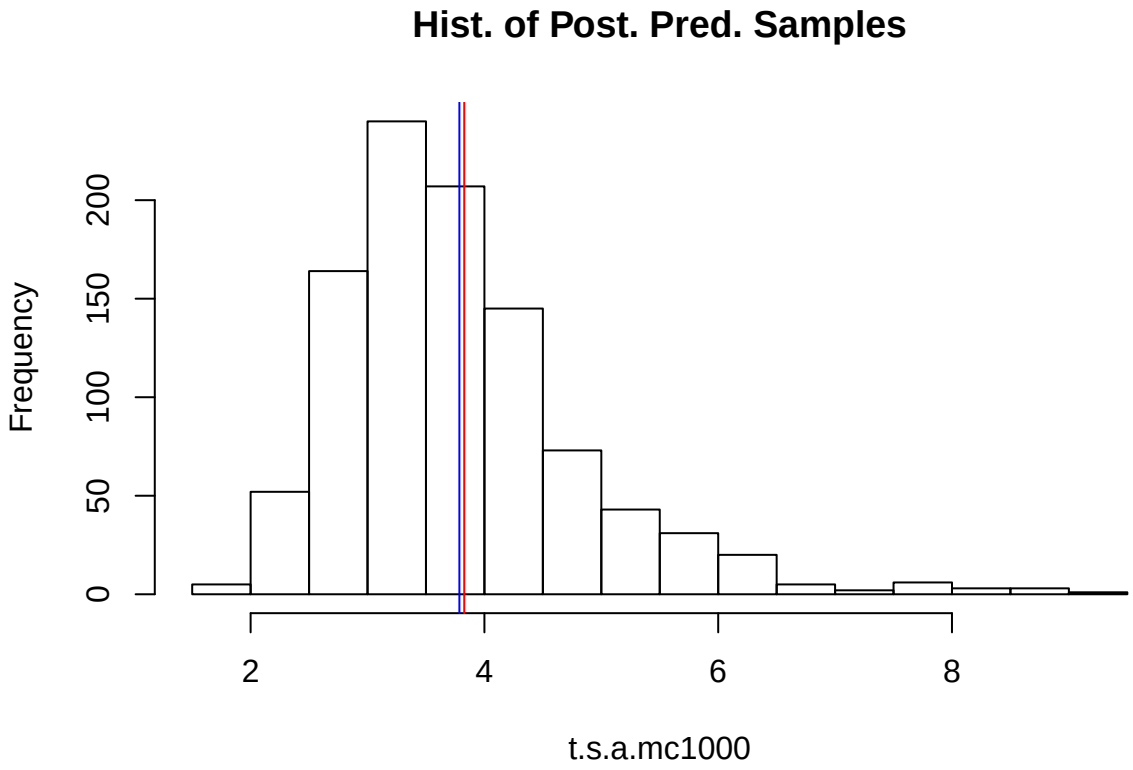
Homework 4

Thomas Fleming- Lab Time: 1:25

September 29, 2016

4.3
a)

