Lecture 19

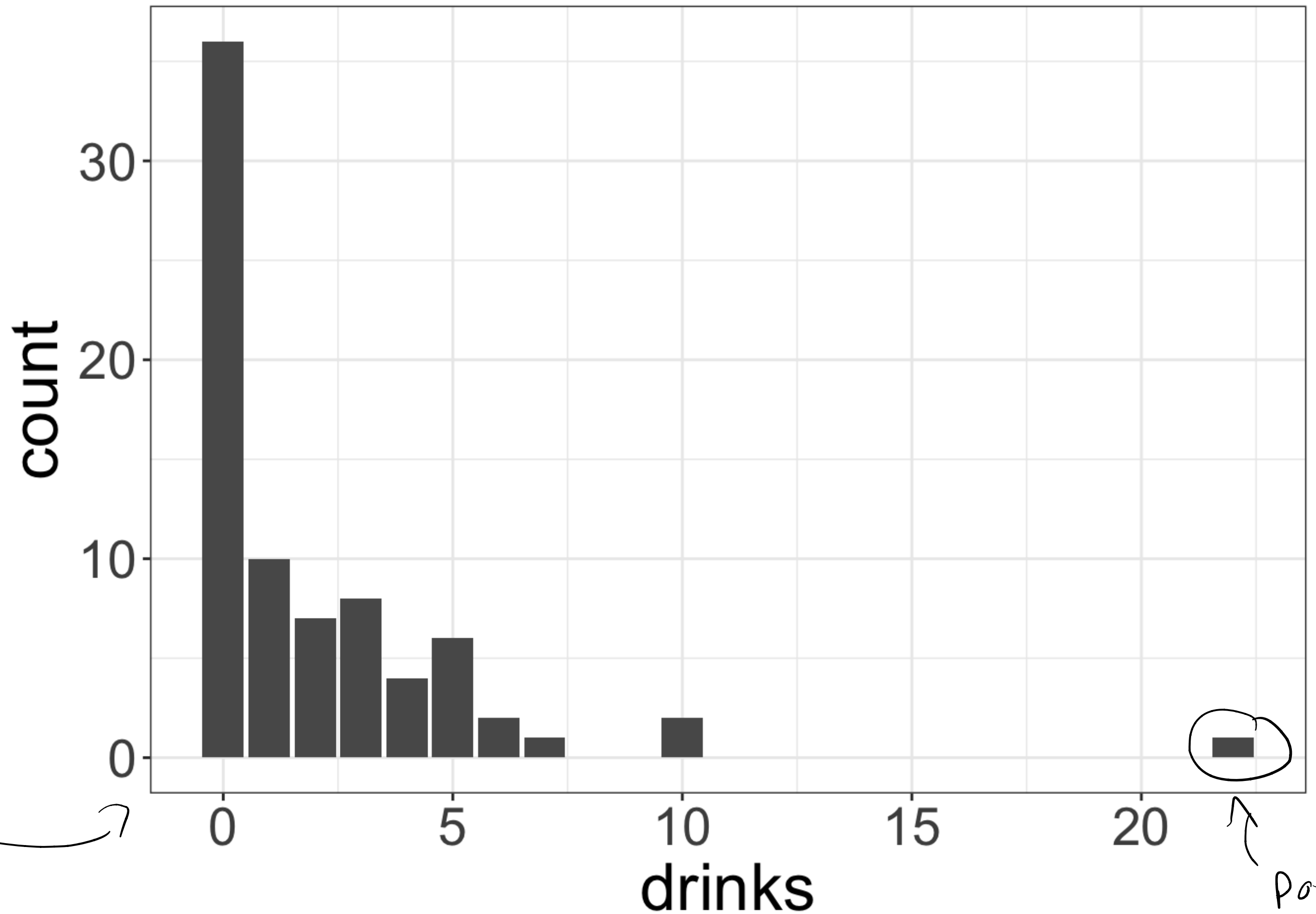
New data

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

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

