BDA - Assignment 2

Anonymous

Contents

Load packages

```
library(aaltobda)
data("algae")
```

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

Exercise 1)

a)

- π: the probability of a monitoring site having detectable blue-green algae levels: $\pi \rightarrow Beta(2, 10)$
- y: observations in algae: $y \rightarrow Binomial(n, p)$, where n is the number of observations having detectable blue-green algae levels and p is the probability of getting one of those.

```
#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 <- 0
positives <- 0

for (i in algae) {
  total_samples <- total_samples +1
  if (i == 1){
    positives <- positives + 1
  }
}

print(positives)
```

```
## [1] 44
```

```
print(total_samples)
```

```
## [1] 274
```

The expression for the Bayes' Rule is the following:

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

In the given case, the likelihood, using a binomial model for the observations, would be as following: $p(y|\pi) \propto \binom{n}{y} \pi^y (1-\pi)^{n-y} = Binomial(n, y) = \text{Binomial}(274, 44)$.

The prior distribution is the following: $p(\pi) \propto \pi^{\alpha-1}(1-\pi)^{\beta-1} = Beta(\alpha, \beta) = \text{Beta}(2{,}10)$

The expression for the posterior would be: $p(\pi|y) \propto Beta(\theta|\alpha + y, \beta + n - y)$

For the given case, the final expression would be: $p(\pi|y) = \text{Beta}(\theta|\ 2 + 44,\ 10 + 274 - 44)$

The result would be:

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

b)

The point estimate such that $E(\pi|y)$ can be interpreted as *the posterior probability of success for a future draw from the population* (BDA3), so:

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

In the given case, the result would be:

$$E(\pi|y) = \frac{2 + 44}{2 + 10 + 274} = 0.1608$$

The result is contained between $\frac{y}{n} = \frac{44}{274} = 0.1606$ and the prior mean $\frac{\alpha}{\alpha+\beta} = \frac{2}{2+10} = 0.1667$. Now, the calculations using R:

```
#Function computing the point estimate, given the beta distribution hyperparameters of
#the prior and the data
beta_point_est <- function(prior_alpha, prior_beta, data){
  total_samples <- 0
  positives <- 0
  for (j in data) {
  total_samples <- total_samples + 1
  if (j == 1){
    positives <- positives + 1
  }
  }
  point_estimate <- (prior_alpha + positives)/(prior_alpha+prior_beta+total_samples)
  return(point_estimate)
  }

estimate <- beta_point_est(2,10,algae)
estimate
```

```
## [1] 0.1608392
```

The obtained result, 0.1608392, using R is the same than the one previously computed analytically.

```r
#In order to get the posterior interval, the following function can be used
beta_interval <- function(prior_alpha, prior_beta, data, prob){
  total_samples <- 0
  positives <- 0
  for (k in data) {
  total_samples <- total_samples + 1
  if (k == 1){
    positives <- positives + 1
  }
  }
  posterior_alpha = prior_alpha + positives
  posterior_beta = prior_beta + total_samples - positives

  x <- seq(from = 0, to = 1, by = 0.01)
  posterior <- dbeta(x, posterior_alpha, posterior_beta)

  samples <- rbeta(n = 1000, posterior_alpha, posterior_beta)

  limit_1 <- (1-prob)/2
  limit_2 <- limit_1 + prob

  a <- quantile(samples, probs = limit_1)
  b <- quantile(samples, probs = limit_2)
  results <- c(a, b)
  return(results)
}

interval <- beta_interval(2, 10, algae, 0.9)
interval
```

```
##        5%       95%
## 0.1258066 0.1967656
```

The 90% posterior interval obtained is [0.1258066, 0.1967656].

c)

```r
#The following function provides the probability that the proportion of monitoring sites
#with detectable algae levels pi is smaller than pi_0
beta_low <- function(prior_alpha, prior_beta, data, pi_0){
  total_samples <- 0
  positives <- 0
  for (k in data) {
  total_samples <- total_samples + 1
  if (k == 1){
    positives <- positives + 1
  }
  }
  posterior_alpha = prior_alpha + positives
  posterior_beta = prior_beta + total_samples - positives

  x <- seq(from = 0, to = 1, by = 0.01)
  posterior <- dbeta(x, posterior_alpha, posterior_beta)
```

```
  probability <- pbeta(pi_0, posterior_alpha, posterior_beta)
  return(probability)

}

prob_result <- beta_low(2, 10, algae, 0.2)
prob_result
```

```
## [1] 0.9586136
```