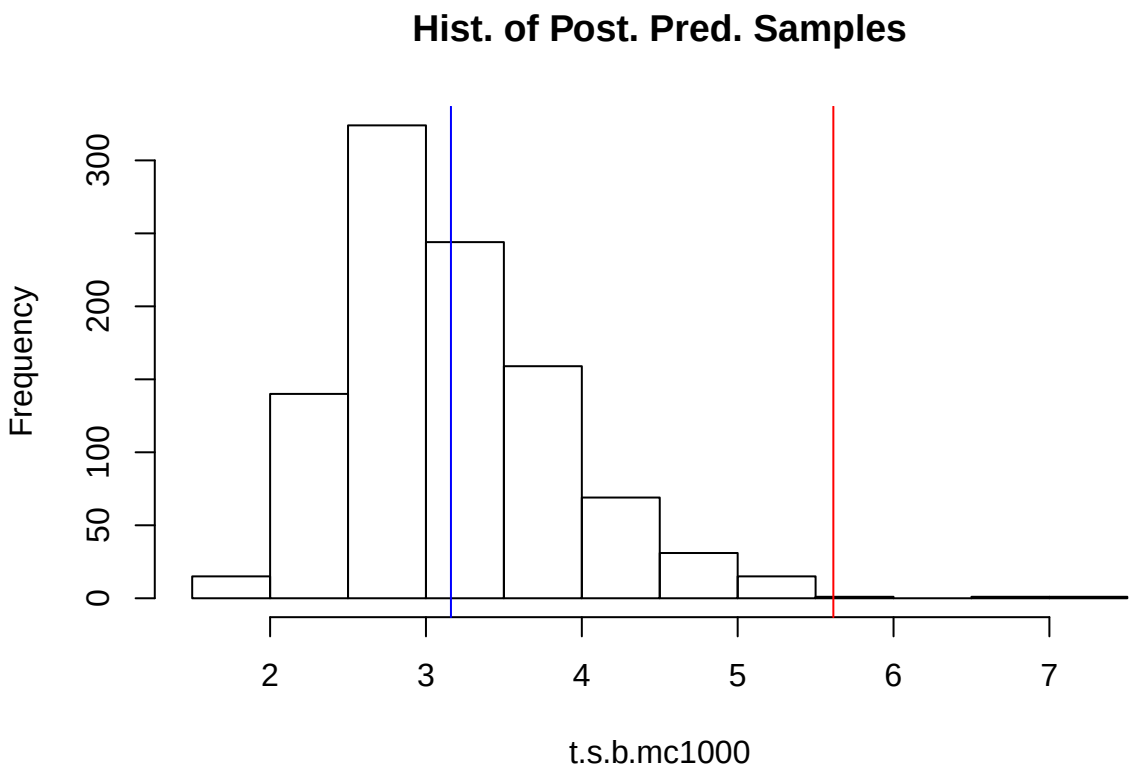
```
set.seed(1)
y.a <- c(12, 9, 12, 14, 13, 13, 15, 8, 15, 6)
y.b <- c(11, 11, 10, 9, 9, 8, 7, 10, 6, 8, 8, 9, 7)
sigma.y.a <- sum(y.a)
sigma.y.b <- sum(y.b)
a.a <- 120
b.a <- 10
a.b <- 12
b.b <- 1
n.a <- 10
n.b <- 13
#Posterior Predictive Check
S <- 1000
stheta.a.mc1000 <- t.s.a.mc1000 <- numeric(S)
for (s in 1:S) {
  stheta.a.mc1000[s] <- rgamma(1, a.a + sigma.y.a, b.a + n.a)
  sy.a <- rpois(10, stheta.a.mc1000[s])
  t.s.a.mc1000[s] <- mean(sy.a)/sd(sy.a, na.rm = FALSE)
}
hist(t.s.a.mc1000, main = "Hist. of Post. Pred. Samples")
abline(v = mean(t.s.a.mc1000), col = "blue")
abline(v = mean(y.a)/sd(y.a), col = "red")
```



From the plot, Poisson model seems to represent the empirical distribution pretty adequately, since the sampled t(s) and the observed t are very close to one another.

b)

