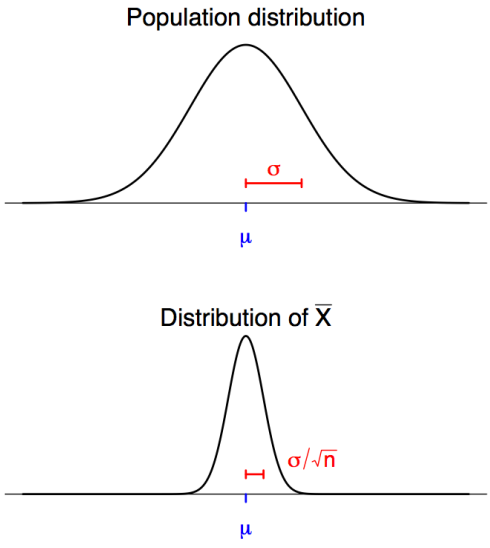
Statistical Modeling 1

Means and two-group comparisons

The sample mean

If $X_1, X_2, \ldots, X_n$ are iid Normal(μ,σ), then

$\longrightarrow \quad \overline{X} \sim \text{Normal}(\mu, \sigma/\sqrt{n})$.



The central limit theorem

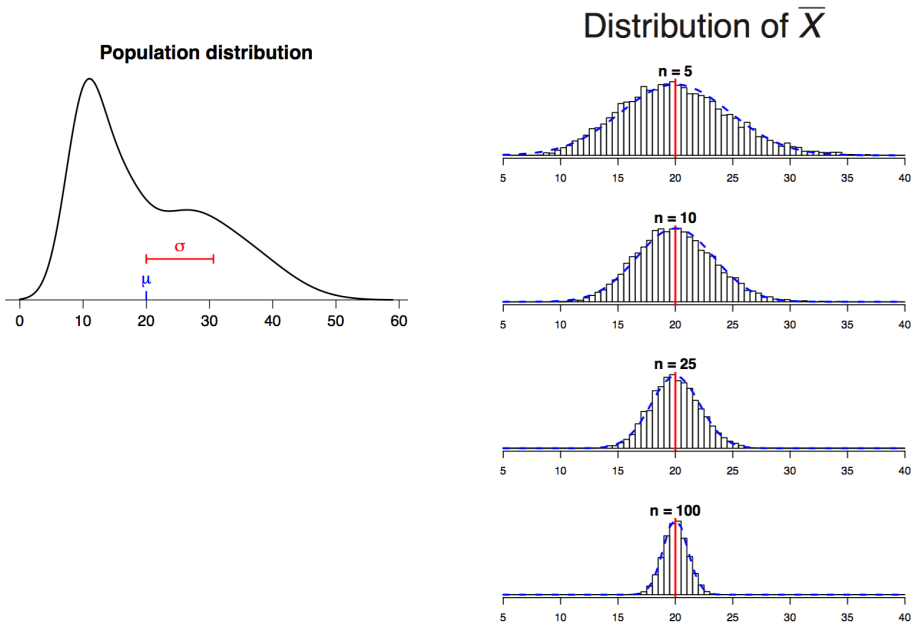
⟶ If $X_1, X_2, \ldots, X_n$ are iid with mean μ and SD σ, and the sample size (n) is large, then

$\overline{X}$ is approximately Normal(μ, $\sigma/\sqrt{n}$).

⟶ How large is large?

It depends on the population distribution.
(But, generally, not too large.)

