NAME: Kaifa Lu UFID: 5205-0501
STA6348 – Bayesian Analysis for Machine Learning and Uncertainty Quantification

Problem 1: Finding Bayes estimators numerically.

Since let Y ~ Bin(n, θ), n = 10, observed y = 3, and the prior is θ ~ Beta(1,1), the posterior is θ|y ~ Beta(4,8). Then we implement the posterior expected loss functions using numerical integration of the loss function with respect to the posterior to define objective functions, and minimize those numerically.

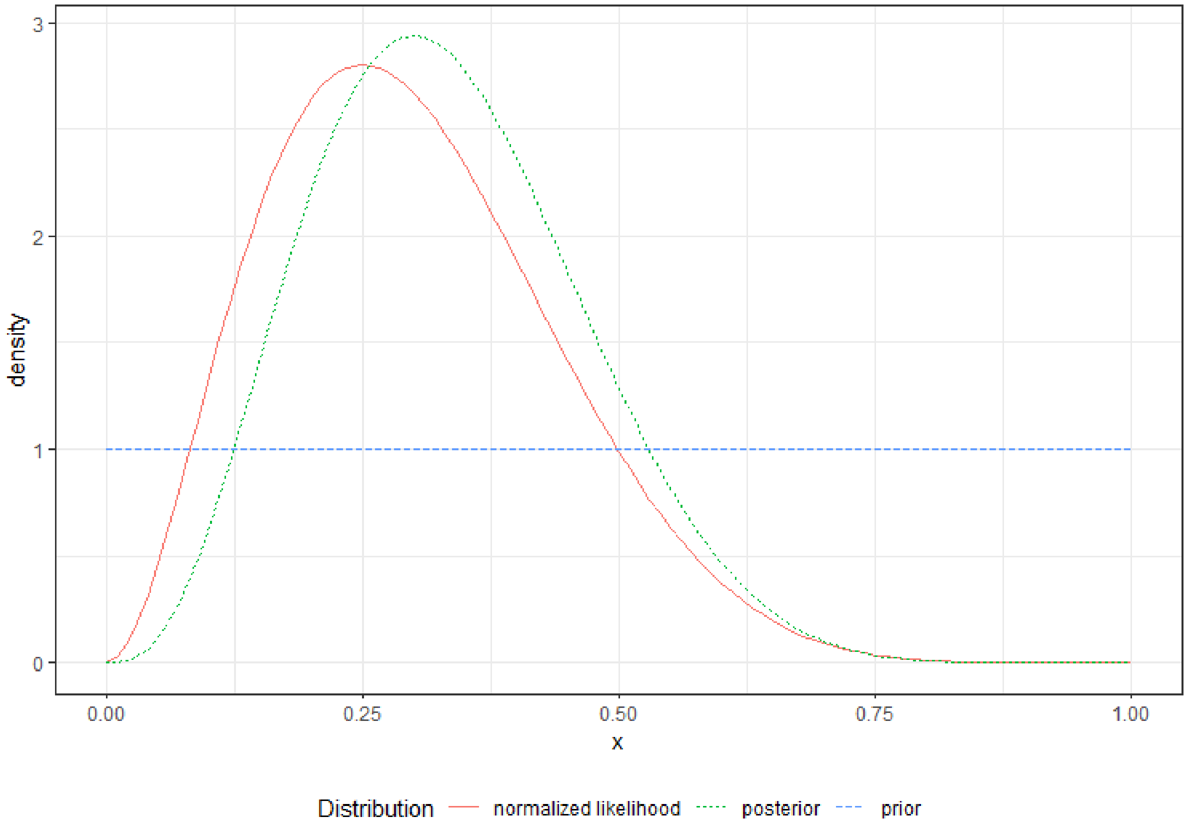
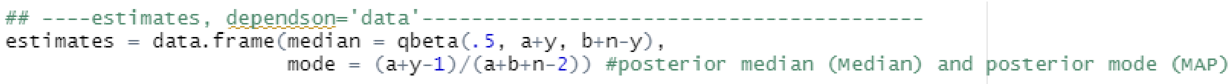


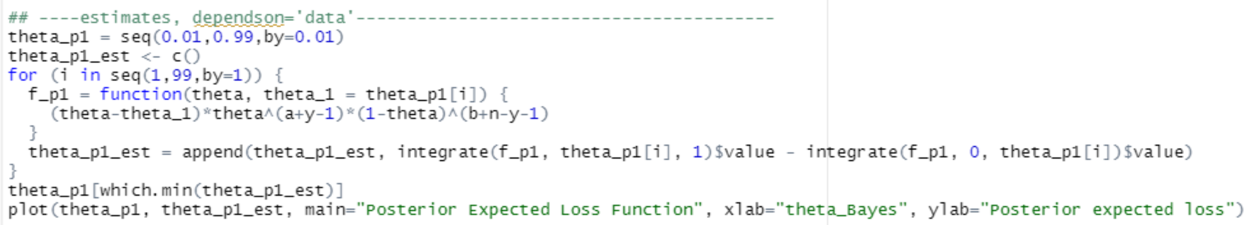
Figure 1: Plots of prior, normalized likelihood, and posterior

1.1 When the loss function is $L(\theta,\hat{\theta}) = \left|\theta - \hat{\theta}\right|$, the best solution of Bayes estimate is the posterior median, and we can use the following R codes to obtain it: $\hat{\theta} = 0.3238045$.



```
## ----estimates, dependson='data'-----------------------------------------
estimates = data.frame(median = qbeta(.5, a+y, b+n-y),
                       mode = (a+y-1)/(a+b+n-2)) #posterior median (Median) and posterior mode (MAP)
```

Then we attempt to find the best solution numerically using the following codes:



```
## ----estimates, dependson='data'----------------------------------------
theta_p1 = seq(0.01,0.99,by=0.01)
theta_p1_est <- c()
for (i in seq(1,99,by=1)) {
  f_p1 = function(theta, theta_1 = theta_p1[i]) {
    (theta-theta_1)*theta^(a+y-1)*(1-theta)^(b+n-y-1)
  }
  theta_p1_est = append(theta_p1_est, integrate(f_p1, theta_p1[i], 1)$value - integrate(f_p1, 0, theta_p1[i])$value)
}
theta_p1[which.min(theta_p1_est)]
plot(theta_p1, theta_p1_est, main="Posterior Expected Loss Function", xlab="theta_Bayes", ylab="Posterior expected loss")
```

Specifically, we employ the methods of grid search and numerical integration to return the Bayes estimate of $\hat{\theta}$ that corresponds to the minimum of the posterior expected loss function from an array of 0.01:0.99. The form of the posterior expected loss function is shown below:

$$\hat{\theta}_{Bayes} = \underset{\hat{\theta}}{\operatorname{argmin}} \left\{ \int_0^{\hat{\theta}} (\hat{\theta} - \theta) p(\theta|y) d\theta + \int_{\hat{\theta}}^1 (\theta - \hat{\theta}) p(\theta|y) d\theta \right\}$$

Here, we plot the values of posterior expected loss function when $\hat{\theta}$ varies from 0.01 to 0.99 as follows. We find that when $\hat{\theta} = 0.32$, the posterior expected loss function reachhes the minimum. Also, we can see that my best solution found numerically ($\hat{\theta} = 0.32$) is very close to the posterior median ($\hat{\theta} = 0.3238045$).

