SISG Brisbane Module 10: Statistical & Quantitative Genetics of Disease

Lecture 3
Single locus model of disease risk

Naomi Wray

Aims of Lecture 3

Theory

- Single locus disease model
- Power calculations

Single locus disease model

Single locus disease model:

G = genotype; D=disease; K = overall disease risk in population;

p = risk allele frequency;

	P(G)
aa	$(1\text{-}p)^2$
Aa	$2p(1\text{-}p)$
AA	p^2

Single locus disease model

Single locus disease model:

G = genotype; D=disease; K = overall disease risk in population;

p = risk allele frequency;

f0 = baseline risk for homozygote non-risk allele – UNKNOWN

R = relative risk for heterozygote; assume risk is multiplicative (on this scale)

	P(G)	P(D|G)
aa	$(1\text{-}p)^2$	f_0
Aa	$2p(1\text{-}p)$	f_0R
AA	p^2	f_0R^2

Single locus disease model

Single locus disease model:

G = genotype; D=disease; K = overall disease risk in population;

p = risk allele frequency;

f0 = baseline risk for homozygote non-risk allele – UNKNOWN

R = relative risk for heterozygote; assume risk is multiplicative (on this scale)

	P(G)	P(D|G)	P(D) =P(D|G)p(G)
aa	$(1\text{-}p)^2$	f_0	$(1\text{-}p)^2 f_0$
Aa	$2p(1\text{-}p)$	f_0R	$2p(1\text{-}p) f_0R$
AA	p^2	f_0R^2	$p^2 f_0R^2$
			Sum= K

$$P(\text{Disease})=K=f_0(1\text{-}p)^2 + f_0R2p(1\text{-}p) + f_0R^2p^2 = f_0(1+p(R\text{-}1))^2$$

$$f_0=K/(1+p(R\text{-}1))^2$$

Single locus disease model

Single locus disease model:

G = genotype; D=disease; K = overall disease risk in population;

p = risk allele frequency;

f0 = baseline risk for homozygote non-risk allele – UNKNOWN

R = relative risk for heterozygote; assume risk is multiplicative (on this scale)

	P(G)	P(D|G)	P(D) =P(D|G)p(G)	P(G|D) =P(G)/P(D)
aa	$(1\text{-}p)^2$	f_0	$(1\text{-}p)^2 f_0$	$(1\text{-}p)^2 f_0/K$
Aa	$2p(1\text{-}p)$	f_0R	$2p(1\text{-}p) f_0R$	$2p(1\text{-}p) f_0R/K$
AA	p^2	f_0R^2	$p^2 f_0R^2$	$p^2 f_0R^2/K$
			Sum= K	

$$P(\text{Disease})=K=f_0(1\text{-}p)^2 + f_0R2p(1\text{-}p) + f_0R^2p^2 = f_0(1+p(R\text{-}1))^2$$

$$f_0=K/(1+p(R\text{-}1))^2$$

Practical

Single locus disease model:

G = genotype; D=disease; K = overall disease risk in population = 0.01;

p = risk allele frequency = 0.2;

f0 = baseline risk for homozygote non-risk allele – UNKNOWN

R = relative risk for heterozygote; assume risk is multiplicative (on this scale) = 1.2

	P(G)	P(D|G)	P(D) =P(D|G)p(G)	P(G|D) =P(G)/P(D)
aa				
Aa				
AA				

$$P(\text{Disease}) = K = f_0(1-p)^2 + f_0 R 2p(1-p) + f_0 R^2 p^2 = f_0(1+p(R-1))^2$$

$$f_0 = K/(1+p(R-1))^2$$

Using the single locus disease model to calculate power in an association study

What is power?

When we set up a statistical test

- The null hypothesis is EITHER
  - true
  - false
- With the data available we EITHER
  - reject the null hypothesis
  - fail to reject the null hypothesis

	Null hypothesis is true	**Null hypothesis is false**
Reject the null hypothesis	Type I error False positive	Correct Outcome True positive
Fail to reject the null hypothesis	Correct Outcome True negative	Type II error False Negative

Power = probability of rejecting the null hypothesis when the null hypothesis is false

