Zero-inflated Poisson Model

STAT 600B

David Yang

Table of content

Poisson distribution

The Poisson distribution is a discrete probability distribution used to model the number of times a given event occurs in a fixed time interval.

This is denoted $Y \sim Poisson(\theta)$, where θ is the average number times the event occurs in the given time interval

A real-life example of something that might follow the Poisson distribution is:

How many times does the average person visit the doctor each year in the United States?

Poisson distribution: Probability Mass Function

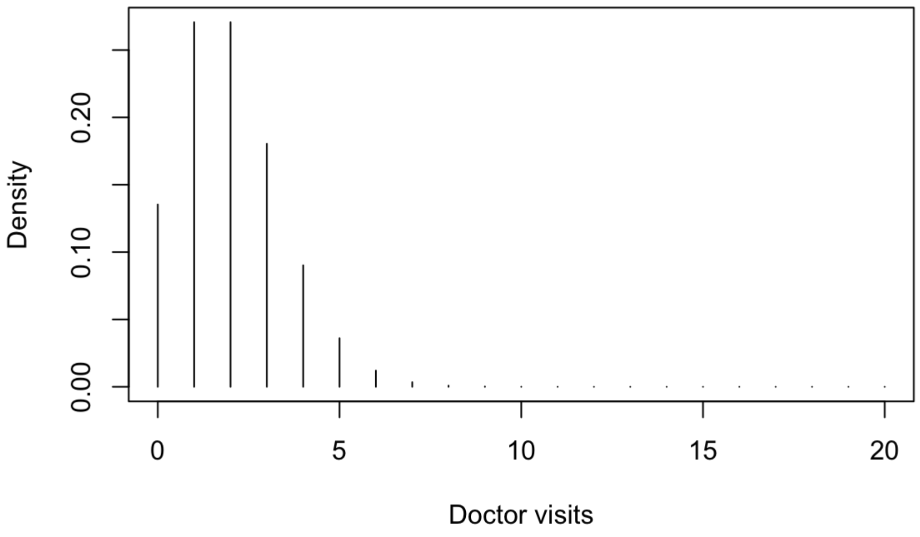
Suppose a random variable $Y1$,...,Yn follows a Poisson distribution with a mean θ then the probability mass function is,

$$f(y_i) = \prod_{i=1}^{n} \theta^{y}\, e^{-\theta}/y_i!$$

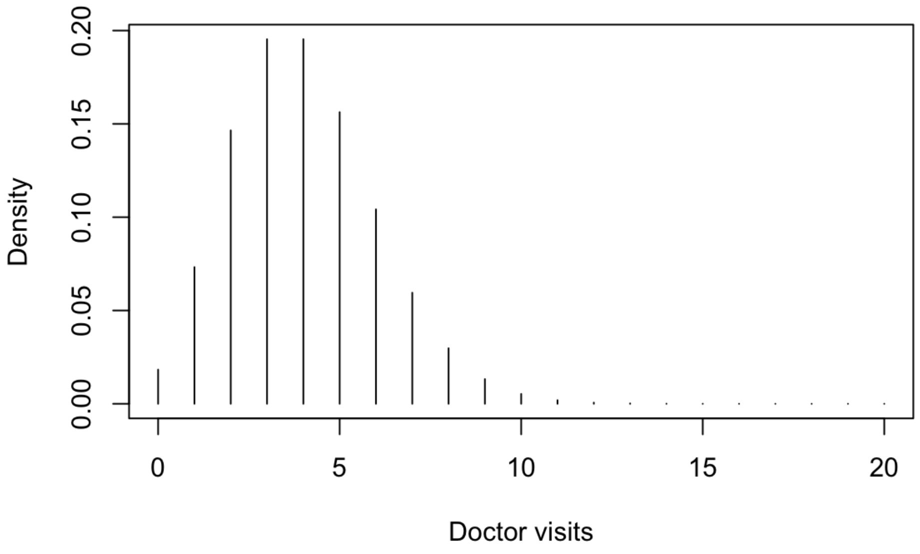
where Yi = {0, 1, 2, 3, …}, and so the likelihood of Yi|θ is