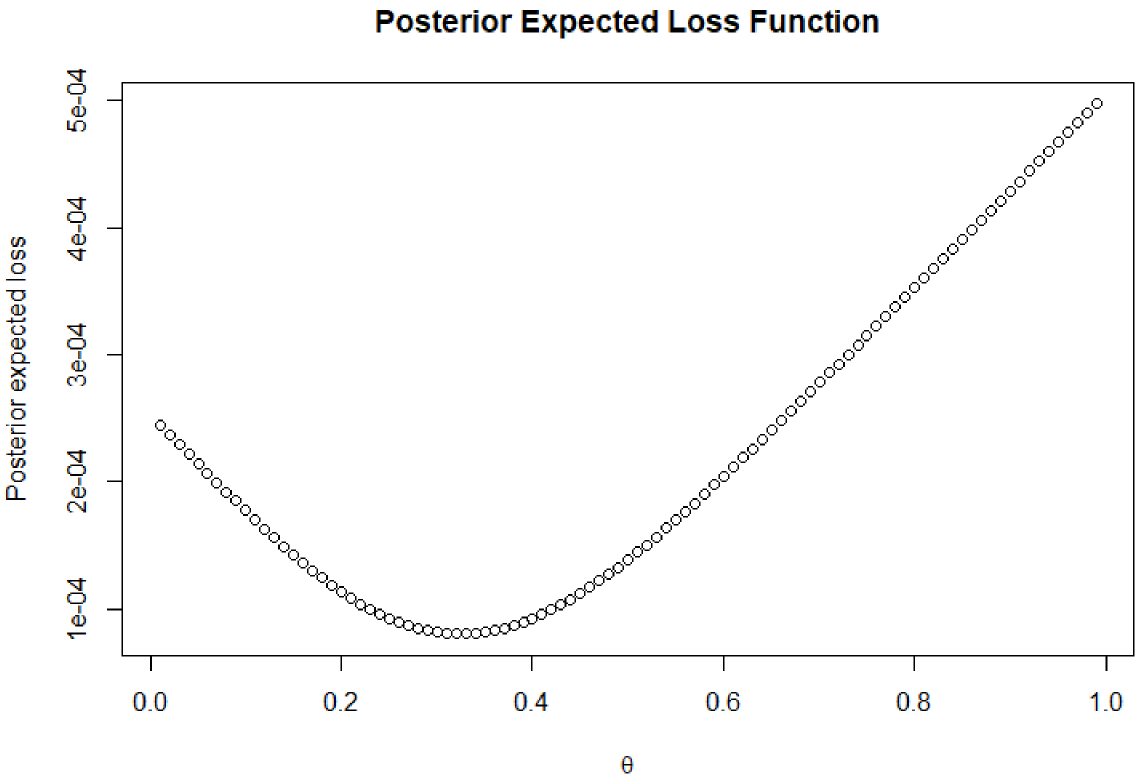
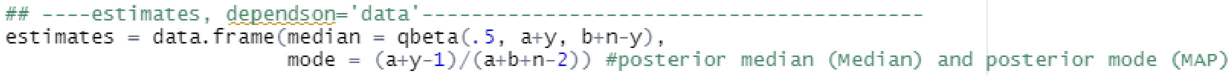


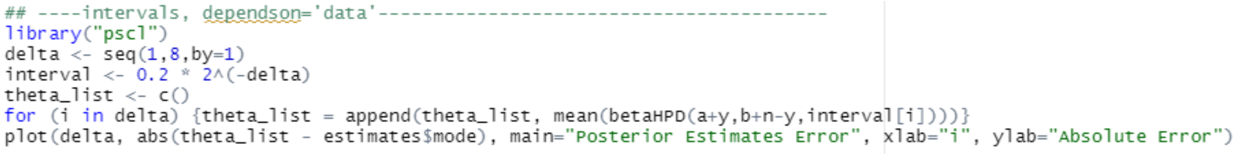
Figure 2: The values of posterior expected loss function when $\hat{\theta}$ varies from 0.01 to 0.99

1.2 When the loss function is $L_\delta(\theta, \hat{\theta}) = -\mathbb{I}\left(\left|\theta - \hat{\theta}\right| < \delta\right)$, the best solution of Bayes estimate is the posterior mode, and we can use the following R codes to obtain it: $\hat{\theta}_{MAP} = 0.3$.



```
## ----estimates, dependson='data'-----------------------------------------
estimates = data.frame(median = qbeta(.5, a+y, b+n-y),
                       mode = (a+y-1)/(a+b+n-2)) #posterior median (Median) and posterior mode (MAP)
```

Then we attempt to find the best solution corresponding to different δ ($\delta = 0.1 * 2^{-i}, for\ i = 1,2,\cdots,8$) by using the following codes:



```
## ----intervals, dependson='data'----------------------------------------
library("pscl")
delta <- seq(1,8,by=1)
interval <- 0.2 * 2^(-delta)
theta_list <- c()
for (i in delta) {theta_list = append(theta_list, mean(betaHPD(a+y,b+n-y,interval[i])))}
plot(delta, abs(theta_list - estimates$mode), main="Posterior Estimates Error", xlab="i", ylab="Absolute Error")
```

Specifically, we employ the method of Highest Posterior Density (HPD) to estimate the shortest interval for $\hat{\theta}$ and then average the lower bound and upper bound to return the Bayes estimate of $\hat{\theta}$. This is feasible due to the form of posterior expected loss function as follows:

$$\hat{\theta}_{Bayes} = \underset{\hat{\theta}}{\operatorname{argmax}} \left\{ \int_{\hat{\theta}-\delta}^{\hat{\theta}+\delta} p(\theta|y) d\theta \right\}$$

Here, when $\delta = 0.1 * 2^{-i}, for\ i = 1,2,\cdots,8$, we separately obtain the Bayes estimate of $\hat{\theta}$ and then plot $\left|\hat{\theta}_{MAP} - \hat{\theta}_\delta\right|$ as a function of i as follows. We can find that when i increases from 1 to 8, the δ gradually decreases, and the Bayes estimate of $\hat{\theta}$ gradually approaches $\hat{\theta}_{MAP}$. Actually, when $i = 5$, the Bayes estimate of $\hat{\theta}$ is equal to $\hat{\theta}_{MAP}$.

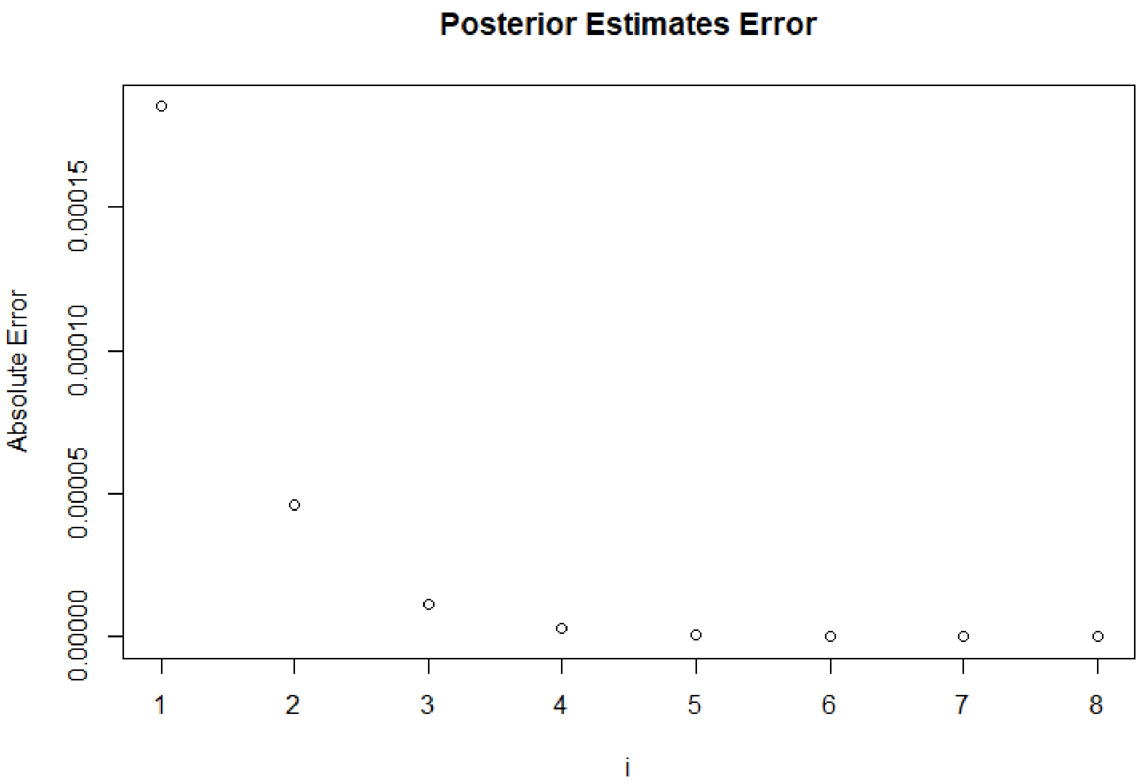
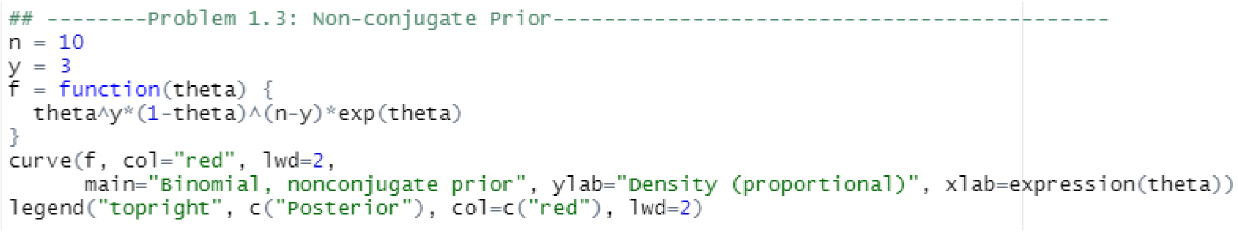