EDA: drinks

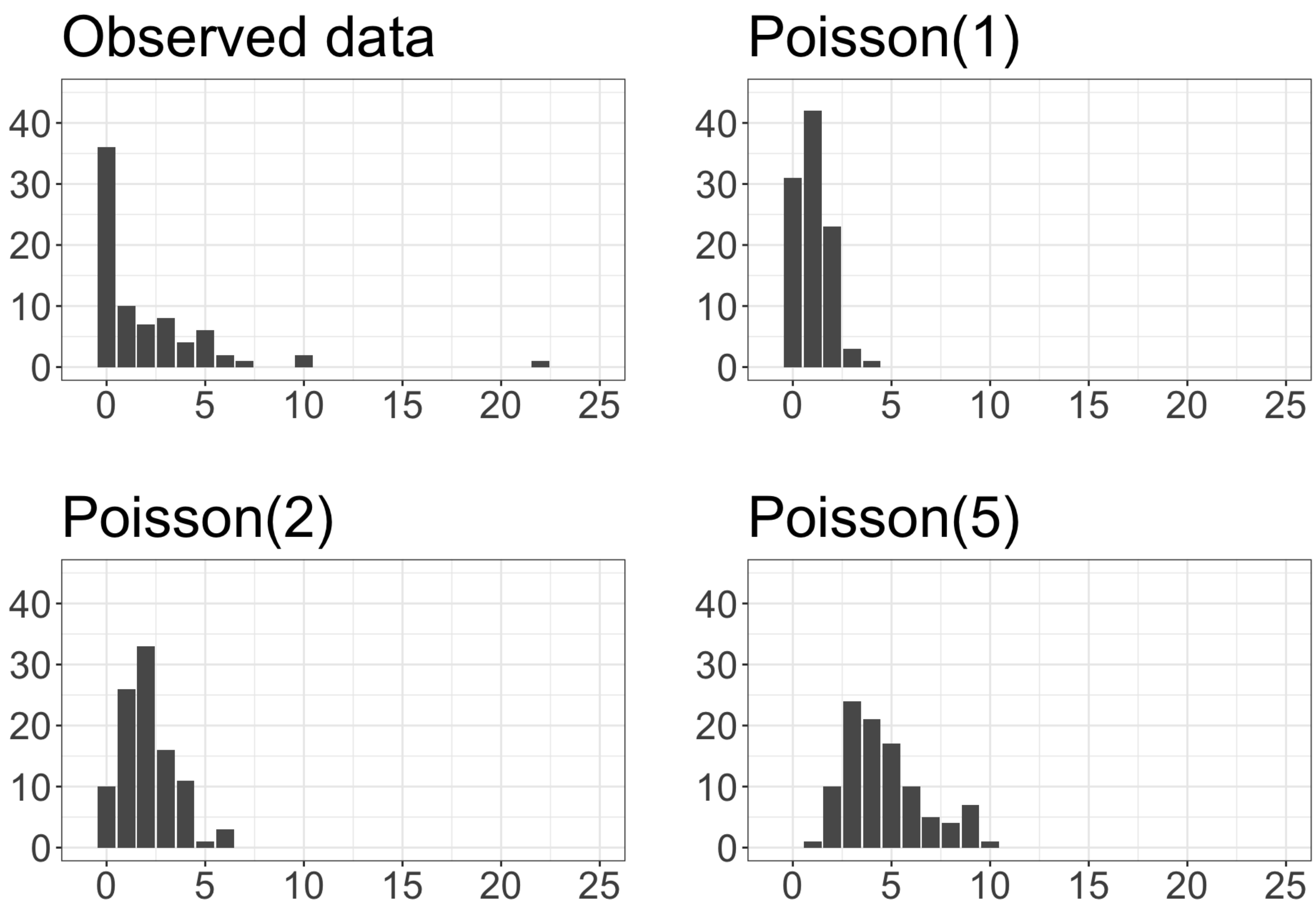


potential outlier?

many students report 0 drinks

What do you notice about this distribution?

