lab 6

Ben Tankus

5/5/2021

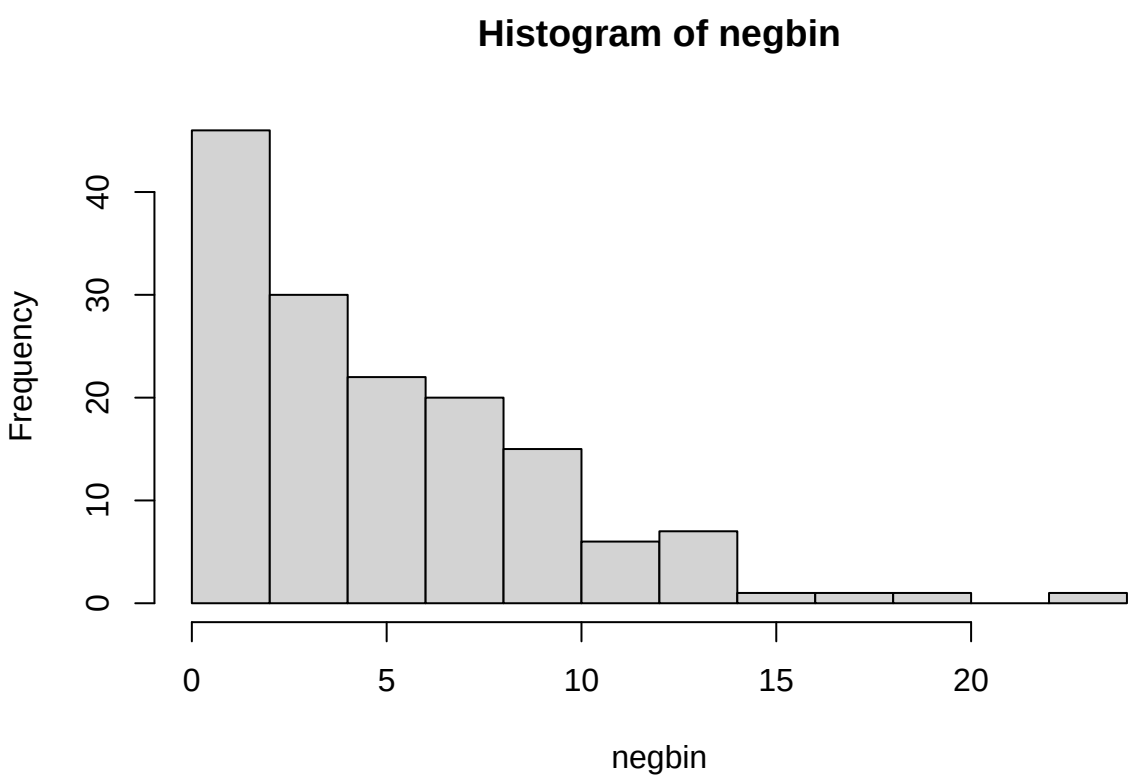
Lab Questions

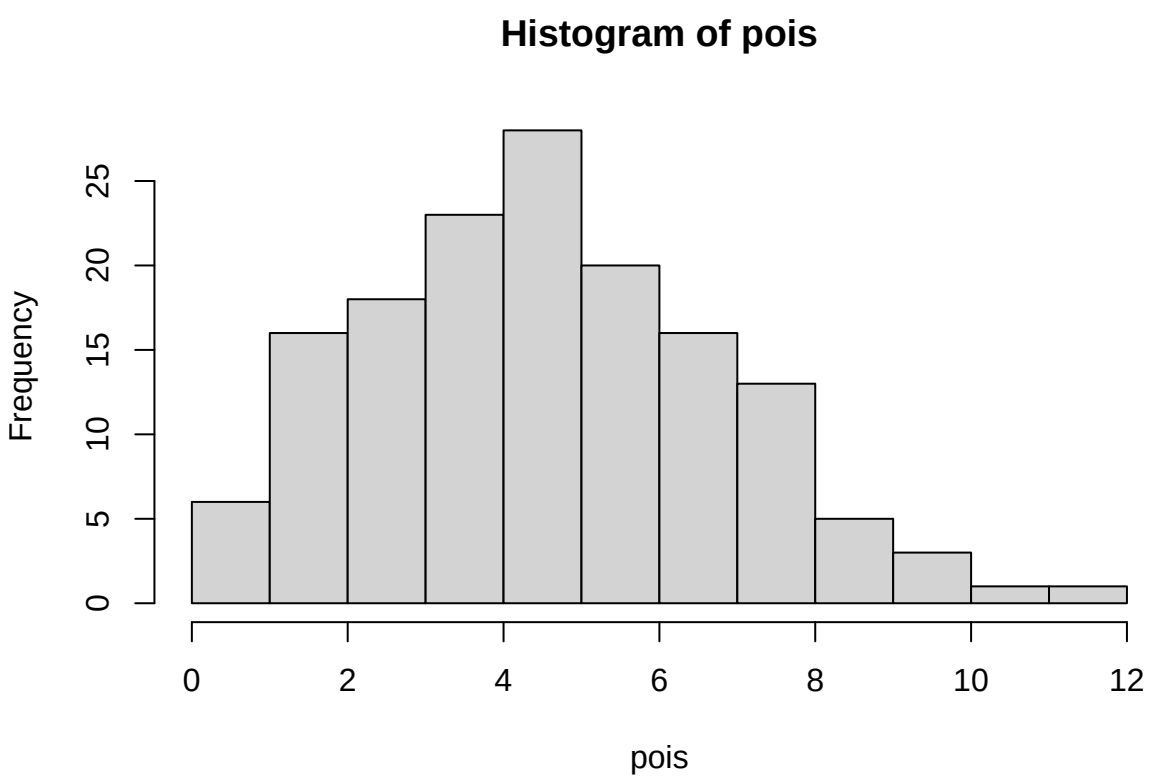
0. Compare the coefficients and their standard errors for the "extra zero" parts of the `mod.nb0` and `mod.nb.hurdle` models. Are the coefficients and standard errors similar or different? Explain.

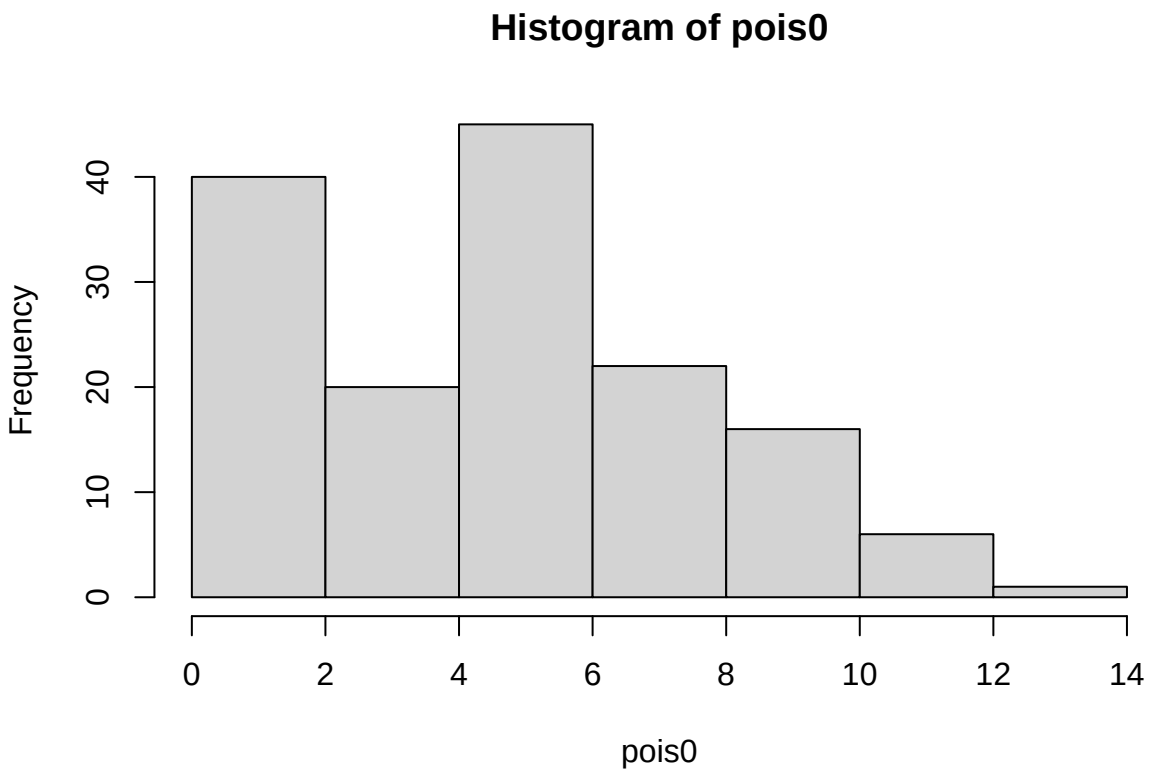
Coefficients - The coefficients are very similar for the β model, typically only 1 or 2% different. The two zero models have very different coefficients, with EthN for example, going from 14 to -2. The std error has an even larger change, going from 736 in mod.nb0 to 1 in the zero terms hurdle model. The β model again sees little change from switching methods. This is because the change in methodology only changes how zero terms are handled, not how the other terms are handled.