Figure 3: Plot of $\left|\hat{\theta}_{MAP} - \hat{\theta}_{\delta}\right|$ as a function of i when i varies from 1 to 8

1.3 Suppose the prior $p(\theta) \propto e^{\theta}$, then the posterior $p(\theta|y) \propto f(\theta) = \theta^{y}(1-\theta)^{n-y}e^{\theta}$, which is not a known distribution. The plot of $f(\theta)$ is shown below:



```
## --------Problem 1.3: Non-conjugate Prior---------------------------------------------
n = 10
y = 3
f = function(theta) {
  theta^y*(1-theta)^(n-y)*exp(theta)
}
curve(f, col="red", lwd=2,
      main="Binomial, nonconjugate prior", ylab="Density (proportional)", xlab=expression(theta))
legend("topright", c("Posterior"), col=c("red"), lwd=2)
```

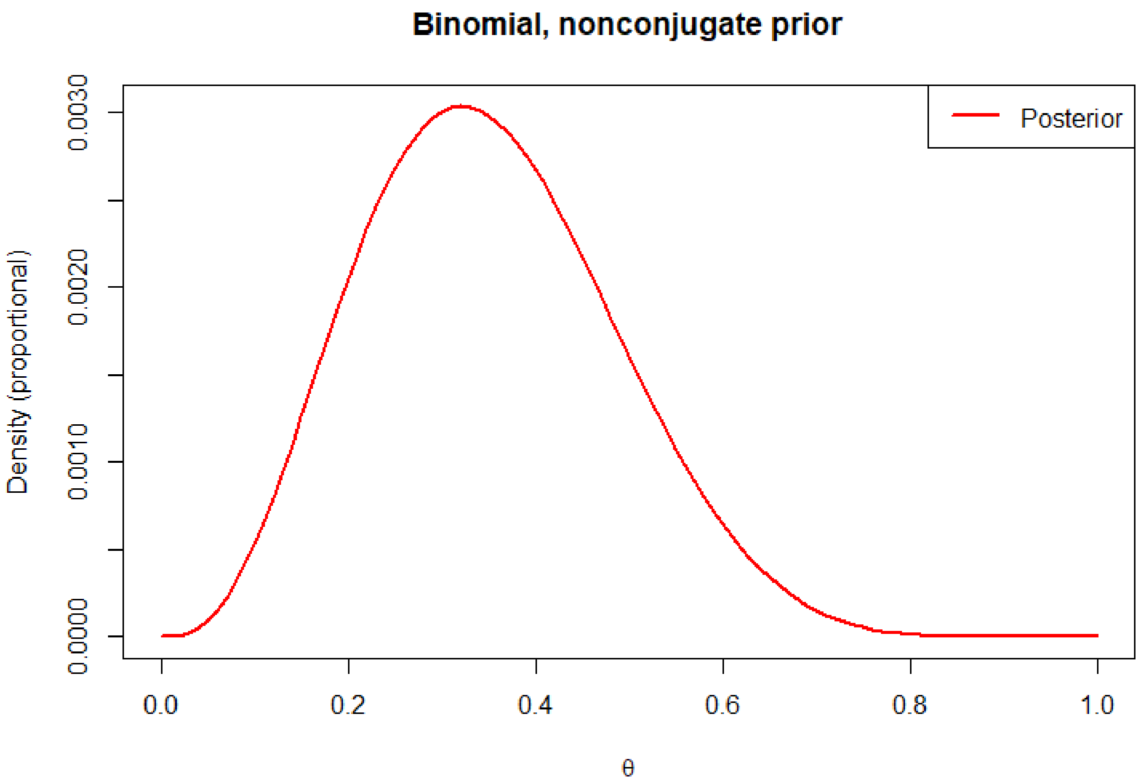
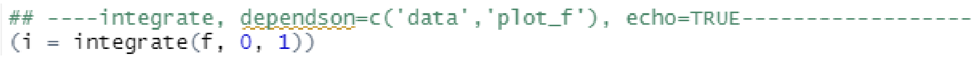


Figure 4: Plot of $f(\theta)$

Then we calculate the normalizing constant $i = \int_0^1 f(\theta)d\theta = 0.0010665$ by using the following code, so that $p(\theta|y) = f(\theta)/i$ as shown below:



```
## ----integrate, dependson=c('data','plot_f'), echo=TRUE------------------
(i = integrate(f, 0, 1))
```

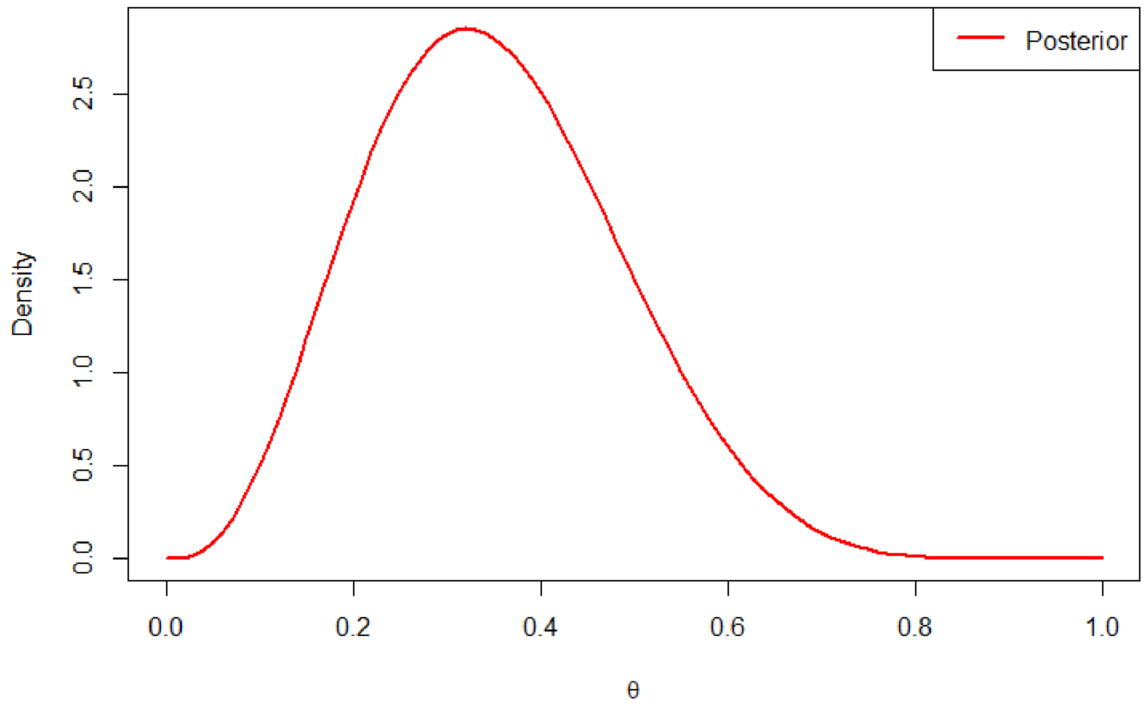
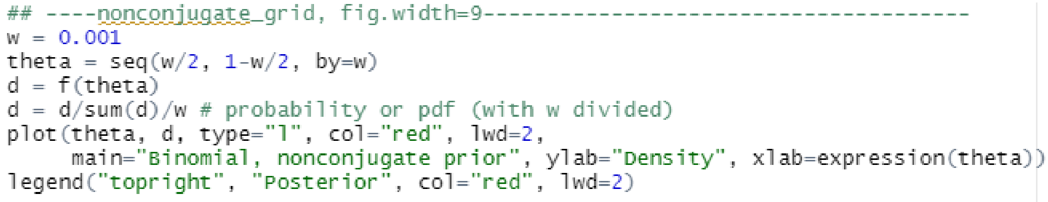


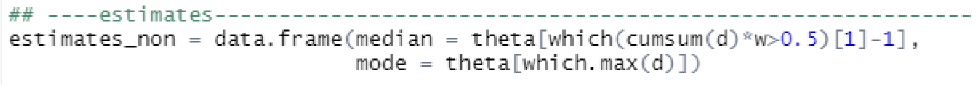
Figure 5: Plot of posterior probability $p(\theta|y)$

Then we use the grid spacing method to discretize the posterior probability $p(\theta|y)$ by using the following code:



