Homework 4 Solution

Math 315, Fall 2019

4.

(a)

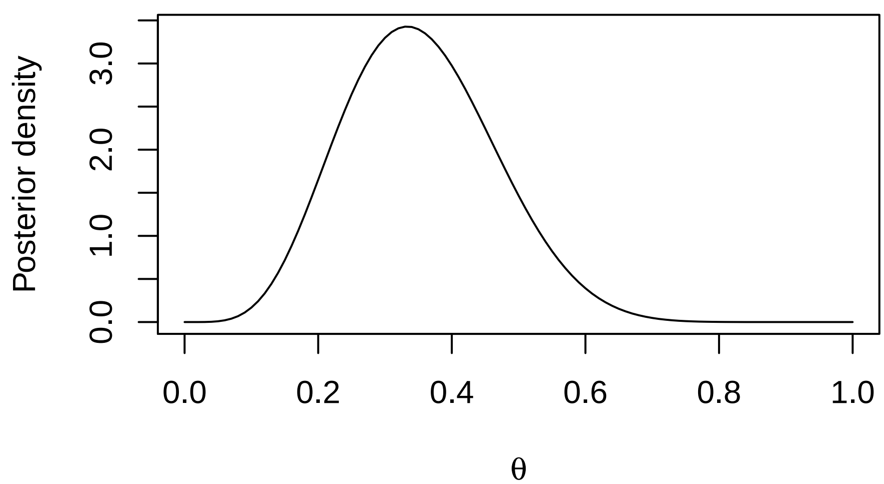
The posterior is proportional to

$$
\begin{aligned}
p(\theta|Y) \propto f(Y|\theta)\pi(\theta) &\propto \left[\theta^m (1-\theta)^Y\right] \cdot \left[\theta^{a-1}(1-\theta)^{b-1}\right] \\
&= \theta^{m+a-1}(1-\theta)^{Y+b-1}
\end{aligned}
$$

which is the kernel of the Beta$(m+a, Y+b)$ distribution.

(b)

```
m <- 5
Y <- 10
a <- 1
b <- 1
curve(dbeta(x, m + a, Y + b), xlab = expression(theta), ylab = "Posterior density")
```




```
qbeta(c(.025, .975), m + a, Y + b)
```

```
## [1] 0.1519837 0.5866206
```

Above I use the "equal-tail" CI, but you could also find the HPDI via simulation:

```
library(HDInterval)
thetas <- rbeta(1e5, m + a, Y + b)
hdi(thetas, credMass = 0.95)
```

```
##     lower     upper
## 0.1406629 0.5740755
## attr(,"credMass")
```

```
## [1] 0.95
```

5.

We need to find a and b such that two conditions hold:

$$Var(\lambda) = a/b^2 = 100 \tag{1}$$
$$F(\lambda_0) = \int_0^{75} \pi(\lambda) d\lambda = \int_{75}^{\infty} \pi(\lambda) d\lambda = 0.5 \tag{2}$$

We can solve this problem in R using the **BB** library, just like our class examples:

```
library(BB)

# define function for the mean of Gamma(a, b) dsn
gamma.var <- function(x) x[1] / x[2]^2

# set up a system of equations to solve for elicited values
gamma.solver <- function(x){
  obj1 <- gamma.var(x) - 100             # var - desired value
  obj2 <- qgamma(0.5, x[1], x[2]) - 75  # quantile - desired value
  return(c(obj1, obj2))
}

# BBsolve(<intial guess>, <system of eqns as function>)
gamma.params <- BBsolve(c(20, .4), gamma.solver, quiet = TRUE)
gamma.params$par
```

```
## [1] 56.9165564  0.7544339
```

Thus, we set our prior to be Gamma$(56.9165564, 0.7544339)$. (Note: I would generally round these, this seems overly precise.)

Now that we have an answer, it's wise to check that the conditions are actually met to avoid a silly error.

