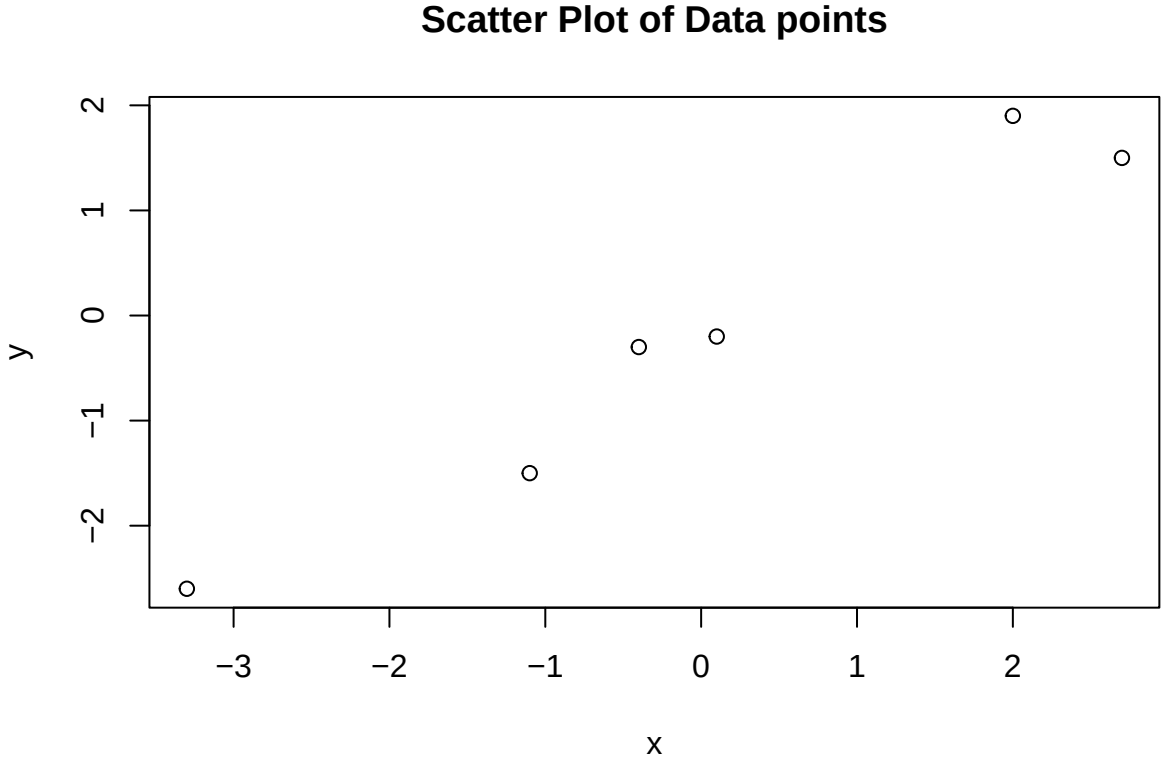
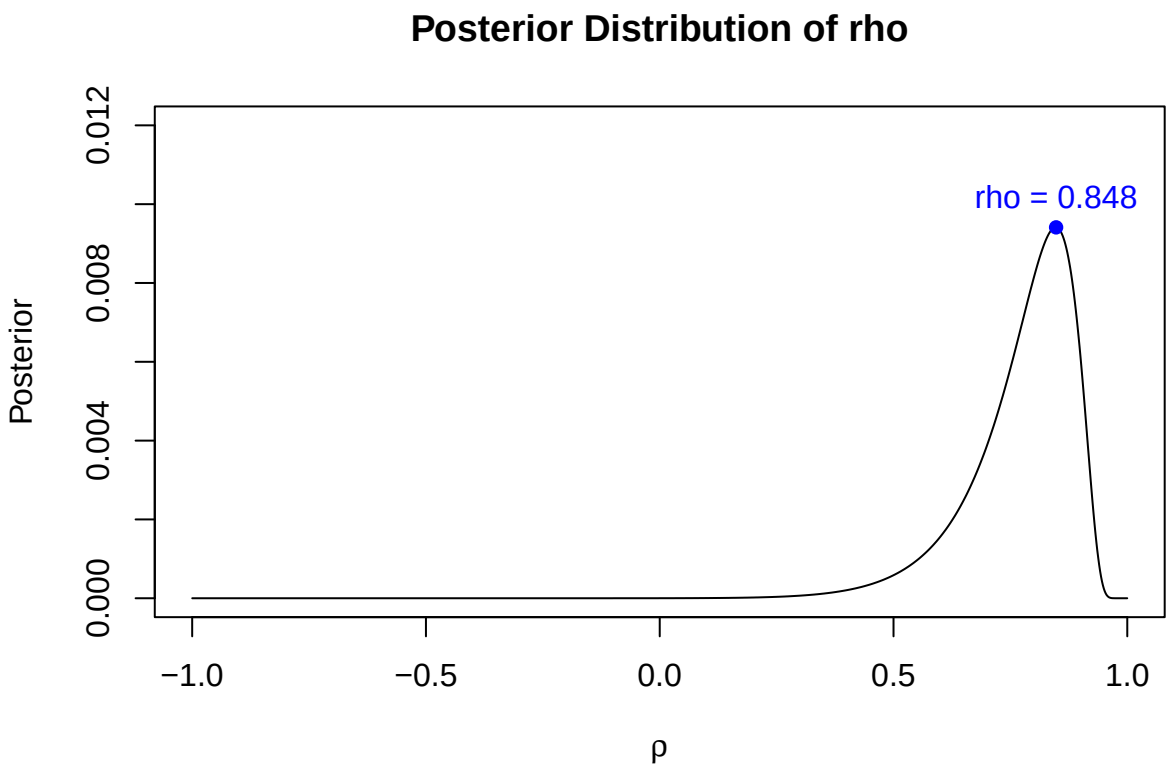
ST 440 Lab #2

Christian Nielsen

(1) Chapter 1, Problem 12

Assume that (X, Y) follow the bivariate normal distribution and that both X and Y have marginal mean zero and marginal variance one. We observe six independent and identically distributed data points: $(-3.3, -2.6), (0.1, -0.2), (-1.1, -1.5), (2.7, 1.5), (2.0, 1.9)$, and $(-0.4, -0.3)$. Make a scatter plot of the data and assuming the correlation parameter ρ has a Uniform$(-1, 1)$ prior, plot the posterior distribution of ρ.





