BDA - Assignment 1

9/15/2021

Contents

Inference for binomial proportion

a)

We begin by importing the data and setting up the test data.

```
library(aaltobda)
library(knitr)
rm(list = ls())
algae_test = c(0, 1, 1, 0, 0, 0)
data("algae")
```

We we know that $\pi = Beta(2, 10)$ and that y follows a binomial model. We can formulate the likelihood $p(y|\pi)$ as

$$p(y|\pi) = \binom{n}{y}\pi^y(1-\pi)^{n-y}$$

and the prior $p(\pi)$ as

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

Using this information, we can formulate the posterior as

$$p(\pi|y) = \binom{n}{y}\pi^y(1-\pi)^{n-y} \cdot p(\pi) = Beta(\pi|2+y, 10+n-y).$$

We use create a function to calculate the posterior alpha and betas, which will also be useful later.

```
prior_to_post = function(prior_alpha, prior_beta, data) {
  n = length(data)
  y = sum(data)
  post_alpha = prior_alpha+y
  post_beta = prior_beta+n-y
  res = list(alpha=post_alpha, beta = post_beta)
  return(res)
}
```

```
res = prior_to_post(2,10,algae)
alpha = res$alpha
beta = res$beta
paste0("alpha: ", alpha)
```

```
## [1] "alpha: 46"
```

```
paste0("beta: ", beta)
```

```
## [1] "beta: 240"
```

Using the above calculations we get that the posterior is

$$p(\pi|y) = Beta(46, 240)$$

b)

We know that the posterior mean can be calculated using the posterior alphas and betas as $E[\pi|y] = \frac{\alpha}{\alpha+\beta}$

```
beta_point_est = function( prior_alpha, prior_beta, data){
  res = prior_to_post(prior_alpha,prior_beta,data)
  alpha = res$alpha
  beta = res$beta
  return(alpha/(alpha+beta))
}
beta_point_est(prior_alpha = 2, prior_beta = 10, data = algae)
```

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

The mean was used as the point estimate and has a value of 0.161

```
beta_interval = function(prior_alpha, prior_beta, data, prob){
  lower = (1-prob)/2
  upper = prob+lower
  interval = c(lower,upper)
  res = prior_to_post(prior_alpha,prior_beta,data)
  alpha = res$alpha
  beta = res$beta
  res = qbeta(interval, alpha, beta)
  return(res)
}
beta_interval(prior_alpha = 2, prior_beta = 10, data = algae, prob = 0.9)
```

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

For the 90% posterior interval estimate we receive a lower bound of 0.127 and a upper bound of 0.198.

c)

```
beta_low = function(prior_alpha, prior_beta, data, pi_0){
  res = prior_to_post(prior_alpha,prior_beta,data)
  alpha = res$alpha
  beta = res$beta
  res = pbeta(pi_0, alpha, beta)
  return(res)
}
beta_low(prior_alpha = 2, prior_beta = 10, data = algae, pi_0 = 0.2)
```

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

There's approximately a $95,9\%$ chance of being less than 0.2.

