Bayesian Models 3

Charlie Qu

June 25, 2017

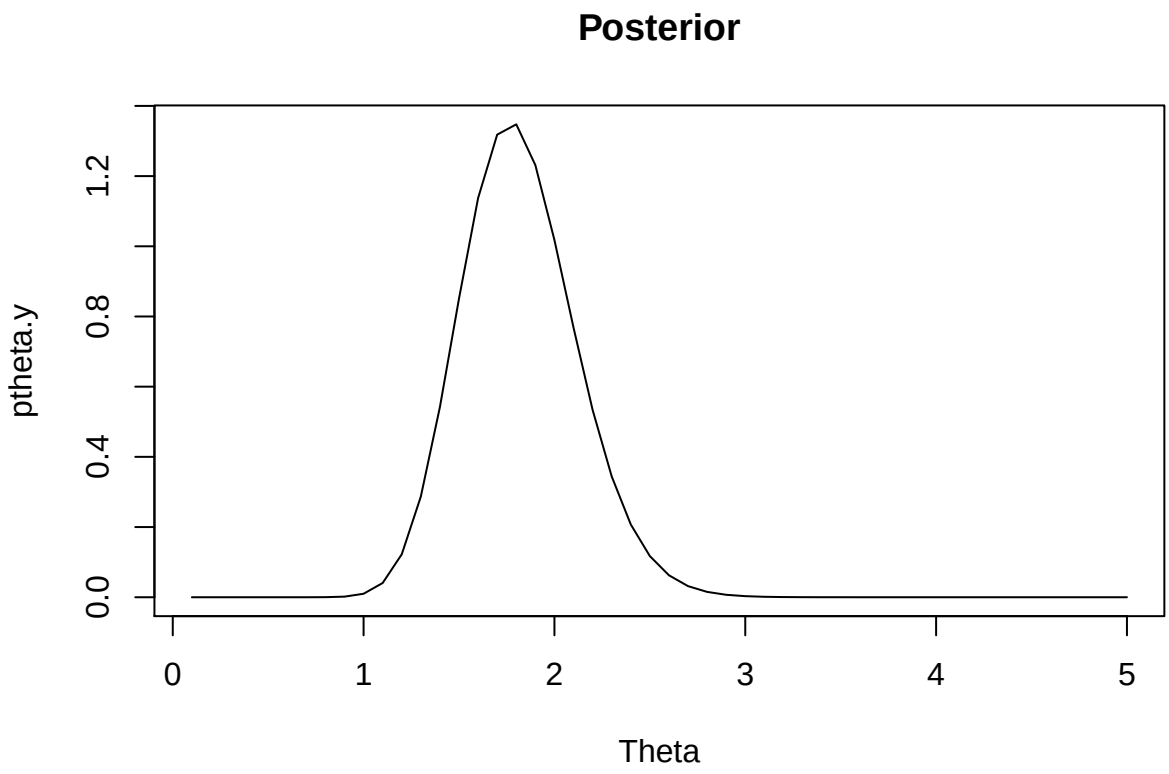
Outline

This is a further of the last write-up, a short writing about posterior predictions/predictive checks, and reading in data from a website

1. Interested in θ, mean number of eggs per nest;
2. Observe Y_i, number of eggs laid in nest i, for i=1,.,n;
3. Assume $Y_i|\theta \sim Poisson(\theta)$;
4. Therefore $Y|\theta = \Sigma_{i=1}^{n} \sim Poisson(n\theta)$;
5. Our data consists of 20 nests and 36 total eggs (n=20n=20, y=36y=36)
6. prior $\theta \sim$ Gamma(a=1.25, b=0.5);
7. Posterior $\theta|y \sim$Gamma(a+y, b+n)

Setup

```
y <- 36; n <- 20
a <- 1.25; b <- 0.5
a.y <- a + y; b.y <- b + n
Theta <- seq(0.1, 5, by = 0.1)
ptheta.y <- dgamma(Theta, a.y, b.y)
plot(Theta, ptheta.y, type = "l", main="Posterior")
```



Reminder to Set a Seed

We'll be taking random draws in this lab. To get the same answers each time you run the code, you need to set a seed.

```
set.seed(1)
```

