Assignment #4

Yutae Lee (626005947)

2022-10-03

Problem 4.1

Posterior comparisons: Reconsider the sample survey in Exercise 3.1. Suppose you are interested in comparing the rate of support in that county to the rate in another county. Suppose that a survey of sample size 50 was done in the second county, and the total number of people in the sample who supported the policy was 30. Identify the posterior distribution of θ_2 assuming a uniform prior. Sample 5,000 values of each of θ_1 and θ_2 from their posterior distributions and estimate $Pr(\theta_1 < \theta_2|\text{the data and prior})$.

We know that $p(y_2|\theta_2) \sim Bin(p_2|n, \theta_2)$ $p(\theta_2) \sim beta(\theta_2|1, 1)$

Now to get posterior distribution:

$$p(\theta_2|y_2) = p(y_2|\theta_2) * p(\theta_2) = \binom{n}{y_2}\theta_2^{y_2} * (1-\theta_2)^{n-y_2} * \frac{\Gamma(2)}{\Gamma(1)\Gamma(1)} \propto \theta_2^{y_2} * (1-\theta_2)^{n-y_2} \propto beta(\theta_2|y+1, n-y+1)$$

In our case n = 50 and $y_2 = 30$ meaning that $p(\theta_2|y_2) \propto beta(\theta_2|31, 21)$

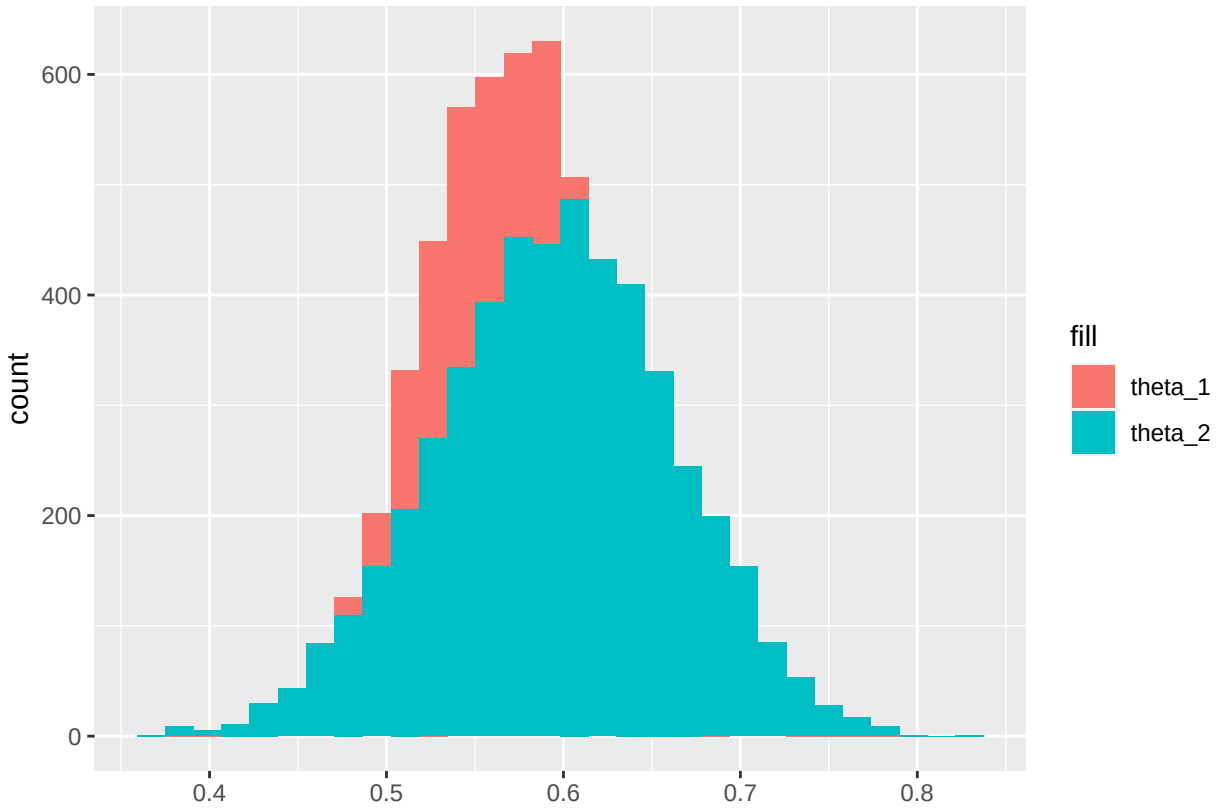
likewise for first country $p(\theta_1|y_1) \propto beta(\theta_1|58, 44)$

```
library(ggplot2)
```

```
## Warning:    'ggplot2'  R   4.1.3
```

```
set.seed(32)
theta_1 = rbeta(5000,58,44)
theta_2 = rbeta(5000,31,21)
df = data.frame(theta_1,theta_2)
ggplot(df) +
    geom_histogram(aes(x = theta_1, fill = "theta_1")) +
    geom_histogram(aes(x = theta_2, fill = "theta_2")) +
    labs(x = "")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
mean(theta_1 < theta_2)
```

```
## [1] 0.6276
```

From this using Monte Carlo approximation I get $Pr(\theta_1 < \theta_2|\text{the data and prior})$ approximately 0.6276.

Problem 4.2

Tumor count comparisons: Reconsider the tumor count data in Exercise 3.3:

a) For the prior distribution given in part a) of that exercise, obtain $Pr(\theta_B < \theta_A|y_A, y_B)$ via Monte Carlo sampling.

prior distribution from exercise 3.3 was given as following:

$\theta_A \sim gamma(120, 10)$, $\theta_B \sim gamma(12, 1)$, $p(\theta_A, \theta_B) = p(\theta_A) * p(\theta_B)$

```
mean_yA = (12+9+12+14+13+13+15+8+15+6)/10
mean_yB = (11+11+10+9+9+8+7+10+6+8+8+9+7)/13
print(mean_yA)
```

```
## [1] 11.7
```

```
print(mean_yB)
```

```
## [1] 8.692308
```