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

d)

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:

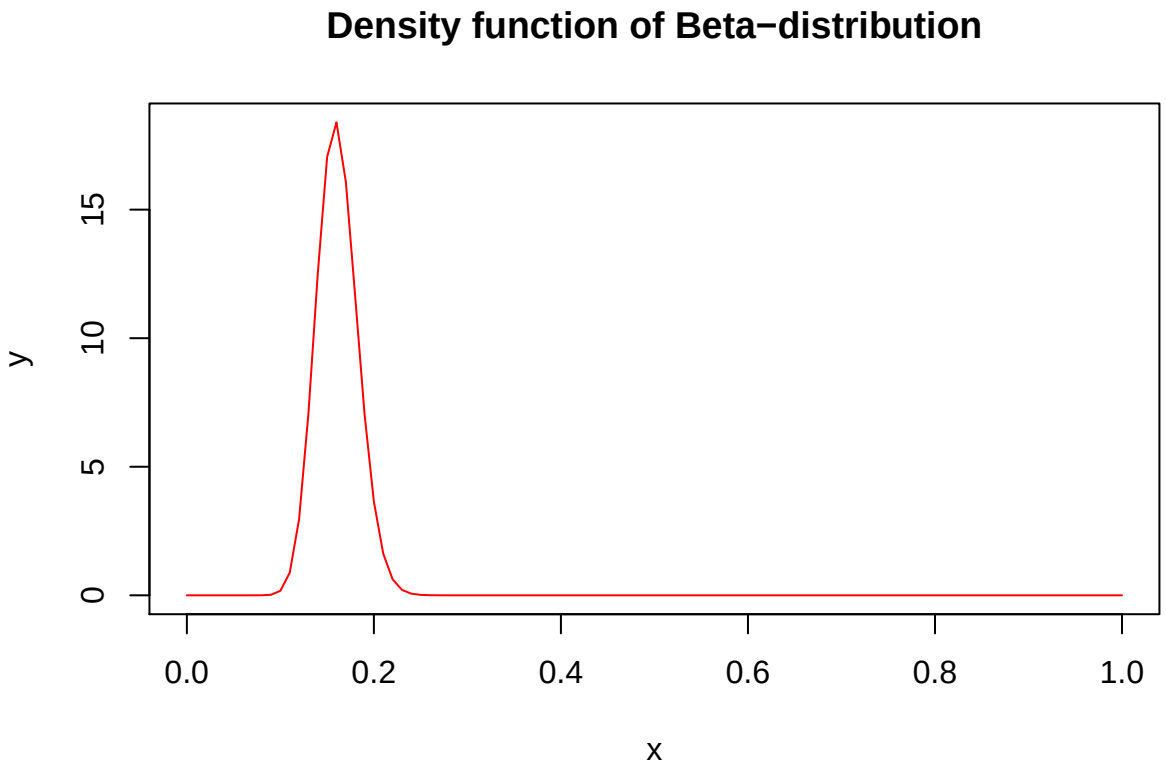
- $E(\pi) = E(E(\pi|y))$: *The prior mean of π is the average of all possible posterior means over the distribution of possible data* (BDA3).
- $var(\pi) = E(var(\pi|y)) + var(E(\pi|y))$: *The posterior variance is on average smaller than the prior variance, by an amount that depends on the variation in posterior means over the distribution of possible data* (BDA3).

e)

```
#The following function returns the posterior function
plot_posterior <- function(prior_alpha, prior_beta, data, prob){
  total_samples <- 0
  positives <- 0
  for (k in data) {
  total_samples <- total_samples + 1
  if (k == 1){
    positives <- positives + 1
  }
  }
  posterior_alpha = prior_alpha + positives
  posterior_beta = prior_beta + total_samples - positives
  x <- seq(from = 0, to = 1, by = 0.01)
  posterior <- dbeta(x, posterior_alpha, posterior_beta)
  #Plotting operation
  y <- posterior
  plot(x, y, type = "l", col = "red", main = "Density function of Beta-distribution")

  #90% posterior interval
  samples <- rbeta(n = 1000, posterior_alpha, posterior_beta)
  limit_1 <- (1-prob)/2
  limit_2 <- limit_1 + prob
  a <- quantile(samples, probs = limit_1)
  b <- quantile(samples, probs = limit_2)
  prior_proportion <- prior_alpha / (prior_alpha+prior_beta)
  amount_information <- prior_alpha + prior_beta
```

```
  limit_sup <- b
  limit_inf <- a
  posterior_median <- (b-a)/2 + a
  result <- c(a, b, posterior_median, prior_proportion, amount_information)
  return(result)
}
info <- "5%, 95%, posterior median, prior proportion, amount of information"
#Results for the original prior
results_prior_1 <- plot_posterior(2, 10, algae, 0.9)
```



