Homework 2

Bingying Liu

January 28, 2020

Question 1

Using the inverse cdf method, generate 1,000 random realizations from the Beta(5,10) distribution truncated to the interval (0.4,0.75).

* What is the mean of your random draws (rounded to 2 decimal places)?

* What is the variance of your random draws (rounded to 2 decimal places)?

```
# generate 1000 random realizations from Beta(5,10) truncated to (0.4, 0.75)
set.seed(123)
u <- runif(1000,0,1)
fa <- pbeta(0.4, shape1=5, shape2 = 10)
fb <- pbeta(0.75, shape1=5, shape2 = 10)
u_trun <- (fb-fa)*u+fa
theta <- qbeta(u_trun, shape1=5, shape2 = 10)

# mean of the random draw
theta_mean <- sum(theta)/1000
round(theta_mean,2)
```

```
## [1] 0.48
```

```
# variance of the random draw
theta_var <- sum((theta-theta_mean)^2)/(1000-1)
round(theta_var,2)
```

```
## [1] 0
```

Question 2:

1. Find the conjugate family of priors for b.

$$\begin{aligned} p(b|y) &\propto p(y|b)p(b) \\ &= p(b)\frac{b^a}{\Gamma(a)}y^{a-1}e^{-by} \\ &\propto p(b)b^a e^{-b} \end{aligned}$$

Therefore, our prior should include terms like $b^{c_1}e^{-b}$. If we have,

$$p(b) = \text{Gamma}(a_0, b_0) = \frac{b_0^{a_0}}{\Gamma(a_0)}b^{a_0-1}e^{-b_0b}$$

then we'll have

$$\begin{aligned} p(b|y) &\propto b^{a_0-1}e^{-b_0b}b^a e^{-by} \\ &\propto b^{a_0+a-1}e^{-b(b_0+y)} \end{aligned}$$

which is also a Gamma distribution: $\text{Gamma}(a_0+a, b_0+y)$.

2. Find the corresponding posterior given the prior you identified in the previous part.

Since we only care about b, we could let a be a constant in this case. If we have a prior

$$p(b) = \text{Gamma}(a_0, b_0) = \frac{b_0^{a_0}}{\Gamma(a_0)}b^{a_0-1}e^{-b_0b}$$

Then the posterior will be

$$\begin{aligned} p(b|y_1,\dots,y_n) &\propto p(y_1,\dots y_n|b)p(b) \\ &= \frac{b_0^{a_0}}{\Gamma(a_0)} b^{a_0-1} e^{-b_0 b} \sum_{i=1}^{n} \frac{b^a}{\Gamma(a)} y_i^{a-1} e^{-by_i} \\ &\propto b^{a_0-1} e^{-b_0 b} b^{na} e^{-b\sum_{i=1}^n y_i} \\ &= b^{a_0+an-1} e^{-b(b_0+\sum_{i=1}^n y_i)} \end{aligned}$$

Therefore, we find that

$$p(b|y_1,\dots,y_n) \propto \text{Gamma}(a_0+an, b_0+\sum_{i=1}^{n} y_i)$$

3. Give an interpretation of the prior parameters as things like “prior mean”, “prior variance”, “prior sample size”, etc.

Since posterior is a Gamme distribution, we can get the posteror mean as following:

$$\begin{aligned} E(b|y_1,\dots,y_n) &= \frac{a_0+an}{b_0+\sum_{i=1}^n y_i} \\ &= \frac{a_0}{b_0+\sum_{i=1}^n y_i} * \frac{b_0}{b_0} + \frac{an}{b_0+\sum_{i=1}^n y_i} \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n y_i} \\ &= \frac{b_0}{b_0+\sum_{i=1}^n y_i} \frac{a_0}{b_0} + \frac{\sum_{i=1}^n y_i}{b_0+\sum_{i=1}^n y_i} \frac{a}{\frac{\sum_{i=1}^n y_i}{n}} \\ &= \frac{b_0}{b_0+\sum_{i=1}^n y_i} * \text{prior mean} + \frac{\sum_{i=1}^n y_i}{b_0+\sum_{i=1}^n y_i} * \text{MLE of } b \end{aligned}$$

Interpretation: The posterior mean is a weighted average of the prior mean and the MLE of b (parameter of the posterior Gamma distribution), with weights proportional to b_0 and $\sum_{i=1}^n y_i$ respectively. Therefore, MLE of b depends on sum of samples and prior mean depends on b_0 which is our prior belief.

Question 3: Hoff 3.3

1. Find the posterior distributions, means, variances and 95% quantile-based confidence intervals for θ_A and θ_B

```
aA <- 120; bA <- 10
y_a <- c(12,9,12,14,13,13,15,8,15,6)
y_b <- c(11,11,10,9,9,8,7,10,6,8,8,9,7)
sum(y_a);sum(y_b)
```

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

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

The posterior distribution for θ_A is

$$\begin{aligned} p(\theta_A|y_A) &\propto gamma(a+\sum_i y_{ai}, b+n) \\ &= gamma(120+117, 10+10) \\ &= gamma(237, 20) \end{aligned}$$

The posterior mean and variance for θ_A is $\frac{237}{20} = 11.85$ and $\frac{237}{20^2} = 0.5925$ respectively.