To conjugate to poisson we have to use $gamma(a + n\bar{y}, b + n)$

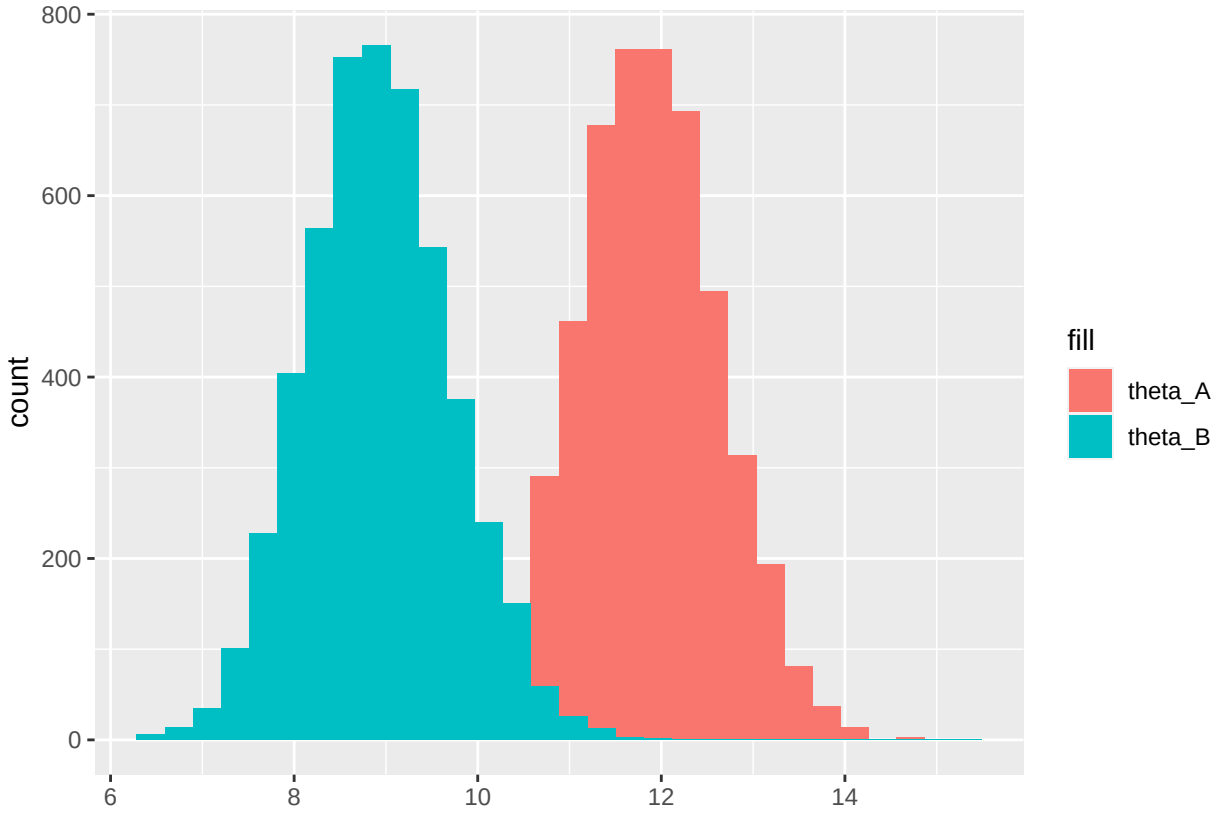
In case of θ_A, n = 10 and $\bar{y} = 11.7$ which means it conjugates to gamma(237,20).

In case of θ_B, n = 13 and $\bar{y} = 8.692308$ which means it conjugates to gamma(125,14).

Using r let's make thes samples

```
set.seed(32)
theta_A = rgamma(5000,237,scale = 1/20)
theta_B = rgamma(5000,125,scale = 1/14)
df2 = data.frame(theta_A,theta_B)
ggplot(df2) +
    geom_histogram(aes(x = theta_A, fill = "theta_A")) +
    geom_histogram(aes(x = theta_B, fill = "theta_B")) +
    labs(x = "")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
mean(theta_B < theta_A)
```

