Negative binomial regression

Recap: inference with negative binomial models

```
...
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.877771   0.123477  23.306  < 2e-16 ***
## male         0.459148   0.027641  16.611  < 2e-16 ***
## age         -0.007010   0.001731  -4.050 5.12e-05 ***
## education2   0.024518   0.032534   0.754    0.451
## education3   0.009252   0.040802   0.227    0.821
## education4  -0.027732   0.044825  -0.619    0.536
## diabetes    -0.010124   0.099126  -0.102    0.919
## BMI          0.003693   0.003573   1.033    0.301
...
```

Likelihood ratio test

```
m2 <- glm.nb(cigsPerDay ~ male + age + education +
               diabetes + BMI, data = smokers)
m3 <- glm.nb(cigsPerDay ~ male + age +
               diabetes + BMI, data = smokers)
m2$twologlik - m3$twologlik
```

```
## [1] 1.423055
```

```
pchisq(1.423, df=3, lower.tail=F)
```

```
## [1] 0.7001524
```

Likelihood ratio test

> Why can I use the residual deviance to perform a likelihood ratio test for a Poisson regression model, but not for a negative binomial model?

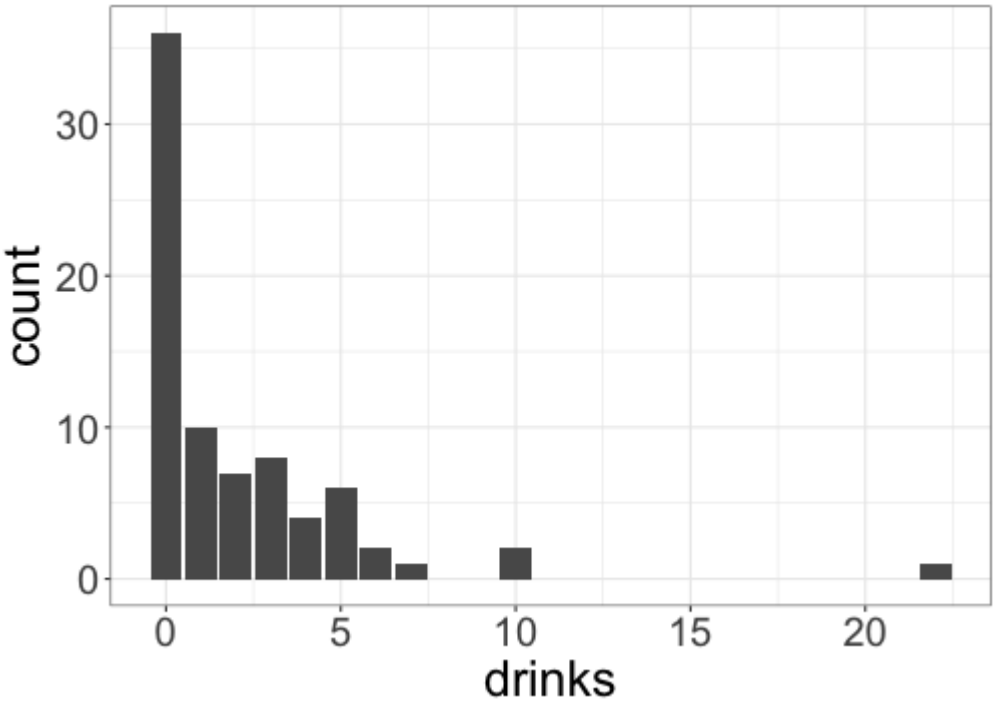
New data

Survey data from 77 college students on a dry campus (i.e., alcohol is prohibited) in the US. Survey asks students "How many alcoholic drinks did you consume last weekend?"

