STA 601 Homework 4

Lingyun Shao

sep. 30, 2018

4.5

Cancer deaths: Suppose for a set of counties $i \in \{1, ..., n\}$ we have information on the population size $X_i = number\ of\ people\ in\ 10,000s$, and $Y_i = number\ of\ cancer\ fatalities$. One model for the distribution of cancer fatalities is that, given the cancer rate θ, they are independently distributed with $Y_i \sim Poisson(\theta X_i)$.

a) Identify the posterior distribution of θ given data $(Y_1, X_1), ..., (Y_n, X_n)$ and a $gamma(a, b)$ prior distribution.

$Y_i's$ are independently distributed given θ. Given the data and prior distribution, we can have the posterior

$$
\begin{aligned}
&p(\theta|(Y_1, X_1), ...(Y_n, X_n)) \\
\propto &p(\theta)\prod_{i=1}^{n} p((Y_i, X_i)|\theta) \\
\propto &\theta^{a-1}e^{-b\theta}\theta^{\sum_{i=1}^n Y_i}e^{-(\sum_{i=1}^n X_i)\theta} \\
= &\theta^{a+\sum_{i=1}^n Y_i-1}e^{-(b+\sum_{i=1}^n X_i)\theta}
\end{aligned}
$$

From the kernal of the posterior, we know that the posterior distribution of θ should be a $gamma(a + \sum_{i=1}^n Y_i, b + \sum_{i=1}^n X_i)$

The file `cancer_react.dat` contains 1990 population sizes (in 10,000s) and number of cancer fatalities for 10 counties in a Midwestern state that are near nuclear reactors. The file `cancer_noreact.dat` contains the same data on counties in the same state that are not near nuclear reactors. Consider these data as samples from two populations of counties: one is the population of counties with no neighboring reactors and a fatality rate of θ_1 deaths per 10,000, and the other is a population of counties having nearby reactors and a fatality rate of θ_2. In this exercise we will model beliefs about the rates as independent and such that $\theta_1 \sim gamma(a_1, b_1)$ and $\theta_2 \sim gamma(a_2, b_2)$.

```
dt1 = read.table(file = 'http://www2.stat.duke.edu/~pdh10/FCBS/Exercises/cancer_noreact.dat',
                 header = TRUE)
dt2 = read.table(file = 'http://www2.stat.duke.edu/~pdh10/FCBS/Exercises/cancer_react.dat',
                 header = TRUE)
```

b) Using the numerical values of the data, identify the posterior distributions for θ_1 and θ_2 for any values of (a_1, b_1, a_2, b_2).

```
(sum1 = colSums(dt1)) # sum of data for theta1
```

```
##    x    y
## 1037 2285
```

```
(sum2 = colSums(dt2)) # sum of data for theta2
```

```
##   x   y
##  95 256
```

As is shown in a), the posterior should be $gamma(a + \sum_{i=1}^{n} Y_i, b + \sum_{i=1}^{n} X_i)$. For θ_1, the sum of $X_i's$ is 1037 and the sum of $Y_i's$ is 2285, so the posterior should be $gamma(a_1 + 2285, b_1 + 1037)$. For θ_2, the sum of $X_i's$ is 95 and the sum of $Y_i's$ is 256, so the posterior should be $gamma(a_2 + 256, b_2 + 95)$.

c) Suppose cancer rates from previous years have been roughly $\tilde{\theta} = 2.2$ per 10,000 (and note that most counties are not near reactors). For each of the following three prior opinions, compute $E[\theta_1|data]$, $E[\theta_2|data]$, 95% quantile-based posterior intervals for θ_1 and θ_2, and $Pr(\theta_2 > \theta_1|data)$. Also plot the posterior densities (try to put $p(\theta_1|data)$ and $p(\theta_2|data)$ on the same plot). Comment on the differences across posterior opinions.

