Genetic association testing with unrelated individuals

Quantitative and dichotomous/binary (e.g. disease) traits

Heather Cordell
Newcastle University

heather.cordell@ncl.ac.uk

Acknowledgement: thanks to Jonathan Marchini who originally created some of these slides

Testing for Association in a Case-Control Design

Let's suppose we want to test whether a given genetic factor – e.g. a SNP - is involved in (or associated with) a disease.

We can sample *N* unrelated cases (who have the disease) and *N* unrelated controls (who do not have the disease) and genotype the SNP in this sample.

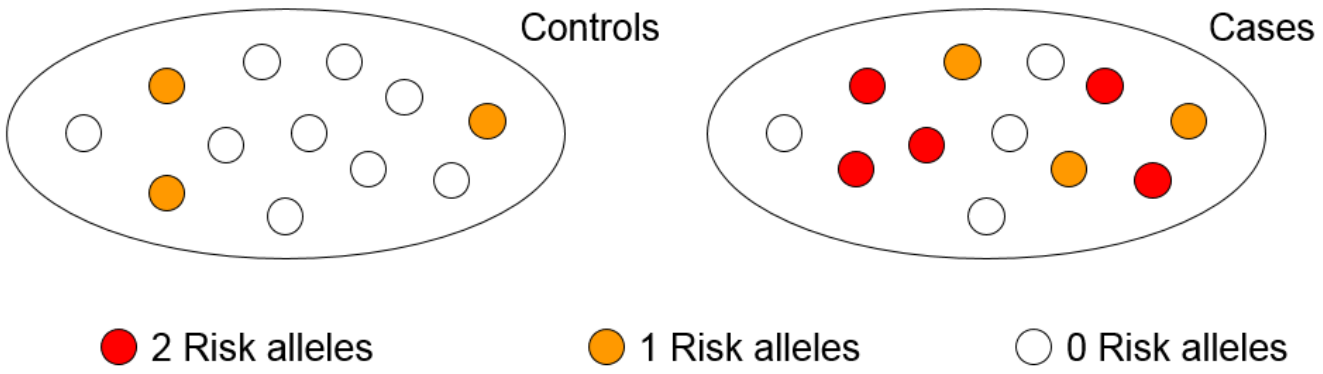
If the SNP is the disease locus, what should we expect to observe in the sample?

Testing for Association in a Case-Control Design

Let's suppose we want to test whether a given genetic factor – e.g. a SNP - is involved in (or associated with) a disease.

We can sample *N* unrelated cases (who have the disease) and *N* unrelated controls (who do not have the disease) and genotype the SNP in this sample.

If the SNP is the disease locus, what should we expect to observe in the sample?



Example

Genotype	Case	Control
aa	100	130
aA	400	390
AA	500	480

In this example, we can see that the distribution of genotypes is different in cases and controls.

Is this difference significant?

This is just a 3x2 contingency table and we could use either a Pearson χ^2 test statistic or a Maximum Likelihood test statistic to test the null hypothesis of no difference in the genotype distribution between cases and controls.

Chi-squared (χ^2) tests

Data

Genotype	Case	Control	
aa	$n_0 = 100$	$m_0 = 130$	$a_0=230$
aA	$n_1 = 400$	$m_1 = 390$	$a_1=790$
AA	$n_2 = 500$	$m_2 = 480$	$a_2=980$
	$N = 1000$	$M = 1000$	$T=2000$

Work out row and column totals (margins), overall total T

Expected value in each cell = (row total × col total)/T

- Corresponds to expected values if there is no difference in the distribution of genotypes between cases and controls

χ^2 test statistic on 2 df = $\sum_{i=1}^{6} \frac{(O_i - E_i)^2}{E_i}$ where O_i and E_i are the observed and expected values in cell i.

Alternative approach: counting alleles

Data

Allele	Case	Control	
a	n_0 = 600	m_0 = 650	a_0=1250
A	n_1 = 1400	m_1 = 1350	a_1=1750
	N = 2000	M = 2000	T=4000

