Assignment 2

anonymous

1 General information

I did not use AI for solving this exercise.

2 Inference for binomial proportion

Loading the library and the data.

```
library(aaltobda)
data("algae")
# The data are now stored in the variable `algae`.
# These are the values for the prior required in the assignment
prior_alpha = 2
prior_beta = 10
```

The below data is **only for the tests**, you need to change to the full data `algae` when reporting your results.

```
algae_test <- c(0, 1, 1, 0, 0, 0)
```

2.1 (a)

```
# The following loop computes the number of sites in which the observations gave that the
# algae was present (1) and the total number of observations
total_samples <- length(algae)
positives <- sum(algae)

cat('Total Samples: ', total_samples, '\n')
```

```
Total Samples:  274
```

```
cat('Positives: ', positives, '\n')
```

```
Positives:  44
```

The Bayes' Rule can be expressed as follows:

$$p(\pi|y) = \frac{p(y|\pi)p(\pi)}{p(y)} \rightarrow posterior = \frac{likelihood \cdot prior}{evidence}$$

In this specific scenario, where we're using a binomial model for the observations, the likelihood is given by:

$p(y|\pi) \propto \binom{n}{y} \pi^y (1-\pi)^{n-y} = Binomial(n, y) = \text{Binomial}(274,\ 44)$.

The prior distribution takes the form:

$p(\pi) \propto \pi^{\alpha-1}(1-\pi)^{\beta-1} = \text{Beta}(\alpha, \beta) = \text{Beta}(2, 10)$

This leads us to the posterior distribution:

$p(\pi|y) \propto Beta(\theta|\alpha + y, \beta + n - y)$

For this specific case, the posterior can be represented as:

$p(\pi|y) = \text{Beta}(\theta|2 + 44, 10 + 274 - 44)$

The final result can be summarized as:

$$Posterior \rightarrow p(\pi|y) \propto \text{Beta}(46, 240)$$

```
# These are not the actual values for the posterior!
# You will have to compute those from the data!
posterior_alpha = prior_alpha + positives
posterior_beta = prior_beta + total_samples - positives
```

2.2 (b)

The point estimate represented by $E(\pi|y)$ can be interpreted as the posterior probability of success for a future draw from the population, as described in BDA3. This estimate is calculated as:

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

In this specific case, the calculation becomes:

$$E(\pi|y) = \frac{2 + 44}{2 + 10 + 274} = 0.1608392$$

The result lies between the sample proportion $\frac{y}{n} = \frac{44}{274} = 0.1605839$ and the prior mean $\frac{\alpha}{\alpha+\beta} = \frac{2}{2+10} = 0.1666667$.

```
# Useful function: qbeta()

beta_point_est <- function(prior_alpha, prior_beta, data) {
    pos <- sum(data)
    total_samples <- length(data)

    posterior_alpha <- prior_alpha + pos
    posterior_beta <- prior_beta + total_samples - pos
    point_estimate <- posterior_alpha /
      (posterior_alpha + posterior_beta)
```

```r
    return(point_estimate)
}

beta_interval <- function(prior_alpha, prior_beta, data, prob = 0.9) {
    pos <- sum(data)
    total_samples <- length(data)

    posterior_alpha <- prior_alpha + pos
    posterior_beta <- prior_beta + total_samples - pos

    lower_quantile <- qbeta((1 - prob) / 2,
                            posterior_alpha,
                            posterior_beta)

    upper_quantile <- qbeta(1 - (1 - prob) / 2,
                            posterior_alpha,
                            posterior_beta)

    return(c(lower_quantile, upper_quantile))
}


info <- function(prior_alpha, prior_beta, data, prob = 0.9) {
    pos <- sum(data)
    total_samples <- length(data)

    posterior_alpha <- prior_alpha + pos
    posterior_beta <- prior_beta + total_samples - pos

    lower_quantile <- qbeta((1 - prob) / 2,
                            posterior_alpha,
                            posterior_beta)

    upper_quantile <- qbeta(1 - (1 - prob) / 2,
                            posterior_alpha,
                            posterior_beta)

    mean <- posterior_alpha / (posterior_alpha + posterior_beta)
    cat("Mean: ", mean,"\n")

    median <- (lower_quantile + upper_quantile) / 2
    cat("Median: ", median,"\n")

    if (posterior_alpha > 1 && posterior_beta > 1) {
      mode <- (posterior_alpha - 1) /
        (posterior_alpha + posterior_beta - 2)
    } else {
      mode <- NA
    }
    cat("Mode: ", mode,"\n")
}
```

```r
cat(beta_point_est (prior_alpha, prior_beta, algae))
```

```
0.1608392
```