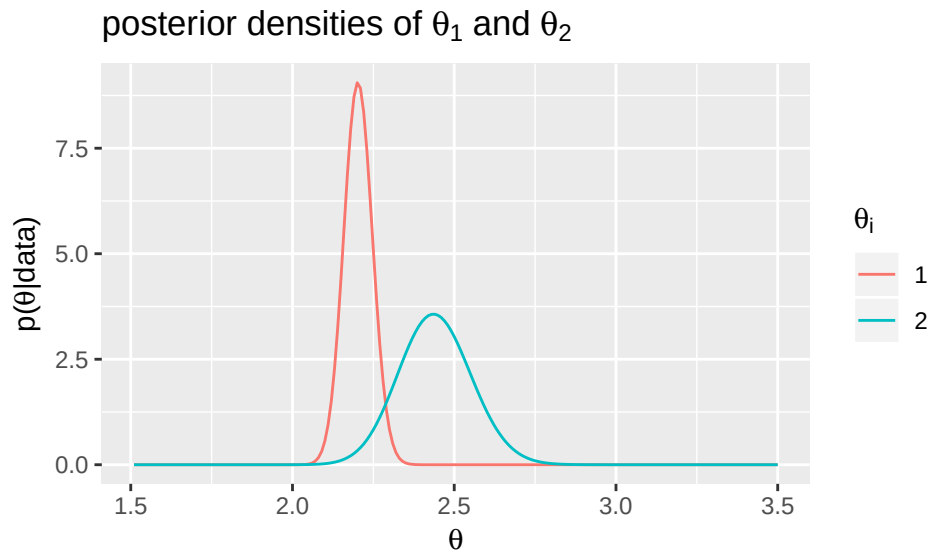
First of all, I defined a function `th_cpr(a1, b1, a2, b2)` to calculate and print out the posterior means, 95% posterior intervals, $Pr(\theta_2 > \theta_1|data)$ and draw the plot of posterior densities of θ_1 and θ_2

```
th_cpr = function(a1, b1, a2, b2) {
  a1.pos = a1 + sum1[2]
  b1.pos = b1 + sum1[1]
  a2.pos = a2 + sum2[2]
  b2.pos = b2 + sum2[1]
  th1.mean = a1.pos/b1.pos # posterior mean
  th1.ci_l = qgamma(0.025, a1.pos, b1.pos) # lower bound of CI
  th1.ci_u = qgamma(0.975, a1.pos, b1.pos) # upper bound of CI
  th2.mean = a2.pos/b2.pos #posterior mean
  th2.ci_l = qgamma(0.025, a2.pos, b2.pos) # lower bound of CI
  th2.ci_u = qgamma(0.975, a2.pos, b2.pos) # upper bound of CI
  set.seed(1)
  pr = mean(rgamma(10000, a2.pos, b2.pos) > rgamma(10000, a1.pos, b1.pos))
  S = cbind(c(th1.mean, th2.mean),
            c(th1.ci_l, th2.ci_l),
            c(th1.ci_u, th2.ci_u),
            c(pr, pr))
  colnames(S) = c('posterior mean', 'CI.lower', 'CI.upper', 'Pr(theta2>theta1|data)')
  rownames(S) = c('theta1', 'theta2')


  x = (151:350)/100
  df1 = data.frame(x = c(x, x), y = c(dgamma(x, a1.pos, b1.pos),
                                      dgamma(x, a2.pos, b2.pos)),
                   pop = rep(c(1,2), each = 200))
  p = ggplot(df1) +
    geom_line(aes(x, y, group = pop, col = factor(pop))) +
    labs(x = expression(theta), y = expression(paste('p(', theta, '|data)', sep = '')),
         title = expression(paste('posterior densities of ', theta[1],
                                  ' and ', theta[2], sep = '')),
         color = expression(theta[i]))
  print(p)
  print(S)
}
```

i. Opinion 1: $(a_1 = a_2 = 2.2 \times 100, b_1 = b_2 = 100)$. Cancer rates for both types of counties are similar to the average rates across all counties from previous years.

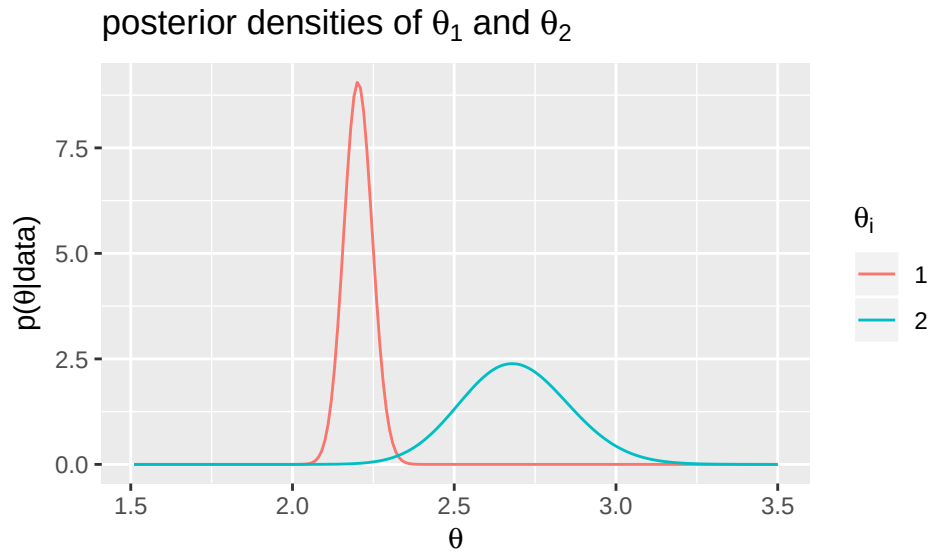
```
th_cpr(a1 = 220, b1 = 100, a2 = 220, b2 = 100)
```




```
##        posterior mean CI.lower CI.upper Pr(theta2>theta1|data)
## theta1       2.203166 2.117726 2.290273                 0.9764
## theta2       2.441026 2.226633 2.665131                 0.9764
```

ii. Opinion 2: $(a_1 = 2.2 \times 100, b_1 = 100, a_2 = 2.2, b_2 = 1)$. Cancer rates in this year for nonreactor counties are similar to rates in previous years in nonreactor counties. We don't have much information on reactor counties, but perhaps the rates are close to those observed previously in nonreactor counties.

```
th_cpr(a1 = 220, b1 = 100, a2 = 2.2, b2 = 1)
```

