Bayesien_hw2

Ching-Tsung_Deron_Tsai

2022/10/9

```
library(tidyverse)
```

4.2

a

The count and summation are 10 and 117 for group y_A, and 13 and 113 for group y_B. Given Gamma conjugate priors for these two Poisson models, the posteriors are:

$$p(\theta_A|y_A) = \Gamma(120+117, 10+10) = \Gamma(237, 20)$$
$$p(\theta_B|y_B) = \Gamma(12+113, 1+13) = \Gamma(125, 14)$$

```
yA = c(12, 9, 12, 14, 13, 13, 15, 8, 15, 6)
yB = c(11, 11, 10, 9, 9, 8, 7, 10, 6, 8, 8, 9, 7)
length(yA) ; sum(yA)    # count & sum of group yA
```

```
## [1] 10
```

```
## [1] 117
```

```
length(yB); sum(yB)    # count & sum of group yB
```

```
## [1] 13
```

```
## [1] 113
```

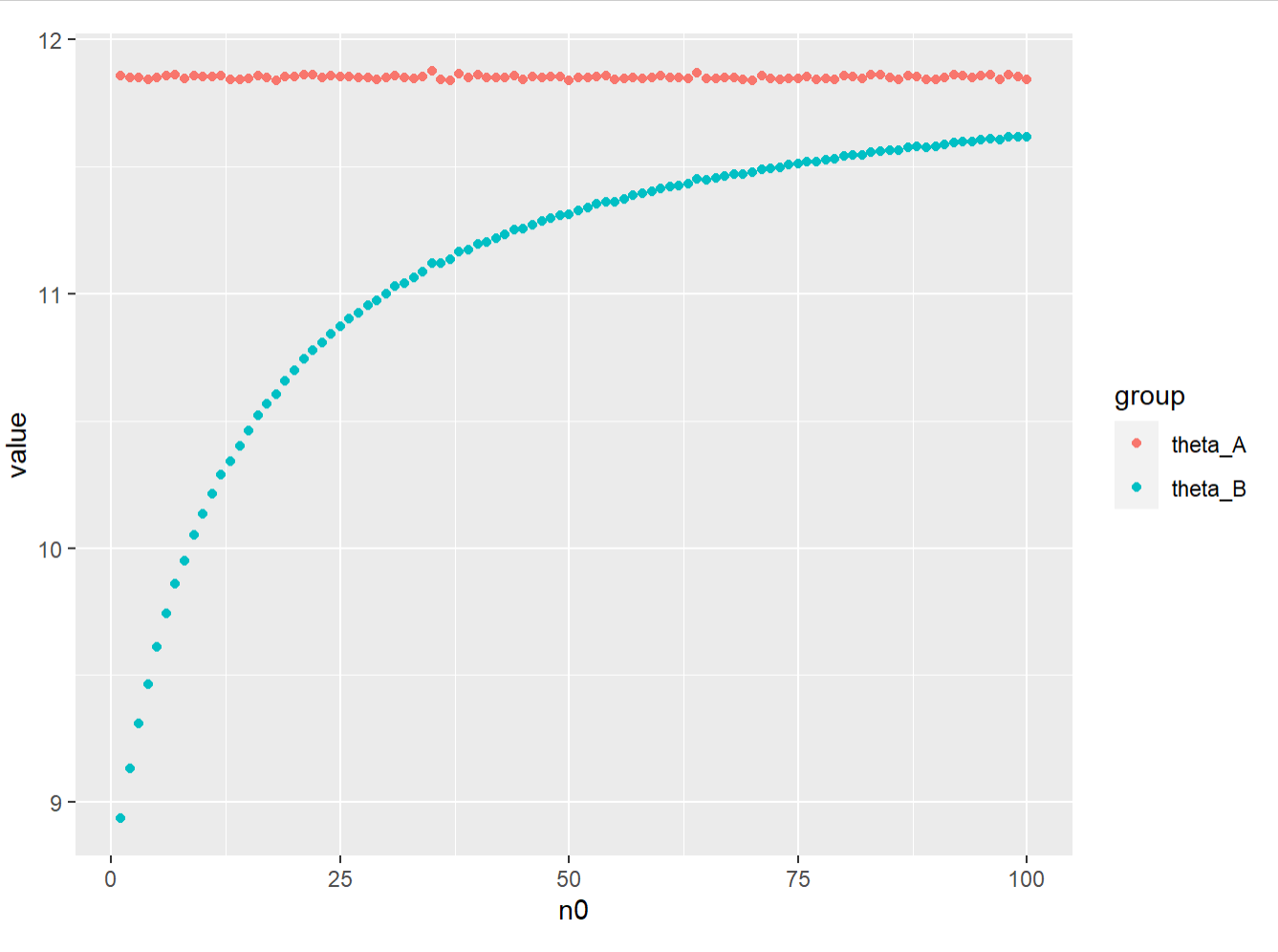
```
# Monte Carlo approximation with B=10000
b=1e4
set.seed(1)
post_A = rgamma(b,120+sum(yA),10+length(yA))
set.seed(2)
post_B = rgamma(b,12+sum(yB),1+length(yB))
(ans = mean(post_B < post_A))
```

```
## [1] 0.9949
```

