Statistical models for polygenic traits

Daniel Crouch
daniel.crouch@well.ox.ac.uk
Wellcome Centre for Human Genetics,
University of Oxford

Introduction

- These sessions build on what you've learned about genome-wide association studies

- Most GWAS implicate large numbers of SNPs in both common diseases (e.g. type II diabetes, bipolar disorder) and normally distributed phenotypes (e.g. height, BMI and IQ)

- What can we do with SNP data to analyse the genetics of these *complex* phenotypes?

- We will implement a specific model for the relationship between SNPs and phenotype, and see what inferences we can use it to make.

Some issues with straightforward GWAS

1) Performing 1000s of separate simple analyses and correcting P-values for multiple tests afterwards

- Can control false-positives, but what about estimating effect sizes? ("Winner's curse")

2) Often, many SNPs will have true associations but with very small effect sizes

- Significant associations of specific SNPs become less important (versus, e.g. cumulative genetic effects in a biological category)
- Lack power to obtain significant associations with these specific variants of very small effect

Odds ratios of genome-wide significant GWAS 'hits' in 7 diseases originally studied by the WTCCC
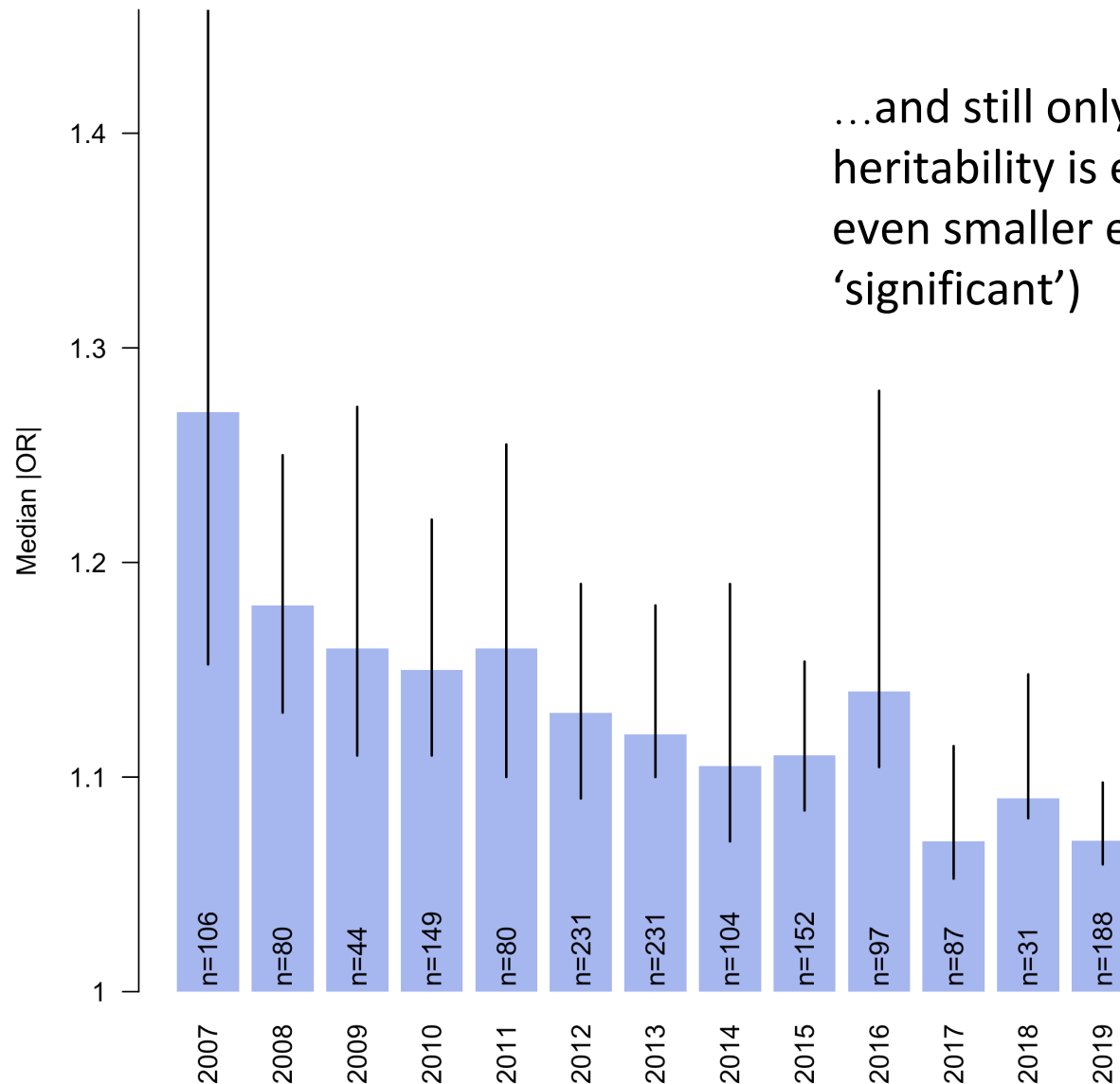
…and still only a fraction of disease heritability is explained (must be many even smaller effects we can't detect as 'significant')

