Statistics I
The multiple problems of multiple testing

Einar Andreas Rødland

7 September 2009

Outline

Outline

Hypothesis testing

Multiple hypothesis testing

P-value correction

Multiple comparisons

The hypothesis test

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

The hypothesis test

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

1. Toss coin N times.

The hypothesis test

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

1. Toss coin N times.
2. Count the number of heads and tails.

The hypothesis test

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

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

The hypothesis test

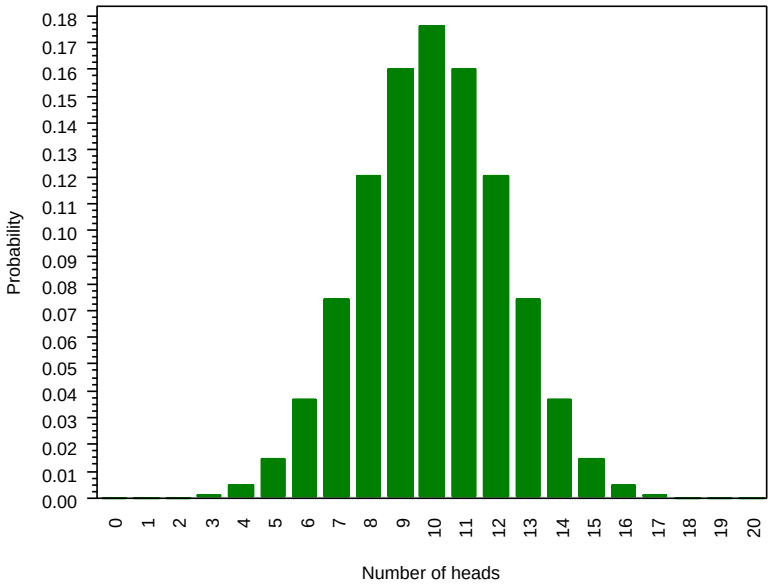
Example: Is the coin
fair, or is either head or tail more likely?

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 retained; if not, the null-hypothesis should be
rejected.

The hypothesis test

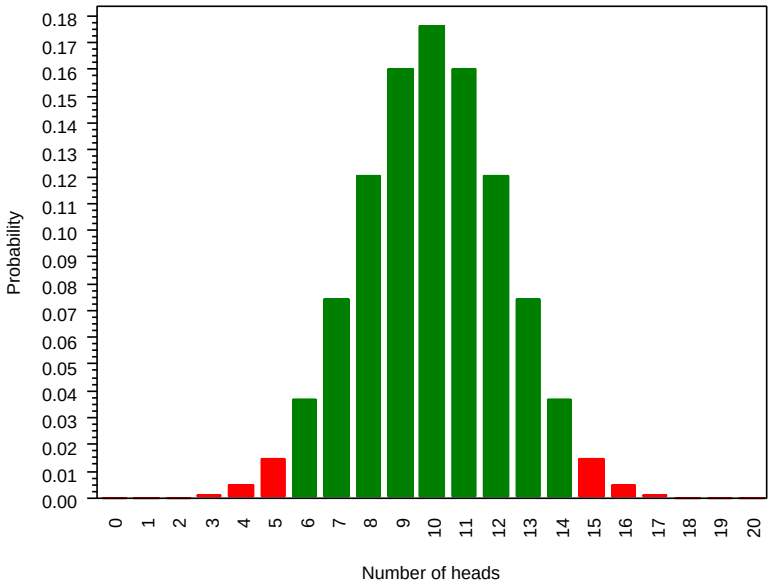
Example:
If we toss a fair coin 20 times, we can compute the probability of getting x heads ($x = 0, \ldots, 20$).