First, evaluate the prior CDF at $\lambda_0 = 75$ to see if it is approximately 0.5:

```
pgamma(75, gamma.params$par[1], gamma.params$par[2])
```

```
## [1] 0.4999454
```

An alternative approach: Reread the two conditions and notice that

- the first condition implies that $a = 100b^2$; thus, we only need to search over the possible values of b.
- the second condition can be restated: the median of the prior should be $\lambda_0 = 75$.

We can find a and b through a grid search for b.

```
# Set up the grid for b
b_seq <- seq(0.1, 1, by = 0.0001)

# Find the median for each b
median_seq <- qgamma(0.5, 100 * b_seq^2, b_seq)

# Which b is closest to 75?
```

```r
b <- b_seq[which.min(abs(median_seq - 75))]
b
```

```
## [1] 0.7544
```

```r
# What's a?
a <- 100 * b^2
a
```

```
## [1] 56.91194
```

8.

(a)

Here we have the model

$$Y_1, \ldots, Y_n | \lambda \overset{iid}{\sim} \text{Poisson}(\lambda) \tag{3}$$
$$\lambda \sim \text{Gamma}(a, b) \tag{4}$$

Thus, the posterior is proportional to

$$p(\lambda | y_1, \ldots, y_n) \propto f(y_1, \ldots, y_n | \lambda) \pi(\lambda) \tag{5}$$
$$= \prod \frac{e^{-\lambda} \lambda^{y_i}}{y_i} \cdot \frac{b^a}{\Gamma(a)} e^{-b\lambda} \lambda^a \tag{6}$$
$$\propto e^{-\lambda(n+b)} \lambda^{\Sigma y_i + a} \tag{7}$$

which resembles the Gamma($\sum y_i + a,\ n + b$). (Thus, gamma is a conjugate prior, as expected.)

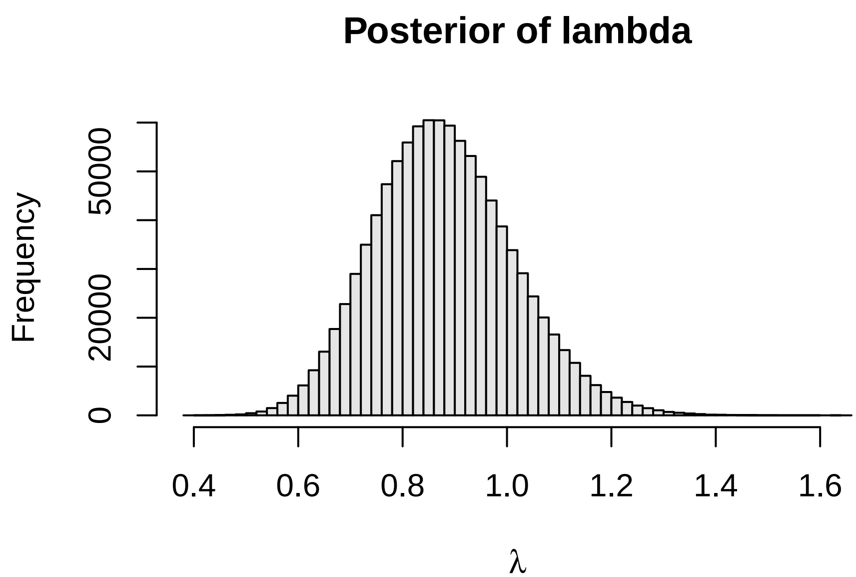
(b)

Note: I use simulation and histograms to display the posterior, but you could also draw the density curve from the `dgamma()` results, or even use grid approximation. All are correct approaches.

```r
# Observed data
n <- 50
y <- c(rep(0, 30), rep(1, 12), rep(2, 6), rep(10, 2))

# Prior specification
a <- b <- 0.01

# Posterior
lambdas <- rgamma(1e6, shape = sum(y) + a, rate = n + b)
hist(lambdas, breaks = 50, main = "Posterior of lambda",
     xlab = expression(lambda), col = "gray90")
```



The posterior mean is

```
(sum(y) + a) / (n + b)
```

```
## [1] 0.880024
```

and a 95% credible interval is

```
qgamma(c(0.025, 0.975), shape = sum(y) + a, rate = n + b)
```

