Homework 03

Joseph Blubaugh
jblubau1@tamu.edu

3.14

a)

$$
\begin{aligned}
p(y|\theta) =& \theta^y(1-\theta)^{(n-y)} \\
log[p(\theta|y)] = \sum log[p(y|\theta)] =& ylog[\theta] + (n-y)log[1-\theta] \\
\frac{dlog[p(\theta|y)]}{d\theta} =& \frac{y}{\theta} + \frac{n-y}{1-\theta} \\
\texttt{MLE estimate}: 0 =& \frac{y}{\theta} + \frac{n-y}{1-\theta} \\
\frac{n}{y} =& \frac{1}{\theta} \\
\theta =& \frac{y}{n} \\
\frac{d^2log[p(\theta|y)]}{d\theta^2} =& \frac{y}{\theta^2} + \frac{(n-y)}{(1-\theta)^2} \\
J(\theta) =& -\frac{y}{\theta^2} - \frac{(n-y)}{(1-\theta)^2}
\end{aligned}
$$

b)

$$
\begin{aligned}
p_u(\theta) =& \theta^y(1-\theta)^{(n-y)}/exp[n] \\
-\frac{d^2P_u(\theta)}{d\theta^2} =& -\frac{y}{n\theta^2} - \frac{n-y}{n(1-\theta)^2}
\end{aligned}
$$

c)

$$
p_u(\theta|y) = \theta^y(1-\theta)^{n-y}[-\frac{y}{n\theta^2} - \frac{n-y}{n(1-\theta)^2}]
$$

No because it no longer retains the form of the binomial density function

d)

$$
\begin{aligned}
p(y|\theta) =& \Pi_{i=1}^{n} \frac{e^{-\theta}\theta^{y}}{y!} \\
=& \frac{e^{-n\theta}\theta^{y}}{y!} \\
lp(\theta|y) =& ylog[\theta] - n\theta \\
\frac{lp(\theta|y)}{d\theta} =& \frac{y}{\theta} - n \\
\text{MLE} : 0 =& \frac{y}{\theta} - n \rightarrow \theta = \frac{y}{n} \\
-\frac{d^2 lp(\theta|y)}{d\theta^2} =& -\frac{y}{\theta^2}
\end{aligned}
$$

$$
\begin{aligned}
p_u(\theta) =& \frac{e^{-n\theta}\theta^{y}}{y!}/n \\
=& exp[\frac{y}{n}]\theta - \theta \\
\frac{d^2 log[p(\theta|y)]}{d\theta^2} =& \frac{y}{n\theta} - 1
\end{aligned}
$$