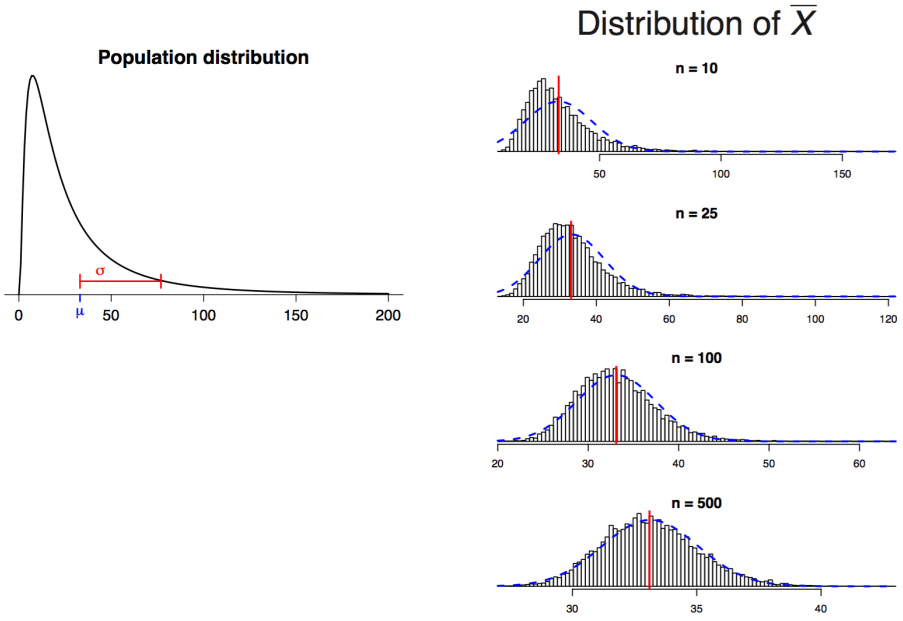
The central limit theorem



Ingo Ruczinski | Asian Institute in Statistical Genetics and Genomics | July 21-22, 2017

[140.615]

The central limit theorem



[140.615]

Means and variances of sums and means

⟶ Mean and SD of sums of random variables:

$E(\sum_i X_i) = \sum_i E(X_i)$ no matter what

$SD(\sum_i X_i) = \sqrt{\sum_i \{SD(X_i)\}^2}$ if the X_i are independent

⟶ Mean and SD of means of random variables:

$E(\sum_i X_i / n) = \sum_i E(X_i)/n$ no matter what

$SD(\sum_i X_i/n) = \sqrt{\sum_i \{SD(X_i)\}^2}/n$ if the X_i are independent

⟶ If the X_i are iid with mean μ and SD σ:

$E(\sum_i X_i / n) = \mu$ and $SD(\sum_i X_i / n) = \sigma/\sqrt{n}$

Baselines

Assume that you have three independendent measurements X_1, X_2, X_3, with $\text{var}(X_i) = \sigma^2$.

$\longrightarrow$ It follows that $\text{var}(\bar{X}) = \sigma^2/3$.

Assume that you have to subtract a baseline B from each measurement. What is $\text{var}(X_i - B)$?

If B is just a constant, then

$\longrightarrow \text{var}(X_i - B) = \text{var}(X_i) = \sigma^2$.

If B is a measurement with $\text{var}(B) = \sigma_B^2$, then

$\longrightarrow \text{var}(X_i - B) = \text{var}(X_i) + \text{var}(B) = \sigma^2 + \sigma_B^2$.

Baselines

What is the variance of the average of those values, i.e. what is the variance of $\sum_i (X_i - B)/3$ .. ?

$$\begin{aligned} \text{var}\left(\sum_i (X_i - B)/3\right) &= \text{var}\left(\frac{1}{3}(X_1 + X_2 + X_3 - 3 \times B)\right) \\ &= \text{var}\left(\bar{X} - B\right) = \text{var}(\bar{X}) + \text{var}(B) \\ &= \frac{\sigma^2}{3} + \sigma_B^2. \end{aligned}$$

In other words: if you use the same baseline in all experiments, you will never be able to get rid of the baseline error, regardless how many replicates you use!

Baselines

If each experiment has its own baseline B_i, measured independently, with $\text{var}(B_i) = \sigma_B^2$, then

$$\begin{aligned} \text{var}\left(\sum_i (X_i - B_i)/3\right) &= \text{var}\left(\bar{X} - \bar{B}\right) = \text{var}(\bar{X}) + \text{var}(\bar{B}) \\ &= \frac{\sigma^2}{3} + \frac{\sigma_B^2}{3} = \frac{\sigma^2 + \sigma_B^2}{3} \end{aligned}$$

In other words: if each experiment has its own baseline, the baseline error will average out.

t-tests

Suppose that

- $X_1, X_2, \ldots, X_n$ are iid Normal(mean=μ_A, SD=σ), and
- $Y_1, Y_2, \ldots, Y_m$ are iid Normal(mean=μ_B, SD=σ).

