BDA - Assignment 2

Anonymous

Table of Contents

Exercise 1: Inference for binomial proportion

```
library(aaltobda)
data("algae")
algae_test <- c(0, 1, 1, 0, 0, 0)
```

Let π be the probability of a monitoring site having detectable blue-green algae levels and y the observations in algae. Use a binomial model for the observations y and a Beta(2; 10) prior for binomial model parameter π to formulate a Bayesian model. Here it is not necessary to derive the posterior distribution for π as it has already been done in the book and it suffices to refer to that derivation. Also, it is not necessary to write out the distributions; it is sufficient to use label-parameter format, e.g. Beta(·,·).

a)

First, we need to determine n and y:

```
print("y is: ")

## [1] "y is: "

y <- sum(algae == 1)
print(y)

## [1] 44

print("n is: ")

## [1] "n is: "
```

```
n <- length(algae)
print(n)

## [1] 274
```

(1) the likelihood $p(\pi|y)$ as a function of π The likelihood is given as:
$p(\pi|y) = \binom{n}{y}\pi^y(1-\pi)^{n-y} = \binom{274}{44}\pi^{44}(1-\pi)^{230}$ (answer)

(2) the prior $p(\pi)$

$$p(\pi) = Beta(2,10)$$

(3) the resulting posterior $p(\pi|y)$. Report the posterior in the format Beta(·,·), where you replace ·'s with the correct numerical values.
The posterior is $Beta(\theta|\alpha + y; \beta + n - y) = Beta(2 + 44; 10 + 274 - 44) = Beta(46,240)$ (answer)

b)

What can you say about the value of the unknown π according to the observations and your prior knowledge? Summarize your results with a point estimate (i.e. $E(\pi|y)$) and a 90% posterior interval.
Note! Posterior intervals are also called credible intervals and are different from confidence intervals.
Note! In your report, use the values from the data algae, not algae_test.

```
beta_point_est <- function(prior_alpha, prior_beta, data){
  count_y = sum(data == 1)
  posterior_mean <- (prior_alpha + count_y) / (prior_alpha + prior_beta + len
gth(data))
  return(posterior_mean)
}

print("The point estimate for the algae data is: ")

## [1] "The point estimate for the algae data is: "

beta_point_est(prior_alpha = 2, prior_beta = 10, data = algae)

## [1] 0.1608392

print("The point estimate for the algae_test data is: ")

## [1] "The point estimate for the algae_test data is: "

beta_point_est(prior_alpha = 2, prior_beta = 10, data = algae_test)

## [1] 0.2222222
```

```r
beta_interval <- function(prior_alpha, prior_beta, data, prob){
  count_y <- sum(data == 1)
  n <- length(data)
  sequence <- seq(0, 1, length=100)
  posterior_distribution = dbeta(sequence, prior_alpha + count_y, prior_beta
+ n - count_y)
  sample <- rbeta(10000000, prior_alpha + + count_y, prior_beta + n - count_y
)
  upper <- 1 - (1 - prob)/2
  lower <- (1 - prob)/2
  range <- quantile(sample, probs=c(lower, upper))
  #plot(seq(0, 1, length=100), posterior_distribution, ylab='density',
     #type ='l', col='purple', main='Beta Distribution')
  return(range)
}

print("The posterior intervals for the algae data is: ")

## [1] "The posterior intervals for the algae data is: "

beta_interval(prior_alpha = 2, prior_beta = 10, data = algae, prob = 0.9)

##        5%       95%
## 0.1265498 0.1978128

print("The posterior intervals for the algae_test data is: ")

## [1] "The posterior intervals for the algae_test data is: "

beta_interval(prior_alpha = 2, prior_beta = 10, data = algae_test, prob = 0.9
)

##         5%        95%
## 0.08467976 0.39564800

# mark_my_assignment()
```

The value of the unknown π according to the observations and the prior knowledge is that, with the right posterior, mean, median, and mode are all approximately 0.1608392. Additionally, the posterior 90% probability intervals, as calculated above, is approximately [0.1265652, 0.1978060]

c)

What is the probability that the proportion of monitoring sites with detectable algae levels π is smaller than $\pi_0 = 0.2$ that is known from historical records?