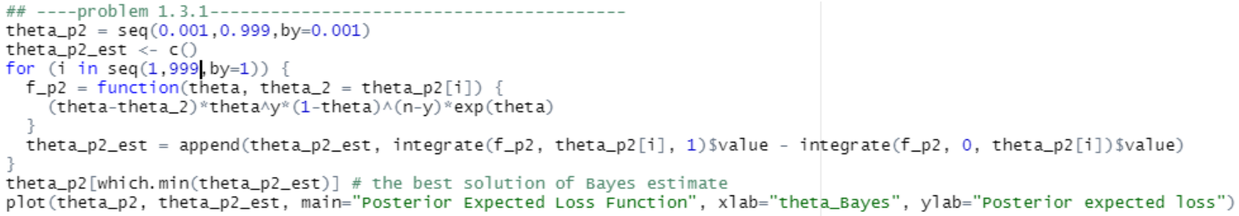
```
## ----nonconjugate_grid, fig.width=9---------------------------------------
w = 0.001
theta = seq(w/2, 1-w/2, by=w)
d = f(theta)
d = d/sum(d)/w # probability or pdf (with w divided)
plot(theta, d, type="l", col="red", lwd=2,
     main="Binomial, nonconjugate prior", ylab="Density", xlab=expression(theta))
legend("topright", "Posterior", col="red", lwd=2)
```

Then we obtain the posterior median $\hat{\theta} = 0.3415$ and posterior mode $\hat{\theta}_{MAP} = 0.3215$ by using the following code:



```
## ----estimates-------------------------------------------------------------
estimates_non = data.frame(median = theta[which(cumsum(d)*w>0.5)[1]-1],
                           mode = theta[which.max(d)])
```

For Problem 1.1, here, we use the similar method as **1.1**, and find that when $\hat{\theta} = 0.343$, the posterior expected loss function reaches the minimum. We can see that my best solution found numerically ($\hat{\theta} = 0.343$) is very close to the posterior median ($\hat{\theta} = 0.3415$).



```
## ----problem 1.3.1----------------------------------------
theta_p2 = seq(0.001,0.999,by=0.001)
theta_p2_est <- c()
for (i in seq(1,999,by=1)) {
  f_p2 = function(theta, theta_2 = theta_p2[i]) {
    (theta-theta_2)*theta^y*(1-theta)^(n-y)*exp(theta)
  }
  theta_p2_est = append(theta_p2_est, integrate(f_p2, theta_p2[i], 1)$value - integrate(f_p2, 0, theta_p2[i])$value)
}
theta_p2[which.min(theta_p2_est)] # the best solution of Bayes estimate
plot(theta_p2, theta_p2_est, main="Posterior Expected Loss Function", xlab="theta_Bayes", ylab="Posterior expected loss")
```

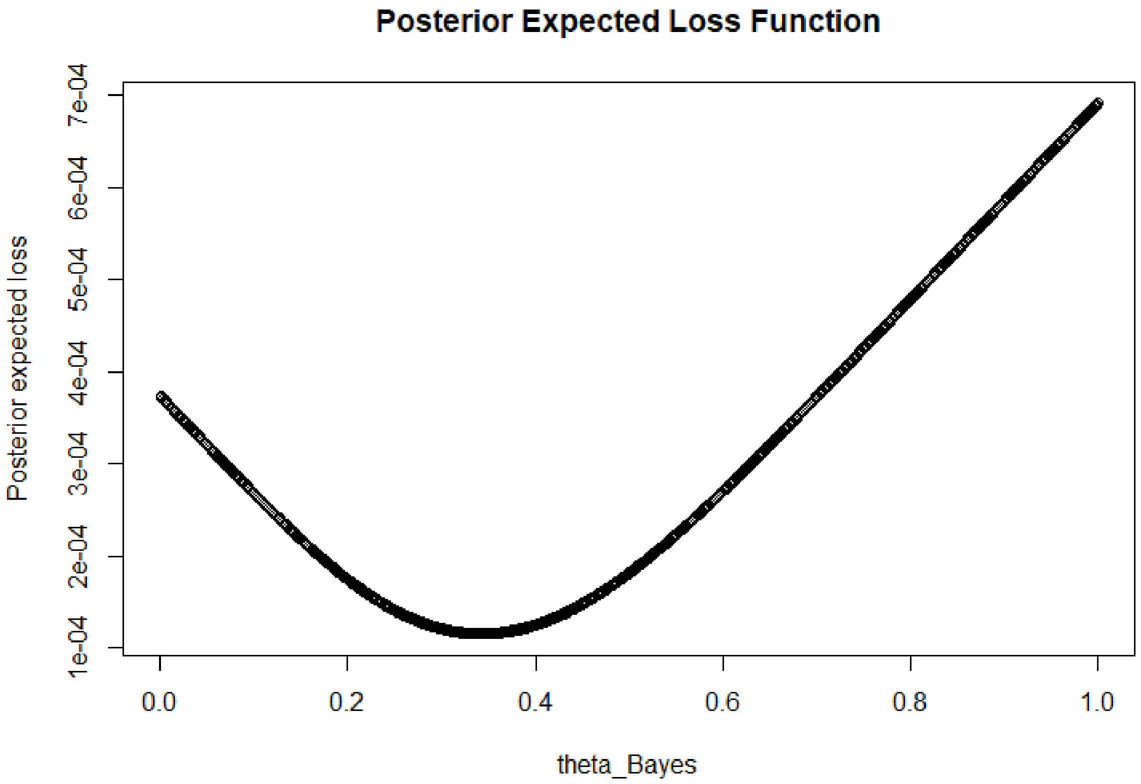


Figure 6: The values of posterior expected loss function when $\hat{\theta}$ varies from 0.001 to 0.999

For Problem 1.2, here, we use the similar method as **1.2**, and when $\delta = 0.1 * 2^{-i}, for\ i = 1,2,\cdots,8$, we separately obtain the Bayes estimate of $\hat{\theta}$ and then plot $\left|\hat{\theta}_{MAP} - \hat{\theta}_{\delta}\right|$ as a function of i as follows. We can find that when i increases from 1 to 8, the δ gradually decreases, and the Bayes estimate of $\hat{\theta}$ is always equal to $\hat{\theta}_{MAP} = 0.3215$.