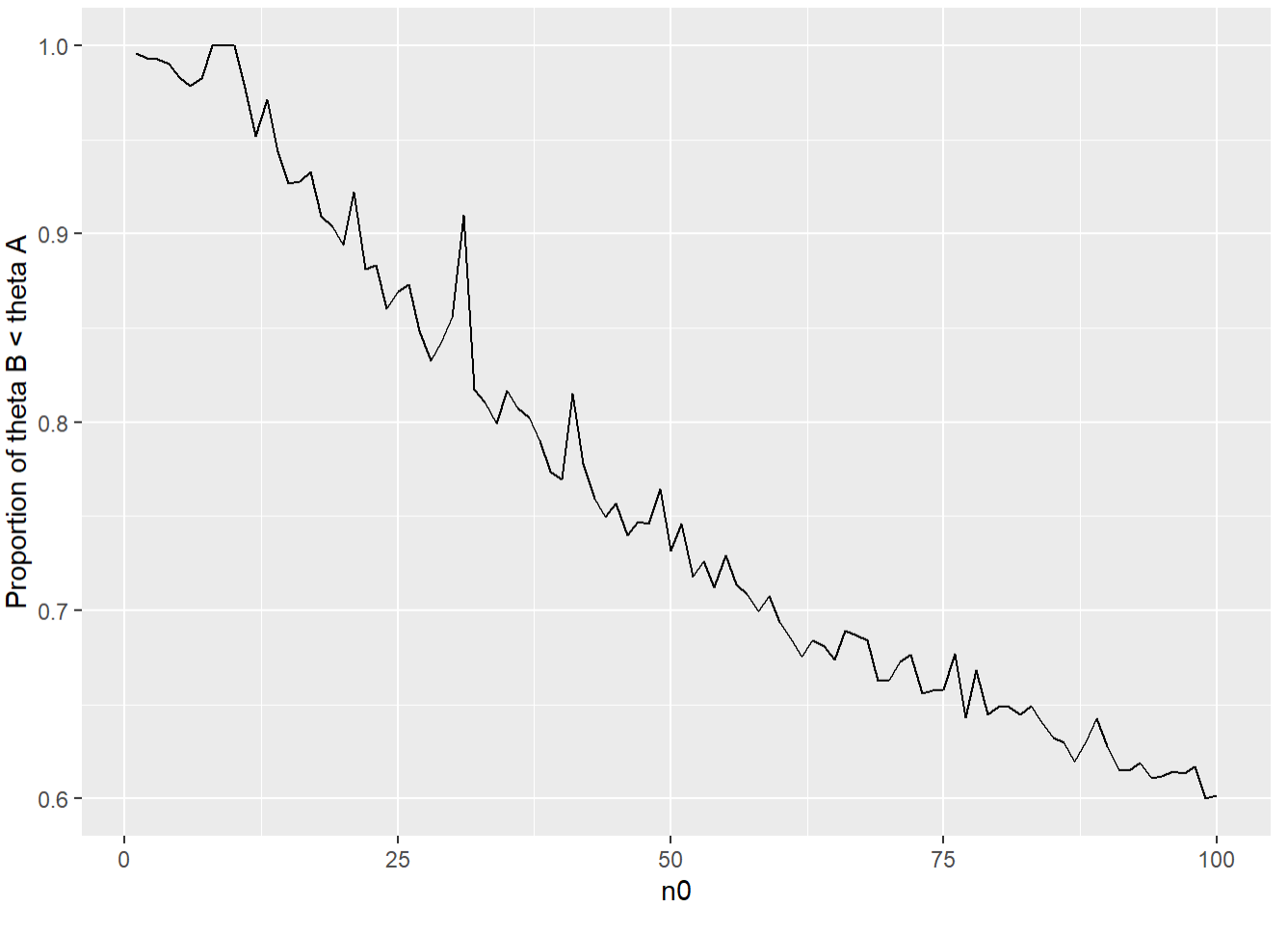
$Pr(\theta_A < \theta_B|y_A, y_B)$ via Monte Carlo sampling with 10000 sample values is 0.9949

b

```
posterior <- t(sapply(1:100, function(n0){
  set.seed(n0)
  theta_a = rgamma(b,120+sum(yA),10+length(yA))
  set.seed(n0)
  theta_b = rgamma(b,12*n0+sum(yB), n0+length(yB))
  c(n0, mean(theta_a), mean(theta_b), mean(theta_b<theta_a))
})) %>% data.frame()
colnames(posterior) <- c("n0", "theta_A", "theta_B", "b_smaller")
posterior %>%
  pivot_longer(cols = c("theta_A", "theta_B"), names_to = "group") %>%
  ggplot() +
  geom_point(aes(x=n0, y=value, col=group))
```




```
ggplot(posterior) +
  geom_line(aes(x=n0, y=b_smaller)) +
  ylab("Proportion of theta B < theta A")
```



Bigger n_0 gives a decreasing posterior difference. This means stronger prior belief will be less sensitive to the conclusion.