The hypothesis test

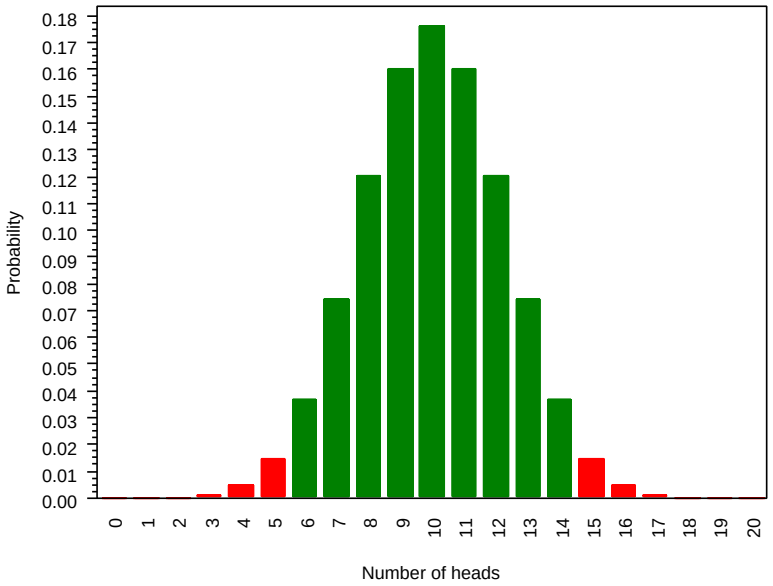
Example:
If we toss a fair coin 20 times, we can compute the probability of getting x heads ($x = 0, \ldots, 20$).

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



The hypothesis test

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



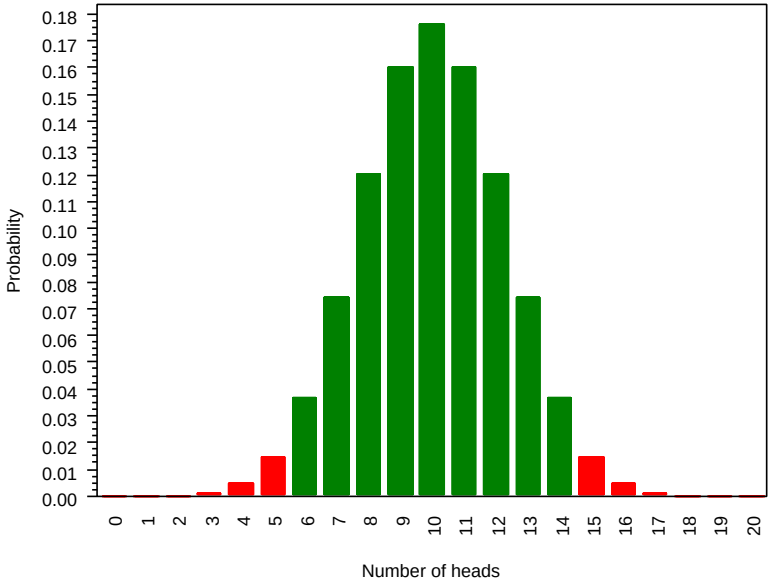
The probability of getting at most 5 heads is appr. 2%; that of 15 more is also appr. 2%.

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

Null-hypothesis:
The coin is fair.

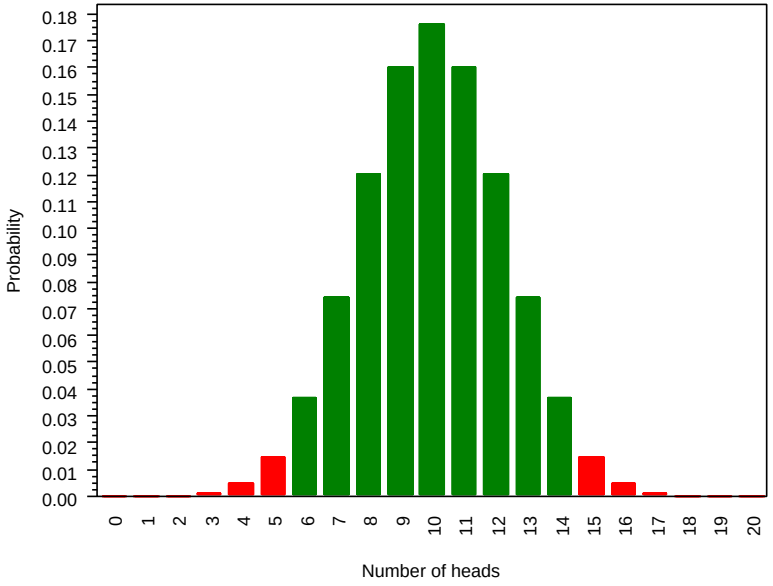
Our test: Toss 20 times.
Reject null-hypothesis
if number of heads
in not between 6 and 14.



