

Power, Sample Size and Replication

David Evans^{1,2,3}

1 Institute for Molecular Bioscience, University of Queensland

2 University of Queensland Diamantina Institute

3 MRC Integrative Epidemiology Unit, University of Bristol

Outline

1. Aims
2. Statistical power
3. Estimate the power of association analysis
 - Analytically
 - Empirically
4. Multiple Testing
5. Replication

1. Aims

1. Know what type-I error and power are

2. Know that you can/should estimate the power of your association analyses (analytically or empirically)

3. Know that there a number of tools that you can use to estimate power

4. Be aware that there are many factors that increase type-I error and decrease power

5. Be able to understand strategy and criteria for replication

2. Statistical power

H₀: There is NO association between a marker and a trait

H₁: There is association between a marker and a trait

		In reality...	
		H ₀ is true	H ₁ is true
We decide...	H ₀ is true	$1 - \alpha$	β Type-2 error
	H ₁ is true	α Type-1 error	$1 - \beta$ Power

Power: probability of detecting association when H₁ is true.

Definitions

▶ Power

The probability of detecting a given size effect in a population from a sample size N , using significance criterion α

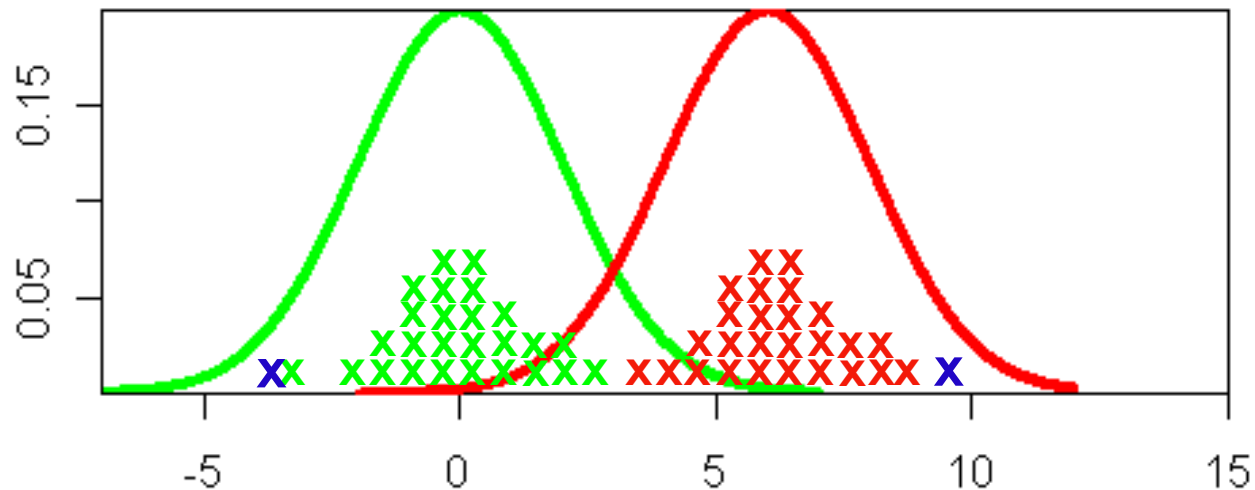
▶ Type I error

The probability of incorrectly rejecting the null hypothesis of no association

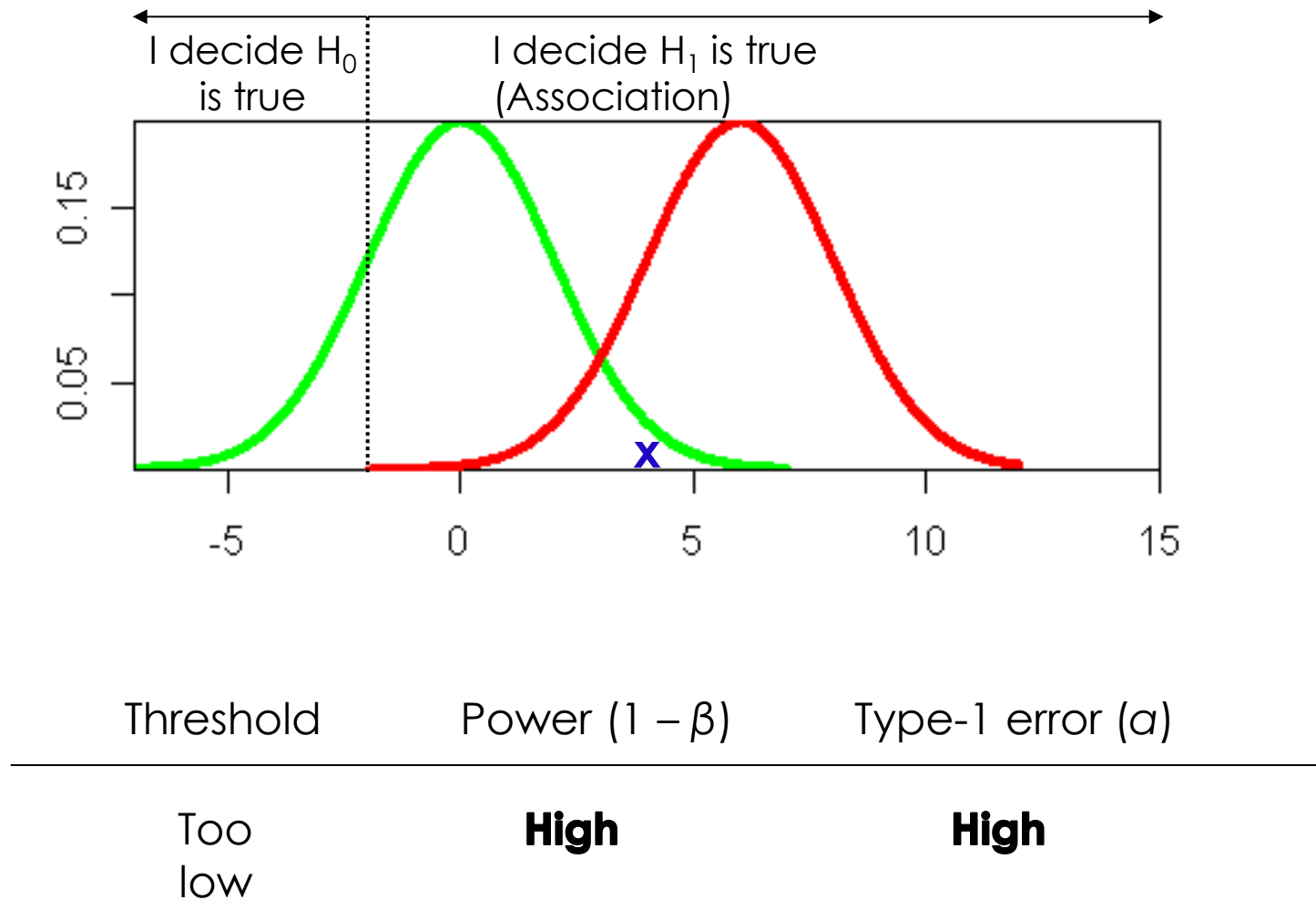
H_0 : There is NO association between a marker and a trait