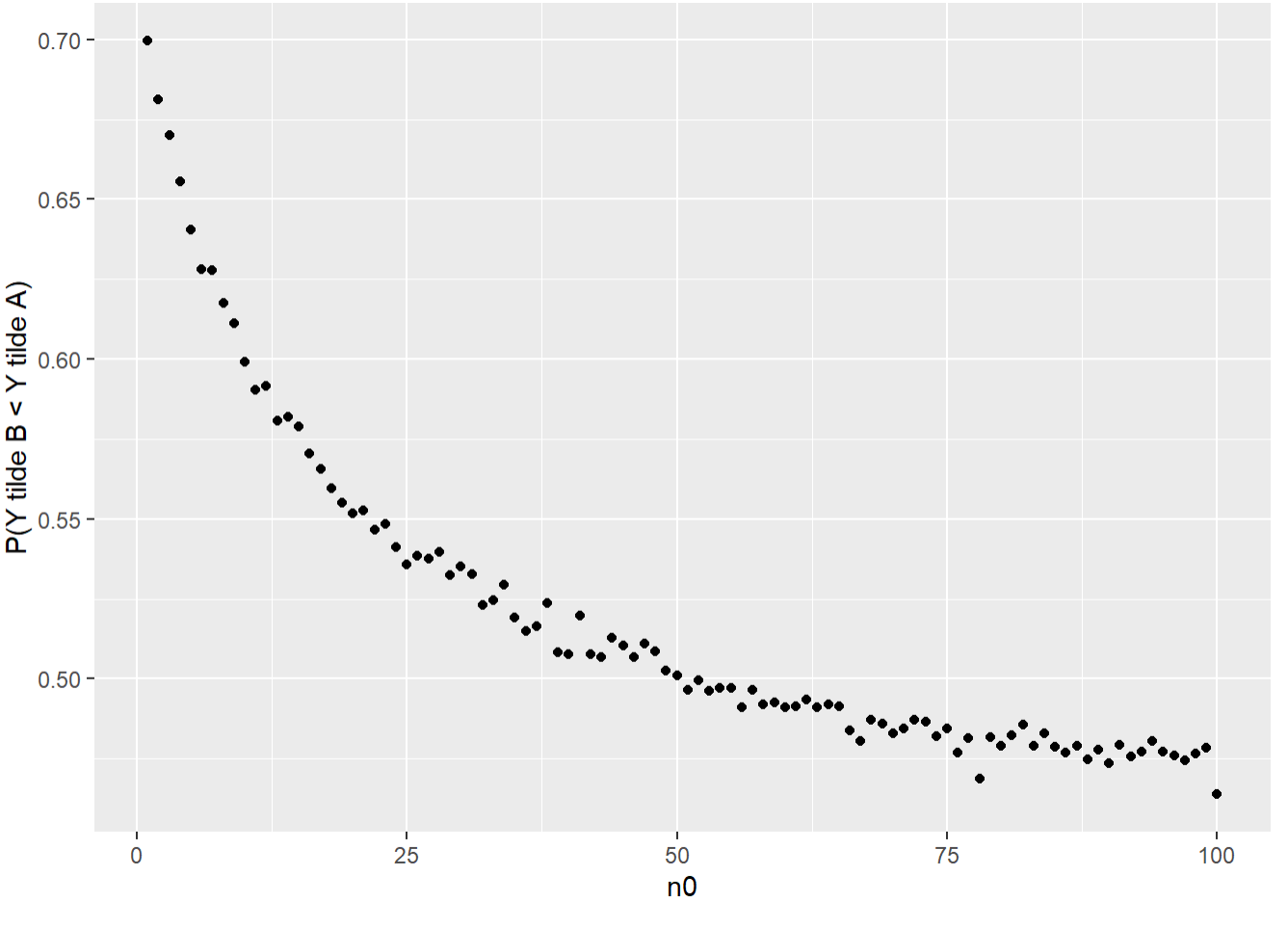
c

According to lecture4 code2 in class, we can use the Monte Carlo sampling thetas to calculate the posterior predictive Y.

```
# repeat part a
set.seed(1)
ytildeA = rpois(b,post_A)
set.seed(2)
ytildeB = rpois(b,post_B)
(ans_2c <- mean(ytildeB<ytildeA))      # Pr(theta_A<theta_B|y_A,y_B)
```

```
## [1] 0.693
```

