STAT 638 Homework 5

Jack Cunningham

4.6)

In order to understand what prior $\gamma = \log \frac{\theta}{1-\theta}$ will have given a uniform distribution of θ we need to find $p_\gamma(\gamma)$. This is:

$$p_\gamma(\gamma) = p_\theta(h(\gamma)) \times |\frac{dh}{d\gamma}|$$

Where $h(\gamma)$ is the inverse of $\gamma = g(\theta) = \log \frac{\theta}{1-\theta}$.

After some algebra we have $h(\gamma) = \frac{e^\gamma}{1+e^\gamma}$.

We can find the first derivative more easily if rewrite $h(\gamma)$ as:

$$h(\gamma) = \frac{1}{e^{-\gamma}+1}$$

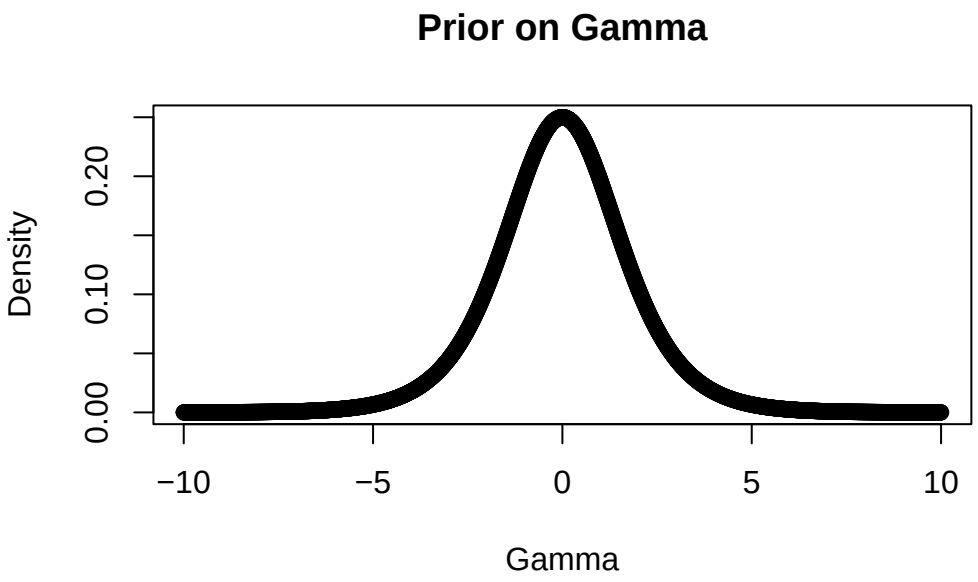
Then we have:

$$\frac{dh}{d\gamma} = \frac{1}{e^\gamma(1+e^{-\gamma})^2}$$

Since the uniform prior is placed on θ we have $p_\theta(\theta) = 1$ for all θ . So the prior $p_\gamma(\gamma)$ is:

$$p_\gamma(\gamma) = \frac{1}{e^\gamma(1+e^{-\gamma})^2}$$

```
gamma = seq(-10,10,.01)
density = 1/(exp(gamma)*(1 + exp(-gamma))^2)
plot(gamma, density, xlab = "Gamma", ylab = "Density", main = "Prior on Gamma")
```



We can see that this prior is clearly informative about γ .

4.8)

Reading Data:

```
bach <- read.csv("bach.csv")
no_bach <- read.csv("nobach.csv")
```