The result obtained through R, denoted as 0.1608392, matches the analytical computation performed earlier.

```r
interval <- beta_interval (prior_alpha, prior_beta, algae, prob = 0.9)
cat(interval)
```

```
0.1265607 0.1978177
```

The 90% posterior interval: [0.1265607, 0.1978177].

```r
info (prior_alpha, prior_beta, algae, prob = 0.9)
```

```
Mean:  0.1608392
Median:  0.1621892
Mode:  0.1584507
```

2.3 (c)

```r
# Useful function: pbeta()

beta_low <- function(prior_alpha, prior_beta, data, pi_0 = 0.2) {
    pos <- sum(data)
    total_samples <- length(data)

    posterior_alpha <- prior_alpha + pos
    posterior_beta <- prior_beta + total_samples - pos

    prob_below_pi_0 <- pbeta(pi_0, posterior_alpha, posterior_beta)
    return(prob_below_pi_0)
}

prob_res <- beta_low(prior_alpha, prior_beta, algae, 0.2)

cat(prob_res)
```

```
0.9586136
```

The probability that the proportion π of monitoring sites with detectable algae levels is less than π_0 is 0.9586136.

2.4 (d)

One the one hand, as it is exposed in the book BDA3, the main assumptions to pass from a prior distribution $p(\pi)$ to a posterior distribution $p(\pi|y)$ are:

- $E(\pi) = E(E(\pi|y))$: *The prior mean of* π is the average of all possible posterior means over the distribution of possible data.
- $var(\pi) = E(var(\pi|y)) + var(E(\pi|y))$: *The posterior variance is on average smaller than the prior variance, by an amount that depends on the variation in posterior means over the distribution of possible data.*

On the other hand,

Binomial Nature of Data:

- Each observation has two possible outcomes, which are coded as 1 and 0.
- This implies a binary classification or a success-failure type of scenario.

Independence of Observations:

- Each observation is assumed to be independent from all others.
- The outcome of one observation does not influence or affect the outcomes of the other observations.
- This assumption is crucial for various statistical analyses, such as logistic regression.

Identically Distributed Observations:

- All observations follow the same probability distribution. In the context of binary data, this means that the probability of success (1) and failure (0) is the same for each observation.
- This assumption ensures that the data is consistent and can be modeled with a single set of parameters.

Prior Knowledge Modeled as Beta Distribution:

- Prior knowledge or beliefs about the probability of success (1) can be described using a beta distribution.
- The beta distribution is a probability distribution that is often used as a prior distribution in Bayesian statistics.

Posterior Distribution as Beta:

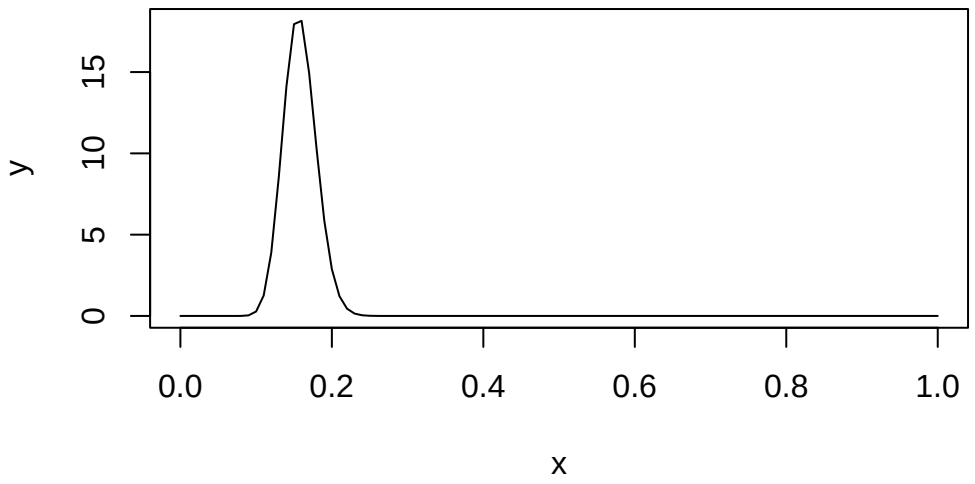
- The model assumes that the posterior distribution, which represents updated knowledge after observing the data, can also be expressed as a beta distribution.
- This means that the model incorporates Bayesian methods, where prior beliefs are updated with observed data to compute a posterior distribution.

2.5 (e)

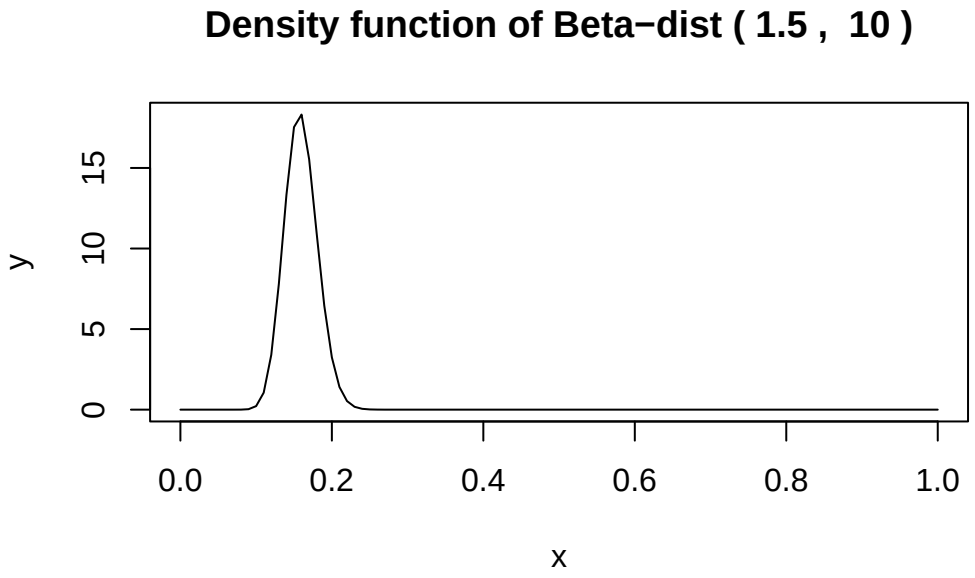
Plot the PDFs here. Explain shortly what you do.

```r
# Useful function: dbeta()
plot_posterior <- function(prior_params, data, prob){
  total_samples <- length(data)
  pos <- sum(data)
  x <- seq(from = 0, to = 1, by = 0.01)

  for (prior_name in names(prior_params)) {
    prior_alpha <- prior_params[[prior_name]][1]
```