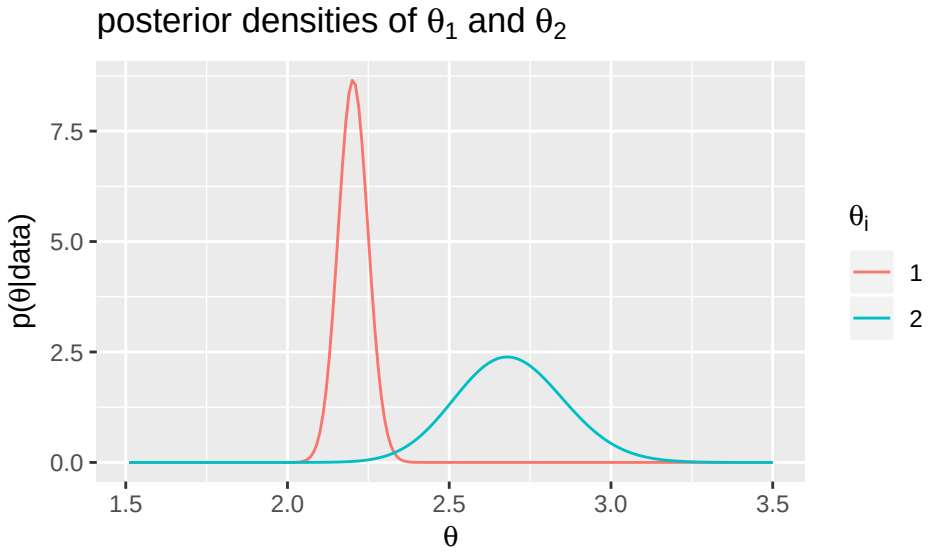



```
##        posterior mean CI.lower CI.upper Pr(theta2>theta1|data)
```

```
## theta1       2.203166 2.117726 2.290273                  0.9981
## theta2       2.689583 2.371497 3.027397                  0.9981
```

iii. Opinion 3: $(a_1 = a_2 = 2.2, b_1 = b_2 = 1)$. Cancer rates in this year could be different from rates in previous years, for both reactor and nonreactor counties.

```
th_cpr(a1 = 2.2, b1 = 1, a2 = 2.2, b2 = 1)
```




```
##        posterior mean CI.lower CI.upper Pr(theta2>theta1|data)
## theta1       2.203468 2.114081 2.294680                  0.9981
## theta2       2.689583 2.371497 3.027397                  0.9981
```

Comment:

For opinion 1, its result is very different from the other 2 for:

1. The sample size is small. Thus the prior here has very large influence on our posterior.
2. The sample mean (2.69) and prior mean (2.2) of θ_2 is very different. If we change our beliefs about θ_2, then there will be huge difference in our posterior inference results.

While for opinion 2 and opinion 3, we can find that there is not much difference between the posteriors, which is basically due to:

1. the huge sample size of data of θ_1. For θ_1, the effect of samples take dominance and thus whether we have large degree of belief ($b_2 = 100$) in our prior or not ($b_2 = 1$) has very limited impact on the posterior results.
2. Another reason might be that the sample mean (2.203) and prior mean (2.2) of θ_1 are very close, so changing our degree of belief in prior makes little difference.

d) In the above analysis we assumed that population size gives no information about fatality rate. Is this reasonable? How would the analysis have to change if this is not reasonable?

I think it might be reasonable to exclude the impact of population size on fatality rate since we already include the effect of population in the sampling model where $Y_i \sim Poisson(\theta X_i)$. Besides, I did search some

reports on cancer fatality rate and population, but there is no obvious relationship between them. Thus, I think the fatality rate is more reasonable to be the same across the countries with different populations in our prior beliefs.

However, if it is not reasonable, then we need to regard the fatality rate θ as from different prior distributions with different parameters like $\theta_{small\ pop}, \theta_{mid\ pop}, \theta_{large\ pop}$ and do bayesian inference for all of them respectively.

e) We encoded our beliefs about θ_1 and θ_2 such that they gave no information about each other (they were a priori independent). Think about why and how you might encode beliefs such that they were a priori dependent.

