Homework 7

Zachary Lazerick

21 March 2023

1) The Monty Hall Problem – At the end of the television show Let's Make a Deal, a contestant is asked to choose one of three doors numbered 1, 2, and 3. Behind one of the doors is a new car, and behind the other two are a gag prize, such as a goat. After the contestant chooses a door, Monty Hall, the host of the show, opens one of the two doors which contain a goat (but never the door that the contestant chose, as that would ruin the fun). Monty offers the contestant the opportunity to switch from his chosen door to the remaining unopened door or keep the prize that he originally chose.

a) Should the contestant switch or stay with the original choice? It would seem that the odds of winning the car are now 50/50, right? Write a simulation to determine the probability of the car being behind each of the remaining doors. In other words, randomly select the placement of the car, the choice of the contestant, and the door that Monty will open. If the contestant keeps their initial selection, how often does (s)he win? If the contestant switches, how often does (s)he win?

```
GoatorNoGoat <- function(iterations, num.doors = c(1, 2, 3), switch = T) {
  output <- rep(0, iterations); counter <- 1
  while (counter < iterations) {
    ## Pick Door for Car
    CarDoor <- sample(num.doors, size = 1)

    ## Pick Door for Contestant
    ContestantDoor <- sample(num.doors, size = 1)

    ## Pick Door for Monty
    MontyDoor <- sample(setdiff(num.doors, union(CarDoor, ContestantDoor)), size = length(num.doors)-2,

    ## If Switch = T, Contestant Switches Their Choice
    if (switch == T) {
      ContestantDoor <- num.doors[-c(MontyDoor, ContestantDoor)]
    }

    ## Check if Contestant Picked Car
    if (ContestantDoor == CarDoor) {
      output[counter] <- 1
      counter <- counter + 1
    }
    else {
      counter <- counter + 1
    }
  }
  return(output)
}
```

```
##set.seed(1932); results1a.1 <- GoatorNoGoat(1000, num.doors = 1:3);
##set.seed(232); results1a.2 <- GoatorNoGoat(1000, num.doors = 1:3, switch = F)
##mean(results1a.1); mean(results1a.2)
```

Unable to fix a bug with code where Monty repeatedly 'opened' a door that was already assigned to the car or the contestant. The theoretical probabilities for the results when the contestant does switch is $P(Success) = \frac{2}{3}$ and when the contestant does no switch is $P(Success) = \frac{1}{3}$. A 'success' is defined as the contestant picking the door that the car is behind.

b) Suppose that there are 10 doors instead, 1 car and 9 goats. The contestant chooses a door, Monty opens 8 doors containing a goat, and then gives the contestant the opportunity to switch before the final reveal. Now what is the probability that the contestant wins if (s)he switches to the unopened door?

```
num.doors <- c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10)
set.seed(732); results1b.1 <- GoatorNoGoat(1000, num.doors = 1:10);
set.seed(1432); results1b.2 <- GoatorNoGoat(1000, num.doors = 1:10, switch = F)
mean(results1b.1); mean(results1b.2)
```

```
## [1] 0.908
```

```
## [1] 0.102
```

Hint: You want to randomly select doors which are not the door with the car and not the door the contestant picks. If X = 1:10, you can set doors = X[-c(contestant, car)] and then sample(doors) to generate a random permutation of integers representing the doors that Monty opens each time. By default, the sample() command samples from a set of integers (here, doors) without replacement. Another approach is a while loop – sample until you get something that is not the car and not the contestant's pick. Note: Having Monty "open" doors is not essential to solving either problem, but it adds to the reality of the simulation. For fun, you can type messages which output which door Monty is opening. But I wouldn't do that when you are running 1000+ simulations in order to calculate the probability. How does the Choice of a Prior Distribution Affect the Posterior Probability?

2) The Binomial Distribution – Suppose that you have the following binary data set where '1' indicates 'Success' and '0' indicates 'Failure'

0 0 1 0 0 1 0 0 0 0 1 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 1

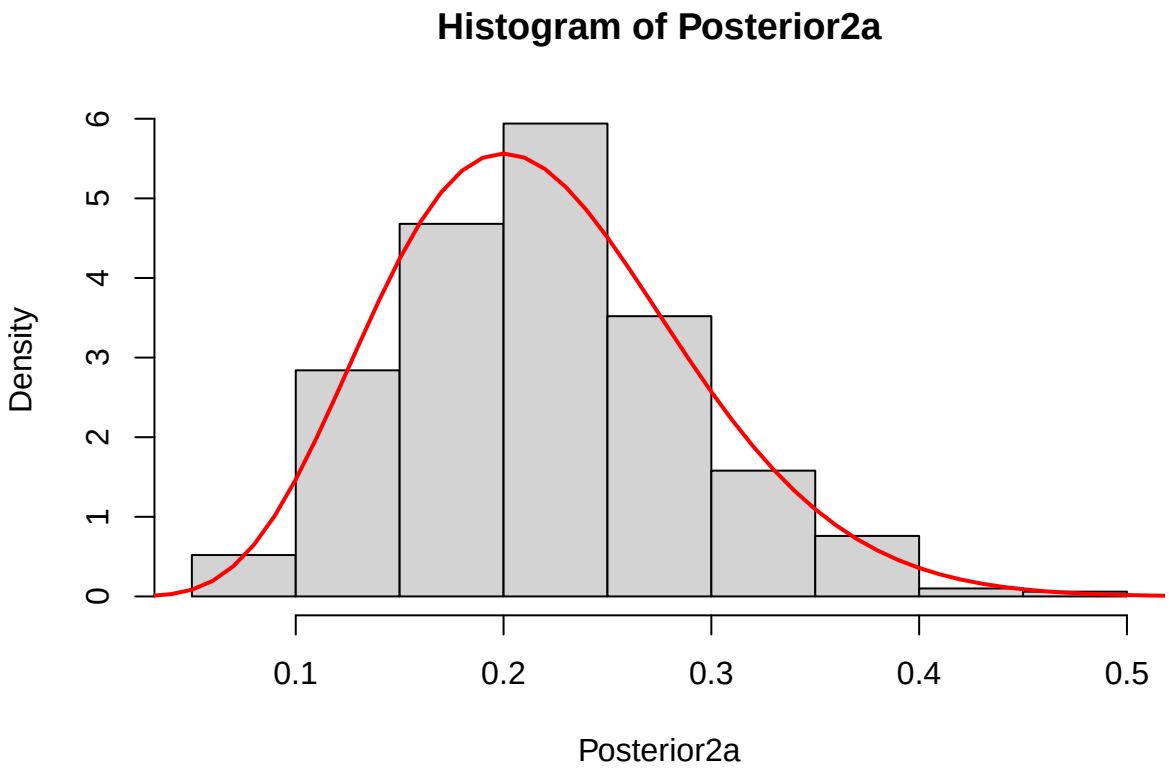
Let θ = the probability of success in a given trial. For each prior distribution on θ given below, plot the posterior distribution, determine the posterior mean of θ, and give 95% credibility limits for θ Note: You can do this via sampling or using some of R's built-in functions. When sampling, if you generate 1000 values from the posterior distribution then your 95% credibility limits are the values in positions 25 and 975 of the sorted list of observations.

```
## Likelihood ~ Binomial(n = 30, p = 6/30) --> 6 success, 24 failure
```

a) $\theta \sim$ Uniform(0, 1)

```
num.samp = 1000; set.seed(1001)
Posterior2a <- rbeta(num.samp, 6+1, 24 + 1)

hist(Posterior2a, freq = F)
X <- seq(0, 1, .01); Beta <- dbeta(X, 6+1, 24+1)
lines(X, Beta, col = 'red', lwd = 2)
```




```
mean(Posterior2a)
```

```
## [1] 0.2187697
```