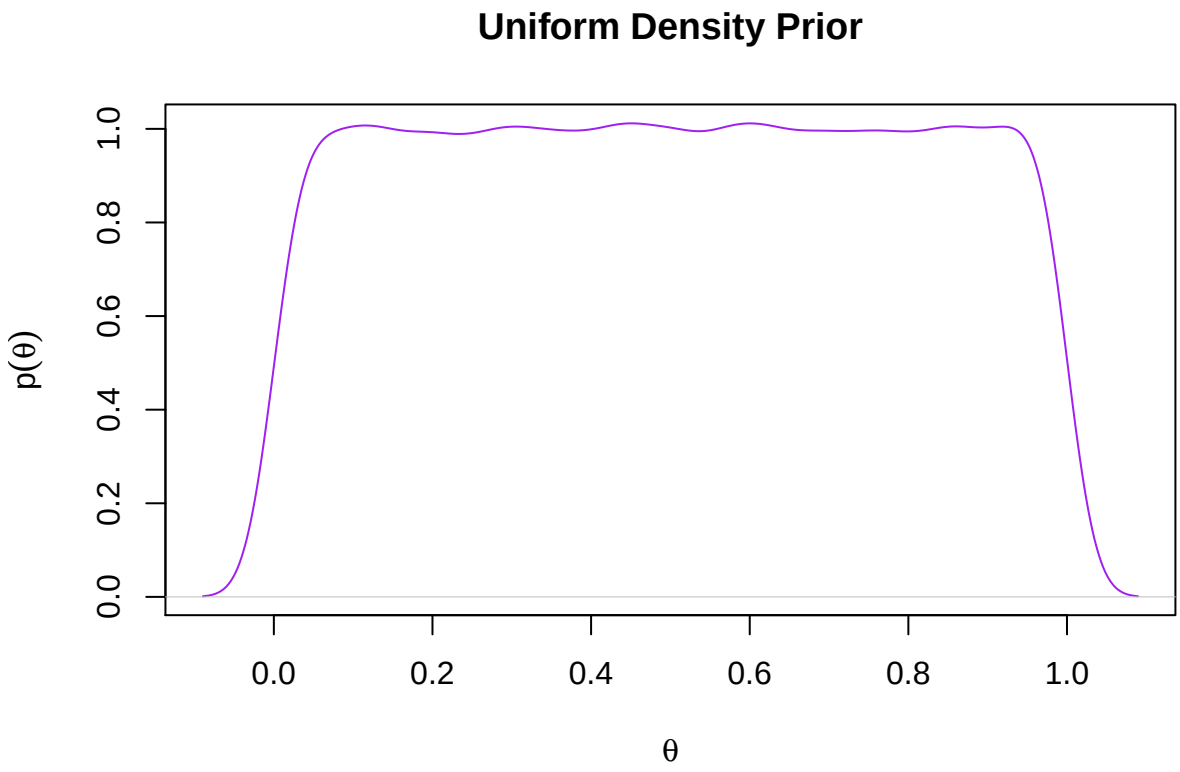
```
stheta.b.mc1000 <- t.s.b.mc1000 <- numeric(S)
for (s in 1:S) {
  stheta.b.mc1000[s] <- rgamma(1, a.b + sigma.y.b, b.b + n.b)
  sy.b <- rpois(13, stheta.b.mc1000[s])
  t.s.b.mc1000[s] <- mean(sy.b)/sd(sy.b, na.rm = FALSE)
}
hist(t.s.b.mc1000, main = "Hist. of Post. Pred. Samples")
abline(v = mean(t.s.b.mc1000), col = "blue")
abline(v = mean(y.b)/sd(y.b), col = "red")
```



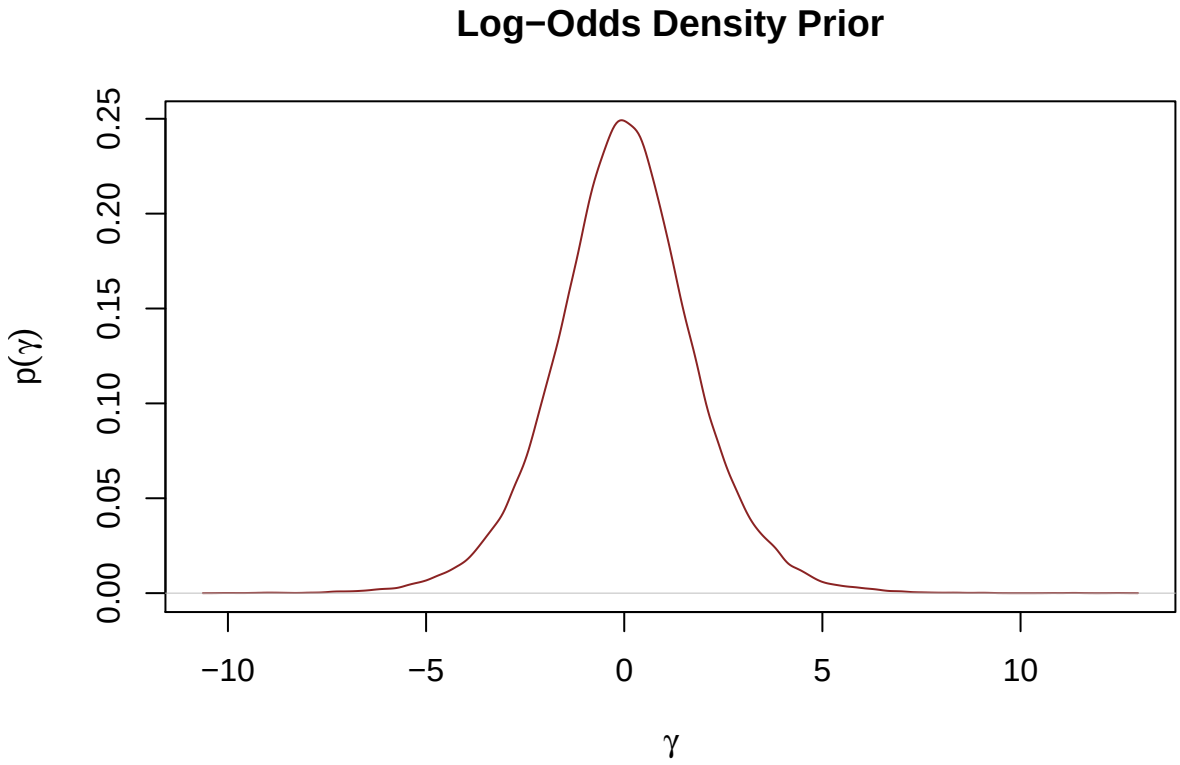
Since the sampled t(s) is a little above 3 and the observed t is a little less than 6, the two means are very far apart. In this case, the Poisson model seems to be less adequate.

