BDA Assignment 2

Contents

Exercise 1

Data: Observations for algae status from 274 monitor sites in Finland from 2008. The data is in indicator values where '0': no algae and '1': algae present.

Using binomial model:

- π: probability of monitoring site having detectable algae levels
- y: number of 'algae present' observations

a)

(1) Likelihood: For binomial model, the likelihood is the binomial distribution $p(y|\pi) = \text{Bin}(n, \pi)$

(2) Prior: Prior is given as $p(\pi) = \text{Beta}(2, 10)$

(3) Posterior: In Bayesian binomial model when prior is $\text{Beta}(\alpha, \beta)$ then the posterior is $\text{Beta}(\alpha+y, \beta+n-y)$. In our model the posterior is then $p(\pi|y) = \text{Beta}(2 + y, 10 + n - y)$

From the data we have that

```
n <- length(algae)
y <- sum(algae)
p_alpha <- 2+y
p_beta <- 10+n-y
cat("$n=$", n, "$y=$", y)
```

$n =$ 274 $y =$ 44

So now our model has

$$p(y|\pi) = \text{Beta}(274, \pi)$$
$$p(\pi) = \text{Beta}(2, 10)$$
$$p(\pi|y) = \text{Beta}(46, 328)$$

b)

When looking at the data, there are about five times more lakes and rivers that do not have algae than those that do. As for my prior knowledge, it would make sense that some lakes or rivers would have blue-green algae but most wouldn't.

The posterior mean for binomial model with prior $\text{Beta}(\alpha, \beta)$ is $E[\theta|y] = \frac{\alpha+y}{\alpha+\beta+n}$.

A 90% posterior interval corresponds to the $[0.05, 0.95]$ quantile.

```
beta_point_est <- function(prior_alpha, prior_beta, data) {
  n <- length(data)
  y <- sum(data)
  (prior_alpha + y)/(prior_alpha + prior_beta + n)
}

beta_interval <- function(prior_alpha, prior_beta, data, prob) {
  n <- length(data)
  y <- sum(data)
  p <- c((1-prob)/2, (1+prob)/2)
  qbeta(p, prior_alpha+y, prior_beta+n-y)
}

E <- beta_point_est(prior_alpha = 2, prior_beta = 10, data = algae)
q <- beta_interval(prior_alpha = 2, prior_beta = 10, data = algae, prob = 0.9)

cat("The posterior mean is", E, " and the posterior interval is [", q, "]")
```

The posterior mean is 0.1608392 and the posterior interval is [0.1265607 0.1978177]

