Introduction to statistical inference and multiple hypothesis testing

Clara-Cecilie Günther

clara-cecilie.gunther@nr.no

MBV-INF4410/9410/9410A

21.11.2012

What is NR?

- Private non-profite foundation
- Applied research within
  - Statistical modelling and analysis
  - Information and communication technology
- Founded in 1952
- 65 researchers

Statistics is important in many fields

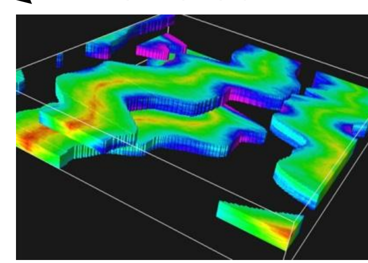
Finance

Marine

Statistical modelling and analysis

Medicine

Petroleum



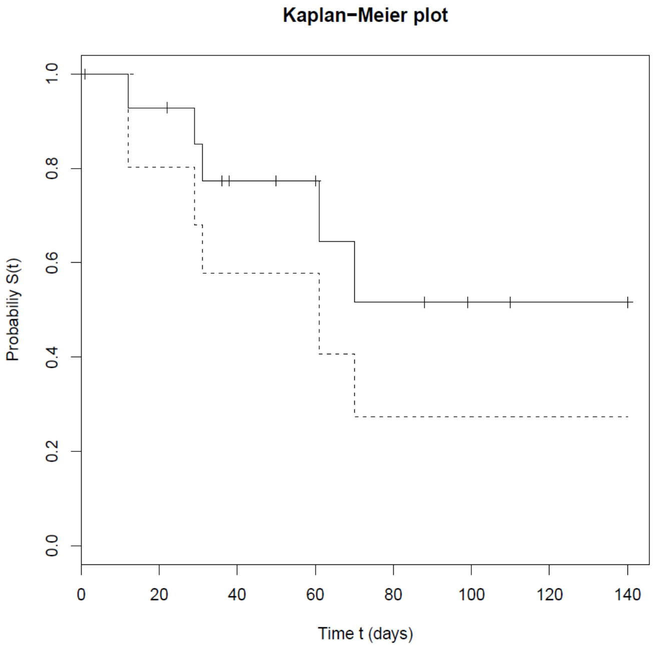
We also do statistical genomics

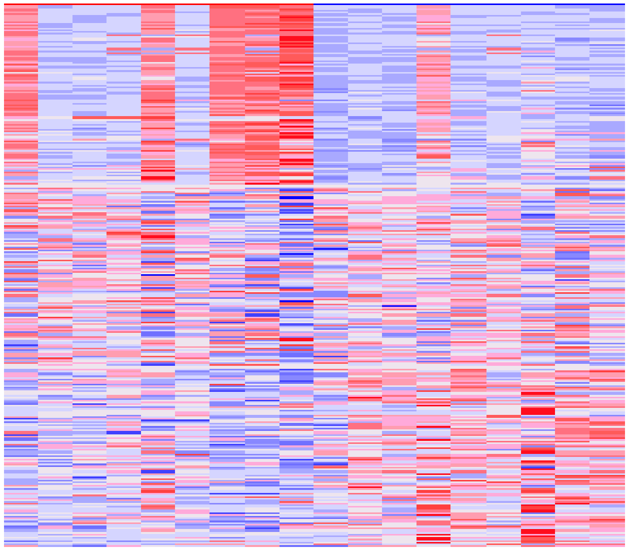
- NR participates in the bioinformatics core facility.
- Data: Microarrays, SNP, copy number, sequencing, methylation.
- Typical tasks:
  - Find list of differentially expressed genes.
  - Pathway analyses.
  - Correlation analyses.
  - Survival analyses.
  - Sample size calculations.
  - Multiple testing adjustment.
  - Clustering.

www.ancestor-dna.com

Projects

- Survival and colorectal cancer
- Survival and cervical cancer
- Oxygen-induced complications of prematurity
- Association between SNPs and low back pain
- Genes associated with BMD and osteoporosis
- Antioxidant-rich food and gene expression in blood cells





Outline

- Hypothesis testing – the general idea.
- Important aspects to know
  - Null and alternative hypothesis.
  - P-value.
  - Type I and type II errors.
  - Rejection area.
  - Significance level and power.
- Some common tests.
- Alternative ways to calculate p-values.
- Multiple hypothesis testing.

Statistical inference

Population:

The collection of subjects that we would like to draw conclusions about.