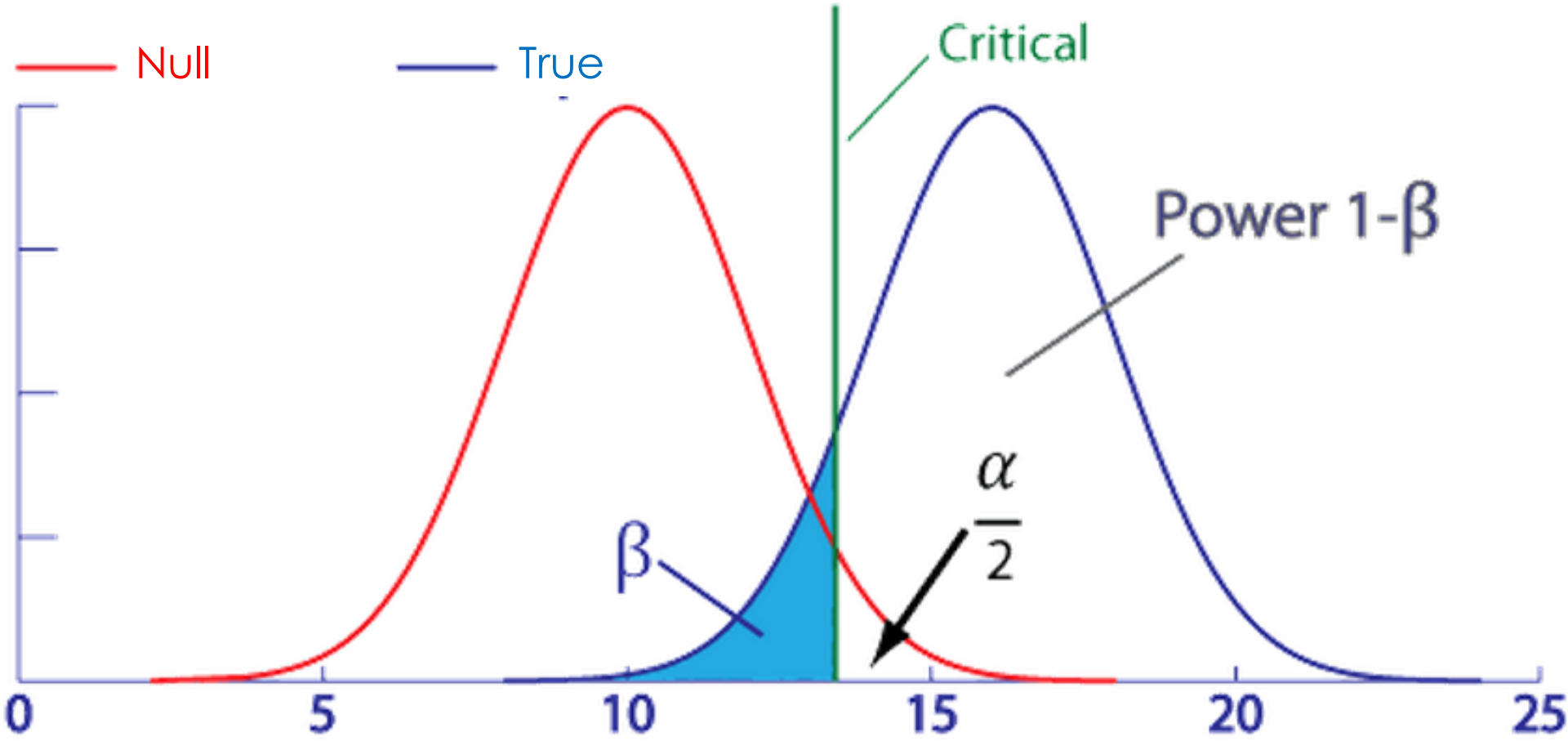
=1 –probability of failing to reject the null hypothesis when the null hypothesis is false

= 1- probability(Type II error)

Power depends on statistical test, effect size to be detected, sample size, acceptable level of Type I error