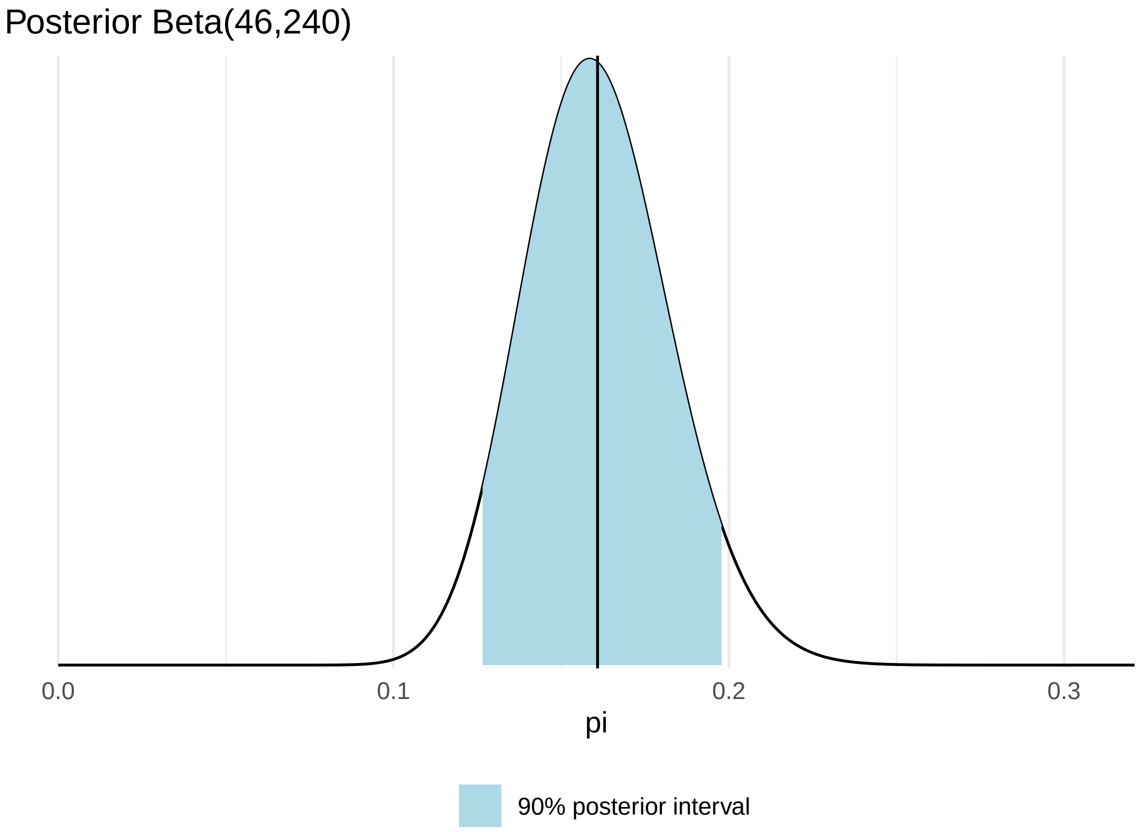
Visualization of the posterior distribution:

```
df1 <- data.frame(pi = seq(0, E*2, 0.001))
df1$p <- dbeta(df1$pi, p_alpha, p_beta)

df2 <- data.frame(pi = seq(q[1], q[2], length.out = 100))
df2$p <- dbeta(df2$pi, p_alpha, p_beta)

ggplot(mapping = aes(pi, p)) +
  geom_line(data = df1) +
  geom_area(data = df2, aes(fill='1')) +
  geom_vline(xintercept = E, linetype='solid') +

  labs(title='Posterior Beta(46,240)', y = '', x = 'pi') +
  scale_y_continuous(expand = c(0, 0.1), breaks = NULL) +
  scale_fill_manual(values = 'lightblue', labels = '90% posterior interval') +
  theme(legend.position = 'bottom', legend.title = element_blank())
```



c)

The probability that the proportion π is smaller than $\pi_0 = 0.2$ can be obtained from the CDF of our posterior.

```
beta_low <- function(prior_alpha, prior_beta, data, pi_0) {
  n <- length(data)
  y <- sum(data)
  pbeta(pi_0, prior_alpha+y, prior_beta+n-y)
}

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

cat("The probability that $\\pi < \\pi_0 = 0.2$ is", p)
```

The probability that $\pi < \pi_0 = 0.2$ is 0.9586136

d)

The binomial model assumes that the observations are i.i.d., that is, the observation points do not affect one another and they have an equal probability π to have algae. It is also assumed that the $\text{Beta}(2, 10)$ is an reasonable prior.

e)

To test how different priors would affect the results the following priors were chosen:

(1) Uniform prior (no prior knowledge assumed)