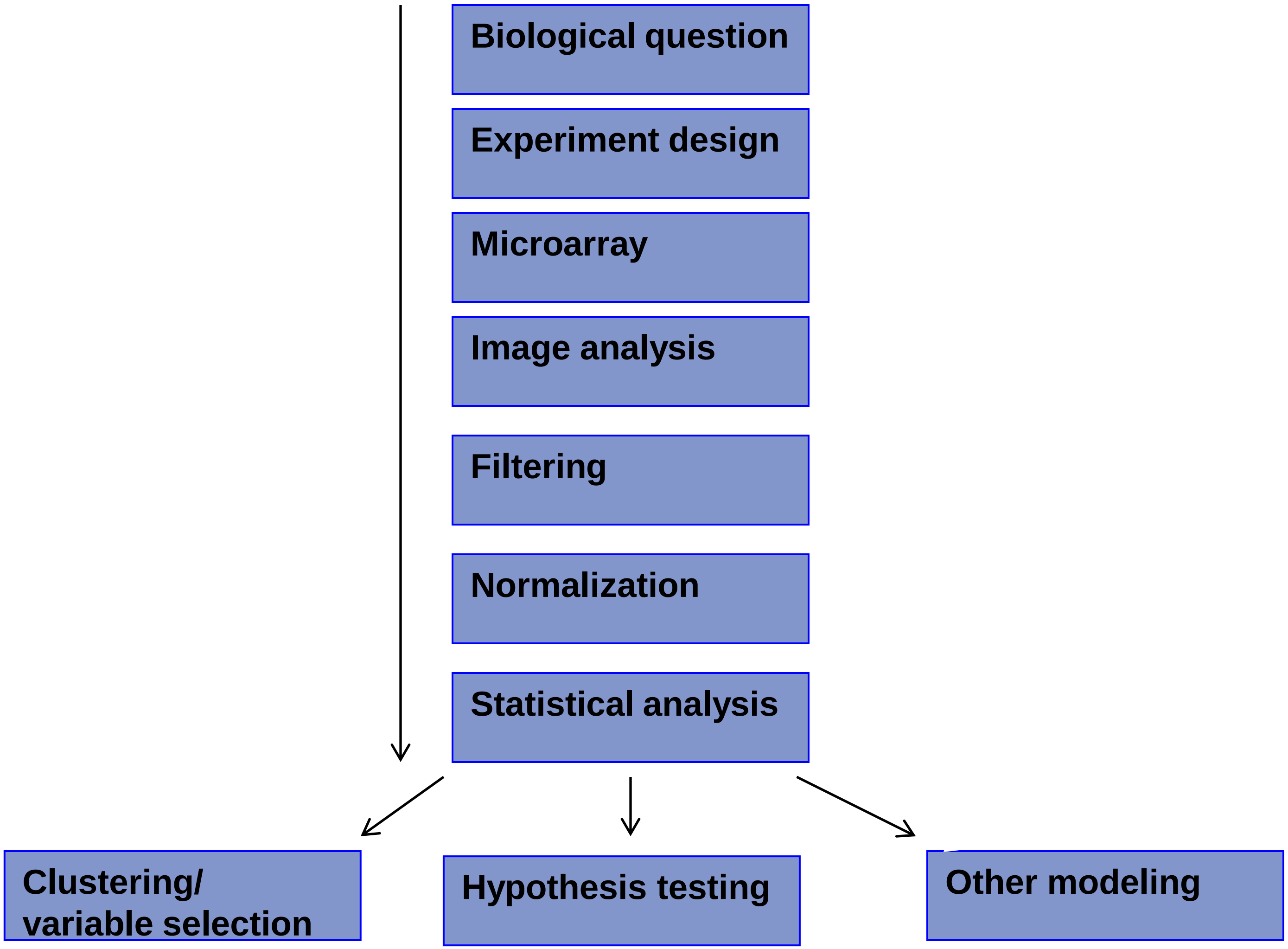
Sample:

The subcollection considered in the study

Statistical inference:

Draw sample-based conclusions about the population, controlling for the probability of making false claims.

Example: Analysis of microarray data



Hypothesis testing

- The results of an experiment can be summarized by various measures
  - average
  - standard deviation
  - diagrams

- But often the aim is to choose between competing hypotheses concerning these measures.

Hypothesis testing

- Typical: have data and information
  - Uncertainty attached to these
  - Must draw a conclusion
  - Examples
    - Is the new medicine better than the old one?
    - Are these genes differentially expressed in tumor and normal cells?

- Hypothesis testing
  - Method to draw conclusions from uncertain data
  - Can say something about the uncertainty in the conclusion

Statistical tests (the idea)

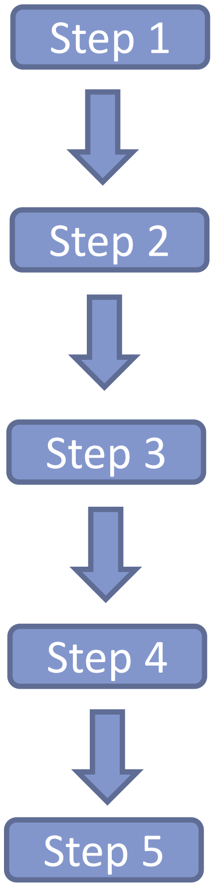
1) A population has individuals with an observable feature X that follows X ~ F(θ). We seek if (say) $\theta = 0$ is violated.

2) We obtain X-values $X_1, ... X_N$ on a random sample.

3) A test statistic $Z = Z(X_1, ... X_N)$ is defined. The observed Z is denoted z_{obs}. Large $|z_{obs}|$ supports violations.

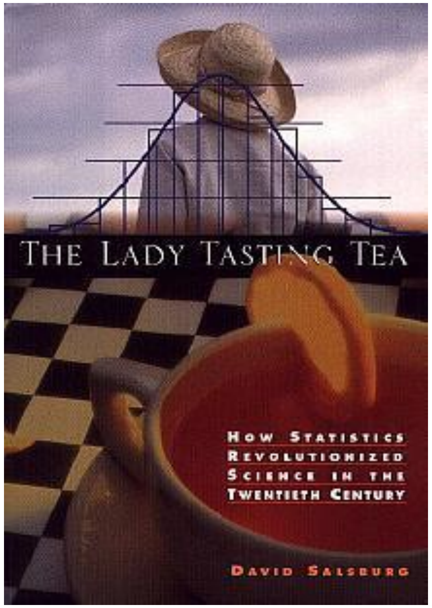
4) Calculate the probability that $|Z| \geq |z_{obs}|$ (= p-value)

5) Conclude that $\theta = 0$ is violated if p-value is small.



