BDA - Assignment 2

Anonymous

21 9 2019

Contents

Used libraries:

```
library(dplyr)
library(ggplot2)
library(aaltobda)
data("algae")
```

Exercise 1

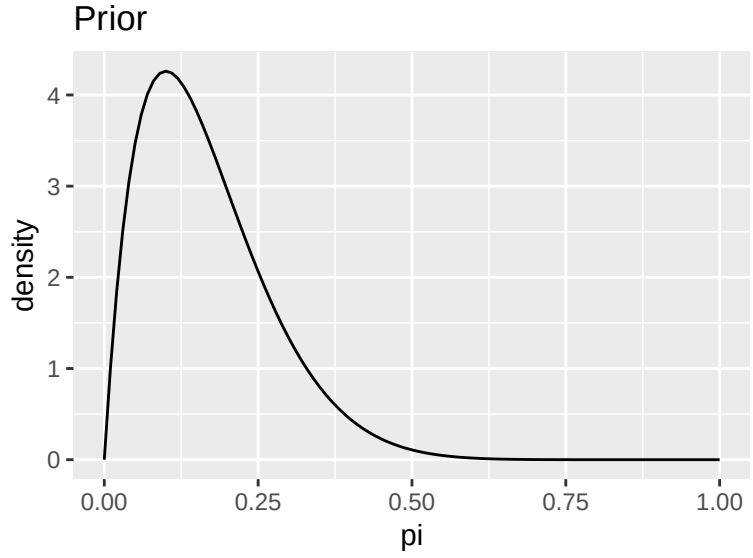
a) What can you say about the value of the unknown π according to the observations and your prior knowledge? Summarize your results with a point estimate (i.e. $E(\pi|y)$) and a 90% interval estimate.

Prior:

$$\pi \sim Beta(2, 10)$$

(this can be interpreted as binomial distribution with n=9, y=1)

```
qplot(x=seq(0,1,by = 0.01),dbeta(seq(0,1,by = 0.01),2,10), geom='line', xlab = 'pi', ylab='density', ma
```



Purely based on observations, one might conclude: $\pi = \frac{y}{n} \sim$ `0.1605839`, and from the prior alone (mean) $\frac{\alpha}{\alpha+\beta} \sim$ `0.1666667`. That is to say, they are somewhat close. However, a more informative estimate can be

acquired by using $E(\pi)$ over posterior distribution. Since that was already derived in the course materials, I am taking the results as given here:

$$E(\pi|y,n) = \int_0^1 \pi p(\pi|y,n)d\pi = \frac{\alpha + y}{\alpha + \beta + n}$$

I'll write a simple function in the form *markmyassignment* accepts:

```
algae_test <- c(0,1,1,0,0,0) #just for testing

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

  (prior_alpha + sum(data))/(prior_alpha + prior_beta + length(data))

}

res_a1 <- beta_point_est(data=algae)
print(paste('Point estimate for the whole algae data: ', res_a1))
```

```
## [1] "Point estimate for the whole algae data:  0.160839160839161"
```

To calculate the 90% interval (assuming centered), I can use the non-scaled (normalized) posterior distribution:

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

Again, I'll write a simple function in the form *markmyassignment* accepts:

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

  p_low <- (1 - prob)/2
  p_high <- 1 - p_low

  new_alpha <- prior_alpha + sum(data)
  new_beta <- prior_beta + length(data) - sum(data)

  qbeta(p = c(p_low, p_high), new_alpha, new_beta)

}

res_a2 <- beta_interval(data=algae)
print(paste('Interval for the whole algae data: ', res_a2))
```

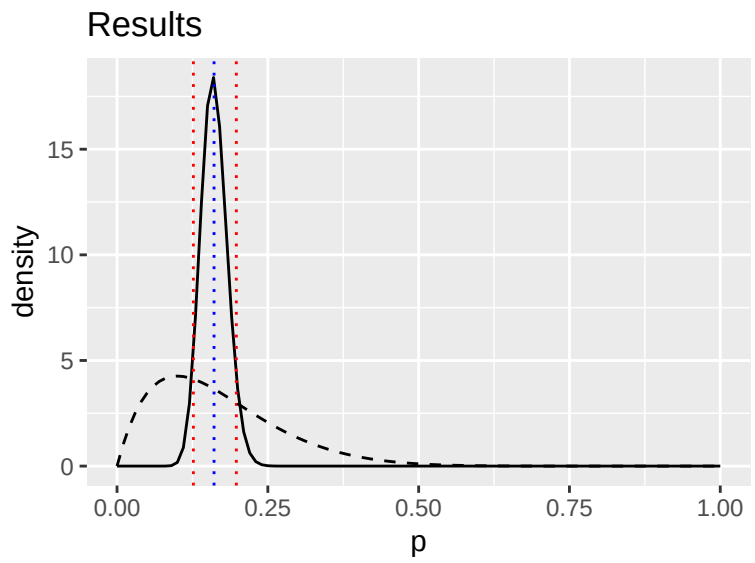
```
## [1] "Interval for the whole algae data:  0.126560711878773"
## [2] "Interval for the whole algae data:  0.197817667316324"
```

Plot for the results:

