Lab Assignment 5

Nora Quick

```
library(arm)
library(Sleuth3)
library(tidyverse)
library(vcdExtra)
library(magrittr)
library(MASS)
library(pscl)
```

0. Compare the coefficients and their standard errors for the "extra zero" parts of the `mod.nb0` and `mod.nb.hurdle` models. Are the coefficients and standard errors similar or different? Explain.

We can see that we have two sets of coefficients and very sidderent standard errors. We get this difference because of how the two models handel the zeros (zero-inflated lets Poisson and nb models create zero as well).

```
library(VGAM)
n <- 150 ## set the sample size
pois <- rpois(n, lambda = 5)
negbin <- rnbinom(n, mu = 5, size = 1.4)
pois0 <- rzipois(n, lambda = 6.25, pstr0 = 0.2)
sim.df <- as.data.frame(cbind(pois, negbin, pois0))
```

1. Run the above code and then make three histograms of the three sets of simulated data. Also, get summary statistics for each of the three sets of simulations.

```
#1
summary(pois)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   3.000   5.000   4.947   6.000  12.000
```

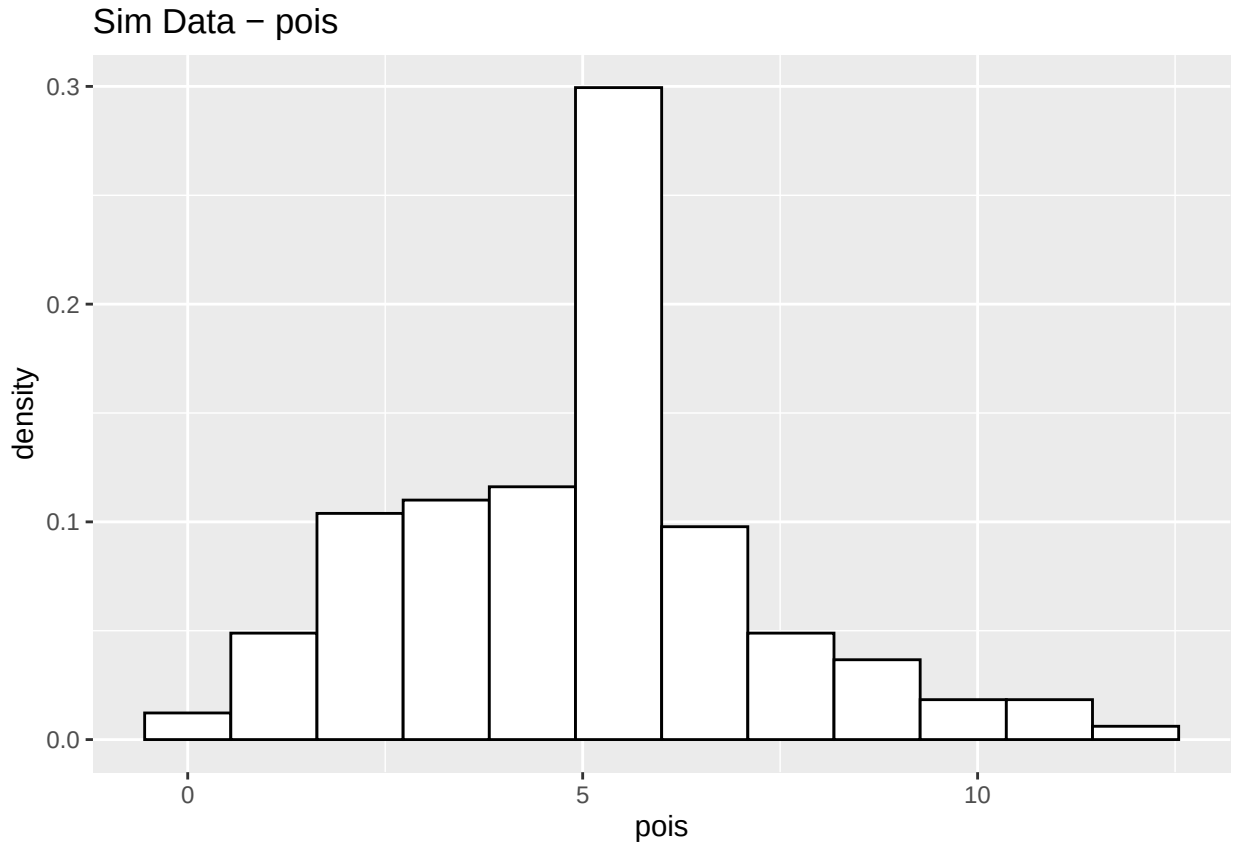
```
#2
summary(negbin)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   2.000   5.000   5.607   8.000  22.000
```