Median |OR|

Crouch and Bodmer (*PNAS* 2020)

What's the model?

- In science we're essentially trying to model things.

- In statistics we're usually calculating evidence for which 'fit' of a model (i.e. which parameter values) is most likely correct

- The standard GWAS model is simple and makes very few assumptions. This has pros and cons

Example: estimating SNP effects

- Suppose we want the best estimate of SNP j's effect size, given the simple estimate we calculated in a GWAS, z_j. Some form of posterior expectation is really what we want:

$$E[\theta_j \mid z_j].$$

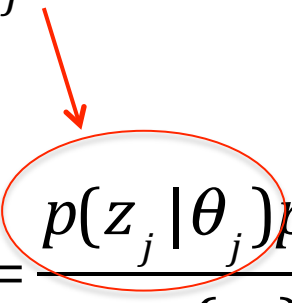- How to work this out? Use Bayes' Theorem and take expectation of resulting posterior distribution.

$$\text{posterior} = p(\theta_j \mid z_j) = \frac{p(z_j \mid \theta_j) p(\theta_j)}{p(z_j)}$$

Example: estimating SNP effects

- Usually the best model for the effect size, assuming it has been standardised (i.e. variance 1), is:

$$z_j \sim \text{Normal}(\theta_j, 1)$$

…which we use for $p(z_j | \theta_j)$, but what about the other two probabilities?

$$\text{posterior} = p(\theta_j | z_j) = \frac{p(z_j | \theta_j)p(\theta_j)}{p(z_j)}$$

Example: estimating SNP effects

- What's a sensible model? First, we could assume that all SNPs j have effects drawn from the same prior distribution, e.g:

$$\theta_j \sim \text{Normal}(\text{mean} = M, \text{var} = A)$$

  ….which gives us $p(\theta_j)$, but we can only use it if we know M and A.

- The **marginal** we can get from the other two probabilities by integration:

$$p(z_j) = \int_{-\infty}^{\infty} p(z_j \mid \theta_j) p(\theta_j) d\theta_j$$

Example: estimating SNP effects

$$\theta_j \sim \mathrm{Normal}(M, A)$$

- The fully **Bayesian** approach is to use your prior knowledge about M and A to make an informed guess.

- With the **empirical Bayes** approach, we estimate M and A from across the full range of SNPs we have, like a **frequentist**, and plug the estimates in, which gives approximately

$$E[\theta_j \,|\, z_j] \approx \hat{M} + \frac{\hat{A}}{1 + \hat{A}}(z_j - \hat{M}).$$

See Efron and Hastie *Computational Age Statistical Inference* Chapter 7

Example: estimating SNP effects

$$E[\theta_j \,|\, z_j] \approx \hat{M} + \frac{\hat{A}}{1+\hat{A}}(z_j - \hat{M}).$$

- As $\hat{A}$ goes to infinity we end up with

$$E[\theta_j \,|\, z_j] = z_j.$$

- Therefore, using the simple estimate we got from our GWAS is equivalent to assuming that the prior variance *A* for the SNP effects is infinite (i.e. no prior knowledge used **- frequentism**)

- This might be justified (e.g. for 'stand out' SNPs with unusually large effects), but the better estimate is usually the first one.