I think the reason why prior beliefs might not be independent is that other than **whether near reactor**, the cancer fatality rate may also be subject to some globally common factors like **ultraviolet radiation** or **atmospheric composition**. So the cancer fatality rates of countries having nearby reactors and those with no neighboring reactors should share some common baseline, and the effect of having reactors can be interpreted as having extra increases upon the baseline, but not completely independent.

So I think one way to describe the dependence here can be to treat one of prior θ as the location shift of the other. The other possible way can be using a joint distribution $p(\theta_1, \theta_2)$ to describe our prior beliefs.

4.7

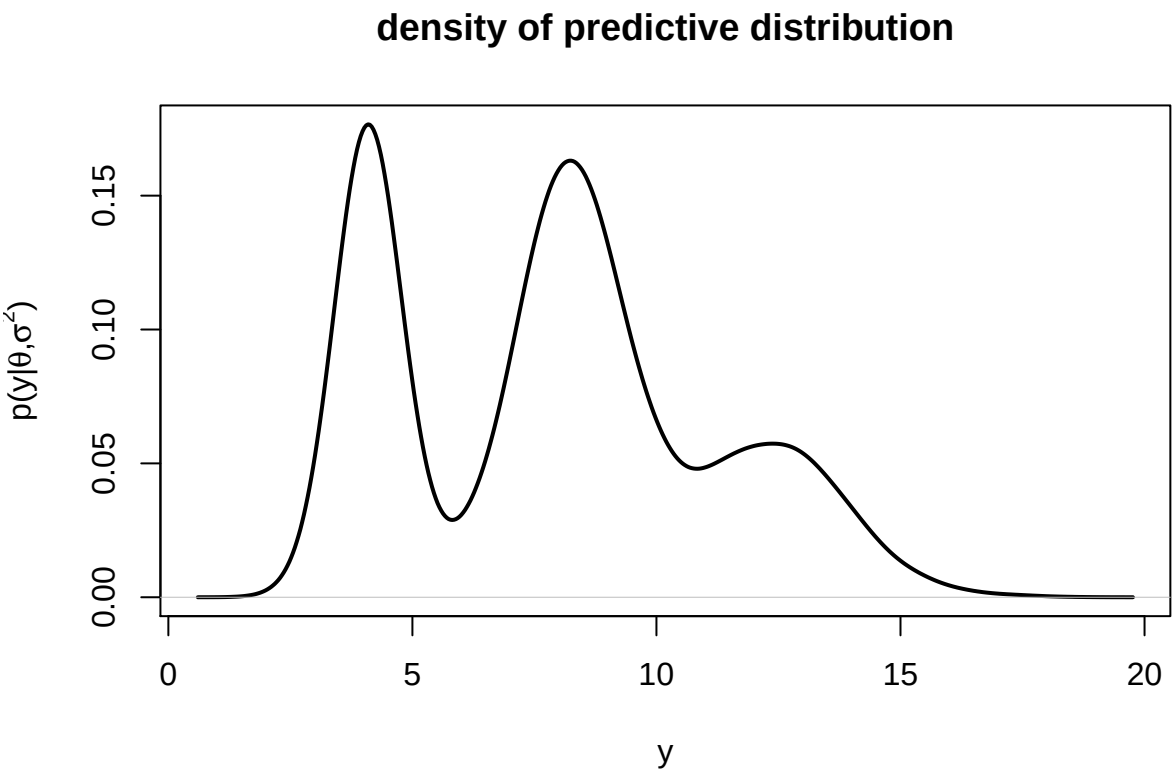
Mixture models: After a posterior analysis on data from a population of squash plants, it was determined that the total vegetable weight of a given plant could be modeled with the following distribution:

$$p(y|\theta, \sigma^2) = .31dnorm(y, \theta, \sigma) + .46dnorm(y, 2\theta_1, 2\sigma) + .23dnorm(y, 3\theta_1, 3\sigma)$$

where the posterior distributions of the parameters have been calculated as $1/\sigma^2 \sim gamma(10, 2.5)$, and $\theta|\sigma^2 \sim normal(4.1, \sigma^2/20)$.

a) Sample at least 5,000 y values from the posterior predictive distribution.

```
n = 10000
set.seed(1)
inv_sig2 = rgamma(n, 10, 2.5)
sig = sqrt(1/inv_sig2)
th = rnorm(n, 4.1, sqrt(1/inv_sig2/20) )
r = runif(n, 0, 1)
y = (r<0.31) * rnorm(n, th, sig) +
    (r>=0.31 & r<0.77) * rnorm(n, 2*th, 2*sig) +
    (r>=0.77) * rnorm(n, 3*th, 3*sig)
plot(density(y), lwd = 2, xlab = 'y',
     ylab = expression(paste('p(y|', theta, ',', sigma^2, ')', sep='')),
     main = 'density of predictive distribution')
```



b) Form a 75% quantile-based confidence interval for a new value of Y .

```
(CI.q = quantile(y, c(0.125, 0.875)))
```

```
##     12.5%     87.5%
##  3.965226 12.181244
```

c) Form a 75% HPD region for a new Y as follows:

i. Compute estimates of the posterior density of Y using the density command in R, and then normalize the density values so they sum to 1.
ii. Sort these discrete probabilities in decreasing order.
iii. Find the first probability value such that the cumulative sum of the sorted values exceeds 0.75. Your HPD region includes all values of y which have a discretized probability greater than this cutoff. Describe your HPD region, and compare it to your quantile-based region.