Then

$$\longrightarrow \quad E(\overline{X} - \overline{Y}) = E(\overline{X}) - E(\overline{Y}) = \mu_A - \mu_B$$

$$\longrightarrow \quad SD(\overline{X} - \overline{Y}) = \sqrt{SD(\overline{X})^2 + SD(\overline{Y})^2} =$$

$$\sqrt{\left(\frac{\sigma}{\sqrt{n}}\right)^2 + \left(\frac{\sigma}{\sqrt{m}}\right)^2} = \sigma\sqrt{\frac{1}{n} + \frac{1}{m}}$$

Note: If n = m, then $SD(\overline{X} - \overline{Y}) = \sigma\sqrt{2/n}$.

t-tests

$$\widehat{\mathrm{SD}}(\overline{X}-\overline{Y}) = \hat{\sigma}_{\text{pooled}}\sqrt{\frac{1}{n}+\frac{1}{m}}$$

$$= \sqrt{\left[\frac{S_A^2(n-1)+S_B^2(m-1)}{n+m-2}\right]\cdot\left[\frac{1}{n}+\frac{1}{m}\right]}$$

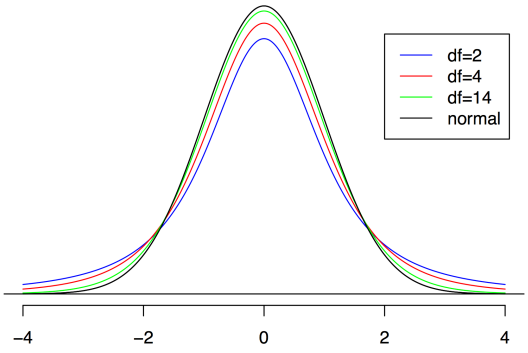
In the case n = m,

$$\widehat{\mathrm{SD}}(\overline{X}-\overline{Y}) = \sqrt{\frac{S_A^2+S_B^2}{n}}$$