Famous example:

- *The Design of Experiments* (1935), Sir Ronald A. Fisher
  - A tea party in Cambridge, the 1920ties
  - A lady claims that she can taste whether milk is poured in cup before or after the tea
  - All professors agree: impossible
  - Fisher: this is statistically interesting!
  - Organised a test



http://www.maa.org/reviews/ladytea.html

The lady tasting tea

- Test with 8 trials, 2 cups in each trial
  - In each trial: guess which cup had the milk poured in first

- Binomial experiment
  - Independent trials
  - Two possible outcomes, she guesses right cup (success), wrong cup (failure)
  - Constant probability of success in each trial

- X=number of correct guesses in 8 trials, each with probability of success p
  - X is Binomially (8,p) distributed $P(X=x)=\binom{8}{x}p^x(1-p)^{(8-x)}$

The lady tasting tea, cont.

- The null (conservative) hypothesis H_0
  - The one we initially believe in
- The alternative hypothesis H_1
  - The new claim we wish to test

- H_0 She has no special ability to taste the difference $p = 0.5$
- H_1 She has a special ability to taste the difference $p > 0.5$

How many right to be convinced?

- We expect maybe 3, 4 or 5 correct guesses if she has no special ability
- Assume 7 correct guesses
  - Is there enough evidence to claim that she has a special ability? If 8 correct guesses this would have been even more obvious!
  - What if only 6 correct guesses?
    - Then it is not so easy to answer YES or NO
- Need a rule that says something about what it takes to be convinced.

How many right to be convinced?

- Rule: We reject H_0 if the observed data have a small probability under H_0 (given H_0 is true).

- Compute the p-value.
  - The probability to obtain the observed value or something more extreme, given that H_0 is true
  - **NB!** The P-value is NOT the probability that H_0 is true

Small p-value: reject the null hypothesis
Large p-value: keep the null hypothesis

The lady tasting tea, cont.

- Say: she identified 6 cups correctly

- P-value

$$
\begin{aligned}
&\boldsymbol{P(X \geq 6 | H_0 \text{ true})} \\
&\quad \boldsymbol{= P(X = 6 | p = 0.5) + P(X = 7 | p = 0.5) + P(X = 8 | p = 0.5)} \\
&\quad \boldsymbol{= 0.1094 + 0.0313 + 0.0039 = 0.1443}
\end{aligned}
$$

- Is this enough to be convinced?
- Need a limit.
  - To set it, we must know about the types of errors we can make.

P-value

The probability to obtain the observed value or something more extreme, given that H_0 is true

Two types of errors

	H_0 true	H_1 true
Accept H_0	OK	Type II error
Reject H_0	Type I error	OK

- Type I error most serious
  - Wrongly reject the null hypothesis
  - Example:
    - H_0 : person is not guilty
    - H_1 : person is guilty
    - To say a person is guilty when he is not is far more serious than to say he is not guilty when he is.

When to reject

- Decide on the hypothesis' level of significance
  - Choose a level of significance α
  - This guarantees P(type I error) ≤ α
  - Example
    - Level of significance at 0.05 gives 5 % probability to reject a true H_0

- Reject H_0 if P-value is less than α

Important parameters in hypothesis testing

- Null hypothesis
- Alternative hypothesis
- Level of significance

Must be decided upon before we know the results of the experiment

The lady tasting tea, cont.

- Choose 5 % level of significance

- Conduct the experiment
  - Say: she identified 6 cups correctly
  - Is this evidence enough?

- P-value

$$\begin{aligned} & \boldsymbol{P(X \geq 6 | H_0 \text{ true})} \\ & \boldsymbol{= P(X = 6 | p = 0.5) + P(X = 7 | p = 0.5) + P(X = 8 | p = 0.5)} \\ & \boldsymbol{= 0.1094 + 0.0313 + 0.0039 = 0.1443} \end{aligned}$$

P-value

The probability to obtain the observed value or something more extreme, given that H_0 is true

The lady tasting tea, cont.

- We obtained a p-value of 0.1443

- The rejection rule says
  - Reject H_0 if p-value is less than the level of significance α
  - Since α = 0.05 we do NOT reject H_0

Small p-value: reject the null hypothesis
Large p-value: keep the null hypothesis

The lady tasting tea, cont.

- In the tea party in Cambridge:
  - The lady got every trial correct!

- Comment:
  - Why does it taste different?
    - Pouring hot tea into cold milk makes the milk curdle, but not so pouring cold milk into hot tea*

*http://binomial.csuhayward.edu/applets/appletNullHyp.html
Curdle = å skille seg

Area of rejection

Reject H_0 if p-value $\leq \alpha$

Reject H_0 if observed x-value $\geq$ critical value

P(type I error)=P(reject H_0 | H_0 true)

$$= P(X \geq x_c | p = 0.5) = \sum_{x=x_c}^{8} \binom{8}{x} \left(\frac{1}{2}\right)^x \left(1 - \frac{1}{2}\right)^{8-x}$$

$$x_c = 7 \rightarrow P(\text{type I error}) = 0.03516 < \alpha$$

$$x_c = 6 \rightarrow P(\text{type I error}) = 0.1443 > \alpha$$

Area of rejection: $\{x: x \geq x_c\} \rightarrow \{x: x \geq 7\}$

NB! X's distribution discrete → no $x_c : P(X \geq x_c | H_0)$ exactly α

x_c lowest possible x-value such that $P(X \geq x_c | H_0) \leq \alpha$

Type II error

$P(\text{Type I error}) \leq \alpha$ $\qquad$ $P(\text{Type II error}) = \beta$

Want both errors as small as possible, especially type I.

β is not explicitly given, depends on H_1.

There is one β for each possible value of *p* under H_1.