The posterior distribution for θ_B is

$$\begin{aligned} p(\theta_B|y_B) &\propto gamma(a+\sum_i y_{bi}, b+n) \\ &= gamma(12+113, 1+13) \\ &= gamma(125, 14) \end{aligned}$$

The posterior mean and variance for θ_B is $\frac{125}{14} = 8.93$ and $\frac{125}{14^2} = 0.638$ respectively.

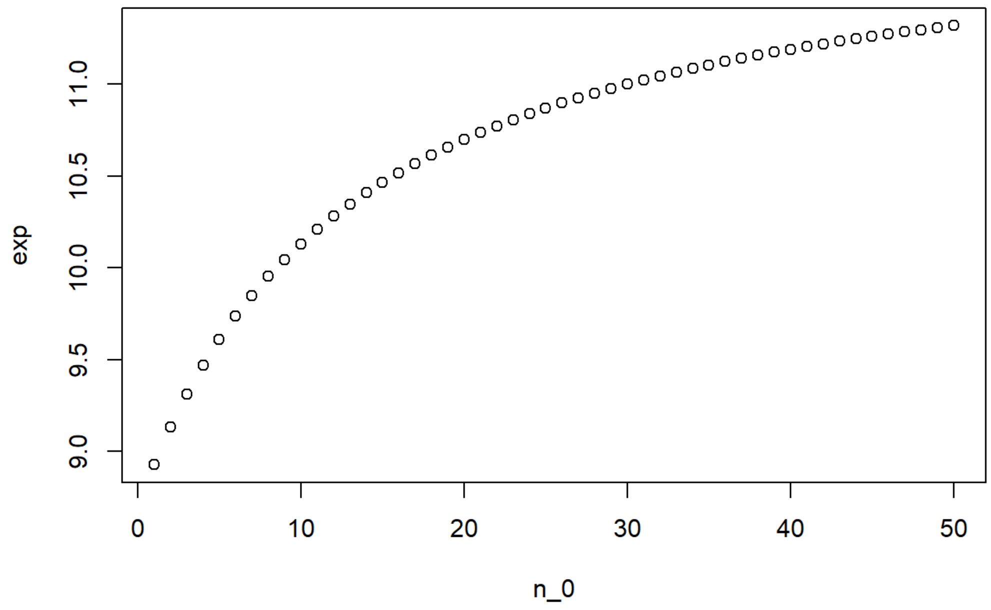
```
A_quantile <- qgamma(c(0.025,0.975), 237, 20)
B_quantile <- qgamma(c(0.025,0.975), 125, 14)
A_quantile;B_quantile
```

```
## [1] 10.38924 13.40545
## [1]  7.432064 10.560308
```

The 95% quantile-based confidence intervals for θ_A is $(10.38924, 13.40545)$. The 95% quantile-based confidence intervals for θ_B is $(7.43206410.560308)$.

2. Compute and plot the posterior expectation of θ_B under the prior distribution $\theta_B \sim gamma(12 * n_0, n_0)$ for each value of $n_0 \in \{1, 2, \ldots, 50\}$. Describe what sort of prior beliefs about θ_B would be necessary in order for the posterior expectation of θ_B tp e close to that of θ_A.

```
n_0 <- seq(1,50,by=1)
exp <- (12*n_0 + 113)/(n_0 + 13)
plot(n_0,exp)
```



Since $E(\theta_a|y) = 11.85$, in order for $E(\theta_b)$ to be close to 11.85, we need n_0 to be close to or even more than more than 50, which means we need a very strong prior belief of θ_b.

3. Should knowledge about population A tells us anything about population B? Discuss whether or not it makes sense to have $p(\theta_A, \theta_B) = p(\theta_A) * p(\theta_B)$.

If study shows that type A mice and type B mice are completely different in tumor counts, then we can assume that $p(\theta_A, \theta_B) = p(\theta_A) * p(\theta_B)$, which means type A mice and type B mice are independent with regard to the number of tumor counts.

However,if there is some relationship between type A mice and type B mice on number of tumor counts, we need to be more careful when making independent assumption. In the case above, type A mice and type B mice might not be independent because their prior mean is the same. If we know they're related, then our prior beliefs about the tumor rate in A should not be independent of tumor rate in B.

Question 4: Hoff 4.1

Given θ_2 has a uniform prior $\text{Beta}(1, 1)$, sampling distribution is binomal and there are 30 people out of 50 supported the policy, we have $p(\theta_2|y) = \text{Beta}(1 + 30, 1 + 50 - 30) = \text{Beta}(31, 21)$.

```
# estimate $Pr(\theta_1 < \theta_2 | the data and prior)
set.seed(123)
theta1.mc <- rbeta(5000,58,44)
theta2.mc <- rbeta(5000,31,21)
mean(theta1.mc<theta2.mc) #0.636
```