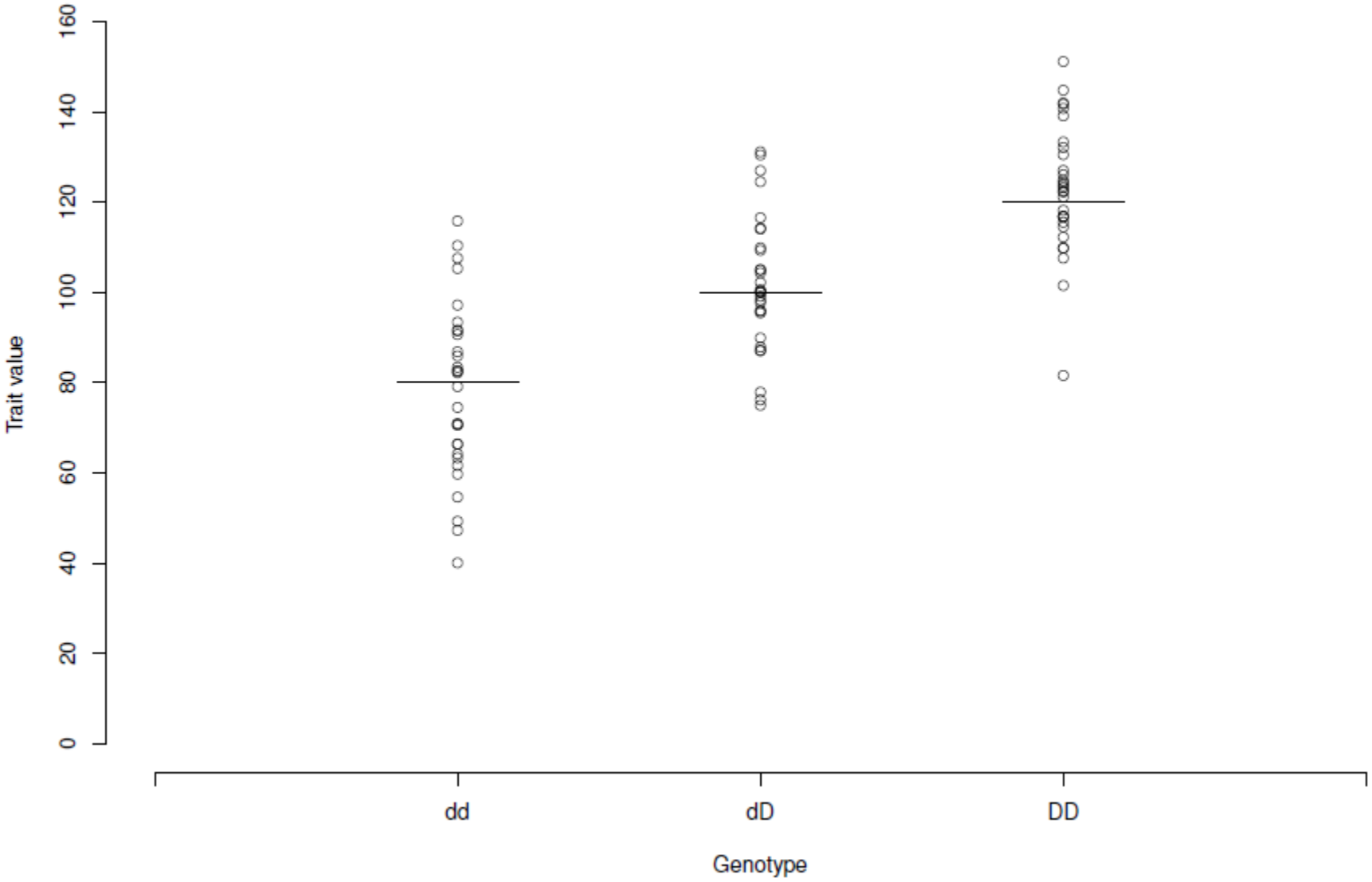
Again work out row and column totals (margins), overall total T
Again expected value in each cell = (row total × col total)/T

χ^2 test statistic on 1 df = $\sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i}$ where O_i and E_i are the observed and expected values in cell i.

- Assumes HWE under null and multiplicative allelic effects under alternative: considers chromosomes as independent units
- Better (equivalent) approach that avoids the HWE assumption: use the counts in the previous genotype table to perform a Cochran-Armitage trend test

Linear and logistic regression

- A more sophisticated approach is to use logistic regression
  - Closely related to linear regression – arguably the most natural way to test for association with a quantitative trait


Trait value
0
20
40
60
80
100
120
140
160
dd
dD
DD
Genotype

Linear regression

- Linear regression fits a "best-fit" line $y=mx+c$
  - Where m is the slope and c the intercept
  - $mx+c$ is the expected value of y (each y_i actually = $mx_i+c+\varepsilon_i$)
  - x is a coded genotype variable taking values (0, 1, 2) for (dd, dD, DD) or (aa, aA, AA)
  - m and c can be estimated via least squares or maximum likelihood
- Tests the null hypothesis that m=0 against the alternative that m ≠ 0

Linear and logistic regression

- $y=mx+c$ could also be written $y= \beta_0 +\beta_1 x$
  - Where β_1 is the slope and β_0 the intercept
- Logistic regression fits a model

$$\log \frac{p}{1-p} = \beta_0 +\beta_1 x \qquad \equiv c + mx$$

  - p represents the probability of being a case rather than a control
  - So we allow the log odds of disease $\log \frac{p}{1-p}$ to vary according to genotype (as opposed to allowing the expected value of y to vary according to genotype)
  - The probability of disease $p = \frac{e^{\beta 0+\beta 1\, x}}{1+e^{\beta 0+\beta 1\, x}}$ therefore also varies according to genotype
- Standard method used in epidemiological studies of non-genetic risk factors (such as smoking in lung cancer)

Details of regression model

- Our coding scheme x=(0,1,2) assumes two copies of allele A has twice the effect of a single copy on log odds scale
  - Corresponds to 'additive' allelic effects on log odds scale, or 'multiplicative' allelic effects on odds scale

Genotype	x	Log odds	Odds	OR
aa	0	β_0	$e^{\beta 0}$	1
aA	1	$\beta_0 + \beta_1$	$e^{\beta 0 + \beta 1}$	$e^{\beta 1}$
AA	2	$\beta_0 + 2\beta_1$	$e^{\beta 0 + 2\beta 1}$	$e^{2\beta 1}=(e^{\beta 1})^2$

- The odds ratio (OR) is the factor by which your odds are multiplied if you have 1 (or 2) copies of the risk allele, compared to none
- If genotype has no effect on the odds of disease, then $\beta_1 = 0$ and all ORs=1
- A test of $\beta_1 = 0$ (e.g. via maximum likelihood) is conceptually equivalent to the 1df χ^2 test for the 2 by 2 table given on page 6.