4.6

```
a <- 1
b <- 1
theta.prior.mc <- rbeta(50000, a, b)
gamma.prior.mc <- log(theta.prior.mc/(1 - theta.prior.mc))
plot(density(theta.prior.mc), lwd = 1, xlab = expression(theta), ylab = expression(p(theta)), col = "pu
```



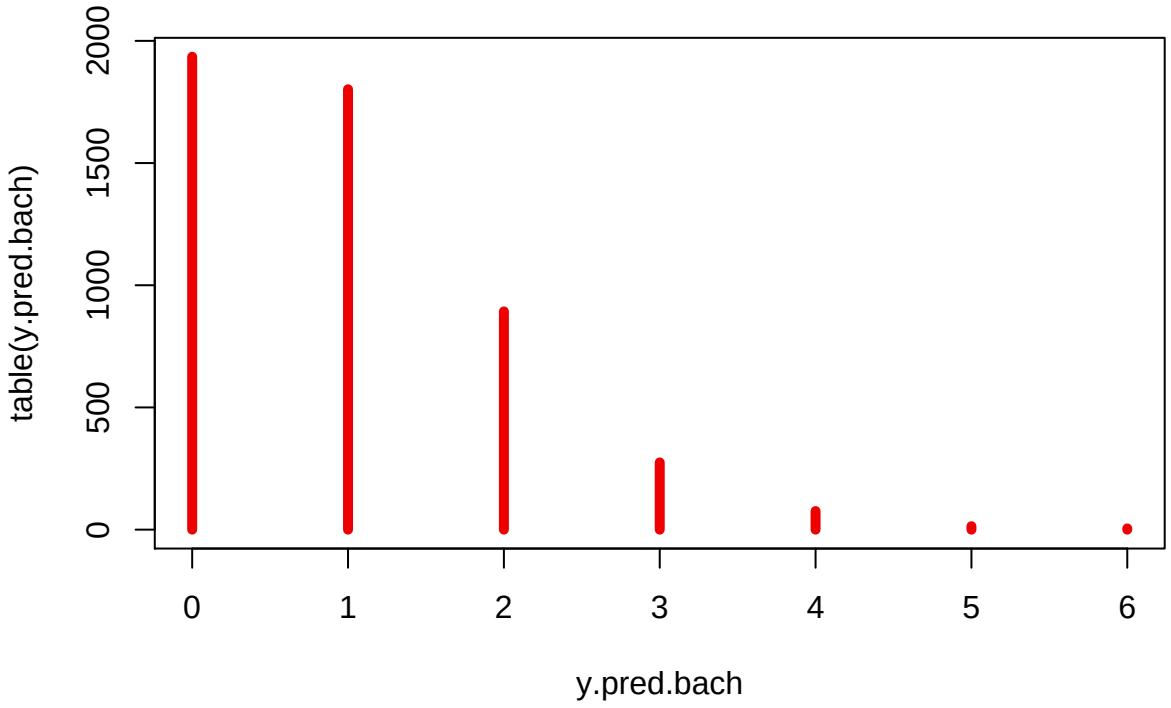

```
plot(density(gamma.prior.mc), lwd = 1, xlab = expression(gamma), ylab = expression(p(gamma)), col = "br
```