H_1 : There is association between a marker and a trait

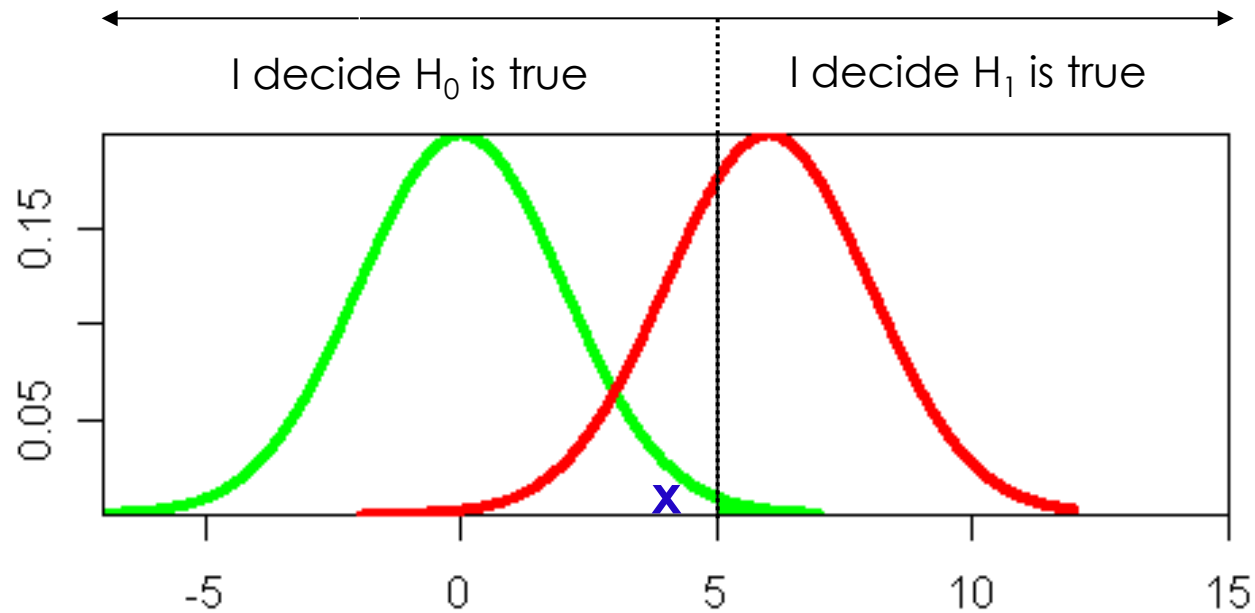
Association test statistic has different
distributions under H_0 and H_1



Where should I set the threshold to determine significance?



Where should I set the threshold to determine significance?



Threshold

Power ($1 - \beta$)

Type-1 error (α)