$$P(y_i|\theta) \propto \theta^{\sum_{i=1}^{n} y_i} e^{-n\theta}$$

Poisson model

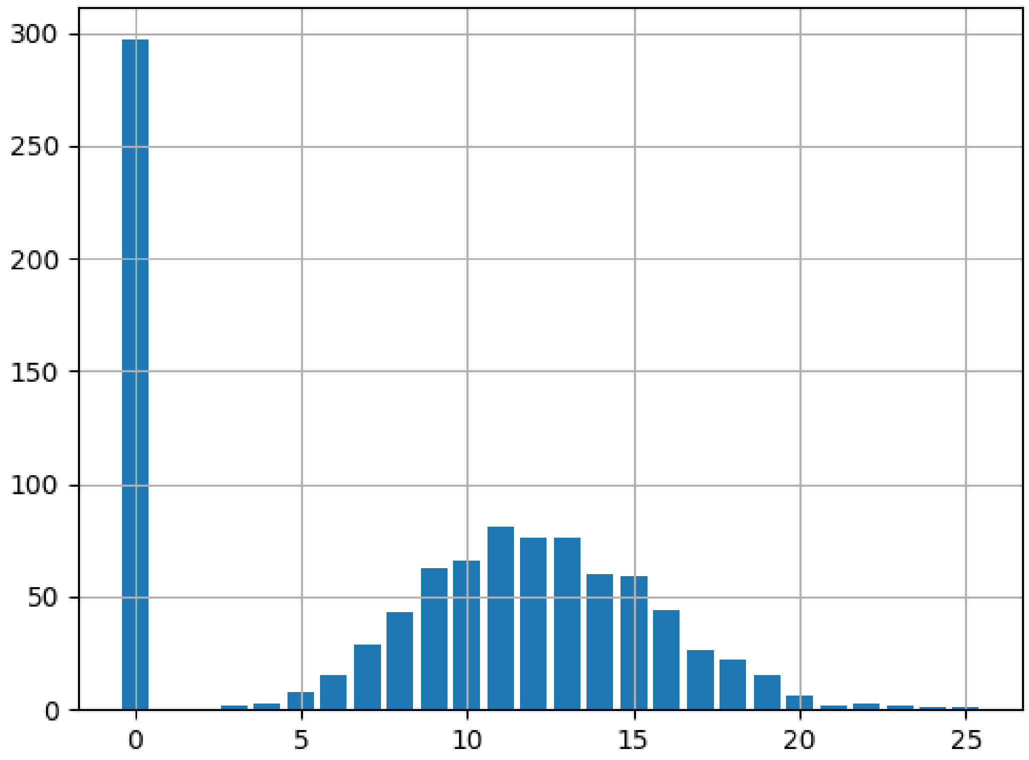


$\theta = 2$



$\theta = 4$

Problem: Excess number of zeros



Why does this happen?

Examples:

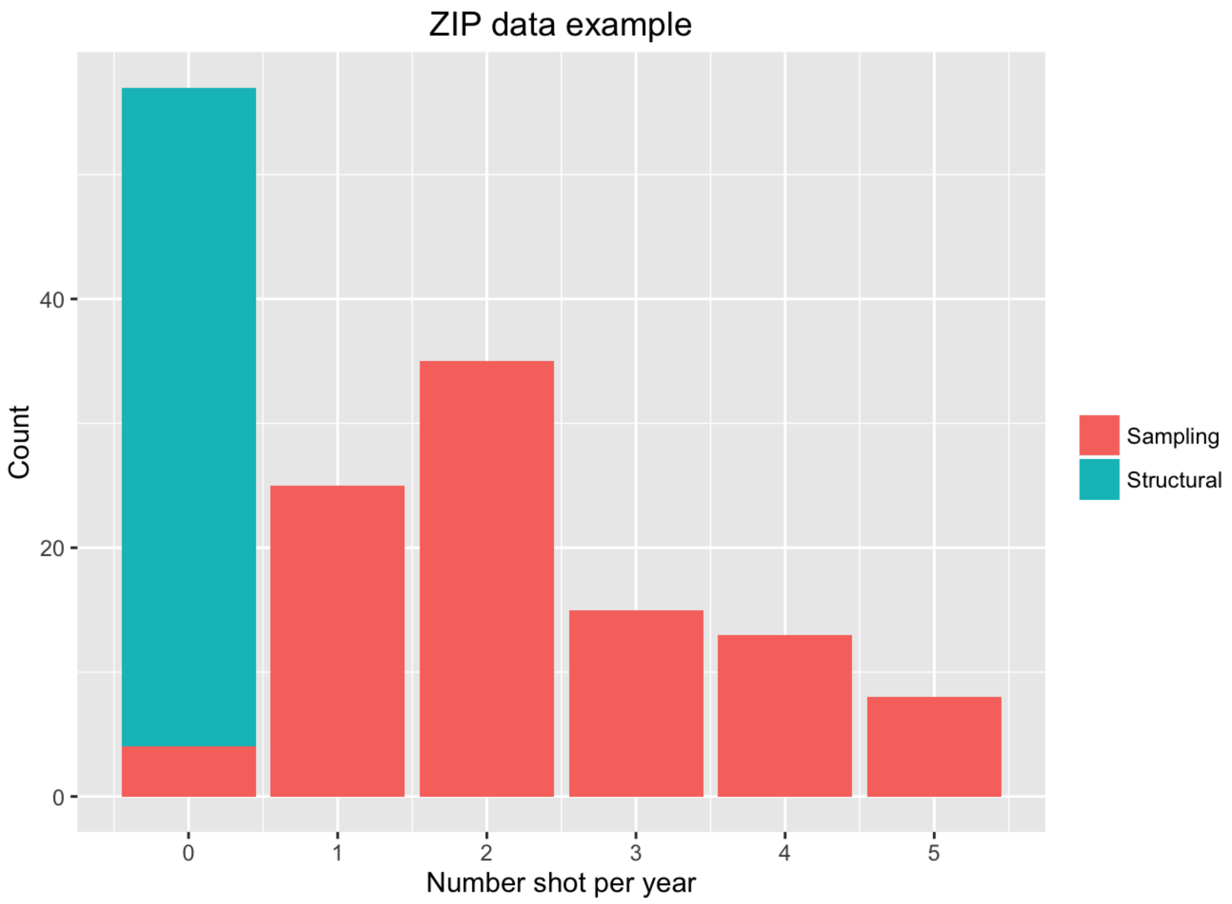
1) Average number of doctor visits per year a person makes in the United States
2) Average number of hours of television a person watches in a single day

