = 0.31299$
14	0.25134	* $19 / 14 = 0.34110$
15	0.41526	* $19 / 15 = 0.52600$
16	0.46761	* $19 / 16 = 0.55529$
17	0.57738	* $19 / 17 = 0.64531$
18	0.75464	* $19 / 18 = 0.79656$
19	0.89514	* $19 / 19 = 0.89514$

Example: Adjusting to control the FDR

Multiple testing

C.C. Günther

Outline

Hypothesis testing

Multiple testing

P-value correction

Multiple
comparisons

Rank	P-value	FDR (5%)
1	0.00082	* 19 / 3 = 0.01083
2	0.00143	* 19 / 3 = 0.01083
3	0.00171	* 19 / 3 = 0.01083
4	0.00242	* 19 / 4 = 0.01150
5	0.00538	* 19 / 5 = 0.02044
6	0.00905	* 19 / 6 = 0.02867
7	0.01241	* 19 / 7 = 0.03368
8	0.03512	* 19 / 8 = 0.08341
9	0.04366	* 19 / 9 = 0.09217
10	0.07431	* 19 / 10 = 0.014119
11	0.14253	* 19 / 11 = 0.024619
12	0.15675	* 19 / 12 = 0.24819
13	0.21415	* 19 / 13 = 0.31299
14	0.25134	* 19 / 14 = 0.34110
15	0.41526	* 19 / 15 = 0.52600
16	0.46761	* 19 / 16 = 0.55529
17	0.57738	* 19 / 17 = 0.64531
18	0.75464	* 19 / 18 = 0.79656
19	0.89514	* 19 / 19 = 0.89514

The Benjamini-Hochberg approach

Multiple testing

C.C. Günther

Outline

Hypothesis testing

Multiple testing

P-value correction

Multiple
comparisons

- ▶ Controls the FDR.
- ▶ Assume independent p -values.
- ▶ Commonly used.
- ▶ Applies to a set of genes, not to individual genes.
- ▶ Does not tell you which p -values are false positives, only how many that are.

Correction of p -values in R

Function `p.adjust` is easy to use.

```
p.adjust(p, method = p.adjust.methods)
```

Input:

- ▶ Vector of p -values.
- ▶ Method is e.g. "holm", "bonferroni", "BH".
- ▶ Returns the adjusted p -values.

Correction of p -values in R

Multiple testing

C.C. Günther

Outline

Hypothesis testing

Multiple testing

P-value correction

Multiple
comparisons

Many BioConductor packages return corrected p -values themselves.

Example: The 'limma' package by Smyth et al.

Tests for differential expression between groups.

The function `topTable` returns a table of top-ranked genes with unadjusted and adjusted p -values. Default correction method is Benjamini-Hochberg.

Another approach to multiple testing

Ideally, one should perform one test only, and decide on the test prior to analysing the data.

In reality, data is scarce, and one wants to perform more analyses, get more results and test more hypotheses.

One compromise is to divide analyses into two parts:

Hypothesis testing: As rigorous as can be done! Want reliable conclusions.

Hypothesis generating: Less rigorous, allowing data mining, multiple testing, etc. Conclusions are not expected to be reliable in themselves, but give good ideas/candidates for further research.

Another approach to multiple testing

Multiple testing

C.C. Günther

Outline

Hypothesis testing

Multiple testing

P-value correction

Multiple
comparisons

Decide whether you want to control the FWER or the FDR.

Example microarrays:

- ▶ Are you most afraid of having gene on your significant list that should not have been there.
 - ▶ Choose FWER.
- ▶ Are you most afraid of missing out on interesting genes.
 - ▶ Choose FDR.

Another approach to multiple testing

Multiple testing

C.C. Günther

Outline

Hypothesis testing

Multiple testing

P-value correction

Multiple
comparisons

A summary of the methods:

Bonferroni
Bonferroni Step-Down
Westfall and Young Permutation
Benjamini and Hochberg False Discovery Rate
None

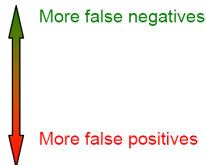


Figure from Multiple Testing Corrections, Agilent Technologies

Outline

Hypothesis testing

Multiple hypothesis testing

P-value correction

Multiple comparisons

Multiple testing

C.C. Günther

Outline

Hypothesis testing

Multiple testing

P-value correction

Multiple
comparisons

Multiple comparisons

Multiple testing

C.C. Günther

Outline

Hypothesis testing

Multiple testing

P-value correction

Multiple
comparisons

One special case of multiple testing is pairwise comparisons of groups.

Example: A doctor is comparing 6 different treatments to find which reduces blood pressure the most by giving each treatment to 10 different patients.

Can use ANOVA (Analysis of Variance) to check if there is any variation between the treatments, and t-tests to compare each pair of treatments. There are 15 pairs, so p -values need to correct for multiple testing.