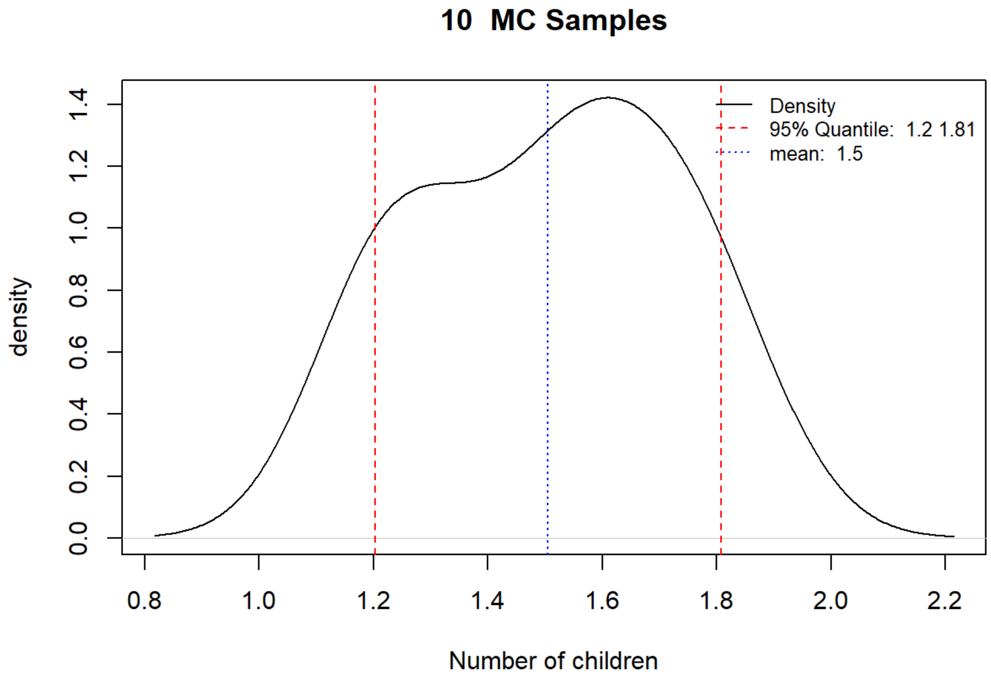
```
## [1] 0.636
```

The estimated probability of $\Pr(\theta_1 < \theta_2 | \text{the data and prior}) = 0.636$.

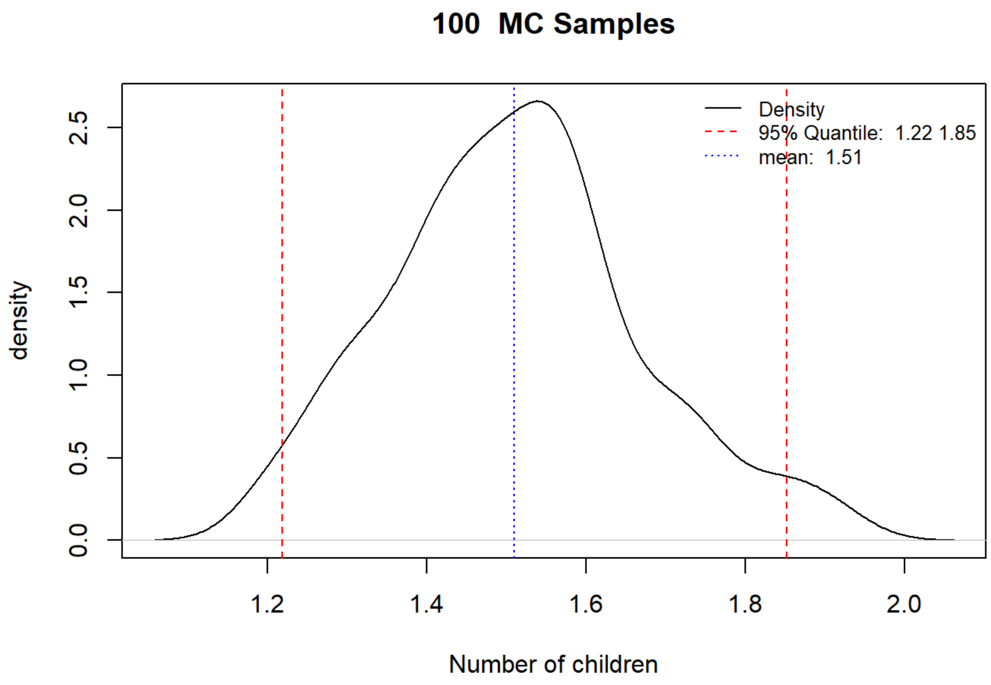
Question 5: How many samples is enough?

1. For each m, make a plot of the random draws and on the same plot, mark the points correspoding to the posterior mean and the 95% equal-tailed credible interval (quantile-based). How do those compare to the true posterior mean and 95% quantile-based CI?