Sampling zeros: Actual zeros that occur by chance in the probability distribution

Structural zeros: zero responses by subjects whose response will always be zero.

Example:

Number of officer-involved shootings in each county per year:



Overdispersion

Given $Y \sim Poisson(\theta)$ we assume that Var(Y) = E(Y) = θ

However, in most count datasets the variance is much greater than the mean

This is described as **over-dispersion**

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}.$$

Results in underestimation of standard error

Zero-inflated Poisson model

The zero-inflated Poisson model deals with an excess number of zeros and overdispersion

The zero-inflated Poisson model has two parts:

1. A process which generates data according to a Poisson probability mass function
2. Another underlying process which determines if the data point in the previous process is zero or non-zero

Zero-inflated Poisson model: Likelihood

Accordingly, the probability mass function of the zero-inflated Poisson model has two parts:

$$P(y_i = 0) = \pi_i + (1 - \pi_i)e^{-\mu_i}$$

$$P(y_i > 0) = (1 - \pi_i)\frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!}$$

where πi is the proportion of zeros in the data and μi is the Poisson mean

Application:

A comparison between Poisson and zero-inflated Poisson regression models with an application to number of black spots in Corriedale sheep

Hugo NAYA[1,2,3*], Jorge I. URIOSTE[2], Yu-Mei CHANG[3], Mariana RODRIGUES-MOTTA[3], Roberto KREMER[4], Daniel GIANOLA[3]

Introduction

- Studying the presence of black-brown fibres in wool of sheep which is a fault that reduces the competitiveness of the wool
- Black-brown fibres in wool can be caused by environmental factors such as urine dyeing
- Black-brown fibres can also be caused by genetic factors. For example, dark skin spots or isolated pigmented fibres
- Dark skin spots are positively correlated with black-brown fibres and age appears to be the main source of variability in the number of spots
- The goal of the paper was to study the effects of these variables on the presence of dark spots