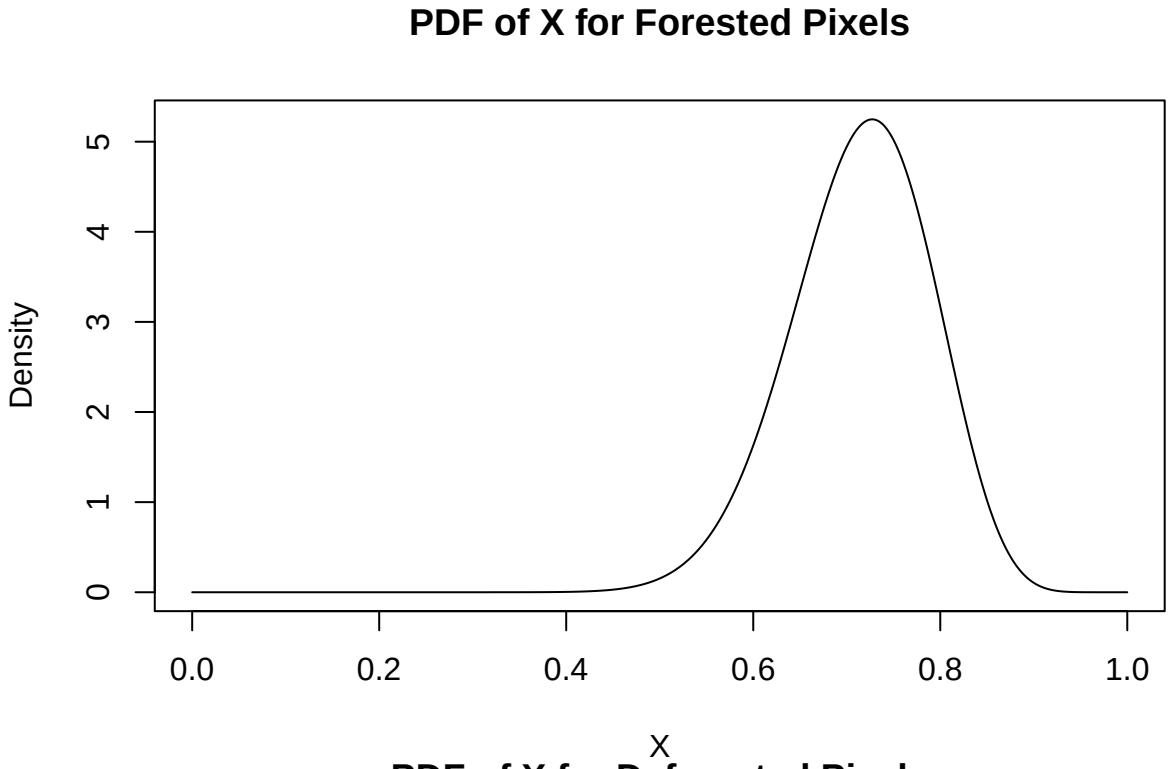
The MAP estimate for the given data is $\hat{\rho} = 0.848$. This seems to be a reasonable estimate given the positive correlation seen in the scatter plot.