Genotype model

- A more general "genotype" model allows the odds (or probability) of disease to vary arbitrarily in all 3 categories
  - This is achieved by designating x as a "factor"

Genotype	X (factor)	Log odds	Odds	OR
aa	0	β_0	$e^{\beta 0}$	1
aA	1	$\beta_0 + \beta_1$	$e^{\beta 0 + \beta 1}$	$e^{\beta 1}$
AA	2	$\beta_0 + \beta_2$	$e^{\beta 0 + \beta 2}$	$e^{\beta 2}$

- If genotype has no effect on the odds of disease, then $\beta_1 = \beta_2 = 0$ and again all ORs=1

- A test of $\beta_1 = \beta_2 = 0$ (e.g. via maximum likelihood) is conceptually equivalent to the 2df χ^2 test for the 3 by 2 table given on page 5.

Dominant and recessive models

Genotype	x	Log odds	Odds	OR
aa	0	β_0	$e^{\beta 0}$	1
aA	1	$\beta_0 + \beta_1$	$e^{\beta 0 + \beta 1}$	$e^{\beta 1}$
AA	1	$\beta_0 + \beta_1$	$e^{\beta 0 + \beta 1}$	$e^{\beta 1}$

Genotype	x	Log odds	Odds	OR
aa	0	β_0	$e^{\beta 0}$	1
aA	0	β_0	$e^{\beta 0}$	1
AA	1	$\beta_0 + \beta_1$	$e^{\beta 0 + \beta 1}$	$e^{\beta 1}$

Relative Risks

- If a disease is reasonably rare, the odds ratio approximates the genotype relative risk (GRR, RR)
  - the factor by which your *probability* of disease is multiplied if you have genotype AA (or aA) as opposed to aa

Genotype	Penetrance	GRR	Odds	OR
aa	0.01	1	0.01/0.99 = 0.0101	1.00
aA	0.02	2	0.02/0.98 = 0.0204	2.02
AA	0.05	5	0.05/0.95 = 0.0526	5.21

- If genotype has no effect on the probability of disease, then all GRRs=1

Fitting the logistic regression model

Data

Genotype	Case	Control	
aa	$n_0 = 100$	$m_0 = 130$	$a_0=230$
aA	$n_1 = 400$	$m_1 = 390$	$a_1=790$
AA	$n_2 = 500$	$m_2 = 480$	$a_2=980$
	N = 1000	M = 1000	T=2000

What are the probabilities of being a case (affected with disease) or a control (unaffected with disease) in the three genotype categories?

Genotype Model

Data

Genotype	Case	Control	
aa	$n_0 = 100$	$m_0 = 130$	$a_0=230$
aA	$n_1 = 400$	$m_1 = 390$	$a_1=790$
AA	$n_2 = 500$	$m_2 = 480$	$a_2=980$
	$N = 1000$	$M = 1000$	$T=2000$

Null	Case	Control
aa	p_0	$1-p_0$
aA	p_0	$1-p_0$
AA	p_0	$1-p_0$

Genotype Model

Data

Genotype	Case	Control	
aa	$n_0 = 100$	$m_0 = 130$	$a_0=230$
aA	$n_1 = 400$	$m_1 = 390$	$a_1=790$
AA	$n_2 = 500$	$m_2 = 480$	$a_2=980$
	$N = 1000$	$M = 1000$	$T=2000$

Null	Case	Control
aa	p_0	$1-p_0$
aA	p_0	$1-p_0$
AA	p_0	$1-p_0$

Alternative	Case	Control
aa	p_0	$1-p_0$
aA	q_0	$1-q_0$
AA	r_0	$1-r_0$

Genotype Model

Data

Genotype	Case	Control	
aa	$n_0 = 100$	$m_0 = 130$	$a_0=230$
aA	$n_1 = 400$	$m_1 = 390$	$a_1=790$
AA	$n_2 = 500$	$m_2 = 480$	$a_2=980$
	N = 1000	M = 1000	T=2000