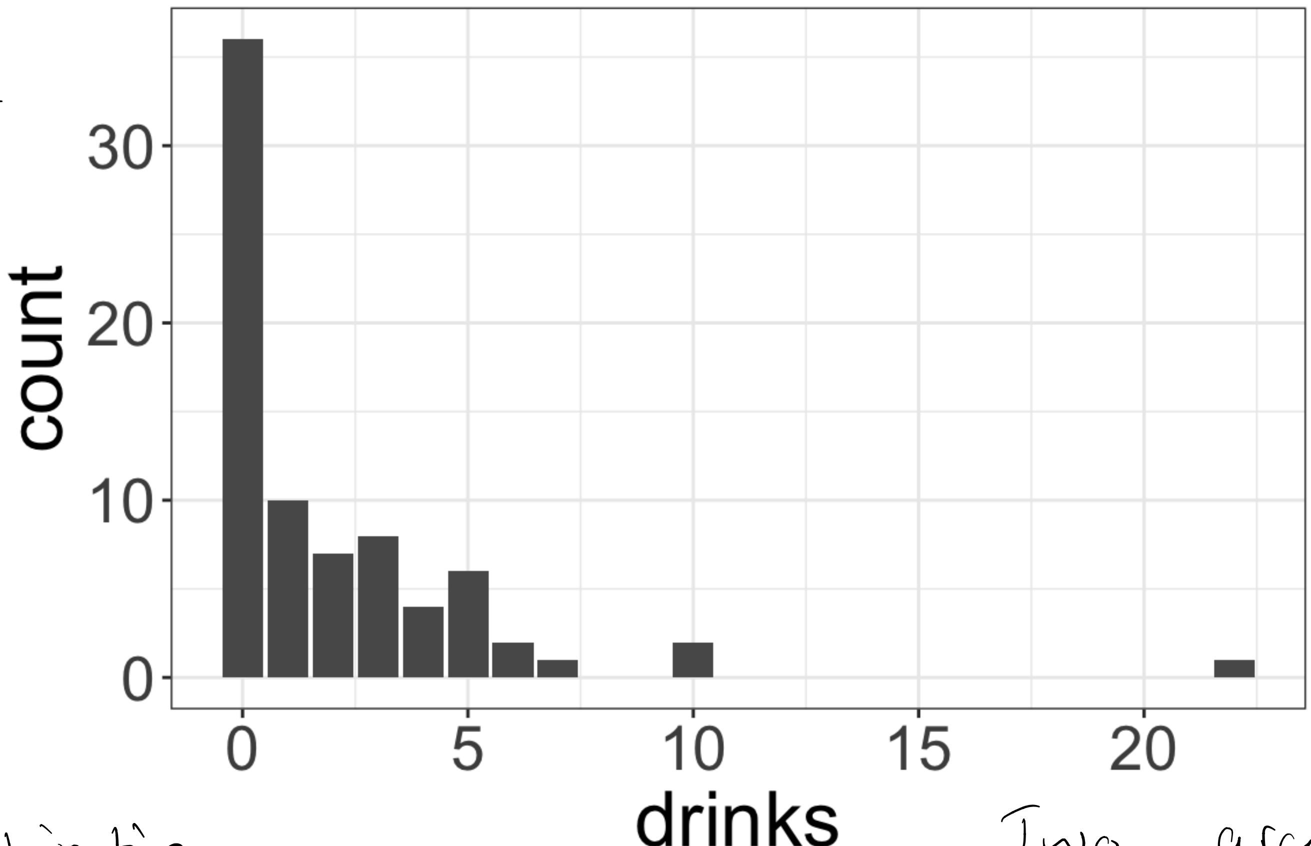
Comparisons with Poisson distributions



Excess zeros

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

Idea:
① Did a student drink?
② If so, how much?



Poisson distribution might not be a good choice
(=> problems with estimation and inference)

Two groups of students:
· drinkers
· did not drink last weekend

Hurdle models: model the zeros separately

either $Y_i = 0$ or $Y_i > 0$

$$P(Y_i > 0) = p_i$$

$$Y_i \mid (Y_i > 0) \sim \text{PosPoisson}(\lambda_i)$$

$$\left\{ \begin{array}{ll} V \sim \text{PosPoisson}(\lambda) & \text{if } \text{support}(V) = \{1, 2, 3, \ldots\} \\ P(V = v) = \dfrac{\lambda^v e^{-\lambda}}{v!\,(1 - e^{-\lambda})} & \end{array} \right\}$$

$$\Rightarrow \mathbb{E}[Y_i \mid (Y_i > 0)] = \frac{\lambda_i}{1 - e^{-\lambda_i}}$$

$$\text{var}(Y_i \mid Y_i > 0) = \frac{\lambda_i + \lambda_i^2}{1 - e^{-\lambda_i}} - \frac{\lambda_i^2}{(1 - e^{-\lambda_i})^2}$$

Y_i is a count variable

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

Hurdle model:

$$\log\left(\frac{p_i}{1-p_i}\right) = \gamma^T X_i$$

$$\log(\lambda_i) = \beta^T X_i$$

(if we want, we can also use separate explanatory variables for the two components)

$V \sim \text{PosPoisson}(\lambda)$, then

$$f(v;\lambda) = \frac{\lambda^v e^{-\lambda}}{v!\,(1-e^{-\lambda})}$$

$$= \frac{1}{v!} \exp\left(v \log \lambda + \log\left(\frac{e^{-\lambda}}{1-e^{-\lambda}}\right)\right)$$

$$= \frac{1}{v!} \exp\left(v \log \lambda - \log(e^{\lambda} - 1)\right)$$

$\phi = 1$ $\quad a(v,\phi) = \frac{1}{v!}$ $\quad \theta = \log(\lambda)$ $\quad \kappa(\theta) = \log(e^{\lambda} - 1)$

canonical link

$$\kappa(\theta) = \log(e^{\lambda} - 1) = \log(e^{e^{\theta}} - 1)$$

$$\frac{\partial \kappa(\theta)}{\partial \theta} = \frac{1}{e^{e^{\theta}} - 1} \cdot e^{e^{\theta}} e^{\theta} = \frac{e^{\lambda} \cdot \lambda}{e^{\lambda} - 1} = \frac{\lambda}{1 - e^{-\lambda}}$$

$$\underset{\sim}{I}(\beta, \gamma) = \begin{bmatrix} X^T W_\beta X & 0 \\ 0 & X^T W_\gamma X \end{bmatrix}$$

$X =$ design matrix

$0s =$ matrices of $0s$

$$W_\gamma = \operatorname{diag}(p_i(1-p_i))$$

$$W_\beta = \operatorname{diag}\left(\frac{\lambda_i e^{\lambda_i}(-\lambda_i + e^{\lambda_i} - 1)}{(e^{\lambda_i} - 1)^2} \quad p_i \right)$$

$$\underset{\sim}{I}^{-1}(\beta, \gamma) = \begin{bmatrix} (X^T W_\beta X)^{-1} & 0 \\ 0 & (X^T W_\gamma X)^{-1} \end{bmatrix}$$

Class activity

https://sta712-f23.github.io/class_activities/ca_lecture_19.html