What's the model?

- We have shown how the choice of prior parameters (estimated or infinite variance) affects our conclusions about the effect of a SNP

- This is a general point. There are any number of distributions and parameters one might choose, both for the **prior** $p(\theta_j)$ and the **likelihood** $p(z_j|\theta_j)$

Practical: implement the estimator

- Go to
  https://github.com/whg-gms/statistics-course/tree/main/
  7_Statistical_models_for_polygenic_traits/practicals for dataset (zStats and MAF)

- Read in zStats data using read.table(…)

- Plot the GWAS estimates. Is the data normal? Use 'hist()'.

- Fit estimator
  - Estimate A (crudely) as sd(x)^2-1
  - Estimate M as mean(x).

- Compare with the basic GWAS estimates. Use plot(zStats,est).

- What do you think are the pros and cons of each estimate (without knowing the true underlying parameter values) ?

- Convert Zs into log ORs by multiplying by the SE. Same pattern or different?

Other types of prior

- Double exponential (laplacian)

- Spike and slab

- Mixed normals (see e.g. BOLT-LMM)

- 'Non parametric' approaches

Part 2: Improving the model

We will now look at how a similar model can be used to…

a) Work out how much trait variation is explained by the SNP data

b) Do better primary GWAS analaysis (tomorrow)

c) Predict phenotype from genotype (tomorrow)

….in the presence of **genetic relatedness.** For this we will model **individual level data** as opposed to **summary statistics.**

Gene/environment model

$$y_i = \sum_{j=1}^{N} W_{ij} u_j + \sum_{k=1}^{K} X_{ik} \beta_k + \epsilon_i$$

y_i = phenotype for individual i

W_{ij} = allele at SNP j for individual i

u_j = effect of SNP j

X_{ik} = covariate k for individual i

β_k = effect of covariate k

ϵ_i = residual

(Yang et al. American Journal of Human Genetics, 2011)

Gene/environment model

$$y_i = \sum_{j=1}^{N} W_{ij} u_j + \sum_{k=1}^{K} X_{ik} \beta_k + \epsilon_i$$

Can write in matrix form as:

$$\mathbf{y} = \mathbf{W}\mathbf{u} + \mathbf{X}\beta + \epsilon$$

(Yang et al. American Journal of Human Genetics, 2011)

Gene/environment model

$$y_i = \sum_{j=1}^{N} W_{ij} u_j + \sum_{k=1}^{K} X_{ik} \beta_k + \epsilon_i$$

There is more to decide about the model beyond this equation, e.g:

    -Which variables are random are which are known constants?

    -What are the probability distributions of the random variables?

Any suggestions?