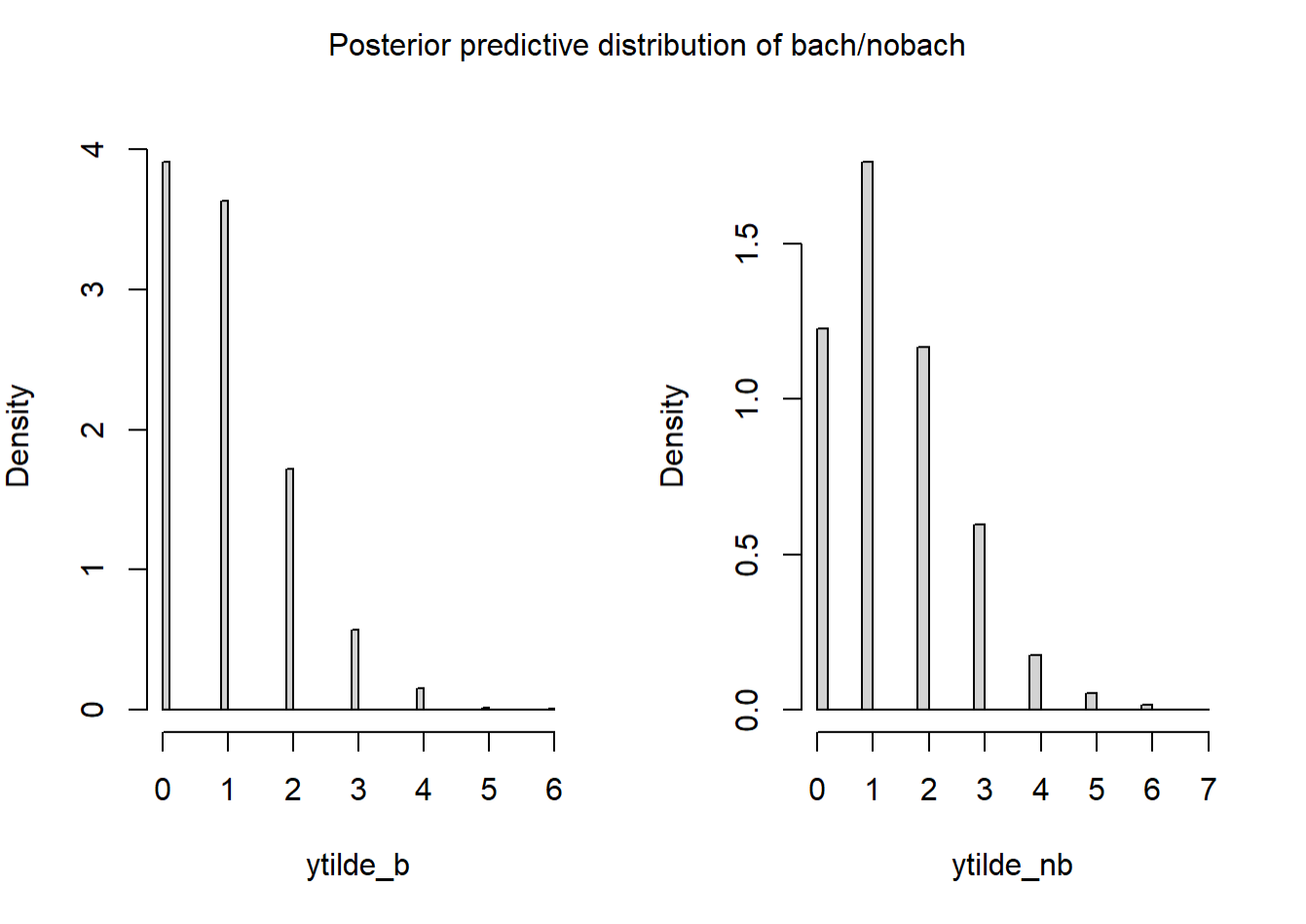
```
# repeat part b
t(sapply(1:100, function(n0){
  set.seed(n0)
  theta_a = rgamma(b,120+sum(yA),10+length(yA))
  set.seed(n0)
  theta_b = rgamma(b,12*n0+sum(yB), n0+length(yB))
  set.seed(n0)
  ytilde_a = rpois(b,theta_a)
  set.seed(n0)
  ytilde_b = rpois(b,theta_b)
  c(n0, mean(ytilde_b<ytilde_a))
})) %>% data.frame() %>%
  ggplot(aes(x=X1, y=X2)) +
  geom_point() +
  xlab("n0") + ylab("P(Y tilde B < Y tilde A)")
```



For repeating in part a, $Pr(_B<_A|y_A,y_B)=$0.693. For repeating in part b, we can find a same trend but non-linear relationship.

4.8

```
# read files:
read_as_vector <- function(path) {
  readLines(path) %>%
    str_split(pattern = " ", simplify = T) %>% t() %>% as.numeric() %>% discard(is.na)
}
bach <- read_as_vector("../HW_Q/menchild30bach.dat")
nobach <- read_as_vector("../HW_Q/menchild30nobach.dat")
B=5000
a_bach = 2 + sum(bach) ; b_bach = 1 + length(bach)
a_nobach = 2 + sum(nobach) ; b_nobach = 1 + length(nobach)
set.seed(1)
theta_bach <- rgamma(B, a_bach, b_bach)
set.seed(2)
theta_nobach <- rgamma(B, a_nobach, b_nobach)
set.seed(3)
ytilde_b <- rpois(B, theta_bach)
set.seed(4)
ytilde_nb <- rpois(B, theta_nobach)
par(mfrow=c(1,2))
hist(ytilde_b, prob=TRUE, breaks = 50, main="")
hist(ytilde_nb, prob=TRUE, breaks = 50, main="")
mtext("Posterior predictive distribution of bach/nobach", side = 3, line = -2, outer = TRUE)
```




```
par(mfrow=c(1,1))
```

```
theta_diff <- theta_nobach - theta_bach
y_diff <- ytilde_nb - ytilde_b
quantile(theta_diff,c(0.025, 0.975))  # theta B - theta A
```

