General statistics

From study design to multiple testing

Eva Freyhult

Department of Medical Sciences, Uppsala University
National Bioinformatics Infrastructure Sweden
Science for Life Laboratory

February 14, 2017

Overview

Study design

Hypothesis testing

Power analysis (sample size calculation)

Multiple testing

Study design

Important concepts

Nuisance variables Variables that influence the experiment (increase variability), but are not of primary interest.

Study design

Important concepts

Nuisance variables Variables that influence the experiment (increase variability), but are not of primary interest.

- batch
- run order
- temperature
- time of day/year
- age
- gender
- ...

Study design

Important concepts

Nuisance variables Variables that influence the experiment (increase variability), but are not of primary interest.

Confounding variable A nuisance variable that changes as the primary variable changes.

Study design

Example

	batch	runorder	treatment
1	1	1	1
2	1	2	1
3	1	3	1
4	1	4	1
5	1	5	1
6	1	6	1
7	2	7	2
8	2	8	2
9	2	9	2
10	2	10	2
11	2	11	2
12	2	12	2

Bad design!

Avoid confounding!

Study design

Important concepts

Nuisance variables	Variables that influence the experiment (increase variability), but are not of primary interest.
Confounding variable	A nuisance variable that changes as the primary variable changes.
Randomization	randomize as much as possible to avoid bias.

Study design

Randomization

	batch	runorder	treatment
1	1	4	2
2	1	2	1
3	1	11	2
4	1	3	2
5	1	7	1
6	1	6	2
7	2	8	1
8	2	9	1
9	2	1	2
10	2	12	1
11	2	5	1
12	2	10	2

Study design

Important concepts

Nuisance variables	Variables that influence the experiment (increase variability), but are not of primary interest.
Confounding variable	A nuisance variable that changes as the primary variable changes.
Randomization	randomize as much as possible to avoid bias.
Blocking	use if the nuisance variable is known and controllable to reduce experimental variability.

Study design

Blocking

Within each block (batch), randomize treatment

	batch	runorder	treatment
1	1	4	1
2	1	2	2
3	1	11	1
4	1	3	2
5	1	7	1
6	1	6	2
7	2	8	2
8	2	9	2
9	2	1	1
10	2	12	2
11	2	5	1
12	2	10	1

General rule: *Block what you can; randomize what you cannot.*

Study design

Important concepts

Nuisance variables	Variables that influence the experiment (increase variability), but are not of primary interest.
Confounding variable	A nuisance variable that changes as the primary variable changes.
Randomization	randomize as much as possible to avoid bias.
Blocking	use if the nuisance variable is known and controllable to reduce experimental variability.
Replication	more biological replicates $\rightarrow$ higher power.

Hypothesis testing

H_0 the null hypothesis, e.g. $t = 0$, $m_A = m_B$, *"no difference"*, *"no change"*.

H_1 the alternative hypothesis, e.g. $t \neq 0$, $m_A \neq m_B$, *"there is a difference/change"*.

Hypothesis testing

H_0 the null hypothesis, e.g. $t = 0$, $m_A = m_B$, *"no difference"*, *"no change"*.

H_1 the alternative hypothesis, e.g. $t \neq 0$, $m_A \neq m_B$, *"there is a difference/change"*.

