SISG
2014

SISG Module 4:
Bayesian Statistics for Genetics

19th Summer Institute in Statistical Genetics

UNIVERSITY *of* WASHINGTON

(This page left intentionally blank.)

2014 SISG MODULE 4: Bayesian Statistics for Genetics
Introduction and Overview

Peter Hoff and Jon Wakefield

Departments of Statistics and Biostatistics
University of Washington

Logistics

- Course Text: P.D. Hoff (2009), *A First Course in Bayesian Statistical Methods*, Springer.
- Supplementary Text: J.C. Wakefield (2013), *Bayesian and Frequentist Regression Methods*, Springer.
- Both texts can be downloaded (for free!) from UW libraries.
- Demonstrations of methods via R implementations will be carried out in class. Students are encouraged to follow along.
- Code and other materials (course notes, papers) are available at the course website:
  http://www.stat.washington.edu/hoff/SISG/
- There will be a lab session at the end of day 2, in which students will have the opportunity to go over some exercises and/or reproduce examples from the class notes.

Course Outline

1. Lecture 1 (Hoff): Why Bayes?
2. Lecture 2 (Hoff): Review of Probability.
3. Lecture 3 (Wakefield) Binomial Sampling.
4. Lecture 4 (Wakefield) Multinomial Sampling.
5. Lecture 5 (Hoff) Linear Regression.
6. Lecture 6 (Wakefield) Bayesian and Frequentist Multiple Testing.
7. Lecture 7 (Hoff) Model Selection and Averaging.
8. Lecture 8 (Wakefield) Generalized Linear Modeling.
9. Lecture 9 (Hoff) Prior Specification including Empirical Bayes.
10. Lecture 10 (Wakefield) Imputation, Hierarchical Mixture Models, Model Comparison.

Module 4: Bayesian Methods
Lecture 1: Why Bayes?

Peter Hoff

Departments of Statistics and Biostatistics
University of Washington

Outline

Bayesian learning

Probability of a rare event

Predictive models

Probability and information

We often use "probability" informally to express belief or information.

This use can be made mathematically formal via *Bayesian theory*:

- Probability can numerically quantify rational beliefs
- There is a relationship between probability and information
- Bayes' rule is a rational method for updating information

Inductive learning via Bayes' rule is referred to as *Bayesian inference*.

Bayesian methods

Bayesian methods are data analysis tools that are derived from the principles of Bayesian inference.

Bayesian methods provide:

- parameter estimates with good statistical properties;
- parsimonious descriptions of observed data;
- predictions for missing data and forecasts of future data;
- a computational framework for model estimation, selection and validation.

Statistical induction

Induction: Reasoning from specific cases to a general principle.

Statistical induction: Using a data sample to infer population characteristics.

Notation:

Parameter: θ quantifies unknown population characteristics

Data: y quantifies the outcome of a survey or experiment

Our goal is to make inference about θ given y.

Ingredients of a Bayesian analysis

Parameter and sample spaces:

sample space: $\mathcal{Y}$ is the set of all possible datasets

parameter space: Θ is the set of all possible θ-values

Quantifying information:

prior distribution: $p(\theta)$, defined for all $\theta \in \Theta$, describes our belief that θ is the true value of the population parameter.

sampling model: $p(y|\theta)$, defined for $\theta \in \Theta, y \in \mathcal{Y}$, describes our belief that y will be the experimental outcome, for each θ.

Updating information:

Bayes' rule: After obtaining data y, the posterior distribution is calculated

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int_\Theta p(y|\tilde{\theta})p(\tilde{\theta})\ d\tilde{\theta}}\,.$$

Role of prior information

$$
\begin{aligned}
p(\theta|y) &= \frac{p(y|\theta)p(\theta)}{\int_\Theta p(y|\tilde{\theta})p(\tilde{\theta})\, d\tilde{\theta}} \\
\frac{p(\theta_a|y)}{p(\theta_b|y)} &= \frac{p(y|\theta_a)}{p(y|\theta_b)}\frac{p(\theta_a)}{p(\theta_b)}.
\end{aligned}
$$

Bayes' rule does not tell us what our beliefs should be.

It tells us how they should change after seeing new information.

Probability of a rare event

Suppose we are interested in the prevalence of a rare genetic mutation in a human subpopulation.

A random sample of 20 individuals from the population are sampled and genotyped.

Parameter and sample space

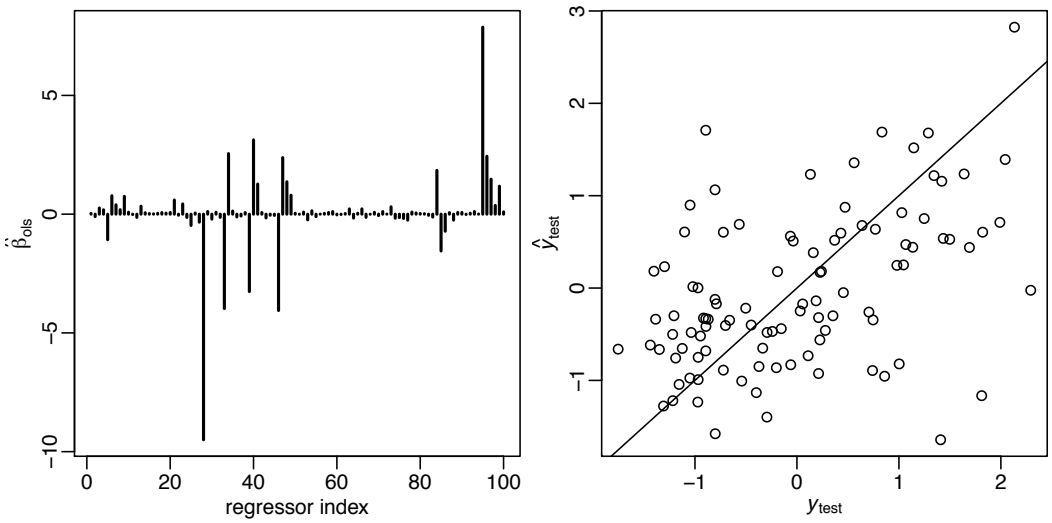
$$\Theta = [0, 1] \qquad \mathcal{Y} = \{0, 1, \ldots, 20\}\,.$$

Sampling model

Before the sample is obtained, the number of individuals with the mutation is unknown. Let Y denote this to-be-determined value.

Sampling model: If the value of θ were known, a reasonable sampling model for Y would be a binomial$(20, \theta)$ probability distribution:

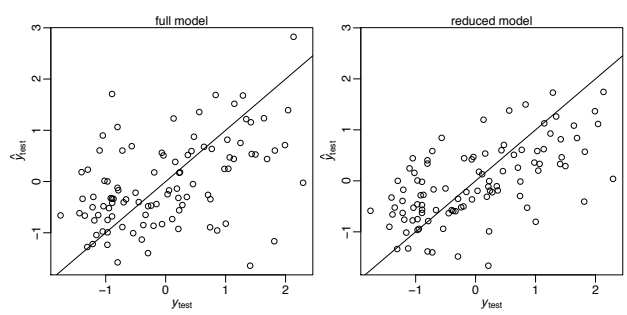
$$Y|\theta \sim \text{binomial}(20, \theta)\,.$$



Prior distribution

$$\theta \sim \text{beta}(2, 20)$$

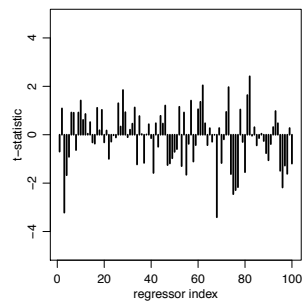
$$\begin{aligned} \text{E}[\theta] &= 0.09 \\ \text{mode}[\theta] &= 0.05 \\ \Pr(0.01 < \theta < 0.24) &= 0.95 \end{aligned}$$



Posterior distribution

$$\{\theta|Y=0\} \sim \text{beta}(2,40)$$

$$\begin{aligned} \text{E}[\theta|Y=0] &= 0.048 \\ \text{mode}[\theta|Y=0] &= 0.025 \\ \Pr(.006<\theta<.129|Y=0) &= 0.95. \end{aligned}$$



Sensitivity analysis

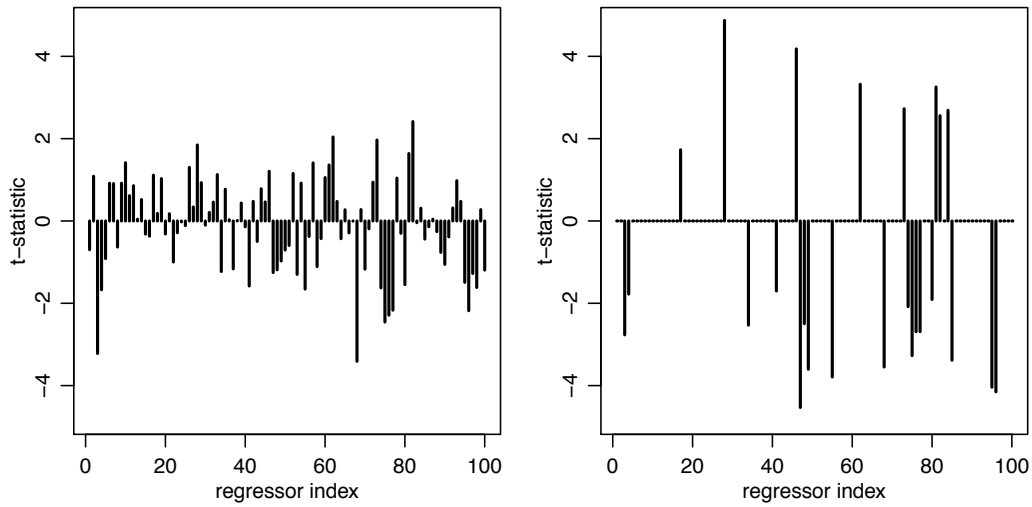
$$\theta \sim \text{beta}(a,b) \Rightarrow \{\theta|Y=y\} \sim \text{beta}(a+y, b+n-y)$$

$$\begin{aligned} \text{E}[\theta|Y=y] &= \frac{a+y}{a+b+n} \\ &= \frac{n}{w+n}\bar{y} + \frac{w}{w+n}\theta_0, \end{aligned}$$

where $\theta_0 = a/(a+b)$ is the prior expectation of θ and $w = a+b$.

Sensitivity analysis

$$
\begin{aligned}
E[\theta|Y=0] &= \frac{n}{w+n}\bar{y} + \frac{w}{w+n}\theta_0 \\
&= \frac{20}{w+20} \times 0 + \frac{w}{w+20} \times \theta_0 \\
&= \frac{w}{w+20} \times \theta_0
\end{aligned}
$$



Comparison to non-Bayesian methods

Non-Bayesian estimate:

$$\hat{\theta} = \bar{y} = y/n$$

For our data, $\hat{\theta} = 0$.

Non-Bayesian confidence interval:

$$\bar{y} \pm 1.96\sqrt{\bar{y}(1-\bar{y})/n} \quad \text{(Wald interval)}$$

For our data, 0 ± 0.

"Adjusted Wald interval" :

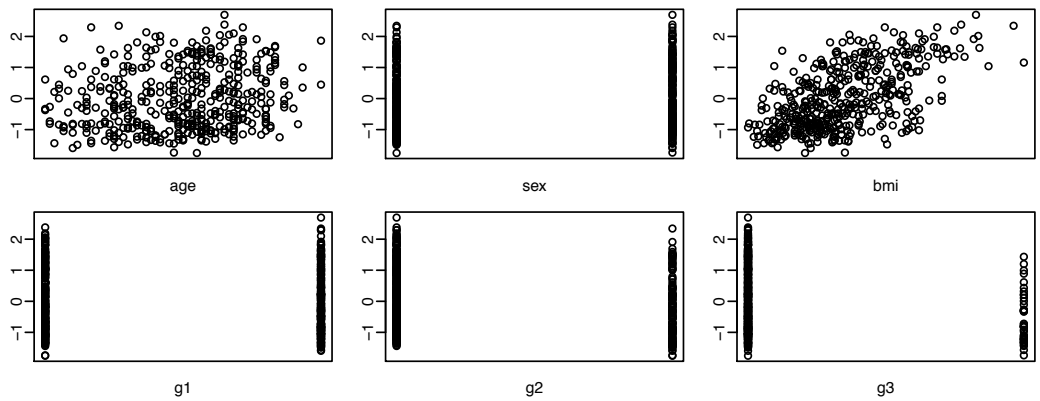
$$
\begin{aligned}
\hat{\theta} &\pm 1.96\sqrt{\hat{\theta}(1-\hat{\theta})/n}\,, \text{ where} \\
\hat{\theta} &= \frac{n}{n+4}\bar{y} + \frac{4}{n+4}\frac{1}{2}.
\end{aligned}
$$

Can be seen as approximately Bayesian.

Example: diabetes progression

Goal:
Predict diabetes progression Y as a function of explanatory variables $\mathbf{x}$.

- Y is a subject's yearly diabetes progression index
- $\mathbf{x}$ is a 100-dimensional vector, including sex, age, genotype, ...



Building a predictive model

We need to

- build a predictive model,
- evaluate its predictive accuracy.

We will do this using out-of-sample validation.

Data on 442 subjects. We divide these into

- $n = 342$ training cases, with which to fit the model;
- $n_{\text{test}} = 100$ test cases, with which to evaluate the model.

Sampling model and parameter space

Data:

- item Y_i is a subject i's progression;
- $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,100})$ are i's explanatory variables.

Linear regression model:

$$Y_i = \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_{100} x_{i,100} + \sigma \epsilon_i.$$

Parameters to estimate include

- $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{100})$
- σ

Prior distribution

The role of the Bayesian prior:

Idealistic: $p(\boldsymbol{\beta})$ should exactly describe your prior information about $\boldsymbol{\beta}$.

Realistic: $p(\boldsymbol{\beta})$ should capture the gross features of our information about $\boldsymbol{\beta}$.

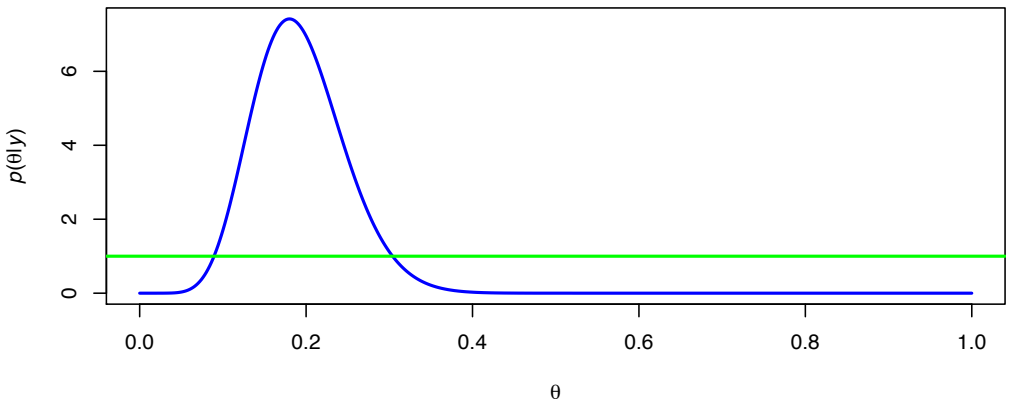
In our example we believe that $\beta_j \approx 0$ for many j.

We will use a prior such that

$$\Pr(\beta_j = 0) = 1/2 \text{ for each } j.$$

Posterior distribution

$$\begin{aligned} \Pr(\beta_j \neq 0) &= 1/2 \text{ for each } j \in \{1, \ldots 100\} \\ \Pr(\beta_j \neq 0 | \mathbf{Y}, \mathbf{X}) &\geq 1/2 \text{ for only six } j \end{aligned}$$



Out of sample validation

How well does the model perform?

Out of sample predictive performance: Compare $y_{i,\text{test}}$ to $\hat{y}_{i,\text{test}}$, where

$$\hat{y}_{i,\text{test}} = \hat{\boldsymbol{\beta}}^T \mathbf{x}_{i,\text{test}}$$

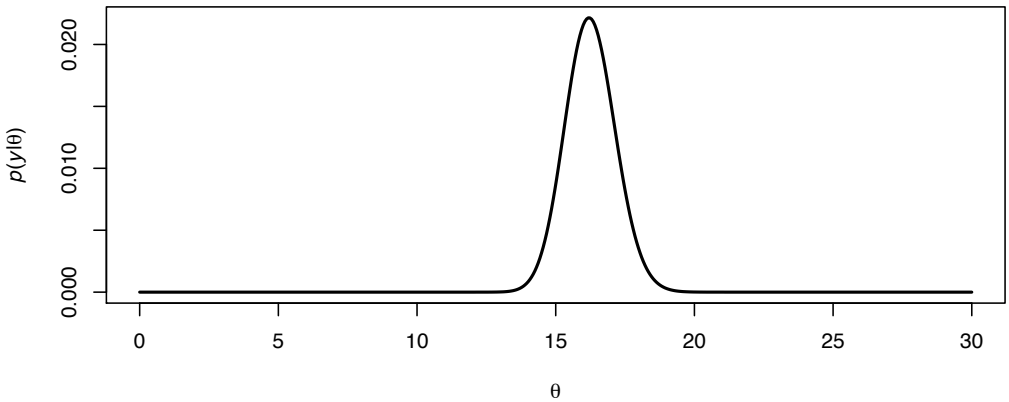
Important! $\hat{\boldsymbol{\beta}}$ is estimated from the 342 training subjects, not the test subjects.

Prediction error

$$SPE(\hat{\boldsymbol{\beta}}) = \frac{1}{100} \sum (y_{i,\text{test}} - \hat{y}_{i,\text{test}})^2$$

Out of sample validation

$$SPE(\hat{\beta}_{\text{bayes}}) = 0.4855 \qquad SPE(\hat{\beta}_{\text{ols}}) = 0.9263$$



Summary

The Bayesian approach provides

- models for rational, quantitative learning;
- estimators that work for small and large sample sizes;
- methods for generating statistical procedures in complicated problems.

Module 4: Bayesian Methods
Lecture 2: Review of Probability

Peter Hoff

Departments of Statistics and Biostatistics
University of Washington

Outline

Beliefs and Probability

Bayes rule

Standard distributions

Posterior summaries

Belief functions

Let F, G, and H be three statements about the world. For example:

$F = \{$ a woman will have breast cancer $\}$
$G = \{$ a woman will undergo HRT $\}$
$H = \{$ a woman has a mutation of BRAC1 $\}$

A *belief function* Be() assigns numbers to statements representing our beliefs:

$$\text{Be}(F|G) \overset{?}{>} \text{Be}(F)$$
$$\text{Be}(F|H) \overset{?}{>} \text{Be}(F)$$

Axioms of belief

B1 $\text{Be}(\text{not } H|H) \leq \text{Be}(F|H) \leq \text{Be}(H|H)$
B2 $\text{Be}(F \text{ or } G|H) \geq \max\{\text{Be}(F|H), \text{Be}(G|H)\}$
B3 $\text{Be}(F \text{ and } G|H)$ can be derived from $\text{Be}(G|H)$ and $\text{Be}(F|G \text{ and } H)$

Do these axioms make sense?

B1: $\text{Be}(F|H)$ is between complete disbelief and complete belief.

B2: Belief in a statement is increasing in the number of possibilities

B3: Deciding on F and G can be done by first deciding on G, then F given G.

Axioms of probability

$F \cup G$ means "F or G," $F \cap G$ means "F and G" and $\emptyset$ is the empty set.

P1 $0 = \Pr(\text{not } H|H) \leq \Pr(F|H) \leq \Pr(H|H) = 1$
P2 $\Pr(F \cup G|H) = \Pr(F|H) + \Pr(G|H)$ if $F \cap G = \emptyset$
P3 $\Pr(F \cap G|H) = \Pr(G|H)\Pr(F|G \cap H)$

A probability function, satisfying **P1, P2** and **P3**, also satisfies **B1, B2** and **B3**.

Examples

[**P2**] $\Pr(F \cup G|H) = \Pr(F|H) + \Pr(G|H)$ if $F \cap G = \emptyset$

- Genotype is $\{aa, Aa, AA\}$ with probability $\{1/4, 1/2, 1/4\}$.
- $F = \{\text{genotype is } aa\}$, $G = \{\text{genotype is } Aa\}$

$$\begin{aligned} \Pr(\text{one or more } a \text{ alleles}) &= \Pr(F \cup G) \\ &= \Pr(F) + \Pr(G) \\ &= 1/4 + 1/2 = 3/4 \end{aligned}$$

[**P3**] $\Pr(F \cap G|H) = \Pr(G|H)\Pr(F|G \cap H)$

- Parents are AB/AB and AB/ab
- $F = \{$ child is A/a at first locus $\}$, $G = \{$ child is B/B at second locus $\}$

$$\begin{aligned} \Pr(AB/aB) &= \Pr(F \cap G) \\ &= \Pr(G) \times \Pr(F|G) \\ &= \Pr(F|G)/2 \end{aligned}$$

Events and partitions

Definition (Partition)

A collection of sets $\{H_1, \ldots, H_K\}$ is a partition of another set $\mathcal{H}$ if

1. *the events are disjoint, which we write as $H_i \cap H_j = \emptyset$ for $i \neq j$;*
2. *the union of the sets is $\mathcal{H}$, which we write as $\cup_{k=1}^{K} H_k = \mathcal{H}$.*

If $\mathcal{H}$ is the set of all possible truths and $\{H_1, \ldots, H_K\}$ is a partition of $\mathcal{H}$, then exactly one out of $\{H_1, \ldots, H_K\}$ contains the truth.

Examples:

- $\mathcal{H}$=someone's number of children
  - {0, 1, 2, 3 or more};
  - {0, 1, 2, 3, 4, 5, 6, ...}.
- $\mathcal{H}$ = the relationship between a genotype and heart disease
  - {some relationship, no relationship};
  - {negative correlation, zero correlation, positive correlation}.

Bayes' rule

The axioms of probability imply the following:

Rule of total probability : $$\sum_{k=1}^{K} \Pr(H_k) = 1$$

Rule of marginal probability : $$\begin{aligned} \Pr(E) &= \sum_{k=1}^{K} \Pr(E \cap H_k) \\ &= \sum_{k=1}^{K} \Pr(E|H_k)\Pr(H_k) \end{aligned}$$

Bayes' rule : $$\begin{aligned} \Pr(H_j|E) &= \frac{\Pr(E|H_j)\Pr(H_j)}{\Pr(E)} \\ &= \frac{\Pr(E|H_j)\Pr(H_j)}{\sum_{k=1}^{K} \Pr(E|H_k)\Pr(H_k)} \end{aligned}$$

An example

A sonogram indicates someone will have twin girls.

What is the probability that the girls are monozygotic?

- $\mathcal{H} = \{H_1, H_2\} = \{$ monozygotic , dizygotic $\}$
- $E = \{$ twins are girls $\}$

An example

- $\mathcal{H} = \{H_1, H_2\} = \{$ monozygotic , dizygotic $\}$
- $E = \{$ twins are girls $\}$

Observed data: twins are girls.
Prior information: One third of twins are monozygotic (identical).

$$
\begin{aligned}
\Pr(H_1|E) &= \frac{\Pr(E|H_1)\Pr(H_1)}{\Pr(E)} \\
&= \frac{\Pr(E|H_1)\Pr(H_1)}{\Pr(E|H_1)\Pr(H_1) + \Pr(E|H_2)\Pr(H_2)} \\
&= \frac{1/2 \times 1/3}{1/2 \times 1/3 + 1/4 \times 2/3} \\
&= \frac{1/6}{1/6 + 1/6} \\
&= 1/2 \quad > \quad 1/3 \quad = \quad \Pr(H_1)
\end{aligned}
$$

Bayesian inference

$\{H_1, \ldots, H_K\}$ often refer to disjoint hypotheses or states of nature

E refers to the outcome of a survey, study or experiment.

Post-experimental evaluation of hypotheses are via the posterior odds ratio:

$$\begin{aligned}
\frac{\Pr(H_i|E)}{\Pr(H_j|E)} &= \frac{\Pr(E|H_i)\Pr(H_i)/\Pr(E)}{\Pr(E|H_j)\Pr(H_j)/\Pr(E)} \\
&= \frac{\Pr(E|H_i)\Pr(H_i)}{\Pr(E|H_j)\Pr(H_j)} \\
&= \frac{\Pr(E|H_i)}{\Pr(E|H_j)} \times \frac{\Pr(H_i)}{\Pr(H_j)} \\
&= \text{“Likelihood ratio”} \times \text{“prior odds”}
\end{aligned}$$

Twin example

Prior odds:

$$\frac{\Pr(H_1)}{\Pr(H_2)} = \frac{1/3}{2/3} = 1/2$$

Prior favors H_2

Likelihood ratio:

$$\frac{\Pr(E|H_1)}{\Pr(E|H_2)} = \frac{1/2}{1/4} = 2$$

Data is better explained by H_1

Posterior odds:

$$\frac{\Pr(H_1)}{\Pr(H_2)}\frac{\Pr(E|H_1)}{\Pr(E|H_2)} = 1$$

Independence

F and G are conditionally independent given H if

$$\Pr(F \cap G|H) = \Pr(F|H)\Pr(G|H)$$

Interpretation: By Axiom **P3**, the following is always true:

$$\Pr(F \cap G|H) = \Pr(G|H)\Pr(F|H \cap G).$$

If F and G are conditionally independent given H, then we must have

$$\begin{aligned}
\Pr(G|H)\Pr(F|H \cap G) &\overset{\text{always}}{=} \Pr(F \cap G|H) \overset{\text{independence}}{=} \Pr(F|H)\Pr(G|H) \\
\Pr(G|H)\Pr(F|H \cap G) &= \Pr(F|H)\Pr(G|H) \\
\Pr(F|H \cap G) &= \Pr(F|H).
\end{aligned}$$

Conditional independence therefore implies that $\Pr(F|H \cap G) = \Pr(F|H)$.

If H is known and F and G are conditionally independent given H, then knowing G does not change our belief about F.

Independence

Example:

$F = \{$ a patient will develop cancer $\}$
$G = \{$ the parents' genotypes $\}$
$H = \{$ a patient's genotype $\}$

$$\Pr(F|H) \overset{?}{=} \Pr(F|G, H)$$

Discrete random variables

Let Y be a *random variable*, an unknown numerical quantity.

Let $\mathcal{Y}$ be the set of all possible values of Y.

Y is *discrete* if the set of possible outcomes is *countable*, meaning that $\mathcal{Y}$ can be expressed as $\mathcal{Y} = \{y_1, y_2, \ldots\}$.

Examples

- Y = number of people in a population with a specific allele
- Y = number of children of a randomly sampled person
- Y = number of years of education of a randomly sampled person

Probability distributions and densities

$\Pr(Y = y)$ is the probability that the outcome Y takes on the value y.

PDF: $\Pr(Y = y) = p(y)$ is often called the *probability density function* (pdf) of Y

1. $0 \leq p(y) \leq 1$ for all $y \in \mathcal{Y}$;
2. $\sum_{y \in \mathcal{Y}} p(y) = 1$.

We can derive various probabilities from $p(y)$:

$$\Pr(Y \in A) = \sum_{y \in A} p(y)$$

If A and B are disjoint subsets of $\mathcal{Y}$, then

$$\begin{aligned} \Pr(Y \in A \text{ or } Y \in B) \equiv \Pr(Y \in A \cup B) &= \Pr(Y \in A) + \Pr(Y \in B) \\ &= \sum_{y \in A} p(y) + \sum_{y \in B} p(y). \end{aligned}$$

Probability densities

If (to a rough approximation) $\mathcal{Y} = \mathbb{R}$, then we cannot define $\Pr(Y \leq 5)$ as equal to $\sum_{y \leq 5} p(y)$ because the sum does not make sense.

Instead, we define a probability density as a function $p(y)$ such that

$$\Pr(Y \in A) = \int_A p(y)\ dy$$

1. $0 \leq p(y)$ for all $y \in \mathcal{Y}$;
2. $\int_{\mathbb{R}} p(y)\ dy = 1$.

As in the discrete case, probability statements about Y can be derived from the pdf: $\Pr(Y \in A) = \int_{y \in A} p(y)\ dy$, and if A and B are disjoint subsets of $\mathcal{Y}$, then

$$\begin{aligned} \Pr(Y \in A \text{ or } Y \in B) \equiv \Pr(Y \in A \cup B) &= \Pr(Y \in A) + \Pr(Y \in B) \\ &= \int_{y \in A} p(y)\ dy + \int_{y \in B} p(y)\ dy. \end{aligned}$$

Notes: Unlike the discrete case,

- $p(y)$ can be larger than 1;
- $p(y)$ is not "the probability that $Y = y$."

However, if $p(y_1) > p(y_2)$ we will sometimes informally say that y_1 "has a higher probability" than y_2.

The binary distribution

Let $\mathcal{Y} = \{0, 1\}$.

The outcome Y has a *binary distribution with probability* θ if

$$\Pr(Y = y|\theta) = p(y|\theta) = \begin{cases} \theta & \text{if } y = 1 \\ (1-\theta) & \text{if } y = 0 \end{cases}$$

Alternatively, we can write

$$\Pr(Y = y|\theta) = p(y|\theta) = \theta^y (1-\theta)^{1-y}$$

Independent binary outcomes

Suppose the prevalence of an allele in a population is θ.

Let $Y_1, \ldots, Y_n$ indicate the presence of the allele for n individuals randomly sampled from the population.

$$\begin{aligned}
\Pr(Y_1 = y_1, \ldots, Y_n = y_n|\theta) &= p(y_1, \ldots, y_n|\theta) \\
&= \left(\theta^{y_1}(1-\theta)^{1-y_1}\right) \times \cdots \times \left(\theta^{y_n}(1-\theta)^{1-y_n}\right) \\
&= \theta^{\sum y_i}(1-\theta)^{n-\sum y_i}
\end{aligned}$$

Note that $p(y_1, \ldots, y_n|\theta)$ depends only on $\sum y_i$.

Often, we only record $y = \sum y_i$ and n.

The binomial distribution

What is the probability that y people in a sample of size n will have the allele?

Consider all n-sequences with y 1's:

$$\begin{aligned}
\Pr(Y_1 = 0, Y_2 = 1, Y_3 = 0, \ldots, Y_n = 1|\theta) &= \theta^y(1-\theta)^{n-y} \\
\vdots \qquad & \quad \vdots \\
\Pr(Y_1 = 1, Y_2 = 0, Y_3 = 1, \ldots, Y_n = 0|\theta) &= \theta^y(1-\theta)^{n-y}
\end{aligned}$$

There are $\binom{n}{y}$ such sequences, so

$$\Pr(\sum Y_i = y|\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}$$

The binomial distribution

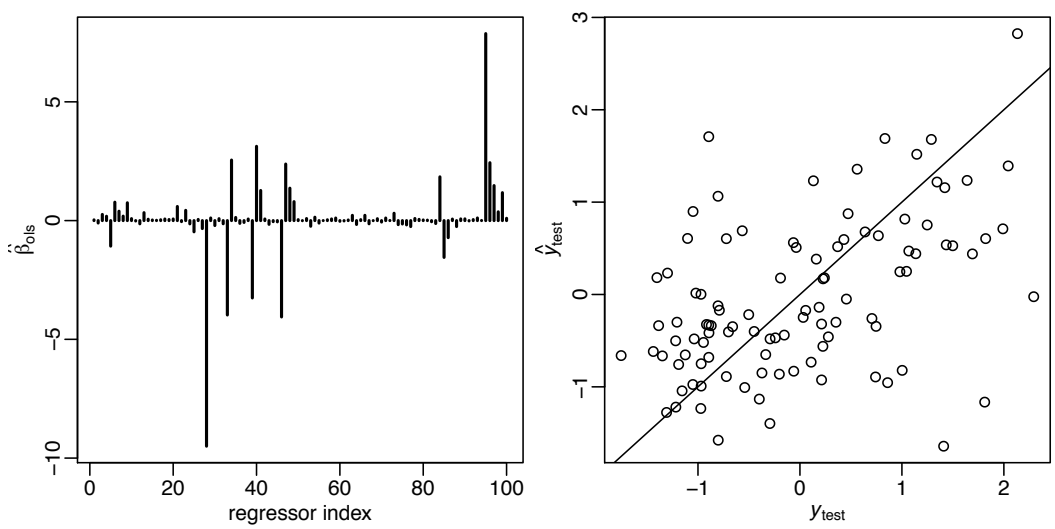
Let $\mathcal{Y} = \{0, 1, 2, \ldots, n\}$ for some positive integer n. The outcome $Y \in \mathcal{Y}$ has a *binomial distribution with probability* θ if

$$\Pr(Y = y|\theta) = \text{dbinom}(y, n, \theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}.$$

For example, if $\theta = .25$ and $n = 4$, we have:

$$\begin{aligned}
\Pr(Y = 0|\theta = .25) &= \binom{4}{0}(.25)^0(.75)^4 = .316 \\
\Pr(Y = 1|\theta = .25) &= \binom{4}{1}(.25)^1(.75)^3 = .422 \\
\Pr(Y = 2|\theta = .25) &= \binom{4}{2}(.25)^2(.75)^2 = .211 \\
\Pr(Y = 3|\theta = .25) &= \binom{4}{3}(.25)^3(.75)^1 = .047 \\
\Pr(Y = 4|\theta = .25) &= \binom{4}{4}(.25)^4(.75)^0 = .004\,.
\end{aligned}$$

The binomial distribution



Likelihood inference for θ

$y = 9$ out of $n = 50$ randomly sampled subjects have a particular allele.

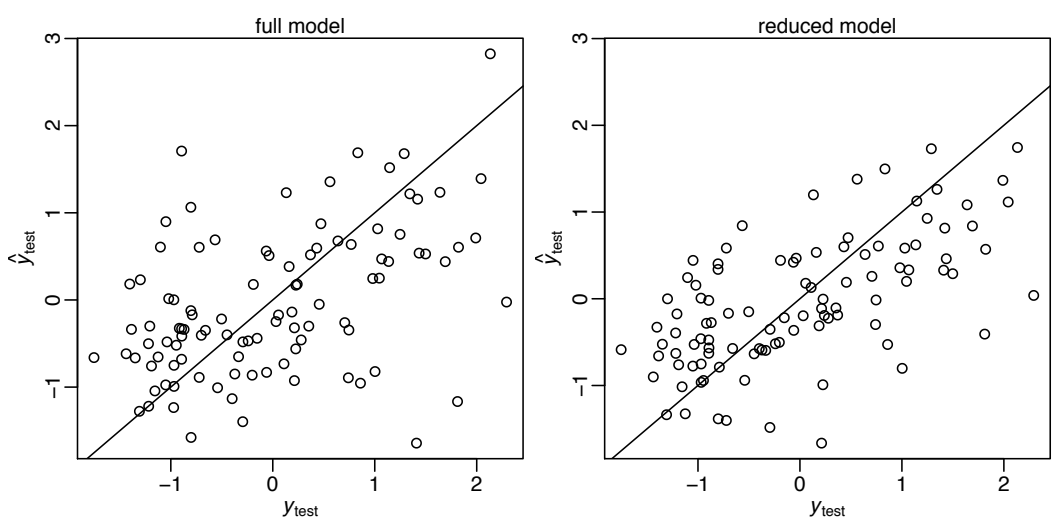
What is the allele frequency in the general population?

The likelihood:

$$\begin{aligned}
\Pr(Y = 9|\theta = 0.15) &= 0.033 \\
\Pr(Y = 9|\theta = 0.35) &= 0.004 \\
\Pr(Y = 9|\theta = 0.65) &\approx 10^{-11} \\
\Pr(Y = 9|\theta = 0.90) &\approx 10^{-32}
\end{aligned}$$

As a function of θ, $\Pr(Y = y|\theta) = p(y|\theta)$ is called the *likelihood of* θ.

The binomial likelihood



Posterior inference via Bayes' rule

Suppose

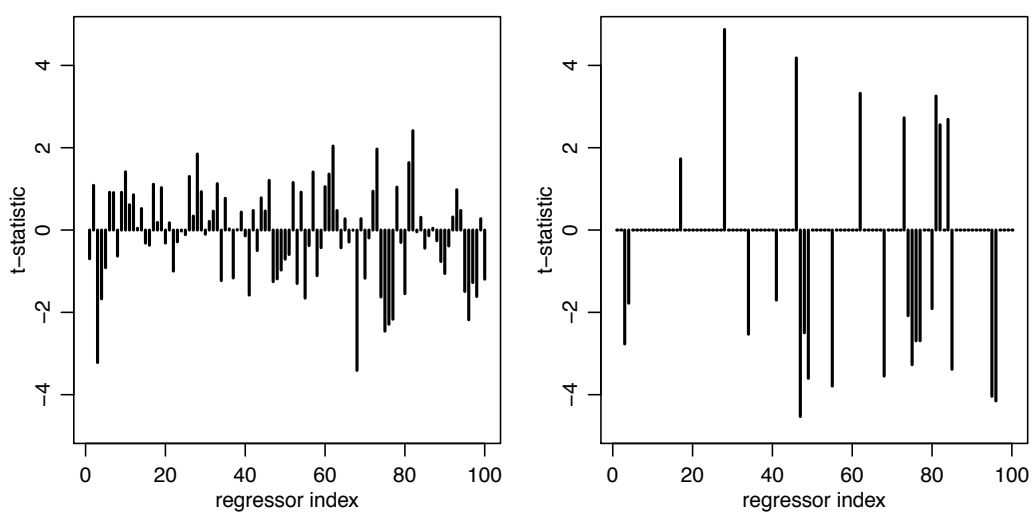
- $\Theta = \{0.01, 0.02, 0.03, \ldots, 0.99, 1.00\}$ and
- $p(\theta) = 1/100$ for each $\theta \in \Theta$.

Bayes' rule says

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\sum_{\theta\in\Theta} p(y|\theta)p(\theta)}$$

```
n<-50
y<-9

theta<-seq(.01,1,length=100)

py.theta<-dbinom(y,n,theta)
ptheta<-rep(1/100,length=length(theta))

ptheta.y<- py.theta*ptheta / sum( py.theta*ptheta )
```


The posterior distribution



The posterior distribution

Now suppose

- $\Theta = [0, 1]$ and
- $p(\theta) = 1$ for each $\theta \in \Theta$.

Bayes' rule says

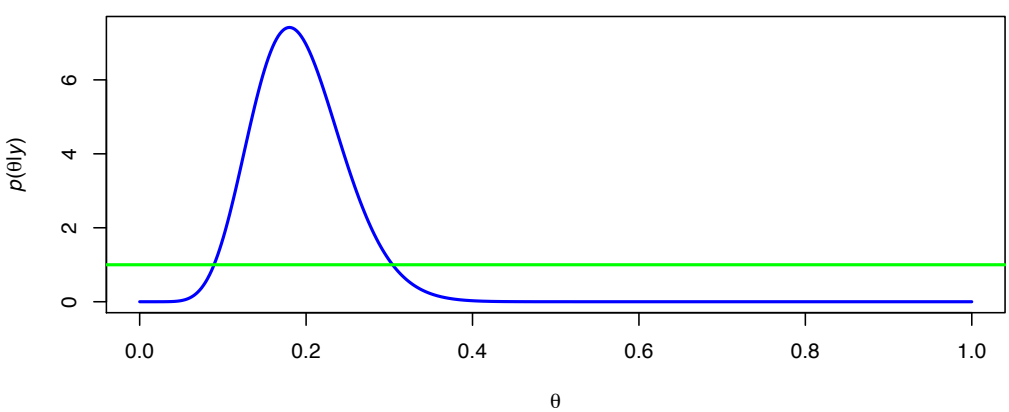
$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int_0^1 p(y|\theta)p(\theta)}$$

The next lecture shows $p(\theta|y)$ is a *beta* $(1 + y, 1 + n - y)$ *distribution*, and

$$\mathrm{E}[\theta|y] = \frac{1}{2}\left(\frac{2}{2+n}\right) + \frac{y}{n}\left(\frac{n}{2+n}\right)$$

```
n<-50
y<-9

theta<-seq(0,1,length=500)
ptheta.y<-dbeta(theta,1+y,1+n-y)
```


The posterior distribution



The Poisson distribution

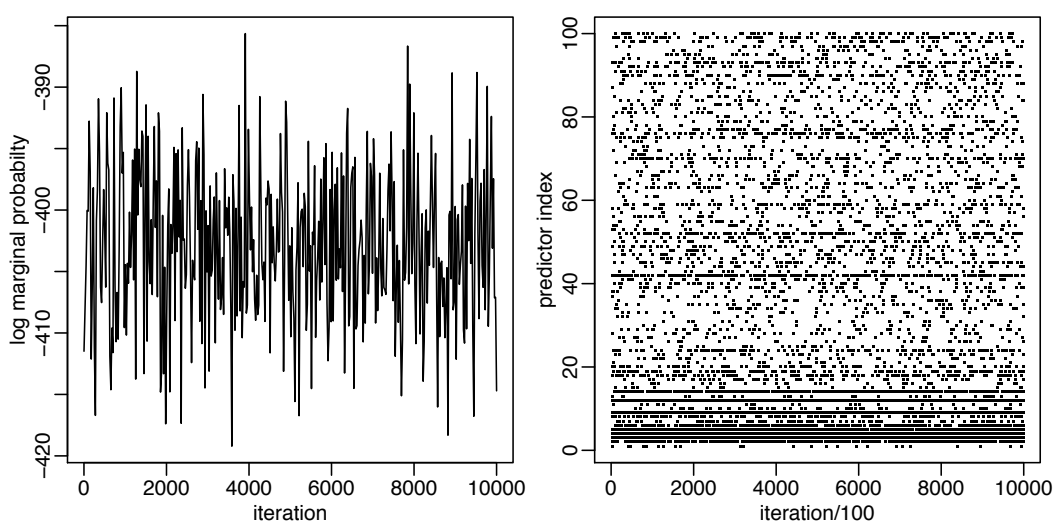
Let $\mathcal{Y} = \{0, 1, 2, \ldots\}$. The outcome $Y \in \mathcal{Y}$ has a *Poisson distribution with mean* θ if

$$\Pr(Y = y|\theta) = \mathrm{dpois}(y, \theta) = \theta^y e^{-\theta}/y!.$$

For example, if $\theta = 2.1$ (the 2006 U.S. fertility rate),

$$\begin{array}{lll} \Pr(Y = 0|\theta = 2.1) = & (2.1)^0 e^{-2.1}/(0!) = & .12 \\ \Pr(Y = 1|\theta = 2.1) = & (2.1)^1 e^{-2.1}/(1!) = & .26 \\ \Pr(Y = 2|\theta = 2.1) = & (2.1)^2 e^{-2.1}/(2!) = & .27 \\ \Pr(Y = 3|\theta = 2.1) = & (2.1)^3 e^{-2.1}/(3!) = & .19 \\ \vdots & \vdots & \vdots \end{array}$$

The Poisson distribution



Poisson distributions with means of 2.1 and 21.

The Poisson likelihood

Let Y_i be the number of intestinal tumors for mouse i, $i = 1, \ldots n$.

What is the mean tumor count in this population?

The likelihood:

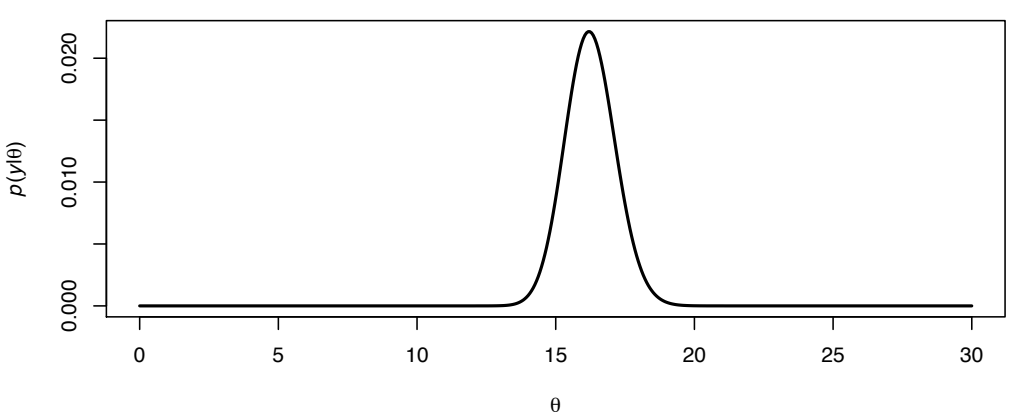
$$\begin{aligned}
\Pr(Y_1 = y_1, \ldots, Y_n = y_n|\theta) &= p(y_1, \ldots, y_n|\theta) \\
&= \prod_{i=1}^{n} p(y_i|\theta) \\
&= \prod_{i=1}^{n} \theta^{y_i} e^{-\theta}/y_i! \\
&= \theta^{\sum y_i} e^{-n\theta} \times (\prod y_i!)^{-1}
\end{aligned}$$

Simplification: Let $Y = \sum Y_i$. Then $\{Y|\theta\} \sim \text{Poisson}(n\theta)$ and so

$$\Pr(Y = y|\theta) = \theta^y e^{-n\theta} \times (n^y/y!)$$

The Poisson likelihood

Suppose $n = 20$ and $y = \sum y_i = 324$ $(y/n = 16.2)$.

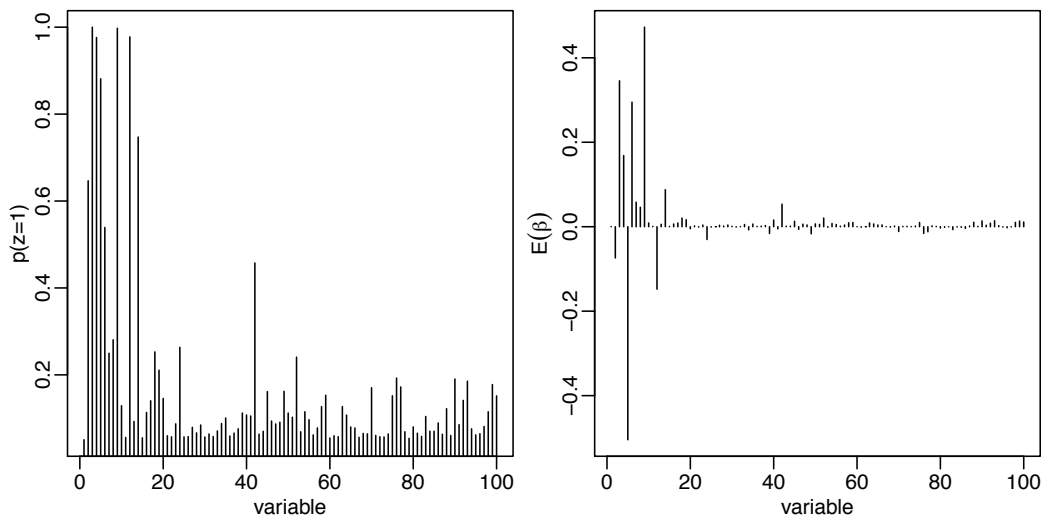


Posterior inference

Similar populations of mice suggest $\theta \approx 10$.

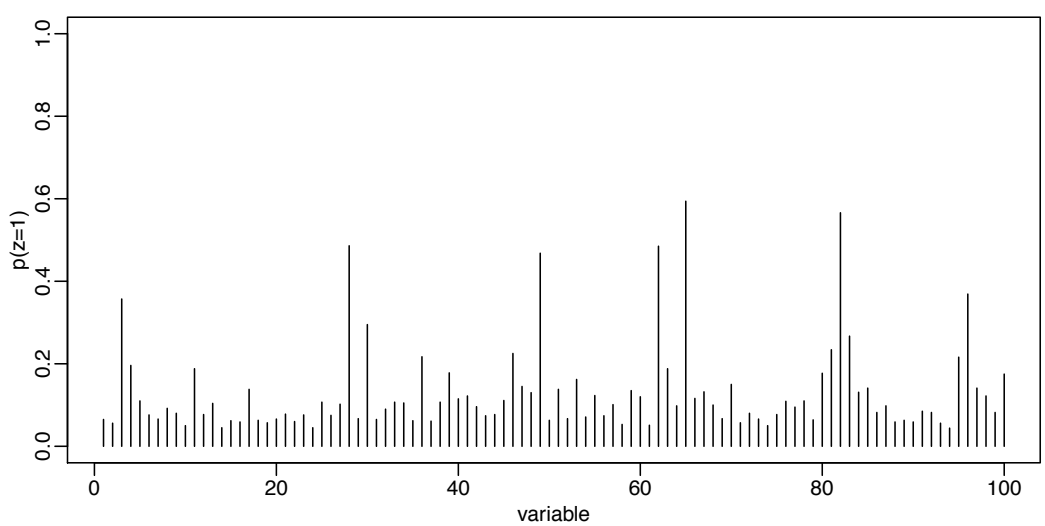
A suitable prior distribution for θ would reflect this:

$$\begin{aligned} \theta &\sim \text{gamma}(10, 1) \\ \text{E}[\theta] &= 10 \\ \text{SD}[\theta] &= \sqrt{10} \approx 3.16 \end{aligned}$$



Discrete approximation to the posterior

```
n<-20
sy<-324

theta<-seq(0,30,length=500)

psy.theta<-dpois(sy, n*theta)
p.theta<-dgamma(theta,10,1) ; p.theta<-p.theta/sum(p.theta)
p.theta.y<- psy.theta * p.theta / sum( psy.theta * p.theta)
```



The gamma posterior distribution

It can be shown that if

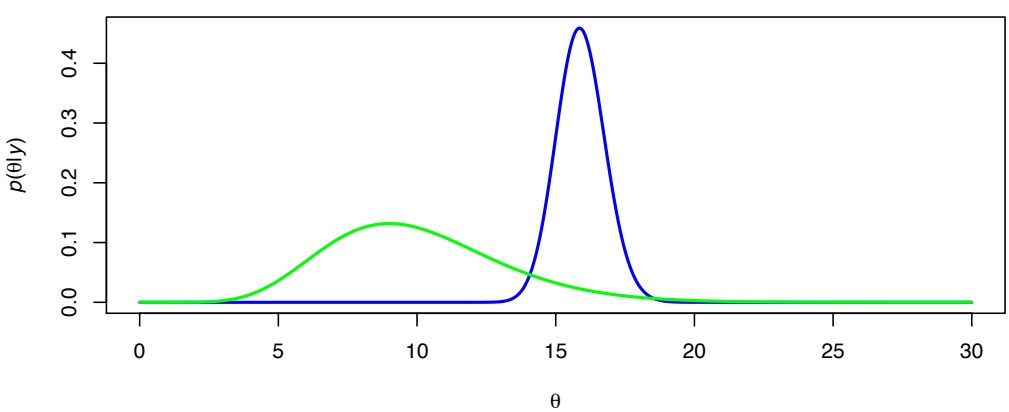
- $\theta \sim \text{gamma}(a, b)$ and
- $Y_1, \ldots, Y_n | \theta \sim \text{Poisson}(\theta)$ then

$$\{\theta | \sum Y_i = y\} \sim \text{gamma}(a + y, b + n)$$

$$\begin{aligned} E[\theta | \sum Y_i = y] &= \frac{a+y}{b+n} \\ &= \frac{a}{b}\left(\frac{b}{b+n}\right) + \frac{y}{n}\left(\frac{n}{b+n}\right) \end{aligned}$$

The gamma posterior distribution

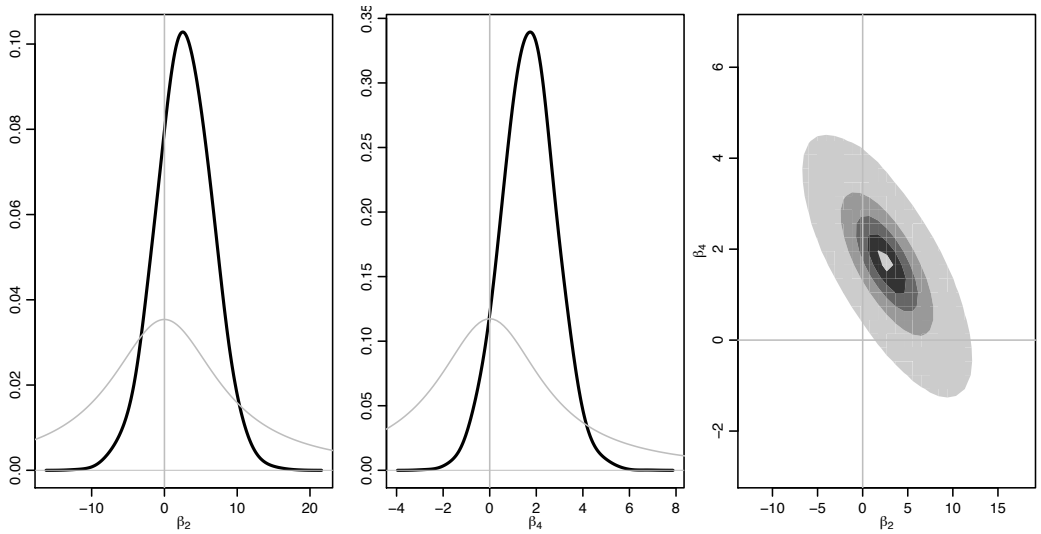
```
a<-10 ; b<-1                          # prior

sy<-324 ; n<-20                       # data

theta<-seq(0,30,length=500)

p.theta.y<- dgamma(theta, a+sy, b+n)  # posterior
```



The normal distribution

Let $\mathcal{Y} = (-\infty, \infty)$. The outcome $Y \in \mathcal{Y}$ has a *normal distribution with mean* θ *and variance* σ^2 if

$$p(y|\theta, \sigma^2) = \text{dnorm}(y, \theta, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}\left(\frac{y-\theta}{\sigma}\right)^2\right\}.$$



The normal posterior distribution

It can be shown that if

- $\theta \sim \text{normal}(\mu_0, \tau_0^2)$ and
- $Y_1, \ldots, Y_n|\theta \sim \text{normal}(\theta, \sigma^2)$

then

$$\{\theta|y_1, \ldots, y_n\} \sim \text{normal}(\mu_n, \tau_n^2)$$

where

$$\begin{aligned} \text{Var}[\theta|y_1, \ldots, y_n] = \tau_n^2 &= [1/\tau_0^2 + n/\sigma^2]^{-1} \\ 1/\tau_n^2 &= 1/\tau_0^2 + n/\sigma^2 \end{aligned}$$

and

$$\begin{aligned} \text{E}[\theta|y_1, \ldots, y_n] = \mu_n &= \frac{\mu_0/\tau_0^2 + \bar{y}n/\sigma^2}{1/\tau_0^2 + n/\sigma^2} \\ &= \mu_0\left(\frac{1/\tau_0^2}{1/\tau_0^2 + n/\sigma^2}\right) + \bar{y}\left(\frac{n/\sigma^2}{1/\tau_0^2 + n/\sigma^2}\right) \end{aligned}$$

Describing posterior location

The posterior *mean* or *expectation* of an unknown quantity θ is given by

$$\mathrm{E}[\theta|y] = \int_{\theta\in\Theta} \theta p(\theta|y)\ d\theta$$

The mean is the center of mass of the distribution.