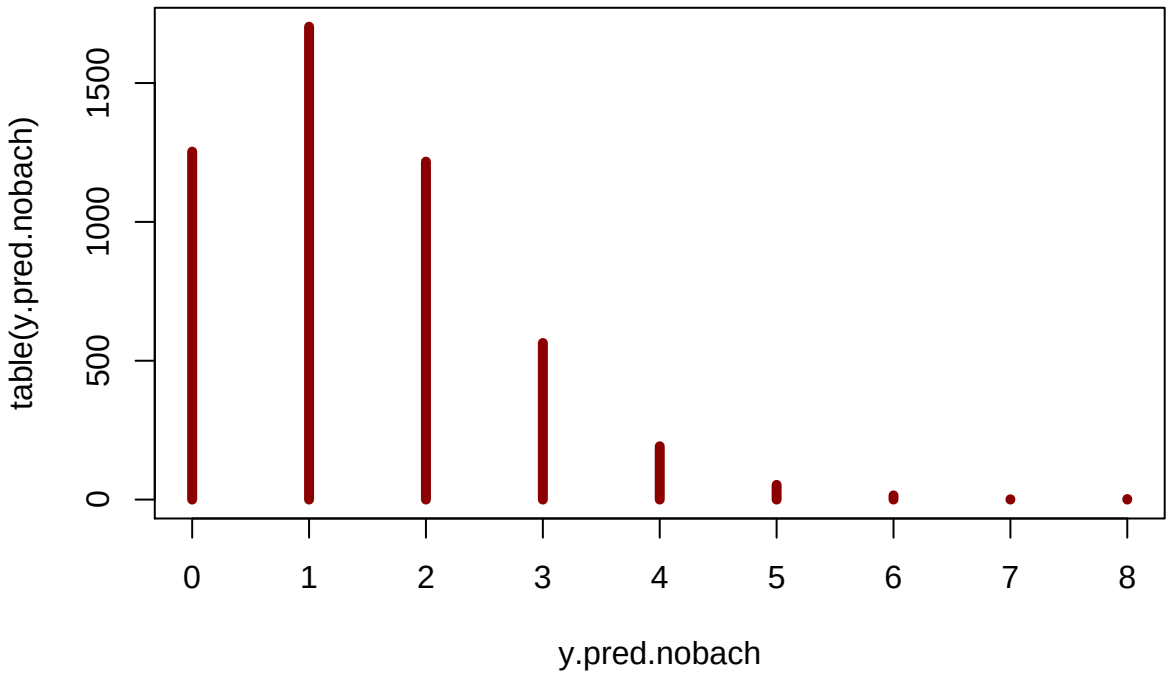
After plotting the distribution on the log-odds scale, it appears that prior is informative if we parameterize it in this way. It is an informative prior for lower-case gamma because the distribution is fairly concentrated from about -5 to 5 and around a particular value (in this case around 0), so that we are more sure of the mean and range of the distribution.

4.8
a)

```
y.bach <- 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
y.nobach <- 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, 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,
 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,
 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,
 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,
 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)
sum.bach <- sum(y.bach)
sum.nobach <- sum(y.nobach)
n.bach <- length(y.bach)
n.nobach <- length(y.nobach)
y.pred.bach <- rnbinom(5000, size = 2 + sum.bach, mu = (2 + sum.bach)/(1 + n.bach))
y.pred.nobach <- rnbinom(5000, size = 2 + sum.nobach, mu = (2 + sum.nobach)/(1 + n.nobach))
plot(table(y.pred.bach), type = "h", lwd = 5, col = "red2")
```




```
plot(table(y.pred.nobach), type = "h", lwd = 5, col = "darkred")
```



b)

```
theta.y.bach.samp <- rgamma(5000, 2 + sum.bach, 1 + n.bach)
theta.y.nobach.samp <- rgamma(5000, 2 + sum.nobach, 1 + n.nobach)
quantile((theta.y.nobach.samp - theta.y.bach.samp), c(.025, .975))
```

```
##      2.5%     97.5%
## 0.1522763 0.7407917
```

```
quantile((y.pred.nobach - y.pred.bach), c(.025, .975))
```

```
##  2.5% 97.5%
##    -2     3
```

```
mean(theta.y.nobach.samp) - mean(theta.y.bach.samp)
```

```
## [1] 0.4500002
```

```
mean(y.pred.nobach) - mean(y.pred.bach)
```