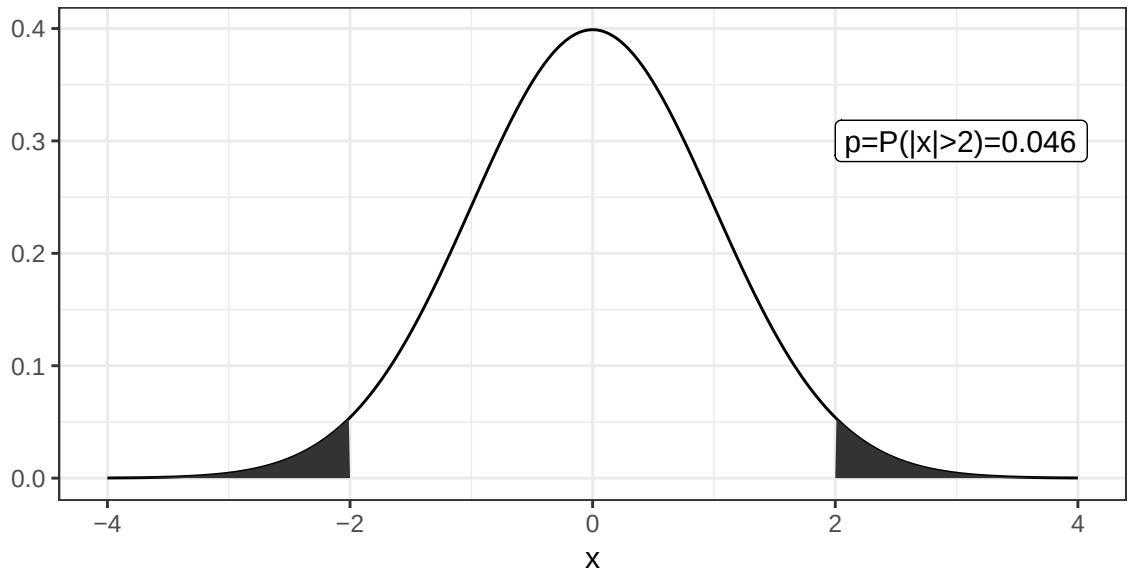
The p-value is the probability of obtaining an effect at least as extreme as the observed, given that the null hypothesis is true.
$p = P(\text{observation or more extreme} | H_0)$

Hypothesis testing

H_0 the null hypothesis, e.g. $t = 0$, $m_A = m_B$, *"no difference"*, *"no change"*.

H_1 the alternative hypothesis, e.g. $t \neq 0$, $m_A \neq m_B$, *"there is a difference/change"*.

The p-value is the probability of obtaining an effect at least as extreme as the observed, given that the null hypothesis is true.
$p = P(\text{observation or more extreme} | H_0)$

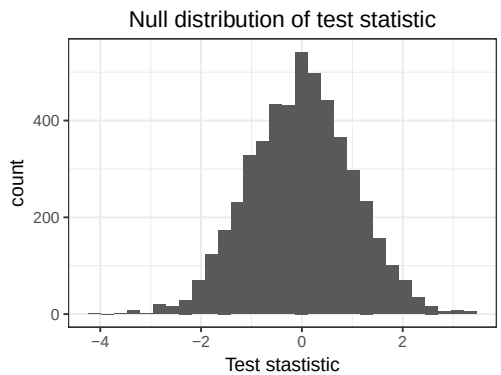


Hypotheis testing

Permutation p-value

When the null distribution of a test statistic is unknown, resampling methods can be used and a permutation p-value can be calculated:

1. Define null and alternative hypothesis, choose test statistic.
2. Calculate the test statistic for the original (unpermuted) data.
3. Permute the labels (e.g. "patient" and "control") and recalculate the test statistic.
4. Repeat 3 many times. (For an exact test do all possible permutations or perform only a subset of all the permutations (Monte Carlo test).)



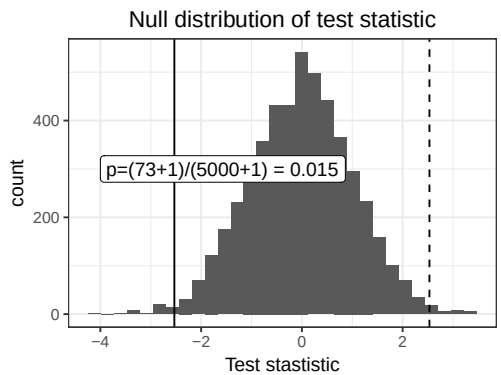
5. Calculate the permutation p-value as (number of permutations with a more extreme test statistic than original + 1)/(number of permutations + 1)

Hypotheis testing

Permutation p-value

When the null distribution of a test statistic is unknown, resampling methods can be used and a permutation p-value can be calculated:

1. Define null and alternative hypothesis, choose test statistic.
2. Calculate the test statistic for the original (unpermuted) data.
3. Permute the labels (e.g. "patient" and "control") and recalculate the test statistic.
4. Repeat 3 many times. (For an exact test do all possible permutations or perform only a subset of all the permutations (Monte Carlo test).)