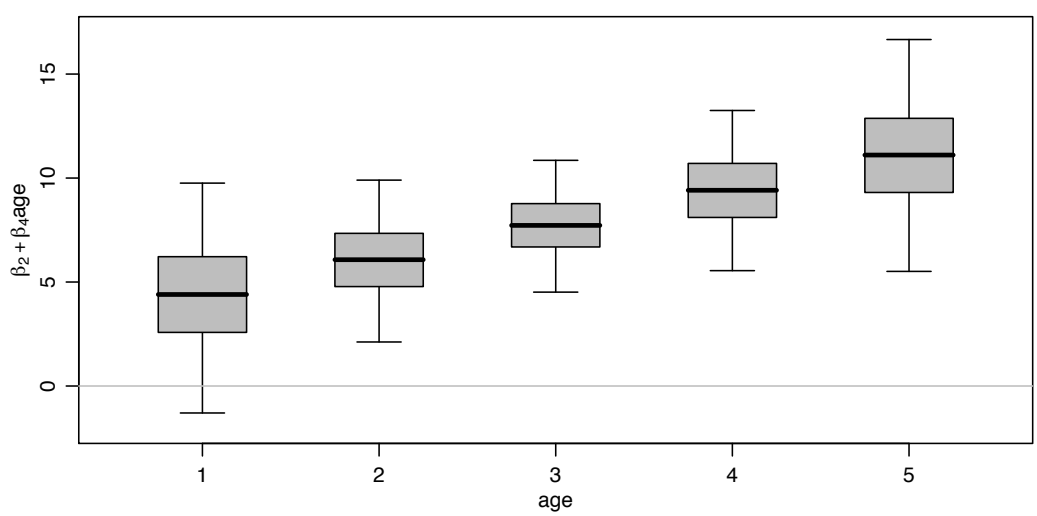
However, it is not in general equal to either of

the *mode*: "the most probable value of θ," or

the *median*: "the value of θ in the middle of the distribution."

For skewed distributions the mean can be far from a "typical" sample value

Mean, median and mode



Variance and SD

How spread out is $p(\theta|y)$?

Popular measures of spread are the *variance* and *standard deviation* :

$$\begin{aligned} \text{Var}[\theta] &= \text{E}[(\theta - \text{E}[\theta|y])^2] \\ &= \text{E}[\theta^2|y] - \text{E}[\theta|y]^2. \end{aligned}$$

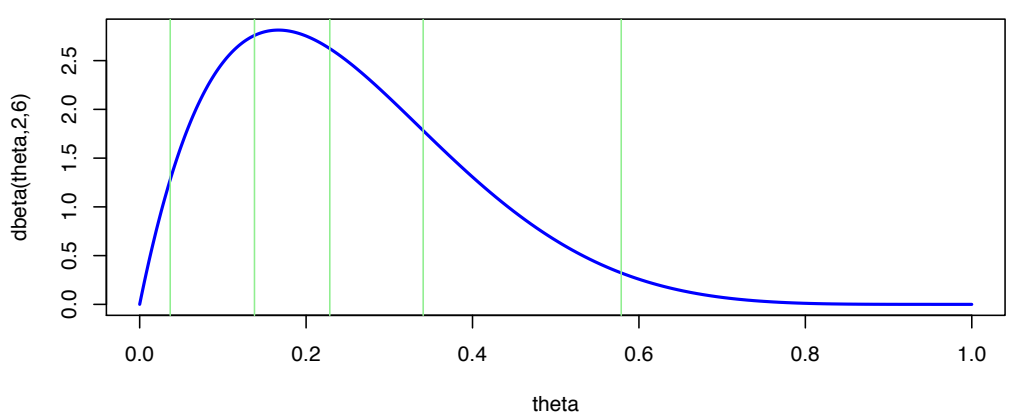
The variance is the average squared distance of θ to $\text{E}[\theta|y]$.

The *standard deviation* is the square root of the variance,
and has the same units as θ.

Quantiles

Alternative measures of spread are based on *quantiles*.

The α-quantile is the value θ_α such that $\Pr(\theta \leq \theta_\alpha|y) = \alpha$.

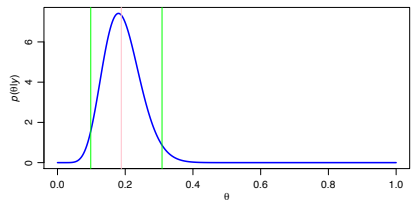


2.5, 25, 50 75 and 97.5 quantiles of a beta distribution

Quantile-based confidence intervals

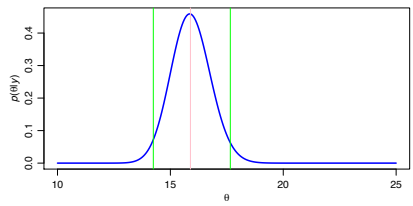
A popular way to construct a confidence interval is with quantiles:

Allele frequency example: $\theta|y \sim \text{beta}(1+y, 1+n-y)$




```
a<-1 ; b<-1    #prior
n<-50 ; y<-9   #data

qbeta(c(.025,.5,.975),a+y,b+n-y)

## [1] 0.09824 0.18834 0.30873
```

Tumor count example: $\theta|y \sim \text{gamma}(a+y, b+n)$




```
a<-10  ; b<-1     #prior
y<-324 ; n<-20    #data

qgamma(c(.025,.5,.975),a+y,b+n)

## [1] 14.24 15.89 17.66
```


Monte Carlo approximation

Posterior descriptions involve integrals we'd often like to avoid:

$$
\begin{aligned}
\text{E}[\theta|y] &= \int \theta p(\theta|y)\ d\theta \\
\text{median}[\theta|y] &= \theta_{1/2} : \int_{-\infty}^{\theta_{1/2}} p(\theta|y) = 1/2 \\
\Pr(\theta_1 < \theta_2 | y_1, y_2) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\theta_2} p(\theta_1, \theta_2 | y_1, y_2)\ d\theta_1\ d\theta_2
\end{aligned}
$$

We can easily approximate such integrals arbitrarily closely with **Monte Carlo approximation**.

Monte Carlo approximation

The basic principle: If $\theta^{(1)}, \ldots, \theta^{(S)} \sim$ iid $p(\theta|y)$, then

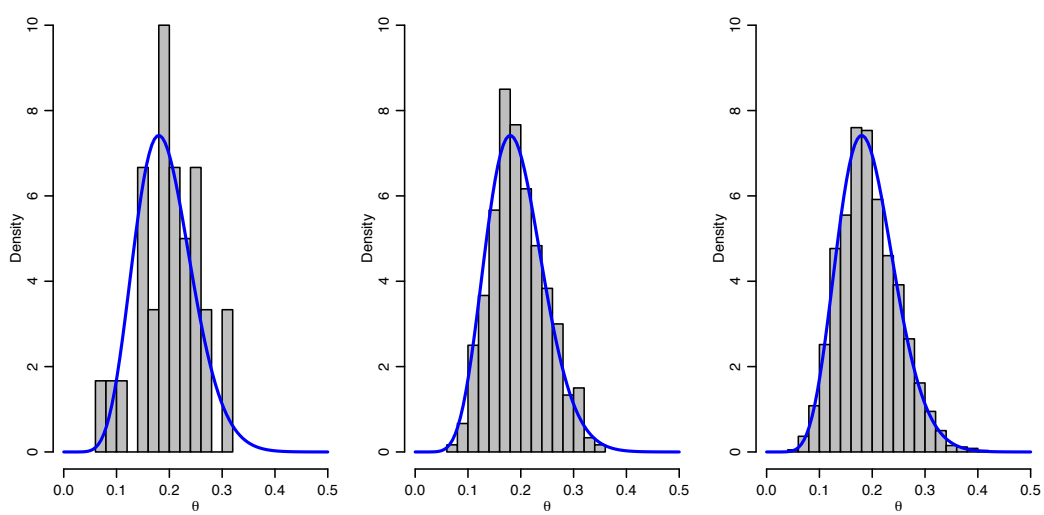
$$\text{histogram}\{\theta^{(1)}, \ldots, \theta^{(S)}\} \approx p(\theta|y)$$

This implies that

$$\begin{aligned} \frac{1}{S}\sum \theta^{(s)} &\approx \mathrm{E}[\theta|y] \\ \frac{\#\{\theta^{(s)} < c\}}{S} &\approx \Pr(\theta < c|y) \\ \text{median}\{\theta^{(1)}, \ldots, \theta^{(S)}\} &\approx \text{median}[\theta|y] \end{aligned}$$

etc. with the approximation improving with increasing S.

Monte Carlo approximation



Monte Carlo approximation

```
## correct values
qbeta(c(.025,.5,.975),a+y, b+n-y)

## [1] 0.09824 0.18834 0.30873

## Monte Carlo simulated data
theta.sim<-rbeta(3000,a+y,b+n-y)

## MC estimate based on 30 sims
quantile(theta.sim[1:30],c(.025,.5,.975))

##   2.5%    50%  97.5%
## 0.1273 0.1763 0.2831

## MC estimates based on 300 sims
quantile(theta.sim[1:300],c(.025,.5,.975))

##   2.5%    50%  97.5%
## 0.1019 0.1901 0.3313

## MC estimates based on 3000 sims
quantile(theta.sim[1:3000],c(.025,.5,.975))

##   2.5%    50%  97.5%
## 0.1002 0.1866 0.3164
```


Monte Carlo for two sample comparisons

Suppose we want to compare the prevalence of a gene in two populations.

- $\theta_1 =$ prevalence rate in population 1
- $\theta_2 =$ prevalence rate in population 2

Prior:

$$\theta_1, \theta_2 \sim \text{uniform}(0, 1)$$

Data:

- $n_1 = 50,\ y_1 = 10,\ y/n = 0.200$
- $n_2 = 27,\ y_2 = 8\ y/n = 0.296$

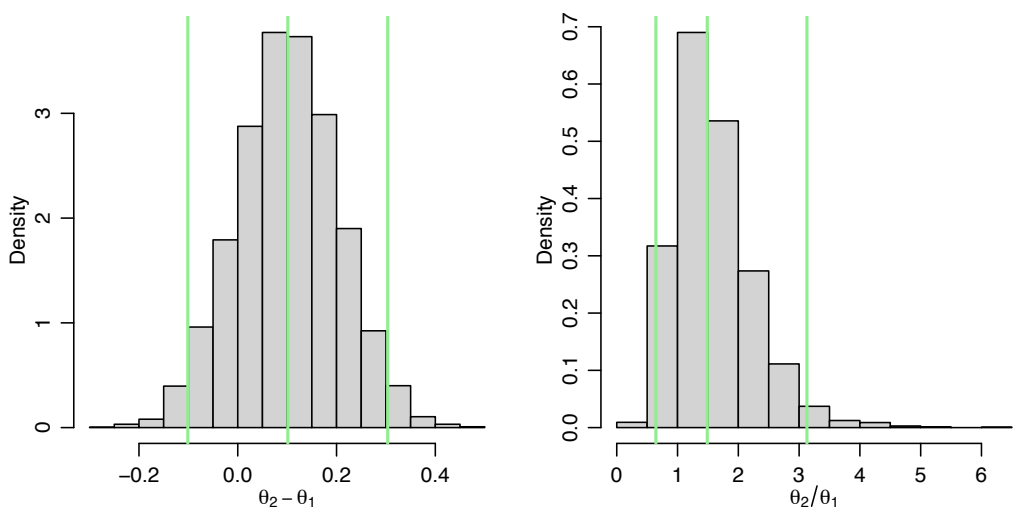
Posterior:

- $\theta_1 \sim \text{beta}(1 + y_1, 1 + n_1 - y_1) = \text{beta}(11, 41)$
- $\theta_2 \sim \text{beta}(1 + y_2, 1 + n_2 - y_2) = \text{beta}(9, 20)$

Monte Carlo for two sample comparisons

```
a<-1 ; b<-1
n1<-50 ; y1<-10
n2<-27 ; y2<-8

theta1.sim<-rbeta(5000,a+y1,b+n1-y1)
theta2.sim<-rbeta(5000,a+y2,b+n2-y2)

mean(theta1.sim<theta2.sim)

## [1] 0.8368

quantile( theta1.sim-theta2.sim,prob=c(.025,.5,.975))

##    2.5%     50%   97.5%
## -0.3036 -0.1014  0.1011
```


Monte Carlo for two sample comparisons



Summary

- Probability distributions encapsulate information
  - $p(\theta)$ describes prior information
  - $p(y|\theta)$ describes information about y for each θ
  - $p(\theta|y)$ describes posterior information
- Posterior distributions can be calculated via Bayes' rule

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta)\ d\theta}$$

- Posteriors can be plotted/summarized with commands in R
  - dbeta, dgamma, dnorm
  - qbeta, qgamma, qnorm
- Posterior integrals can be approximated with Monte Carlo methods.
  1. simulate $\theta^{(1)}, \ldots, \theta^{(S)} \sim$ iid $p(\theta|y)$
  2. approximate integrals with simulation averages.

2014 SISG Module 4: Bayesian Statistics for Genetics
Lecture 3: Binomial Sampling

Jon Wakefield

Departments of Statistics and Biostatistics
University of Washington

Outline

Introduction and Motivating Example

Bayesian Analysis of Binomial Data
- Derivation of the Posterior
- The Beta Prior
- Specifics of Prior Choice
- Bayes Factors

Analysis of ASE Data

Conclusions

Introduction

- In this lecture we will consider the Bayesian modeling of binomial data.
- The analysis of allele specific expression data will be used to motivate the binomial model.
- Conjugate priors will be introduced.
- Sampling from the posterior will be emphasized as a method for flexible inference.

Motivating Example: Allele Specific Expression

- Gene expression variation is an important contribution to phenotypic variation within and between populations.
- Expression variation may be due to genetic or environmental sources.
- Genetic variation may be due to cis- or trans-acting mechanisms.
- Polymorphisms that act in cis affect expression in an allele specific manner.
- RNA-Seq is a high throughput technology that allows allele-specific expression (ASE) to be measured.

Motivating Example: An Example of ASE

- Consider a gene with one exon and five SNPs within that exon.
- Suppose the BY allele of the gene is expressed at a high level.
- In contrast, the RM allele has a mutation in a transcription factor binding site upstream of the gene that greatly reduces expression of this allele.
- Then, in the mRNA isolated from the yeast, when we look just at this gene, there are lots more BY mRNA molecules than RM mRNA molecules.

Example of ASE

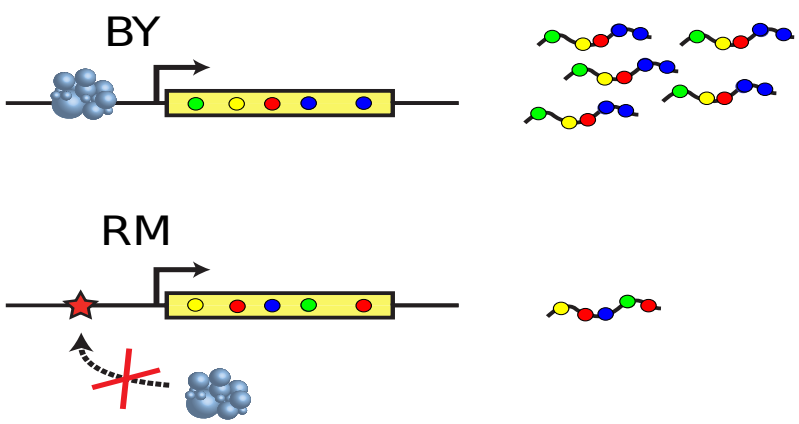


Figure 1 : In the top figure the transcription factor (blue) leads to high transcription. In the bottom figure an upstream polymorphism (red star) prevents the transcription factor from binding.

Specifics of ASE Experiment

Details of the data:

- Two "individuals" from genetically divergent yeast strains, BY and RM, are mated to produce a diploid hybrid.
- Three replicate experiments: same individuals, but separate samples of cells.
- Two technologies: Illumina and ABI SOLiD.
- Each of a few trillion cells are processed.
- Pre- and post-processing steps are followed by fragmentation to give millions of 200–400 base pair long molecules, with short reads obtained by sequencing.
- Need SNPs since otherwise the reference sequence is identical and so we cannot tell which strain the read arises from.
- Strict criteria to call each read as a match are used, to reduce read-mapping bias.
- Data from 25,652 SNPs within 4,844 genes.
- More details in Skelly *et al.* (2011).

Simple Approach to Testing for ASE

- Let N be the total number of counts at a particular gene, and Y the number of reads to the BY strain.
- Let θ be the probability of a map to BY.
- A simple approach is to assume:

$$Y|\theta \sim \text{Binomial}(N, \theta),$$

  and carry out a test of $H_0 : \theta = 0.5$, which corresponds to no allele specific expression.
- A non-Bayesian approach would use an exact test, i.e. enumerate the probabaility, under the null, of all the outcomes that are equal to or more extreme than that observed.
- Issues:
  - p-values are not uniform under the null due to discreteness of Y.
  - How to pick a threshold? In general and when there are multiple tests.
  - Do we really want a point null, i.e. $\theta = 0.5$?

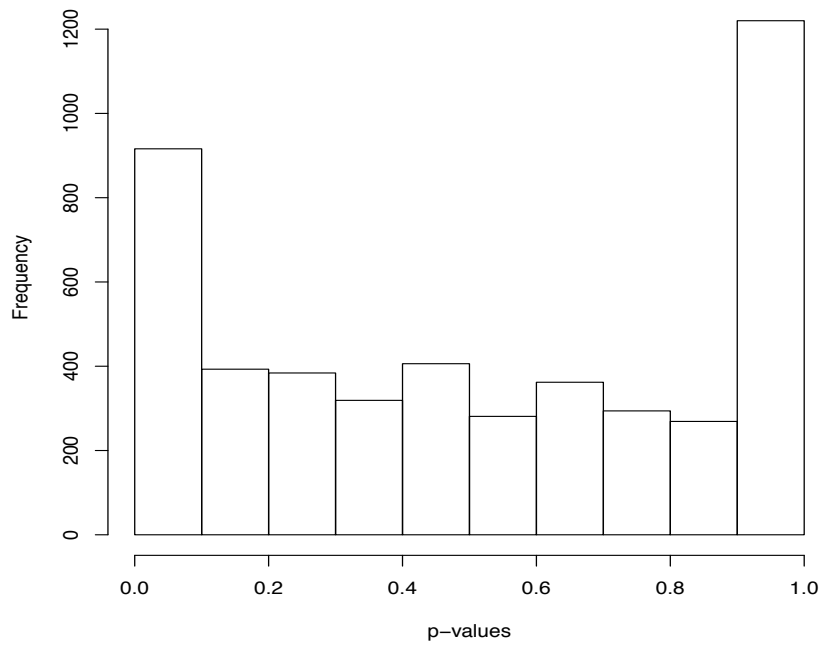


Figure 2 : p-values from 4,844 exact tests.

Bayes Theorem Recap

- We derive the posterior distribution via Bayes theorem:

$$p(\theta|y) = \frac{\Pr(y|\theta) \times p(\theta)}{\Pr(y)}.$$

- The denominator:

$$\Pr(y) = \int \Pr(y|\theta) \times p(\theta) d\theta$$

  is a normalizing constant to ensure the RHS integrates to 1.

- More colloquially:

$$\begin{aligned} \text{Posterior} \quad &\propto \quad \text{Likelihood} \times \text{Prior} \\ &= \quad \Pr(y|\theta) \times p(\theta) \end{aligned}$$

  since in considering the posterior we only need to worry about terms that depend on the parameter θ.

Overview of Bayesian Inference

- Simply put, to carry out a Bayesian analysis one must specify a likelihood (probability distribution for the data) and a prior (beliefs about the parameters of the model).
- The approach is therefore model-based, in contrast to approaches in which only the mean and the variance of the data are specified (e.g. weighted least squares).
- To carry out inference, integration is required, and a large fraction of the Bayesian research literature focusses on this aspect.
- Bayesian summaries:
  1. Estimation: marginal posterior distributions on parameters of interest.
  2. Hypothesis Testing: Bayes factors giving the evidence in the data with respect to two or more hypotheses.
  3. Prediction: via the predictive distribution.
- These three objectives will now be described in the context of a binomial model.

Elements of Bayes Theorem for a Binomial Model

- We assume independent responses with a common "success" probability θ.
- In this case, the contribution of the data is through the binomial probability distribution:

$$\Pr(Y = y|\theta) = \binom{N}{y} \theta^y (1-\theta)^{N-y} \tag{1}$$

  and tells us the probability of seeing $Y = y$, $y = 0, 1, ..., N$ given the probability θ.
- For fixed y, we may view (1) as a function of θ – this is the likelihood function.
- The maximum likelihood estimate (MLE) is that value

$$\widehat{\theta} = y/n$$

  that gives the highest probability to the observed data, i.e. maximizes the likelihood function.

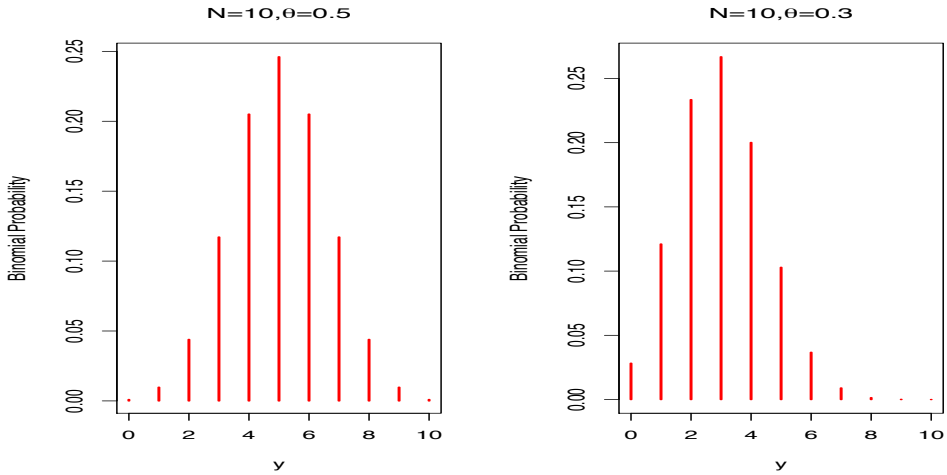


Figure 3 : Binomial distributions for two values of θ with $N = 10$.

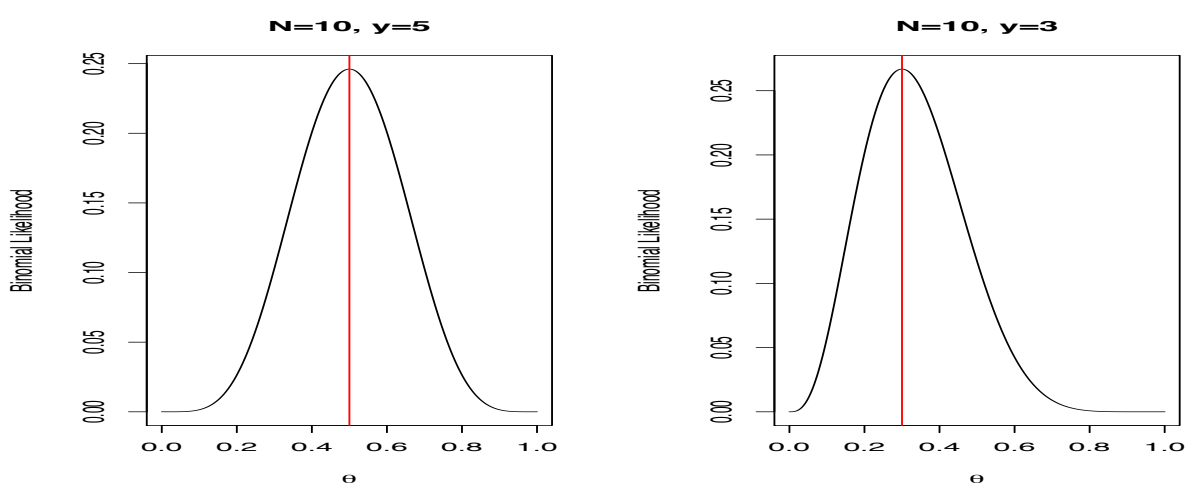


Figure 4 : Binomial likelihoods for values of $y = 5$ (left) and $y = 10$ (right), with $N = 10$. The MLEs are indicated in red.

The Beta Distribution as a Prior Choice for a Binomial θ

- Bayes theorem requires the likelihood, which we have already specified as binomial, and the prior.
- For a probability $0 < \theta < 1$ an obvious candidate prior is the uniform distribution on (0,1): but this is too restrictive in general.
- The beta distribution, beta(a, b), is more flexible and so may be used for θ, with a and b specified in advance. The uniform distribution is a special case with $a = b = 1$.
- The form of the beta distribution is

$$p(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}$$

  for $0 < \theta < 1$, where $\Gamma(\cdot)$ is the gamma function[1].
- The distribution is valid[2] for $a > 0, b > 0$.

[1] $\Gamma(z) = \int_0^\infty t^{z-1}\mathrm{e}^{-t}dt$
[2] A distribution is valid if it is non-negative and integrates to 1

The Beta Distribution as a Prior Choice for a Binomial θ

- How can we think about specifying a and b?
- For the normal distribution the parameters μ and σ^2 are just the mean and variance, but for the beta distribution a and b have no such simple interpretation.
- The mean and variance are:

$$\begin{aligned} \mathrm{E}[\theta] &= \frac{a}{a+b} \\ \mathrm{var}(\theta) &= \frac{\mathrm{E}[\theta](1-\mathrm{E}[\theta])}{a+b+1}. \end{aligned}$$

  Hence, increasing a and/or b concentrates the distribution about the mean.
- The quantiles, e.g. the median or the 10% and 90% points, are not available as a simple formula, but are easily obtained within software such as R using the function `qbeta(p,a,b)`.

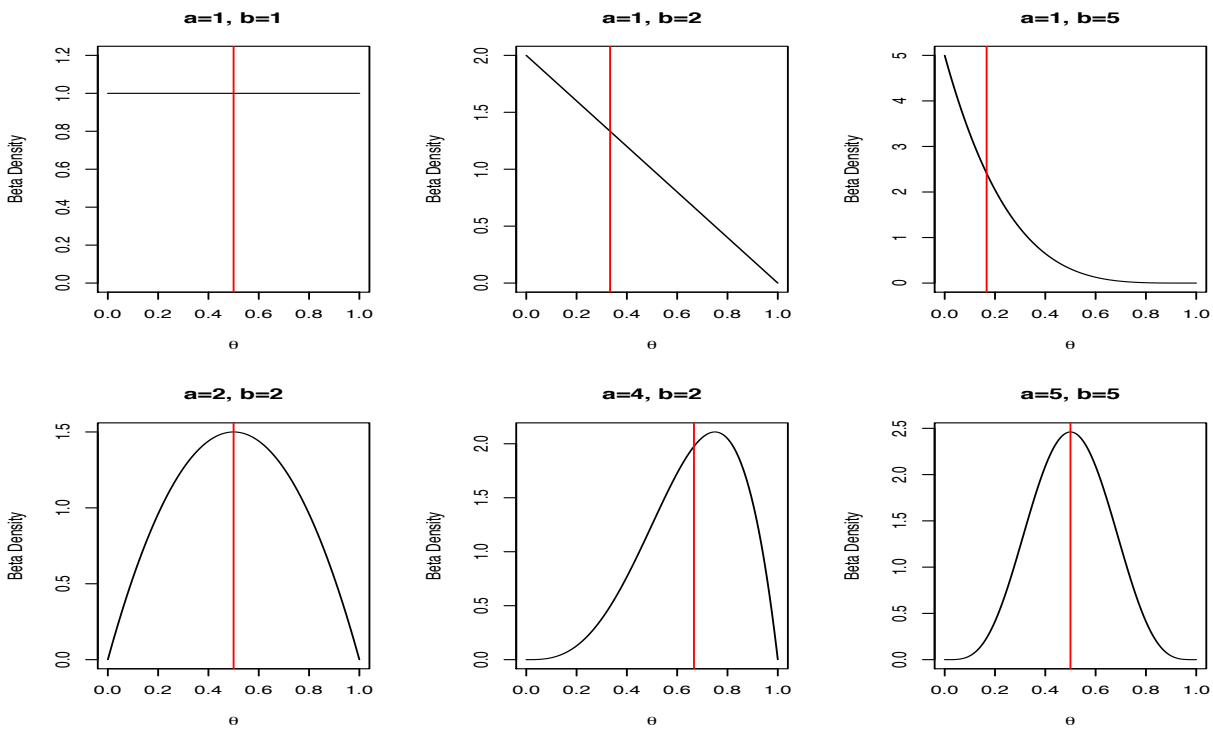


Figure 5 : Beta distributions, beta(a, b), the red lines indicate the means.

Samples to Summarize Beta Distributions

- Probability distributions can be investigated by generating samples and then examining histograms, moments and quantiles.

```
#
# First look at the theoretical quantiles of a uniform, a beta(1,1)
#
> qbeta(p=c(0.05,.1,.5,.9,.95),1,1)
[1] 0.05 0.10 0.50 0.90 0.95
#
# Now find the mean and quantiles from a large sample from a uniform
#
> nsim <- 5000
> samp <- rbeta(nsim,1,1)
> mean(samp)
[1] 0.504371
> quantile(samp,p=c(0.05,.1,.5,.9,.95))
        5%        10%        50%        90%        95%
0.04857267 0.10749531 0.50531835 0.90295985 0.95282366
#
# These differ slightly from the theoretical quantiles because of
# sampling variability
#
```

Samples to Summarize Beta Distributions

- In Figure 6 we show histograms of beta distributions for different choices of a and b.
- The code below creates the first and fifth plots on the figure.

```
#
# Now we will examine a histogram representation of a
# uniform distribution
#
> hist(samp,xlab=expression(theta),ylab="Beta Density",
      main="a=1, b=1",freq=F,nclass=10)
> abline(v=mean(samp),col="red")
#
# Now we do the same for a beta(4,2) distribution
#
> qbeta(p=c(0.05,.1,.5,.9,.95),4,2)
[1] 0.3425917 0.4161096 0.6861898 0.8877650 0.9235596
> samp <- rbeta(nsim,4,2)
> mean(samp)
[1] 0.6654911
> quantile(samp,p=c(0.05,.1,.5,.9,.95))
      5%      10%      50%      90%      95%
0.3394967 0.4096720 0.6838093 0.8842126 0.9217181
> hist(samp,xlab=expression(theta),ylab="Beta Density",
      main="a=4, b=2",freq=F,nclass=10)
> abline(v=mean(samp),col="red")
```

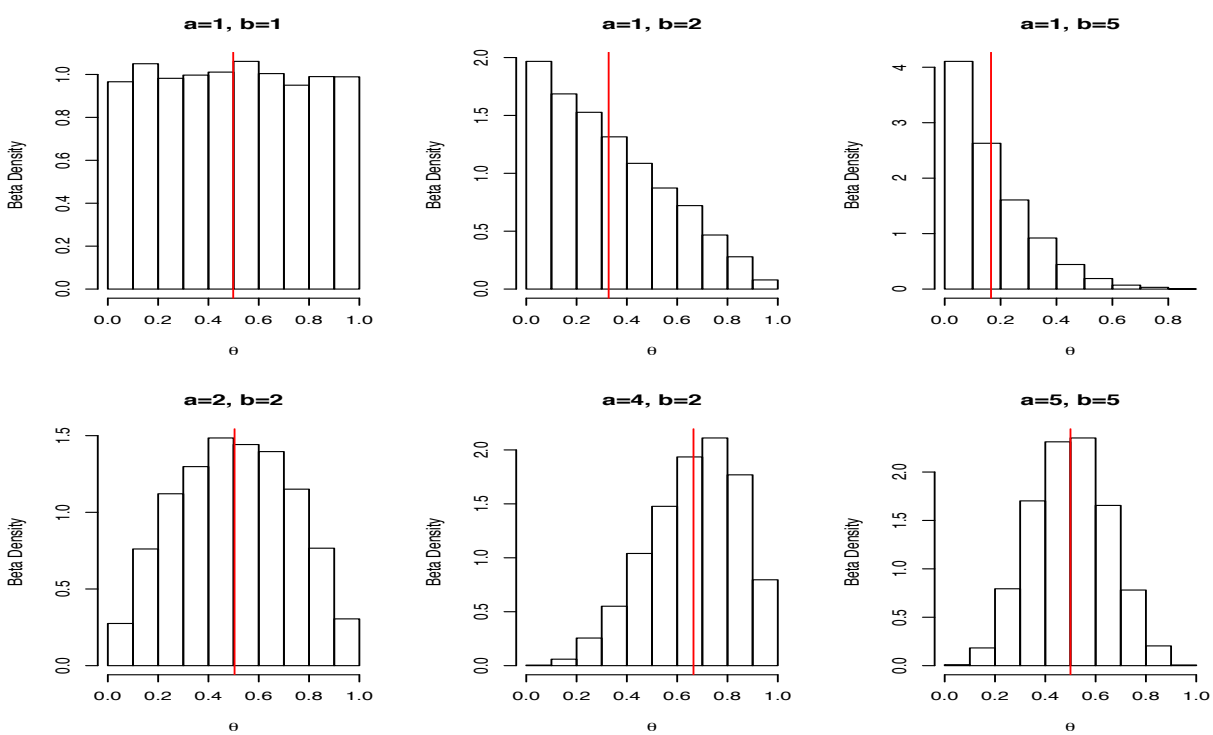



Figure 6 : Random samples from beta distributions; sample means as red lines.

Samples for Describing Weird Parameters

- So far the samples we have generated have produced summaries we can easily obtain anyway.
- But what about functions of the probability θ, such as the odds $\theta/(1-\theta)$?
- Once we have samples for θ we can simply transform the samples to the functions of interest.
- We may have clearer prior opinions about the odds, than the probability.
- The code below displays a histogram representation of the prior on the odds $\theta/(1-\theta)$ when θ is beta(10,10).

```
> nsim <- 5000
> samp <- rbeta(nsim,10,10)
> odds <- samp/(1-samp)
> hist(odds,xlab="Odds",
   main=expression(paste("Odds with ",theta," from a beta(10,10)")))
> abline(v=mean(odds),col="red")
```

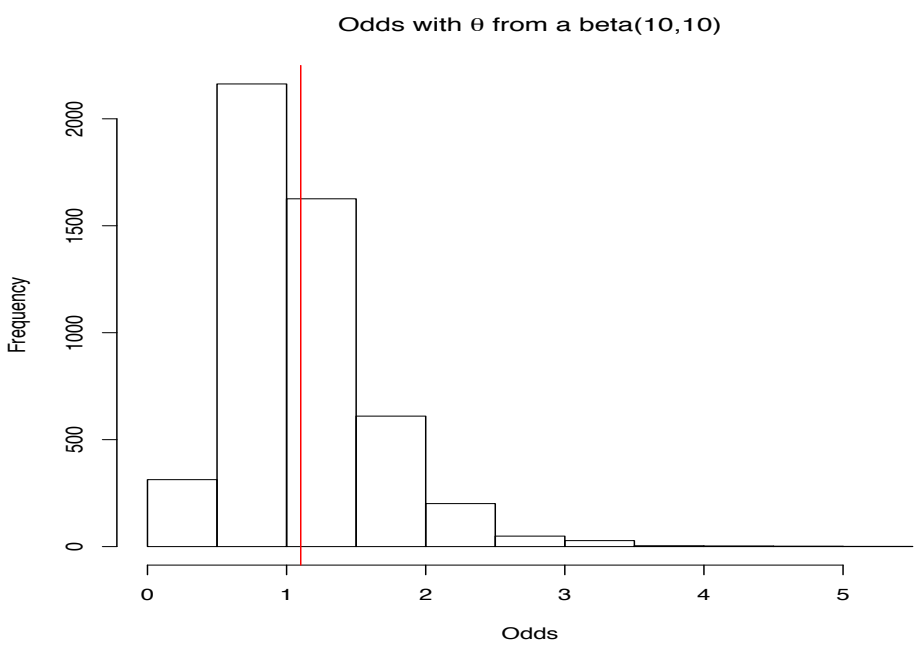



Figure 7 : Samples from the prior on the odds $\theta/(1-\theta)$ with $\theta \sim$ beta$(10, 10)$, the red line indicates the sample mean.

Are Priors Really Uniform?

- We might think that if we have little prior opinion about a parameter then we can simply assign a uniform prior, i.e. a prior

$$p(\theta) \propto \text{ const}$$

- There are two problems with this strategy:
  - we can't be uniform on all scales and,
  - if the parameter is not on a finite range, an improper distribution will result (that is, the form will not integrate to 1). This can lead to an improper posterior distribution, and without a proper posterior we can't do inference.

Are Priors Really Uniform?

- We illustate the first (non-uniform on all scales) point.
- In the binomial example a uniform prior for θ seems a natural choice.
- But suppose we are going to model on the logistic scale so that

$$\phi = \log\left(\frac{\theta}{1-\theta}\right)$$

  is a quantity of interest.
- A uniform prior on θ produces the very non-uniform distribution on ϕ in Figure 8.

```
> nsim <- 5000
> theta <- rbeta(nsim,1,1)
> phi <- log(theta/(1-theta))
> hist(phi,xlab=expression(paste("Log Odds ",phi)),nclass=30,
  main=expression(paste("Log Odds with ",theta," from a beta(1,1)")))
> abline(v=0,col="red")
```

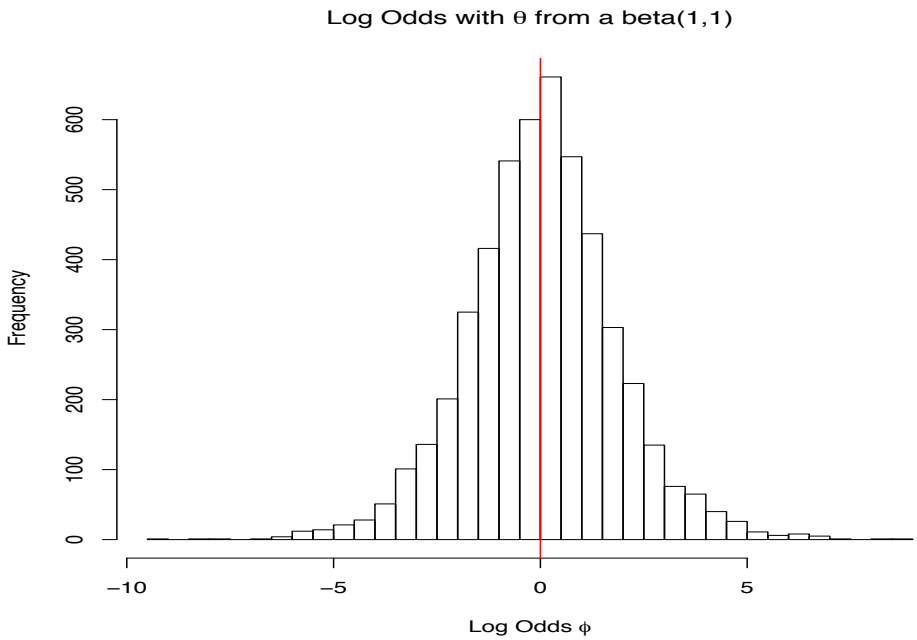


Figure 8 : Samples from the prior on the odds $\phi = \log[\theta/(1-\theta)]$ with $\theta \sim \text{beta}(1,1)$, the red line indicates the sample mean.

Posterior Derivation: The Quick Way

- When we want to identify a particular probability distribution we only need to concentrate on terms that involve the random variable.
- For example, if the random variable is x and we see a density of the form

$$p(x) \propto \exp\left(c_1 x^2 + c_2 x\right),$$

for constants c_1 and c_2, then we know x must have a normal distribution.

Posterior Derivation: The Quick Way

- For the binomial-beta model we concentrate on terms that only involve θ.
- The posterior is

$$\begin{aligned} p(\theta|y) &\propto \Pr(y|\theta) \times p(\theta) \\ &= \theta^y (1-\theta)^{N-y} \times \theta^{a-1}(1-\theta)^{b-1} \\ &= \theta^{y+a-1}(1-\theta)^{N-y+b-1} \end{aligned}$$

- We recognize this as the important part of a beta$(y+a, N-y+b)$ distribution.
- We know what the normalizing constant must be, because we have a distribution which must integrate to 1.

Posterior Derivation: The Long (and Unnecessary) Way

- The posterior can also be calculated by keeping in all the normalizing constants:

$$\begin{aligned} p(\theta|y) &= \frac{\Pr(y|\theta) \times p(\theta)}{\Pr(y)} \\ &= \frac{1}{\Pr(y)} \binom{N}{y} \theta^y (1-\theta)^{N-y} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1}(1-\theta)^{b-1}. \quad (2) \end{aligned}$$

- The normalizing constant is

$$\begin{aligned} \Pr(y) &= \int_0^1 \Pr(y|\theta) \times p(\theta) d\theta \\ &= \binom{N}{y} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 \theta^{y+a-1}(1-\theta)^{N-y+b-1} d\theta \\ &= \binom{N}{y} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(y+a)\Gamma(N-y+b)}{\Gamma(N+a+b)} \end{aligned}$$

- The integrand on line 2 is a beta$(y+a, N-y+b)$ distribution, up to a normalizing constant, and so we know what this constant has to be.

Posterior Derivation: The Long (and Unnecessary) Way

- The normalizing constant is therefore:

$$\Pr(y) = \binom{N}{y} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(y+a)\Gamma(N-y+b)}{\Gamma(N+a+b)}$$

- This is a probability distribution, i.e. $\sum_{y=0}^{N} \Pr(y) = 1$ with $\Pr(y) > 0$.
- For a particular y value, this expression tells us the probability of that value given the model, i.e. the likelihood and prior we have selected: this will reappear later in the context of hypothesis testing.
- Substitution of $\Pr(y)$ into (2) and canceling the terms that appear in the numerator and denominator gives the posterior:

$$p(\theta|y) = \frac{\Gamma(N+a+b)}{\Gamma(y+a)\Gamma(N-y+b)} \theta^{y+a-1}(1-\theta)^{N-y+b-1}$$

which is a beta$(y+a, N-y+b)$.

The Posterior Mean: A Summary of the Posterior

- Recall the mean of a beta(a, b) is $a/(a+b)$.
- The posterior mean of a beta$(y+a, N-y+b)$ is therefore

$$\begin{aligned} \mathrm{E}[\theta|y] &= \frac{y+a}{N+a+b} \\ &= \frac{y}{N+a+b} + \frac{a}{N+a+b} \\ &= \frac{y}{N} \times \frac{N}{N+a+b} + \frac{a}{a+b} \times \frac{a+b}{N+a+b} \\ &= \text{MLE} \times \text{W} + \text{Prior Mean} \times (1\text{-W}). \end{aligned}$$

- The weight W is

$$\text{W} = \frac{N}{N+a+b}.$$

- As N increases, the weight tends to 1, so that the posterior mean gets closer and closer to the MLE.
- Notice that the uniform prior $a = b = 1$ gives a posterior mean of

$$\mathrm{E}[\theta|y] = \frac{y+1}{N+2}.$$

The Posterior Mode

- First, note that the mode of a beta(a, b) is

$$\text{mode}(\theta) = \frac{a-1}{a+b-2}.$$

- As with the posterior mean, the posterior mode takes a weighted form:

$$\begin{aligned}
\text{mode}(\theta|y) &= \frac{y+a-1}{N+a+b-2} \\
&= \frac{y}{N} \times \frac{N}{N+a+b-2} + \frac{a-1}{a+b-2} \times \frac{a+b-2}{N+a+b-2} \\
&= \text{MLE} \times \text{W}^\star + \text{Prior Mode} \times (1\text{-W}^\star).
\end{aligned}$$

- The weight $\text{W}^\star$ is

$$\text{W}^\star = \frac{N}{N+a+b-2}.$$

- Notice that the uniform prior $a = b = 1$ gives a posterior mode of

$$\text{mode}(\theta|y) = \frac{y}{N},$$

the MLE. Which makes sense, right?

Other Posterior Summaries

- We will rarely want to report a point estimate alone, whether it be a posterior mean or posterior median.
- Interval estimates are obtained in the obvious way.
- A simple way of performing testing of particular parameter values of interest is via examination of interval estimates.
- For example, does a 95% interval contain the value $\theta_0 = 0$?

Other Posterior Summaries

- In our beta-binomial running example, a 90% posterior credible interval (θ_l, θ_u) results from the points

$$
\begin{aligned}
0.05 &= \int_0^{\theta_l} p(\theta|y)\ d\theta \\
0.95 &= \int_0^{\theta_u} p(\theta|y)\ d\theta
\end{aligned}
$$

- The quantiles of a beta are not available in closed form, but easy to evaluate in R:

```
y <- 7; N <- 10; a <- b <- 1
qbeta(c(0.05,0.5,0.95),y+a,N-y+1)
[1] 0.4356258 0.6761955 0.8649245
```

- The 90% credible interval is (0.44,0.86) and the posterior median is 0.68.

Prior Sensitivity

- For small datasets in particular it is a good idea to examine the sensitivity of inference to the prior choice, particularly for those parameters for which there is little information in the data.
- An obvious way to determine the latter is to compare the prior with the posterior, but experience often aids the process.
- Sometimes one may specify a prior that reduces the impact of the prior.
- In some situations, priors can be found that produce point and interval estimates that mimic a standard non-Bayesian analysis, i.e. have good frequentist properties.
- Such priors provide a baseline to compare analyses with more substantive priors.
- Other names for such priors are objective, reference and non-subjective.
- We now describe another approach to specification, via subjective priors.

Choosing a Prior, Approach One

- To select a beta, we need to specify two quantities, a and b.
- The posterior mean is
$$\mathrm{E}[\theta|y] = \frac{y+a}{N+a+b}.$$
- Viewing the denominator as a sample size suggests a method for choosing a and b within the prior.
- We need to specify two numbers, but rather than a and b, which are difficult to interpret, we may specify the mean $m_{\text{prior}} = a/(a+b)$ and the prior sample size $N_{\text{prior}} = a+b$
- We then solve for a and b via
$$\begin{aligned} a &= N_{\text{prior}} \times m_{\text{prior}} \\ b &= N_{\text{prior}} \times (1 - m_{\text{prior}}). \end{aligned}$$
- Intuition: a is like a prior number of successes and b like the prior number of failures.

Choosing a Prior, Approach One

An Example:

- Suppose we set $N_{\text{prior}} = 5$ and $m_{\text{prior}} = \frac{2}{5}$. It is as if we saw 2 successes out of 5.
- Suppose we obtain data with $N = 10$ and $\frac{y}{N} = \frac{7}{10}$.
- Hence $\mathrm{W} = 10/(10+5)$ and
$$\begin{aligned} \mathrm{E}[\theta|y] &= \frac{7}{10} \times \frac{10}{10+5} + \frac{2}{5} \times \frac{5}{10+5} \\ &= \frac{9}{15} = \frac{3}{5}. \end{aligned}$$
- Solving:
$$\begin{aligned} a &= N_{\text{prior}} \times m_{\text{prior}} = 5 \times \frac{2}{5} = 2 \\ b &= N_{\text{prior}} \times (1 - m_{\text{prior}}) = 5 \times \frac{3}{5} = 3 \end{aligned}$$
- This gives a $\text{beta}(y+a, N-y+b) = \text{beta}(7+2, 3+3)$ posterior.

Beta Prior, Likelihood and Posterior

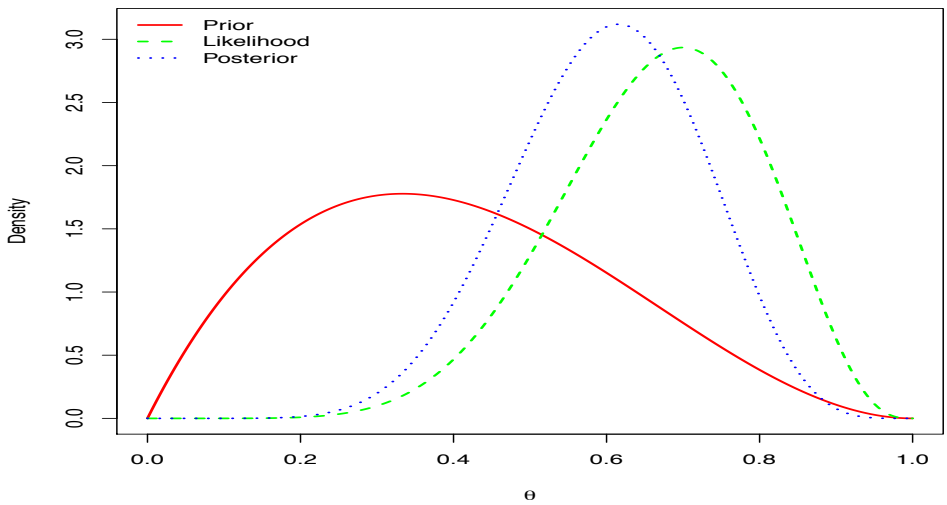


Figure 9 : The prior is beta(2,3) the likelihood is proportional to a binomial(7,3) and the posterior is beta(7+2,3+3).

Beta Prior, Likelihood and Posterior: R code

- The code below produces Figure 9.

```
> a <- 2
> b <- 3
> N <- 10
> y <- 7
> thetaseq <- seq(0,1,.001)
> prior <- dbeta(thetaseq,a,b)
> likelihood <- dbeta(thetaseq,y+1,N-y+1)
> posterior <- dbeta(thetaseq,a+y,b+N-y)
> par(mfrow=c(1,1))
> plot(posterior~thetaseq,xlab=expression(theta),type="n",
    ylab="Density")
> lines(prior~thetaseq,type="l",col="red",lwd=2,lty=1)
> lines(likelihood~thetaseq,type="l",col="green",lwd=2,lty=2)
> lines(posterior~thetaseq,type="l",col="blue",lwd=2,lty=3)
> legend("topleft",legend=c("Prior","Likelihood","Posterior"),
    col=c("red","green","blue"),lwd=2,bty="n",lty=1:3)
```

Choosing a Prior, Approach Two

- An alternative convenient way of choosing a and b is by specifying two quantiles for θ with associated (prior) probabilities.
- For example, we may wish $\Pr(\theta < 0.1) = 0.05$ and $\Pr(\theta > 0.6) = 0.05$.
- The values of a and b may be found numerically.
- For example, we may solve

$$[p_1 - \Pr(\theta < q_1|a, b)]^2 + [p_2 - \Pr(\theta < q_2|a, b)]^2 = 0 \tag{3}$$

for a, b.

R code for Beta Prior Specification

- Example: The R code below finds the beta distribution with 5% and 95% points of 0.1 and 0.6. Running the code gives $a = 2.73$ and $b = 5.67$.
- The `optim` function produces the solution to (3).

```
# Function to find a and b
priorch <- function(x,q1,q2,p1,p2){
(p1-pbeta(q1,x[1],x[2]))^2 + (p2-pbeta(q2,x[1],x[2]))^2 }
#
> p1 <- 0.05
> p2 <- 0.95
> q1 <- 0.1
> q2 <- 0.6
> opt <- optim(par=c(1,1),fn=priorch,q1=q1,q2=q2,p1=p1,p2=p2,
          control=list(abstol=1e-8))
> cat("a and b are ",opt$par,"\n")
a and b are  2.73 5.67
> probvals <- seq(0,1,.001)
> plot(probvals,dbeta(probvals,shape1=opt$par[1],shape2=opt$par[2]),
     type="l", xlab=expression(theta),ylab="Beta Density")
> abline(v=q1,col="red")
> abline(v=q2,col="red")
```

Beta Prior Choice via Quantile Specification

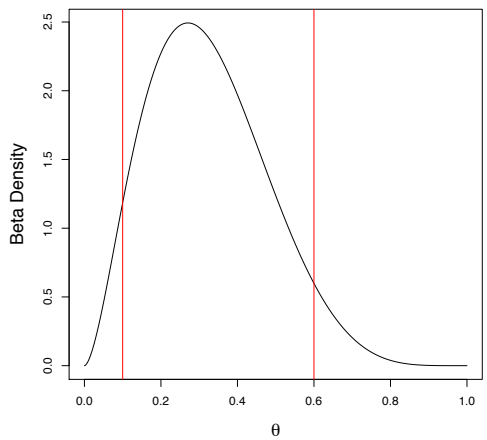


Figure 10 : beta(2.73,5.67) prior with 5% and 95% quantiles highlighted.

Bayesian Sequential Updating

- We show how probabilistic beliefs are updated as we receive more data.
- Suppose the data arrives sequentially via two experiments:
  1. Experiment 1: (y_1, N_1).
  2. Experiment 2: (y_2, N_2).
- Prior 1: $\theta \sim \text{beta}(a, b)$.
- Likelihood 1: $y_1|\theta \sim \text{binomial}(N_1, \theta)$.
- Posterior 1: $\theta|y_1 \sim \text{beta}(a + y_1, b + N_1 - y_1)$.
- This posterior forms the prior for experiment 2.
- Prior 2: $\theta \sim \text{beta}(a^\star, b^\star)$ where $a^\star = a + y_1$, $b^\star = b + N_1 - y_1$.
- Likelihood 2: $y_2|\theta \sim \text{binomial}(N_2, \theta)$.
- Posterior 2: $\theta|y_1, y_2 \sim \text{beta}(a^\star + y_2, b^\star + N_2 - y_2)$.
- Substituting for $a^\star, b^\star$:

$$\theta|y_1, y_2 \sim \text{beta}(a + y_1 + y_2, b + N_1 - y_1 + N_2 - y_2).$$

Bayesian Sequential Updating

- Schematically:

$$(a, b) \to (a + y_1, b + N_1 - y_1) \to (a + y_1 + y_2, b + N_1 - y_1 + N_2 - y_2)$$

- Suppose we obtain the data in one go as $y^\star = y_1 + y_2$ successes from $N^\star = N_1 + N_2$ trials.
- The posterior is

$$\theta | y^\star \sim \text{beta}(a + y^\star, b + N^\star - y^\star),$$

which is the same as when we receive in two separate instances.

Predictive Distribution

- Suppose we see y successes out of N trials, and now wish to obtain a predictive distribution for a future experiment with M trials.
- Let $Z = 0, 1, ..., M$ be the number of successes.
- Predictive distribution:

$$\begin{aligned} \Pr(z|y) &= \int_0^1 p(z, \theta|y) d\theta \\ &= \int_0^1 \Pr(z|\theta, y) p(\theta|y) d\theta \\ &= \int_0^1 \Pr(z|\theta) p(\theta|y) d\theta \end{aligned}$$

because of conditional independence.

Predictive Distribution

- Continuing with the calculation:

$$
\begin{aligned}
\Pr(z|y) &= \int_0^1 \Pr(z|\theta) \times p(\theta|y) d\theta \\
&= \int_0^1 \binom{M}{z} \theta^z (1-\theta)^{M-z} \\
&\times \frac{\Gamma(N+a+b)}{\Gamma(y+a)\Gamma(N-y+b)} \theta^{y+a-1}(1-\theta)^{N-y+b-1} d\theta \\
&= \binom{M}{z} \frac{\Gamma(N+a+b)}{\Gamma(y+a)\Gamma(N-y+b)} \int_0^1 \theta^{y+a+z-1}(1-\theta)^{N-y+b+M-z-1} d\theta \\
&= \binom{M}{z} \frac{\Gamma(N+a+b)}{\Gamma(y+a)\Gamma(N-y+b)} \frac{\Gamma(a+y+z)\Gamma(b+N-y+M-z)}{\Gamma(a+b+N+M)}
\end{aligned}
$$

for $z = 0, 1, ..., M$.

- A likelihood approach would take the predictive distribution as binomial$(M, \widehat{\theta})$ with $\widehat{\theta} = y/N$.

