STA 601/360 Homework 4

Yi Mi

25 September, 2019

Exercise 1

Hoff 4.3

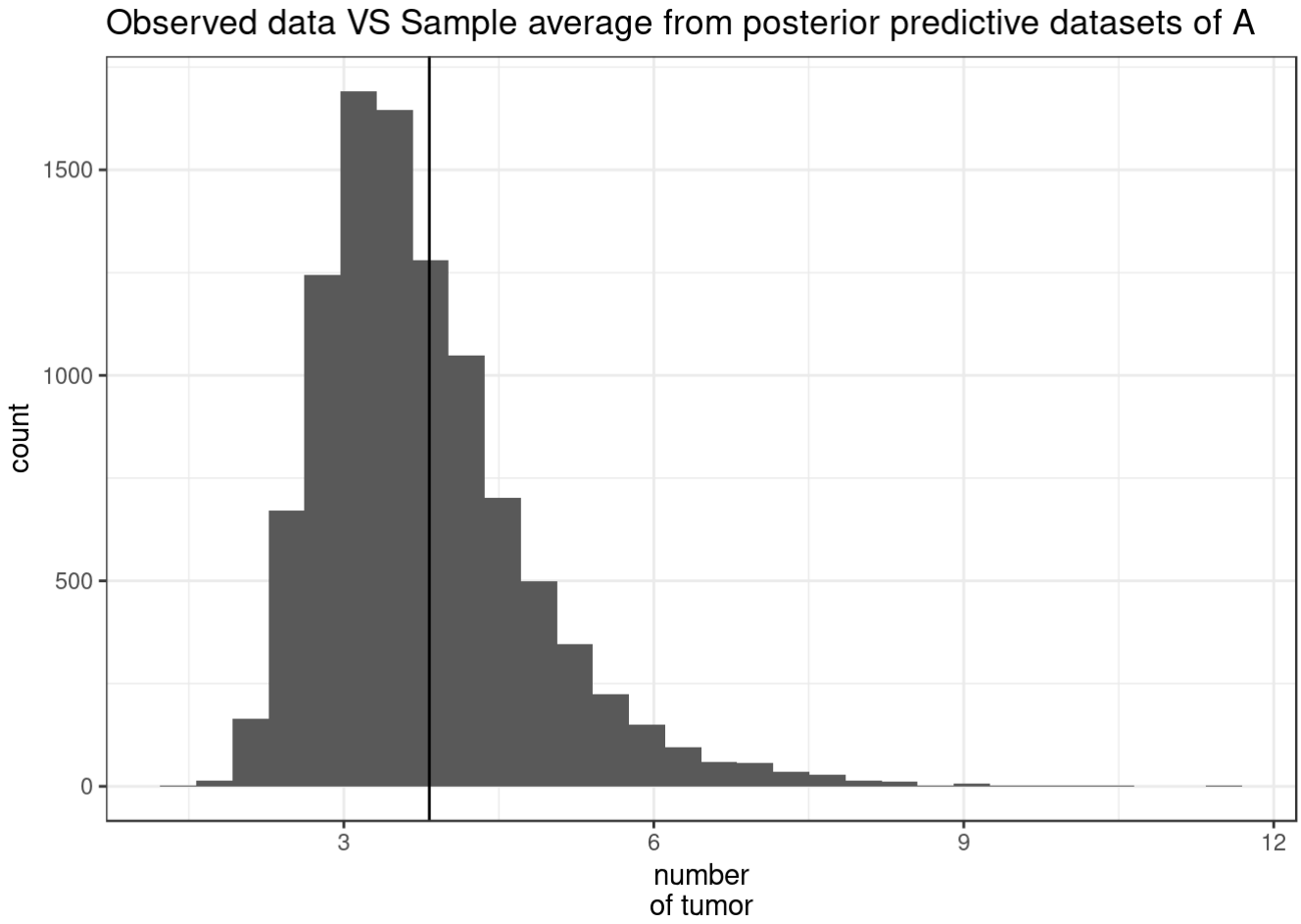
Part (a)

```
yA = c(12, 9, 12, 14, 13, 13, 15, 8, 15, 6)
yB = c(11, 11, 10, 9, 9, 8, 7, 10, 6, 8, 8, 9, 7)

set.seed(20)
thetaA.1 = rgamma(10000, 237, 20)
ts.A=c()
for(i in 1:10000){
  y=rpois(10, thetaA.1[i])
  ts.A[i]=mean(y)/sd(y)
}

t.A=mean(yA)/sd(yA)
ggplot(data.frame(ts.A), aes(x = ts.A)) + geom_histogram() + geom_vline(xintercept=t.A) + lab
  s(x='number
of tumor', title='Observed data VS Sample average from posterior predictive datasets of A')
```

