Homework 4

Zach White

9/24/2016

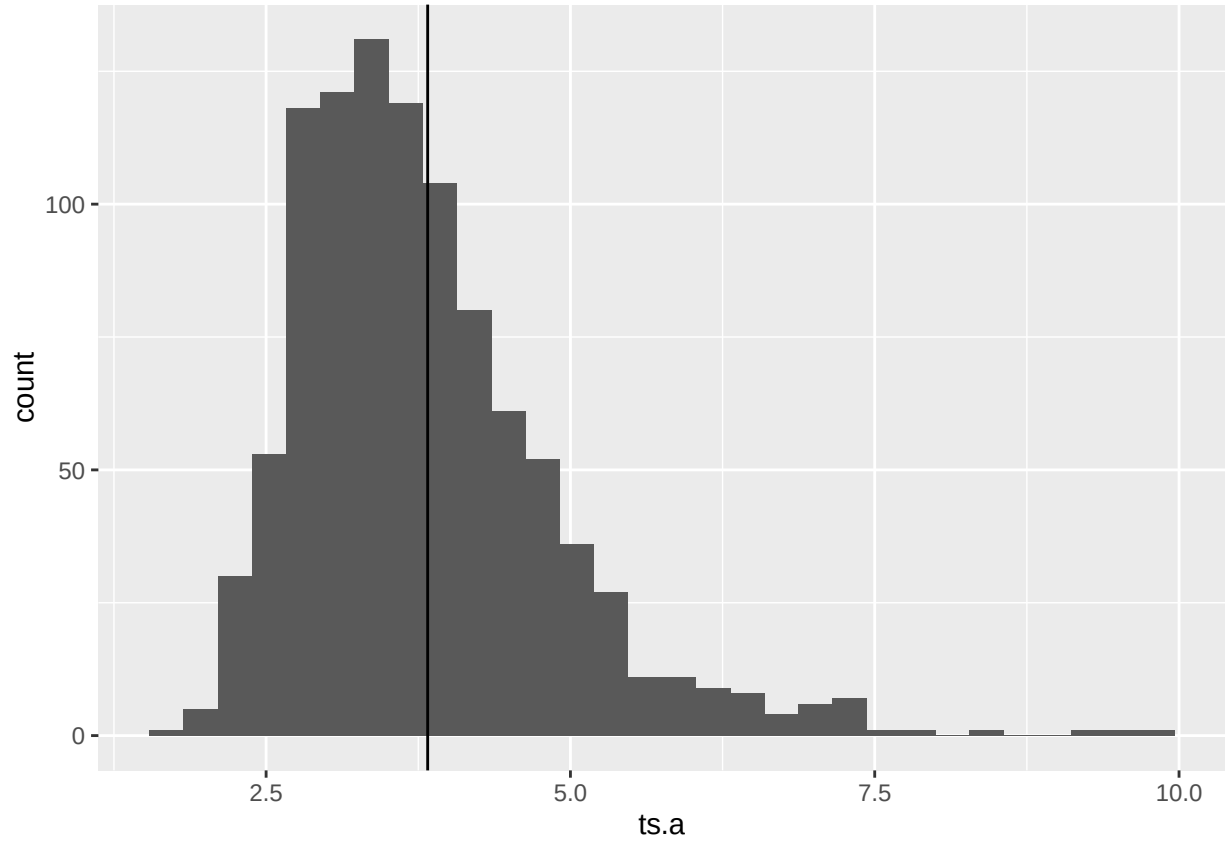
Exercise 4.3

```
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)

a.a = 120; a.b = 10
b.a = 12 ; b.b = 1
n.y.a = length(y.a)
n.y.b = length(y.b)
sumy.a = sum(y.a)
sumy.b = sum(y.b)
post.a.a = a.a + sumy.a
post.a.b = n.y.a + a.b

post.b.a = b.a + sumy.b
post.b.b = n.y.b + b.b

post.pred.a = matrix(NA,1000,10)
post.pred.b = matrix(NA, 1000,10)

theta.a.draws = rgamma(1000,post.a.a,post.a.b)
theta.b.draws = rgamma(1000,post.b.a,post.b.b)
for(i in 1:1000){
  post.pred.a[i,] = rpois(10,theta.a.draws[i])
  post.pred.b[i,] = rpois(10,theta.b.draws[i])
}

ts.a = apply(post.pred.a,1,function(x) mean(x)/sd(x))
ts.b = apply(post.pred.b,1,function(x) mean(x)/sd(x))
```

Part A

```
t.a.plot = ggplot(data = as.data.frame(ts.a), aes(x = ts.a))
t.a.plot + geom_histogram() + geom_vline(xintercept = mean(y.a) / sd(y.a))
```