R Code for Predictive Predictions

```
binomialpred <- function(a,b,y,N,z,M){
 lchoose(M,z) + lgamma(a+b+N) - lgamma(a+y) - lgamma(b+N-y) +
                lgamma(a+y+z) + lgamma(b+N-y+M-z) - lgamma(a+b+N+M)
}
> a <- b <- 1
> y <- 2
> N <- 20
> M <- 10
> binpred <- NULL
> z <- seq(0,M)
> sumcheck <- 0
> for (i in 1:(M+1)){
    binpred[i] <- exp(binomialpred(a,b,y,N,z[i],M))
    sumcheck <- sumcheck + binpred[i]
}
> likpred <- dbinom(z,M,prob=y/N)
> cat("Sum of probs = ",sumcheck,"\n")
> plot(binpred~z,type="h",col="red",ylim=c(0,max(likpred,binpred)),
     ylab="Predictive Distribution")
> points(z+.2,likpred,type="h",col="blue",lty=2)
> legend("topright",legend=c("Likelihood Prediction",
        "Bayesian Prediction"),lty=2:1,col=c("blue","red"),bty="n")
```

Predictive Distribution

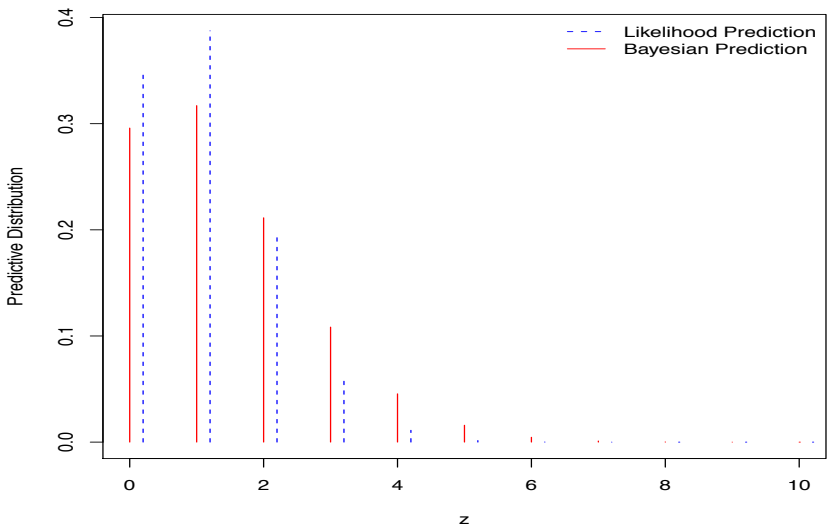


Figure 11 : Likelihood and Bayesian predictive distribution of seeing $z = 0, 1, \ldots, M = 10$ successes, after observing $y = 2$ out of $N = 20$ successes (with $a = b = 1$).

Predictive Distribution

- The posterior and sampling distributions won't usually combine so conveniently.
- In general, we may form a Monte Carlo estimate of the predictive distribution:

$$\begin{aligned} p(z|y) &= \int p(z|\theta)p(\theta|y)d\theta \\ &= \mathrm{E}_{\theta|y}[p(z|\theta)] \\ &\approx \frac{1}{S}\sum_{s=1}^{S} p(z|\theta^{(s)}) \end{aligned}$$

  where $\theta^{(s)} \sim p(\theta|y)$, $s = 1, ..., S$, is a sample from the posterior.
- This provides an estimate of the distribution at the point z.
- Alternatively, we may sample from $p(z|\theta^{(s)})$ a large number of times to reconstruct the predictive distribution.

Difference in Binomial Proportions

- It is straightforward to extend the methods presented for a single binomial sample to a pair of samples.
- Suppose we carry out two binomial experiments:

$$\begin{aligned} Y_1|\theta_1 &\sim \text{binomial}(N_1, \theta_1) \quad \text{for sample 1} \\ Y_2|\theta_2 &\sim \text{binomial}(N_2, \theta_2) \quad \text{for sample 2} \end{aligned}$$

- Interest focuses on $\theta_1 - \theta_2$, and often in examing the possibitlity that $\theta_1 = \theta_2$.
- With a sampling-based methodology, and independent beta priors on θ_1 and θ_2, it is straightforward to examine the posterior $p(\theta_1 - \theta_1|y_1, y_2)$.

Difference in Binomial Proportions

- Savage *et al.* (2008) give data on allele frequencies within a gene that has been linked with skin cancer.
- It is interest to examine differences in allele frequencies between populations.
- We examine one SNP and extract data on Northern European (NE) and United States (US) populations.
- Let θ_1 and θ_2 be the allele frequencies in the NE and US population from which the samples were drawn, respectively.
- The allele frequencies were 10.69% and 13.21% with sample sizes of 650 and 265, in the NE and US samples, respectively.
- We assume independent beta(1,1) priors on each of θ_1 and θ_2.
- The posterior probability that $\theta_1 - \theta_2$ is greater than 0, is 0.12, so there is little evidence of a difference in allele frequencies between the NE and US samples.

Difference in Binomial Proportions

- These data were reconstructed from figures in the original paper (hence the `floor` function.

```
> N1 <- 650
> y1 <- floor(N1*.1069)
> N2 <- 265
> y2 <- floor(N2*.1321)
> nsamp <- 10000
> a <- b <- 1
> post1 <- rbeta(nsamp,y1+a,N1-y1+b)
> post2 <- rbeta(nsamp,y2+a,N2-y2+b)
> par(mfrow=c(1,3))
> hist(post1,xlab=expression(theta[1]),main="",cex.lab=1.5)
> hist(post2,xlab=expression(theta[2]),main="",cex.lab=1.5)
#
# This is the key step: constructing a sample estimate of the
# difference in probabilities
#
> hist(post1-post2,xlab=expression(paste(theta[1],"-",
    theta[2])),main="",cex.lab=1.5)
> abline(v=0,col="red")
> sum(post1-post2>0)/nsamp
[1] 0.1248
```


Binomial Two Sample Example

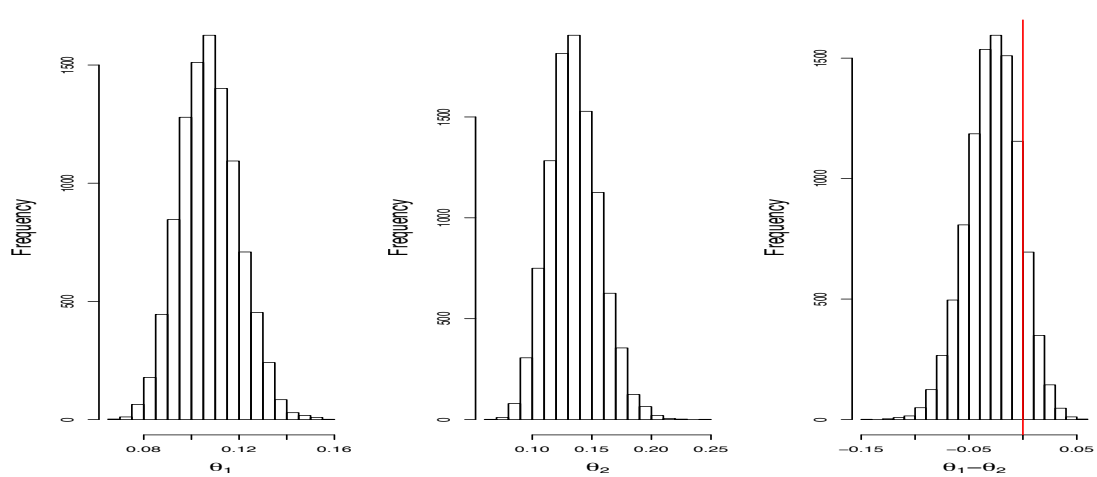


Figure 12 : Histogram representations of $p(\theta_1|y_1)$, $p(\theta_2|y_2)$ and $p(\theta_1 - \theta_2|y_1, y_2)$. The red line in the right plot is at the reference point of zero.

Bayes Factors for Hypothesis Testing

- The Bayes factor provides a summary of the evidence for a particular hypothesis (model) as compared to another.
- The Bayes factor is

$$\mathrm{BF} = \frac{\Pr(y|H_0)}{\Pr(y|H_1)}$$

  and so is simply the probability of the data under H_0 divided by the probability of the data under H_1.
- Values of BF > 1 favor H_0 while values of BF < 1 favor H_1.
- Note the similarity to the likelihood ratio

$$\mathrm{LR} = \frac{\Pr(y|H_0)}{\Pr(y|\widehat{\theta})}$$

  where $\widehat{\theta}$ is the MLE under H_1.
- If there are no unknown parameters in H_0 and H_1 (for example, $H_0 : \theta = 0.5$ versus $H_1 : \theta = 0.3$), then the Bayes factor is identical to the likelihood ratio.

Calibration of Bayes Factors

- Kass and Raftery (1995) suggest intervals of Bayes factors for reporting:

1/Bayes Factor	Evidence Against H_0
1 to 3.2	Not worth more than a bare mention
3.2 to 20	Positive
20 to 150	Strong
>150	Very strong

- These provide a guideline, but should not be followed without question.

Bayes Factors for Binomial Data

An Example:

- For each gene in the ASE dataset we may be interested in $H_0 : \theta = 0.5$ versus $H_1 : \theta \neq 0.5$.
- The numerator and denominator of the Bayes factor are:

$$\begin{aligned}
\Pr(y|H_0) &= \binom{N}{y} 0.5^y 0.5^{N-y} \\
\Pr(y|H_1) &= \int_0^1 \binom{N}{y} \theta^y (1-\theta)^{N-y} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1}(1-\theta)^{b-1} d\theta \\
&= \binom{N}{y} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(y+a)\Gamma(N-y+b)}{\Gamma(N+a+b)}
\end{aligned}$$

- We have already seen the denominator calculation, when we normalized the posterior.

Values Taken by the Log Bayes Factor, as a Function of y

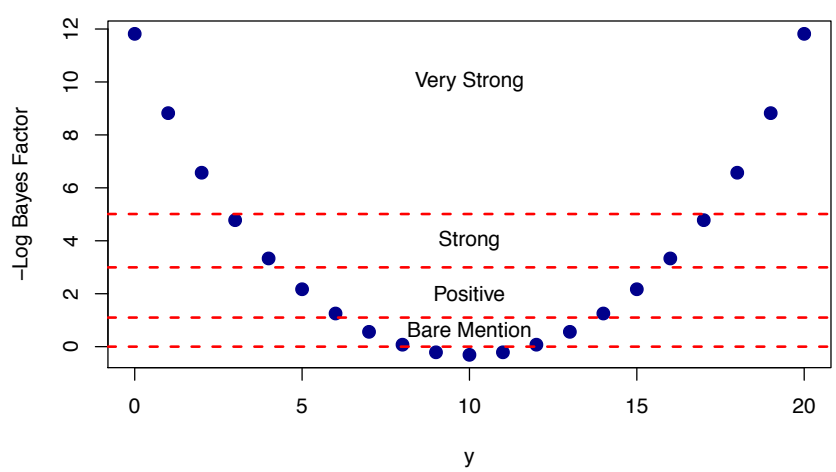


Figure 13 : Negative Log Bayes factor as a function of $y|\theta \sim \text{Binomial}(20, \theta)$ for $y = 0, 1, \ldots, 20$. High values indicate evidence against the null.

R Code for Bayes Factor Calculation

```
BFbinomial <- function(N,y,a,b,p0){
    logPrH0 <- lchoose(N,y) + y*log(p0) + (N-y)*log(1-p0)
    logPrH1 <- lchoose(N,y) + gamma(a+b) - lgamma(a) - lgamma(b) +
               lgamma(y+a) + lgamma(N-y+b) -lgamma(N+a+b)
    logBF <- logPrH0 - logPrH1
    list(logPrH0=logPrH0,logPrH1=logPrH1,logBF=logBF)
}
> N <- 20
> y <- seq(0,N,1)
> a <- b <- 1
> p0 <- 0.5
> logBFr <- NULL
> for (i in 1:(N+1)){
    BFcall <- BFbinomial(N,y[i],a,b,p0)
    logBFr[i] <- -BFcall$logBF # Take log of reciprocal (so evidence
                               # in favor of H1
}
> plot(logBFr~y,type="p",pch=20,col="darkblue",
    ylab="-Log Bayes Factor",cex=2)
> abline(h=log(150),col="red",lty=2,lwd=2); text(10,10,"Very Strong")
> abline(h=log(20),col="red",lty=2,lwd=2); text(10,4,"Strong")
> abline(h=log(3),col="red",lty=2,lwd=2); text(10,2,"Positive")
> abline(h=log(1),col="red",lty=2,lwd=2); text(10,0.65,"Bare Mention")
```


Bayesian Analysis of the ASE Data

Three approaches to inference:

1. Posterior Probabilities:
   - A simple approach to testing is to calculate the posterior probability that $\theta < 0.5$.
   - We can then pick a threshold for indicating worthy of further study, e.g. if $\Pr(\theta < 0.5|y) < 0.01$ or $\Pr(\theta < 0.5|y) < 0.99$
2. Bayes Factors:
   - Calculating the Bayes factor.
   - Pick a threshold for indicating worthy of further study, e.g. if the Bayes factor is greater than 150.
3. Decision theory:
   - Place priors on the null and alternative hypotheses.
   - Calculate the posterior odds:

$$\begin{aligned} \frac{\Pr(H_0|y)}{\Pr(H_1|y)} &= \frac{\Pr(y|H_0)}{\Pr(y|H_1)} \times \frac{\Pr(H_0)}{\Pr(H_1)} \\ \text{Posterior Odds} &= \text{Bayes Factor} \times \text{Prior Odds} \end{aligned}$$

   - Pick a threshold R, so that if the Posterior Odds $<$ R we choose H_1.

Bayesian Analysis of the ASE Data

- In Figure 14 we give a histogram of the posterior probabilities $\Pr(\theta < 0.5|y)$ and we see large numbers of genes have probabilities close to 0 and 1, indicating allele specific expression (ASE).
- In Figure 15 we plot $\Pr(\theta < 0.5|y)$ versus the p-values and the general pattern is what we would expect — small p-values have posterior probabilities close to 0 and 1.
- The strange lines in this plot are due to the discreteness of the outcome y.
- In Figure 16 we plot the -Log Bayes Factor against $\Pr(\theta < 0.5|y)$. Large values of the former correspond to strong evidence of ASE; again we see an aggreement in inference, with large values of the negative log Bayes factor corresponding with $\Pr(\theta < 0.5|y)$ close to 0 and 1.

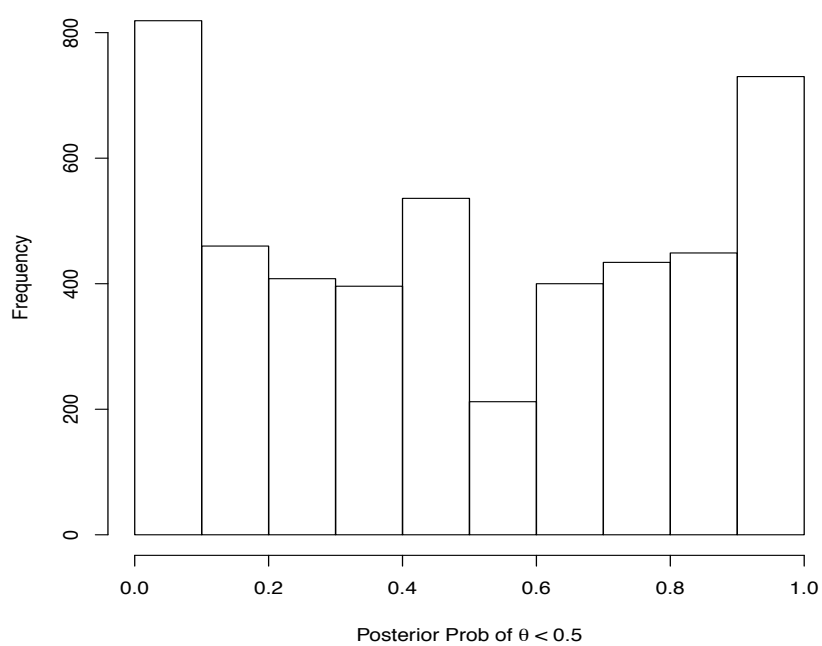


Figure 14 : Histogram of 4,844 posterior probabilities of $\theta < 0.5$.

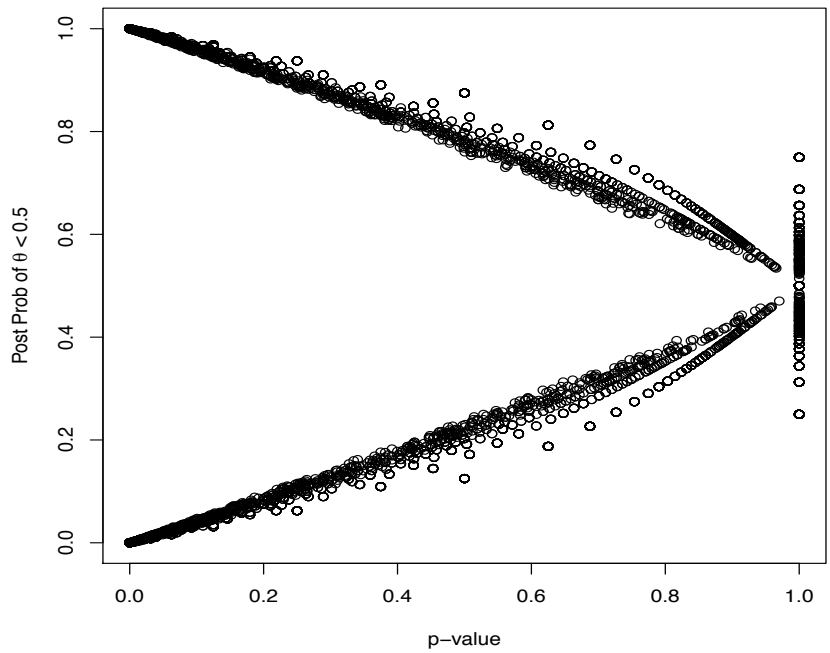


Figure 15 : Posterior probabilities of $\theta < 0.5$ and p-values from exact tests.

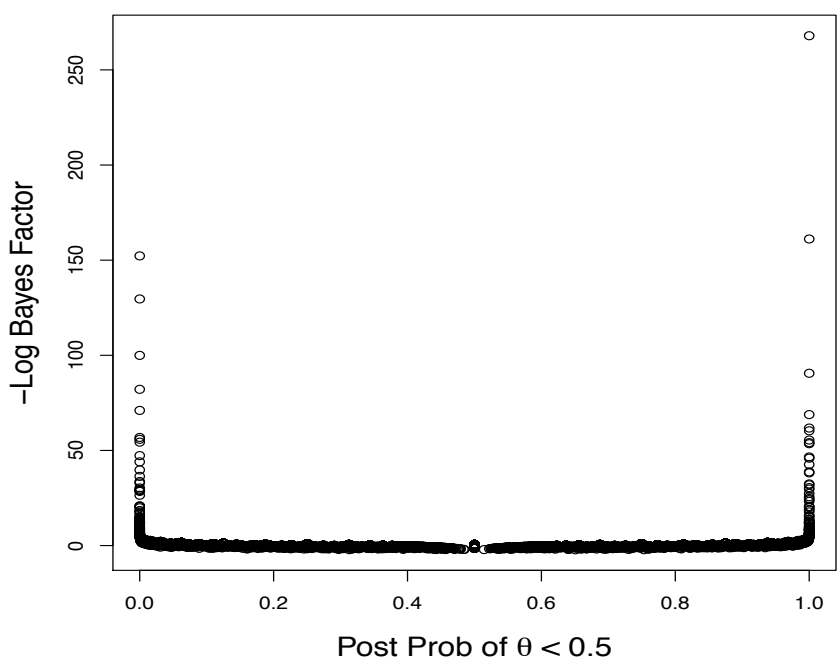


Figure 16 : Negative Log Bayes factor versus posterior probabilities of $\theta < 0.5$.

ASE Example

- Applying a Bonferroni correction to control the family wise error rate at 0.05, gives a p-value threshold of $0.05/4844 = 10^{-5}$ and 111 rejections. More on this later!
- There were 278 genes with $\Pr(\theta < 0.5|y) < 0.01$ and 242 genes with $\Pr(\theta < 0.5|y) > 0.99$.
- Following the guideline of requiring very strong evidence, there were 197 genes with the Bayes factor greater than 150.
- Requiring less stringent evidence, i.e. strong only, there were 359 genes.
- We consider a formal decision theory approach to testing in Lecture 8.

ASE Output Data

- Below are some summaries from the ASE analysis – we order with respect to the variable `logBFr`, which is the reciprocal Bayes factor (so that high numbers correspond to strong evidence against the null).
- The `postprob` variable is the posterior probability of $\theta < 0.5$.

```
> allvals <- data.frame(Nsum,ysum,pvals,postprob,logBFr)
> oBF <- order(-logBFr)
> orderallvals <- allvals[oBF,]
> head(orderallvals)
     Nsum ysum          pvals     postprob   logBFr
4751  437    6 5.340324e-119 1.000000e+00 267.9572
4041  625   97  1.112231e-72 1.000000e+00 161.1355
2370  546  468  8.994944e-69 2.621622e-69 152.2517
2770  256  245  1.127211e-58 2.943484e-59 129.6198
2291  150  150  1.401298e-45 3.503246e-46  99.9548
1328  228   19  1.224323e-41 1.000000e+00  90.5573
> tail(orderallvals)
     Nsum ysum      pvals  postprob     logBFr
824   761  382 0.9422103 0.4567334 -2.086604
2163  776  390 0.9142477 0.4429539 -2.091955
3153  769  384 1.0000000 0.5143722 -2.097079
2860 1076  546 0.6474878 0.3129473 -2.146555
2028 1440  707 0.5100331 0.7532969 -2.176356
395  1123  555 0.7202938 0.6508932 -2.211576
```

Conclusions

- Monte Carlo sampling provides flexibility of inference.
- All this lecture considered Binomial sampling, for which there is only a single parameter. For more parameters, prior specification and computing becomes more interesting...as we shall see.
- Multiple testing is considered in Lecture 8.
- For estimation and with middle to large sample sizes, conclusions from Bayesian and non-Bayesian approaches often coincide.
- For testing it is a different story, as discussed in Lecture 8.

Conclusions

Benefits of a Bayesian approach:

- Inference is based on probability and output is very intuitive.
- Framework is flexible, and so complex models can be built.
- Can incorporate prior knowledge!

Challenges of a Bayesian analysis:

- Require a likelihood and a prior, and inference is only as good as the appropriateness of these choices.
- Computation can be daunting, though software is becoming more user friendly and flexible (later we will use `INLA`).
- One should be wary of model becoming too complex – we have the technology to contemplate complicated models, but do the data support complexity?

References

Kass, R. and Raftery, A. (1995). Bayes factors. *Journal of the American Statistical Association*, **90**, 773–795.

Savage, S. A., Gerstenblith, M. R., Goldstein, A., Mirabello, L., Fargnoli, M. C., Peris, K., and Landi, M. T. (2008). Nucleotide diversity and population differentiation of the melanocortin 1 receptor gene, MC1R. *BMC Genetics*, **9**, 31.

Skelly, D., Johansson, M., Madeoy, J., Wakefield, J., and Akey, J. (2011). A powerful and flexible statistical framework for testing hypothesis of allele-specific gene expression from RNA-Seq data. *Genome Research*, **21**, 1728–1737.

2014 SISG Module 4: Bayesian Statistics for Genetics
Lecture 4: Multinomial Sampling

Jon Wakefield

Departments of Statistics and Biostatistics
University of Washington

Outline

Introduction and Motivating Examples

Bayesian Analysis of Multinomial Data
- Derivation of the Posterior and Prior Specfication
- Inference for Parameters of Interest

Analysis of HWE Data

Conclusions

Introduction

- In this lecture we will consider the Bayesian modeling of multinomial data.
- The examination of Hardy-Weinberg equilibrium will be used to motivate the multinomial model.
- Again, conjugate priors will be used, though sampling from the posterior will be emphasized as a method for flexible inference.
- Bayes factors will be used as a measure of evidence for hypothesis testing.

Motivating Example: Testing for HWE

- For simplicity we consider a diallelic marker, and suppose we obtain a random sample of genotypes for n individuals.
- The form of the data is

	Genotype			Total
	A_1A_1	A_1A_2	A_2A_2	
Count	n_1	n_2	n_3	n
Population Frequency	q_1	q_2	q_3	1

- So the model contains 3 probabilities (which sum to 1) q_1, q_2, q_3; hence, there are 2 free parameters.
- Suppose the proportions of alleles A_1 and A_2 in a given generation are p_1 and $p_2 = 1 - p_1$.
- In terms of q_1, q_2, q_3:

$$\begin{aligned} p_1 &= q_1 + \frac{q_2}{2} \\ p_2 &= \frac{q_2}{2} + q_3 \end{aligned}$$

Motivating Example: Testing for HWE

- HWE is the statistical independence of an individual's alleles at a locus.
- Under HWE, the probability distribution for the genotype of an individual in the next generation is:

	Genotype			
	A_1A_1	A_1A_2	A_2A_2	
Proportion	p_1^2	$2p_1p_2$	p_2^2	1

- Reasons for deviation from HWE include: small population size, selection, inbreeding and population structure.

A Toy Example

Example:

- Consider the data $n_{11} = 88, n_{12} = 10, n_{22} = 2$.
- Are these frequencies consistent with HWE?
- The MLEs are:

$$\begin{aligned} \widehat{q}_1 &= 0.88 \quad \widehat{q}_2 = 0.10 \quad \widehat{q}_3 = 0.02 \\ \widehat{p}_1 &= 0.93 \quad \widehat{p}_2 = 0.07 \end{aligned}$$

- For these data the exact p-value for

$$H_0 : q_1 = p_1^2, \quad q_2 = 2p_1p_2, \quad q_3 = p_2^2$$

is 0.0654.

```
> library(hwde)
> n1 <- 88
> n2 <- 10
> n3 <- 2
> exact <- hwexact(n1,n2,n3)
> exact
[1] 0.06544
```

Critique of Non-Bayesian Approach

- Testing for HWE is carried out via (asymptotic, i.e., large sample) χ^2 tests or exact tests.
- χ^2 tests require very large sample sizes for accurate p-values.
- The exact test can be computationally expensive to perform, when there are many alleles/samples.
- Under the null of HWE, the discreteness of the test statistic causes difficulties.
- In general, how to decide on a significance level? The level should be a function of sample size (and in particular should decrease as sample size increases), but how should it be chosen?
- Estimation depends on asymptotic approximations (i.e., large sample sizes).
- Estimation also difficult due to awkward constraints on parameters (particularly with many alleles).

Bayes Theorem

	Genotype			Total
	A_1A_1	A_1A_2	A_2A_2	
Count	n_1	n_2	n_3	n
Population Frequency	q_1	q_2	q_3	1

- The multinomial with three counts is known as the trinomial distribution.
- We have three parameters, q_1, q_2, q_3, but they sum to 1, so that effectively we have two parameters.
- We write $\mathbf{q} = (q_1, q_2, q_3)$ to represent the vector of probabilities, and $\mathbf{n} = (n_1, n_2, n_3)$ for the data vector.
- Via Bayes Theorem:

$$p(\mathbf{q}|\mathbf{n}) = \frac{\Pr(\mathbf{n}|\mathbf{q}) \times p(\mathbf{q})}{\Pr(\mathbf{n})}$$

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}$$

Elements of Bayes Theorem: The Likelihood

- We assume n independent draws with common probabilities $\mathbf{q} = (q_1, q_2, q_3)$.
- In this case, the distribution of n_1, n_2, n_3 is multinomial:

$$\Pr(n_1, n_2, n_3 | q_1, q_2, q_3) = \frac{n!}{n_1! n_2! n_3!} q_1^{n_1} q_2^{n_2} q_3^{n_3}. \tag{1}$$

- For fixed $\mathbf{n}$, we may view (1) as a function of $\mathbf{q}$ – this is the likelihood function.
- The maximum likelihood estimate (MLE) is

$$\widehat{\mathbf{q}} = \left(\frac{n_1}{n}, \frac{n_2}{n}, \frac{n_3}{n} \right).$$

- The MLE gives the highest probability to the observed data, i.e. maximizes the likelihood function.

The Dirichlet Distribution as a Prior Choice for a Multinomial $\mathbf{q}$

- Once the likelihood is specified we need to think about the prior distribution.
- We require a prior distribution over (q_1, q_2, q_3) — not straightforward since the three probabilities all lie in [0,1], and must sum to 1.
- A distribution that satisfies these requirements is the dirichlet distribution, denoted dirichlet(v_1, v_2, v_3) and has density:

$$\begin{aligned} p(q_1, q_2, q_3) &= \frac{\Gamma(v_1 + v_2 + v_3)}{\Gamma(v_1)\Gamma(v_2)\Gamma(v_3)} \times q_1^{v_1 - 1} q_2^{v_2 - 1} q_3^{v_3 - 1} \\ &\propto q_1^{v_1 - 1} q_2^{v_2 - 1} q_3^{v_3 - 1} \end{aligned}$$

where $\Gamma(\cdot)$ denotes the gamma function.

The Dirichlet Distribution as a Prior Choice for a Multinomial $\mathbf{q}$

- The dirichlet(v_1, v_2, v_3) prior:

$$\begin{aligned} p(q_1, q_2, q_3) &= \frac{\Gamma(v_1 + v_2 + v_3)}{\Gamma(v_1)\Gamma(v_2)\Gamma(v_3)} \times q_1^{v_1-1} q_2^{v_2-1} q_3^{v_3-1} \\ &\propto q_1^{v_1-1} q_2^{v_2-1} q_3^{v_3-1}. \end{aligned}$$

- $v_1, v_2, v_3 > 0$ are specified to reflect prior beliefs about (q_1, q_2, q_3).
- The dirichlet distribution can be used with general multinomial distributions (i.e. for $k = 2, 3, \ldots$ categories).
- The beta distribution is a special case of the dirichlet when there are two categories only.

Dirichlet Prior

- The mean and variance are

$$\begin{aligned} \mathrm{E}[q_i] &= \frac{v_i}{v} \\ \mathrm{var}(q_i) &= \frac{\mathrm{E}[q_i](1 - \mathrm{E}[q_i])}{v + 1} \end{aligned}$$

  for $i = 1, 2, 3$, where $v = v_1 + v_2 + v_3$.
- Large values of v increase the influence of the prior.
- The dirichlet has a single parameter only (v) to control the spread for all of the dimensions, which is a deficiency.
- The quartiles may be empirically calculated from samples.

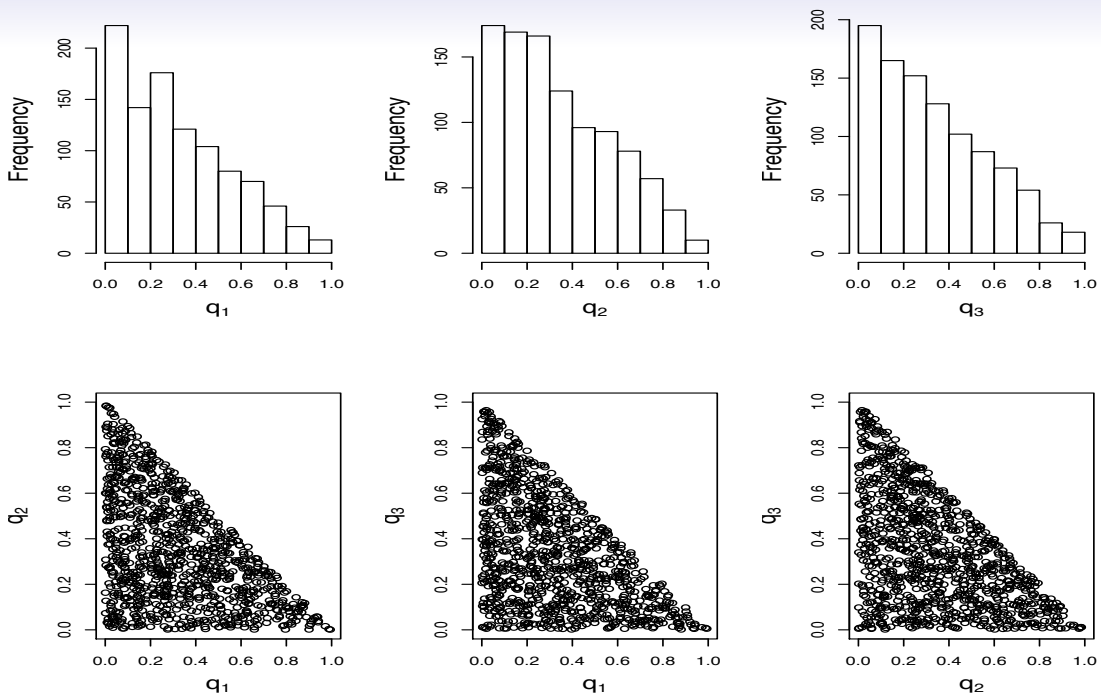


Figure 1 : Samples from a dirichlet$(1, 1, 1)$ distribution. The mean is $\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$.

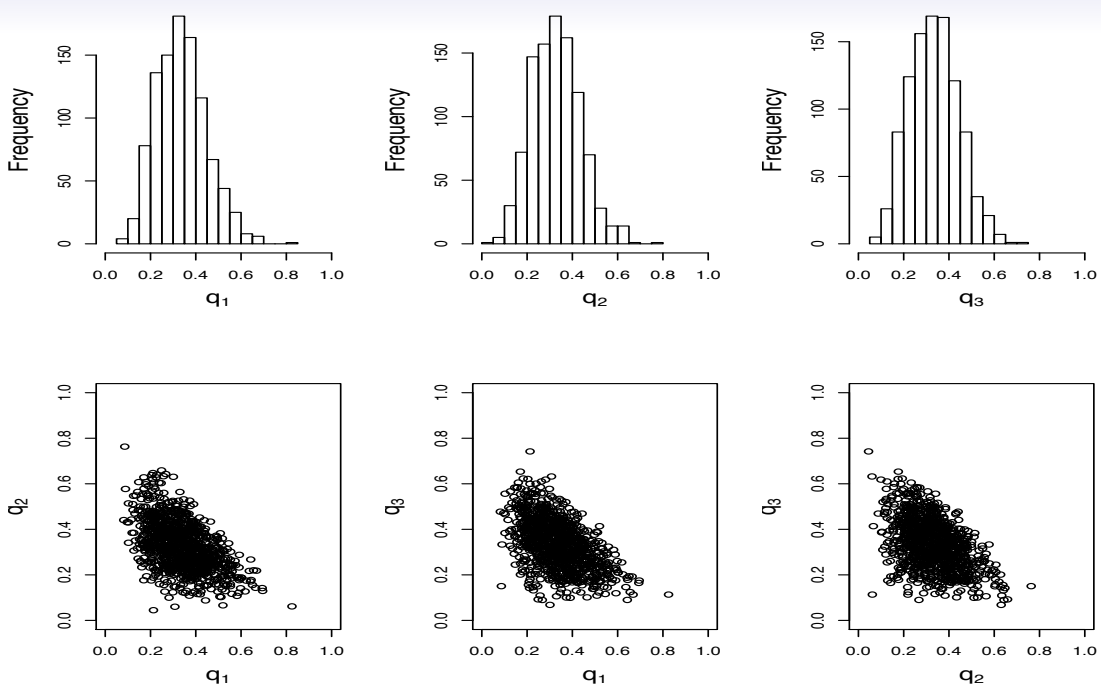


Figure 2 : Samples from a Dirichlet$(6, 6, 6)$ distribution. The mean is $\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$.

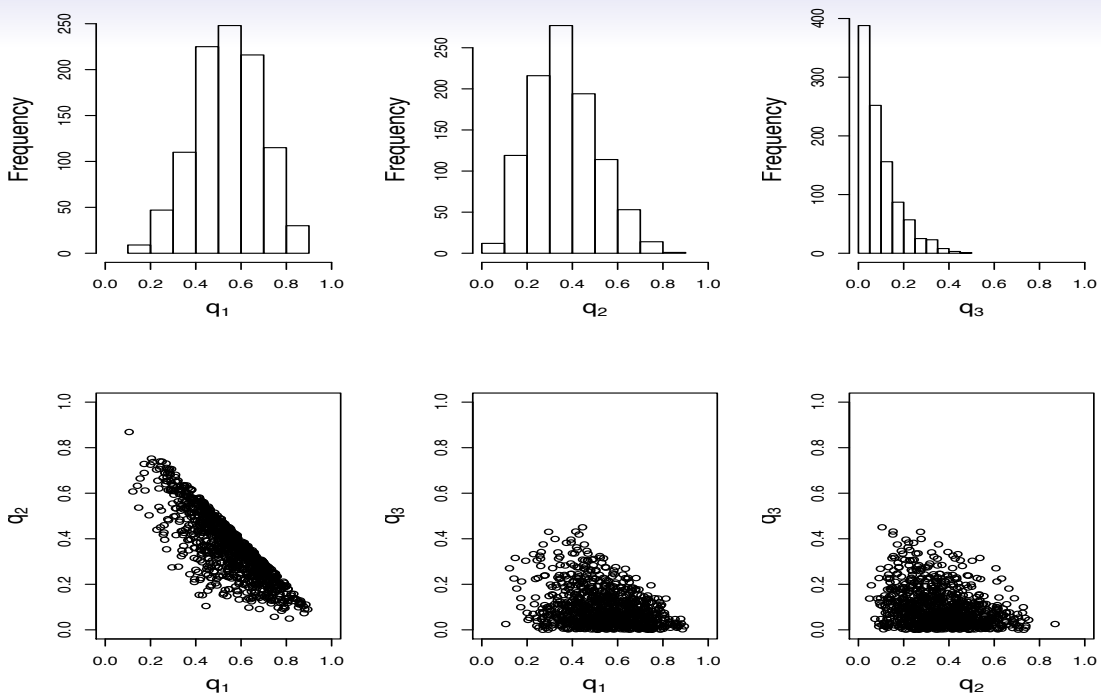


Figure 3 : Samples from a Dirichlet$(6, 4, 1)$ distribution. The mean is $\left(\frac{6}{11}, \frac{4}{11}, \frac{1}{11}\right) = (0.55, 0.36, 0.09)$.

R Code for dirichlet(1,1,1) Sampling

- The code below produces Figure 1 which is the dirichlet(1,1,1) case.

```
> library(VGAM) # To access the rdiric function
> nsim <- 1000
> par(mfrow=c(2,3))
> q <- rdiric(nsim,c(1,1,1))
#
# Univariate marginal representations
#
> hist(q[,1],xlab=expression(q[1]),main="",cex.lab=1.5,xlim=c(0,1))
> hist(q[,2],xlab=expression(q[2]),main="",cex.lab=1.5,xlim=c(0,1))
> hist(q[,3],xlab=expression(q[3]),main="",cex.lab=1.5,xlim=c(0,1))
#
# Bivariate representations
#
> plot(q[,1],q[,2],xlim=c(0,1),ylim=c(0,1),xlab=expression(q[1]),
     ylab=expression(q[2]),cex.lab=1.5)
> plot(q[,1],q[,3],xlim=c(0,1),ylim=c(0,1),xlab=expression(q[1]),
     ylab=expression(q[3]),cex.lab=1.5)
> plot(q[,2],q[,3],xlim=c(0,1),ylim=c(0,1),xlab=expression(q[2]),
     ylab=expression(q[3]),cex.lab=1.5)
```

Posterior Distribution

- Combining the Dirichlet prior, dirichlet(v_1, v_2, v_3), with the multinomial likelihood gives the posterior:

$$\begin{aligned} p(q_1, q_2, q_3|\mathbf{n}) &\propto \Pr(\mathbf{n}|\mathbf{q}) \times p(\mathbf{q}) \\ &\propto q_1^{n_1} q_2^{n_2} q_3^{n_3} \times q_1^{v_1-1} q_2^{v_2-1} q_3^{v_3-1} \\ &= q_1^{n_1+v_1-1} q_2^{n_2+v_2-1} q_3^{n_3+v_3-1}. \end{aligned}$$

- This distribution is another Dirichlet:

$$\text{dirichlet}(n_1 + v_1, n_2 + v_2, n_3 + v_3).$$

- Notice: "as if" we had observed counts $(n_1 + v_1, n_2 + v_2, n_3 + v_3)$.

Choosing a Prior

- The posterior mean for the expected proportion of counts in cell i is, for $i = 1, 2, 3$:

$$\begin{aligned} \mathrm{E}[q_i|\mathbf{n}] &= \frac{n_i + v_i}{n + v} \\ &= \frac{n_i}{n}\frac{n}{n+v} + \frac{v_i}{v}\frac{v}{n+v} \\ &= \text{MLE} \times \mathrm{W} + \text{Prior Mean} \times (1 - \mathrm{W}) \end{aligned}$$

  where $n = n_1 + n_2 + n_3$, $v = v_1 + v_2 + v_3$.

- The weight W is

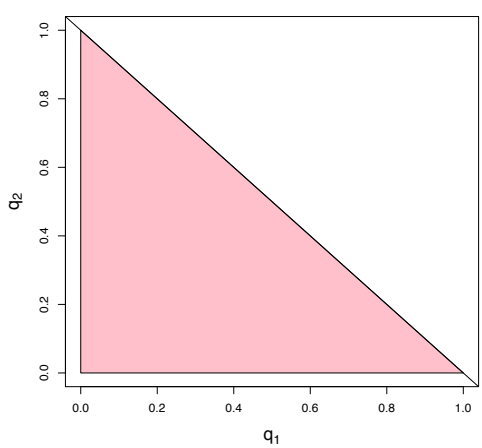
$$\mathrm{W} = \frac{n}{n+v}$$

  which is the proportion of the total information $(n + v)$ that is contributed by the data (n).

Choosing a Prior

- Recall the prior mean is
$$\left(\frac{v_1}{v}, \frac{v_2}{v}, \frac{v_3}{v}\right)$$
- These forms help to choose v_1, v_2, v_3.
- As with the beta distribution we may specify the prior means, and the relative weight that the prior and data contribute: n and v are on a comparable scale.
- For example, suppose we believe that event 1 is four times as likely as each of event 2 or event 3.
- Then we may specify the means in the ratios 4:1:1.
- Suppose $n = 24$ and we wish to allow the prior contribution to be a half of this total (and therefore a third of the complete information). Then the prior sample size is $v = 12$ and the prior mean requirement gives
$$v_1 = 8, v_2 = 2, v_3 = 2.$$

A Uniform Prior



An obvious choice of parameters is $v_1 = v_2 = v_3 = 1$ to give a prior that is uniform over the simplex:
$$\pi(q_1, q_2, q_3) = 2$$
for
$$0 < q_1, q_2, q_3 < 1, \quad q_1 + q_2 + q_3 = 1$$

Note: not uniform over all parameter of interests, as we see shortly

Simple HWE Example

- The data is
$$n_{11} = 88, n_{12} = 10, n_{22} = 2.$$
- We assume a flat dirichlet prior on the allowable values of $\mathbf{q}$:
$$v_1 = v_2 = v_3 = 1.$$
- This gives the posterior as dirichlet$(88+1, 10+1, 2+1)$ with posterior means:
$$\begin{aligned} \mathrm{E}[q_1|\mathbf{n}] &= \frac{1+88}{3+100} = \frac{89}{103} \\ \mathrm{E}[q_2|\mathbf{n}] &= \frac{1+10}{3+100} = \frac{11}{103} \\ \mathrm{E}[q_3|\mathbf{n}] &= \frac{1+2}{3+100} = \frac{3}{103}. \end{aligned}$$
- Note the similarity to the MLEs of
$$\left(\frac{88}{100}, \frac{10}{100}, \frac{2}{100}\right).$$

Simple HWE Example

- We continue with this example and now examine posterior distributions.
- We generate samples from
$$\text{dirichlet}(88+1, 10+1, 2+1).$$
- As posterior summaries we display, in Figure 4:
  - Histograms of the 3 univariate marginal distributions $p(q_1|\mathbf{y})$, $p(q_2|\mathbf{y})$, $p(q_3|\mathbf{y})$.
  - Scatterplots of the 3 bivariate marginal distributions $p(q_1, q_2|\mathbf{y})$, $p(q_1, q_3|\mathbf{y})$, $p(q_2, q_3|\mathbf{y})$.
- On each plot we indicate the MLEs for the general model, i.e. the non-HWE model (in red) and under the assumption of HWE (in blue).

Samples from the Posterior

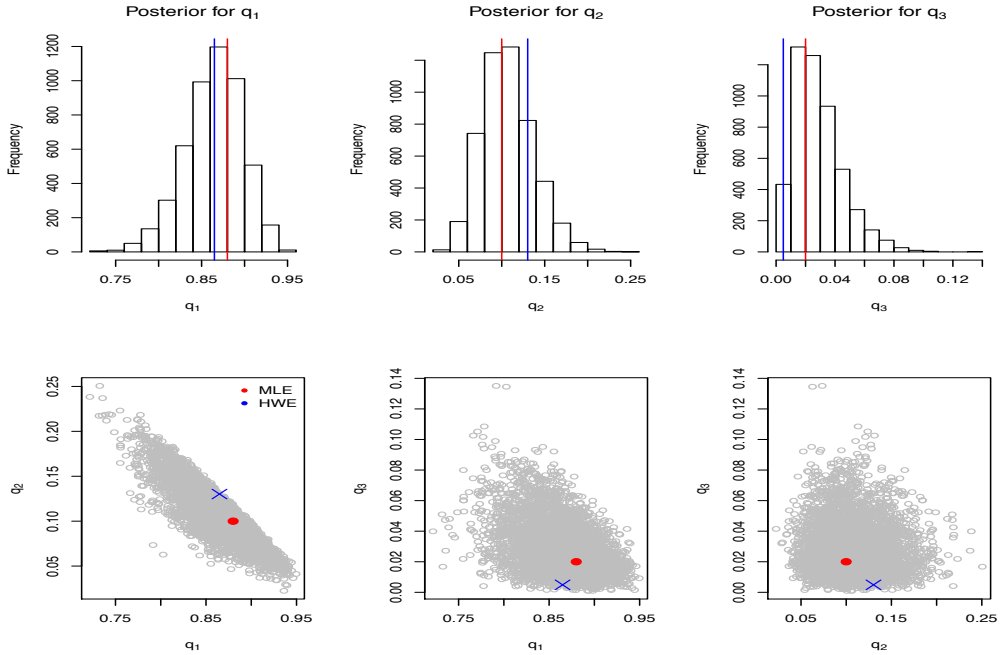


Figure 4 : Univariate and bivariate posterior distributions for $\mathbf{n} = (88, 10, 2)$. MLEs in red for the general model and in blue for the HWE model.

R Code for Histograms in Figure 4

```
> n1 <- 88; n2 <- 10; n3 <- 2
> p1 <- 88/100 + 0.5*10/100 # Estimated allele frequencies
> p2 <- 2/100 + 0.5*10/100  # for A1 and A2
> v1 <- v2 <- v3 <- 1
> nsim <- 5000
> q <- rdiric(nsim,c(n1+v1,n2+v2,n3+v3))
> par(mfrow=c(2,3))
> hist(q[,1],xlab=expression(q[1]),
    main=expression(paste("Posterior for ",q[1])))
> abline(v=n1/(n1+n2+n3),col="red")
> abline(v=p1^2,col="blue")
> hist(q[,2],xlab=expression(q[2]),
    main=expression(paste("Posterior for ",q[2])))
> abline(v=n2/(n1+n2+n3),col="red")
> abline(v=2*p1*p2,col="blue")
> hist(q[,3],xlab=expression(q[3]),
    main=expression(paste("Posterior for ",q[3])))
> abline(v=n3/(n1+n2+n3),col="red")
> abline(v=p2^2,col="blue")
```

- As expected with a sample size of $n = 100$ and a flat prior, the MLEs lie close to the center of the posteriors.
- Note the asymmetry of the posteriors. Asymptotic confidence intervals of the form $\widehat{q}_i \pm 1.96 \times \text{se}(\widehat{q}_i)$ would be symmetric.

R Code for Scatterplots in Figure 4

```
> plot(q[,2]~q[,1],xlab=expression(q[1]),ylab=expression(q[2]),
    col="grey")
> points(n1/(n1+n2+n3),n2/(n1+n2+n3),col="red",pch=20,cex=2)
> points(p1^2,2*p1*p2,col="blue",pch=4,cex=2)
> legend("topright",legend=c("MLE","HWE"),col=c("red","blue"),
      pch=c(20,20),bty="n")
> plot(q[,3]~q[,1],xlab=expression(q[1]),ylab=expression(q[3]),
    col="grey")
> points(n1/(n1+n2+n3),n3/(n1+n2+n3),col="red",pch=20,cex=2)
> points(p1^2,p2^2,col="blue",pch=4,cex=2)
> plot(q[,3]~q[,2],xlab=expression(q[2]),ylab=expression(q[3]),
    col="grey")
> points(n2/(n1+n2+n3),n3/(n1+n2+n3),col="red",pch=20,cex=2)
> points(2*p1*p2,p2^2,col="blue",pch=4,cex=2)
```

- In the context of a binomial sampling model and interest in a particular point (for example, $\theta = 0.5$) we could examine intervals for θ.
- In a multinomial context the situation is more complex; shortly we will examine Bayes factors to carry out hypothesis testing.

Parameters of Interest

	Genotype			Total
	A_1A_1	A_1A_2	A_2A_2	
Population Frequency	q_1	q_2	q_3	1

- Rather than q_1, q_2, q_3, we may be interested in other parameters of interest.
- In the HWE context: Let X_1 and X_2 be 0/1 indicators of A_1 for the two alleles at a locus.
- The covariance between X_1 and X_2 is the disequilibrium coefficent:

$$D = q_1 - p_1^2$$

  Under HWE $q_1 = p_1^2$, and the covariance is zero.
- Another quantity of interest (Shoemaker *et al.*, 1998) is

$$\psi = \frac{q_2^2}{q_1 q_3}.$$

  Under HWE, $\psi = 4$.

Parameters of Interest

- The inbreeding coefficient is

$$f = \frac{q_1 - p_1^2}{p_1 p_2}$$

- The variance of X_1 and X_2 is $p_1(1-p_1) = p_1 p_2$ and so f is the correlation.
- We may express q_1, q_2, q_3 as

$$\begin{aligned} q_1 &= p_1^2 + p_1(1-p_1)f \\ q_2 &= 2p_1(1-p_1)(1-f) \\ q_3 &= (1-p_1)^2 + p_1(1-p_1)f \end{aligned}$$

- Positive values of f indicate an excess of homozygotes (and may indicate inbreeding), while negative values indicate an excess of heterozygotes.
- Each of D, ψ and f are complex functions of q_1, q_2, q_3 and given a Dirichlet prior for the latter do not have known posterior forms.
- The "flat" prior for $\mathbf{q}$, dirichlet$(1,1,1)$, does not correspond to a flat prior for D, f, ψ, as Figure 5 shows.

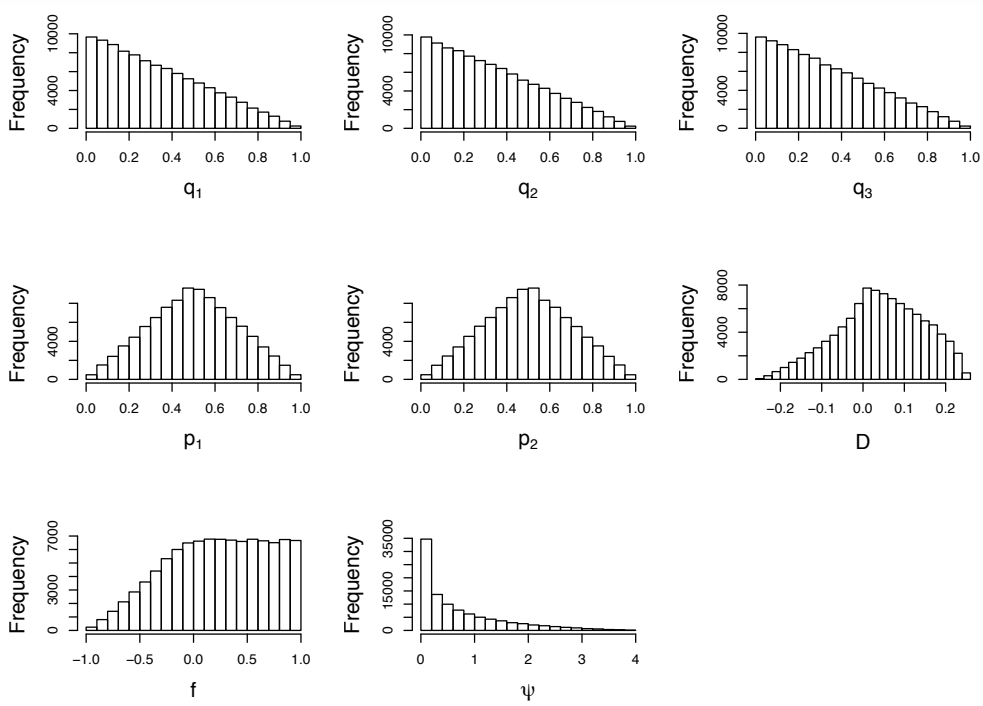


Figure 5 : Samples from a dirichlet(1,1,1) for various functions.

R Code for Functions of Interest

- Figure 5 was created via this code, recall we simply transform the samples:

```
> v1 <- v2 <- v3 1
> nsim <- 100000
> samps <- rdiric(nsim,c(v1,v2,v3))
> q1 <- samps[,1]
> q2 <- samps[,2]
> q3 <- samps[,3]
> par(mfrow=c(3,3))
> hist(q1,main="",xlab=expression(q[1]),cex.lab=1.5)
> hist(q2,main="",xlab=expression(q[2]),cex.lab=1.5)
> hist(q3,main="",xlab=expression(q[3]),cex.lab=1.5)
> p1 <- q1+q2/2
> p2 <- q3+q2/2;
> f <- (q1-p1^2)/(p1*p2)
> D <- q1-p1^2
> psi <- q2^2/(p1*p2)
> hist(p1,main="",xlab=expression(p[1]),cex.lab=1.5)
> hist(p2,main="",xlab=expression(p[2]),cex.lab=1.5)
> hist(D,main="",xlab=expression(D),cex.lab=1.5)
> hist(f,main="",xlab="f",cex.lab=1.5)
> hist(psi,main="",xlab=expression(psi),cex.lab=1.5)
```


Implied Prior on Functions of Interest

- Very important point: As we saw in the binomial development, you can't be "flat" on every scale, but the implications of particular priors can be assessed via simulation.
- We emphasize that we are not uniform on the marginal distributions for q_i, since these follow

$$\text{beta}(v_i, v - v_i)$$

  distributions, i.e. beta(1,2) if $v_1 = v_2 = v_3 = 1$.
- The priors on the measures of distance from HWE are far from uniform.
- For example, with a "flat" Dirichlet prior Dir(1,1,1) the prior probability that $f > 0$ is 0.67.

```
> cat("Prior prob f>0: ",sum(f>0)/nsim,"\n")
Prior prob f>0:  0.66762
> cat("Prior prob D>0: ",sum(D>0)/nsim,"\n")
Prior prob D>0:  0.66762
```

Bayesian Approaches to HWE

Previous approaches include:

- Altham (1971).
- Pereira and Rogatko (1984).
- Lindley (1988).
- Shoemaker *et al.* (1998).
- Montoya-Delgado *et al.* (2001).
- Consonni *et al.* (2008).
- Wakefield (2010).

Bayes factors for HWE

- Recall that Bayes factors measure the evidence in a sample for one hypothesis, as compared to an alternative.
- We derive the Bayes factor for multinomial data in the context of testing for HWE.
- We wish to test

$$H_0 : \text{HWE} \quad \text{versus} \quad H_1 : \text{Not HWE}.$$

- We need to specify priors on the null and alternatives, and then calculate the Bayes factor:

$$\frac{\Pr(\mathbf{n}|H_0)}{\Pr(\mathbf{n}|H_1)} = \frac{\int \Pr(\mathbf{n}|p_1)p(p_1)dp_1}{\int \Pr(\mathbf{n}|q_1, q_2)p(q_1, q_2)dq_1 dq_2}$$

  where p_1 and (q_1, q_2) are the parameters under the null and alternative, respectively.
- Under the null we have a single parameter, and under the alternative two.
- Important point: When Bayes factors are evaluated we need to include the normalizing constants.

