BSDA: Assignment 2

September 8, 2023

```
knitr::opts_chunk$set(echo = TRUE)
```

Times used for reading and self-study exercise: 7 hours
Time used for the assignment: 16 hours
Good with the assignment: Clear instructions and good to include some analysis
Things to improve in the assignment: Maybe some question with an uncommon

Question 1

Q1a, Prior distribution, posterior distribution and likelihood function

For the first subtask for the Assignment we will formulate the likelihood function, the prior distribution and lastly the resulting posterior distribution. The prior distribution will be presented in the format of $\beta(\cdot, \cdot)$ where the dots will be replace with correct numerical values for the data set algae.

Below we are stating the prior distribution $P(\pi)$, likelihood function $P(y|\pi)$ and the posterior distribution $P(\pi|y)$ in that order. We are treating the parameter π as unknown and the number of observations in data set algae corresponds to n = 274. The derivation for the posterior can be found in the course literature. (Reference: Bayesian Data Analysis Third edition, chapter: 2 , page: 35-36)

The prior distribution

$$p(\pi) = \beta(\alpha, \beta) = \beta(\alpha = 2, \beta = 10)$$

The likelihood function

$$p(y|\pi) = \prod_{i=1}^{n} Bernoulli(\pi)$$

The posterior distribution

$$p(\pi|y) = \beta(\alpha + \sum_{i=1}^{n} y_i, n + \beta - \sum_{i=1}^{n} y_i) = \beta(\alpha + y, n + \beta - y)$$

where $y = \sum_{i=1}^{n} y_i$.

Because we will report the posterior distribution with correct numerical values we get the following:

```
library(bsda)
data("algae")

alpha = 2
beta = 10
# alpha for posterior distribution
alpha_upd <- sum(algae) + alpha
alpha_upd

## [1] 46

# beta for posterior distribution
beta_upd <- length(algae) + beta - sum(algae)
beta_upd

## [1] 240
```

Hence, the posterior distribution with correct numerical values becomes $p(\pi|y) = \beta(46, 240)$.

Q1b, Point estimate and 0.9 postior interval

For subtask Q1b we will calculate the point estimate and the 90 % posterior interval. We start by calculating the point estimate $E(\pi|y)$ which formula looks like the following:

$E(\pi|y) = \frac{\alpha^*}{\alpha^*+\beta^*}$

where α^* and β^* is the parameter element α and β for the posterior distribution, respective. Now we create a function in R to calculate the point estimate.

```
beta_point_est <- function(prior_alpha = 2,
                           prior_beta = 10,
                           data = algae_test) {

  post_alpha = prior_alpha + sum(data)
  post_beta = length(data) + prior_beta - sum(data)

  exp.value <- ( post_alpha ) / ( post_alpha + post_beta)

  return(exp.value)
}
beta_point_est(data = algae)

## [1] 0.1608392
```

So the point estimate for data set algae corresponds to $E(\pi|y) = 0.1608392$. Now when the point estimate for the parameter element π given the data y has been calculated we compute the 90 % posterior interval.

```r
beta_interval <- function(prior_alpha = 2, prior_beta = 10,
                          data = algae_test, prob = 0.9) {

  lowerbound <- qbeta( ( 1-prob)/2,
                       prior_alpha + sum(data),
                       length(data) + prior_beta - sum(data))

  upperbound <- qbeta(1-( 1-prob)/2,
                      prior_alpha + sum(data),
                      length(data) + prior_beta - sum(data))

  return(c(lowerbound, upperbound))

}
beta_interval(data = algae)

## [1] 0.1265607 0.1978177
```

The 90 % posterior interval is $[0.1265607, 0.1978177]$ such that $P(0.1265607 < \Theta < 0.1978177) = 0.9$ for this case.

Q1c, What is the probability that pi is less than pi0?