	H_0 true	H_1 true
Accept H_0	OK	Type II error
Reject H_0	Type I error	OK

Example, type II error

$P(\text{type II error}) = P(\text{not reject } H_0 \mid H_1 \text{ true})$

$\mathrm{p} = 0.7$:

$$= P(\text{not reject } H_0 | p = 0.7) = 1 - P(\text{reject } H_0 | p = 0.7)$$

$$= 1 - P(X \geq 7 | p{=}0.7) = 1 - (1 - P(X < 7 | p = 0.7))$$

$$= P(X \leq 6 | p = 0.7) = \sum_{x=0}^{6} \binom{8}{x} 0.7^x (1 - 0.7)^{8-x} = 0.7447$$

If $\mathrm{p}=0.7 \rightarrow$ wrongly accept H_0 in 74.47% of times.

Power of the test

The probability that a false H_0 is rejected

$$P(\text{reject } H_0 | H_1 \text{ true}) = 1 - P(\text{accept } H_0 | H_1 \text{ true}) = 1 - \beta$$

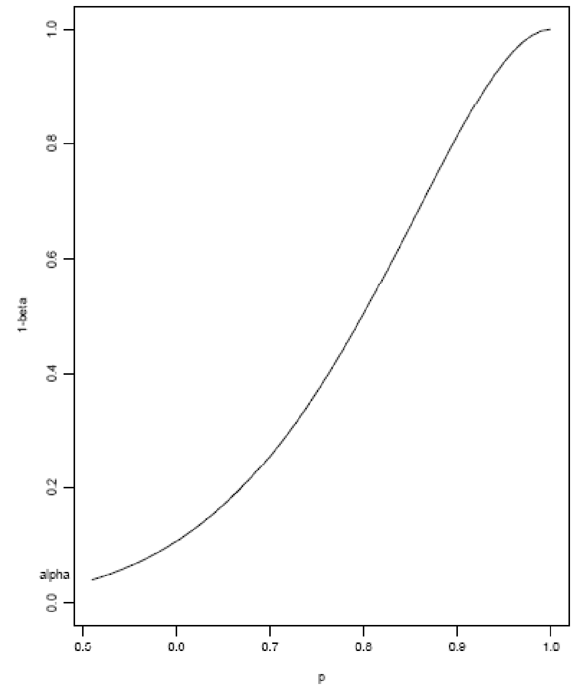
Test with large power:
larger probability to draw the right conclusion
to reject a false null hypothesis than a test with low power

Because of the connection between α and β will decreasing α also decrease the power of the test.

Power function

Probability	β	Power
0.6	0.8936243	0.10637568
0.7	0.7447017	0.25529833
0.8	0.4966835	0.50331648
0.9	0.1868953	0.81310473
0.99	0.002690078	0.9973099

Power function



Expand the number of trials to 16

Assume she guesses 12 correct (12 of 16, before 6 of 8)

P-value $= P(X \geq 12 | H_0 \text{ true}) = 0.038 \rightarrow H_0$ rejected!

Significance probability dropps from 0.1443 til 0.038

Point: we tend to think "proportionally" $\rightarrow$ wrong!
The lower number of trials, the more often we register biased outcomes

The proportionally equal good result becomes more significant

Expand the number of trials, cont.

$n = 16 \rightarrow x_c = 12$

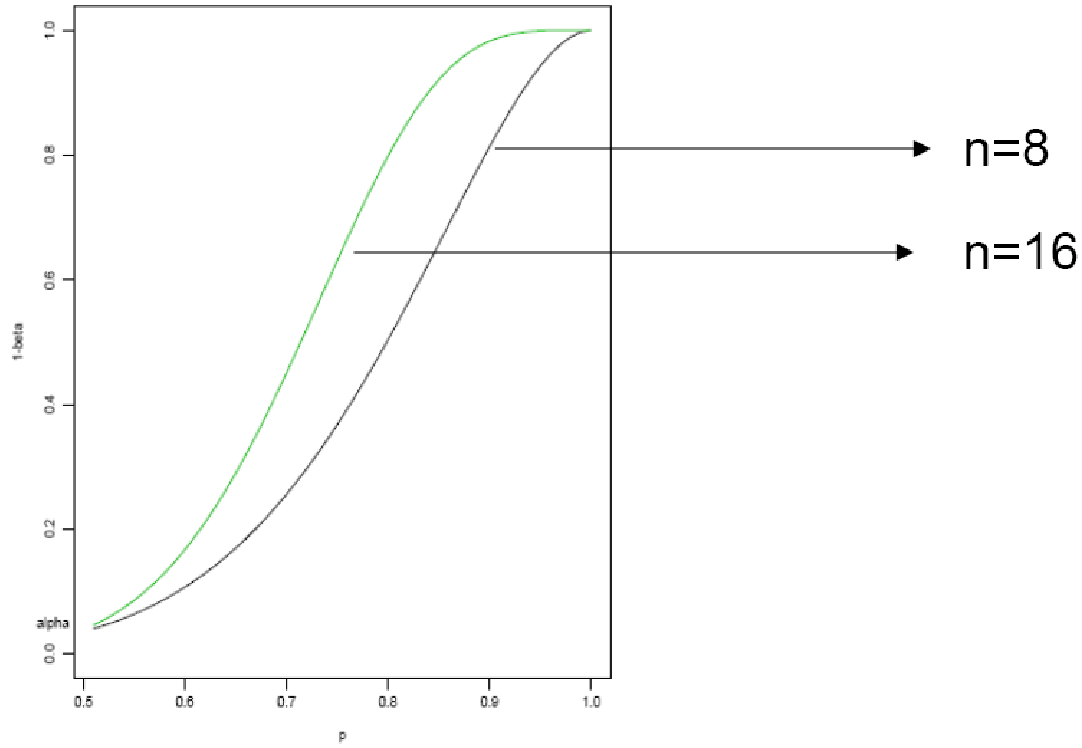
$\mathrm{p} = 0.7$

$P(\text{type II error}) = \sum_{x=0}^{11} \binom{16}{x} (0.7)^x (1-0.7)^{16-x} = 0.5501$

Probability for type II error at $p = 0.7$ dropps from 0.7747 to 0.5501

Expand the number of trials, cont.