5. Calculate the permutation p-value as (number of permutations with a more extreme test statistic than original + 1)/(number of permutations + 1)

Hypothesis testing

Errors

	H_0 is true	H_0 is false (H_1 is true)
Reject H_0	Type I error false positive	Correct true positive
Accept H_0	Correct true negative	Type II error false negative

Significance level:
$P(\text{reject } H_0 | H_0 \text{ is true}) = P(\text{type I error}) = \alpha$

Statistical power:
$P(\text{reject } H_0 | H_1 \text{ is true}) = 1 - P(\text{type II error}) = 1 - \beta$

Power analysis (sample size calculation)

Total number of samples n for the two class problem (equally sized classes) can be calculated based on the following:

- The significance level, α
- The power, $1 - \beta$
- The effect size, δ
- The standard deviation, σ

[1] Simon R, Dobbin K. Experimental design of DNA microarray experiments. Biotechniques 34:1-5, 2002

Power analysis (sample size calculation)

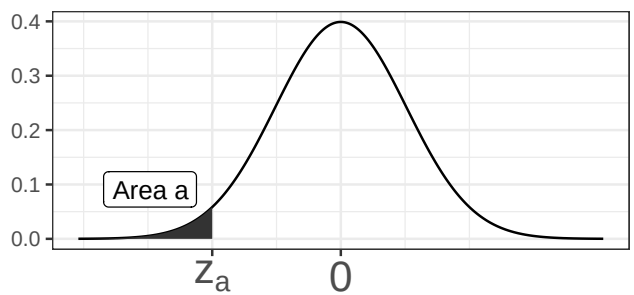
Total number of samples n for the two class problem (equally sized classes) can be calculated based on the following:

- The significance level, α
- The power, $1 - \beta$
- The effect size, δ
- The standard deviation, σ

Assume normal distribution (for larger sample sizes):

$$n \approx \frac{4(z_{\alpha/2} + z_{\beta})^2}{(\delta/\sigma)^2},$$

where z_a denote the value along the x-axis such that the area under the standard normal curve to the left of z_a is a.



[1] Simon R, Dobbin K. Experimental design of DNA microarray experiments. Biotechniques 34:1-5, 2002

Power analysis (sample size calculation)

Total number of samples n for the two class problem (equally sized classes) can be calculated based on the following:

- The significance level, α
- The power, $1 - \beta$
- The effect size, δ
- The standard deviation, σ

or use t-distribution (for small sample sizes):

$$n \approx \frac{4(t_{\alpha/2} + t_{\beta})^2}{(\delta/\sigma)^2},$$

where the t-distribution has $n - 2$ degrees of freedom.

[1] Simon R, Dobbin K. Experimental design of DNA microarray experiments. Biotechniques 34:1-5, 2002