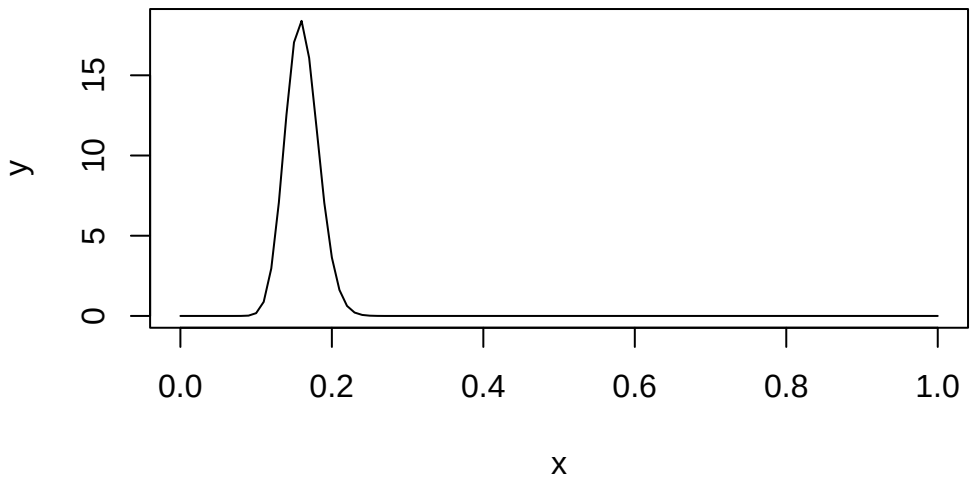
```r
    prior_beta <- prior_params[[prior_name]][2]

    posterior_alpha <- prior_alpha + pos
    posterior_beta <- prior_beta + total_samples - pos
    posterior <- dbeta(x, posterior_alpha, posterior_beta)

    #Plotting operation
    y <- posterior
    plot(x, y, type = "l", main =
         paste("Density function of Beta-dist (", prior_alpha, ", ", prior_beta, ")")
       )

    lower_quantile <- qbeta((1 - prob) / 2, posterior_alpha, posterior_beta)
    upper_quantile <- qbeta(1 - (1 - prob) / 2, posterior_alpha, posterior_beta)

    prior_proportion <- prior_alpha / (prior_alpha + prior_beta)
    amount_information <- prior_alpha + prior_beta
    posterior_median <- (upper_quantile + lower_quantile) / 2

    print(paste("alpha", prior_alpha))
    print(paste("beta", prior_beta))
    print(paste("95% Credible Interval:", lower_quantile, "-", upper_quantile))
    print(paste("Posterior Median:", posterior_median))
    print(paste("Prior Proportion of Success:", prior_proportion))
    print(paste("Amount of Prior Information:", amount_information))
    print("----------------------------")
  }
}

# Example usage
prior_params1 <- list(
  prior1 = c(alpha = 1, beta = 10),
  prior2 = c(alpha = 1.5, beta = 10),
  prior3 = c(alpha = 2, beta = 10),
  prior4 = c(alpha = 2.5, beta = 10),
  prior5 = c(alpha = 3, beta = 10)
)

plot_posterior(prior_params1, algae, 0.9)
```



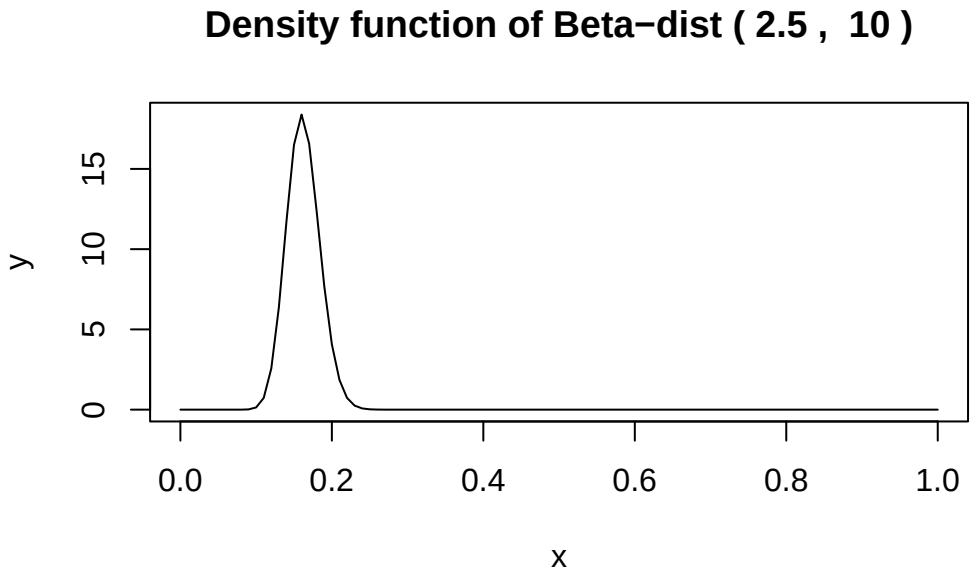

```
[1] "alpha 1"
[1] "beta 10"
[1] "95% Credible Interval: 0.123838541943476 - 0.194683938122638"
[1] "Posterior Median: 0.159261240033057"
[1] "Prior Proportion of Success: 0.0909090909090909"
[1] "Amount of Prior Information: 11"
[1] "----------------------------"
```