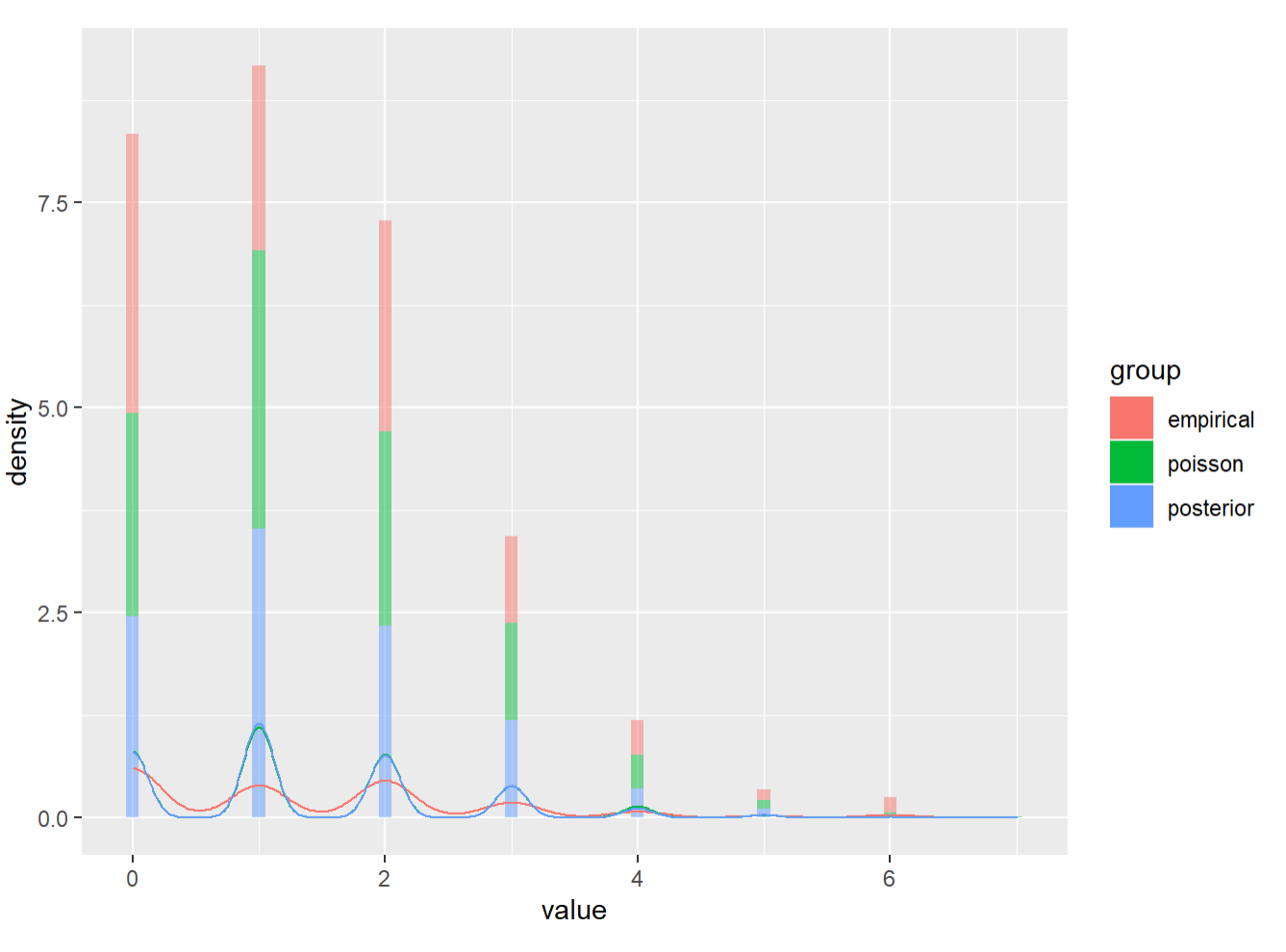
```
##      2.5%     97.5%
## 0.1552025 0.7361474
```

```
quantile(y_diff,c(0.025, 0.975))      # y tilde B - y tilde A
```

```
##  2.5% 97.5%
##    -2     4
```

We can be 95% confident that there exist differences between θ_B and θ_A. But we cannot say $\tilde{Y}_B$ and $\tilde{Y}_A$ have differences according to the 95% credible interval.

```
set.seed(2)
pois <- rpois(5000,1.4)
empirical<-rep(nobach, round(5000/length(nobach)))    # replicate the empirical would provide very close summary
 statistics:
simu <- data.frame(empirical=empirical[1:5000], poisson=pois, posterior=ytilde_nb)
simu %>%
  pivot_longer(cols=c("empirical", "poisson", "posterior"), names_to = "group") %>%
  ggplot(aes(x=value)) +
  geom_histogram(aes(y = ..density.., fill=group), alpha=0.5, binwidth = .1) +
  geom_density(aes(col=group))
```




```
summary(simu)
```

```
##    empirical        poisson        posterior
##  Min.   :0.000   Min.   :0.000   Min.   :0.000
##  1st Qu.:0.000   1st Qu.:1.000   1st Qu.:1.000
##  Median :1.000   Median :1.000   Median :1.000
##  Mean   :1.398   Mean   :1.415   Mean   :1.394
##  3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.000
##  Max.   :6.000   Max.   :7.000   Max.   :7.000
```