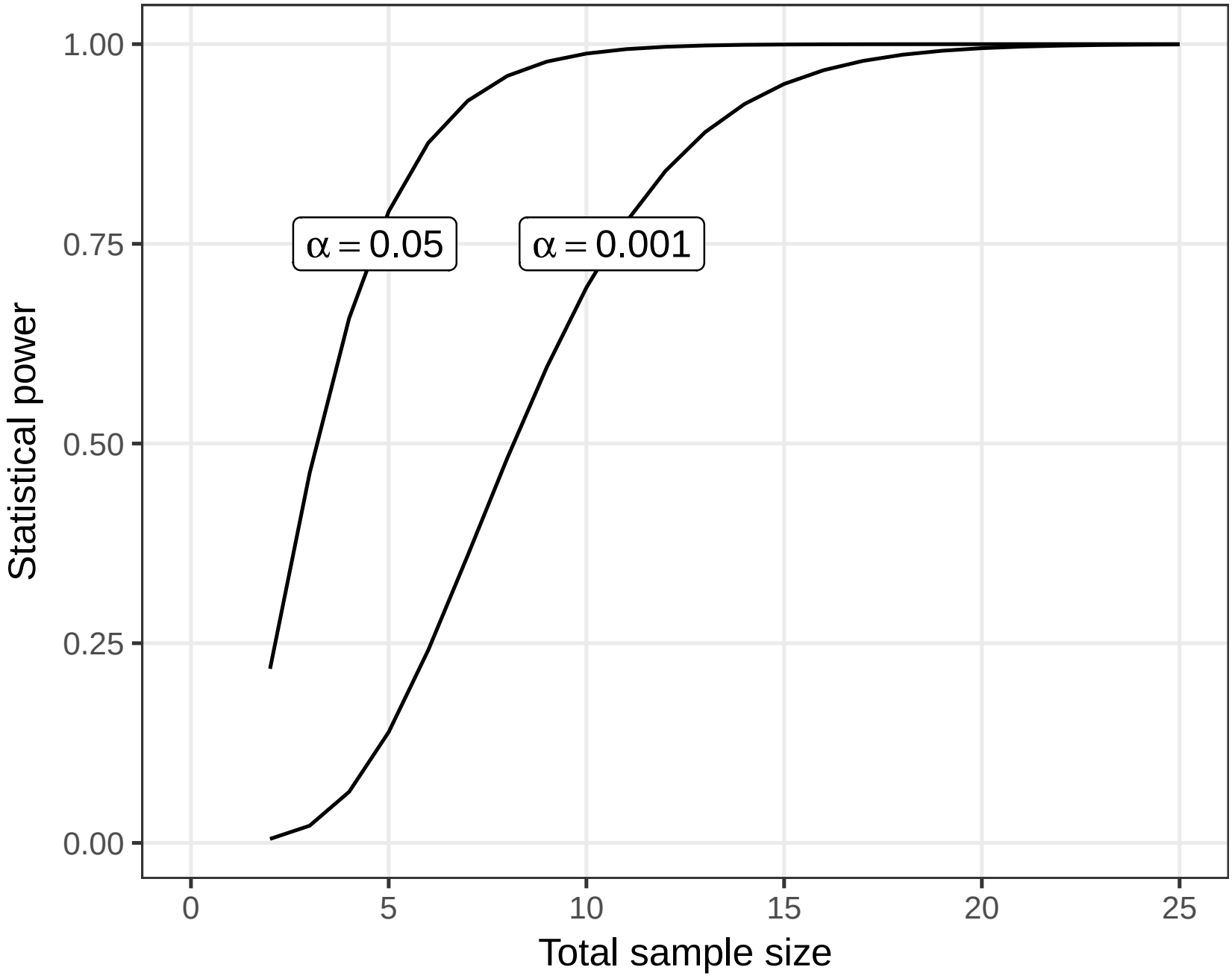
Sample size calculation, example

Let

- $\alpha = 0.001$ (0.05 is too high if we are testing many genes)
- power 95%, $\beta = 0.05$
- $\sigma = 0.5$ (should be estimated from previous studies of similar sample types, use e.g. median over all genes.
- An interesting effect size might be $\delta = 1$ (a 2-fold change if working on $\log_2$-scale).

This gives a total sample size of approximately 29, i.e. 15 samples per class.