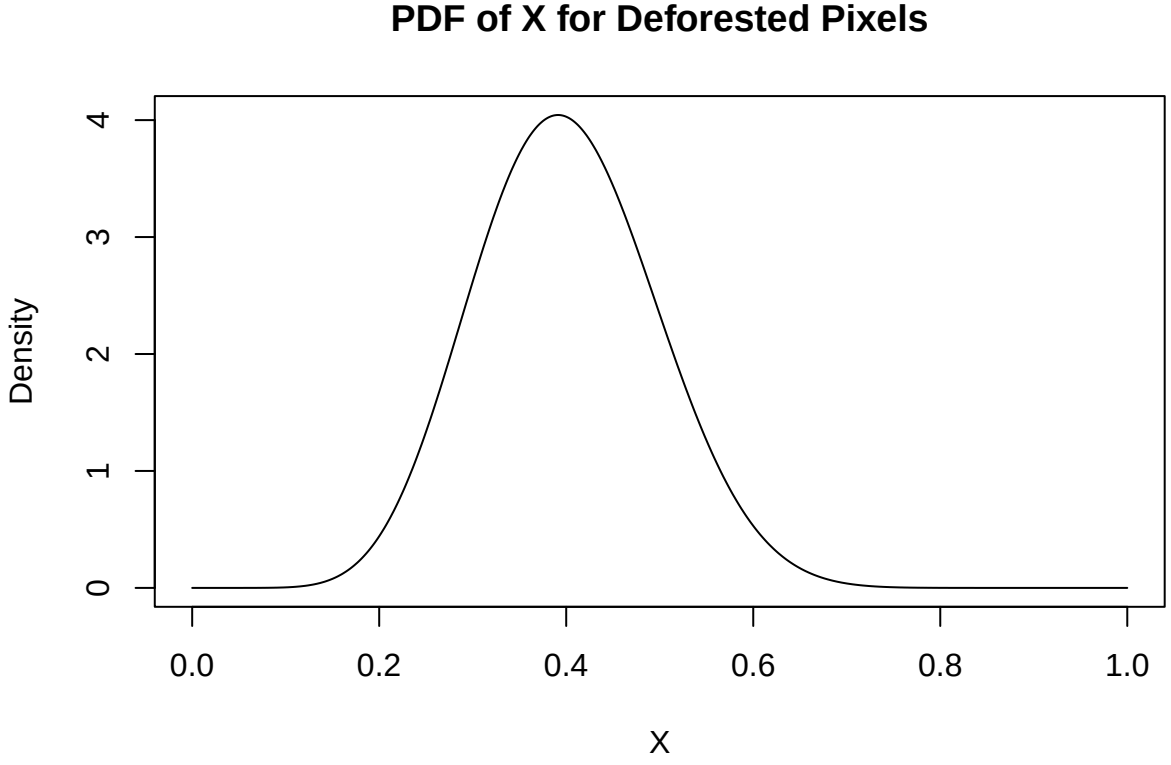
(2) Chapter 1, Problem 13

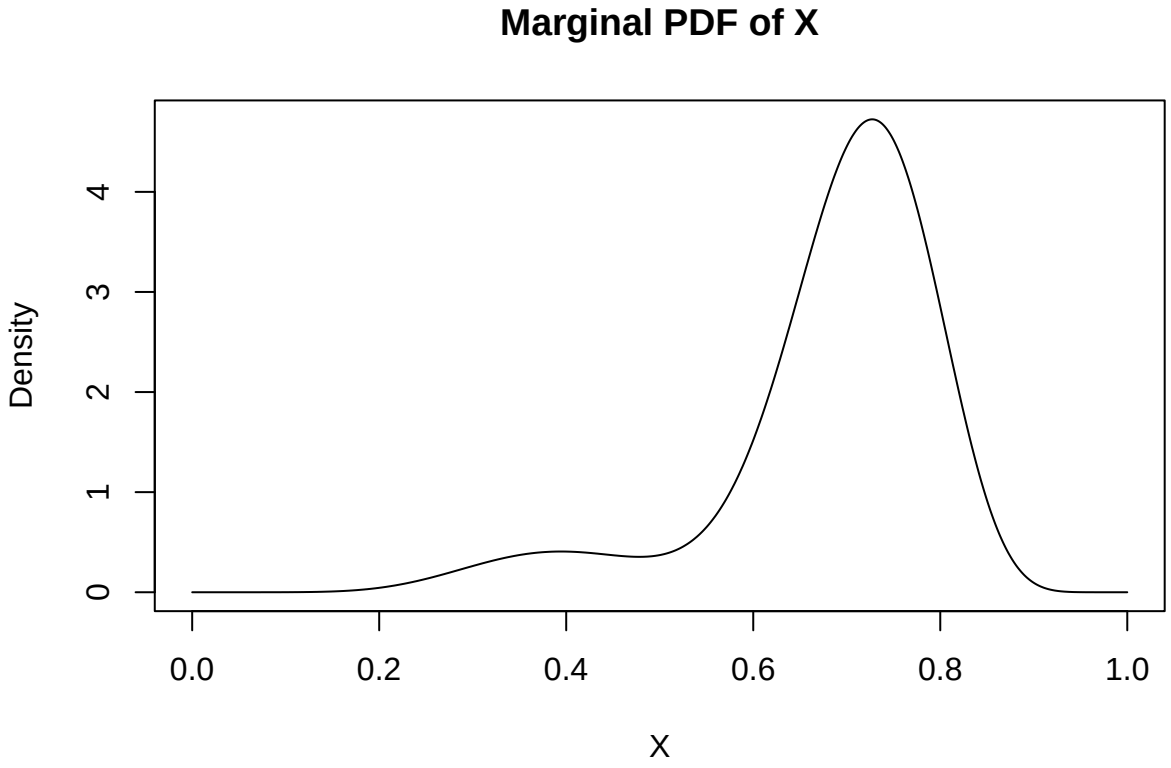
The normalized difference vegetation index (NDVI) is commonly used to classify land cover using remote sensing data. Hypothetically, say that NDVI follows a Beta(25, 10) distribution for pixels in a rain forest, and a Beta(10, 15) distribution for pixels in a deforested area now used for agriculture. Assuming about 10% of the rain forest has been deforested, your objective is to build a rule to classify individual pixels as deforested based on their NDVI.

(a) Plot the PDF of NDVI for forested and deforested pixels, and the marginal distribution of NDVI averaging over categories.

First, let X denote NDVI levels.