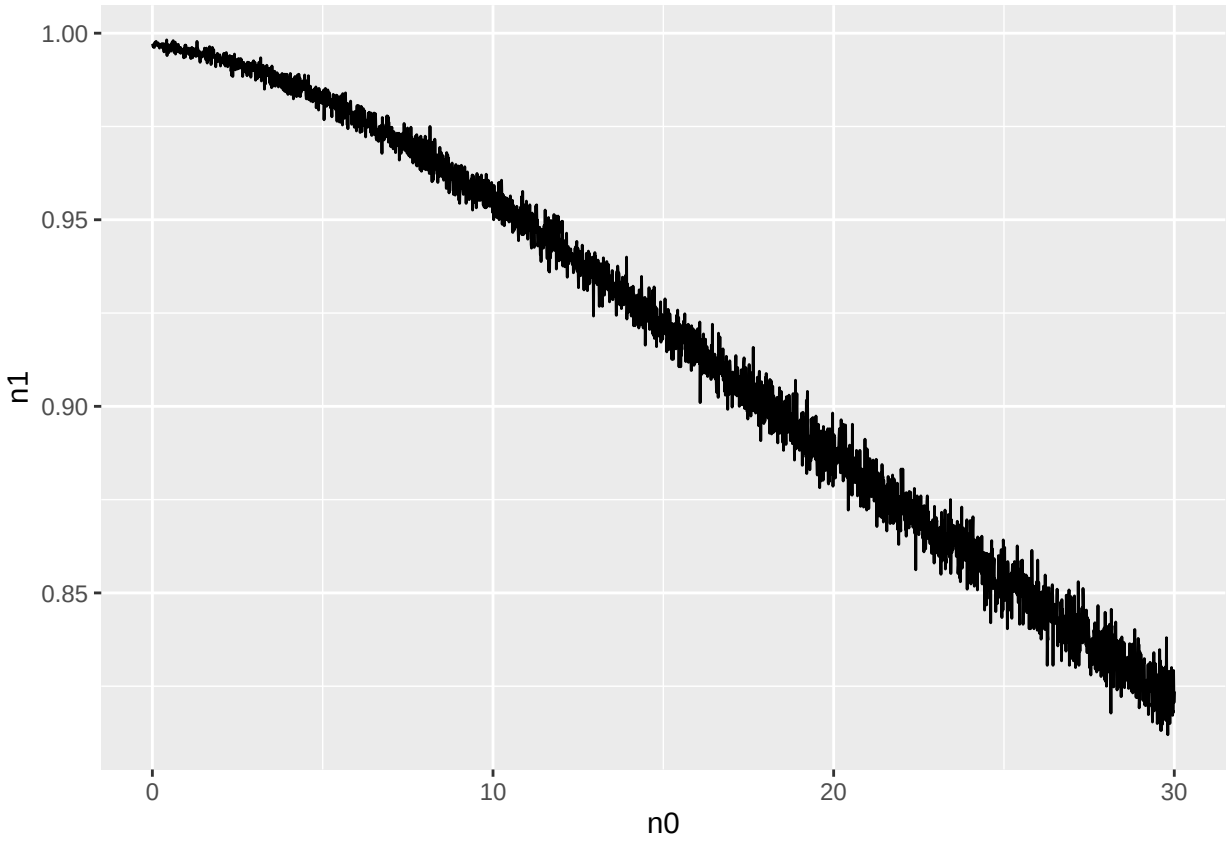
```
## [1] 0.9946
```

Using Monte Carlo approximation, I get 0.9946 as our mean.

b) For a range of values of n_0, obtain $Pr(\theta_B < \theta_A|y_A, y_B)$ for $\theta_A \sim gamma(120, 10)$ and $\theta_B \sim gamma(12 \times n_0, n_0)$. Describe how sensitive the conclusions about the event $\theta_B < \theta_A$ are to the prior distribution on θ_B.

```
probability_calculation <- function(n0){
  return(mean(rgamma(5000, 12 * n0 + 113, scale = 1/(n0+13)) < rgamma(5000,237,scale = 1/20)))
}
```

```
n0 <- seq(0, 30, length.out = 3001)
n1 <- c()
for (i in n0){
  n1 <- c(n1,probability_calculation(i))}
df3 <- data.frame(n0,n1)
ggplot(df3) +
  geom_line(aes(x = n0, y = n1))
```



As we can see from the plot as n0 grow from 0 to 30, it is obvious that $\theta_B < \theta_A$ is dropping by 20%. Intuitively speaking, it must be true since we are having more n0 which represent our confidence in the prior of θ_B. Also, we can see that rate of dropping is increasing i.e the line is concave.

c) Repeat parts a) and b), replacing the event $\theta_B < \theta_A$ with the event $\tilde{Y}_B < \tilde{Y}_A$, where $\tilde{Y}_A$ and $\tilde{Y}_B$ are samples from the posterior predictive distribution.

c)a)

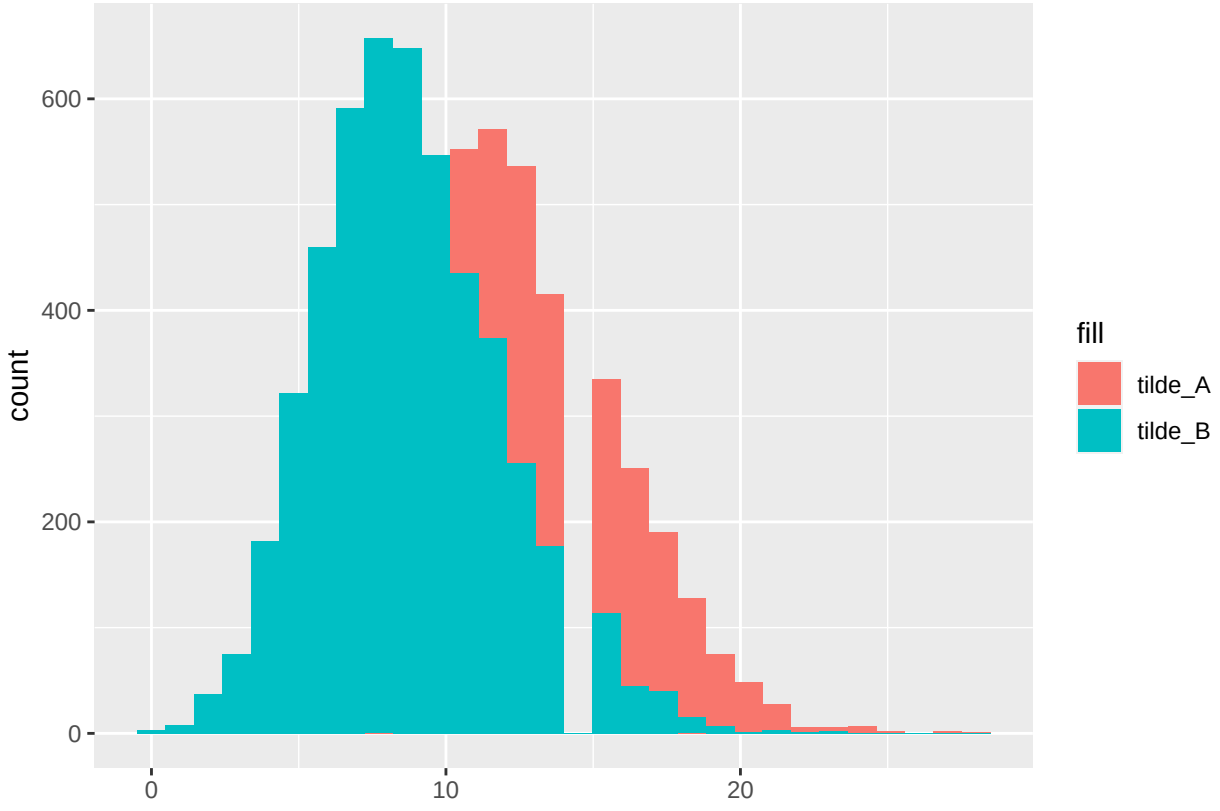
Doing the steps again using the samples from the posterior predictive distribution:

we know that $p(\theta_A|y_A) \propto gamma(237, 20)$ and $p(\theta_B|y_B) \propto gamma(125, 14)$ which are our conjugated value from a).