Too low

High

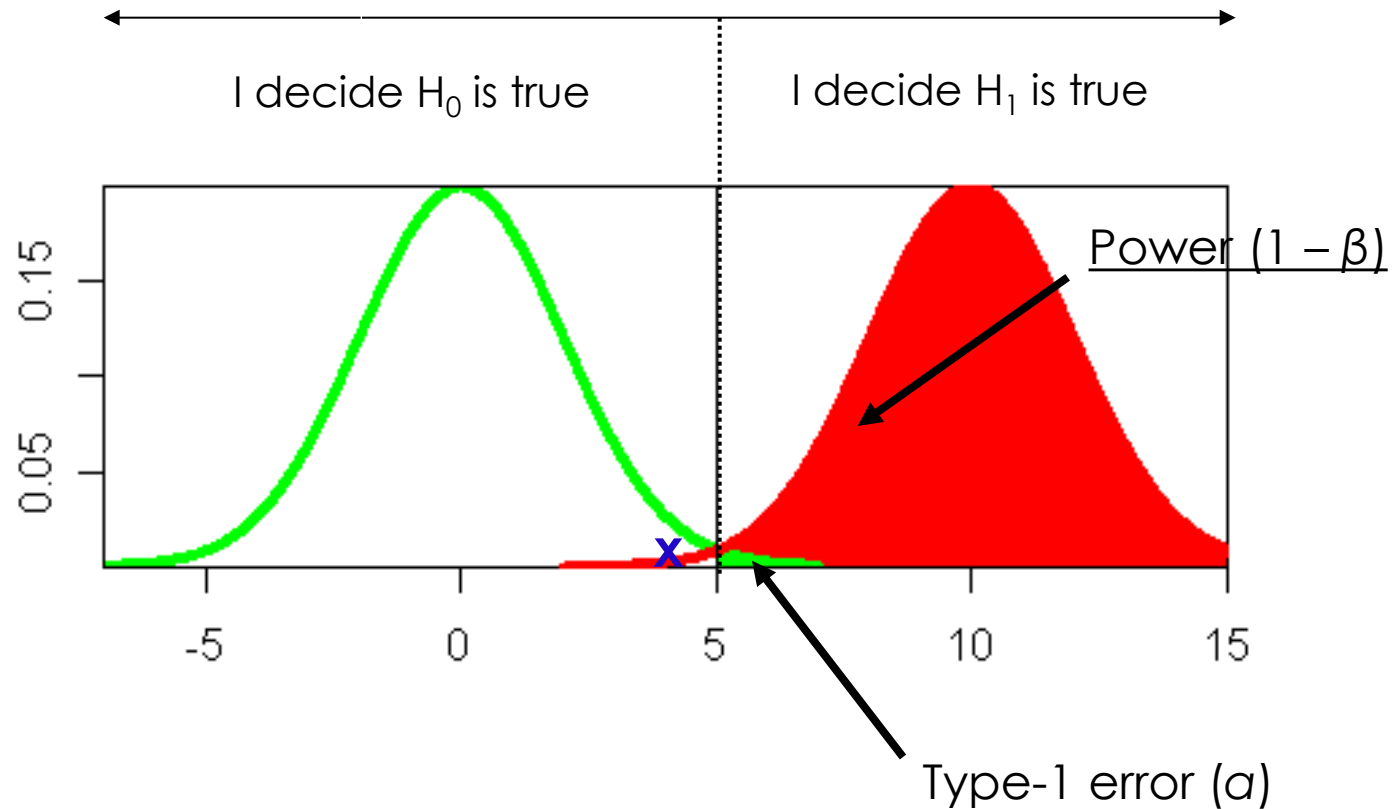
High

Too high

Low

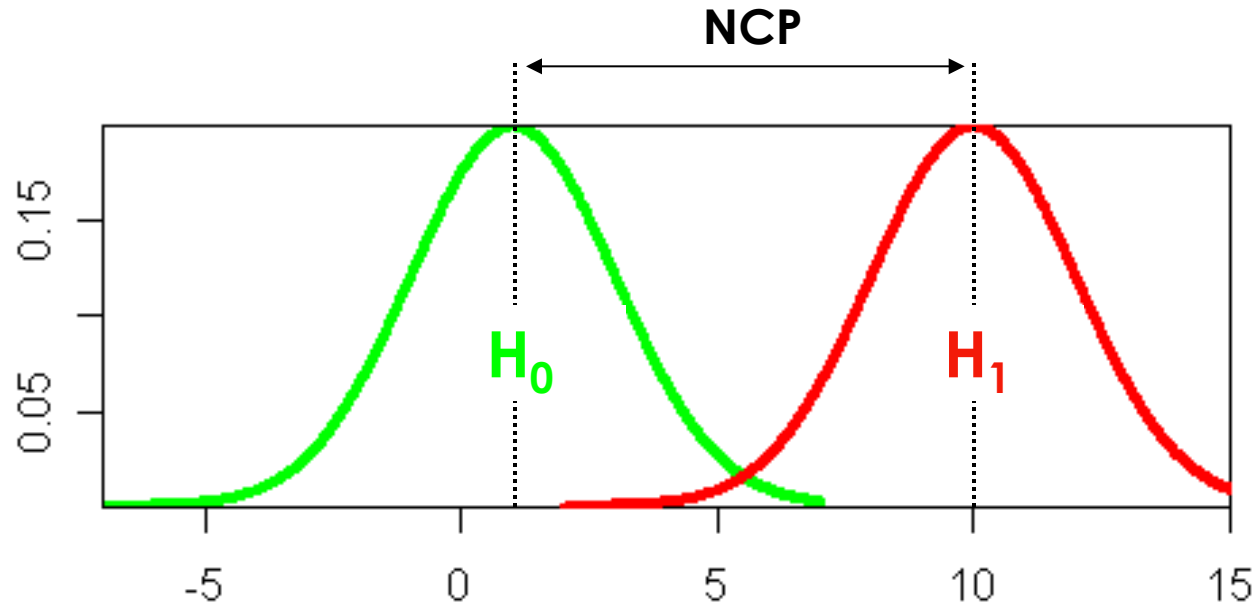
Low

How do I maximise Power while minimising Type-1 error rate?



1. **Set a high threshold for significance** (i.e. results in low α [e.g. 0.05-0.00002])
2. **Try to shift the distribution of the association test statistic when H_1 is true as far as possible from the distribution when H_0 is true.**

Non-centrality parameter

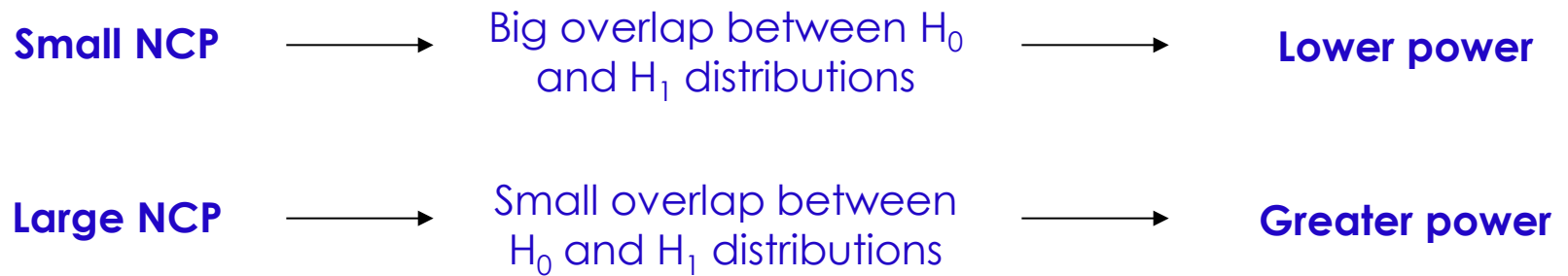
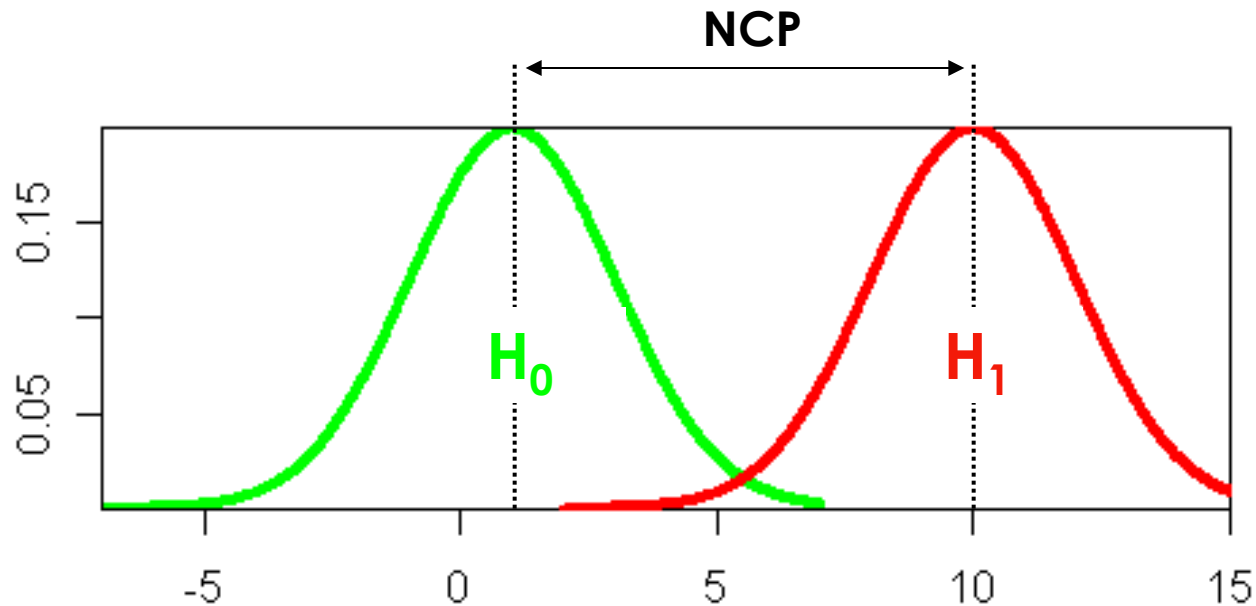