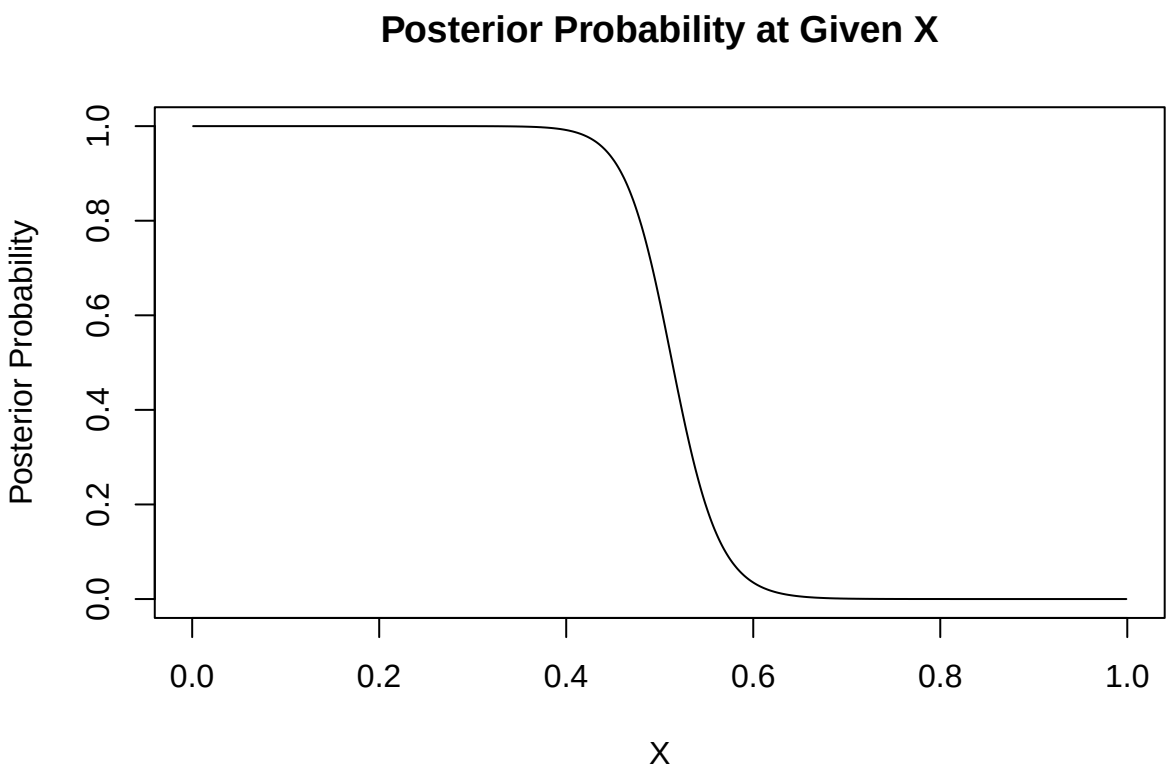

PDF of X for Forested Pixels
Density
0
1
2
3
4
5
0.0
0.2
0.4
0.6
0.8
1.0
X


PDF of X for Deforested Pixels
Density
0
1
2
3
4
0.0
0.2
0.4
0.6
0.8
1.0
X

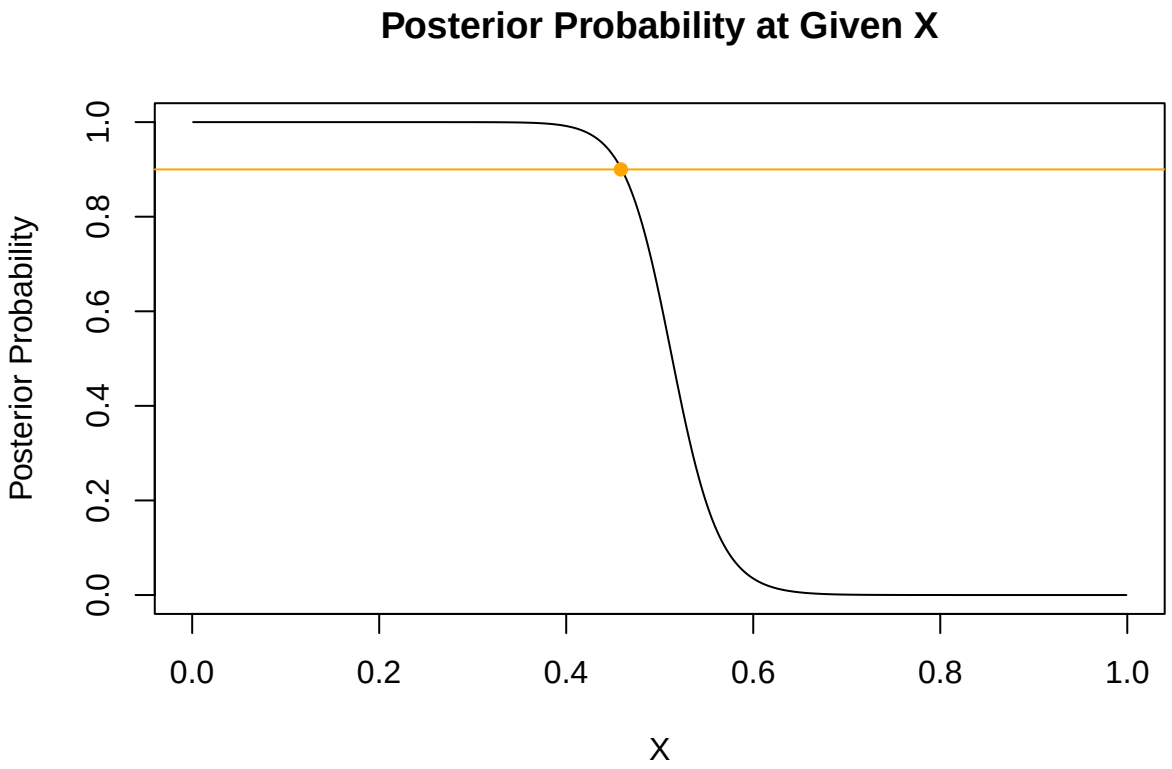

Marginal PDF of X
Density
0
1
2
3
4
0.0
0.2
0.4
0.6
0.8
1.0
X

(b) Give an expression for the probability that a pixel is deforested given its NDVI value, and plot this probability by NDVI.

$$P(\text{Deforested} \mid X) = \frac{P(X \mid \text{Deforested})P(\text{Deforested})}{P(X)}$$



(c) You will classify a pixel as deforested if you are at least 90% sure it is deforested. Following this rule, give the range of NDVI that will lead to a pixel being classified as deforested.



Classifier:

Deforested if $x \in [0, 0.4585]$

Forested if $x \in (0.4585, 1]$

(3) Chapter 1, Problem 17

The table below has the overall free throw proportion and results of free throws taken in pressure situations, defined as "clutch" (https://stats.nba.com/), for ten National Basketball Association players (those that received the most votes for the Most Valuable Player Award) for the 2016–2017 season. Since the overall proportion is computed using a large sample size, assume it is fixed and analyze the clutch data for each player separately using Bayesian methods. Assume a uniform prior throughout this problem.

Player	Overall proportion	Clutch makes	Clutch attempts
Russell Westbrook	0.845	64	75
James Harden	0.847	72	95
Kawhi Leonard	0.880	55	63
LeBron James	0.674	27	40
Isaiah Thomas	0.909	75	83
Stephen Curry	0.898	24	26
Giannis Antetokounmpo	0.770	28	36
John Wall	0.801	66	82
Anthony Davis	0.802	40	54
Kevin Durant	0.875	13	16

(a) Describe your model for studying the clutch success probability including the likelihood and prior.

