BDA - Assignment 2

Anonymous

```
library(ggplot2)
theme_set(theme_minimal())
library(tidyr)
library(aaltobda)
data("algae")
head(algae)
```

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

a) (Value of the unknown π)

Following page 35 in the book, for an informative prior ($p(\pi) = Beta(2, 10)$):

$p(\pi|y) = \frac{p(y|\pi)p(\pi)}{p(y)}$

Where $p(y|\pi) = \pi^y \times (1-\pi)^{n-y}$ and $p(\pi) = \pi^{(\alpha-1)} \times (1-\pi)^{(\beta-1)}$.

When we multiply these (and ignore the normalization factor $p(y)$), we get: $p(\pi|y) \propto \pi^{(y+\alpha-1)} \times (1-\pi)^{(n-y+\beta-1)} = Beta(y+\alpha, n-y+\beta)$.

As expected, the product of conjugate Beta distributions is also a Beta distribution. $\alpha = 2$, $\beta = 10$, $y = 44$ and $n = 274$, so $p(\pi|y)$ is proportional to $Beta(\pi|46, 240)$.

Therefore, the process of Bayesian inference in this problem involves passing from a prior distribution, $p(\pi) = Beta(2, 10)$, to a posterior distribution, $p(\pi|y) = Beta(46, 240)$. The likelihood is $p(y|\pi) = binomial(\pi)$. For binomial model, the posterior mean of π, which may be interpreted as the posterior probability of algae present for a future draw from the population can be calculated as follows. The posterior interval can be computed directly from cumulative distribution function.

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

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


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

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

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

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

Based on the above formulation, the posterior expectation of the parameter π is 0.1608392 which always lies between the sample proportion, $y/n = 44/274 = 0.1605839$, and the prior mean, $\alpha/(\alpha+\beta) = 2/12 = 0.1666667$; and the posterior uncertainty with central 90% interval is $[0.1265607, 0.1978177]$.

b) (Probability of π less than historical record)

It can be seen that 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 is 95%. So we can be somehow certain that proportion of algae present in monitoring sites is smaller than historical record.

```r
beta_low <- function(prior_alpha, prior_beta, data, pi_0){
  y <- sum(data)
  n <- length(data)
  # cumulative distribution function
  cdf <- pbeta(pi_0, prior_alpha+y, prior_beta+n-y)
  return(cdf)
}

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

cat("the probability of algae levels pi smaller than 0.2 is: ", beta_smaller )
```

```
## the probability of algae levels pi smaller than 0.2 is:  0.9586136
```

c) (Assumptions required for this model and data)

One assumption is that the samples arei.i.d, meaning that the lakes and rivers have no correlation between them regarding if they get algae or not. This may not be true, since lakesthat are closer to eachother, or closer to Helsinki (or to Russia or Sweden or farther North or farther South) might be more likely to be contaminated.

We also have to assume that the posterior distribution follows beta distribution. This makes the prior a conjugate prior, which means that the posterior and prior distributions follow the same form. We should assume that our sample size is reasonably big to make accurate inference regarding the algae levels of lakes and rivers. Considering that there are over 150,000 lakes in Finland, the sample size of 274 is quite small.

We also assume that the levels in the lakes are binary: either there is algae or there isn't, while in realitythis isn't the case.

d) (Prior sensitivity analysis)

```r
"Uniform prior "
```

```
## [1] "Uniform prior "
```

```r
N <- 2
apr <- 0.5
alpha <- N*apr
beta <- N*(1-apr)
post_mean_u = beta_point_est(prior_alpha = alpha , prior_beta = beta, data = algae)
post_int_u = beta_interval(prior_alpha = alpha, prior_beta = beta, data = algae, prob = 0.9)

post_mean_u
```

```
## [1] 0.1630435
```

```r
post_int_u
```

```
## [1] 0.1279681 0.2008987
```

```r
"Different beta prior"
```

```
## [1] "Different beta prior"
```

```r
N <- c(2, 5, 10, 20, 100, 200)
apr = 0.2
alpha <- N*apr
beta <- N*(1-apr)
post_mean = c()
post_int = matrix(rep(0,2*length(N)), ncol=2)
for(i in 1:length(N)){
  post_mean[i] = beta_point_est(prior_alpha = alpha[i] , prior_beta = beta[i], data = algae)
  post_int[i, ] = beta_interval(prior_alpha = alpha[i], prior_beta = beta[i], data = algae, prob = 0.9)
}

post_mean
```

```
## [1] 0.1608696 0.1612903 0.1619718 0.1632653 0.1711230 0.1772152
```

```r
post_int
```

```
##           [,1]      [,2]
## [1,] 0.1260013 0.1985356
## [2,] 0.1265613 0.1987836
## [3,] 0.1274718 0.1991819
## [4,] 0.1292119 0.1999265
## [5,] 0.1401590 0.2040887
## [6,] 0.1491877 0.2067925
```

The prior sensitivity analysis is shown in the following table. The first row is related to the uniform prior $\alpha = 1$, $\beta = 1$ and the other rows use prior distributions concentrated around 0.2 (the proportion of algae present in hitorical records).The columns are the prior mean, the amount of prior information, posterioir mean and posterioir intervals, respectively. As it can be seen when the amount of prior information is increased, the posterioirs are pulled toward the prior distribution and the 90% posterior intervals include the prior mean..

Parameters of the prior distribution		Summaries of the posterior distribution	
$\frac{\alpha}{\alpha+\beta}$	$\alpha+\beta$	mean of π	90% posterior interval for π
0.5	2	0.1630435	[0.1279681, 0.2008987]
0.2	2	0.1608696	[0.1260013, 0.1985356]
0.2	5	0.1612903	[0.1265613, 0.1987836]
0.2	10	0.1619718	[0.1274718, 0.1991819]
0.2	20	0.1632653	[0.1292119, 0.1999265]
0.2	100	0.1711230	[0.1401590, 0.2040887]
0.2	200	0.1772152	[0.1491877, 0.2067925]