Zero inflated models

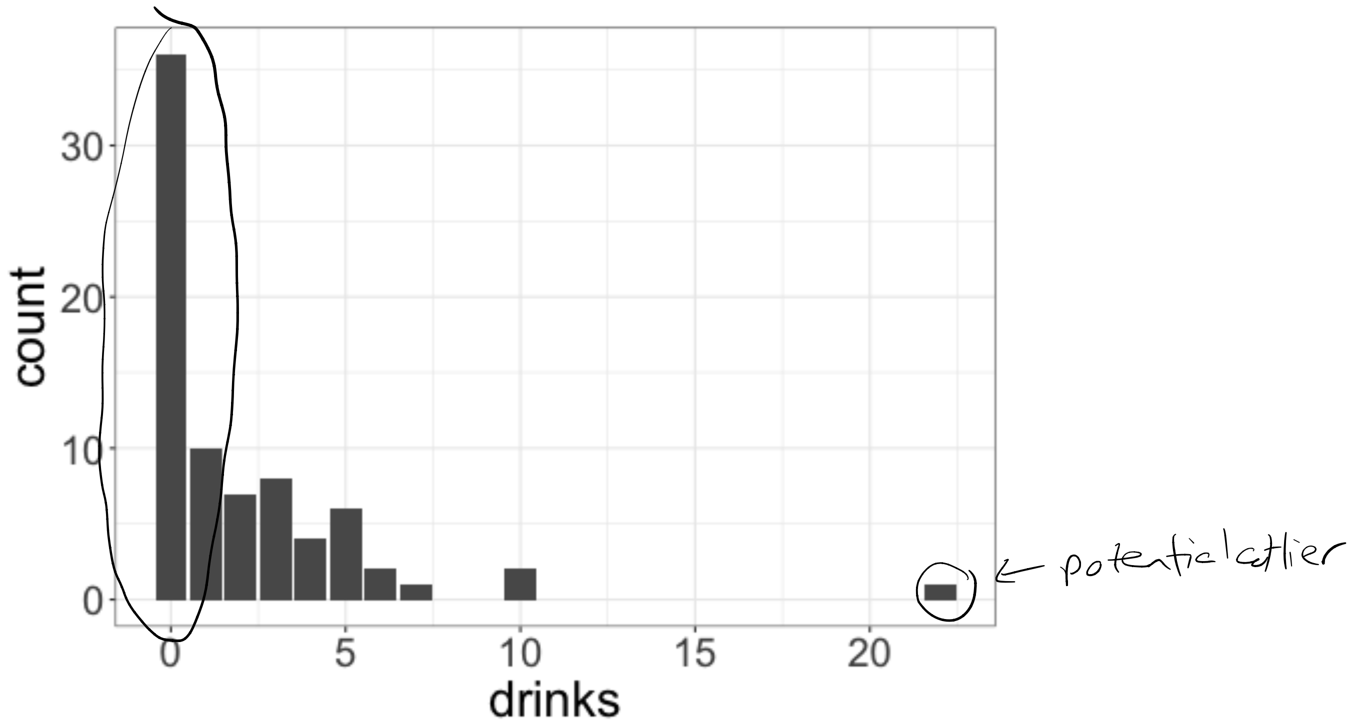
Data: College drinking

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

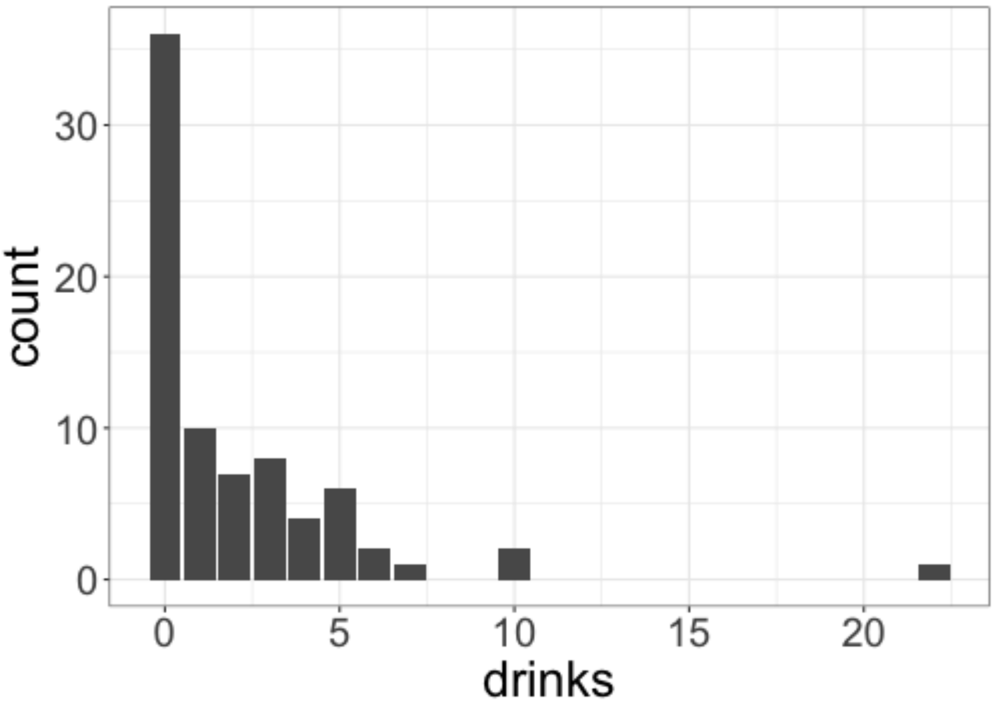


What do you notice about this distribution?

