2019 SISG Module 8: Bayesian Statistics for Genetics
Lecture 3: Binomial Sampling

Jon Wakefield

Departments of Statistics and Biostatistics
University of Washington

Outline

Introduction

Introduction

In this lecture we will consider the Bayesian modeling of binomial data.

Two motivations for a binomial model:

- a so-called allele specific expression (ASE) experiment will be considered.
- a time series of counts, in order to model prevalence of a condition.

Conjugate priors will be described in detail.

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- (local) or trans (distant)-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

- The data we consider is in yeast, and is a controlled experiment in which two strains, BY and RM, are hybridized.
- 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

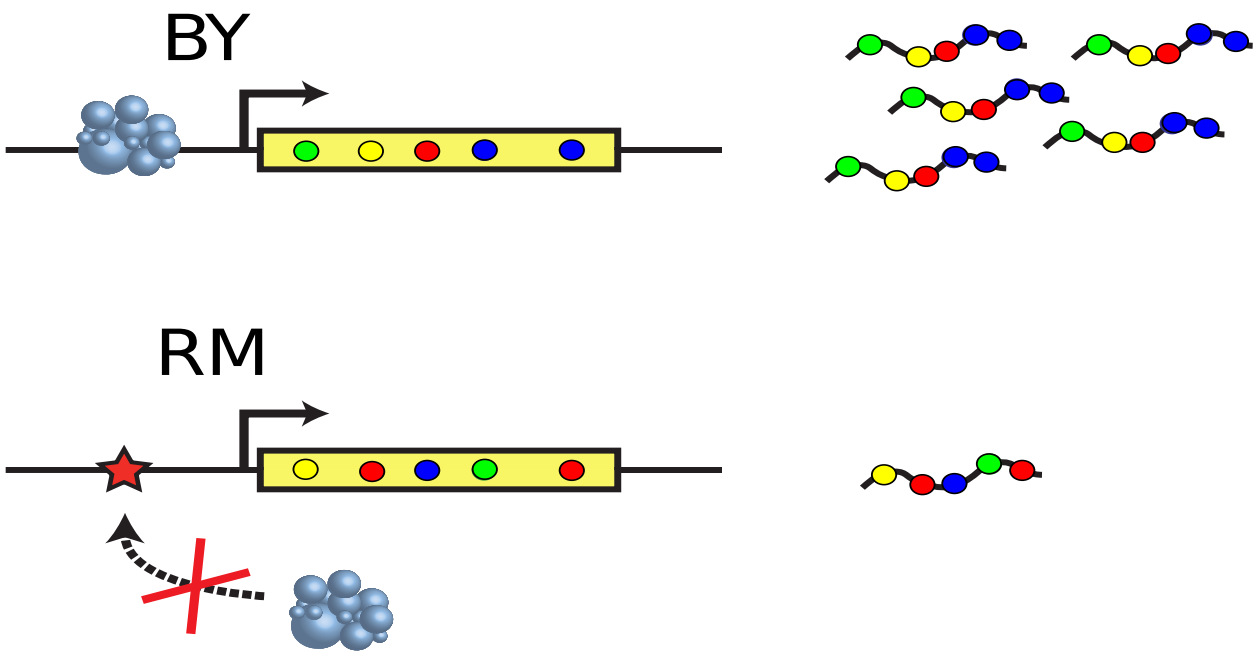


Figure: In the top figure the transcription factor (blue) leads to high transcription. In the bottom figure an upstream polymorphism (red star) prevents the transcription factor from binding.

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).

The Data

Table: First few rows of ASE data.

BY Count	Total Count	MLE $\widehat{\theta}$
62	107	0.58
33	59	0.56
658	1550	0.42
14	61	0.23
57	153	0.37
218	451	0.48
10	19	0.53
⋮	⋮	⋮

Simple Approach to Testing for ASE

For a generic gene:

- 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 might use an exact test, i.e. enumerate the probability, 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. $\theta = 0.5$?
  - How would a Bayesian perform inference for this problem?

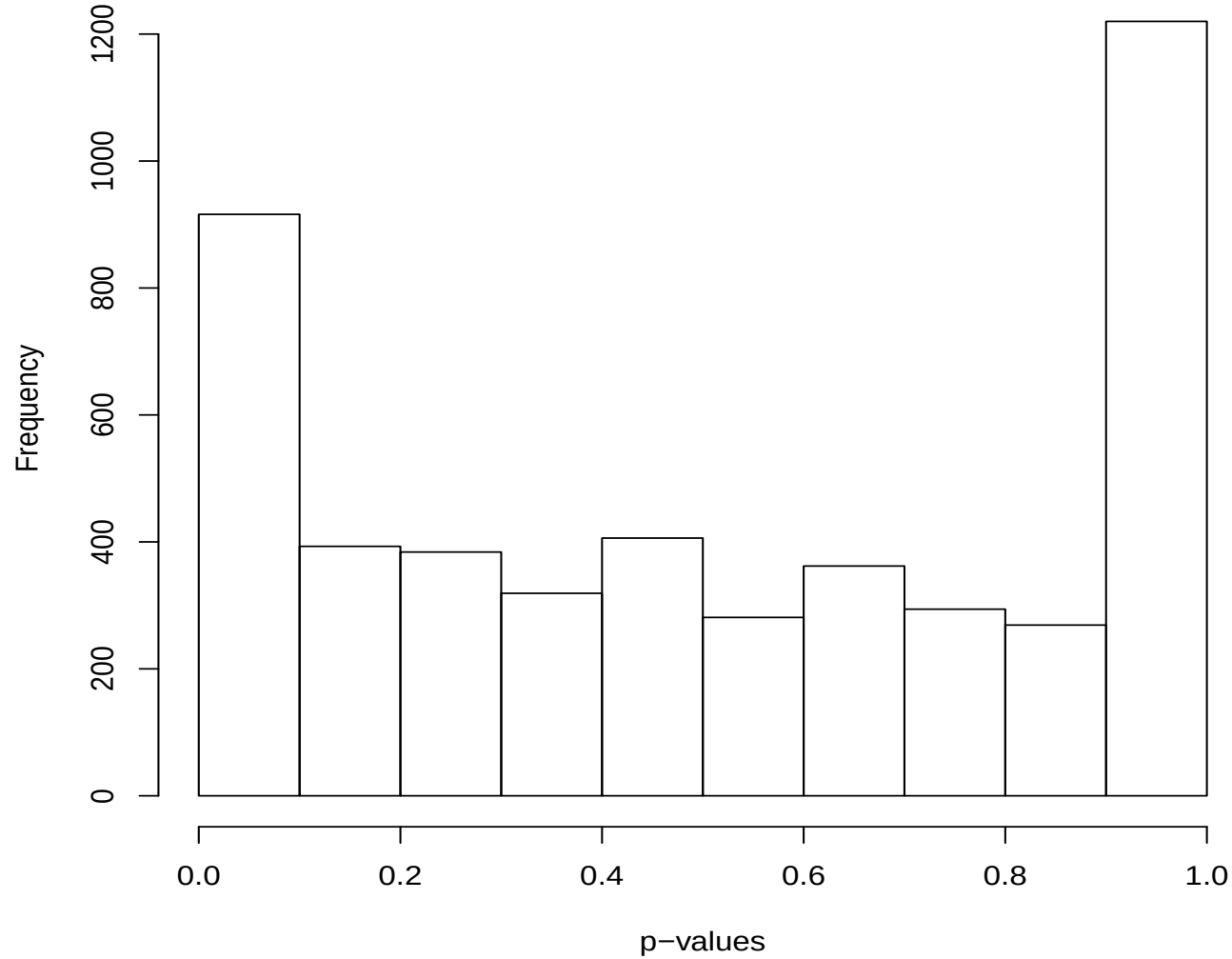


Figure: *p*-values from 4,844 exact tests.

Motivating Example: Smoothing/Penalization

- When looking at estimates over space or time, we want to know if the differences we see are “real”, or simply reflecting sampling variability.
- In data sparse situations, when one expects similarity smoothing local patterns (in time, space, or both) can be highly beneficial.
- This can equivalently be thought of penalization, in which large deviations from “neighbors”, suitably defined, are discouraged.
- In the examples that follow we will generically think of modeling prevalence.
- We give an example of temporal modeling.