This part requires the computation of the CDF of the posterior Beta function. It can be implemented as follows

```
beta_low <- function(prior_alpha, prior_beta, data, pi_0){
  step <- 0.01
  sequence <- seq(0, 1, by = step)
  index <- round((pi_0/step)) + 1
  count_y <- sum(data == 1)
  n = length(data)
  posterior_pbeta <- pbeta(sequence, prior_alpha + count_y, prior_beta + n -
count_y)
  #print(x_pbeta)
  #print(posterior_pbeta)
  return(posterior_pbeta[index])
}

print("The posterior intervals for the algae data is: ")

## [1] "The posterior intervals for the algae data is: "

beta_low(prior_alpha = 2, prior_beta = 10, data = algae, pi_0 = 0.2)

## [1] 0.9586136

print("The posterior intervals for the algae_test data is: ")

## [1] "The posterior intervals for the algae_test data is: "

beta_low(prior_alpha = 2, prior_beta = 10, data = algae_test, pi_0 = 0.2)

## [1] 0.4511238

# mark_my_assignment()
```

The probability that the proportion of monitoring sites with detectable algae levels smaller than $\pi_0 = 0.2$ is therefore 0.9586136 (answer)

d)

What assumptions are required in order to use this kind of a model with this type of data? (No need to discuss exchangeability yet)

The required assumptions to use the binomial model is:
Assumption 1: Each observation only has two possible outcomes.
In this case, each lakes has algae or does not have algae. If they have more than two outcomes, then the observations no longer follow the binomial model

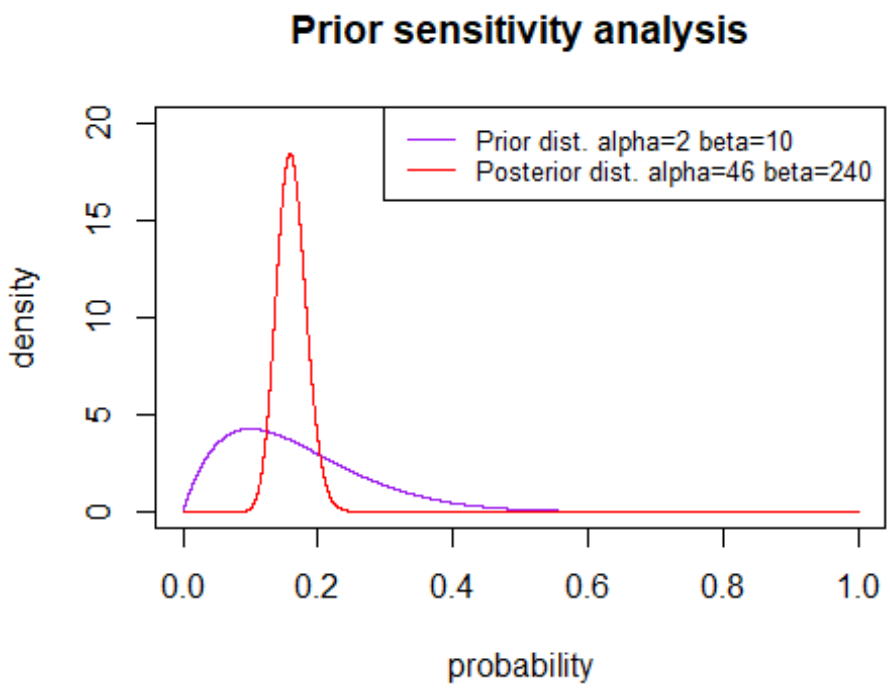
Assumption 2: The probability of success is the same for each observation.
In other words, the chance of having algae is equal for all lakes. Otherwise, π will vary depending on the lakes and can't be formulated with the binomial model

Assumption 3: Each observation is independent.
The independence of each observation means that whether some lakes have algae does not affect the chance of other lakes having algae. If there is interdependence between the observations, the formula $\pi^y(1-\pi)^{n-y}$ cannot be applied.

