2016 SISG Bayesian Statistics for Genetics
R Notes: Binomial Sampling

Jon Wakefield
Departments of Statistics and Biostatistics, University of Washington

2016-07-28

Introduction

As we have seen there is an important duality between probability distributions and samples.

In many approaches to implementation, Bayesian inference is carried out via samples from the posterior distibution.

In this lecture we demonstrate this approach in the context of binomial sampling.

Samples to Summarize Beta Distributions

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

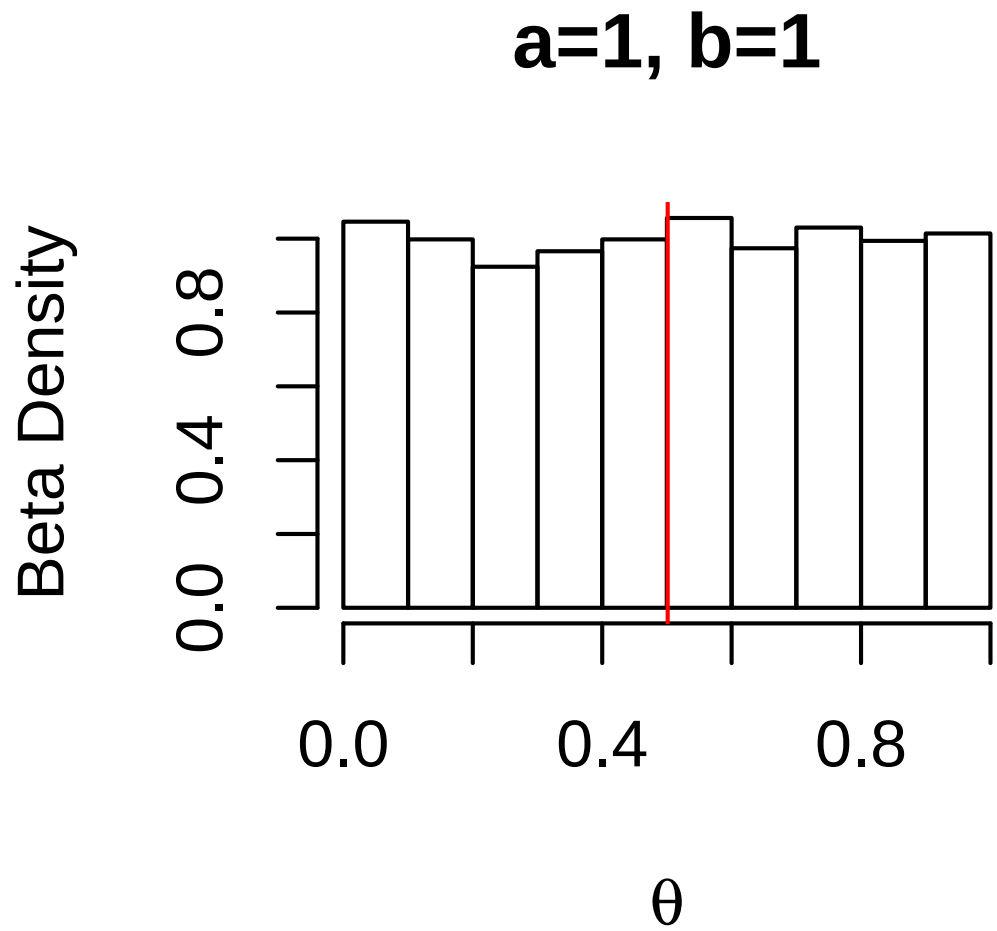
```
# First look at the theoretical quantiles of a
# uniform distribution, a beta(1,1)
qbeta(p = c(0.05, 0.1, 0.5, 0.9, 0.95), 1, 1)
## [1] 0.05 0.10 0.50 0.90 0.95
# Now find the mean and quantiles from a large
# sample from a uniform
nsim <- 5000
samp <- rbeta(nsim, 1, 1)
mean(samp)
## [1] 0.5011864
quantile(samp, p = c(0.05, 0.1, 0.5, 0.9, 0.95))
##         5%        10%        50%        90%        95%
## 0.04911498 0.09462013 0.50701693 0.90183073 0.95010043
# These differ slightly from the theoretical
# quantiles because of sampling variability
```

Samples to Summarize Beta Distributions

We now examine a histogram representation of a random variable θ with a uniform distribution, and then add a vertical line at the mean.

```
hist(samp, xlab = expression(theta), ylab = "Beta Density",
    main = "a=1, b=1", freq = F, nclass = 10)
abline(v = mean(samp), col = "red")
```