Compare power curves



Parallel to experiments: do replications to increase power!

Statistical tests (the idea)

1) A population has individuals with an observable feature X that follows X ~ F(θ). We seek if (say) $\theta = 0$ is violated.

2) We obtain X-values $X_1, \ldots X_N$ on a random sample.

3) A test statistic $Z = Z(X_1, \ldots X_N)$ is defined. The observed Z is denoted z_{obs}. Large $|z_{obs}|$ supports violations.

4) Calculate the probability that $|Z| \geq |z_{obs}|$ (= p-value)

5) Conclude that $\theta = 0$ is violated if p-value is small.

Step 1

Step 2

Step 3

Step 4

Step 5

Common tests

One sample location tests

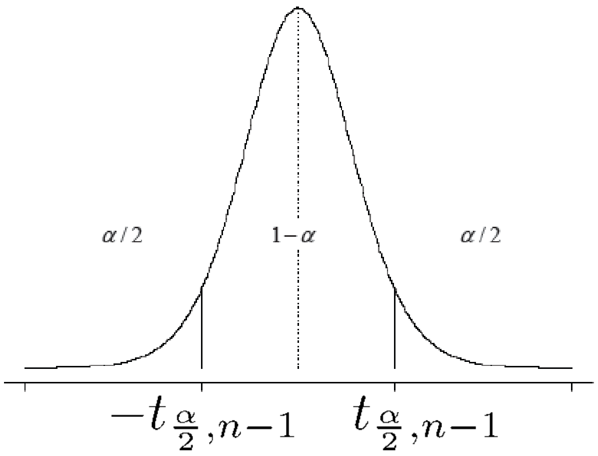
- Purpose: Compare location parameter of a population to a known constant value.
- Example:
- One sample z-test
- One sample t-test
- One sample Wilcoxon signed ranked test (when normality cannot be assumed)

The one sample t-test

- Data:
  - Y = log intensity value of gene.
  - Assume

$$Y_1, ..., Y_n \sim N(\mu, \sigma^2)$$

- Test: $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$



- Under H_0, the test statistic $t = \frac{\bar{Y}-\mu_0}{s/\sqrt{n}}$ is T distributed with n-1 degrees of freedom

$$H_0 \text{ is rejected if } t_{\text{obs,n}-1} \leq -t_{\frac{\alpha}{2},n-1} \text{ or } t_{\text{obs,n}-1} \geq t_{\frac{\alpha}{2},n-1}$$

$$p-\text{value}=P\left(T \geq t_{obs,n-1} \middle| H_0\right) + P\left(T \leq -t_{obs,n-1} \middle| H_0\right) = 2 \cdot P(T \geq t_{obs,n-1} | H_0)$$

Two sample tests

- Purpose: Compare the mean of two different groups.
- Two types of problems:
  - Two treatments – same subjects:
    - Measure cholesterol level before and after diet
    - Measure gene expression in tumor cell before and after radiation.
  - Same treatment – two groups of subjects:
    - Measure cholesterol level in men and women.
    - Intervention study: One group given antioxidant enriched diet, another antioxidant deprived diet. Measure difference in change in gene expression.

Two-sample problems: Paired data

Ex.: Measurements of cholesterol level

- H_0: no effect of the diet

Person no.	Before diet	After diet	D (difference)
1	5.69	2.39	5.69-2.39=3.30
2	5.90	5.40	5.90-5.40=0.50
3	4.65	4.05	
4	4.09	2.31	
5	6.38	5.79	
6	5.38	4.34	
7	6.55	5.74	
8	6.39	5.48	
9	7.00	6.01	
10	8.31	5.41	8.31-5.41=2.90

- t=4.247
- Degrees of freedom: 10-1=9
- P-value (two-sided test)
  $2 \cdot P(T_9 \geq 4.247) \approx 0.002 < 0.05$
- Conclusion: reject H_0

$$X_{1i} = \text{measure person } i \text{ before diet}$$
$$X_{2i} = \text{measure person } i \text{ after diet}$$

$$X_{1i} \sim N(\mu_1, \sigma_1^2) \; X_{2i} \sim N(\mu_2, \sigma_2^2)$$

$$H_0 : \mu_1 - \mu_2 = 0 \; H_1 : \mu_1 - \mu_2 \neq 0$$

Test statistic

$$t = \frac{\bar{D} - 0}{sd(\bar{D})} = \frac{\bar{D}}{sd(D)/\sqrt{n}}, \; D = X_1 - X_2$$

is T-distributed under H_0

with *n-1* degrees of freedom (*n=n1=n2*)

Two-sample problems: different samples

Ex.: Measurements of cholesterol level, 12 men and 9 women

	Men ($X1$)	Women ($X2$)
1	9.65	6.11
2	5.17	4.70
3	6.48	6.87
4	7.58	7.20
5	6.50	8.49
6	6.09	7.07
7	5.75	6.58
8	7.99	7.02
9	5.63	6.62
10	8.05	
11	8.88	
12	6.28	