· unimodal right-skewed

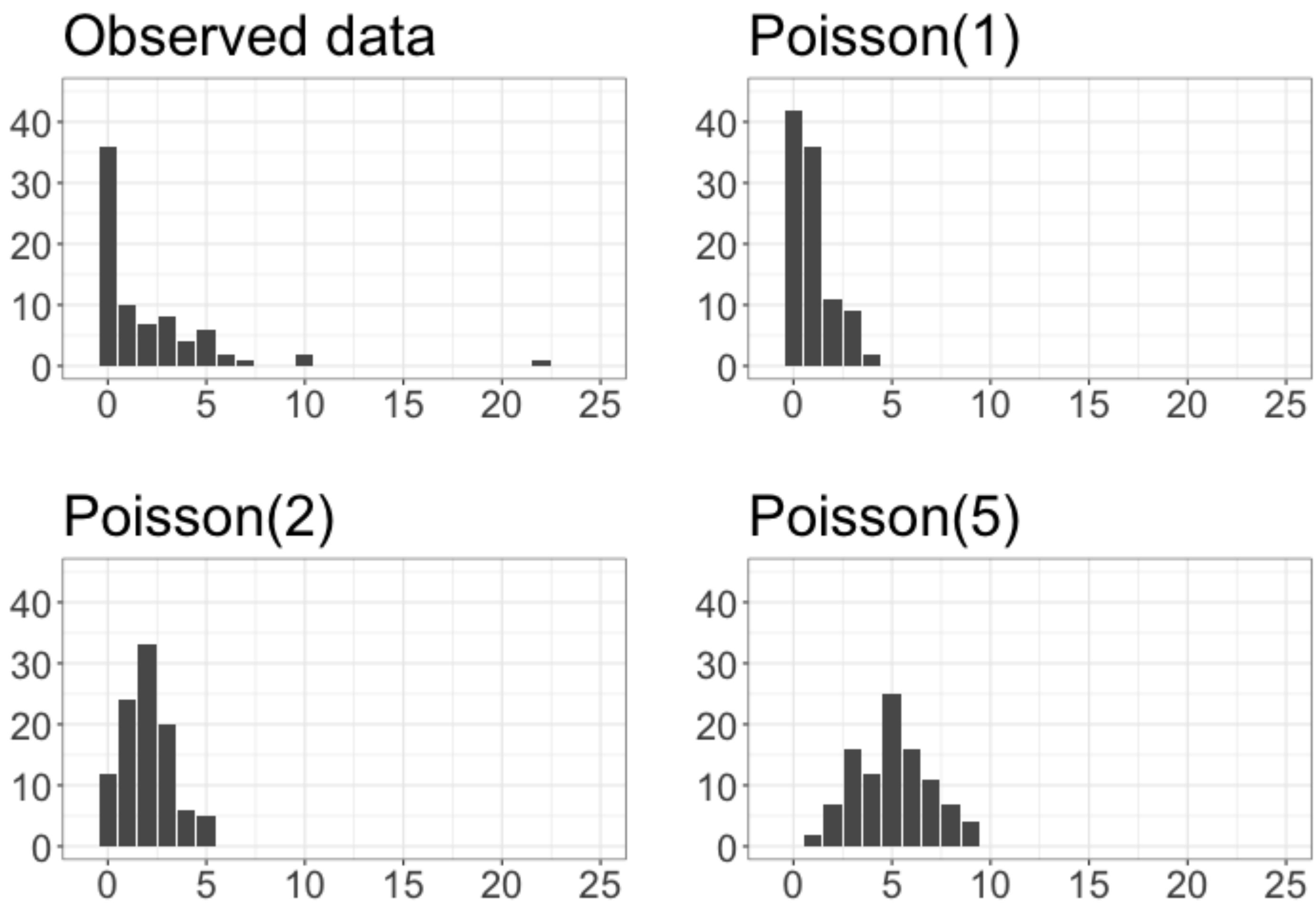
· spike at 0

EDA: drinks



- The distribution is right skewed and unimodal
- There is an outlier near 20
- *There are more zeros than we would expect from a Poisson distribution!*

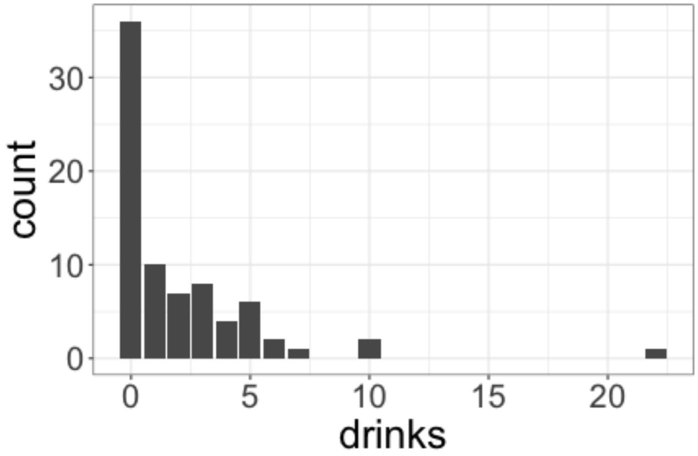
Comparisons with Poisson distributions



· too many 0s to be a Poisson sample

Excess zeros

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



Excess zeros

The problem:

- There are two groups of people contributing 0s to the data: those who never drink, and those who sometimes drink but didn't drink last weekend
- By itself, a Poisson distribution doesn't do a good job modeling data that is a mixture of these two groups

> Why don't I just include whether or not the student drinks as a variable in the model?

· If we observed this variable, we could include it

But our data doesn't include this variable

· Plan: create separate models for drinkers and for non-drinkers then combine

Modeling

Let

- Z_i denote whether student i is a non-drinker (1 = never drinks, 0 = sometimes drinks)
- $\alpha_i = P(Z_i = 1)$

(Zi is not observed in the data, but we can still imagine trying to model Zi)

We believe that α_i depends on whether or not student i is a first year.

> What model can I use for the relationship between being a first year student and being a non-drinker?