For this subtask we will look at the probability that the proportion of monitoring sites with detectable algae level π is less than some value π_0 from the historical records where $\pi|y \sim \beta(46, 240)$ for the data set algae. The function that computes this probability uses the pbeta function.

```r
beta_low <- function(prior_alpha = 2, prior_beta = 10,
                     data = algae_test, pi_0 = 0.2) {

  output <- pbeta(pi_0,
                  shape1 = prior_alpha + sum(data),
                  shape2 = length(data) + prior_beta - sum(data))

  return(output)
}
beta_low(data = algae)

## [1] 0.9586136
```

The probability of π being less or equal to $\pi_0 = 0.2$ given our data algae corresponds to 0.9586136.

Q1d, State the assumption required for the model

This model consist of a likelihood from the binomial distribution and a conjugate prior from the beta distribution. With this model and the data set some assumptions is required to be included and these are being stated according to below:

Assumption 1 [Assumption of a beta prior distribution for the parameter]

Assumption 2 [Assume for the data to be binary]

Assumption 3 [Y1, ..., Yn are conditionally independent and identically distributed, iid, given the parameter]

Assumption 4 [Y1, ..., Yn is a finite set of random elements]

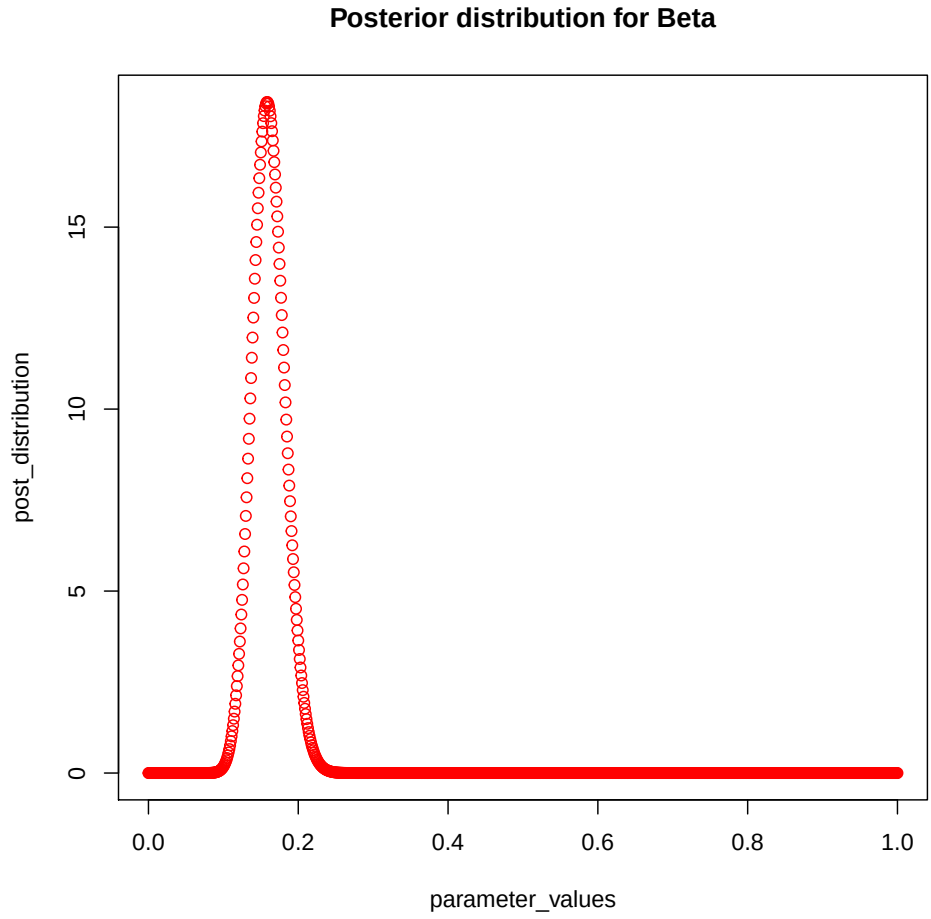
Worth mentioning is that we decided not to state the exchangeability assumption for this assignment.