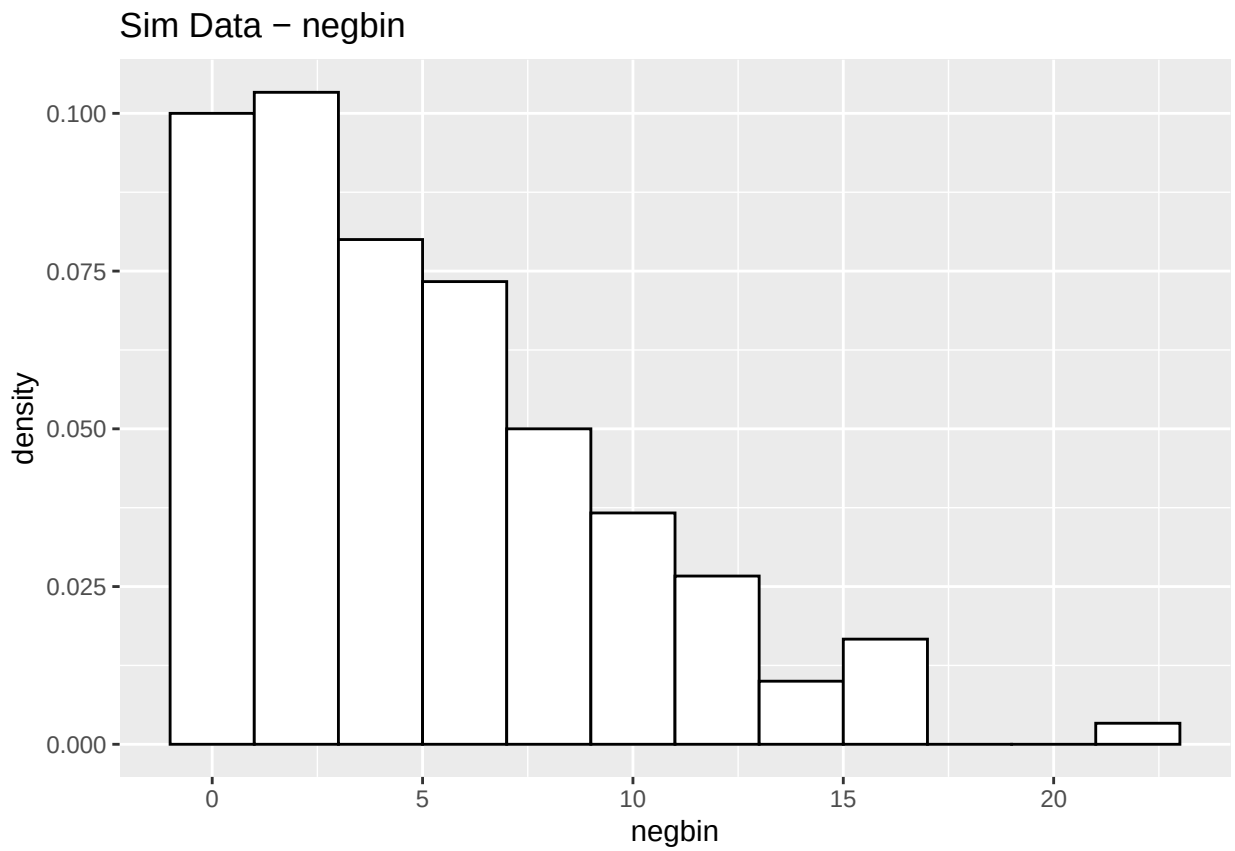
```r
#3
summary(pois0)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0     2.0     5.0     4.9     8.0    13.0
```