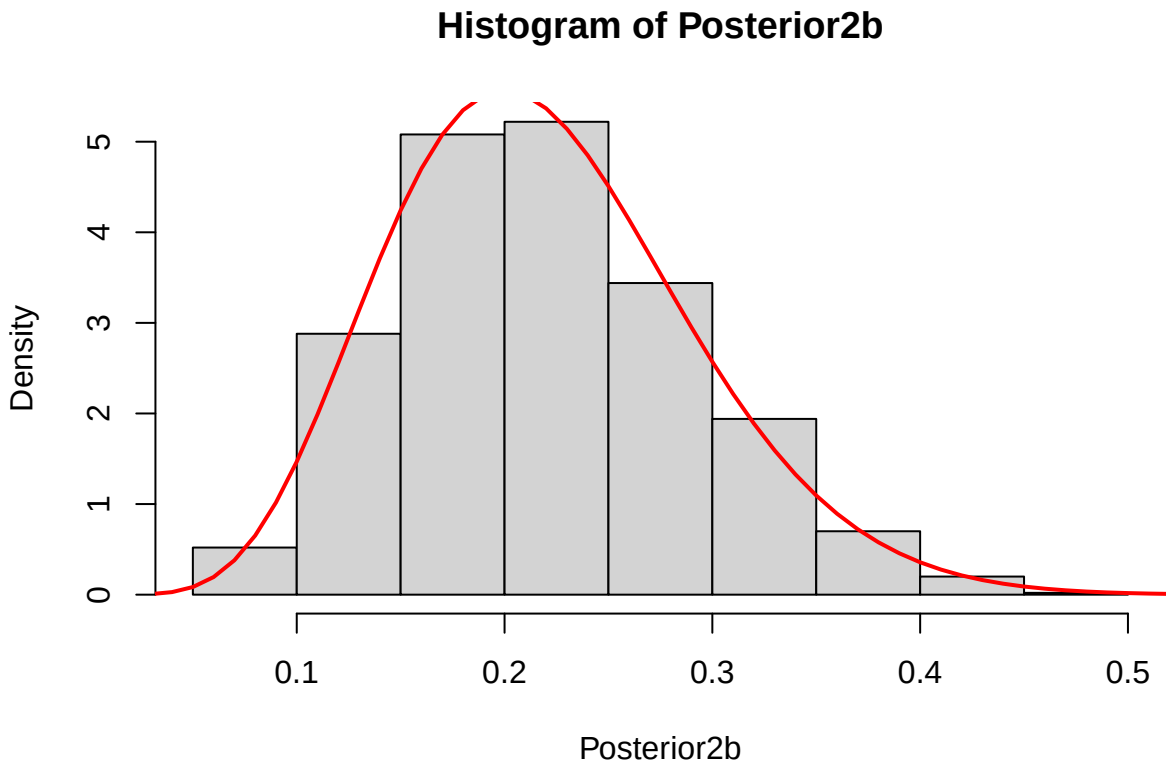
```
quantile(Posterior2a, c(0.025, 0.975))
```

```
##       2.5%      97.5%
## 0.09843336 0.37415750
```

b) $\theta \sim \text{Beta}(2, 6)$

```
set.seed(1002)
Posterior2b <- rbeta(num.samp, 6+1, 24 + 1)

hist(Posterior2b, freq = F)
X <- seq(0, 1, .01); Beta <- dbeta(X, 6+1, 24+1)
lines(X, Beta, col = 'red', lwd = 2)
```




```
mean(Posterior2b)
```

```
## [1] 0.2199696
```

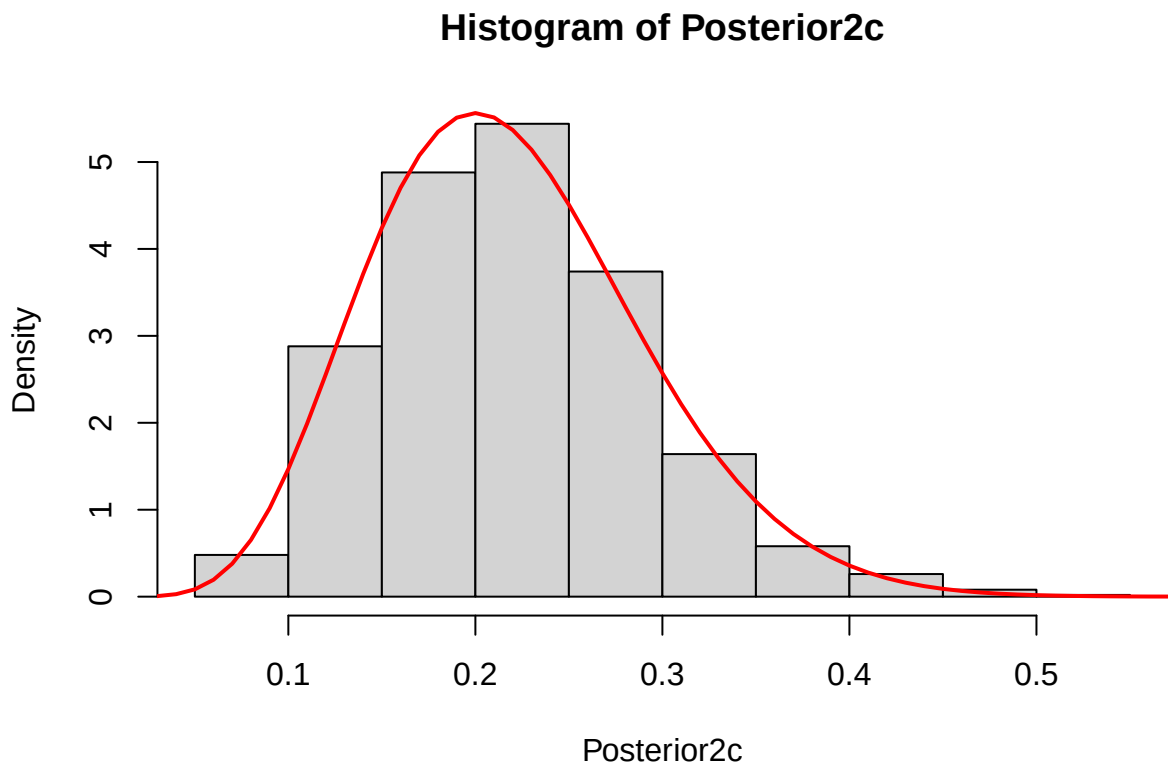
```
quantile(Posterior2b, c(0.025, 0.975))
```

```
##       2.5%      97.5%
## 0.09959845 0.37131951
```

c) $\theta \sim \text{Beta}(6, 2)$

```
set.seed(1000)
Posterior2c <- rbeta(num.samp, 6+1, 24 + 1)

hist(Posterior2c, freq = F)
X <- seq(0, 1, .01); Beta <- dbeta(X, 6+1, 24+1)
lines(X, Beta, col = 'red', lwd = 2)
```




```
mean(Posterior2c)
```

```
## [1] 0.2198084
```