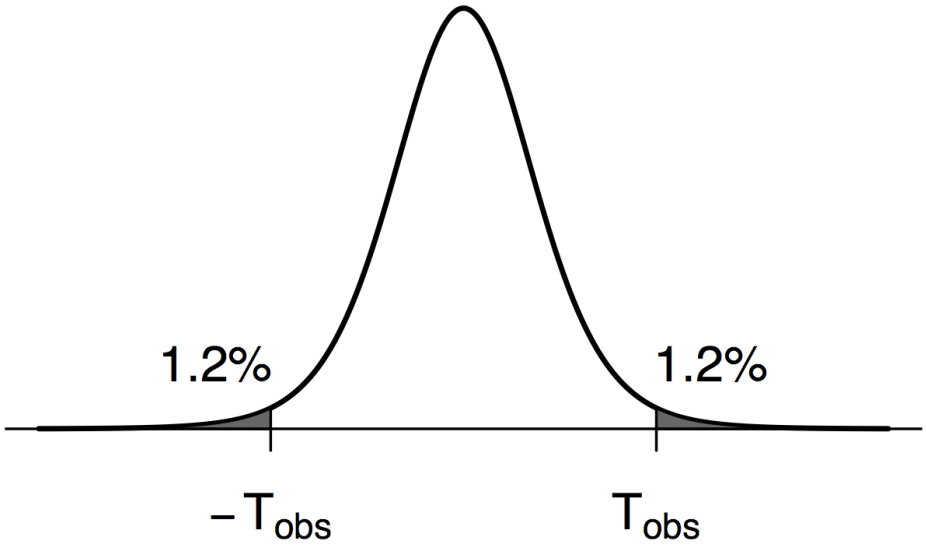
t-tests

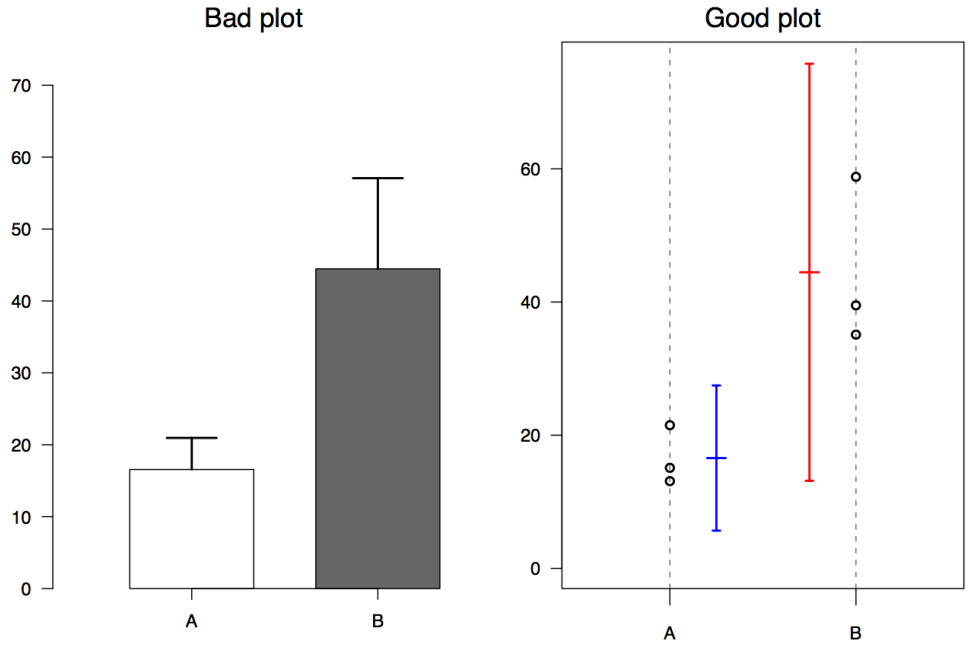
If the null hypothesis is true and there there are no differences in group means, then

$$\frac{\bar{X}-\bar{Y}}{\widehat{\mathrm{SD}}(\bar{X}-\bar{Y})} \sim t_{n+m-2}$$

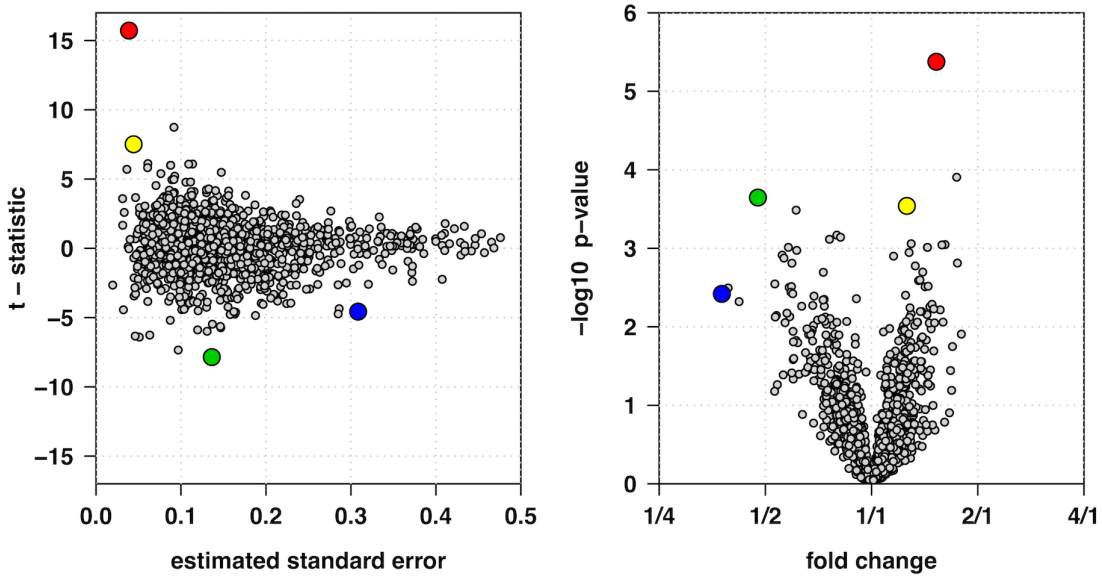


t-tests


1.2%
1.2%
$-T_{obs}$
T_{obs}


Bad plot
Good plot
70
60
50
40
30
20
10
0
A
B
60
40
20
0
A
B

Problems with t-tests in genomics



Hypothesis testing

We consider two hypotheses:

Null hypothesis, H_0: $\mu = 0$ Alternative hypothesis, H_a: $\mu \neq 0$

Type I error:	Reject H_0 when it is true	(false positive)
Type II error:	Fail to reject H_0 when it is false	(false negative)

We set things up so that a Type I error is a worse error (and so that we are seeking to prove the alternative hypothesis). We want to control the rate (the significance level, α) of such errors.