HWE Bayes Factor

- Under H_0 and H_1 we must take care to evaluate the probability of the same data, n_1, n_2, n_3.
- Under the null,

$$\Pr(\mathbf{n}|p_1) = \Pr(n_1, n_2, n_3|p_1) = \frac{n!2^{n_1}}{n_2!n_{12}!n_3!} p_1^{2n_1+n_2}(1-p_1)^{n_2+2n_3}.$$

- With a $\text{Be}(w_1, w_2)$ prior on p_1:

$$\begin{aligned} \Pr(n_1, n_2, n_3|H_0) &= \int \Pr(\mathbf{n}|p_1) \times p(p_1)dp_1 \\ &= \frac{n!2^{n_2}\Gamma(w)\Gamma(2n_1+n_2+w_1)\Gamma(n_2+2n_3+w_2)}{n_1!n_2!n_3!\Gamma(w_1)\Gamma(w_2)\Gamma(2n+w)} \end{aligned} \tag{2}$$

- This is the probability of the observed data under the null.

HWE Bayes Factor

- The Bayes factor is

$$\frac{\Pr(\mathbf{n}|H_0)}{\Pr(\mathbf{n}|H_1)}$$

  and we have just given the form of the numerator.
- We now turn to the denominator.
- Under the alternative we assume $\mathbf{q} \sim \text{dirichlet}(v_1, v_2, v_3)$.
- The probability of the data under the alternative is:

$$\begin{aligned} \Pr(n_1, n_2, n_3|H_1) &= \int \Pr(\mathbf{n}|q_1, q_2) \times p(q_1, q_2)dq_1dq_2 \\ &= \frac{n!\Gamma(v)\Gamma(n_1+v_1)\Gamma(n_2+v_2)\Gamma(n_3+v_3)}{n_1!n_2!n_3!\Gamma(v_1)\Gamma(v_2)\Gamma(v_3)\Gamma(n+v)}. \end{aligned} \tag{3}$$

- Again, just a probability distribution, which we may evaluate for any realization of (n_1, n_2, n_3).

The HWE Bayes Factor

- Hence, the Bayes factor, measuring the evidence in the data for the null, as compared to the alternative is:

$$\begin{aligned} \text{BF} &= \frac{\Pr(n_1, n_2, n_3 | H_0)}{\Pr(n_1, n_2, n_3 | H_1)} \\ &= \frac{2^{n_2}\Gamma(w)\Gamma(2n_1 + n_2 + w_1)\Gamma(v_1)\Gamma(v_2)\Gamma(v_3)\Gamma(n_2 + 2n_3 + w_2)\Gamma(n + v)}{\Gamma(w_1)\Gamma(w_2)\Gamma(2n + w)\Gamma(v)\Gamma(n_1 + v_1)\Gamma(n_2 + v_2)\Gamma(n_3 + v_3)} \end{aligned}$$

  which is (2) divided by (3).
- This appears complex, but is just a function of the observed data, and the prior inputs, and can be easily evaluated[1].
- If BF $> 1 (< 1)$ the data are more (less) likely to have come from the null.
- Can be readily extended to $k > 2$ alleles.
- We next consider a formal decision rule.

[1] When we work out a χ^2 tail area we don't worry about the form of the distribution we just use the relevant function in our favorite software

Bayesian Decision Theory

- Decision as to reject H_0 in favor of H_1 depends on the costs of making the two types of error:

		Decision	
		Report H_0	Report H_1
Truth	H_0	0	C_I
	H_1	C_{II}	0

- Costs of making the two types of error C_I is the cost of a type I error and C_{II} the cost of a type II error.
- The decision theory solution is to report H_1 if:

$$\text{Posterior Odds of } H_0 = \text{BF} \times \text{Prior Odds} < \frac{C_{II}}{C_I} = R$$

  so that we only need to consider the ratio of costs R.
- If $\frac{C_{II}}{C_I} = 4$ (type II errors four times as bad as type I errors) then report H_1 if

$$\text{Posterior Odds of } H_0 < 4,$$

  i.e. if

$$\Pr(H_1 | \text{ data }) > 0.2.$$

A Simple Example

- We again consider the data $n_{11} = 88, n_{12} = 10, n_{22} = 2$.
- These data give a *p*-value of 0.0654.
- With "flat" conjugate Dirichlet priors ($w_1 = w_2 = v_1 = v_2 = v_3 = 1$) we obtain a Bayes factor of 1.54 so that the data are 50% more likely under the null than the alternative, so the evidence in favor of H_0 is not strong.
- With a prior probability of the null π_0, to give a prior odds of $\pi_0/(1-\pi_0)$, we have

$$\text{Posterior Odds of } H_0 = \text{BF} \times \frac{\pi_0}{1-\pi_0}$$

- Hence, with $\pi_0 = 0.5$ the posterior odds equal the Bayes factor, i.e. 1.54.
- The posterior probability of the null is

$$\frac{1.54}{1+1.54} = 0.61.$$

- This probability is very sensitive to the prior on the null. For example, with $\pi_0 = 2/3$ we obtain a posterior odds of $1.54 \times 2 = 3.08$ to give a posterior probability on the null of

$$\frac{3.08}{1+3.08} = 0.75.$$

Posterior Inference for Functions of Interest

- The code below produces Figure 6.

```
> q <- rdiric(nsim,c(n1+1,n2+1,n3+1))
> q1 <- q[,1]; q2 <- q[,2]; q3 <- q[,3]
> par(mfrow=c(3,3))
> hist(q1,main="",xlab=expression(q[1]),cex.lab=1.5)
> hist(q2,main="",xlab=expression(q[2]),cex.lab=1.5)
> hist(q3,main="",xlab=expression(q[3]),cex.lab=1.5)
> p1 <- q1+q2/2; p2 <- q3+q2/2;
> f <- (q1-p1^2)/(p1*p2); D <- q1-p1^2; psi <- q2^2/(p1*p2)
> hist(p1,main="",xlab=expression(p[1]),cex.lab=1.5)
> hist(p2,main="",xlab=expression(p[2]),cex.lab=1.5)
> hist(D,main="",xlab=expression(D),cex.lab=1.5)
> hist(f,main="",xlab="f",cex.lab=1.5)
> hist(psi,main="",xlab=expression(psi),cex.lab=1.5)
```

Posterior Inference for Functions of Interest

- Note that the asymptotic confidence interval for f contains inadmissible negative values.

```
#
# MLE of f is 0.23 with asymptotic standard error 0.17
# 95% interval is
> .23-1.96*.17
[1] -0.1032
> .23+1.96*.17
[1] 0.5632
#
# Bayesian posterior quantiles are
> quantile(f,c(.025,.5,.975))
      2.5%        50%      97.5%
0.02159864 0.27112179 0.58320757
```

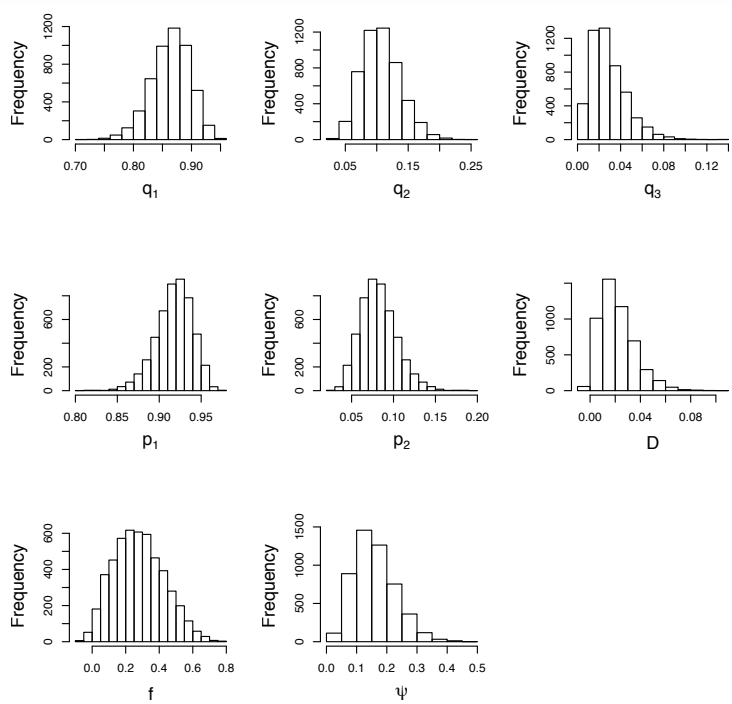



Figure 6 : Posterior inference for simple HWE example.

The HWEBayes Package

- The R package HWEBayes implements the rejection algorithm and importance sampling (a numerical integration technique), for testing and estimation in the HWE context:

  http://cran.r-project.org/web/packages/HWEBayes/index.html

- The *vignette* contains a worked example.
- Code for a four-allele example is here:

  http://faculty.washington.edu/jonno/HWEBayesFourAllele.R

- More details of the methodology: Wakefield (2010).

Conclusions

- The dirichlet distribution is convenient but quite inflexible as a prior distribution.
- Alternative priors are more difficult to specify since they are on scales that are more difficult to interpret (e.g. the logistic-normal distribution).
- Bayes factors are sensitive to the prior.
- Monte Carlo sampling is a powerful tool for inference.
- For multiple alleles computation is slow whether the approach is frequentist or Bayesian.

References

Altham, P. (1971). Exact Bayesian analysis of an intraclass 2×2 table. *Biometrika*, **58**, 679–680.

Consonni, G., Gutierrez-Pena, E., and Veronese, P. (2008). Compatible priors for Bayesian model comparison with an application to the Hardy-Weinberg equilibrium model. *Test*, **17**, 585–605.

Lindley, D. (1988). Statistical inference concerning Hardy-Weinberg equilibrium. In J. Bernardo, M. DeGroot, D. Lindley, and A. Smith, editors, *Bayesian Statistics 3*, pages 307–326, Oxford University Press, Oxford.

Montoya-Delgado, L., Irony, T., de B. Pereira, C., and Whittle, M. (2001). An unconditional exact test for the Hardy-Weinberg equilibrium law: sample-space ordering using the Bayes factor. *Genetics*, **158**, 875–883.

Pereira, C. and Rogatko, A. (1984). The Hardy-Weinberg equilibrium under a Bayesian perspective. *Revista Brasieira de Genetica*, **4**, 689–707.

Shoemaker, J., Painter, I., and Weir, B. (1998). A Bayesian characterization of Hardy-Weinberg disequilibrium. *Genetics*, **149**, 2079–2088.

Wakefield, J. (2010). Bayesian methods for examining Hardy-Weinberg equilibrium. *Biometrics*, **66**, 257–265.

Module 4: Bayesian Methods
Lecture 5: Linear regression

Peter Hoff

Departments of Statistics and Biostatistics
University of Washington

Outline

The linear regression model

Bayesian estimation

Regression models

How does an outcome Y vary as a function of $\mathbf{x} = \{x_1, \ldots, x_p\}$?

- Which x_j's have an effect?
- What are the effect sizes?
- Can we predict Y as a function of $\mathbf{x}$?

These questions can be assessed via a **regression model** $p(y|\mathbf{x})$.

Regression data

Parameters in a regression model can be estimated from data:

$$\begin{pmatrix} y_1 & x_{1,1} & \cdots & x_{1,p} \\ \vdots & \vdots & & \vdots \\ y_n & x_{n,1} & \cdots & x_{n,p} \end{pmatrix}$$

These data are often expressed in matrix/vector form:

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_n \end{pmatrix} = \begin{pmatrix} x_{1,1} & \cdots & x_{1,p} \\ \vdots & & \vdots \\ x_{n,1} & \cdots & x_{n,p} \end{pmatrix}$$

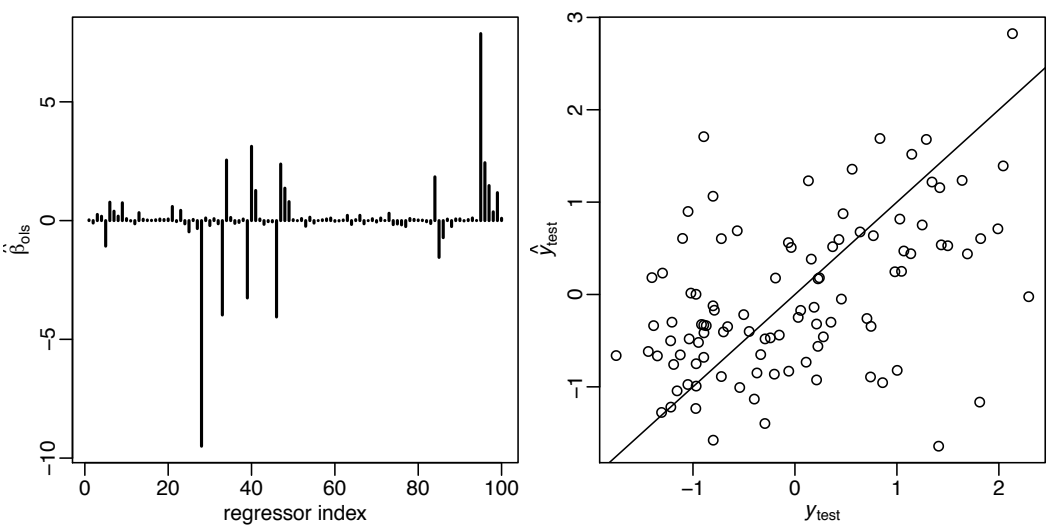
FTO experiment

FTO gene is hypothesized to be involved in growth and obesity.

Experimental design:

- 10 $fto+/-$ mice
- 10 $fto-/-$ mice
- Mice are sacrificed at the end of 1-5 weeks of age.
- Two mice in each group are sacrificed at each age.

FTO Data



Data analysis

- y = weight
- x_g = fto heterozygote $\in \{0, 1\}$ = number of "+" alleles
- x_a = age in weeks $\in \{1, 2, 3, 4, 5\}$

How can we estimate $p(y|x_g, x_a)$?

Cell means model:

genotype	*age*				
$-/-$	$\theta_{0,1}$	$\theta_{0,2}$	$\theta_{0,3}$	$\theta_{0,4}$	$\theta_{0,5}$
$+/-$	$\theta_{1,1}$	$\theta_{1,2}$	$\theta_{1,3}$	$\theta_{1,4}$	$\theta_{1,5}$

Problem: Only two observations per cell.

Linear regression

Solution: Assume smoothness as a function of age. For each group,

$$y = \alpha_0 + \alpha_1 x_a + \epsilon$$

This is a *linear regression model*.
Linearity means "linear in the parameters".

We could also try the model

$$y = \alpha_0 + \alpha_1 x_a + \alpha_2 x_a^2 + \alpha_3 x_a^3 + \epsilon,$$

which is also a linear regression model.

Multiple linear regression

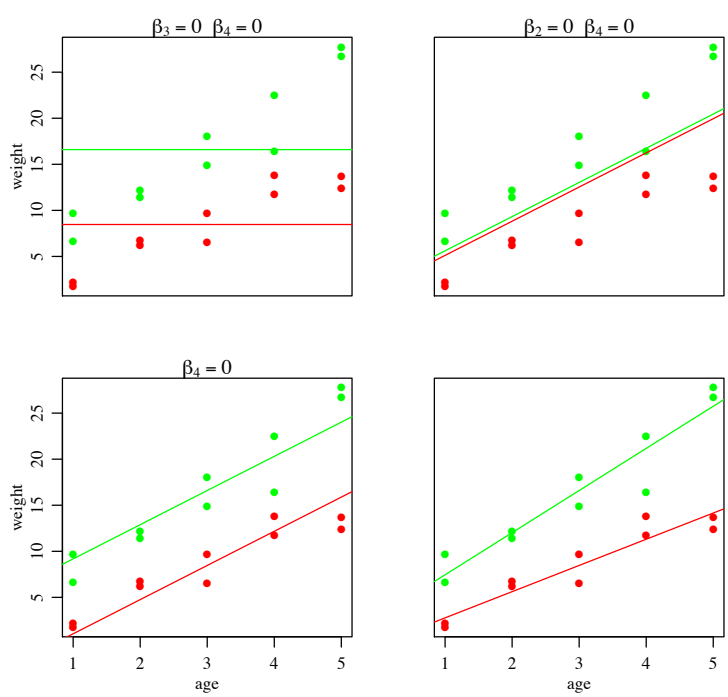
We can estimate the regressions for both groups simultaneously:

$$
\begin{aligned}
Y_i &= \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,4} + \epsilon_i \ , \text{ where} \\
x_{i,1} &= 1 \text{ for each subject } i \\
x_{i,2} &= 0 \text{ if subject } i \text{ is homozygous, 1 if heterozygous} \\
x_{i,3} &= \text{age of subject } i \\
x_{i,4} &= x_{i,2} \times x_{i,3}
\end{aligned}
$$

Under this model,

$$
\begin{aligned}
\mathrm{E}[Y|\mathbf{x}] &= \beta_1 + \beta_3 \times \mathrm{age} \ \text{ if } x_2 = 0, \text{ and} \\
\mathrm{E}[Y|\mathbf{x}] &= (\beta_1 + \beta_2) + (\beta_3 + \beta_4) \times \mathrm{age} \ \text{ if } x_2 = 1.
\end{aligned}
$$

Multiple linear regression



Normal linear regression

How does each Y_i vary around $\mathrm{E}[Y_i|\boldsymbol{\beta}, \mathbf{x}_i]$?

Assumption of normal errors:

$$\begin{aligned} \epsilon_1, \ldots, \epsilon_n &\sim \text{ i.i.d. normal}(0, \sigma^2) \\ Y_i &= \boldsymbol{\beta}^T \mathbf{x}_i + \epsilon_i \,. \end{aligned}$$

This completely specifies the probability density of the data:

$$\begin{aligned} & p(y_1, \ldots, y_n | \mathbf{x}_1, \ldots \mathbf{x}_n, \boldsymbol{\beta}, \sigma^2) \\ &= \prod_{i=1}^{n} p(y_i | \mathbf{x}_i, \boldsymbol{\beta}, \sigma^2) \\ &= (2\pi\sigma^2)^{-n/2} \exp\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \boldsymbol{\beta}^T \mathbf{x}_i)^2\}. \end{aligned}$$

Matrix form

- Let $\mathbf{y}$ be the n-dimensional column vector $(y_1, \ldots, y_n)^T$;
- Let $\mathbf{X}$ be the $n \times p$ matrix whose ith row is $\mathbf{x}_i$.

Then the normal regression model is that

$$\{\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^2\} \sim \text{ multivariate normal } (\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}),$$

where $\mathbf{I}$ is the $p \times p$ identity matrix and

$$\mathbf{X}\boldsymbol{\beta} = \begin{pmatrix} \mathbf{x}_1 \rightarrow \\ \mathbf{x}_2 \rightarrow \\ \vdots \\ \mathbf{x}_n \rightarrow \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} = \begin{pmatrix} \beta_1 x_{1,1} + \cdots + \beta_p x_{1,p} \\ \vdots \\ \beta_1 x_{n,1} + \cdots + \beta_p x_{n,p} \end{pmatrix} = \begin{pmatrix} \mathrm{E}[Y_1|\boldsymbol{\beta}, \mathbf{x}_1] \\ \vdots \\ \mathrm{E}[Y_n|\boldsymbol{\beta}, \mathbf{x}_n] \end{pmatrix}.$$

Ordinary least squares estimation

What values of $\boldsymbol{\beta}$ are consistent with our data $\mathbf{y}, \mathbf{X}$?

Recall

$$p(y_1, \ldots, y_n | \mathbf{x}_1, \ldots \mathbf{x}_n, \boldsymbol{\beta}, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \boldsymbol{\beta}^T \mathbf{x}_i)^2\}.$$

This is big when $\text{SSR}(\boldsymbol{\beta}) = \sum (y_i - \boldsymbol{\beta}^T \mathbf{x}_i)^2$ is small.

$$\begin{aligned} \text{SSR}(\boldsymbol{\beta}) &= \sum_{i=1}^{n} (y_i - \boldsymbol{\beta}^T \mathbf{x}_i)^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \mathbf{y}^T \mathbf{y} - 2\boldsymbol{\beta}^T \mathbf{X}^T \mathbf{y} + \boldsymbol{\beta}^T \mathbf{X}^T \mathbf{X} \boldsymbol{\beta} \,. \end{aligned}$$

What value of $\boldsymbol{\beta}$ makes this the smallest?

Calculus

Recall from calculus that

1. a minimum of a function $g(z)$ occurs at a value z such that $\frac{d}{dz} g(z) = 0$;
2. the derivative of $g(z) = az$ is a and the derivative of $g(z) = bz^2$ is $2bz$.

$$\begin{aligned} \frac{d}{d\boldsymbol{\beta}} \text{SSR}(\boldsymbol{\beta}) &= \frac{d}{d\boldsymbol{\beta}} \left(\mathbf{y}^T \mathbf{y} - 2\boldsymbol{\beta}^T \mathbf{X}^T \mathbf{y} + \boldsymbol{\beta}^T \mathbf{X}^T \mathbf{X} \boldsymbol{\beta} \right) \\ &= -2\mathbf{X}^T \mathbf{y} + 2\mathbf{X}^T \mathbf{X} \boldsymbol{\beta} \,, \end{aligned}$$

Therefore,

$$\begin{aligned} \frac{d}{d\boldsymbol{\beta}} \text{SSR}(\boldsymbol{\beta}) = 0 \quad &\Leftrightarrow \quad -2\mathbf{X}^T \mathbf{y} + 2\mathbf{X}^T \mathbf{X} \boldsymbol{\beta} = 0 \\ &\Leftrightarrow \quad \mathbf{X}^T \mathbf{X} \boldsymbol{\beta} = \mathbf{X}^T \mathbf{y} \\ &\Leftrightarrow \quad \boldsymbol{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \,. \end{aligned}$$

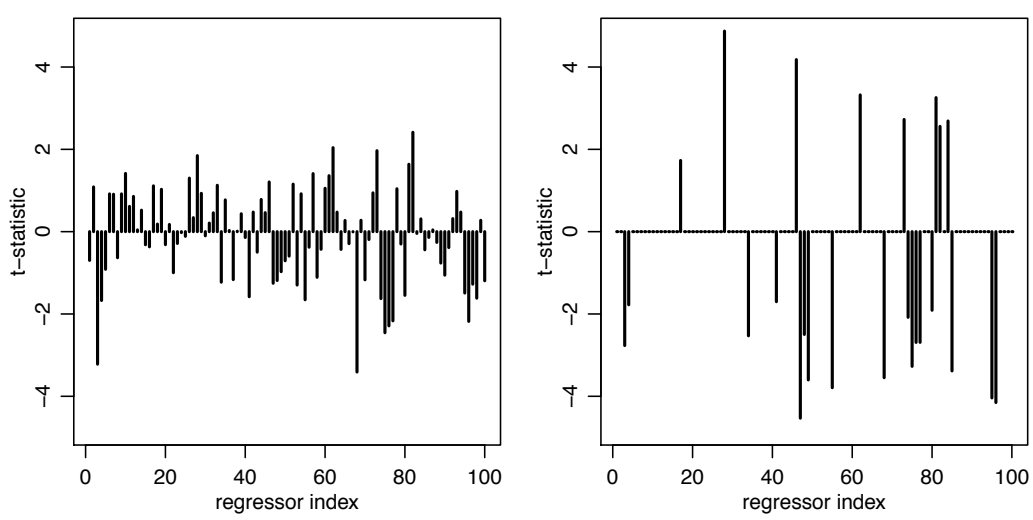
$\hat{\boldsymbol{\beta}}_{\text{ols}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$ is the *OLS estimator* of $\boldsymbol{\beta}$.

OLS estimation in R

```
### OLS estimate
beta.ols<- solve( t(X)%*%X )%*%t(X)%*%y

c(beta.ols)

## [1] -0.06822  2.94485  2.84421  1.72948
```

```
### using lm
fit.ols<-lm(y~ X[,2] + X[,3] +X[,4] )

summary(fit.ols)$coef

##             Estimate Std. Error  t value  Pr(>|t|)
## (Intercept) -0.06822     1.4223 -0.04796 9.623e-01
## X[, 2]       2.94485     2.0114  1.46406 1.625e-01
## X[, 3]       2.84421     0.4288  6.63235 5.761e-06
## X[, 4]       1.72948     0.6065  2.85171 1.154e-02
```


OLS estimation




```
summary(fit.ols)$coef

##             Estimate Std. Error  t value  Pr(>|t|)
## (Intercept) -0.06822     1.4223 -0.04796 9.623e-01
## X[, 2]       2.94485     2.0114  1.46406 1.625e-01
## X[, 3]       2.84421     0.4288  6.63235 5.761e-06
## X[, 4]       1.72948     0.6065  2.85171 1.154e-02
```

Bayesian inference for regression models

$$y_i = \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p} + \epsilon_i$$

Motivation:

- Posterior probability statements: $\Pr(\beta_j > 0 | \mathbf{y}, \mathbf{X})$
- OLS tends to overfit when p is large, Bayes more conservative.
- Model selection and averaging

Prior and posterior distribution

prior	$\boldsymbol{\beta}$	$\sim$	$\text{mvn}(\boldsymbol{\beta}_0, \Sigma_0)$
sampling model	$\mathbf{y}$	$\sim$	$\text{mvn}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$
posterior	$\boldsymbol{\beta}\|\mathbf{y}, \mathbf{X}$	$\sim$	$\text{mvn}(\boldsymbol{\beta}_n, \Sigma_n)$

where

$$\begin{aligned} \Sigma_n = \mathrm{Var}[\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}, \sigma^2] &= (\Sigma_0^{-1} + \mathbf{X}^T\mathbf{X}/\sigma^2)^{-1} \\ \boldsymbol{\beta}_n = \mathrm{E}[\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}, \sigma^2] &= (\Sigma_0^{-1} + \mathbf{X}^T\mathbf{X}/\sigma^2)^{-1}(\Sigma_0^{-1}\boldsymbol{\beta}_0 + \mathbf{X}^T\mathbf{y}/\sigma^2). \end{aligned}$$

Notice:

- If $\Sigma_0^{-1} \ll \mathbf{X}^T\mathbf{X}/\sigma^2$, then $\boldsymbol{\beta}_n \approx \hat{\boldsymbol{\beta}}_{\text{ols}}$
- If $\Sigma_0^{-1} \gg \mathbf{X}^T\mathbf{X}/\sigma^2$, then $\boldsymbol{\beta}_n \approx \boldsymbol{\beta}_0$

The g-prior

How to pick $\boldsymbol{\beta}_0, \Sigma_0$?

g-prior:

$$\boldsymbol{\beta} \sim \text{mvn}(\mathbf{0}, g\sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$$

Idea: The variance of the OLS estimate $\hat{\boldsymbol{\beta}}_{\text{ols}}$ is

$$\text{Var}[\hat{\boldsymbol{\beta}}_{\text{ols}}] = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} = \frac{\sigma^2}{n}(\mathbf{X}^T\mathbf{X}/n)^{-1}$$

This is roughly the uncertainty in $\boldsymbol{\beta}$ from n observations.

$$\text{Var}[\boldsymbol{\beta}]_{\text{gprior}} = g\sigma^2(\mathbf{X}^T\mathbf{X})^{-1} = \frac{\sigma^2}{n/g}(\mathbf{X}^T\mathbf{X}/n)^{-1}$$

The g-prior can roughly be viewed as the uncertainty from n/g observations.

For example, $g = n$ means the prior has the same amount of info as 1 obs.

Posterior distributions under the g-prior

$$\{\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}, \sigma^2\} \sim \text{mvn}(\boldsymbol{\beta}_n, \boldsymbol{\Sigma}_n)$$

$$\begin{aligned}
\Sigma_n = \text{Var}[\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}, \sigma^2] &= \frac{g}{g+1}\sigma^2(\mathbf{X}^T\mathbf{X})^{-1} \\
\boldsymbol{\beta}_n = \text{E}[\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}, \sigma^2] &= \frac{g}{g+1}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}
\end{aligned}$$

Notes:

- The posterior mean estimate $\boldsymbol{\beta}_n$ is simply $\frac{g}{g+1}\hat{\boldsymbol{\beta}}_{\text{ols}}$.
- The posterior variance of $\boldsymbol{\beta}$ is simply $\frac{g}{g+1}\text{Var}[\hat{\boldsymbol{\beta}}_{\text{ols}}]$.
- g shrinks the coefficients and can prevent overfitting to the data
- If $g = n$, then as n increases, inference approximates that using $\hat{\boldsymbol{\beta}}_{\text{ols}}$.

Monte Carlo simulation

What about the error variance σ^2?

prior	$1/\sigma^2$	$\sim$	$\text{gamma}(\nu_0/2, \nu_0\sigma_0^2/2)$
sampling model	$\mathbf{y}$	$\sim$	$\text{mvn}(\mathbf{X}\beta, \sigma^2\mathbf{I})$
posterior	$1/\sigma^2\|\mathbf{y}, \mathbf{X}$	$\sim$	$\text{gamma}([\nu_0 + n]/2, [\nu_0\sigma_0^2 + \text{SSR}_g]/2)$

where SSR_g is somewhat complicated.

Simulating the joint posterior distribution:

joint distribution	$p(\sigma^2, \boldsymbol{\beta}\|\mathbf{y}, \mathbf{X})$	$=$	$p(\sigma^2\|\mathbf{y}, \mathbf{X}) \times p(\boldsymbol{\beta}\|\mathbf{y}, \mathbf{X}, \sigma^2)$
simulation	$\{\sigma^2, \boldsymbol{\beta}\} \sim p(\sigma^2, \boldsymbol{\beta}\|\mathbf{y}, \mathbf{X})$	$\Leftrightarrow$	$\sigma^2 \sim p(\sigma^2\|\mathbf{y}, \mathbf{X}), \boldsymbol{\beta} \sim p(\boldsymbol{\beta}\|\mathbf{y}, \mathbf{X}, \sigma^2)$

To simulate $\{\sigma^2, \boldsymbol{\beta}\} \sim p(\sigma^2, \boldsymbol{\beta}|\mathbf{y}, \mathbf{X})$,

1. First simulate σ^2 from $p(\sigma^2|\mathbf{y}, \mathbf{X})$
2. Use this σ^2 to simulate $\boldsymbol{\beta}$ from $p(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}, \sigma^2)$

Repeat 1000's of times to obtain MC samples: $\{\sigma^2, \boldsymbol{\beta}\}^{(1)}, \ldots, \{\sigma^2, \boldsymbol{\beta}\}^{(S)}$.

FTO example

Priors:

$$\begin{aligned} 1/\sigma^2 &\sim \text{gamma}(1/2, 3.6781/2) \\ \boldsymbol{\beta}|\sigma^2 &\sim \text{mvn}(\mathbf{0}, g \times \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}) \end{aligned}$$

Posteriors:

$$\begin{aligned} \{1/\sigma^2|\mathbf{y}, \mathbf{X}\} &\sim \text{gamma}((1+20)/2, (3.6781 + 251.7753)/2) \\ \{\boldsymbol{\beta}|\mathbf{Y}, \mathbf{X}, \sigma^2\} &\sim \text{mvn}(.952 \times \hat{\boldsymbol{\beta}}_{\text{ols}}, .952 \times \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}) \end{aligned}$$

where

$$(\mathbf{X}^T\mathbf{X})^{-1} = \begin{pmatrix} 0.55 & -0.55 & -0.15 & 0.15 \\ -0.55 & 1.10 & 0.15 & -0.30 \\ -0.15 & 0.15 & 0.05 & -0.05 \\ 0.15 & -0.30 & -0.05 & 0.10 \end{pmatrix} \quad \hat{\boldsymbol{\beta}}_{\text{ols}} = \begin{pmatrix} -0.0682 \\ 2.9449 \\ 2.8442 \\ 1.7295 \end{pmatrix}$$

R-code

```
##  data dimensions
n<-dim(X)[1] ; p<-dim(X)[2]

## prior parameters
nu0<-1
s20<-summary(lm(y~-1+X))$sigma^2
g<-n

## posterior calculations
Hg<- (g/(g+1)) * X%*%solve(t(X)%*%X)%*%t(X)
SSRg<- t(y)%*%( diag(1,nrow=n)  - Hg ) %*%y

Vbeta<- g*solve(t(X)%*%X)/(g+1)
Ebeta<- Vbeta%*%t(X)%*%y

## simulate sigma^2  and beta
s2.post<-beta.post<-NULL
for(s in 1:5000)
{
  s2.post<-c(s2.post,1/rgamma(1, (nu0+n)/2, (nu0*s20+SSRg)/2 ) )
  beta.post<-rbind(beta.post, rmvnorm(1,Ebeta,s2.post[s]*Vbeta))
}
```


MC approximation to posterior

```
s2.post[1:5]

## [1]  9.737 13.002 15.284 14.528 14.818
```

```
beta.post[1:5,]

##        [,1]    [,2]  [,3]  [,4]
## [1,]  1.701  1.2066 1.649 2.841
## [2,] -1.868  1.2554 3.216 1.975
## [3,]  1.032  1.5555 1.909 2.338
## [4,]  3.351 -1.3819 2.401 2.364
## [5,]  1.486 -0.6652 2.032 2.977
```

MC approximation to posterior

```
quantile(s2.post,probs=c(.025,.5,.975))

##   2.5%    50%  97.5%
##  7.163 12.554 24.774

quantile(sqrt(s2.post),probs=c(.025,.5,.975))

##  2.5%   50% 97.5%
## 2.676 3.543 4.977
```

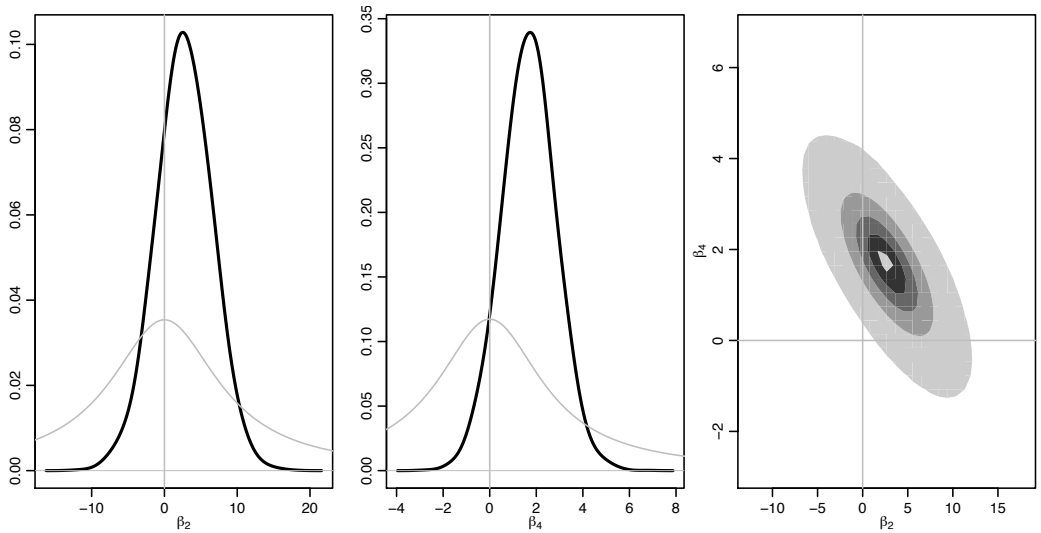
```
apply(beta.post,2,quantile,probs=c(.025,.5,.975))

##           [,1]   [,2]  [,3]    [,4]
## 2.5%  -5.26996 -4.840 1.065 -0.5929
## 50%   -0.01051  2.698 2.678  1.6786
## 97.5%  5.20650  9.992 4.270  3.9071
```


OLS/Bayes comparison

```
apply(beta.post,2,mean)

## [1] 0.0133 2.7080 2.6796 1.6736

apply(beta.post,2,sd)

## [1] 2.6637 3.7726 0.8055 1.1429
```

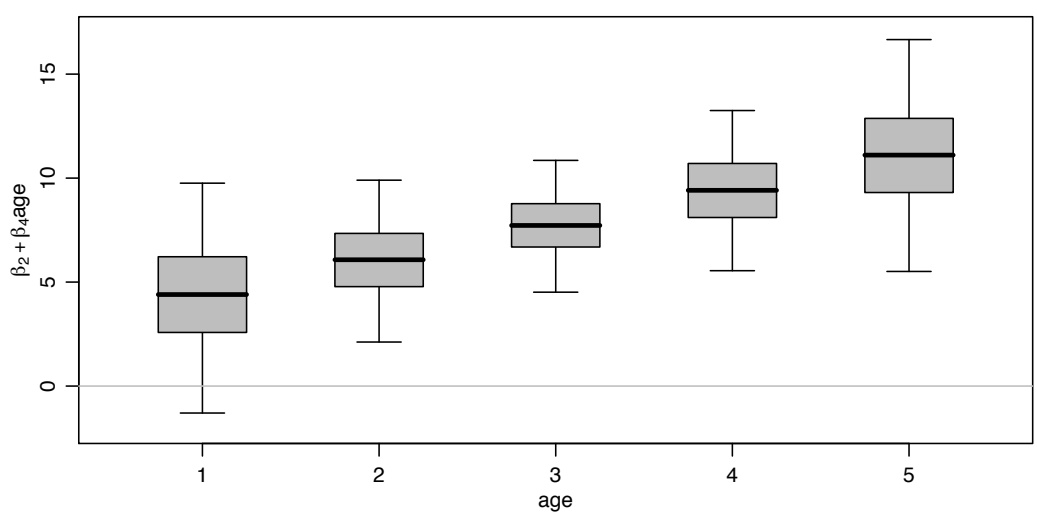
```
summary(fit.ols)$coef

##     Estimate Std. Error  t value  Pr(>|t|)
## X   -0.06822     1.4223 -0.04796 9.623e-01
## Xxg  2.94485     2.0114  1.46406 1.625e-01
## Xxa  2.84421     0.4288  6.63235 5.761e-06
## X    1.72948     0.6065  2.85171 1.154e-02
```

Posterior distributions



Summarizing the genetic effect

$$
\begin{aligned}
\text{Genetic effect} &= \mathrm{E}[y|age, +/-] - \mathrm{E}[y|age, -/-] \\
&= [(\beta_1 + \beta_2) + (\beta_3 + \beta_4) \times \text{age}] - [\beta_1 + \beta_3 \times \text{age}] \\
&= \beta_2 + \beta_4 \times \text{age}
\end{aligned}
$$



2014 SISG MODULE 4: Bayesian Statistics for Genetics
Lecture 6: Bayesian and Frequentist Multiple Testing

Jon Wakefield

Departments of Statistics and Biostatistics
University of Washington

Outline

Introduction and Motivating Examples

Motivation: Homocysteine Example

Connection Between p-Values and Bayes Factors
- Review of p-Values
- A Bayesian Perspective

Multiple Testing
- Review of Non-Bayesian Procedures
- An Example
- Bayesian Approach

Methodological Framework
- The Model

Homocysteine Example

Substantive Prior Information

Conclusions

Introduction

- With the advent of high-throughput technologies in genomics there is now the possibility of carrying out millions of tests, and so the implications of such multiple testing must be carefully considered.
- In this lecture we will review the rationale for p-values.
- We then explore the connection between p-values and Bayes factors.
- Multiple testing will be reviewed and a Bayesian perspective presented.
- An example in the context of a pharmacogenomics GWAS will be presented.
- The use of substantive prior information will also be demonstrated.

Motivating Data Description

- The Vitamin Intervention for Stroke Prevention (VISP) trial is an NIH-funded, multi-center, double-blind, randomized, controlled clinical trial.
- More detail in Wakefield *et al.* (2014).
- The aim is to determine whether a daily intake of high dose folic acid and vitamins B6 and B12 was associated with cardiovascular endpoints.
- We examine data on $n = 1670$ individuals, with 837 randomized to the high dose and 833 to the low dose.
- The outcome is the intermediary variable homocysteine level: high levels in blood are associated with cardiovascular disease.
- In the VISP trial, levels were measured longitudinally but for simplicity we take as outcome the difference between the baseline and the first post-baseline measurements: Y will represent this difference.
- The change was -0.37 μmol/L in the low dose group versus -2.36 μmol/L in the high dose group, i.e., a difference of -1.99 μmol/L ($p < 2 \times 10^{-16}$).

Marker-Specific Treatment Effects

- An increasingly important venture is examining treatment effects by marker (e.g. SNP): a particular type of gene-environment interaction.
- Historically, candidate gene studies were popular, but now genome-wide scans are also being performed, see Daly (2010) for a review.
- Pharmacogenomics-related traits: Drug response, susceptibility to adverse drug reactions,...
- **Key Statistical Point:** The estimated interactions are based on subgroups of varying sizes, so that the power varies substantially across tests.
- In the **VISP** trial, there are $J = 803,122$ SNPs and suppose we define subgroups as having at least one copy of the minor allele.
- The number in this subgroup ranges between 21 and 1564 across SNPs.

The Pharmacogenomics Journal (2012), 1–9
© 2012 Macmillan Publishers Limited. All rights reserved 1470-269X/12
www.nature.com/tpj

ORIGINAL ARTICLE

Beyond single-marker analyses: mining whole genome scans for insights into treatment responses in severe sepsis

M Man[1], SL Close[2], AD Shaw[3], GR Bernard[4], IS Douglas[5], RJ Kaner[6], D Payen[7], J-L Vincent[8], S Fossceco[1], JM Janes[1], AG Leishman[1], L O'Brien[1], MD Williams[9] and JGN Garcia[10]

[1]*Lilly Research Laboratories, Eli Lilly and Company, Lilly Corporate Center, Indianapolis, IN, USA;* [2]*Department of Clinical Pharmacology, Indiana University School of Medicine, Indianapolis, IN, USA;* [3]*Department of Anesthesiology, Duke University Medical Center, Durham, NC, USA;* [4]*Division of Allergy, Pulmonary and Critical Care, Vanderbilt University School of Medicine, Nashville, TN, USA;* [5]*Denver Health Medical Center, University of Colorado, School of Medicine, Denver, CO,*

Management of severe sepsis, an acute illness with high morbidity and mortality, suffers from the lack of effective biomarkers and largely empirical predictions of disease progression and therapeutic responses. We conducted a genome-wide association study using a large randomized clinical trial cohort to discover genetic biomarkers of response to therapy and prognosis utilizing novel approaches, including combination markers, to overcome limitations of single-marker analyses. Sepsis prognostic models were dominated by clinical variables with genetic markers less informative. In contrast, evidence for gene–gene interactions were identified for sepsis treatment responses with genetic biomarkers dominating models for predicting therapeutic responses, yielding candidates for replication in other cohorts.
The Pharmacogenomics Journal advance online publication, 7 February 2012; doi:10.1038/tpj.2012.1

Keywords: drotrecogin alfa (activated); epistasis; genetic markers; genome-wide association study; polymorphism; severe sepsis

- Aim: To identify marker-defined populations with improved response to DAA (for treatment of severe sespis).

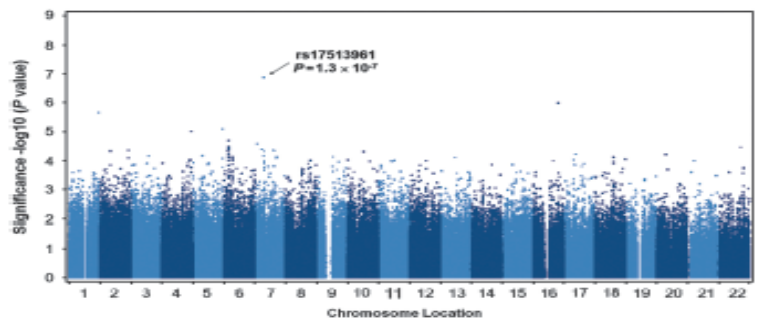


Figure 1 Representative Manhattan plot from GWAS in the entire cohort for genotype BB vs not BB. All possible combinations of genotypes representing a dominant, heterozygous and recessive inheritance were evaluated. Three pairs of comparisons AA vs not AA, AB (heterozygous) vs not AB and BB vs not BB were completed. This figure represents the plot of chromosome position and *P*-values for homozygous genotype (BB) vs heterozygous (AB) or homozygous for the other allele (AA), or BB vs not BB. (Manhattan plots of GWAS results for genotype AA vs not AA and AB vs not AB are shown in Supplementary Figure 1). GWAS = genome-wide association study.

The Statistical Set-Up

- We begin with a very simple situation in which we have a single parameter of interest θ.
- Assume the null of interest is

$$H_0 : \theta = 0$$

with θ, for example, a treatment difference, or a log odds ratio, or a log hazard ratio.

- We assume an analysis yields a statistic T for which large values indicate departures from the null.
- For example, the squared Wald statistic, $T = \widehat{\theta}^2/V$, with V the asymptotic variance of the MLE[1].
- An alternative is the likelihood ratio statistic.

[1] $T = Z^2$ where Z is the Z-score

Types of Testing

- The observed p-value is given by:

$$p = \Pr(T > t_{\text{obs}} | H_0)$$

  where t_{obs} is a number that is evaluated for the data at hand.
- To report p only, gives a pure significance test.
- A small p-value can arise because:
  - H_0 is true but we were "unlucky".
  - H_0 is not true.

  – to decide which explanation is responsible depends crucially on the prior belief on whether H_0 is true or not.

Key question: How small is small?

Types of Testing

- A test of significance sets a cut-off value (e.g. $\alpha = 0.05$) and rejects H_0 if $p < \alpha$.

Again: How to pick α?

- A type I error is to reject H_0 when it is true, and a test of significance controls the type I error (whereas a pure significance test does not).
- A type II error occurs when H_1 is true but H_0 is not rejected.
- A hypothesis test goes one step further and specifies an alternative hypothesis.
- A decision is then taken as to which of H_0 and H_1 is chosen.
- The celebrated Neyman-Pearson lemma shows that for fixed α-level the likelihood ratio statistic maximizes the power.
- Wouldn't it be more reasonable to balance type I and type II errors?

The Dangers of Fixed Significance Levels

- Example: Sample, $Y_1, ..., Y_n$ of size n from normal$(\theta, 1)$,

$$H_0 : \theta = 0, \quad H_1 : \theta = 1.$$

  Obvious that we should reject H_0 for $\overline{Y} > k$, a constant.
- The table below illustrates the problems of choosing a fixed α, regardless of sample size — imbalance in α and β as a function of n.

n	α	β	k
1	0.01	0.91	6.21
25	0.01	0.0038	0.42
100	0.01	8×10^{-15}	2.5×10^{-12}

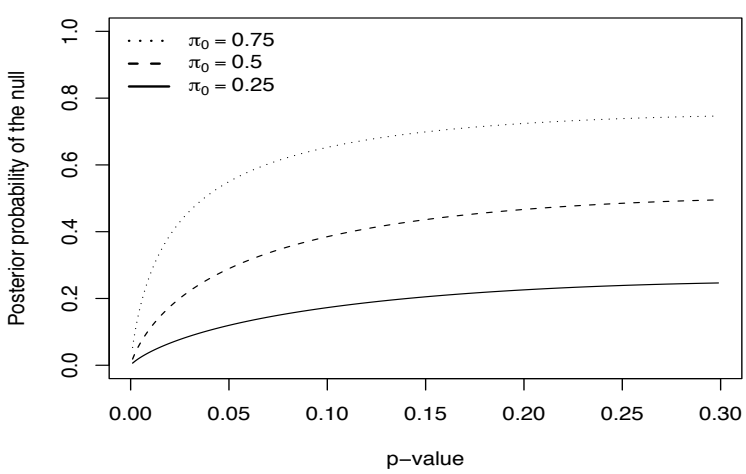
- Also: Statistical versus practical significance.
- For both p-values and α levels we need thresholds that decrease as a function of the sample size n. Pearson (1953, p. 68), "...the quite legitimate device of reducing α as n increases".

Calibrating p-values

- With $\pi_0 = \Pr(H_0)$, Sellke, Bayarri and Berger (2001) show that:

$$\Pr(H_0 | \text{ data }) \geq \left\{ 1 - \frac{1}{2.72\ p \log p} \times \frac{1 - \pi_0}{\pi_0} \right\}^{-1} \tag{1}$$

- A small p-value doesn't translate to a small probability that the null is not true.



Why does anyone use *p*-values?

- Historically, it was usual to carry out well-powered (single) experiments, and the prior on the alternative was not small.
- With respect to (1) and with $\pi_0 = 0.5$:
  - p-value $= 0.05$ gives $\Pr(H_0|$ data $) > 0.29$.
  - p-value $= 0.01$ gives $\Pr(H_0|$ data $) > 0.11$.
- Scientists well-calibrated in their own discipline?
- Perhaps, but if you're going to be subjective, why not be formal about it?
- Aside: Reason for lack of replication in observational epidemiology? Along with confounding, data dredging, measurement error,...

Calibrating α-Levels

- We want $\Pr(H_0|$ data $)$, where "data" corresponds to the event $T > t_{\text{fix}}$, but to obtain this we must specify alternatives – consider a simple alternative, say $H_1 : \theta = \theta_1$.
- Then,

$$\begin{aligned} \text{Posterior Odds of } H_0 &= \frac{\Pr(T > t_{\text{fix}}|H_0)}{\Pr(T > t_{\text{fix}}|H_1)} \times \frac{\Pr(H_0)}{\Pr(H_1)} \\ &= \frac{\alpha}{1-\beta} \times \text{ Prior Odds of } H_0 \end{aligned}$$

- For ranking associations (which does not involve the prior odds if constant across tests): must consider the power, $\Pr($ data $|H_1)$.
- For calibration: must consider the prior odds of H_0.

A Sanity Check via a Simple Example

- The model:
$$Y_i|\theta \sim_{iid} \text{normal}(\theta, \sigma^2), \qquad \sigma^2 \text{ known},$$
$i = 1, ..., n$.
- The distribution of the MLE is:
$$\widehat{\theta} = \overline{Y} \sim \text{normal}(\theta, V)$$
with $V = \sigma^2/n$,
$$T = \frac{n\overline{Y}^2}{\sigma^2}.$$
- Null and alternative hypotheses are
$$H_0 : \theta = 0, \quad H_1 : \theta \neq 0,$$

A Sanity Check via a Simple Example

- Under H_1 assume the prior $\theta \sim \text{normal}(0, W)$.
- Recall from Lectures 3 and 4, that the evidence in the data for a pair of hypotheses is summarized in the Bayes factor.
- The Bayes factor is
$$\text{BF} = \frac{p(\mathbf{y}|H_0)}{p(\mathbf{y}|H_1)} = \frac{\prod_{i=1}^n \text{dnorm}(y_i|0, \sigma^2)}{\int_\theta \prod_{i=1}^n \text{dnorm}(y_i|\theta, \sigma^2) \times \text{dnorm}(\theta|0, W) d\theta}$$
where dnorm is shorthand for the density of a normal random variable.
- We take $W = \sigma^2$, which corresponds to the "unit information prior" of Kass and Wasserman (1995) (this choice not so important).
- With a prior odds, PO, and ratio of costs R this gives the decision rule to reject H_0:
$$\text{BF} \times \text{PO} = \sqrt{1+n} \times \exp\left(-\frac{T}{2}\frac{n}{1+n}\right) \times \text{PO} < R$$
- Notice how this depends on T and n.

A Bayesian Test Statistics Threshold

- Rearrangement gives a threshold for rejection of:

$$T > \frac{2(1+n)}{n} \log\left(\frac{\mathrm{PO}}{\mathrm{R}}\sqrt{1+n}\right)$$

- For relatively large prior odds on the null PO: require T to be larger (more evidence).
- For relatively large cost of Type II errors R (so that we are averse to type II error, i.e. missing signals): require T to be smaller (less evidence).
- Not such a simply summarization for n but, beyond a certain point, as n gets larger, we require larger T (more evidence).
- The above should be contrasted with the usual frequentist approach of

$$T > \text{const}$$

with the constant usually chosen to control the type I error.

A Bayesian Test Statistic Threshold

- The table below evaluates the probability of rejection given H_0. We assume $R = 1$.
- For $\pi_0 = 0.5$ and $n = 20, 50, 100$ the thresholds give ≈ 0.05 — the situation in which this infamous threshold was first derived?

	$\pi_0 = 0.25$	$\pi_0 = 0.50$	$\pi_0 = 0.95$
$n = 10$	0.64	0.10	0.0025
$n = 20$	0.35	0.074	0.0022
$n = 50$	0.18	0.045	0.0016
$n = 100$	0.12	0.031	0.0011
$n = 1000$	0.030	0.0085	0.00034

Calibration with p-values

Interesting question: When do Bayes and frequentist p-value inference coincide?

- Consider an approximate Bayes approach. We have parameter of interest θ with
  - Data: MLE $\widehat{\theta}$ and standard error $\sqrt{V}$ to give likelihood $\widehat{\theta}|\theta \sim \text{normal}(\theta, V)$,
  - Prior: $\theta \sim \text{normal}(0, W)$.
- The null and alternative hypotheses of interest are

$$H_0 : \theta = 0 \qquad H_1 : \theta \neq 0.$$

- This leads to the approximate Bayes factor (ABF)

$$\begin{aligned} \text{ABF} &= \frac{1}{\sqrt{1-r}} \exp\left(-\frac{Z^2}{2} r\right) \\ &= \sqrt{\frac{V+W}{V}} \exp\left(-\frac{Z^2}{2} \frac{W}{V+W}\right) \end{aligned}$$

where $r = W/(V+W)$ and $Z = \widehat{\theta}/\sqrt{V}$.

Calibration with p-values

- $\text{ABF} = \sqrt{\frac{V+W}{V}} \exp\left(-\frac{Z^2}{2} \frac{W}{V+W}\right)$, which depends on n, because V does.
- Recall we reject if ABF $\times$ PO $>$ threshold R.
- We are trying to find a Bayes factor that does not depend on n, to correspond with a p-value rule.
- We can reverse engineer a version of ABF that does not depend on n by taking the prior variance $W = K \times V$, where K is a constant.
- Then we have approximate Bayes factor

$$\text{ABF}^p = \sqrt{1+K} \exp\left(-\frac{Z^2}{2} \frac{K}{1+K}\right).$$

- Important point: No dependence on n, i.e. it depends on Z only, and therefore on the p-value only.
- If we use the above prior and Bayes factor in multiple tests, then the rankings of p-values and ABF^p will be identical. The problem is that this prior makes no sense.

Calibration with *p*-values

- The ABF with W not depending on n is consistent (you get the right answer with a lot of data), whereas the "p-value" Bayes factor is not.
- The original ABF can be inverted to give a rule for Z^2 that depends on PO, R and n (as with the simple example presented previously).
- For more details, see Wakefield (2009).
- Figure 1 shows the behavior of this rule as a function of the sample size n, and for different choices of the prior on the alternative π_1 and the ratio of costs of type II to type I errors.
- Larger values on the y axis correspond to less extreme test statistics.
- The curves have the expected ordering and, as n gets large, a greater and greater level of evidence is required.
- This is as we would expect because as the sample size increases we want both Type I and Type II errors to go to zero.