This time we are going to have condition on the samples from a):

```
set.seed(32)
tilde_A = rpois(5000,theta_A)
tilde_B = rpois(5000,theta_B)
df4 = data.frame(tilde_A,tilde_B)
ggplot(df4) +
    geom_histogram(aes(x = tilde_A, fill = "tilde_A")) +
    geom_histogram(aes(x = tilde_B, fill = "tilde_B")) +
    labs(x = "")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
mean(tilde_B < tilde_A)
```

```
## [1] 0.6902
```

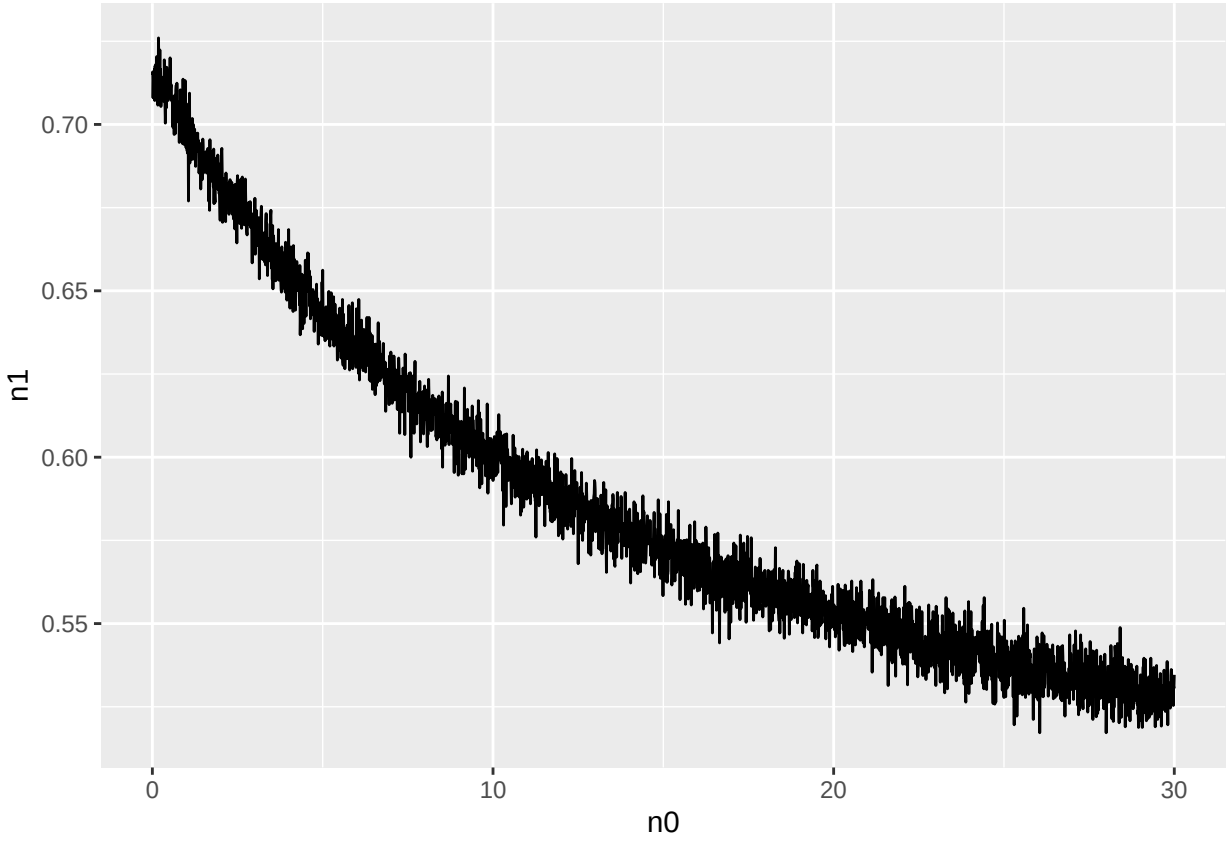
We can see that through monte carlo approximation it $\tilde{Y}_B < \tilde{Y}_A$ is about 0.6902.

c)b)

Doing the same process from b) using $\tilde{Y}_A, \tilde{Y}_B$

```
prob_calc<- function(n0){
  return(mean(rpois(5000,rgamma(5000, 12 * n0 + 113, scale = 1/(n0+13))) < tilde_A))
}
```

```
n0 <- seq(0, 30, length.out = 3001)
n1 <- c()
for (i in n0){
  n1 <- c(n1,prob_calc(i))}
df5 <- data.frame(n0,n1)
ggplot(df5) +
  geom_line(aes(x = n0, y = n1))
```



Similar to part b) we can see that as n0 grows the probability of $\tilde{Y}_B < \tilde{Y}_A$ decreases. However unlike part b) we can see the rate of decreasing decreases as n0 grows i.e it is convex.