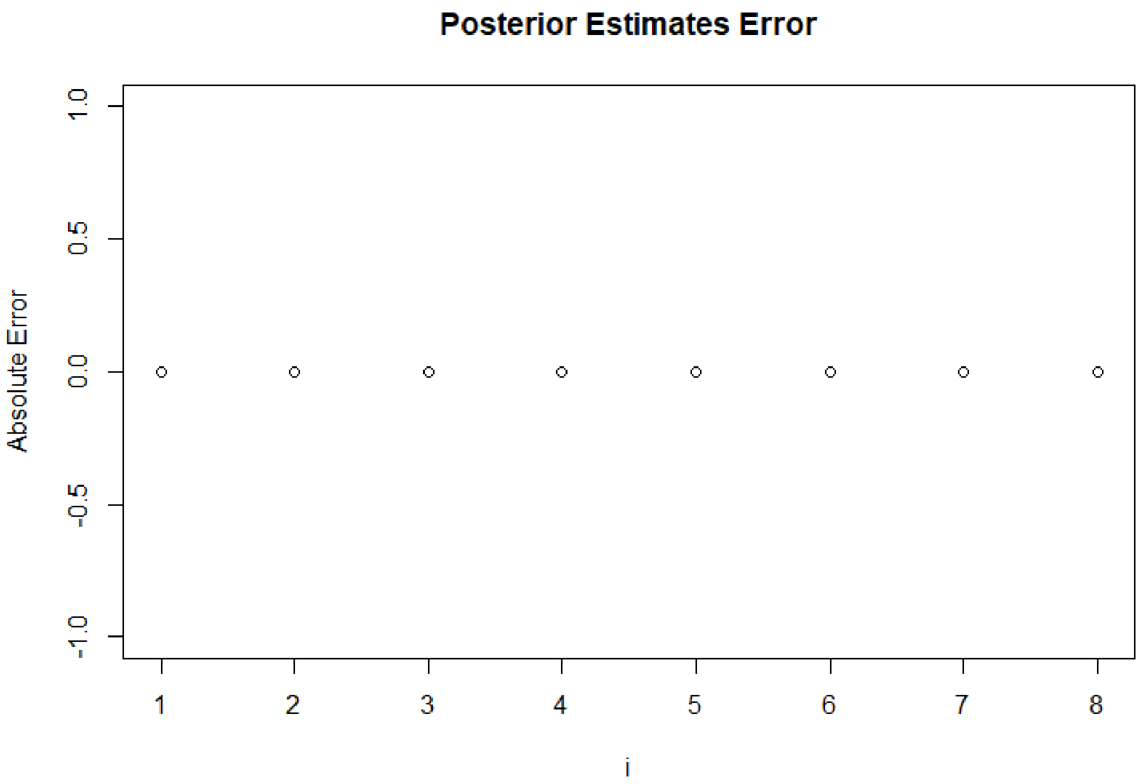
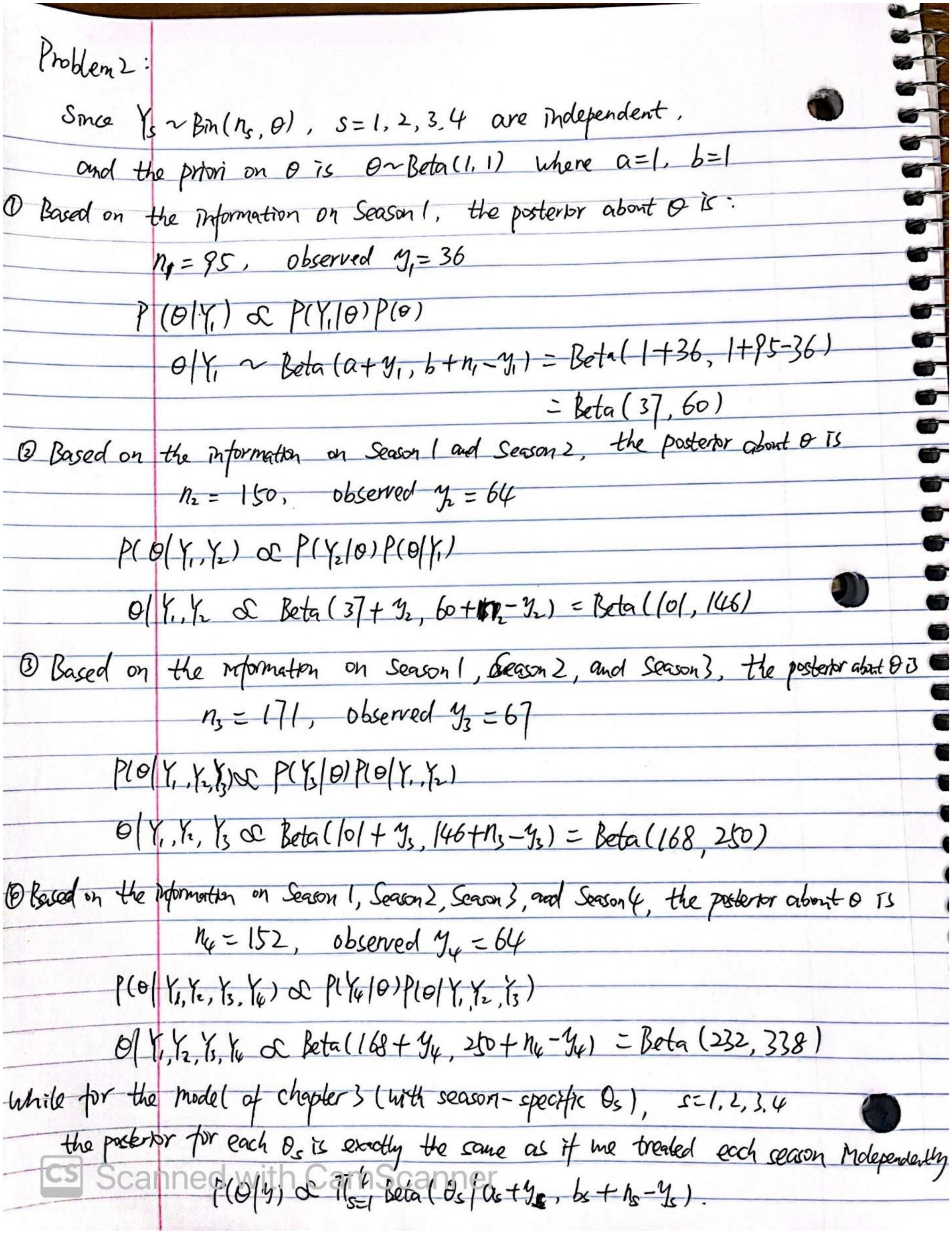


Figure 7: Plot of $\left|\hat{\theta}_{MAP} - \hat{\theta}_{\delta}\right|$ as a function of i when i varies from 1 to 8

Problem 2: Sequential Bayesian learning for independent data.



Problem 2:

Since $Y_s \sim Bin(n_s, \theta)$, $s=1,2,3,4$ are independent,

and the priori on θ is $\theta \sim Beta(1,1)$ where $a=1$, $b=1$

① Based on the information on Season 1, the posterior about θ is:

$n_1 = 95$, observed $y_1 = 36$

$P(\theta|Y_1) \propto P(Y_1|\theta)P(\theta)$

$\theta|Y_1 \sim Beta(a+y_1, b+n_1-y_1) = Beta(1+36, 1+95-36)$

$= Beta(37, 60)$

② Based on the information on Season 1 and Season 2, the posterior about θ is

$n_2 = 150$, observed $y_2 = 64$

$P(\theta|Y_1,Y_2) \propto P(Y_2|\theta)P(\theta|Y_1)$

$\theta|Y_1,Y_2 \propto Beta(37+y_2, 60+n_2-y_2) = Beta(101, 146)$

③ Based on the information on Season 1, Season 2, and Season 3, the posterior about θ is

$n_3 = 171$, observed $y_3 = 67$

$P(\theta|Y_1,Y_2,Y_3) \propto P(Y_3|\theta)P(\theta|Y_1,Y_2)$

$\theta|Y_1,Y_2,Y_3 \propto Beta(101+y_3, 146+n_3-y_3) = Beta(168, 250)$

④ Based on the information on Season 1, Season 2, Season 3, and Season 4, the posterior about θ is

$n_4 = 152$, observed $y_4 = 64$

$P(\theta|Y_1,Y_2,Y_3,Y_4) \propto P(Y_4|\theta)P(\theta|Y_1,Y_2,Y_3)$

$\theta|Y_1,Y_2,Y_3,Y_4 \propto Beta(168+y_4, 250+n_4-y_4) = Beta(232, 338)$

while for the model of chapter 3 (with season-specific θ_s), $s=1,2,3,4$

the posterior for each θ_s is exactly the same as if we treated each season independently

$P(\theta|y) \propto \prod_{s=1}^{4} Beta(\theta_s|a_s+y_s, b_s+n_s-y_s)$.

The comparison between the posterior densities for θ conditional on $y_1, \cdots, y_s$ for $s = 1,2,\cdots,4$ with the solution for the joint posterior model of Chapter 3 is separately shown in Figure 8 and 9. The main difference lies in: (1) The joint posterior model for each θ_s is exactly the same as if we treated each season independently. In other words, information about the other three seasons does not contribute to the posterior θ_s for the season. (2) In contrast, the posterior density curve for θ conditional on $y_1, \cdots, y_s$ for $s = 1,2,\cdots,4$ is more and more concentrated on the mean value when gradually adding information in Season 1, 2, 3, 4.

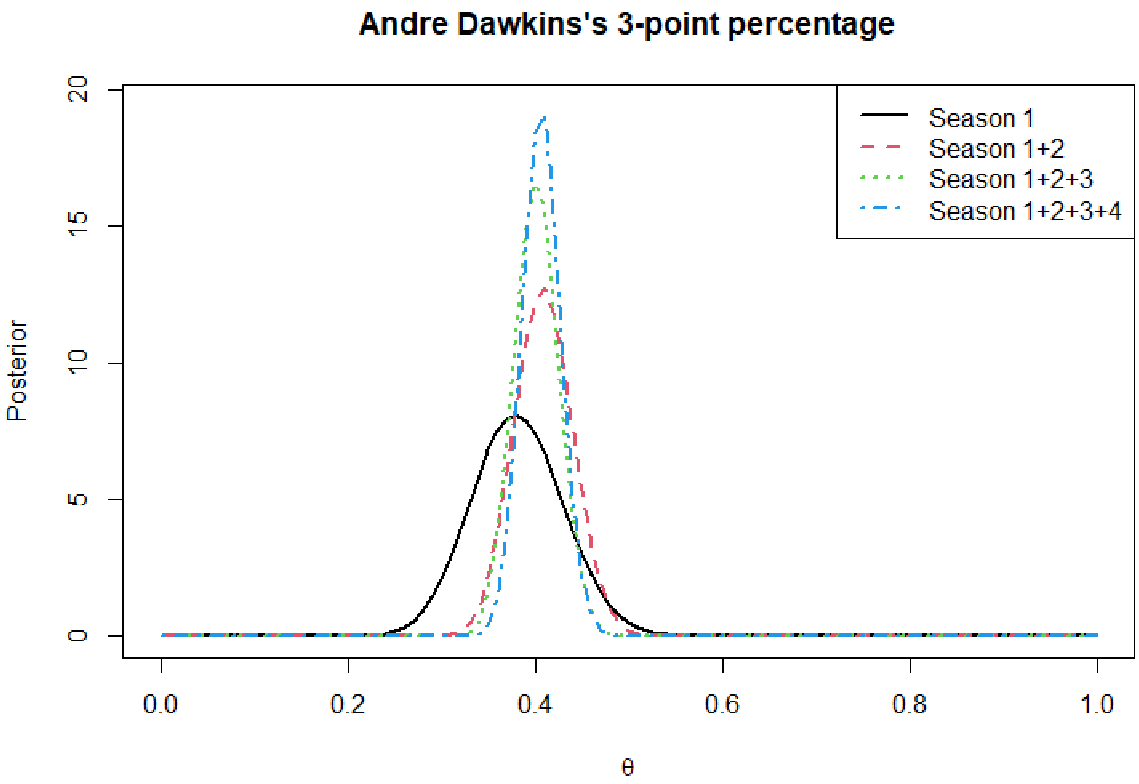


Figure 8: Posterior densities for θ conditional on $y_1, \cdots, y_s$ for $s = 1,2,\cdots,4$