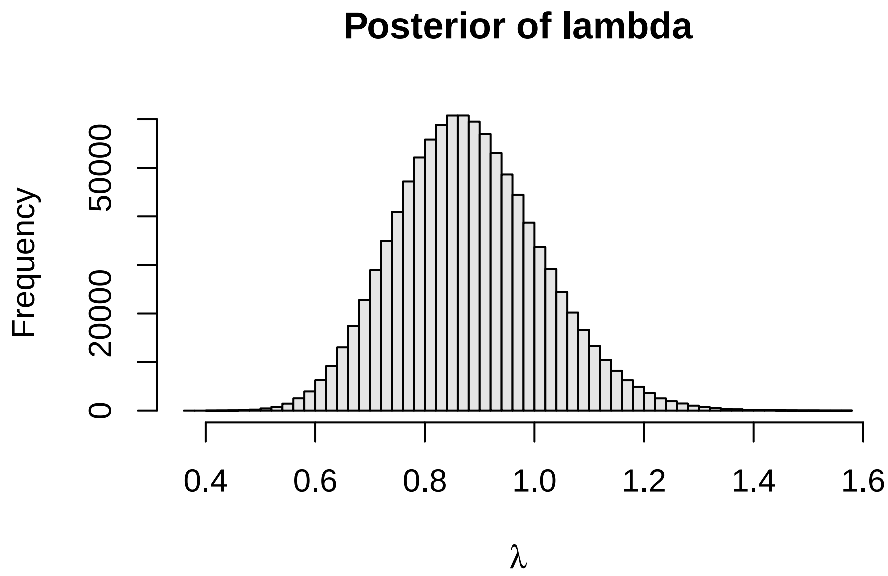
```
## [1] 0.6394519 1.1584123
```

(c)

Gamma(0.1, 0.1) prior

```
# Prior specification
a <- b <- 0.1

# Posterior
hist(rgamma(1e6, shape = sum(y) + a, rate = n + b), breaks = 50,
     main = "Posterior of lambda", xlab = expression(lambda), col = "gray90")
```



The posterior mean is

```
(sum(y) + a) / (n + b)
```

```
## [1] 0.8802395
```

and a 95% credible interval is

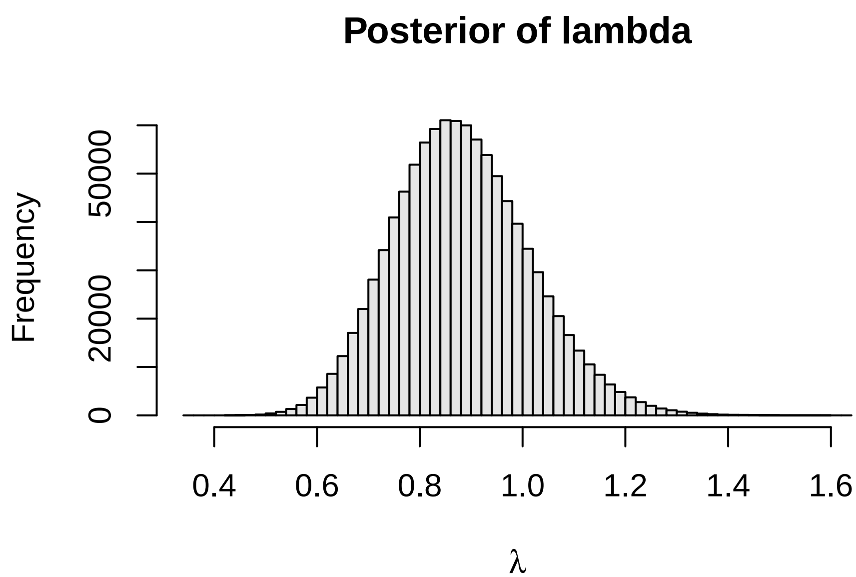
```
qgamma(c(0.025, 0.975), shape = sum(y) + a, rate = n + b)
```

```
## [1] 0.6398339 1.1583935
```