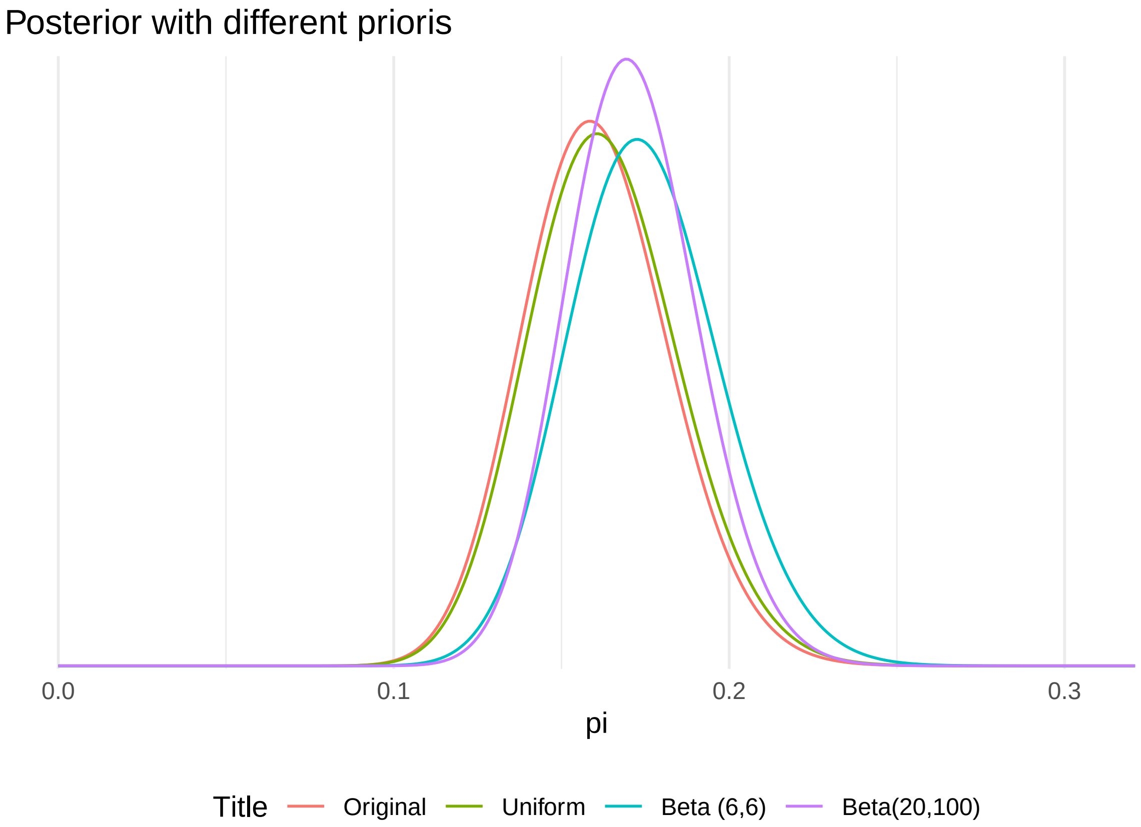
(2) Beta prior with as many observations (2+10-2=10) as original prior, but with proportion 0.5: Beta(6,6)
(3) Beta prior with same mean (1/5=0.2) as original prior, but with about ten times more observations: Beta(20,80)

Comparison of the different posteriors:

```
df3 <- data.frame(pi = seq(0, E*2, 0.001))
df3$p_0 <- dbeta(df3$pi, p_alpha, p_beta)   # original prior
df3$p_1 <- dbeta(df3$pi, y+1, n-y+1)        # uniform prior
df3$p_2 <- dbeta(df3$pi, 6+y, 6+n-y)        # Beta (6,6) prior
df3$p_3 <- dbeta(df3$pi, 20+y, 80+n-y)      # Beta(20,80) prior


df4 <- df3 %>%                                         # preparing the dataframe
  select(pi, p_0, p_1, p_2, p_3) %>%
  gather(key = "priori", value = "value", -pi)

ggplot(df4, aes(x = pi, y = value)) +                  # plotting results
  geom_line(aes(color = priori)) +
  scale_color_discrete(name = "Title",
                       labels = c("Original", "Uniform", "Beta (6,6)", "Beta(20,100)")) +
  labs(title='Posterior with different prioris', y = '', x = 'pi') +
  scale_y_continuous(expand = c(0, 0.1), breaks = NULL) +
  theme(legend.position = 'bottom')
```



From selected priors, the uniform prior has the smallest effect to the posterior compared to the Beta(2,10) prior. The Beta(20,80) prior makes the posterior more narrow, and changes the mean. It seems to make too many assumptions. That posterior is however still closer to the original posterior than Beta(6,6) prior which makes less assumptions but is presumably according to prior knowledge more “wrong”.