Null	Case	Control
aa	p_0	$1-p_0$
aA	p_0	$1-p_0$
AA	p_0	$1-p_0$

Alternative	Case	Control
aa	p_0	$1-p_0$
aA	q_0	$1-q_0$
AA	r_0	$1-r_0$

Model	Null	Alternative
Likelihood	$L(p) = p_0^{n_0+n_1+n_2}(1-p_0)^{m_0+m_1+m_2}$	$L(p,q,r) = p_0^{n_0}(1-p_0)^{m_0} q_0^{n_1}(1-q_0)^{m_1} r_0^{n_2}(1-r_0)^{m_2}$
Maximized Likelihood	$L_1 = \left(\frac{N}{T}\right)^N \left(\frac{M}{T}\right)^M$	$L_2 = \left(\frac{n_0}{a_0}\right)^{n_0}\left(\frac{m_0}{a_0}\right)^{m_0}\left(\frac{n_1}{a_1}\right)^{n_1}\left(\frac{m_1}{a_1}\right)^{m_1}\left(\frac{n_2}{a_2}\right)^{n_2}\left(\frac{m_2}{a_2}\right)^{m_2}$

Genotype Model

Data

Genotype	Case	Control	
aa	$n_0 = 100$	$m_0 = 130$	$a_0=230$
aA	$n_1 = 400$	$m_1 = 390$	$a_1=790$
AA	$n_2 = 500$	$m_2 = 480$	$a_2=980$
	N = 1000	M = 1000	T=2000

Null	Case	Control
aa	p_0	$1-p_0$
aA	p_0	$1-p_0$
AA	p_0	$1-p_0$

Alternative	Case	Control
aa	p_0	$1-p_0$
aA	q_0	$1-q_0$
AA	r_0	$1-r_0$

Model	Null	Alternative
Likelihood	$L(p) = p_0^{n_0+n_1+n_2}(1-p_0)^{m_0+m_1+m_2}$	$L(p,q,r) = p_0^{n_0}(1-p_0)^{m_0} q_0^{n_1}(1-q_0)^{m_1} r_0^{n_2}(1-r_0)^{m_2}$
Maximized Likelihood	$L_1 = \left(\frac{N}{T}\right)^N \left(\frac{M}{T}\right)^M$	$L_2 = \left(\frac{n_0}{a_0}\right)^{n_0}\left(\frac{m_0}{a_0}\right)^{m_0}\left(\frac{n_1}{a_1}\right)^{n_1}\left(\frac{m_1}{a_1}\right)^{m_1}\left(\frac{n_2}{a_2}\right)^{n_2}\left(\frac{m_2}{a_2}\right)^{m_2}$

$$\text{Maximum Likelihood Test Statistic} = -2\log\left(\frac{L_1}{L_2}\right) = 4.458993 \sim \chi_2^2 \text{ under the Null. p-value} = 0.108$$

Logistic Regression

Logistic regression is a very general (and so useful) model for testing association.

Write Likelihood as $L(\beta) = \prod_{i=1}^{T} {p_i}^{Y_i}(1-p_i)^{Y_i}$

where p_i = P(individual i has disease), and Y_i = 0/1 indicates disease

Model log odds of disease $\log \frac{p_i}{1-p_i}$ as some linear function of genotype

Logistic Regression

Logistic regression is a very general (and so useful) model for testing association.

Write Likelihood as $L(\beta) = \prod_{i=1}^{T} p_i^{Y_i}(1-p_i)^{Y_i}$

where p_i = P(individual i has disease), and Y_i = 0/1 indicates disease

Model log odds of disease $\log \frac{p_i}{1-p_i}$ as some linear function of genotype

Additive model: $\log \frac{p_i}{1-p_i} = \beta_0 + \beta_1 x_i$ $\quad p = \frac{e^{\beta 0+\beta 1\, x}}{1+e^{\beta 0+\beta 1\, x}}$

Genotype model: $\log \frac{p_i}{1-p_i} = \beta_0 + \beta_1 I(x_i=1) + \beta_2 I(x_i=2)$ $\quad p = \frac{e^{\beta 0+\beta 1\, I(x=1)+\beta 2\, I(x=2)}}{1+e^{\beta 0+\beta 1\, I(x=1)+\beta 2\, I(x=2)}}$

Recessive model: $\log \frac{p_i}{1-p_i} = \beta_0 + \beta_1 I(x_i=2)$ $\quad p = \frac{e^{\beta 0+\beta 1\, I(x=2)}}{1+e^{\beta 0+\beta 1\, I(x=2)}}$

where x_i takes values (0, 1, 2) for genotypes (aa, aA, AA)

The Additive Model

The so-called Additive Model is the most widely used model to test for association in genome-wide scans. It assumes the log-odds of disease increase additively with number of alleles.

Confusingly, it is sometimes called the multiplicative model, as the ***odds of disease*** increase multiplicatively with number of alleles.

A way of fitting this model (known as a Score Test) leads to the test statistic known as the **Cochran-Armitage Trend Test**.

Logistic Regression

Logistic regression is a **very general** (and so useful) model for testing association.

The log odds can be modelled by a linear function of any measured variables.