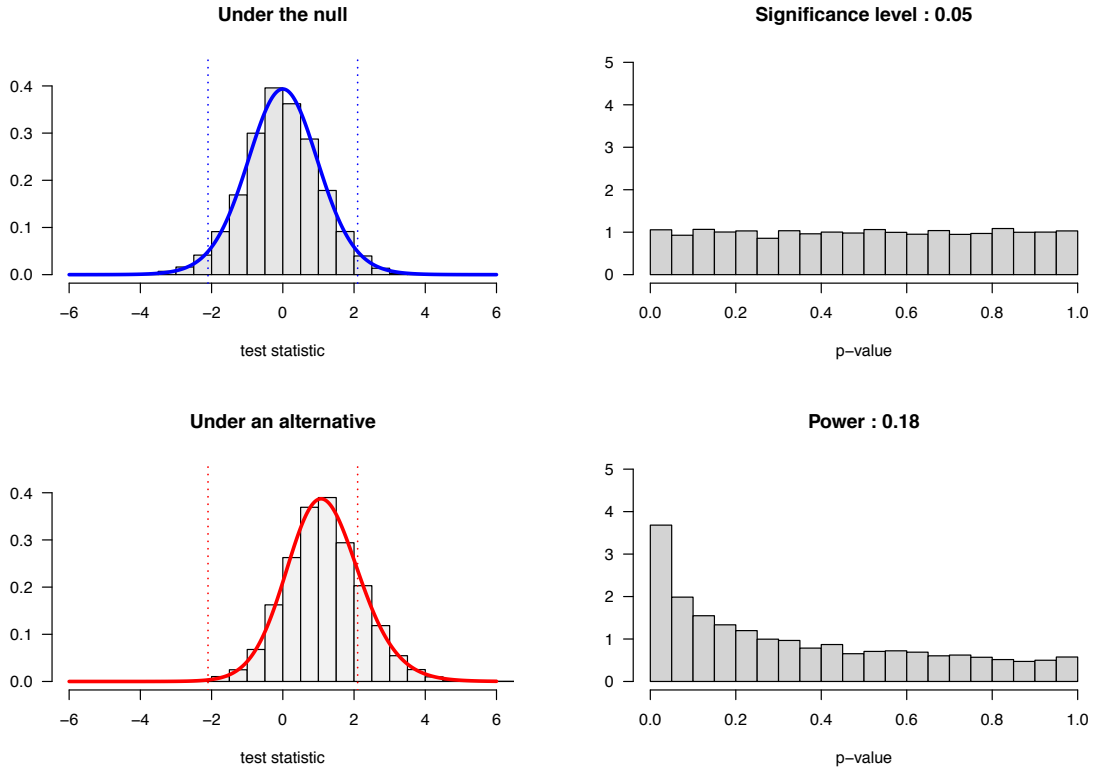
Two-sample t-test

$X_1, \ldots, X_n$ iid Normal(μ_A, σ) $\qquad$ $Y_1, \ldots, Y_m$ iid Normal(μ_B, σ)

Test $H_0 : \mu_A = \mu_B$ vs $H_a : \mu_A \neq \mu_B$

Test statistic: $T = \dfrac{\overline{X} - \overline{Y}}{s_p \sqrt{\frac{1}{n} + \frac{1}{m}}}$ where $s_p = \sqrt{\dfrac{s_A^2(n-1) + s_B^2(m-1)}{n+m-2}}$

$\longrightarrow$ Compare to the t distribution with n + m – 2 d.f.



10 observations per group.
Alternative: difference in means is one half within-group standard deviation.

