BSDA: Assignment 2

Anonymous student

Contents

General Information to include **1**

1. Inference for binomial proportion (Computer) **1**

Appendix: Code for plots in e) **6**

General Information to include

- **Time used for reading and self-study exercises**: ~11 hours.
- **Time used for the assignment**: ~8 hours
- **Good with assignment**: The level of the questions were just right and really pedagogical to understand the concepts, I was reading the book while answering the assignment and everything made sense, even the questions were in the same order as the book presented the concepts.
- **Things to improve in the assignment**: I feel like the instructions in the last question were a little vague I tried to plot exactly this: "testing a couple of different reasonable priors and plot the different posteriors".

1. Inference for binomial proportion (Computer)

```
library(bsda)
library(reshape2)
library(tidyverse)
data("algae")
head(algae)
```

```
## [1] 0 1 1 0 0 0
```

a) formulate (1) the likelihood $p(y|\pi)$ as a function of π, (2) the prior $p(\pi)$, and (3) the resulting posterior $p(\pi|y)$. Report the posterior in the format Beta$(\cdot, \cdot)$, where you replace $\cdot$'s with the correct numerical values.

1. Likelihood:

Considered as a function of π, the likelihood is of the form

$$p(y|\pi, n) = p(y|\pi) = Bin(y|n, \pi) = \binom{n}{y} \pi^y (1-\pi)^{n-y}$$
$$\propto \pi^y (1-\pi)^{n-y} \propto \pi^{44}(1-\pi)^{230}$$

where on the second equality and in the rest of the answers we can suppress the dependence on n because it is considered fixed.

2. Prior:

Let's assume that the prior distribution for π is the suggested beta. Then

$$p(\pi|n) = p(\pi) = Beta(\pi|\alpha = 2, \beta = 10) = \frac{\pi^{\alpha-1}(1-\pi)^{\beta-1}}{B(\alpha,\beta)} = \frac{\pi(1-\pi)^9}{B(2,10)}$$
$$\propto \pi^{\alpha-1}(1-\pi)^{\beta-1} \propto \pi(1-\pi)^9$$

where

$$B(\alpha = 2, \beta = 10) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} = \frac{\Gamma(2)\Gamma(10)}{\Gamma(12)}$$

3. Posterior:

Using Bayes' Rule and substituting we can derive the posterior

$$p(\pi|y,n) = p(\pi|y) = \frac{p(y|\pi)}{p(y)}$$
$$\propto \pi^y(1-\pi)^{n-y}\pi^{\alpha-1}(1-\pi)^{\beta-1} \propto \pi^{y+\alpha-1}(1-\pi)^{n-y+\beta-1} \propto \pi^{45}(1-\pi)^{239}$$

where

$$p(y) = \int_0^1 p(y|\pi)p(\pi)\,d\pi$$

After normalization

$$p(\pi|y) = Beta(\pi|\alpha + y, \beta + n - y) = Beta(\pi|46, 240)$$

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.

```
#define function to get E(pi|y)
beta_point_est <- function(prior_alpha, prior_beta, data) {

  n <- length(data)
  y <- sum(data)

  posterior_mean <- (prior_alpha + y)/(prior_alpha + prior_beta + n)

  return(posterior_mean)
}
point_est <- beta_point_est(prior_alpha = 2, prior_beta = 10, data = algae)
point_est
```

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

```
#define function to get posterior interval
beta_interval <- function(prior_alpha, prior_beta, data, prob) {

  posterior_alpha <- prior_alpha + sum(data)
  posterior_beta <- prior_beta + length(data) - sum(data)
```

```
  lower <- qbeta((1-prob)/2, shape1 = posterior_alpha, shape2 = posterior_beta)
  upper <- qbeta(1-(1-prob)/2, shape1 = posterior_alpha, shape2 = posterior_beta)

  return(c(lower, upper))
}

interval <- beta_interval(prior_alpha = 2, prior_beta = 10, data = algae, prob = 0.9)
interval
```

```
## [1] 0.1265607 0.1978177
```