Samples to Summarize Beta Distributions



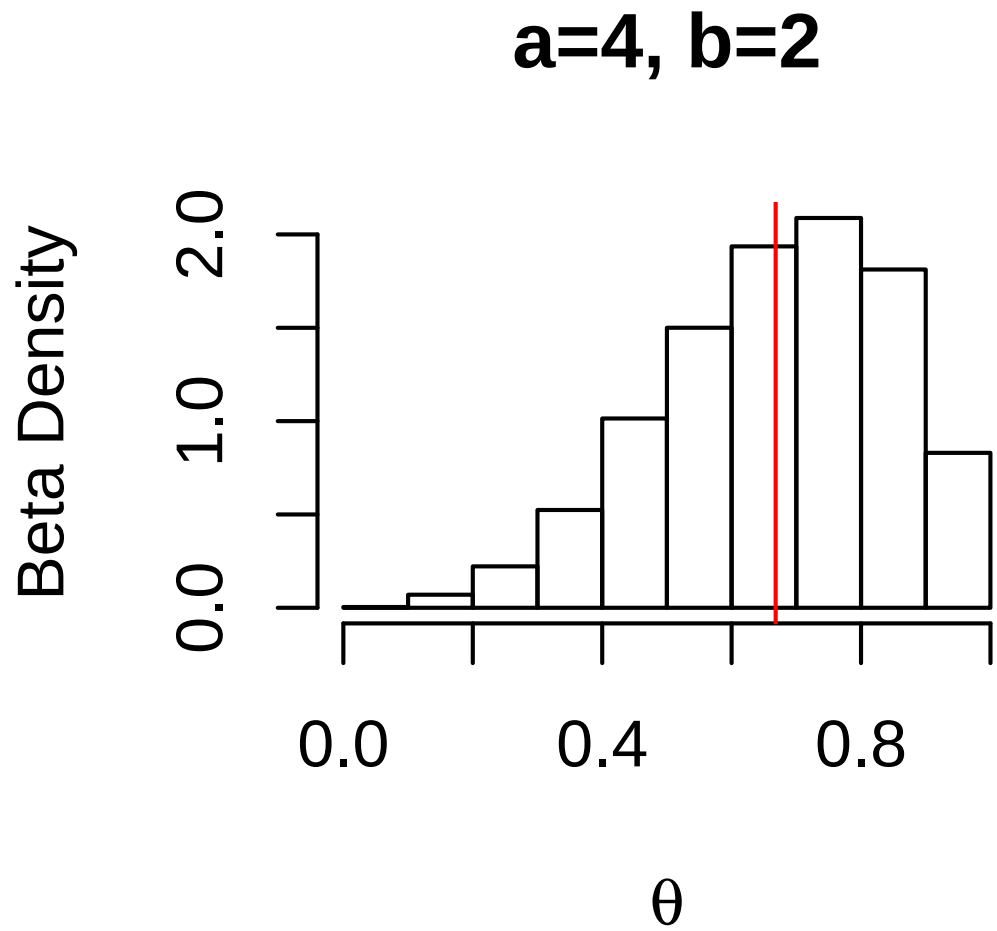
Samples to Summarize Beta Distributions

Now we examine a beta(4,2) distribution.

We first look at the theoretical quantiles (using the qbeta function), and then simulate a sample and evaluate the empirical quantiles.

```
qbeta(p = c(0.05, 0.1, 0.5, 0.9, 0.95), 4, 2)
## [1] 0.3425917 0.4161096 0.6861898 0.8877650 0.9235596
samp <- rbeta(nsim, 4, 2)
mean(samp)
## [1] 0.6637057
quantile(samp, p = c(0.05, 0.1, 0.5, 0.9, 0.95))
##        5%       10%       50%       90%       95%
## 0.3405334 0.4102922 0.6830611 0.8888653 0.9245403
hist(samp, xlab = expression(theta), ylab = "Beta Density",
    main = "a=4, b=2", freq = F, nclass = 10)
abline(v = mean(samp), col = "red")
```

Samples to Summarize Beta Distributions



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.

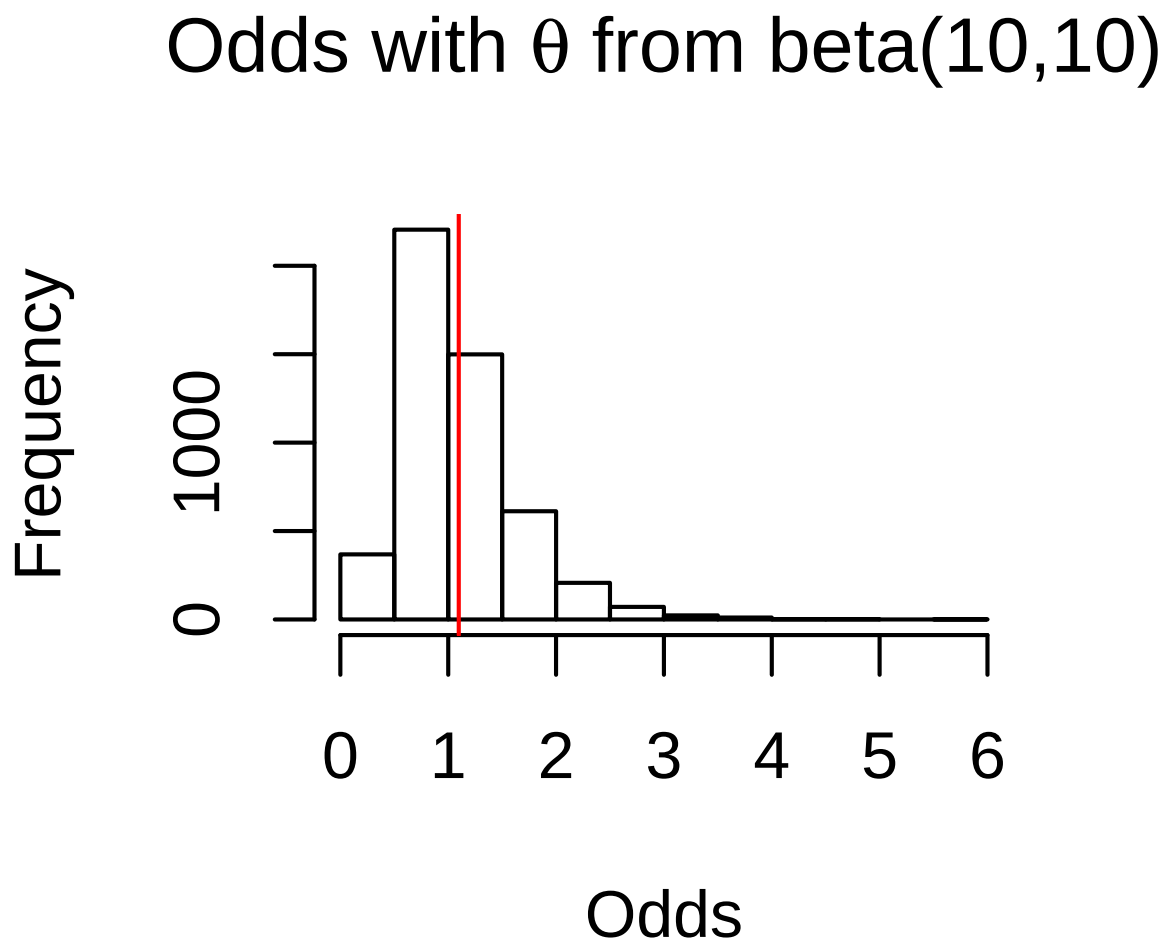
In a prior specification context, we may have clearer prior opinions about the odds, than the probability.