- `drinks`: the number of drinks the student reports consuming
- `sex`: an indicator for whether the student identifies as male
- `OffCampus`: an indicator for whether the student lives off campus
- `FirstYear`: an indicator for whether the student is a first-year student

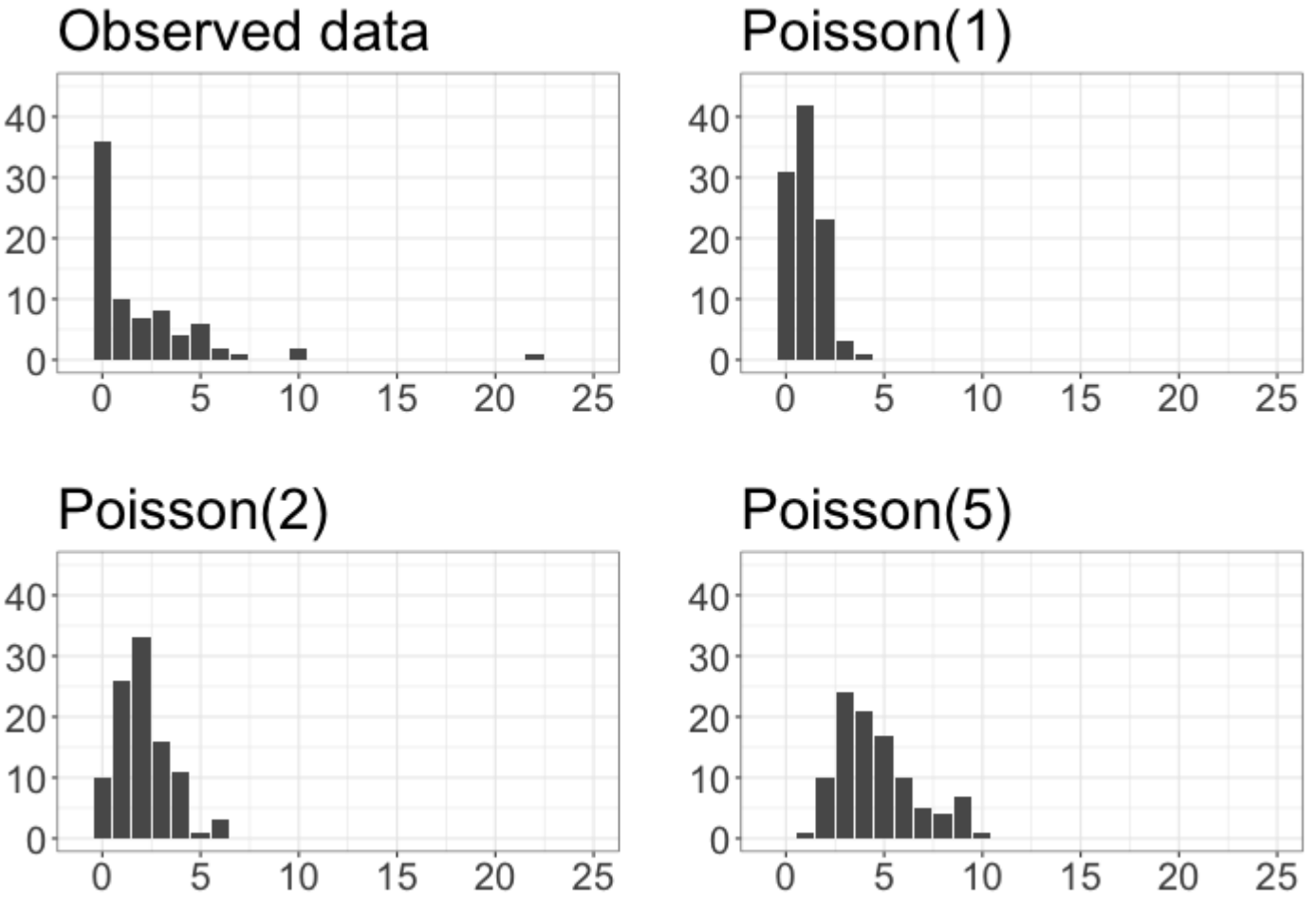
> Our goal: model the number of drinks students report consuming.