```
dat <- cbind(p=seq(0,1,by=0.01), density=dbeta(seq(0,1,by=0.01), 2 + sum(algae), 10 + length(algae) - su
dat_prior <- cbind(p=seq(0,1,by=0.01), density=dbeta(seq(0,1,by=0.01), 2, 10))


p <- ggplot(data = dat %>% as.data.frame()) +
  geom_line(aes(x=p, y=density)) +
  geom_vline(xintercept = res_a2[1], linetype = 'dotted', color = 'red') +
  geom_vline(xintercept = res_a2[2], linetype = 'dotted', color = 'red') +
  geom_vline(xintercept = res_a1, linetype = 'dotted', color = 'blue') +
  ggtitle('Results')
```

```
p + geom_line(data = dat_prior %>% as.data.frame(), aes(x=p, y=density), color = 'black', linetype = 'da
```



(Black dashed line is the prior, reds are the quantiles, and blue is the point estimate)

Results: Point estimate for π ~0.1608392 is a compromise between data (mean 0.1605839) and prior (mean 0.1666667). And since the prior can be seen as a distribution of prior experiment with n = 9, y = 1, it is no surprise to see the result being more strongly influenced by the data (n = 274, y = 44). Interval [0.1265607,0.1978177] seems to hold both prior and data estimates as well.

b) What is the probability that the proportion of monitoring sites with detectable algae levels π is smaller than $\text{pi}_0 = 0.2$ that is known from historical records?

By using posterior probability $\pi|y,n \propto Beta(\alpha+y, \beta+n-y)$, the 0.2 confidence can be calculated as a normal integral $\int_0^{0.2} p(\pi|y,n)d\pi$. The normalization coefficient could be calculated approximately in r as a sum but since we have a statistical program at our disposal, it is easier to just use it. Here is a simple function that takes advatage of r's pbeta(...) function:

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

  new_alpha <- prior_alpha + sum(data)
  new_beta <- prior_beta + length(data) - sum(data)

  pbeta(pi_0, new_alpha, new_beta)

}

res_b <- beta_low(data=algae)
print(paste('probability for pi < 0.2 for the whole algae data: ', res_b))
```

```
## [1] "probability for pi < 0.2 for the whole algae data:  0.958613587194853"
```

Result: The probability of $\pi < 0.2$ (assuming given data and prior) is roughly 0.96, or 96%. And using the previous results (a), one could quickly tell based on 90 quantile, that the result would have to be more than 95%.

c) What assumptions are required in order to use this kind of a model with this type of data?

Use of binomial model (series of Bernoulli samples) necessitates independence of samples, i.e., the sites where the algae samples are taken should have no dependencies; sample from site a has no effect on sample at site b. Further more, the sites are considered "equal" (exchangable) and each having the same likelihood of having algae (label 1). In other words, the samples are Bernoulli and i.i.d.

d) Make prior sensitivity analysis by testing a couple of different reasonable priors. Summarize the results by one or two sentences.

I will use uniform prior (this corresponds to giving more weight to data), and some Beta priors with similar mean but increasing amount of samples (this corresponds to giving increasingly more weight to prior). And finally I will look at two 0.5 centered Beta priors with both small and large α, β (respectively: less/more weight to prior).