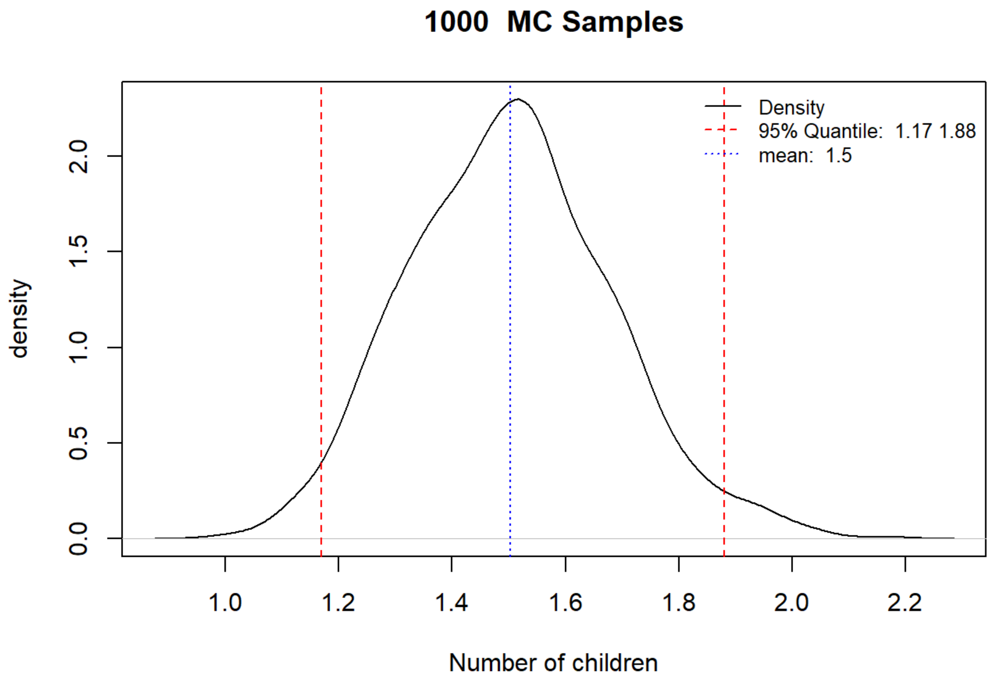
```r
sample_plot <- function(samples){
  set.seed(123)
  theta.mc <- rgamma(samples, 68, 45)
  mean.mc <- mean(theta.mc)
  probs.mc <- quantile(theta.mc, c(0.025,0.975))

  plot(density(theta.mc), xlab="Number of children", ylab="density", col="black",lty=1,main=paste(samples, " MC Samples"))
  abline(v = probs.mc, col=c("red"), lty=2, lwd=c(1))
  abline(v = mean.mc, col="blue", lty=3,lwd=1)
  legend("topright", legend=c("Density", paste("95% Quantile: ", round(probs.mc[[1]],2), round(probs.mc[[2]],2)), paste("mean: ", round(mean.mc,2))),
         col=c("black", "red","blue"), lty=1:3, cex=0.8, bg="transparent", bty = "n")
}

sample_plot(10)
```



```r
sample_plot(100)
```



```
sample_plot(1000)
```



```
# true mean and quantile-based CI
true_mean <- 68/45
true_quantile <- qgamma(c(0.025,0.975),68,45)
true_mean; true_quantile
```

```
## [1] 1.511111
```

```
## [1] 1.173437 1.890836
```

The true mean is 1.51 and 95% quantile-based CI is $(1.1734371.890836)$.
As long as number of samples for MC approximation is big enough (in this case 1000 samples is the most accurate), the result of mean and 95% CI is very close to the true mean and CI.

2. In addition, calculate the posterior probability that $\theta_2 < 1.5$ in each case.

```
# to ensure samples are exactly the same as in sample_plot function, we set the same seed
set.seed(123)
theta.mc10 <- rgamma(10, 68, 45)
theta.mc100 <- rgamma(100, 68, 45)
theta.mc1000 <- rgamma(1000, 68, 45)

mean(theta.mc10 < 1.5);mean(theta.mc100 < 1.5); mean(theta.mc1000 < 1.5)
```

```
## [1] 0.4
```

```
## [1] 0.48
```

```
## [1] 0.509
```

Therefore, the posterior probability $P(\theta_2 < 1.5|Y)$ for 10, 100 and 1000 MC samples are 0.4, 0.48, 0.509 respectively.

3. How large should m be if 95% of the time we want the difference between the Monte Carlo estimate of the posterior mean and the true posterior mean to be <= 0.001?

```
var.estimate <- 68/(45^2)
s.min = var.estimate/(0.001/2)^2
s.min
```

