2021 SISG Bayesian Statistics for Genetics R Notes: Probability and Bayes Theorem

Jon Wakefield
Departments of Statistics and Biostatistics, University of Washington

2021-07-13

Introduction

As we have saw in lectures there is an important duality between probability distributions and samples.

In many approaches to implementation, Bayesian inference is carried out via samples obtained from the posterior distibution, either via direct sampling, or from MCMC.

In these notes we demonstrate the direct sampling 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 Beta(1,1) (uniform).

```
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 <- 20000
samp <- rbeta(nsim, 1, 1)
mean(samp)
## [1] 0.4972254
quantile(samp, p = c(0.05, 0.1, 0.5, 0.9, 0.95))
##         5%        10%        50%        90%        95%
## 0.05222843 0.10382260 0.49495851 0.89678360 0.94787768
```

These differ slightly from the theoretical quantiles because of sampling variability.

Samples to Summarize Beta Distributions

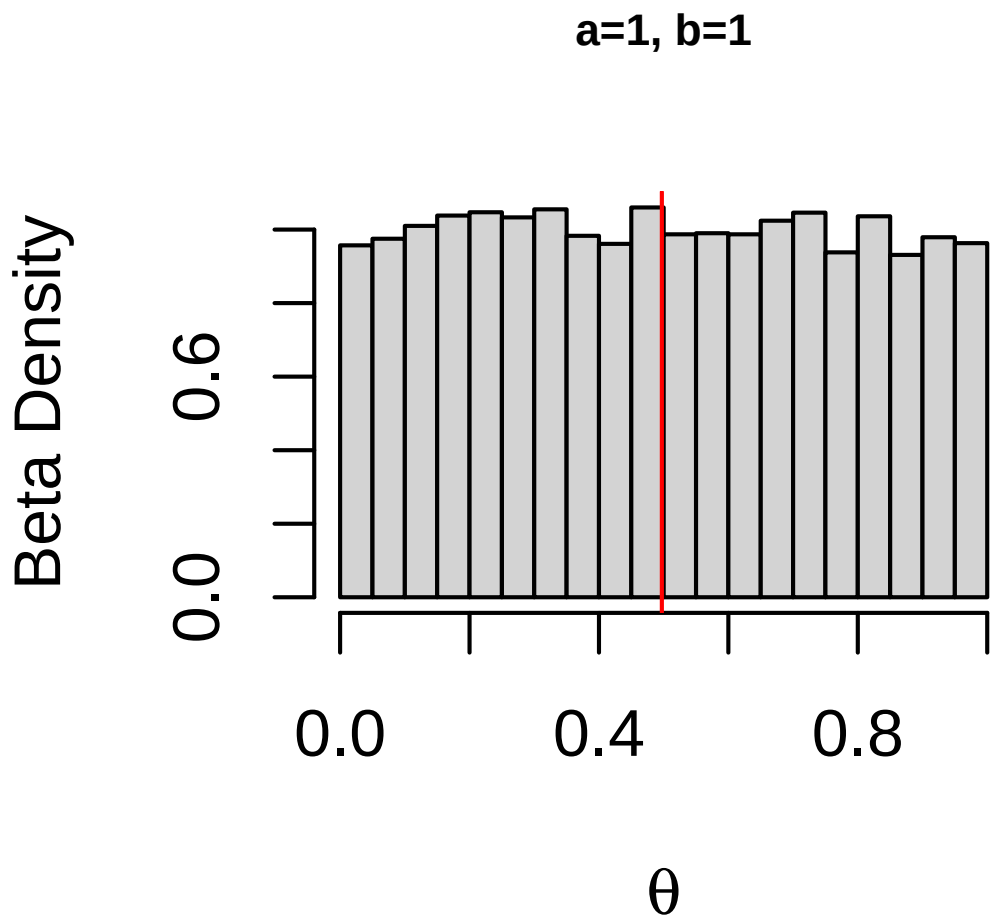
We now examine a histogram representation of a uniform random variable, i.e.,