Logistic regression!

Modeling non-drinkers

Z_i denote whether student i is a non-drinker (1 = never drinks, 0 = sometimes drinks)

$$Z_i \sim Bernoulli(\alpha_i)$$

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

Modeling drinks

$Y_i =$ number of drinks consumed by student i

> If $Z_i = 1$ (the student never drinks), what is the probability of consuming 0 drinks?

$p(Y_i = 0 \mid Z_i = 1) = 1$

$y \in \{1, 2, 3, \ldots\}$

$p(Y_i = y \mid Z_i = 1) = 0$

Modeling drinks

- $Y_i =$ number of drinks consumed by student i
- Suppose that whether or not a student identifies as male and whether or not a student lives off campus has some relationship with the number of drinks consumed.

> If $Z_i = 0$ (the student sometimes drinks), how could I model Y_i ?

If $Z_i = 0$, Y_i (# drinks) is a count variable

=> Poisson distribution?

$$Y_i \mid (Z_i = 0) \sim \text{Poisson}(\lambda_i) \Rightarrow P(Y_i = y \mid Z_i = 0) = \frac{e^{-\lambda_i} \lambda_i^{y}}{y!}$$

$$\log(\lambda_i) = \beta_0 + \beta_1 \text{OffCampus}_i + \beta_2 \text{Male}_i$$

So far:

=1 non-drinker

=0 drinker

$$Z_i \sim Bernoulli(\alpha_i) \qquad \log\left(\frac{\alpha_i}{1-\alpha_i}\right) = \gamma_0 + \gamma_1 FirstYear_i$$

$$P(Y_i = 0 | Z_i = 1) = 1$$

$$Y_i | Z_i = 0 \sim Poisson(\lambda_i)$$

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

> Can we fit these models?