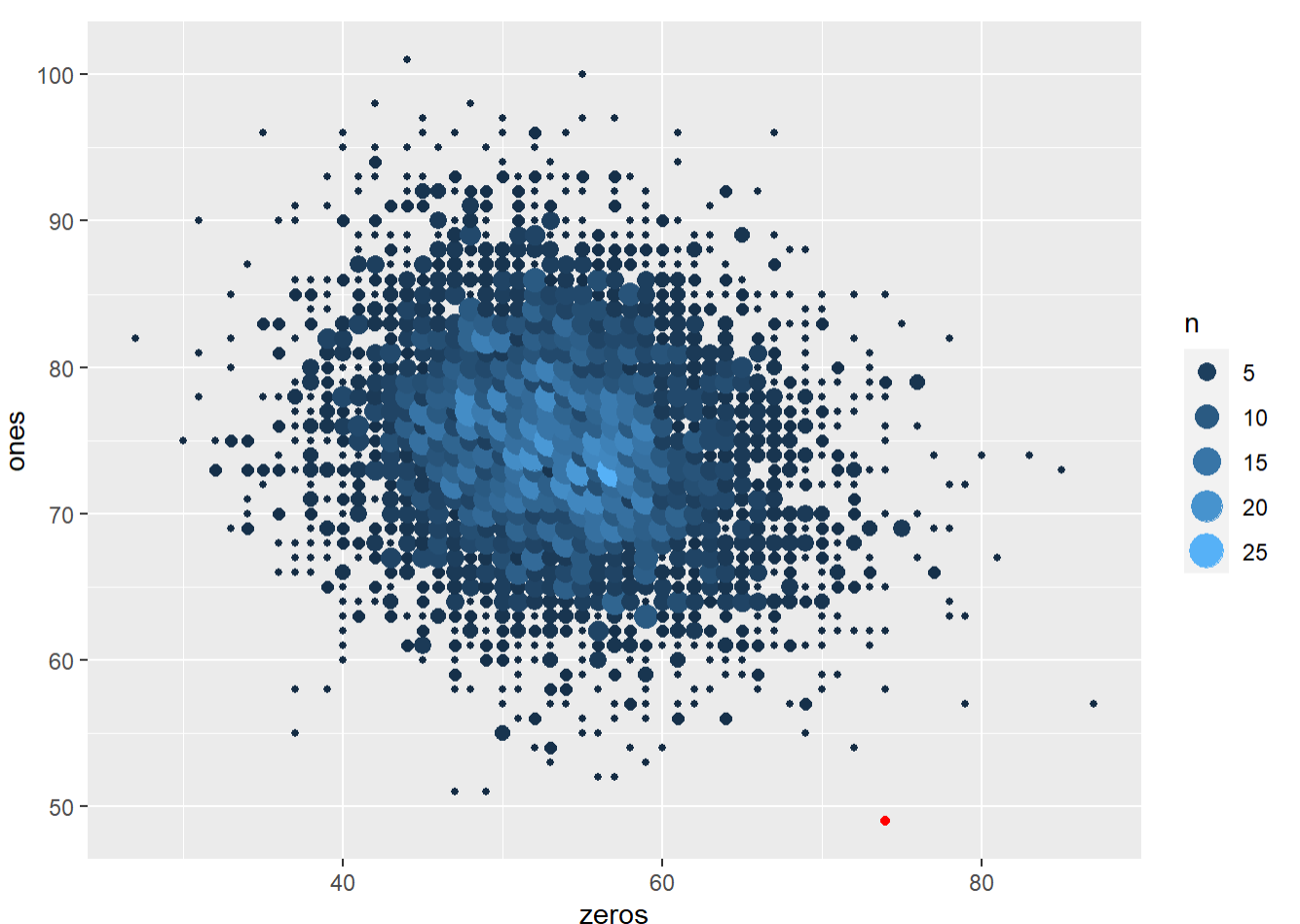
```
apply(simu,2,function(x) names(sort(-table(x)))[1])  # the mode of each group
```

```
## empirical   poisson posterior
##       "0"       "1"       "1"
```