Gene/environment model
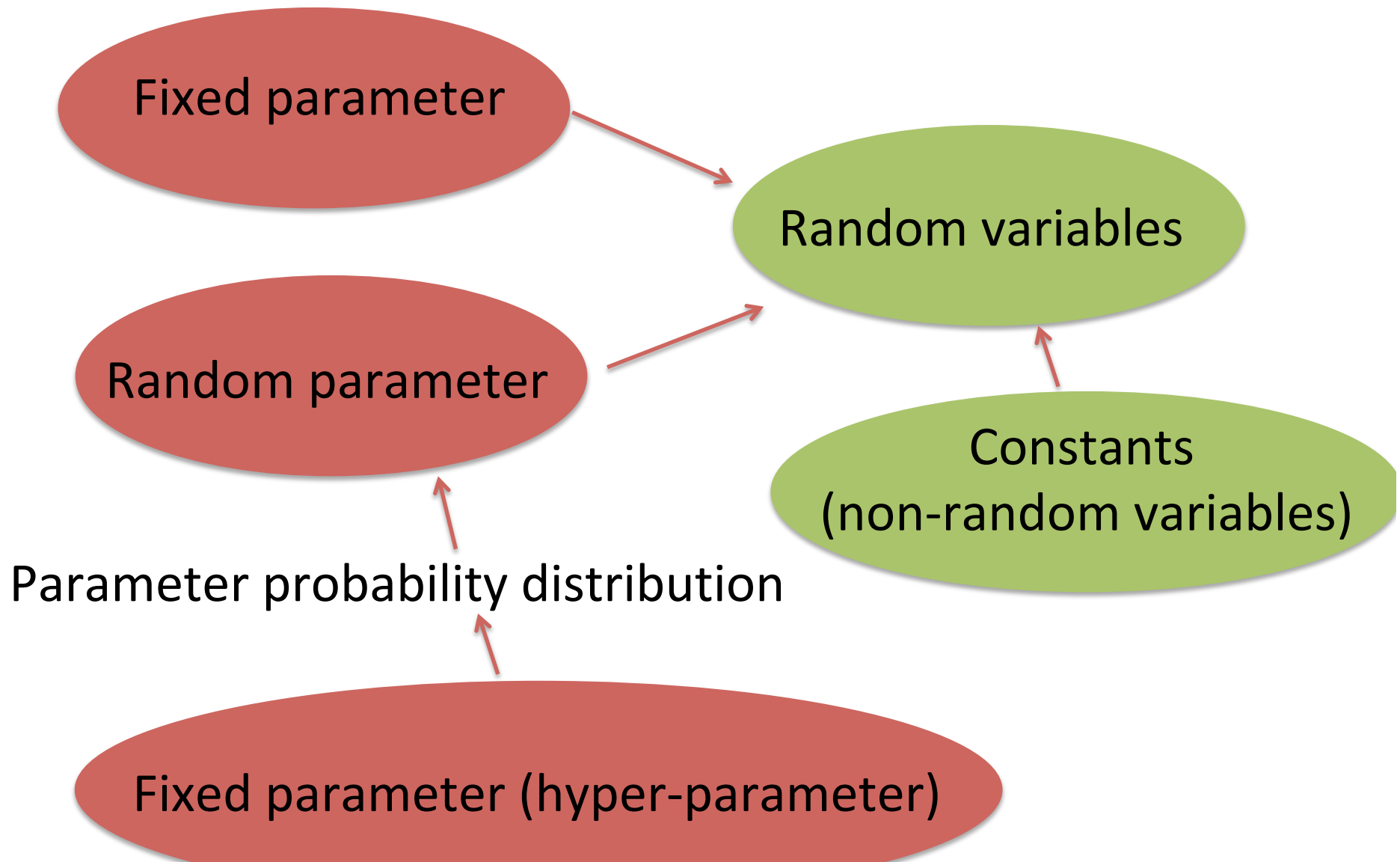
$$y_i = \sum_{j=1}^{N} W_{ij} u_j + \sum_{k=1}^{K} X_{ik} \beta_k + \epsilon_i$$

- We'll treat X and W as known **constants** – transfer the randomness onto other variables like in a linear regression

- y_i is also observed in the data, but we'll assume it is related to everything in the summations by the **random variable** ϵ_i

- Which other variables are **random** (assume an underlying distribution) and which are **fixed** (assume a fixed underlying true value)?

'Fixed' and 'random'

- **Parameters** can be random or fixed
  - Parameters are things like 'true effect of an allele', or 'true average height difference between men and women': things that control the probability of data and that we often want to infer. Can be fixed or random

- **Constants** are non-random variables. Data we do not apply probability distributions to

- **Random variable** typically refers to data that are assumed to be generated by an underlying distribution

'Fixed' and 'random':
Example depiction of a probability model

Fixed parameter

Random variables

Random parameter

Constants
(non-random variables)

Parameter probability distribution

Fixed parameter (hyper-parameter)

$$y_i = \sum_{j=1}^{N} W_{ij} u_j + \sum_{k=1}^{K} X_{ik} \beta_k + \epsilon_i$$

- It is standard to assume that the environmental component, ϵ is normally distributed

- X is constant and β is a vector of fixed effects: they do not vary so they have no variance.