Q1e, Prior sensitity analysis and plotting the different posteriors

We will do an sensitivity analysis and plotting posteriors for different reasonable priors. To do this analysis we will create a function which has an visualization of the density function, the point estimate and the 90 % posterior interval. Even though Jeffreys prior and uniform prior, which is a $\beta(0.5, 0.5)$ distributed and $\beta(1, 1)$ distributed respectively, is non-informative priors it will be included in the analysis.

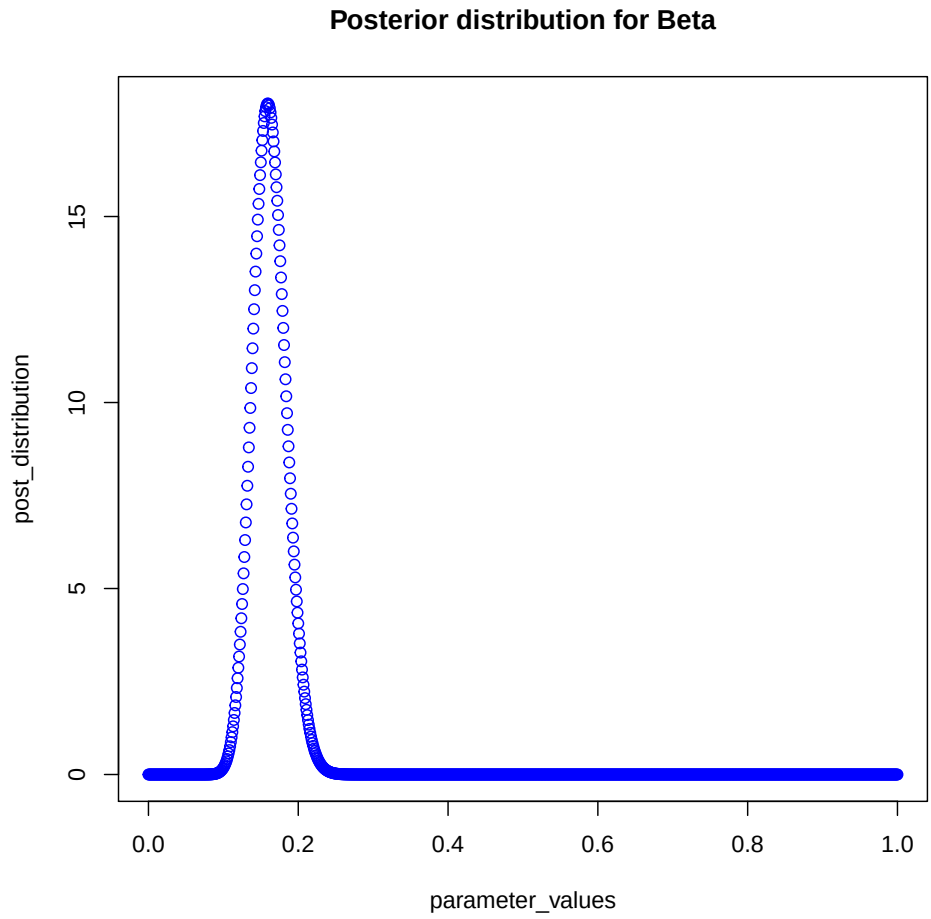
```
sensitity_analysis <- function(prior_alpha = 2, prior_beta = 10,
                               data = algae, prob = 0.9,
                               color ="blue") {

  post_alpha <- sum(data) + prior_alpha
  post_beta <- length(data) + prior_beta - sum(data)

  # Posterior interval
  lowerbound <- qbeta( ( 1-prob)/2, post_alpha, post_beta)
  upperbound <- qbeta(1-( 1-prob)/2, post_alpha, post_beta)
  post_inteval <- c(lowerbound, upperbound)

  # Point estimate
  point_est <- post_alpha/  ( post_alpha +  post_beta)

  # Plotting the distribution
  parameter_values <- seq(0,1, by = 0.001)
  post_distribution <- dbeta(parameter_values,
                             shape1 = post_alpha,
                             shape2 = post_beta)

  plotting <- plot(x = parameter_values, y = post_distribution,
                   main = "Posterior distribution for Beta",
                   col = color)

  output <- round(c(post_alpha, post_beta,
                    post_inteval, point_est), digits = 4)

  names(output) <- c("Post_Alpha", "Post_Beta",
                     "Interval_lower", "Interval_higher",
                     "Point_est")

  return(output)
}
# Posterior distribution when having Beta(2, 10),
```

```r
# ie the same distribution we used earlier in this assignment.
sensitity_analysis(prior_alpha = 2,
                   prior_beta = 10,
                   data = algae,
                   color = "red")
```