Motivation for Smoothing: Temporal Case

- Temporal setting: Even if the underlying prevalence is the same over time, we will see differences in the empirical estimates.
- Figure 3 demonstrates: We sampled binomial data with $n = 10, 20, 200$ and $p = 0.2$ (shown in blue) in all cases.
- In the top plot in particular, we might conclude large temporal variation, but all we are seeing is sampling variation.
- Figure 4 summarizes estimates from a second simulation in which there is a real temporal pattern – here we would not want to oversmooth and remove the trend.
- Later (Lecture 5) I will apply temporal smoothing models to these two sets of data.

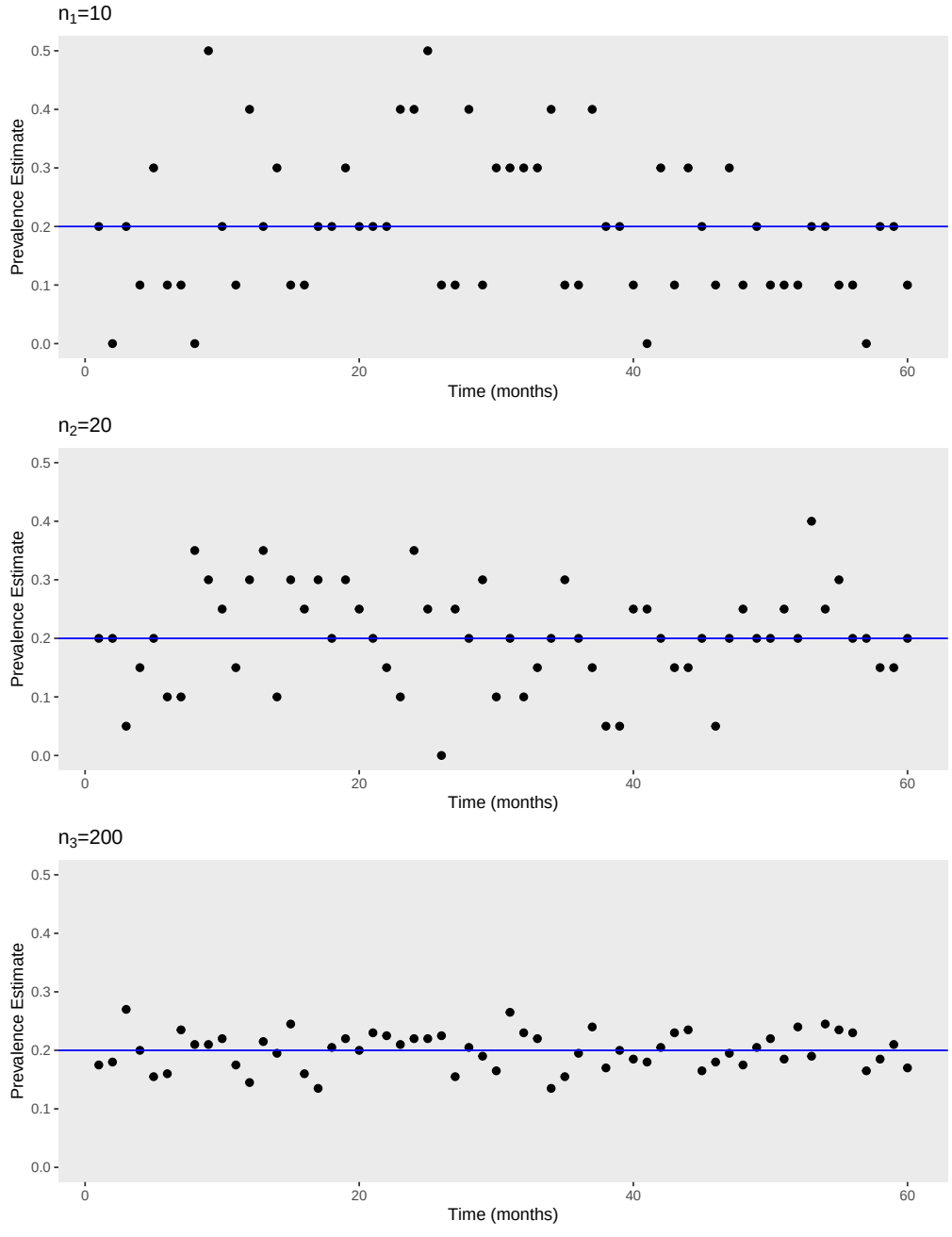


Figure: Prevalence estimates over time from simulated data with true prevalence of $p = 0.2$ (blue solid lines).

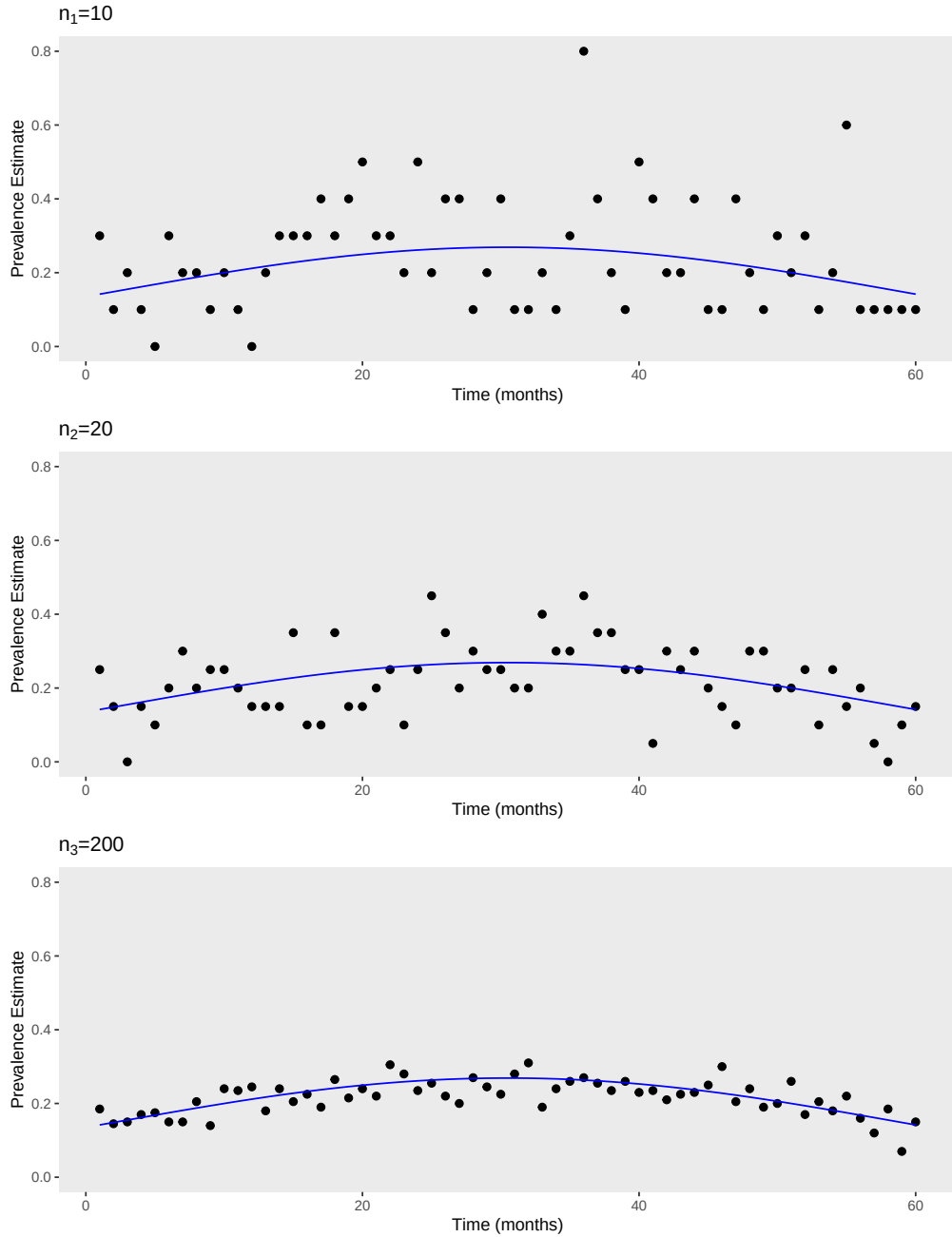


Figure: Prevalence estimates over time from simulated data, true prevalence corresponds to curved blue solid line.