The data $X \in \{0, 1, \cdots, n\}$ is the number of successes (clutch makes) in n independent trials (clutch attempts) each with success probability θ. Thus, the likelihood is

$$X \mid \theta \sim \text{Binomial}(n, \theta).$$

We will also pick the prior

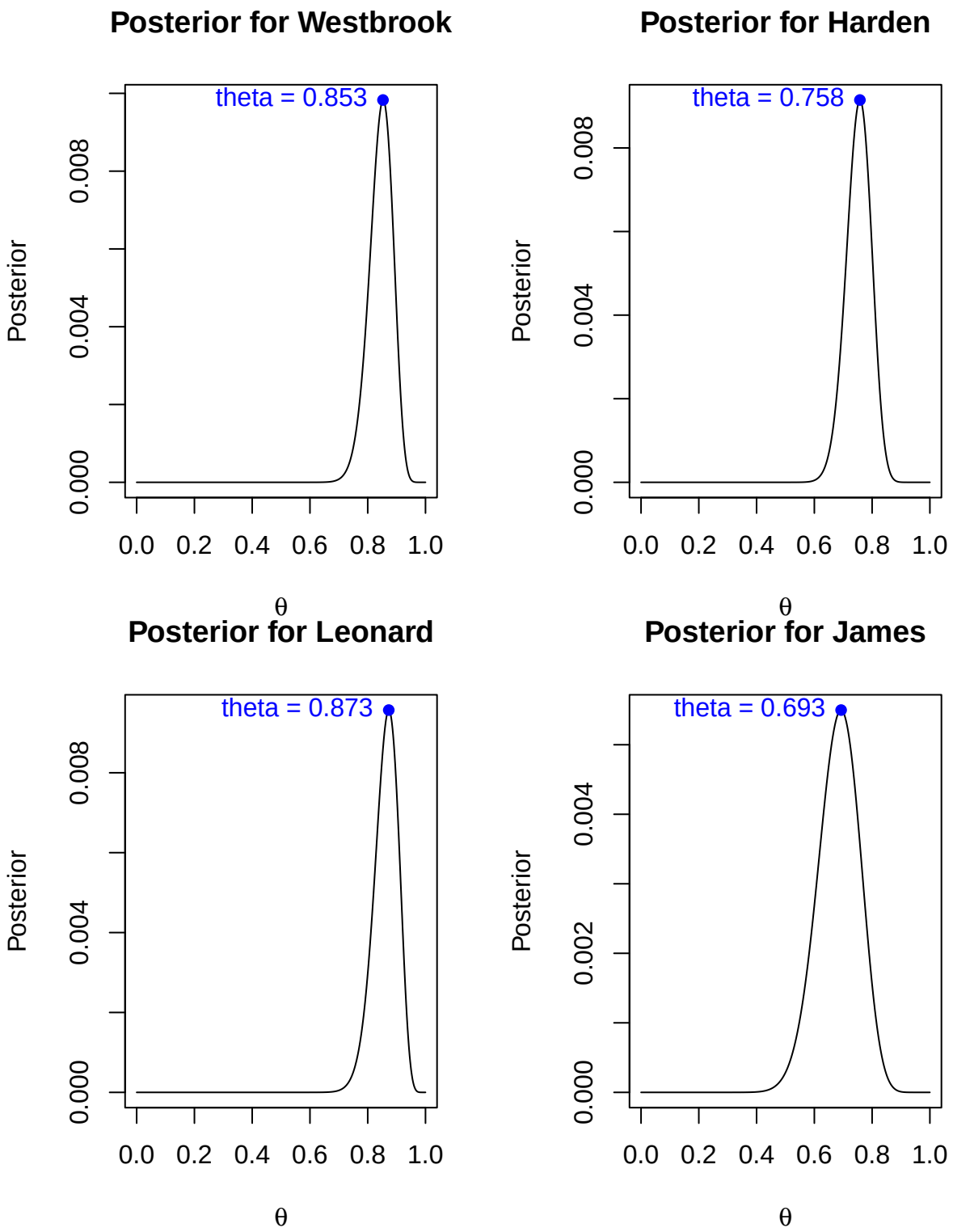
$$\theta \sim \text{Beta}(a, b).$$

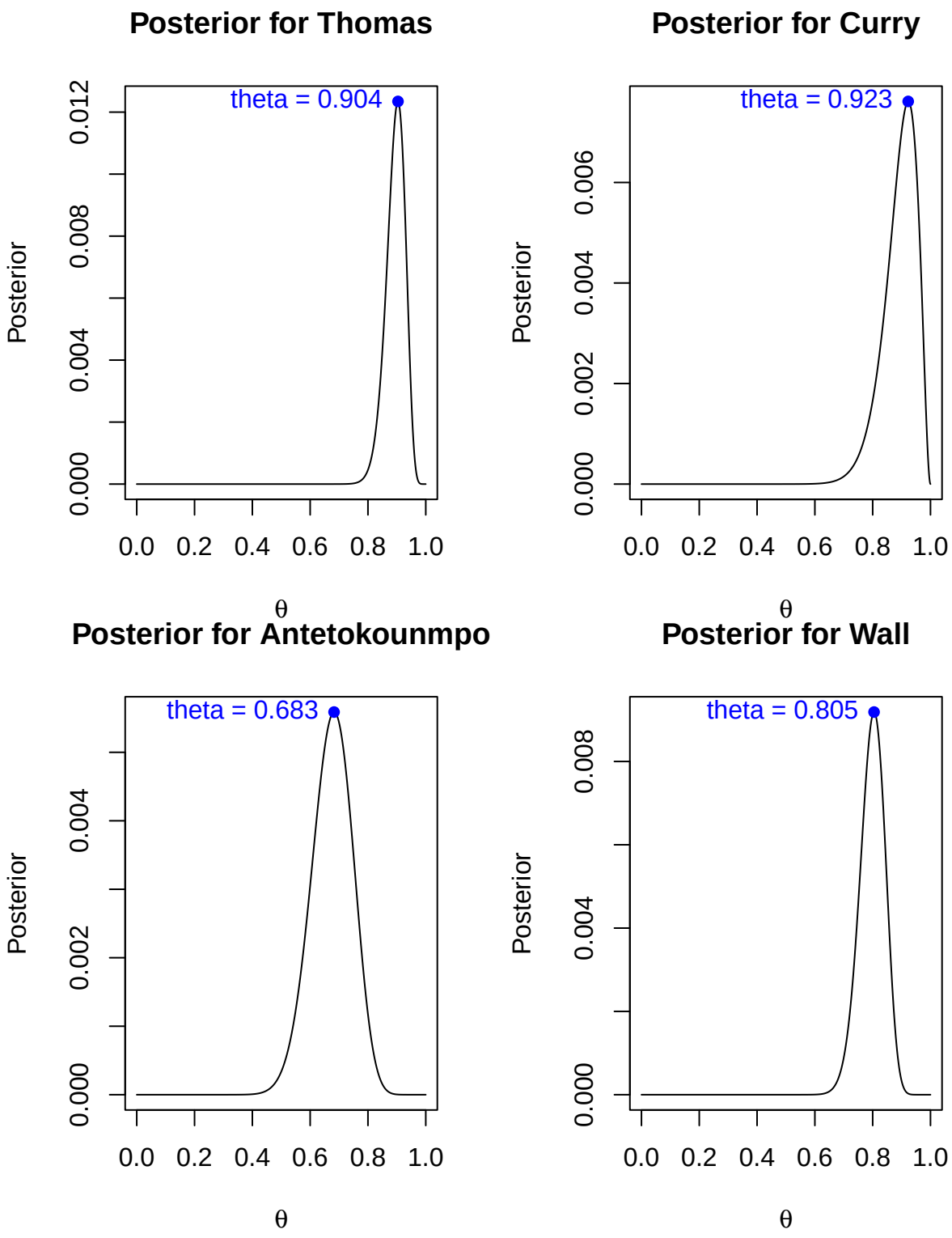
Which will lead to the posterior distribution