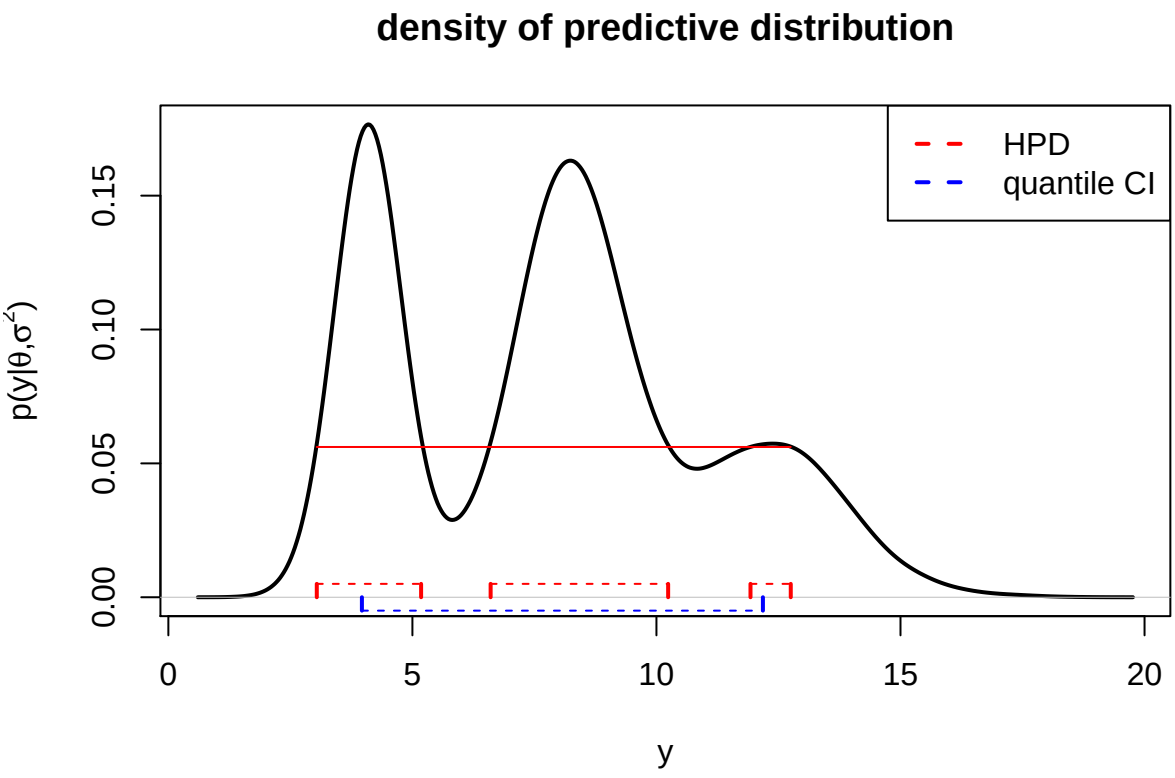
```
y.d = (density(y)$y)/sum(density(y)$y)
x.d = density(y)$x
y.dens = y.d %>% sort(decreasing = TRUE)
c_value = (cumsum(y.dens) > 0.75) %>% which.max() %>% y.dens[.] # critical value
x.HPD = (y.d > c_value) * x.d
R=NULL
r=NULL
x.old = 0
for(x in x.HPD) {
```

```
  if(x > 0){
    r = c(r, x)
  } else if (x.old > 0){
    R = c(R, list(r))
    r = NULL
  }
  x.old = x
}
HPD = sapply(R, function(x) c(x[1], x[length(x)])) %>% t()
rownames(HPD) = c('Int 1', 'Int 2', 'Int 3')
colnames(HPD) = c('lower', 'upper')
HPD
```

```
##           lower     upper
## Int 1  3.041085  5.178148
## Int 2  6.602856 10.239611
## Int 3 11.926766 12.751597
```

```
plot(density(y), lwd = 2, xlab = 'y',
     ylab = expression(paste('p(y|', theta, ',', sigma^2, ')', sep='')),
     main = 'density of predictive distribution')
for(i in 1:nrow(HPD)) {
  arrows(HPD[i,1],0,HPD[i,1],0.005,length = 0,col=2, lwd=2)
  arrows(HPD[i,2],0,HPD[i,2],0.005,length = 0,col=2, lwd=2)
  segments(HPD[i,1], 0.005, HPD[i,2], 0.005, col = 2, lty =2)
}
arrows(CI.q[1],-0.005,CI.q[1],0,length = 0,col=4, lwd=2)
arrows(CI.q[2],-0.005,CI.q[2],0,length = 0,col=4, lwd=2)
segments(CI.q[1], -0.005, CI.q[2], -0.005, col = 4, lty = 2)
segments(HPD[1,1], c_value * sum(density(y)$y),
         HPD[3,2], c_value * sum(density(y)$y), col = 2)

legend('topright',legend = c('HPD','quantile CI'), col = c(2,4), lty = 2, lwd = 2)
```