Bayesian Analysis of Binomial Data

Bayes Theorem Recap

- We derive the posterior distribution via Bayes theorem:

$$p(\theta|y) = \frac{\Pr(y|\theta) \times p(\theta)}{\Pr(y)}. \tag{1}$$

- The denominator:

$$\Pr(y) = \int \Pr(y|\theta) \times p(\theta) d\theta = \mathrm{E}[\Pr(y|\theta)]$$

  is a normalizing constant to ensure the RHS of (1) integrates to 1 (we assume a continuous parameter θ).

- More colloquially:

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

  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).

And then do some computation... and interpretation...

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, quasi-likelihood).

Overview of Bayesian Inference

To carry out inference, integration is required, and a large fraction of the Bayesian research literature focusses on this aspect. Bayesian approaches to:

1. Estimation: marginal posterior distributions on parameters of interest.
2. Hypothesis Testing: Bayes factors give the evidence in the data with respect to two or more hypotheses, and provide one approach.
3. Prediction: via the predictive distribution.

These three endeavors 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:

$$\Pr(Y = y|\theta) = \left(\begin{array}{c} N \\ y \end{array} \right) \theta^y (1-\theta)^{N-y} \tag{2}$$

and tells us the probability of seeing $Y = y$, $y = 0, 1, \ldots, N$ given the probability θ.

For fixed y, we may view (2) 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.

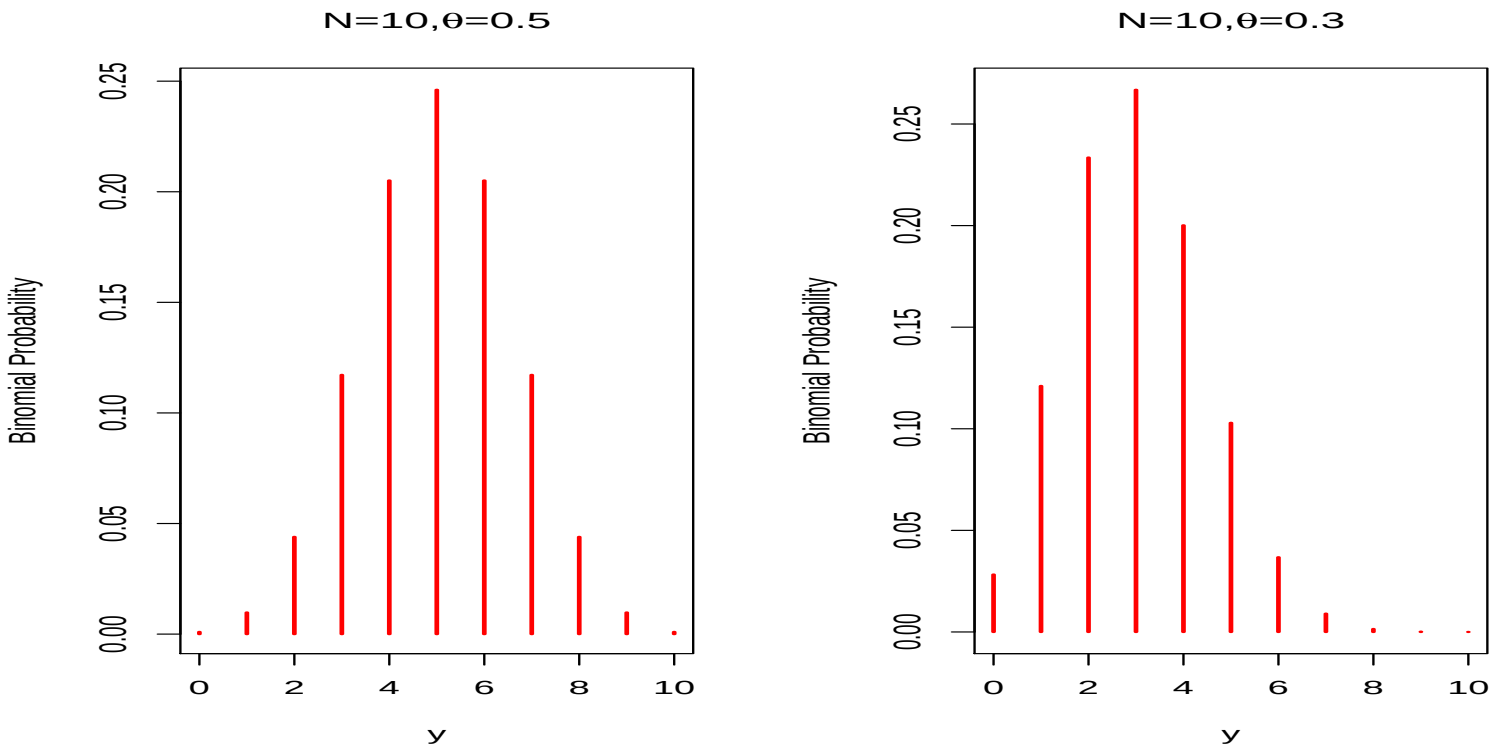
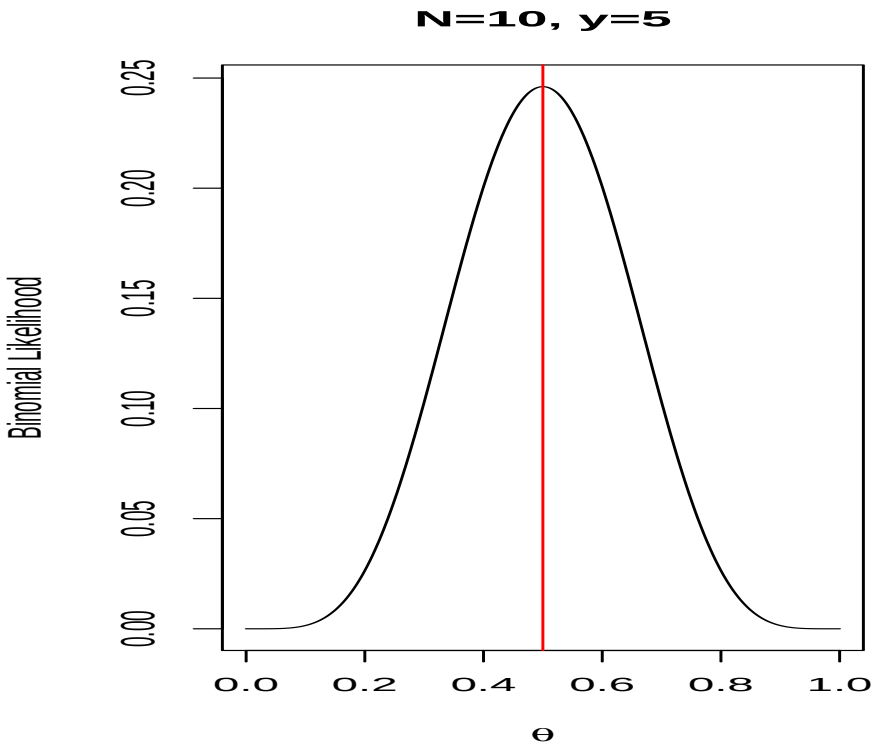


Figure: Binomial distributions for two values of θ with $N = 10$.



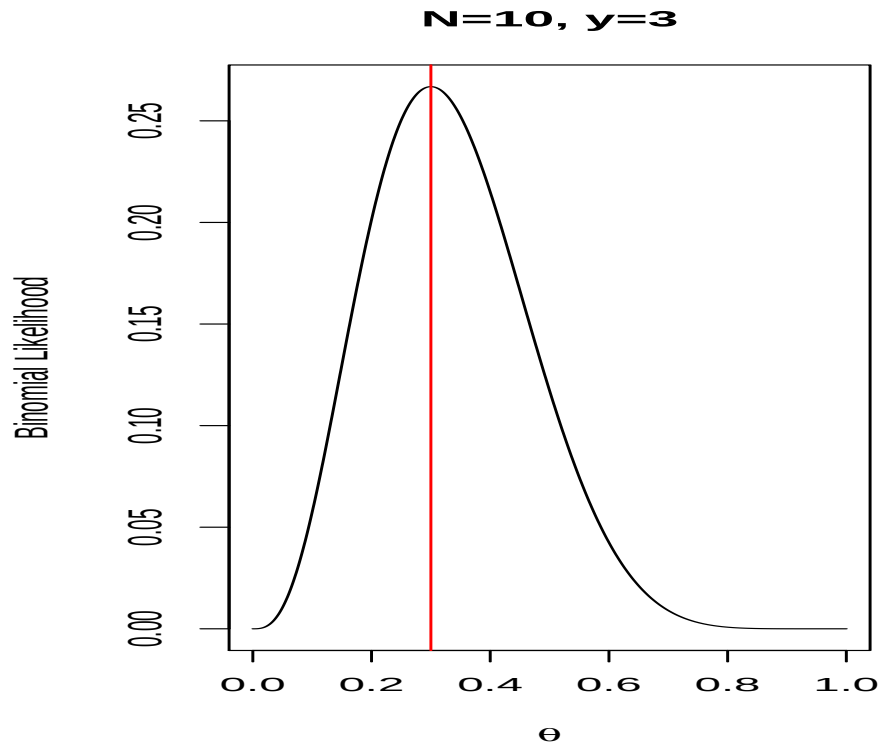


Figure: Binomial likelihoods for values of $y = 5$ (left) and $y = 10$ (right), with $N = 10$. The MLEs are indicated in red.