So, for example, if we want to carry out tests of a genetic factor x conditional upon (i.e. allowing for) Age, Sex and Population of Origin of the individuals, we could compare the following 2 models:

$$\text{Null:} \quad \log \frac{p_i}{1-p_i} = \beta_0 + \beta_1 \text{Age}_i + \beta_2 \text{Sex}_i + \beta_3 \text{Pop}_i$$

$$\text{Alternative:} \quad \log \frac{p_i}{1-p_i} = \beta_0 + \beta_1 \text{Age}_i + \beta_2 \text{Sex}_i + \beta_3 \text{Pop}_i + \beta_4 x_i$$

Interactions

- Logistic (or linear) regression also offers a very convenient way to model (statistical) interactions between variables (either GxG or GxE)

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1$$

Marginal effect of factor 1

$$\log \frac{p}{1-p} = \beta_0 + \beta_2 x_2$$

Marginal effect of factor 2

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Main effects of factors 1 and 2

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$$

Main effects and interaction term

- For quantitative traits, use linear regression (i.e. replace log p/(1-p) with y)

Haplotype Association Analysis

- We may want to construct x variables that code for the haplotypes present in an individual (the **combination** of alleles inherited from the same parent)
  - For some diseases, there are genuine haplotype effects
  - Haplotypes may better `mark' an untyped causal variant

Haplotype Association Analysis

- We may want to construct x variables that code for the haplotypes present in an individual (the **combination** of alleles inherited from the same parent)
  - For some diseases, there are genuine haplotype effects
  - Haplotypes may better `mark' an untyped causal variant
- In principle we can fit a regression model for these effects
  - However, haplotypes are not observed: may be inferred probabilistically from unphased genotype data
  - Either prior to, or at same time as, fitting regression model

Haplotype Association Analysis

- We may want to construct x variables that code for the haplotypes present in an individual (the **combination** of alleles inherited from the same parent)
  - For some diseases, there are genuine haplotype effects
  - Haplotypes may better `mark' an untyped causal variant
- In principle we can fit a regression model for these effects
  - However, haplotypes are not observed: may be inferred probabilistically from unphased genotype data
  - Either prior to, or at same time as, fitting regression model
- Various methods/software implementations:
  - Missing data likelihoods/EM algorithm e.g. UNPHASED
  - Score test (Schaid 2004 Genet Epid 27:348-364)
  - Weighted regression e.g. --hap-logistic in PLINK
  - Bayesian partition models e.g. GeneBPM (Morris 2006 AJHG 79:679-694)
- All methods involve (in some sense) averaging over the possible haplotype configurations, while testing association of haplotypes with disease

Population stratification

- A potential problem, particularly with case-control studies, is that of population stratification (= population substructure)
  - Population sampled actually consists of several 'sub-populations' that do not really intermix
- Can lead to spurious false positives (type 1 errors) in case/control studies
  - If cases and controls not well-matched for ancestry
  - And if disease rates and marker allele frequencies are different between ancestral populations
- Problem could be avoided if one knew which population each person was from, and stratified by this (e.g. included 'population' as a variable in logistic regression)

Principal Components Analysis

- Price et al. (2006) Nature Genetics 38:904-909; Patterson et al. (2006); PLoS Genetics 2(12):e190
- Idea is to compute the *eigenvectors* and *eigenvalues* of matrix of correlations between individuals (based on genome-wide SNP data, pruned/thinned for LD)
- Include principal component scores from top 10 (say) eigenvectors as covariates in a logistic regresssion analysis
- Popular approach, but underlying 'sub-population' model may be unrealistic
  - Has arguably been superseded by linear mixed models (Kang et al. (2010) Nat Genet 42:348-354)

Genomic control

- A simpler approach was proposed by Devlin and Roeder (1999) Biometrics 55: 997-1004
- Devlin and Roeder used theoretical arguments to propose that with population structure, the distribution of χ^2 tests is inflated by a constant multiplicative factor λ
- Use a set of "null" loci (not expected to be associated with phenotype) – or indeed all loci from a GWAS – to estimate λ
- Then at any test locus t, we can divide our χ^2 value by λ to get an adjusted χ^2 test

Bayes Factors

Bayes Factors have recently emerged in the GWAS literature as an alternative to p-values.

Balding (2006) Nature Reviews Genetics 7: 781-791

Servin and Stephens (2007) PLoS Genet 3(7): e114

Marchini, Howie et al. (2007) Nature Genetics 39 : 906-913

Stephens and Balding (2009) Nature Reviews Genetics 681-690

For some people ("Bayesians") the use of Bayes Factors is philosophically preferable to the p-values preferred by "Frequentists"

For others (e.g. me!) it is more a matter of convenience

- e.g. various downstream analysis approaches have been developed (especially for fine-mapping) that require Bayes Factors as input

Bayes Factors

Bayes Factors have recently emerged in the GWAS literature as an alternative to p-values.

Balding (2006) Nature Reviews Genetics 7: 781-791

Servin and Stephens (2007) PLoS Genet 3(7): e114

Marchini, Howie et al. (2007) Nature Genetics 39 : 906-913

Stephens and Balding (2009) Nature Reviews Genetics 681-690

For some people ("Bayesians") the use of Bayes Factors is philosophically preferable to the p-values preferred by "Frequentists"

For others (e.g. me!) it is more a matter of convenience

- e.g. various downstream analysis approaches have been developed (especially for fine-mapping) that require Bayes Factors as input

- What are Bayes Factors?
- What is their relationship to p-values?
- Assessing significance in GWAS using Bayes Factors?