Central χ^2

Non-central χ^2

Mean (μ)	df	df + NCP
Variance (σ^2)	$2*(df)$	$2*(df) + 4*NCP$

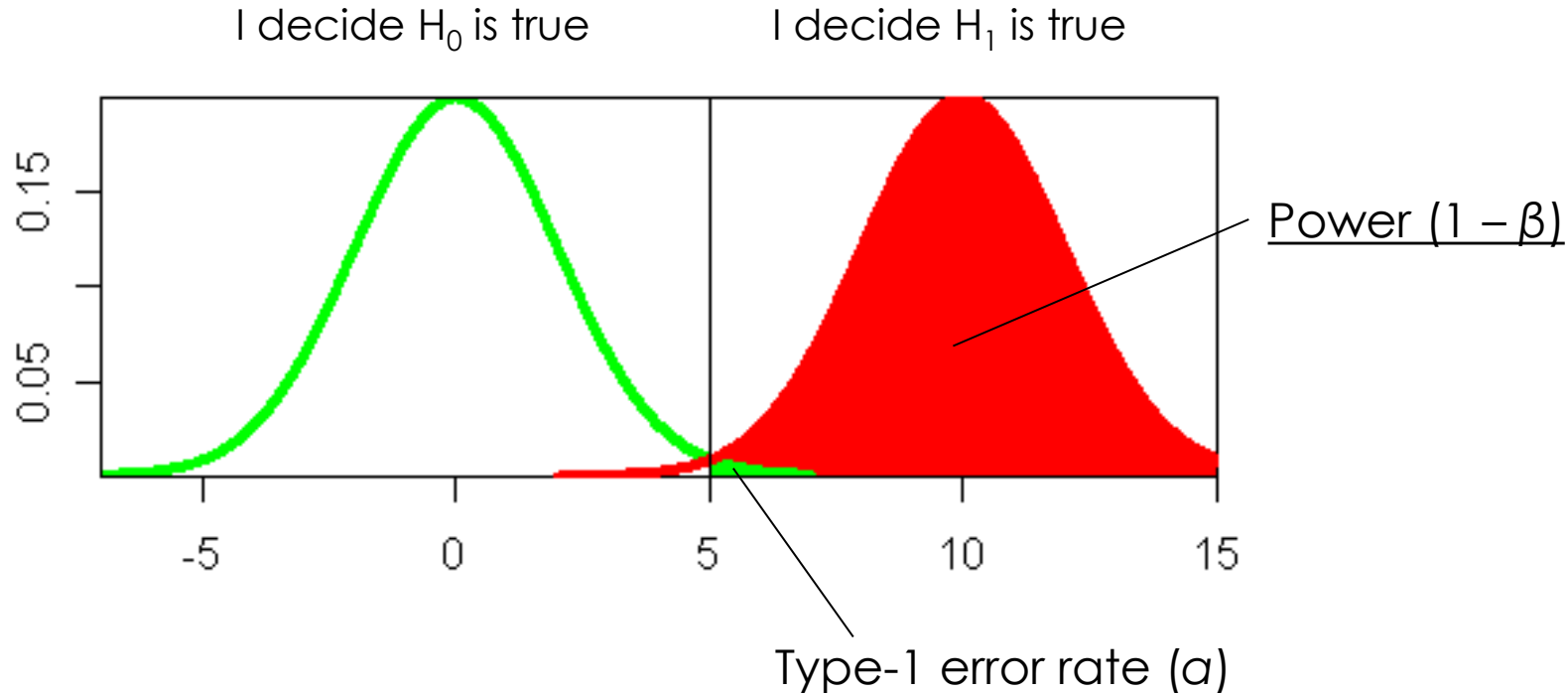
These distributions ARE NOT chi-sq with 1df!! Just for illustration... (Question- why do we use chi-sq?)



- ▶ Sample size does NOT scale linearly with Power
- ▶ But, sample size scales linearly with NCP

3. Estimate power for association

Theoretical power estimation



1. Set type 1 error rate (e.g. $\alpha = 0.05$)
2. Determine what critical value this corresponds to on the X axis
3. Work out the non-centrality parameter of the test ($NCP = E(H_1) - E(H_0)$)
4. Calculate the area to the right of the threshold under H_1

Trivial Example: OLS Linear Regression

Under H_0 :

$$(\beta = 0) \quad \frac{b - 0}{SE} \longrightarrow Z(0,1) \quad (*\text{in large samples})$$

$$\frac{(b - 0)^2}{SE^2} \longrightarrow \chi^2_1 \quad \text{Central chi-square distribution}$$

Under H_1 :

$$(\beta \neq 0) \quad \frac{b - 0}{SE} \longrightarrow Z\left(\frac{\beta}{SE}, 1\right)$$

$$\frac{(b - 0)^2}{SE^2} \longrightarrow \chi^2_1\left(\frac{\beta^2}{SE^2}\right) \quad \text{Non-central chi-square distribution}$$

Trivial Example: OLS Linear Regression

$$\frac{(b - 0)^2}{SE^2} \longrightarrow \chi^2_1\left(\frac{\beta^2}{SE^2}\right)$$

1. Set type 1 error rate (e.g. $\alpha = 0.05$)
2. Determine what critical value this corresponds to on the X axis

```
qchisq(p = 0.05, df = 1, ncp = 0, lower.tail = FALSE, log.p = FALSE)
```

```
[1] 3.841459
```

3. Work out the non-centrality parameter of the test $\frac{\beta^2}{SE^2}$

$$\beta = 0.1$$

$$SE \approx 1/\sqrt{N} \quad (\text{Assume } N = 1000)$$

4. Calculate the area to the right of the threshold under H_1

```
pchisq(q=3.84, df=1, ncp = 0.1^2/(1/1000), lower.tail = FALSE, log.p = FALSE)
```

```
[1] 0.8854512
```

Factors that influence power and type-1 error

Association

Quantitative

Case-control

1. Disease model

Effect size, MAF,
disease prevalence



2. Genome coverage (r^2)



3. Sample size



4. Ascertainment



5. Deviations in trait distribution



6. Measurement error*



7. Genotyping errors*



8. Missing data*



*Assume random

Theoretical power estimation: Exercise

- ▶ What case control sample size do we need to achieve 80% power for genome-wide significance for an odds ratio of 1.2 in a multiplicative model and an allele frequency of 20% when we directly type the locus for a disease with 5% prevalence?

Practical Exercise

<http://zzz.bwh.harvard.edu/gpc/> (Note new location!)

Genetic Power Calculator - Windows Internet Explorer

http://pngu.mgh.harvard.edu/~purcell/gpc/

File Edit View Favorites Tools Help

Common tests, suggestions, comments, etc to [Purcell S, Cherny SS](#).

If you use this site, please reference the following [Bioinformatics article](#):

Purcell S, Cherny SS, Sham PC. (2003) Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics*, 19(1):149-150.

Modules

Case-control for discrete traits	Notes
Case-control for threshold-selected quantitative traits	Notes
QTL association for sibships and singletons	Notes
TDT for discrete traits	Notes
TDT and parenTDT with ascertainment	Notes
TDT for threshold-selected quantitative traits	Notes
Epistasis power calculator	Notes
QTL linkage for sibships	Notes
Probability Function Calculator	Notes

Instructions for power calculations

VC model calculations are based upon formula derived in Sham et al (2000) [\[AJHG, 66, 1616-1630\]](#). Users of this site who are unsure of the nature of the VC

start Internet 100% 23:12

Practical Exercise

Statistical Genetics Group - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://pngu.mgh.harvard.edu/~purcell/gpc/cc2.html

Most Visited Smart Bookmarks Free Hotmail Amazon.com: Statistic... The Arreat Summit - It... Share on Facebook

Gmail - Update: ... IBG workshop:2009:... IBG Index of /worksh... Wiley InterScien... Hulu - Videos Cypriots and U.N... Statistical Ge...

Genetic Power Calculator

Case - control for discrete traits

Allele frequency at the risk locus

High risk allele frequency (A)	: 0.2	(0 - 1)
Prevalence	: .05	(0.0001 - 0.9999)
Genotype relative risk Aa	: 1.2	(>1)
Genotype relative risk AA	: 1.44	(>1)
D-prime	: 1	(0 - 1)
Marker allele frequency (B)	: 0.2	(0 - 1)
Number of cases	: 1000	(0 - 10000000)
Control : case ratio	: 1	(>0) (1 = equal number of cases and controls)
<input type="checkbox"/> Unselected controls? (* see below)		
User-defined type I error rate	: 5e-8	(0.00000001 - 0.5)
User-defined power: determine N (1 - type II error rate)	: 0.80	(0 - 1)

Created by [Shaun Purcell](#) 24.Oct.2008

Find: exclude Next Previous Highlight all Match case

Boston: Wed 12:49 UK: Wed 17:49 Netherlands: Wed 18:49 Hong Kong: Thu 01:49 Los Angeles: Wed 09:49 Stopped

start 2 F. 2 W. 2 M. 3 M. 5 M. RG... ma... silv... 92% 12:49 PM

Practical Exercise

Statistical Genetics Group - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://pngu.mgh.harvard.edu/~purcell/gpc/cc2.html

Most Visited Smart Bookmarks Free Hotmail Amazon.com: Statistic... The Arreat Summit - It... Share on Facebook

Gmail - Update: ... IBG workshop:2009:... IBG Index of /worksh... Wiley InterScien... Hulu - Videos Cypriots and U.N... Statistical Ge...