The Beta Distribution as a Prior Choice for 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, i.e., *a priori*. 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] $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt$

[2] A distribution is valid if it is non-negative and integrates to 1

The Beta Distribution as a Prior Choice for 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)`.

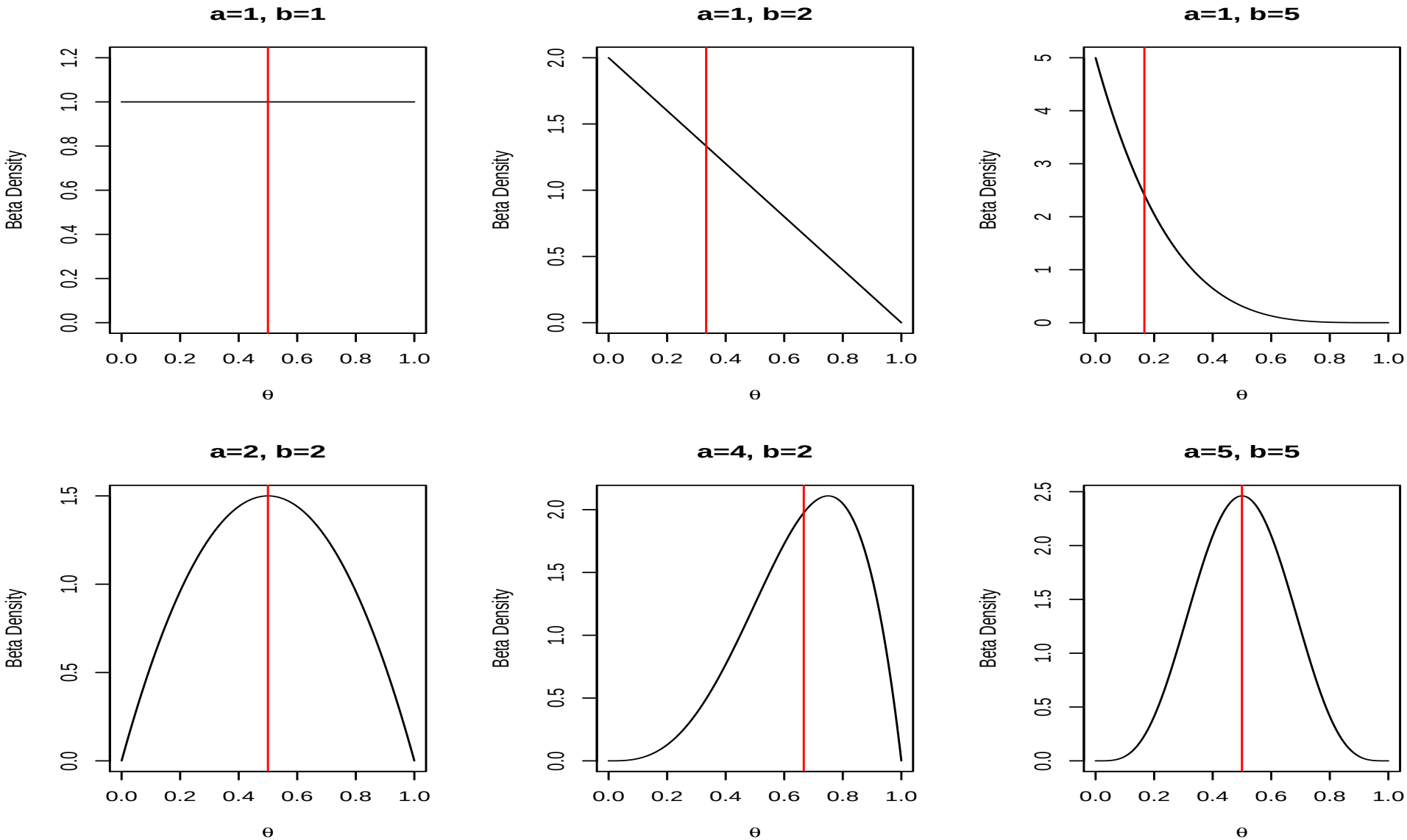


Figure: Beta distributions, beta(a, b), the red lines indicate the means.

Samples to Summarize Beta Distributions

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

In Figure 8 we show histograms of beta distributions for different choices of a and b.

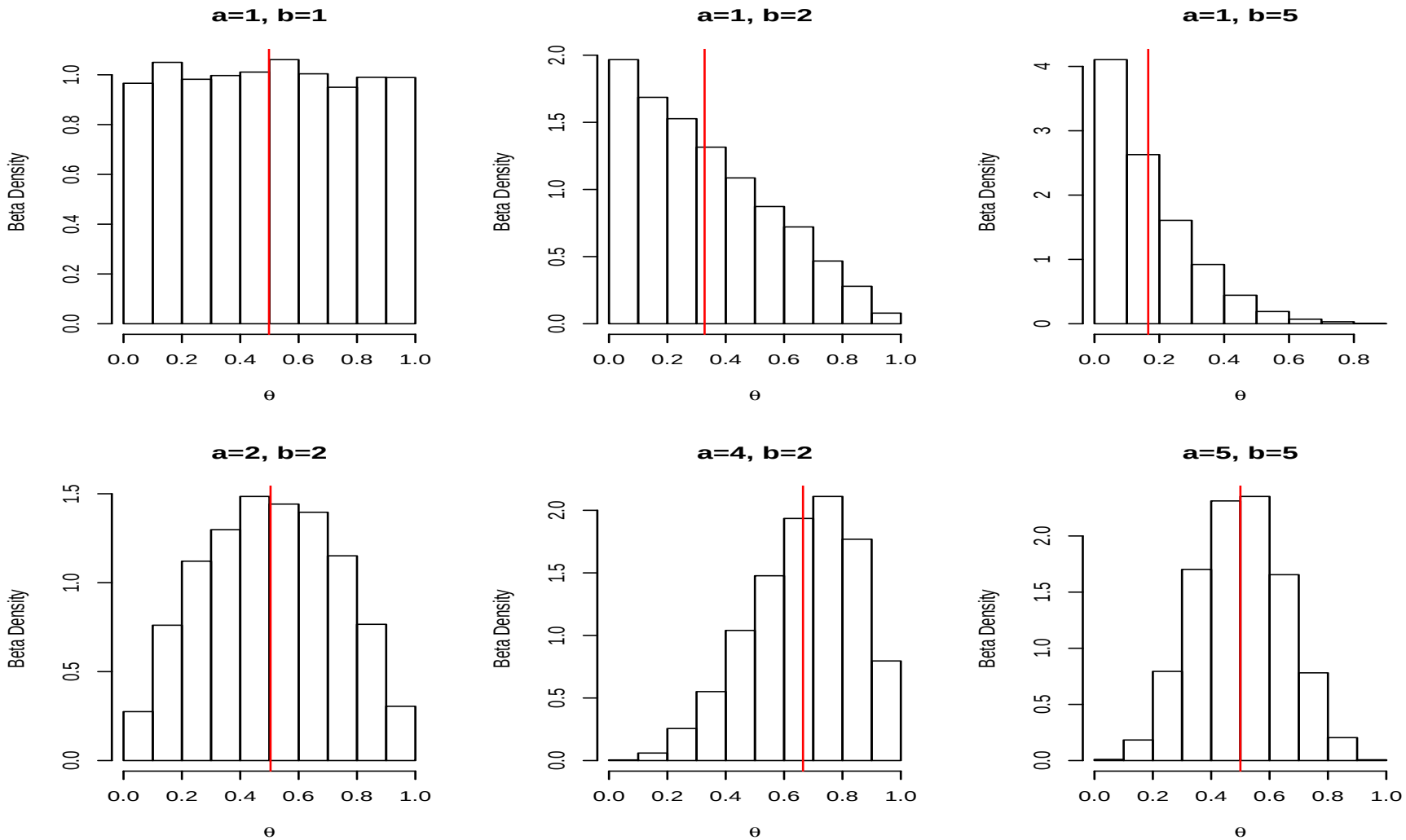


Figure: Random samples from beta distributions; sample means as red lines.