Below we give 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)
```

Samples for Describing Weird Parameters

```
hist(odds, xlab = "Odds", main = expression(paste("Odds with ",
    theta, " from beta(10,10)")))
abline(v = mean(odds), col = "red")
```



Are Priors Really Uniform?

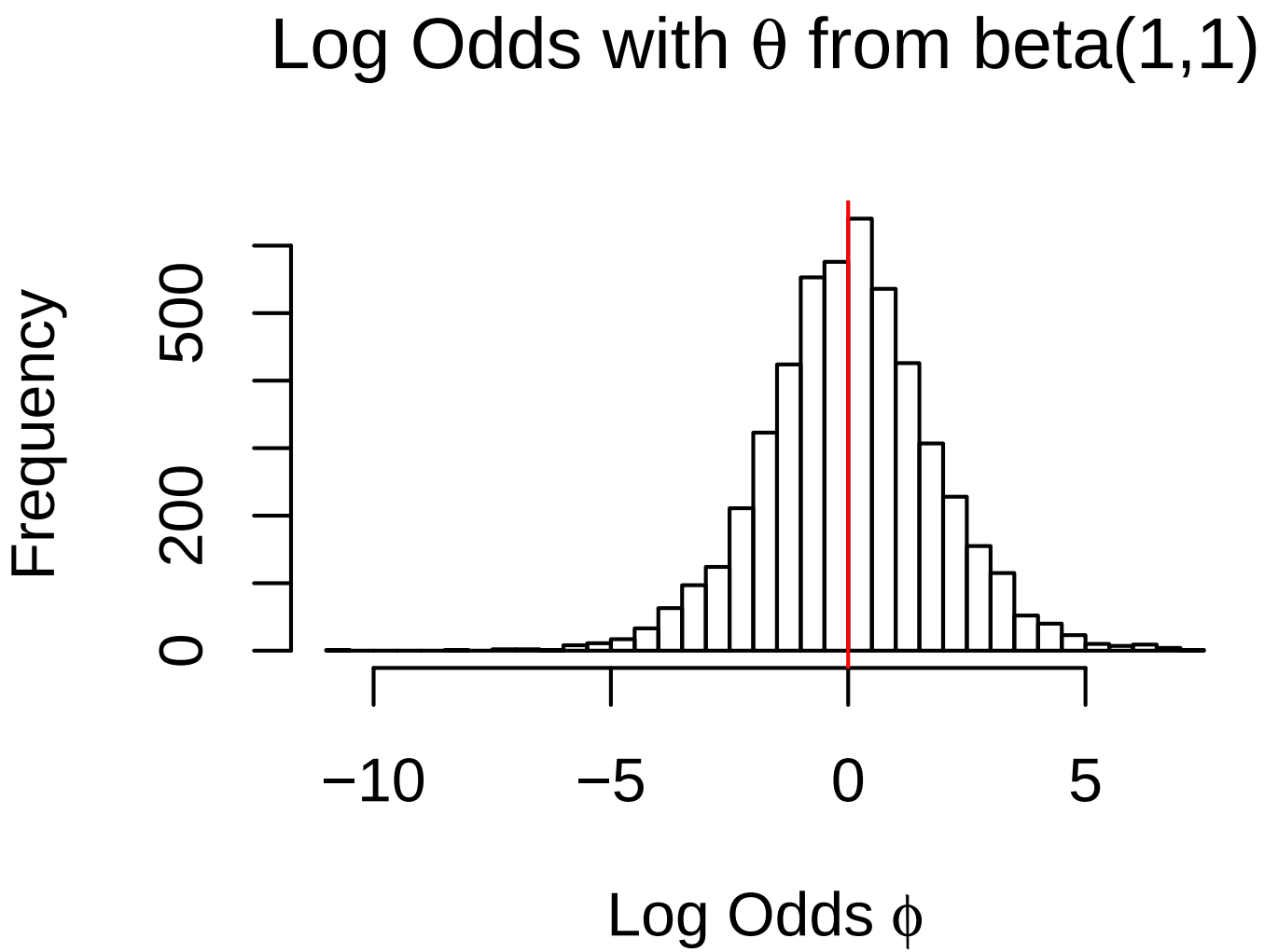
Suppose we have a uniform prior on θ, i.e. $\theta \sim \text{beta}(1, 1)$.

This prior is not uniform on the function

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

```
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")
```

Are Priors Really Uniform?



Beta Prior, Binomial Likelihood and Beta Posterior

We consider a beta prior for a proportion θ and a binomial likelihood and beta posterior that these choices lead to.

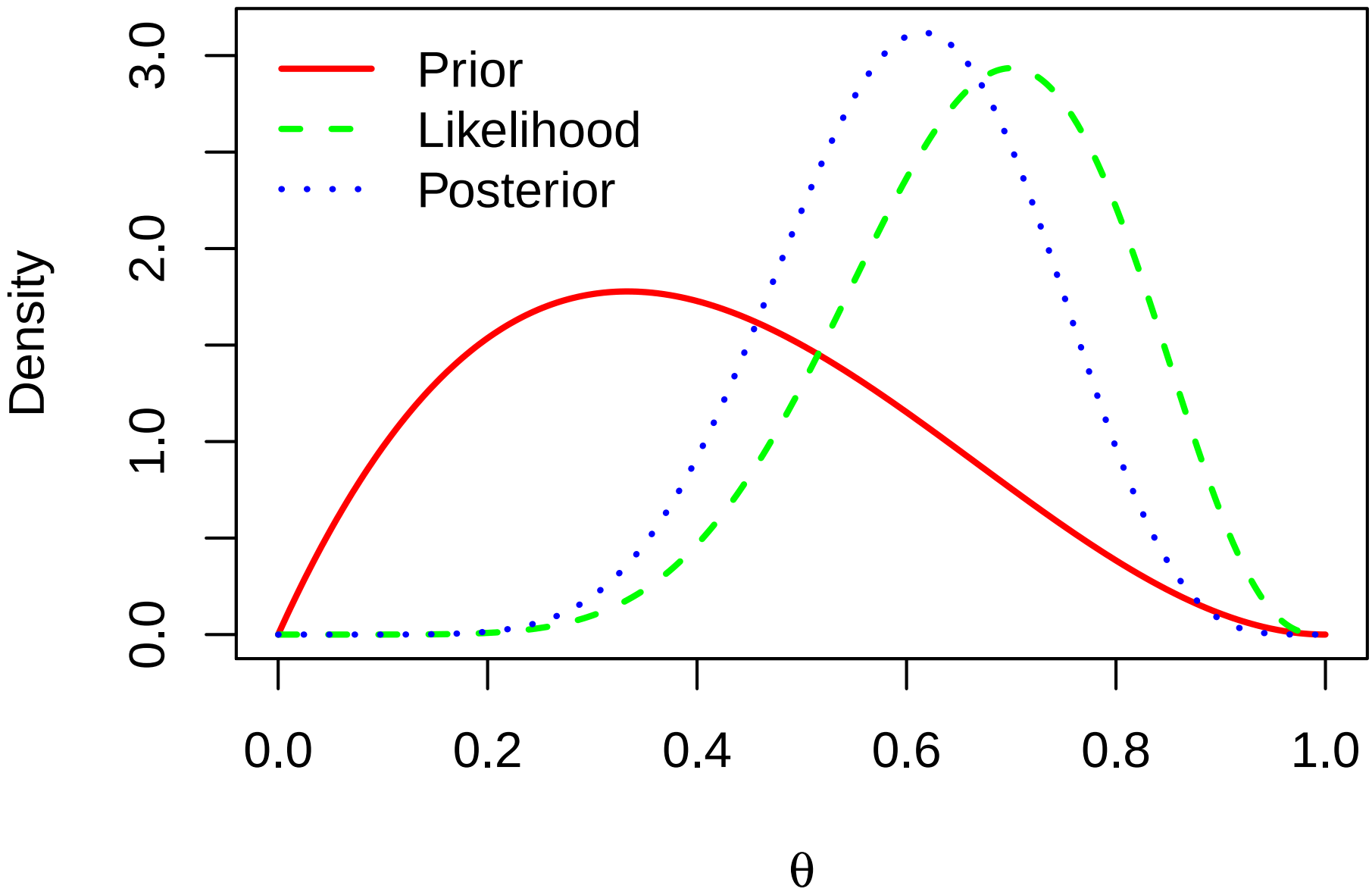
The prior is `beta(2,3)` the likelihood is proportional to a `binomial(7,3)` and the posterior is `beta(7+2,3+3)`.