Sample size calculation, example



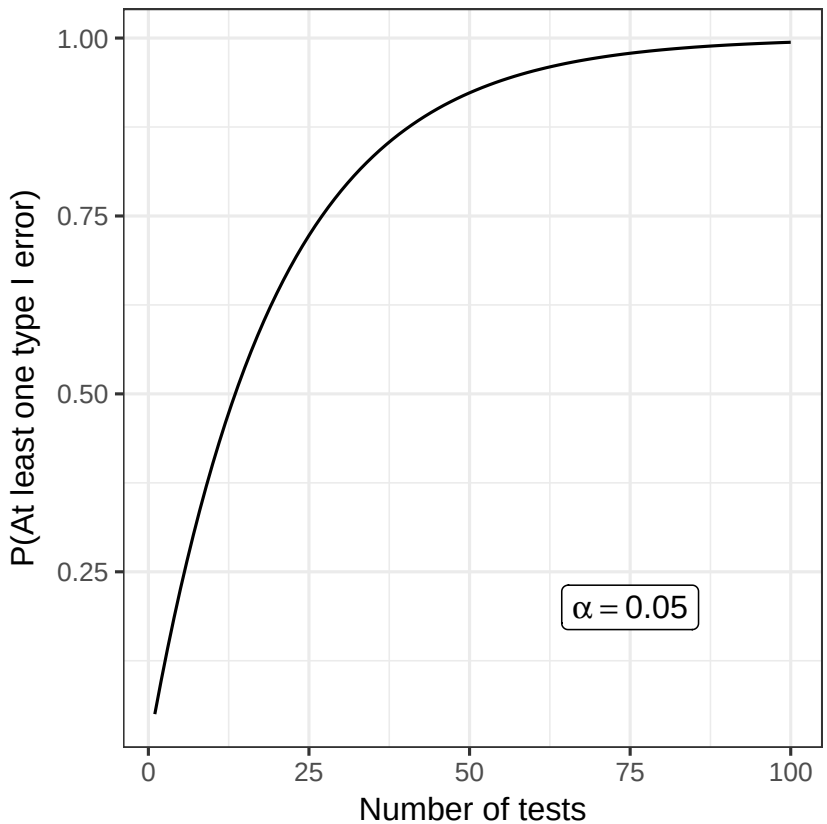
Multiple testing

Perform one test:

- $P(\text{One type I error}) = \alpha$
- $P(\text{No type I error}) = 1 - \alpha$

Perform m independent tests:

- $P(\text{No type I errors in } m \text{ tests}) = (1 - \alpha)^m$
- $P(\text{At least one type I error in } m \text{ tests}) = 1 - (1 - \alpha)^m$



Multiple testing

FWER family-wise error rate, probability of one or more false positive, e.g. Bonferroni, Holm

FDR false discovery rate, proportion of false positives among "hits", e.g. Benjamini-Hochberg, Storey

Multiple testing

Bonferroni correction

To achieve a family-wise error rate of $\leq \alpha$ when performing m tests, declare significance and reject the null hypothesis for any test with $p \leq \alpha/m$.
Objections: too conservative

Multiple testing

Benjamini-Hochberg, FDR

	H_0 is true	H_0 is false (H_1 is true)
Reject H_0	FP	TP
Accept H_0	TN	FN

The false discovery rate is the proportion of false positives among 'hits', i.e. $\frac{FP}{TP+FP}$.

Benjamini-Hochberg's method control the FDR level, γ, when performing m **independent** tests, as follows:

1. Sort the p-values $p_1 \leq p_2 \leq \cdots \leq p_m$.
2. Find the maximum j such that $p_j \leq \gamma \frac{j}{m}$.
3. Declare significance for all tests $1, 2, \ldots, j$.

Multiple testing

'Adjusted' p-values

Sometimes an adjusted significance threshold is not reported, but instead 'adjusted' p-values are reported.

- Using Bonferroni's method the adjusted p-values are: $\tilde{p}_i = \min(mp_i, 1)$.

  A feature's adjusted p-value represents the smallest FWER at which the null hypothesis will be rejected, i.e. the feature will be deemed significant.

- Benjamini-Hochberg's 'adjusted' p-values are called q-values: $q_i = \min(\frac{m}{i} p_i, 1)$

  A feature's q-value can be interpreted as the lowest FDR at which the corresponding null hypothesis will be rejected, i.e. the feature will be deemed significant.

Multiple testing

'Adjusted' p-values

Example, 10000 independent tests (genes)

	p-value	adj p (Bonferroni)	q-value (B-H)
1	1.7e-08	0.00017	0.00017
2	5.8e-08	0.00058	0.00029
3	3.4e-07	0.0034	0.0011
4	9.1e-07	0.0091	0.002
5	1e-06	0.01	0.002
6	2.4e-06	0.024	0.004
7	2.3e-05	0.23	0.033
8	3.6e-05	0.36	0.045
9	0.00022	1	0.23
10	0.00023	1	0.23
11	0.00073	1	0.66
12	0.0032	1	1
13	0.0045	1	1
14	0.0087	1	1
15	0.0089	1	1

Questions?