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

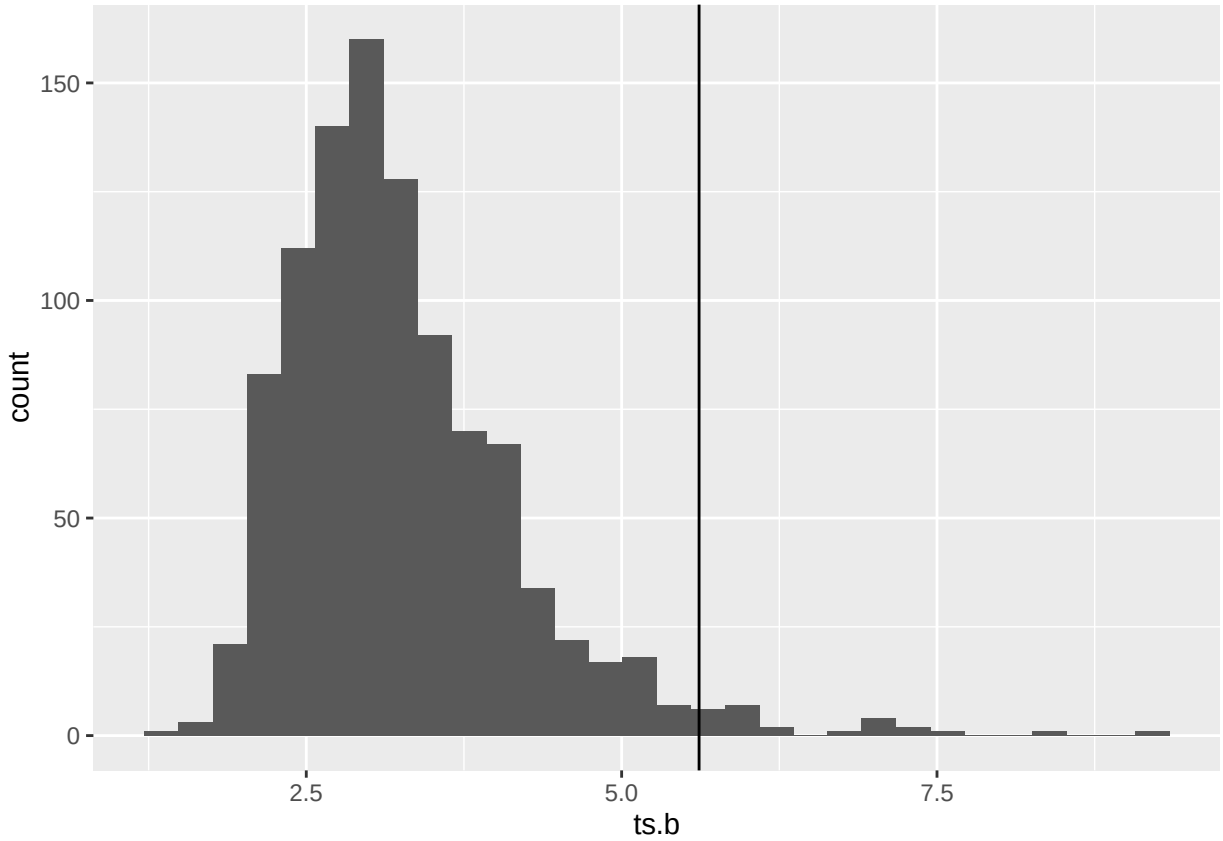


Our model in this case actually seems like a pretty good fit. The mean of the distribution is past the peak of the distribution, which is right skewed. So it isn't a great fit, but it is okay.

Part B

```
t.b.plot = ggplot(data = as.data.frame(ts.b), aes(x = ts.b))
t.b.plot + geom_histogram() + geom_vline(xintercept = mean(y.b) / sd(y.b))
```