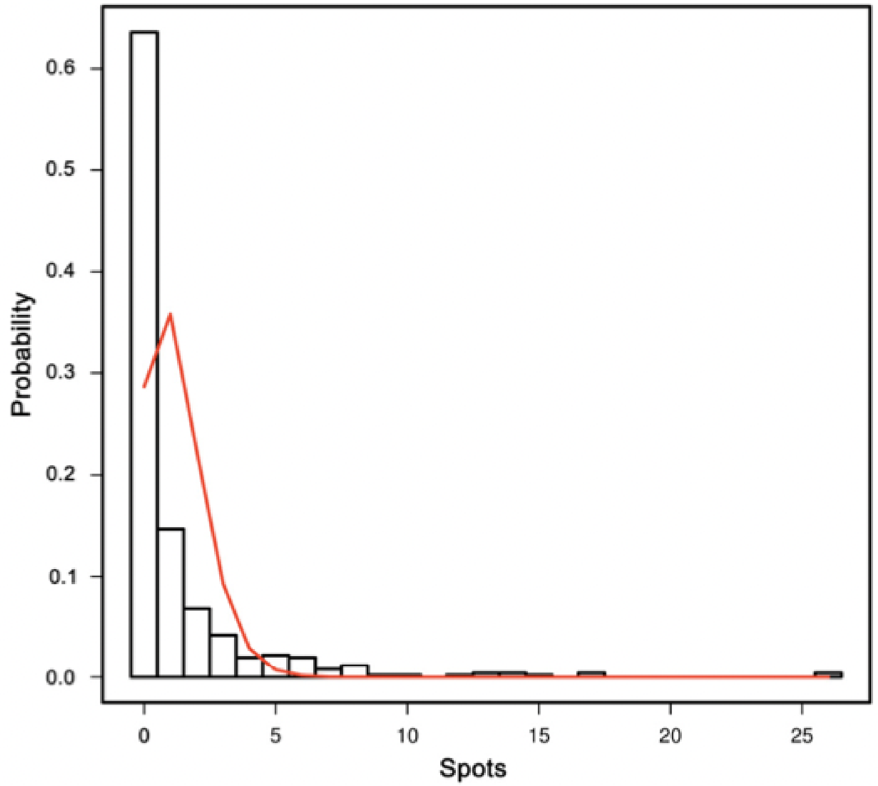


Figure 1. Distribution of the number of black spots in field data (n = 497). The solid line represents the best fit of a Poisson distribution to the observed data, fitted with package "gnlm" (http://popgen.unimaas.nl/~jlindsey/rcode.html) of *R* [26].

1) Excess number of zeros

2) Variance to mean ratio of 6.8 = overdispersion

Modelling of data

- In this study the authors modelled the data with both Poisson models and zero-inflated Poisson model

Table I. Model label, simulated data distribution given the parameters, regression function and name of each scenario (H1, H2, H3, H4).

Model	Distribution	Regression	Scenario
Z	$y_{i,j,k} \sim \mathrm{ZIP}(\theta, \lambda_{i,j})$	$\log(\lambda_{i,j}) = b_0 + b_1 \cdot \mathrm{age}_i + \mathrm{ram}_j$	H1
Ze	$y_{i,j,k} \sim \mathrm{ZIP}(\theta, \lambda_{i,j,k})$	$\log(\lambda_{i,j,k}) = b_0 + b_1 \cdot \mathrm{age}_i + \mathrm{ram}_j + e_{i,j,k}$	H2
P	$y_{i,j,k} \sim \mathrm{Poisson}(\lambda_{i,j})$	$\log(\lambda_{i,j}) = b_0 + b_1 \cdot \mathrm{age}_i + \mathrm{ram}_j$	H3
Pe	$y_{i,j,k} \sim \mathrm{Poisson}(\lambda_{i,j,k})$	$\log(\lambda_{i,j,k}) = b_0 + b_1 \cdot \mathrm{age}_i + \mathrm{ram}_j + e_{i,j,k}$	H4

The b's are unknown regressions.

Methods

Bayesian computation (gamma prior)

Parameter inference through OpenBUGS Software (MCMC 10,000 iterations)

```
model
{
  for(i in 1:N)
  {
    z[i] <- 0
    z[i] ~ dpois(phi[i])

    # likelihood
    phi[i] <- -L[i]

    # prior for p
    p[i] ~ dbeta(1,1)
    L[i] <- zero[i]*(log(p[i]+(1-p[i])*exp(-mu[i])))+(1-zero[1])*(log(1-p[i])-mu[i]+y[i]*log(mu[i])-logfact(y[i]))
    zero[i] <- equals(y[i],0)

    # prior for mu
    log(mu[i]) <- beta_1 + b[i]

    b[i] ~ dnorm(0,1)
  }
  beta_1 ~ dflat()
}
```

Conclusion

- Pe and Ze models were the most competitive in simulation
- Using deviance information criterion (DIC), the Pe model was best in most scenarios
- Ze estimated true parameters well
- With field data, parameters estimates were similar, Pe outperformed Ze under DIC

Thank you!

Question?