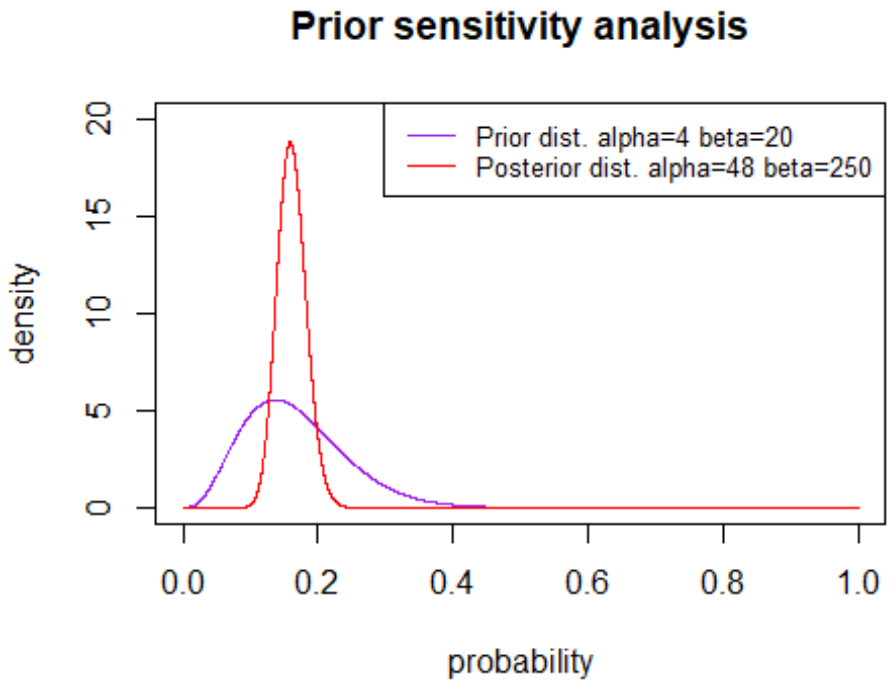
e)

Make prior sensitivity analysis by testing a couple of different reasonable priors and plot the different posteriors. Summarize the results by one or two sentences.
Hint! With a conjugate prior, a closed-form posterior has Beta form (see equations in the book and in the slides). Useful functions: dbeta, pbeta, qbeta in R.