A Bayes Factor Threshold

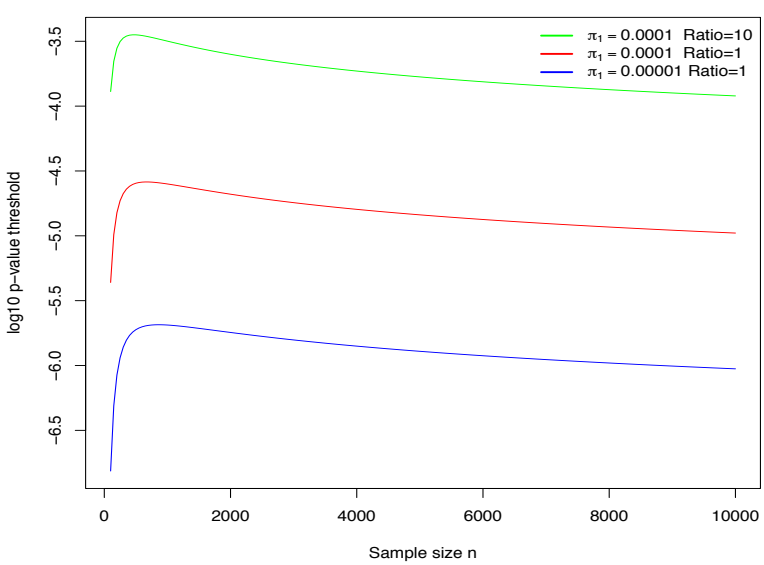


Figure 1 : Threshold for rejection, on the $-\log_{10}(p)$-value scale, versus sample size. Notice how the threshold is decreasing with increasing sample size.

Multiple Testing

The problem: m tests being carried out, often (in a GWAS context, for example) with a tiny probability of any association being non-null.

We describe:

- Family-wise error: Bonferroni and Sidak.
- Control of expected number of false discoveries.
- Control of false discovery rate.
- A Bayesian perspective.

Framework for Multiple Testing

Possibilities with m tests and when K are flagged as requiring further attention:

	Non-Flagged	Flagged	
H_0	A	B	m_0
H_1	C	D	m_1
	$m-K$	K	m

- m_0 is the number of true nulls.
- B is the number of type I errors.
- C is the number of type II errors.

Problem: To select a rule that will determine K.

We discriminate between:

- A sensible criterion.
- How the criterion should depend on sample size.

The Family-Wise Error Rate

	Non-Flagged	Flagged	
H_0	A	B	m_0
H_1	C	D	m_1
	$m-K$	K	m

- The family-wise error rate (FWER) is the probability of making at least one Type I error, i.e.
$$\Pr(B \geq 1 | \text{ all } H_0 \text{ true }).$$
- Let B_i be the event that the i-th null is incorrectly rejected, so that $B = \cup_{i=1}^m B_i$ is the total number of incorrectly rejected nulls.

- The FWER is given by:
$$\begin{aligned} \text{FWER} = \Pr(B \geq 1 | \text{ all } H_0 \text{ true }) &= \Pr\left(\cup_{i=1}^m B_i | \text{ all } H_0 \text{ true }\right) \\ &\leq \sum_{i=1}^m \Pr(B_i | \text{ all } H_0 \text{ true }) \\ &= m\alpha^\star \end{aligned}$$
where $\alpha^\star$ is the level for each test.
- This is true regardless of whether the tests are independent or not.
- Bonferroni takes $\alpha^\star = \alpha/m$ to give FWER $\leq \alpha$.
- Example: For control at $\alpha = 0.05$ with $m = 500K$ tests take $\alpha^\star = 0.05/500,000 = 10^{-7}$.
- Such stringent rules lead to a loss of power, but not ridiculous if you think there is a reasonable chance that all nulls could be true (but α should depend on n, in partcular should decrease as n gets larger and larger).

Sidak Correction

- If all the tests are independent:

$$\begin{aligned}\Pr(B \geq 1) &= 1 - \Pr(B = 0) \\ &= 1 - \Pr(\cap_{i=1}^{m} B_i') \\ &= 1 - \prod_{i=1}^{m} \Pr(B_i') \\ &= 1 - (1 - \alpha^\star)^m\end{aligned}$$

- So to achieve FWER $= \alpha$ take $\alpha^\star = 1 - (1 - \alpha)^{1/m}$ — the Sidak correction (Sidak, 1967).
- Example: with $m = 500K$ tests take

$$\alpha^\star = 1 - (1 - 0.05)^{1/500,000} = 1.03 \times 10^{-7}.$$

Expected Number of False Discoveries

- We describe an alternative criterion.
- For $i = 1, ..., m$ tests let B_i be the 1/0 random variable representing whether the null was incorrectly rejected or not, so that $B = \cup_{i=1}^{m} B_i$.
- The expected number of false discoveries (EFD), with significance level α for each test, is given by

$$\text{EFD} = E[B] = \sum_{i=1}^{m} E[B_i] = m\alpha$$

if all nulls are true.

Expected False Discoveries

For m_0 true nulls: $E[B] = m_0\alpha$, but m_0 is unknown, so all we can say is

$$\text{EFD} = E[B] \leq m\alpha.$$

- In a GWAS context suppose $m = 500K$ and $\alpha = 0.05$; this gives EFD $\leq 25,000$, so conventional levels will clearly not work!
- We can easily put an upper bound on the EFD.
- For example, if we set $\alpha = 1/m$ the expected number of false discoveries is bounded by 1.
- With $\alpha = 5/m$ the expected number of false discoveries is bounded by 5.
- Compare to Bonferroni which controls the FWER via α/m.

False Discovery Rate

- A very popular criterion is the false discovery rate (FDR).

	Non-Flagged	Flagged	
H_0	A	B	m_0
H_1	C	D	m_1
	$m - K$	K	m

- Define the false discovery proportion (FDP) as the proportion of incorrect rejections:

$$\text{FDP} = \begin{cases} \frac{B}{K} & \text{if } K > 0 \\ 0 & \text{if } K = 0 \end{cases}$$

- Then the *false discovery rate* (FDR), the expected proportion of rejected nulls that are actually true nulls, is given by

$$\text{FDR} = \text{E}[\text{FDP}].$$

False Discovery Rate

We describe an algorithm for controlling the FDR.

- Consider the following procedure for independent p-values:
  1. Let $P_{(1)} < ... < P_{(m)}$ denote the ordered p-values.
  2. Define $l_i = i\alpha/m$ and $R = \max\{i : P_{(i)} < l_i\}$ where α is the value at which we would like FDR control.
  3. Then define the p-value threshold as $p_T = P_{(R)}$.
  4. Reject all H_{0i} for which $P_i \leq P_T$.
- Benjamini and Hochberg (1995) show that if this procedure is applied, then regardless of how many nulls are true (m_0) and regardless of the distribution of the p-values when the null is false

$$\text{FDR} \leq \frac{m_0}{m}\alpha < \alpha.$$

False Discovery Rate

- If all the nulls are true then $B = K$ (all rejections are false) and

$$\text{FDR} = E\left[\frac{B}{K}\right] = 1 \times \Pr(B \geq 1) = \text{FWER}.$$

- FDR in this form and with extensions, e.g. Storey and Tibshirani (2003) (description of the q-value methodology) have been successfully used in the microarrays field, where the number of non-null associations is not small.
- Unfortunately less successful in a GWAS, because the proportion of nulls is very close to 1.

Simulated Example

- We illustrate control by the family-wide error rate (FWER), the expected number of false discoveries (EFD) and the false discovery rate (FDR).
- We simulate data for $m = 100$ tests, with $m_1 = 5$ being non-null.
- True table:

	Non-Flagged	Flagged	
H_0	A	B	95
H_1	C	D	5
	$m - K$	K	100

Simulated Example

- We begin by plotting, in Figure 2 the oberved p-values versus those expected under the null, i.e. $i/(m+1)$ for $i = 1, ..., m$.
- Hard to interpret, so we truncate the scales in Figure 3.
- Finally we stretch the scale in Figure 4 to show $-\log_{10}$ the observed p-values versus expected p-values.
- On this scale, a value of 2 corresponds to a p-value of 0.01, and a value of 3 corresponds to a p-value of 0.001.
- We see that the FWER is very conservative ($p = 0.05/100 = 5 \times 10^{-4}$, or $-\log_{10}(p) = 3.30$) and only flags one test as being significant.
- The EFD=1 gives a p-value threshold of 0.01, or $-\log_{10} p = 2$ and picks up all 5 signals.
- The FDR control at 5% gives the green diagonal line and rejects 3 tests.

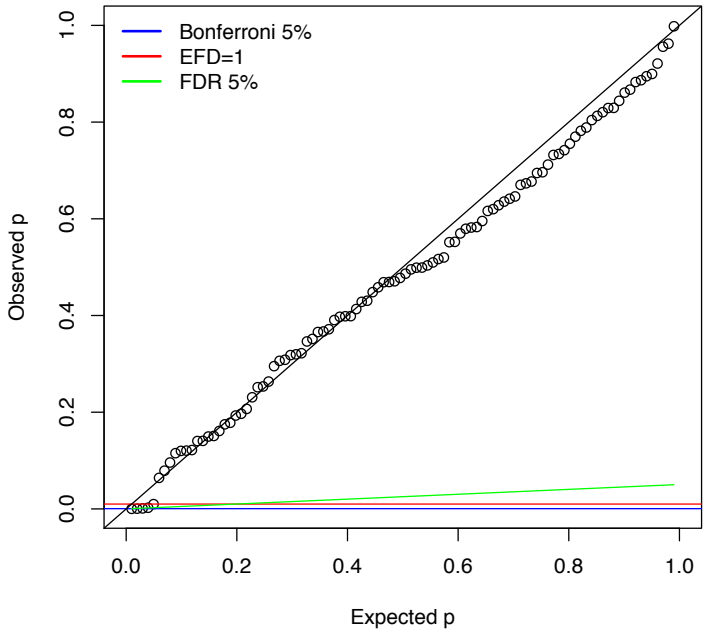


Figure 2 : Observed versus expected *p*-values.

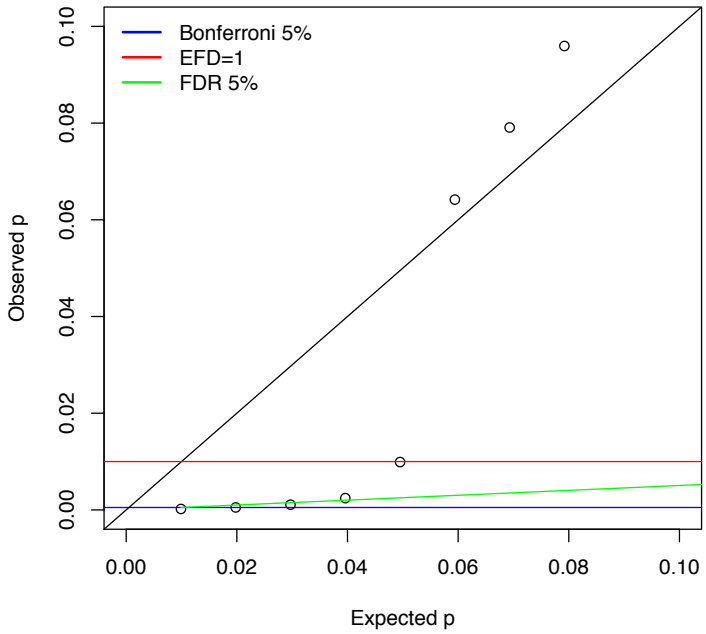


Figure 3 : Observed versus expected *p*-values with truncated scale.

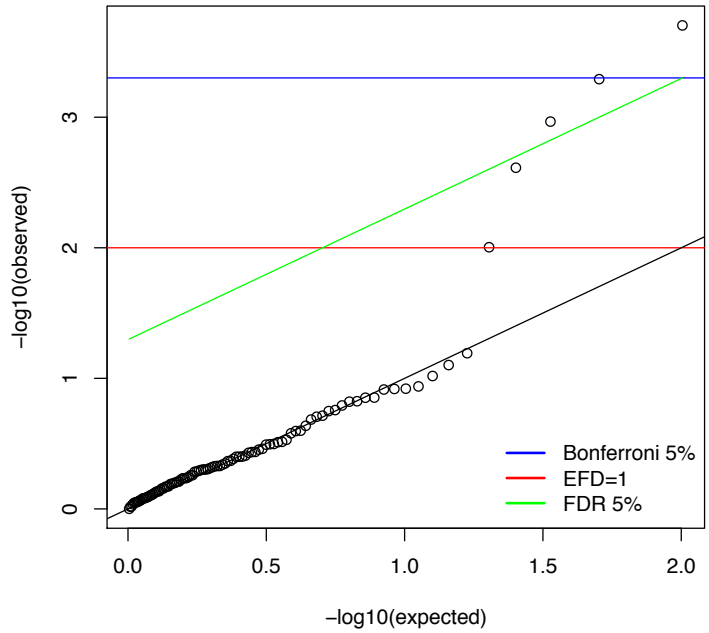


Figure 4 : Observed versus expected p-values, on $-\log_{10}$ scale.

Bayesian False Discoveries/Non-Discoveries

- In a Bayesian approach, based on Bayes factors we have a rule to flag a single association as noteworthy if:

$$\begin{aligned} \text{Posterior Odds} &= \text{Bayes Factor} \times \text{Prior Odds} \\ &< R \end{aligned}$$

  where R is the ratio of costs of type II to type I errors.
- In a multiple testing situation in which m associations are being examined nothing, in principle, changes.
- We simply apply the same rule m times, perhaps changing the priors if we have different priors for different associations.
- The choice of threshold, R, and hence the procedure, does not depend on: the number of tests being carried out[2].

[2] unless the prior on the null, or the ratio of costs of errors depends on the number of tests

Bayesian False Discoveries/Non-Discoveries

- As we have seen, the Bayes factor depends, crucially, on the sample size.
- In contrast, multiple testing based on p-values (e.g. Bonferroni/Sidak) does not depend on the sample size but, crucially, on the number of tests m.
- We have already noted that p-value calibration is very difficult, and we would like a procedure by which p-value thresholds decrease to zero with increasing sample size.
- The same would also be required of EFD or FDR based procedures.
- To summarize in the case of normal test statistics:

  The Bayesian decision is based on the Z score and on the sample size, n, but not on the number of tests, m.

- In contrast:

  The Bonferroni decision is based on the Z score and on the number of tests, m, but not on the sample size, n.

Bayesian Multiple Testing

In a Bayesian context, for a single test:

- If we call a hypothesis noteworthy then $\Pr(H_0|\text{ data })$ is the probability of a false discovery.
- If we call a hypothesis not rejected then $\Pr(H_1|\text{ data })$ is the probability of a false non-discovery.
- In a multiple-hypothesis testing situation (and assuming ordered so the first K are rejected), we have

$$\text{Expected number of false discoveries} = \sum_{i=1}^{K} \Pr(H_{0i}|\text{ data}_i)$$

$$\text{Expected number of false non-discoveries} = \sum_{i=K+1}^{m} \Pr(H_{1i}|\text{ data}_i).$$

A Key Point: A Bayesian analysis of a single SNP alone, or the same SNP from multiple SNPs will produce the same decision (assuming the prior is the same).

Overall Treatment Effect

- We now describe the methodology for the VISP trial.
- Suppose we have two treatments $T = 0/1$ (e.g. low dose/high dose), a continuous response Y and $n/2$ subjects in each treatment group, where n is the number of trial participants.
- Let Y_i be the response for the i-th individual and T_i the treatment indicator.
- To estimate the overall treatment effect we fit the model

$$Y_i = \alpha + \beta T_i + \epsilon_i$$

with $\text{var}(\epsilon_i) = \sigma^2$, so that β is the parameter of interest.
- $H_0 : \beta = 0$ is the null of interest, i.e. no treatment effect?
- Test statistic:

$$Z = \frac{\widehat{\beta}}{\text{s.e.}(\widehat{\beta})} \sim \text{normal}(0,1) \text{ under } H_0$$

where $\widehat{\beta} = \overline{Y}_{HI} - \overline{Y}_{LO}$ and $\text{s.e.}(\widehat{\beta}) = \widehat{\sigma}/\sqrt{n}$.

Treatment-by-Marker Interactions

- Now consider the situation in which we wish to examine the treatment effect by marker.
- To be concrete, define the subgroups relative to a recessive model so that at a generic SNP we have $S = 0$:

No Copies of the Minor Allele

and $S = 1$:

One or Two Copies of the Minor Allele.

- There are therefore m comparisons of interest, with summary data at marker j, as below:

	Marker		
	$S = 0$	$S = 1$	**Sample Size**
$T = 0$	$\overline{Y}_{00}$	$\overline{Y}_{01}$	$n/2$
$T = 1$	$\overline{Y}_{10}$	$\overline{Y}_{11}$	$n/2$
	$n - n_s$	n_s	n

Table 1 : Summary data at a generic marker, under two treatments $T = 0/1$; there are n individuals in total, of which n_s possess the marker of interest.

Treatment-by-Marker Interactions

- Let $S_i = 0/1$ be a marker indicator for individual i and a generic SNP.
- For the treatment effect and at each marker we fit the model

$$Y_i = \alpha + \beta T_i + \gamma S_i + \underbrace{\Delta}_{Interaction} T_i \times S_i + \epsilon_i$$

  with $\text{var}(\epsilon_i) = \sigma^2$.
- $H_0 : \Delta = \Delta_0$ is the null of interest, i.e. is there a differential treatment effect of a certain size at the SNP, e.g. $\Delta_0 = 0$, to compare to the marginal treatment effect.
- Test statistic

$$Z = \frac{\widehat{\Delta} - \Delta_0}{\text{s.e.}(\widehat{\Delta})} \sim \text{normal}(0, 1) \text{ under } H_0.$$

- To emphasize, the same 833/837 responses are used in each of the m comparisons, but they are distributed into the four treatment $\times$ marker cells differently.
- Key Observation: The standard error will vary considerably across SNPs.

VISP Example

- After data cleaning, there were $m = 803,122$ SNPs on which data were available, with at least 5 individuals in each treatment $\times$ marker subgroup.
- Suppose we are interested in detecting marker subgroups for which there is an enhanced effect, i.e. an increased reduction over the marginal treatment effect.
- Figure 5 shows the standard errors in the VISP trial – large variability and so the power ranges considerably also.
- Now refresh memory on the Bayesian approach to testing.

Computation of Bayes Factors

- Recall that

$$\begin{aligned} \widehat{\Delta}|\Delta &\sim N(\Delta, V) \\ \Delta &\sim N(\Delta_0, W). \end{aligned}$$

where $\sqrt{V}$ is the standard error of the estimator leads to a simple form for the Bayes factor:

$$BF = \sqrt{\frac{V+W}{V}} \exp\left(-\frac{Z^2}{2}\frac{W}{V+W}\right)$$

where

$$Z = \frac{\widehat{\Delta} - \Delta_0}{\sqrt{V}}.$$

Bayesian Boundaries

- We again use the Bayes factors as a mechanism by which Z-score boundaries can be calculated, as a function of the standard error $\sqrt{V}$.
- The Bayesian Z^2 score threshold is:

$$Z^2 > z_B^2 = \left(\frac{V+W}{W}\right)\left\{\log\left(\frac{V+W}{V}\right) + 2\log\left(\frac{PO}{R}\right)\right\}$$

to give a threshold which is an explicit function of V, R and PO.
- If the prior odds PO on the null increases, threshold increases: require more evidence.
- If cost of Type II to Type I errors R increase, threshold decreases: require less evidence.

Bayesian Boundaries

- The Bayesian boundary:

$$Z^2 > z_B^2 = \left(\frac{V+W}{W}\right)\left\{\log\left(\frac{V+W}{V}\right) + 2\log\left(\frac{PO}{R}\right)\right\}.$$

- Beyond a certain point, as V decreases the Type I error decreases to zero.
- Specifically, let n denote an appropriate measure of sample size and $V = \sigma^2/n$. Then, as $n \to \infty$,

$$z_B^2 \to \underbrace{\log\left(1 + \frac{nW}{\sigma^2}\right)}_{\to\infty} + 2\log\left(\frac{PO}{R}\right).$$

- Relative to a fixed boundary:
  - For small n/large standard error the Bayesian approach requires more evidence because of the low power.
  - For large n/small standard error the Bayesian approach requires more evidence because of the high power and the comparison with the distribution under H_A.

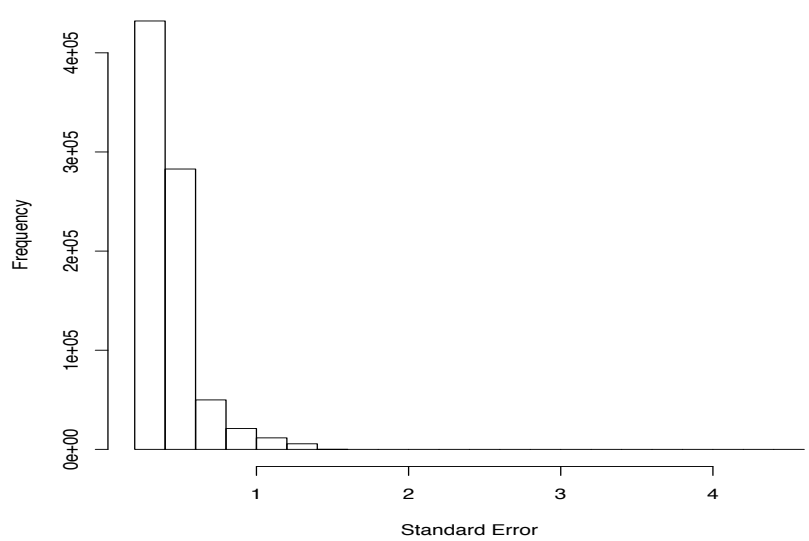


Figure 5 : Histogram of standard errors of the interaction parameter estimates $\widehat{\Delta}$ in the VISP study.

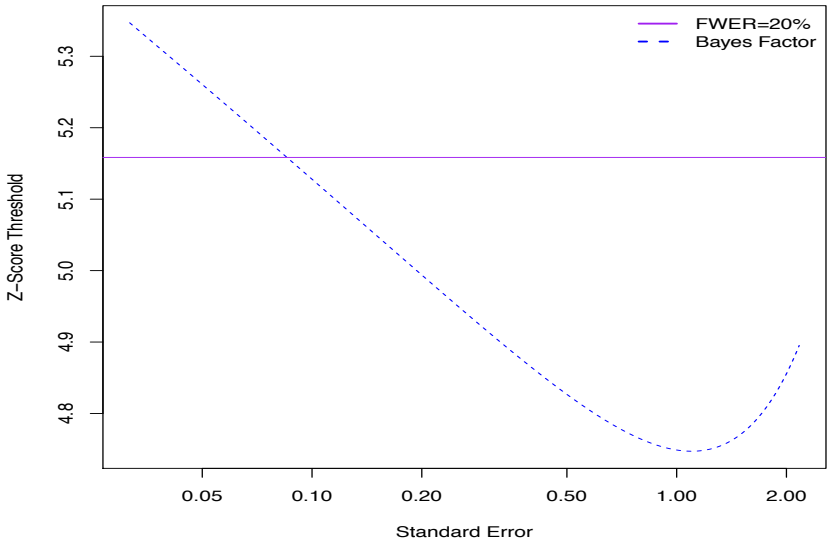


Figure 6 : Bayesian Z-score threshold as a function of the standard error. The Bayesian threshold is based on a prior on the alternative of 0.0001, $R = 1$ and a prior standard deviation on the interaction effect size of $\sqrt{W} = 5.1$; this prior gives a 95% interval on Δ of (-10,10).

A Priori Operating Characteristics

- Ranking is straightforward with Bayes factors, since the only choice is the prior on the effect parameter (W), and inference is relatively insensitive to this value.
- There is much greater sensitivity to the ratio of costs R and the prior odds PO.
- Deciding upon values for R and PO is not straightforward, but only the ratio PO/R is needed.
- We assume $R = 1$ (equal costs of type I and type II errors) and $\pi_1 = 0.001, 0.0001, 0.0001$.
- For $m = 803,122$ SNPs this corresponds to expecting 803, 80 and 8 non-null interactions, respectively.
- These signals will not reflect 803, 80, 8 different causal variants since typically multiple SNPs will tag each causal variant.
- Figure 7 plots various useful operating characteristics.

Operating Characteristics

- To determine the EFD and ETD we require specification of the number of null and non-null signals, which we label as m_0 and m_1, respectively (so that $m = m_0 + m_1$).
- We take the true number of signals as $m_1 = 50$ so that there are $m_0 = 803,072$ null signals.
- Then

$$\begin{aligned} \text{EFD} &= m_0 \times \alpha \\ \text{ETD} &= m_1 \times (1 - \beta) \end{aligned}$$

  where α and β are the type I and type II errors.
- We emphasize that in a GWAS in which the fraction of non-null associations is close to zero, the ETD is highly sensitive to the choice of m_1 (in contrast to EFD, which is insensitive)

Introduction Motivation p-Values and BFs Multiple Testing Methodology for PG PG Example Keyword Prior Conclusions References

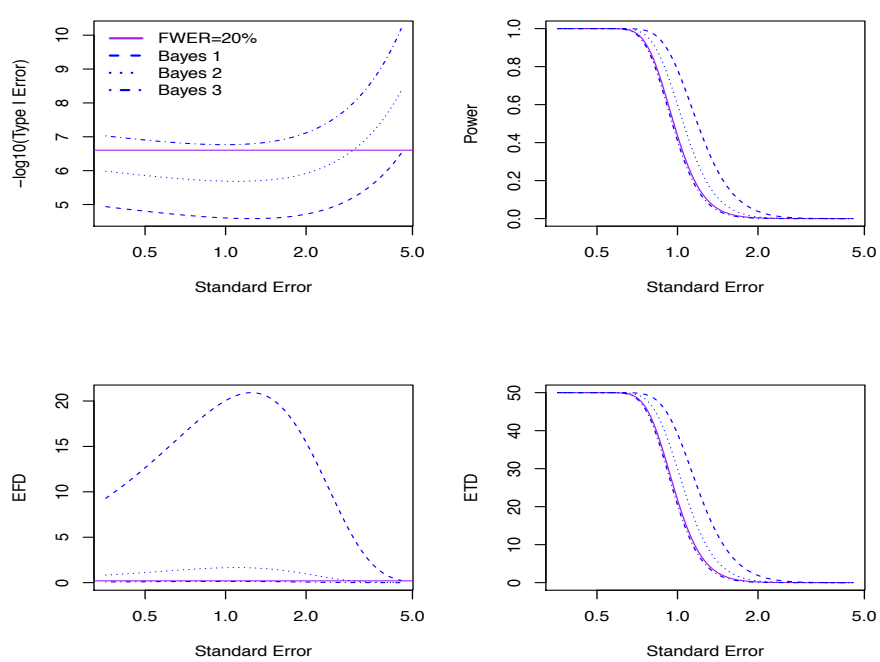


Figure 7 : Operating characteristics of Bayes/Bonferroni. For Bayes boundaries R = 1 and "Bayes 1", "Bayes 2", "Bayes 3" correspond to priors of $\pi_1 = 0.001, 0001, 00001$. Power is to detect a drop of 5 units. For EFD/ETD we set $m_1 = 50$.

Operating Characteristics

- The most liberal prior of $\pi_1 = 0.001$ produces a large number of type I errors (around 20 for standard errors in the mid-range) and might be judged to give unacceptably poor performance.
- The most sceptical prior is more conservative than Bonferroni (with a FWER of 20%) and the prior with $\pi_1 = 0.0001$ is a compromise for this choice of m_1.
- For example: For a standard error of 1, around 2 false discoveries would be expected (as in the lower left panel) but with around 10 more true signals being detected (as seen in the lower right panel), which seems a reasonable trade-off.
- Note, however, that if we think the number of true signals is smaller than $m_1 = 50$ then the number of true signals will fall proportionally.
- For example: At a standard error of 1, if $m_1 = 5$ then we would only expect to detect a single additional signal, when compared to the use of Bonferroni.
- Armed with this information we move to an analysis of the VISP data.

Motivating Homocysteine Example

- We fitted the interaction model with adjustment for age and gender.
- The genetic subgroups are defined as having at least one copy of the minor allele as compared to two copies of the major allele.
- The number in the former subgroup ranges between 21 and 1,564 across SNPs.
- We choose W to give a 95% prior interval for the interactions Δ of ± 10.
- Figure 8 plots the Z-scores versus the standard error, along with boundary corresponding to a FWER of 20%.
- For both the most conservative prior and the Bonferroni approach (with a FWER of 20%, which gives a p-value threshold of 2.5×10^{-7}) two SNPs are flagged.
- With a FWER of 5% the Bonferroni threshold is 6.2×10^{-8} and results in a single SNP being deemed significant.
- With the more optimistic prior of $\pi_1 = 0.0001$, a further signal is flagged (and these are not significant using Bonferroni).

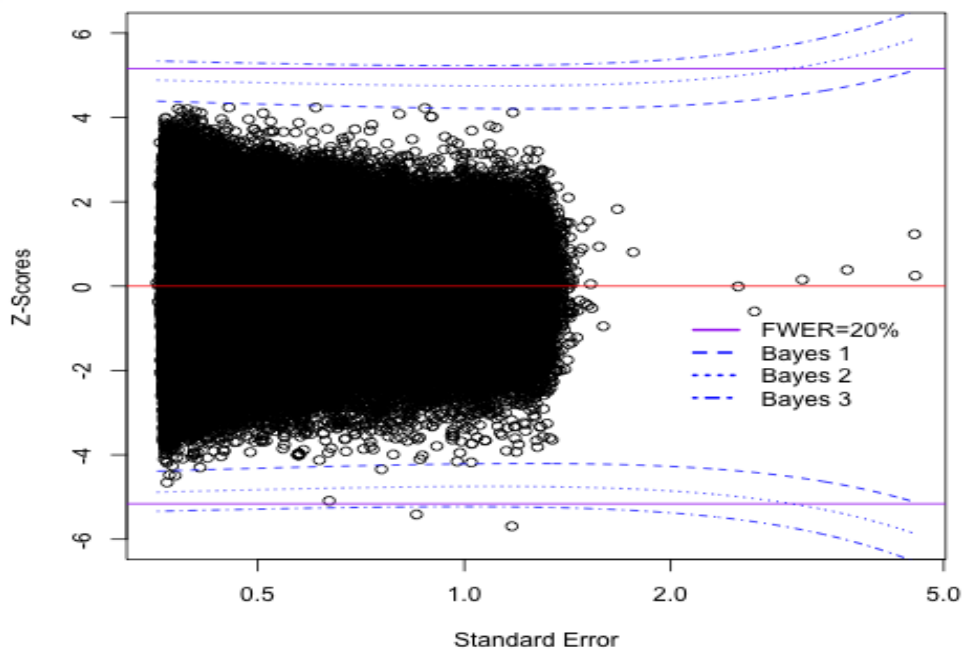


Figure 8 : Z-score threshold as a function of the standard error for the VISP data, ratio of costs of type II to type I errors R = 1 and varying priors on the alternative of $\pi_1 = 0.001, 0001, 00001$ (to give Bayes 1, Bayes 2, Bayes 3 boundaries).

Flagged Signals

SNP ID	Chrom	$\widehat{\Delta}$	$\widehat{\text{s.e.}}(\widehat{\Delta})$	*p*-value	Bayes Factor	Post Prob
rs3736238	17	-6.68	1.38	1.5×10^{-8}	9.3×10^{-7}	0.99
rs16893296	6	-4.61	0.85	7.1×10^{-8}	3.9×10^{-6}	0.96
rs1739317	6	-3.23	0.64	4.0×10^{-7}	2.3×10^{-5}	0.81
rs11819196	10	-1.72	0.37	3.5×10^{-6}	2.9×10^{-4}	0.26

Table 2 : The SNPs in the VISP study that had posterior probabilities on the alternative of greater than 0.25 ($R = 3$), with a prior on the alternative of $\pi_1 = 0.0001$ and under the equal variances recessive genetic model.

VISP Results

- Figure 9 plots the posterior probabilities of the alternative hypothesis (with $\pi_1 = 0.0001$) versus chromosomal position (this is similar to a Manhattan plot in which $-\log_{10}$ p-values are plotted against position).
- The 3 SNPs that fall outside of the boundary in Figure 8 are highlighted.
- The strongest signal is for SNP rs3736238 on chromosome 17. For this SNP there are 42 individuals in the $M = 1$ subgroup, of which 24 and 18 are in the low and high dose groups, respectively.
- The probability of this signal being a false discovery is 0.01 *under* our assumed prior.

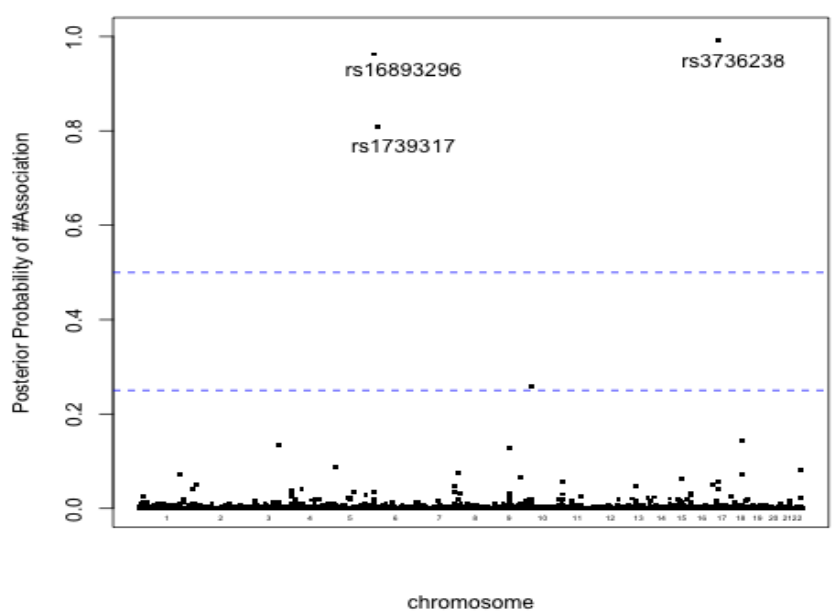


Figure 9 : Posterior probability on the alternative plotted versus genomic position for the VISP data. The prior on the alternative is $\pi_1 = 0.0001$.

VISP Example

- Figure 10 shows that the *p*-values and Bayes factors differ in their rankings due to the differing sample sizes/standard errors.
- The points are color-coded by the size of the standard error and we see that the points with larger standard errors are consistently ranked as giving greater evidence for the alternative under the Bayesian approach.
- This behavior occurs here because of the association between the Z^2 boundary and the standard error for these priors, as shown in Figure 6.
- Specifically, the majority of the signals occur in that portion of the latter curve in which the Bayes boundary lies below the FWER boundary.
- Figure 11 shows an example in which distinctly different behavior occurs.

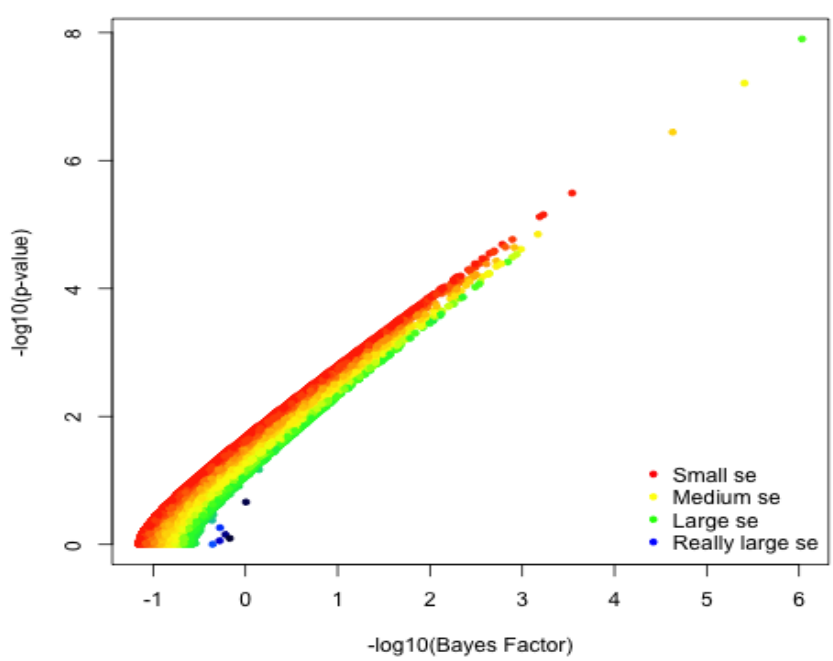


Figure 10 : -$\log_{10}$BFs vs -$\log_{10}$ *p*-values, color-coded by standard error with $W = 10$.

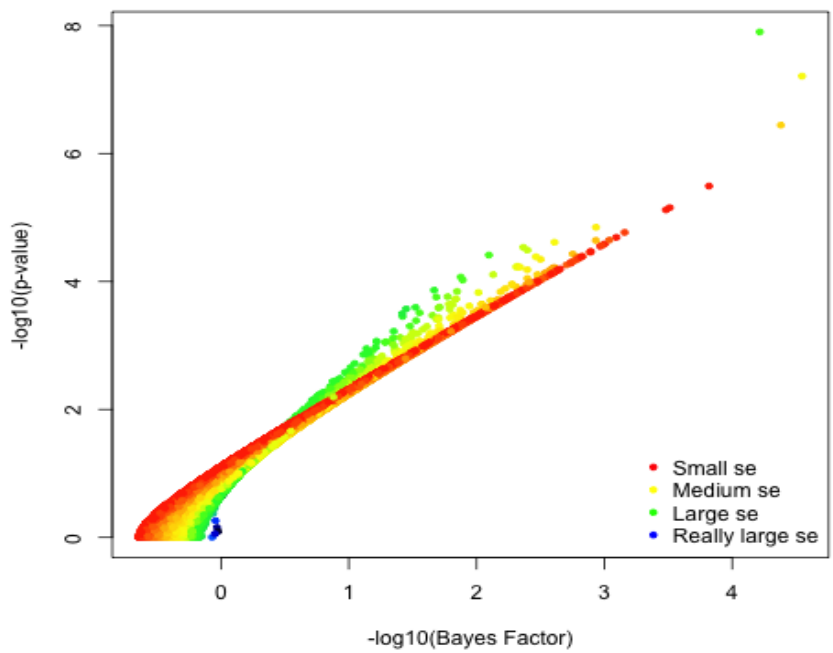


Figure 11 : -$\log_{10}$BFs vs -$\log_{10}$ p-values, color-coded by standard error with $W = 3$.

VISP Example

- A related interesting exercise is to simulate the distribution of observed effect sizes under our assumed priors (on both the proportion of non-null signals and the effect sizes), using the observed distribution of standard errors.
- The distribution of effect sizes is $N(\Delta, V + W)$ for the non-null signals and normal$(0, V)$ for the null signals.
- We can then evaluate the power, and hence determine the number of signals we would expect to detect given our prior assumptions.
- For the VISP data, with a proportion of non-null signals $\pi_1 = 0.0001$, R$=$ 1 and 95% range for the effect sizes of ± 10, we would expect to see 52 true positives and one false positive.
- Given we only observed three non-null signals, this implies that either the range of effect sizes (as defined through W) was too wide or, more probably, that our estimate of π_1 was optimistic.
- Repeating this exercise with $\pi_1 = 0.00001$ gives 5 true positives and close to 0 false positives, which is more consistent with that which was observed.
- Figure 12 gives the posterior probabilities for this prior.

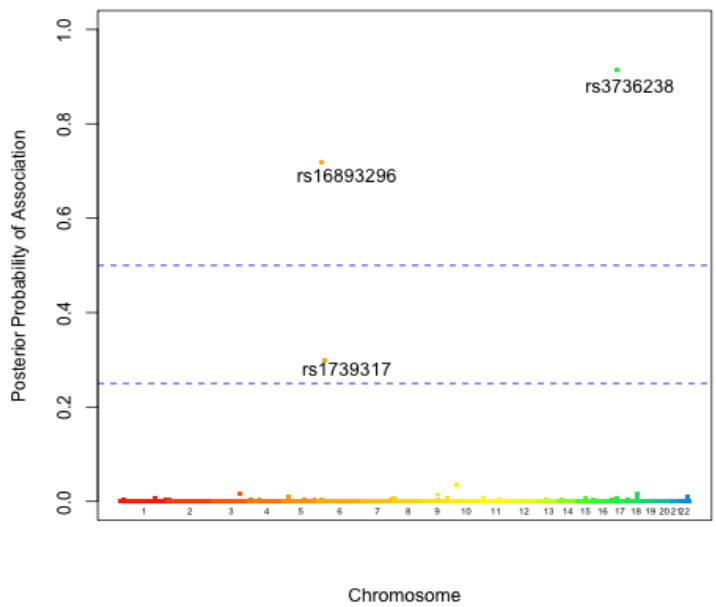


Figure 12 : Posterior probability on the alternative plotted versus genomic region for the VISP data. The prior on the alternative is the more conservative choice of $\pi_1 = 0.00001$.).

VISP Discussion

- We chose the value $\pi_1 = 0.0001$ by examining frequentist summaries before the real data analysis was performed.
- We define π_1 as the proportion of SNPs that would be associated with the disease, if the power were 1.
- After the data are analyzed we can, for those SNPs declared as null (i.e. all but 3 SNPs in the VISP trial), sum up the posterior probabilities of being non-null, and this gives the expected number of false non-discoveries .
- For the VISP data, this expected number is 24.6 so that we are missing a large number of signals, with lack of power being a major issue.
- For the three significant signals, at the 0.5 threshold, the probabilities of the null being true are 0.01, 0.04 and 0.19, so that the expected number of false discoveries is 0.24.
- Taking the threshold of significance as 0.25 gives an additional SNP as being declared significant.
- The sum of the posterior probabilities of the null is 0.98 in this case and so, under this prior, we would expect one of the reports signals to be a false discovery.

Sensitivity to π_1

- The posterior probability of the alternative is highly dependent on the choice of prior on the null π_0, and a sensitivity analysis is always warranted.
- Ideally, rather than fix π_0 as we have done, one would estimate of π_0 from the totality of data (i.e. over all m SNPs), but this is difficult because in a GWAS the proportion of *detectable* null signals is typically very close to 1; there may be many thousands of small but non-zero effects, but the power to detect these signals is low, with the usual sample sizes.
- In other contexts, such as the analysis of gene expression data (Storey and Tibshirani, 2003), the data can be used to estimate π_0 more reliably.
- If the same prior on the null is used for all the tests, the rankings based on the Bayes factor will remain the same as the ranking based on posterior probabilities.
- However, calibrating the Bayes factors to the probability scale requires prior probabilities.
- Within a sensitivity exercise one may include an analysis in which any available information on particular SNPs may be included.

An Alternative Approach to Significance

- The posterior probability (and the Z-score threshold) is equally sensitive to R as to π_1.
- The form of the latter suggests that all we need to do is to fix PO/R.
- As mentioned above, in the VISP analysis we selected π_1 by examining the frequentist operating characteristics.
- An alternative method (Wakefield, 2012) for obtaining PO/R is to specify a value for the Z^2 boundary, z_{B}^2, at a particular V (for example, at a MAF and sample size that one is familiar with) and then solve for $U = \log(\mathrm{PO/R})$ via

$$\widehat{U} = \frac{z_{\mathrm{B}}^2 \times W}{2(V+W)} - \frac{1}{2}\log\left(\frac{V+W}{V}\right).$$

- With this value of $\widehat{U} = \mathrm{PO/R}$ one can then proceed to use

$$Z^2 > z_{\mathrm{B}}^2 = \left(\frac{V+W}{W}\right)\left\{\log\left(\frac{V+W}{V}\right) + 2\log(\widehat{U})\right\}$$

across the observed range of standard errors.

Incorporating Prior Information

- We now briefly describe a method for assigning priors to SNPs based on substantive information.
- In collaboration with scientists at IARC and at the Department of Computer Science at Sheffield University a method had been developed that searches through PubMed abstracts for pre-assigned keywords and key concepts.
- More details inJohansson *et al.* (2012).
- This information is used to assign prior probabilities of association with the phenotype for each SNP of interest.
- Three prior groups were assigned, depending on the number of hits.
- The priors can subsequently be incorporated with the association results of GWAS using the previously described Bayesian framework.
- The method has acronym: Adjusting Association Priors with Text (AdAPT).
- Details of the method can be found in Johansson *et al.* (2012).
- The `AdAPT` software is available here:
  http://services.gate.ac.uk/lld/gwas/service/config

Incorporating Prior Information in a GWAS

- SNPs are assigned to a group, based on the number of keywords that were found to be associated with this SNP.
- For the priors, keywords were ranked by priority: In the:
  - 1st group G_1: were words specific to lung cancer (eg, smoking, lung carcinoma).
  - 2nd group G_2: were more general words specifcally relevant to lung cancer (smoking, nicotine, non-small cell carcinoma),
  - 3rd group G_3: were more general words (carcinogen, DNA damage).
- Each SNP was then placed in one of three prior categories:
  1. $C_1 = \{\text{not } G_1, \text{not } G_2, \text{not } G_3\}$.
  2. $C_2 = \{$ at least one of G_1, G_2, G_3 but not all $\}$.
  3. $C_3 = \{G_1, G_2, G_3\}$.
- We then assigned prior odds (PO) to $\Pr(H_0|C_j)/\Pr(H_1|C_j)$. Specifically for the three categories, the PO was set to 7874 (C_1), 899 (C_2) and 224 (C_3).
- These were used in the analysis to obtain the posterior odds on H_0:

$$\frac{\Pr(H_0|y, C_j)}{\Pr(H_1|y, C_j)} = \frac{\Pr(y|H_0)}{\Pr(y|H_1)} \times \frac{\Pr(H_0|C_j)}{\Pr(H_1|C_j)}.$$

Incorporating Prior Information

- First, the power was evaluated for the three categories, see Figure 13.
- The method was tested by comparing rankings of known susceptibility alleles in a previous lung cancer GWAS of 1989 cases and 2625 controls in 6 central European countries.
- The rankings of 6 SNPs that have been independently replicated in multiple studies were calculated.

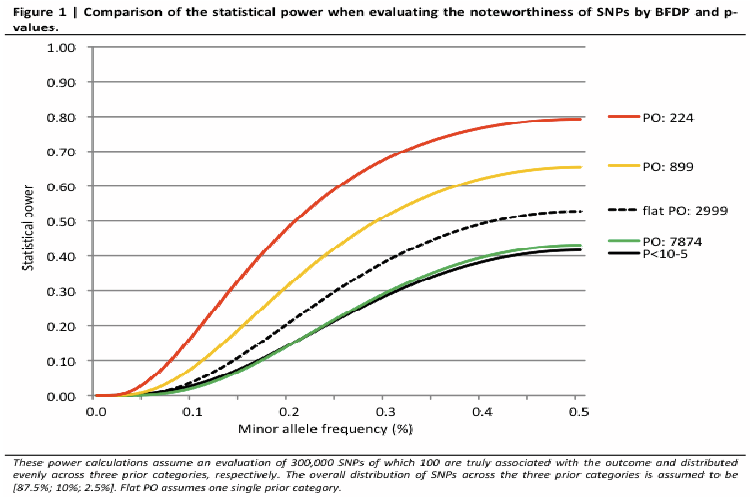


Figure 13 : Power as a function of MAF, for three prior categories, for a single prior, and for a *p*-value approach.

Incorporating Prior Information: Proof of Principle Results

- The results below show that known susceptibility SNPs were ranked more highly by AdAPT BFDPs than by p-values.
- Rankings based on initial data with informative priors for the Bayes rankings:

SNP	*p*-value ranking	Bayes ranking
rs8034191	1	1
rs1051730	2	2
rs4324798	4	5
rs401681	73	30
rs2736100	76	32
rs3117582	121	34

Incorporating Prior Information: New Study Results

- Subsequently, the method was applied on a novel two phase GWAS of oral cancer, with 791 cases and 7,012 controls included in the discovery phase.
- A Bayes threshold on the null of 0.8 was assigned and 6 SNPs passed this test.
- One of these was already replicated, the replication was carried out for the remaining 5 AdAPT ranked SNPs in 1,046 cases and 2,131 controls from 4 case-control studies.
- rs991316, located in the ADH gene region of 4q23, displayed a statistically significant association with oral cancer risk in the replication phase (per-rare-allele log additive *p*-value $=2.5 \times 10^{-3}$).
- This SNP was ranked 76th in the *p*-value list and so would not have been selected to carry forward, but was ranked 4th in the BFDP list.
- The combined odds ratio associated with having one additional rare allele was 0.84 (95% CI: 0.75–0.94).

Conclusions

- Bayesian analysis is attractive in a multiple testing context, but the results are very sensitive to the prior on the proportion of nulls, π_0.
- Fast methods are required for large m (e.g. in a GWAS context) of tests, which is still a drawback for many Bayesian approaches.
- As we have seen, priors can be made a function of characteristics of the SNP (e.g. non-synonymous, previously implicated,...).
- Such priors can have a major impact on rankings and posterior probabilities.
- In genetics, journals are sympathetic to Bayes analyses (not true in all disciplines).
- A huge GWAS enterprise used p-values and Bayes factors to assess significance (Wellcome Trust Case Control Consortium, 2007).
- Stephens and Balding (2009) provide a review of Bayesian approaches in GWAS.

References

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, **57**, 289–300.

Daly, A. (2010). Genome-wide association studies in pharmacogenomics. *Nature Reviews Genetics*, **11**, 214–246.

Johansson, M., Roberts, A., Chen, D., Li, Y., Delahaye-Sourdeix, M., Aswani, N., Greenwood, M., Benhamou, S., Lagiou, P., Holcáová, I., Richiardi, L., Kjaerheim, K., Agudo, A., Castellsagué, X., Macfarlane, T., Barzan, L., Canova, C., Thakker, N. S., ans A. Znaor, D. C., Healy, C., Ahrens, W., Zaridze, D., Szeszenia-Dabrowska, N., Lissowska, J., Fabiánová, E., Mates, I., Bencko, V., Foretova, L., Janout, V., Curado, M., Koifman, S., Menezes, A., W unsch-Filho, V., Eluf-Neto, J., Boffetta, P., Franceschi, S., Herrero, R., Garrote, L., Talamini, R., Boccia, S., Galan, P., Vatten, L., Thomson, P., Zelenika, D., Lathrop, M., Byrnes, G., Cunningham, H., Brennan, P., Wakefield, J., and Mckay, J. (2012). Using prior information from the medical literature in GWAS of oral cancer identifies novel susceptibility variant on chromosome 4 – the AdAPT method. *PLoS One*, **7**.

Kass, R. and Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the schwarz criterion. *Journal of the American Statistical Association*, **90**, 928–934.

Sellke, T., Bayarri, M., and Berger, J. (2001). Calibration of *p* values for testing precise null hypotheses. *The American Statistician*, **55**, 62–71.

Sidak, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, **62**, 626–633.

Stephens, M. and Balding, D. (2009). Bayesian statistical methods for genetic association studies. *Nature Reviews Genetics*, **10**, 681–690.

Storey, J. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Science*, **100**, 9440–9445.

Wakefield, J. (2009). Bayes factors for genome-wide association studies: comparison with *p*-values. *Genetic Epidemiology*, **33**, 79–86.

Wakefield, J. (2012). Commentary: Genome-wide significance thresholds via bayes factors. *International Journal of Epidemiology*, **42**, 286–291.

Wakefield, J., Skrivankova, V., Hsu, F.-C., Sale, M., and Heagerty, P. (2014). Detecting signals in pharmacogenomic studies. *The Pharmacogenomics Journal*. Published online, January 2014.

Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**, 661–678.

Module 4: Bayesian Methods
Lectures 7: Model selection and averaging

Peter Hoff

Departments of Statistics and Biostatistics
University of Washington

Outline

Model selection

Stochastic search

Model selection and averaging

Diabetes example:

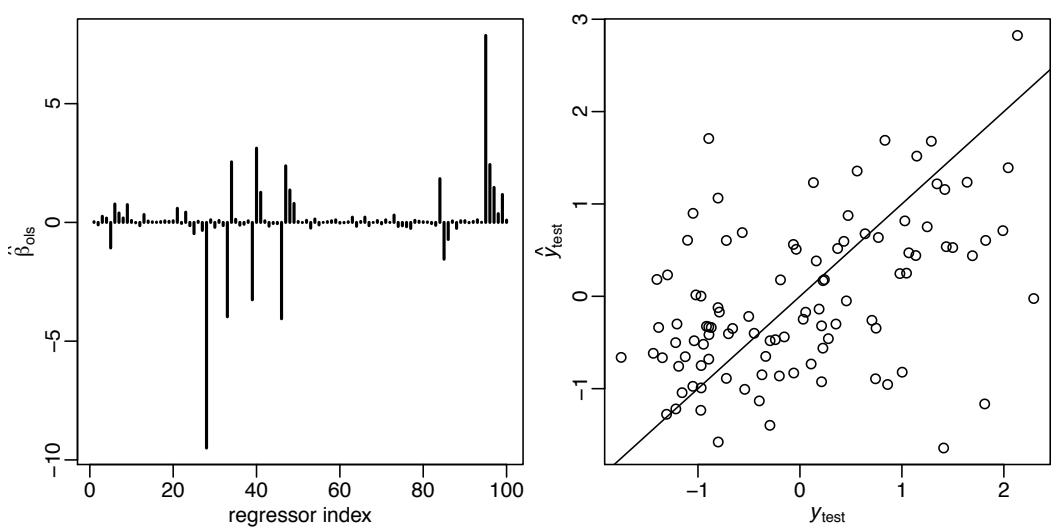
- 342 subjects
- y_i = diabetes progression
- $\mathbf{x}_i$ = explanatory variables.

Each $\mathbf{x}_i$ includes

- 13 subject specific measurements ($x_{\text{age}}, x_{\text{sex}}, \ldots$);
- $78 = \binom{13}{2}$ interaction terms ($x_{\text{age}} \cdot x_{\text{sex}}, \ldots$) ;
- 9 quadratic terms (x_{sex} and three genetic variables are binary)

100 explanatory variables total!

OLS regression



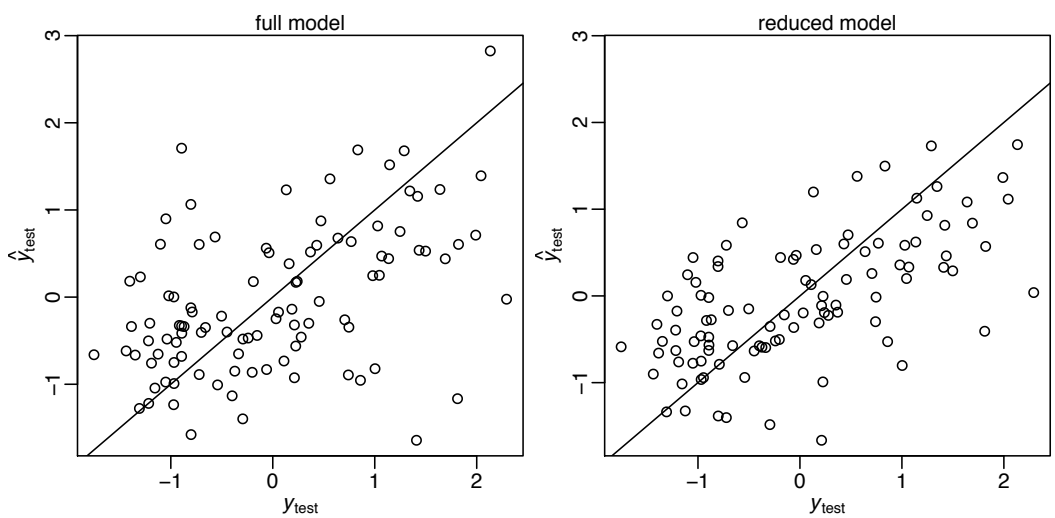
$$\frac{1}{100}\sum(y_{\text{test},i} - \hat{y}_{\text{test},i})^2 = 0.9263$$

$$\frac{1}{100}\sum(y_{\text{test},i} - 0)^2 = \sum y_{\text{test},i}^2 = 1.0095$$

Backwards elimination

1. Obtain the estimator $\hat{\beta}_{\text{ols}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ and its t-statistics.
2. If there are any regressors j such that $|t_j| < t_{\text{cutoff}}$,
   2.1 find the regressor j_{min} having the smallest value of $|t_j|$ and remove column j_{min} from $\mathbf{X}$.
   2.2 return to step 1.
3. If $|t_j| > t_{\text{cutoff}}$ for all variables j remaining in the model, then stop.