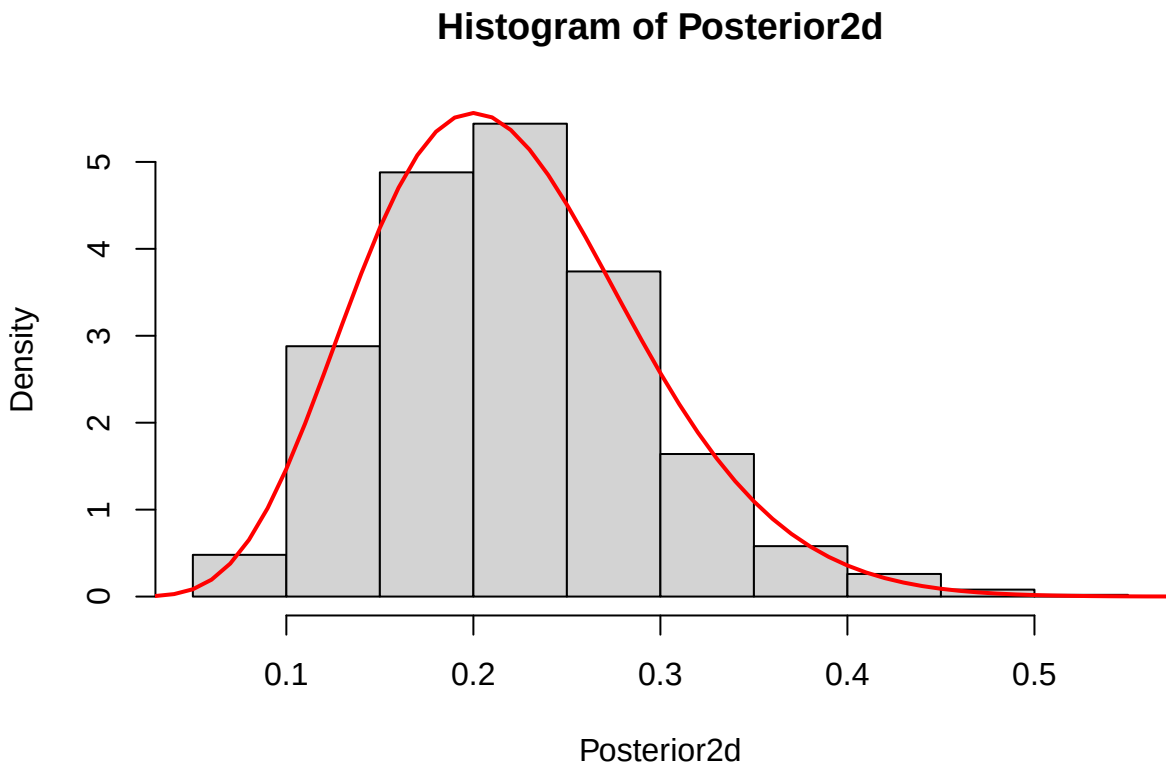
```
quantile(Posterior2c, c(0.025, 0.975))
```

```
##      2.5%     97.5%
## 0.1008735 0.3846856
```

d) $\theta \sim \text{Beta}(20, 60)$

```
set.seed(1000)
Posterior2d <- rbeta(num.samp, 6+1, 24 + 1)
```

```r
hist(Posterior2d, freq = F)
X <- seq(0, 1, .01); Beta <- dbeta(X, 6+1, 24+1)
lines(X, Beta, col = 'red', lwd = 2)
```




```r
mean(Posterior2d)
```

```
## [1] 0.2198084
```

```r
quantile(Posterior2d, c(0.025, 0.975))
```

```
##      2.5%     97.5%
## 0.1008735 0.3846856
```

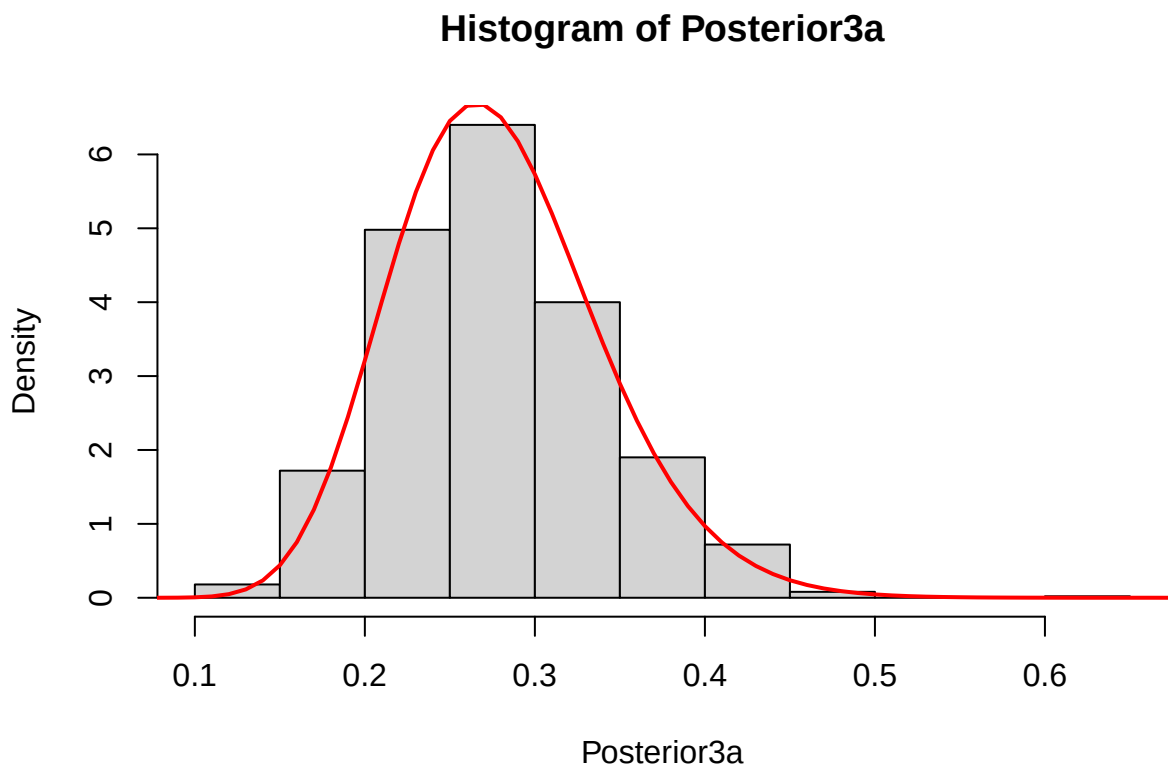
3) The Exponential Distribution – Suppose that you have the following data points which are assumed to follow an Exponential distribution with mean $\beta = 1/\theta$. | 5.149 | 1.544 | 1.692 | 7.266 | 8.515 | 2.786 | 0.164 | 4.311 | 2.143 | 5.988 | | 1.144 | 5.465 | 2.725 | 1.432 | 0.462 | 0.166 | 2.412 | 7.904 | 7.607 | 5.427 | For each prior distribution on θ given below, plot the posterior distribution, determine the posterior mean of θ, and give 95% credibility limits for θ

```
## Likelihood --> n = 20, sum.x = 74.302
```

