Lab 2 – Beta-Binomial Distribution

Sara Shao

2021-08-26

In class, you saw the Binomial-Beta model. We will now use this to solve a very real problem! Suppose I wish to determine whether the probability that a worker will fake an illness is truly 1%. Your task is to assist me! Tasks 1–3 will be completed in lab and tasks 3–5 should be completed in your weekly homework assignment. You should still upload task 3 even though this will be worked through in lab!

Task 1

Let's start by quickly deriving the Beta-Binomial distribution.

We assume that

$$X \mid \theta \sim \text{Binomial}(\theta)$$

,

$$\theta \sim \text{Beta}(a, b),$$

where a, b are assumed to be known parameters. What is the posterior distribution of $\theta \mid X$?

$$\begin{aligned} p(\theta \mid X) &\propto p(X \mid \theta) p(\theta) && (1) \\ &\propto \theta^x (1-\theta)^{(n-x)} \times \theta^{(a-1)} (1-\theta)^{(b-1)} && (2) \\ &\propto \theta^{x+a-1} (1-\theta)^{(n-x+b-1)}. && (3) \end{aligned}$$

This implies that

$$\theta \mid X \sim \text{Beta}(x + a, n - x + b).$$

Task 2

Simulate some data using the rbinom function of size $n = 100$ and probability equal to 1%. Remember to set.seed(123) so that you can replicate your results.

The data can be simulated as follows:

```
# set a seed
set.seed(123)
# create the observed data
obs.data <- rbinom(n = 100, size = 1, prob = 0.01)
# inspect the observed data
head(obs.data)
```

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

```
tail(obs.data)
```

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

```r
length(obs.data)
```

```
## [1] 100
```

Task 3

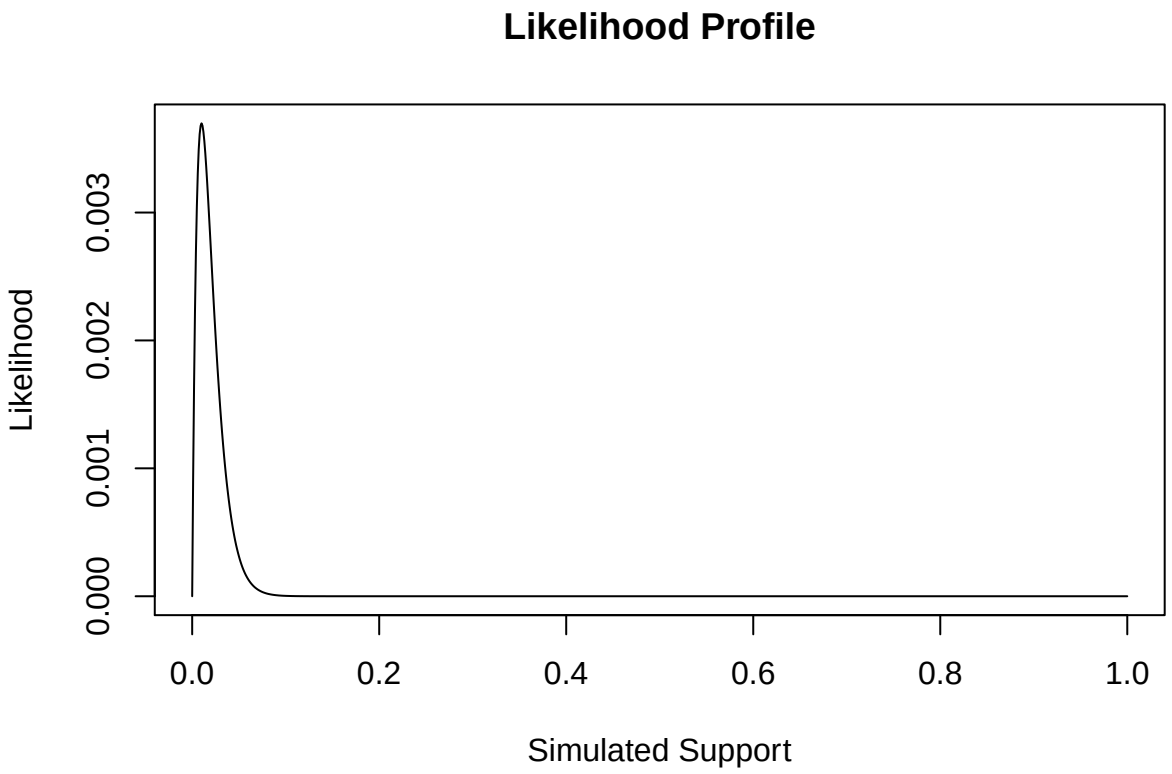
Write a function that takes as its inputs that data you simulated (or any data of the same type) and a sequence of θ values of length 1000 and produces Likelihood values based on the Binomial Likelihood. Plot your sequence and its corresponding Likelihood function.