- Assume each u_j is random, with same prior normal distribution as in our previous model, then the variance of y is

$$\text{var}(y_i) = \sigma_u^2 \sum_{j=1}^{N} W_{ij}^2 + \sigma_\epsilon^2$$

….and its expected value (mean) is

$$E(y_i) = \sum_{k=1}^{K} X_{ik} \beta_k$$

Nice properties of this model

$$y \sim \text{Multivariate-normal}\left(\mathbf{X}\beta, \sigma_u^2 \mathbf{WW'} + \mathbf{I}\sigma_\epsilon^2\right)$$

- Not too many parameters to fit ($\beta, \sigma_u^2, \sigma_\epsilon^2$). Don't have to fit an effect for each SNP, which is hard.

- σ_u^2 tells us directly the variance in *y* that is explained by the genetic data. The 'Chip heritability' is then:

$$H_{chip}^2 = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_\epsilon^2}$$

…which we could for example use to find how much heritability is explained via a certain portion of genes.

Nice properties of this model

$$\mathbf{y} \sim \text{Multivariate-normal}\left(\mathbf{X}\beta, \sigma_u^2 \mathbf{W}\mathbf{W}' + \mathbf{I}\sigma_\epsilon^2\right)$$

-We're accounting for relatedness

$$\text{cov}[y_i, y_l] = \sigma_u^2 \sum_{j=1}^{N} W_{ij} W_{lj}$$

WW' is known as the genetic relatedness matrix
Measures genetic covariance

What is this estimate of heritability?

$$\mathrm{var}[y_i] = \sigma_u^2 + \sigma_\epsilon^2$$

- The genetic variance we can explain using the data we have genotyped (and imputed if applicable)

- This tends to much lower than the estimate of heritability we can from classical (non-molecular) genetics, usually **twin studies** – can you think why?

$$\sigma_A^2 = 2(r_{MZ} - r_{DZ})$$

- **This is changing** (Wainschtain et al. 2019, bioRxiv, *Recovery of trait heritability from whole genome sequence data*)

Random or Fixed?
Bayesian or Frequentist?

- The real parameters of interest in this use of the model are

$$\sigma^2_u \text{ and } \sigma^2_\epsilon \text{ , and to a lesser extent } \beta$$

- I've stated that we'll assume β is fixed: is this best? Why?

- What about σ^2_u and σ^2_ϵ ?

Random or Fixed?
Bayesian or Frequentist?

- I've stated that we'll assume β is fixed – is this best? Why?

- What about σ^2_u and σ^2_ϵ ?

Answer: there is no right answer, but the one you chose depends how you wish to utilise prior information

- We've seen how an empirical Bayesian analysis can work for random SNP effects (compute the posterior expectation)

- For fixed parameters we use **frequentist** estimation, usually maximum likelihood

Fit the model in R using maximum likelihood

- Read in data using read.table($...$)

- Scale and center each variable, use scale($...$)

- Calculate genetic relatedness matrix:
  GRM=genoScaled%*%t(genoScaled)/N

- Write a likelihood function (i.e. the multivariate normal shown before)
  NegLogLikelihood<-function(x){$...$}

- Maximise log-liklihood function using constrOptim

Part 3:
Improving GWAS association analysis

- Two parts:

  a) Fit association model in presence of relatedness by using yesterday's model (**mixed-model** association analysis)

  b) Is there a better way than Bonferroni correction to establish significance? Yes: the **False Discovery Rate**

Back to our linear model

$$y_i = \sum_{j=1}^{N} W_{ij} u_j + \sum_{k=1}^{K} X_{ik} \beta_k + \epsilon_i$$

- Second summation is over the **fixed effects**, in contrast with the **random effects** in the first

- Fixed effects are good for estimating specific effects of interest

- We already know how to fit this model – but how to get a P-value out of it?