$$\theta \sim \text{Beta}(1, 1),$$

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 = 20)
abline(v = mean(samp), col = "red")
```

Samples to Summarize Beta Distributions



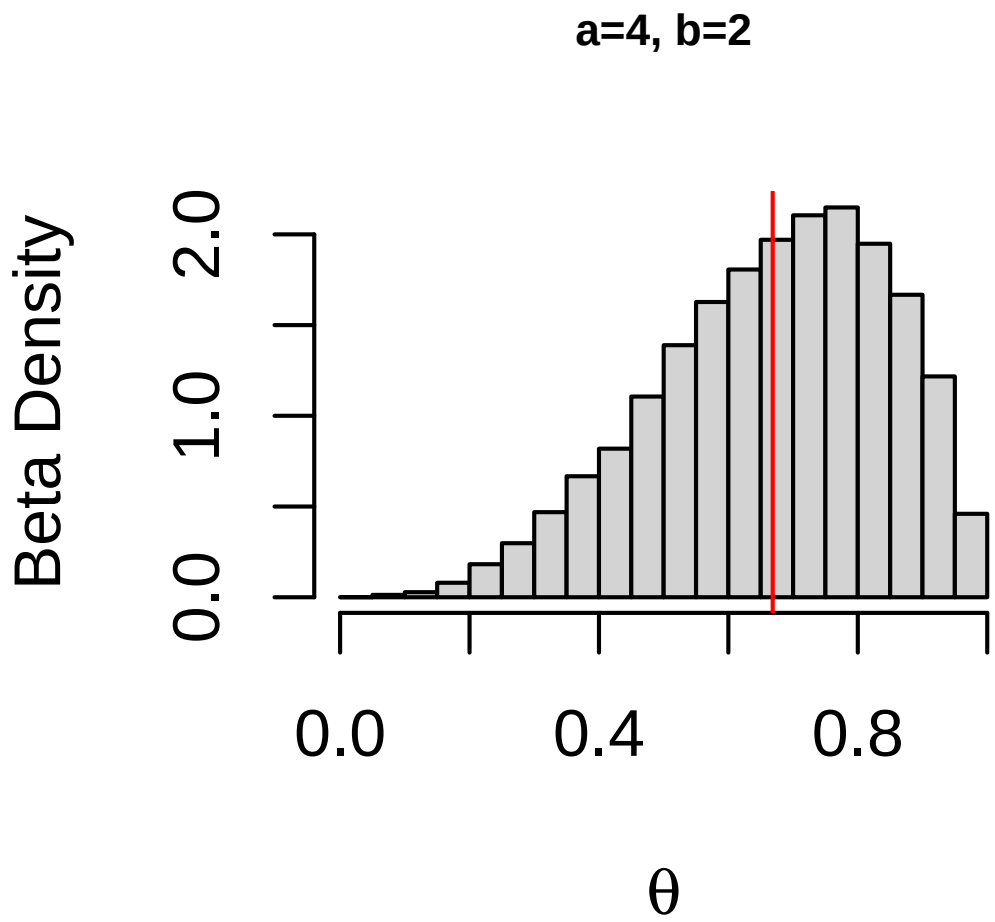
Samples to Summarize Beta Distributions

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] 0.6671744
quantile(samp, p = c(0.05, 0.1, 0.5, 0.9, 0.95))
##        5%       10%       50%       90%       95%
## 0.3438353 0.4176879 0.6852210 