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.

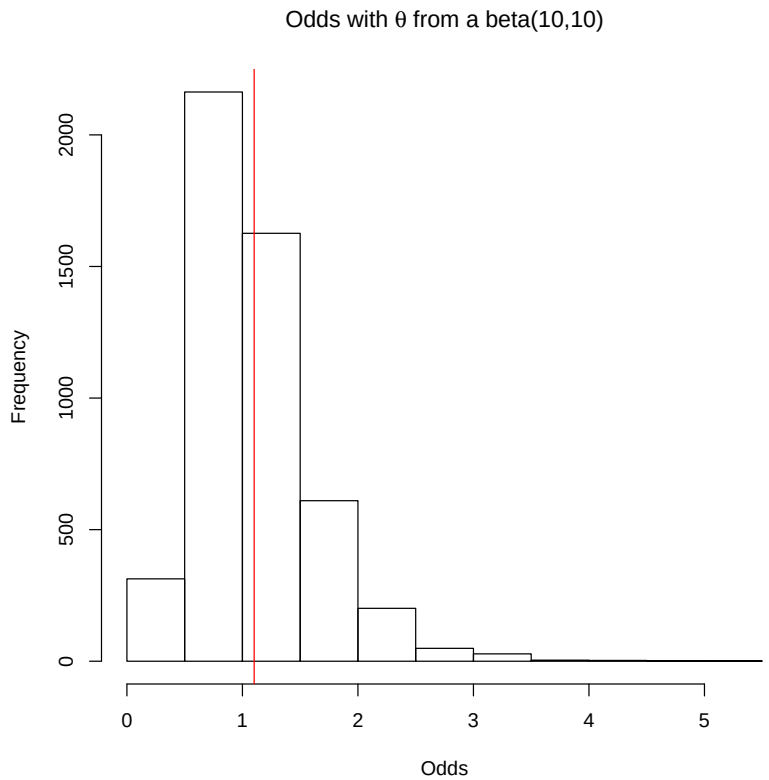


Figure: Samples from the prior on the odds $\theta/(1-\theta)$ with $\theta \sim \text{beta}(10, 10)$, the red line indicates the sample mean.

Issues with Uniformity

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 since, if $\phi = g(\theta)$:

$$\underbrace{p_\phi(\phi)}_{\text{Prior for } \phi} = \underbrace{p_\theta(g^{-1}(\phi))}_{\text{Prior for } \theta} \times \underbrace{\left|\frac{d\theta}{d\phi}\right|}_{\text{Jacobian}}$$

  and so if $g(\cdot)$ is a nonlinear function, the Jacobian will be a function of ϕ and hence not uniform.
- 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 illustrate 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 10.
- Not being uniform on all scales is not necessarily a problem, and is correct probabilistically, but one should be aware of this characteristic.

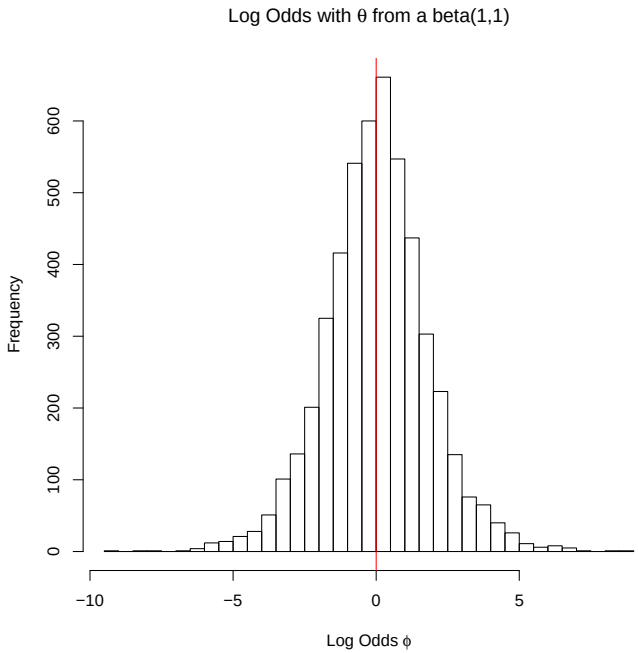


Figure: Samples from the prior on the odds $\phi = \log[\theta/(1-\theta)]$ with $\theta \sim \text{beta}(1,1)$, the red line indicates the sample mean.

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

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

for constants c_1 and c_2, then we know that the random variable 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 $\text{Beta}(y+a, N-y+b)$ distribution.
- We know what the normalizing constant must be, because we have a distribution which must integrate to 1.

Posterior Derivation: The Long (Unnecessary) Way

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

$$\begin{aligned} p(\theta|y) &= \frac{\Pr(y|\theta) \times p(\theta)}{\Pr(y)} \\ &= \frac{1}{\Pr(y)} \left(\begin{array}{c} N \\ y \end{array} \right) \theta^y (1-\theta)^{N-y} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1} (1-\theta)^{b-1}. \quad (3) \end{aligned}$$

- The normalizing constant is

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

- 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 (3) 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 $\theta_0 = 0.5$?

Other Posterior Summaries

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

$$0.05 = \int_0^{\theta_L} p(\theta|y)\, d\theta$$

$$0.95 = \int_0^{\theta_U} p(\theta|y)\, d\theta$$

- 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+b)
[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.

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 $W = 10/(10+5)$ 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