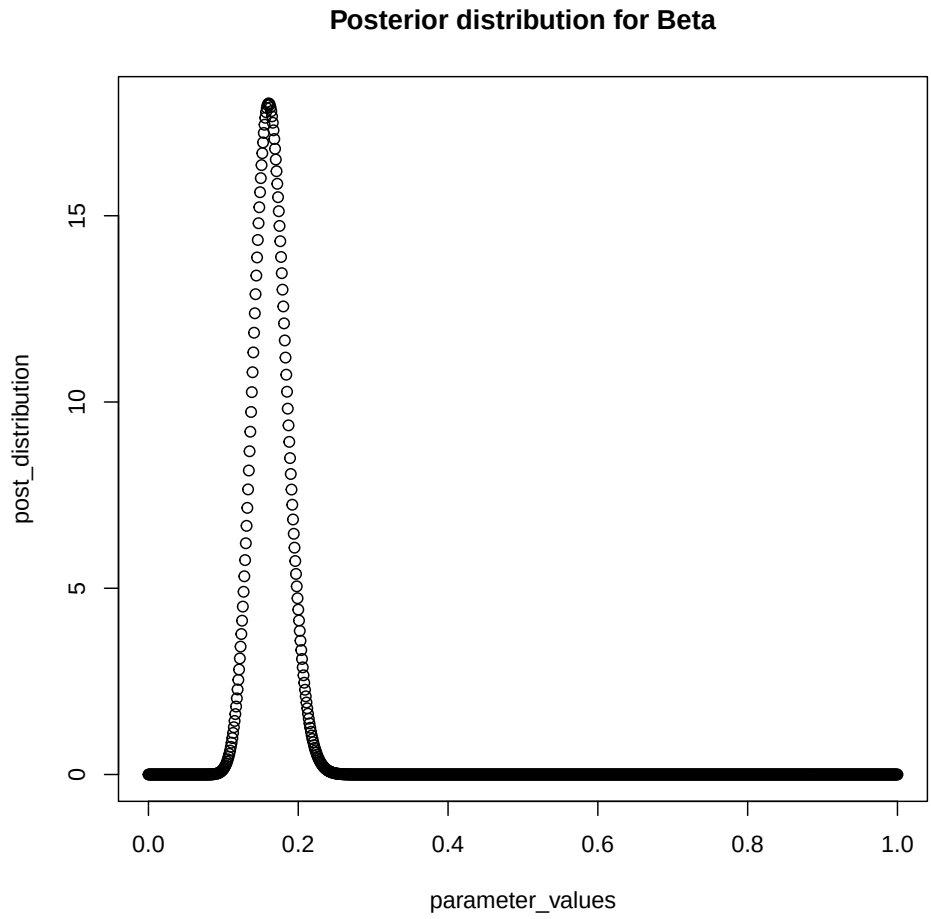

```
##      Post_Alpha       Post_Beta  Interval_lower Interval_higher       Point_est
##         46.0000        240.0000          0.1266          0.1978          0.1608
```

```r
# Posterior distribution when having Jeffreys prior,
# ie Beta(0.5, 0.5)
sensitity_analysis(prior_alpha = 0.5,
                   prior_beta = 0.5,
                   data = algae,
                   color = "blue")
```