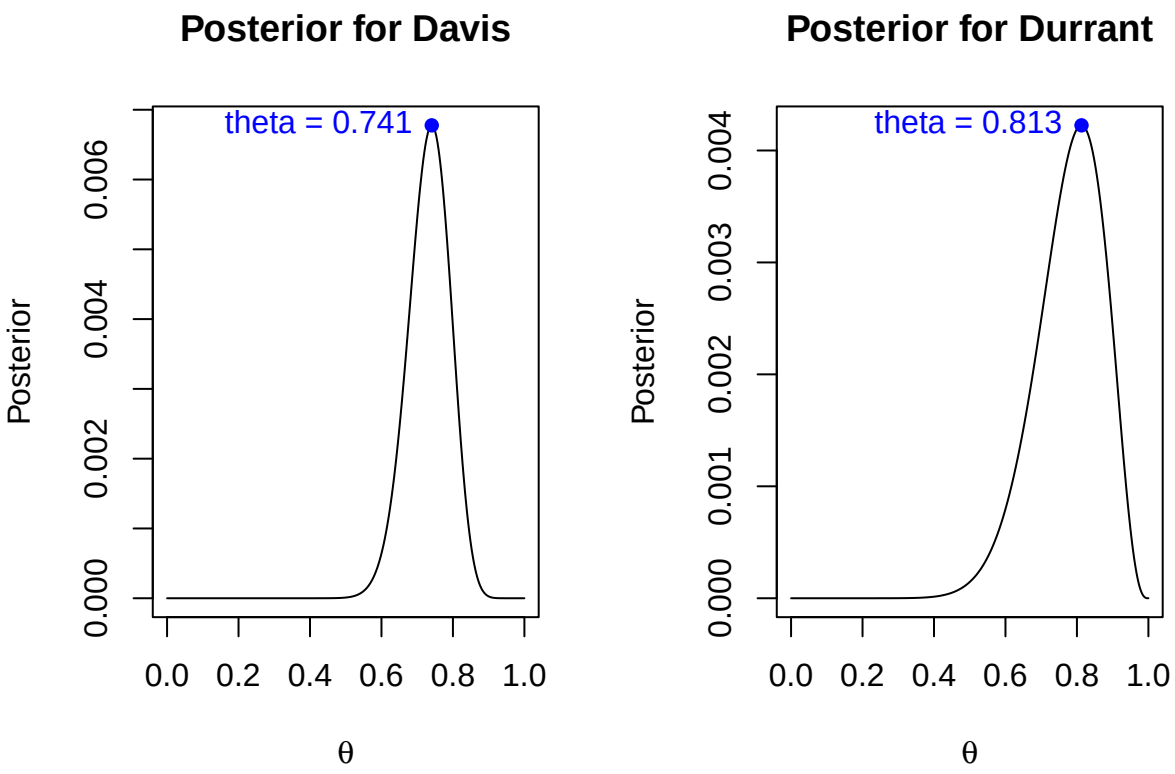
$$\theta \mid X \sim \text{Beta}(A, B),$$

where $A = X + a$ and $B = n - X + b$.

(b) Plot the posteriors of the clutch success probabilities.




Posterior for Thomas
theta = 0.904
Posterior
0.000
0.004
0.008
0.012
0.0
0.2
0.4
0.6
0.8
1.0
θ
Posterior for Curry
theta = 0.923
Posterior
0.000
0.002
0.004
0.006
0.0
0.2
0.4
0.6
0.8
1.0
θ
Posterior for Antetokounmpo
theta = 0.683
Posterior
0.000
0.002
0.004
0.0
0.2
0.4
0.6
0.8
1.0
θ
Posterior for Wall
theta = 0.805
Posterior
0.000
0.004
0.008
0.0
0.2
0.4
0.6
0.8
1.0
θ



(c) Summarize the posteriors in a table.

The following table was calculated using Monte Carlo approximation.

player	posterior_mean	posterior_SD	lower_95_Crit	upper_95_Crit
Westbrook	0.8442	0.0409	0.7558	0.9158
Harden	0.7526	0.0436	0.6624	0.8324
Leonard	0.8613	0.0424	0.7685	0.9329
James	0.6832	0.0717	0.5350	0.8138
Thomas	0.8941	0.0331	0.8216	0.9497
Curry	0.8929	0.0575	0.7570	0.9764
Antetokounmpo	0.6741	0.0704	0.5290	0.8041
Wall	0.7979	0.0433	0.7065	0.8758
Davis	0.7322	0.0585	0.6104	0.8381
Durrant	0.7777	0.0958	0.5647	0.9319

(d) Do you find evidence that any of the players have a different clutch percentage than overall percentage?

There seems to be evidence that James Harden has a different clutch percentage than overall percentage. His overall percentage is 0.847, which is higher than his 95% critical region for clutch percentage $(0.6624, 0.8324)$. There may be evidence to suggest that James Harden gets worse at free throws in clutch situations.

(e) Are the results sensitive to your prior? That is, do small changes in the prior lead to substantial changes in the posterior?