The posterior mean of π, can be interpreted as the posterior probability of a monitoring site having detectable blue-green algae levels for a future observation from the population

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

which will lie between the sample proportion, $y/n = 0.1606$, and the prior mean, $\alpha/(\alpha + \beta) = 0.1667$

We can construct a posterior interval in the following way:

Let F_{Beta} denote the CDF of our $Beta(46, 240)$ distribution. Then, we can choose $F^{-1}_{Beta}(0.05) \approx 0.1266$ and $F^{-1}_{Beta}(0.95) \approx 0.1978$, then our credible interval is $[0.1266, 0.1978]$. There is a 90% probability that π falls in [0.1266, 0.1978].

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?

```
beta_low <- function(prior_alpha, prior_beta, data, pi_0) {

  posterior_alpha <- prior_alpha + sum(data)
  posterior_beta <- prior_beta + length(data) - sum(data)

  prob <- pbeta(pi_0, shape1 = posterior_alpha, shape2 = posterior_beta)

  return(prob)
}

prob_pi <- beta_low(prior_alpha = 2, prior_beta = 10, data = algae, pi_0 = 0.2)
prob_pi
```

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

To compute $Pr(\pi < 0.2|y)$ we can use F_{Beta} which we defined as the CDF of our $Beta(46, 240)$ distribution. Then, $F_{Beta}(0.2) \approx 0.9586$, so we have

$$Pr(\pi < 0.2|y) = 0.9586$$

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

To use a binomial model with a Beta prior for the parameter π in this context, some assumptions are required, I will try to mention the most relevant ones:

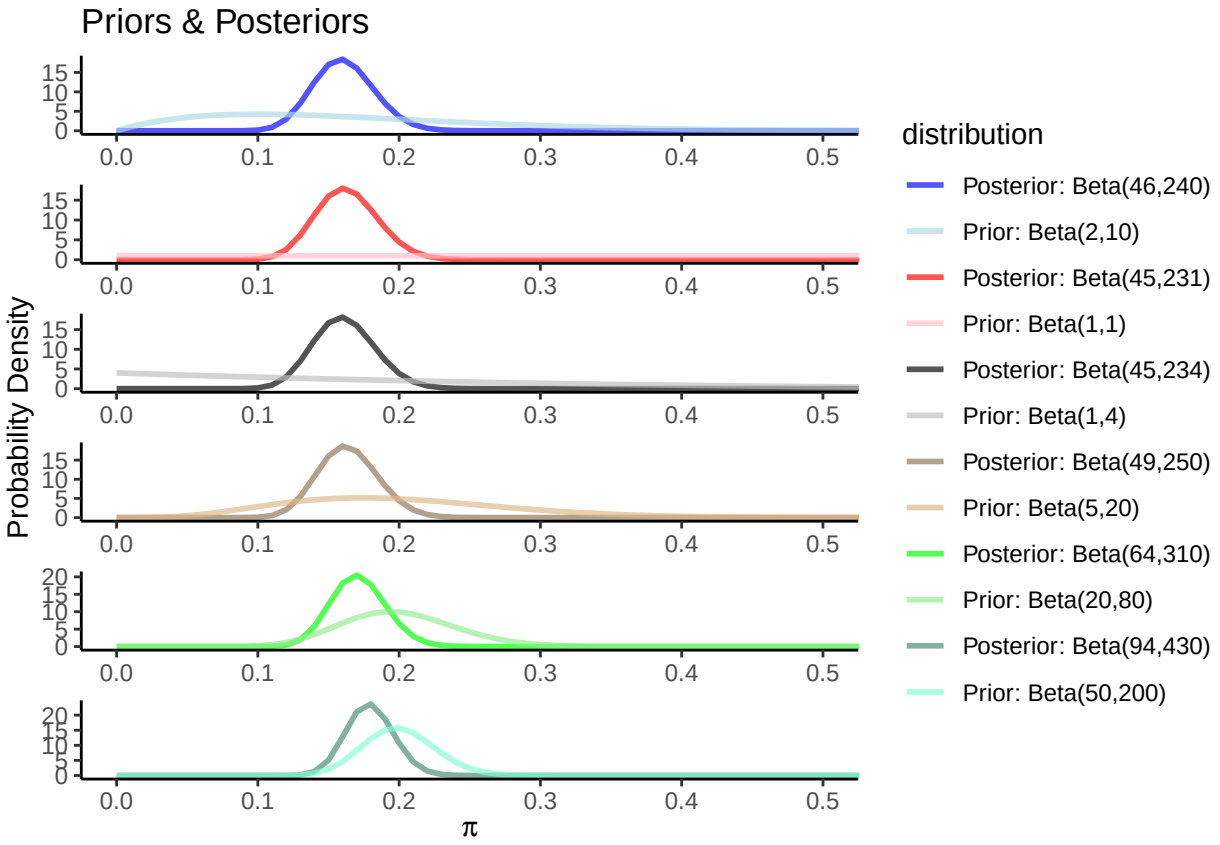
1. The chosen binomial likelihood function is considered a reasonable approximation to describe the data.