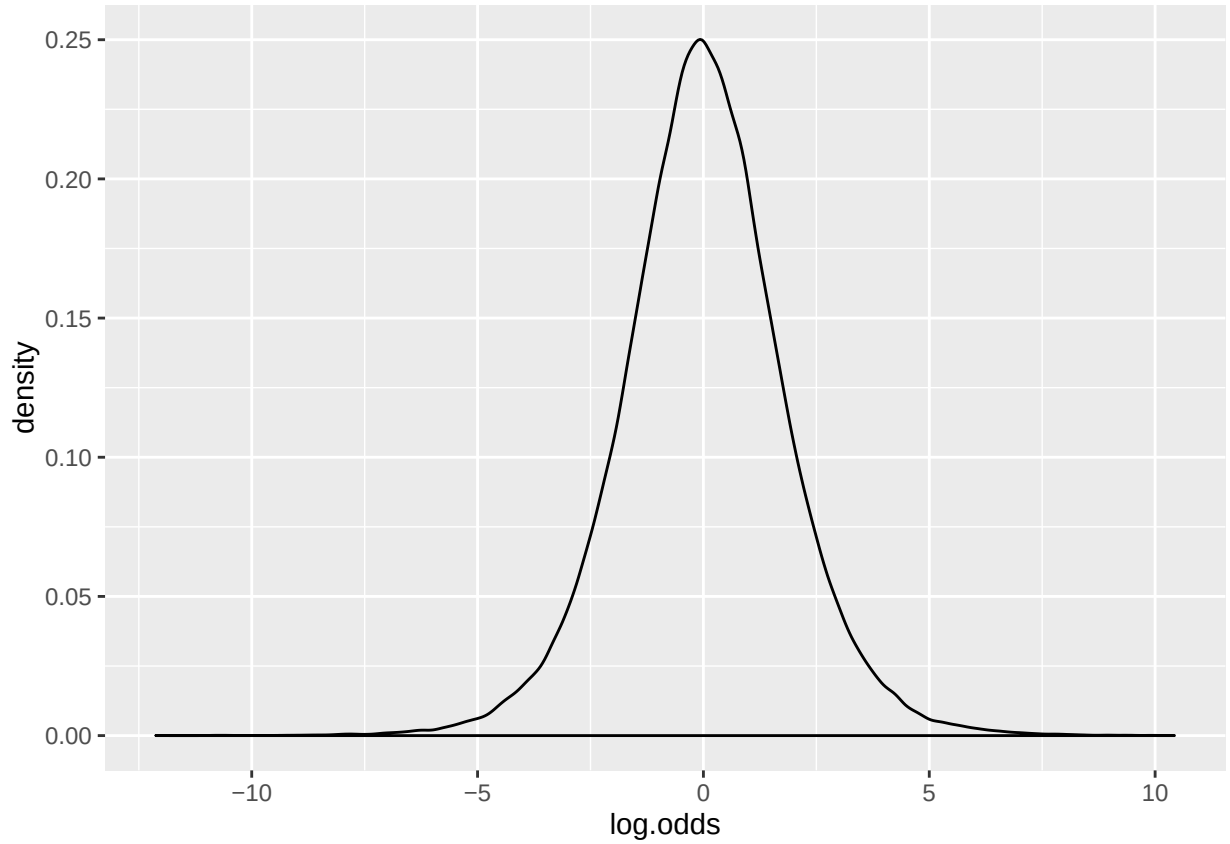
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The fit in this case is not good. The mean of the distribution is past the bulk of the distribution, which means that for this case, it is not the best. When comparing it to the fit of part A, it clearly isn't as good.

Exercise 4.6

```
a = 1
b = 1
theta.vals = rbeta(100000,a,b)
log.odds = log(theta.vals/(1-theta.vals))

ggplot(data = as.data.frame(log.odds), aes(x = log.odds)) + geom_density()
```



The uniform prior for $log(\frac{\theta}{1-\theta})$ would follow the distribution shown above. I didn't calculate the closed form of the prior, but it would look like this.