Genetic Power Calculator

Case - control for discrete traits

How common disease is

High risk allele frequency (A) : 0.2 (0 - 1)

Prevalence : 0.05 (0.0001 - 0.9999)

Genotype relative risk Aa : 1.2 (>1)

Genotype relative risk AA : 1.44 (>1)

D-prime : 1 (0 - 1)

Marker allele frequency (B) : 0.2 (0 - 1)

Number of cases : 1000 (0 - 10000000)

Control : case ratio : 1 (>0)
(1 = equal number of cases and controls)

☐ Unselected controls? (* see below)

User-defined type I error rate : 5e-8 (0.00000001 - 0.5)

User-defined power: determine N : 0.80 (0 - 1)
(1 - type II error rate)

Created by [Shaun Purcell](#) 24.Oct.2008

X Find: exclude Next Previous Highlight all Match case

Boston: Wed 12:49 UK: Wed 17:49 Netherlands: Wed 18:49 Hong Kong: Thu 01:49 Los Angeles: Wed 09:49 Stopped

start 2 F. 2 W. 2 M. 3 M. 5 M. RG... ma... silv... 92% 12:49 PM

Practical Exercise

Statistical Genetics Group - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://pngu.mgh.harvard.edu/~purcell/gpc/cc2.html

Most Visited Smart Bookmarks Free Hotmail Amazon.com: Statistic... The Arreat Summit - It... Share on Facebook

Gmail - Update: ... IBG workshop:2009:... IBG Index of /worksh... Wiley InterScien... Hulu - Videos Cypriots and U.N... Statistical Ge...

Genetic Power Calculator

Case - control for discrete traits

High risk allele frequency (A) : 0.2 (0 - 1)

Prevalence : 0.05 (0.0001 - 0.9999)

Genotype relative risk Aa : 1.2 (> 1)

Genotype relative risk AA : 1.44 (> 1)

D-prime : 1 (0 - 1)

Marker allele frequency (B) : 0.2 (0 - 1)

Number of cases : 1000 (0 - 10000000)

Control : case ratio : 1 (> 0)
(1 = equal number of cases and controls)

☐ Unselected controls? (* see below)

User-defined type I error rate : 5e-8 (0.00000001 - 0.5)

User-defined power: determine N : 0.80 (0 - 1)
(1 - type II error rate)

Created by [Shaun Purcell](#) 24.Oct.2008

Find: exclude Next Previous Highlight all Match case

Boston: Wed 12:49 UK: Wed 17:49 Netherlands: Wed 18:49 Hong Kong: Thu 01:49 Los Angeles: Wed 09:49 Stopped

start 2 F. 2 W. 2 M. 3 M. 5 M. RG... ma... silv... 92% 12:49 PM

This is the relative risk—not the odds ratio. The OR is approximately equivalent to the RR for small values of RR.

Practical Exercise

Statistical Genetics Group - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://pngu.mgh.harvard.edu/~purcell/gpc/cc2.html

Most Visited Smart Bookmarks Free Hotmail Amazon.com: Statistic... The Arreat Summit - It... Share on Facebook

Gmail - Update: ... IBG workshop:2009:... IBG Index of /worksh... Wiley InterScien... Hulu - Videos Cypriots and U.N... Statistical Ge...

Genetic Power Calculator

Case - control for discrete traits

High risk allele frequency (A)	: 0.2	(0 - 1)
Prevalence	: .05	(0.0001 - 0.9999)
Genotype relative risk Aa	: 1.2	(> 1)
Genotype relative risk AA	: 1.44	(> 1)
D-prime	: 1	(0 - 1)
Marker allele frequency (B)	: 0.2	(0 - 1)
Number of cases	: 1000	(0 - 10000000)
Control : case ratio	: 1	(> 0) (1 = equal number of cases and controls)
<input type="checkbox"/> Unselected controls? (* see below)		
User-defined type I error rate	: 5e-8	(0.00000001 - 0.5)
User-defined power: determine N (1 - type II error rate)	: 0.80	(0 - 1)

Created by [Shaun Purcell](#) 24.Oct.2008

X Find: exclude Next Previous Highlight all Match case

Boston: Wed 12:49 UK: Wed 17:49 Netherlands: Wed 18:49 Hong Kong: Thu 01:49 Los Angeles: Wed 09:49 Stopped

start 2 F. 2 W. 2 M. 3 M. 5 M. RG... ma... silv... 92% 12:49 PM

Risk of the AA genotype. Note that the model of risk is defined by the relationship between Aa and AA. We have a multiplicative model because $1.44 = 1.2 \times 1.2$.

Practical Exercise

Statistical Genetics Group - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://pngu.mgh.harvard.edu/~purcell/gpc/cc2.html

Most Visited Smart Bookmarks Free Hotmail Amazon.com: Statistic... The Arreat Summit - It... Share on Facebook

Gmail - Update: ... IBG workshop:2009:... IBG Index of /worksh... Wiley InterScien... Hulu - Videos Cypriots and U.N... Statistical Ge...

Genetic Power Calculator

Case - control for discrete traits

High risk allele frequency (A)	: 0.2	(0 - 1)
Prevalence	: .05	(0.0001 - 0.9999)
Genotype relative risk Aa	: 1.2	(>1)
Genotype relative risk AA	: 1.44	(>1)
D-prime	: 1	(0 - 1)
Marker allele frequency (B)	: 0.2	(0 - 1)
Number of cases	: 1000	(0 - 10000000)
Control : case ratio	: 1	(>0) (1 = equal number of cases and controls)
<input type="checkbox"/> Unselected controls? (* see below)		
User-defined type I error rate	: 5e-8	(0.00000001 - 0.5)
User-defined power: determine N (1 - type II error rate)	: 0.80	(0 - 1)

Created by [Shaun Purcell](#) 24.Oct.2008

X Find: exclude Next Previous Highlight all Match case

Boston: Wed 12:49 UK: Wed 17:49 Netherlands: Wed 18:49 Hong Kong: Thu 01:49 Los Angeles: Wed 09:49 Stopped

start 2 F. 2 W. 2 M. 3 M. 5 M. RG... ma... silv... 92% 12:49 PM

The LD statistic D' which represents recombination patterns historically. $D' +$ allele frequency at the typed locus information yields r^2

Practical Exercise

Statistical Genetics Group - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://pngu.mgh.harvard.edu/~purcell/gpc/cc2.html

Most Visited Smart Bookmarks Free Hotmail Amazon.com: Statistic... The Arreat Summit - It... Share on Facebook

Gmail - Update: ... IBG workshop:2009:... IBG Index of /worksh... Wiley InterScien... Hulu - Videos Cypriots and U.N... Statistical Ge...