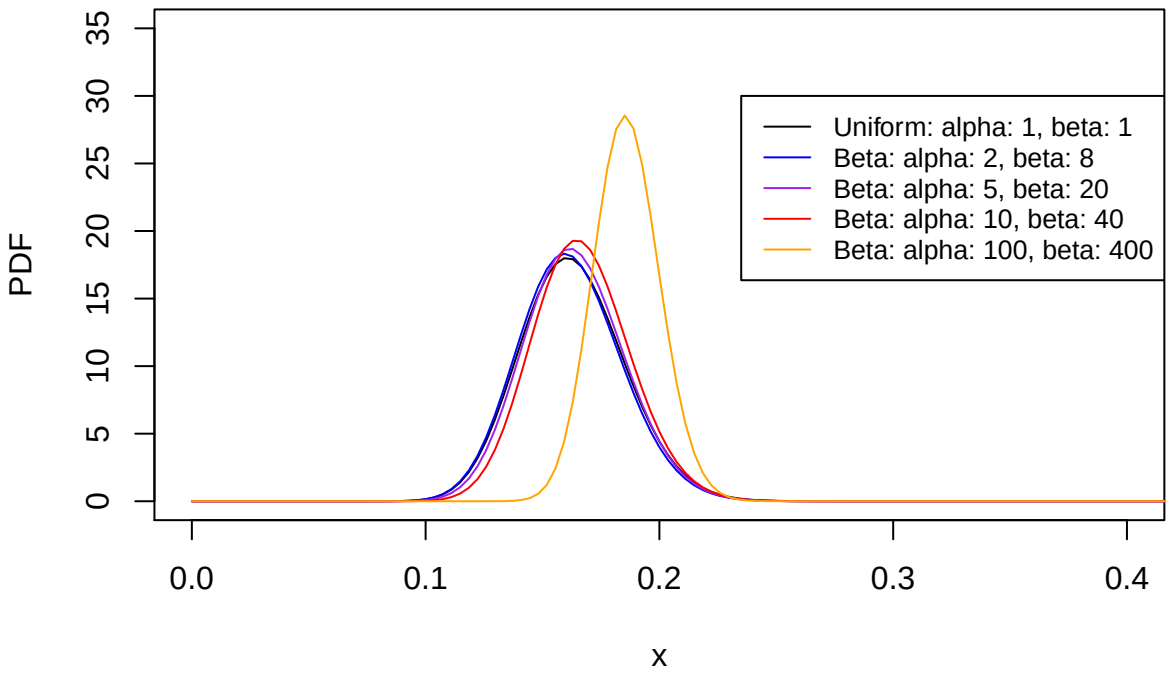
d)

We have assumed that π is continuous. We have, additionally, assumed that all the monitoring sites are independent of each either and behave identically. I.e. we have assumed that our data is i.i.d.. We have also assumed that the algae status is binary, there is or there is not algae in the water. These are the requirements for us to be able to use binomial- and, by extension, beta-distributions on the data.

e)

We will use $\pi_0 = 0.2$ as stated in c) known from historical records to calculated different priors.

```
E = 0.2
prior_alpha = c(1,2,5,10,100)
prior_beta = c(1,-(prior_alpha[-1]/E)*(E-1))
res = prior_to_post(prior_alpha,prior_beta,algae)
alpha = res$alpha
beta = res$beta
x= seq(from=0,to=1, by=1/270)
betapdf1 = dbeta(x,alpha[1],beta[1])
betapdf2 = dbeta(x,alpha[2],beta[2])
betapdf3 = dbeta(x,alpha[3],beta[3])
betapdf4 = dbeta(x,alpha[4],beta[4])
betapdf5 = dbeta(x,alpha[5],beta[5])
par(mfrow=c(1,1))
plot(x, betapdf1,
     type='l', col='black',
     ylim=c(0,35),
     xlim=c(0,0.4),
     ylab="PDF")
leg1 = paste0("Uniform: alpha: ",prior_alpha[1], ", beta: ", prior_beta[1])

lines(x, betapdf2,
     type='l', col ='blue')
leg2 = paste0("Beta: alpha: ",prior_alpha[2], ", beta: ", prior_beta[2])

lines(x, betapdf3,
     type='l', col ='purple')
leg3 = paste0("Beta: alpha: ",prior_alpha[3], ", beta: ", prior_beta[3])

lines(x, betapdf4,
     type='l', col ='red')
leg4 = paste0("Beta: alpha: ",prior_alpha[4], ", beta: ", prior_beta[4])

lines(x, betapdf5,
     type='l', col ='orange')
leg5 = paste0("Beta: alpha: ",prior_alpha[5], ", beta: ", prior_beta[5])

legend(0.235,30, legend=c(leg1, leg2, leg3, leg4, leg5),
       col=c("black", "blue", "purple", "red", "orange"), lty=1, cex=0.8)
```



We can see that increasing the prior knowledge makes the beta posterior distribution narrower and shifts towards the prior mean. From the table below, we can see that the posterior means are mostly still within the interval we got in the b part. When the prior knowledge, $\alpha + \beta$, is of the magnitude 500 only then does our earlier calculated posterior mean fall outside the 90% interval.

Prior alpha + beta	Prior mean	Post mean	90% interval lower bound	upper bound
2	0.5	0.163	0.128	0.201
10	0.2	0.162	0.127	0.199
25	0.2	0.164	0.130	0.200
50	0.2	0.167	0.134	0.202
500	0.2	0.186	0.164	0.209