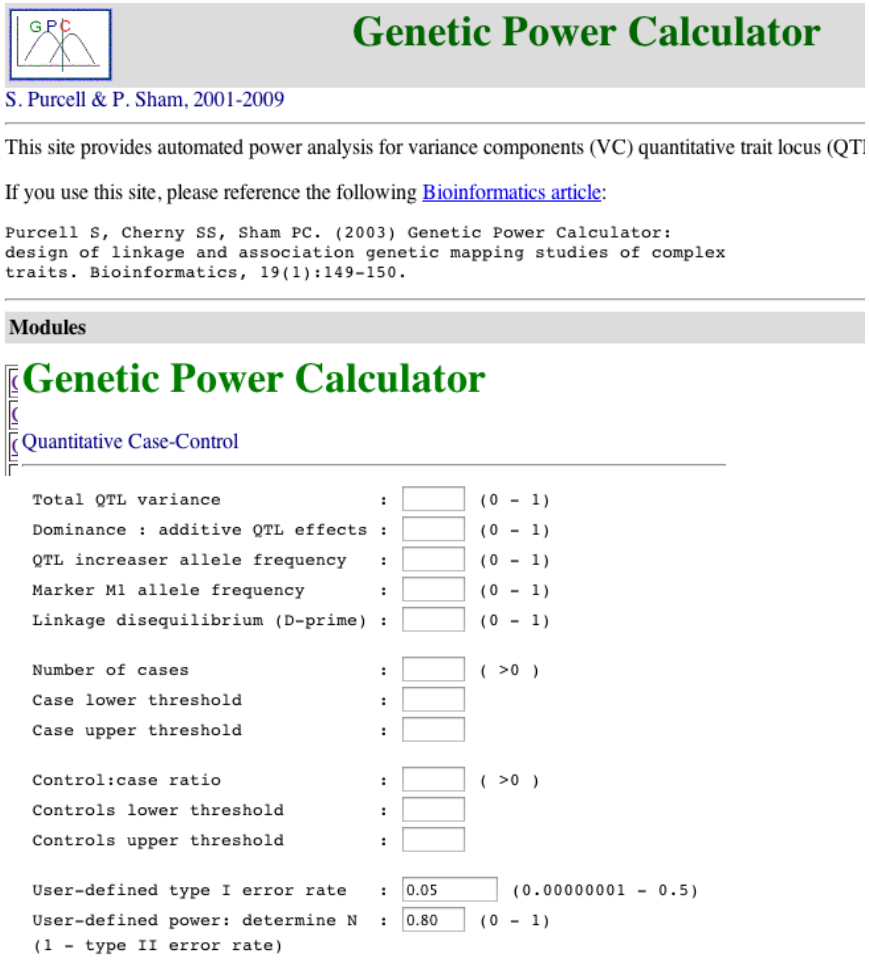
Non-centrality parameter depends on statistical test, effect size to be detected, sample size

Power



β = probability of rejecting the null hypothesis when the alternative hypothesis is true
α = probability of rejecting the null hypothesis when the null hypothesis is true
Variance about mean values depends on sample size

Genetic Power Calculator



Genetic Power Calculator

S. Purcell & P. Sham, 2001-2009

This site provides automated power analysis for variance components (VC) quantitative trait locus (QT

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

Genetic Power Calculator

Quantitative Case-Control

Total QTL variance	:		(0 - 1)
Dominance : additive QTL effects	:		(0 - 1)
QTL increaser allele frequency	:		(0 - 1)
Marker M1 allele frequency	:		(0 - 1)
Linkage disequilibrium (D-prime)	:		(0 - 1)
Number of cases	:		(>0)
Case lower threshold	:		
Case upper threshold	:		
Control:case ratio	:		(>0)
Controls lower threshold	:		
Controls upper threshold	:		
User-defined type I error rate	:	0.05	(0.00000001 - 0.5)
User-defined power: determine N (1 - type II error rate)	:	0.80	(0 - 1)

Send Clear Form

Genetic Power Calculator

Case - control for discrete traits

High risk allele frequency (A)	:		(0 - 1)
Prevalence	:		(0.0001 - 0.9999)
Genotype relative risk Aa	:		(>1)
Genotype relative risk AA	:		(>1)
D-prime	:		(0 - 1)
Marker allele frequency (B)	:		(0 - 1)
Number of cases	:		(0 - 10000000)
Control : case ratio	:		(>0) (1 = equal number of cases and controls)
Unselected controls? (* see below)			
User-defined type I error rate	:	0.05	(0.00000001 - 0.5)
User-defined power: determine N (1 - type II error rate)	:	0.80	(0 - 1)

Process Reset

Created by Shaun Purcell 24.Oct.2008

Genetic Power Calculator

Case - control for discrete traits

High risk allele frequency (A) : .2 (0 - 1)
Prevalence : .01 (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) : .2 (0 - 1)

Number of cases : 5000 (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)