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



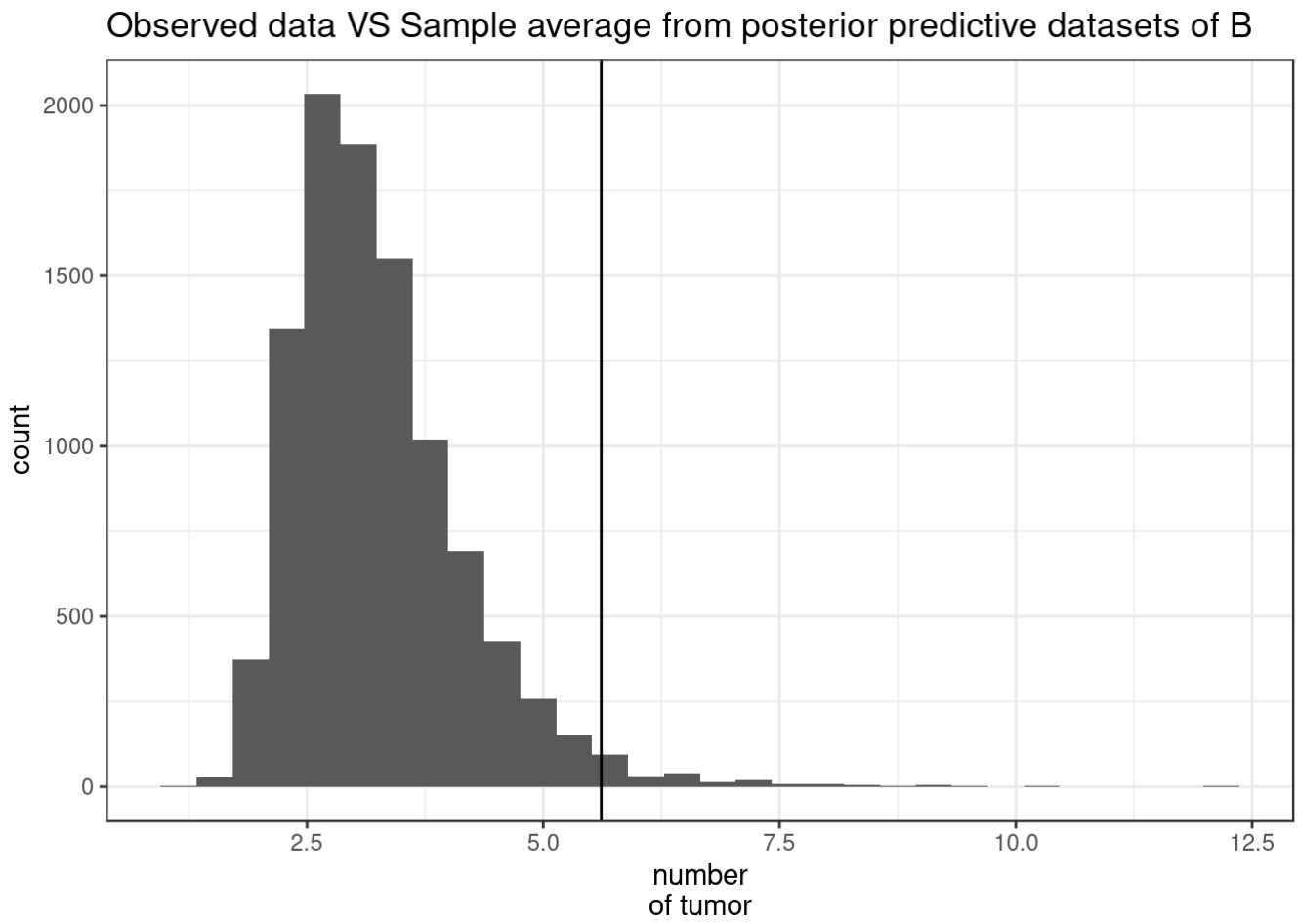
- The statistic of observed value lies near the mode of posterior predictive distribution of statistics(mean/sd), so the Poisson model is not a bad model for fitting the data.

Part (b)

```
set.seed(20)
thetaB.1 = rgamma(10000, 125, 14)
ts.B=c()
for(i in 1:10000){
  y=rpois(10, thetaB.1[i])
  ts.B[i]=mean(y)/sd(y)
}

t.B=mean(yB)/sd(yB)
ggplot(data.frame(ts.B), aes(x = ts.B)) + geom_histogram() + geom_vline(xintercept=t.B) + lab
  s(x='number
of tumor', title='Observed data VS Sample average from posterior predictive datasets of B')
```

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



- The statistic of observed value lies far from the mode of posterior predictive distribution of statistics(mean/sd), so the Poisson model is not a good model for fitting the data and maybe we should think about other models to fit.