0.8898567 0.9253087
hist(samp, xlab = expression(theta), ylab = "Beta Density",
    main = "a=4, b=2", freq = F, nclass = 20, cex.main = 0.7)
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.

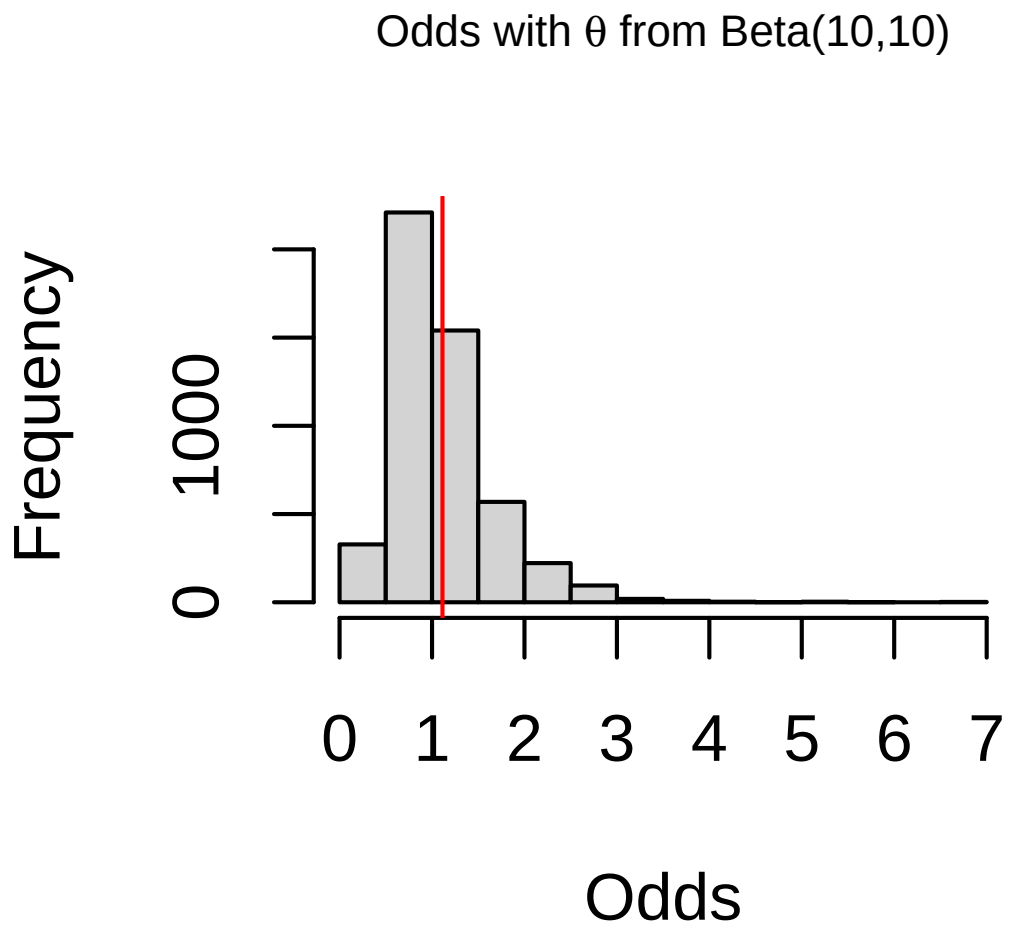
In a prior specification context, we may have clearer prior opinions about the odds, than the probability.

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)")), cex.main = 0.7)
abline(v = mean(odds), col = "red")
```