Type I and Type II errors

Null-hypothesis:
The coin is fair.

Our test: Toss 20 times.
Reject null-hypothesis
if number of heads
in not between 6 and 14.

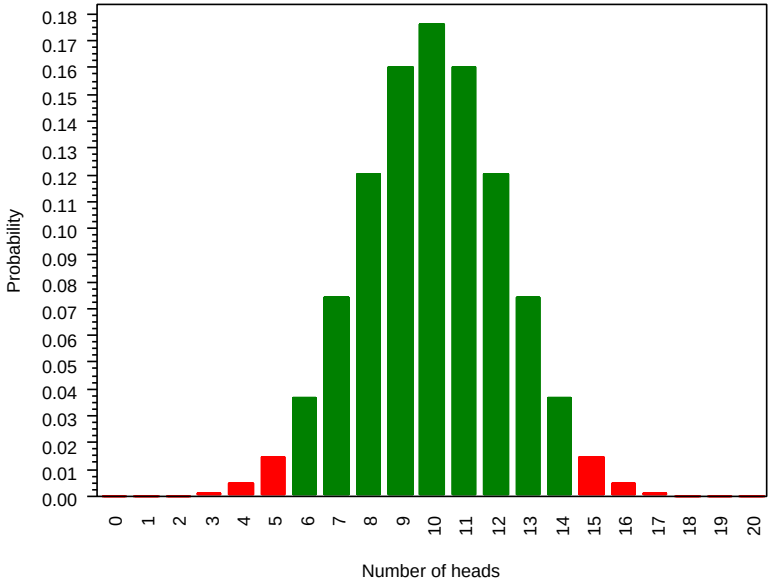


Type I error:
False positive. Even
if the coin is fair, we have
4% likelihood of rejecting the null-hypothesis.

Type I and Type II errors

Null-hypothesis:
The coin is fair.

Our test: Toss 20 times.
Reject null-hypothesis
if number of heads
in not between 6 and 14.



Type I error:
False positive. Even
if the coin is fair, we have
4% likelihood of rejecting the null-hypothesis.

Type II error: False negative. Even if the coin is biased,
we may end up retaining the null-hypothesis.

Significance level of a test

Significance level: The risk of Type I error (false positive) of a given test.

Significance level of a test

Significance level: The risk of Type I error (false positive) of a given test.

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

Significance level of a test

Significance level: The risk of Type I error (false positive) of a given test.

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

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

Significance level of a test

Significance level: The risk of Type I error (false positive) of a given test.

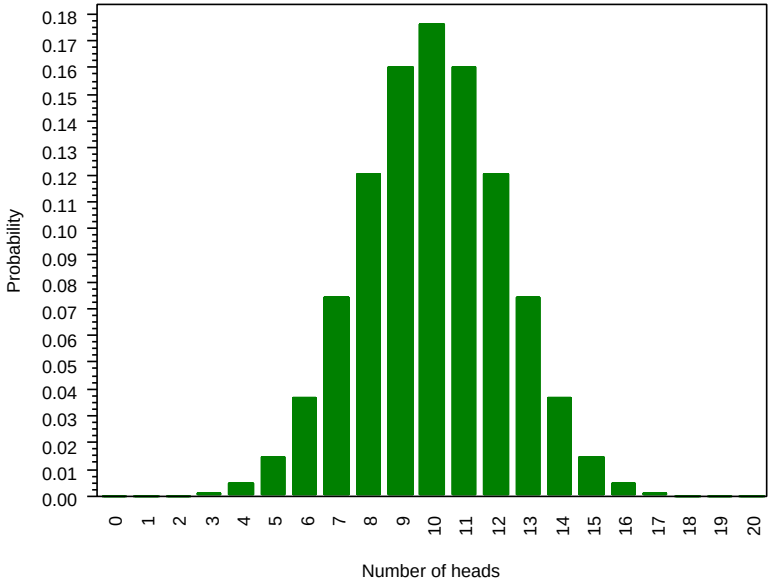
It is very common to make tests at the 5% significance level: i.e. so that false positive risk is *at most* 5%.