Genetic Power Calculator

Case - control for discrete traits

Sample size for cases

High risk allele frequency (A)	: 0.2	(0 - 1)
Prevalence	: .05	(0.0001 - 0.9999)
Genotype relative risk Aa	: 1.2	(>1)
Genotype relative risk AA	: 1.44	(>1)
D-prime	: 1	(0 - 1)
Marker allele frequency (B)	: 0.2	(0 - 1)
Number of cases	: 1000	(0 - 10000000)
Control : case ratio	: 1	(>0) (1 = equal number of cases and controls)
<input type="checkbox"/> Unselected controls? (* see below)		
User-defined type I error rate	: 5e-8	(0.00000001 - 0.5)
User-defined power: determine N (1 - type II error rate)	: 0.80	(0 - 1)

Created by [Shaun Purcell](#) 24.Oct.2008

Find: exclude Next Previous Highlight all Match case

Boston: Wed 12:49 UK: Wed 17:49 Netherlands: Wed 18:49 Hong Kong: Thu 01:49 Los Angeles: Wed 09:49 Stopped

start 2 F. 2 W. 2 M. 3 M. 5 M. RG... ma... silv... 92% 12:49 PM

Practical Exercise

Statistical Genetics Group - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://pngu.mgh.harvard.edu/~purcell/gpc/cc2.html

Most Visited Smart Bookmarks Free Hotmail Amazon.com: Statistic... The Arreat Summit - It... Share on Facebook

Gmail - Update: ... IBG workshop:2009:... IBG Index of /worksh... Wiley InterScien... Hulu - Videos Cypriots and U.N... Statistical Ge...

Genetic Power Calculator

Case - control for discrete traits

Ratio of Controls to Cases

High risk allele frequency (A)	: 0.2	(0 - 1)
Prevalence	: .05	(0.0001 - 0.9999)
Genotype relative risk Aa	: 1.2	(>1)
Genotype relative risk AA	: 1.44	(>1)
D-prime	: 1	(0 - 1)
Marker allele frequency (B)	: 0.2	(0 - 1)
Number of cases	: 1000	(0 - 10000000)
Control : case ratio	: 1	(>0) (1 = equal number of cases and controls)
<input type="checkbox"/> Unselected controls? (* see below)		
User-defined type I error rate	: 5e-8	(0.00000001 - 0.5)
User-defined power: determine N (1 - type II error rate)	: 0.80	(0 - 1)

Created by [Shaun Purcell](#) 24.Oct.2008

Find: exclude Next Previous Highlight all Match case

Boston: Wed 12:49 UK: Wed 17:49 Netherlands: Wed 18:49 Hong Kong: Thu 01:49 Los Angeles: Wed 09:49 Stopped

start 2 F. 2 W. 2 M. 3 M. 5 M. RG... ma... silv... 92% 12:49 PM

Practical Exercise

Statistical Genetics Group - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://pngu.mgh.harvard.edu/~purcell/gpc/cc2.html

Most Visited Smart Bookmarks Free Hotmail Amazon.com: Statistic... The Arreat Summit - It... Share on Facebook

Gmail - Update: ... IBG workshop:2009:... IBG Index of /worksh... Wiley InterScien... Hulu - Videos Cypriots and U.N... Statistical Ge...

Genetic Power Calculator

Case - control for discrete traits

High risk allele frequency (A) : 0.2 (0 - 1)
Prevalence : .05 (0.0001 - 0.9999)
Genotype relative risk Aa : 1.2 (>1)
Genotype relative risk AA : 1.44 (>1)

D-prime : 1 (0 - 1)
Marker allele frequency (B) : 0.2 (0 - 1)

Number of cases : 1000 (0 - 10000000)
Control : case ratio : 1 (>0)
(1 = equal number of cases and controls)

☐ Unselected controls? (* see below)

User-defined type I error rate : 5e-8 (0.00000001 - 0.5)
User-defined power: determine N : 0.80 (0 - 1)
(1 - type II error rate)

Created by [Shaun Purcell](#) 24.Oct.2008

X Find: exclude Next Previous Highlight all Match case

Boston: Wed 12:49 UK: Wed 17:49 Netherlands: Wed 18:49 Hong Kong: Thu 01:49 Los Angeles: Wed 09:49 Stopped

start 2 F. 2 W. 2 M. 3 M. 5 M. RG... ma... silv... 92% 12:49 PM

Genome-wide significance threshold
We'll learn about this later in the session

Practical Exercise

Statistical Genetics Group - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://pngu.mgh.harvard.edu/~purcell/gpc/cc2.html

Most Visited Smart Bookmarks Free Hotmail Amazon.com: Statistic... The Arreat Summit - It... Share on Facebook

Gmail - Update: ... IBG workshop:2009:... IBG Index of /worksh... Wiley InterScien... Hulu - Videos Cypriots and U.N... Statistical Ge...

Genetic Power Calculator

Case - control for discrete traits

High risk allele frequency (A) : 0.2 (0 - 1)

Prevalence : .05 (0.0001 - 0.9999)

Genotype relative risk Aa : 1.2 (>1)

Genotype relative risk AA : 1.44 (>1)

D-prime : 1 (0 - 1)

Marker allele frequency (B) : 0.2 (0 - 1)

Number of cases : 1000 (0 - 10000000)

Control : case ratio : 1 (>0)
(1 = equal number of cases and controls)

☐ Unselected controls? (* see below)

User-defined type I error rate : 5e-8 (0.00000001 - 0.5)

User-defined power: determine N : 0.80 (0 - 1)
(1 - type II error rate)

Created by [Shaun Purcell](#) 24.Oct.2008

Find: exclude Next Previous Highlight all Match case

Boston: Wed 12:49 UK: Wed 17:49 Netherlands: Wed 18:49 Hong Kong: Thu 01:49 Los Angeles: Wed 09:49 Stopped

start 2 F. 2 W. 2 M. 3 M. 5 M. RG... ma... silv... 92% 12:49 PM

Power level—what we're interested in observing

Practical Exercise

Statistical Genetics Group - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://pngu.mgh.harvard.edu/~purcell/gpc/cc2.html

Most Visited Smart Bookmarks Free Hotmail Amazon.com: Statistic... The Arreat Summit - It... Share on Facebook

Gmail - Update: ... IBG workshop:2009:... IBG Index of /worksh... Wiley InterScien... Hulu - Videos Cypriots and U.N... Statistical Ge...

Genetic Power Calculator

Case - control for discrete traits

High risk allele frequency (A) : 0.2 (0 - 1)

Prevalence : .05 (0.0001 - 0.9999)

Genotype relative risk Aa : 1.2 (>1)

Genotype relative risk AA : 1.44 (>1)

D-prime : 1 (0 - 1)

Marker allele frequency (B) : 0.2 (0 - 1)

Number of cases : 1000 (0 - 10000000)

Control : case ratio : 1 (>0)
(1 = equal number of cases and controls)

☐ Unselected controls? (* see below)

User-defined type I error rate : 5e-8 (0.00000001 - 0.5)

User-defined power: determine N : 0.80 (0 - 1)
(1 - type II error rate)

Click here to process

Created by [Shaun Purcell](#) 24.Oct.2008

Find: exclude Next Previous Highlight all Match case

Boston: Wed 12:49 UK: Wed 17:49 Netherlands: Wed 18:49 Hong Kong: Thu 01:49 Los Angeles: Wed 09:49 Stopped

start 2 F. 2 W. 2 M. 3 M. 5 M. RG... ma... silv... 92% 12:49 PM

Practical Exercise

Genetic Power Calculator - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://pngu.mgh.harvard.edu/~purcell/cgi-bin/cc2k.cgi

Most Visited Smart Bookmarks Free Hotmail Amazon.com: Statistic... The Arreat Summit - It... Share on Facebook

Gmail - Update: ... IBG workshop:2009:... IBG Index of /worksh... Wiley InterScien... Hulu - Videos Cypriots and U.N... Genetic Powe...