a) $\theta \sim$ Gamma(α=1, λ=1)

```
set.seed(1005)
Posterior3a <- rgamma(num.samp, 20 + 1, 74.302 + 1)

hist(Posterior3a, freq = F)
X <- seq(0, 1, .01); Gamma <- dgamma(X, 20 + 1, 74.302+1)
lines(X, Gamma, col = 'red', lwd = 2)
```




```
mean(Posterior3a)
```

```
## [1] 0.2798423
```

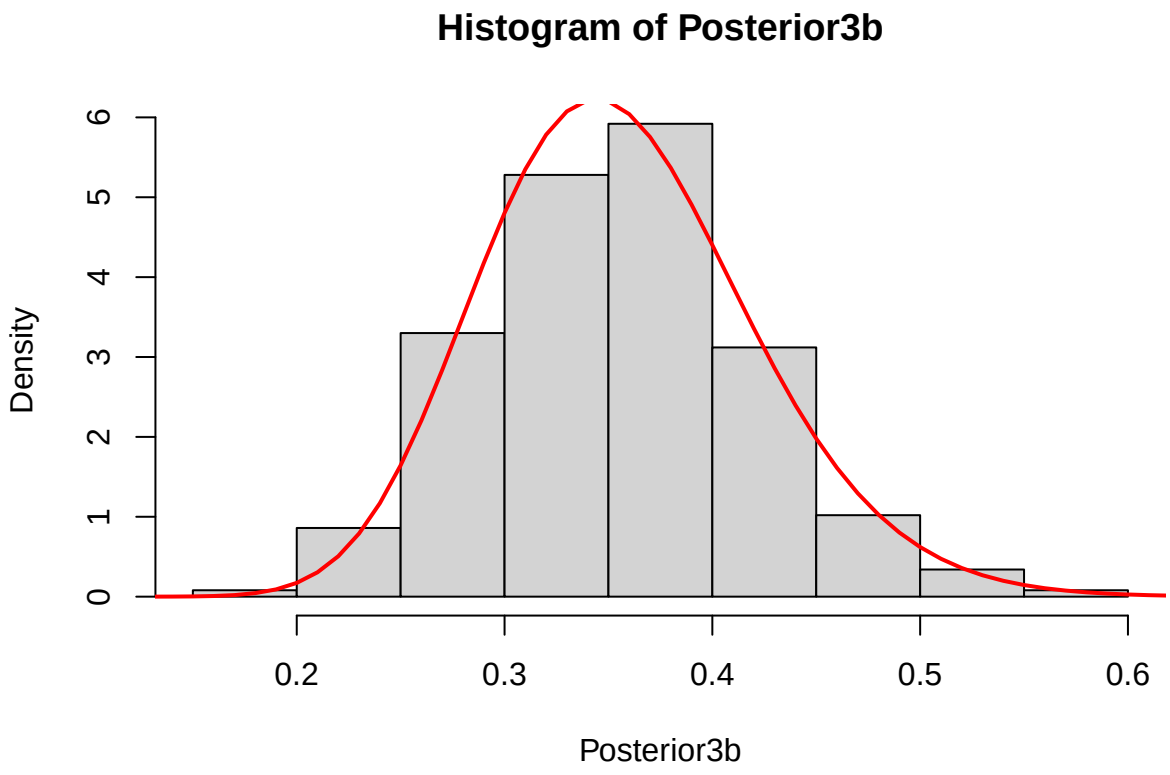
```
quantile(Posterior3a, c(0.025, 0.975))
```

```
##      2.5%     97.5%
## 0.1733945 0.4170736
```

b) $\theta \sim$ Gamma(α=10, λ=10)

```
set.seed(1006)
Posterior3b <- rgamma(num.samp, 20 + 10, 74.302 + 10)

hist(Posterior3b, freq = F)
X <- seq(0, 1, .01); Gamma <- dgamma(X, 20 + 10, 74.302+10)
lines(X, Gamma, col = 'red', lwd = 2)
```




```
mean(Posterior3b)
```

```
## [1] 0.3542217
```

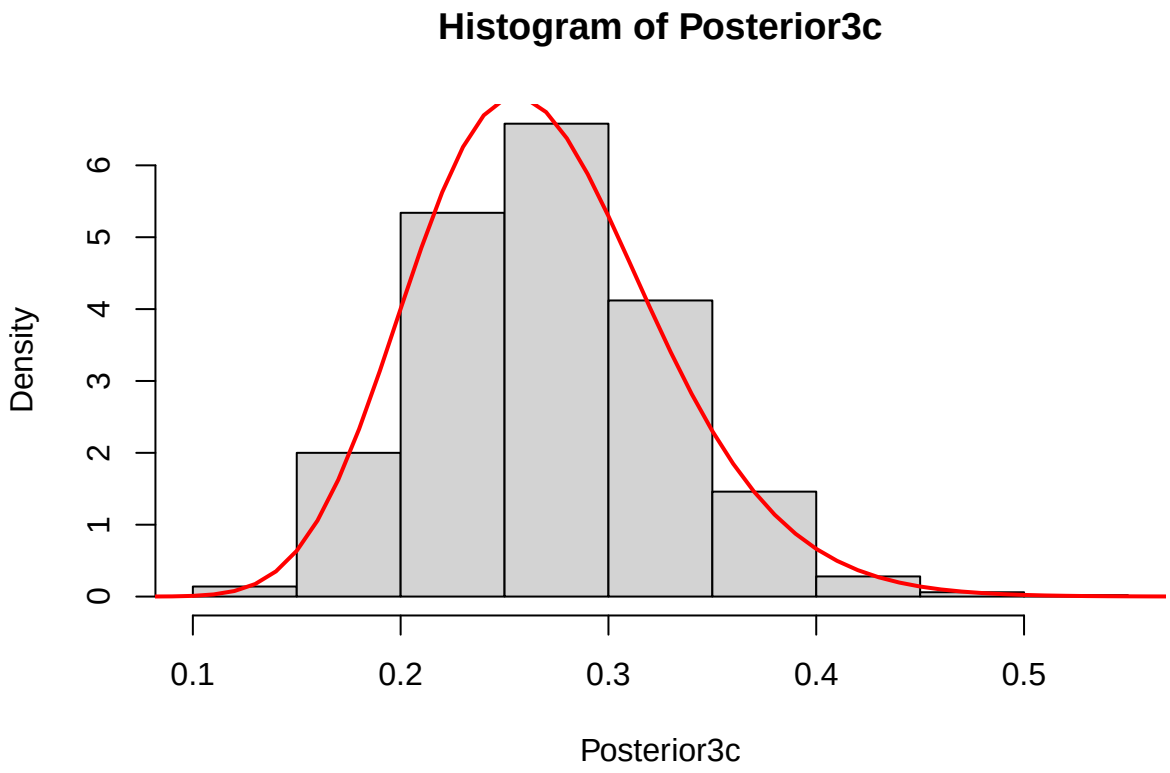
```
quantile(Posterior3b, c(0.025, 0.975))
```

```
##      2.5%     97.5%
## 0.2402608 0.4875288
```