Yes, the results are sensitive to the prior, tweaking the parameters of Beta(a, b) causes the posterior distribution to change shape. This makes sense because a and b are nested in the parameters of the posterior distribution Beta(A, B), recall that $A = X + a$ and $B = n - X + b$.

(4) Chapter 2, Problem 2

The Major League Baseball player Reggie Jackson is known as "Mr. October" for his outstanding performances in the World Series (which takes place in October). Over his long career he played in 2820 regular-season games and hit 563 home runs in these games (a player can hit 0, 1, 2, ... home runs in a game). He also played in 27 World Series games and hit 10 home runs in these games. Assuming uninformative conjugate priors, summarize the posterior distribution of his home-run rate in the regular season and World Series. Is there sufficient evidence to claim that he performs better in the World Series?

For these data, we will assume a model

$$Y \mid \theta \sim \text{Poisson}(N\theta).$$

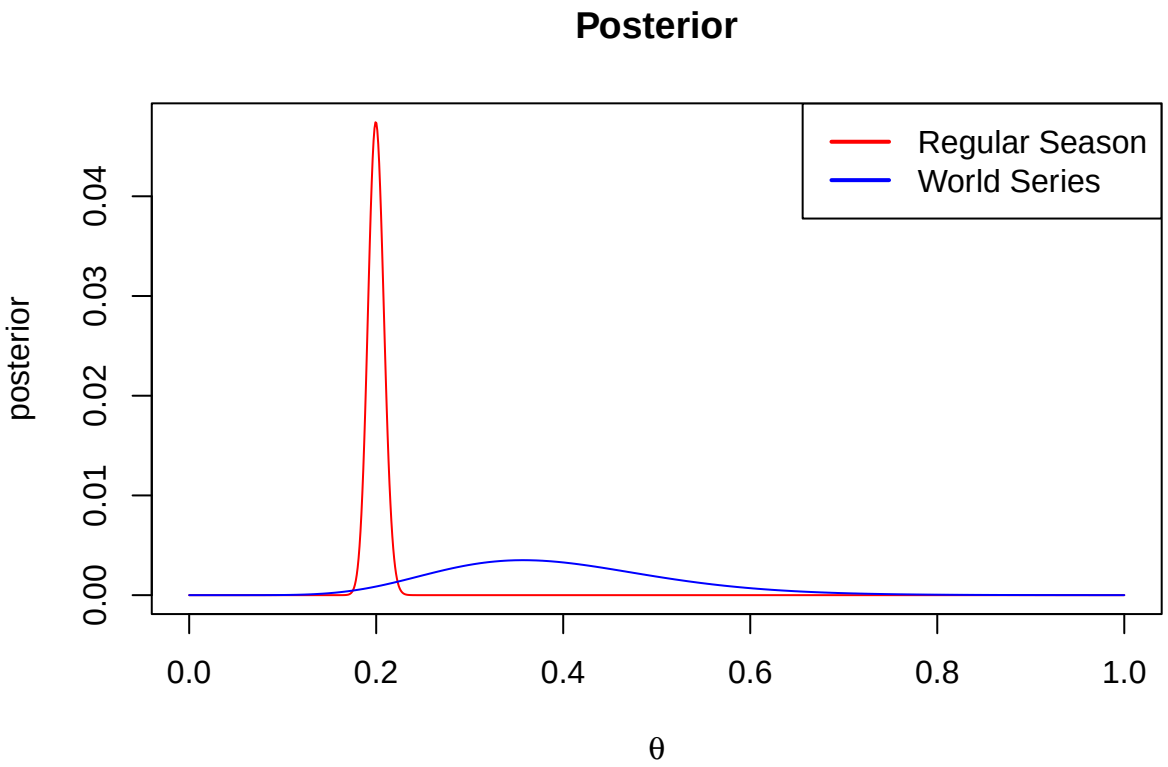
We pick the prior

$$\theta \sim \text{Gamma}(a, b).$$

The resulting posterior distribution is

$$\theta \mid Y \sim \text{Gamma}(A, B)$$

where $A = Y + a$ and $B = N + b$. We will set the parameters of the prior to $a = b = 1$ in order to make it uninformative.



Based on the plot of the posterior distributions, there seems to be evidence that Reggie Jackson is indeed "Mr. October" and performs better in the World Series. Specifically, the MAP estimate for his home run rate is significantly higher in World Series games compared to his games in the regular season.