Stratification/relatedness

- Standard one-variable at a time GWAS uses a combination of PC covariates and removing individuals with relatedness below about 10% or 20%

- The mixed model approach is to picture a bivariate normal distribution for the phenotypes of two individuals with correlated genotypes, but 'expand' to dimensions equal to the sample size:
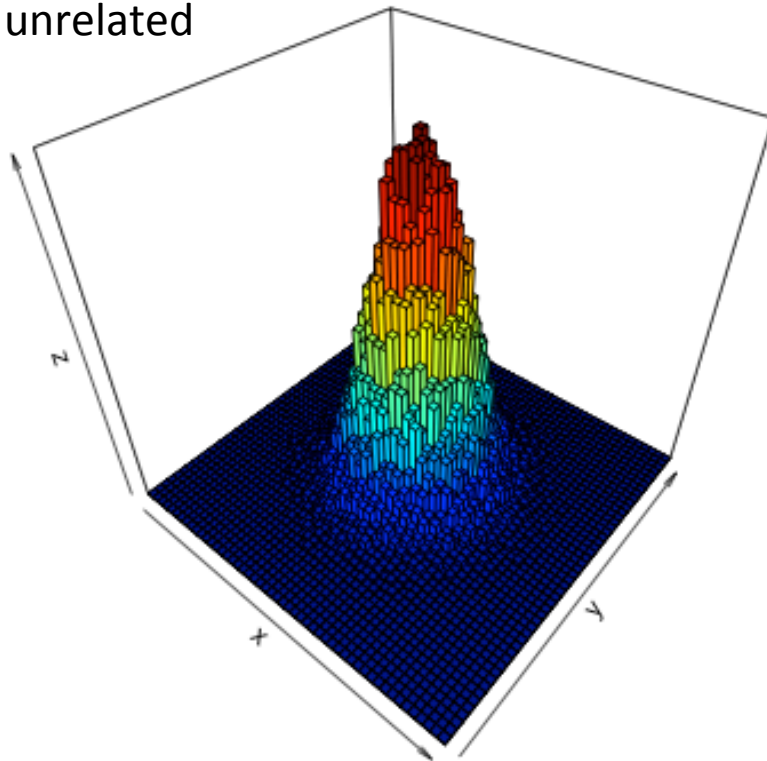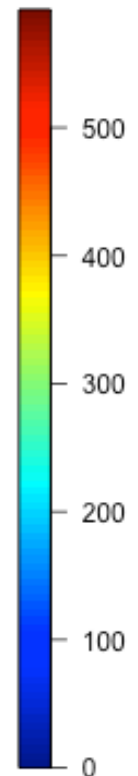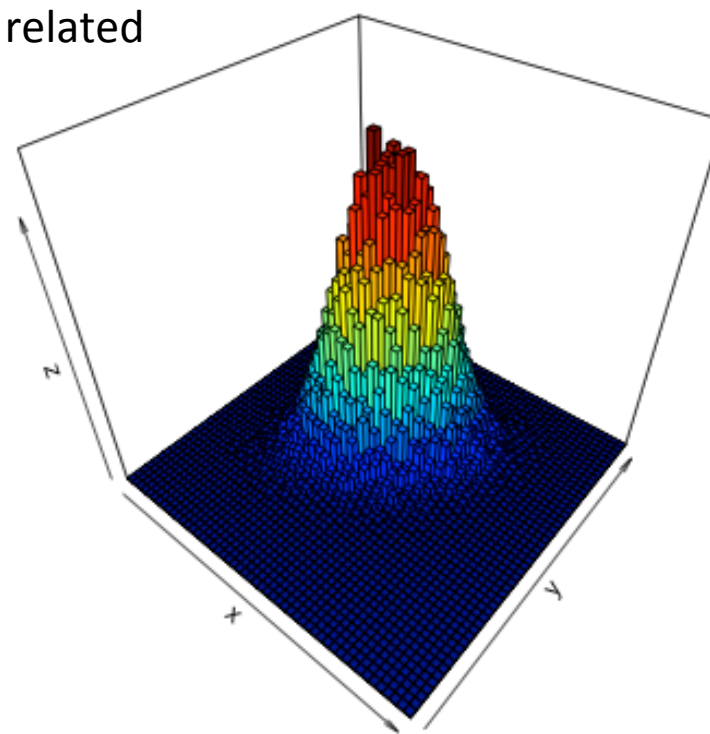
x=ind1 phenoype, y=ind2 phenotype, z=density