For plotting the results, I will use the following function (pretty much copy-pasted from the earlier task a):

```
plot_comparison <- function(A,B) {
  dat <- cbind(p=seq(0,1,by=0.01), density=dbeta(seq(0,1,by=0.01), A + sum(algae), B + length(algae) - s
  dat_prior <- cbind(p=seq(0,1,by=0.01), density=dbeta(seq(0,1,by=0.01), A, B))

  ival <- beta_interval(A,B,data=algae)
  point_est <- beta_point_est(A,B,data=algae)

  p <- ggplot(data = dat %>% as.data.frame()) +
    geom_line(aes(x=p, y=density)) +
    geom_vline(xintercept = ival[1], linetype = 'dotted', color = 'red') +
    geom_vline(xintercept = ival[2], linetype = 'dotted', color = 'red') +
    geom_vline(xintercept = point_est, linetype = 'dotted', color = 'blue') +
    ggtitle(paste('Results for alpha = ', A, ' beta = ', B, sep=''))

  p + geom_line(data = dat_prior %>% as.data.frame(), aes(x=p, y=density), color = 'black', linetype =
}
```

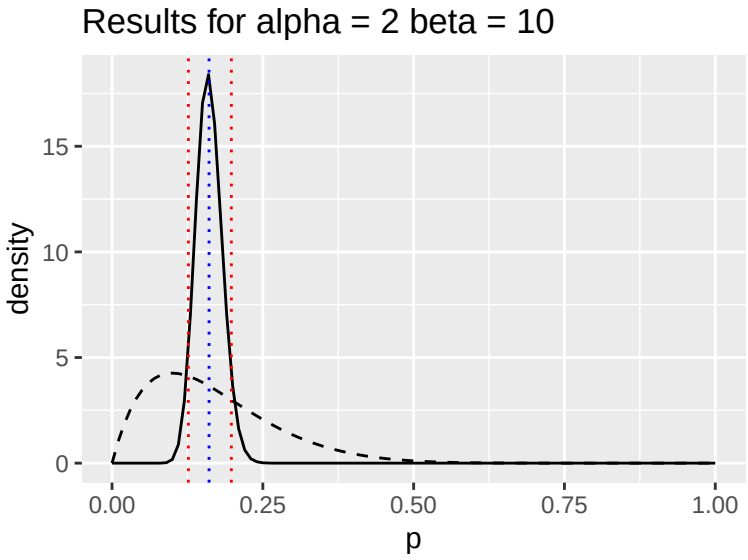
Data:

- n = $length(algae)$
- y = $sum(algae)$
- data mean = $\frac{y}{n}$=0.1605839

Original setup, $\alpha = 2, \beta = 10$**:**

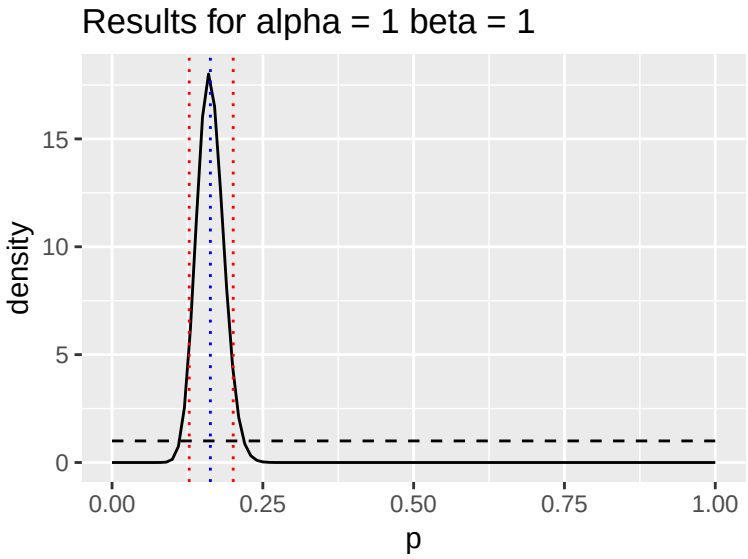
- prior mean = $\frac{\alpha}{\alpha+\beta} = \frac{2}{12}$ =0.1666667
- $E(\pi|y,n) = \frac{\alpha+y}{\alpha+\beta+n} \sim 0.1608$