c) $\theta \sim$ Gamma (α=1, λ=4)

```
set.seed(1007)
Posterior3c <- rgamma(num.samp, 20 + 1, 74.302 + 4)
```

```
hist(Posterior3c, freq = F)
X <- seq(0, 1, .01); Gamma <- dgamma(X, 20 + 1, 74.302+4)
lines(X, Gamma, col = 'red', lwd = 2)
```




```
mean(Posterior3c)
```

```
## [1] 0.2710907
```

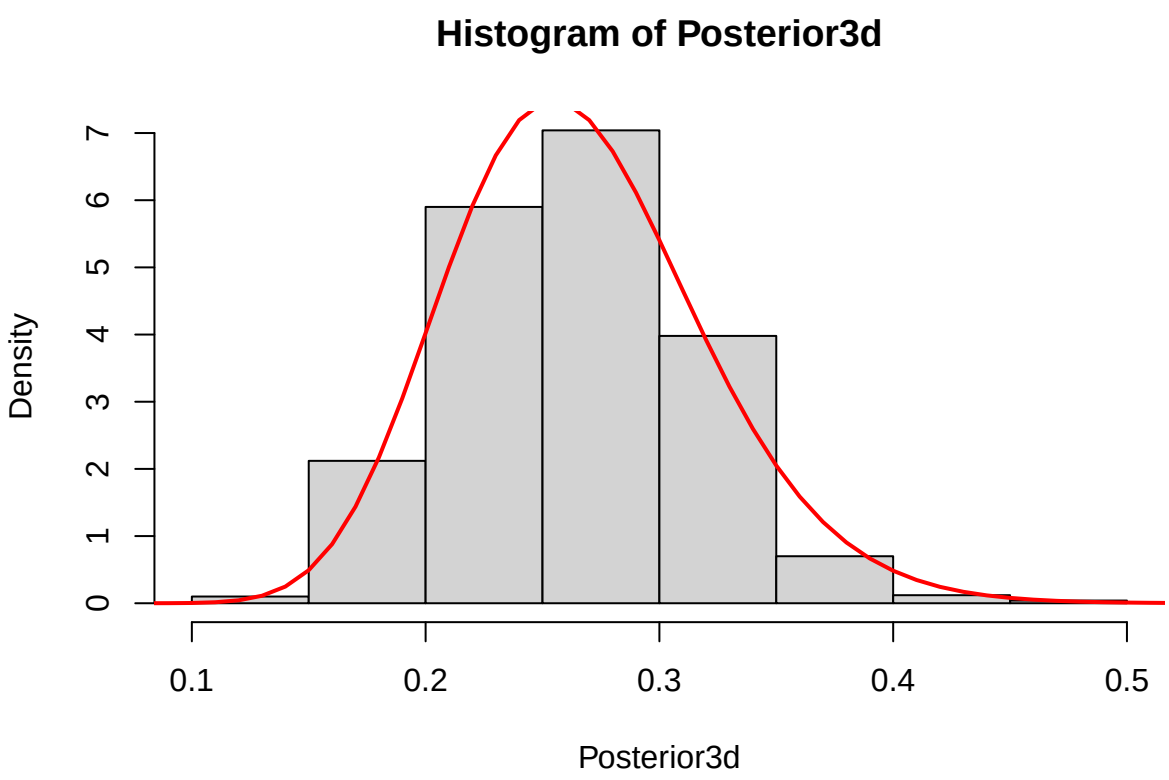
```
quantile(Posterior3c, c(0.025, 0.975))
```

```
##      2.5%     97.5%
## 0.1674650 0.3904462
```

d) $\theta \sim \text{Gamma}(\alpha{=}4, \lambda{=}16)$

```
set.seed(1008)
Posterior3d <- rgamma(num.samp, 20 + 4, 74.302 + 16)

hist(Posterior3d, freq = F)
X <- seq(0, 1, .01); Gamma <- dgamma(X, 20 + 4, 74.302+16)
lines(X, Gamma, col = 'red', lwd = 2)
```




```
mean(Posterior3d)
```

```
## [1] 0.2636385
```

```
quantile(Posterior3d, c(0.025, 0.975))
```

```
##      2.5%     97.5%
## 0.1697694 0.3643467
```

Note: There are two different parameterizations for the gamma distribution. Make sure you are using the correct one for each calculation. In R's dgamma/qgamma functions, the input parameter scale is equivalent to β, rate is equivalent to λ.

e) Using the R function rgamma(), simulate 1000 values from the posterior distribution of θ in part (d).

```
set.seed(1140)
results3e <- rgamma(1000, 20 + 4, 74.302 + 16)
```

f) By transforming (i.e. inverting) these simulated draws, we can obtain a simulated sample from the posterior distribution of β. Calculate the mean and standard deviation of the posterior distribution on β, then make a histogram.