```
a <- 2
b <- 3
N <- 10
y <- 7
thetaseq <- seq(0, 1, 0.001)
prior <- dbeta(thetaseq, a, b)
likelihood <- dbeta(thetaseq, y + 1, N - y + 1)
posterior <- dbeta(thetaseq, a + y, b + N - y)
```

Beta Prior, Binomial Likelihood and Beta Posterior

```
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)
```

Beta Prior, Likelihood and Posterior

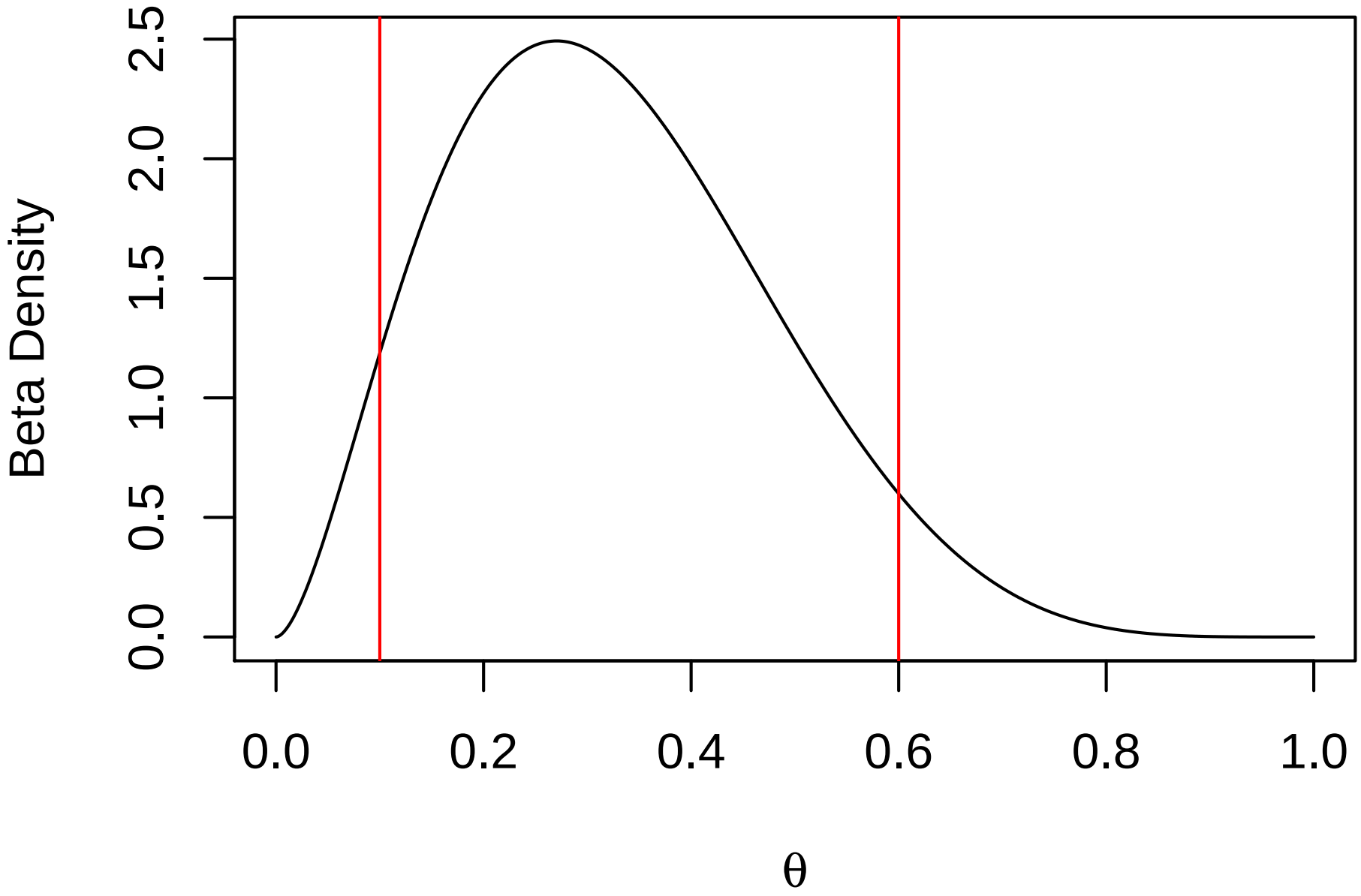


Specifying a prior distribution

The code below finds the beta distribution, i.e. the a and the b, with 5% and 95% points of 0.1 and 0.6.

```
# 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-08))
cat("a and b are ", opt$par, "\n")
## a and b are  2.730616 5.667462
probvals <- seq(0, 1, 0.001)
plot(probvals, dbeta(probvals, shape1 = opt$par[1],
    shape2 = opt$par[2]), type = "l", xlab = expression(theta),
    ylab = "Beta Density")
abline(v = q1, col = "red")
abline(v = q2, col = "red")
```

Specifying a prior distribution



Predictions from a Binomial Distribution

We now consider prediction.

Assume $y|\theta \sim \text{binomial}(N, \theta)$ and $\theta \sim \text{beta}(a, b)$.

We suppose we wish to predict the number of successes Z from M trials.

The predictive distribution is

$$\Pr(z|y) = \left(\begin{array}{c} M \\ z \end{array} \right) \frac{\Gamma(N+a+b)}{\Gamma(y+a)\Gamma(N-y+b)} \frac{\Gamma(a+y+z)\Gamma(b+N-y+M-z)}{\Gamma(a+b+N+M)}$$

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

Predictions from a Binomial Distribution

We demonstrate with a uniform prior and observing $y = 2$ successes from $N = 20$ trials, and suppose we wish to predict the number of successes we will see in 10 additional trials.