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**!

P-values

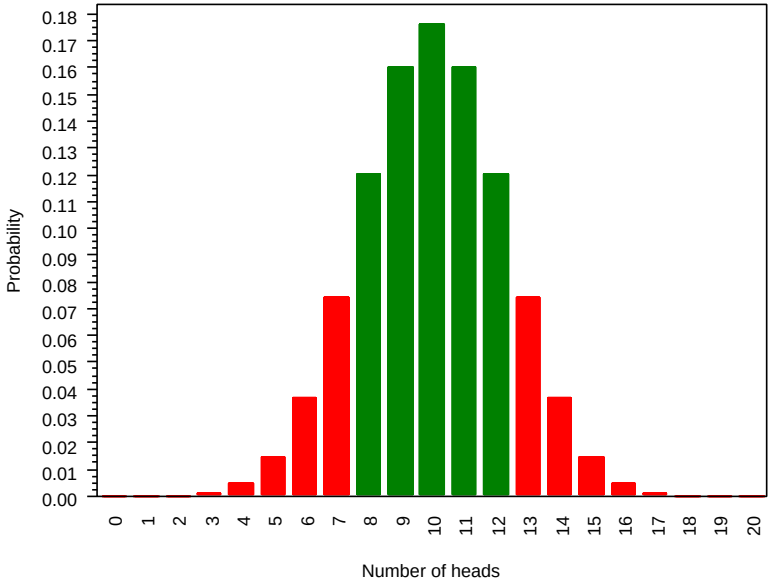
Our experiment:
We toss the coin
20 times and get 7 heads.



P-values

Our experiment:
We toss the coin
20 times and get 7 heads.

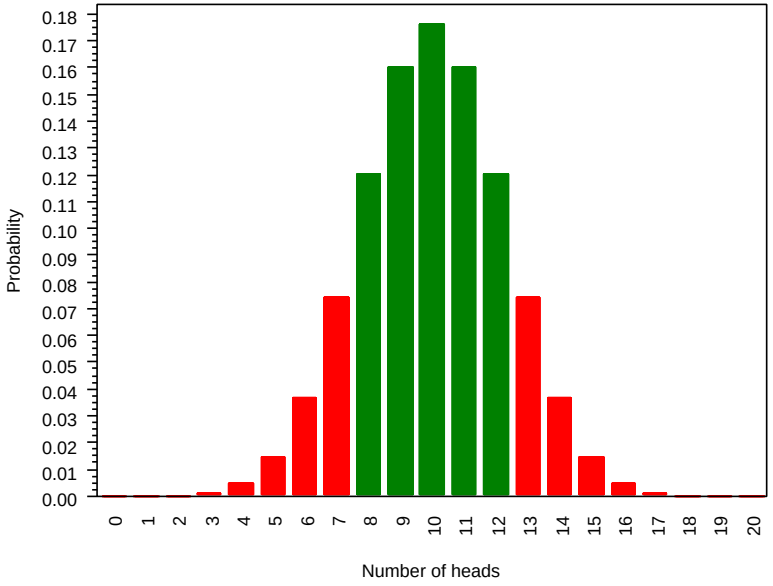
P-value:
The likelihood of getting
this outcome or one
that deviates even more
from what is expected
under the null-hypothesis.



P-values

Our experiment:
We toss the coin
20 times and get 7 heads.