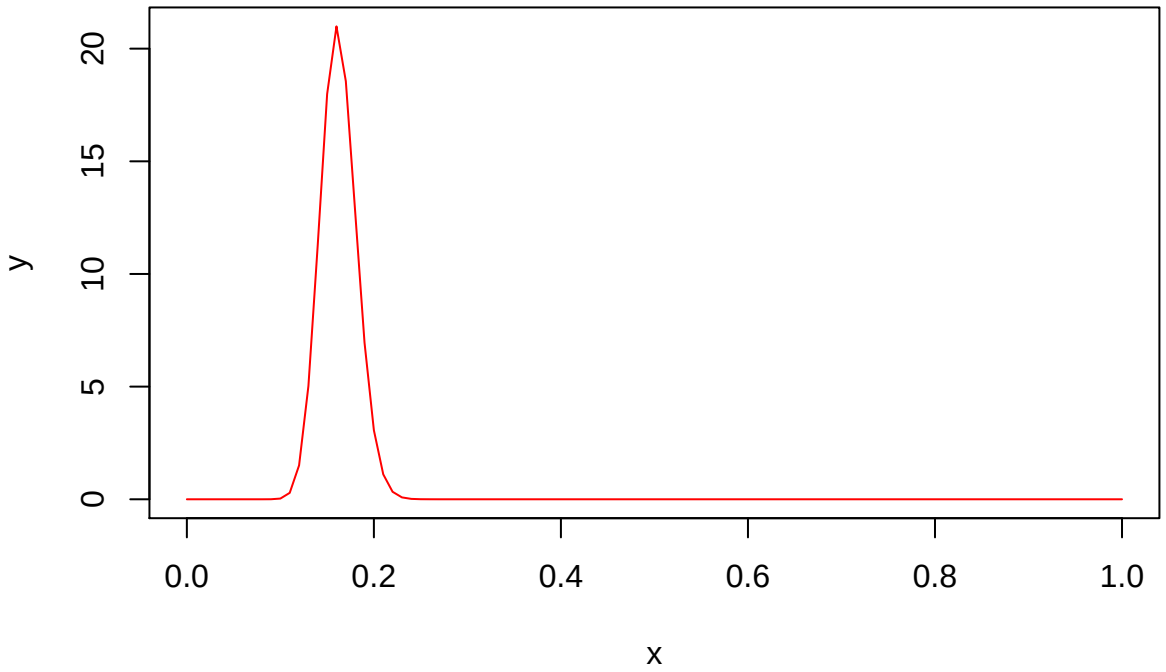

```
info[1]
```

```
## [1] "5%, 95%, posterior median, prior proportion, amount of information"
```

```
results_prior_1
```

```
##         5%        95%        95%
##  0.1281584  0.1979120  0.1630352  0.1666667 12.0000000
```

```
#Results for a new prior: Beta(16.66667, 83.333333)
results_prior_2 <- plot_posterior(16.666667, 83.33333333, algae, 0.9)
```