```
# User written function
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)
```

Predictions from a Binomial Distribution

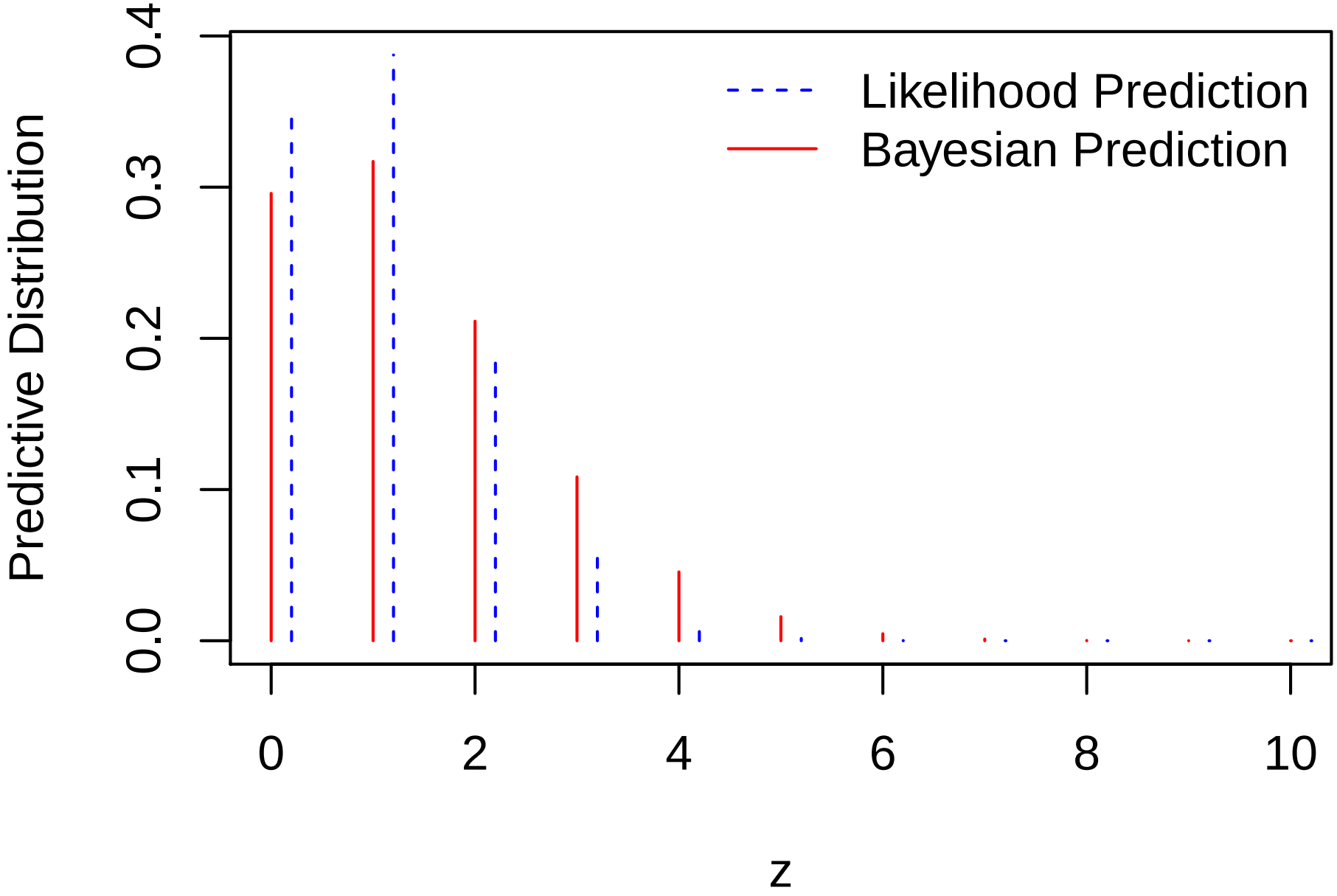
Along with the Bayesian predictive distribution, we also include a simple approach in which we assume simply take a `binomial(M,y/N)` distribution, i.e. assuming the probability is known to be the sample fraction.

```
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")
## Sum of probs =  1
```

Predictions from a Binomial Distribution