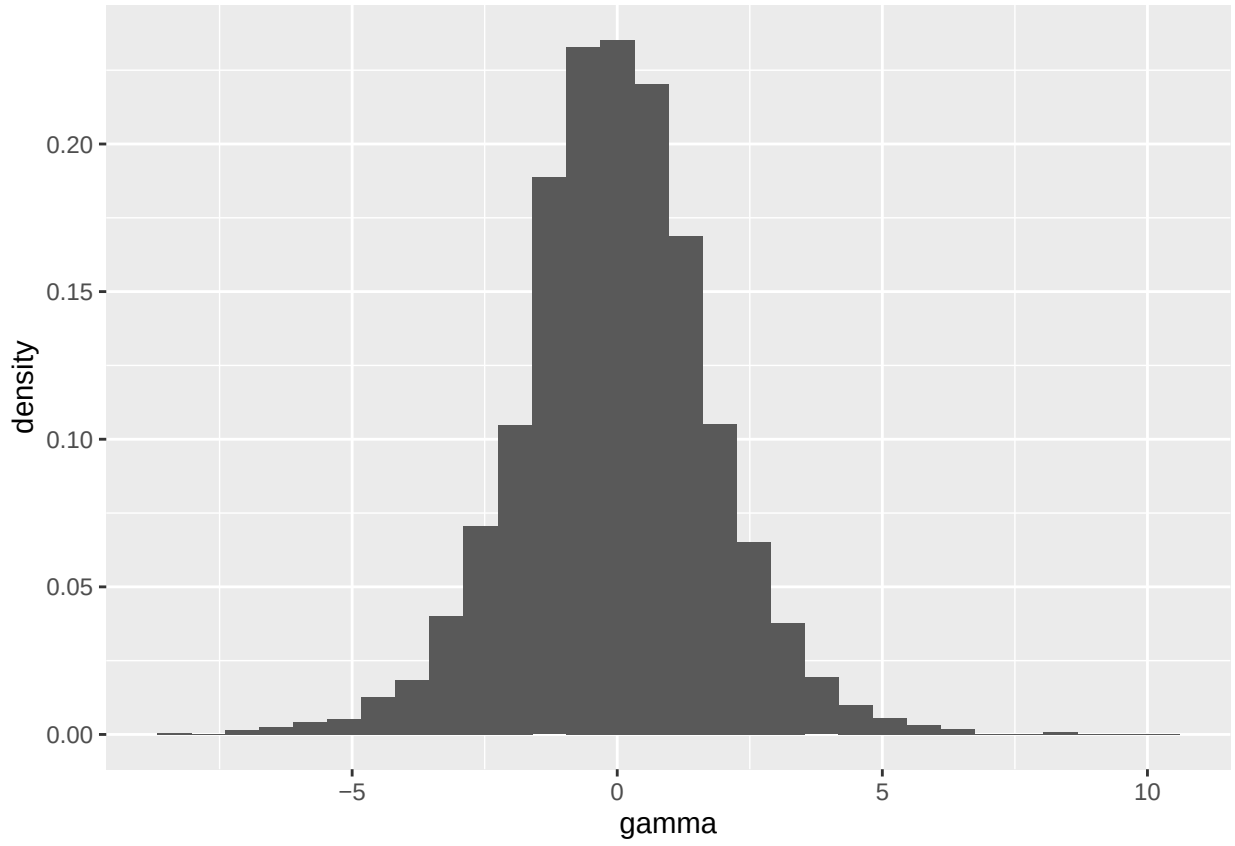
Problem 4.6

Non-informative prior distributions: Suppose for a binary sampling problem we plan on using a uniform, or beta(1,1), prior for the population proportion θ. Perhaps our reasoning is that this represents "no prior information about θ." However, some people like to look at proportions

on the log-odds scale, that is, they are interested in $\gamma = log\frac{\theta}{1-\theta}$. Via Monte Carlo sampling or otherwise, find the prior distribution for γ that is induced by the uniform prior for θ. Is the prior informative about γ?

```
set.seed(32)
theta = runif(5000)
gamma_func <- function(x){
  return(log(x/(1-x)))
}
gamma <-c()
for (i in theta){
  gamma <- c(gamma,gamma_func(i))
}
df_gam <- data.frame(gamma)
ggplot(df_gam) +
  geom_histogram(aes(x = gamma, y = ..density..))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



It appears that this fairly uniform, however this is not absolutely uniform centered on 0. So we can say that this is actually informative rather than non-informative. So a better way to have non-informative prior maybe a Jeffrey's prior.

Problem 4.8

More posterior predictive checks: Let A and B be the average number of children of men in their 30s with and without bachelor's degrees, respectively.

a) Using a Poisson sampling model, a gamma(2,1) prior for each θ and the data in the files menchild30bach.dat and menchild30nobach.dat, obtain 5,000 samples of $\tilde{Y}_A$ and $\tilde{Y}_B$ from the posterior predictive distribution of the two samples. Plot the Monte Carlo approximations to these two posterior predictive distributions.

```
menchild30bach = scan(url('https://www2.stat.duke.edu/courses/Fall09/sta290/datasets/Hoffdata/menchild30
menchild30nobach = scan(url('https://www2.stat.duke.edu/courses/Fall09/sta290/datasets/Hoffdata/menchil
```

```
no_bach = menchild30nobach
yes_bach = menchild30bach
sum_no_bach = sum(no_bach)
sum_yes_bach = sum(yes_bach)
num_no_bach = length(no_bach)
num_yes_bach = length(yes_bach)
```

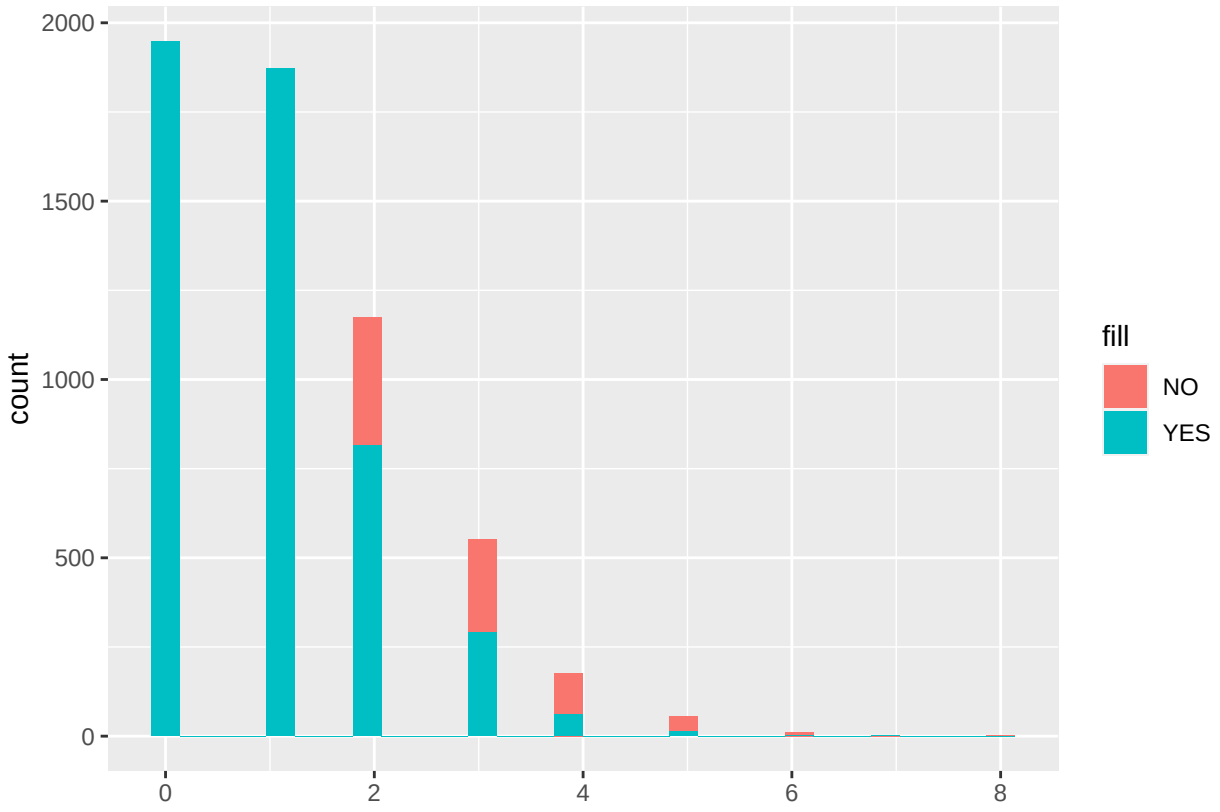
Now we have a gamma(2,1) prior for each θ so..