Posterior Predictive Checks

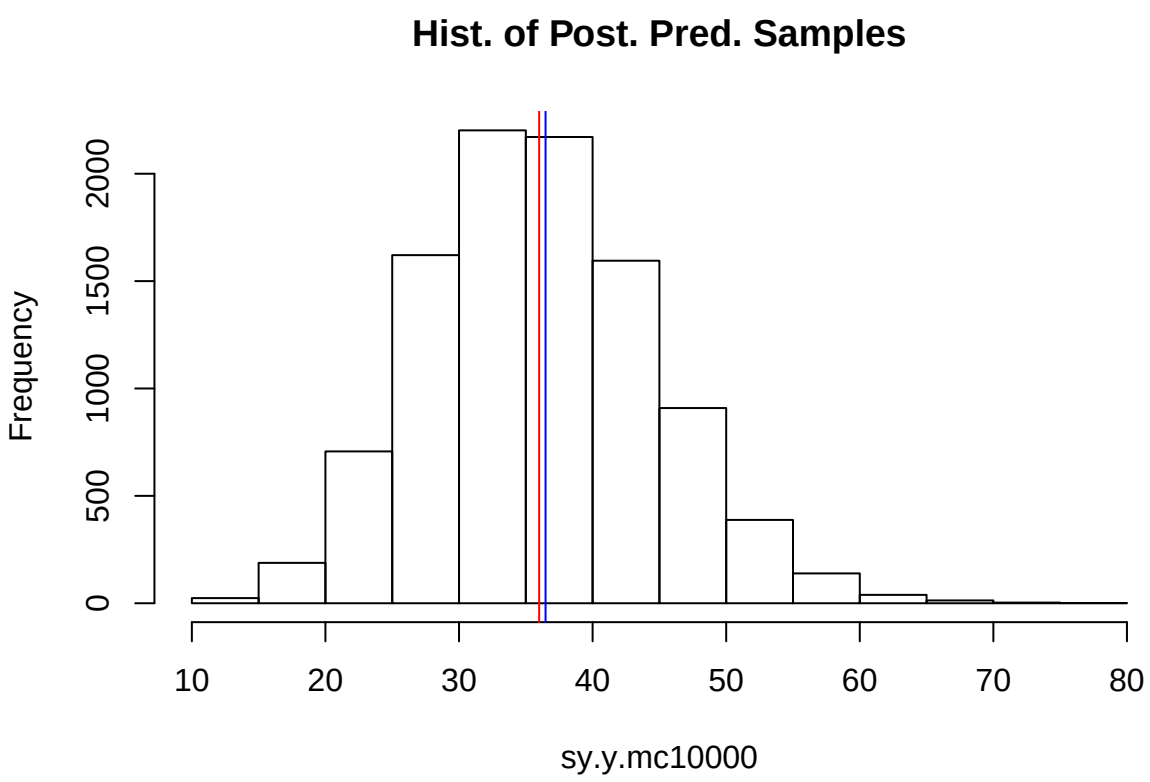
With the given information, we can compute the mean of the posterior predictive distribution for the total number of eggs, to get a Monte Carlo approximation of E[$\tilde{y}$|y], we iteratively

(1) Sample $\theta^{(s)}$ from p(θ|y);
(2) Sample $\tilde{y}^{(s)}$???Poisson($20\theta^{(s)}$).

Then we can examine the distribution of y_y and compute $E[\tilde{y}|y] \approx 1/s\Sigma_{s=1}^{S}\tilde{y}^{(s)}$

Let's take S=10,000S=10,000 Monte Carlo samples to compute the approximation.

```
S <- 10000
stheta.y.mc10000 <- sy.y.mc10000 <- numeric(S)
for (s in 1:S) {
  stheta.y.mc10000[s] <- rgamma(1, a.y, b.y)
  sy.y.mc10000[s] <- rpois(1, 20*stheta.y.mc10000[s])
}
hist(sy.y.mc10000, main = "Hist. of Post. Pred. Samples")
abline(v = mean(sy.y.mc10000), col = "blue")
abline(v = y, col = "red")
```



Read in delimited data If we had observed each Y_i separately, we could compare more features of the observed data distribution to features of the posterior predictive distribution.