Are Priors Really Uniform?

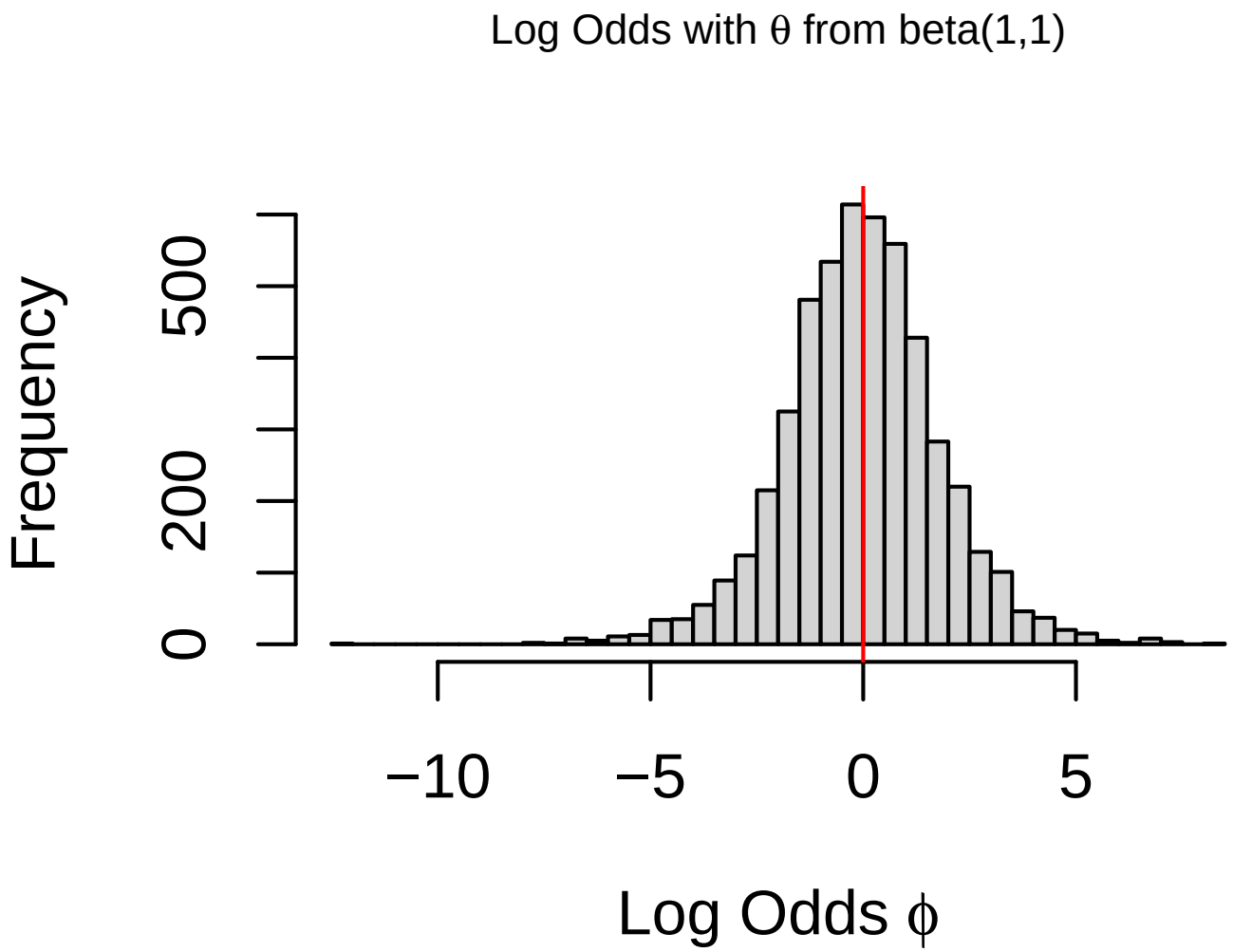
Suppose we have a uniform prior on θ, i.e. $\theta \sim \text{Beta}(1, 1)$.

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)")), cex.main = 0.7)
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.