As is shown in the result, my HPD region is composed of three intervals, which are (3.04,5.18), (6.60,10.24) and (11.93, 12.75). Compared to the quantile-based region (4.00,12.18), we can see from the figure that it's not one consecutive interval, but three intervals. Besides, it consists of the regions with highest probability densities.

d) Can you think of a physical justification for the mixture sampling distribution of Y ?

One possibility is that the squash plants that we are doing analyses on have three subspecies, with each one following a respective normal distribution. Thus, using a mixture sampling model is justifiable.

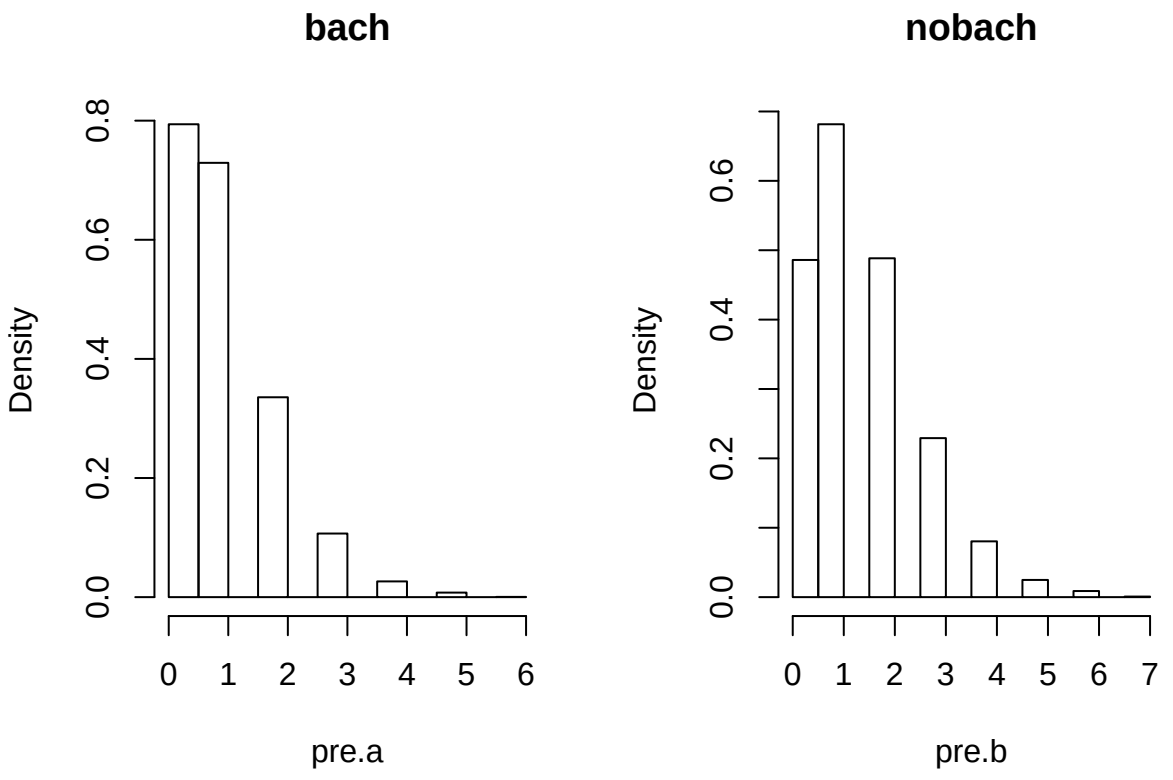
4.8

More posterior predictive checks: Let θ_A and θ_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.

Based on our knowledge of conjugate priors, we can easily find that the posterior $\theta|data \sim gamma(2 + \sum_{i=1}^{n} y_i, 1 + n)$

```
y.a = scan(file =
           'http://www2.stat.duke.edu/~pdh10/FCBS/Exercises/menchild30bach.dat')
y.b = scan(file =
           'http://www2.stat.duke.edu/~pdh10/FCBS/Exercises/menchild30nobach.dat')
a0 = 2
b0 = 1
ya = sum(y.a)
na = length(y.a)
yb = sum(y.b)
nb = length(y.b)
a1.a = a0 + ya
b1.a = b0 + na
a1.b = a0 + yb
b1.b = b0 + nb
n = 5000
set.seed(1)
sam.a = rgamma(n, a1.a, b1.a)
sam.b = rgamma(n, a1.b, b1.b)
pre.a = rpois(n, sam.a)
pre.b = rpois(n, sam.b)
par(mfrow=c(1,2))
hist(pre.a, freq = FALSE, main = 'bach')
hist(pre.b, freq = FALSE, main = 'nobach')
```



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.