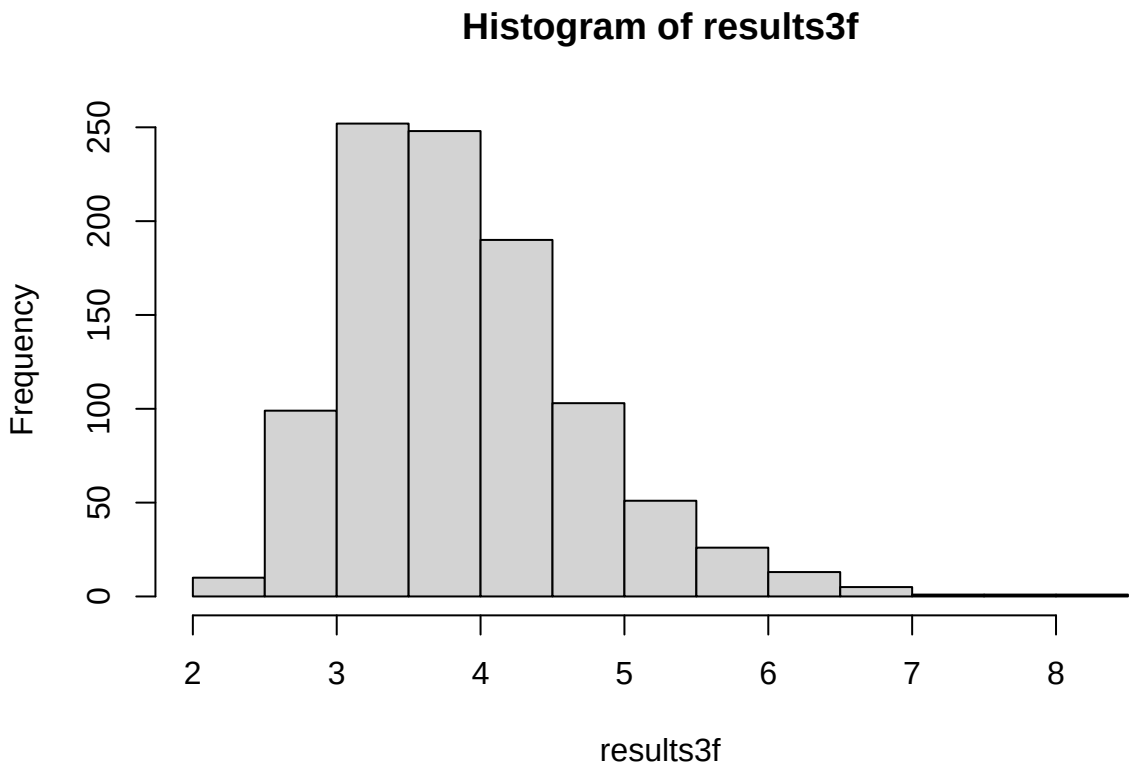
```
results3f <- 1/results3e

mean(results3f); var(results3f)
```

```
## [1] 3.897398
```

```
## [1] 0.6903288
```

```
hist(results3f)
```



g) Estimate the posterior probability that β exceeds 4.

```
mean(results3f > 4)
```

```
## [1] 0.391
```

4) The Poisson Distribution (Source: Bayesian Data Analysis, Ch. 2, Question 13) Note: You only have to turn in answers to parts (a)-(c) and (f). Parts (d) and (e) are optional. The table below gives the number of fatal accidents and deaths on scheduled airline flights per year over a ten-year period: | Year | Fatal Accidents | Passenger Deaths | Death Rate | |:—-:|:————:|:————-:|:———-:| | 1976 | 24 | 734 | 0.19 | | 1977 | 25 | 516 | 0.12 | | 1978 | 31 | 754 | 0.15 | | 1979 | 31 | 877 | 0.16 | | 1980 | 22 | 814 | 0.14 | | 1981 | 21 | 362 | 0.06 | | 1982 | 26 | 764 | 0.13 | | 1983 | 20 | 809 | 0.13 | | 1984 | 16 | 223 | 0.03 | | 1985 | 22 | 1066 | 0.15 |

Note: Death rate is passenger deaths per 100 million passenger miles

a) Assume that the numbers of fatal accidents in each year are independent with a Poisson(θ) distribution. Select a prior distribution for θ and determine the posterior distribution based on the data from 1976 through 1985. Be sure to give a short justification for your choice of parameter values. Then determine the mean and 95% credibility limits of your posterior distribution.

```
## Prior --> Gamma(47.6, 2) (Averaged 23.8 fa/year, add 2 years of 'past' data)

set.seed(1036)
Posterior4a <- rgamma(num.samp, 238 + 47.6, 10 + 2)

mean(Posterior4a)
```

```
## [1] 23.82017
```

```
quantile(Posterior4a, c(0.025, 0.975))
```

```
##     2.5%    97.5%
## 21.13771 26.53937
```

b) Under this model, give a 95% predictive interval for the number of fatal accidents in 1986. Note: Predictive probability is given by:

$$P(\hat{Y}|Y) = \int P(\hat{Y}|\theta, Y)P(\theta|Y)d\theta$$

and this integral reduces to a Negative Binomial distribution! In other words, $Y \sim$ Neg-Bin(α, β) where the parameters α and β correspond to the parameters α and β in the posterior gamma distribution on θ. Specifically:

$$P(\hat{Y}|Y) = \binom{\alpha + \hat{y} - 1}{\hat{y}} (\frac{\beta}{\beta+1})^{\alpha} (\frac{1}{\beta+1})^{\hat{y}}$$

Here, α = the number of "Successes" (size parameter), y = the number of "Failures", and the probability of "Success" is given by $\beta/(1+\beta)$ (prob parameter). Thus, the predictive interval can easily be computed via simulation or using some of R's built-in functions (see Homework #5, Q1). Just remember that R defines a negative binomial random variable as the number of failures until the rth success (as opposed to the number of trails until the rth success), which is convenient in this case since that is what $\hat{y}$ represents. See pages 52-53 in the handout from Bayesian Data Analysis for more details.

```
results4b <- rnbinom(num.samp, size = 238 + 47.6, prob = (10+2)/(10+2+1))

quantile(results4b, c(0.025, 0.975))
```

```
##  2.5% 97.5%
##    14    34
```

c) Repeat parts (a) and (b) above, replacing 'fatal accidents' with 'passenger deaths'.

```
## Prior --> Gamma(1383.8, 2) (691.9 Averaged pd/year, add 2 years of 'past' data)

set.seed(1145)
Posterior4c <- rgamma(num.samp, 6919 + 1383.8, 10 + 2)

mean(Posterior4c)
```

```
## [1] 691.817
```

```
quantile(Posterior4c, c(0.025, 0.975))
```

```
##     2.5%    97.5%
## 677.2435 705.6652
```

```
results4c <- rnbinom(1000, size = 6919 + 1383.8, prob = (10+2)/(10+2+1))

quantile(results4c, c(0.025, 0.975))
```

```
##    2.5%   97.5%
## 639.975 743.000
```

d) For a challenge, try: Assume that the number of fatal accidents each year follow independent Poisson distributions with a constant rate and exposure each year proportional to the number of passenger miles flown. In other words, your rate parameter, θ, is now the number of 'fatal accidents' per 100 million miles flown, and the exposure each year is X(θ), where X is the number of hundreds of millions of miles flown in that particular year [Note: this changes our formula for β that we derived in class]. You can estimate the number of passenger miles flown in each year by dividing the 'deaths' column by the 'rate' column, ignoring round-off errors. Set a prior distribution for θ and determine the posterior distribution based on the data for 1976-1985. Give a 95% predictive interval for the number of fatal accidents in 1986 under the assumption that $8x10^{11}$ passenger miles (x = 8000 hundred million miles) are flown that year. Here,

$$P(\hat{Y}|Y) = \binom{\alpha + \hat{y} - 1}{\hat{y}} (\frac{\beta}{\beta + \hat{x}})^{\alpha} (\frac{\hat{x}}{\beta + \hat{x}})^{\hat{y}}$$

[Answer should be (22, 46)]

e) For a challenge, try: Repeat (d) above, replacing 'fatal accidents' with 'passenger deaths'. [Answer should be (904, 1034)]

f) In which of the cases (b) – (e) above does the Poisson model seem more or less reasonable? Why? Discuss based on general principals, without specific reference to the numbers in the table. Incidentally, in 1986, there were 22 fatal accidents, 546 passenger deaths, and a death rate of 0.06 per 100 million miles flown.