Likelihood ratio test

- There are three common types of significance test
  - 1) Wald test
  - 2) Score test
  - 3) Likelihood ratio test

Use likelihood function

- Likelihood ratio tests are generally the best.
  - Likelihood is maximised and compared to the maximised likelihood from a **null model**.
  - Usually the null model is one in which the parameter of interest is forced to be zero

$$\chi_k^2 \approx 2\ln \frac{\text{Maximised alternative Likelihood}}{\text{Maximised null likelihood}}$$

k = degrees of freedom equal to difference in number of parameters between models

Practical: estimate and test fixed effects for each SNP in turn

- Use functions from yesterday

- Fit null and alternative models

- Use apply() to loop over SNPs

- Implement likelihood ratio tests to get P-values. Use pchisq()

- This will be slow. Can you think of a trick to speed things up?

- Can you compare estimates for some SNPs with the GWAS Z stats?

- Do you think we're OK calculating the GRM using all SNPs?

Mixed model simplifications

- Estimate variance components σ_u^2 and σ_ϵ^2 first, then fix during association testing

  e.g. EMMAX (2010)

- Find analytical maxima for some parameters, then maximise the likelihood for the others

  e.g. GEMMA (2012) – analytical maxima for all parameters except:

  $$\lambda = \frac{\sigma_u^2}{\sigma_\epsilon^2}$$

Aside: significance testing in 'big data' scenarios

- Need to correct for large number of tests in order to control false positives

- We can usually do better than bonferroni correction using **false discovery rates**

- These have an empirical Bayes' interpretation. (Chapter 15, Efron and Hastie 2017)

Empirical Bayes False-discovery rate
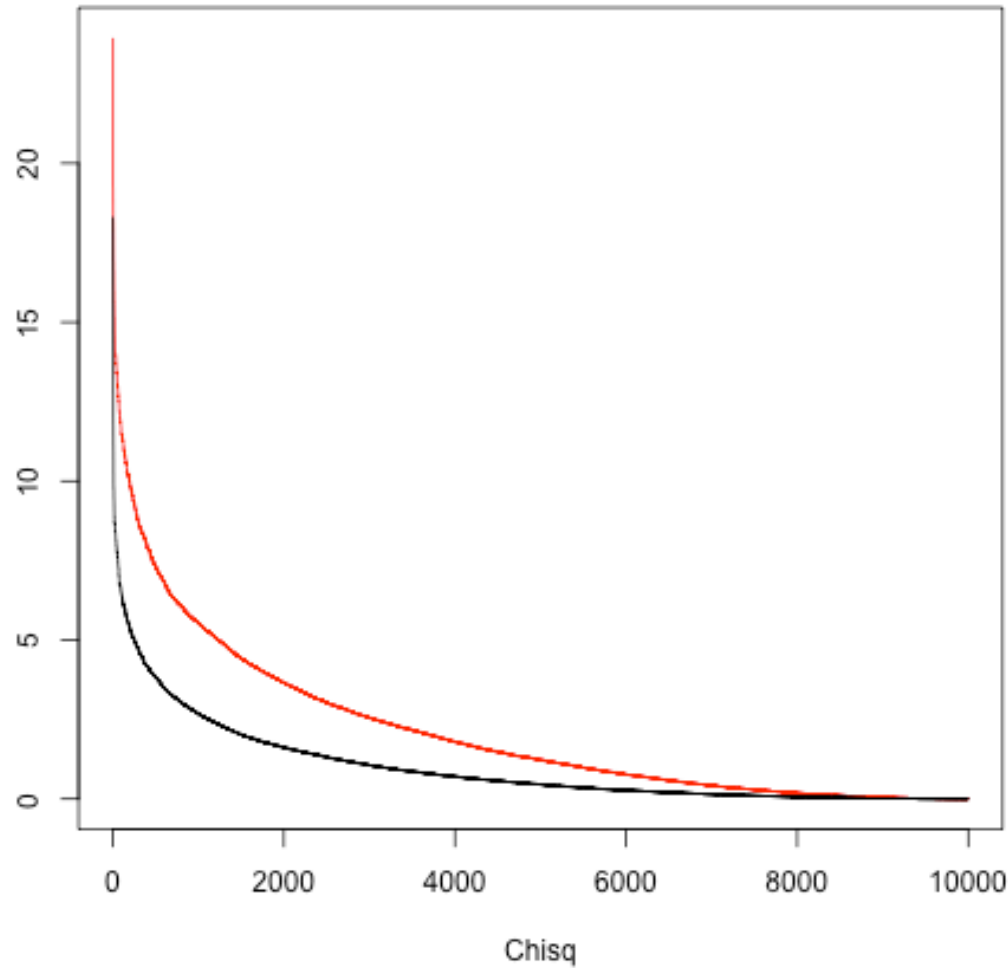
$$\text{fdr} = p(\text{null true} | \chi^2)$$