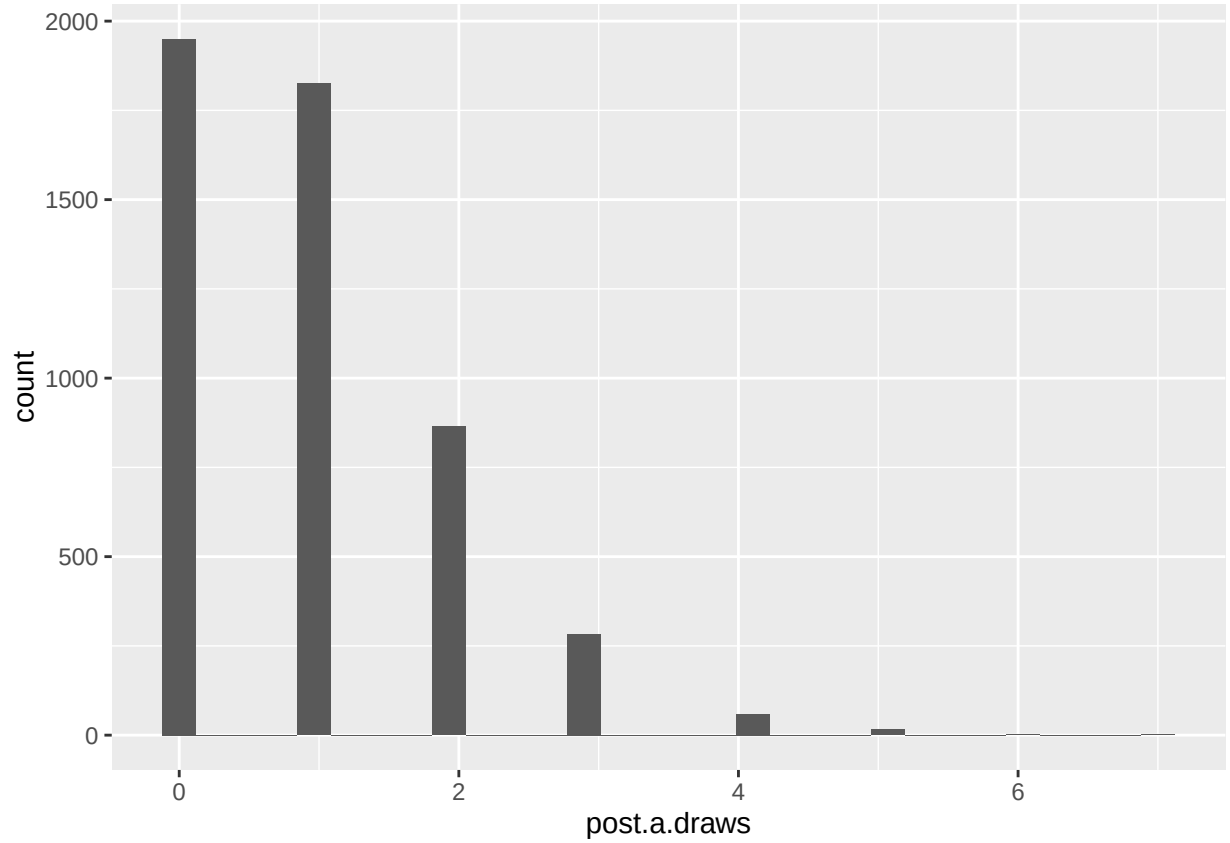
Exercise 4.8

Part A

```
menchild30 = scan("menchild30bach.dat", sep = " ")
menchild30not = scan("menchild30nobach.dat", sep = " ")
a = 2
b = 1
sum.y.a = sum(menchild30)
sum.y.b = sum(menchild30not)
n.a = length(menchild30)
n.b = length(menchild30not)
a.a.prime = sum.y.a + a
a.b.prime = n.a + b
b.a.prime = sum.y.b + b
b.b.prime = n.b + b

theta.a.draws = rgamma(5000,a.a.prime,a.b.prime)
theta.b.draws = rgamma(5000,b.a.prime,b.b.prime)

post.a.draws = rpois(5000, theta.a.draws)
post.b.draws = rpois(5000,theta.b.draws)
```

```
post.a.plot = ggplot(data = as.data.frame(post.a.draws), aes(x = post.a.draws))
post.a.plot + geom_histogram()
```

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




```
post.b.plot = ggplot(data = as.data.frame(post.b.draws), aes(x = post.b.draws))
post.b.plot = geom_histogram()
```