Exercise 2

Hoff 4.7

Part (a)

```
set.seed(20)
sigma2=1/rgamma(10000,10,2.5)
theta=rnorm(10000,4.1,sigma2/20)
y.tilde=0.31*rnorm(10000,theta,sqrt(sigma2))+0.46*rnorm(10000, 2*theta+2*sqrt(sigma2))+0.23*r
  norm(10000, 3*theta+3*sqrt(sigma2))
```

Part (b)

```
quantile(y.tilde, c(0.125, 0.875))
```

```
##    12.5%    87.5%
## 8.060307 9.346233
```

Part (c)

i.

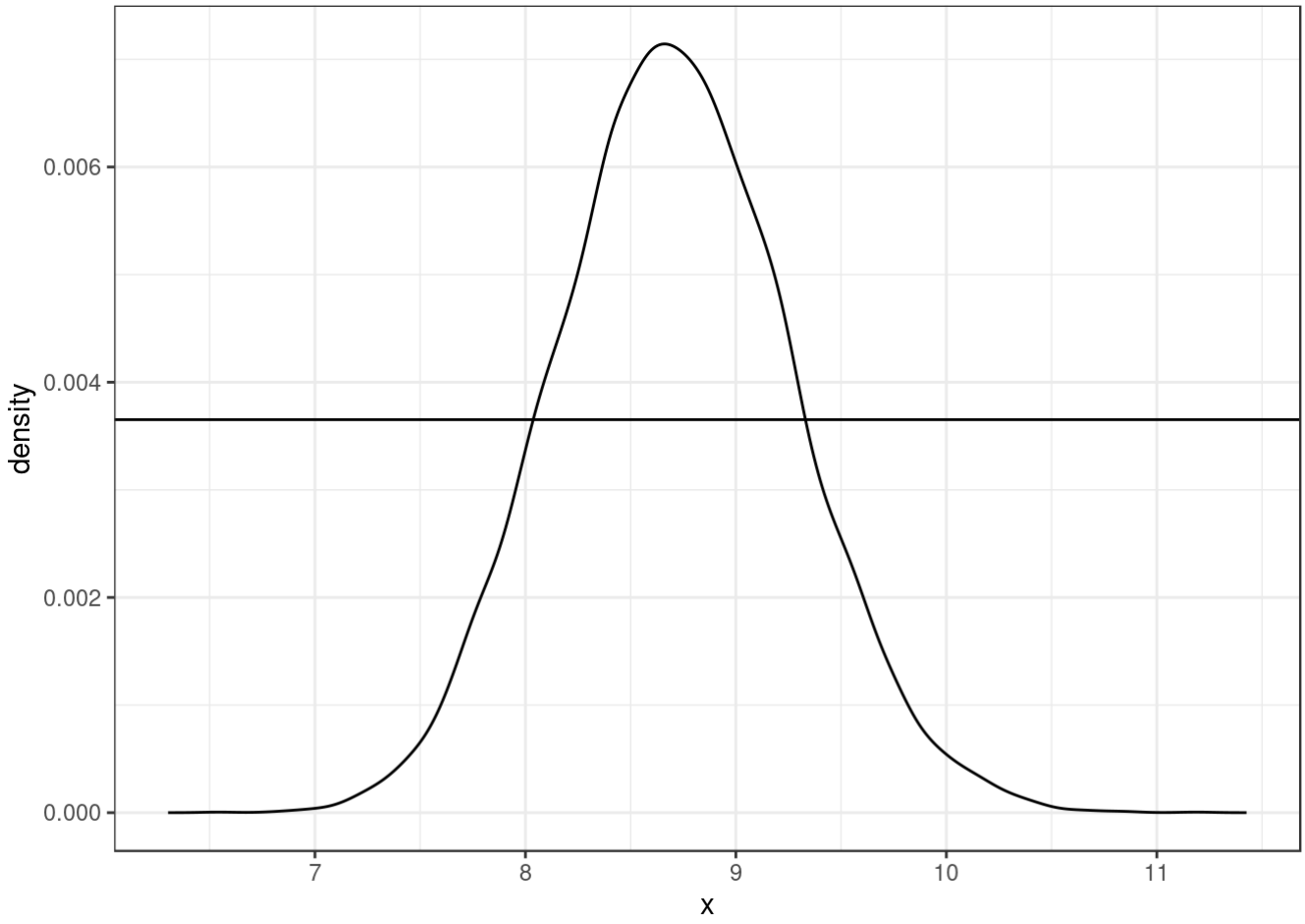
```
d=density(y.tilde)
d.norm=d$y/sum(d$y)
```