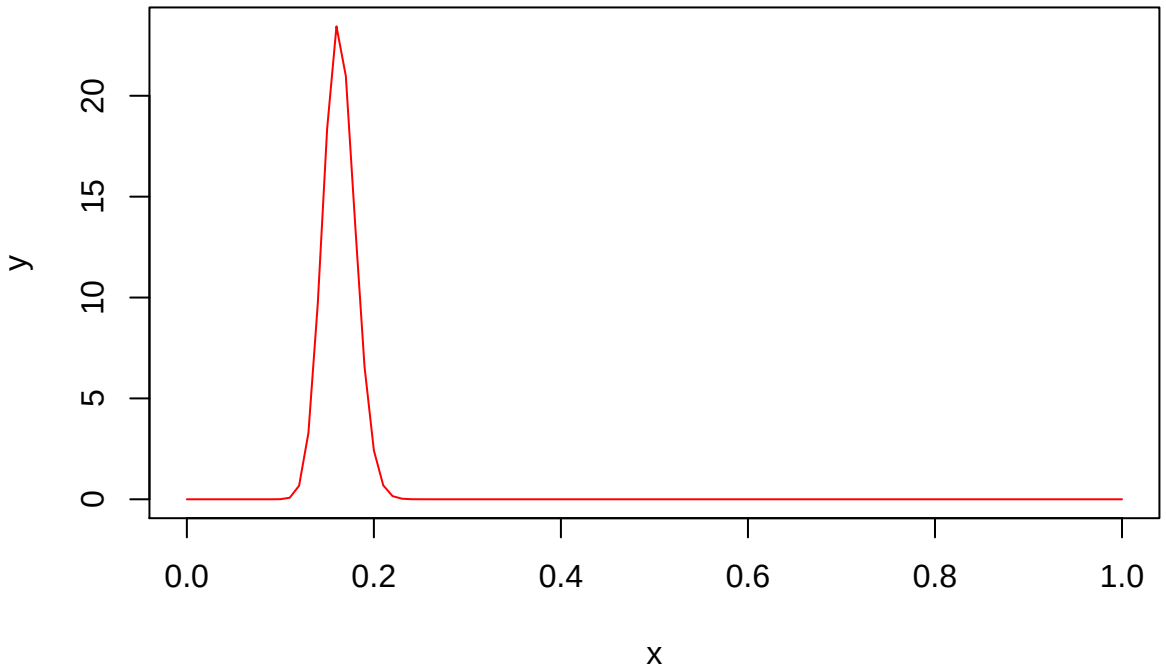

```
info[1]
```

```
## [1] "5%, 95%, posterior median, prior proportion, amount of information"
```

```
results_prior_2
```

```
##          5%         95%         95%
##   0.1307667   0.1939687   0.1623677   0.1666667 100.0000003
```

```
#Results for a new prior: Beta(33.3333333, 166.666667)
results_prior_3 <- plot_posterior(33.33333333, 166.6666667, algae, 0.9)
```



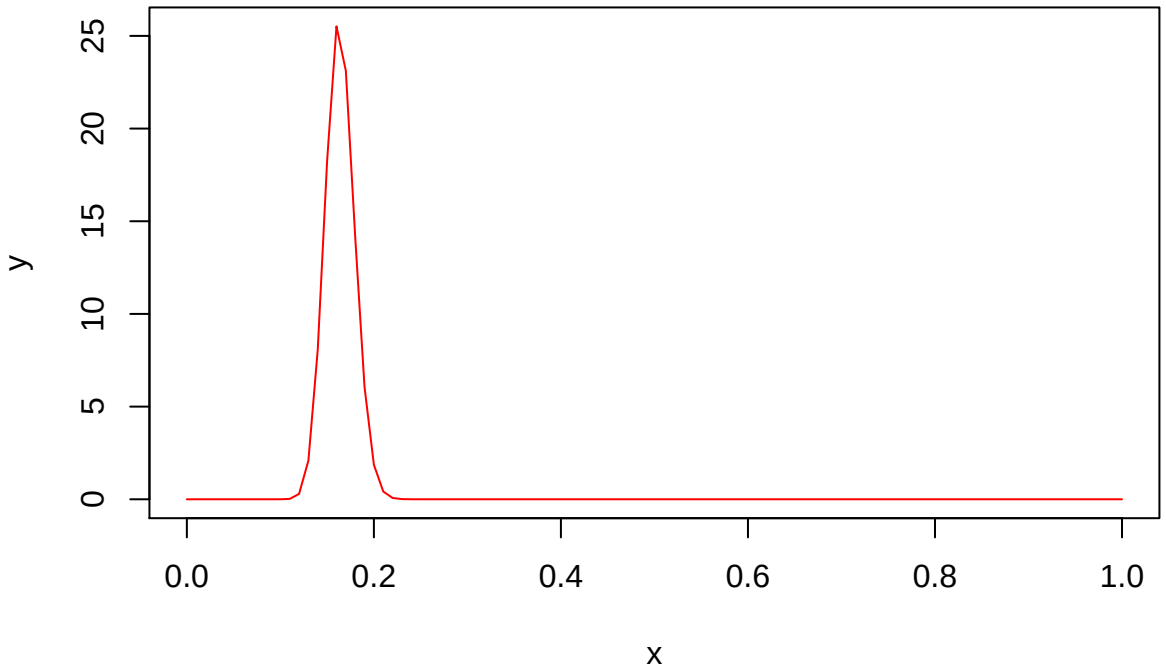

```
info[1]
```

```
## [1] "5%, 95%, posterior median, prior proportion, amount of information"
```

```
results_prior_3
```

```
##          5%         95%         95%
##   0.1366320   0.1922911   0.1644615   0.1666667 200.0000000
```

```
#Results for a new prior: Beta(50, 250)
results_prior_4 <- plot_posterior(50, 250, algae, 0.9)
```