What is a p-value?

	0	1	2
Cases	n_0	n_1	n_2
Controls	m_0	m_1	m_2

H_0 : No Association $\quad P(Data \mid \theta_0, M_0)$

H_1 : Association $\quad P(Data \mid \theta_1, M_1)$

What is a p-value?

	0	1	2
Cases	n_0	n_1	n_2
Controls	m_0	m_1	m_2

H_0 : No Association $\quad P(Data \mid \theta_0, M_0)$

H_1 : Association $\quad P(Data \mid \theta_1, M_1)$

$$LR = \frac{\max_{\theta_1} P(Data \mid \theta_1, M_1)}{\max_{\theta_0} P(Data \mid \theta_0, M_0)}$$

What is a p-value?

	0	1	2
Cases	n_0	n_1	n_2
Controls	m_0	m_1	m_2

H_0 : No Association $\quad P(Data \mid \theta_0, M_0)$

H_1 : Association $\quad P(Data \mid \theta_1, M_1)$

$$LR = \frac{\max_{\theta_1} P(Data \mid \theta_1, M_1)}{\max_{\theta_0} P(Data \mid \theta_0, M_0)} \qquad 2\log(LR) \sim \chi_1^2$$

What is a p-value?

	0	1	2
Cases	n_0	n_1	n_2
Controls	m_0	m_1	m_2

$$H_0: \text{ No Association} \qquad P(Data \mid \theta_0, M_0)$$

$$H_1: \text{ Association} \qquad P(Data \mid \theta_1, M_1)$$

$$LR = \frac{\max_{\theta_1} P(Data \mid \theta_1, M_1)}{\max_{\theta_0} P(Data \mid \theta_0, M_0)} \qquad 2\log(LR) \sim \chi_1^2$$

$$p - value = P\left(\chi_1^2 \geq 2\log(LR) \mid H_0\right)$$

What is a Bayes Factor?

	0	1	2
Cases	n_0	n_1	n_2
Controls	m_0	m_1	m_2

H_0 : No Association $\quad P(Data \mid \theta_0, M_0)$

H_1 : Association $\quad P(Data \mid \theta_1, M_1)$

What is a Bayes Factor?

	0	1	2
Cases	n_0	n_1	n_2
Controls	m_0	m_1	m_2

H_0 : No Association $\quad P(Data \mid \theta_0, M_0)$

H_1 : Association $\quad P(Data \mid \theta_1, M_1)$

Bayes Factor Prior Odds

$$\text{Posterior Odds} = \frac{P(M_1 \mid Data)}{P(M_0 \mid Data)} = \frac{P(Data \mid M_1)}{P(Data \mid M_0)} \frac{P(M_1)}{P(M_0)}$$

What is a Bayes Factor?

	0	1	2
Cases	n_0	n_1	n_2
Controls	m_0	m_1	m_2

H_0 : No Association $\quad P(Data \mid \theta_0, M_0)$

H_1 : Association $\quad P(Data \mid \theta_1, M_1)$

Bayes Factor Prior Odds

$$\text{Posterior Odds} = \frac{P(M_1 \mid Data)}{P(M_0 \mid Data)} = \frac{P(Data \mid M_1)}{P(Data \mid M_0)} \frac{P(M_1)}{P(M_0)}$$

Choose Model 1 if Posterior Odds > 1 i.e. if Bayes Factor > 1/ Prior Odds

What is a Bayes Factor?

	0	1	2
Cases	n_0	n_1	n_2
Controls	m_0	m_1	m_2

H_0 : No Association $\quad P(Data \mid \theta_0, M_0)$

H_1 : Association $\quad P(Data \mid \theta_1, M_1)$

Bayes Factor Prior Odds

$$\text{Posterior Odds} = \frac{P(M_1 \mid Data)}{P(M_0 \mid Data)} = \frac{P(Data \mid M_1)}{P(Data \mid M_0)} \frac{P(M_1)}{P(M_0)}$$

Choose Model 1 if Posterior Odds > 1 i.e. if Bayes Factor > 1/ Prior Odds

$$\text{Bayes Factor} = \frac{\int P(Data \mid \theta_1, M_1) P(\theta_1 \mid M_1) d\theta_1}{\int P(Data \mid \theta_0, M_0) P(\theta_0 \mid M_0) d\theta_0} \neq \text{LR} = \frac{\max_{\theta_1} P(Data \mid \theta_1, M_1)}{\max_{\theta_0} P(Data \mid \theta_0, M_0)}$$

What is a Bayes Factor?