Code Appendix

```
knitr::opts_chunk$set(echo = F,
                      message = F,
                      warning = F)
library(tidyverse)
library(knitr)

data_points <- data.frame(x1 = c(-3.3, 0.1, -1.1, 2.7, 2.0, -0.4),
                          x2 = c(-2.6, -0.2, -1.5, 1.5, 1.9, -0.3))

plot(data_points$x1, data_points$x2, main = "Scatter Plot of Data points",
     xlab = "x", ylab = "y")
bivariate_normal_pdf <- function(x1, x2, mu1, mu2, sigma1, sigma2, rho) {
  # Compute the joint PDF using the formula
  denom <- 2 * pi * sigma1 * sigma2 * sqrt(1 - rho^2)
  exponent <- -1 / (2 * (1 - rho^2)) * (
    (x1 - mu1)^2 / sigma1^2 +
    (x2 - mu2)^2 / sigma2^2 -
    2 * rho * (x1 - mu1) * (x2 - mu2) / (sigma1 * sigma2)
  )
  return(exp(exponent) / denom)
}

# Set mu and sigma values for the joint pdf based on info from marginal distributions.
mu1 <- mu2 <- 0
sigma1 <- sigma2 <- 1

rho_vals <- seq(-0.9999, 0.9999, length.out = 1000) # A subset of possible rho values
prior <- dunif(rho_vals, -1, 1) # Defining the prior distribution
likelihoods <- numeric(length(rho_vals)) # Setting up an empty list for the likelihood values

for (i in 1:length(rho_vals)) {
  rho <- rho_vals[i] # Looping through the subset of rho values
  likelihood <- 1 # Initialize likelihood to 1 at the start of each loop
  for (j in 1:nrow(data_points)) { # Loop through the data points
    x1 <- data_points[j, 1]
    x2 <- data_points[j, 2]
    likelihood <- likelihood * bivariate_normal_pdf(x1, x2, mu1, mu2, sigma1, sigma2, rho)
  }
  likelihoods[i] <- likelihood # return the likelihood for each rho value
}

posterior_unnormalized <- likelihoods * prior
posterior <- posterior_unnormalized / sum(posterior_unnormalized)

plot(rho_vals, posterior, type = "l", xlab = expression(rho), ylab = "Posterior",
     main = "Posterior Distribution of rho", xlim = c(-1, 1), ylim = c(0, 0.012))

argmax_index <- which.max(posterior) # argmax of the posterior
points(rho_vals[argmax_index], posterior[argmax_index], col = "blue", pch = 16)
text(rho_vals[argmax_index], posterior[argmax_index],
     labels = paste("rho =", round(rho_vals[argmax_index], 3)),
```

```
     pos = 3, col = "blue", offset = 0.5)
X <- seq(0, 1, length.out = 1000)
pdf_forested <- dbeta(X, 25, 10)
pdf_deforested <- dbeta(X, 10, 15)

plot(X, pdf_forested, type = "l",
     main = "PDF of X for Forested Pixels", ylab = "Density")
plot(X, pdf_deforested, type = "l",
     main = "PDF of X for Deforested Pixels", ylab = "Density")

p_forested <- 0.9
p_deforested <- 0.1

pdf_marginal <- p_forested * pdf_forested + p_deforested * pdf_deforested
plot(X, pdf_marginal, type = "l", main = "Marginal PDF of X", ylab = "Density")
posterior_deforest <- pdf_deforested * p_deforested / pdf_marginal
plot(X, posterior_deforest, type = "l",
     ylab = "Posterior Probability", main = "Posterior Probability at Given X")
threshold <- 0.9
deforested_range <- X[posterior_deforest >= threshold]

deforested_class_max <- max(deforested_range, na.rm=T)
argmax_index <- which.max(deforested_range)

plot(X, posterior_deforest, type = "l",
     ylab = "Posterior Probability", main = "Posterior Probability at Given X")
abline(h = 0.9, col = "orange")
points(X[argmax_index], 0.9, col = "orange", pch = 16)
theta <- seq(0, 1, length.out=1000)

calc_posterior <- function(x, size){
  likelihood <- dbinom(x, size, theta)
  prior <- dbeta(theta, 1, 1)

  posterior_unnormalized <- likelihood * prior
  posterior <- posterior_unnormalized/sum(posterior_unnormalized)
  return(posterior)
}

plot_posterior <- function(player_name) {
  player_data <- nba_data[nba_data$player == player_name, ]
  x <- player_data$clutch_makes
  size <- player_data$clutch_attempts

  theta <- seq(0, 1, length.out = 1000)

  posterior <- calc_posterior(x, size)
  plot(theta, posterior, type = "l", main = paste("Posterior for", player_name),
       xlab = expression(theta), ylab = "Posterior")

  argmax_index <- which.max(posterior)
  points(theta[argmax_index], posterior[argmax_index], col = "blue", pch = 16)
```

```
  text(theta[argmax_index], posterior[argmax_index],
       labels = paste(expression(theta), "=", round(theta[argmax_index], 3)),
       pos = 2, col = "blue", offset = 0.5)
}
nba_data <- data.frame(
  player = c("Westbrook", "Harden", "Leonard", "James", "Thomas",
             "Curry", "Antetokounmpo", "Wall", "Davis", "Durrant"),
  overall_prop = c(0.845, 0.847, 0.88, 0.674, 0.909,
                   0.898, 0.77, 0.801, 0.802, 0.875),
  clutch_makes = c(64, 72, 55, 27, 75, 24, 28, 66, 40, 13),
  clutch_attempts = c(75, 95, 63, 39, 83, 26, 41, 82, 54, 16))
par(mfrow=c(1,2))
for (j in 1:nrow(nba_data)){
  plot_posterior(nba_data$player[j])
}
set.seed(23)
# Initialize an empty data frame to store the results
all_player_summaries <- data.frame()

# Loop over each player and calculate the posterior summary
for (player_name in unique(nba_data$player)) {
  player_data <- nba_data[nba_data$player == player_name, ]
  X <- player_data$clutch_makes
  n <- player_data$clutch_attempts
  a <- 1
  b <- 1
  A <- X + a
  B <- n - X + b

  # Generate posterior samples using the beta distribution
  S <- 100000
  samples <- rbeta(S, A, B)

  # Calculate posterior mean, standard deviation, and 95% CI
  Posterior_mean <- mean(samples)
  Posterior_SD <- sd(samples)
  lower_CI <- quantile(samples, 0.025)
  upper_CI <- quantile(samples, 0.975)

  # Create a data frame for the player summary
  player_summary <- data.frame(player = player_name,
                               posterior_mean = round(Posterior_mean, 4),
                               posterior_SD = round(Posterior_SD, 4),
                               lower_95_Crit = round(lower_CI, 4),
                               upper_95_Crit = round(upper_CI, 4))

  # Append the result to all_player_summaries
  all_player_summaries <- rbind(all_player_summaries, player_summary)
}

# reset row names
rownames(all_player_summaries) <- NULL
```

```r
kable(all_player_summaries)
Y_reg <- 563
n_reg <- 2820
Y_ws <- 10
n_ws <- 27

# Posterior parameters for regular season (Gamma posterior)
alpha_post_reg <- 1 + Y_reg
beta_post_reg <- 1 + n_reg

# Posterior parameters for World Series (Gamma posterior)
alpha_post_ws <- 1 + Y_ws
beta_post_ws <- 1 + n_ws

theta <- seq(0, 1, length.out = 1000)

posterior_reg <- dgamma(theta, alpha_post_reg, beta_post_reg)
posterior_ws <- dgamma(theta, alpha_post_ws, beta_post_ws)
plot(theta, posterior_reg/sum(posterior_reg), type = "l", col = "red", main = "Posterior",
     ylab = "posterior", xlab = expression(theta))
lines(theta, posterior_ws/sum(posterior_ws), col = "blue")
legend("topright", legend = c("Regular Season", "World Series"),
       col = c("red", "blue"), lwd = 2)
```