We saw above that, after accounting for over dispersion, there wasn't really any evidence of zero-inflation in the `quine` data set. From looking at exploratory plots of the data, however, it can be somewhat difficult to tell whether there is over dispersion, zero-inflation or both. Let's run the following code to simulate what counts would look like under a Poisson model (no zero-inflation, no over dispersion), a zero-inflated Poisson model (zero-inflation, no over dispersion), and a negative binomial model (over dispersion, no zero-inflation). We will need to install and load the `VGAM` package to easily simulate from a zero-inflated Poisson model.

```
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))



hist(negbin); hist(pois); hist(pois0)
```


Histogram of negbin
Frequency
0
10
20
30
40
0
5
10
15
20
negbin


Histogram of pois
Frequency
0
5
10
15
20
25
0
2
4
6
8
10
12
pois




```
summary(sim.df)
```

```
##       pois            negbin          pois0
##  Min.   : 0.000   Min.   : 0.00   Min.   : 0.000
##  1st Qu.: 3.000   1st Qu.: 2.00   1st Qu.: 2.000
##  Median : 5.000   Median : 4.00   Median : 5.000
##  Mean   : 5.047   Mean   : 5.42   Mean   : 4.813
##  3rd Qu.: 7.000   3rd Qu.: 8.00   3rd Qu.: 7.000
##  Max.   :12.000   Max.   :23.00   Max.   :13.000
```

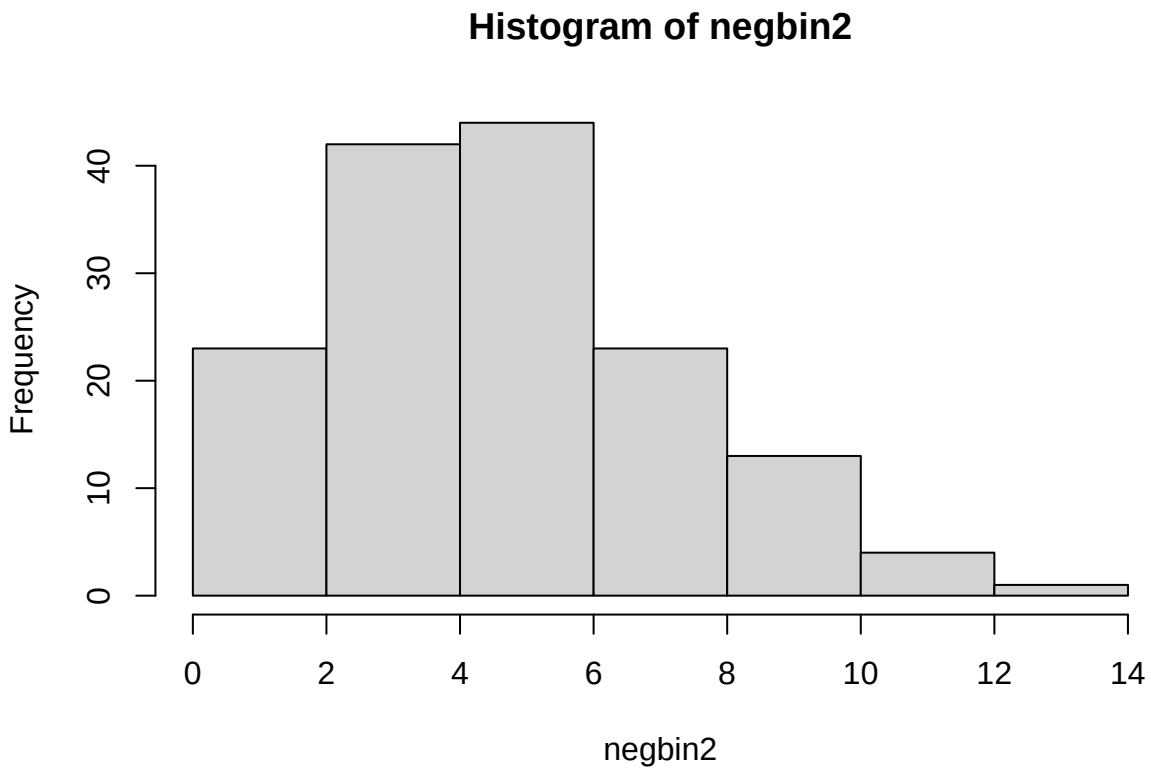
1. Run the above code and then make three histograms of the three sets of simulated data. Also, get summary statistics for each of the three sets of simulations.