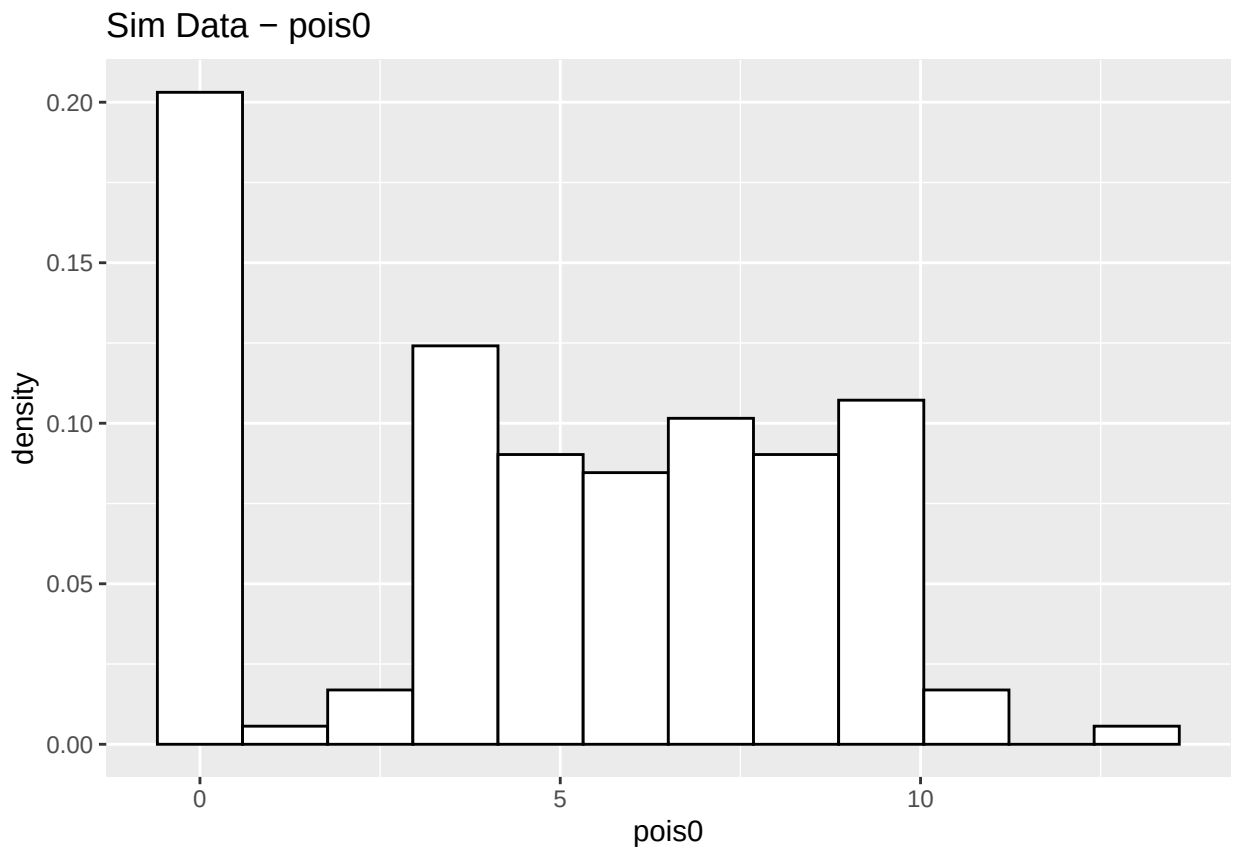
```r
#1
ggplot(data = sim.df, aes(x = pois, y = ..density..)) +
    geom_histogram(bins = 12, colour = "black", fill = "white") +
    ggtitle("Sim Data - pois")
```




```r
#2
ggplot(data = sim.df, aes(x = negbin, y = ..density..)) +
    geom_histogram(bins = 12, colour = "black", fill = "white") +
    ggtitle("Sim Data - negbin")
```




```
#3
ggplot(data = sim.df, aes(x = pois0, y = ..density..)) +
    geom_histogram(bins = 12, colour = "black", fill = "white") +
    ggtitle("Sim Data - pois0")
```



2. What are some of the differences between the three histograms? Also, what do you notice about the negative binomial simulated data? (which applies to our data analysis above)

The distribusion is the biggest difference between the three histograms. Both pois and pois0 have semi-consitant density across the plot while negbin has an decreasing density. I noticed that the negative binomial simulated data seems to be the best fit for the data with its pattern.

3. Repeat the simulation of the negative binomial data, but try changing the `size` parameter to a few different values. What does the size parameter seem to control?

```
negbin1 <- rnbinom(n, mu = 5, size = 1.0)
negbin2 <- rnbinom(n, mu = 5, size = 1.8)
negbin3 <- rnbinom(n, mu = 5, size = 2.4)

summary(negbin1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    1.00    3.00    4.98    7.00   30.00
```

```r
summary(negbin2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    2.00    3.50    4.48    6.75   20.00
```

```r
summary(negbin3)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   2.000   4.000   5.087   7.000  19.000
```