- t=0.48
- P-value (two-sided test) $2{*}P(T_{19} \geq 0.48) \approx 0.64 > 0.05$
- Conclusion: Do not reject H_0

$$X_{1i} = \text{measure man } i$$
$$X_{2j} = \text{measure woman } j$$

$$X_{1i} \sim N(\mu_1, \sigma_1^2) \; X_{2i} \sim N(\mu_2, \sigma_2^2)$$

$$\text{Assume } \sigma_1^2 = \sigma_2^2$$

$$H_0 : \mu_1 - \mu_2 = 0 \; H_1 : \mu_1 - \mu_2 \neq 0$$

Test statistic

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_f \sqrt{\frac{1}{n_1} \frac{1}{n_2}}}$$

$$\text{where } s_f = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$$

is under H_0 t-distributed with n_1+n_2-2 degrees of freedom
s_f is a common std.dev. for both groups
s1 and *s2* are the empirical std.dev. of X_1 and X_2, respectively

More ways to calculate p-values

- So far, all p-values have been calculated from

  $P(|Z| \geq |z_{obs}| \mid H_0)$.

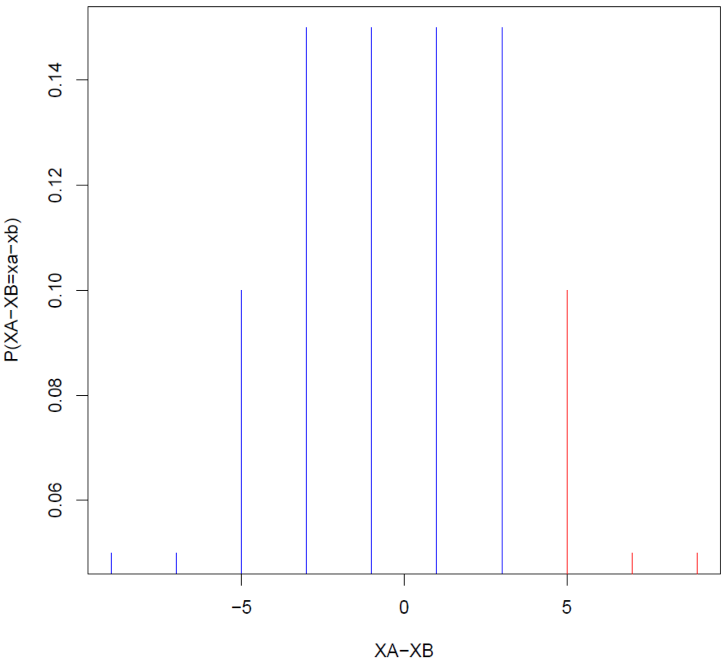
- This is easy when the distribution of Z is known (e.g. binomial, normal, student t).
- Often the distribution of Z is not known.
- Can use permutation tests instead.
  - Find the distribution of Z by permutations.

Simple example

- Two groups, three measurements in each group.
- X_A: 8, 11, 12. X_B: 7, 9, 10.
- We want to test if Z=$\sum_{i=1}^{3} X_{A,i} - \sum_{i=1}^{3} X_{B,i}$=0 vs Z>0.
- The observed value: Z=31-26=5.
- How likely is Z=5 under the null hypothesis?
  - Do not know the distribution of Z.
- Solution: Permute the labels A and B.
  - Find all possible ways to permute the measurements in two groups with three observations in each group.

Simple example: Permutation p-value

A	B	Z	A	B	Z
8, 11,12	7, 9, 10	5	7, 9, 10	8, 11, 12	-5
7, 11, 12	8, 9, 10	3	8, 9, 10	7, 11, 12	-3
7, 8, 12	9, 10,11	-3	9, 10,11	7, 8, 12	3
7, 8, 11	9, 10, 12	-5	9, 10, 12	7, 8, 11	5
9, 11, 12	7, 8, 10	7	7, 8, 10	9, 11, 12	-7
8, 9, 12	7, 10, 11	1	7, 10, 11	8, 9, 12	-1
8, 9, 11	7, 10, 12	-1	7, 10, 12	8, 9, 11	1
10, 11, 12	7, 8, 9	9	7, 8, 9	10, 11, 12	-9
8, 10, 12	7, 9, 11	3	7, 9, 11	8, 10, 12	-3
8, 10, 11	7, 9, 12	1	7, 9, 12	8, 10, 11	-1



$$p - \text{value}{=}\mathrm{P}(Z \geq 5) = 0.1 + 0.05 + 0.05 = 0.20$$

Do not reject the null hypothesis.

This p-value is exact.

Often, the number of possible permutations is huge

- Example: 30 individuals, 15 cases and 15 controls.
- Number of possible permutations $\binom{30}{15}$ = 155 117 520.
- Impossible to calculate test statistic for alle permutations.
- Instead we can sample the case/control labels randomly a large number of times.
- Get approximate p-value.
- This is called Monte Carlo sampling

Permutation tests – general example

- Data: Gene set measurements for cases and control group.
- For each gene i=1,…,n, a test statistic t_i is calculated.
- Permute the case and control labels → new dataset
- Calculate new ${t_{i,b}}^*$ for the permuted sample.
- Repeat B times, B=10 000 or 100 000.
- The ${t_{i,b}}^*$, *b=1,…,B* now form a distribution for t_i under the null hypothesis.
- The p-value of t_i can be calculated as

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

General example - illustration

Original data

genes

	cases				controls				
	1	2	3	…	16	17	…	30	
1	53	42	11	…	135	69	…	88	t_1
2	256	34	143	…	57	29	…	192	t_2
.									
.									
n	72	153	86	…	120	134	…	356	t_n

$$p_1 = \frac{\#|t_{1,b}{}^*| \geq |t_1|}{B}$$

$$p_2 = \frac{\#|t_{2,b}{}^*| \geq |t_2|}{B}$$

.
.
.