Let's import a file that contains individual observations Y_i such that y=36, and n=20, as before.

```
nesteggs <- read.table(
"http://www2.stat.duke.edu/courses/Fall16/sta601.001//files/nesteggs.dat",
header=TRUE)
y.i <- nesteggs$Eggs
```

read.table and read.csv can also read files located on your computer: (1) Mac: read.table("/mydrive/myfolder/file.dat") (2) Windows: read.table("/C:/Users/myname/myfolder/file.dat")

Posterior Predictive Checks

As a posterior predictive check, let's try comparing $t = |Y_i \geq 5|/20$ with $E[\tilde{t}|Y_1, ...Y_{20}] = E[|\tilde{Y} \geq 5|/20|Y_1, ...Y_{20}]$

```
r   t <- mean(y.i >= 5)
```

To get a Monte Carlo approximation of $E[|\tilde{Y} \geq 5|/20|Y_1, ...Y_{20}]$, we iteratively (1) Sapmple $\theta^{(s)}$ from $p(\theta|y)$; (2) Sample $\tilde{Y_1}^{(s)}, \ldots, \tilde{Y_2}0^{(s)} \sim$ Possion$(\theta^{(s)})$; (3) Compute $\tilde{t}^{(s)} = \frac{|\tilde{Y} \geq 5|}{20}$.

Then we can examine the distribution of $\tilde{t}$ and compute

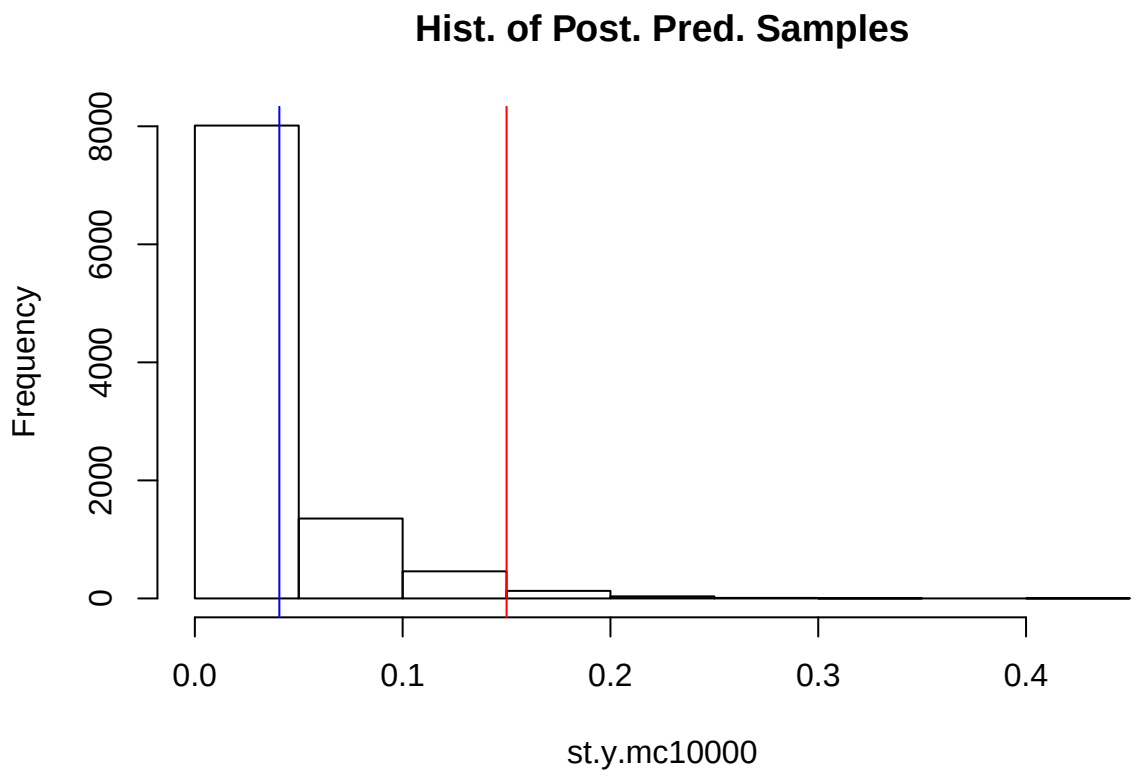
$E[\tilde{t}|Y_1, ...Y_{20}] \approx \frac{1}{S}\Sigma_{s=1}^{s}\tilde{t}^{(s)}$