```
plot(binpred ~ z, type = "h", col = "red", ylim = c(0,
    max(likpred, binpred)), ylab = "Predictive Distribution")
points(z + 0.2, likpred, type = "h", col = "blue",
    lty = 2)
legend("topright", legend = c("Likelihood Prediction",
    "Bayesian Prediction"), lty = 2:1, col = c("blue",
    "red"), bty = "n")
```

Predictions from a Binomial Distribution



Differences in Binomial Proportions

We consider an example in which we wish to compare allele frequencies between two 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.

Differences in Binomial Proportions

The $y1$ and $y2$ data (i.e. the numbers with the allele in the two populations) were reconstructed from figures in the original paper in which only the denominators and the frequencies were given, hence the `floor` function.

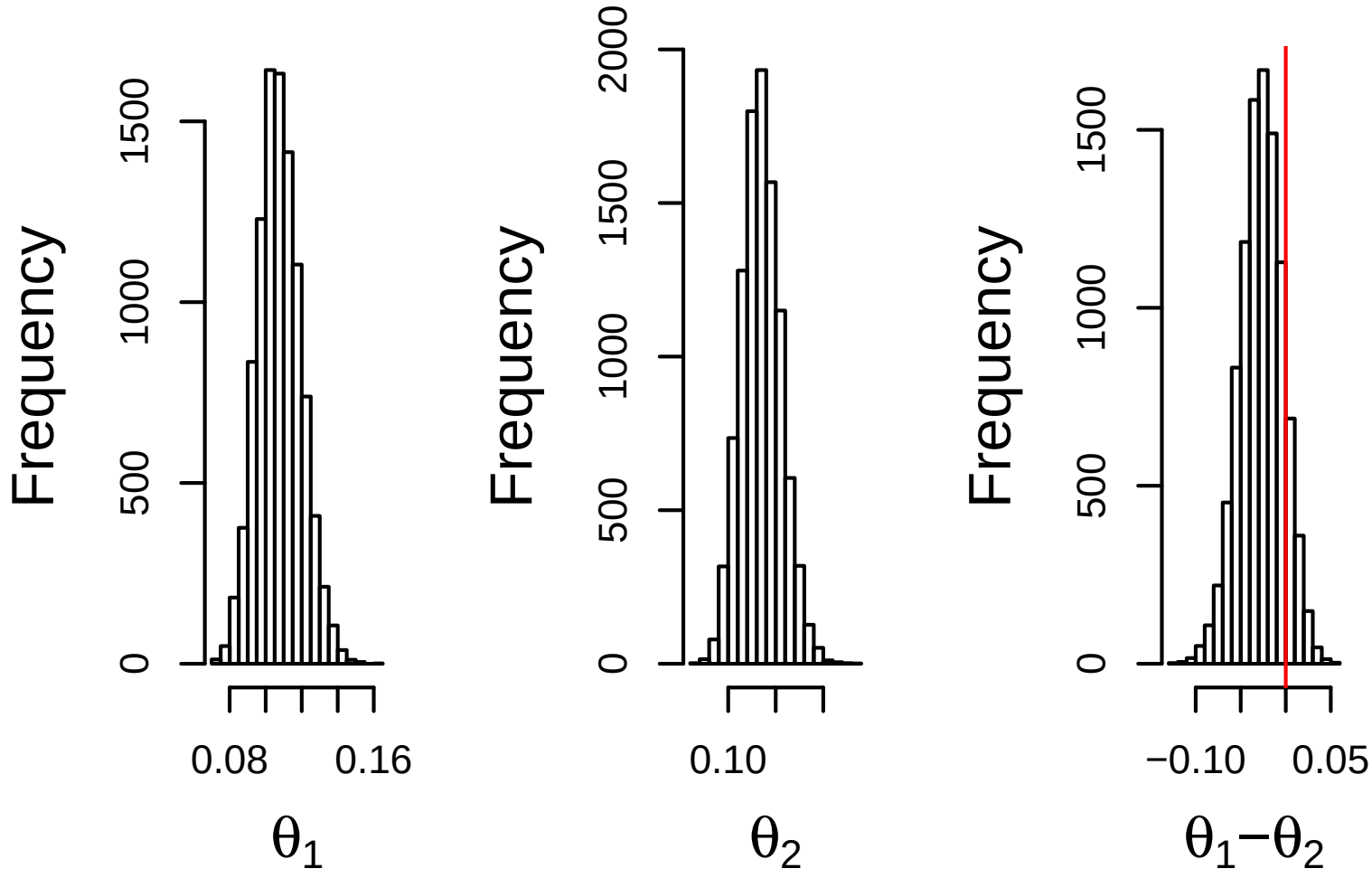
```
N1 <- 650
y1 <- floor(N1 * 0.1069)
N2 <- 265
y2 <- floor(N2 * 0.1321)
nsamp <- 10000
a <- b <- 1
post1 <- rbeta(nsamp, y1 + a, N1 - y1 + b)
post2 <- rbeta(nsamp, y2 + a, N2 - y2 + b)
```

Differences in Binomial Proportions

The key step is in constructing a sample estimate of the difference in probabilities $\theta_1 - \theta_2$.

```
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)
#
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] 0.1259
```

Differences in Binomial Proportions



“‘

ASE data

Now to analyze the ASE yeast data

First read in data

```
ASEdat <- read.table("ASEgene.txt", header = TRUE)
head(ASEdat)
##      Y    N
## 1   62  107
## 2   33   59
## 3 658 1550
## 4   14   61
## 5   57  153
## 6 218  451
dim(ASEdat)
## [1] 4844    2
ngenes <- dim(ASEdat)[1]
```

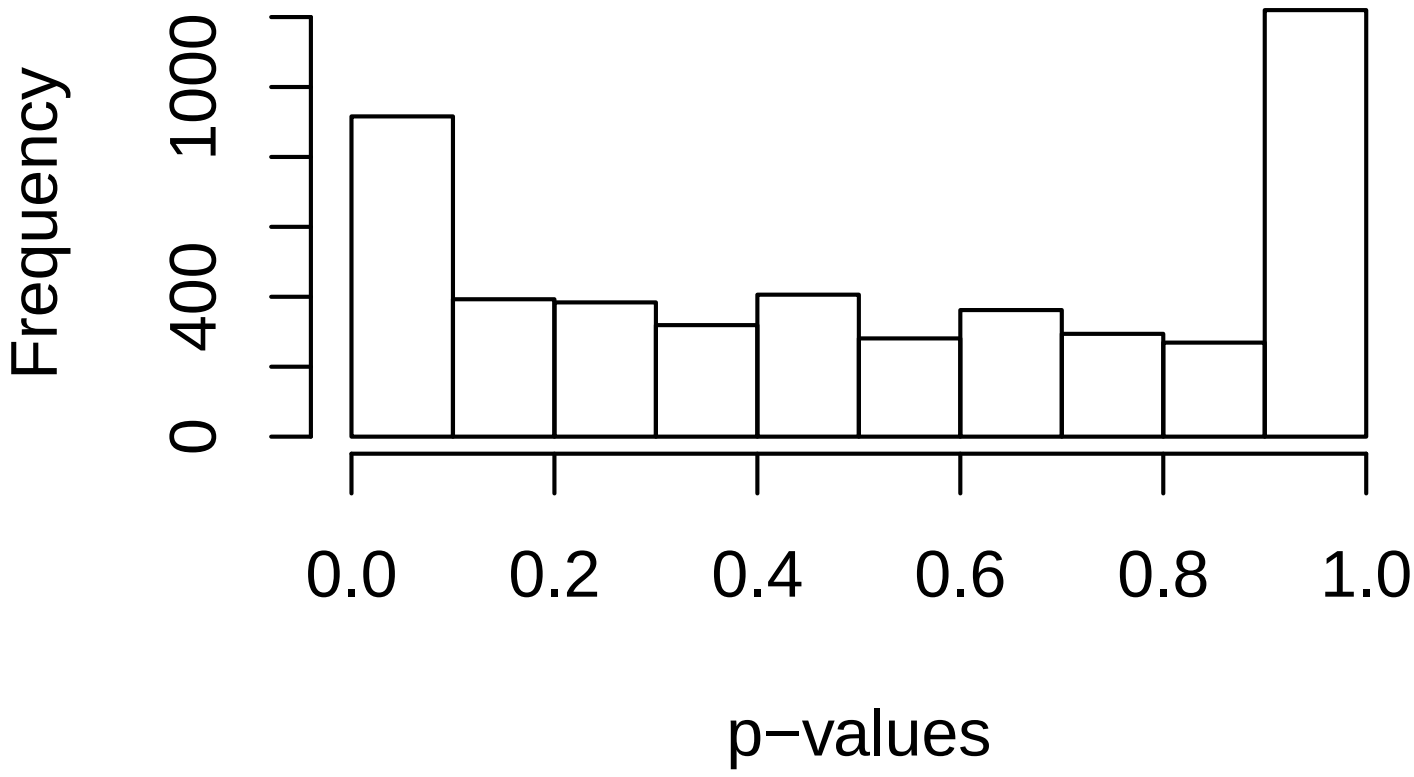
ASE data