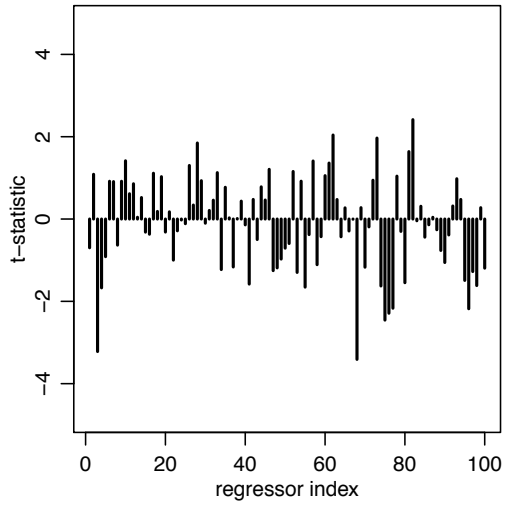
Backwards elimination



$$\frac{1}{100}\sum(y_{\text{test},i} - \hat{y}_{\text{test}^{bel},i})^2 = 0.6392$$

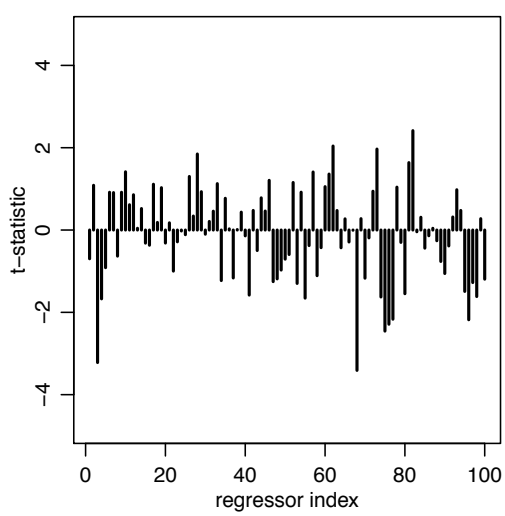
Spurious associations

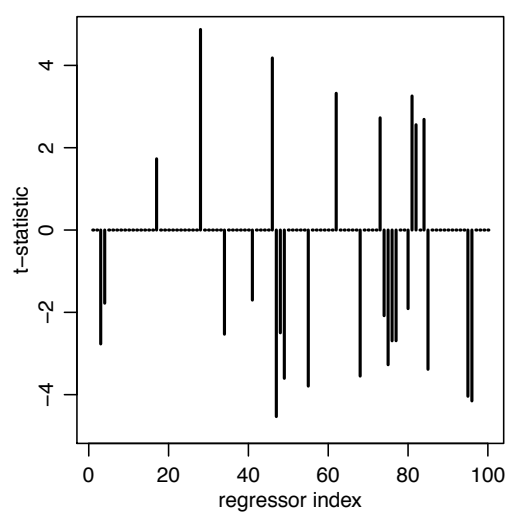
Now try modeling $y_{\pi i} = \boldsymbol{\beta}^T \mathbf{x}_i + \epsilon_i$



Spurious associations

Now try modeling $y_{\pi i} = \boldsymbol{\beta}^T \mathbf{x}_i + \epsilon_i$





Spurious associations

```
sum(abs(t.bslperm)>2 )

## [1] 21

sum(abs(t.bslperm)>3 )

## [1] 12

sum(abs(t.bslperm)>4 )

## [1] 5
```

- 21 regressors have t-stats > 2 ($p \approx 0.05$)
- 12 5 regressors have t-stats > 3 ($p \approx 0.003$)
- 5 regressors have t-stats > 4 ($p \approx 0.00006$)

Bayesian model selection

Prior belief: $\beta_j \approx 0$ for many j's.

Formulation: Write $\beta_j = z_j \times b_j$, where $z_j \in \{0, 1\}$ and $b_j \in \mathbb{R}$.

$$y_i = z_1 b_1 x_{i,1} + \cdots + z_p b_p x_{i,p} + \epsilon_i.$$

For example, in the FTO experiment,

$$\begin{aligned}
\mathrm{E}[Y|\mathbf{x}, \mathbf{b}, \mathbf{z} = (1,0,1,0)] &= b_1 x_1 + b_3 x_3 \\
&= b_1 + b_3 \times \mathrm{age} \\
\mathrm{E}[Y|\mathbf{x}, \mathbf{b}, \mathbf{z} = (1,1,0,0)] &= b_1 x_1 + b_2 x_2 \\
&= b_1 + b_2 \times \mathrm{group} \\
\mathrm{E}[Y|\mathbf{x}, \mathbf{b}, \mathbf{z} = (1,1,1,0)] &= b_1 x_1 + b_2 x_2 + b_3 x_3 \\
&= b_1 + b_2 \times \mathrm{group} + b_3 \times \mathrm{age}.
\end{aligned}$$

Each value of $\mathbf{z} = (z_1, \ldots, z_p)$ corresponds to a *different model*.

Bayesian model comparison

Posterior probability

$$p(\mathbf{z}|\mathbf{y},\mathbf{X}) = \frac{p(\mathbf{z})p(\mathbf{y}|\mathbf{X},\mathbf{z})}{p(\mathbf{y}|\mathbf{X})}$$

Model comparison

$$\begin{array}{rcl} \dfrac{p(\mathbf{z}_a|\mathbf{y},\mathbf{X})}{p(\mathbf{z}_b|\mathbf{y},\mathbf{X})} & = & \dfrac{p(\mathbf{z}_a)}{p(\mathbf{z}_b)} \times \dfrac{p(\mathbf{y}|\mathbf{X},\mathbf{z}_a)}{p(\mathbf{y}|\mathbf{X},\mathbf{z}_b)} \\ \text{posterior odds} & = & \text{prior odds} \times \text{“Bayes factor”} \end{array}$$

Parsimony

The formula for $p(\mathbf{y}|\mathbf{X},\mathbf{z})$ is messy, but

$$\frac{p(\mathbf{y}|\mathbf{X},\mathbf{z}_a)}{p(\mathbf{y}|\mathbf{X},\mathbf{z}_b)} = (1+n)^{(p_{z_b}-p_{z_a})/2}\left(\frac{s^2_{z_a}}{s^2_{z_b}}\right)^{1/2} \times \left(\frac{s^2_{z_b}+\mathrm{SSR}^{z_b}_g}{s^2_{z_a}+\mathrm{SSR}^{z_a}_g}\right)^{(n+1)/2}.$$

A model $\mathbf{z}_a$ is penalized if

- it is too complex (p_A is large)
- it doesn't fit well (SSR^a_g is large)

FTO example

$$\begin{aligned} \mathrm{E}[Y_i|\boldsymbol{\beta}, \mathbf{x}_i] &= \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,4} \\ &= \beta_1 + \beta_2 \times \text{grp}_i + \beta_3 \times \text{age}_i + \beta_4 \times \text{grp}_i \times \text{age}_i \,. \end{aligned}$$

effect of group $\Leftrightarrow$ one of more of β_2, β_4 not zero

z	model	$\log p(\mathbf{y}\|\mathbf{X}, \mathbf{z})$	$p(\mathbf{z}\|\mathbf{y}, \mathbf{X})$
(1,0,0,0)	β_1	-71.82	0
(1,1,0,0)	$\beta_1 + \beta_2 \times \text{grp}_i$	-70.04	0
(1,0,1,0)	$\beta_1 + \beta_3 \times \text{age}_i$	-67.04	0
(1,1,1,0)	$\beta_1 + \beta_2 \times \text{grp}_i + \beta_3 \times \text{age}_i$	-67.04	0.63
(1,1,1,1)	$\beta_1 + \beta_2 \times \text{grp}_i + \beta_3 \times \text{age}_i + \beta_4 \times \text{grp}_i \times \text{age}_i$	-61.72	0.37

$$\begin{aligned} \Pr(\beta_2 \text{ or } \beta_4 \neq 0) &= 0.60 \\ \Pr(\beta_2 \text{ or } \beta_4 \neq 0|\mathbf{y}, \mathbf{X}) &\approx 1 \end{aligned}$$

High dimensional regression

Diabetes example: $p = 100 \Rightarrow 2^{100} \approx 10^{30}$ models to consider.

We can't compute $p(\mathbf{z}|\mathbf{y}, \mathbf{X})$ for each $\mathbf{z}$. Instead, we hope to

- search for models $\mathbf{z}$ with high posterior probability;
- approximate $\beta_j = z_j \times b_j$ for each j;
- build a predictive model for $\mathbf{y}$.

This can be achieved via a Monte Carlo method known as *Gibbs sampling*.

The Gibbs sampler

Goal: A Monte Carlo approximation to $p(x, y, z)$

Given $\{x^{(s)}, y^{(s)}, z^{(s)}\}$,

1. simulate $x^{(s+1)} \sim p(x|y^{(s)}, z^{(s)})$,
2. simulate $y^{(s+1)} \sim p(y|x^{(s+1)}, z^{(s)})$,
3. simulate $z^{(s+1)} \sim p(z|x^{s+1)}, y^{(s+1)})$.

This generates $\{x^{(s+1)}, y^{(s+1)}, z^{(s+1)}\}$.

The Gibbs sampler

Repeated many times, this generates $\{x^{(1)}, y^{(1)}, z^{(1)}\}, \ldots, \{x^{(S)}, y^{(S)}, z^{(S)}\}$

The distribution of this sequence *approximates* $p(x, y, z)$:

$$\begin{aligned}
\frac{1}{S}\sum x^{(s)} &\approx \mathrm{E}[x] = \int x \; p(x, y, z) \; dx \; dy \; dz \\
\frac{\#(x^{(s)} \in A)}{S} &\approx \Pr(x \in A) = \int\int\int_A p(x, y, z) \; dx \; dy \; dz \\
\frac{\#(\{x^{(s)}, y^{(s)}, z^{(s)}\} \in B)}{S} &\approx \int\int_B\int p(x, y, z) \; dx \; dy \; dz
\end{aligned}$$

By necessity, the sequence will frequently visit regions where $p(x, y, z)$ is large.

Gibbs sampling for model selection

Goal Approximate $p(z_1, \ldots, z_p | \mathbf{y}, \mathbf{X})$.

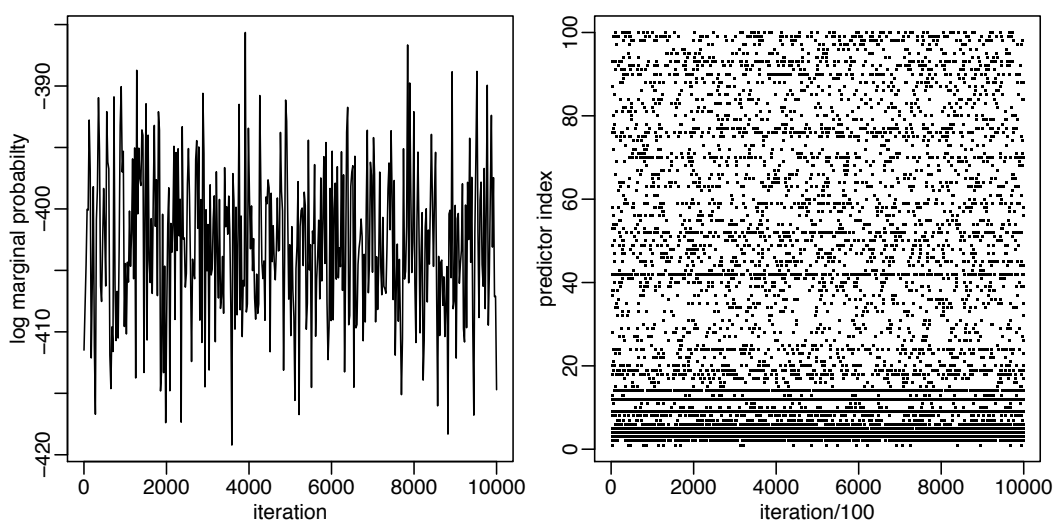
Gibbs sampler: Given $\mathbf{z}^{(s)} = (z_1^{(s)}, \ldots, z_p^{(s)})$,

$$\begin{aligned} z_1^{(s+1)} &\sim p(z_1 | z_2^{(s)}, \ldots, z_p^{(s)}, \mathbf{y}, \mathbf{X}) \\ z_2^{(s+1)} &\sim p(z_2 | z_1^{(s+1)}, z_3^{(s)}, \ldots, z_p^{(s)}, \mathbf{y}, \mathbf{X}) \\ &\vdots \\ z_p^{(s+1)} &\sim p(z_p | z_1^{(s+1)}, \ldots, z_{p-1}^{(s+1)}, \mathbf{y}, \mathbf{X}) \end{aligned}$$

This generates $\mathbf{z}^{(s+1)}$ from $\mathbf{z}^{(s)}$.

Repeating this generates $z^{(1)}, \ldots, z^{(S)}$ with which to approximate $p(\mathbf{z} | \mathbf{y}, \mathbf{X})$.

Diabetes example



Marginal inference

What is the estimate of $\boldsymbol{\beta}$?

Recall

$$\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p) = (b_1 z_1, \ldots, b_p, z_p)$$

Our Monte Carlo samples are

$$\begin{array}{rcccccccc}
\boldsymbol{\beta}^{(1)} & = & (0 & -.299 & 0 & .427 & \cdots & .845) \\
\boldsymbol{\beta}^{(2)} & = & (0 & -.235 & .834 & .374 & \cdots & 0) \\
\vdots & & & & \vdots & & & \\
\boldsymbol{\beta}^{(S)} & = & (0 & -.315 & 0 & .536 & \cdots & 0)
\end{array}$$

A posterior mean for $\boldsymbol{\beta}$ is obtained in the usual way:

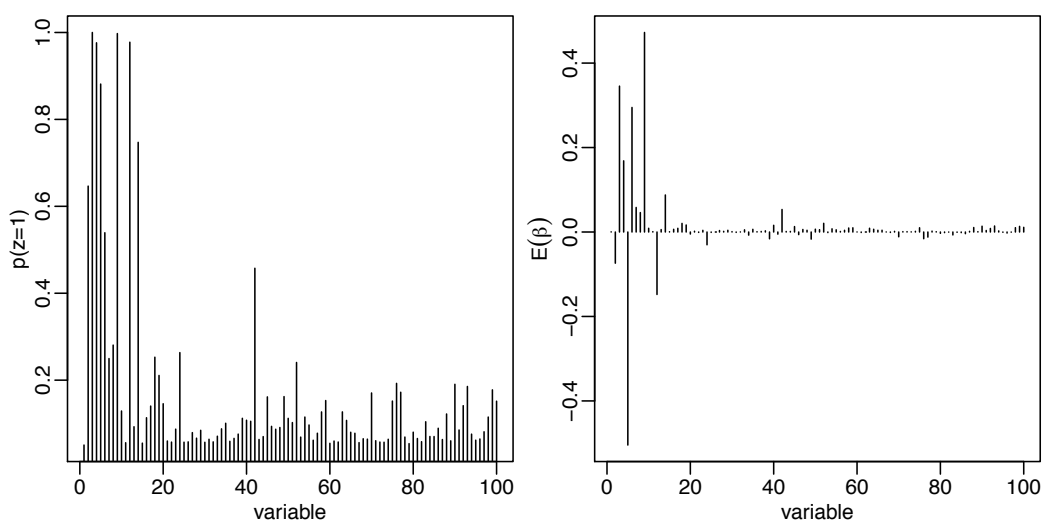
$$\hat{\boldsymbol{\beta}}^{\text{bayes}} = \frac{1}{S} \sum \boldsymbol{\beta}^{(s)} \approx \mathrm{E}[\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}]$$

Out of sample predictions can be made with $\hat{\boldsymbol{\beta}}_{\text{bayes}}$:

$$\hat{y}^{bayes}_{\text{test,i}} = \hat{\boldsymbol{\beta}}^T_{\text{bayes}} \mathbf{x}_{\text{test,i}}$$

Out of sample prediction error: $\frac{1}{S} \sum (y_{\text{test,i}} - \hat{y}^{bayes}_{\text{test,i}})^2 = 0.4853$

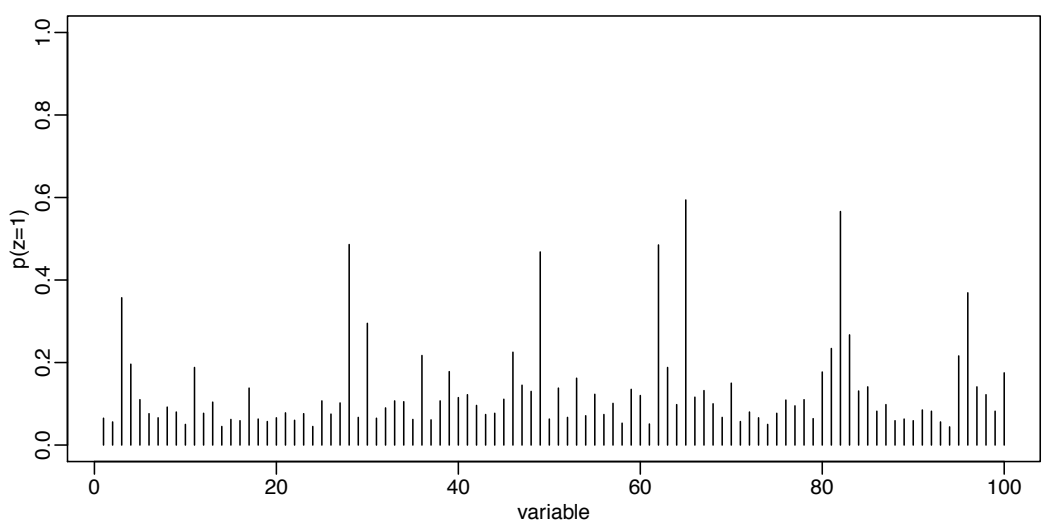
Marginal inference



Important variables

```
colnames(X)[ order(z.pmean,decreasing=TRUE)[1:10] ]

##  [1] "bmi"     "ltg"     "g2"      "map"     "tc"      "sex.age" "sex"
##  [8] "ldl"     "ltg.age" "tch"

colnames(X)[ order(b.pmean,decreasing=TRUE)[1:10] ]

##  [1] "ltg"     "bmi"     "ldl"     "map"     "sex.age" "hdl"     "ltg.age"
##  [8] "tch"     "glu.bmi" "map.sex"
```


Checking the null



2014 SISG Module 4: Bayesian Statistics for Genetics
Lecture 8: Generalized Linear Modeling

Jon Wakefield

Departments of Statistics and Biostatistics
University of Washington

Outline

Introduction and Motivating Examples

Generalized Linear Models
- Definition
- Bayes Linear Model
- Bayes Logistic Regression
- Generalized Linear Mixed Models

Approximate Bayes Inference
- The Approximation
- Case-Control Example

Conclusions

Introduction

- In this lecture we will discuss Bayesian modeling in the context of Generalized Linear Models (GLMs).
- This discussion will include the addition of random effects, i.e. the class of Generalized Linear Mixed Models (GLMMs).
- Estimation via the quick INLA technique will be demonstrated, along with its R implementation.
- An approximation technique that is useful in the context of Genome Wide Association Studies (GWAS) (in which the number of tests is large) will also be introduced.

Motivating Example I: Logistic Regression

- We consider case-control data for the disease Leber Hereditary Optic Neuropathy (LHON) disease with genotype data for marker rs6767450:

	CC $x = 0$	CT $x = 1$	TT $x = 2$	Total
Cases	6	8	75	89
Controls	10	66	163	239
Total	16	74	238	328

- Let $x = 0, 1, 2$ represent the number of T alleles, and $p(x)$ the probability of being a case, given x copies of the T allele.

Motivating Example I: Logistic Regression

- For such case-control data one may fit the multiplicative odds model:

$$\frac{p(x)}{1-p(x)} = \exp(\alpha) \times \exp(\theta x),$$

  with a binomial likelihood.
- Interpretation:
  - $\exp(\alpha)$ is of little interest given the case-control sampling.
  - $\exp(\theta)$ is the odds ratio describing the multiplicative change in risk for one T allele versus zero T alleles.
  - $\exp(2\theta)$ is the odds ratio describing the multiplicative change in risk for two T alleles versus zero T alleles.
  - Odds ratios approximate the relative risk for a rare disease.

R code for Logistic Regression Estimation via MLE

```
> x <- c(0,1,2)
# Case data for CC CT TT
> y <- c(6,8,75)
# Control data for CC CT TT
> z <- c(10,66,163)
#
# Fit logistic regression model, the default choice for
# binomial data
#
> logitmod <- glm(cbind(y,z)~x,family="binomial")
> thetahat <- logitmod$coeff[2]            # Log odds ratio
thetahat
        x
0.4787428
> exp(thetahat) # Odds ratio
       x      # An extra T allele is associated with an increase
1.614044      # of 61% in risk
> V <- vcov(logitmod)[2,2]                 # standard error^2
# Asymptotic confidence interval for odds ratio
> exp(thetahat-1.96*sqrt(V))
       x
0.987916
> exp(thetahat+1.96*sqrt(V))
       x
2.637004
# So 95% interval (just) contains 1.
```

R code for Logistic Regression Hypothesis Testing via a LRT

```
#
# Let's look at a likelihood ratio test of H0: theta=0
#
> logitmod
Call:  glm(formula = cbind(y, z) ~ x, family = "binomial")
Coefficients:
(Intercept)            x
    -1.8077       0.4787
Degrees of Freedom: 2 Total (i.e. Null);  1 Residual
Null Deviance:      15.01
Residual Deviance: 10.99   AIC: 27.79
> pchisq(15.01-10.99,1,lower.tail=F)
[1] 0.04496
# So just significant at the 5% level.
```

- So for these data both estimation and testing point towards borderline significance, at conventional levels.

Motivating Example II: FTO Data Revisited

Linear Model Example

- y = weight
- x_g = fto heterozygote $\in \{0, 1\}$
- x_a = age in weeks $\in \{1, 2, 3, 4, 5\}$

We examine the fit of the model

$$E[Y|x_g, x_a] = \beta_0 + \beta_g x_g + \beta_a x_a + \beta_{int} x_g x_a.$$

```
> fto<-
  read.table("http://www.stat.washington.edu/~hoff/SISG/fto_data.txt",
                    header=TRUE)
> liny <- fto$y
> linxg <- fto$xg
> linxa <- fto$xa
> linxint <- fto$xgxa
> ftodf <- list(liny=liny,linxg=linxg,linxa=linxa,linxint=linxint)
> ols.fit <- lm(liny~linxg+linxa+linxint,data=ftodf)
> summary(ols.fit)
Coefficients:
            Estimate  Std. Error  t value  Pr(>|t|)
(Intercept) -0.06822    1.42230   -0.048    0.9623
linxg        2.94485    2.01143    1.464    0.1625
linxa        2.84421    0.42884    6.632  5.76e-06 ***
linxint      1.72948    0.60647    2.852    0.0115 *
```

Motivating Example III: RNA Seq with Replicates

- We report an experiment carried out in a collaboration with Caitlin Connelly and Josh Akey (UW Genome Sciences), see Connelly *et al.* (2014) for further details.
- Start with two haploid yeast strains (individuals).
- From these we obtain RNA-Seq data, where we isolate RNA from the two individuals, fragment and sequence it using next-generation sequencing, and map the sequencing reads back to the genome to generate RNA levels in the form of counts of the number of sequencing reads mapping at each gene.
- Also mate the two haploid yeast strains together to form a diploid hybrid. We again isolate RNA, fragment, and sequence it.
- Then take advantage of polymorphisms between the two strains in order to map reads to either of the two haploid individuals, giving us counts for the number of reads mapping to either one of the parental genomes in the diploid hybrid for each gene.

Motivating Example III: RNA Seq with Replicates

- We are interested in two questions from this data. First, we want to look for evidence of trans effects at each gene; in biological terms, this means that polymorphisms located far from the gene are responsible for differences in RNA levels.
- To detect this, look for genes where the difference between RNA levels in the haploids differs from the difference between RNA levels for the two parental strains in the diploid.
- Also interested in looking for cis effects, meaning polymorphisms near the gene itself are responsible for differences in RNA levels. We can detect cis effects as a difference in the count of reads mapping to each of the parental strains in the diploid at a gene.

Motivating Example III: RNA Seq Data, Statistical Model

- There are two replicates and so for each of N genes we obtain two sets of counts.
- For the diploid hybrid let Y_{ij} be the number of A alleles for gene i and replicate j, and N_{ij} is the total number of counts, so that $N_{ij} - Y_{ij}$ is the number of T alleles $j = 1, 2$.
- We fit a hierarchical logistic regression model starting with first stage:

$$Y_{ij}|N_{ij}, p_{ij} \sim \text{binomial}(N_{ij}, p_{ij})$$

  so that p_{ij} is the probability of seeing an A read for gene i and replicate j.
- At the second stage:

$$\text{logit } p_{ij} \quad = \quad \theta_i + \epsilon_{ij}$$

  where $\epsilon_{ij} \sim \text{normal}(0, \sigma^2)$ represent random effects that allow for excess-binomial variation.
- In the model θ_i is a parameter of interest – if a (say) 95% posterior interval estimate contains 0 then we have evidence of cis effects.

Generalized Linear Models

- Generalized Linear Models (GLMs) provide a very useful extension to the linear model class.
- GLMs have three elements:
  1. The responses follow an exponential family.
  2. The mean model is linear in the covariates on some scale.
  3. A link function relates the mean of the data to the covariates.
- In a GLM the response y_i are independently distributed and follow an exponential family[1], $i = 1, \ldots, n$.
- Examples: Normal, Poisson, binomial.

[1] so that the distribution is of the form $p(y_i|\theta_i, \alpha) = \exp(\{y_i\theta_i - b(\theta_i)\}/\alpha + c(y_i, \alpha))$, where θ_i and α are scalars

Generalized Linear Models

- The link function $g(\cdot)$ provides the connection between the mean $\mu = \mathrm{E}[Y]$ and the linear predictor $\mathbf{x}\boldsymbol{\beta}$, via

$$g(\mu) = \mathbf{x}\boldsymbol{\beta},$$

  where $\mathbf{x}$ is a vector of explanatory variables and $\boldsymbol{\beta}$ is a vector of regression parameters.
- For normal data, the usual link is the identity

$$g(\mu) = \mu = \mathbf{x}\boldsymbol{\beta}.$$

- For binary data, a common link is the logistic

$$g(\mu) = \log\left(\frac{\mu}{1-\mu}\right) = \mathbf{x}\boldsymbol{\beta}.$$

- For Poisson data, a common link is the log

$$g(\mu) = \log(\mu) = \mathbf{x}\boldsymbol{\beta}.$$

Bayesian Modeling with GLMs

- For a generic GLM, with regression parameters $\boldsymbol{\beta}$ and a scale parameter α, the posterior is

$$p(\boldsymbol{\beta}, \alpha|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\beta}, \alpha) \times p(\boldsymbol{\beta}, \alpha).$$

- An immediate question is: How to specify a prior distribution $p(\boldsymbol{\beta}, \alpha)$?
- How to perform the computations required to summarize the posterior distribution (including the calculation of Bayes factors)?

Bayesian Computation

Various approaches to computation are available:

- Conjugate analysis — the prior combines with likelihood in such a way as to provide analytic tractability (at least for some parameters).
- Analytical Approximations — asymptotic arguments used (e.g. Laplace).
- Numerical integration.
- Direct (Monte Carlo) sampling from the posterior, as we have already seen.
- Markov chain Monte Carlo — very complex models can be implemented, for example within the free software `WinBUGS`.
- Integrated nested Laplace approximation (INLA). Cleverly combines analytical approximations and numerical integration: we illustrate the use of this method in some detail.

Integrated Nested Laplace Approximation (INLA)

- To download INLA:

```
> source("http://www.math.ntnu.no/inla/givemeINLA.R")
> inla.upgrade()
```

- Alternatively, on a mac you can type

```
> install.packages("INLA.tgz", repos=NULL, type="source")
```

- The homepage of the software is here:
  http://www.r-inla.org/home
- There are also lots of example links at this website.
- The fitting of many common models is described here:
  http://www.r-inla.org/models/likelihoods
- INLA can fit GLMs, GLMMs and many other useful model classes.

INLA for the Linear Model

- We first fit a linear model to the FTO data with the default prior settings.

```
> liny <- fto$y
> linxg <- fto$xg
> linxa <- fto$xa
> linxint <- fto$xgxa
#
# Data should be input to INLA as either a list or a dataframe
#
> ftodf <-list(liny=liny,linxg=linxg,linxa=linxa,linxint=linxint)
> formula <- liny~linxg+linxa+linxint
> lin.mod <- inla(formula,data=ftodf,family=''gaussian'')
```

- We might wonder, where are the priors?
- We didn't specify any...but INLA has default choices (more on this later).

INLA for the Linear Model

```
> summary(lin.mod)
Fixed effects:
               mean     sd 0.025quant 0.5quant 0.975quant kld
(Intercept) -0.0609 1.3709    -2.7681  -0.0611     2.6496   0
linxg        2.9323 1.9367    -0.8943   2.9326     6.7593   0
linxa        2.8423 0.4134     2.0255   2.8424     3.6593   0
linxint      1.7329 0.5841     0.5795   1.7328     2.8878   0
Model hyperparameters:
                              mean   sd     0.025quant 0.5quant
Precision for the Gaus obsers 0.3055 0.1018 0.1457     0.2930
                                     0.975quant
Precision for the Gaus observations 0.5402
> summary(ols.fit) # From before!
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.06822    1.42230  -0.048   0.9623
linxg        2.94485    2.01143   1.464   0.1625
linxa        2.84421    0.42884   6.632 5.76e-06 ***
linxint      1.72948    0.60647   2.852   0.0115 *
Residual standard error: 1.917 on 16 degrees of freedom
```

- The posterior means and standard deviations are in very close agreement with the OLS fits presented earlier.

INLA for the Linear Model

```
> summary(lin.mod)
Fixed effects:
                 mean     sd 0.025quant 0.5quant 0.975quant
(Intercept) -0.0609 1.3709     -2.7681  -0.0611      2.6496
linxg        2.9323 1.9367     -0.8943   2.9326      6.7593
linxa        2.8423 0.4134      2.0255   2.8424      3.6593
linxint      1.7329 0.5841      0.5795   1.7328      2.8878
Model hyperparameters:
                                  mean    sd      0.025quant 0.5quant 0.975quant
Precision for Gaus obsers 0.3055 0.1018 0.1457      0.2930   0.5402
```

- The model is

$$Y = \mathrm{E}[Y|x_g, x_a] = \beta_0 + \beta_g x_g + \beta_a x_a + \beta_{\text{int}} x_g x_a + \epsilon$$

  where $\epsilon|\sigma^2 \sim_{iid} N(0, \sigma^2)$.
- The four fixed effects are $\beta_0, \beta_g, \beta_a, \beta_{\text{int}}$ and for each the posterior mean and standard deviation are given along with the 2.5%, 50% and 97.5% quantiles of the posterior.
- The model hyperparameter is the precision σ^{-2}.
- Note that posterior quantiles are invariant to transformation so, for example, the posterior median for σ is $1/\sqrt{0.2930} = 1.85$ (compare with 1.92 from OLS fit).

R Code for Marginal Distributions

- It is straightforward to create plots of the marginal distributions using INLA.
- The code below sends the output to a file, `plot(lin.mod)` sends to a separate window.

```
> plot(lin.mod, pdf=T, prefix="linmod")
```

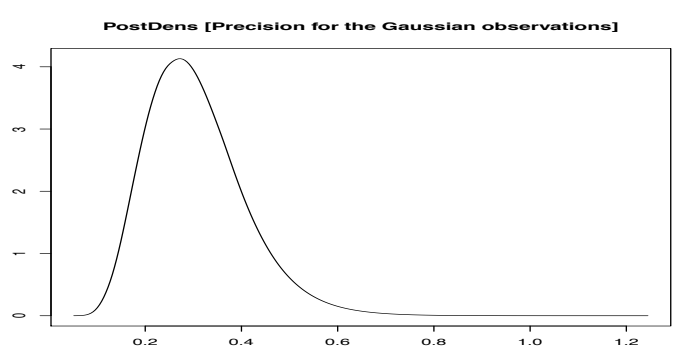


Figure 1 : Marginal distribution of the error precision.

FTO Posterior Marginal Distributions

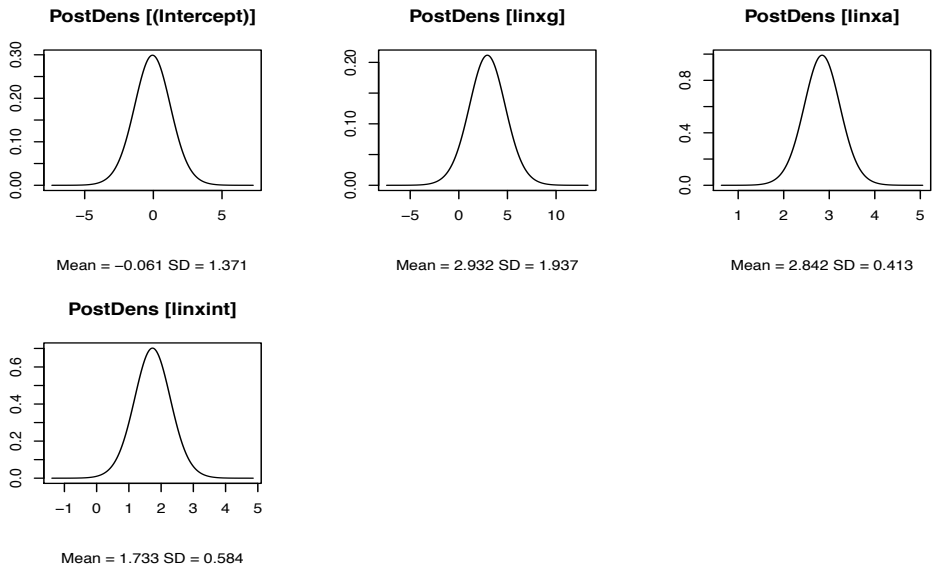


Figure 2 : Marginal distributions of the regression coefficients.

FTO Extended Analysis

- In order to carry out model checking we rerun the analysis, but now switch on a flag to obtain fitted values.

```
> lin.mod <- inla(liny~linxg+linxa+linxint,data=ftodf,
      family="gaussian",control.predictor=list(compute=TRUE))
> names(lin.mod) # Type this to see what can be extracted
> fitted <- lin.mod$summary.fitted.values[,1]
#
# Now extract the posterior median of the measurement error sd
> sigmamed <- 1/sqrt(lin.mod$summary.hyperpar[,4])
```

- With the fitted values we can examine the fit of the model. In particular:
  - Normality of the errors (sample size is relatively small).
  - Errors have constant variance (and are uncorrelated).
  - Linear model is adequate.

Assessing the Model

- The code below forms residuals and then forms
  1. a QQ plot to assess normality,
  2. a plot of residuals versus age, to assess linearity,
  3. a plot of residuals versus fitted values, to see if an unmodeled mean-variance relationship) and
  4. a plot of fitted versus observed for an overall assessment of fit.

```
> residuals <- (liny-fitted)/sigmamed
> par(mfrow=c(2,2))
> qqnorm(residuals,main="")
> title("(a)")
> abline(0,1,lty=2,col="red")
> plot(residuals~linxa,ylab="Residuals",xlab="Age")
> title("(b)")
> abline(h=0,lty=2,col="red")
> plot(residuals~fitted,ylab="Residuals",xlab="Fitted")
> title("(c)")
> abline(h=0,lty=2,col="red")
> plot(fitted~liny,xlab="Observed",ylab="Fitted")
> title("(d)")
> abline(0,1,lty=2,col="red")
```


FTO Diagnostic Plots

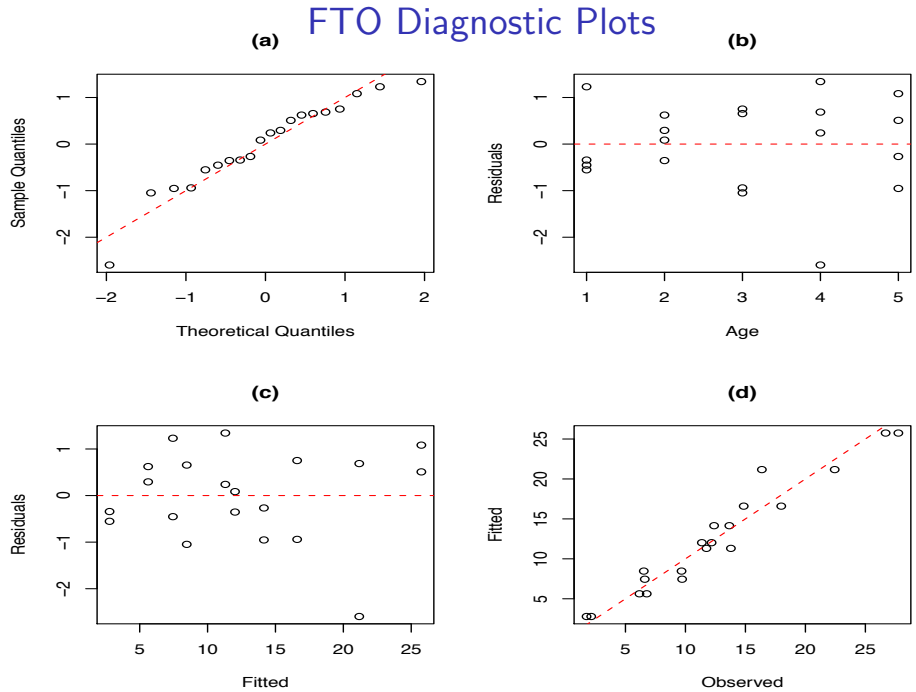


Figure 3 : Plots to assess model adequacy: (a) Normal QQ plot, (b) residuals versus age, (c) residuals versus fitted, (d) fitted versus observed.

Bayes Logistic Regression

- The likelihood is

$$Y(x)|p(x) \sim \text{Binomial}(\, N(x), p(x)\,), \qquad x = 0, 1, 2.$$

- Logistic link:

$$\log\left(\frac{p(x)}{1-p(x)}\right) = \alpha + \theta x$$

- The prior is

$$p(\alpha, \theta) = p(\alpha) \times p(\theta)$$

  with
  - $\alpha \sim \text{normal}(\mu_\alpha, \sigma_\alpha)$ and
  - $\theta \sim \text{normal}(\mu_\theta, \sigma_\theta)$. where $\mu_\alpha, \sigma_\alpha, \mu_\theta, \sigma_\theta$ are constant that are specified to reflect prior beliefs.

Bayes Logistic Regression

- We perform two analyses.
- The first analysis uses the default priors in INLA (which are relatively flat).

```
> x <- c(0,1,2)
> y <- c(6,8,75)
> z <- c(10,66,163)
> cc.dat <- as.data.frame(rbind(y,z,x))
> cc.mod <- inla(y~x,family="binomial",data=cc.dat,Ntrials=y+z)
> summary(cc.mod)
Fixed effects:
              mean     sd 0.025quant 0.5quant 0.975quant
(Intercept) -1.807 0.4554    -2.7487  -1.7903    -0.9593
x            0.480 0.2505     0.0088   0.4726     0.9930
```

Prior Choice for Poisson-Lognormal Models

- It is convenient to specify lognormal priors for a positive parameter $\exp(\beta)$, since one may specify two quantiles of the distribution, and directly solve for the two parameters of the lognormal.
- Denote by $\text{LogNormal}(\mu, \sigma)$ the lognormal distribution for a generic parameter θ with $\text{E}[\theta] = \mu$ and $\text{var}(\log\theta) = \sigma^2$, and let θ_1 and θ_2 be the q_1 and q_2 quantiles of this prior.
- In our example, $\theta = \exp(\beta)$.
- Then it is straightforward to show that

$$\mu = \log(\theta_1)\left(\frac{z_{q_2}}{z_{q_2} - z_{q_1}}\right) - \log(\theta_2)\left(\frac{z_{q_1}}{z_{q_2} - z_{q_1}}\right), \ \sigma = \frac{\log(\theta_1) - \log(\theta_2)}{z_{q_1} - z_{q_2}}. \tag{1}$$

Prior Choice for Poisson-Lognormal Models

- As an example, suppose that for the odds ratio e^{β} we believe there is a 50% chance that the odds ratio is less than 1 and a 95% chance that it is less than 5; with $q_1 = 0.5, \theta_1 = 1.0$ and $q_2 = 0.95, \theta_2 = 5.0$, we obtain lognormal parameters $\mu = 0$ and $\sigma = (\log 5)/1.645 = 0.98$.
- There is a function in the `SpatialEpi` R package to find the parameters, as we illustrate.

```
library(SpatialEpi}
lnprior <- LogNormalPriorCh(1,5,0.5,0.95)
lnprior
$mu
[1] 0
$sigma
[1] 0.9784688
plot(seq(0,7,.1),dlnorm(seq(0,7,.1),meanlog=lnprior$mu,
 sdlog=lnprior$sigma),type="l",xlab="x",ylab="LogNormal Density")
```

The density is shown in Figure 4.

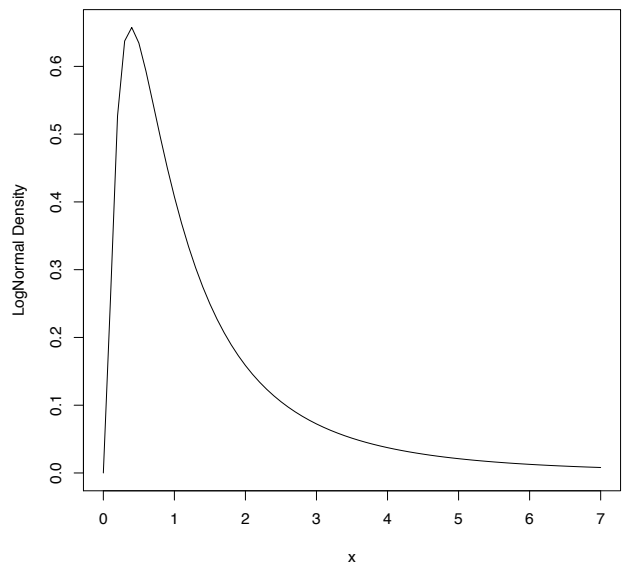


Figure 4 : Lognormal density with 50% point 1 and 95% point 5.

R Code for Logistic Regression Example

- In the second analysis we specify

$$\begin{aligned} \alpha &\sim \text{normal}(0, 1/0.1) \\ \theta &\sim \text{normal}(0, W) \end{aligned}$$

where W is such that the 97.5% point of the prior is log(1.5), i.e. we believe the odds ratio lies between 2/3 and 3/2 with probability 0.95.

```
# Now with informative priors
> W <- LogNormalPriorCh(1,1.5,0.5,0.975)$sigma^2
> cc.mod2 <- inla(y~x,family="binomial",data=cc.dat,Ntrials=y+z,
    control.fixed=list(mean.intercept=c(0),prec.intercept=c(.1),
                       mean=c(0),prec=c(1/W)))
> summary(cc.mod2)
Fixed effects:
                mean     sd 0.025quant 0.5quant 0.975quant
(Intercept) -1.3227 0.2896    -1.9005  -1.3194    -0.7641
x            0.1986 0.1536    -0.0999   0.1977     0.5027
plot(cc.mod2,pdf=T,prefix=''logistic'')
```

The quantiles for θ can be translated to odds ratios by exponentiating.

Logistic Marginal Plots

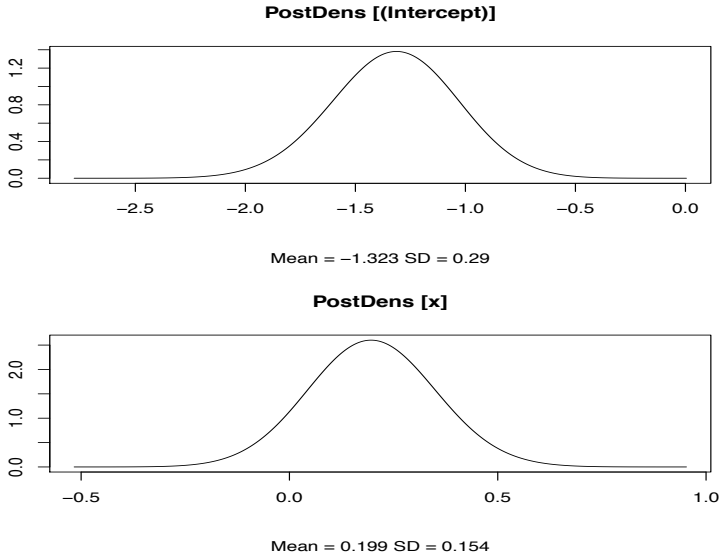


Figure 5 : Posterior marginals for the intercept α and the log odds ratio θ.

A Simple ANOVA Example

- We begin with simulated data from the simple one-way ANOVA model example:

$$\begin{aligned} Y_{ij}|\beta_0, b_i &= \beta_0 + b_i + \epsilon_{ij} \\ \epsilon_{ij}|\sigma_\epsilon^2 &\sim_{iid} \text{normal}(0, \sigma_\epsilon^2) \\ b_i|\sigma_b^2 &\sim_{iid} \text{normal}(0, \sigma_b^2) \end{aligned}$$

  $i = 1, ..., 10; j = 1, ..., 5$, with $\beta_0 = 0.5$, $\sigma_\epsilon^2 = 0.2^2$ and $\sigma_b^2 = 0.3^2$.
- b_i are random effects and ϵ_{ij} are measurement errors and there are two variances to estimate, σ_ϵ^2 and σ_b^2.
- In a fixed effects Bayesian model, the variance σ_b^2 would be fixed in advance.
- Simulation:

```
> sigma.b <- 0.3
> sigma.e <- 0.2
> m <- 10
> ni <- 5
> beta0 <- 0.5
> b <- rnorm(m, mean=0, sd=sigma.b)
> e <- rnorm(m*ni, mean=0, sd=sigma.e)
> Yvec <- beta0 + rep(b, each=ni) + e
> simdata <- data.frame(y=Yvec, ind=rep(1:m, each=ni))
```

A Simple ANOVA Example

- We fit the one-way ANOVA model and see reasonable recovery of the true values that were used to simulate the data.
- Not a big surprise, since we have fitted the model that was used to simulate the data!

```
> result <- inla(y ~ f(ind, model="iid"), data = simdata)
> summary(result)
Fixed effects:
                 mean          sd 0.025quant  0.5quant   0.975quant
(Intercept) 0.3780418 0.09710096 0.1828018   0.3780368  0.5734883
Model hyperparameters:
                      mean   sd      0.025quant 0.5quant 0.975quant
Precision for Gaus obs 22.816 4.943  14.439     22.389   33.755
Precision for ind      13.983 6.857  4.784      12.632   31.164
> sigma.est <- 1/sqrt(result$summary.hyperpar[,4])
> sigma.est
Gaus obs    ind          # Extract the posterior medians
0.2119460   0.2795877    # of the precision and invert
```

- `sigma.est` correspond to the posterior medians of σ_ϵ and σ_b, respectively.

The RNA-Seq Data: INLA Analysis

- Recall there are two replicates and so for each of N genes we obtain two sets of counts.
- For the diploid hybrid, let Y_{ij} be the number of A alleles for gene i and replicate j, and N_{ij} is the total number of counts, $j = 1, 2$.
- We fit a hierarchical logistic regression model starting with first stage:

$$Y_{ij}|N_{ij}, p_{ij} \sim \text{binomial}(N_{ij}, p_{ij})$$

  so that p_{ij} is the probability of seeing an A read for gene i and replicate j.
- At the second stage:

$$\text{logit } p_{ij} \quad = \quad \theta_i + \epsilon_{ij}$$

  where $\epsilon_{ij}|\sigma^2 \sim \text{normal}(0, \sigma^2)$ represent random effects that allow for excess-binomial variation; there are a pair for each gene.
- The θ_i parameters are taken as fixed effects with a relatively flat prior (the default choice in `INLA`).
- $\exp(\theta_i)$ is the odds of seeing an A read for gene i.

INLA Code for RNA-Seq Data: Data Setup

- Rows 1 and 2 represent the two replicates for gene 1, rows 3 and 4 represent the two replicates for gene 2, etc...
- `rep1` is the variable that defines the random effects.
- `xvar` is the gene number, there are 10 genes in this dataset.

```
repdat <- read.table("RNA-SeqN10.txt",header=T)
repdat
      y     n rep1 xvar
1  1963  7617    1    1
2  3676 10413    2    1
3   249   308    3    2
4   110   114    4    2
...
17  397   810   17    9
18  480   928   18    9
19  242   466   19   10
20  174   313   20   10
```


INLA Code for RNA-Seq Data: Model Fitting

- Below is the code for fitting the random effects model.
- The `-1` in the model specification removes the intercept, so that the factor levels are defined with one level for each gene.

```
> RNAdat <- data.frame(repdat)
> RNAfit <- inla(y~as.factor(xvar)-1+f(rep1,model="iid"),
          family="binomial",data=RNAdat,Ntrials=n)
> summary(RNAfit)
Fixed effects:
                      mean     sd 0.025quant 0.5quant 0.975quant
as.factor(xvar)1   -0.8316 0.2298    -1.2923  -0.8315    -0.3712
as.factor(xvar)2    2.0222 0.2932     1.4748   2.0105     2.6364
as.factor(xvar)3    0.1730 0.2559    -0.3351   0.1728     0.6822
as.factor(xvar)4    0.4302 0.2317    -0.0342   0.4303     0.8938
as.factor(xvar)5    0.5964 0.2304     0.1348   0.5964     1.0580
as.factor(xvar)6    0.5456 0.2347     0.0760   0.5456     1.0150
as.factor(xvar)7    1.3631 0.2826     0.8126   1.3599     1.9316
as.factor(xvar)8    0.0280 0.2308    -0.4345   0.0280     0.4903
as.factor(xvar)9    0.0149 0.2342    -0.4537   0.0150     0.4834
as.factor(xvar)10   0.1497 0.2406    -0.3302   0.1496     0.6303
```

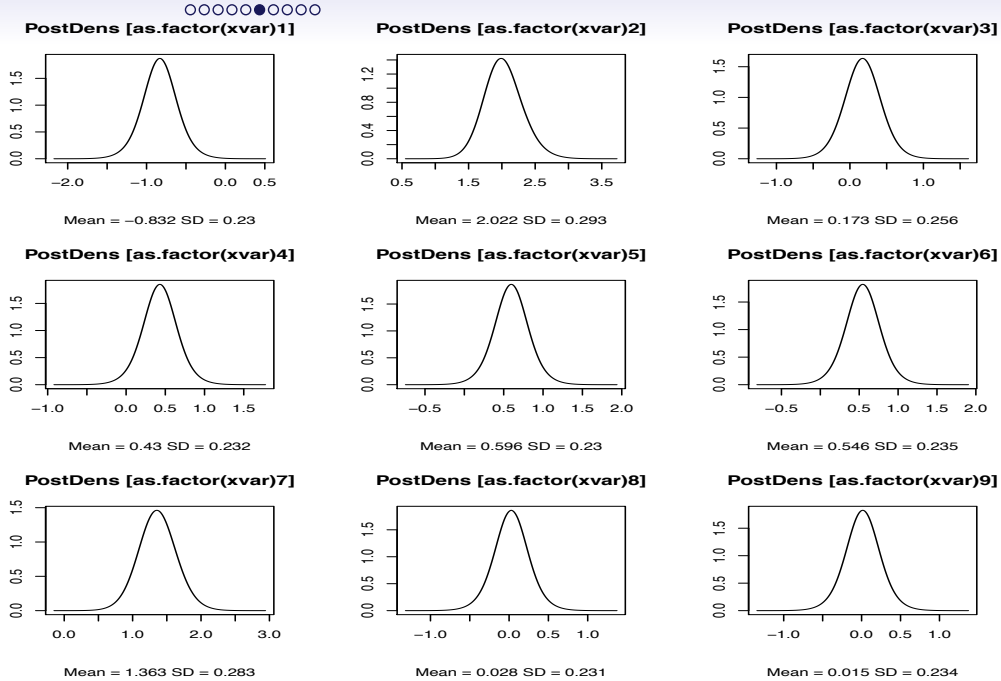


Figure 6 : Posterior marginals for the first 9 gene effects (compare with zero for evidence of cis effects). We plot 9 rather than all 10 for display purposes.

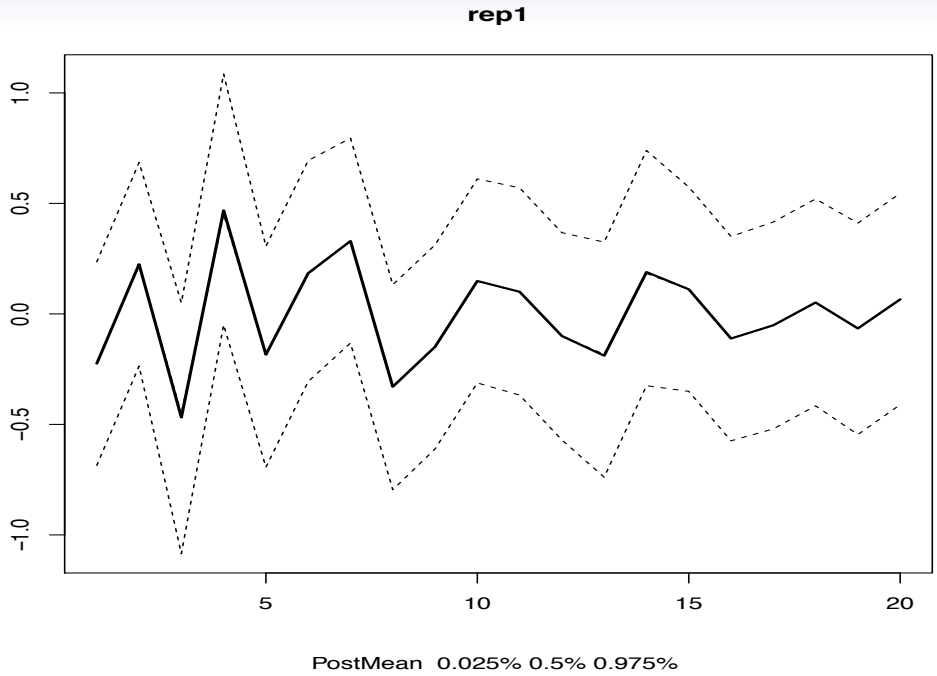


Figure 7 : Posterior quantiles for 20 random effects, which allow excess-binomial variation.