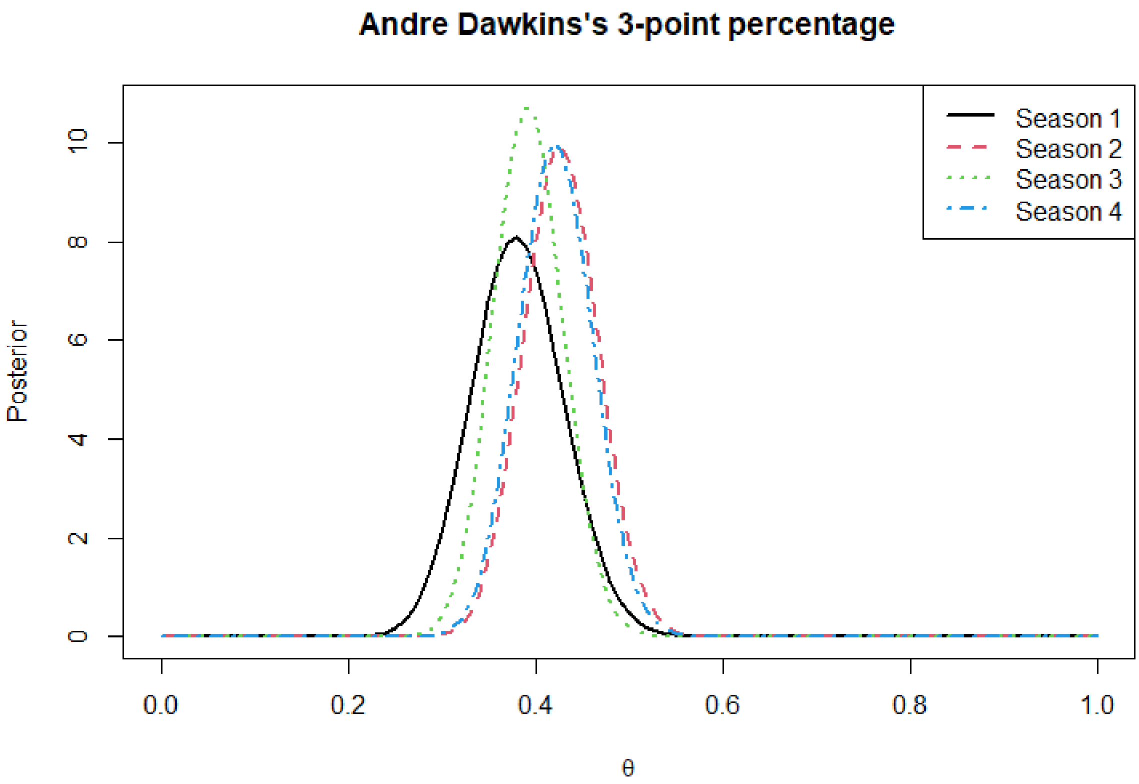


Figure 9: The joint posterior model of Chapter 3

Problem 3: Mixture of conjugate priors for coin flips.

3.1 For one student flipping a coin 5 times and getting 4 heads and 1 tail, we can assume the priori about the probability of heads θ as $\theta \sim Beta(5,2)$; Similarly, for the other student flipping a coin 5 times and getting 1 head and 4 tails, we can assume the priori about the probability of heads θ as $\theta \sim Beta(2,5)$. Since each student's performance is "equally credible", the mixture of conjugate priors about θ for the coin filp experiment should be $\theta \sim 0.5Beta(5,2) + 0.5Beta(2,5)$. Then we plot each student's likelihood/prior and the equal weight mixture prior as follows:

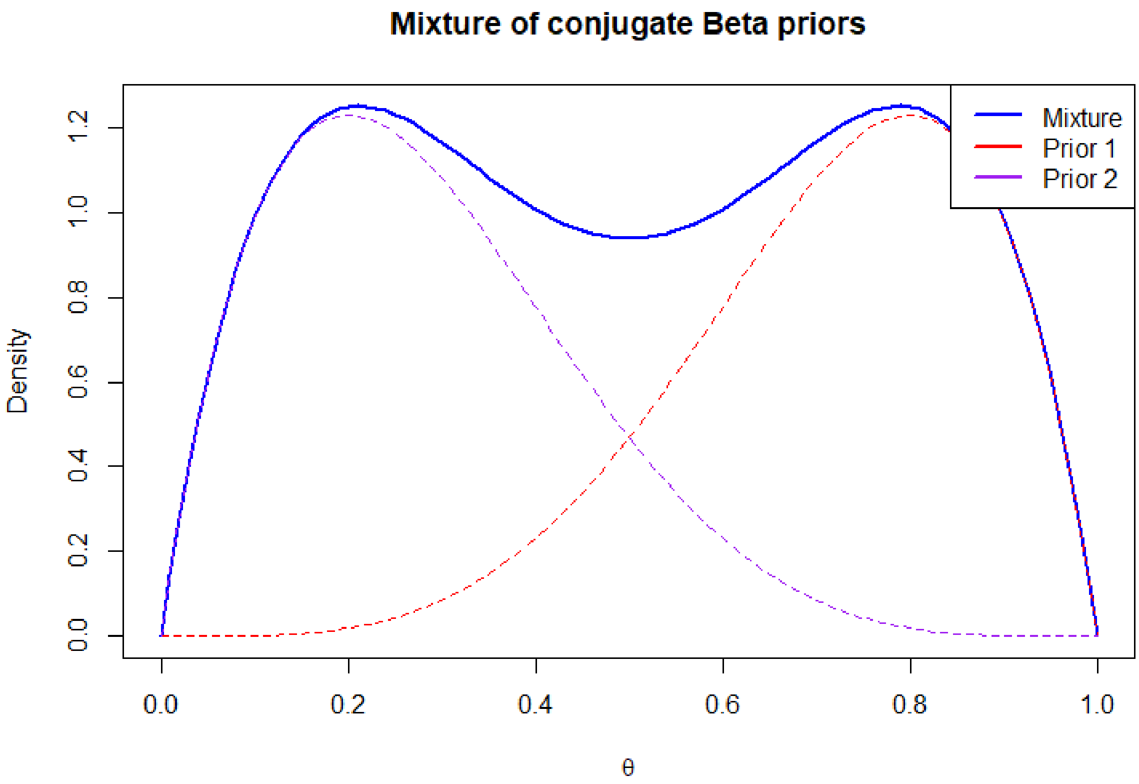
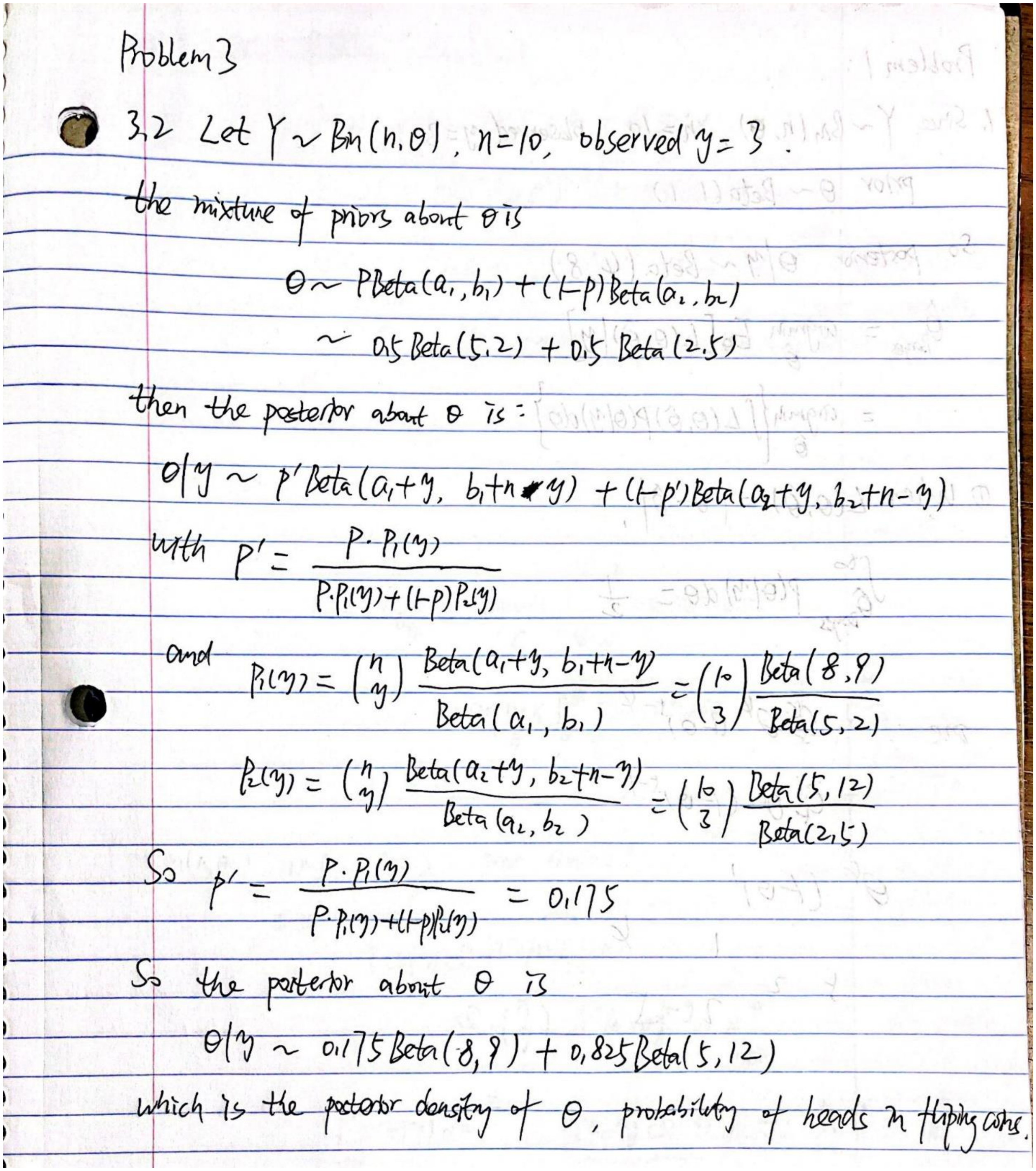