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

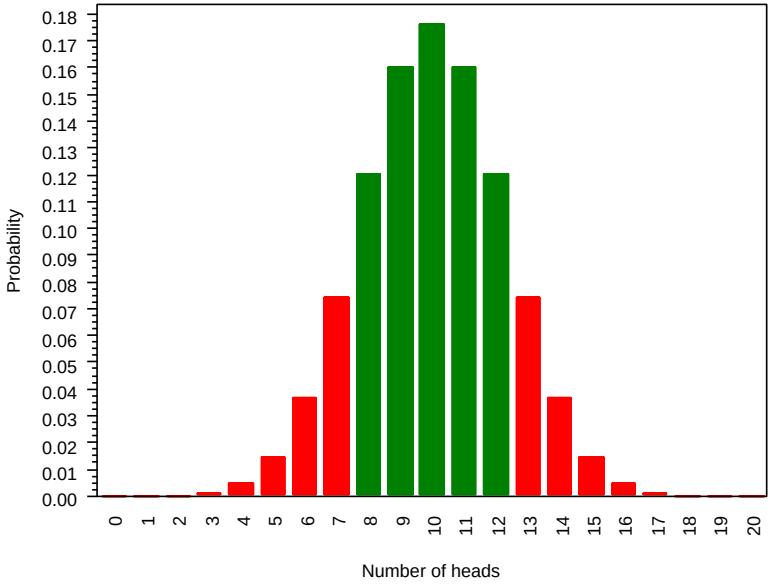


$$P = Pr[X \leq 7 \text{ or } X \geq 13 \mid \text{null-hyp.}] = 0.263 \text{ (or 26.3\%)}.$$

P-values

Our experiment:
We toss the coin
20 times and get 7 heads.

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



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

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

P-values

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

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.

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?

Confidence intervals

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

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

Confidence intervals

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

Hypothesis p**:** 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 retained as possible values. If tests are at the 5% significance level, the retained values of p form the *95% confidence interval*.

Confidence intervals

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

Hypothesis p**:** 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 retained as possible values. If tests are at the 5% significance level, the retained values of p form the *95% confidence interval*.

The null-hypothesis that the coin is fair ($p = 1/2$) is retained 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

Testing multiple hypotheses at one time

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

Testing multiple hypotheses at one time

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

Toss each coin 20 times, and use our test.

Testing multiple hypotheses at one time

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% likelihood of a Type I error.

Testing multiple hypotheses at one time

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% likelihood of a Type I error.

There is appr. 20% risk of making at least one Type I error.

The problem of multiple hypothesis testing

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

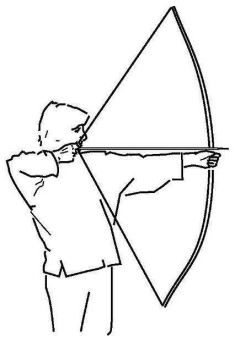
The problem of multiple hypothesis testing

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

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

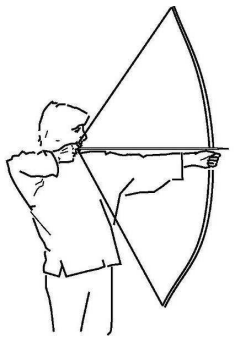
Bad solution to the multiple testing problem

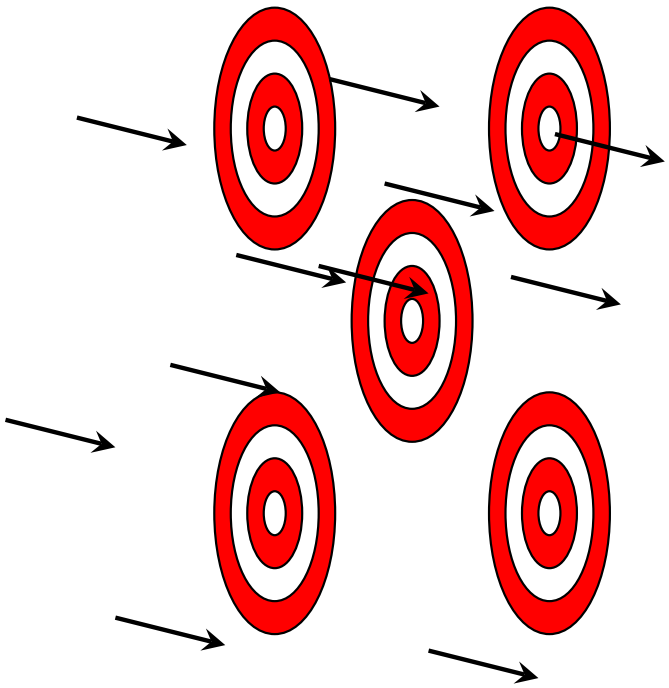
The big DON'T: It is **not** permissible to perform several tests and only present those that gave the desired outcome.