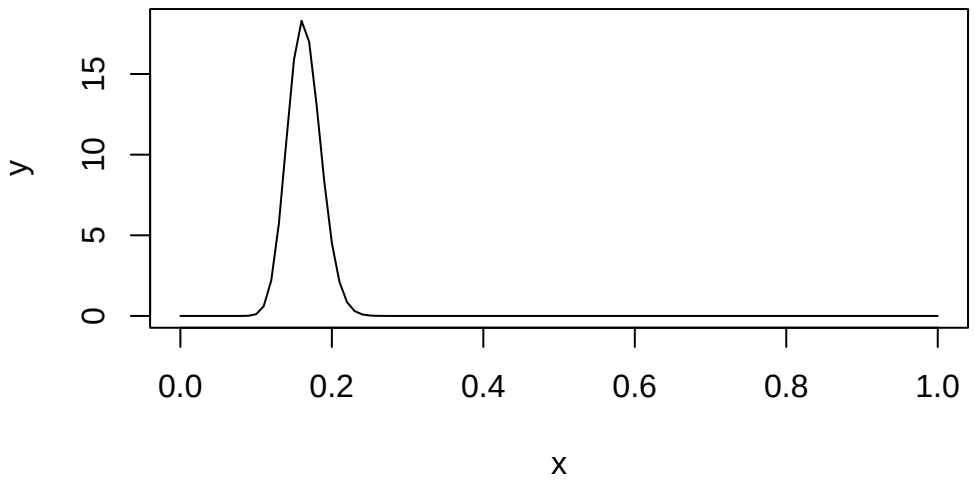

```
[1] "alpha 1.5"
[1] "beta 10"
[1] "95% Credible Interval: 0.125201327681062 - 0.19625420962928"
[1] "Posterior Median: 0.160727768655171"
[1] "Prior Proportion of Success: 0.130434782608696"
[1] "Amount of Prior Information: 11.5"
[1] "----------------------------"
```




```
[1] "alpha 2"
[1] "beta 10"
[1] "95% Credible Interval: 0.126560711878773 - 0.197817667316324"
[1] "Posterior Median: 0.162189189597549"
[1] "Prior Proportion of Success: 0.166666666666667"
[1] "Amount of Prior Information: 12"
[1] "----------------------------"
```