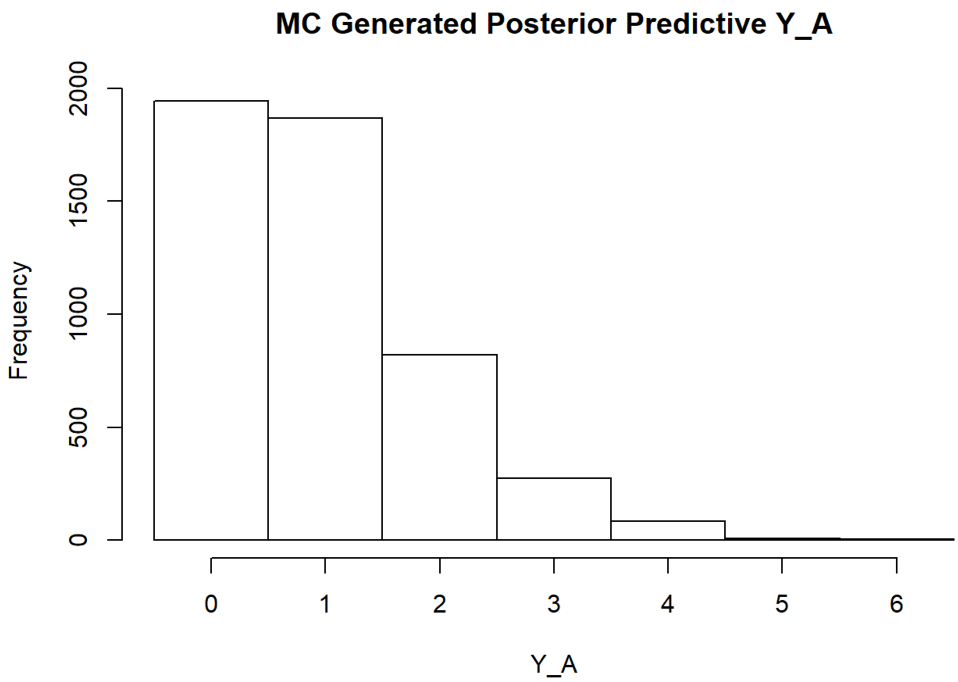
```
## [1] 134321
```

Using central limit theorem, after calculating variance of $\text{Gamma}(68, 45)$ and using MC standard error formula (to ensure posterior mean is within 2 std devs from the true mean), we need at least 134321 MC samples.