Part B

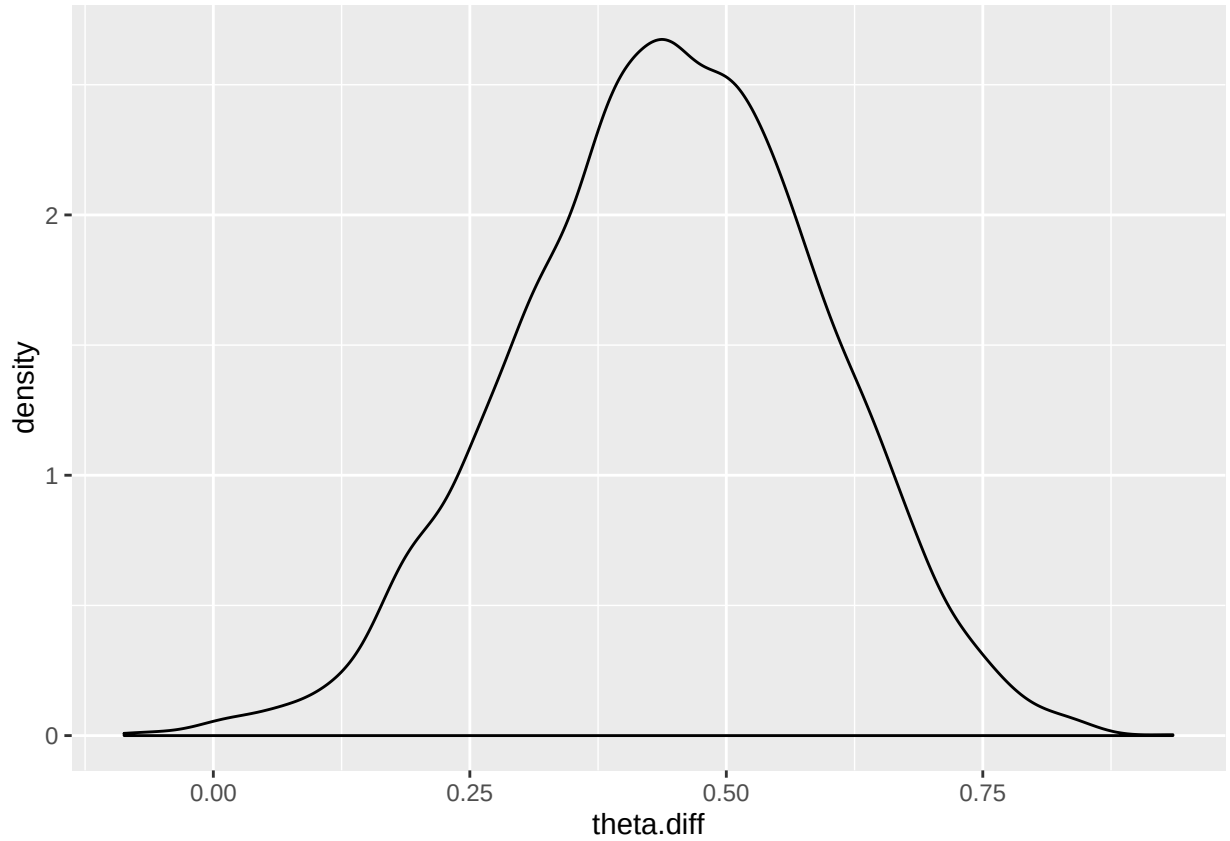
```
theta.diff = theta.b.draws - theta.a.draws
post.pred.diff = post.b.draws - post.a.draws
quantile(theta.diff,c(.025,.975))
```

```
##      2.5%     97.5%
## 0.1577145 0.7222290
```

```
quantile(post.pred.diff,c(.025,.975))
```

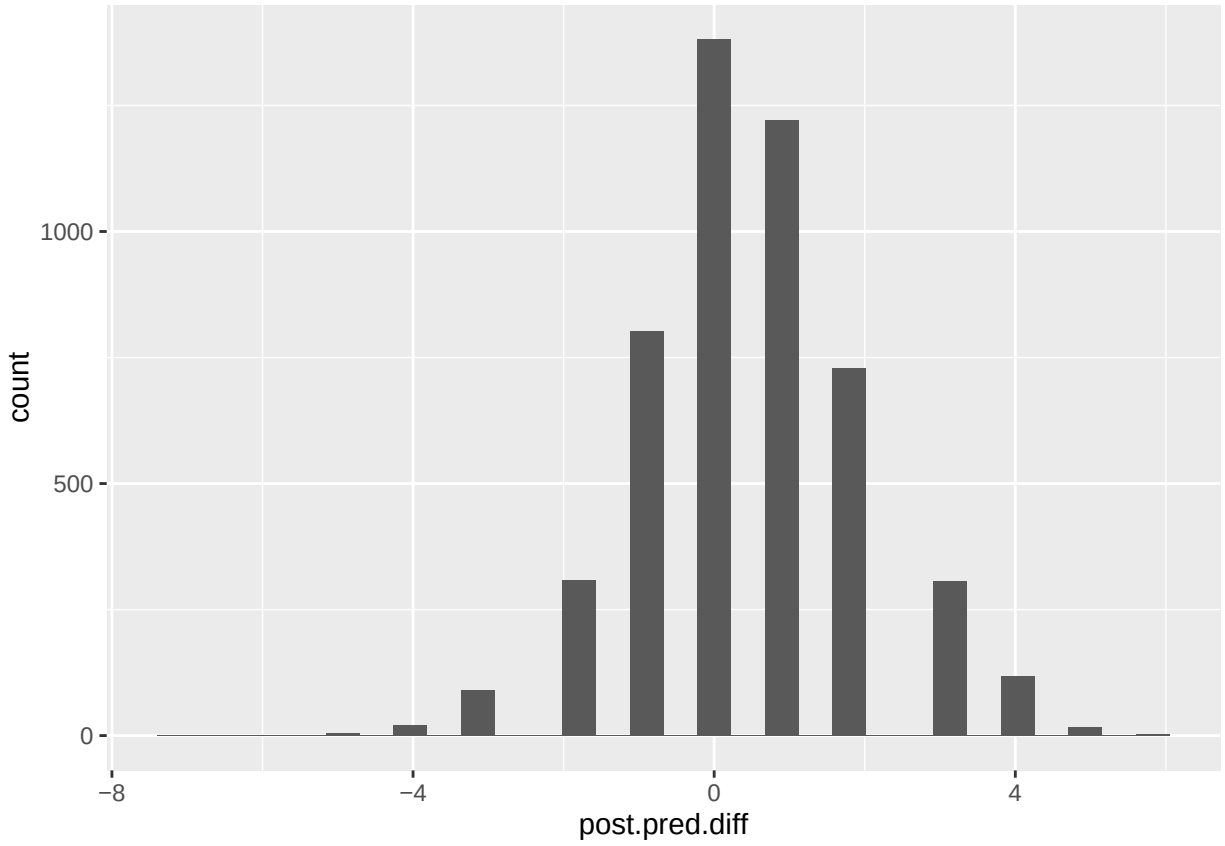
```
##  2.5% 97.5%
##    -2     4
```

```
theta.diff.int = quantile(theta.diff,c(.025,.975))
pred.diff.int = quantile(post.pred.diff,c(.025,.975))
ggplot(as.data.frame(theta.diff), aes(x = theta.diff)) + geom_density()
```




```
ggplot(as.data.frame(post.pred.diff), aes(x = post.pred.diff)) + geom_histogram()
```

