Lab 2 – Binomial-Beta Distribution

YQ

10/12/2021

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
```

```
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.

Solution:

1. Write a function with:

- **input:** simulated data and sequence of θ values.
- **output:** binomial likelihood of the data corresponding to each θ value.

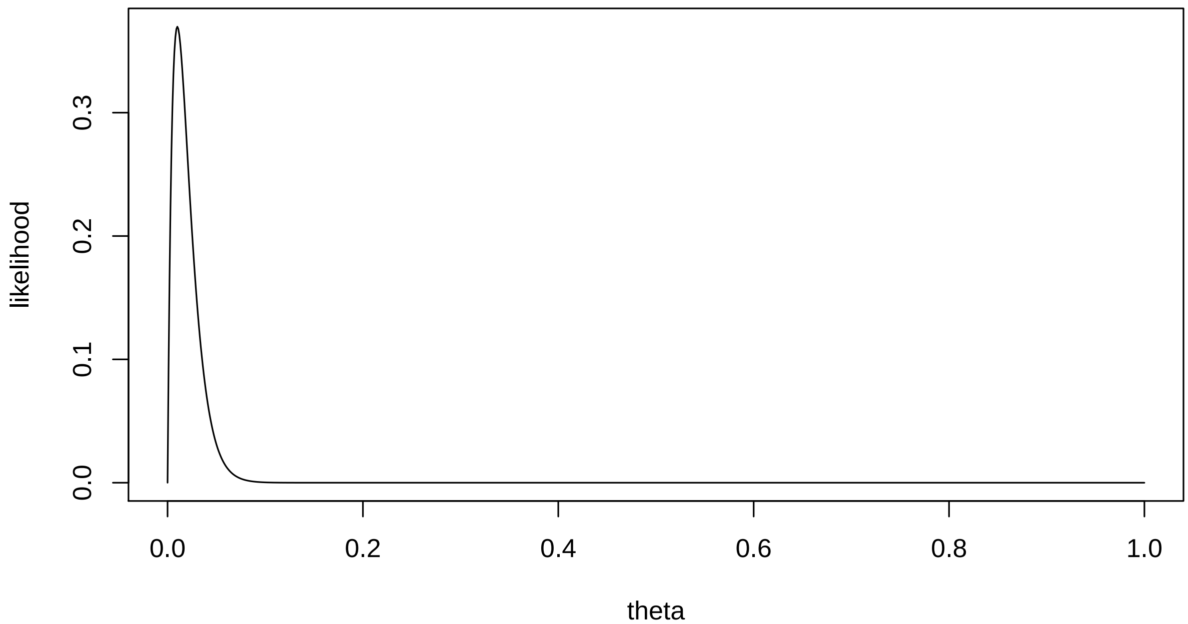
```
theta = c(0.01, 0.1)
N = length(obs.data)
X = sum(obs.data)
LH = choose(N, X) * theta^(X) * (1-theta)^(N-X)
LH
```

```
## [1] 0.3697296376 0.0002951267
```

```
likelihood <- function(obs.data, theta) {
  N = length(obs.data)
  X = sum(obs.data)
  LH = choose(N, X) * theta^(X) * (1-theta)^(N-X)
  return(LH)
}
```

2. Plot the likelihood over a grid of θ values

```
theta = seq(0,1,length.out = 1000)
plot(theta, likelihood(obs.data, theta), type = "l",
     ylab = "likelihood", xlab = "theta")
```



Task 4

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)}.

Solution:

1. Write a function with:

- **input:** prior parameters a, b, and the observed data.
- **output:** parameters of the Beta posterior distribution of θ.

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.

```
post_parameters <- function(a, b, obs.data){
  N = length(obs.data)
  X = sum(obs.data)
  a.post = a + X
  b.post = N - X + b
  return(c(a.post, b.post))
}
```

2. **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).

```r
post_parameters(1,1, obs.data)
```

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

```r
post_parameters(3,1, obs.data)
```

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

Task 5

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.

Solution:

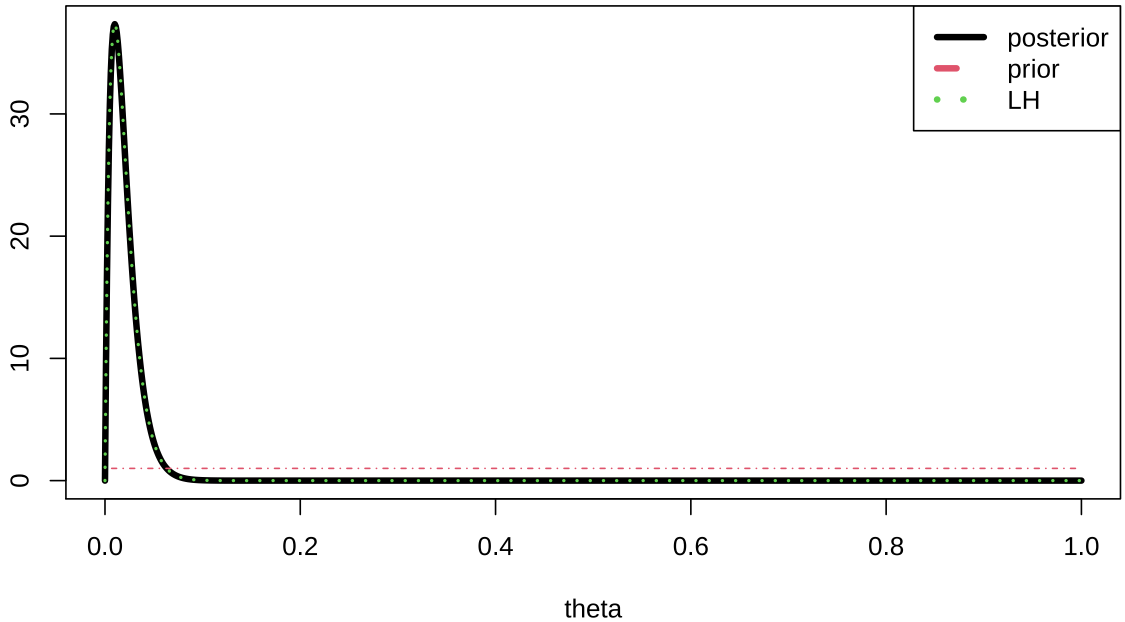
1. non-informative prior

```r
params1 = post_parameters(1,1, obs.data)

# Plot posterior distribution
theta = seq(0,1, length.out = 1000)
plot(theta, dbeta(theta, shape1 = params1[1], shape2 = params1[2]),
     type = "l", xlab = "theta", ylab = "")
# Plot prior
lines(theta, dbeta(theta, shape1 = 1, shape2 = 1), col = 2, ltyp = 2)
# Plot likelihood
LH = likelihood(obs.data, theta)
lines(theta, 1000*LH/sum(LH), col = 3, lty = 3)

legend("topright", legend = c("posterior", "prior", "LH"),
       lty = c(1,2,3), col = c(1,2,3))
```

Here's the result:



Observation: The posterior is almost the same as the normalized likelihood.

Interpretation: With a non-informative prior, the *likelihood* drives inference.