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)

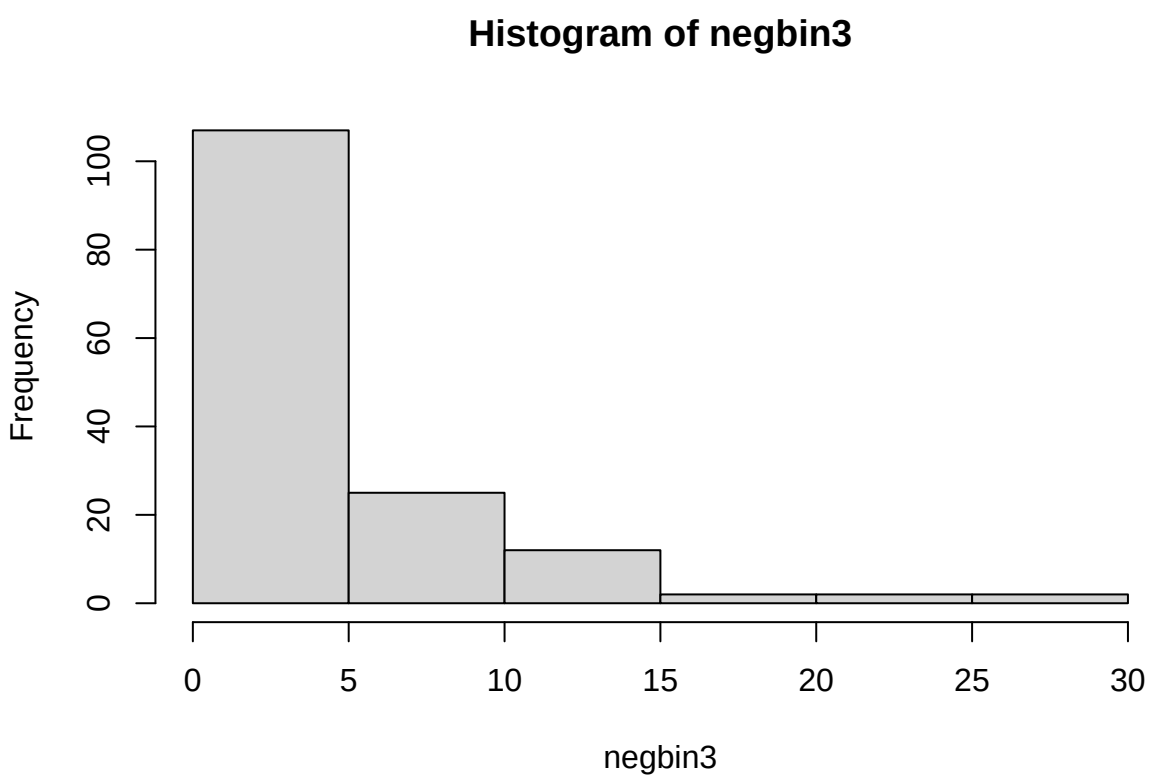
The negbin model has many zeros and appears to be almost an exponential distributuion with a much wider spread than the other two plots. the pois and pois0 plots are obviously similar, with the zero inflated model (suprise) having more zeros.

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?

```r
negbin2 <- rnbinom(n, mu = 5, size = 14)
negbin3 <- rnbinom(n, mu = 5, size = 1)

hist(negbin2)
```




```r
hist(negbin3)
```