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

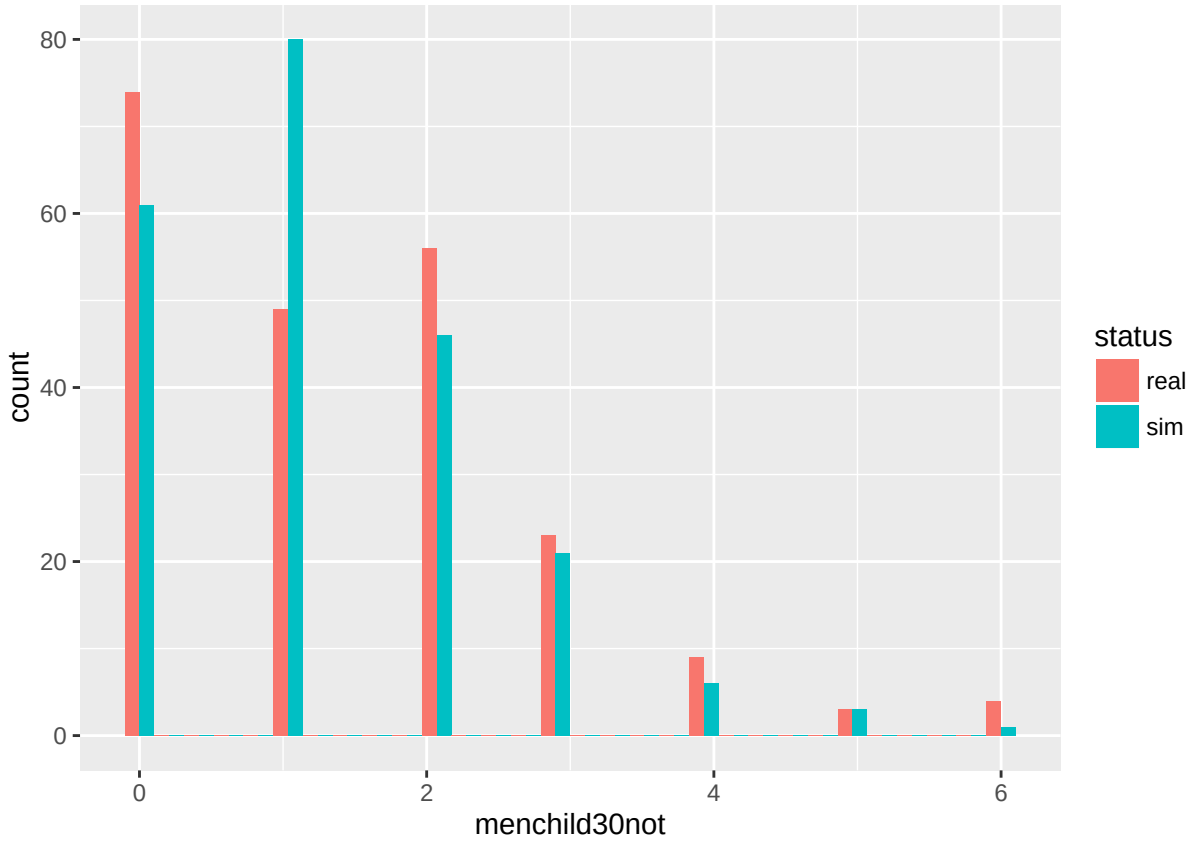


According to this simulation, there is a .95 probability that men without bachelor degrees have on average between 0.1577145, 0.722229 more kids than men with bachelor degrees. There is .95 probability that a man without a bachelor degree will have between -2, 4 more kids than a man with a bachelor degree. Thus, it seems like this is significant on average, but when it comes to prediction and actual counts, the difference isn't as significant.