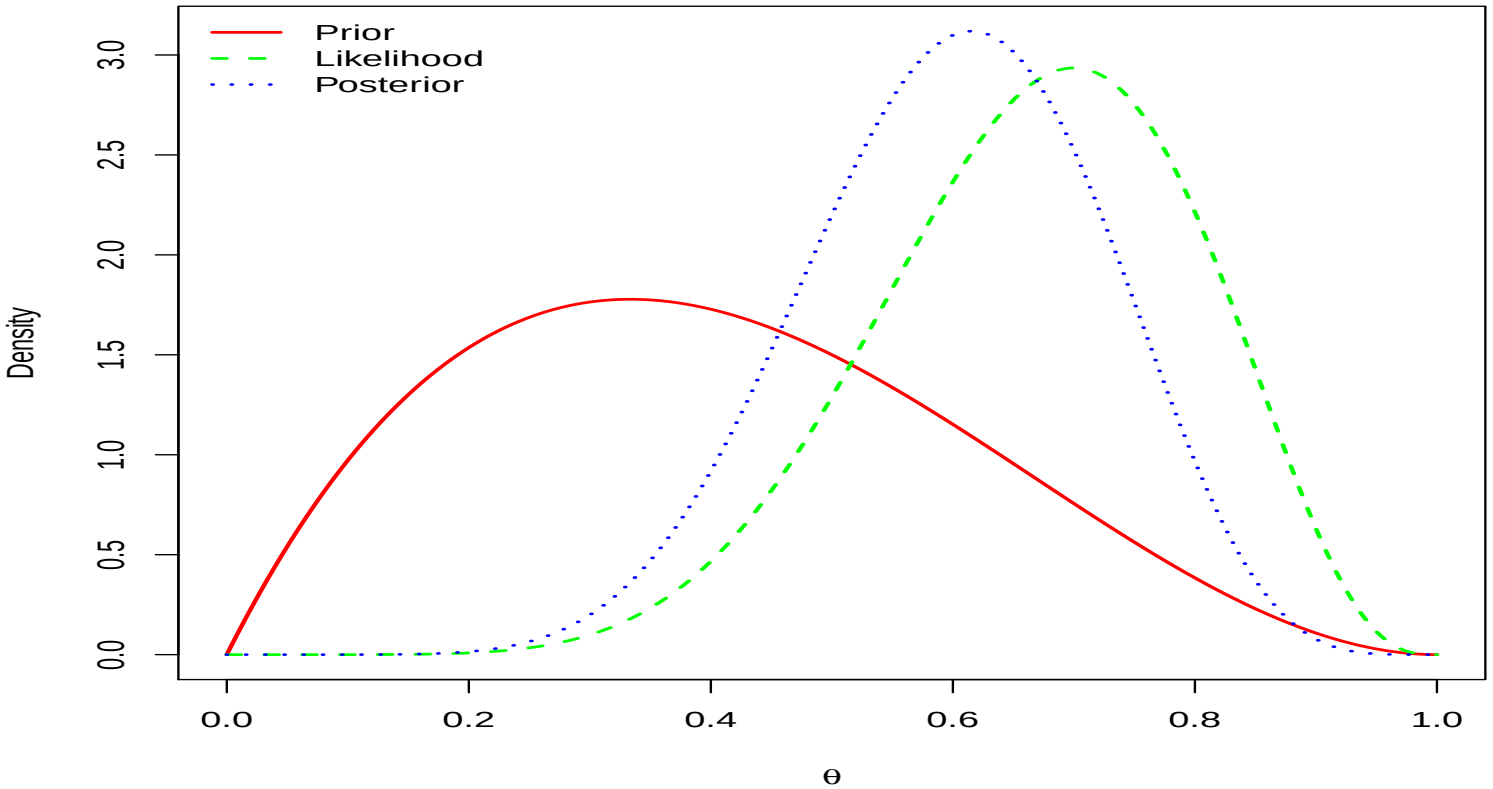


Figure: The prior is Beta(2,3) the likelihood is proportional to a Beta(7,3) and the posterior is Beta(7+2,3+3).

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

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

for a, b.

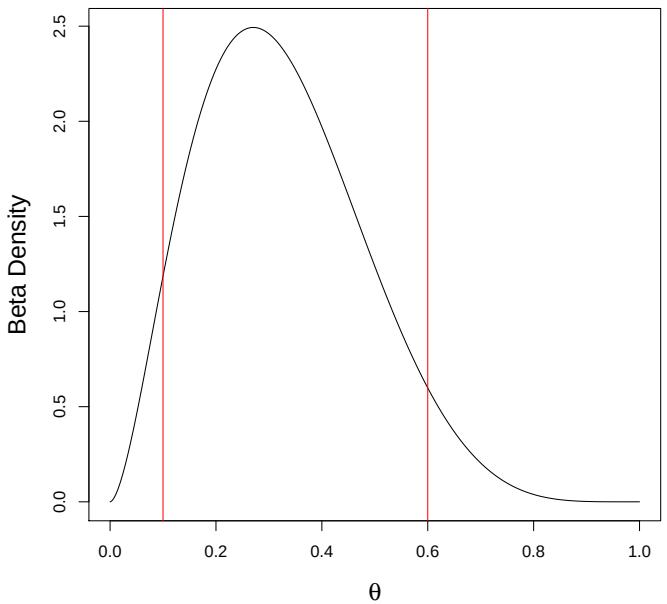


Figure: Beta(2.73,5.67) prior with 5% and 95% quantiles highlighted.

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 $Z = 0, 1, \ldots, M$ be the number of successes.
- Predictive distribution:

$$\begin{aligned}
\Pr(z|y) &= \int_0^1 p(z, \theta|y) d\theta \\
&= \int_0^1 \Pr(z|\theta, y) p(\theta|y) d\theta \\
&= \int_0^1 \underbrace{\Pr(z|\theta)}_{\text{binomial}} \times \underbrace{p(\theta|y)}_{\text{posterior}} d\theta
\end{aligned}$$

where we move between lines 2 and 3 because z is conditionally independent of y given θ.

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 $z = 0, 1, \ldots, M$.

A likelihood approach would take the predictive distribution as binomial$(M, \widehat{\theta})$ with $\widehat{\theta} = y/N$: this does not account for estimation uncertainty.

In general, we have sampling uncertainty (which we can't get away from) and estimation uncertainty.

Predictive Distribution

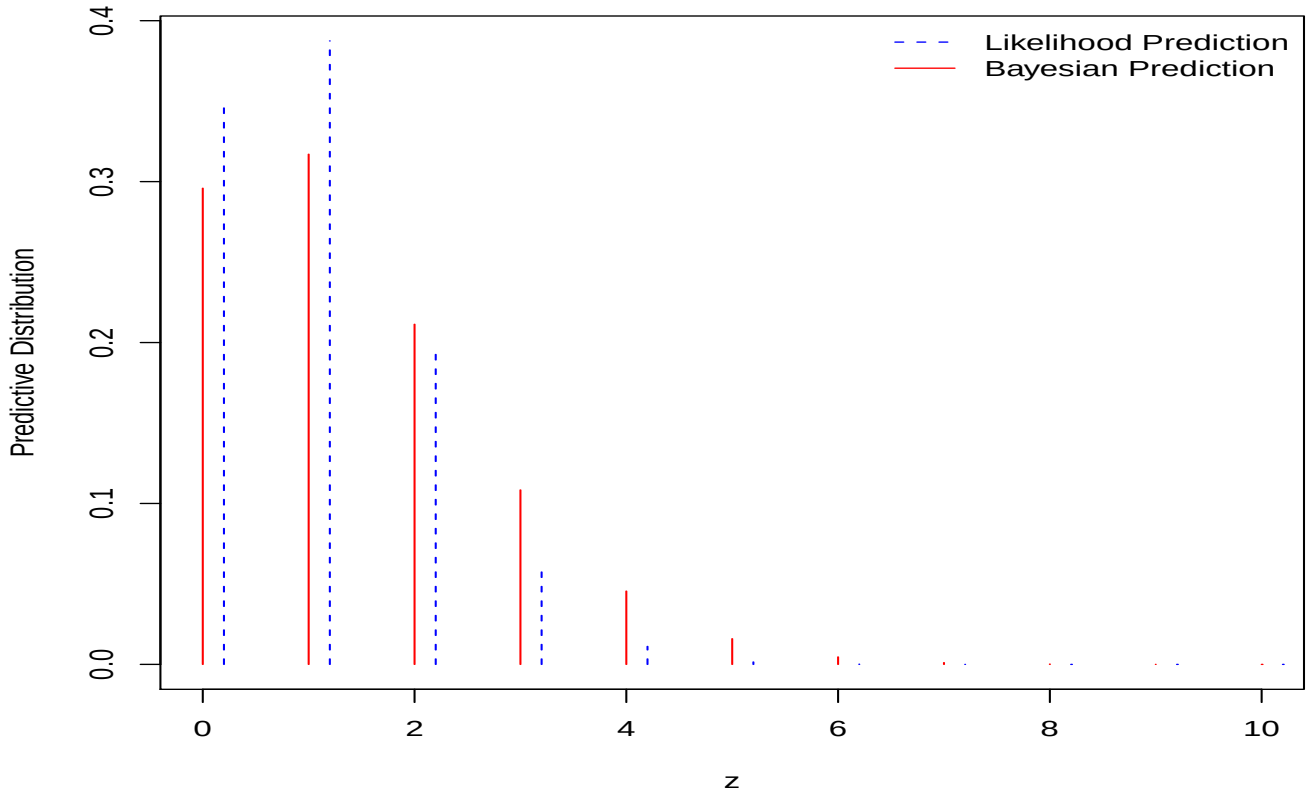


Figure: Likelihood and Bayesian predictive distribution of seeing $z = 0, 1, \ldots, M = 10$ successes, after observing $y = 2$ out of $N = 20$ successes (with $a = b = 1$).

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 $\theta^{(s)} \sim p(\theta|y)$, $s = 1, \ldots, S$, is a sample from the posterior.

This provides an estimate of the predictive distribution at the point z.

Predictive Distribution

- Alternatively, we may sample from $p(z|\theta^{(s)})$ a large number of times to reconstruct the predictive distribution.
- First sample from the posterior:

$$\theta^{(s)}|y \sim p(\theta|y).$$

- Next sample from the likelihood:

$$z^{(s)}|\theta^{(s)} \sim p(z|\theta^{(s)}),$$

for $s = 1, \ldots, S$.

- To give a sample $z^{(s)}$ from the posterior, this is illustrated to the right.