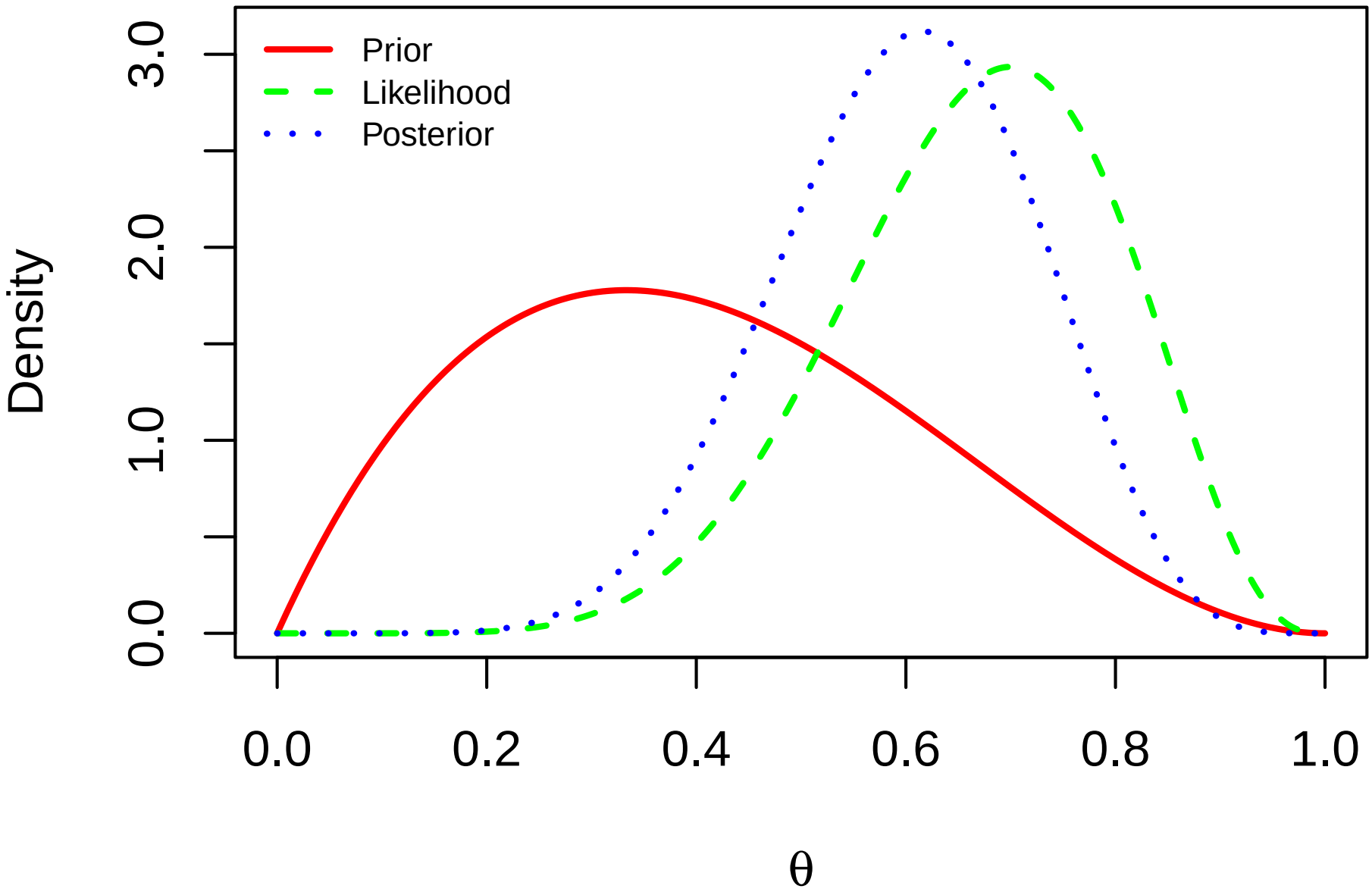
The prior is $\theta \sim \text{Beta}(2, 3)$ the likelihood is $y = 7|\theta \sim \text{Binomial}(N = 10, \theta)$, and the posterior is Beta(2+7,3+3).

```
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, cex = 0.7)
```