Case-control statistics: allelic 1 df test (B versus b)
Sample NCP = 28.59

Alpha	Power	N cases for 80% power
0.1	0.9999	1081
0.05	0.9996	1372
0.01	0.9972	2042
0.001	0.9802	2985
5e-08	0.4586	6924

Power of a case-control study

Power of a disease trait

p = frequency of risk allele in population

p_{case} = frequency of risk allele in cases

p_{cont} = frequency of risk allele in controls

v = proportion of a sample of N that are cases

$\bar{p}$ = mean allele frequency across cases and controls

= v p_{case} + (1-v) $p_{control}$

Yang et al (2009)

Power of a case-control study

Power of a disease trait

p = frequency of risk allele in population

p_{case} = frequency of risk allele in cases

p_{cont} = frequency of risk allele in controls

v = proportion of a sample of N that are cases

$\bar{p}$ = mean allele frequency across cases and controls

$\bar{p}$ = v p_{case} + (1-v) $p_{control}$

Z-Test statistic of association = test of difference of two proportions =

$$\frac{p_{case} - p_{cont}}{s.e.(pooled\ sample\ p)} = \frac{p_{case} - p_{cont}}{s.e.(\bar{p})}$$

χ^2 non-centrality parameter $= NCP_{01} = \dfrac{(p_{case} - p_{cont})^2}{var(\bar{p})}$

$$var(\bar{p}) = \frac{1}{2}\bar{p}(1-\bar{p})\left(\frac{1}{Nv} + \frac{1}{N(1-v)}\right)$$

Allele Frequency in Cases

Single locus disease model:
G = genotype; D=disease; K = overall disease risk in population

	P(G)	P(D|G)	P(D) =P(D|G)p(G)	P(G|D) =P(G)/P(D)
aa	$(1-p)^2$	f_0	$(1-p)^2 f_0$	$(1-p)^2 f_0/K$
Aa	$2p(1-p)$	f_0R	$2p(1-p) f_0R$	$2p(1-p) f_0R/K$
AA	p^2	f_0R^2	$p^2 f_0R^2$	$p^2 f_0R^2/K$
			Sum= K	

$P(\text{Disease})=K=f_0(1-p)^2 + f_0R2p(1-p) + f_0R^2p^2 = f_0(1+p(R-1))^2$

$f_0=K/(1+p(R-1))^2$

p_{case} = ½ P(Aa|D)+P(AA|D) Allele frequency in cases

$$= f_0 pR\big((1-p)+pR\big)/K = \frac{pR}{(1+p(R-1))}$$

Find allele frequency in controls in the same way

$$p_{cont} = \frac{p}{1-K}\left(1-\frac{KR}{(1+p(R-1))}\right)$$

Allele Frequency in Controls

Single locus disease model:
G = genotype; D=disease; K = overall disease risk in population

	P(G)	P(D'|G)	P(D') =P(D'|G)p(G)	P(G|D') =P(G)/P(D')
aa	$(1-p)^2$	$(1-f_0)$	$(1-p)^2\ (1-f_0)$	$(1-p)^2\ (1-f_0)/(1-K)$
Aa	$2p(1-p)$	$(1-f_0R)$	$2p(1-p)\ (1-f_0R)$	$2p(1-p)\ (1-f_0R)/(1-K)$
AA	p^2	$(1-f_0R^2)$	$p^2\ (1-f_0R^2)$	$p^2\ (1-f_0R^2)/(1-K)$
			Sum= 1-K	

$f_0=K/(1+p(R-1))^2$

$p_{control}$ = ½ P(Aa|D')+P(AA|D') Allele frequency in controls

$$= \frac{p}{1-K}\left(1-\frac{KR}{(1+p(R-1))}\right)$$

Power of a case-control study

$NCP_{01} = \dfrac{(p_{case} - p_{cont})^2}{var(\bar{p})}$

α = significance level - acceptable level of type I error

$t = \Phi^{-1}\left(\frac{\alpha}{2}\right)$ Normal distribution threshold above which null hypothesis will be rejected

$Power = \Phi\left(\sqrt{NCP_{01}} + t\right)$

N=10000,v=0.5,p=0.2,R=1.2,K=0.01,a=5e-8,K=0.01, power = 0.46

Agrees with the genetic power calculator

Yang et al (2009) Comparing Apples and Oranges: Equating the Power of Case-Control and Quantitative Trait Association Studies. Genetic Epidemiology

Yang et al (2009) Comparing Apples and Oranges: Equating the Power of Case-Control and Quantitative Trait Association Studies. Genetic Epidemiology

Research Question in 2009:
We had GWAS success with height but not with disease.