$$
p_u(\theta|y) = \frac{e^{-n\theta}\theta^{y-1}y}{y!} - 1
$$

Yes because it still maintains the general form of the poisson density function

4.1

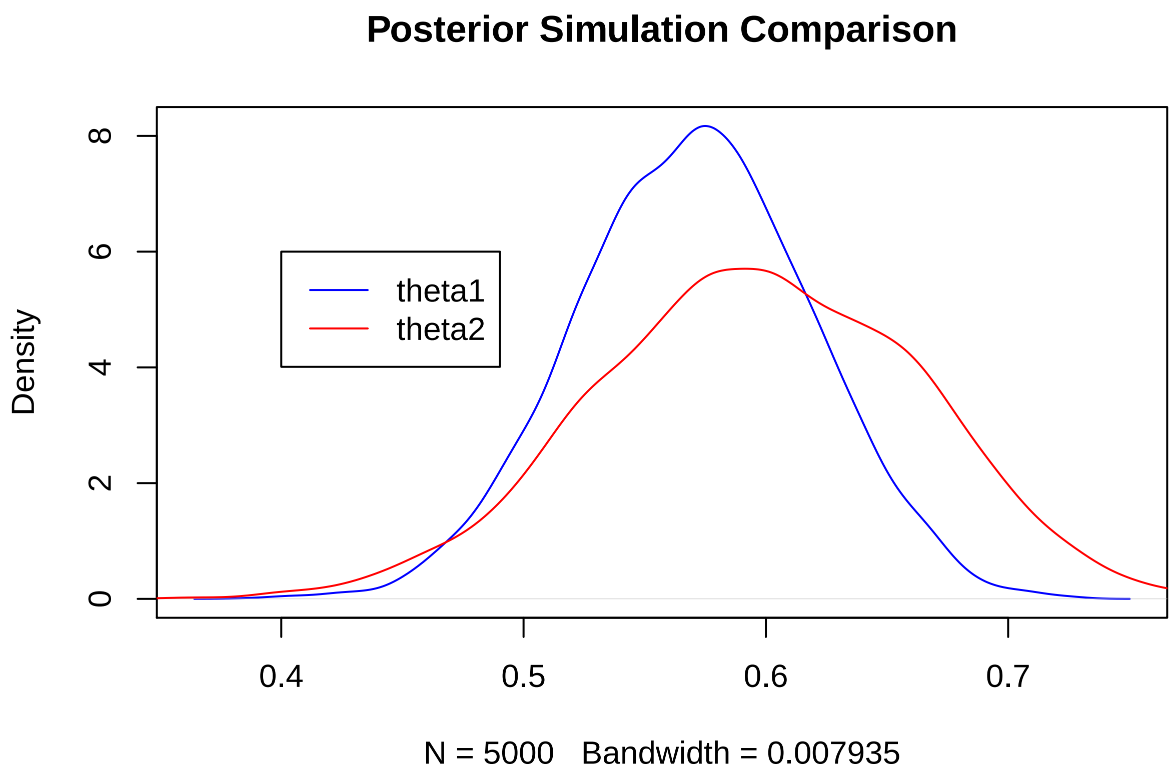
Posterior Distributions: $P(\theta_1|Y_1 = 57) = beta(58, 47)$
$P(\theta_2|Y_2 = 30) = beta(31, 21)$

```
theta.1 = rbeta(n = 5000, shape1 = 58, shape2 = 44)
theta.2 = rbeta(n = 5000, shape1 = 31, shape2 = 21)

## Estimated Probability that theta.1 < theta.2
length(which(theta.1/theta.2 < 1) == TRUE) / 5000
```

```
[1] 0.6214
```

```
plot(density(theta.1), col = "blue", main = "Posterior Simulation Comparison")
lines(density(theta.2), col = "red")
legend(x = .4, y = 6, legend = c("theta1", "theta2"), lty = c(1,1), col = c("blue", "red"))
```



4.2

a)

$P(\theta_a|Y_a) = P(Y_a|\theta)P(\theta) = Poisson(Y_a/N)|\theta)Gamma(120, 10) = Gamma(120 + 237, 10 + 10)$
$P(\theta_b|Y_b) = P(Y_b|\theta)P(\theta) = Poisson(Y_b/N))Gamma(12, 1) = Gamma(12 + 125, 1 + 13)$

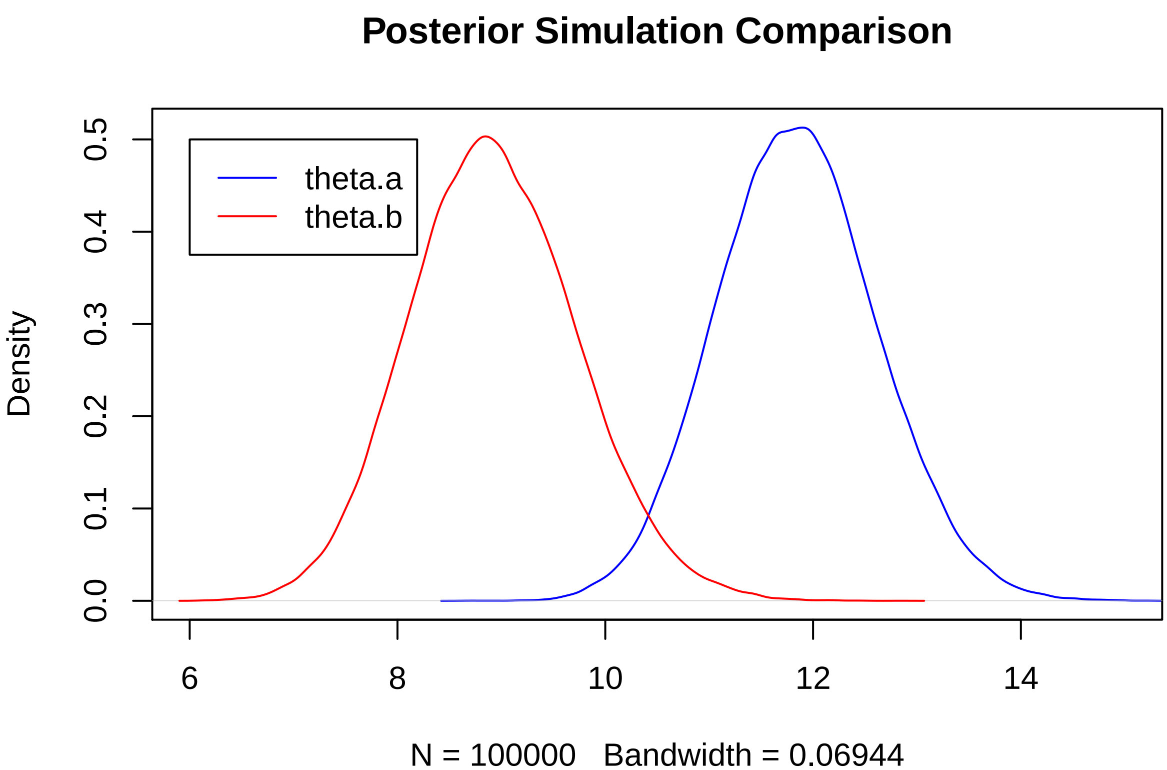
```
Y.a = c(12, 9, 12, 14, 13, 13, 15, 8, 15, 6)
Y.b = c(11, 11, 10, 9, 9, 8, 7, 10, 6, 8, 8, 9, 7)

theta.a = rgamma(n = 100000, shape = 120 + sum(Y.a), rate = 10 + 10)
theta.b = rgamma(n = 100000, shape = 12 + sum(Y.b), rate = 1 + 13)

## Probability that theta.a is > theta.b
mean(theta.a > theta.b)
```

```
[1] 0.99494
```

```
plot(density(theta.a), col = "blue", main = "Posterior Simulation Comparison",
     xlim = c(6, 15))
lines(density(theta.b), col = "red")
legend(x = 6, y = .5, legend = c("theta.a", "theta.b"), lty = c(1,1), col = c("blue", "red"))
```