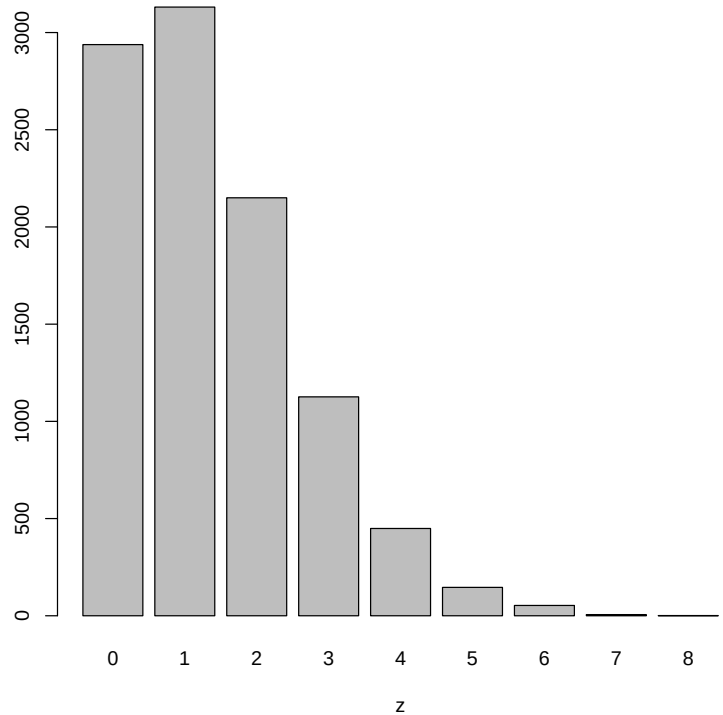


Figure: Sampling version of prediction in Figure 13, based on $S = 10,000$ samples.

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 (computed as the proportion of the samples $\theta_1^{(s)} - \theta_2^{(s)}$ that are greater than 0), so there is little evidence of a difference in allele frequencies between the NE and US samples.

Binomial Two Sample Example

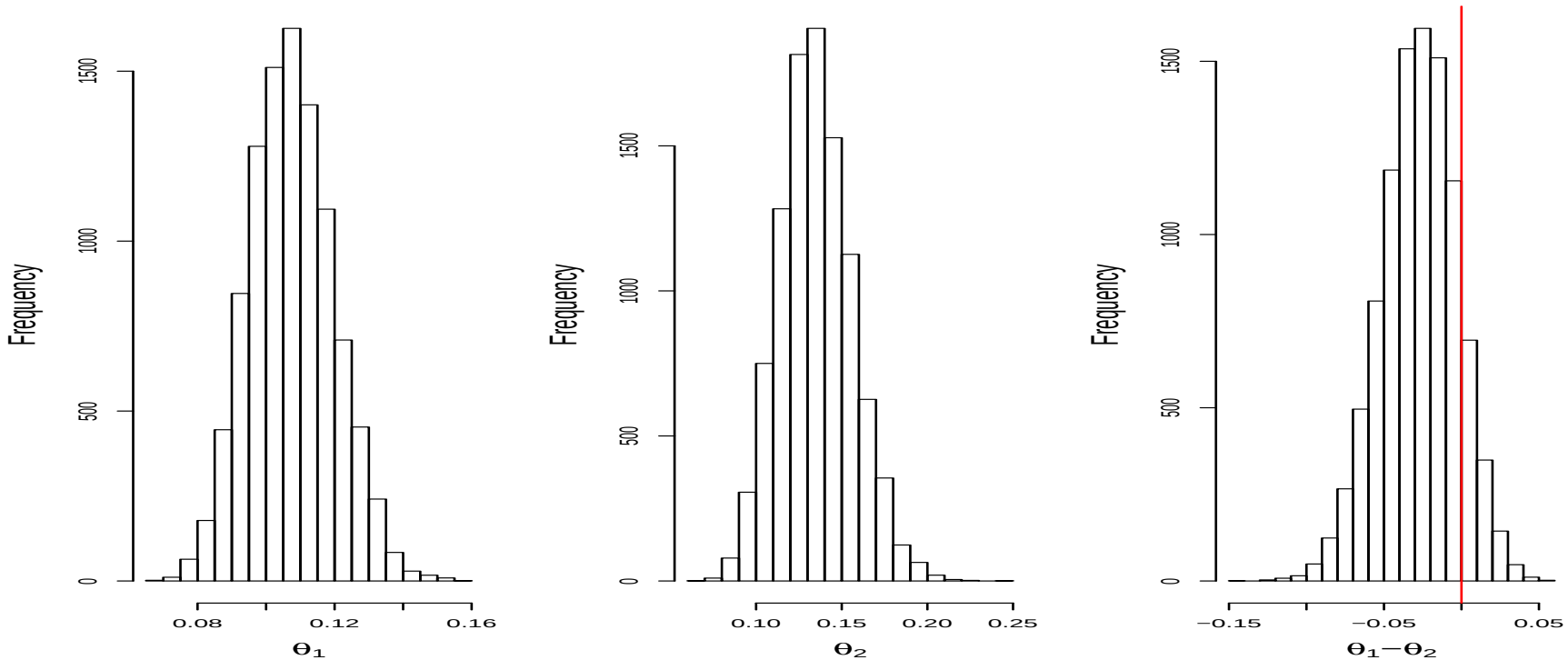


Figure: Histogram representations of $p(\theta_1|y_1)$, $p(\theta_2|y_2)$ and $p(\theta_1 - \theta_2|y_1, y_2)$. The red line in the right plot is at the reference point of zero.

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

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

  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

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

  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.

Example: Bayes Factors for Binomial Data

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 Negative Log Bayes Factor, as a Function of y

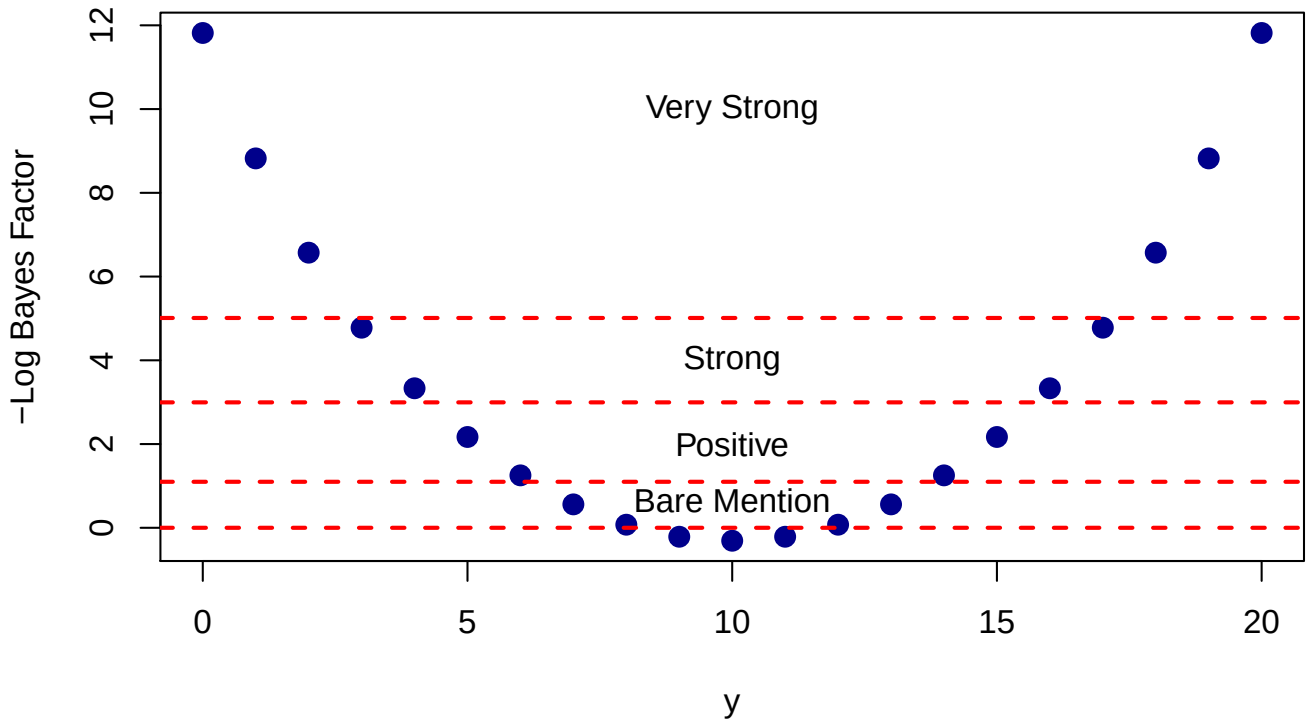


Figure: Negative Log Bayes factor as a function of $y|\theta \sim \text{Binomial}(20, \theta)$ for $y = 0, 1, \ldots, 20$ and $a = b = 1$. High values indicate evidence against the null.