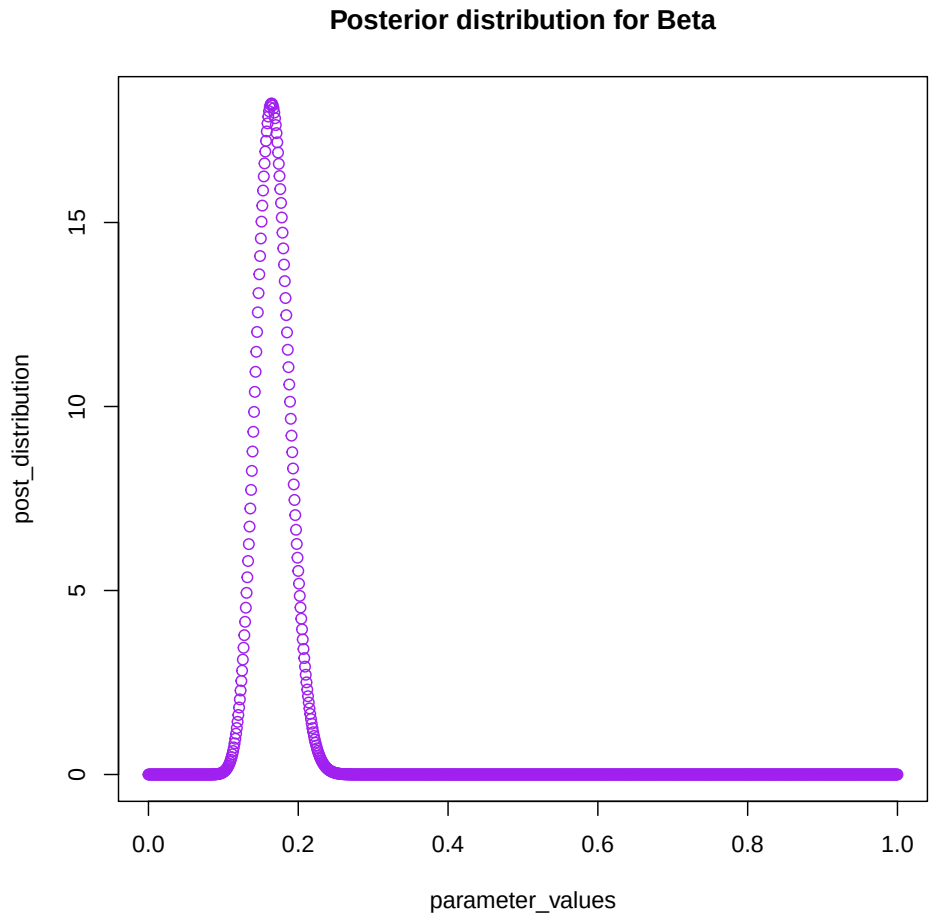

```
##      Post_Alpha       Post_Beta  Interval_lower Interval_higher       Point_est
##         44.5000        230.5000          0.1268          0.1996          0.1618
```

```
# Posterior distribution when having uniform prior, ie Beta(1,1)
sensitity_analysis(prior_alpha = 1,
                   prior_beta = 1,
                   data = algae,
                   color = "black")
```




```
##      Post_Alpha       Post_Beta  Interval_lower Interval_higher       Point_est
##         45.0000        231.0000          0.1280          0.2009          0.1630
```

```
# Posterior distribution when having Beta(4, 10)
sensitity_analysis(prior_alpha = 4,
                   prior_beta = 10,
                   data = algae,
                   color = "purple")
```




```
##      Post_Alpha       Post_Beta  Interval_lower Interval_higher       Point_est
##         48.0000        240.0000          0.1320          0.2040          0.1667
```

For the analysis we included $\beta(2, 10)$, $\beta(0.5, 0.5)$,$\beta(1, 1)$ and $\beta(4, 10)$ as prior distribution. When looking at the point estimate, the density function and the posterior interval, the different reasonable priors resulted in very similar posteriors. We used a large number of parameter values for this analysis and the different posterior might be similar each other because a posterior that is based on a large sample are not particularly sensitive to the prior. (Reference: Bayesian Data Analysis Third edition, chapter: 2 , page: 38)