2. It is assumed that observations at different monitoring sites are conditionally independent given π, with the probability of presence of algae equal to π for all cases. This assumption implies that the presence or absence of algae at one site does not influence the presence or absence of algae at another site.

3. Exchangeability (not discussed)

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.

I used a uniform prior and then priors increasingly concentrated around 0.2 (which is derived from historical records according to question c)).



Prior distribution		Posterior distribution	
$\frac{\alpha}{\alpha+\beta}$	$\alpha+\beta$	Posterior mean of π	90% posterior interval for π
0.1666667	12	0.1608392	[0.1265607, 0.1978177]
0.5	2	0.1630435	[0.1279681, 0.2008987]
0.2	5	0.1612903	[0.1265613, 0.1987836]
0.2	25	0.1638796	[0.1300439, 0.2002747]
0.2	100	0.171123	[0.140159, 0.2040887]
0.2	250	0.1793893	[0.152556, 0.207615]

The posterior seems not to be very sensible to the prior distribution, we can see in the plots that the posterior is barely changing. Only when the priors contain information equivalent to more than 100 observations of

algae status, our posterior distribution and intervals start to be pulled towards 0.2; therefore, prior mean of 0.2 seems quite unlikely.

Appendix: Code for plots in e)

```
new_alpha_1 <- 1
new_beta_1 <- 4

new_alpha_2 <- 2
new_beta_2 <- 8

new_alpha_2 <- 5
new_beta_2 <- 20

new_alpha_3 <- 20
new_beta_3 <- 80

new_alpha_4 <- 50
new_beta_4 <- 200


priors <- list(
  dbeta(seq(0, 1, by = 0.01), 2, 10),#original prior
  dbeta(seq(0, 1, by = 0.01), 1, 1), #uniform(0,1)
  dbeta(seq(0, 1, by = 0.01), new_alpha_1, new_beta_1),
  dbeta(seq(0, 1, by = 0.01), new_alpha_2, new_beta_2),
  dbeta(seq(0, 1, by = 0.01), new_alpha_3, new_beta_3),
  dbeta(seq(0, 1, by = 0.01), new_alpha_4, new_beta_4)

)

posteriors <- list(
  dbeta(seq(0, 1, by = 0.01), 2+y, 10+n-y),#original posterior
  dbeta(seq(0, 1, by = 0.01), 1+y, 1+n-y), #posterior using uniform(0,1)
  dbeta(seq(0, 1, by = 0.01), new_alpha_1+y, new_beta_1+n-y),
  dbeta(seq(0, 1, by = 0.01), new_alpha_2+y, new_beta_2+n-y),
  dbeta(seq(0, 1, by = 0.01), new_alpha_3+y, new_beta_3+n-y),
  dbeta(seq(0, 1, by = 0.01), new_alpha_4+y, new_beta_4+n-y)

)

# save results in tibble
results <- tibble(pi = seq(0, 1, by = 0.01))

# add posteriors for each prior
for (i in 1:length(priors)) {
  results[paste("Prior",i)] <- priors[[i]]
  results[paste("Posterior",i)] <- posteriors[[i]]
}

# create string of new_prior and new_posterior
new_prior_1 <- paste("Prior: Beta(", new_alpha_1, ",", new_beta_1, ")", sep = "")
new_prior_2 <- paste("Prior: Beta(", new_alpha_2, ",", new_beta_2, ")", sep = "")
new_prior_3 <- paste("Prior: Beta(", new_alpha_3, ",", new_beta_3, ")", sep = "")
new_prior_4 <- paste("Prior: Beta(", new_alpha_4, ",", new_beta_4, ")", sep = "")
```

```r
new_posterior_1 <- paste("Posterior: Beta(", new_alpha_1 + y,
                         ",", new_beta_1 + n - y, ")", sep = "")
new_posterior_2 <- paste("Posterior: Beta(", new_alpha_2 + y,
                         ",", new_beta_2 + n - y, ")", sep = "")
new_posterior_3 <- paste("Posterior: Beta(", new_alpha_3 + y,
                         ",", new_beta_3 + n - y, ")", sep = "")