```
set.seed(32)
theta_no = rgamma(5000, 2 + sum_no_bach, scale = 1/(1+num_no_bach))
theta_yes = rgamma(5000, 2 + sum_yes_bach, scale = 1/(1+num_yes_bach))
```

Now we have to get samples conditioned on our posterior samples

```
set.seed(32)
new_theta_no = rpois(5000, theta_no)
new_theta_yes = rpois(5000, theta_yes)
df_bach = data.frame(new_theta_no,new_theta_yes)
ggplot(df_bach) +
    geom_histogram(aes(x = new_theta_no, fill = "NO")) +
    geom_histogram(aes(x = new_theta_yes, fill = "YES")) +
    labs(x = "")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



b) Find 95% quantile-based posterior confidence intervals for $\theta_B - \theta_A$ and $\tilde{Y}_B - \tilde{Y}_A$. Describe in words the differences between the two populations using these quantities and the plots in a), along with any other results that may be of interest to you.

```
theta_difference = theta_no-theta_yes
print(quantile(theta_difference, c(0.025, 0.975)))
```

```
##      2.5%     97.5%
## 0.1524051 0.7297068
```

```
print('')
```

```
## [1] ""
```

```
print('')
```

```
## [1] ""
```

```
new_difference = new_theta_no-new_theta_yes
print(quantile(new_difference, c(0.025, 0.975)))
```

```
##  2.5% 97.5%
##    -3     3
```

we can see that the 95% c.i for $\theta_B - \theta_A$ is 0.1524051 and 0.7297068

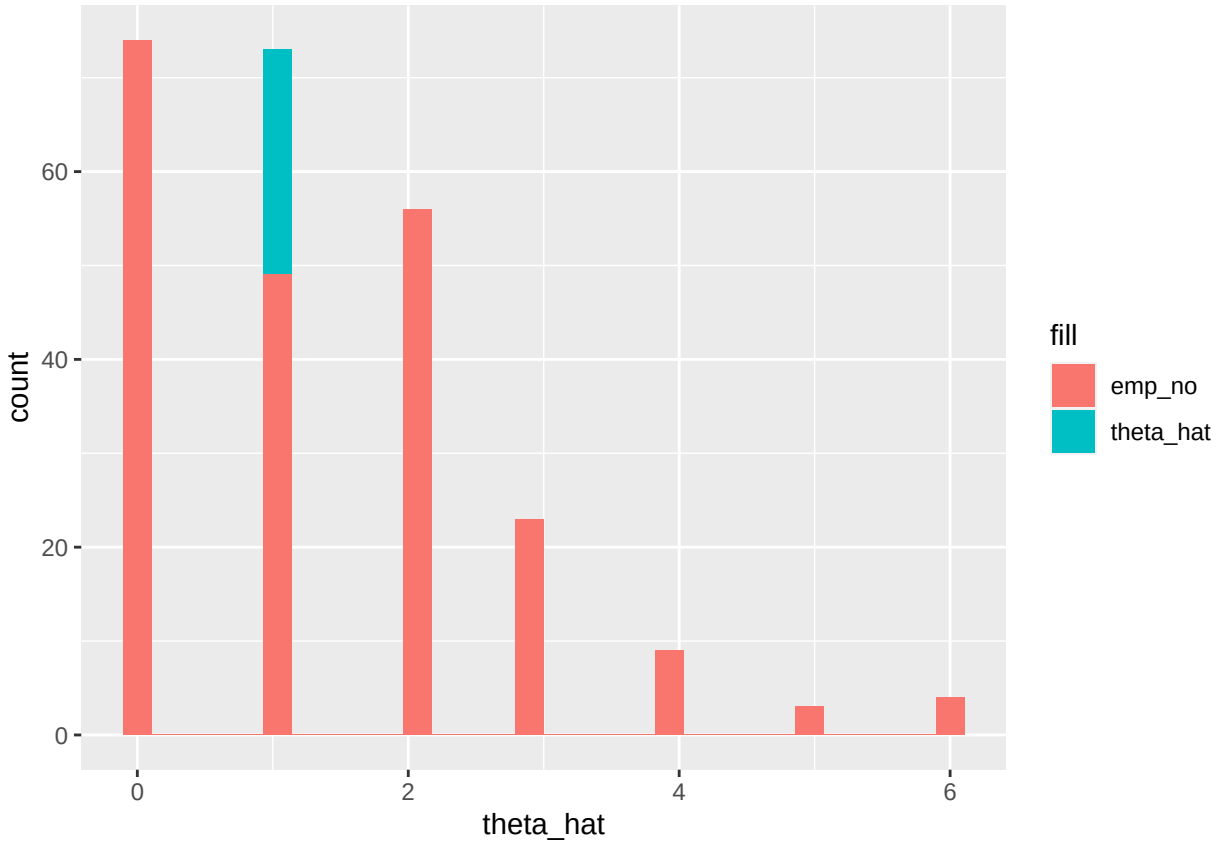
And 95% c.i for $\tilde{Y}_B - \tilde{Y}_A$ is actually -3 and 3.