Not yet – we don't have Z_i

Combining models

We can calculate $\overbrace{P(Y_i = y|Z_i = 0)}^{\frac{e^{-\lambda_i}\lambda_i^{y}}{y!}}$ and $\overbrace{P(Y_i = y|Z_i = 1)}^{=\begin{cases} 1 & y=0 \\ 0 & y>0 \end{cases}}$.
Using the fact that

(law of total probability)

$$P(Y_i = y) = P(Y_i = y|Z_i = 0)P(Z_i = 0) + P(Y_i = y|Z_i = 1)P(Z_i = 1),$$

write down an equation for $P(Y_i = y)$ involving λ_i and α_i.
Hint: it will help to separate the cases $y = 0$ *and* $y > 0$

Combining models

Case 1: $y = 0$

$$P(Y_i=0) = \underbrace{P(Y_i=0 \mid Z_i=0)}_{= e^{-\lambda_i}} \underbrace{P(Z_i=0)}_{1-\alpha_i} + \underbrace{P(Y_i=0 \mid Z_i=1)}_{=1} \underbrace{P(Z_i=1)}_{=\alpha_i}$$

$$\Rightarrow P(Y_i=0) = e^{-\lambda_i}(1-\alpha_i) + \alpha_i$$

Case 2: $y > 0$:

$$P(Y_i=y) = \underbrace{P(Y_i=y \mid Z_i=0)}_{\frac{e^{-\lambda_i}\lambda_i^y}{y!}} \underbrace{P(Z_i=0)}_{1-\alpha_i} + \underbrace{P(Y_i=y \mid Z_i=1)}_{0} \underbrace{P(Z_i=1)}_{\alpha_i}$$

$$\Rightarrow P(Y_i=y) = \frac{e^{-\lambda_i}\lambda_i^y}{y!}(1-\alpha_i) \qquad y = 1, 2, 3, \ldots$$

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

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

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

> This is called a *mixture* model (it is a mixture of two different models). We *can* fit this model on the observed data (we don't need to observe Z_i)

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

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

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

> What do α_i and λ_i represent?

α_i = P(student i does not drink)

λ_i = average # of drinks consumed by a student who does drink

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

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

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

> What do α_i and λ_i represent?

$\alpha_i =$ probability the student doesn't drink, $\lambda_i =$ average number of drinks if the student *does* drink

Class activity

https://sta214-s23.github.io/class_activities/ca_lecture_25.html

Class activity: The fitted model

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

$$\log\left(\frac{\widehat{\alpha}_i}{1-\widehat{\alpha}_i}\right) = -0.60 + 1.14 FirstYear_i$$

$$\log(\widehat{\lambda}_i) = 0.75 + 0.42\, OffCampus_i + 1.02\, Male_i$$

> What is the estimated probability that a first year student never drinks? $\hat{\alpha}_i$

$\log\left(\frac{\hat{\alpha}_i}{1-\hat{\alpha}_i}\right) = -0.60 + 1.14$

$\Rightarrow \quad \hat{\alpha} = 0.63$

The fitted model

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

$$\log\left(\frac{\widehat{\alpha}_i}{1-\widehat{\alpha}_i}\right) = -0.60 + 1.14 FirstYear_i$$

$$\log(\widehat{\lambda}_i) = 0.75 + 0.42\, OffCampus_i + 1.02\, Male_i$$

> What is the estimated average number of drinks for a male student who lives off campus and sometimes drinks? $\hat{\lambda}_i$

$\hat{\lambda}_i = \exp\{0.75 + 0.42 + 1.02\} = 8.93$

The fitted model

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

$$\log\left(\frac{\widehat{\alpha}_i}{1-\widehat{\alpha}_i}\right) = -0.60 + 1.14 FirstYear_i$$

$$\log(\widehat{\lambda}_i) = 0.75 + 0.42\, OffCampus_i + 1.02\, Male_i$$

> What is the estimated probability that a male first year student who lives off campus had at least one drink last weekend?

$P(Y_i > 0) = 1 - P(Y_i = 0)$

$= 1 - (e^{-\hat{\lambda}_i}(1-\hat{\alpha}_i) + \hat{\alpha}_i)$

$= 0.37$

$\hat{\alpha}_i = 0.63$

$\hat{\lambda}_i = 8.93$