Figure 10: Plot of each student's likelihood/prior and the equal weight mixture prior

3.2 Based on the mixture of conjugate Beta priors about $\theta \sim 0.5Beta(5,2) + 0.5Beta(2,5)$, we obtain the posterior density for the experiment $Y \sim Bin(n, \theta)$ where $n = 10$ and observed $y = 3$ by using the following steps as shown below:



Problem 3

3.2 Let $Y \sim Bin(n, \theta)$, $n=10$, observed $y=3$.

the mixture of priors about θ is

$$\theta \sim P Beta(a_1, b_1) + (1-p) Beta(a_2, b_2)$$
$$\sim 0.5 Beta(5,2) + 0.5 Beta(2,5)$$

then the posterior about θ is:

$$\theta|y \sim p' Beta(a_1+y,\ b_1+n-y) + (1-p') Beta(a_2+y,\ b_2+n-y)$$

with $p' = \dfrac{P \cdot P_1(y)}{P \cdot P_1(y) + (1-p) P_2(y)}$

and $P_1(y) = \binom{n}{y} \dfrac{Beta(a_1+y,\ b_1+n-y)}{Beta(a_1, b_1)} = \binom{10}{3} \dfrac{Beta(8,9)}{Beta(5,2)}$

$P_2(y) = \binom{n}{y} \dfrac{Beta(a_2+y,\ b_2+n-y)}{Beta(a_2, b_2)} = \binom{10}{3} \dfrac{Beta(5,12)}{Beta(2,5)}$

So $p' = \dfrac{P \cdot P_1(y)}{P \cdot P_1(y) + (1-p) P_2(y)} = 0.175$

So the posterior about θ is

$$\theta|y \sim 0.175 Beta(8,9) + 0.825 Beta(5,12)$$

which is the posterior density of θ, probability of heads in flipping coins.

Then we plot the posterior alongside with the mixture of conjugate priors in Figure 11. The prior curve is composed of two equal weighted beta priors, which is expressed by $\theta \sim 0.5 Beta(5,2) + 0.5 Beta(2,5)$, while the posterior is also composed of two weighted beta posteriors, but not equal, which is denoted by $\theta|y \sim 0.175 Beta(8,9) + 0.825 Beta(5,12)$.

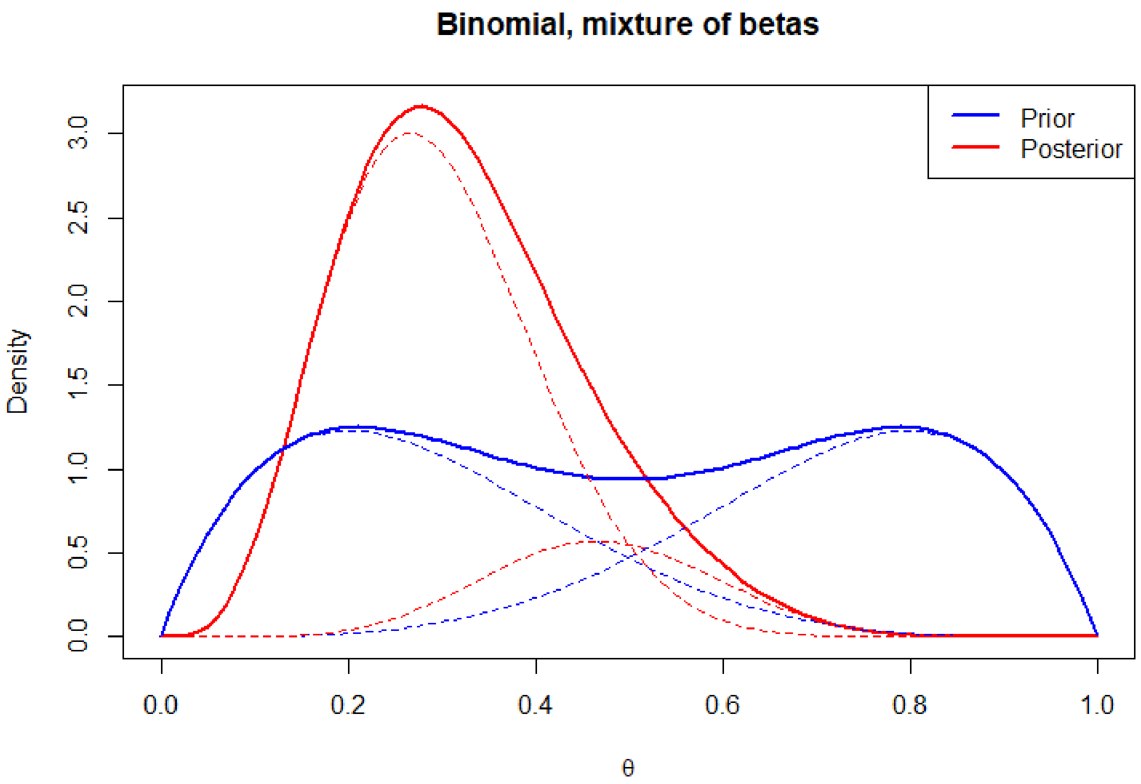


Figure 11: Plot of the posterior alongside with the prior

Problem 4: Shortest-length credible intervals.

We need to write a function to compute the shortest-length $100(1-\alpha)\%$ credible interval for a posterior distribution with an (unnormalized) bounded pdf f. To simplify this task, we make the following assumptions about f:

(1) f is continuous so that we can directly and easily use "integrate" function to calculate the probability of parameter θ falling into the interval (l,u): $\int_l^u p(\theta|y)d\theta$

(2) Support of f is (a,b) so that we can restrict the searching space of credible interval within (a,b).

(3) f is unimodal so that we only need to focus on one variable, which can significantly reduce the complexity of the task.