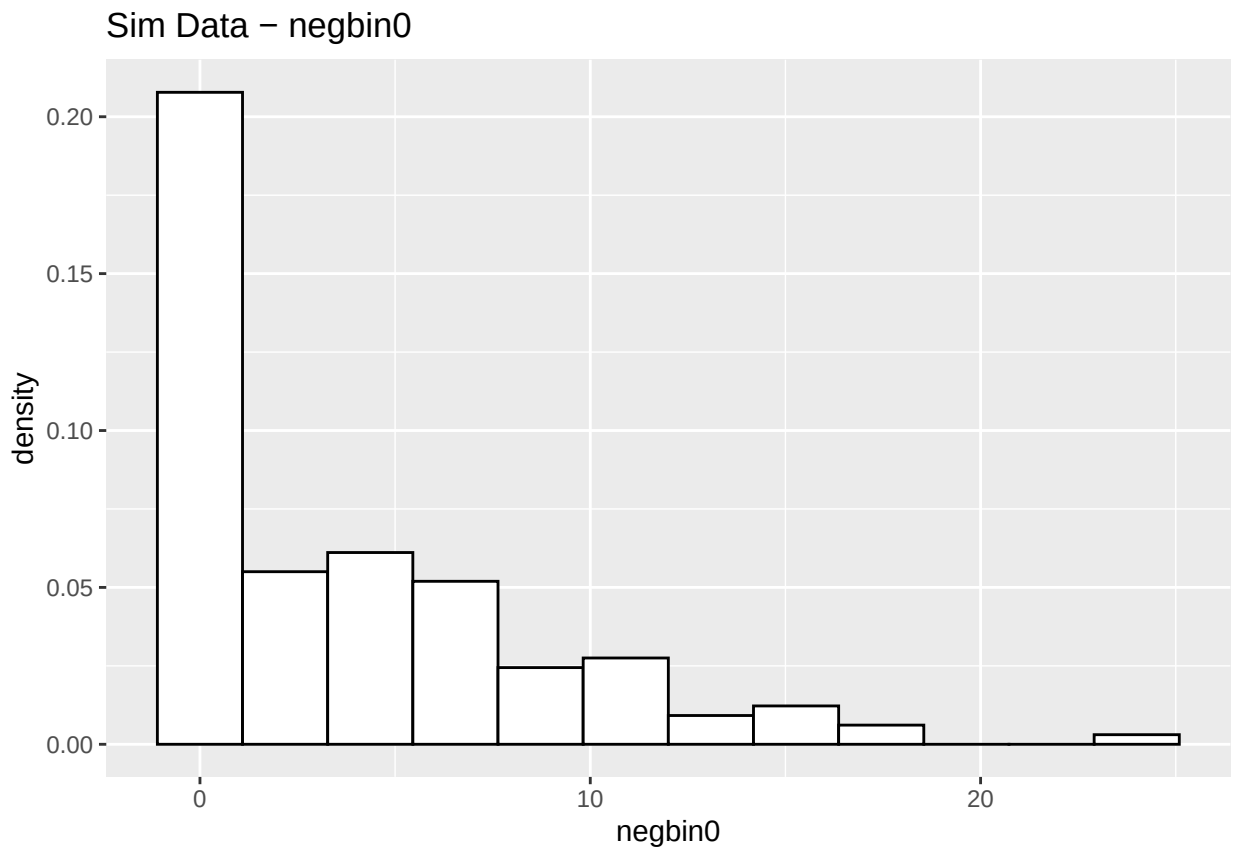
It changes the mean and max quite a bit between the different sizes. I cannot find the pattern in the the max but a smaller and larger size result in a larger max and mean.

4. Finally, simulate some data from a zero-inflated negative binomial distribution using the `rzinegbin` function in the `VGAM` package.

```r
?rzinegbin
sim.df$negbin0 <- rzinegbin(n = n, size = 1.4, munb = 5, pstr0 = 0.2)
```

Try plotting a histogram of the zero-inflated negative binomial data. Is it easy or difficult to tell based on these histograms whether data come from the poisson, negative binomial, zero-inflated poisson, or zero-inflated negative binomial models?

```r
ggplot(data = sim.df, aes(x = negbin0, y = ..density..)) +
    geom_histogram(bins = 12, colour = "black", fill = "white") +
    ggtitle("Sim Data - negbin0")
```



For me, I find it difficult based on the graphs. From the lab I would say echo that it has too many zeros to be a poisson or negative binomial. Based on that and the new graph I would say that the zero-inflated negative binomial model would be the best but I am not confident on that using the histograms.