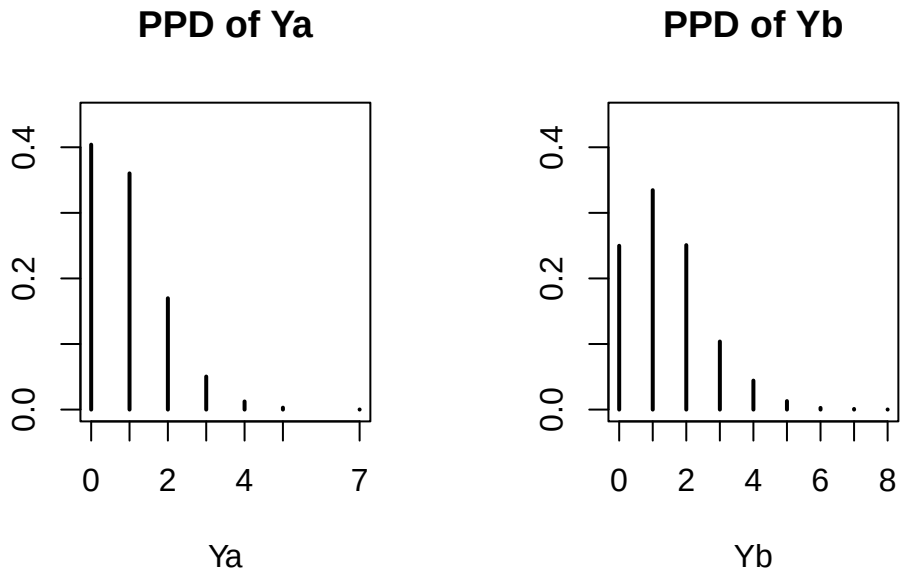
Monte Carlo:

```
set.seed(2)
a <- 2; b <- 1
sy1 = sum(bach); n1 <- dim(bach)[1]
sy2 = sum(no_bach); n2 <- dim(no_bach)[1]

theta1.mc <- rgamma(5000, a + sy1, b + n1)
theta2.mc <- rgamma(5000, b + sy2, b + n2)
y1.mc <- rpois(5000, theta1.mc)
y2.mc <- rpois(5000, theta2.mc)
```

Plotting:

```
par(mfrow = c(1,2))
y1.ppd = table(y1.mc)/length(y1.mc)
y2.ppd = table(y2.mc)/length(y2.mc)
plot(y1.ppd, xlab = "Ya", ylab = "",main = "PPD of Ya", ylim = c(0,.45))
plot(y2.ppd, xlab = "Yb", ylab = "",main = "PPD of Yb", ylim = c(0,.45))
```



b)

$\theta_B - \theta_A$ confidence interval:

```
theta_diff = theta2.mc - theta1.mc
quantile(theta_diff,probs = c(.025,.975))
```

```
     2.5%     97.5%
0.1407746 0.7314419
```

$\widetilde{Y}_B - \widetilde{Y}_A$:

```
y_diff = y2.mc - y1.mc
quantile(y_diff,probs = c(.025,.975))
```

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

Although the hypothesis $\theta_B - \theta_A = 0$ is rejected at a p-value of .05 we see that $Y_B - Y_A$ does have 0 included in its 95% probability interval. One main difference between the two populations is the fact that significantly more men with a bachelors degree are predicted to have no children than men without bachelors degrees. The population of men without a bachelors degree also has a longer and fatter tail than men with bachelors, they are more likely to have larger families.

It seems that the population of men with bachelors degrees may be best fit by a zero inflated poisson model which can account for the high probability placed on zero kids.

c)

Plotting a comparison:

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.2     v tibble    3.2.1
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.0.4
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to beco
```