Alpha	Power	N cases for 80% power
0.1	0.4374	2803
0.05	0.3177	3559
0.01	0.1377	5296
0.001	0.0355	7742
5e-08	3.674e-05	17958

Case-control statistics: general 2 df test (BB versus Bb versus bb)
Sample NCP = 6.216

Alpha	Power	N cases for 80% power
0.1	0.716	1240
0.05	0.6002	1550
0.01	0.3609	2233
0.001	0.1451	3163
5e-08	0.0007464	6920

Case-control statistics: allelic 1 df test (B versus b)
Sample NCP = 6.224

Scroll to the bottom for answer

Alpha	Power	N cases for 80% power
0.1	0.8024	993
0.05	0.7037	1260
0.01	0.4677	1876
0.001	0.2131	2743
5e-08	0.001557	6362

Controls are selected (i.e. screened for not being a case)

Find: exclude Next Previous Highlight all Match case

Bosoton: Wed 13:02 UK: Wed 18:02 Netherlands: Wed 19:02 Hong Kong: Thu 02:02 Los Angeles: Wed 10:02 Done

start 2 F... 2 W... 2 M... 2 M... 2 M... RG... mac... silv... 94% 1:02 PM

Practical Exercise

- ▶ What power do we have to detect a locus with an odds ratio of 1.5 at genome-wide significance assuming a multiplicative disease model, a disease allele frequency of 5%, a disease prevalence of 0.5% and 3000 cases and controls?

Genetic Power Calculator

Case - control for discrete traits

High risk allele frequency (A) : 0.05 (0 - 1)
Prevalence : 0.005 (0.0001 - 0.9999)
Genotype relative risk Aa : 1.5 (>1)
Genotype relative risk AA : 2.25 (>1)

D-prime : 1 (0 - 1)
Marker allele frequency (B) : 0.05 (0 - 1)

Number of cases : 3000 (0 - 10000000)
Control : case ratio : 1 (>0)
(1 = equal number of cases and controls)

☐ Unselected controls? (* see below)

User-defined type I error rate : 0.00000005 (0.00000001 - 0.5)
User-defined power: determine N : 0.80 (0 - 1)
(1 - type II error rate)

Process Reset

Created by [Shaun Purcell](#) 24.Oct.2008

Note : unselected controls indicates a true random population sample (e.g. for a 1% disease, 1% of controls would also, by chance, have the disease); if this is unchecked (the

0.01	0.2124	11085
0.001	0.06518	16206
$5e-08$	0.0001204	37587

Case-control statistics: general 2 df test (BB versus Bb versus bb)

Sample NCP = 28.11

Alpha	Power	N cases for 80% power
0.1	0.9995	822
0.05	0.9986	1028
0.01	0.9916	1481
0.001	0.9553	2098
$5e-08$	0.3425	4590

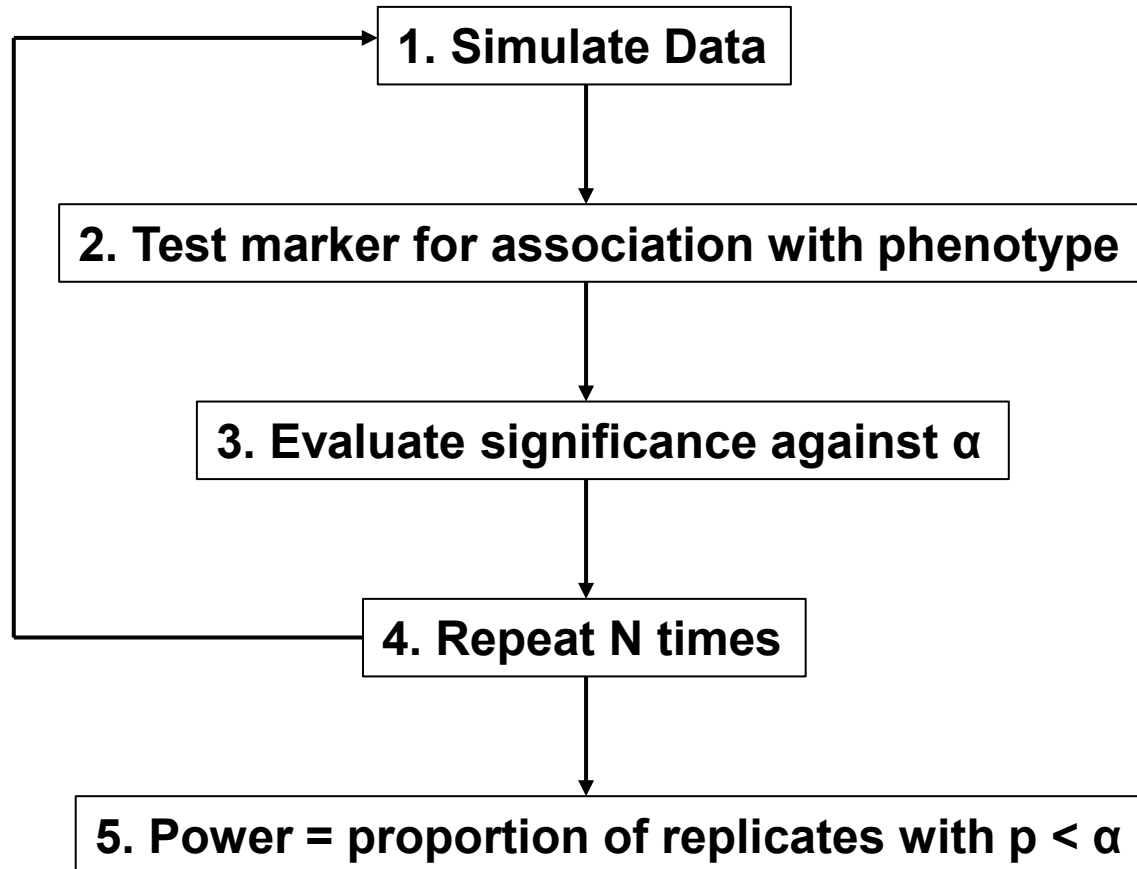
Case-control statistics: allelic 1 df test (B versus b)

Sample NCP = 28.18

Alpha	Power	N cases for 80% power
0.1	0.9999	658
0.05	0.9996	835
0.01	0.9969	1243
0.001	0.9782	1818
$5e-08$	0.443	4216

Controls are selected (i.e. screened for not being a case)

Empirical power estimation



```
rm(list=ls())
set.seed(12345) #Set seed to enable repeatability
p <- 0.5 #(Decreaser) Allele frequency
q <- 1 - p #Other allele
Nsample <- 1000 #Set sample size
Nrep <- 1000 #Set number of replications
Nsig <- 0 #Count variable for number of significant replicates
threshold <- 0.05 #Threshold for declaring significance
beta <- 0.1 #Effect size
a <- sqrt(1/(2*p*q)) #Additive value
residvar <- 1 - beta^2 #Residual variance
for (i in seq(1,Nrep)) { #Loop over replicates
#Simulate genotypes assuming HWE
  x <- sample(x=c(-a,0,a) ,size=Nsample, replace=TRUE, prob=c(p^2,2*p*q,q^2))
  y <- beta*x + rnorm(n=Nsample, mean=0, sd=sqrt(residvar))
  if(summary(lm(y~x))$coefficients[2,4] < threshold) {Nsig=Nsig+1}
}
power = Nsig/Nrep #Power is proportion of significant replicates
Power #0.885
```