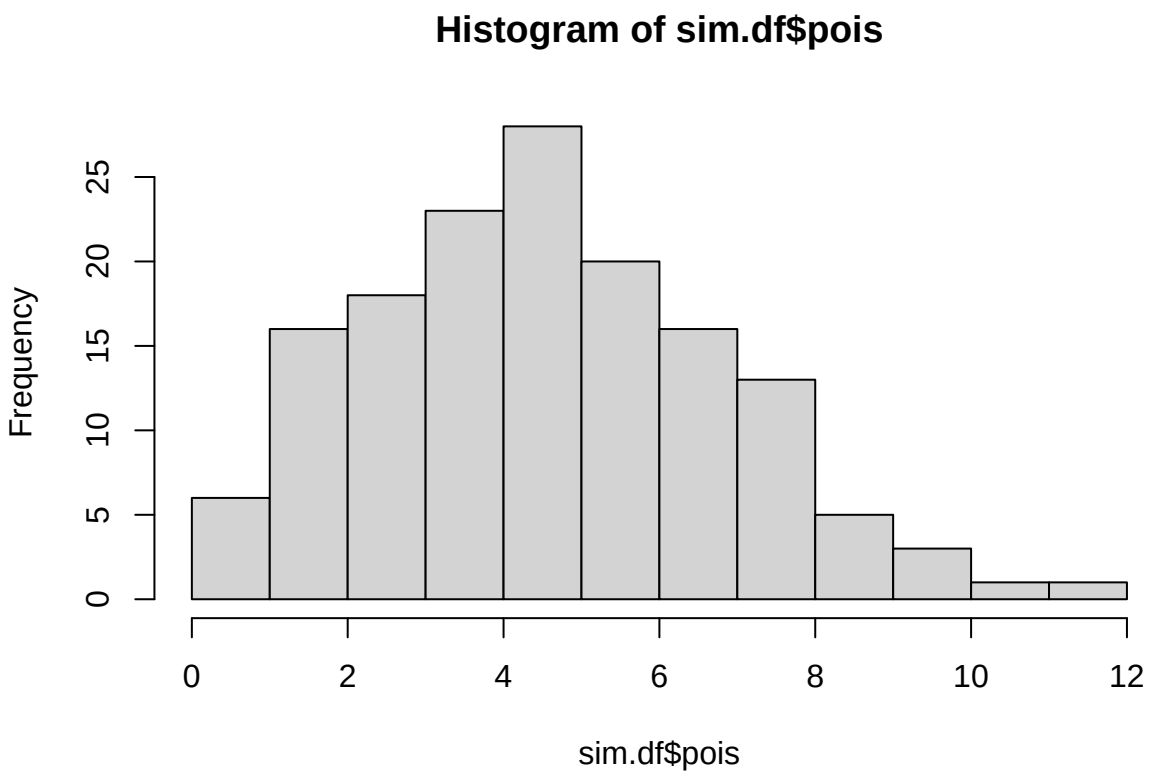

```
summary(negbin2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    3.00    5.00    5.16    7.00   13.00
```

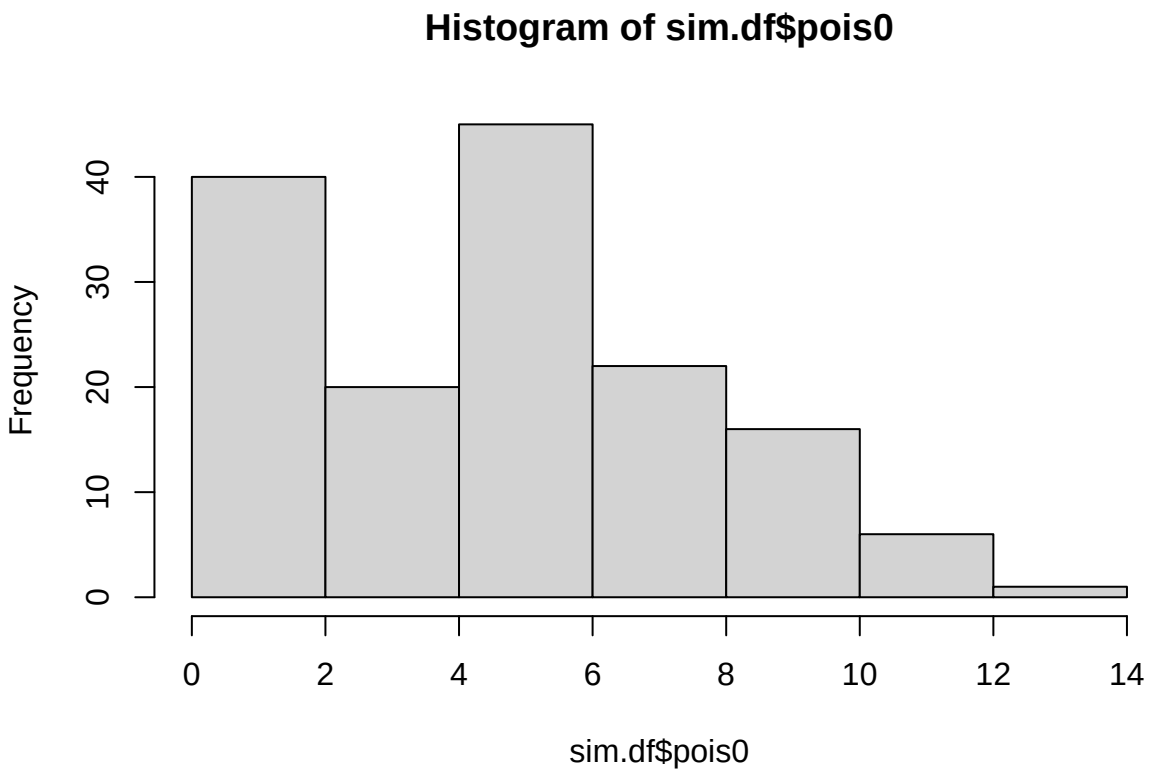
The size paramter controls the dispersion parameter. It *seems* to contol sample spread.