Gamma(1, 1) prior

```
# Prior specification
a <- b <- 1

# Posterior
hist(rgamma(1e6, shape = sum(y) + a, rate = n + b), breaks = 50,
     main = "Posterior of lambda", xlab = expression(lambda), col = "gray90")
```



The posterior mean is

```
(sum(y) + a) / (n + b)
```

```
## [1] 0.8823529
```

and a 95% credible interval is

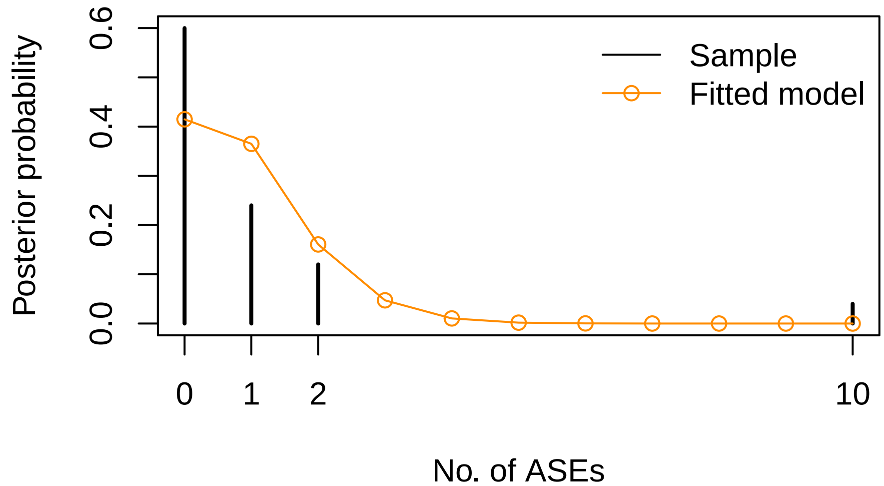
```
qgamma(c(0.025, 0.975), shape = sum(y) + a, rate = n + b)
```

```
## [1] 0.6435943 1.1581950
```

(d)

```
a <- b <- 0.01
lambda_hat <- (sum(y) + a) / (n + b)

plot(table(y) / length(y), ylab = "Posterior probability", xlab = "No. of ASEs")
points(x = 0:10, y = dpois(0:10, lambda_hat), col = "darkorange")
lines(x = 0:10, y = dpois(0:10, lambda_hat), col = "darkorange")
legend("topright", c("Sample", "Fitted model"), col = c("black", "darkorange"),
       pch = c(NA,1), lty=1 , bty="n")
```



No, the Poisson likelihood does not fit these data particularly well. The likelihood has negligible probability at 10, whereas the data contain 2 such cases. Also, more small observations would be expected.

Note: There are multiple acceptable graphics here. The key is a clear comparison between the observed number of ASEs and what the model expects. The legend was not required, but it certainly makes your plot easier to read!

(e)

We can use Monte Carlo sampling to answer this question:

```
# for a = b = 0.01
ys_draw1 <- rgamma(1e6, shape = sum(y) + 0.01, rate = n + 0.01)
mean(ys_draw1 > 1)
```

```
## [1] 0.179678
```

```
# for a = b = 0.1
ys_draw2 <- rgamma(1e6, shape = sum(y) + 0.1, rate = n + 0.1)
mean(ys_draw2 > 1)
```

```
## [1] 0.180423
```

```
# for a = b = 1
ys_draw3 <- rgamma(1e6, shape = sum(y) + 1, rate = n + 1)
mean(ys_draw3 > 1)
```

```
## [1] 0.182191
```

The posterior probability that the new medication has a higher side effect rate than the previous medication is about 0.18, and this is quite robust to prior specification.

Note: You could also calculate these using the `1 - pgamma(1, sum(y) + a, n + b)`, I am simply reinforcing the idea of Monte Carlo sampling here.

9.

(a)

A negative binomial assumes that

1. each day is independent,
2. each day has the same probability of a relapse, and
3. that the experiment stops after a prespecified number of failures (in this case one).