ii.

```
d.norm.dec=d.norm[order(d.norm, decreasing = T)]
```

iii.

```
position=min(which(cumsum(d.norm.dec)>=0.75))
cutoff=d.norm.dec[position]
df2=data.frame(x=d$x,density=d.norm)
ggplot(df2, aes(x=x,y=density)) + geom_line(aes(x = x, y = density)) + geom_hline(yintercept=
  cutoff)
```




```
HPD=d.norm>cutoff
d$x[min(which(HPD))]
```

```
## [1] 8.046928
```

```
d$x[max(which(HPD))]
```

```
## [1] 9.330452
```

- The cutoff line of HPD region is the horizontal line showing in the plot and all values of y in HPD region have a discretized probability greater than the cutoff. HPD region is similar to quantile-based region.

Part (d)

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

- Sorry I don't know which physical justification is proper.

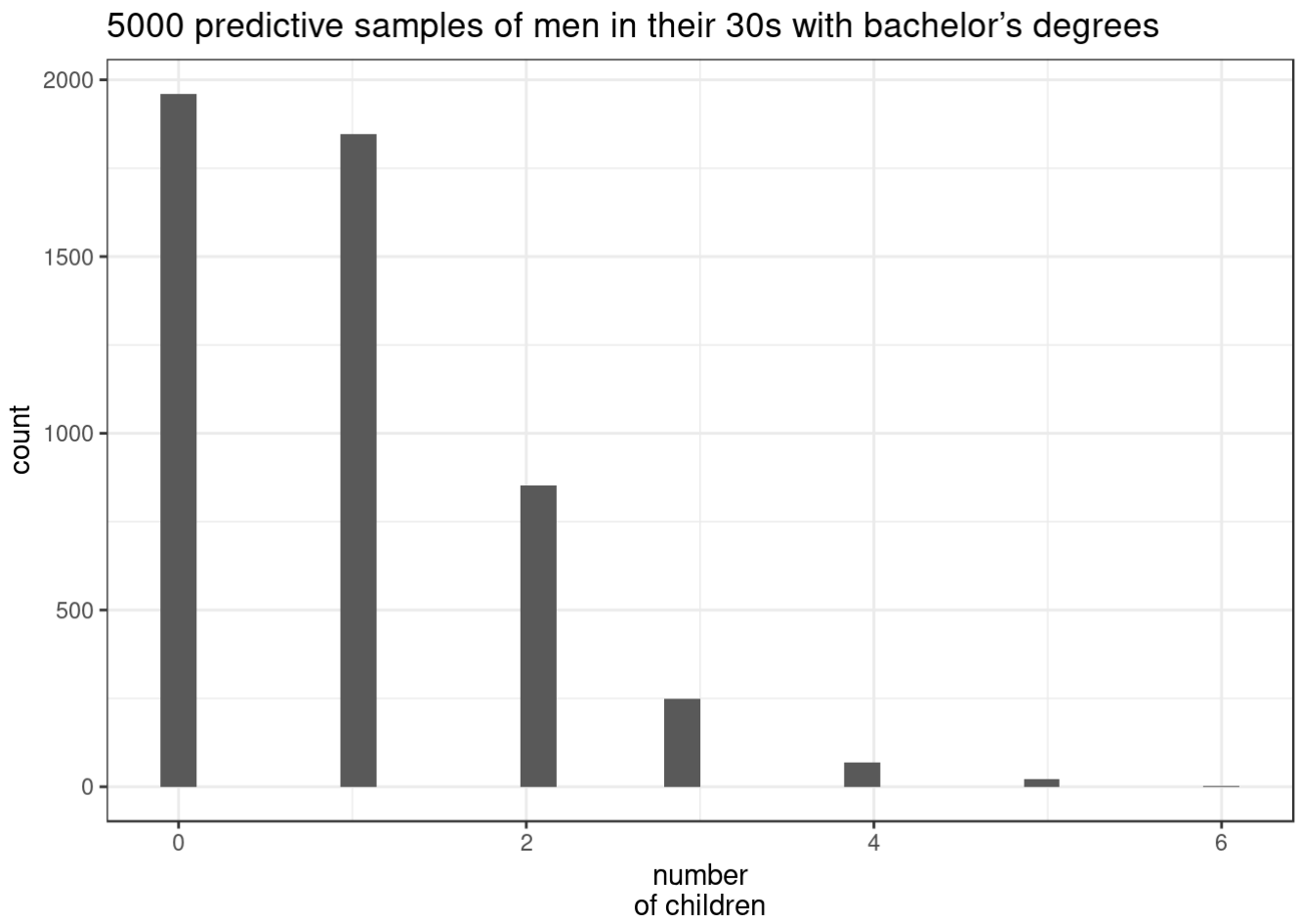
Exercise 3

Hoff 4.8

Part (a)