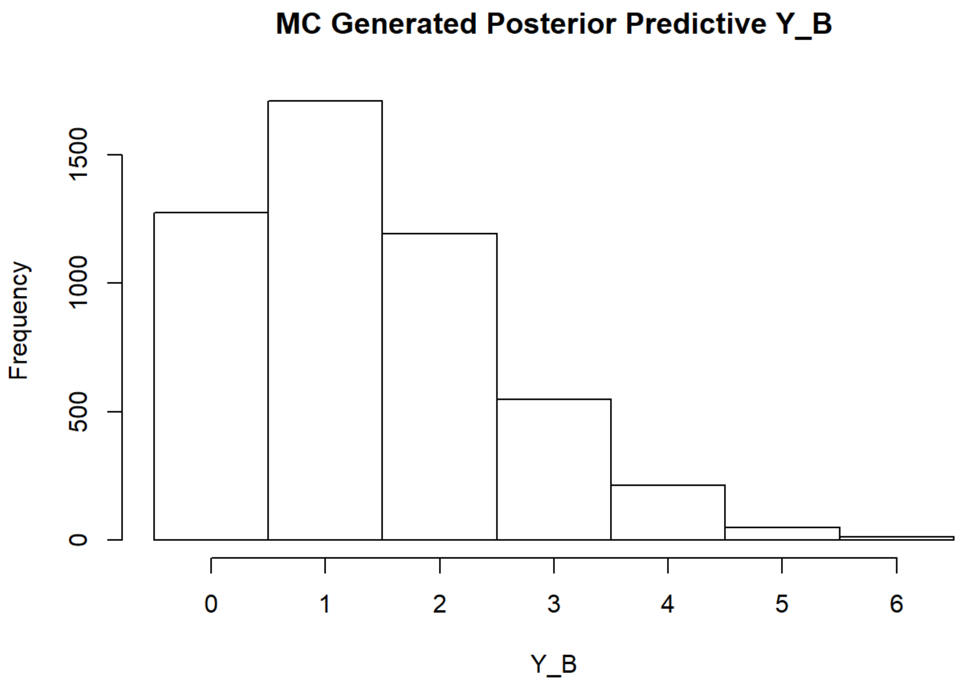
Question 6: Hoff 4.8

a. Obtain 5000 samples from the posterior predictive distribution. Plot the MC approximations to these two posterior predictive distributions.

```
man_A <- c(1, 0, 0, 1, 2, 2, 1, 5, 2, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 1, 1, 2, 1, 3,
            2, 0, 0, 3, 0, 0, 0, 2, 1, 0, 2, 1, 0, 0, 1, 3, 0, 1, 1, 0, 2, 0, 0, 2, 2, 1,
            3, 0, 0, 0, 1, 1)
man_B <- c(2, 2, 1, 1, 2, 2, 1, 2, 1, 0, 2, 1, 1, 2, 0, 2, 2, 0, 2, 1, 0, 0, 3, 6, 1, 6,
           4, 0, 3, 2, 0, 1, 0, 0, 0, 3, 0, 0, 0, 0, 0, 1, 0, 4, 2, 1, 0, 0, 1, 0, 3, 2,
           5, 0, 1, 1, 2, 1, 2, 1, 2, 0, 0, 0, 2, 1, 0, 2, 0, 2, 4, 1, 1, 1, 2, 0, 1, 1,
           1, 1, 0, 2, 3, 2, 0, 2, 1, 3, 1, 3, 2, 2, 3, 2, 0, 0, 0, 1, 0, 0, 0, 1, 2, 0,
           3, 3, 0, 1, 2, 2, 2, 0, 6, 0, 0, 0, 2, 0, 1, 1, 1, 3, 3, 2, 1, 1, 0, 1, 0, 0,
           2, 0, 2, 0, 1, 0, 2, 0, 0, 2, 2, 4, 1, 2, 3, 2, 0, 0, 0, 1, 0, 0, 1, 5, 2, 1,
           3, 2, 0, 2, 1, 1, 3, 0, 5, 0, 0, 2, 4, 3, 4, 0, 0, 0, 0, 0, 0, 2, 2, 0, 0, 2,
           0, 0, 1, 1, 0, 2, 1, 3, 3, 2, 2, 0, 0, 2, 3, 2, 4, 3, 3, 4, 0, 3, 0, 1, 0, 1,
           2, 3, 4, 1, 2, 6, 2, 1, 2, 2)
a <- 2
b <- 1
sum_A <- sum(man_A)
sum_B <- sum(man_B)
n_A <- length(man_A)
n_B <- length(man_B)

set.seed(123)
Y_A <- rnbinom(5000, size = a+sum_A, prob = 1-1/(b+n_A+1))
Y_B <- rnbinom(5000, size = a+sum_B, prob = 1-1/(b+n_B+1))

# plot the MC approximations to these two posterior predictive dist
hist(Y_A, breaks = seq(-0.5, max(Y_A)+0.5,1), main = 'MC Generated Posterior Predictive Y_A')
```



```
hist(Y_B, breaks = seq(-0.5, max(Y_B)+0.5,1), main = 'MC Generated Posterior Predictive Y_B')
```



b. Find 95% quantile-based posterior confidence interval for $\theta_B - \theta_A$ and $Y_B - Y_A$. Describe in words the difference between the two populations using these quantities.

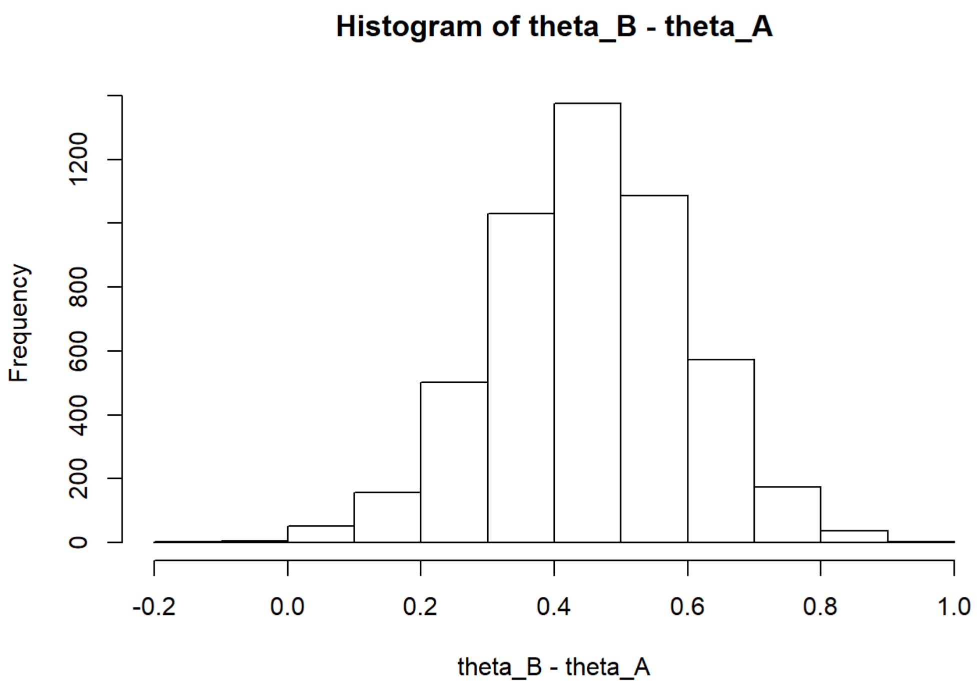
```
set.seed(123)
theta_A <- rgamma(5000,a+sum_A, b+n_A)
theta_B <- rgamma(5000,a+sum_B, b+n_B)

quantile(theta_B - theta_A, c(0.025,0.975))
```

```
##      2.5%     97.5% 
## 0.1559684 0.7359099
```