$$p_n = \frac{\#|t_{n,b}{}^*| \geq |t_n|}{B}$$

Permutation data

	cases				controls				
	7	4	29	...	1	18	9	…	
1	35	93	45	…	53	103	68		$t_{1,1}{}^*$
2	189	103	38	…	256	39	97		$t_{2,1}{}^*$
.									
.									
n	238	255	108	…	72	194	86		$t_{n,1}{}^*$

.
.
.

	cases				controls				
	16	3	23	…	2	25	8	…	
1	135	11	98	…	42	103	293		$t_{1,B}{}^*$
2	57	143	115	…	34	204	142		$t_{2,B}{}^*$
.									
.									
n	120	86	53	…	153	122	94		$t_{n,B}{}^*$

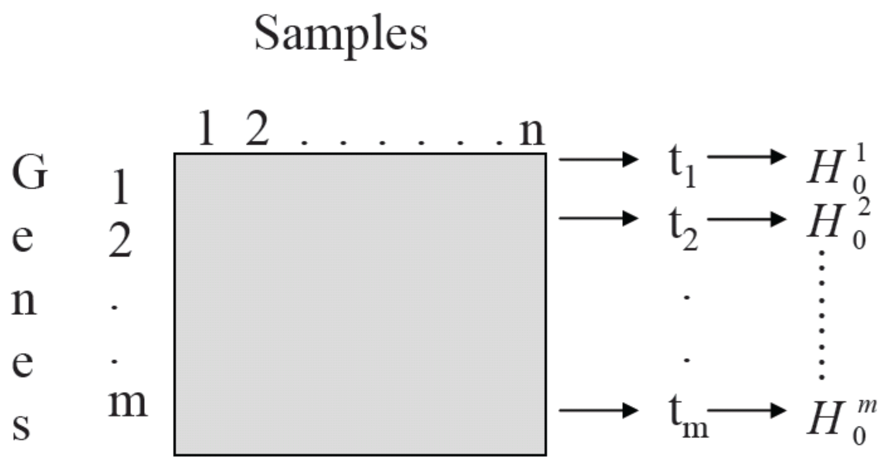
Examples of use of permutation tests

- SAM
  - Differential expression.
- GSEA
  - Enrichment of gene sets.
- Hyperbrowser
  - Many different applications.

Multiple hypothesis testing

Often we don't test just one hypothesis

- Instead
  - Large number of hypotheses tested simultaneously



- Many genes $\rightarrow$ many hypotheses tested simultaneously
  - H_0^i gene number *i* is not differentially expressed
  - $p_1, \ldots, p_m$ are the *p*-values associated with each test statistic

Example: 10 000 genes

- Q: is gene g, g = 1, …, 10 000, differentially expressed?
- Gives 10 000 null hypothesis: $H_0^1, \ldots, H_0^{10\,000}$
  - H_0^1 : gene 1 not differentially expressed
  - …
- Assume: no genes differentially expressed, i.e. H_0^g true for all g
- Significance level $\alpha = 0.01$
  - The probability to incorrectly conclude that one gene is differentially expressed is 0.01.

$\text{Level of significance } \alpha \text{ guarantee that } P(\text{type I error}) \leq \alpha$
$P(\text{type I error}) = P(\text{reject } H_0 | H_0 \text{ true})$

Example: 10 000 genes, cont.

- Significance level $\alpha = 0.01$
  - When 10 000 tests:
    - Expect $10\,000 \cdot \alpha = 10\,000 \cdot 0.01 = 100$ genes to have p-value below 0.01 by chance

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

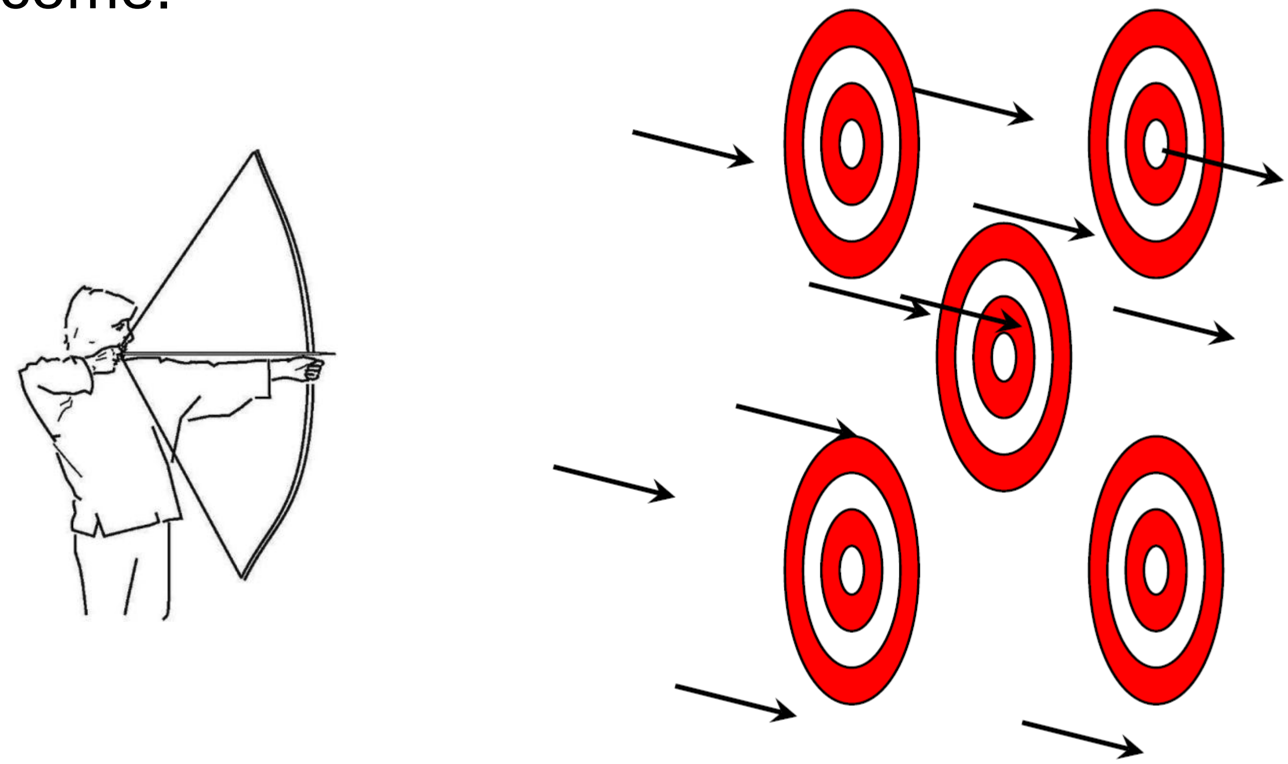
- Many tests → many false positive conclusions
  - This is not good!