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

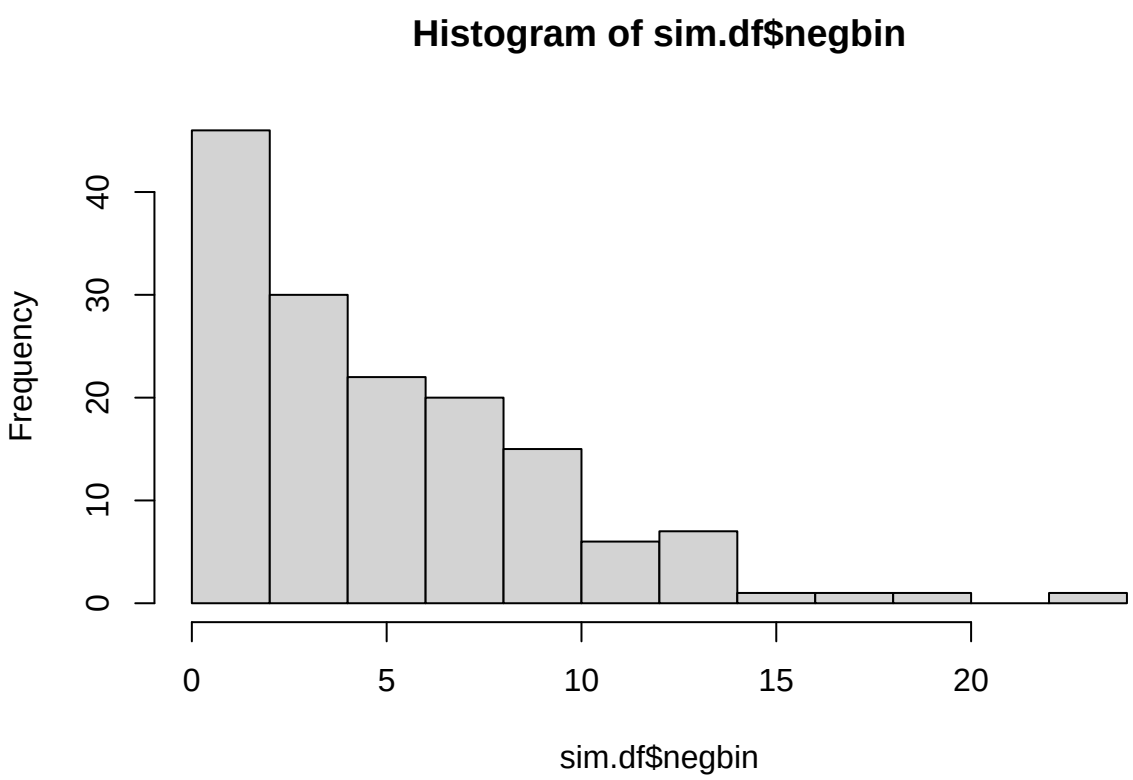
```
#?rzinegbin
sim.df$negbin0 <- rzinegbin(n = n, size = 1.4, munb = 5, pstr0 = 0.2)

hist(sim.df$pois)
```




```
hist(sim.df$pois0)
```




```
hist(sim.df$negbin)
```



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?

It's pretty easy to tell population here, but out of context it is likely a bit more difficult.