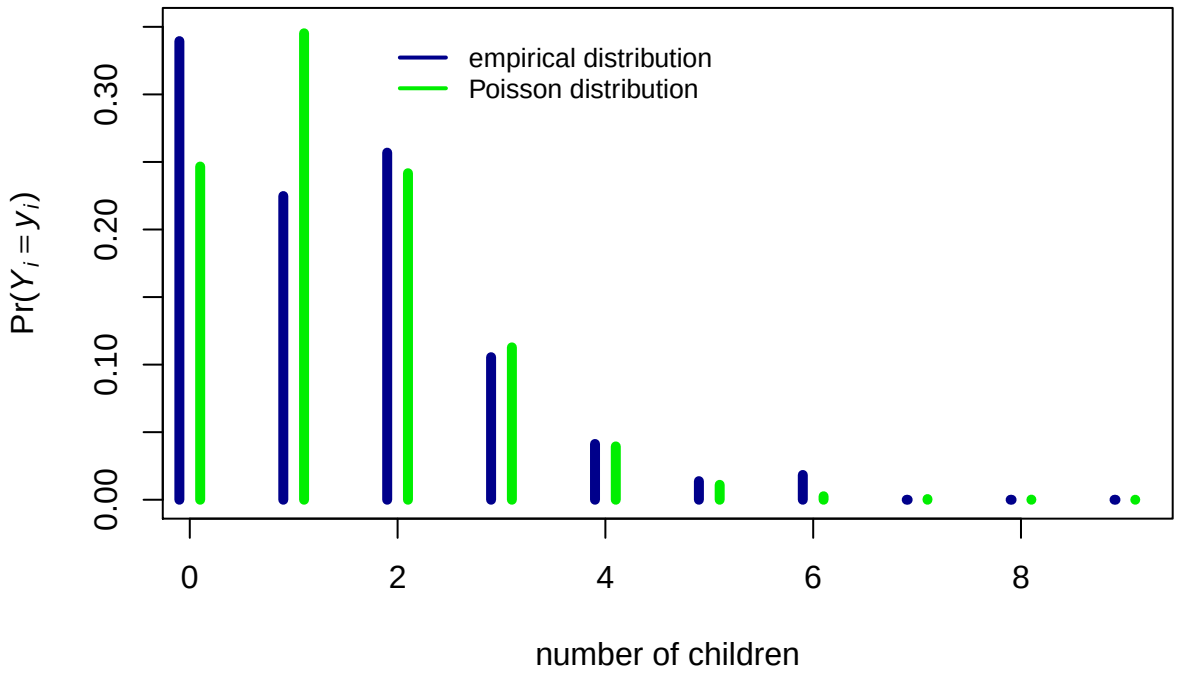
```
## [1] 0.4316
```

The means from both the posterior samples and posterior predictive samples are fairly similar, with the difference in means being ~.45 in the former and ~.43 in the latter, suggesting that men without bachelor's degrees have about .44 more children than those with bachelor's degrees. However, the posterior samples provide more certainty that the true difference in means is close to this value, as its confidence interval is more narrow. This difference in mean number of children is reflected in the histograms in a), as we can see that the bars spike at 0 for those with bachelors degrees and at 1 for those without bachelors degrees.

c)

```
d.emp <-(table(c(y.nobach,0:9))-1 )/sum(table(y.nobach))
dpois.samp <- dpois(0:9, 1.4, log = FALSE)
plot(0:9+.1, dpois.samp,type="h",lwd=5,xlab="number of children",
     ylab=expression(paste("Pr(",italic(Y[i]==y[i]),")",sep="")),col="green2",
      ylim=c(0,.35))
points(0:9-.1, d.emp,lwd=5,col="darkblue",type="h")
```

```
legend(1.8,.35,
     legend=c("empirical distribution","Poisson distribution"),
     lwd=c(2,2),col=
       c("darkblue","green2"),bty="n",cex=.8)
```