The likelihood function is given below. Since this is a probability and is only valid over the interval from $[0, 1]$ we generate a sequence over that interval of length 1000.

You have a rough sketch of what you should do for this part of the assignment. Try this out in lab on your own.

```r
### Bernoulli LH Function ###
# Input: obs.data, theta
# Output: bernoulli likelihood
myBernLH <- function(obs.data, theta) {
  N <- length(obs.data)
  x <- sum(obs.data)
  LH <- (theta^x)*((1-theta)^(N-x))
  return(LH)
}

### Plot LH for a grid of theta values ###
# Create the grid #
theta.sim <- seq(from = 0, to = 1, length.out = 1000)
# Store the LH values
sim.LH <- myBernLH(obs.data, theta.sim)
# Create the Plot
plot(theta.sim, sim.LH, type = "l", main = "Likelihood Profile",
     xlab = "Simulated Support", ylab = "Likelihood")
```



Task 4 (To be completed for homework)

Write a function that takes as its inputs prior parameters `a` and `b` for the Beta-Bernoulli model and the observed data, and produces the posterior parameters you need for the model. **Generate and print** the posterior parameters for a non-informative prior i.e. (a,b) = (1,1) and for an informative case (a,b) = (3,1)}.

```
postParam <- function(a, b, obs.data) {
  n <- length(obs.data)
  x <- sum(obs.data)
  post_a <- x + a
  post_b <- n - x + b
  return(c(post_a, post_b))
}
```

Non-informative

```
postParam(1, 1, obs.data)
```

```
## [1]   2 100
```

Informative

```
postParam(3, 1, obs.data)
```

```
## [1]   4 100
```

Task 5 (To be completed for homework)

Create two plots, one for the informative and one for the non-informative case to show the posterior distribution and superimpose the prior distributions on each along with the likelihood. What do you see? Remember to turn the y-axis ticks off since superimposing may make the scale non-sense.

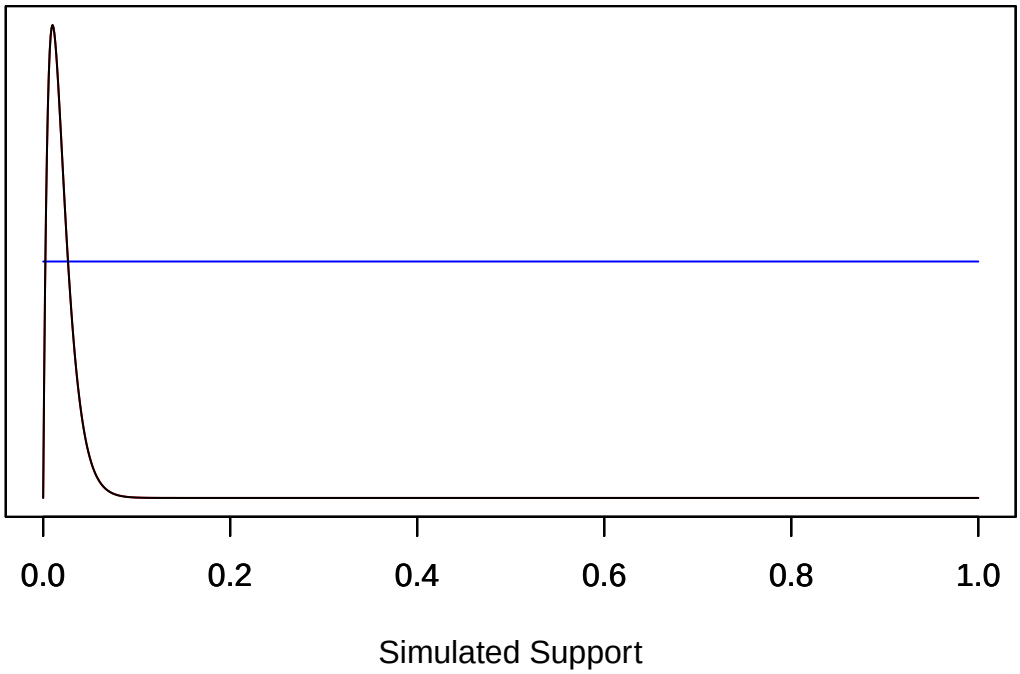
Non-informative

```
theta.sim <- seq(from = 0, to = 1, length.out = 1000)

non_a <- postParam(1, 1, obs.data)[1]
non_b <- postParam(1, 1, obs.data)[2]

sim.LH <- myBernLH(obs.data, theta.sim)

#posterior is red
plot(theta.sim, dbeta(theta.sim, non_a, non_b), type = "l", col = "red",
     yaxt = "n", xlab = "", ylab = "")
par(new = TRUE)
#prior is blue
plot(theta.sim, dbeta(theta.sim, 1, 1), type = "l", col = "blue",
     yaxt = "n", xlab = "", ylab = "")
par(new = TRUE)
#likelihood is black
plot(theta.sim, sim.LH, type = "l",
     yaxt = "n", xlab = "Simulated Support", ylab = "")
```