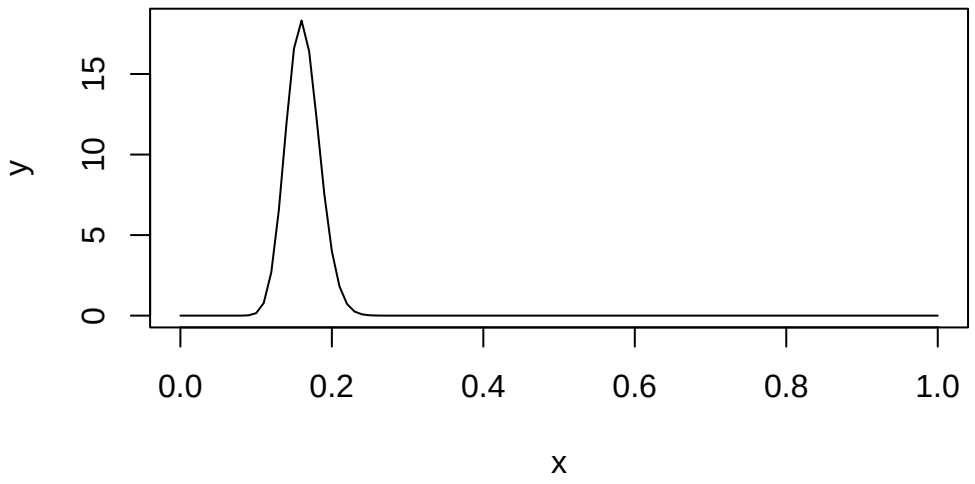



```
[1] "alpha 2.5"
[1] "beta 10"
[1] "95% Credible Interval: 0.127916690816538 - 0.199374368209427"
[1] "Posterior Median: 0.163645529512982"
[1] "Prior Proportion of Success: 0.2"
[1] "Amount of Prior Information: 12.5"
[1] "----------------------------"
```



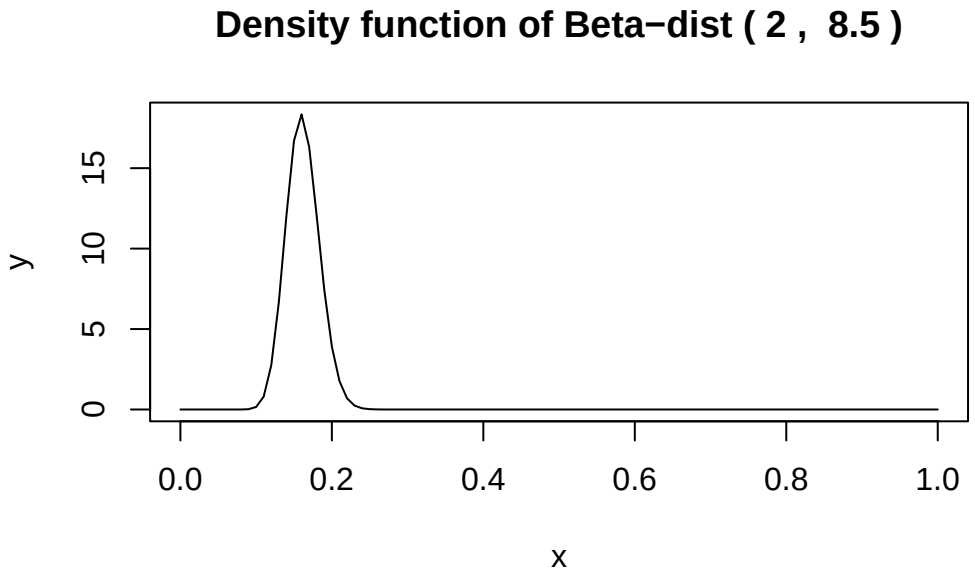

```
[1] "alpha 3"
[1] "beta 10"
[1] "95% Credible Interval: 0.129269261355371 - 0.200924368381563"
[1] "Posterior Median: 0.165096814868467"
[1] "Prior Proportion of Success: 0.230769230769231"
[1] "Amount of Prior Information: 13"
[1] "----------------------------"
```

```r
prior_params2 <- list(
  prior1 = c(alpha = 2, beta = 8),
  prior2 = c(alpha = 2, beta = 8.5),
  prior3 = c(alpha = 2, beta = 9),
  prior4 = c(alpha = 2, beta = 9.5),
  prior5 = c(alpha = 2, beta = 10)
)

plot_posterior(prior_params2, algae, 0.9)
```