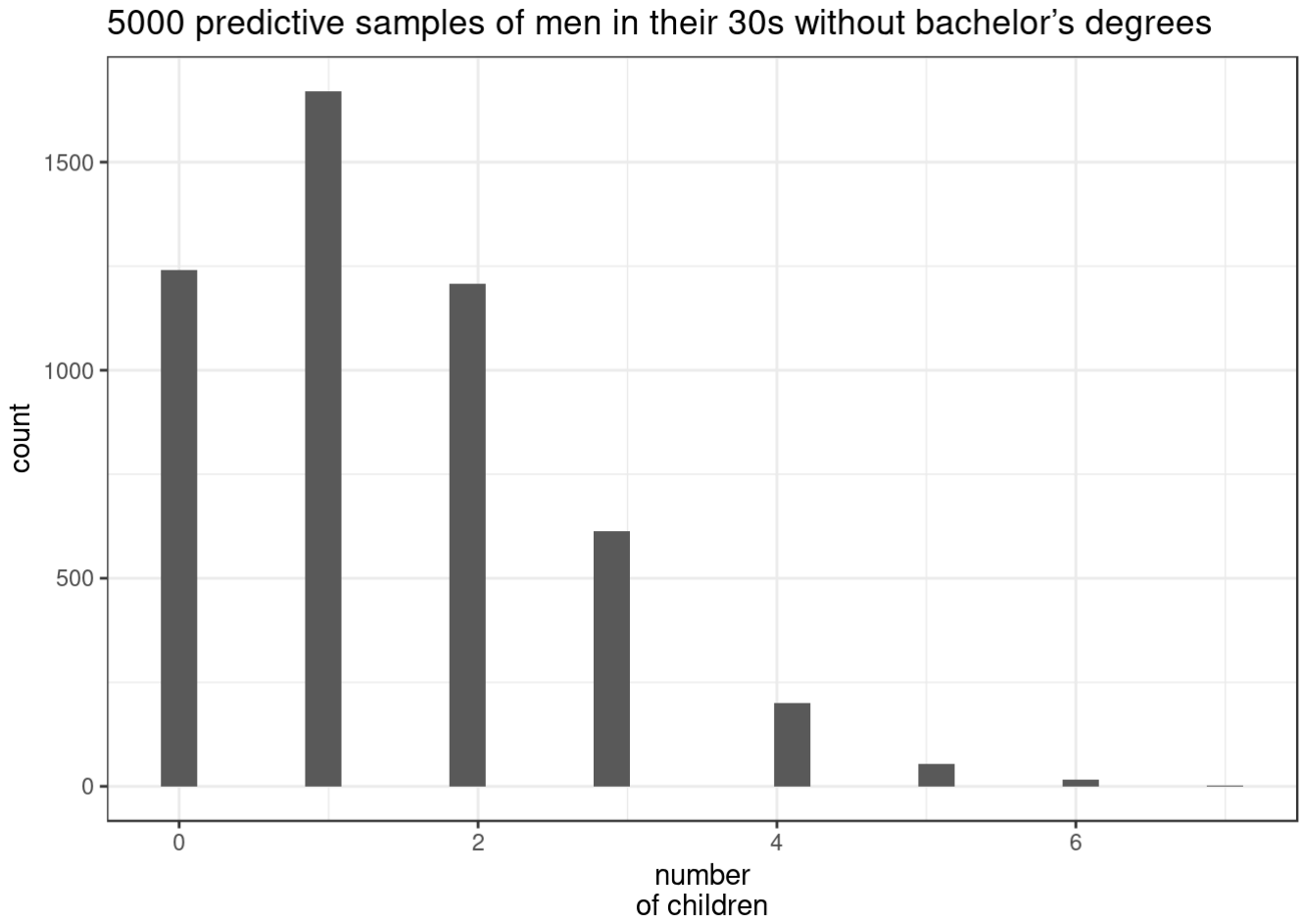
```
A=scan(file=url("http://www.stat.washington.edu/~pdhoff/Book/Data/hwdata/menchild30bach.dat"
  ))
B=scan(file=url("http://www.stat.washington.edu/~pdhoff/Book/Data/hwdata/menchild30nobach.da
  t"))
A.sum=sum(A)
A.n=length(A)
B.sum=sum(B)
B.n=length(B)
thetaA.2=rgamma(5000,2+A.sum,1+A.n)
thetaB.2=rgamma(5000,2+B.sum,1+B.n)
ytilde.A=rpois(5000,thetaA.2)
ytilde.B=rpois(5000,thetaB.2)
ggplot(data.frame(ytilde.A), aes(x = ytilde.A)) + geom_histogram() + labs(x='number
of children', title='5000 predictive samples of men in their 30s with bachelor's degrees')
```