Since, confidence interval for $\theta_B - \theta_A$ does not contain 0 we can confidently say that θ_B is greater than θ_A. However in the case of $\tilde{Y}_B - \tilde{Y}_A$, we can see that it is centralized at 0, so it does not hold the same explanation.

c) Obtain the empirical distribution of the data in group B. Compare this to the Poisson distribution with mean $\hat{\theta} = 1.4$. Do you think the Poisson model is a good fit? Why or why not?

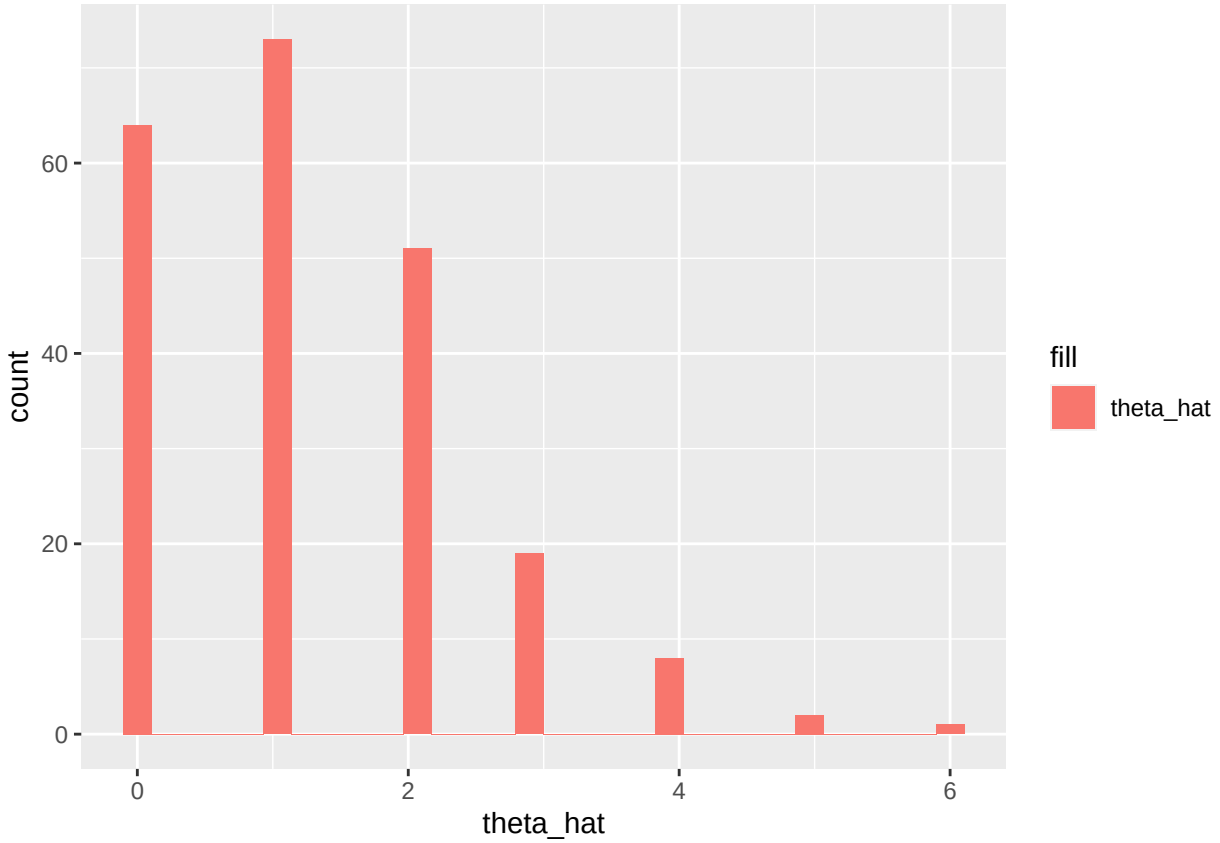
```
set.seed(32)
theta_hat = rpois(218, 1.4)
df_hat_no <- data.frame(theta_hat, no_bach)
ggplot(df_hat_no) +
  geom_histogram(aes(x = theta_hat, fill = 'theta_hat')) +
  geom_histogram(aes(x = no_bach, fill ='emp_no'))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



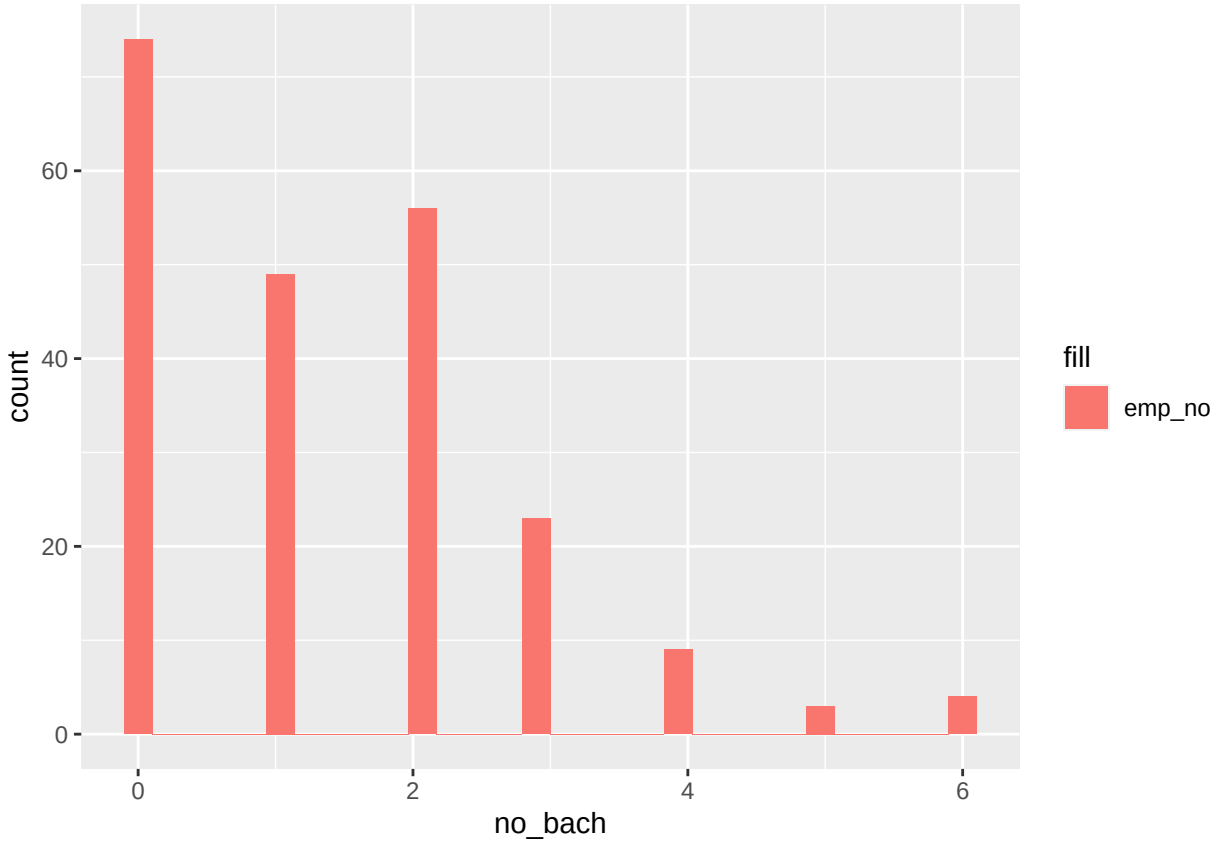