In case (b), the poisson model seems fairly reasonable. We are not tracking which planes crash, just how many total crash. In case (c), the poisson model does not seem reasonable because the number of passenger deaths from the fatal accidents would likely be affected by which types of planes (short domestic, long domestic, international) crash, since different types of planes can seat a varying amount of people. Parts (d) and (e) fall into similar predicaments as parts (b) and (c) respectively. However, these are slightly better than their counterparts due to the normalization of setting a 'time interval' by calculating per 100 million passenger miles.

5) Election Day. An election is approaching, and you are interested in knowing whether people in Massachusetts prefer candidate A or candidate B. A recently published newspaper poll states of that of 100 randomly sampled people, 58 preferred candidate A and the remainder preferred candidate B.

```
## Binomial Likelihood --> Binom(100, 58/100)
```

a) Suppose that before the newspaper poll, your prior belief was a uniform distribution (which would make sense if you didn't know which candidate had stronger support in the state). What is the mean, variance, and 95% credible interval on your beliefs after learning of the newspaper poll results?

```
set.seed(1317)
Posterior5a <- rbeta(num.samp, 58 + 1, 42 + 1)

mean(Posterior5a); var(Posterior5a)
```

```
## [1] 0.5776616
```

```
## [1] 0.002349291
```

```
quantile(Posterior5a, c(0.025, 0.975))
```

```
##      2.5%     97.5%
## 0.4803360 0.6699415
```

b) You want to conduct a follow-up poll to narrow down your estimate of the population's preference. In your follow-up poll, you randomly sample 100 other people and find that 57 prefer candidate A and the remainder prefer candidate B. Assuming that peoples' opinions have not changed between polls, what is the mean, variance, and 95% credible interval on your beliefs after conducting this second poll? How does this compare to your answer to part (a)?

```
set.seed(1348)
Posterior5b <- rbeta(num.samp, 57 + 58 + 1, 43 + 42 + 1)

mean(Posterior5b); var(Posterior5b)
```

```
## [1] 0.5736816
```

```
## [1] 0.001169789
```

```
quantile(Posterior5b, c(0.025, 0.975))
```

```
##      2.5%     97.5%
## 0.5106481 0.6401739
```

After the follow-up poll, the 95% credible interval states that candidate A has majority support. (i.e. all values in the interval $> .5$).

c) Now, suppose you are interested in comparing the rate of support for candidate A in two different states. A survey of sample size 50 was done in the second state (say, New York), and the total number of people in the sample who supported candidate A was 30. Using the data from the original newspaper survey as the data for Massachusetts (58/100), identify the posterior distribution of both θ_{MA} and θ_{NY} assuming a uniform prior. Sample 1,000 values of each of θ_{MA} and θ_{NY} from their posterior distributions and estimate $\mathrm{P}(\theta_{MA} < \theta_{NY})$.

```
## NY ~ Beta(30 + 1, 20 + 1), MA ~ Beta(58 + 1, 42 + 1)

set.seed(1350)
PosteriorNY <- rbeta(num.samp, 30 + 1, 20 + 1); PosteriorMA <- rbeta(num.samp, 58 + 1, 42 + 1)

mean(PosteriorMA < PosteriorNY)
```

```
## [1] 0.604
```

6) Three-Point Shooting. The number of three-point shots made by a team in an NCAA college basketball game can be modeled by a Poisson distribution with unknown rate θ. The purpose of this question is to estimate the rate of three-point shots made by the Holy Cross Women's basketball team in the 2022-2023 season. Note: Overall, NCAA teams make approximately 6 three-point shots per game.

a) What would be an appropriate prior distribution in this situation?

An appropriate prior distribution would be a gamma distribution with a ratio for $\alpha : \beta$ of 6:1 for 6 'successful' three-pointers per 1 game.

b) In their last 12 games, the Crusaders have made the following number of three-point shots: | 4 | 7 | 6 | 2 | 3 | 6 | 2 | 8 | 5 | 8 | 3 | 4 | |:-:|:-:|:-:|:-:|:-:|:-:|:-:|:-:|:-:|:-:|:-:|:-:|:-:|:-:| Use this data to identify the parameters of the posterior distribution. What is the mean, variance, and 95% credible interval for the average number of three-point shots made in a game (theta)?

```
## Prior Ratio 6:1, add 2 'past' games --> Gamma(12, 2)

set.seed(1402)
Posterior6b <- rgamma(num.samp, 12 + 58, 2 + 12)

mean(Posterior6b); var(Posterior6b)
```

```
## [1] 4.982761
```

```
## [1] 0.3429205
```

```
quantile(Posterior6b, c(0.025, 0.975))
```

```
##     2.5%    97.5%
## 3.899365 6.209897
```

c) On Sunday, the Crusaders beat BU to win the Patriot League championship and will advance to the NCAA tournament! Their opening round game on Friday is against Maryland. Let's try and predict how many three-point field goals they'll make! To do this, sample a value from the posterior distribution on θ, and then use this value of θ to simulate a value from the Poisson distribution. Repeat the process 1000 times. Save the sampled values in a vector and then call the table() function to see the distribution of three-point shots made across a large number of games. In this scenario, what is the mean number of three-point shots made? How does it compare to your answer from part (b)?

```
set.seed(1404)
Theta6c <- rgamma(num.samp, 12 + 58, 2 + 12)
Poisson6c <- rpois(num.samp, Theta6c)

table(Poisson6c)
```

```
## Poisson6c
##   0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16
##   9  34  68 160 177 163 136  94  72  35  27  11   8   3   1   1   1
```

```
mean(Poisson6c)
```

```
## [1] 5.109
```

The mean is slightly higher in this scenario

d) Maryland has made the following number of three-point field goals across their last 12 games: | 8 | 5 | 7 | 11 | 4 | 12 | 8 | 6 | 2 | 6 | 7 | 4 | |—|—|—|—|-|—|—-|—|—|—|—|—|—| Using the same prior that you identified in part (a), what is the posterior distribution for the number of three-point field goals made by the Maryland?

```
set.seed(1420)
Posterior6d <- rgamma(num.samp, 12 + 80, 2 + 12)

mean(Posterior6d); var(Posterior6d)
```

```
## [1] 6.558677
```

```
## [1] 0.4652078
```

```
quantile(Posterior6d, c(0.025, 0.975))
```

```
##     2.5%    97.5%
## 5.312940 7.959982
```