b)

```r
results = data.frame()

n = 20; sims = 100000

for (i in 1:n) {
  theta.a = rgamma(n = sims, shape = 120, rate = 10)
  theta.b = rgamma(n = sims, shape = 12 * i, rate = i)

  x = mean(theta.a < theta.b)

  results = rbind(results, data.frame(N = i, Prob = x))

}

## The Probability of N where Theta.b < Theta.a ~ 10+
t(round(results, 4))
```

```
          1      2      3      4      5      6      7      8      9
N    1.0000 2.0000 3.0000 4.0000 5.0000 6.0000 7.0000 8.0000 9.0000
Prob 0.4691 0.4826 0.4841 0.4892 0.4934 0.4939 0.4973 0.4985 0.4991
          10      11      12      13      14      15      16      17
N    10.0000 11.0000 12.0000 13.0000 14.0000 15.0000 16.0000 17.0000
Prob  0.4984  0.5034  0.4977  0.4999  0.5048  0.5046  0.5022  0.5048
          18      19     20
N    18.0000 19.0000 20.000
Prob  0.4988  0.5046  0.503
```

c)

```r
theta.a = rgamma(n = sims, shape = 120, rate = 10 + 10)
theta.b = rgamma(n = sims, shape = 12 + sum(Y.b), rate = 1 + 13)

pois.a = rpois(n = sims, lambda = theta.a)
pois.b = rpois(n = sims, lambda = theta.b)

mean(pois.b < pois.a)
```

```
[1] 0.19368
```

```r
results = data.frame()

sims = 100000

for (i in seq(.0, 2, .05)) {

  theta.a = rgamma(n = sims, shape = 120, rate = 10 + 10)
```

```r
  theta.b = rgamma(n = sims, shape = 12 * i, rate = i)

  pois.a = rpois(n = sims, lambda = theta.a)
  pois.b = rpois(n = sims, lambda = theta.b)

  x = mean(pois.b < pois.a)

  results = rbind(results, data.frame(N = i, Prob = x))

}

## The Probability of N where Theta.b < Theta.a for approximately all N
t(round(results, 4))
```

```
         1      2      3    4      5      6      7      8      9     10
N    0.000 0.0500 0.1000 0.15 0.2000 0.2500 0.3000 0.3500 0.4000 0.4500
Prob 0.997 0.4542 0.3396 0.28 0.2404 0.2123 0.1933 0.1797 0.1665 0.1546
       11     12     13     14     15     16    17     18     19     20
N    0.50 0.5500 0.6000 0.6500 0.7000 0.7500 0.800 0.8500 0.9000 0.9500
Prob 0.15 0.1408 0.1373 0.1319 0.1257 0.1227 0.121 0.1157 0.1134 0.1119
         21    22     23     24     25     26     27     28     29     30
N    1.0000 1.050 1.1000 1.1500 1.2000 1.2500 1.3000 1.3500 1.4000 1.4500
Prob 0.1088 0.106 0.1049 0.1034 0.1007 0.0991 0.0976 0.0983 0.0949 0.0953
         31     32     33     34     35     36     37     38     39     40
N    1.5000 1.5500 1.6000 1.6500 1.7000 1.7500 1.8000 1.8500 1.9000 1.9500
Prob 0.0942 0.0932 0.0911 0.0903 0.0901 0.0892 0.0882 0.0888 0.0851 0.0857
         41
N    2.0000
Prob 0.0858
```