new_posterior_4 <- paste("Posterior: Beta(", new_alpha_4 + y,
                         ",", new_beta_4 + n - y, ")", sep = "")

#reshape df to long version
results_long <- melt(results, id.vars = "pi",
                     variable.name = "distribution",
                     value.name = "density") %>%
  mutate(number = str_extract(distribution, "\\d+"),
         distribution = case_when(
           distribution == 'Prior 1' ~ 'Prior: Beta(2,10)',
           distribution == 'Prior 2' ~ 'Prior: Beta(1,1)',
           distribution == 'Prior 3' ~ new_prior_1,
           distribution == 'Prior 4' ~ new_prior_2,
           distribution == 'Prior 5' ~ new_prior_3,
           distribution == 'Prior 6' ~ new_prior_4,
           distribution == 'Posterior 1' ~ 'Posterior: Beta(46,240)',
           distribution == 'Posterior 2' ~ 'Posterior: Beta(45,231)',
           distribution == 'Posterior 3' ~ new_posterior_1,
           distribution == 'Posterior 4' ~ new_posterior_2,
           distribution == 'Posterior 5' ~ new_posterior_3,
           distribution == 'Posterior 6' ~ new_posterior_4
         ))

results_long$distribution <- factor(results_long$distribution,
                                    levels=c("Posterior: Beta(46,240)",
                                             "Prior: Beta(2,10)",
                                             "Posterior: Beta(45,231)",
                                             "Prior: Beta(1,1)",
                                             new_posterior_1,
                                             new_prior_1,
                                             new_posterior_2,
                                             new_prior_2,
                                             new_posterior_3,
                                             new_prior_3,
                                             new_posterior_4,
                                             new_prior_4
                                             ))

# create plot with priors and posteriors
ggplot(results_long, aes(x = pi, y = density, color = distribution)) +
  geom_line(size = 1, alpha = 0.67) +
  labs(title = "Priors & Posteriors",
       x = expression(pi),
       y = "Probability Density") +
  scale_color_manual(values = c("Prior: Beta(2,10)" = "lightblue",
                                "Prior: Beta(1,1)" = "pink",
                                "Prior: Beta(1,4)" = "grey",
```

```
                                "Prior: Beta(5,20)" = "burlywood",
                                "Prior: Beta(20,80)" = "lightgreen",
                                "Prior: Beta(50,200)" = "aquamarine",
                                "Posterior: Beta(46,240)" = "blue",
                                "Posterior: Beta(45,231)" = "red",
                                "Posterior: Beta(45,234)"= "black",
                                "Posterior: Beta(49,250)"= "burlywood4",
                                "Posterior: Beta(64,310)"= "green",
                                "Posterior: Beta(94,430)"= "aquamarine4"
                                )) +  coord_cartesian(xlim = c(0.00, 0.5)) +
facet_wrap(~number,ncol=1, scales = "free") + theme_classic() +
theme(strip.background = element_blank(),
      strip.text.x = element_blank()
)
```