Let's take S=10,000S=10,000 samples to compute the Monte Carlo approximation.

```
S <- 10000
stheta.y.mc10000 <- st.y.mc10000 <- numeric(S)
for (s in 1:S) {
stheta.y.mc10000[s] <- rgamma(1, a.y, b.y)
sy.y <- rpois(20, stheta.y.mc10000[s])
st.y.mc10000[s] <- mean(sy.y >= 5)
}
```

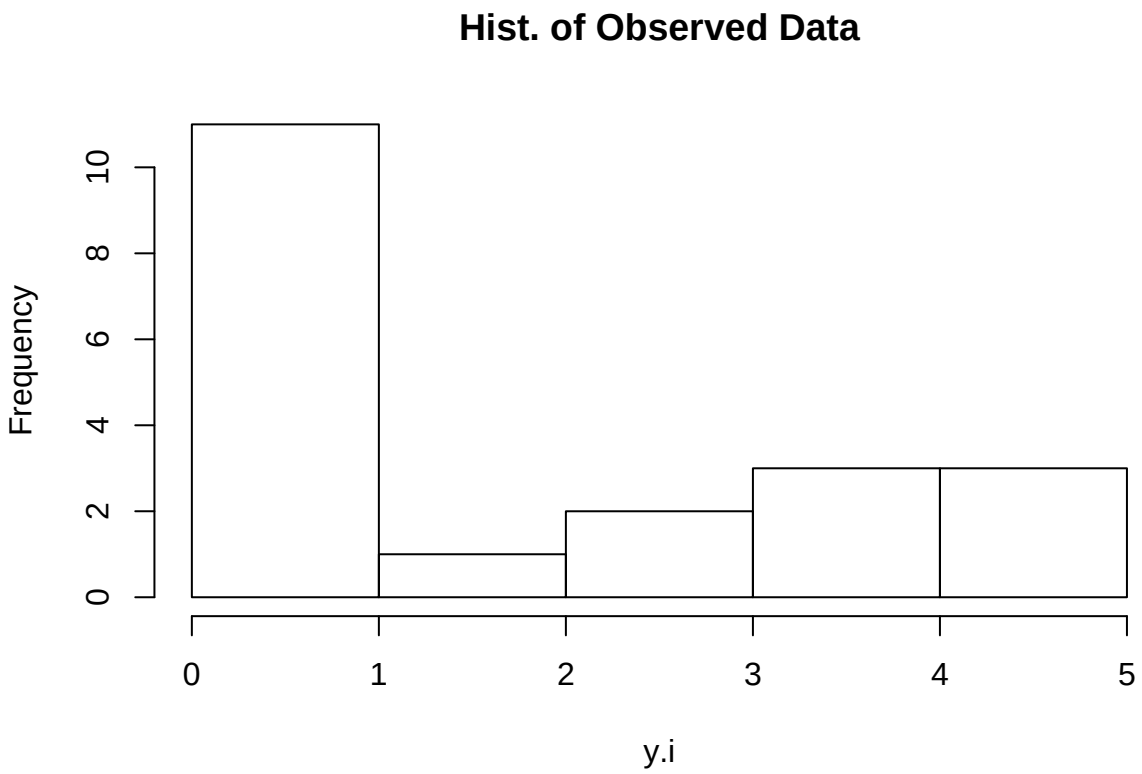
We can compare the distribution of $\tilde{t}^{(s)}$ to t.

```
hist(st.y.mc10000, main = "Hist. of Post. Pred. Samples")
abline(v = mean(st.y.mc10000), col = "blue")
abline(v = 0.15, col = "red")
```



The difference between the posterior predictive distribution and observed data distributions of the number of nests with more than 55 eggs could be due to: (1) small sample size of n=20; (2) inappropriateness of Poisson model.

```
hist(y.i, main = "Hist. of Observed Data")
```



We can directly compute the posterior mean from what we know about the beta distribution.The nesteggs.dat data was constructed by drawing Y_1,.,Y_2,.,$Y_1$0 from a uniform distribution on 1,.,5 subject to the constraint y=36y=36 and setting $Y_1$1,.,$Y_2$0=0, so in this case the Poisson model is incorrect.

As we see, the Poisson model may be "good enough" for drawing inference on the posterior mean, but inappropriate for drawing inference on other features of the data distribution.