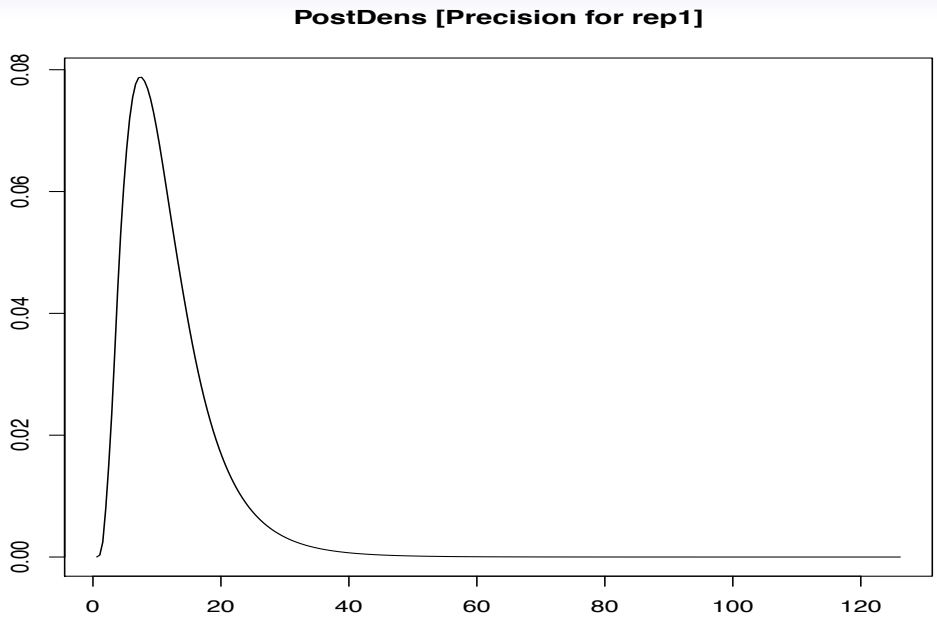


Figure 8 : Posterior marginal for precision of random effects.

An Informative Summary for the RNA-Seq Data

- We extract the 95% intervals and posterior medians for the log odds of being an A allele.
- Comparison with 0 (in Figure 9) gives an indication of cis effects.
- Genes 1, 2, 5, 6, 7 show evidence of cis effects.

```
thetasum <- RNAfit$summary.fixed[,3:5]
par(mfrow=c(1,1))
plot(thetasum[,2],seq(1,10),xlim=c(min(thetasum),max(thetasum)),
      ylab="Gene",xlab="Log Odds")
for (i in 1:10){
   lines(x=c(thetasum[i,1],thetasum[i,3]),y=c(i,i))
}
abline(v=0,col="red") # Intervals to the left/right of this line?
```

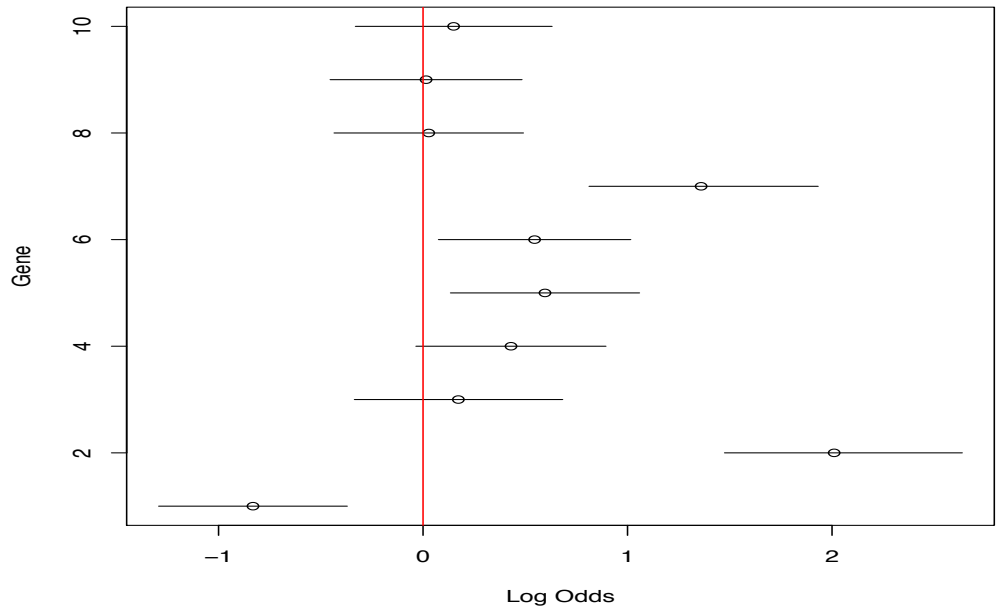


Figure 9 : Posterior marginal intervals for posterior of interest. Genes with posterior intervals that do not include zero, show evidence of cis effects.

Approximate Bayes Inference

- Particularly in the context of a large number of experiments, a quick and accurate model is desirable.
- We describe such a model in the context of a GWAS.
- This model is relevant when the sample size in each experiment is large.
- We first recap the normal-normal Bayes model.
- Subsequently, we describe the approximation and provide an example.

Recall: The Normal-Normal Model

- For the model
  - Prior: $\theta \sim$ normal(μ_0, τ_0^2) and
  - Likelihood: $Y_1, ..., Y_n|\theta \sim$ normal(θ, σ^2).
- Posterior

$$\theta|y_1, ..., y_n \sim \text{normal}(\mu, \tau^2)$$

where

$$\begin{aligned} \text{var}(\theta|y_1, ..., y_n) = \tau^2 &= [1/\tau_0^2 + n/\sigma^2]^{-1} \\ \text{Precision} = 1/\tau^2 &= 1/\tau_0^2 + n/\sigma^2 \end{aligned}$$

and

$$\begin{aligned} \text{E}[\theta|y_1, ..., y_n] = \mu &= \frac{\mu_0/\tau_0^2 + \bar{y}n/\sigma^2}{1/\tau_0^2 + n/\sigma^2} \\ &= \mu_0 \left(\frac{1/\tau_0^2}{1/\tau_0^2 + n/\sigma^2} \right) + \bar{y} \left(\frac{n/\sigma^2}{1/\tau_0^2 + n/\sigma^2} \right) \end{aligned}$$

A Normal-Normal Approximate Bayes Model

- Consider again the logistic regression model

$$\text{logit } p_i = \alpha + x_i\theta$$

with interest focusing on θ.

- We require priors for α, θ, and some numerical/analytical technique for estimation/Bayes factor calculation.
- As discussed in Lecture 6 Wakefield (2007, 2009) considered replacing the likelihood by the approximation

$$p(\theta|\widehat{\theta}) \propto p(\widehat{\theta}|\theta)p(\theta)$$

where
  - $\widehat{\theta}|\theta \sim$ normal(θ, V) – the asymptotic distribution of the MLE,
  - $\theta \sim$ normal$(0, W)$ – the prior on the log RR. Can choose W so that 95% of relative risks lie in some range, e.g. [2/3,1.5].

Posterior Distribution

- Under the alternative, the posterior distribution for the log odds ratio θ is

$$\theta|\widehat{\theta} \sim \text{normal}(r\widehat{\theta}, rV)$$

where

$$r = \frac{W}{V+W}.$$

- Hence, we have shrinkage to the prior mean of 0.
- The posterior median for the odds ratio is $\exp(r\widehat{\theta})$ and a 95% credible interval is

$$\exp(r\widehat{\theta} \pm 1.96\sqrt{rV}).$$

- Note that as $W \to \infty$ and/or $V \to 0$ (which occurs as we gather more data) the non-Bayesian point and interval estimates are recovered (since $r \to 1$).

A Normal-Normal Approximate Bayes Model

- We are interested in the hypotheses: $H_0 : \theta = 0$, $H_1 : \theta \neq 0$ and evaluation of the Bayes factor

$$\text{BF} = \frac{p(\widehat{\theta}|H_0)}{p(\widehat{\theta}|H_1)}.$$

- Using the approximate likelihood and normal prior we obtain:

$$\text{Approximate Bayes Factor } = \frac{1}{\sqrt{1-r}} \exp\left(-\frac{Z^2}{2} r\right),$$

with $Z = \frac{\widehat{\theta}}{\sqrt{V}}$, $r = \frac{W}{V+W}$.

A Normal-Normal Approximate Bayes Model

- The approximation can be combined with a Prior Odds $= \pi_0/(1-\pi_0)$ to give
$$\text{Posterior Odds on } H_0 = \frac{\text{BFDP}}{1-\text{BFDP}} = \text{ABF} \times \text{Prior Odds}$$
where BFDP is the Bayesian False Discovery Probability.
- BFDP depends on the power, through r.
- For implementation, all that we need from the data is the Z-score and the standard error $\sqrt{V}$, or a confidence interval.
- Hence, published results that report confidence intervals can be converted into Bayes factors for interpretation (see later lecture).
- The approximation relies on large sample sizes, so the normal distribution of the estimator provides a good summary of the information in the data.

Bayesian Logistic Regression: Estimation

- We return to the example presented at the start of the lecture.
- Case-control data for the disease Leber Hereditary Optic Neuropathy (LHON) disease with genotype data for marker rs6767450:

	CC $x=0$	CT $x=1$	TT $x=2$	Total
Cases	6	8	75	89
Controls	10	66	163	239
Total	16	74	238	328

```
> x <- c(0,1,2)
> y <- c(6,8,75)
> z <- c(10,66,163)
```

Bayesian Logistic Regression: Estimation

- Below we construct the posterior "by hand".

```
> logitmod <- glm(cbind(y,z)~x,family="binomial")
> thetahat <- logitmod$coef[2]
> V <- vcov(logitmod)[2,2]
#
# 97.5 point of prior is log(1.5) so that we with prob
# 0.95 we think theta lies in (2/3,1.5)
#
> W <- LogNormalPriorCh(1,1.5,0.5,0.975)$sigma^2
> r <- W/(V+W)
> r
[1] 0.405545  # Not so much data here, so weight on prior is high.
# Bayesian posterior median
> exp(r*thetahat)
        x
1.214281   # Shrunk towards prior median of 1
# Note: INLA estimate (with same prior) is 1.22 and approximate
# posterior SD here is sqrt(rV)=0.159, INLA version is 0.154.
# Bayesian approximate 95% credible interval
> exp(r*thetahat-1.96*sqrt(r*V))
        x
0.888
> exp(r*thetahat+1.96*sqrt(r*V))
        x
1.660
```


Bayesian Logistic Regression: Hypothesis Testing

- Now we turn to testing using Bayes factors.
- We examine the sensitivity to the prior on the alternative, π_1.

```
> pi1 <- c(1/2,1/100,1/1000,1/10000,1/100000) # 5 prior probs on the null
> source("http://faculty.washington.edu/jonno/BFDP.R")
> BFcall <- BFDPfunV(thetahat,V,W,pi1)
> BFcall
$BF
        x
0.5110967
$pH0
        x
0.256323
$pH1
        x
0.5015156
$BFDP # Corresponding 5 posterior probs of the null
[1] 0.3382290 0.9806196 0.9980453 0.9998044 0.9999804
```

- So data are twice as likely under the alternative (0.502) as compared to the null (0.256).
- Apart from under the 0.5 prior, under these priors the overall evidence is of no association.

Combination of Data Across Studies

- Suppose we wish to combine data from two studies where we assume a common log odds ratio θ.
- The estimates from the two studies are $\widehat{\theta}_1, \widehat{\theta}_2$ with standard errors $\sqrt{V_1}$ and $\sqrt{V_2}$.
- The Bayes factor is
$$\frac{p(\widehat{\theta}_1, \widehat{\theta}_2 | H_0)}{p(\widehat{\theta}_1, \widehat{\theta}_2 | H_1)}.$$
- The approximate Bayes factor is
$$\mathrm{ABF}(\widehat{\theta}_1, \widehat{\theta}_2) = \mathrm{ABF}(\widehat{\theta}_1) \times \mathrm{ABF}(\widehat{\theta}_2 | \widehat{\theta}_1) \tag{2}$$
where
$$\mathrm{ABF}(\widehat{\theta}_2 | \widehat{\theta}_1) = \frac{p(\widehat{\theta}_2 | H_0)}{p(\widehat{\theta}_2 | \widehat{\theta}_1, H_1)}$$
and
$$p(\widehat{\theta}_2 | \widehat{\theta}_1, H_1) = \mathrm{E}_{\theta | \widehat{\theta}_1} \left[p(\widehat{\theta}_2 | \theta) \right]$$
so that the density is averaged with respect to the posterior for θ.
- Important Point: The Bayes factors are not independent.

Combination of Data Across Studies

- This leads to an approximate Bayes factor (which summarizes the data from the two studies) of
$$\mathrm{ABF}(\widehat{\theta}_1, \widehat{\theta}_2) = \sqrt{\frac{W}{RV_1V_2}} \exp\left\{ -\frac{1}{2} \left(Z_1^2 RV_2 + 2Z_1Z_2R\sqrt{V_1V_2} + Z_2^2 RV_1 \right) \right\}$$
where
  - $R = W/(V_1W + V_2W + V_1V_2)$
  - $Z_1 = \frac{\widehat{\theta}_1}{\sqrt{V_1}}$ and
  - $Z_2 = \frac{\widehat{\theta}_2}{\sqrt{V_2}}$ are the usual Z statistics.
- The ABF will be small (evidence for H_1) when the absolute values of Z_1 and Z_2 are large and they are of the same sign.

Combination of Data Across Studies: The General Case

- Suppose we have K studies with estimates $\widehat{\theta}_k$ and asymptotic variances V_k, $k = 1, ..., K$.
- Assume a common underlying parameter θ.
- The Bayes factor is given by

$$\begin{aligned}
\mathrm{BF}_K &= \frac{p(\widehat{\theta}_1, ..., \widehat{\theta}_K | H_0)}{p(\widehat{\theta}_1, ..., \widehat{\theta}_K | H_1)} \\
&= \frac{\prod_{k=1}^{K} (2\pi V_k)^{-1/2} \exp\left(-\frac{\widehat{\theta}_k^2}{2V_k}\right)}{\int \prod_{k=1}^{K} (2\pi V_k)^{-1/2} \exp\left(-\frac{(\widehat{\theta}_k^2 - \theta)^2}{2V_k}\right) (2\pi W)^{-1/2} \exp\left(-\frac{\theta^2}{2V_k}\right) d\theta} \\
&= \sqrt{W\left(W^{-1} + \sum_{k=1}^{K} V_k^{-1}\right)} \exp\left[-\frac{1}{2}\left(\sum_{k=1}^{K} \frac{\widehat{\theta}_k}{V_k}\right)^2 \left(W^{-1} + \sum_{k=1}^{K} V_k^{-1}\right)^{-1}\right]
\end{aligned}$$

Combination of Studies: The General Case

- The posterior is given by

$$\theta | \widehat{\theta}_1, ..., \widehat{\theta}_K \sim \mathrm{normal}(\mu, \sigma^2)$$

where

$$\begin{aligned}
\mu &= \left(\sum_{k=1}^{K} \frac{\widehat{\theta}_k}{V_k}\right) \left(W^{-1} + \sum_{k=1}^{K} V_k^{-1}\right)^{-1} \\
\sigma^2 &= \left(W^{-1} + \sum_{k=1}^{K} V_k^{-1}\right)^{-1}
\end{aligned}$$

Example of Combination of Studies in a GWAS

- We illustrate how reported confidence intervals can be converted to Bayesian summaries.
- Frayling *et al.* (2007) report a GWAS for Type II diabetes.
- For SNP rs9939609:

Stage	Estimate (CI)	p-value	$-\log_{10}$ BF	$\Pr(H_0\|\text{data})$ with prior: 1/5,000	1/50,000
1st	1.27 (1.16–1.37)	6.4×10^{-10}	7.28	0.00026	0.0026
2nd	1.15 (1.09–1.23)	4.6×10^{-5}	2.72	0.905	0.990
Combined	–	–	13.8	8×10^{-11}	8×10^{-10}

- Combined evidence is stronger than each separately since the point estimates are in agreement.
- For summarizing inference the (5%, 50%, 95%) points for the RR are:

Prior	1.00 (0.67–1.50)
First Stage	1.26 (1.17–1.36)
Combined	1.21 (1.15–1.27)

Conclusions

- Computationally GLMs and GLMMs can now be fitted in a relatively straightforward way.
- `INLA` is very convenient and is being constantly improved.
- As with all analyses, it is crucial to check modeling assumptions (and there are usually more in a Bayesian analysis).
- For binary observations `INLA` can produce inaccurate estimates.
- Markov chain Monte Carlo provides an alternative for computation. `WinBUGS` is one popular implementation.
- Other MCMC possibilities include: `JAGS`, `BayesX`, Stan.

References

Connelly, C., Wakefield, J., and Akey, J. (2014). Evolution and architecture of chromatin accessibility and function in yeast. *PLoS Genetics*. To appear.

Frayling, T., Timpson, N., Weedon, M., Zeggini, E., Freathy, R., and et al., C. L. (2007). A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science*, **316**, 889–894.

Wakefield, J. (2007). A Bayesian measure of the probability of false discovery in genetic epidemiology studies. *American Journal of Human Genetics*, **81**, 208–227.

Wakefield, J. (2009). Bayes factors for genome-wide association studies: comparison with *p*-values. *Genetic Epidemiology*, **33**, 79–86.

Module 4: Bayesian Methods
Lecture 9 A: Default prior selection

Peter Hoff

Departments of Statistics and Biostatistics
University of Washington

Outline

Jeffreys prior

Unit information priors

Empirical Bayes priors

Independent binary sequence

Suppose researcher A has data of the following type:

$$M_A: \quad y_1, \ldots, y_n \sim \text{ i.i.d. binary}(\theta), \ \theta \in [0, 1].$$

A asks you to do a Bayesian analysis, but either

- doesn't have any prior information about θ, or
- wants you to obtain "objective" Bayesian inference for θ.

You need to come up with some prior $\pi_A(\theta)$ to use for this analysis.

Jeffreys prior Unit information priors Empirical Bayes priors

Independent binary sequence

Suppose researcher B has data of the following type:

$$M_B: \quad y_1, \ldots, y_n \sim \text{ i.i.d. binary}(\tfrac{e^\gamma}{1+e^\gamma}), \ \gamma \in (-\infty, \infty).$$

B asks you to do a Bayesian analysis, but either

- doesn't have any prior information about γ, or
- wants you to obtain "objective" Bayesian inference for γ.

You need to come up with some prior $\pi_B(\gamma)$ to use for this analysis.

Prior generating procedures

Suppose we have a procedure for generating priors from models:

$$\text{Procedure}(M) \to \pi$$

Applying the procedure to model M_A should generate a prior for θ:

$$\text{Procedure}(M_A) \to \pi_A(\theta)$$

Applying the procedure to model M_B should generate a prior for γ:

$$\text{Procedure}(M_B) \to \pi_B(\gamma)$$

What should the relationship between π_A and π_B be?

Induced priors

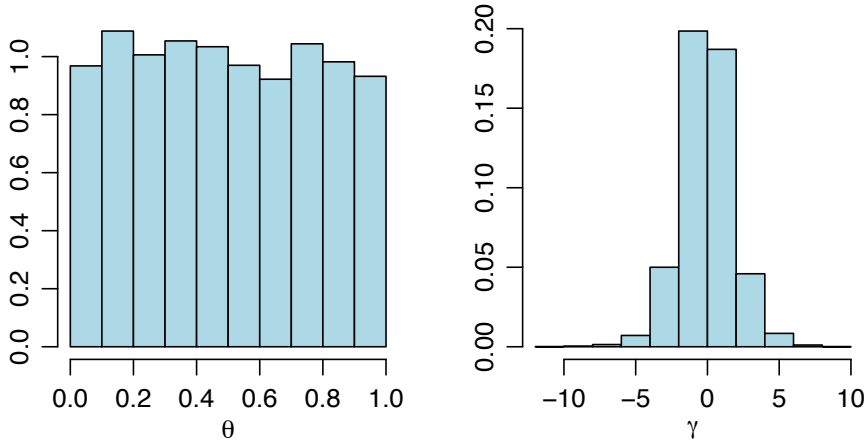
Note that a prior $\pi_A(\theta)$ over θ induces a prior $\pi_A(\gamma)$ over $\gamma = \log \frac{\theta}{1-\theta}$.

This *induced prior* can be obtained via

- calculus;
- simulation.

Induced priors

```
theta<-rbeta(5000,1,1)
gamma<-log(theta/(1-theta))
```



Internally consistent procedures

This fact creates a small conundrum:

We could generate a prior for γ via the induced prior on θ:

$$\text{Procedure}(M_A) \to \pi_A(\theta) \to \pi_A(\gamma)$$

Alternatively, a prior for γ could be obtained directly from M_B:

$$\text{Procedure}(M_B) \to \pi_B(\gamma)$$

Both $\pi_A(\gamma)$ and $\pi_B(\gamma)$ are obtained from the Procedure.

Which one should we use?

Jeffreys' principle

Jeffreys (1949) says that any default Procedure should be *internally consistent* in the sense that the two priors on γ should be the same.

More generally, his principle states if M_B is a reparameterization of M_A, then

$$\pi_A(\gamma) = \pi_B(\gamma).$$

Of course, all of this logic applies to the model in terms of θ:

$$\begin{aligned} \text{Procedure}(M_A) &\to \pi_A(\theta) \\ \text{Procedure}(M_B) &\to \pi_B(\gamma) \to \pi_B(\theta) \\ \pi_A(\theta) &= \pi_B(\theta) \end{aligned}$$

Jeffreys' prior

It turns out that Jeffreys' principle leads to a unique Procedure:

$$\pi_J(\theta) = \sqrt{|\mathrm{E}[(\tfrac{d}{d\theta} \log p(y|\theta))^2]|}$$

Example: Binomial/binary model

$$\begin{aligned} y_1, \ldots, y_n &\sim \text{i.i.d. binary}(\theta) \\ \pi_J(\theta) &\propto \theta^{-1/2}(1-\theta)^{-1/2} \end{aligned}$$

We recognize this prior as a beta(1/2,1/2) distribution:

$$\theta \sim \text{beta}(1/2, 1/2)$$

Default Bayesian inference is then based on the following posterior:

$$\theta | y_1, \ldots, y_n \sim \text{beta}(1/2 + \sum y_i, 1/2 + \sum (1 - y_i)).$$

Jeffreys' prior

Example: Poisson model

$$y_1, \ldots, y_n \sim \text{i.i.d. Poisson}(\theta)$$
$$\pi_J(\theta) \propto 1/\sqrt{\theta}$$

Recall our conjugate prior for θ in this case was a gamma(a, b) density:

$$\theta(\theta|a, b) \propto \theta^{a-1} e^{-\theta/b}$$

For the Poisson model and gamma prior,

$$\theta \sim \text{gamma}(a, b) \rightarrow \theta|y_1, \ldots, y_n \sim \text{gamma}(a + \sum y_i, b + n)$$

What about under the Jeffreys prior?

$\pi_J(\theta)$ "looks like" a gamma distribution with $(a, b) == (1/2, 0)$. It follows that

$$\theta \sim \pi_J \rightarrow \theta|y_1, \ldots, y_n \sim \text{gamma}(1/2 + \sum y_i, n).$$

(Note: π_J is not an actual gamma density - it is not a probability density at all!)

Jeffreys' prior

Example: Normal model

$$y_1, \ldots, y_n \sim \text{i.i.d. Normal}(\mu, \sigma^2)$$
$$\pi_J(\mu, \sigma^2) = 1/\sigma^2$$

(this is a particular version of Jeffreys' prior for multiparameter problems)

It is very interesting to note that the resulting posterior for μ is

$$\frac{\mu - \bar{y}}{s/\sqrt{n}} \sim t_{n-1}$$

This means that a 95% objective Bayesian confidence interval for μ is

$$\mu \in \bar{y} \pm t_{.975,n-1} s/\sqrt{n}$$

This is exactly the same as the usual t-confidence interval for a normal mean.

Notes on Jeffreys' prior

1. Jeffreys' principle leads to Jeffreys' prior.
2. Jeffreys' prior isn't always a "proper" prior distribution.
3. Improper priors can lead to proper posteriors.
   - These often lead to Bayesian interpretations of "frequentist" procedures.

Data-based priors

Recall from the binary/beta analysis:

$$\begin{aligned} \theta &\sim \text{beta}(a, b) \\ y_1, \ldots, y_n &\sim \text{binary}(\theta) \\ \theta | y_1, \ldots, y_n &\sim \text{beta}(a + \sum y_i, b + \sum (1 - y_i) \end{aligned}$$

Under this posterior,

$$\begin{aligned} \mathrm{E}[\theta | y_1, \ldots, y_n] &= \frac{a + \sum y_i}{a + b + n} \\ &= \left(\frac{a + b}{a + b + n} \right) \frac{a}{a + b} + \left(\frac{n}{a + b + n} \right) \bar{y} \end{aligned}$$

- $\frac{a}{a+b} \approx$ guess at what θ is
- $a + b \approx$ confidence in guess.

Data-based priors

We may be reluctant to guess at what θ is. Wouldn't $\bar{y}$ be better than a guess?

Idea: Set $\frac{a}{a+b} = \bar{y}$.

Problem: This is cheating! Using $\bar{y}$ for your prior misrepresents the amount of information you have.

Solution: Cheat as little as possible:

- Set $\frac{a}{a+b} = \bar{y}$.
- Set $a + b = 1$.
- This implies $a = \bar{y}$, $b = 1 - \bar{y}$.

The amount of cheating has the information content of only one observation.

Unit information principle

If you don't have prior information about θ, then

1. Obtain an MLE/OLS estimator $\hat{\theta}$ of θ;
2. Make the prior $\pi(\theta)$
   - weakly centered around $\hat{\theta}$,
   - have the information equivalent of one observation.

Again, such a prior leads to double-use of the information in your sample.

However, the amount of "cheating" is small, and decreases with n.

Poisson example:

$$y_1, \ldots, y_n \sim \text{i.i.d. Poisson}(\theta)$$

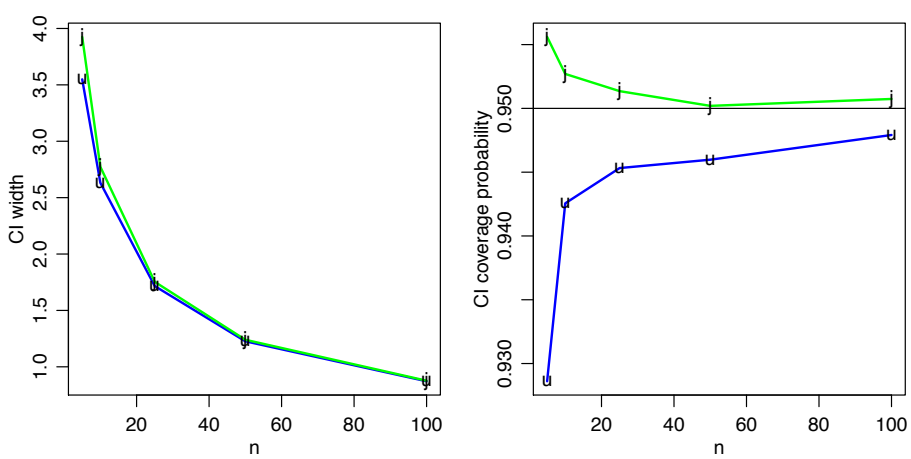
Under the gamma(a, b) prior,

$$\begin{aligned} \mathrm{E}[\theta | y_1, \ldots, y_n] &= \frac{a + \sum y_i}{b + n} \\ &= (\frac{b}{b+n})\frac{a}{b} + (\frac{n}{b+n})\bar{y} \end{aligned}$$

Unit information prior:

$$a/b = \bar{y}, b = 1 \Rightarrow (a, b) = (\bar{y}, 1)$$

Comparison to Jeffreys' prior



Notes on UI priors

1. UI priors weakly concentrate around a data-based estimator.
2. Inference under UI priors is anti-conservative, but this bias decreases with n.
3. Can be used in multiparameter settings, and is related to BIC.

Normal means problem

$$y_j = \theta_j + \epsilon_j, \quad \epsilon_1, \ldots, \epsilon_p \sim \text{i.i.d. normal}(0, 1)$$

Task: Estimate $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$.

An odd problem:

- What does estimation of θ_j have to do with estimation of θ_k?
- There is only one observation y_j per parameter θ_j - how well can we do?

Where the problem comes from:

- Comparison of two groups A and B on p variables (e.g. expression levels)
- For each variable j, construct a two-sample t-statistic

$$y_j = \frac{\bar{x}_{A,j} - \bar{x}_{B,j}}{s_j/\sqrt{n}}$$

- For each j, y_j is approximately normal with
  - mean $\theta_j = \sqrt{n}(\mu_{A,j} - \mu_{B,j})/\sigma_j$
  - variance 1.

Normal means problem

$$y_j = \theta_j + \epsilon_j, \quad \epsilon_1, \ldots, \epsilon_p \sim \text{i.i.d. normal}(0, 1)$$

One obvious estimator of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ is $\mathbf{y} = (y_1, \ldots, y_p)$.

- $\mathbf{y}$ is the MLE;
- $\mathbf{y}$ is unbiased and the UMVUE.

However, it turns out that $\mathbf{y}$ is not so great in terms of risk:

$$R(\mathbf{y}, \boldsymbol{\theta}) = \mathrm{E}[\sum_{j=1}^{p} (y_j - \theta_j)^2]$$

When $p > 2$ we can find an estimator that beats $\mathbf{y}$ for every value of $\boldsymbol{\theta}$, and is much better when p is large. This estimator has been referred to as an *empirical Bayes estimator*.

Bayesian normal means problem

$$y_j = \theta_j + \epsilon_j, \quad \epsilon_1, \ldots, \epsilon_p \sim \text{i.i.d. normal}(0, 1)$$

Consider the following prior on $\boldsymbol{\theta}$:

$$\theta_1, \ldots, \epsilon_p \sim \text{i.i.d. normal}(0, \tau^2)$$

Under this prior,

$$\hat{\theta}_j = \mathrm{E}[\theta_j | y_1, \ldots, y_n] = \frac{\tau^2}{\tau^2 + 1} y_j$$

This is a type of "shrinkage" prior:

- It "shrinks" the estimates towards zero, away from y_j;
- It is particularly good if many of the true θ_j's are very small or zero.

Empirical Bayes

$$\hat{\theta}_j = \frac{\tau^2}{\tau^2+1} y_j$$

We might know we want to shrink towards zero.

We might not know the appropriate amount of shrinkage.

Solution: Estimate τ^2 from the data!

$$\left.\begin{array}{ll} y_j & = \theta_j + \epsilon_j \\ \epsilon_j & \sim N(0,1) \\ \theta_j & \sim N(0,\tau^2) \end{array}\right\} \quad \Rightarrow \quad y_j \sim N(0, \tau^2+1)$$

We should have

$$\sum y_j^2 \approx p(\tau^2+1)$$
$$\sum y_j^2/p - 1 \approx \tau^2$$

Idea: Use $\hat{\tau}^2 = \sum y_j^2/p - 1$ for the shrinkage estimator.

Modification Use $\hat{\tau}^2 = \sum y_j^2/(p-2) - 1$ for the shrinkage estimator.

James-Stein estimation

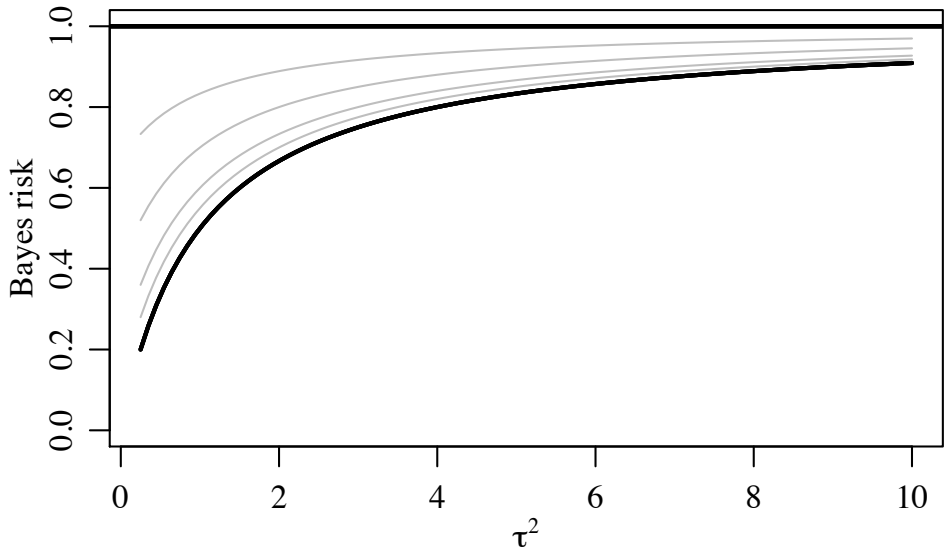
$$\hat{\theta}_j = \frac{\hat{\tau}^2}{\hat{\tau}^2+1} y_j \qquad \hat{\tau}^2 = \sum y_j^2/(p-2) - 1$$

It has been shown theoretically that from a non-Bayesian perspective, this estimator beats $\mathbf{y}$ in terms of risk for all $\boldsymbol{\theta}$.

$$R(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) < R(\mathbf{y}, \boldsymbol{\theta}) \text{ for all } \boldsymbol{\theta}$$

Also, from a Bayesian perspective, this estimator is almost as good as the optimal Bayes estimator, under a known τ^2.

Comparison of risks



The Bayes risk of the JSE is between that of **X** and the Bayes estimator.

Bayes risk functions are plotted for $p \in \{3, 5, 10, 20\}$.

Empirical Bayes in general

Model: $p(y|\theta), \theta \in \Theta$

Prior class: $\pi(\theta|\psi), \psi \in \Psi$

What value of ψ to choose?

Empirical Bayes:

1. Obtain the "marginal likelihood" $p(y|\psi) = \int p(y|\theta)\pi(\theta|\psi)d\theta$;
2. Find an estimator $\hat{\psi}$ based on $p(y|\psi)$;
3. Use the prior $\pi(\theta|\hat{\psi})$.

Notes on empirical Bayes

1. Empirical Bayes procedures are obtained by "estimating" hyperparameters from the data.
2. Often these procedures behave well from both Bayes and frequentist procedures.
3. They work best when the number of parameters is large and hyperparameters are distinguishable.

Module 4: Bayesian Methods
Lecture 9 B: QTL interval mapping

Peter Hoff

Departments of Statistics and Biostatistics
University of Washington

Outline

The F1 Backcross

The mixture model

Marker data

Bayesian estimation

QTLs

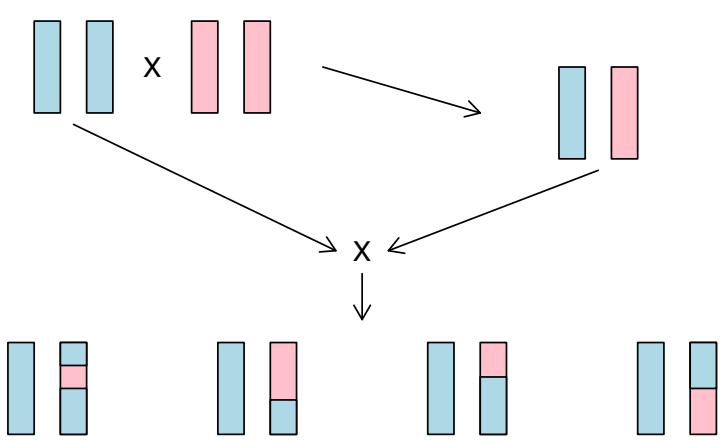
Genetic variation ⇒ quantitative phenotypic variation

QTLs have been associated with many health-related phenotypes:

- cancer
- obesity
- heritable disease

QTL interval mapping: A statistical approach to the identification of QTLs from marker and phenotype data.

F1 Backcross



At any given locus, an animal could be AA or AB.

Two-component mixture model

Suppose there is a single QTL affecting a continuous trait. Let

- x be the location of the QTL;
- $g(x)$ be the genotype at x
  - $g(x) = 0$ if AA at x
  - $g(x) = 1$ if AB at x
- y be a continuous quantitative trait.

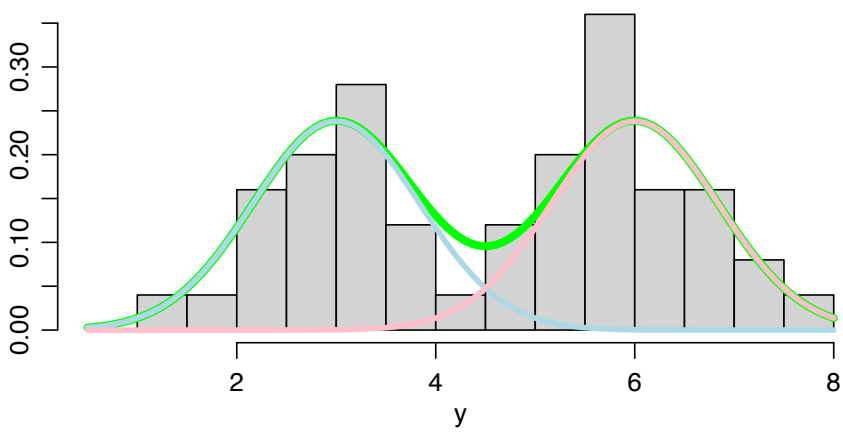
Two-component mixture model:

$$y \sim \begin{cases} \text{normal}(\mu_{AA}, \sigma^2) & \text{if } g(x) = 0 \\ \text{normal}(\mu_{AB}, \sigma^2) & \text{if } g(x) = 1 \end{cases}$$

About half of the animals are $g(x) = 0$ and half are $g(x) = 1$, but we don't know which are which.

Two-component mixture model

Data from 50 animals:

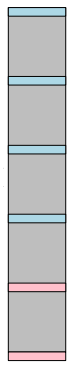


Marker data

If the location of x of the QTL were known, we could genotype it:

- $\mathbf{y}_0 = \{y_i : g_i(x) = 0\}$
- $\mathbf{y}_1 = \{y_i : g_i(x) = 1\}$
- evaluate effect size with a two sample t-test.

Instead of $g(x)$ we have genotype information at a set of markers:

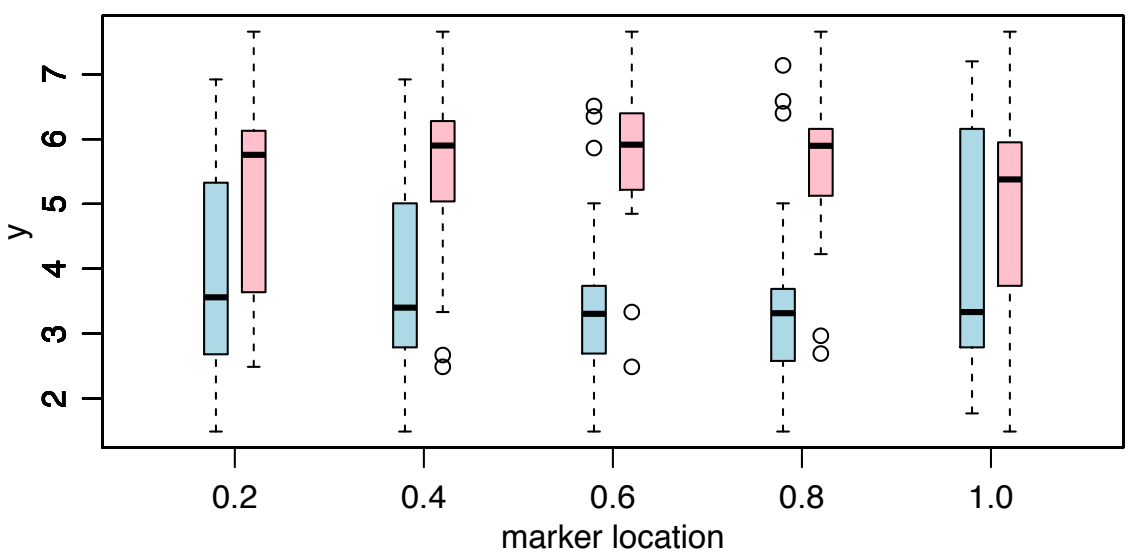


Genotype information at evenly spaced markers $m_1, \ldots, m_K$

$$g_i(m_k) = \begin{cases} 0 & \text{if animal } i \text{ is homozygous at } m_k \\ 1 & \text{if animal } i \text{ is heterozygous at } m_k \end{cases}$$

Comparisons at marker locations

$n = 50$ animals at $K = 6$ equally spaced marker locations:



Comparisons across the genome

Procedure: Move along each chromosome, making comparisons of heterozygotes to homozygotes at each possible QTL location x.

Problem: Genotypes at non-marker locations x are not known.

However, they are known probabilistically: Let

- r = recombination rate between left and right flanking markers;
- r_l = recombination rate between left flanking marker m_l and x;
- r_r = recombination rate between right flanking marker m_r and x.

$$\Pr(g(x) = 1 | g(m_l) = 1, g(m_r) = 1) = \frac{(1 - r_l) \times (1 - r_r)}{1 - r}$$

$$\Pr(g(x) = 1 | g(m_l) = 0, g(m_r) = 1) = \frac{r_l \times (1 - r_r)}{r}$$

etc.

Knowns and unknowns quantities

Unknown quantities in the system include

- QTL location x
- genotypes $G(x) = \{g_1(x), \ldots, g_n(x)\}$
- parameters of the QTL distributions: $\theta = \{\mu_{AA}, \mu_{AB}, \sigma^2\}$

Known quantities include

- quantitative trait data $\mathbf{y} = y_1, \ldots, y_n$
- marker data $\mathbf{M} = \{g_i(m_k), i = 1, \ldots, n, k = 1, \ldots K\}$

Bayesian analysis: Obtain Pr(unknowns|knowns)

$$\Pr(x, G(x), \theta | \mathbf{y}, \mathbf{M})$$

Gibbs sampler

We can approximate $\Pr(x, G(x), \theta|\mathbf{y}, \mathbf{M})$ with a Gibbs sampler:

1. simulate $x \sim p(x|\theta, \mathbf{y}, \mathbf{M})$
2. simulate $G(x) \sim p(G(x)|\theta, x, \mathbf{y}, \mathbf{M})$
3. simulate $\theta \sim p(\theta|x, G(x), \mathbf{y}, \mathbf{M})$

For example, based on marker data alone,

$$\begin{aligned}\Pr(g_i(x) = 1|\mathbf{M}) &= \frac{\Pr(g_i(x) = 1|\mathbf{M})}{\Pr(g_i(x) = 1|\mathbf{M}) + \Pr(g_i(x) = 0|\mathbf{M})} \\ &= \frac{p_{i1}}{p_{i1} + p_{i0}}.\end{aligned}$$

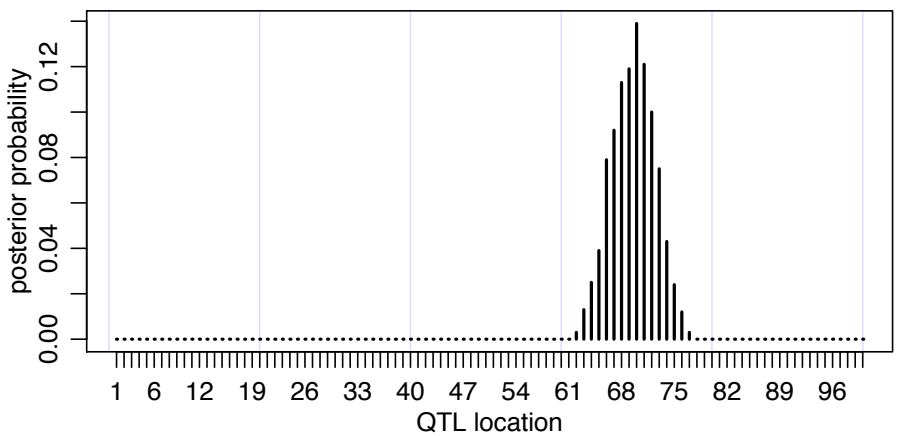
Given phenotype data,

$$\begin{aligned}\Pr(g_i(x) = 1|x, \theta, \mathbf{y}, \mathbf{M}) &= \frac{p_{i1} \times p(y_i|g_i(x) = 1, \theta)}{p_{i1} \times p(y_i|g_i(x) = 1, \theta) + p_{i0} \times p(y_i|g_i(x) = 0, \theta)} \\ &= \frac{p_{i1} \times \text{dnorm}(y_i, \mu_{AB}, \sigma)}{p_{i1} \times \text{dnorm}(y_i, \mu_{AB}, \sigma) + p_{i0} \times \text{dnorm}(y_i, \mu_{AA}, \sigma)}.\end{aligned}$$

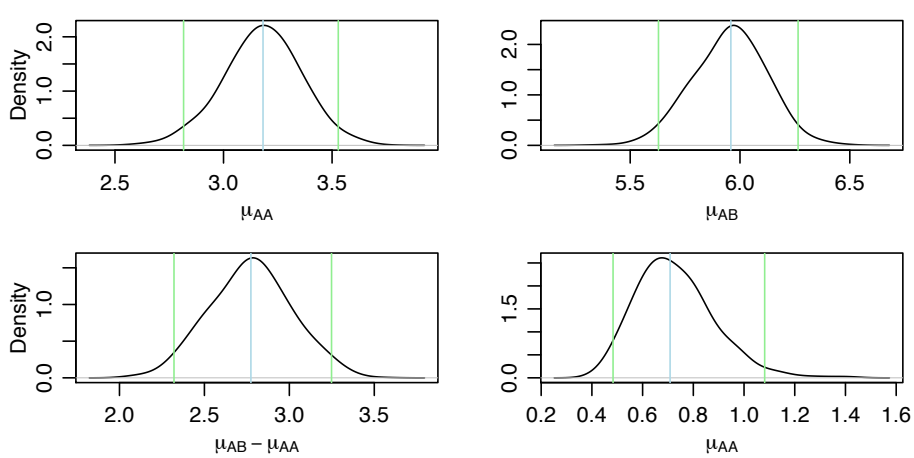
R-code for Gibbs sampler

```
for(s in 1:25000)
{

  ## update x
  lpy.x<-NULL;for(x in 1:100){lpy.x<-c(lpy.x,lpy.theta(y,G,x,mu,s2))}
  x<-sample(1:100,1,prob=exp(lpy.x-max(lpy.x)))

  ## update gx
  pg1.x<-prhet.sG(x,G,mpos)
  py.g1<-dnorm(y,mu[2],sqrt(s2))
  py.g0<-dnorm(y,mu[1],sqrt(s2))
  pg1.yx<- py.g1*pg1.x/( py.g1*pg1.x + py.g0*(1-pg1.x))
  gx<-rbinom(n,1,pg1.yx)

  ## update s2
  s2<-1/rgamma( 1,(nu0+n)/2,(nu0*s20+sum( (y-mu[gx+1])^2) )/2  )

  ## update mu
  mu<-rnorm(2,(mu0*k0+tapply(y,gx,sum))/(k0+table(gx)),
                sqrt(s2/(k0+ table(gx))))
}
```

QTL location



Parameter estimates



Some references

- "Review of statistical methods for QTL mapping in experimental crosses" (Broman, 2001).
- "QTLBIM" - QTL Bayesian interval mapping: R-package.

2014 SISG MODULE 4: Bayesian Statistics for Genetics Lecture 10: Imputation, Hierarchical Mixture Models, Model Comparison

Jon Wakefield

Departments of Statistics and Biostatistics
University of Washington

Outline

Introduction and Description of Data

Methods for Imputation

Hierarchical Modeling of Allele-Specific Expression Data
- Motivation
- Modeling

Model Comparison

Conclusions

Introduction

- In this lecture we consider two topics.
- First, we consider methods for imputation of missing genotypes.
- We describe a number of the more common Bayesian approaches to this problem.
- Second, we return to the Allele-Specific Expression (ASE) data introduced in Lecture 3.
- An in-depth analysis of the ASE data will be given – we will analyze with a Bayes hierarchical mixture model.
- The material in the second part of the lecture is necessarily more advanced than in previous lectures, but will give a sense of the power of Bayesian methods, and will highlight some of the practical issues in the use of these methods.
- Finally, we will briefly review a number of procedures to carry out model comparison.

Motivation for Imputation

- Imputation is the prediction of missing genotypes.
- Imputation is used in both GWAS and in fine-mapping studies.
- The technique is becoming increasingly popular since it can:
  - Increase power in GWAS.
  - Facilitate meta-analysis in which data it is required to combine information from different panels which have different sets of SNPs. In this way power can be increased.
  - Fine-map causal variants, see Figure 1. Imputed SNPs that show large associations can be better candidates for replication studies.
- The key idea in the approaches we describe is the use of data on haplotypes from a relevant population to build a prior model for the missing data, basically the models leverage linkage disequilibrium.

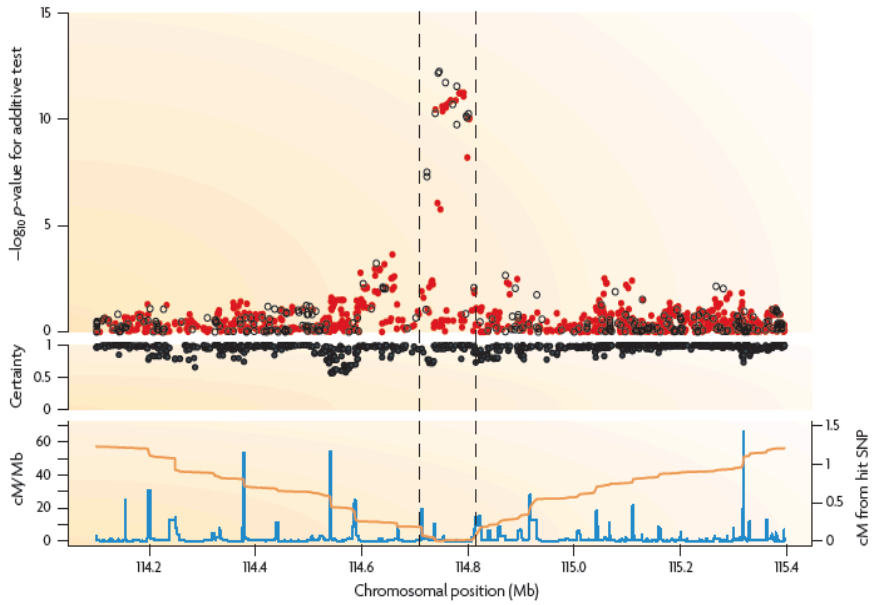


Figure 1 : Imputation for the TCF7L2 gene, from Marchini *et al.* (2007). Imputed SNP signals are in red and observed SNPs in black.

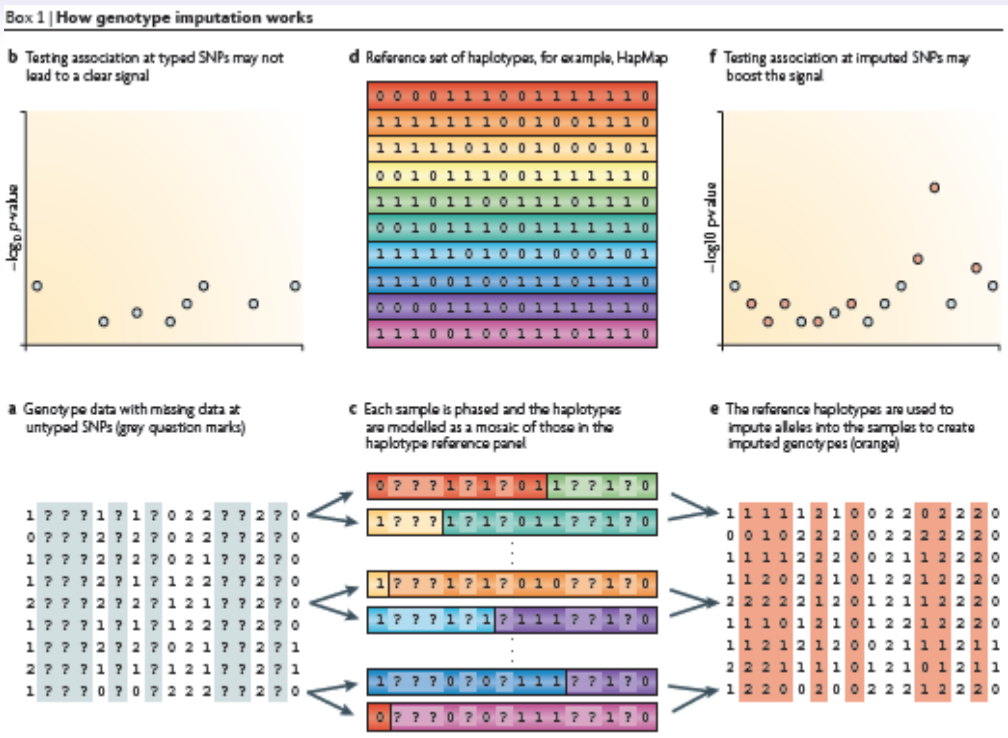


Figure 2 : Imputation overview from Marchini and Howie (2010).

The Statistical Framework

- Suppose we wish to estimate the association between a phenotype and m genetic markers in n individuals.
- Let G_{ij} represent the genotype of individual i at SNP j with G_{ij} unobserved for some SNPs.
- We consider diallelic SNPs so that G_{ij} can take the value 0, 1 or 2 depending on whether the pair of constituent SNPs are $\{0,0\}$, $\{0,1\}$, $\{1,0\}$ or $\{1,1\}$.
- If G_{ij} is observed then for SNP j we simply model

$$p(y_i|G_{ij})$$

- For example, if the phenotype y_i is continuous, we might assume a normal model:

$$\mathrm{E}[Y_i] = \beta_0 + \beta_1 G_{ij},$$

and if y_i is binary, a logistic model is an obvious candidate:

$$\frac{p_i}{1-p_i} = \exp(\beta_0 + \beta_1 G_{ij})$$

where p_i is the probability of disease for individual i.

The Statistical Framework

- Let $\mathbf{H} = (\mathbf{H}_1, ..., \mathbf{H}_N)$ represent haplotype information at m SNPs in a relevant reference-panel, with N distinct haplotypes.
- If G_{ij} is unobserved then for SNP j we have the model

$$p(y_i|\mathbf{H}, \mathbf{G}_i) = \sum_{k=0}^{2} p(y_i|G_{ij} = k) \times \Pr(G_{ij} = k|\mathbf{H}, \mathbf{G}_i)$$

- The big question is how to obtain the predictive distribution

$$\Pr(G_{ij} = k|\mathbf{H}, \mathbf{G}_i).$$

- A common approach is to take as prior a Hidden Markov Model (HMM).
- We digress to discuss HMMs.

Hidden Markov Models

- Example: Poisson Time Series A common problem is how to model count data over time. A Poisson model is the obvious choice but how to introduce:
  1. overdispersion and
  2. dependence over time.
- Consider the model:

$$\begin{aligned}
&\textit{Stage 1: } Y_t|\lambda_t \sim \text{Poisson}(\lambda_t),\ t = 1, 2, \ldots \\
&\textit{Stage 2: } \lambda_t|Z_t \sim_{iid} \begin{cases} \lambda_0 & \text{if } Z_t = 0 \\ \lambda_1 & \text{if } Z_t = 1 \end{cases} \\
&\textit{Stage 3: } Z_t|p \sim_{iid} \text{Bernoulli}(p).
\end{aligned}$$

- An alternative model replaces *Stage 3* with a (first-order) Markov chain model, i.e, $\Pr(Z_t|Z_1, \ldots, Z_{t-1}) = \Pr(Z_t|Z_{t-1})$:

$$\begin{aligned}
\Pr(Z_t = 0|Z_{t-1} = 0) &= p_0 \\
\Pr(Z_t = 1|Z_{t-1} = 1) &= p_1
\end{aligned}$$

- Z_t is an unobserved (hidden) state.
- As an example we consider the number of major earthquakes (magnitude 7 and above) for the years 1990–2006.
- We illustrate the fit of this model with two or three underlying states.