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




```
ggplot(data.frame(ytilde.B), aes(x = ytilde.B)) + geom_histogram() + labs(x='number
of children', title='5000 predictive samples of men in their 30s without bachelor’s degrees')
```

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



Part (b)

```
quantile(thetaB.2-thetaA.2, c(0.025, 0.975))
```

```
##      2.5%     97.5%
## 0.1512869 0.7422191
```

```
quantile(ytilde.B-ytilde.A, c(0.025, 0.975))
```

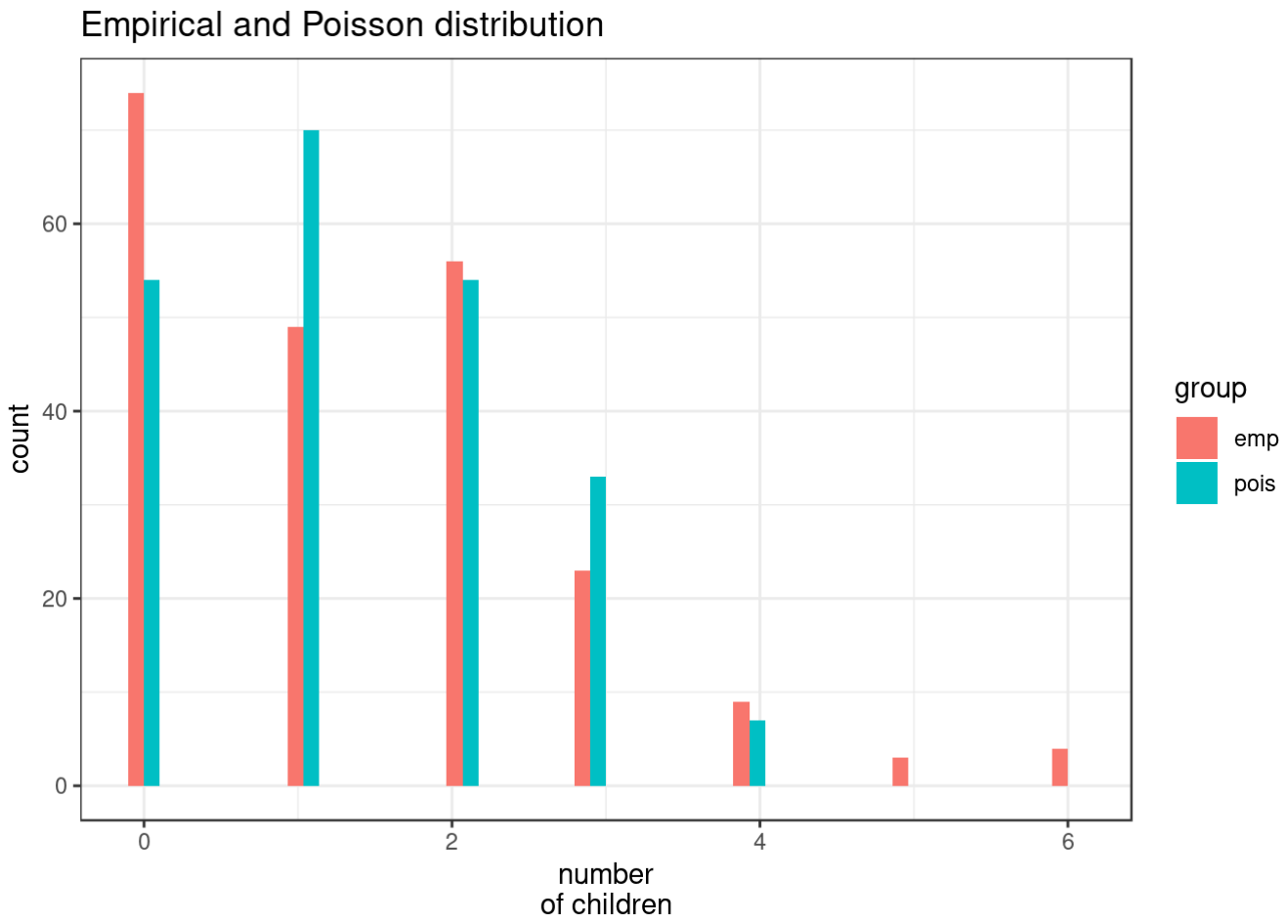
```
##  2.5% 97.5%
##    -2     4
```

- The posterior confidence interval for $\theta_B - \theta_A$ shows that we could be 95% certain that θ_B is greater than θ_A. The posterior confidence interval for $\tilde{y}_B - \tilde{y}_A$ includes 0, indicating that the we can not be sure about the relation between $\tilde{y}_B$ and $\tilde{y}_A$.