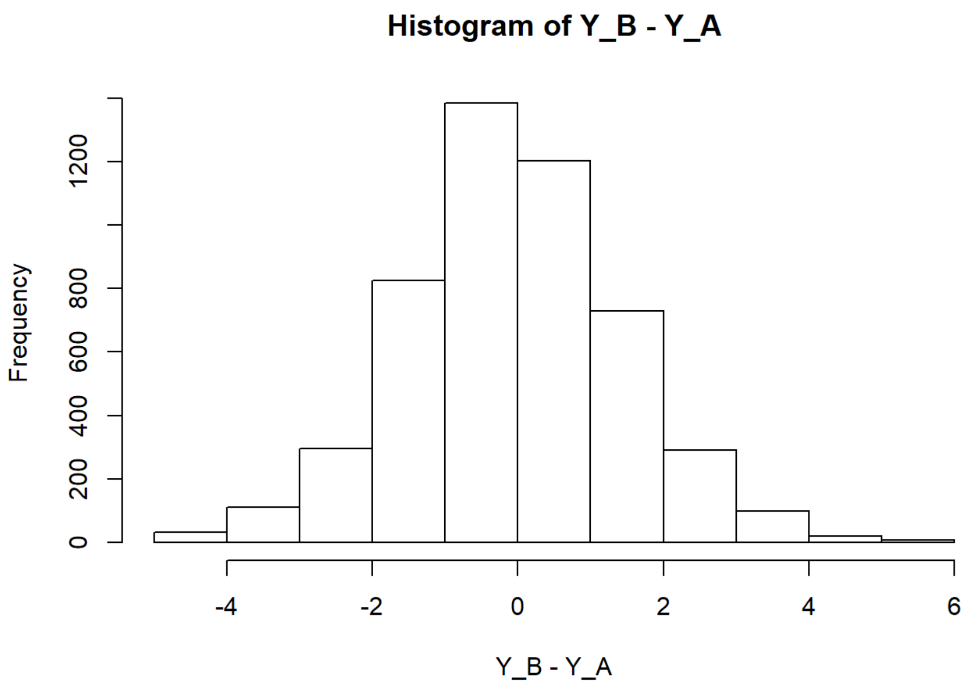
```
quantile(Y_B-Y_A, c(0.025,0.975))
```

```
##  2.5% 97.5%
##    -3     4
```

```
hist(theta_B-theta_A)
```



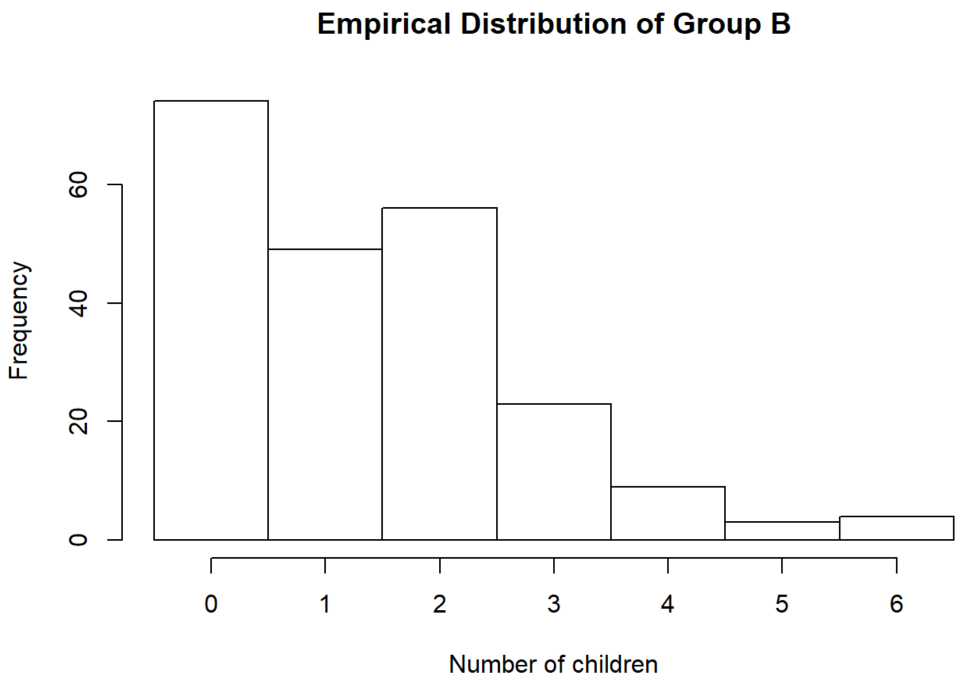
```
hist(Y_B-Y_A)
```



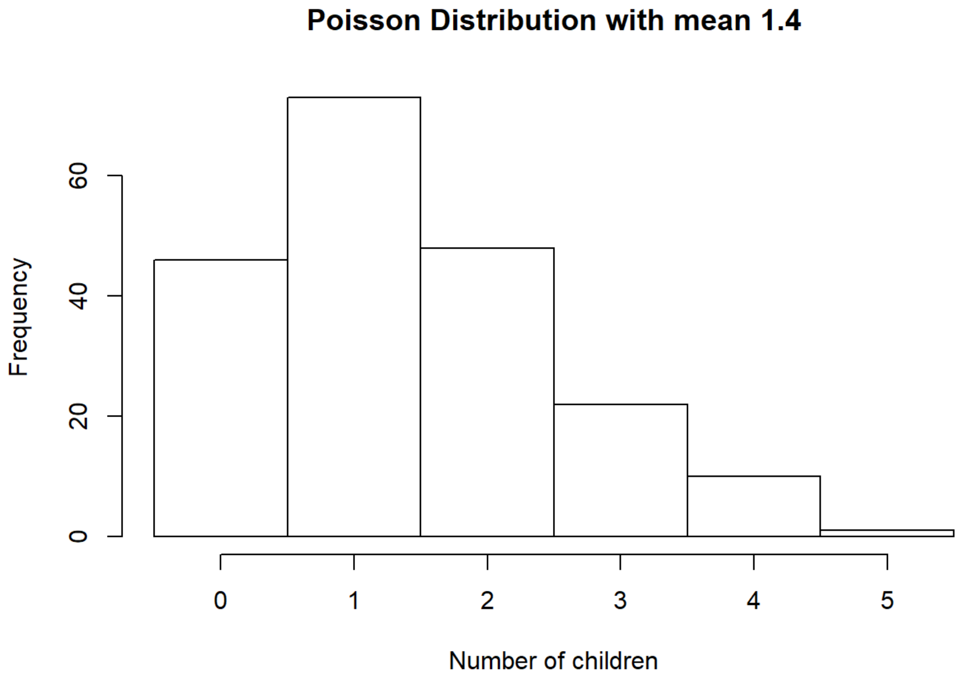
According to the 95% posterior CI and the distribution for $\theta_B - \theta_A$, we find that θ_A is extremely unlikely to be greater than θ_B. The confidence interval for $\theta_B - \theta_A$ doesn't cover zero, and a small portion of it in the histogram is smaller than zero. However, the values of $Y_B - Y_A$ are evenly distributed on either sides of zero in the histogram. The confidence interval contains zero. All above shows that Y_B is more right-shifted than Y_A, but the predictive samples have a greater variance than predictive parameters.