$$= \frac{p(\chi^2 | \text{null true}) p(\text{null true})}{p(\chi^2)}$$

- Plug in observed distribution for the marginal
- Estimate or guess prior

Part 4: Predict phenotype

- Why?
  - Preventative medicine for at-risk patients
  - Lifestyle interventions
  - Livestock and crop selection
  - Find biological links between diseases

- Once more, we'll apply our 'normal prior' model.

Genomic best-unbiased linear predictor (BLUP)

- Similar to before, the best predictor in a general sense is the posterior expectation

$$E[g_i \mid y_{j \neq i}], \text{ where } g_i = \sum_{j=1}^{J} W_{ij} u_j$$

- g_i is often called the 'breeding value', 'additive genetic value' or 'genetic value.

- In our multivariate normal model, this is

$$E[\mathbf{g} \mid \mathbf{y}] = \sigma_u^2 \mathbf{W}\mathbf{W}'[\mathbf{W}\mathbf{W}'\sigma_u^2 + \mathbf{I}\sigma_\epsilon^2]^{-1}\mathbf{y}$$

…and we'll plug in our parameter estimates accordingly, as with our posterior effect size estimates

de los Campos et al., 2013
Prediction of complex human traits using the
genomic best linear unbiased predictor

Genomic best-unbiased linear predictor (BLUP)

Often when we want to predict, we have:
- Genotypes and phenotypes for a **training set**
- Only genotypes for a **test set**

Then, the expectation is:

$$E[\mathbf{g}_i \mid \mathbf{y}_{-i}] = \sigma_u^2 \mathbf{W}_{test} \mathbf{W'}_{train} [\mathbf{W}_{train} \mathbf{W'}_{train} \sigma_u^2 + \mathbf{I}\sigma_\epsilon^2]^{-1} \mathbf{y}_{train}$$

de los Campos et al., 2013
Prediction of complex human traits
using the genomic best linear unbiased
predictor

Aside: polygenic scores (PGS)

- Sometimes called polygenic risk scores (PRS)

- Approach to fitting each SNPs coefficient explicitly – some similarities with machine learning

- Fits model predicting phenotype from genotypes, using additional info from GWAS analysis, then applies to a test set to pick best **tuning parameters**

Practical: phenotype prediction

- Try to predict phenotype from genotype using all our individuals, in the following ways:
  - Using G-BLUP
  - Using posterior effect sizes from yesterday
  - Using maximum likelihood (basic GWAS) estimates (you'll have to be creative!)

- Compare the estimates. Which do you think is best and why?