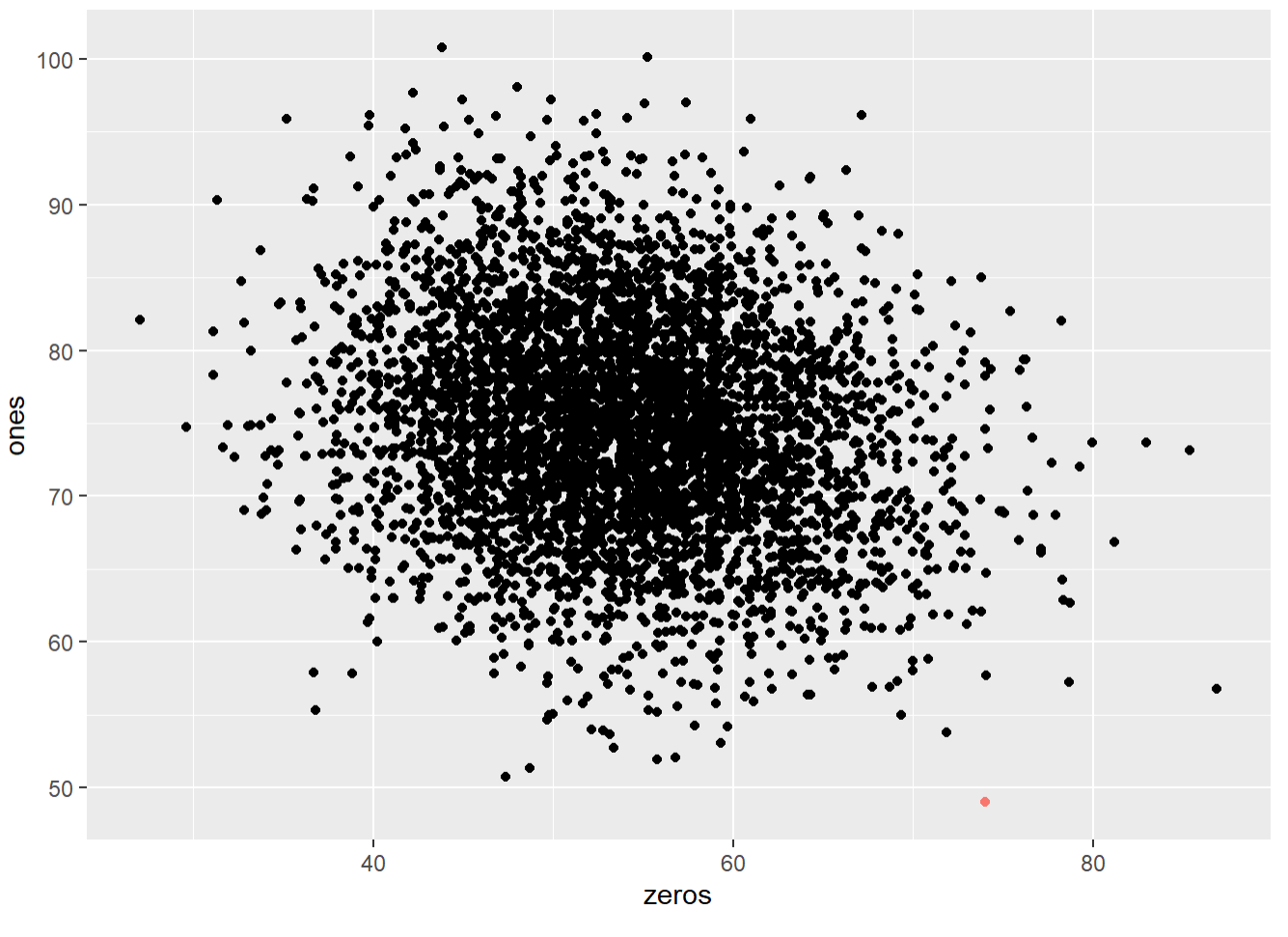
The mean of empirical and posterior are both close to 1.4, but the overall shape of density function and the peak of empirical group is different to the Poisson group. I would say a Poisson model is quite unreasonable.

d

```
df <- sapply(theta_nobach, function(theta){
  dat <- rpois(218, theta)
  ones <- sum(dat==1)
  zeros <- sum(dat==0)
  c(zeros,ones)
}) %>% t() %>% data.frame()
colnames(df) <- c("zeros","ones")
ggplot(df, aes(x=zeros, y=ones)) +
  geom_count(aes(color = ..n.., size = ..n..))+
  annotate('point', x = sum(nobach==0), y = sum(nobach==1), color = 'red') +
  guides(color = 'legend')
```




```
ggplot(df) +
  geom_jitter(aes(x=zeros, y=ones))+
  geom_point(data=data.frame(nobach),aes(x=sum(nobach==0), y=sum(nobach==1),col="red"), show.legend = F)
```