Bad solution to the multiple testing problem

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





All-against-all correlations

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

All-against-all correlations

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

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

Large scale T-testing

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

Large scale T-testing

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

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	

Large scale T-testing

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

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!

Other cases where multiple testing occurs

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.).

Other cases where multiple testing occurs

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.

Other cases where multiple testing occurs

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

False positive rate under multiple tests

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

False positive rate under multiple tests

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.

False positive rate under multiple tests

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.

False positive rate under multiple tests

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

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's P-value correction

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.

Bonferroni's P-value correction

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 likelihood of getting a Bonferroni corrected P-value less than α for a true null-hypothesis is at most α.

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	-

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

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	-

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	

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.

Most micro arrays now contains more than 40.000 probes: too many to test them one by one and hope that they can survive Bonferroni correction.

Bonferroni's P-value correction

Bonferroni correction is the most common multiple testing correction:

- It is very simple.

Bonferroni's P-value correction

Bonferroni correction is the most common multiple testing correction:

- It is very simple.
- It is always correct: no model assumptions, no assumption of independence.

Bonferroni's P-value correction

Bonferroni correction is the most common multiple testing correction:

- It is very simple.
- It is always correct: no model assumptions, no assumption of independence.
- Gives one new P-value for each test.

Bonferroni's P-value correction

Bonferroni correction is the most common multiple testing correction:

- It is very simple.
- It is always correct: no model assumptions, no assumption of independence.
- Gives one new P-value for each test.
- Useable even if some hypotheses are false.

Bonferroni's P-value correction

Bonferroni correction is the most common multiple testing correction:

- It is very simple.
- It is 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).

Bonferroni's P-value correction

Bonferroni correction is the most common multiple testing correction:

- It is very simple.
- It is 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

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.

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.

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 very conservative.

Alternative P-value corrections

Exists less conservative methods.

Alternative P-value corrections

Exists less conservative methods.

Bonferroni–Holm Like Bonferroni, but correct the k-th smallest P-value with a factor $N + 1 - k$.

Alternative P-value corrections

Exists less conservative methods.

Bonferroni–Holm Like Bonferroni, but correct the k-th smallest P-value with a factor $N + 1 - k$.

Simes' procedure If there are k P-values less than q, the *over-all* P-value is at most $q \times N/k$.

Alternative P-value corrections

Exists less conservative methods.

Bonferroni–Holm Like Bonferroni, but correct the k-th smallest P-value with a factor $N + 1 - k$.

Simes' procedure If there are k P-values less than q, the *over-all* P-value is at most $q \times N/k$.

False discovery rate Relaxes the criteria by allowing *some* false positives.

Alternative P-value corrections

Exists less conservative methods.

Bonferroni–Holm Like Bonferroni, but correct the k-th smallest P-value with a factor $N + 1 - k$.

Simes' procedure If there are k P-values less than q, the *over-all* P-value is at most $q \times N/k$.

False discovery rate Relaxes the criteria by allowing *some* false positives.

Some procedures (e.g. Simes') require caution: test the *over-all* hypothesis that *all* the null-hypotheses are true. Need not tell you which of the null-hypotheses are rejected, only that they cannot all be retained.