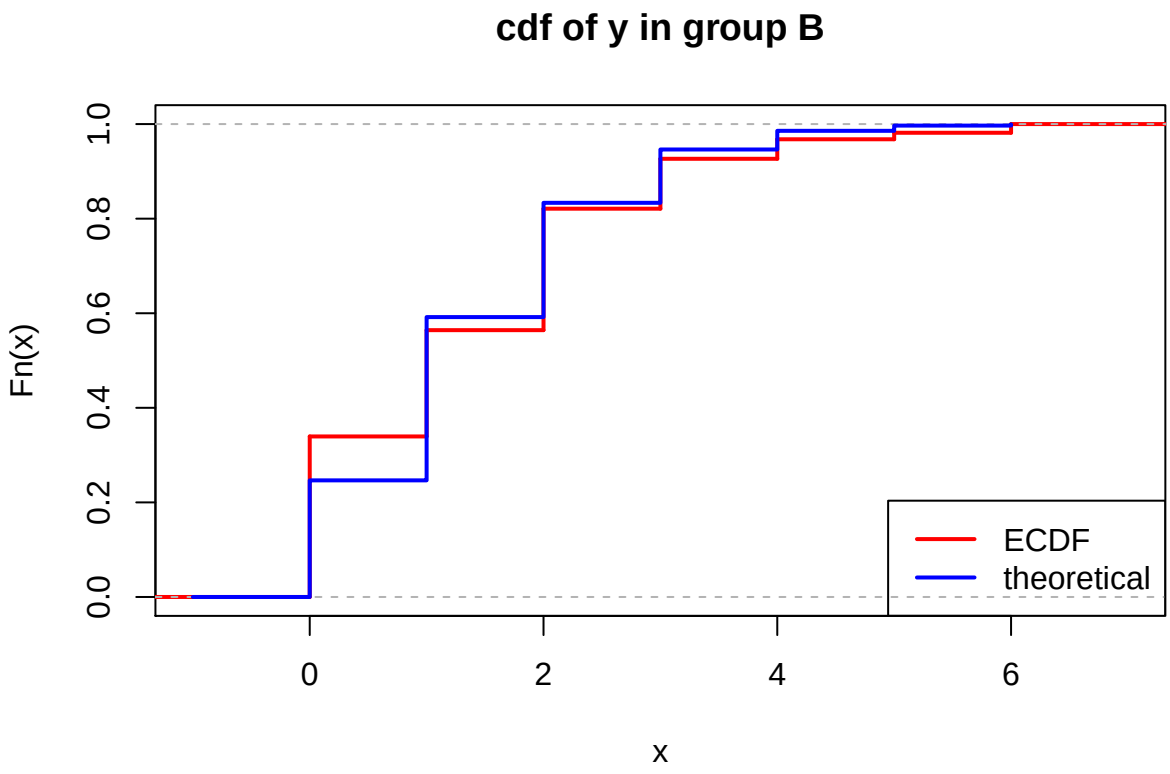
```
CI.theta = quantile(sam.b-sam.a, c(0.025,0.975))
CI.pre = quantile(pre.b-pre.a, c(0.025,0.975), type = 2)
(CI = rbind(CI.theta, CI.pre))
```

```
##                2.5%     97.5%
## CI.theta  0.1454462 0.7337033
## CI.pre   -2.5000000 4.0000000
```

For these two populations, first of all we can find that the confidence interval for $\theta_B - \theta_A$ is all greater than 0, which means that we are more than 95% sure that men in their 30's without bachelor's degree follow a poisson distibution with a larger parameter θ than those with a bachelor's degree. From the confidence interval of posterior predictive values, we know that those men without bachelor's degree tend to have a larger predictive value of number of children. Besides, the plots in a) also suggests those with a Bachelor's degree tend to have more children.