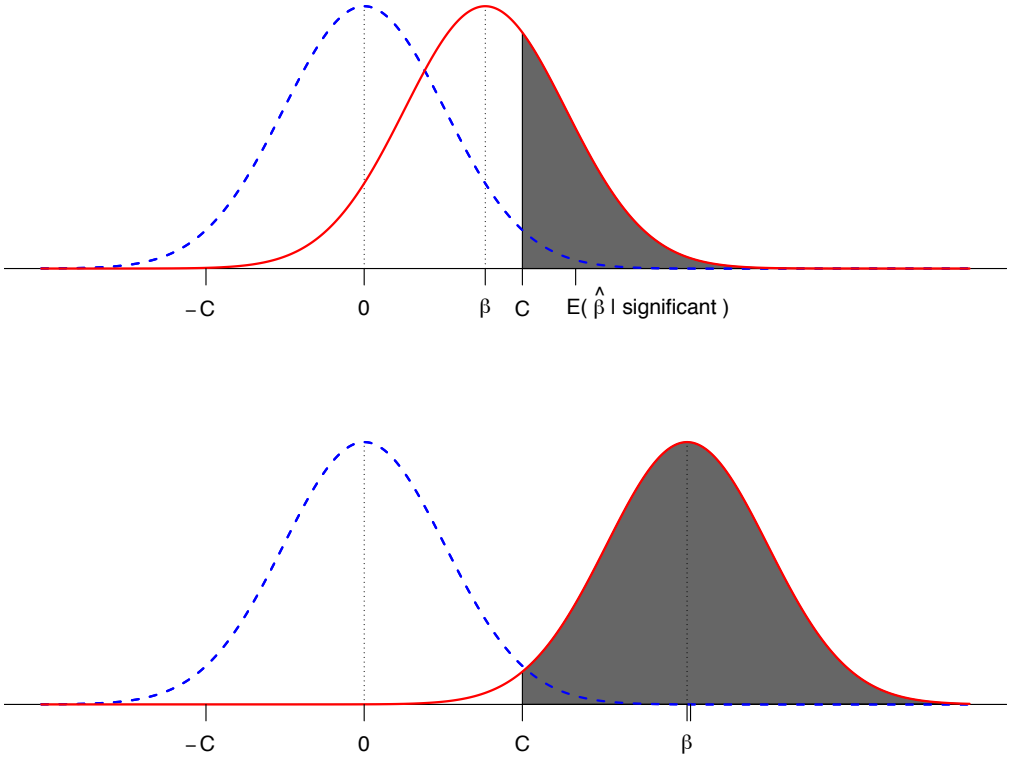
ARTICLE

Overcoming the Winner's Curse: Estimating Penetrance Parameters from Case-Control Data

Sebastian Zöllner and Jonathan K. Pritchard

Genomewide association studies are now a widely used approach in the search for loci that affect complex traits. After detection of significant association, estimates of penetrance and allele-frequency parameters for the associated variant indicate the importance of that variant and facilitate the planning of replication studies. However, when these estimates are based on the original data used to detect the variant, the results are affected by an ascertainment bias known as the "winner's curse." The actual genetic effect is typically smaller than its estimate. This overestimation of the genetic effect may cause replication studies to fail because the necessary sample size is underestimated. Here, we present an approach that corrects for the ascertainment bias and generates an estimate of the frequency of a variant and its penetrance parameters. The method produces a point estimate and confidence region for the parameter estimates. We study the performance of this method using simulated data sets and show that it is possible to greatly reduce the bias in the parameter estimates, even when the original association study had low power. The uncertainty of the estimate decreases with increasing sample size, independent of the power of the original test for association. Finally, we show that application of the method to case-control data can improve the design of replication studies considerably.