In plots, the black dashed line is the prior



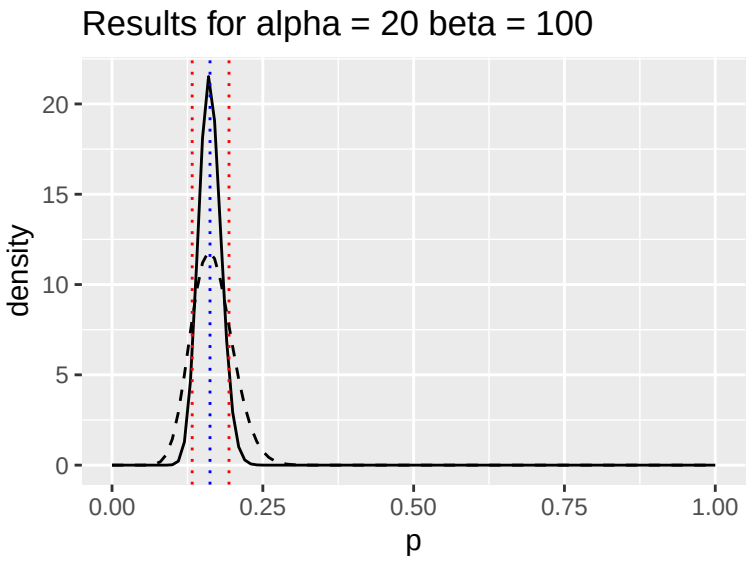
uniform prior, $\alpha = 1, \beta = 1$**:**

- prior mean $= \dfrac{1}{2}$
- $E(\pi|y,n) = \dfrac{y+1}{n+2}$ ~0.1630435



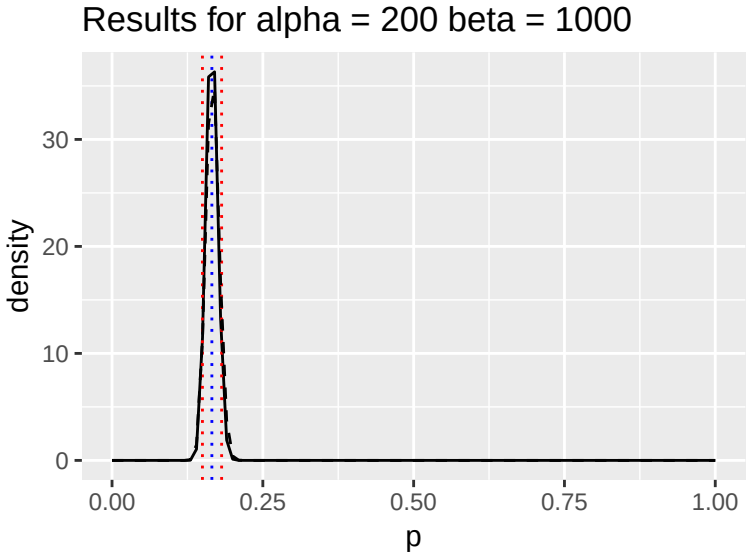
Beta prior, $\alpha = 20, \beta = 100$

- prior mean $= \dfrac{\alpha}{\alpha+\beta} = \dfrac{2}{12}$ =0.1666667
- $E(\pi|y,n) = \dfrac{\alpha+y}{\alpha+\beta+n}$ ~0.1624365



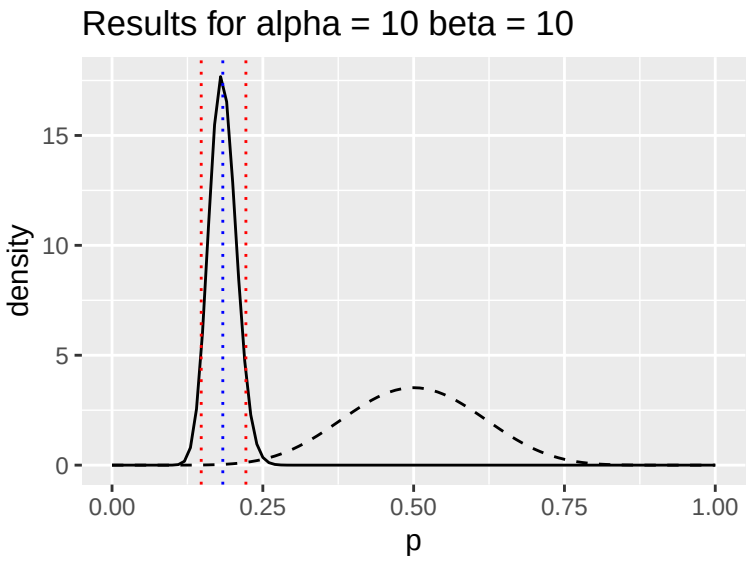
Beta prior, $\alpha = 200, \beta = 1000$

- prior mean $= \dfrac{\alpha}{\alpha+\beta} = \dfrac{2}{12}$ =0.1666667
- $E(\pi|y,n) = \dfrac{\alpha+y}{\alpha+\beta+n}$ ~0.165536