Trivial Example: OLS Linear Regression

$$\frac{(b - 0)^2}{SE^2} \longrightarrow \chi^2_1\left(\frac{\beta^2}{SE^2}\right)$$

1. Set type 1 error rate (e.g. $\alpha = 0.05$)
2. Determine what critical value this corresponds to on the X axis

```
qchisq(p = 0.05, df = 1, ncp = 0, lower.tail = FALSE, log.p = FALSE)
```

```
[1] 3.841459
```

3. Work out the non-centrality parameter of the test $\frac{\beta^2}{SE^2}$

$$\beta = 0.1$$

$$SE \approx 1/\sqrt{N} \quad (\text{Assume } N = 1000)$$

4. Calculate the area to the right of the threshold under H_1

```
pchisq(q=3.84, df=1, ncp = 0.1^2/(1/1000), lower.tail = FALSE, log.p = FALSE)
```

```
[1] 0.8854512
```

4. Multiple Testing

Genome-wide Association



High throughput genotyping

Other Multiple Testing Considerations

- Genome-wide association is really bad
 - At 1 test per SNP for 500,000 SNPs
 - 25,000 expected to be significant at $p < 0.05$, by chance alone

Other Multiple Testing Considerations

- Genome-wide association is really bad
 - At 1 test per SNP for 500,000 SNPs
 - 25,000 expected to be significant at $p < 0.05$, by chance alone
- To make things worse
 - Dominance (additive/dominant/recessive)
 - Epistasis (multiple combinations of SNPs)
 - Multiple phenotype definitions
 - Subgroup analyses
 - Multiple analytic methods

Bonferroni Correction

- For testing 500,000 SNPs
 - 5,000 expected to be significant at $p < 0.01$
 - 500 expected to be significant at $p < 0.001$
 -
 - 0.05 expected to be significant at $p < 0.0000001$
- Suggests setting significance level to $\alpha = 10^{-7}$ *
- Bonferroni correction for m tests
 - set significance level for p-values to $\alpha = 0.05 / m$
 - (or adjust the p-values to $m \times p$, before applying the usual $\alpha = 0.05$ significance level)
- *See Risch and Merikangas 1999

Genome-wide Significance

- Multiple testing theory requires an estimate of the number of ‘independent tests’
- Risch and Merikangas 1996 estimated a threshold of $10^{-6} = (0.05/(5*10,000))$
- HapMap 2005 estimate 10^{-8} based on encode deep sequencing in ENCODE regions
- Dudbridge and Gusnato, and Pe'er et al. 2008 Genetic Epidemiology estimate based on ‘infinite density’ like Lander and Kruglyak 1995 generate 5×10^{-8}

5. Replication

Replication

- Replicating the genotype-phenotype association is the “gold standard” for “proving” an association is genuine
- Most loci underlying complex diseases will not be of large effect
- It is unlikely that a single study will unequivocally establish an association without the need for replication

Winner's Curse

If the location of a variant and its phenotypic effect size are estimated from the same data sets, the effect size will be over-estimated, in many cases substantially. Statistical significance and the *estimated* magnitude of the parameter are highly correlated.

H Göring et al. Am J Hum Genetics 2001;69:1357-69

Guidelines for Replication

Replication studies should be of sufficient size to demonstrate the effect

Replication studies should be conducted in independent datasets

Replication should involve the same phenotype

Replication should be conducted in a similar population

The same SNP should be tested

The replicated signal should be in the same direction

Joint analysis should lead to a lower p value than the original report

Well designed negative studies are valuable

Mendelian randomization Power

<http://cnsngenomics.com/shiny/mRnd/>

The screenshot shows the mRnd web application interface. The browser tab is titled 'mRnd: Power calculation' and the address bar shows 'cnsngenomics.com/shiny/mRnd/'. The page title is 'mRnd: Power calculations for Mendelian Randomization'. The interface is divided into several sections:

- Input:** Contains fields for 'Calculate:' (Power selected, Sample size), 'Provide:' (Sample size: 1000), ' α ' (0.05), 'Type-I error rate', ' β_{yz} ' (0), 'The regression coefficient β_{yz} for the true underlying causal association between the exposure (X) and outcome (Y) variables', ' β_{OLS} ' (0), 'The regression coefficient β_{OLS} for the observational association between the exposure (X) and outcome (Y) variables', and ' R^2_{Zz} ' (0.01), 'Proportion of variance explained for the association between the SNP or allele score (Z) and'.
- Continuous outcome:** Selected tab. Sub-sections include:
 - Two-stage least squares:** Shows Power (0.05), NCP (0.00), Non-Centrality-Parameter, and F-statistic (11.10). The strength of the instrument.
 - Power or sample size calculations for two-stage least squares Mendelian Randomization studies using a genetic instrument Z (a SNP or allele score), a continuous exposure variable X (e.g. body mass index [BMI, $\frac{kg}{m^2}$]) and a continuous outcome variable Y (e.g. blood pressure [mmHg]).**
 - YZ association:** Shows Power (0.05), NCP (0.00), Non-Centrality-Parameter.
 - Power or sample size calculations for the regression association of a genetic instrument Z (e.g. a BMI SNP), with a continuous outcome variable Y (blood pressure).**
 - Working Example:** Provides a detailed example of calculating the minimum required sample size for an MR study. It includes text about the association of BMI and SBP in children, the regression coefficients for BMI and SBP, the SD for SBP, and the calculation of the power of an MR study using the following parameters:
$$\beta_{OLS} = 1.41 \frac{mmHg}{SD}$$
$$\beta_{yz} = 1.3 \frac{mmHg}{SD} |^+$$
$$\sigma^2(x) = 1$$
$$\sigma^2(y) = 10.8^2 = 116.6 mmHg^2$$
For an α of 0.05 and power of 0.8, the calculated minimum sample size for the Mendelian Randomization study is $N = 53,218$. The reason why this sample size is so large is because BMI explains a small amount of variation in SBP in this case and because the genetic instrument explains a small proportion of variance in BMI.

Mendelian randomization Power

<http://cnsgenomics.com/shiny/mRnd/>

mRnd: Power calculation × + -

cnsgenomics.com/shiny/mRnd/

mRnd: Power calculations for Mendelian Randomization

Continuous outcome Binary outcome Binary outcome derivations Citation About

Calculate:

☒ Power
☐ Sample size

Provide:

Sample size

1000

α

0.05

Type-I error rate

K

0

Proportion of cases in the study

OR

0

True odds ratio of the outcome variable per standard deviation of the exposure variable

R^2_{xx}

0.01

Proportion of variance explained for the association between the SNP or allele score (Z) and the exposure variable (X)

Power	NA	
NCP	NA	Non-Centrality-Parameter
F-statistic	11.10	The strength of the instrument

Type here to search

3:28 PM 31/05/2019