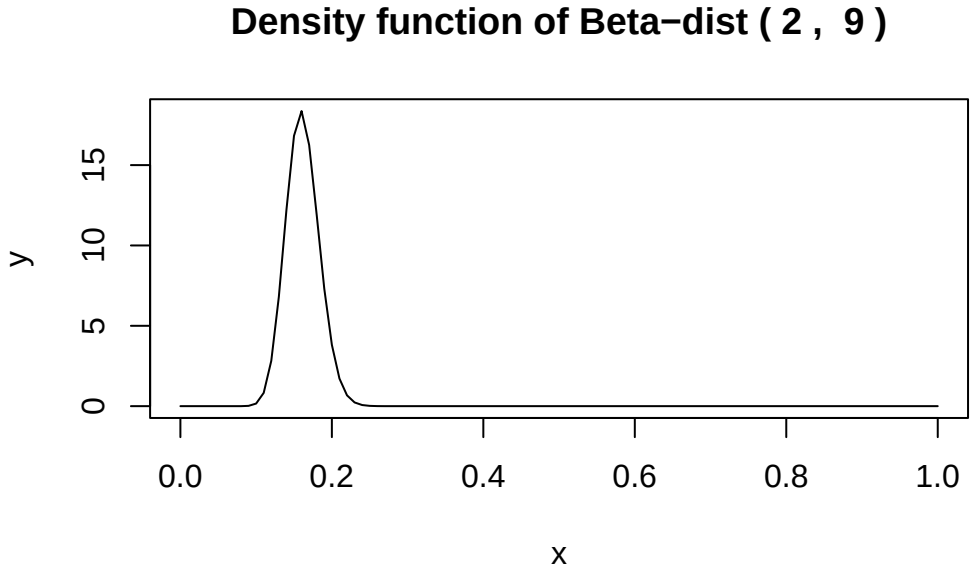



```
[1] "alpha 2"
```

```
[1] "beta 8"
[1] "95% Credible Interval: 0.127471795664355 - 0.199181902747749"
[1] "Posterior Median: 0.163326849206052"
[1] "Prior Proportion of Success: 0.2"
[1] "Amount of Prior Information: 10"
[1] "----------------------------"
```

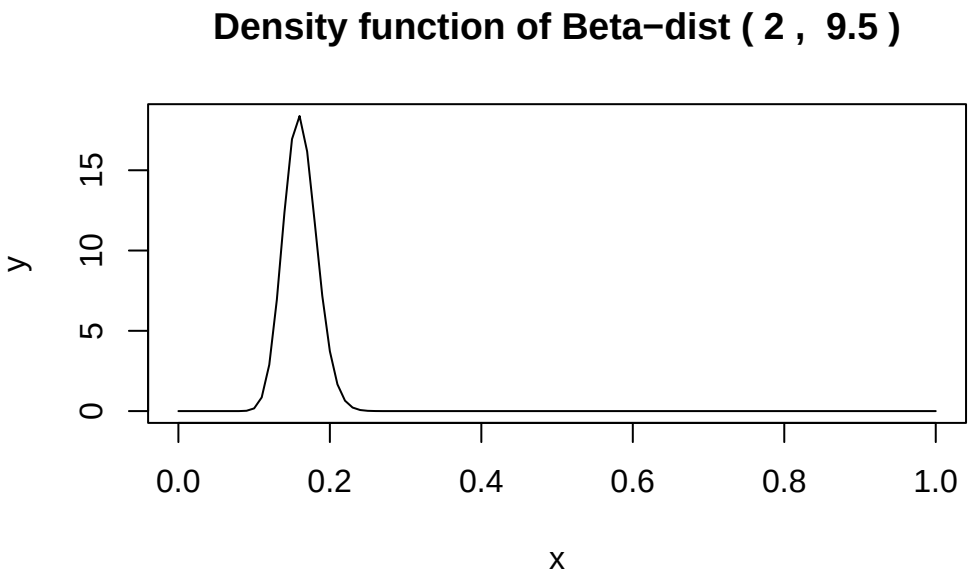



```
[1] "alpha 2"
[1] "beta 8.5"
[1] "95% Credible Interval: 0.127242796201961 - 0.198839085221829"
[1] "Posterior Median: 0.163040940711895"
[1] "Prior Proportion of Success: 0.19047619047619"
[1] "Amount of Prior Information: 10.5"
[1] "----------------------------"
```