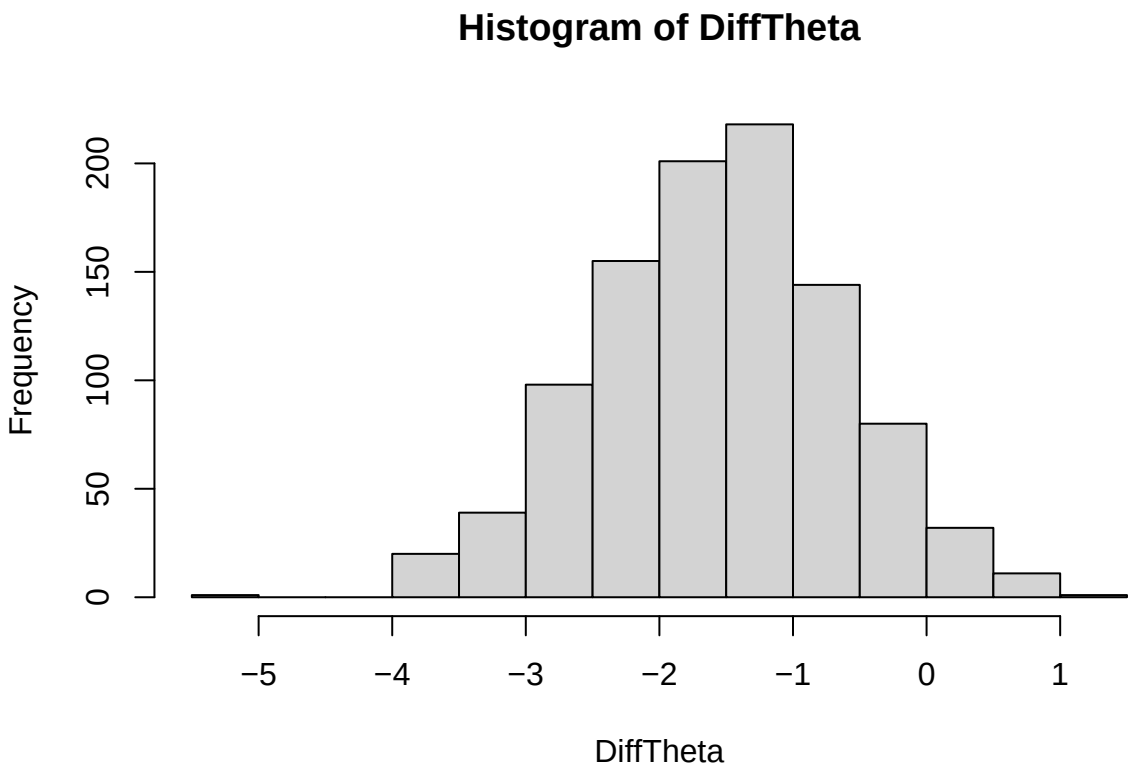
e) Repeat part (c) for Maryland. Use the two sets of sampled values to approximate the distribution of $\theta_{HC} - \theta_M$ and $Y_{HC} - Y_M$. Then construct 95% Bayesian credible intervals for $\theta_{HC} - \theta_M$ and $Y_{HC} - Y_M$. Describe in words the differences between these two teams in terms of their three-point shooting.

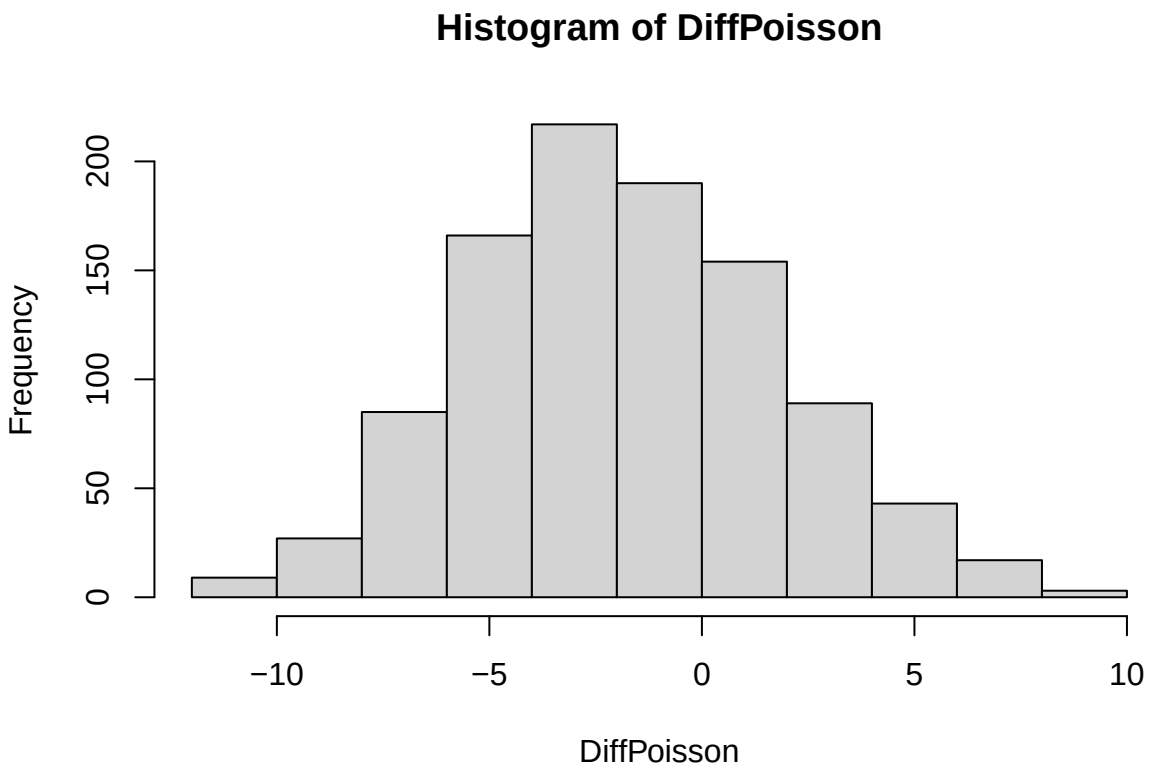
```
## ''6e Refer to Maryland, ''6c Refer to Holy Cross

set.seed(1424)
Theta6e <- rgamma(num.samp, 12 + 80, 2 + 12)
Poisson6e <- rpois(num.samp, Theta6e)

DiffTheta <- Theta6c - Theta6e; DiffPoisson <- Poisson6c - Poisson6e

hist(DiffTheta); hist(DiffPoisson)
```


Histogram of DiffTheta
Frequency
0
50
100
150
200
−5
−4
−3
−2
−1
0
1
DiffTheta




```
quantile(DiffTheta, c(0.025, 0.975)); quantile(DiffPoisson, c(0.025, 0.975))
```

```
##       2.5%      97.5%
## -3.3243691  0.1433257
```

```
##  2.5% 97.5%
##    -8     6
```

Since both distributions for the difference are skewed right, with the concentration of values less than 0. This means that, generally, Maryland is a team with a greater three-point shooting percentage (rate) than Holy Cross. However, as the 95% credible intervals show, there are occasions where HC might over take Maryland in terms of three-point shooting because 0 is included in both intervals.