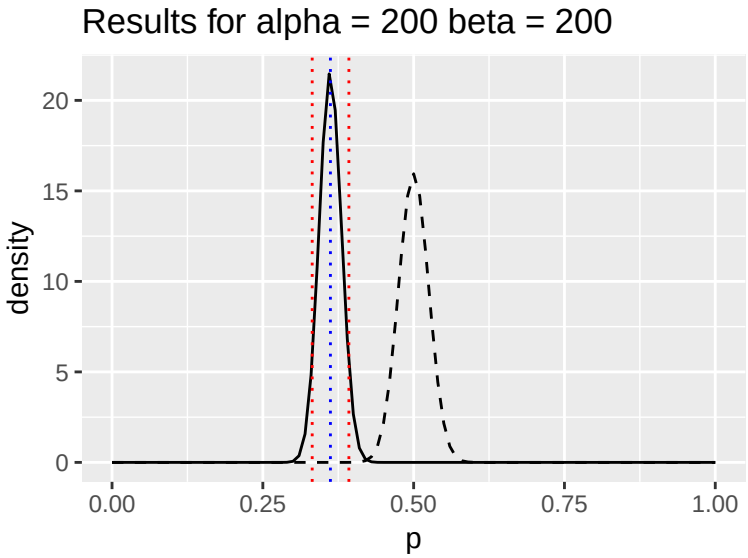
Beta prior, $\alpha = 10, \beta = 10$

- prior mean $= \dfrac{\alpha}{\alpha+\beta} = \dfrac{10}{20}$ =0.5
- $E(\pi|y,n) = \dfrac{\alpha+y}{\alpha+\beta+n}$ ~0.1836735



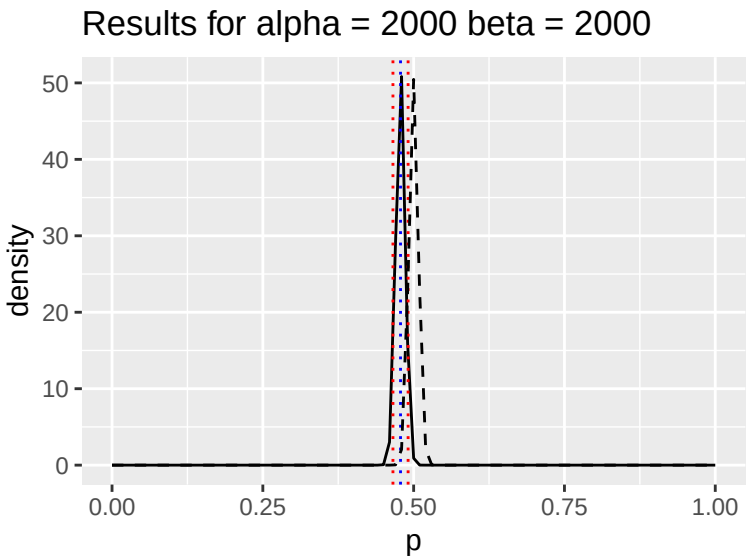
Beta prior, $\alpha = 200, \beta = 200$

- prior mean $= \frac{\alpha}{\alpha+\beta} = \frac{200}{400} =$`0.5`
- $E(\pi|y,n) = \frac{\alpha+y}{\alpha+\beta+n} \sim$`0.3620178`



Beta prior, $\alpha = 2000, \beta = 2000$

- prior mean $= \frac{\alpha}{\alpha+\beta} = \frac{2000}{4000} =$`0.5`
- $E(\pi|y,n) = \frac{\alpha+y}{\alpha+\beta+n} \sim$`0.3620178`



SA, Interpretation

Sensitivity analysis, wrt. to the prior, shows clearly how the prior starts to have strong effect only as α, β begin to grow closer to the size of the data. This is easier to understand if one remembers that $\alpha - 1, \beta - 1$ can be considered to be prior observations - this is direct result of Beta being conjugate prior (the model behaviour stays consistent).

Worth noting is, that even though uniform prior has the same mean 0.5 as does the balanced priors (20,20 and 200,200 and 2000,2000), the latter ones have a considerable effect on the posterior (increasing in effect with "prior" samples), where as the uniform distribution (by definition) has no effect at all.

The more there is agreement between prior and data, the narrower the 90% confidence gets, and more peaked the posterior becomes. This effect can also occur when there is far stronger prior (2000,2000) when it totally dominates the posterior - again, this corresponds to having a large prior sample and hence `2044`, `2274` over all y, n.

All this seems pretty standard behaviour.