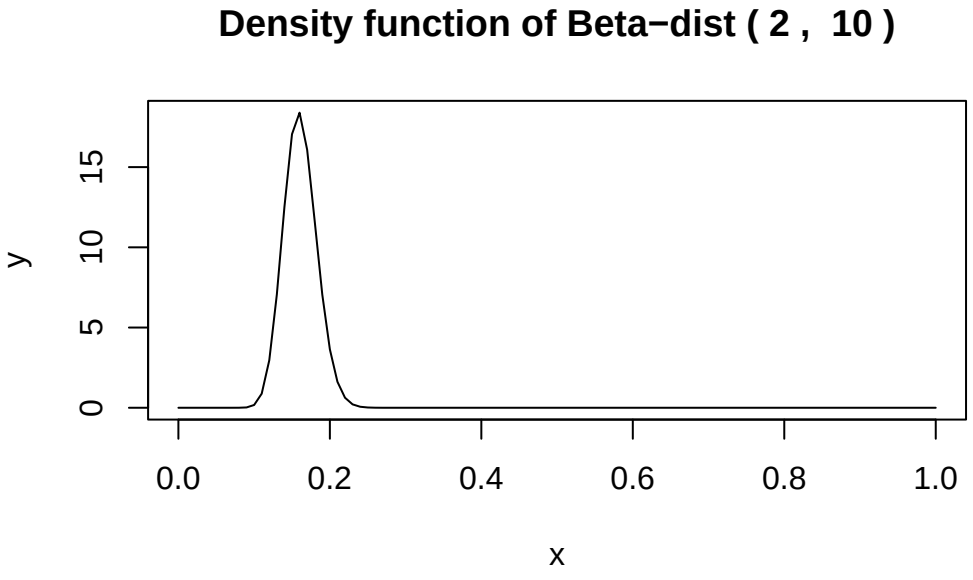



```
[1] "alpha 2"
[1] "beta 9"
[1] "95% Credible Interval: 0.127014618692395 - 0.198497444158788"
[1] "Posterior Median: 0.162756031425592"
```

```
[1] "Prior Proportion of Success: 0.181818181818182"
[1] "Amount of Prior Information: 11"
[1] "----------------------------"
```




```
[1] "alpha 2"
[1] "beta 9.5"
[1] "95% Credible Interval: 0.126787258714495 - 0.198156973522081"
[1] "Posterior Median: 0.162472116118288"
[1] "Prior Proportion of Success: 0.173913043478261"
[1] "Amount of Prior Information: 11.5"
[1] "----------------------------"
```




```
[1] "alpha 2"
[1] "beta 10"
[1] "95% Credible Interval: 0.126560711878773 - 0.197817667316324"
[1] "Posterior Median: 0.162189189597549"
[1] "Prior Proportion of Success: 0.166666666666667"
[1] "Amount of Prior Information: 12"
[1] "----------------------------"
```

In the provided data for each of the plots, we have included essential information, including the 90% posterior interval, the posterior median, the prior proportion calculated as $\frac{\alpha}{\alpha+\beta}$, and the amount of prior information estimated by $\alpha+\beta$. It's expected that when there's a higher amount of prior information, the posterior median tends to be closer to the prior mean. When examining the plots, we can observe that they become more sharply defined as the amount of data increases. The plot shows how it gets a sharper shape when the amount of data is higher, meaning that the interval around its expected mean becomes smaller. So the higher the amount of the data, the more accurate the distribution.