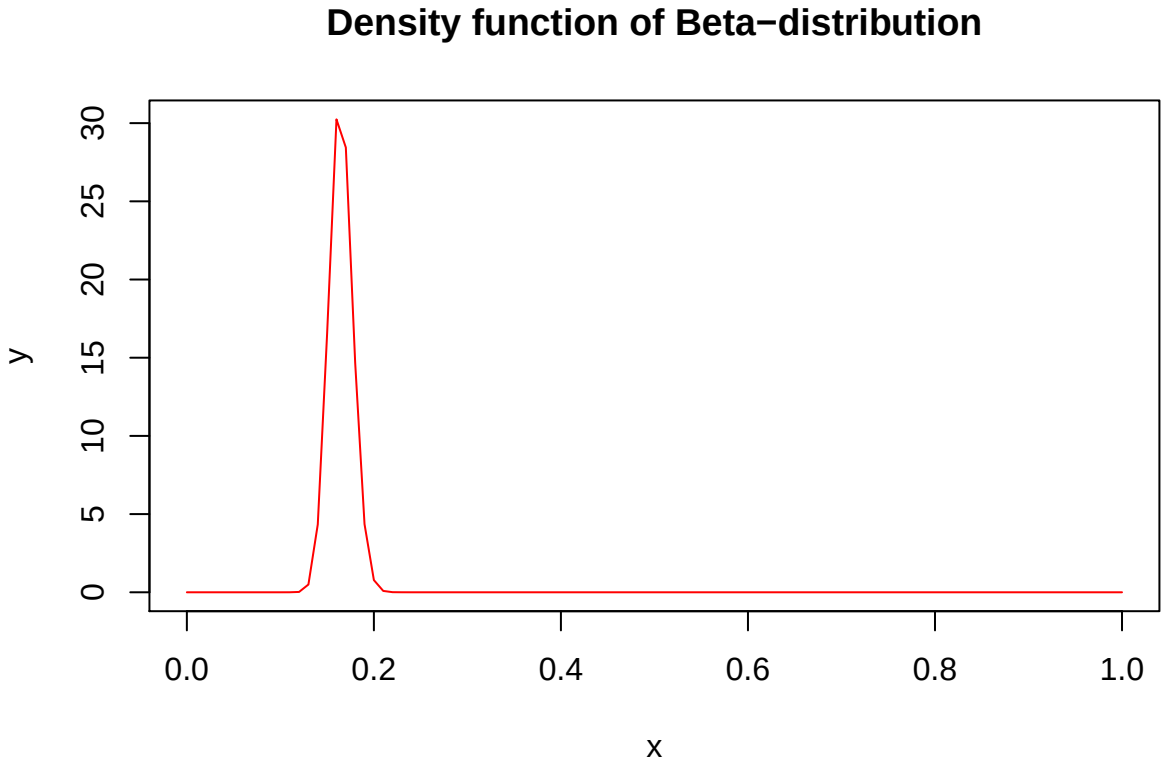

```
info[1]
```

```
## [1] "5%, 95%, posterior median, prior proportion, amount of information"
```

```
results_prior_4
```

```
##          5%         95%         95%
##   0.1391430   0.1901153   0.1646291   0.1666667 300.0000000
```

```
#Results for a new prior: Beta(100, 500)
results_prior_5 <- plot_posterior(100, 500, algae, 0.9)
```




```
info[1]
```

```
## [1] "5%, 95%, posterior median, prior proportion, amount of information"
```

```
results_prior_5
```

```
##          5%         95%         95%
##   0.1451657   0.1850860   0.1651259   0.1666667 600.0000000
```

The previous data, provided for each of the plots, is composed by the 90% posterior interval, the posterior median (the second 95% title can be ignored, it is due to the units), the prior proportion, given by $\frac{\alpha}{\alpha+\beta}$ and the amount of prior information, estimated by $\alpha + \beta$. As it can be expected, the higher the amount of information, the closer the posterior median is to the prior mean. 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.