Over-all tests of multiple hypotheses

Example: We compute all pairwise correlations for 10 variables (that's 45 pairs). The smalles P-values we get are 0.0014, 0.0021, 0.0025 and 0.0031. None of these would survive the Bonferroni correction.

Simes' procedure would give an over-all P-value of $0.0031 \times 45/4 = 0.035$. However, it would be wrong to conclude that all four of these correlations are non-zero at the 5% significance level.

Over-all tests of multiple hypotheses

Example: We compute all pairwise correlations for 10 variables (that's 45 pairs). The smalles P-values we get are 0.0014, 0.0021, 0.0025 and 0.0031. None of these would survive the Bonferroni correction.

Simes' procedure would give an over-all P-value of $0.0031 \times 45/4 = 0.035$. However, it would be wrong to conclude that all four of these correlations are non-zero at the 5% significance level.

Over-all tests are often more powerful than e.g. Bonferroni, but lead to conclusions that are harder to interpret and explain.

One approach to multiple testing

How to interpret and present P-values in a multiple testing setting:

One approach to multiple testing

How to interpret and present P-values in a multiple testing setting:

P-value survives Bonferroni correction: Corrected P-value is reliable.

One approach to multiple testing

How to interpret and present P-values in a multiple testing setting:

P-value survives Bonferroni correction: Corrected P-value is reliable.

Over-all test is not statistically significant: No reason to believe there are any statistically significant P-values.

One approach to multiple testing

How to interpret and present P-values in a multiple testing setting:

P-value survives Bonferroni correction: Corrected P-value is reliable.

Over-all test is not statistically significant: No reason to believe there are any statistically significant P-values.

Conflict: If the uncorrected P-value is statistically significant, but Bonferroni corrected is not, proceed with caution! This may indicate a possible, but unreliable, finding.

Another approach to multiple testing

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

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 performe more analyses, get more results and test more hypotheses.

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 performe more analyses, get more results and test more hypotheses.

One compromise is to divide analyses into two parts:

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 performe 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.

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 performe 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.

Outline

Hypothesis testing

Multiple hypothesis testing

P-value correction

Multiple comparisons

Multiple comparisons

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

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.

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.

Multiple comparisons

Group	x LSMEAN	95% Confidence Limits	
1	1.864723	1.194959	2.534487
2	0.606615	-0.063149	1.276378
3	2.621182	1.951418	3.290945
4	0.789182	0.119418	1.458946
5	1.196442	0.526678	1.866206
6	3.397056	2.727292	4.066820

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Group	5	60.21737137	12.04347427	10.79	<.0001

ANOVA shows that there is variation between the treatments (P<0.0001), but this does not tell us which treatments differ.

All-against-all comparisons

Tukey: Adjustment of P-values for all-against-all T-tests.

Tukey	1	2	3	4	5	6
1		0.0999	0.6014	0.2215	0.7182	0.0236
2	0.0999		0.0011	0.9988	0.8109	<.0001
3	0.6014	0.0011		0.0037	0.0429	0.5749
4	0.2215	0.9988	0.0037		0.9538	<.0001
5	0.7182	0.8109	0.0429	0.9538		0.0003
6	0.0236	<.0001	0.5749	<.0001	0.0003	

Tukey		Mean	N	gr
	A	3.3971	10	6
B	A	2.6212	10	3
B	C	1.8647	10	1
	C	1.1964	10	5
	C	0.7892	10	4
	C	0.6066	10	2

One-against-all comparisons

Dunnet: Adjustment of P-values for one-against-all T-tests.

Group	x LSMEAN	Pr > |t|
1	1.86472306	
2	0.60661459	0.0419
3	2.62118151	0.3680
4	0.78918174	0.1028
5	1.19644178	0.4854
6	3.39705568	0.0090

E.g. if group 1 is placebo or the standard treatment against which the others should be compared.

Some links

StatSoft textbook Good overview of methods and concepts: http://statsoft.com/textbook/stathome.html

SAS manuals Thorough with overview of analysis procedures found in SAS: http://support.sas.com/onlinedoc/913/