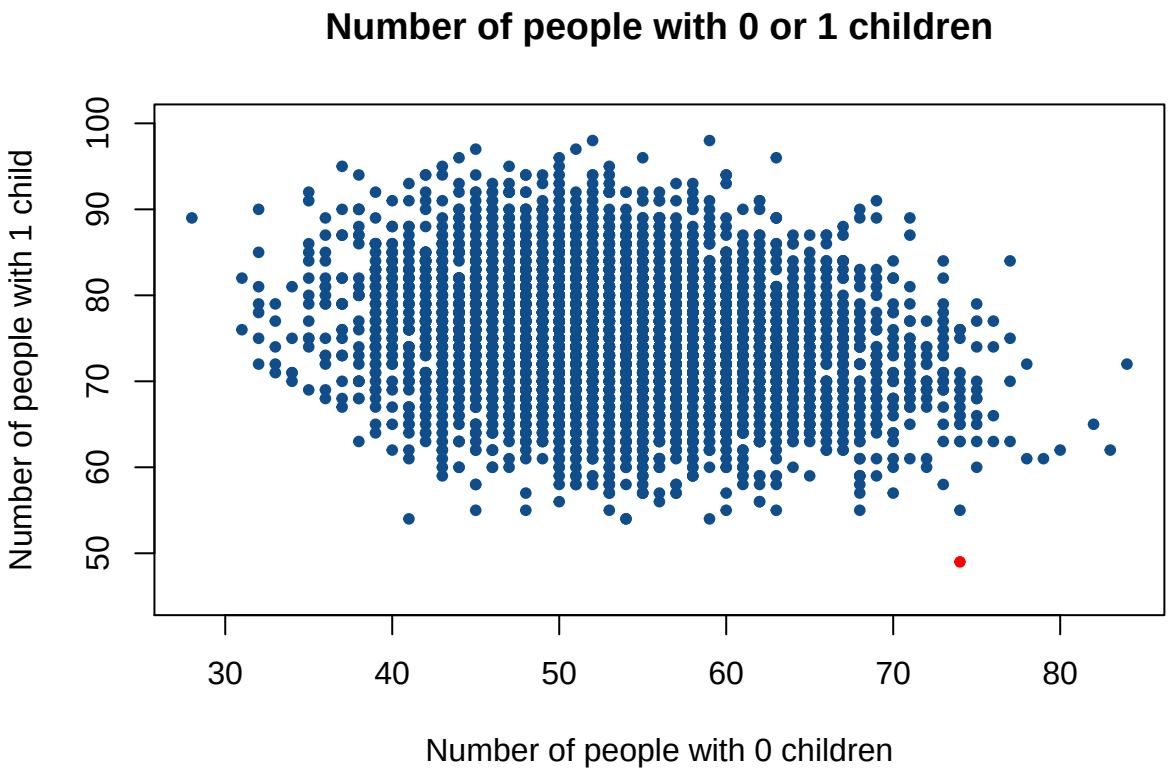

```
mean(y.nobach)
```

```
## [1] 1.399083
```

Compared to the empirical distribution, the Poisson underestimates the probability of 0 children and overestimates the probability of having 1 child. However, for number of children equal to or greater than 2, it seems to be very accurate, at least from viewing this histogram. Moreover, the mean of the empirical is about 1.399, which is very close to that of the Poisson, so the Poisson seems to be a good fit for this distribution.

d)

```
S.nb <- 5000
t.1 <- t.0 <- numeric(S.nb)
for (s in 1:5000) {
  theta.y.nobach.samp[s] <- rgamma(1, 2 + sum.nobach, 1 + n.nobach)
  sy.nb <- rpois(218, theta.y.nobach.samp[s])
  t.1[s] <- sum(sy.nb == 1)
  t.0[s] <- sum(sy.nb == 0)
}
plot(t.0, t.1, type = "p",
     main = "Number of people with 0 or 1 children",
     col = "dodgerblue4", pch = 20, xlab = "Number of people with 0 children",
     ylab = "Number of people with 1 child",
     ylim = c(45, 100))
points(sum(y.nobach == 0), sum(y.nobach == 1), col = "red", pch = 20)
```



Since the point from the empirical distribution (in red) is far away from the middle of the cluster of points, it seems from this plot that the Poisson is not an adequate model for the empirical distribution, at least in modeling men with 0 or 1 children. This conforms to the histogram in part c), as we saw that the estimates were off for 0 and 1 children.