4.8

a)

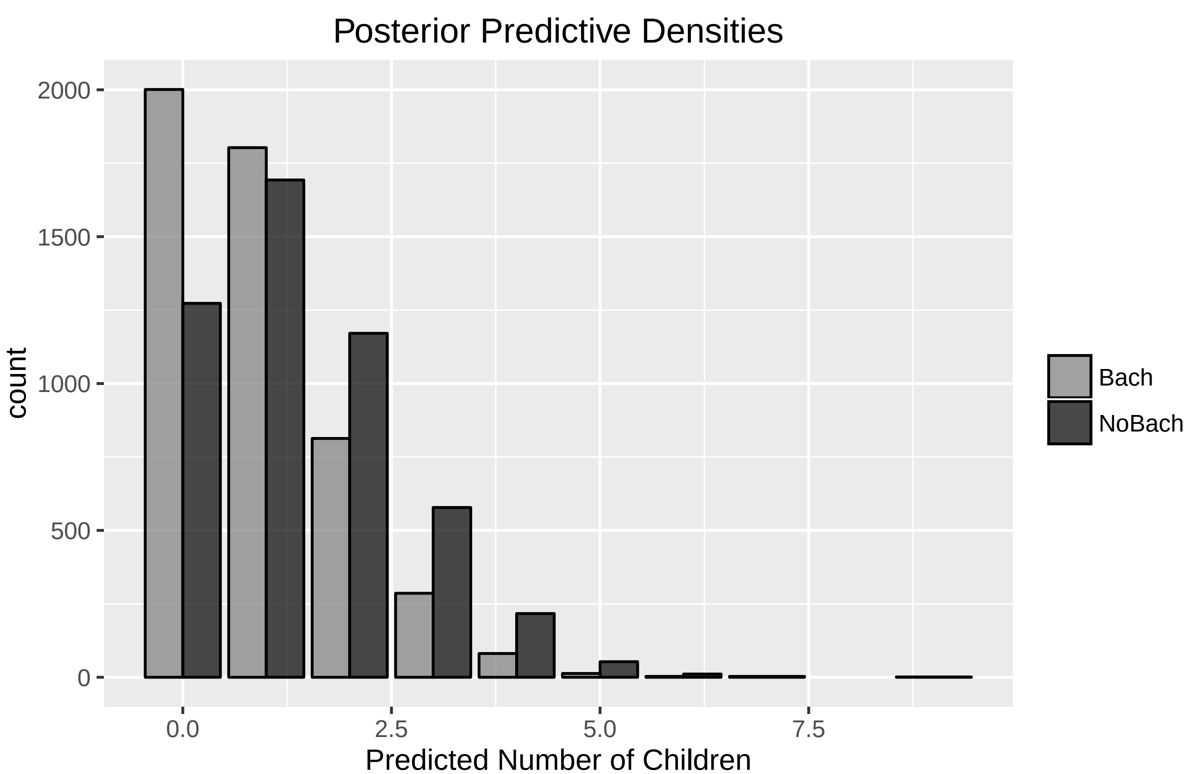
```r
prior.theta.shape = 2; prior.theta.rate = 1

post.theta.a = rgamma(5000, shape = prior.theta.shape + sum(bach),
                      rate = prior.theta.rate + length(bach))
post.theta.b = rgamma(5000, shape = prior.theta.shape + sum(nobach),
                      rate = prior.theta.rate + length(nobach))

pred.theta.a = rpois(n = 5000, lambda = post.theta.a)
pred.theta.b = rpois(n = 5000, lambda = post.theta.b)

dta = data.frame(
  Theta = c(rep("A", 5000), rep("B", 5000)),
  Pred = c(pred.theta.a, pred.theta.b)
)

ggplot(dta) +
  geom_bar(aes(x = Pred, fill = Theta), position = "dodge", alpha = .7, color = "black") +
  scale_x_continuous("Predicted Number of Children") +
  scale_fill_manual("", labels = c("Bach", "NoBach"), values = c("gray50", "black")) +
  ggtitle("Posterior Predictive Densities")
```



b)