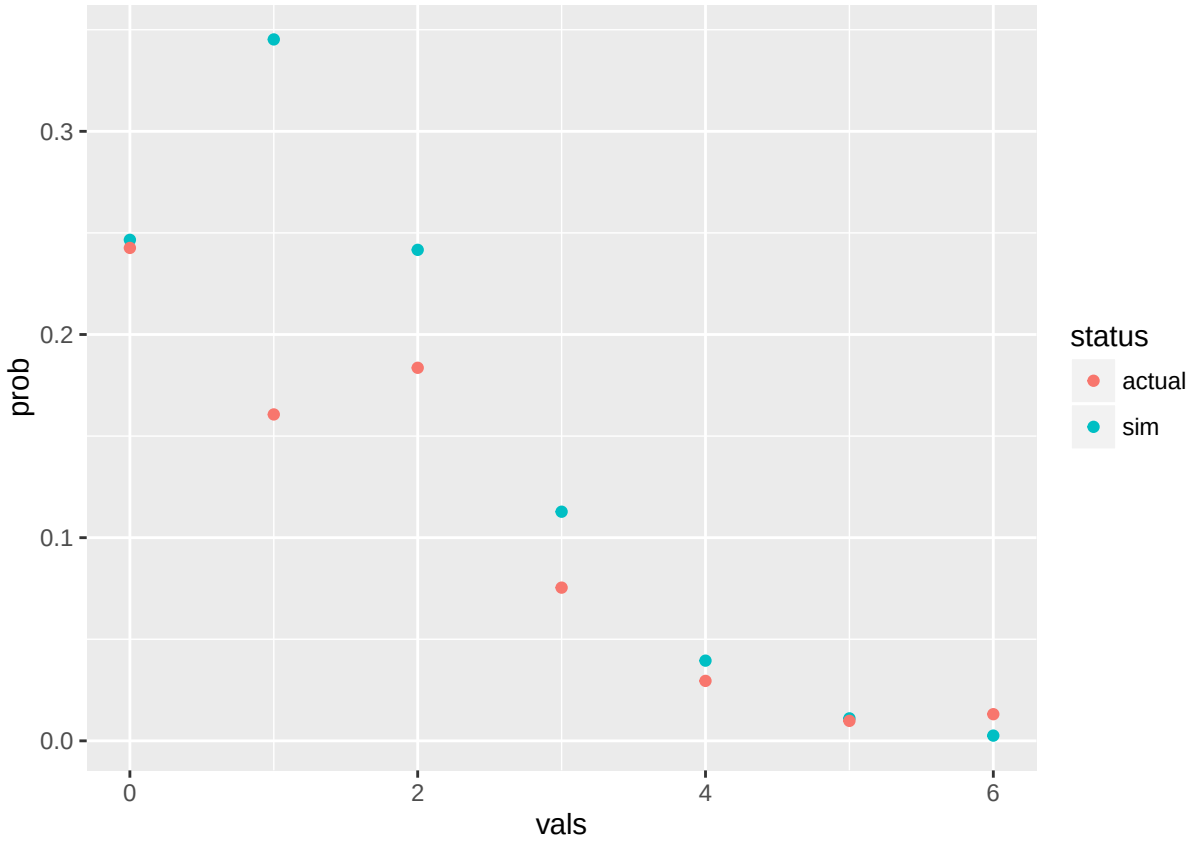
Part C

```
n = length(menchild30not)
param = 1.4
theoret.pois = rpois(n,param)
menchild30not = as.data.frame(menchild30not)
theoret.pois = as.data.frame(theoret.pois)
menchild30not$status = "real"
theoret.pois$status = "sim"
names(theoret.pois) = names(menchild30not)
emp.data.frame = as.data.frame(rbind(menchild30not,theoret.pois))


ggplot(emp.data.frame, aes(x = menchild30not, fill = status)) + geom_histogram(position = "dodge")
```