```
library(reshape2)
```

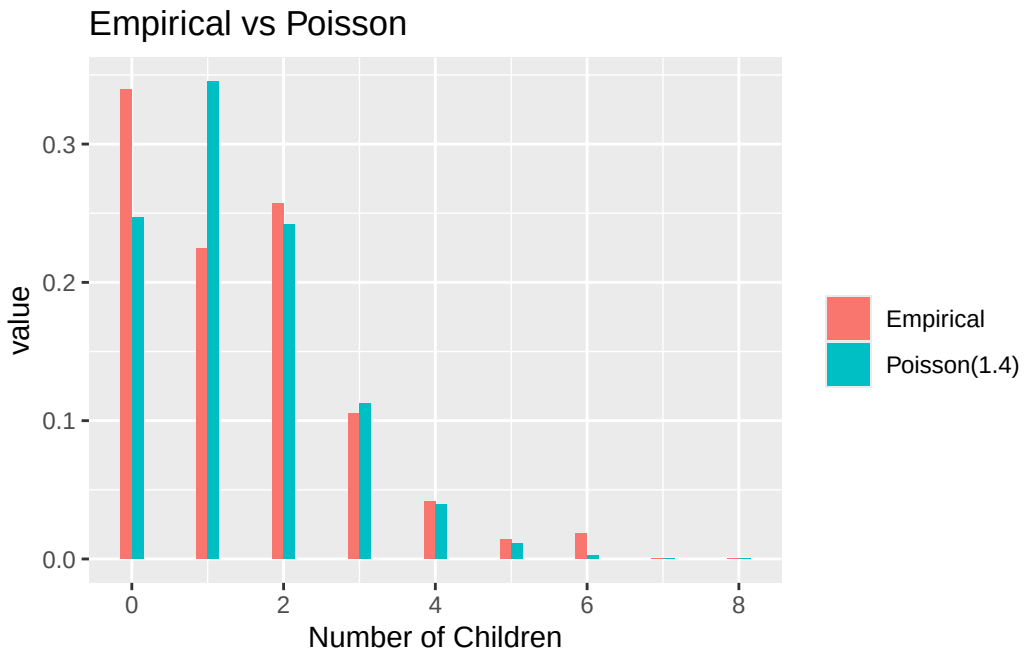
```
Attaching package: 'reshape2'

The following object is masked from 'package:tidyr':

    smiths
```

```r
ed_b = table(no_bach)/dim(no_bach)[1]
pois_b = dpois(x = seq(0,10),lambda = 1.4)
x = seq(0,8)
ed_b = c(ed_b,0,0)
pois_b = pois_b[1:9]
df = data.frame(x, ed_b, pois_b)
df2 = melt(df,id.var = 'x')

df2 |>
  ggplot(aes(x = x, y = value, fill = variable)) +
  geom_bar(stat = 'identity', position = 'dodge',width = .3) +
  scale_fill_discrete(name = "", labels = c("Empirical","Poisson(1.4)")) +
  labs(title = "Empirical vs Poisson",x = "Number of Children")
```



We can see a large discrepancy between the two distributions at the number of children equal to 0 and 1. The poisson distribution expects around 25% of men without bachelors degree to have no children while the observed data is about 35%. The poisson distribution expects 35% of men without bachelors degrees to have one child, while the observed data is about 20%. These two deviations are enough to conclude that this model is not a good fit.

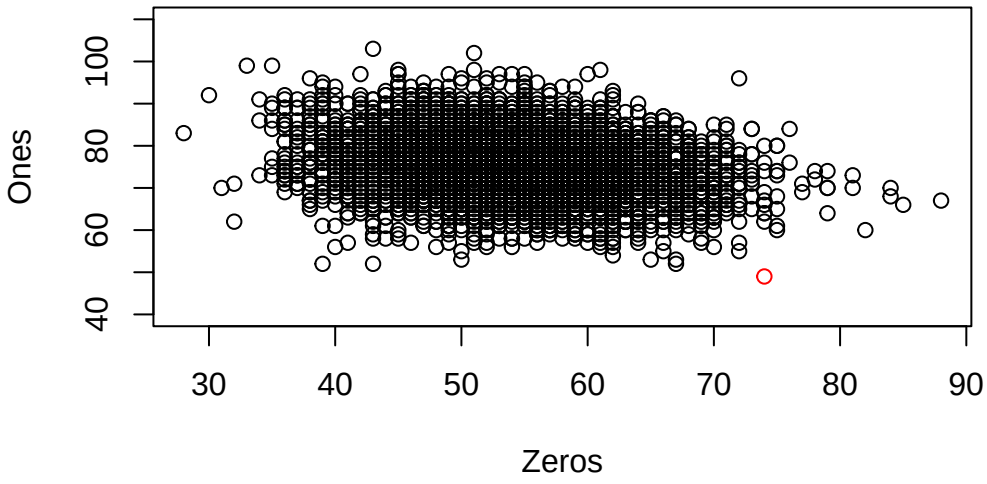
d)

Simulation:

```r
set.seed(99)
counts = matrix(0, nrow = 5000, ncol = 2,dimnames = list(c(),c("Zeros","Ones")))
for(i in 1:length(theta2.mc)) {
  poisson_pull = rpois(218,theta2.mc[i])
  counts[i,"Zeros"] <-  sum(poisson_pull == 0)
  counts[i,"Ones"] <-  sum(poisson_pull == 1)
}
```

Plot:

```r
observed_zeros = sum(no_bach$Count == 0)
observed_ones = sum(no_bach$Count == 1)
plot(counts[,"Zeros"],counts[,"Ones"], ylim = c(40,110),
     xlab = "Zeros",ylab = "Ones")
points(observed_zeros,observed_ones,col = "red")
```



The observed data is the red point on this plot. We can see how none of the simulated data is close to what we are seeing in the actual data set. This makes it clear that the Poisson model is not a good choice to model this data set. Instead we should consider other models, such as the zero inflated Poisson model.

5.5)

a)

The reparameterized normal model is:

$$p(y|\theta,\psi) = \sqrt{\frac{\psi}{2\pi}} e^{-1/2\psi(y-\theta)^2}$$

To get the log likelihood we first take the log:

$$\log(p(y|\theta,\psi)) = \frac{1}{2}\log(\psi) - \frac{1}{2}\log(2\pi) - \frac{1}{2}\psi(y_i-\theta)^2$$

Then we sum over all y_i:

$$l(\theta,\psi|y) = \frac{n}{2}\log(\psi) - \frac{2}{n}\log(2\pi) - \frac{1}{2}\psi\sum_{i=1}^{n}(y_i-\theta)^2$$

b)

We are looking for a probability density where:

$$\log(p_U(\theta,\psi)) = \frac{1}{2}\log(\psi) - \frac{1}{2n}\psi\sum_{i=1}^{n}(y_i-\theta)^2 + c$$

Using our tip we can rewrite as:

$$\log(p_U(\theta,\psi)) = \frac{1}{2}\log(\psi) - \frac{1}{2n}\psi(\sum_{i=1}^{n}(y_i-\bar{y})^2) - \frac{1}{2}\psi(\theta-\bar{y})^2 + c$$

Then using $\log(p_U(\theta,\psi)) = \log(p_U(\theta|\psi)) + \log(p_U(\psi))$ we can say that:

$$p_U(\theta|\psi) \propto e^{-1/2\psi(\theta-\bar{y})^2}$$

We can identify that as $p(\theta|\psi) \sim \text{normal}(\bar{y}, 1/\psi)$.

And we can also say:

$$p_U(\psi) \propto \psi^{1/2}$$

Which we can identify as $p(\psi) \sim$ gamma(3/2,0), which is improper but shouldn't serve to be a problem later.

Having identified both parts of the distribution we can say that:

$$p_U(\theta, \psi) = \text{normal-gamma}(\bar{y}, 1, 3/2, 0)$$

c)

We write out $p_U(\theta, \psi) \times p(y_1, ..., y_n|\theta, \psi)$ as:

$$\psi^{n/2} e^{-\psi/2 \sum_{i=1}^{n} (y_i - \theta)^2} \psi^{1/2} e^{-\psi/2(\theta - \bar{y})^2}$$

That leaves us with:

$$\psi^{(n+1)/2} e^{-\psi/2[\sum_{i=1}^{n} (y_i - \theta)^2 - \psi/2(\theta - \bar{y})]}$$

This joint density can be considered a posterior density, this is proportional to a normal gamma distribution.