Homework

Exercise 6.6

(a)

Under the old protocol, the total number of data n is fixed as 20. But here the protocol is to stop once 13 zeros have appeared. The original posterior distribution of parameter $\theta\theta$, given the data $y_i \, y_i$ s is :

$$p(\theta|y) \propto \theta^{\sum y}(1 - \theta)^{n - \sum y} = Beta(\alpha = 8, \beta = 14)$$

Here we know that the new protocol can be summarized as:

$$p(y, n|\theta) = \prod_{i=1}^{n} \theta^{y_i} (1 - \theta)^{1-y_i} I(\sum_{i=1}^{n}(1 - y_i) = 13)$$

For the data we have here, exactly 13 zero (total 20 data) have been generated and the last element in the data is exact zero which makes no difference to the posterior in the old protocol. Since the kernal of these two posterior distributions have no difference, we know that the posterior distribution is unchanged.

(b)

```r
set.seed(5224)

# new protocol
n_switch <- c()
s = 10000 # number of simulations (same as the textbook)
for (i in 1:s){
  ## simulate the rep data under
  theta <- rbeta(n = 1,shape1 = 8,shape2 = 14)
  y_rep <- rbinom(n = 1,size=1,prob=theta)
  while (sum(y_rep==0) < 13) {   # the new protocol
    y_rep <- c(y_rep,rbinom(n = 1,size=1,prob=theta))
  }
  n_rep <- length(y_rep)
  # calculate the number of switch
  n_switch <- c(n_switch,
                sum(y_rep[2:n_rep] != y_rep[1:(n_rep-1)]))
}


# original one
n_switch_old <- c()
s = 10000 # number of simulations (same as the textbook)
for (i in 1:s){
  ## simulate the rep data under new protocol
  theta <- rbeta(n = 1,shape1 = 8,shape2 = 14)
  y_rep <- rbinom(n = 1,size=1,prob=theta)
  while (sum(y_rep) < 20) {   # the new protocol
    y_rep <- c(y_rep,rbinom(n = 1,size=1,prob=theta))
  }
  # calculate the number of switch
  n_switch_old <- c(n_switch_old,
                sum(y_rep[2:20] != y_rep[1:19]))
}
n_switch <- data.frame(n_switch=n_switch)
n_switch$type <- 'new protocol'
n_switch_old <- data.frame(n_switch=n_switch_old)
n_switch_old$type <- 'old protocol'
plot_data <- rbind(n_switch,n_switch_old)

library(ggplot2)
ggplot(plot_data,aes(n_switch,fill=type))+
  geom_histogram(alpha = 0.5, aes(y = ..density..), position = 'identity',bins = 20)+
  geom_vline(xintercept=3, linetype="dashed", color = "red", size=2)
```
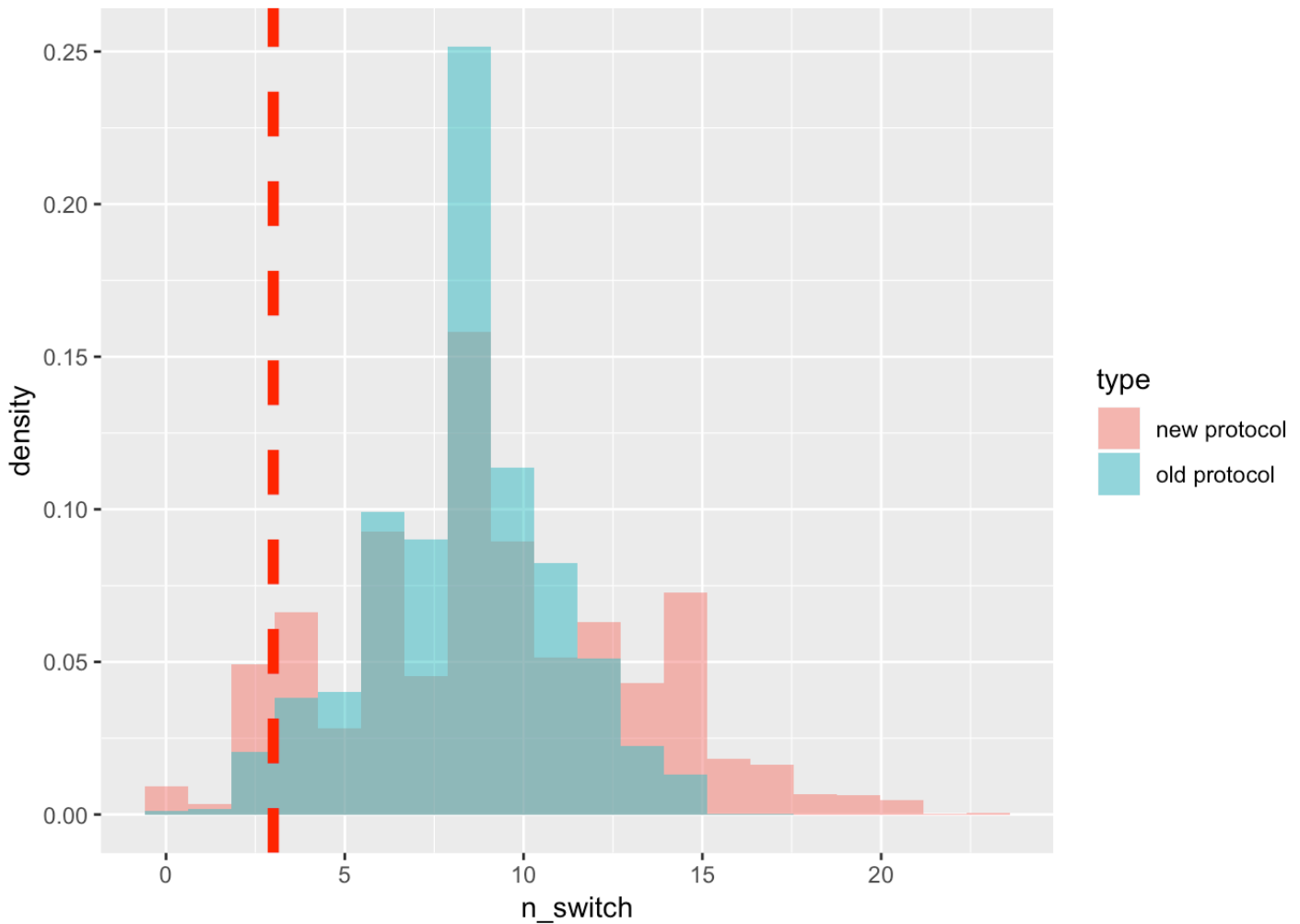
As we can see from this plot compared with the old protocol in the textbook, the new protocol's predictive distribution is :

1. more spread. This is because the total number of data in each simulation n can be bigger than 20.

```
mean(n_switch$n_switch > 3)
```

```
## [1] 0.9249
```

```
mean(n_switch_old$n_switch > 3)
```

```
## [1] 0.9715
```

2. with the new protocol the p value is closer to 0.5 which shows that the new protocol gives , to some extent, a more reliable result and tends to be a better model.