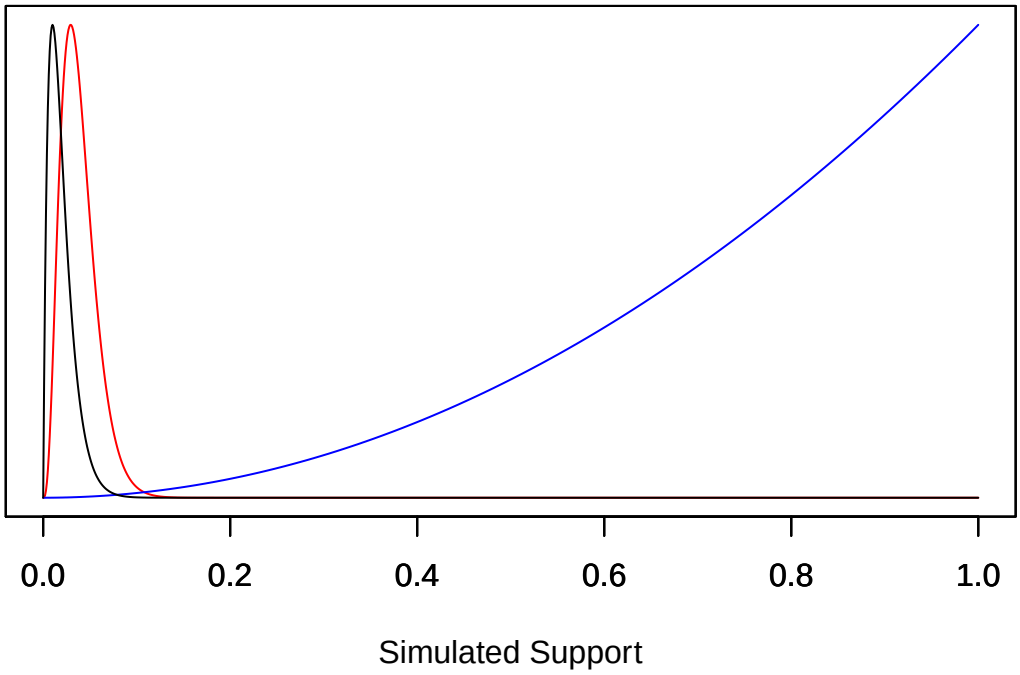


Informative

```
inf_a <- postParam(3, 1, obs.data)[1]
inf_b <- postParam(3, 1, obs.data)[2]

#posterior is red
plot(theta.sim, dbeta(theta.sim, inf_a, inf_b), type = "l", col = "red",
     yaxt = "n", xlab = "", ylab = "")
par(new = TRUE)
#prior is blue
plot(theta.sim, dbeta(theta.sim, 3, 1), type = "l", col = "blue",
     yaxt = "n", xlab = "", ylab = "")
```

```r
par(new = TRUE)
#likelihood is black
plot(theta.sim, sim.LH, type = "l",
     yaxt = "n", xlab = "Simulated Support", ylab = "")
```



In the graph for the non-informative case, the posterior distribution is pretty much exactly proportional to the likelihood distribution. In the informative case, however, we can see that although the posterior distribution is still similar to the likelihood, the mean is shifted a little bit in the the direction of the prior mean.