Example: Earthquake Data

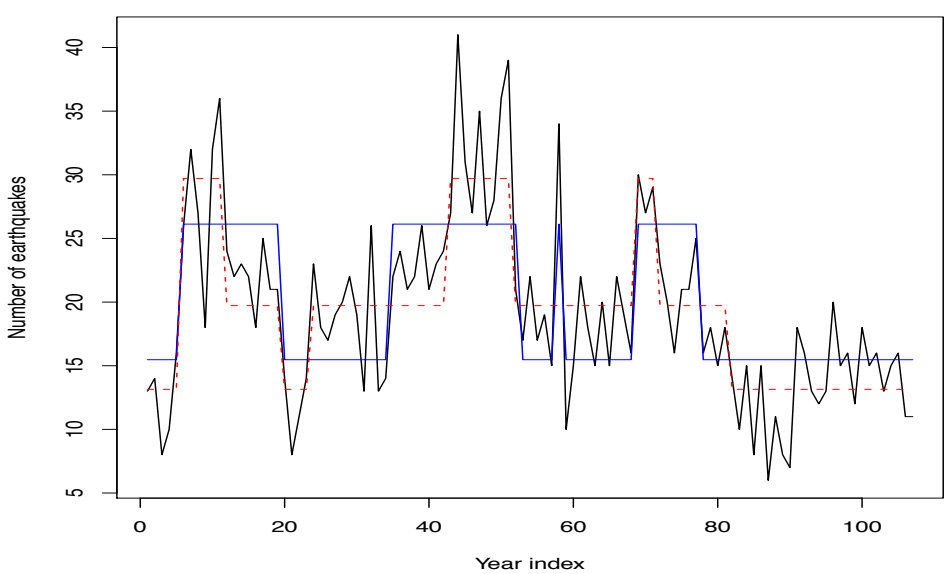


Figure 3 : The earthquake data along with the underlying states for the two and three state HMMs, in blue and red, respectively.

IMPUTE v1

- Marchini *et al.* (2007) consider a HMM for the vector of genotypes for individual i:

$$\Pr(\mathbf{G}_i|\mathbf{H},\theta,\rho) = \sum_{\mathbf{Z}_i=(\mathbf{Z}_i^{(1)},\mathbf{Z}_i^{(2)})} \Pr(\mathbf{G}_i|\mathbf{Z}_i,\theta) \times \Pr(\mathbf{Z}_i|\mathbf{H},\rho)$$

  where $\mathbf{Z}_i^{(1)} = \{Z_{i1}^{(1)},...,Z_{iJ}^{(1)})$ and $\mathbf{Z}_i^{(2)} = \{Z_{i1}^{(2)},...,Z_{iJ}^{(2)})$.
- The $(\mathbf{Z}_i^{(1)},\mathbf{Z}_i^{(2)})$ are the pair of haplotypes for SNP j from the reference panel that are copied to form the genotype vector. These are the hidden states.
- The term $\Pr(\mathbf{Z}_i|\mathbf{H},\rho)$ models how the pair of copied haplotypes for individual i changes along the sequence. This probability changes according to a Markov chain with the switching of states depending on the fine-scale recombination rate ρ.
- The term $\Pr(\mathbf{G}_i|\mathbf{Z},\theta)$ allows the observed genotypes to differ from the pair of copied haplotypes through mutation; the mutation parameter is θ.
- IMPUTE v2 (Howie *et al.*, 2009) is a more flexible version that alternates between phasing and haploid imputation.

fastPHASE and BIMBAM

- We describe the model of Scheet and Stephens (2006).
- A Hidden Markov Model (HMM) is used to determine $\Pr(G_{ij} = k|\boldsymbol{\alpha},\boldsymbol{\theta},r)$.
- The basic idea is that haplotypes tend to cluster into groups of similar haplotypes; suppose there are K clusters.
- The unobserved hidden state is the haplotype cluster from which this SNP arose from. Each cluster has an associated set of allele frequencies $\boldsymbol{\theta}_{kj}$.
- With K underlying states we have, for SNP j, α_{kj} being the probability of arising from haplotype k, with

$$\sum_{k=1}^{K} \alpha_{kj} = 1.$$

- The model is

$$\Pr(\mathbf{G}_i|\boldsymbol{\alpha},\boldsymbol{\theta},r) = \sum_{\mathbf{Z}} \Pr(\mathbf{G}_i|\mathbf{Z}_i,\boldsymbol{\theta}) \times \Pr(\mathbf{Z}_i|\boldsymbol{\alpha},r)$$

  with Z_{ij} the haplotype of origin for individual i and SNP j.
- A Markov chain is constructed for Z_{ij} with the strength of dependence being based on the recombination rate r at a given location.
- Given $Z_{ij} = k$, the genotype assigned depends on the allele frequencies of the k-th haplotype at the j-th SNP.

Use in Association Studies

- The simplest approach to using imputed SNPs is to substitute $\widehat{G}_{ij}$ (a number between 0 and 2) into the phenotype association model.
- A set of probabilities $\Pr(G_{ij} = k|\mathbf{G}, \mathbf{H})$ for $k = 0, 1, 2$ are produced and these may be used to average over the uncertainty in the phenotype model.
- Within BIMBAM the unknown genotype is sampled from its posterior distribution, within an MCMC framework.
- Other approaches:
  - MACH: similar methodology to IMPUTE (Li *et al.*, 2010).
  - Beagle: uses a graphical model for haplotypes Browning and Browning (2009).

Practical Issues

- One may attempt to match the haplotype panel (e.g. from HapMAP 2) with the study individuals.
- An alternative approach is to use all available haplotypes, and assigning equal prior probabilities to each.
- Many studies, for example Huang *et al.* (2009), have examined SNP imputation accuracy in different populations.

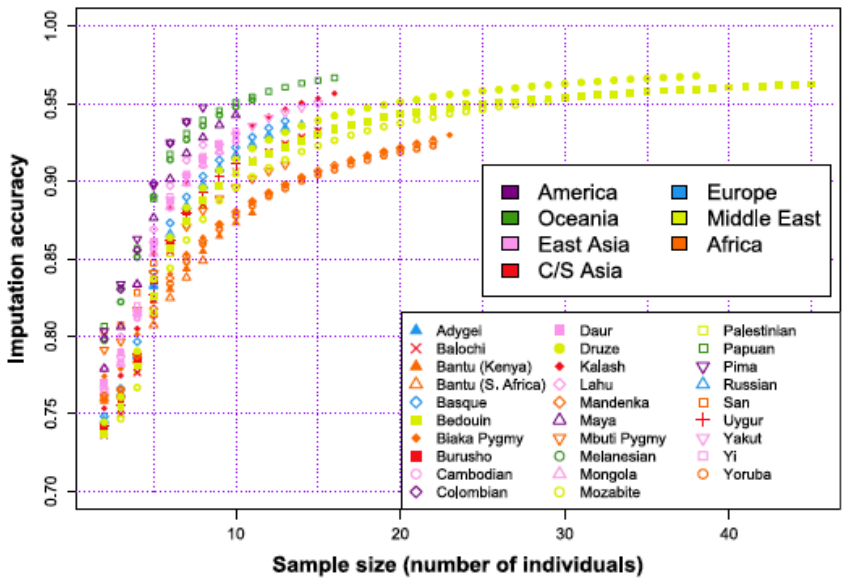


Figure 4 : Imputation accuracy as a function of sample size, from Huang *et al.* (2009).

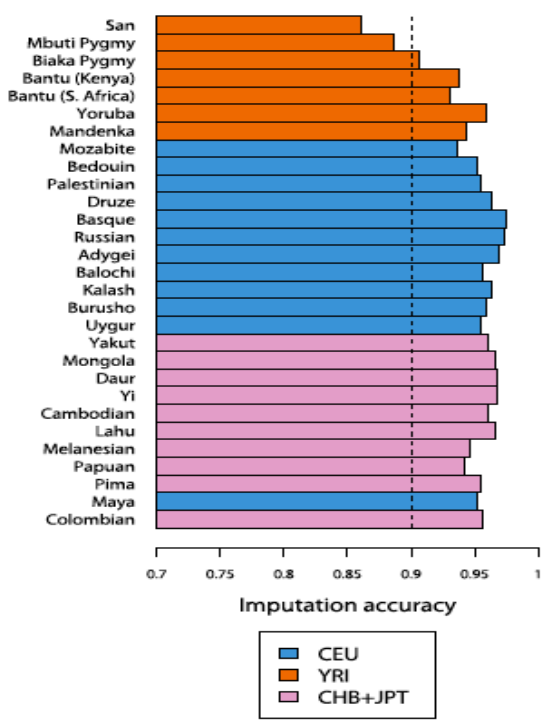


Figure 5 : Imputation accuracy for different populations with a reference-panel of 120 haplotypes. From Huang *et al.* (2009).

Table 1. Association Analysis results.

Locus	SNPname	Type	Effect Allele/ Other	Freq Effect Allele	Effect (SE)[a]	P-value	Genomic Annotation	Variance explained by the locus	Top GWAS SNP	Effect Allele/ Other	Freq Effect Allele	Effect (SE)[a]	P-value	r2	Adjusted P-value	Variance explained by the locus
PCSK9	rs11591147	Metabochip	T/G	0.037	−0.380 (0.048)	2.90×10^{-15}	**missense (R46L)**	1.19%	rs11206510	C/T	0.243	−0.106 (0.023)	5.71×10^{-07}	0.101	0.013	0.23%
	rs2479415	1000G	C/T	0.413	0.076 (0.019)	7.50×10^{-05}	8 Kb from *PCSK9*									
SORT1	rs583104	Metabochip	T/G	0.177	0.149 (0.024)	1.28×10^{-09}	31 Kb from *SORT1*[b]	0.63%	rs599839	G/A	0.276	−0.148 (0.025)	1.43×10^{-09}	0.991	0.90	0.61%
B3GALT4	rs28361085	1000G	C/T	0.073	0.114 (0.036)	0.00169	146 Kb from *B3GALT3*	0.22%	rs2254287	G/C	0.492	0.005 (0.018)	0.771	0.413	0.84	0.02%
B4GALT4	rs34507110	1000G	G/A	0.154	0.122 (0.030)	4.99×10^{-05}	83 Kb from *B4GALT4*	0.48%	rs12695382	A/G	0.075	−0.074 (0.035)	0.035	0.795	0.48	0.03%
APOB	rs547235	1000G	A/G	0.187	−0.144 (0.024)	1.69×10^{-09}	140 Kb from *APOB*	0.51%	rs562338	A/G	0.173	−0.139 (0.025)	1.43×10^{-8}	0. 878	0.98	0.43%
LDLR	rs73015013	Metabochip	T/C	0.138	−0.155 (0.027)	1.12×10^{-08}	9 kb from *LDLR*	1.17%	rs6511720	T/G	0.132	−0.160 (0.027)	1.71×10^{-08}	0.934	0.97	0.59%
	rs72658864	Metabochip	C/T	0.005	0.626 (0.136)	3.90×10^{-06}	**missense (V578A)**									
APOC1/C2/E	rs7412	Metabochip	T/C	0.037	−0.563 (0.048)	1.80×10^{-31}	**missense (R176C)** *APOE*	3.33%	rs4420638[c]	G/A	0.097	0.218 (0.031)	4.67×10^{-12}	0.0003	6.41×10^{-10}	1.07%
	rs429358	Affy+Sanger	C/T	0.071	0.260 (0.036)	5.82×10^{-11}	**missense (C130R)** *APOE*									

The left panel shows the association results at 7 loci. For each gene, the strongest variant is listed first, and any second detected independent signal is listed with results from the conditional analysis (Materials and Methods). The column Type indicates whether the SNP was directly genotyped (Metabochip) or imputed using 1000G reference haplotype (1000G) or the Sardinian reference panel (Affy+Sanger). The right panel shows the association results for the GWAS SNPs previously described [5], the correlation with the top SNP listed in the left panel, and its p-value in the conditional analysis (Adjusted P-value).

[a]Effect sizes are standardized (see Materials and Methods), and represent the change in trait LDL-C values associated with each copy of the reference allele, measured in standard deviation units.

[b]SNP rs583104 is also 1 Kb from *PSRC1* transcript.

[c]r^2 =0.967 with Metabochip second-independent SNP, rs429358. After adjusting for the two independent SNPs, rs7412 and rs429358, the p-value for rs4420638 was 0.5.

doi:10.1371/journal.pgen.1002198.t001

Figure 6 : Example from Sanna *et al.* (2011). Imputation carried out using the `MACH` software.

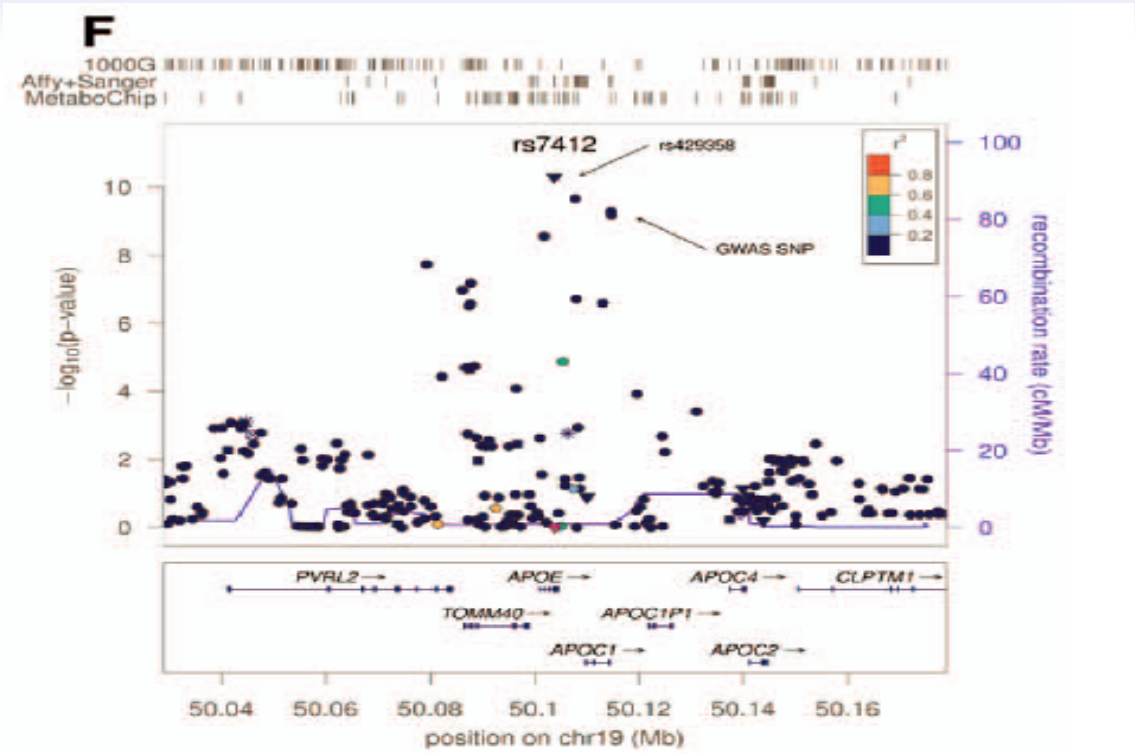


Figure 7 : Example from Sanna *et al.* (2011).

Specifics of ASE Experiment

Details of the data:

- Two "individuals" from genetically divergent yeast strains, BY and RM, are mated to produce a diploid hybrid.
- Three replicate experiments: same individuals, but separate samples of cells.
- Two technologies: Illumina and ABI SOLiD. Each of a few trillion cells are processed.
- Pre- and post-processing steps are followed by fragmentation to give millions of 200–400 base pair long molecules, with short reads obtained by sequencing.
- Strict criteria to call each read as a match are used, to reduce read-mapping bias.
- Data from 25,652 SNPs within 4,844 genes.

Allele Specific Expression via RNA-Seq

Additional data:

- Genomic DNA is sequenced in the diploid hybrid, which has one copy of each gene from BY and from RM.
- The only difference between the genomic DNA and the main experiment is that we expect the genomic DNA to always be present 50:50 (one copy each of BY and RM), whereas for the main experiment it is only 50:50 if there is no ASE.
- For both genomic DNA and RNA we obtain counts at SNPs, at each of BY and RM.

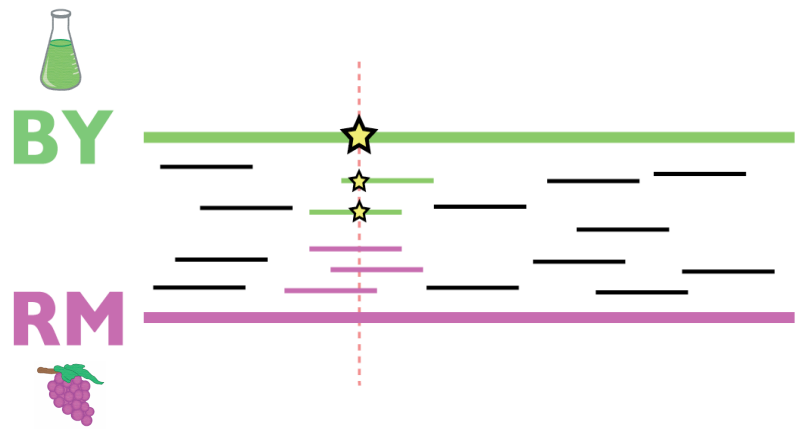


Figure 8 : Mapping of RNA short reads to BY and RM.

Statistical Problem

- Aim of the Experiment: Estimate the proportion of genes that display ASE.
- Let p be the probability of a map to BY at a particular SNP.
- Additionally, we would like to classify genes into:
  - Genes that do not show ASE.
  - Genes that show:
    - Constant ASE across SNPs.
    - Variable ASE across SNPs, i.e. p varies within gene.

  Subsequently, we will examine genes displaying ASE to investigate the mechanism.
- A hierarchical model is feasible since we have within gene and between gene variability.
- Further, a mixture model is suggested, with a mixture of genes that do not display ASE (so there p's are 0.5) and that do display ASE.

Summaries for ASE Data

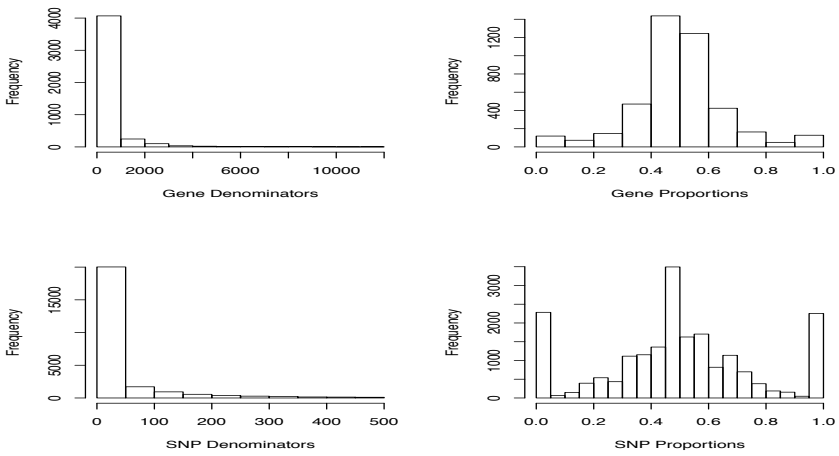


Figure 9 : Summaries for RNA BY/RM yeast data * 739 SNP denominators > 500.

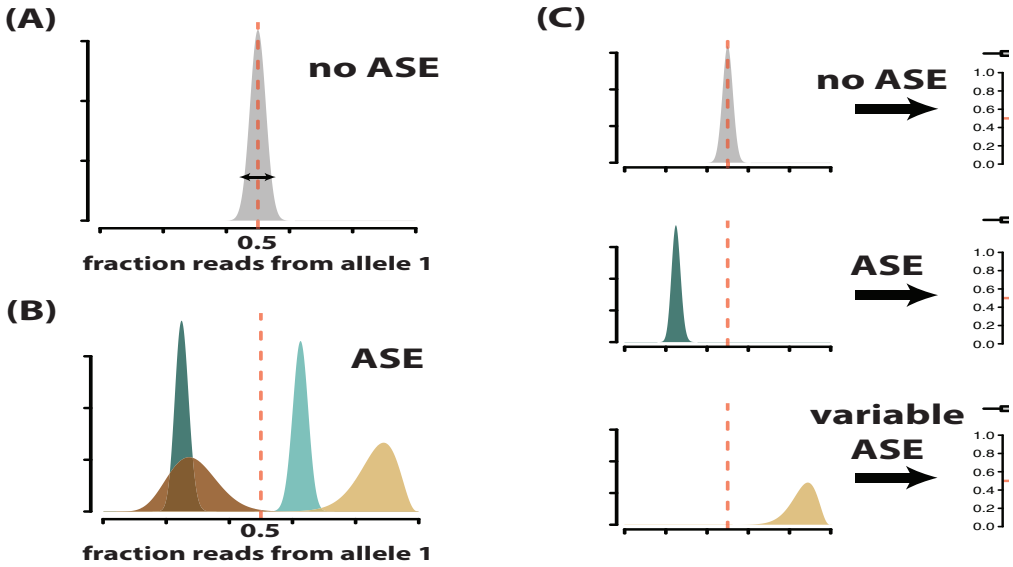


Figure 10 : Schematic of the hierarchical model.

Approach to Modelling RNASeq Data

Overview, three models fitted:

1. Model 1: Two component mixture model to filter out aberrant SNPs using genomic DNA data.
2. Model 2: Using the filtered genomic DNA data, fit a hierarchical SNP within gene model, to determine the "null" distribution of counts. Specifically: "wobble" in p about 0.5, and SNP "wobble" in p within genes.
   Absence of ASE is not experimentally equivalent to $Y_i \sim \text{binomial}(N_i, p = 0.5)$ because of the steps involved in the experiment.
3. Model 3: For the RNA Seq data develop a two-component mixture model where each gene either displays no ASE, or ASE, with null component determined from the analysis of the genomic DNA data (Model 2).

Model 1: Filtering Model for Genomic DNA

Two-component mixture model for SNPs:

1. Majority of SNP counts arise from a beta-binomial distribution with p "close to" 0.5.
2. Minority of SNP counts arise from a beta-binomial distribution with p "not close to" 0.5 due to sequencing bias at these SNPs.

- Data: y_j and N_j are counts at SNP j for $j = 1, ..., m$ SNPs.
- Note: Ignores gene information – don't want to impose too much structure at this point.
- SNPs that are more likely to arise from component 2 are then removed from further analyses.

Filtering Model for Genomic DNA

- *Stage 1: SNP Count Likelihood:*

$$y_j|p_j \sim \text{binomial}(N_j, p_j), \quad j = 1, ..., N.$$

- *Stage 2: Between-SNP Prior:*

$$p_j|a, b, c, \pi_0 = \begin{cases} \text{beta}(a, a) & \text{with probability } \pi_0 \\ \text{beta}(b, c) & \text{with probability } 1 - \pi_0 \end{cases}$$

- *Stage 3: Hyperpriors:* Constrain $b < 1$, $c < 1$ to give U-shaped beta distribution.

$$\begin{aligned} a &\sim \text{lognormal}(4.3, 1.8)^\star \\ b &\sim \text{uniform}(0, 1) \\ c &\sim \text{uniform}(0, 1) \\ \pi_0 &\sim \text{uniform}(0, 1) \end{aligned}$$

* 80% interval for p : $[0.43, 0.57]$. Separate a, b, c, π_0 for each technology.

Implementation for Genomic DNA

- Integrate p_j from model to give:

$$y_j|a, b, c, \pi_0 \sim \pi_0 \times \text{beta-binomial}(N_j, a, a) + (1 - \pi_0) \times \text{beta-binomial}(N_j, b, c).$$

- This is a mixture of two distributions:
  1. The first distribution is for the majority of signals close to 0.5. The size of a denotes how close is close.
  2. The second distribution is for the minority of aberrant SNPs.

Implementation for Genomic DNA

- Likelihood:

$$\Pr(\mathbf{y}|a,b,c,\pi_0) = \prod_{j=1}^{N} \binom{N_j}{Y_j} \left\{ \pi_0 \frac{\Gamma(2a)}{\Gamma(a)^2} \frac{\Gamma(y_j+a)\Gamma(N_j-y_j+a)}{\Gamma(N_j+2a)} + (1-\pi_0)\frac{\Gamma(b+c)}{\Gamma(b)\Gamma(c)} \frac{\Gamma(y_j+b)\Gamma(N_j-y_j+c)}{\Gamma(N_j+b+c)} \right\}$$

- Posterior:

$$p(a,b,c,\pi_0|\mathbf{y}) \propto \Pr(\mathbf{y}|a,b,c,\pi_0) \times p(a)p(b)p(c)p(\pi_0).$$

- *Implementation:* Markov chain Monte Carlo.
  - Recall: Sequencing bias lead to aberrant SNPs, and these errors are likely to be repeated in the main experiment.
  - SNPs falling in the second mixture component were removed from further analyses.

Posterior Distributions

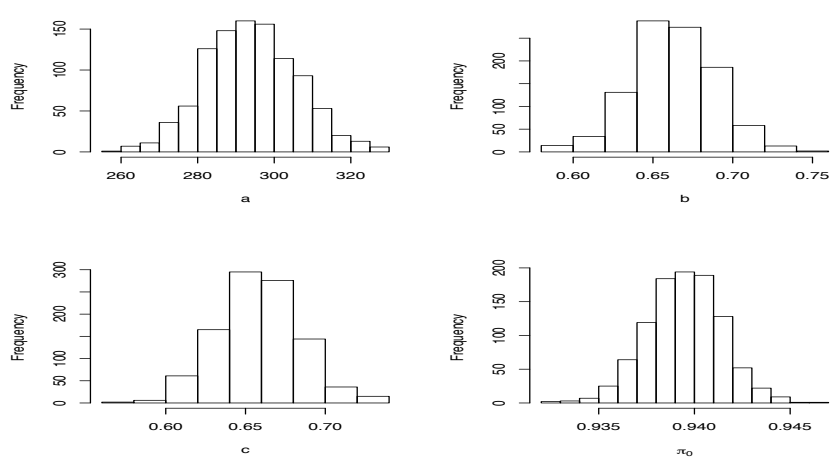


Figure 11 : Posteriors for genomic filtering model for Illumina platform.

Posterior Filter

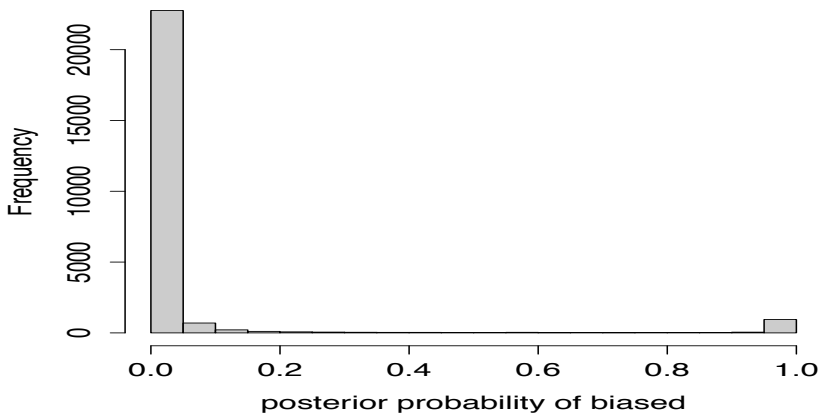


Figure 12 : Posterior probabilities of biased genomic DNA SNPs: 1,295 removed from 25,262.

Effect of Filtering

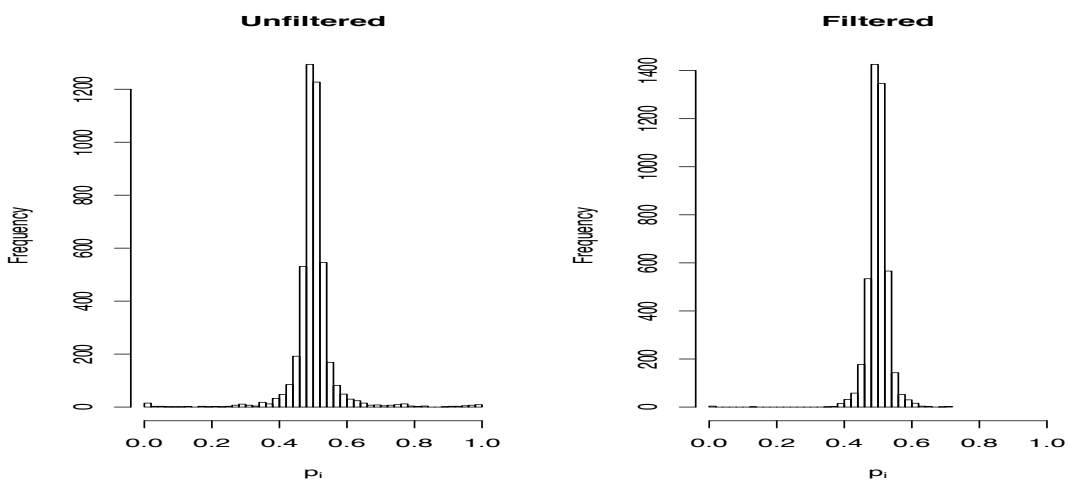


Figure 13 : Original and filtered data, for Illumina platform.

Model 2: Calibration Model for Genomic Data

- With aberrant SNPs removed, the next step is to calibrate the null component.
- *Stage 1: Within-Gene Likelihood:*

$$Y_{ij}|p_{ij} \sim \text{binomial}(N_{ij}, p_{ij}).$$

where p_{ij} is the probability of an outcome from the first genetic background.

- *Stage 2: Within-Gene Prior:*

$$p_{ij}|\alpha_i, \beta_i \sim \text{beta}(\alpha_i, \beta_i)$$

so that α_i, β_i determine the distribution of variants within gene i.

Calibration Model for Genomic Data

- α_i and β_i are not straightforward to interpret (as we saw in Lecture 3).
- We reparameterize $(\alpha_i, \beta_i) \rightarrow (p_i, e_i)$ with mean and dispersion parameters (recall $\alpha_i + \beta_i$ is a prior sample size):

$$p_i = \frac{\alpha_i}{\alpha_i + \beta_i}$$
$$e_i = \frac{1}{1 + \alpha_i + \beta_i}$$

- Moments of ASE parameters:

$$\begin{aligned} E[p_{ij}|p_i, e_i] &= p_i \\ var(p_{ij}|p_i, e_i) &= p_i(1 - p_i)e_i \end{aligned}$$

- Moments of data:

$$\begin{aligned} E[Y_{ij}|p_i, e_i] &= N_{ij}p_i \\ var(Y_{ij}|p_i, e_i) &= N_{ij}p_i(1 - p_i)\left[1 + (N_{ij} - 1)e_i\right] \end{aligned}$$

  - As $e_i \rightarrow 0$ we approach the binomial model.
  - As $e_i \rightarrow 1$ we have increased overdispersion (variability within gene).

Calibration Model for Genomic data

- *Stage 3: Within-Gene Likelihood:*

$$\begin{aligned} p_i|a &\sim \text{beta}(a, a) \\ e_i|d &\sim \text{beta}(1, d) \end{aligned}$$

Note: prior on within-gene dispersion is monotonic decreasing from 0 (corresponding to no variability).

- *Stage 4: Hyperpriors:* Require priors on $a > 0, d > 0$.
- We take

$$\begin{aligned} a &\sim \text{lognormal}(4.3, 1.8) \\ d &\sim \text{exponential}(0.0001) \end{aligned}$$

- The latter prior determines the within-gene variability within-gene variability in genomic DNA – chosen by examination of resultant p_{ij}'s.
- Separate a, d for each technology.

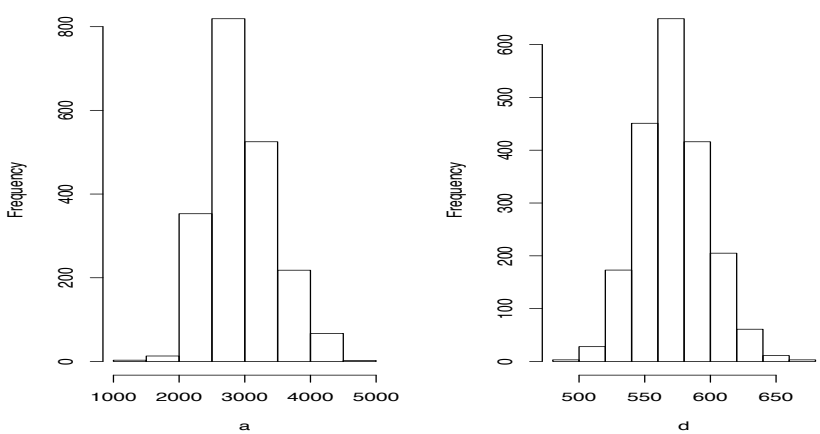


Figure 14 : Posteriors for the RNA-Seq data, Illumina platform.

Model 3: Model for RNA-Seq Data

- Data are modeled as a two-component mixture: the first "null" component having a known distribution, from the genomic DNA analysis on the filtered data.
- *Stage 1: Within-Gene Likelihood:*

$$Y_{ij}|p_{ij} \sim \text{binomial}(N_{ij}, p_{ij}).$$

where p_{ij} is the probability of an outcome from the first genetic background.

- *Stage 2: Within-Gene Prior:*

$$p_{ij}|\alpha_i, \beta_i \sim \text{beta}(\alpha_i, \beta_i)$$

so that α_i, β_i determine the distribution of variants within gene i.

- *Stage 3: Between-Gene Prior:* We again reparameterize $(\alpha_i, \beta_i) \to (p_i, e_i)$:

$$p_i, e_i|f, g, h, \pi_0 \sim \begin{cases} \text{beta}(\widehat{a}, \widehat{a}) \times \text{beta}(1, \widehat{d}) & \text{with probability } \pi_0 \\ \text{beta}(f, g) \times \text{beta}(1, h) & \text{with probability } 1 - \pi_0 \end{cases}$$

with $\widehat{a}, \widehat{d}$ from genomic DNA analysis.

Stage 4: Hyperpriors: Require priors on $\pi_0, f > 0, g > 0, h > 0$.

- Uniform prior on π_0.
- f and g describe beta distribution of p_i for genes displaying ASE – want this distribution to be centered around symmetry.
- Reparameterize as

$$q = \frac{f}{f+g} \qquad r = \frac{1}{1+f+g}$$

so that $E[p_i] = q$, $var(p_i) = q(1-q)r$.

- Through experimentation:

$$q \sim \text{beta}(100, 100) \quad r \sim \text{beta}(1, 20)$$

- For h, the distribution of within-gene variability in ASE:

$$h \sim \text{exponential}(0.03).$$

- Separate f, g, h for each technology.

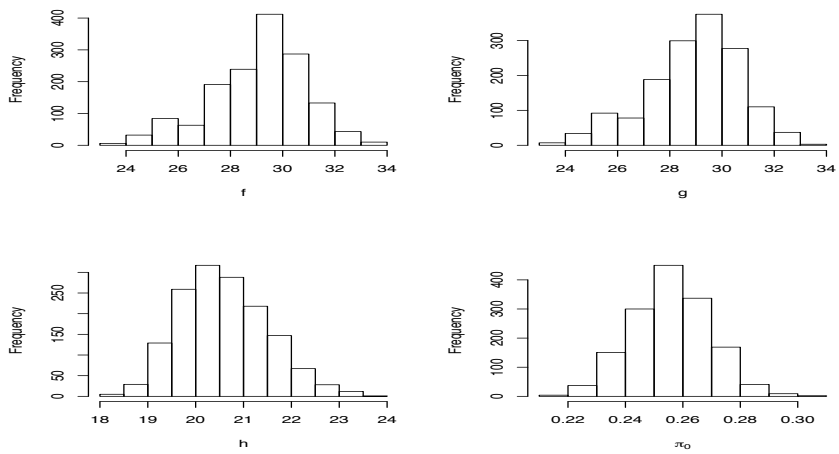


Figure 15 : Posteriors for the RNA-Seq data, Illumina platform.

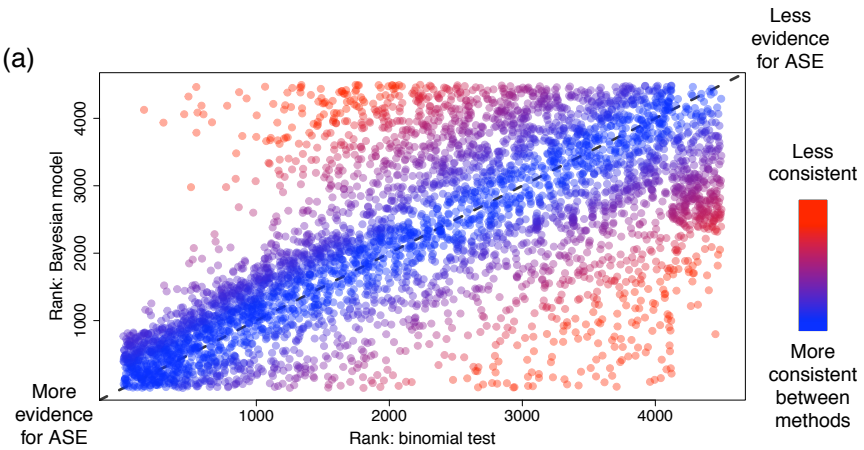


Figure 16 : Comparison of rankings from binomial test and hierarchical model.

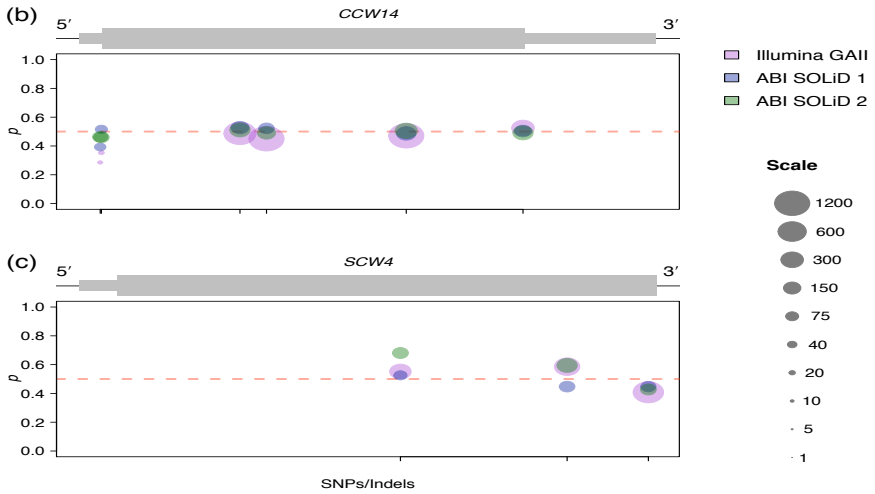


Figure 17 : Examples of opposite conclusions: In (b) the p-value said ASE and Bayes not. (large sample size, Bayes allows wobble). In (c) the p-value said no ASE, Bayes analysis yes.

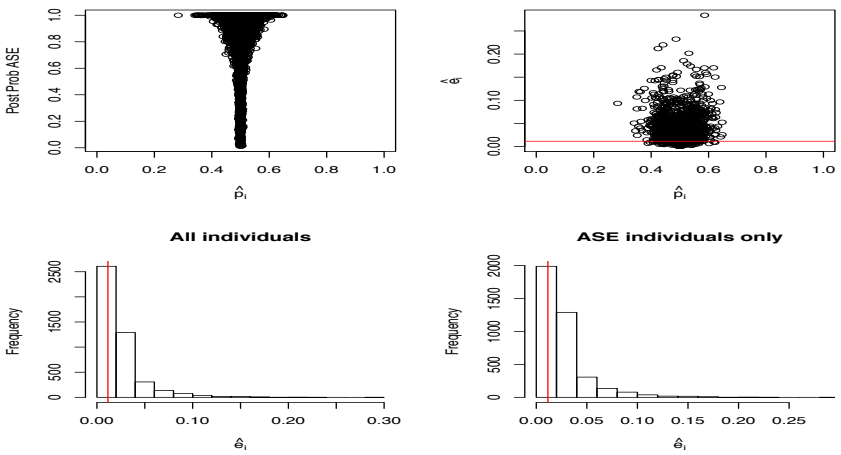


Figure 18 : Between-gene variability p_i and within-gene variability e_i.

Varying ASE within genes

- One mechanism: Imagine a gene with an exon and an intron, and that we have SNPs in both.
- At each exonic SNP we see approximately the same number of BY and RM reads.
- Now suppose the intron is not spliced out for the BY allele, but it is spliced out efficiently for the RM allele. At each intronic SNP we will still see the same number of BY reads as in the exon (everything else being equal), but approximately 0 RM reads, leading to variable ASE across the gene
- In the figure: The "thin" part of the gene (YML024W) is an intron, while the "thick" part is an exon.
- For the RM allele (magenta) the intron is not spliced out, while it is mostly spliced out in the BY allele (green).

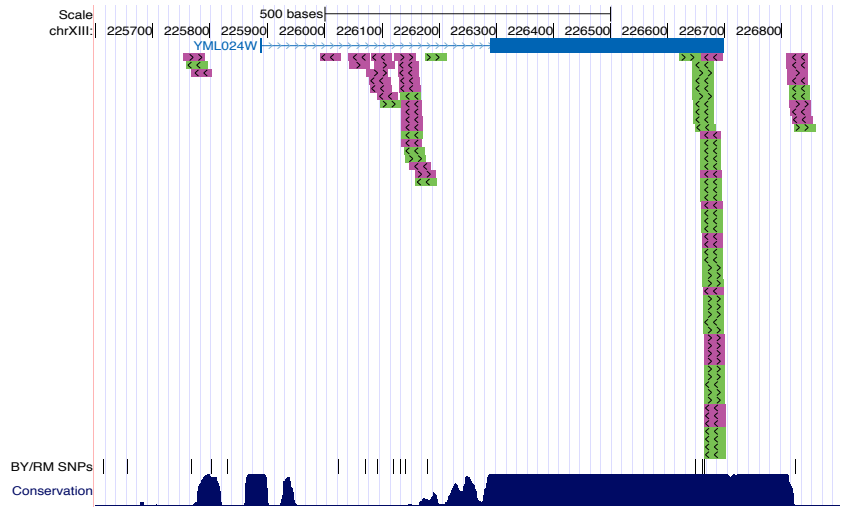


Figure 19 : Example of a gene displaying variable ASE within a gene. Green = RM, magenta = BY.

Conclusions for Mixture Model

- For the ASE data we used the DNA experiment to calibrate the prior.
- More details of this experiment and the model can be found in Skelly *et al.* (2011).
- Implementation was via Markov chain Monte Carlo, but we had to write our own code.

Model Comparison

- Markov chain Monte Carlo in particular has allowed the fitting of more and more complex models, often hierarchical in nature with layers of random effects.
- The search for a method to find the "best" of a set of candidate models has also grown.
- Let $p(\mathbf{y}|\boldsymbol{\theta})$ represent a generic likelihood for $\mathbf{y} = [y_1, \ldots, y_n]$ and let

$$D(\boldsymbol{\theta}) = -2\log[p(\mathbf{y}|\boldsymbol{\theta})]$$

  represent the deviance.
- For example, in an iid normal$(\mu_i(\boldsymbol{\theta}), \sigma^2)$ normal the deviance is

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}[y_i - \mu_i(\boldsymbol{\theta})]^2.$$

- Frequentist model comparison for nested models is often carried out using likelihood ratio statistics, which corresponds to the comparison of deviances in generalized linear models (GLMs), see for example McCullagh and Nelder (1989).

Model Comparison: AIC

- One approach to model comparison is based on a model's ability to make good predictions.
- Such an objective, and predicting the actual observed data, leads to Akaike's an information criterion (AIC), derived in Akaike (1973).
- In AIC one tries to estimate the (Kullback-Leibler) distance between the true distribution of the data, and the modeled distribution of the data.
- AIC is given by
$$\text{AIC} = -2\log[p(y|\widehat{\boldsymbol{\theta}})] + 2k$$
where $\widehat{\theta}$ is the MLE and k is the number of parameters in the model, i.e. the size of $\boldsymbol{\theta}$.
- Small values of the AIC are favored, since they suggest low prediction error.
- The penalty term $2k$ penalizes the double use of the data.
- In general for prediction: overly complex models are penalized since redundant parameters "use up" information in the data.

Model Comparison: BIC

- Another approach is based on trying to identify the "true" model.
- Schwarz (1978) developed the Bayesian Information Criterion (BIC) which is given by
$$\text{BIC} = -2\log[p(y|\widehat{\boldsymbol{\theta}})] + k\log n.$$
- BIC approximates $-2\log p(\mathbf{y}|\boldsymbol{\theta})$ under a certain unit information prior (Kass and Wasserman, 1995).
- BIC is consistent[1] for finding the true model, if that model lies in the set being compared.
- AIC is not consistent for finding the true model, but recall is intended for prediction.

[1] meaning the BIC hones in on the true model as the sample size increases

Model Comparison: DIC

- Spiegelhalter *et al.* (2002) introduced what has proved to be a very popular model comparison statistic, the deviance information criterion (DIC).
- To define the DIC, define an "effective number of parameters as

$$\begin{aligned} p_D &= E_{\theta|y}\{-2\log[p(\mathbf{y}|\boldsymbol{\theta})]\} + 2\log[p(\mathbf{y}|\overline{\boldsymbol{\theta}})] \\ &= \overline{D} + D(\overline{\boldsymbol{\theta}}) \end{aligned}$$

  where $\overline{\boldsymbol{\theta}} = E[\boldsymbol{\theta}|\mathbf{y}]$ is the posterior mean, $D(\overline{\boldsymbol{\theta}})$ is the deviance evaluated at the posterior mean and $\overline{D} = E[D|\mathbf{y}]$.
- Hence, p_D is the

$$\text{posterior mean deviance} - \text{deviance of posterior means.}$$

- The DIC is given by

$$\begin{aligned} \text{DIC} &= D(\overline{\boldsymbol{\theta}}) + 2p_D \\ &= \overline{D} + p_D, \end{aligned}$$

  so that we have a measure of goodness of fit + complexity.
- DIC is straightforward to evaluate using MCMC or INLA.

Model Comparison: DIC

DIC has been heavily criticized (Spiegelhalter *et al.*, 2014):

- p_D is not invariant to parameterization.
- DIC is not consistent for choosing the correct model.
- DIC has a weak theoretical justification and is not universally applicable.
- DIC has been shown to under penalize complex models (Plummer, 2008; Ando, 2007).
- See Spiegelhalter *et al.* (2014) for an interesting discussion of the history of DIC, including a summary of attempts to improve DIC.
- According to Google Scholar, as of June 20th, 2014, Spiegelhalter *et al.* (2002) has 5251 citations...

Model Comparison: CPO

- Another approach based on prediction uses the conditional predictive ordinate (CPO).
- Let
$$\mathbf{y}_{-i} = [y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n]$$
represent the vector of data with the i-th observation removed.
- The idea is to predict the density ordinate of the left-out observation, based on those that remain.
- Specifically, the CPO for observation i is defined as:

$$\begin{aligned} \text{CPO}_i &= p(y_i|\mathbf{y}_{-i}) \\ &= \int p(y_i|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y}_{-i})\, d\boldsymbol{\theta} \\ &= E_{\theta|y_{-i}}\left[p(y_i|\boldsymbol{\theta})\right] \end{aligned}$$

Model Comparison: CPO

- The CPOs can be used to look at local fit, or one can define an overall score for each model:
$$\log(\text{CPO}) = \sum_{i=1}^{n} \log \text{CPO}_i.$$
- Good models will have relatively high values of $\log(\text{CPO})$.
- See Held *et al.* (2010) for a discussion of shortcuts for estimation (i.e. avoidance of fitting the model n times) using MCMC and INLA.

Model Comparison: Illustration, Childhood Mortality in Tanzania

- We illustrate the use of CPO and DIC in a study of estimating childhood (under 5) mortality in regions of Tanzania.
- The data are collected via a series of 8 surveys in 21 regions covering the period 1980–2009.
- Let q_{its} be the childhood mortality in area i, time point t from survey s.
- Based on the surveys we can obtain weighted (Horvitz-Thompson) estimators $\widehat{q}_{its}$ with associated asymptotic variances V_{its}.
- We summarize the data via logit estimates

$$y_{its} = \log\left(\frac{\widehat{q}_{its}}{1-\widehat{q}_{its}}\right).$$

- Let

$$\phi_{its} = \log\left(\frac{q_{its}}{1-q_{its}}\right)$$

represent the logit of the childhood mortality.

Model Comparison: Illustration, Childhood Mortality in Tanzania

We have a three-stage hierarchical model:

- Stage 1: Likelihood:

$$y_{its}|\phi_{its} \sim \text{normal}(\phi_{its}, V_{its}).$$

and we compare the following six models:

$$\begin{aligned}
\text{Model 1: } \phi_{its} &= \mu + \alpha_t + \gamma_t + \theta_i + \eta_i + \delta_{it} \\
\text{Model 2: } \phi_{its} &= \mu + \alpha_t + \gamma_t + \theta_i + \eta_i + \delta_{it} + \nu_s \\
\text{Model 3: } \phi_{its} &= \mu + \alpha_t + \gamma_t + \theta_i + \eta_i + \delta_{it} + \nu_s + \nu_{is} \\
\text{Model 4: } \phi_{its} &= \mu + \alpha_t + \gamma_t + \theta_i + \eta_i + \delta_{it} + \nu_s + \nu_{ts} \\
\text{Model 5: } \phi_{its} &= \mu + \alpha_t + \gamma_t + \theta_i + \eta_i + \delta_{it} + \nu_s + \nu_{ts} + \nu_{is} \\
\text{Model 6: } \phi_{its} &= \mu + \alpha_t + \gamma_t + \theta_i + \eta_i + \delta_{it} + \nu_s + \nu_{ts} + \nu_{is} + \nu_{its}
\end{aligned}$$

where α_t, θ_i, δ_{it} are independent random effects for time, area and the interaction, γ_t and η_i are random effects that carry out local smoothing in time and space and $\nu_s, \nu_{ts}, \nu_{is}, \nu_{its}$ are independent random effects to reflect survey effects.

- Stage 2: Normal random effects Distributions.
- Stage 3: Hyperpriors on μ and the random effects variances.

Model Comparison: Illustration, Childhood Mortality in Tanzania

Table 1 : Model comparison statistics for 6 models for the Tanzania data; "best" in red.

Model	No. Parameters	p_D	$\bar{D}$	DIC	$\log$(CPO)
2	181	75	409	484	-295
2	189	81	382	463	-288
3	313	120	219	339	-193
4	223	91	364	454	-282
5	347	128	202	330	-182
6	920	149	185	334	-184

- Notice how much smaller the effective number of parameters is, when compared with the total number of parameteres; this is because of the shrinkage/penalization of the random effects distributions.
- Both CPO and DIC suggest that model 5 is the best:

$$\text{Model 5: } \phi_{its} = \mu + \alpha_t + \gamma_t + \theta_i + \eta_i + \delta_{it} + \nu_s + \nu_{ts} + \nu_{is}$$

- So survey effects vary across time and across areas (different teams sent out).

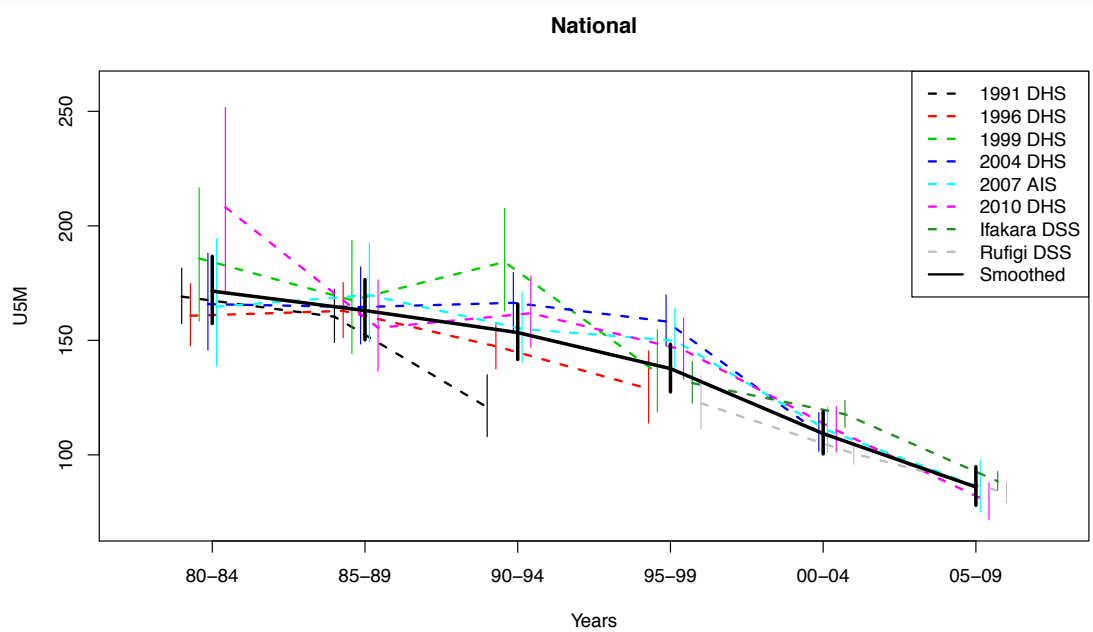


Figure 20 : Smoothed estimates of national under 5 mortality in Tanzania (solid line), different surveys denoted with dashed lines and vertical lines represent 95% interval estimates.

Conclusions

- Hierarchical models allow complex dependencies within data to be modeled.
- Prior specification for variance components is not straightforward, and sensitivity analysis is a good idea.
- No universally agreed upon approach to carrying out model comparison.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In P. B.N. and C. F., editors, *Second International Symposium on Information Theory*, pages 267–281. Akademia Kiadó, Budapest.

Ando, T. (2007). Bayesian predictive information criterion for the evaluation of hierarchical bayesian and empirical bayes models. *Biometrika*, **94**, 443–458.

Browning, B. and Browning, S. (2009). A unified approach to genotype imputation and haplotype-based inference for large data sets of trios and unrelated individuals. *The American Journal of Human Genetics*, **84**, 1084–1097.

Held, L., Schrödle, B., and Rue, H. (2010). Posterior and cross-validatory predictive checks: A comparison of MCMC and INLA. In T. Kneib and G. Tutz, editors, *Statistical Modeling and Regression Structures – Festschrift in Honour of Ludwig Fahrmeir*, pages 91–110. Physica-Verlag.

Howie, B., Donnelly, P., and Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics*, **5**, e1000529.

Huang, L., Yun, L., Singleton, A., Hardy, J., Abecasis, G., Rosenberg, N., and Scheet, P. (2009). Genotype-imputation accuracy across worldwide human populations. *The American Journal of Human Genetics*, **84**, 235–250.

Kass, R. and Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the schwarz criterion. *Journal of the American Statistical Association*, **90**, 928–934.

Li, Y., Willer, C., Fing, J., Scheet, P., and Abecasis, G. (2010). Mach: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic Epidemiology*, **34**, 816–834.

Marchini, J. and Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, **11**, 499–511.

Marchini, J., Howie, B., Myers, S., McVean, G., and Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics*, **39**, 906–913.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models, Second Edition*. Chapman and Hall, London.

Plummer, M. (2008). Penalized loss functions for bayesian model comparison. *Biostatistics*, **9**, 523–539.

Sanna, S., Li, B., and Mulas, A. (2011). Fine mapping of five loci associated with low density lipoprotein choesterol fetects variants that double the explained heritability. *PLoS Genetics*, **7**, 1002198.

Scheet, P. and Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *The American Journal of Human Genetics*, **78**, 629–644.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464.

Skelly, D., Johansson, M., Madeoy, J., Wakefield, J., and Akey, J. (2011). A powerful and flexible statistical framework for testing hypothesis of allele-specific gene expression from RNA-Seq data. *Genome Research*, **21**, 1728–1737.

Spiegelhalter, D., Best, N., Carlin, B., and Linde, A. V. D. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B*, **64**, 583–639.

Spiegelhalter, D., Best, N., Carlin, B., and Linde, A. V. D. (2014). The deviance information criterion: 12 years on (with discussion). *Journal of the Royal Statistical Society: Series B*, **64**, 485–493.