Part (c)

```
ypois.B=rpois(218,1.4)
df3=rbind(data.frame(y=B, group='emp'),data.frame(y=ypois.B, group='pois'))
ggplot(df3, aes(x=y, group=group, fill=group)) + geom_histogram(position = 'dodge') + labs(x=
  'number
of children',title='Empirical and Poisson distribution')
```

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



- I don’t think Poisson model is a good fit because: 1.there are two peaks in emperical distribution but only one peak in Poisson distribution, indicating the shape of these two model does not match; 2.the mode of emperial distribution is 0 and the mode of poisson distribution is 1, indicating the Poisson model may not be a good fit.

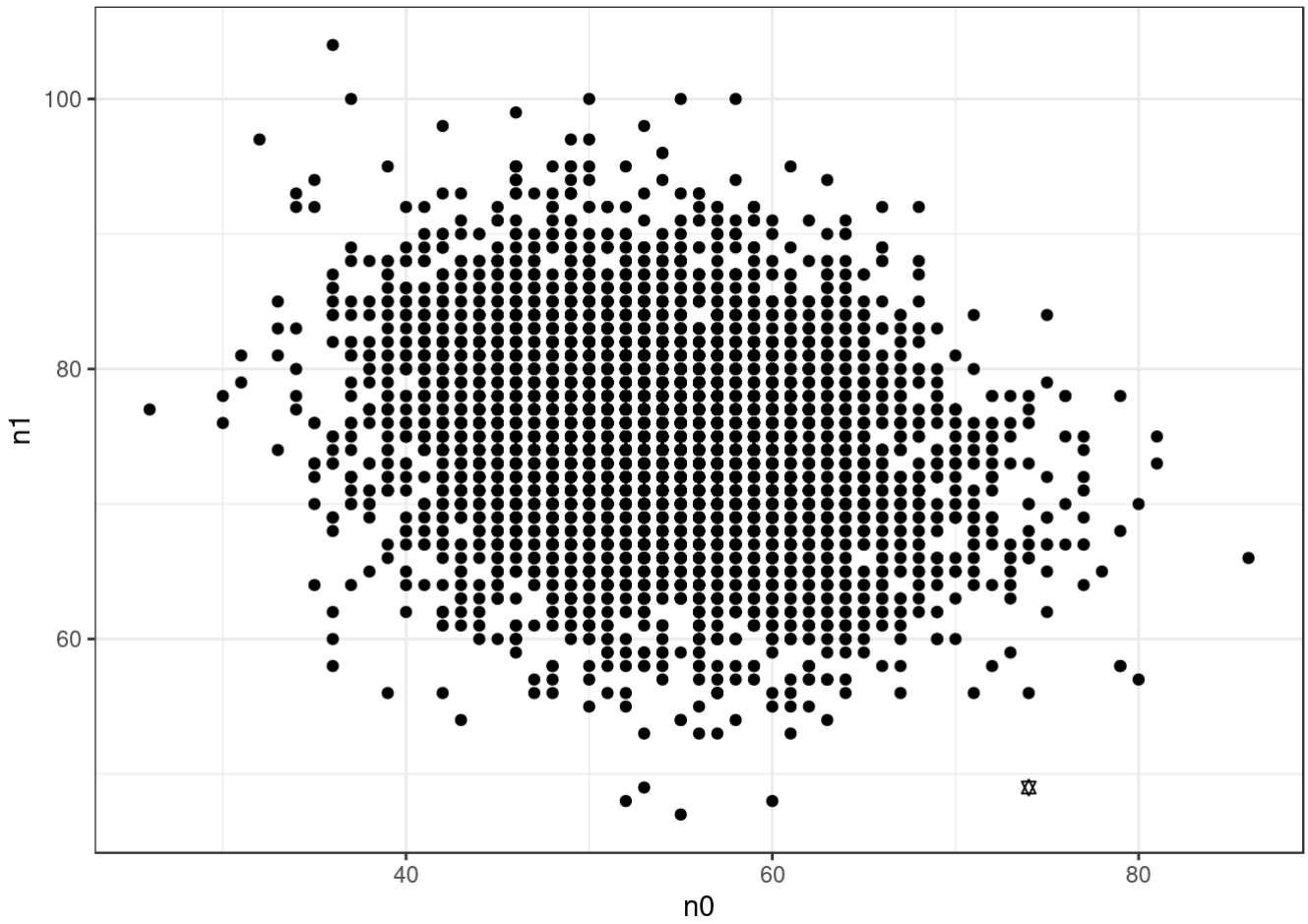
Part (d)

```
n0=c()
n1=c()
for(i in 1:5000){
  theta=thetaB.2[i]
  y=rpois(218,theta)
  n0[i]=0
  n1[i]=0

  for(j in y){
    if(j==0){
      n0[i]=n0[i]+1
    }
    if(j==1){
      n1[i]=n1[i]+1
    }
  }

}

point.x=sum(B==0)
point.y=sum(B==1)
ggplot(data.frame(n0,n1),aes(x=n0,y=n1))+geom_point()+annotate('point',x=point.x,y=point.y,sh
  ape='star')
```