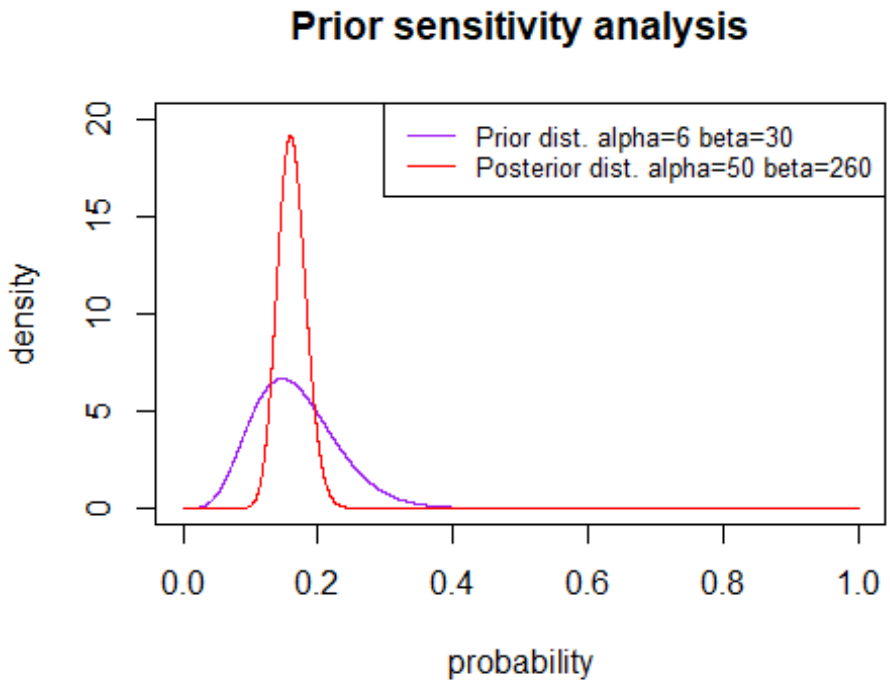
```
library(stringr)
y <- sum(algae == 1)
n <- length(algae)

plotBeta <- function(prior_alpha, prior_beta, n, y){
  posterior_alpha <- prior_alpha + y
  posterior_beta <- prior_beta + n - y
  sequence <- seq(0, 1, length=10000)
  prior_distribution <- dbeta(sequence, prior_alpha, prior_beta)
  posterior_distribution <- dbeta(sequence, posterior_alpha, posterior_beta)
  plot(sequence, prior_distribution, ylab='density',xlab="probability",type =
'l', col='purple', main='Prior sensitivity analysis', ylim = c(0, 20))
  lines(sequence, posterior_distribution, col="red")
  legend(x="topright",y="topright",legend=c(str_glue("Prior dist. alpha={prio
r_alpha} beta={prior_beta}"), str_glue("Posterior dist. alpha={posterior_alph
a} beta={posterior_beta}")),
       col=c("purple", "red"), lty=1:1, cex=0.8)
}
plotBeta(2,10, n, y)
```



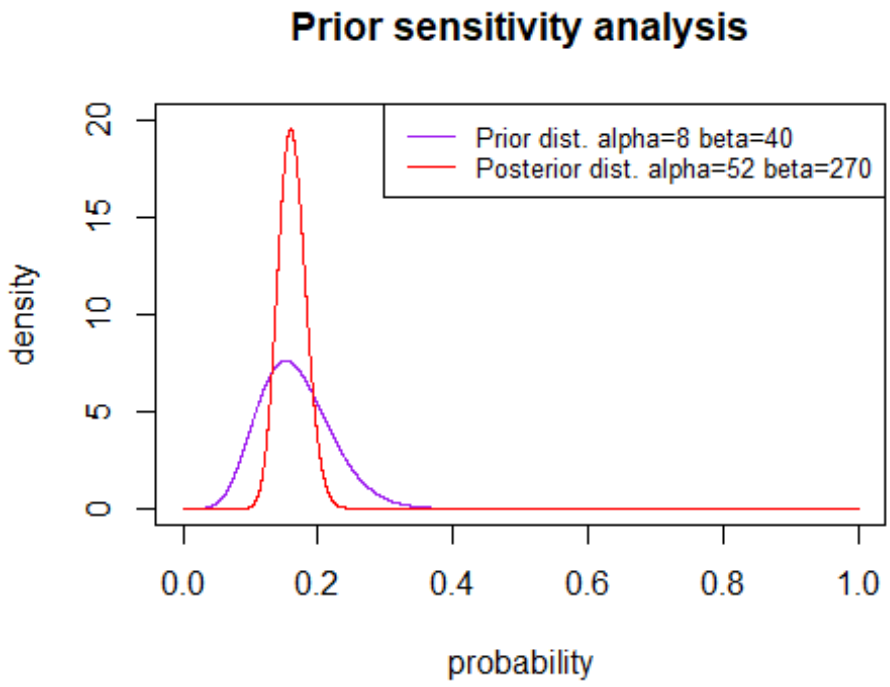

```
plotBeta(4,20, n, y)
```




```
plotBeta(6,30, n, y)
```




```
plotBeta(8,40, n, y)
```



Conclusion: as we increased the samples and keep the same success rate for the prior distributions, the posterior distribution curve gets sharper around the expected value. This means that the posterior distribution increases the density around the expected value, enforcing the belief derived from the prior distribution.