Based on these assumptions, we design the following algorithm about a given posterior distribution fucntion f to return the shortest-length $100(1-\alpha)\%$ credible interval:

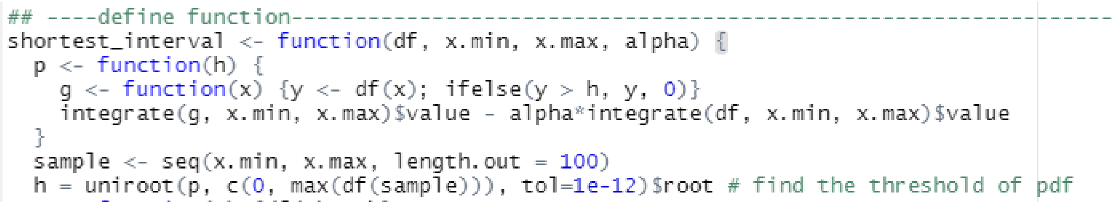
Step 1: Given any posterior distribution function f, all highest posterior density sets are of the form $\{\theta|f(\theta|y)\geq h\}$. The total probability of any such set should be

$$p_f(h)=\int I(f(\theta|y)\geq h)f(\theta|y)\,d\theta$$

Therefore, obtaining the $100(1-\alpha)\%$ highest probability density set is a matter of solving the following equation:

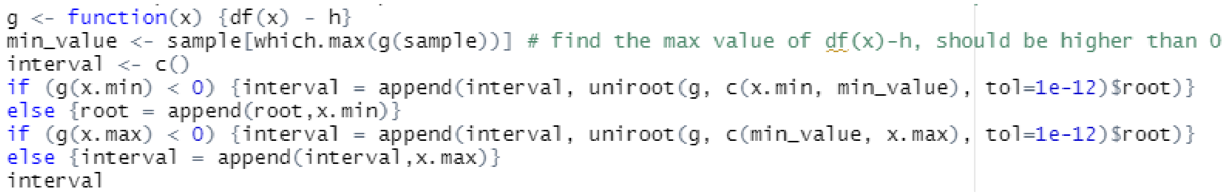
$$p_f(h)-(1-\alpha)=0$$

The goal of the first step is to find the root h of the above equation by using the following codes:



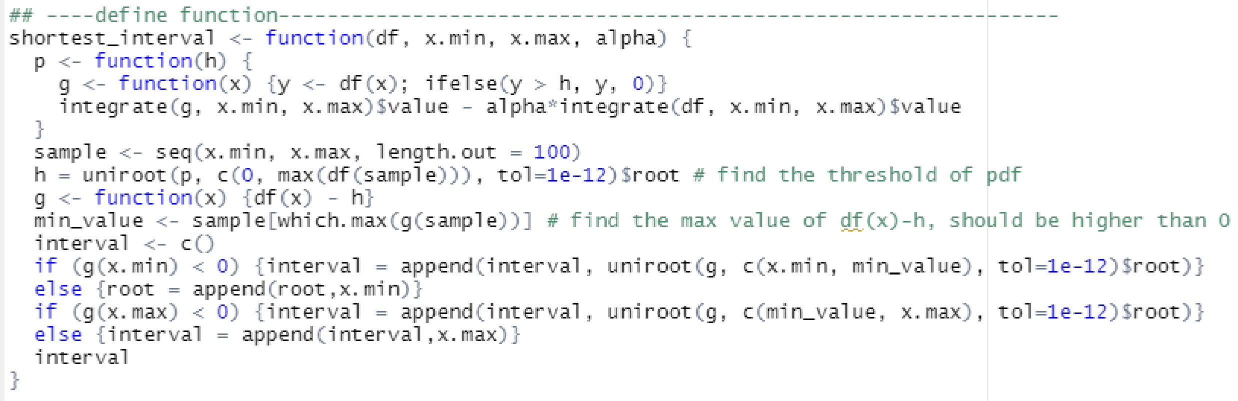
```
## ----define function--------------------------------------------------------------
shortest_interval <- function(df, x.min, x.max, alpha) {
  p <- function(h) {
    g <- function(x) {y <- df(x); ifelse(y > h, y, 0)}
    integrate(g, x.min, x.max)$value - alpha*integrate(df, x.min, x.max)$value
  }
  sample <- seq(x.min, x.max, length.out = 100)
  h = uniroot(p, c(0, max(df(sample))), tol=1e-12)$root # find the threshold of pdf
```

Step 2: Based on the derived h from **Step 1**, we use the equation $f(\theta|y)=h$ to obtain the lower bound and upper bound of θ as the shortest credible interval by using the following codes:



```
  g <- function(x) {df(x) - h}
  min_value <- sample[which.max(g(sample))] # find the max value of df(x)-h, should be higher than 0
  interval <- c()
  if (g(x.min) < 0) {interval = append(interval, uniroot(g, c(x.min, min_value), tol=1e-12)$root)}
  else {root = append(root,x.min)}
  if (g(x.max) < 0) {interval = append(interval, uniroot(g, c(min_value, x.max), tol=1e-12)$root)}
  else {interval = append(interval,x.max)}
  interval
```

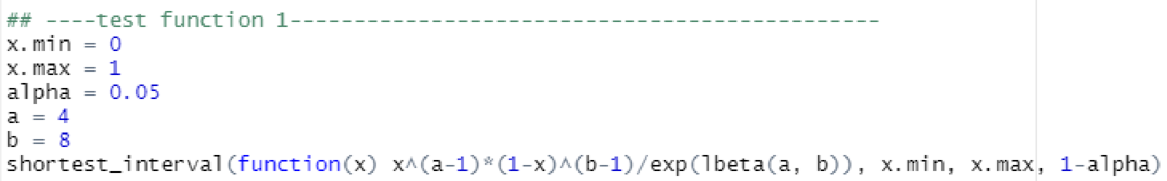
Therefore, the function used to compute the shortest-length $100(1-\alpha)\%$ credible interval for a posterior distribution can be written as follows. As expected, this function takes as inputs: the unnormalized pdf f, the support (a,b) of f, and $\alpha \epsilon (0,1)$, and returns the lower and upper bounds of the credible interval.



```
## ----define function------------------------------------------------------------------
shortest_interval <- function(df, x.min, x.max, alpha) {
  p <- function(h) {
    g <- function(x) {y <- df(x); ifelse(y > h, y, 0)}
    integrate(g, x.min, x.max)$value - alpha*integrate(df, x.min, x.max)$value
  }
  sample <- seq(x.min, x.max, length.out = 100)
  h = uniroot(p, c(0, max(df(sample))), tol=1e-12)$root # find the threshold of pdf
  g <- function(x) {df(x) - h}
  min_value <- sample[which.max(g(sample))] # find the max value of df(x)-h, should be higher than 0
  interval <- c()
  if (g(x.min) < 0) {interval = append(interval, uniroot(g, c(x.min, min_value), tol=1e-12)$root)}
  else {root = append(root,x.min)}
  if (g(x.max) < 0) {interval = append(interval, uniroot(g, c(min_value, x.max), tol=1e-12)$root)}
  else {interval = append(interval,x.max)}
  interval
}
```

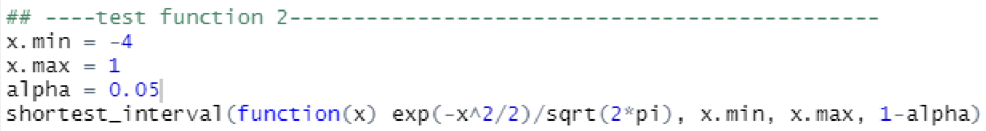
Finally, we test the function for $\alpha = 0.05$ on the following three function:

(1) Function 1: $Beta(4,8)$ pdf, the support (0,1). It returns the $100(1-\alpha)\%$ credible interval as [0.09336928, 0.58795883], which is almost the same as the result of the standard highest posterior density (HPD) function of Beta distribution in R.



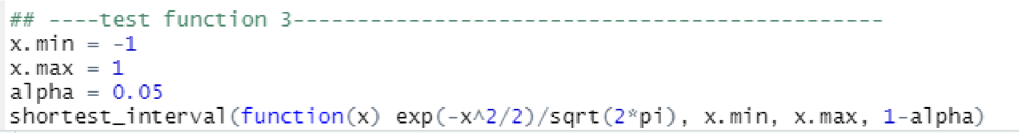
```
## ----test function 1-----------------------------------------------
x.min = 0
x.max = 1
alpha = 0.05
a = 4
b = 8
shortest_interval(function(x) x^(a-1)*(1-x)^(b-1)/exp(lbeta(a, b)), x.min, x.max, 1-alpha)
```

(2) Function 2: standard normal truncated to the interval (-4, 1).). It returns the $100(1-\alpha)\%$ credible interval as [-1.726744, 1.000000].



```
## ----test function 2-----------------------------------------------
x.min = -4
x.max = 1
alpha = 0.05
shortest_interval(function(x) exp(-x^2/2)/sqrt(2*pi), x.min, x.max, 1-alpha)
```

(3) Function 3: standard normal truncated to the interval (-1, 1).). It returns the $100(1-\alpha)\%$ credible interval as [-0.9317972, 0.9317972].



```
## ----test function 3---------------------------------------------
x.min = -1
x.max = 1
alpha = 0.05
shortest_interval(function(x) exp(-x^2/2)/sqrt(2*pi), x.min, x.max, 1-alpha)
```