- The point marking observed data lies in lower right corner of the plot and is not even inside the points simulated from Poisson model, so I don't think the Poisson model is adequate.

Exercise 3

Let Y1,…,Yn|α,β be iid Gamma(α,β) with α known. (1) Find the conjugate family of priors for β; (2) find the posterior; (3) Give an interpretation of the prior parameters as things like “prior mean”, “prior variance”, “prior sample size”, etc.

Part (a)

$$Y \sim Gamma(a, b)$$

$$p(y \mid a, \phi) = h(y \mid \phi)c(\phi)e^{\phi t(y)} = \frac{b^a}{\Gamma(a)} y^{a-1} e^{-by}$$

$$\implies \phi = b, t(y) = y, h(y) = -\frac{y^{a-1}}{\Gamma(a)}, c(\phi) = b^a$$

prior distribution:

$$p(\phi \mid n_0, t_0) \propto c(\phi)e^{n_0 t_0 \phi} \propto b^{an_0} e^{n_0 t_0 b}$$

$$\implies$$

$$p(b \mid n_0, t_0) = p(\phi \mid n_0, t_0)\frac{|d\phi|}{|db|} \propto b^{an_0} e^{n_0 t_0 b}$$

$$b \mid n_0, t_0 \sim Gamma(b, an_0 + 1, -n_0 t_0)$$

Part (b)

posterior distribution:

$$\begin{aligned} p(b \mid y_1, \dots y_n) &\propto p(b)p(y_1, \dots y_n \mid b) \\ &\propto b^{an_0} e^{n_0 t_0 b} * b^{an} e^{-b \sum y_i} \\ &\propto b^{an_0 + an} e^{-(\sum y_i - n_0 t_0)b} \\ b \mid y_1, \dots y_n &\sim Gamma(b, an_0 + an + 1, \sum y_i - n_0 t_0) \end{aligned}$$

Part (c)

$$p(y_1, \dots y_n \mid a, b) = \frac{b^{an}}{(\Gamma(a))^n} (\prod_{i=1}^{n} y)^{a-1} e^{-b \sum_{i=1}^{n} y}$$

$$L = log(p(y_1, \dots y_n \mid a, b)) = an \log b + (a-1) \prod_{i=1}^{n} y - b \sum_{i=1}^{n} y + c$$

$$\frac{dL}{db} = \frac{an}{b} - \sum_{i=1}^{n} y = 0$$

$$\hat{b}_{MLE} = \frac{an}{\sum_{i=1}^{n} y}$$

$$\text{prior mean} = \frac{an_0 + 1}{-n_0 t_0}$$

$$\text{prior mean} = \frac{an_0 + 1}{(n_0 t_0)^2}$$

$\text{prior sample size} = n_0$

n_0 indicates how informative the prior is

t_0 is the prior estimate of t(y), which is y

$$\text{posterior mean} = \frac{an_0 + an + 1}{\sum_{i=1}^{n} y - n_0 t_0} = \underbrace{\frac{-n_0 t_0}{\sum_{i=1}^{n} y - n_0 t_0}}_{\text{weight of prior mean}} * \underbrace{\frac{an_0 + 1}{-n_0 t_0}}_{\text{prior mean}} + \underbrace{\frac{\sum_{i=1}^{n} y}{\sum_{i=1}^{n} y - n_0 t_0}}_{\text{weight of MLE}} * \underbrace{\frac{an}{\sum_{i=1}^{n} y}}_{\text{MLE}}$$

Posterior mean of b is the weighted sum of prior mean of b and MLE of b.