This describes the smoking cessation attempt fairly well. The main assumptions are independence and equal probability. Surely the probability varies with time since the beginning of the quit attempt, but it perhaps reasonable for a short study such as this.

(b)

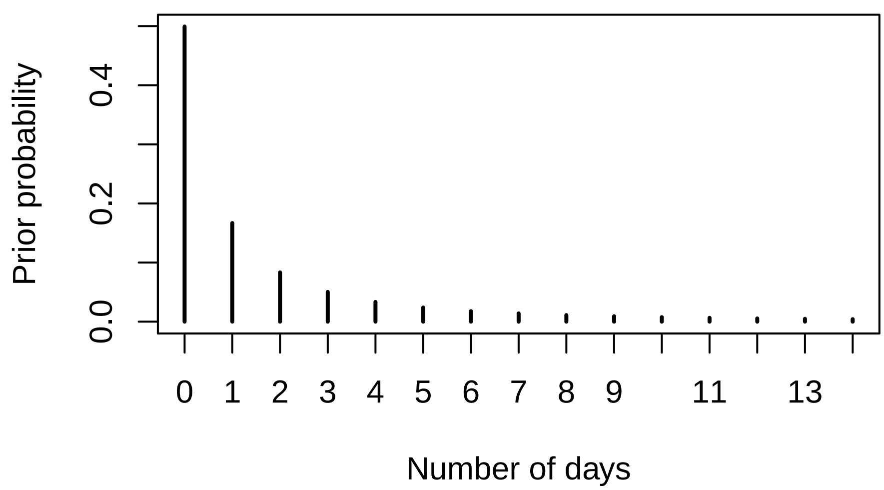
This question requires that you use the prior predictive distribution. You could derive this distribution using calculus, but it's far easier to approximate it using Monte Carlo sampling.

Find the prior probability the smoker will be smoke-free for at least 7 days. (Here, we need to use the prior-predictive distribution... does the book mention this?)

```
S <- 1e6
theta <- runif(S, 0, 1)
y     <- rnbinom(S, size = 1, prob = theta)
```

Let's plot the prior predictive to see what we'd expect (the plot is truncated at 14 because the PMF has a very heavy right tail):

```
plot(table(y)/S, xlab = "Number of days", ylab = "Prior probability", xlim = c(0, 14))
```



The prior probability that the smoker will be smoke-free for at least 7 days is approximately 0.125.

```
mean(y >= 7)
```

```
## [1] 0.12509
```

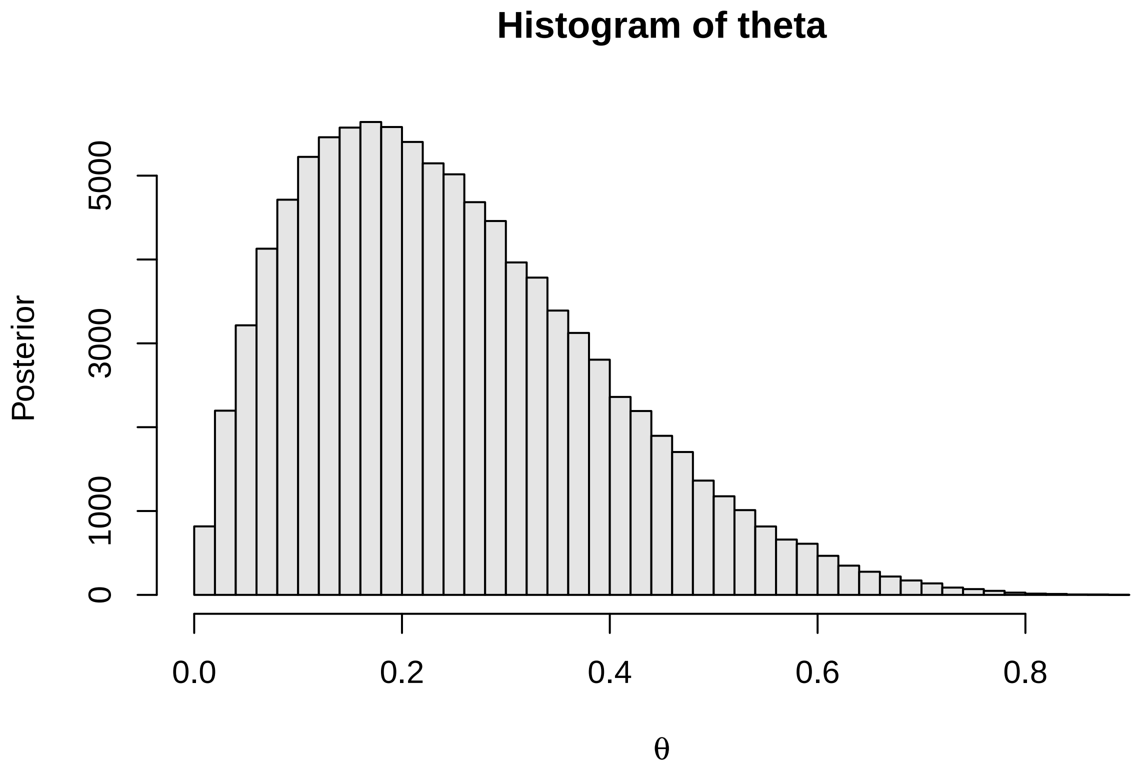
(c)