	0	1	2
Cases	n_0	n_1	n_2
Controls	m_0	m_1	m_2

H_0 : No Association $\quad P(Data \mid \theta_0, M_0)$

H_1 : Association $\quad P(Data \mid \theta_1, M_1)$

Bayes Factor Prior Odds

$$\text{Posterior Odds} = \frac{P(M_1 \mid Data)}{P(M_0 \mid Data)} = \frac{P(Data \mid M_1)}{P(Data \mid M_0)} \frac{P(M_1)}{P(M_0)}$$

Choose Model 1 if Posterior Odds > 1 i.e. if Bayes Factor > 1/ Prior Odds

$$\text{Bayes Factor} = \frac{\int P(Data \mid \theta_1, M_1) P(\theta_1 \mid M_1) d\theta_1}{\int P(Data \mid \theta_0, M_0) P(\theta_0 \mid M_0) d\theta_0} \neq \text{LR} = \frac{\max_{\theta_1} P(Data \mid \theta_1, M_1)}{\max_{\theta_0} P(Data \mid \theta_0, M_0)}$$

The SNPTEST program uses a Laplace approximation to evaluate these integrals

Priors

What priors do people typically use? (E.g. for the WTCCC analysis)

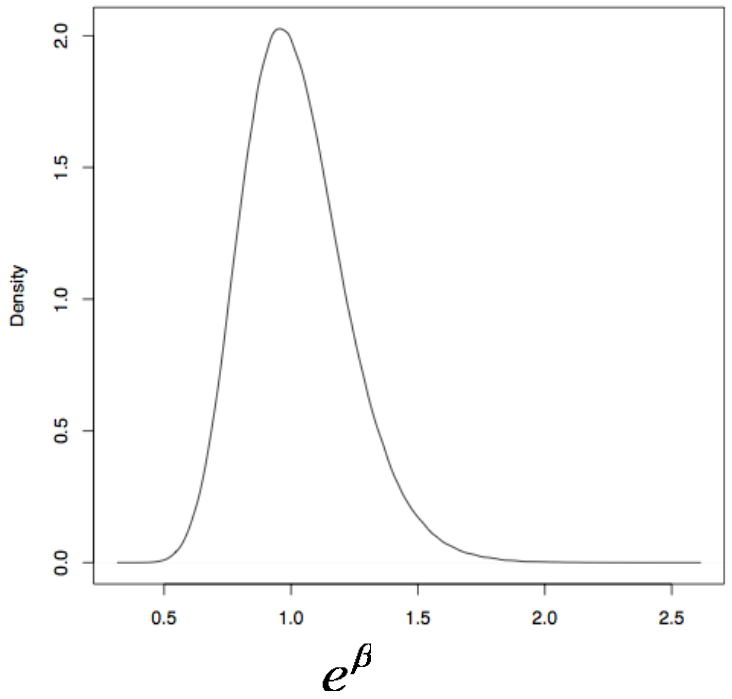
The main parameter θ of interest is the additive effect parameter β

The WTCCC used a Normal prior for this parameter

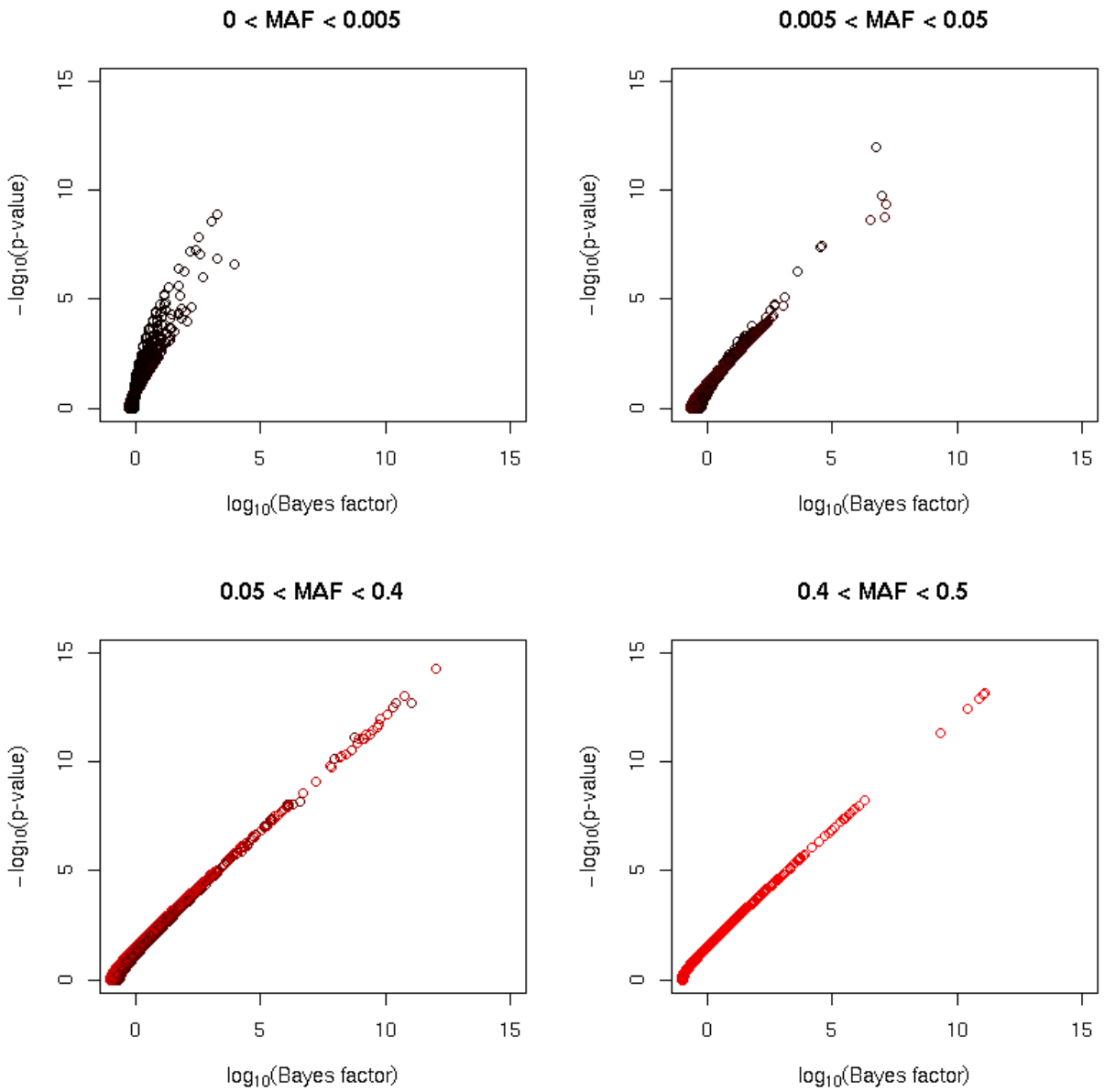
$$\beta \sim \mathrm{N}\left(0, \sigma^2\right) \quad \sigma = 0.2$$