Since the 95% Confidence Interval of the difference in the posterior distributions does not include 0 we could conclude that there are significant differences between two distributions. The 95% Confidence intervals of the difference in posterior predictive distributions does however contain 0 so we not be able to say that these two distributions are significantly different.

```r
## Difference in Posterior Theta Estimate
quantile(post.theta.b - post.theta.a, c(.025, .975))
```

```
     2.5%     97.5%
0.1503111 0.7355482
```

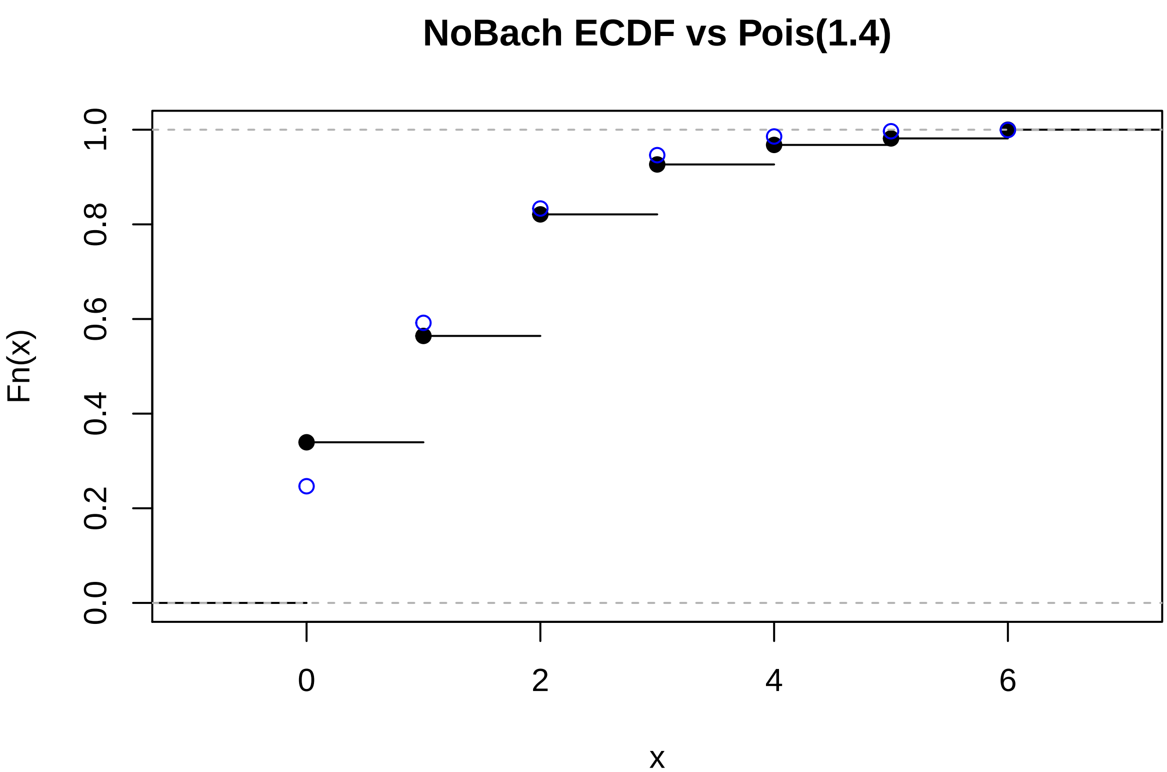
```r
## Difference in Posterior Prediction Estimate
quantile(pred.theta.b - pred.theta.a, c(.025, .975))
```

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

c)

The Poisson(1.4) appears to be a good model for approximating the distribution of the NoBach data set. Except for Y = 0, all of the ECDF points are very close to the points of of the poisson densisty.

```r
plot(ecdf(nobach), main = "NoBach ECDF vs Pois(1.4)")
points(x = 0:6, ppois(q = 0:6, lambda = 1.4), col = "blue")
```



d)

The plot supports the poisson model as an appropriate model for the data. The average count of NoBach that have one child is higher than the those with no children which is consistent with the sampled data and with the distribution function of the Poisson(1.4) model used in the last section. I would expect Ones to be closer than Zero because its most likely closer to the true average.

```
results = data.frame()

for (i in 1:5000) {
  y = pred.theta.b[sample(x = 5000, size = 218)]

  zero = length(which(y == 0))
  ones = length(which(y == 1))

  results = rbind(results, data.frame(zero = zero, ones = ones))
}

ggplot(results) +
  geom_point(aes(x = zero, y = ones)) +
  geom_vline(xintercept = mean(results$zero), lty = 2, color = "blue") +
  geom_hline(yintercept = mean(results$ones), lty = 2, color = "blue") +
  ggtitle("Comparison of Theta_B Posterior Predictive Sampling")
```