PMID 17357068



Ingo Ruczinski | Asian Institute in Statistical Genetics and Genomics | July 21-22, 2017

Wilcoxon rank-sum test

Rank the X's and Y's from smallest to largest (1, 2, . . . , n+m)

R = sum of ranks for X's

(Also known as the Mann-Whitney Test)

X	Y	rank
35.0		1
38.2		2
43.3		3
	46.8	4
	49.7	5
50.0		6
	51.9	7
57.1		8
61.2		9
	74.1	10
	75.1	11
	84.5	12
	90.0	13
	95.1	14
	101.5	15

$R = 1 + 2 + 3 + 6 + 8 + 9 = 29$

P-value = 0.026

$\longrightarrow$ use `wilcox.test()`

Note: The distribution of R (given that X's and Y's have the same dist'n) is calculated numerically

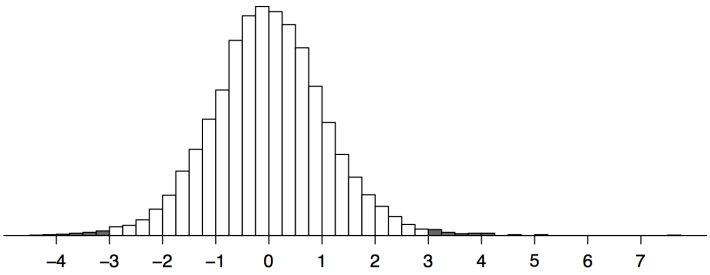
Permutation tests

Response	R	R	•••	NR	NR
	Patient 1	Patient 2	•••	Patient n-1	Patient n
Gene 1	-1.64	-0.42	•••	-1.39	-0.38
Gene 2	-3.12	-3.60	•••	-3.80	-2.82
:	:	:	•••	:	:
:	:	:		:	:
:	:	:	•••	:	:
:	:	:		:	:
Gene m-1	-2.34	-0.22	•••	-1.22	-2.76
Gene m	4.53	3.23	•••	0.29	3.11

Response	NR	R	•••	NR	R
	Patient 1	Patient 2	•••	Patient n-1	Patient n
Gene 1	-1.64	-0.42	•••	-1.39	-0.38
Gene 2	-3.12	-3.60	•••	-3.80	-2.82
:	:	:	•••	:	:
:	:	:		:	:
:	:	:	•••	:	:
:	:	:		:	:
Gene m-1	-2.34	-0.22	•••	-1.22	-2.76
Gene m	4.53	3.23	•••	0.29	3.11

This leaves the relationship between the features (such as genes) unchanged.

Permutation null distribution



P-value = $\Pr(|T^\star| \geq |T_{obs}|)$

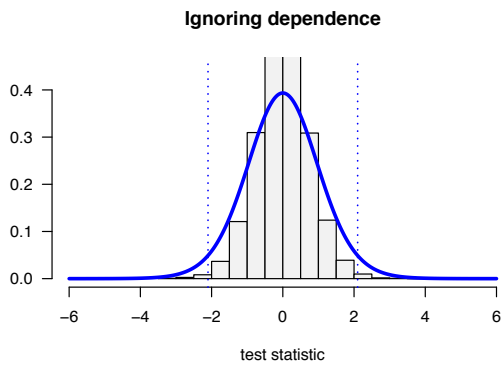
$\longrightarrow$ Small n & m: Look at all $\binom{n+m}{n}$ possible shuffles

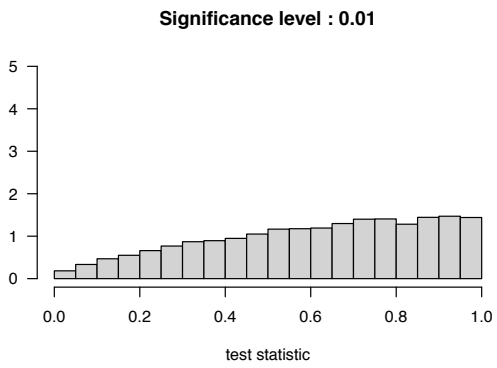
$\longrightarrow$ Large n & m: Look at a sample (w/ repl) of 1000 such shuffles

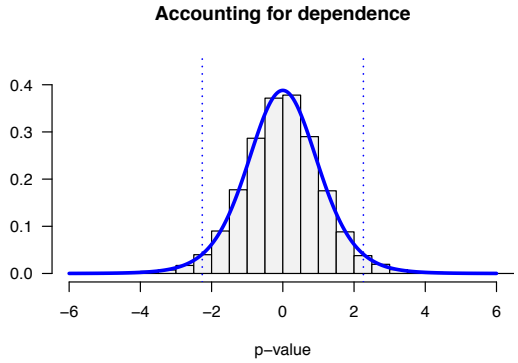
Here, n is the number of responders, and m the number of non-responders.