c. Do you think the Poisson model is a good fit?

```
# Empirical distribution
hist(man_B, breaks = seq(-0.5, max(man_B)+0.5,1), xlab = 'Number of children', main = 'Empirical Distribution of Group B')
```



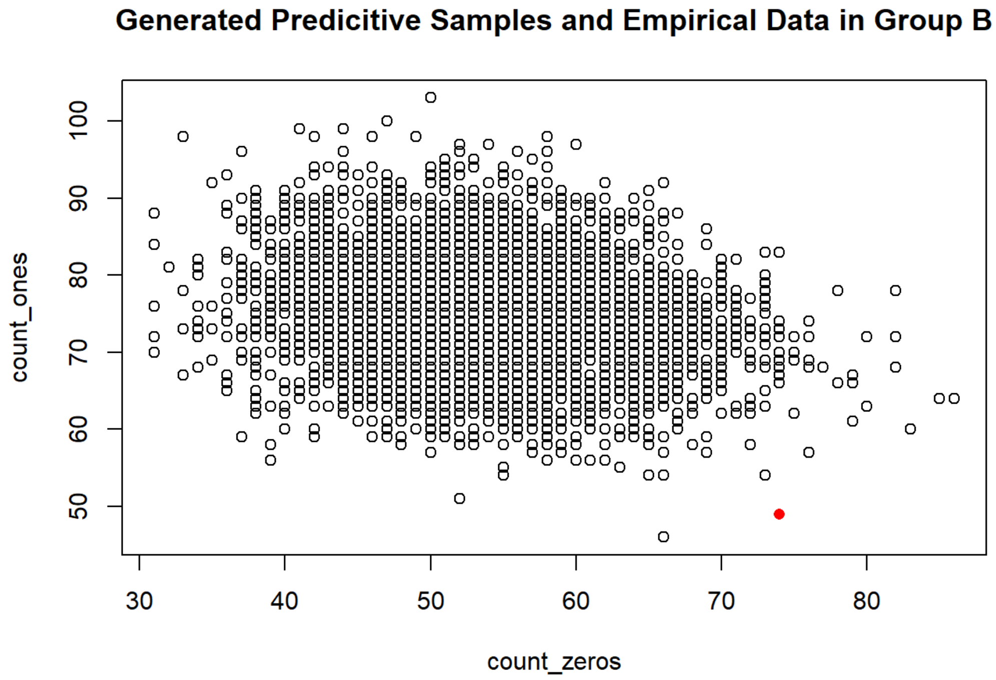
```
# Poisson distribution
set.seed(123)
pos_sam <- rpois(200,1.4)
hist(pos_sam, breaks = seq(-0.5, max(pos_sam)+0.5,1), xlab = 'Number of children', main = 'Poisson Distribution with mean 1.
4')
```



I don’t think it’s a good fit. In the Poisson model with mean 1.4, the frequency of 1’s is almost 1.5 times the frequency of 0s or 2s. While in the empirical distribution for group B, frequency of 0s is 1.3 times the frequency of 1s or 2s. And frequency of 1s is lower than that of 2s. The Poisson model cannot reflect the fact of frequency of 1s is less than both 0s an 2s as well as the frequency of 0s is higher than 1s and 2s.

d. Using this plot, describe the adequacy of the Poisson model

```
count_zeros <- c()
count_ones <- c()
for (theta in theta_B){
  pos_rv <- rpois(218, theta)
  count_zeros <- c(count_zeros, sum(pos_rv==0))
  count_ones <- c(count_ones, sum(pos_rv==1))
}
plot(x=count_zeros, y=count_ones, main="Generated Predicitive Samples and Empirical Data in Group B")
points(x=sum(man_B == 0), y=sum(man_B==1),pch=16, col="red")
```



The Poisson model doesn’t fit data in this statistics. The point from observed data lies outside of the generated samples area. It means that the generated data don’t have the same number of 1s and 0s as in the observed data. This coincides with our conclusion in c.