```
ggplot(df_hat_no) +
  geom_histogram(aes(x = theta_hat, fill = 'theta_hat'))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
ggplot(df_hat_no) +
  geom_histogram(aes(x = no_bach, fill ='emp_no'))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

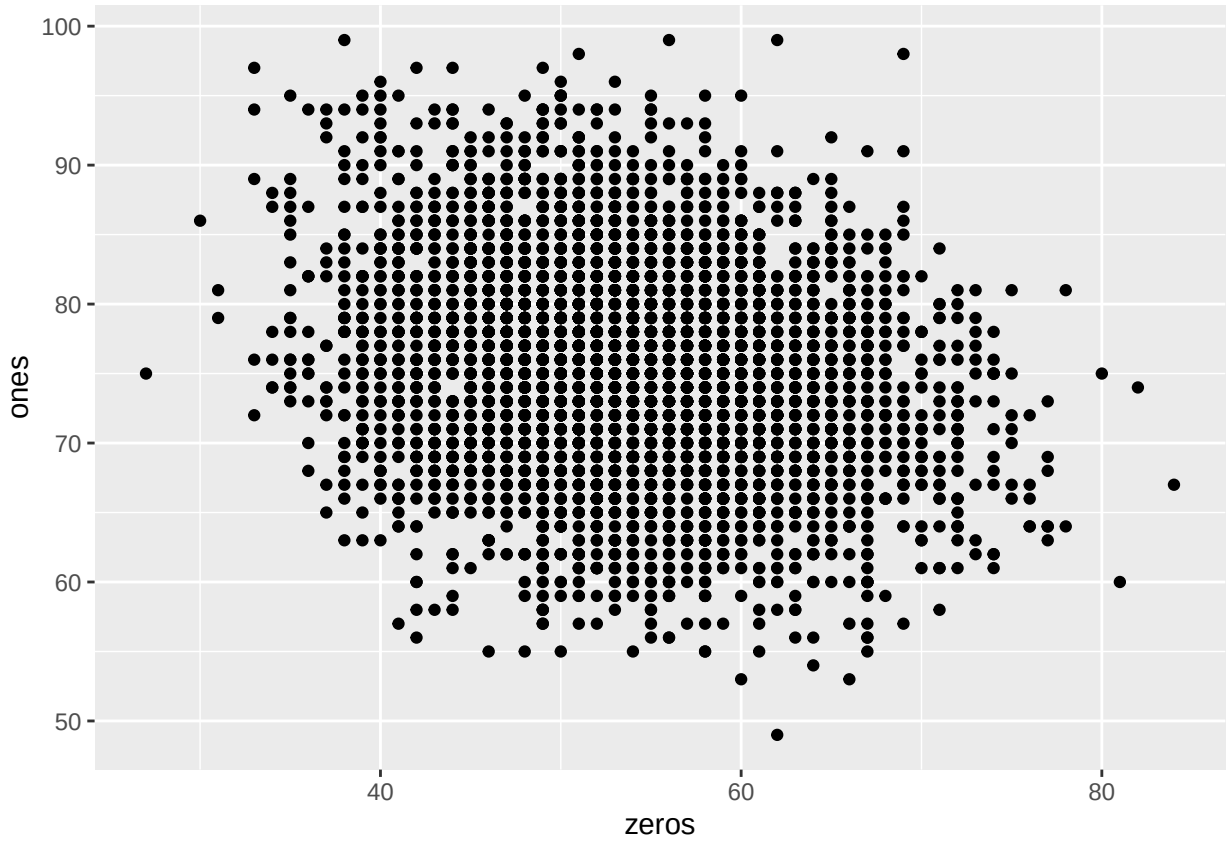


The two histogram lines up actually lines up really close to each other to the point where I have to draw a two different histogram separately again. Thus, the poisson model with 1.4 seems appropriate model.

d) For each of the $5,000\theta_B$ -values you sampled, sample $n_B = 218$ Poisson random variables and count the number of 0s and the number of 1s in each of the 5,000 simulated datasets. You should now have two sequences of length 5,000 each, one sequence counting the number of people having zero children for each of the 5,000 posterior predictive datasets, the other counting the number of people with one child. Plot the two sequences against one another (one on the x-axis, one on the y-axis). Add to the plot a point marking how many people in the observed dataset had zero children and one child. Using this plot, describe the adequacy of the Poisson model.

```
simulator <- c()
for (i in theta_no){
  simulator <- c(simulator,rpois(218,i))
}
zeros <- c()
ones <- c()
for (i in seq(0,4999, length.out = 5000)){
  a = 1 + 218*i
  b = 218 + 218*i
  zeros <- c(zeros,sum(simulator[a:b] == 0))
  ones <- c(ones,sum(simulator[a:b] == 1))
}
```

```r
df_zeros_ones <- data.frame(zeros,ones)
ggplot(df_zeros_ones) +
  geom_point(aes(x = zeros,y = ones))
```



As we can see from here poisson model may not be the best to represent. As we there should be much more zeros than the prediction of the poisson and less ones.