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 control 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.



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
sign_germB Germinal center B cell sign.	1.00000	0.16336 0.0113	-0.05530 0.3938	-0.08362 0.1967
sign_lymph Lymph node signature	0.16336 0.0113	1.00000	-0.31586 <.0001	-0.02660 0.6818
sign_prolif Proliferation signature	-0.05530 0.3938	-0.31586 <.0001	1.00000	0.14079 0.0292
BHP6 BMP6	-0.08362 0.1967	-0.02660 0.6818	0.14079 0.0292	1.00000
MHC MHC class II signature	0.17837 0.0056	0.15082 0.0194	-0.13411 0.0379	0.08650 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. Perform t-test for each gene.
- H_i^0: gene i is not differentially expressed, i=1,…,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-values is inappropriate when we have done 100 tests!

Other cases where multiple testing occurs

- 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.)
- A researcher wants to check if health depends on social status. Both health and status can be measured in many different, although similar ways. He checks all combinations.
- A researcher cannot decide which is more appropriate to use: Pearson correlation or Spearman. He picks the one with the lowest p-value.

Corrected p-values

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

- 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 it is also true when some of the null-hypotheses are actually wrong.
- May use this to formulate a multiple test that controls the overall risk of having a false positive.

Bonferroni's p-value correction

- 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.
- If you run N tests, multiply all the p-values by N to get the Bonferroni corrected p-values.
- The probability of getting a Bonferroni corrected p-value less than α for a true null-hypothesis is at most α.

Large scale t-testing

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

Microarrays now contain more than 40 000 probes: Too many to test them one by one and hope that they can survive the Bonferroni correction.

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

${H_0}^i$: gene i is not differentially expressed, i=1,…,40000.

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

i.e. if $p_i \leq 0.0000025$.

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

Bonferroni correction

- Remember:

The probability that a false H_0 is rejected:
$$P(\text{reject } H_0 | H_1 \text{ true}) = 1 - P(\text{accept } H_0 | H_1 \text{ true}) = 1 - \beta$$

Because of the connection between α and β will decreasing α also decrease the power of the test.

- Problem: very low power!

Summary of Bonferroni correction

It 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.

However, Bonferroni-correction is often conservative!

The problem of conservative corrections

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

Alternative p-value corrections

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

$$V = \text{no. of type I errors [false positives]}$$
$$T = \text{no. of type II errors [false negatives]}$$

$$P(\text{type I error}) = P(\text{reject } H_0 | H_0 \text{ true})$$
$$P(\text{type II error}) = P(\text{accept } H_0 | H_1 \text{ true})$$

Family-wise error rate (FWER)

- The probability of at least one type I error
  - $$\mathrm{FWER} = P(V \geq 1)$$

- Control FWER at a level α
  - Procedures that modify the adjusted p-values separately
    - Single step procedures
  - More powerful procedures adjust sequentially, from the smallest to the largest, or vice versa
    - Step-up and step-down methods

- The Bonferroni correction controls the FWER

$$\boxed{\begin{array}{l} V = \text{no. of type I errors [false positives]} \\ P(\text{type I error}) = P(\text{reject } H_0 | H_0 \text{ true}) \end{array}}$$

Methods that control the FWER

- Bonferroni
- Sidak
- Bonferroni-Holm
- Westfall & Young

False discovery rate (FDR)

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

V = no. of type I errors [false positives]
R = total no. of rejected H_0
$P(\text{type I error}) = P(\text{reject } H_0 | H_0 \text{ true})$

The Benjamini and Hochberg procedure

- Let $p_{(1)}, \ldots, p_{(n)}$ be an ordering of $p_1, \ldots, p_n$
- Let $H_0^{(1)}, \ldots, H_0^{(n)}$ be the corresponding null hypotheses
- The following adjusted p-values $\tilde{p}_{(i)}$ control the FDR when the unadjusted p-values p_i are independently distributed

$$\tilde{p}_{(i)} = min_{k \in \{i, \ldots, n\}} \frac{n \cdot p_{(k)}}{k}$$

- Variations exist (higher power)

Example: Adjusting to control the FDR

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

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

Guidelines

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

- Are you most afraid of getting stuff on your significant list that should not have been there?
  - Choose FWER.
- Are you most afraid of missing out on interesting stuff?
  - Choose FDR.

Summary

- 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.