```
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)
}
nsim <- 5000
a <- 1
b <- 1
p0 <- 0.5
```

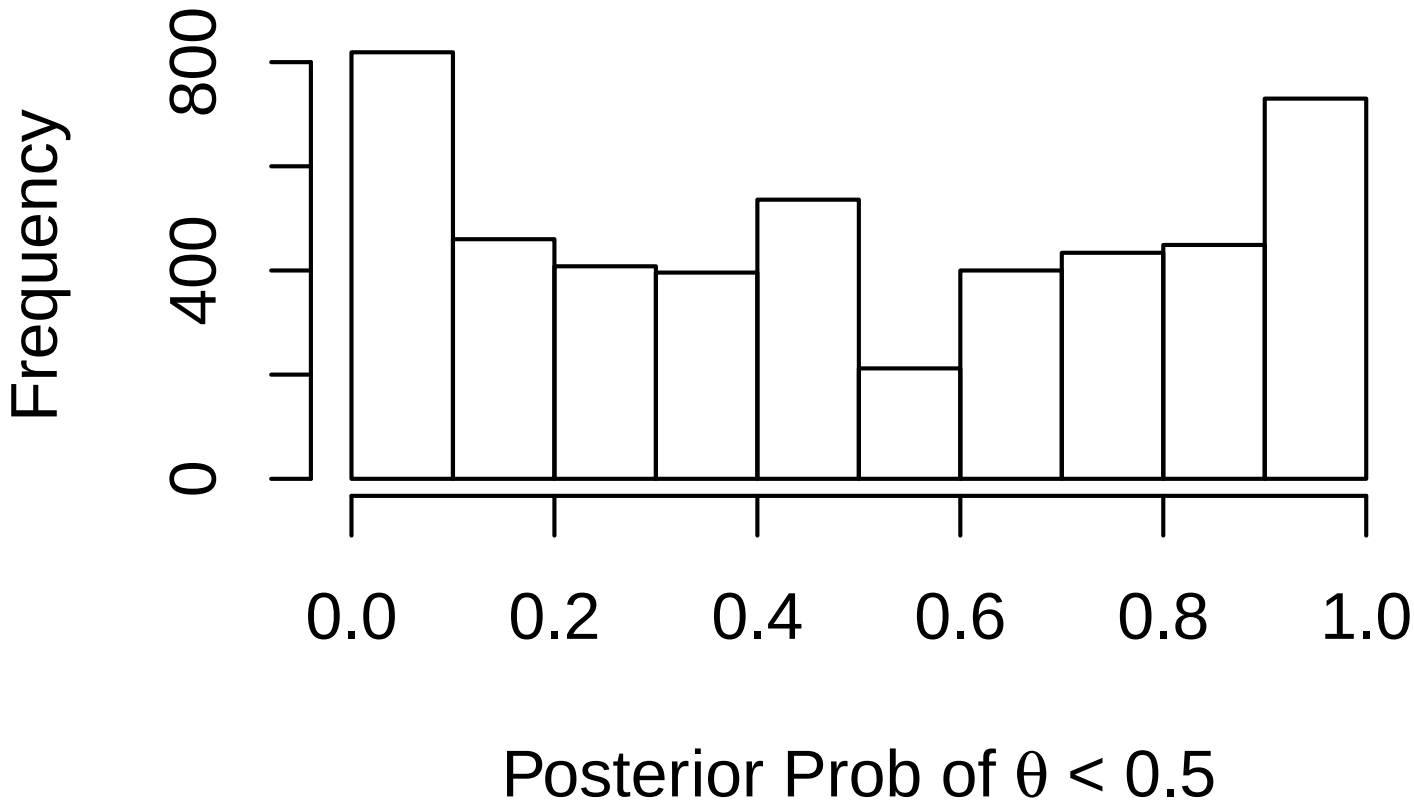
ASE data