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?

```
plot(ecdf(y.b), verticals = TRUE, do.points = FALSE, col = 2, lwd = 2,
     main = 'cdf of y in group B')
lines(-1:6 ,ppois(-1:6, 1.4), type = 's', col = 4, lwd = 2)
legend('bottomright', legend = c('ECDF', 'theoretical'), col = c(2,4), lwd = 2)
```



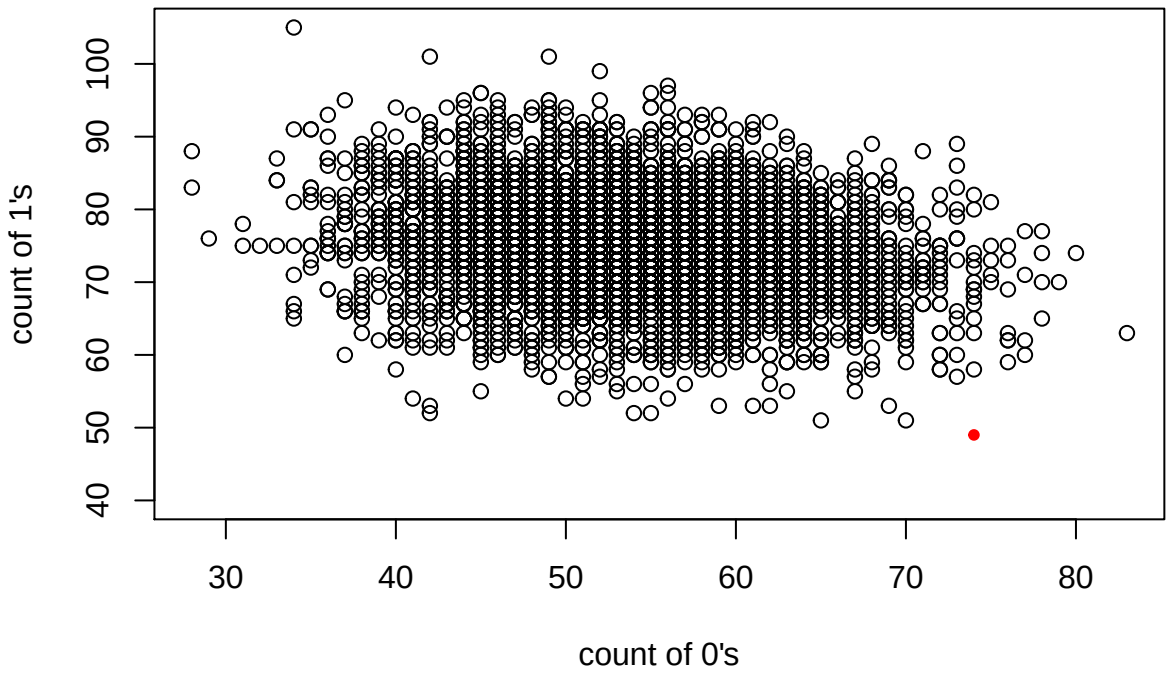
I think the Poisson model with mean $\hat{\theta} = 1.4$ is not a good fit.

The plot shows that at large values of x, the ecdf of data in group B are close to the theoretical cdf of a poisson distribution with a mean 1.4. But we can see a huge deviance in at the value of 0 and 1. In fact, for the theoretical cdf, the probability at 0 is less than at 1 while for the edcf, it's the other way around.

d) For each of the 5,000 θ_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.

```
n.b = 218
set.seed(1)
count01 = function(x) {
  r = rpois(n.b, x)
  count = c(sum(r==0),  sum(r==1))
}

count = sapply(sam.b, FUN=function(x) count01(x)) %>% t()
plot(count[,1], count[,2], ylim = c(40,105),
     xlab = 'count of 0\'s', ylab = 'count of 1\'s')
points(sum(y.b==0), sum(y.b==1), col = 2, pch = 20)
```



From the plot, we can find that the count of 0 and 1 in the observed data in group B deviates dramatically

from the simulated points with a Poisson distribution. Thus, I think It may be not adequate to use Poisson model as the sampling model here.

Math Problem

Let $Y_1, ..., Y_n|p \sim Geometric(p)$, which means that $Pr(Y = k) = (1-p)^k p$. Find the conjugate prior for p. Find the parameters of the posterior distribution. Give an interpretation of the parameters in the prior.

Let $\pi(p)$ be the prior of p and $\pi(p|y)$ be the posterior of p, we have

$$\begin{aligned}\pi(p|y) \propto & \pi(p) \prod_{i=1}^{n}((1-p)^{Y_i} p) \\ \propto & \pi(p)(1-p)^{\sum_{i=1}^{n} Y_i} p^n\end{aligned}$$

Judging from the kernel, we know that the conjugate prior should have the term like $p^{c_1}(1-p)^{c_2}$. This is the kernal of a beta distribution, so our conjugate prior for p should be a beta distribution. Suppose our prior is beta(a,b), then $\pi(p|y) \propto p^{a+n-1}(1-p)^{b+\sum_{i=1}^{n} Y_i - 1}$. Thus, the posterior $p|y \sim beta(a+n, b+\sum_{i=1}^{n})$, the shape parameter is $a+n$, the rate parameter is $b+\sum_{i=1}^{n}$.

The prior parameters can be interpreted as:

Suppose we had a groups of prior experiments, for each group of prior experiment, we kept failing and did the experiment over and over again until we got a success. Then b can be interpreted as the total sum of number of failures we had before a success in every group of our prior experiments.