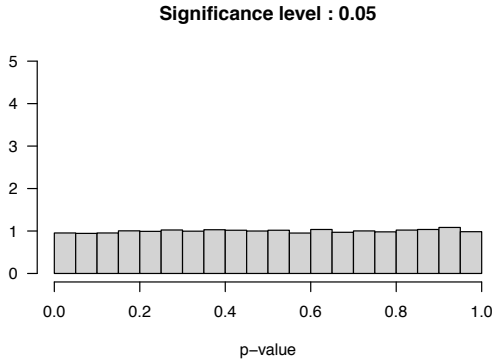
Another reason to like balanced designs

$\downarrow$	σ_1^2/σ_2^2						
n_1/n_2	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{2}$	1	2	4	8
$\frac{1}{2}$	0.011	0.016	0.028	0.050	0.080	0.110	0.133
1	0.050	0.050	0.050	0.050	0.050	0.050	0.050
2	0.133	0.110	0.080	0.050	0.028	0.016	0.011
4	0.237	0.179	0.110	0.050	0.016	0.004	0.001
8	0.331	0.237	0.133	0.050	0.011	0.001	0.000

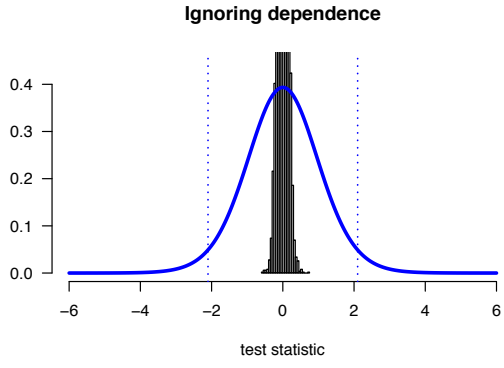


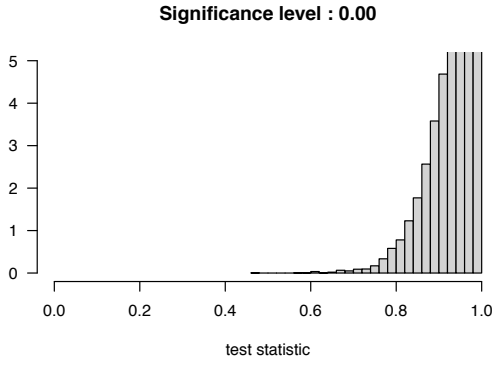


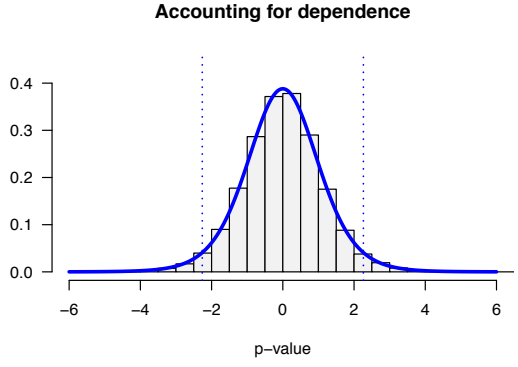


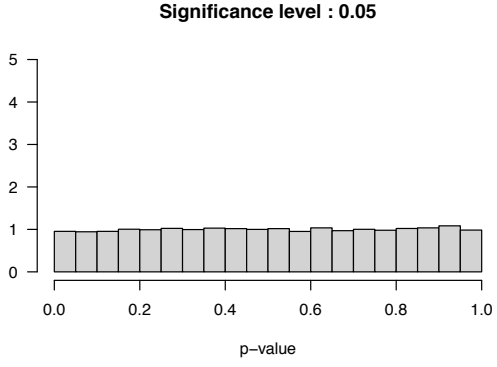


Paired data in ten observations.
Between subject variability (SD) equal to error.









Paired data in ten observations.
Between subject variability (SD) ten times larger than error.