Was this a function of power?

For the same sample size what is the connection between power for a quantitative trait vs case-control?

Answer:

- "So a planned meta-analysis for height on 120,000 individuals has power equivalent to a CC study on 33,100 schizophrenia cases and 33,100 controls, a size not yet achievable for this disease."

Power of a case-control association study expressed in terms of variance explained by the locus

χ^2 *non-centrality parameter* $= NCP_{01} = \dfrac{(p_{case} - p_{cont})^2}{var(\bar{p})}$

$$NCP_{01} = \frac{2\bar{p}(1-\bar{p})(R-1)^2 v(1-v)N}{(1-K)^2(1+p(R-1))^2}$$

If R is small then $(1+\mathrm{p(R\text{-}1)})^2 \approx 1$ e.g., p=0.2, R=1.2, $(1+\mathrm{p(R\text{-}1)})^2 = 1.08$

Variance explained by a locus $= h^2_{L[j]} \approx \dfrac{2p(1-p)(R-1)^2}{i^2}$

$$NCP_{01} \approx \frac{h^2_{L[j]} i^2 v(1-v)N}{(1-K)^2}$$

Yang et al (2009) Comparing Apples and Oranges: Equating the Power of Case-Control and Quantitative Trait Association Studies. Genetic Epidemiology

Approximate variance explained by a locus

Regression of disease on jth SNP, $x_{[j]}$ = 0,1,2

$$y_{01} = K + b_{01}x_{[i]} + \varepsilon$$

When x[j]=0 $\widehat{y_{01}} = K$ = P(Disease | Genotype =aa)

When x[j]=1 $\widehat{y_{01}} = K + b_{01}$ = P(Disease | Genotype =Aa)

Relative Risk = R= P(Disease | Genotype =Aa)/P(Disease | Genotype =Aa)

$$= (\mathrm{K}+\mathrm{b}_{01})/\mathrm{K} \qquad \text{so} \quad \mathrm{b}_{01} = \mathrm{K}(\mathrm{R}\text{-}1)$$

Variance attributable to the locus on the disease scale

$$\sigma^2_{A_{01[j]}} = h^2_{01[j]}K(1-K) = b^2_{01}var(x) = 2p(1-p)b^2_{01}$$

$$h^2_{01[j]} = 2p(1-p)b^2_{01}/K(1-K)$$

$$h^2_{L[j]} = \frac{(1-K)h^2_{01[i]}}{i^2K} = \frac{2p(1-p)b^2_{01}}{i^2K^2} = \frac{2p(1-p)(R-1)^2}{i^2}$$

Assumes a population sample not a case control sample

See Lecture 1: Dempster & Lerner (1950) Appendix by Alan Robertson. Heritability of threshold characters. Genetics 35

Power of a association study of a quantitative trait

χ^2 *non-centrality parameter* $= NCP_{QT} = \dfrac{N_{QT} h^2_{L[i]}}{1 - h^2_{L[i]}}$

When the variance explained is the same in c-c and for quantitative trait

$NCP_{01} \approx \dfrac{h^2_{L[j]} i^2 v(1-v) N_{01}}{(1-K)^2}$

$$\frac{NCP_{01}}{NCP_{QT}} \approx \frac{i^2 v(1-v) N_{01}}{(1-K)^2 N_{QT}}$$

Yang et al (2009) Comparing Apples and Oranges: Equating the Power of Case-Control and Quantitative Trait Association Studies. Genetic Epidemiology

Practical

a) Code for slides 3-6. Done already

b) Power in case-control study design

i) Compare to GPC

ii) Compare power for screened to unscreened controls

iii) Compare the impact on power of screening controls for # schizophrenia K =0.01 and Major depression K = 0.15

iv) For which disorders is screening of controls most recommended

c) Power Graphs in case-control study design

First run code through to see graphs, then look at code. Makes graph with 3 lines based on RAF for one disease risk

i) You make graph with 3 lines based on disease risk K=0.001,0.01,0.1

Q: For a given GRR is the power bigger or smaller as disease prevalence increases

Q: Why does this make sense?

d) Just run code