The posterior is

$$
\begin{aligned}
p(\theta|Y) &\propto f(Y|\theta)\pi(\theta) && (8)\\
&\propto \left[\theta(1-\theta)^Y\right]\cdot 1 && (9)\\
&= \theta^{2-1}(1-\theta)^{(Y+1)-1} && (10)
\end{aligned}
$$

which resembles the Beta$(1+1,\ Y+1)$.

```
S <- 1e5
theta <- rbeta(S, 2, 6)
hist(theta, breaks = 50 , xlab = expression(theta), ylab = "Posterior", col = "gray90")
```



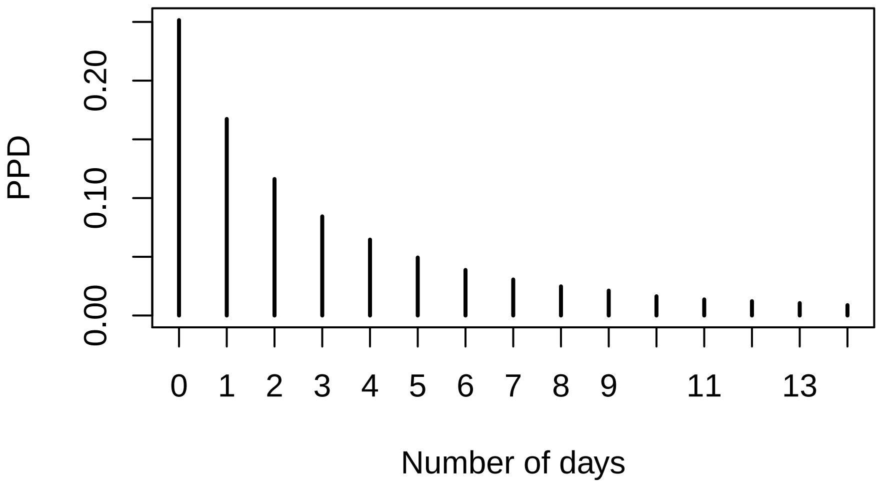
(d)

We can approximate the posterior predictive distribution using Monte Carlo simulation:

```
S <- 1e5
theta   <- rbeta(S, 2, 6)
y_draws <- rnbinom(S, size = 1, prob = theta)
```

Let's plot the prior predictive to see what we'd expect (the plot is truncated at 14 because the PMF has a very heavy right tail):

```
plot(table(y_draws)/S, xlab="Number of days", ylab="PPD", xlim= c (0, 14))
```



Given the data, the probability that the smoker will be smoke-free for at least 7 days is approximately 0.231.

```
mean(y_draws >= 7)
```

```
## [1] 0.2277
```