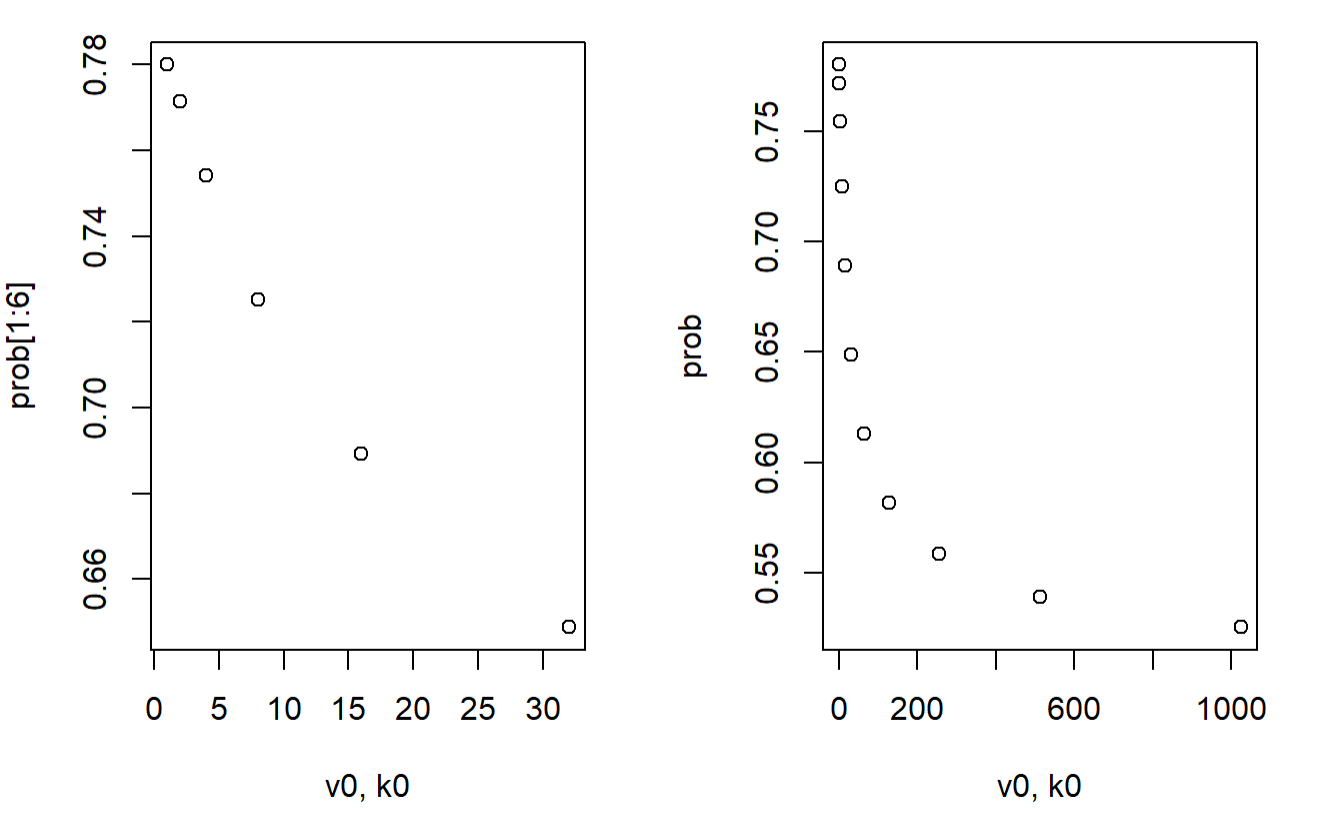
The red point is our observed data. This shows that Poisson model may not be a good fit.

5.2

```
B=1e4
mu_0 = 75 ; var_0 = 100
k0 = v0 = 2**c(0:10)
mc_sampling <- function(k0,v0){
  # group A:
  n_A = 16 ; y_A = 75.2 ; s_A = 7.3
  kn_A = k0 + n_A ; vn_A = v0 + n_A ; mun_A = k0/kn_A*mu_0 + n_A/kn_A*y_A
  varn_A = (v0*var_0+(n_A-1)*s_A**2+(k0*n_A/(k0+n_A))*(y_A-mu_0)**2) / vn_A
  # group B:
  n_B = 16 ; y_B = 77.5 ; s_B = 8.1
  kn_B = k0 + n_B ; vn_B = v0 + n_B ; mun_B = k0/kn_B*mu_0 + n_B/kn_B*y_B
  varn_B = (v0*var_0+(n_B-1)*s_B**2+(k0*n_B/(k0+n_B))*(y_B-mu_0)**2) / vn_B
  # MC method:
  set.seed(1)
  mc_var_A <- 1/(rgamma(B, vn_A/2,vn_A*varn_A/2))
  set.seed(2)
  theta_A <- rnorm(B,mun_A, sqrt(mc_var_A/kn_A))
  set.seed(3)
  mc_var_B <- 1/(rgamma(B, vn_B/2,vn_B*varn_B/2))
  set.seed(4)
  theta_B <- rnorm(B,mun_B, sqrt(mc_var_B/kn_B))
  mean(theta_A<theta_B)
}

prob <- sapply(1:11, function(i) mc_sampling(k0=k0[i], v0=v0[i]))

par(mfrow=c(1,2))
plot(k0[1:6],prob[1:6], xlab = "v0, k0")
plot(k0,prob, xlab = "v0, k0")
```




```
par(mfrow=c(1,1))
```

From the above plot, we can tell that the result is not so sensitive to the different priors. Under the range of 0~32 for v0 and k0, $Pr(\theta_A < \theta_B|y_A, y_B)$ is about 0.6 to 0.8. Given extremely strong prior belief could only turn the probability close to 0.5

5.4