This results in the following prior for the odds ratio:

$$e^{\beta} = \text{Odds Ratio} \approx \text{Relative Risk}$$



p-values vs Bayes Factors



Significance Thresholds

In a frequentist analysis, a threshold is set on the size of the p-value to control the Type I Error that any association will be found across the genome (i.e. to control the multiple testing problem).

For GWAS (where most analyses still use p-values) this is typically in the region 5.0E-08.

For Bayes Factors, it has been argued that we similarly need a Bayes Factor > 10,000 (or $\log_{10}$ Bayes Factor > 4)

Based on the following argument:

Choose Model 1 if Posterior Odds > 1 $\Rightarrow$ Bayes Factor > 1/ Prior Odds

If we think there are about 10,000,000 SNPs in the genome, and if for a given disease about 1,000 of them might be (or be in LD with) disease variants, then

$$\text{Prior Odds} = \frac{P(M_1)}{P(M_0)} = \frac{1}{10{,}000}$$

Polygenic risk scores

- Recall our (additive) logistic regression model:

$$\log \frac{p_i}{1-p_i} = \beta_0 + \beta_1 x_1$$

- We mentioned that an advantage of logistic regression is that we can include multiple predictors in the model

$$\log \frac{p_i}{1-p_i} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5$$

- In a GWAS we may well have identified 5 (or more) independent signals marked by specific ”top SNPs”
  - So we can use the estimated *β’s* for these SNPs to estimate the log odds of disease for a new individual, given their genotypes at these SNPS
    - And thus their probability of disease, if β_0 is known (or can be estimated from the population prevalence)

Polygenic risk scores

- In practice, the expectation is that most if not all complex traits are **polygenic**
  - Influenced by thousands of genetic variants, each having a small effect

- A new person's PRS is

$$x_{PRS} = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \ldots$$

  - Can be used to predict their disease risk
  - But Dudbridge (2018) cautions that **large sample sizes** are required to attain accurate prediction from polygenic models
    - Of the order of 100,000 subjects
    - And the heritability of a trait imposes an upper limit to the prediction performance that can be achieved
    - See Dudbridge (2018) Genet Epid 40(4):268-72

Choice of predictors

- Which variants should be included in the polygenic risk score?
  - Selecting into the risk score only variants that are associated at genome-wide significance, generally explains little heritability
  - And therefore (even based on the combination of variants) shows little predictive accuracy

Choice of predictors

- Which variants should be included in the polygenic risk score?
  - Selecting into the risk score only variants that are associated at genome-wide significance, generally explains little heritability
  - And therefore (even based on the combination of variants) shows little predictive accuracy

- Dudbridge (2013) suggests that prediction is optimized by selecting variants with P-values as high as 0.001

Choice of predictors

- Which variants should be included in the polygenic risk score?
  - Selecting into the risk score only variants that are associated at genome-wide significance, generally explains little heritability
  - And therefore (even based on the combination of variants) shows little predictive accuracy

- Dudbridge (2013) suggests that prediction is optimized by selecting variants with P-values as high as 0.001

- The PRSice software (Euesden et al. 2015, Bioinformatics 31:14661468) computes polygenic risk scores with contributing SNPs selected at different P-value thresholds
  - The final P-value threshold chosen is that which provides the best prediction performance, as assessed via internal cross validation.

Polygenic risk scores

- By definition, if you calculate the PRS in a sample of new people, and test for association of this single x_{PRS} variable with disease, you would expect to find significant association with disease
  - Assuming the original findings were "real"

- An arguably more interesting application is to test for association with a ***different*** disease
  - To uncover genetic connections between the two diseases
  - Purcell et al. (2009) Nature 460(7256):748-752 used this approach to show shared basis for schizophrenia and bipolar disorder