Analysis of ASE Data

Three Approaches to Inference for the ASE Data

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 reciprocal of 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

- Here 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).

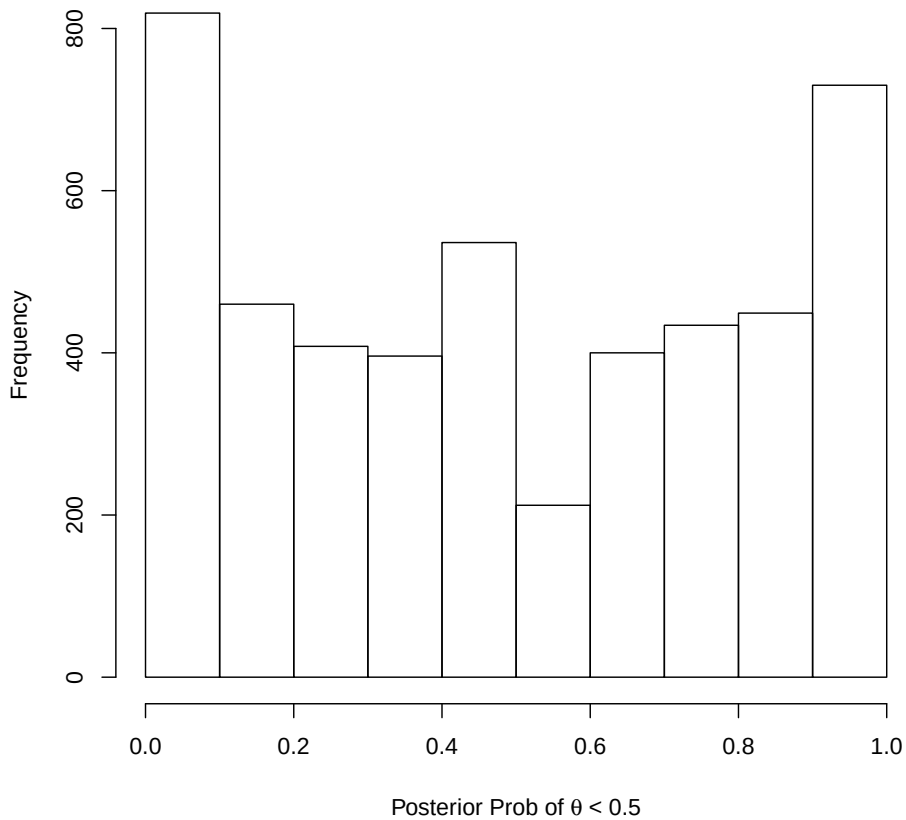


Figure: Histogram of 4,844 posterior probabilities of $\theta < 0.5$.

Bayesian Analysis of the ASE Data

- To the left 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.
- Weird lines are due to discreteness of the data.

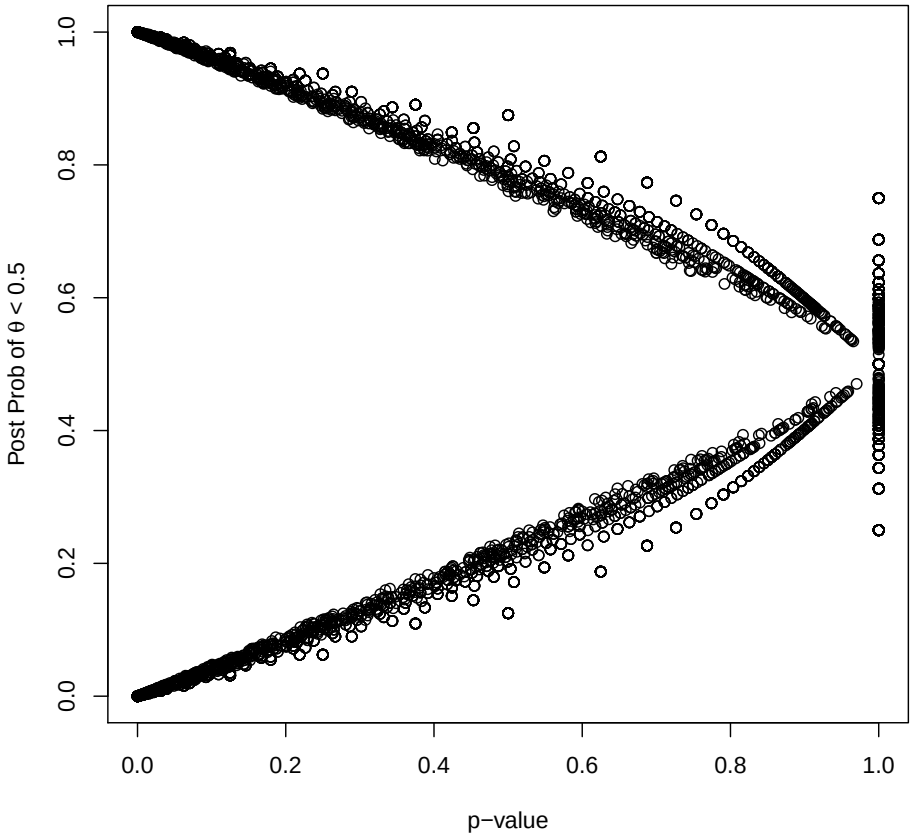


Figure: Posterior probabilities of $\theta < 0.5$ and *p*-values from exact tests.

Bayesian Analysis of the ASE Data

- Here 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.

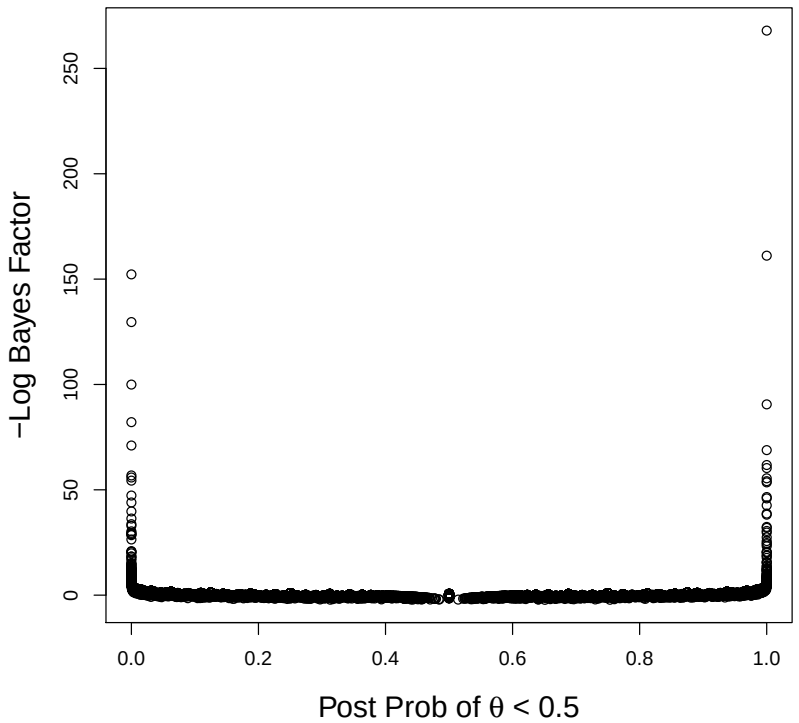


Figure: Negative Log Bayes factor versus posterior probabilities of $\theta < 0.5$.

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 reciprocal Bayes factor greater than 150.

Requiring less stringent evidence, i.e. strong and very strong (reciprocal BF greater than 20), there were 359 genes.

We later consider a formal decision theory approach to testing.

In this example, the rankings of the different approaches are similar, but the calibration, i.e., picking a threshold, is not straightforward.

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
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
```

Conclusions

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 9.

For estimation and with middle to large sample sizes, conclusions from Bayesian and non-Bayesian approaches often coincide.

For testing it's more complex, as discussed in Lecture 9.

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!
- If the sample size is large, prior choice is less crucial.

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 describe and illustrate a number of approaches including `INLA` and `Stan`.
- 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.