Analysis of Variance

Multiple testing

C.C. Günther

Outline

Hypothesis testing

Multiple testing

P-value correction

Multiple
comparisons

Let μ_i be the expected mean blood pressure for patients receiving treatment i , $i = 1, \dots, 6$.

We want to test whether all the means are equal.

If they are not, then some of the variability between observations may be due to the different treatments.

The overall ANOVA test only tells us whether at least one treatment differs from the others, not which treatment does.

Multiple comparisons

ANOVA testing

Step 1: Test if there is any variation between the treatments.

H_0^* : All treatments have the same mean, $\mu_1 = \dots = \mu_6$.

vs

H_1^* : At least one treatment has a different mean.

Step 2: If H_0^* is rejected, then for each pair of treatments i and j , we test the null hypothesis

$H_{0,ij}$: Treatment i and j have the same mean, $\mu_i = \mu_j$.

vs

$H_{1,ij}$: Treatment i and j do not have the same mean, $\mu_i \neq \mu_j$.

Multiple comparisons

Multiple testing

C.C. Günther

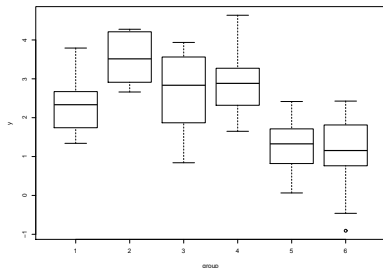
Outline

Hypothesis testing

Multiple testing

P-value correction

Multiple comparisons



Example output from ANOVA in R:

```
group
  1      2      3      4      5      6
2.358 3.543 2.646 2.885 1.327 1.042

      Df Sum Sq Mean Sq F value    Pr(>F)
group    5  45.394   9.0788  12.222 6.098e-08 ***
Residuals 54  40.112   0.7428

--
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Multiple comparisons

The null hypothesis for each pair of treatments can be tested using a t-test.

However, we need to correct for multiple testing.

Two situations:

- ▶ All-against-all comparisons
 - ▶ Tukey
- ▶ One-against-all comparisons
 - ▶ Dunnet

Tukey's procedure

- ▶ Adjustment of p -values for all-against-all T-tests.
- ▶ Controls the FWER.
- ▶ When the sample sizes are equal, the control is exact.

All-against-all comparisons

Multiple testing

C.C. Günther

Output from R (using the TukeyHSD function)

```
Tukey multiple comparisons of means
 95% family-wise confidence level
```

```
Fit: aov(formula = y ~ group)
```

```
$group
```

	diff	lwr	upr	p adj
2-1	1.1846806	0.04591626	2.3234450	0.0369410
3-1	0.2884340	-0.85033032	1.4271984	0.9747363
4-1	0.5272223	-0.61154207	1.6659867	0.7456503
5-1	-1.0312727	-2.17003704	0.1074917	0.0970729
6-1	-1.3157768	-2.45454114	-0.1770124	0.0147094
3-2	-0.8962466	-2.03501095	0.2425178	0.2020111
4-2	-0.6574583	-1.79622270	0.4813060	0.5341050
5-2	-2.2159533	-3.35471767	-1.0771889	0.0000062
6-2	-2.5004574	-3.63922177	-1.3616930	0.0000004
4-3	0.2387883	-0.89997611	1.3775526	0.9891193
5-3	-1.3197067	-2.45847108	-0.1809424	0.0142918
6-3	-1.6042108	-2.74297518	-0.4654465	0.0015222
5-4	-1.5584950	-2.69725934	-0.4197306	0.0022220
6-4	-1.8429991	-2.98176343	-0.7042347	0.0001932
6-5	-0.2845041	-1.42326846	0.8542603	0.9762048

Outline

Hypothesis testing

Multiple testing

P-value correction

Multiple
comparisons

Dunnett's test:

- ▶ Adjustment of p -values for one-against-all T-tests.
- ▶ One group is e.g. placebo or the standard treatment to which the others should be compared.
- ▶ Controls the FWER at level α .

Output from R using the `glht` function in the `multcomp` package.

Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Dunnett Contrasts

Fit: `aov(formula = y ~ group)`

Linear Hypotheses:

	Estimate	Std. Error	t value	Pr(> t)	
2 - 1 == 0	1.1847	0.3854	3.074	0.01438	*
3 - 1 == 0	0.2884	0.3854	0.748	0.91262	
4 - 1 == 0	0.5272	0.3854	1.368	0.51708	
5 - 1 == 0	-1.0313	0.3854	-2.676	0.04042	*
6 - 1 == 0	-1.3158	0.3854	-3.414	0.00548	**

--

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported - single-step method)

Summary

Multiple testing

C.C. Günther

Outline

Hypothesis testing

Multiple testing

P-value correction

Multiple
comparisons

- ▶ Always try to decide what you want to test and how before looking at the results.
- ▶ Always keep multiple testing in mind when you are testing more than one hypothesis.
- ▶ When testing many hypotheses, it is usually desirable to control the FDR.
- ▶ For a smaller number of hypotheses, controlling the FWER may be the right choice.