Beta Prior, Binomial Likelihood, Beta Posterior



Seroprevalence example

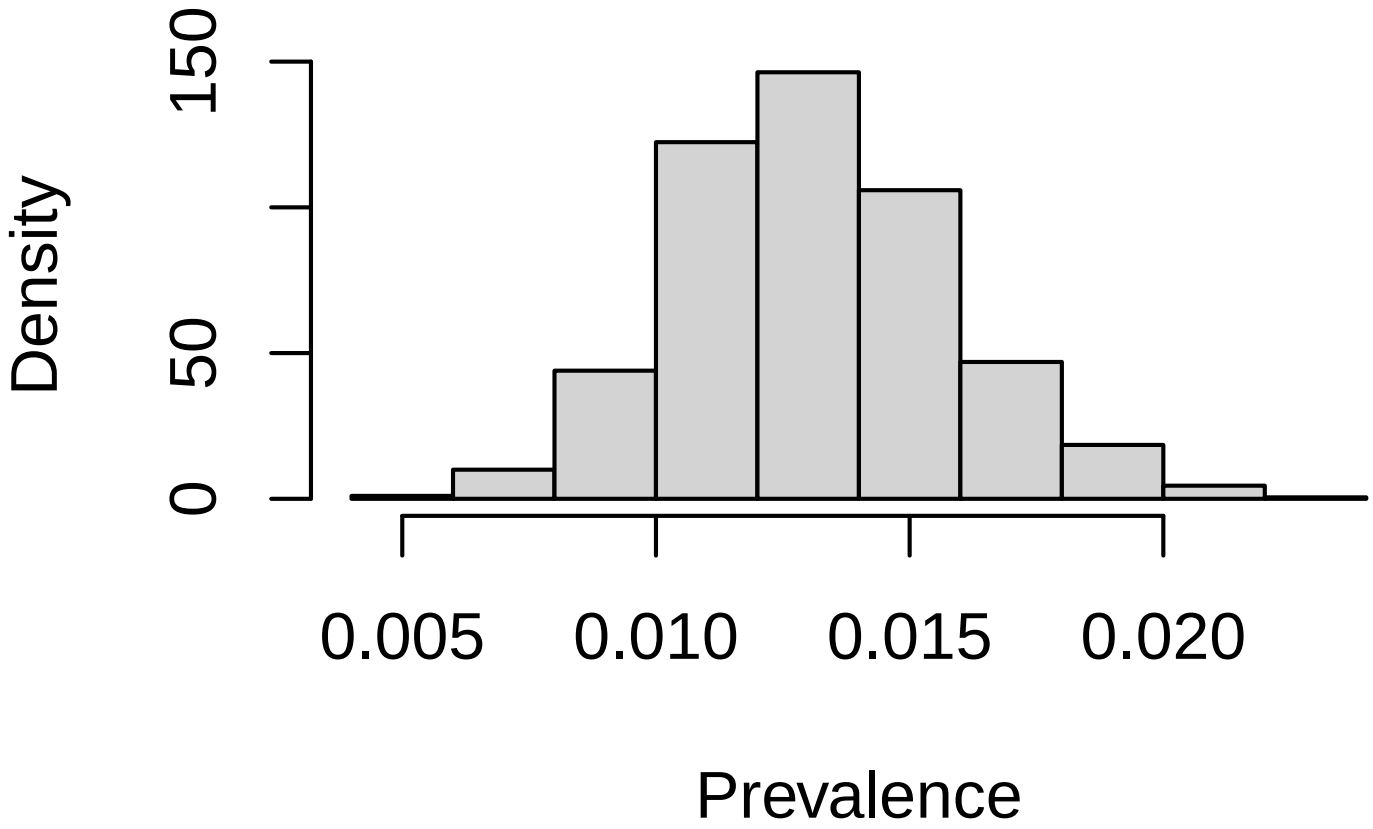
```
y <- 50
n <- 3300
delta <- 0.8
gamma <- 0.995
A <- delta + gamma - 1
B <- 1 - gamma
MLE <- (y - n * B)/(n * A)
loglik <- function(n, y, prev, delta, gamma) {
    A <- delta + gamma - 1
    B <- 1 - gamma
    p <- prev * A + B
    loglik <- y * log(p) + (n - y) * log(1 - p)
    loglik
}
maxl <- loglik(n, y, MLE, delta, gamma)
nsim <- 1000
success <- 0
a <- b <- 1
```

Seroprevalence example: the rejection algorithm

```
post <- NULL
while (success < nsim + 1) {
    U <- runif(1)
    theta <- rbeta(1, a, b)
    test <- loglik(n, y, theta, delta, gamma)
    if (log(U) < test - maxl) {
        success <- success + 1
        post[success] <- theta
    }
}
mean(post)
## [1] 0.01313352
quantile(post, p = c(0.05, 0.5, 0.95))
##          5%         50%         95%
## 0.008807793 0.012979686 0.017677144
```

Seroprevalence example

```
hist(post, xlab = "Prevalence", main = "", freq = F)
```



Exercises

1. Suppose we observe data with $N = 20, y = 20$ and we assume a binomial likelihood with probability θ.
   - What are the MLE and standard error of the MLE?
   - Plot the posterior distribution under a Beta(2,2) prior
   - Simulate from the posterior and draw a histogram of the posterior samples.
   - What is the posterior median, and give a 90% credible interval for θ: evaluate these quantities in two ways, one via the `qbeta` function and the other via sampling.
   - What is the MLE for the odds $\theta/(1-\theta)$?
   - What is the posterior probability that the odds are greater than 100?
2. Redo the seroprevalence example with a Beta(1,9) prior on the prevalence.