```
postprob <- logBFr <- pvalues <- rep(0, ngenes)
pcutoff <- 0.05/ngenes  # Bonferroni
for (i in 1:ngenes) {
    Ysum <- ASEdat$Y[i]
    Nsum <- ASEdat$N[i]
    BFcall <- BFbinomial(Nsum, Ysum, a, b, p0)
    logBFr[i] <- -BFcall$logBF
    postprob[i] <- pbeta(0.5, a + Ysum, b + Nsum -
        Ysum)
    pvalues[i] <- binom.test(Ysum, Nsum, alternative = c("two.sided"))$p.value
}
cat("log BFr > log(150) = ", sum(logBFr > log(150)),
    "\n")
## log BFr > log(150) =  197
cat("log BFr > log(20) = ", sum(logBFr > log(20)),
    "\n")
## log BFr > log(20) =  359
cat("p-values >  ", pcutoff, sum(pvalues < pcutoff),
    "\n")
## p-values >   1.032205e-05 111
cat("postprobs < 0.01 and > 0.99 ", sum(postprob <
    0.01), sum(postprob > 0.99), "\n")
## postprobs < 0.01 and > 0.99  278 242
```

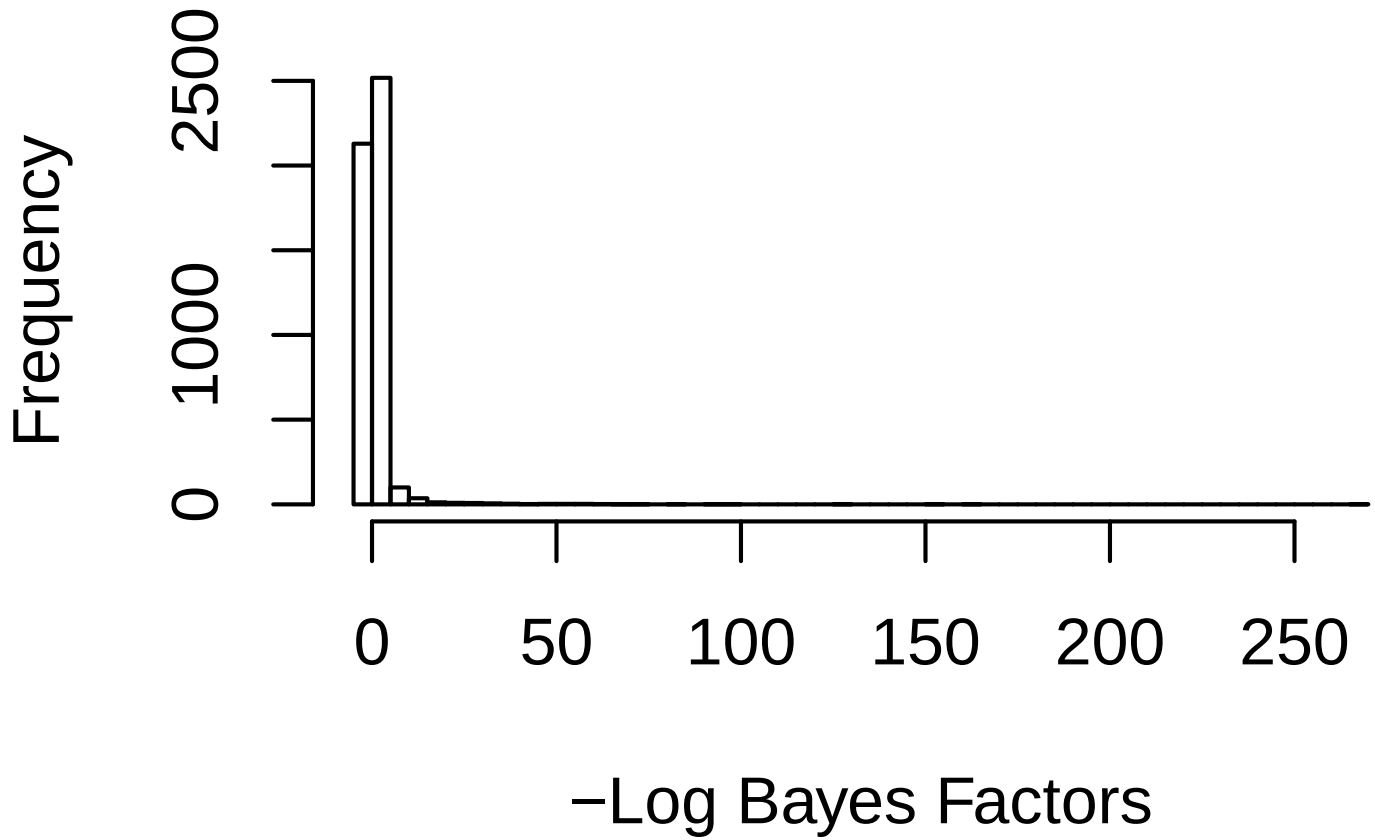
ASE



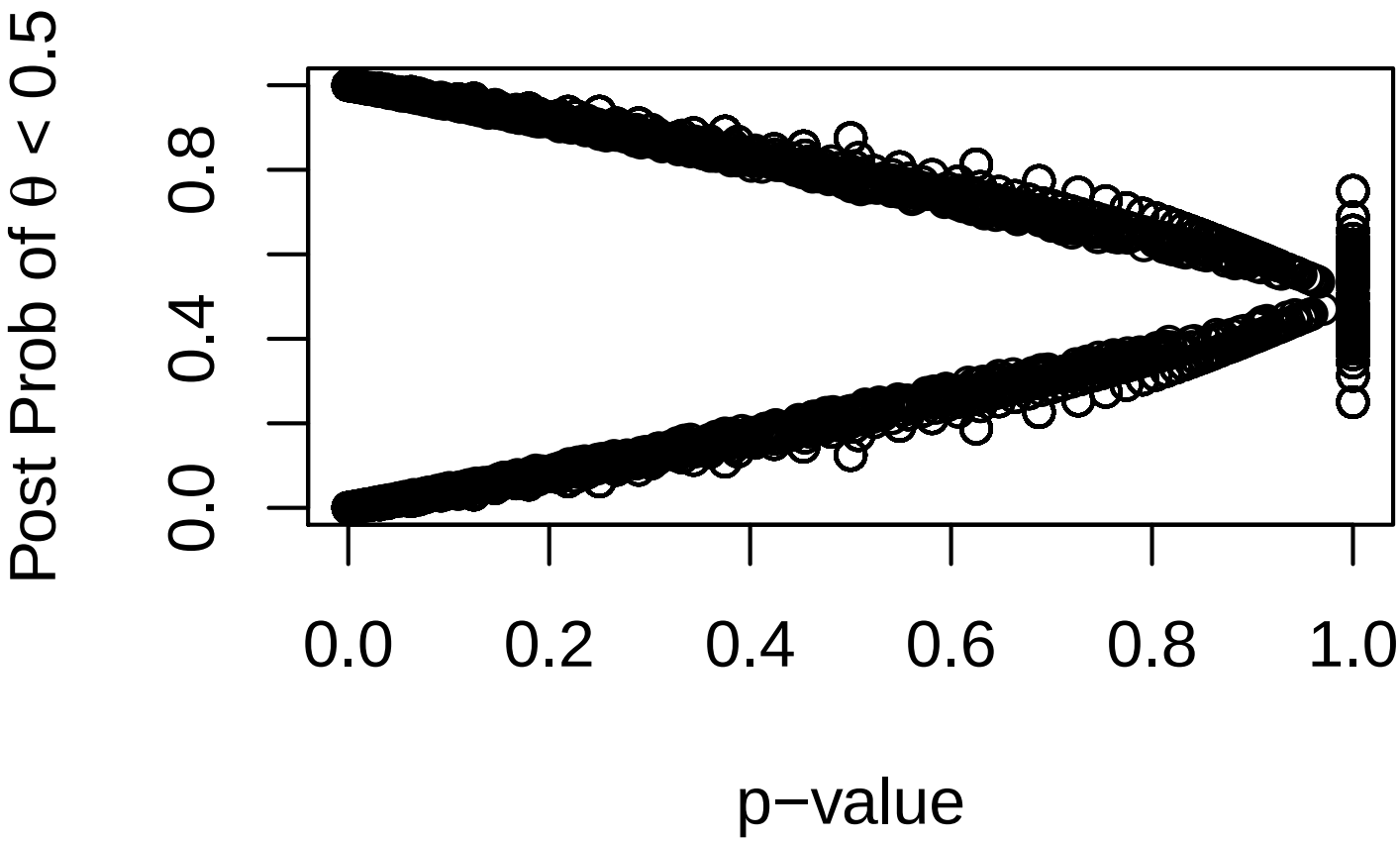
ASE



ASE



ASE



ASE