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



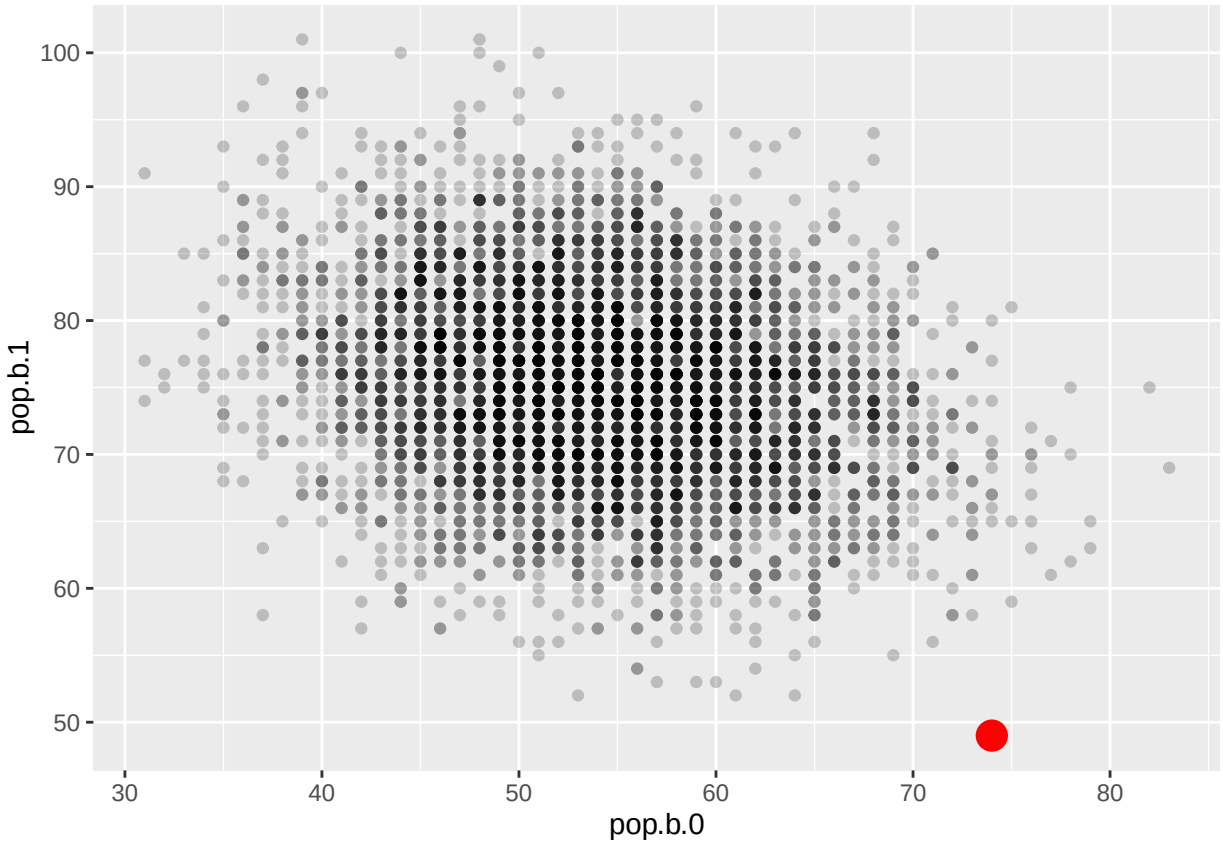

```
menchild30not = scan("menchild30nobach.dat", sep = " ")
props.act = as.data.frame(table(menchild30not) / sum(menchild30not))
props.sim = as.data.frame(dpois(0:6,1.4))
names(props.sim) = c("kids")
props.act$status = "actual"
props.sim$status = "sim"
props.act = props.act[,c(2,3)]
names(props.act) = names(props.sim)

props.data.frame = rbind(props.sim,props.act)
props.data.frame$vals = c(0:6,0:6)

ggplot(props.data.frame, aes(x = vals, y = kids, color = status)) + geom_point() + ylab("prob")
```



Our model follows the trend relatively well, but it's clear that it misses out on some key features. For example, our model estimates much more men with 1 child than was actually observed.

Part D

```
#theta.b.draws
menchild30not = scan("menchild30nobach.dat", sep = " ")
n.samps = 5000

pop.b.0 = rep(NA,5000)
pop.b.1 = rep(NA, 5000)
for(i in 1:n.samps){
  theta.b.draw = theta.b.draws[i]
  sim.kids = rpois(218, theta.b.draw)
  pop.b.0[i] = sum(sim.kids == 0)
  pop.b.1[i] = sum(sim.kids == 1)
}

act.0 = sum(menchild30not == 0)
act.1 = sum(menchild30not == 1)

act.data.frame = as.data.frame(cbind(act.0,act.1))
sim.data.set = data.frame(cbind(pop.b.0,pop.b.1))
ggplot(data = sim.data.set, aes(x = pop.b.0, y = pop.b.1)) + geom_point(alpha = .2) + geom_point(data =
```



Our model basically misses out on this feature, which means that our model is probably lacking in some way.