EDA: drinks



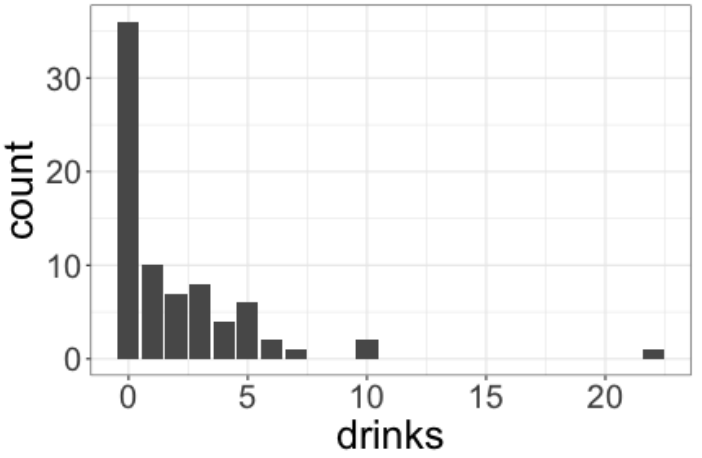
What do you notice about this distribution?

Comparisons with Poisson distributions



Excess zeros

> *Why might there be excess 0s in the data, and why is that a problem for modeling the number of drinks consumed?*



Modeling

Zero-inflated Poisson (ZIP) model

$$P(Y_i = y) = \begin{cases} e^{-\lambda_i}(1-\alpha_i) + \alpha_i & y = 0 \\ \dfrac{e^{-\lambda_i}\lambda_i^y}{y!}(1-\alpha_i) & y > 0 \end{cases}$$

where

$$\log\left(\frac{\alpha_i}{1-\alpha_i}\right) = \gamma_0 + \gamma_1 FirstYear_i + \gamma_2 OffCampus_i + \gamma_3 Male_i$$

$$\log(\lambda_i) = \beta_0 + \beta_1 FirstYear_i + \beta_2 OffCampus_i + \beta_3 Male_i$$

In R

```
library(pscl)
m1 <- zeroinfl(drinks ~ FirstYear + OffCampus + sex |
                 FirstYear + OffCampus + sex,
               data = wdrinks)
summary(m1)
```

```
...
## Count model coefficients (poisson with log link):
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)     0.8010     0.1620   4.945 7.60e-07 ***
## FirstYearTRUE  -0.1619     0.3651  -0.444   0.6574
## OffCampusTRUE   0.3724     0.2135   1.744   0.0811 .
## sexm            0.9835     0.1889   5.205 1.94e-07 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -0.39618    0.39752  -0.997    0.319
## FirstYearTRUE  0.89197    0.65878   1.354    0.176
## OffCampusTRUE -1.69137    1.47761  -1.145    0.252
## sexm          -0.07079    0.58846  -0.120    0.904
...
```

Interpretation

```
...
## Zero-inflation model coefficients (binomial with logit l
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -0.39618    0.39752  -0.997    0.319
## FirstYearTRUE  0.89197    0.65878   1.354    0.176
## OffCampusTRUE -1.69137    1.47761  -1.145    0.252
## sexm          -0.07079    0.58846  -0.120    0.904
...
```

> How would I interpret the estimated coefficient 0.892 in the logistic regression component of the model?

Interpretation

```
...
## Count model coefficients (poisson with log link):
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)      0.8010     0.1620   4.945 7.60e-07 ***
## FirstYearTRUE   -0.1619     0.3651  -0.444   0.6574
## OffCampusTRUE    0.3724     0.2135   1.744   0.0811 .
## sexm             0.9835     0.1889   5.205 1.94e-07 ***
##
...
```

> How would I interpret the estimated coefficient 0.372 in the Poisson regression component of the model?