Theory of Statistics Likelihood Assigment

Sean Soutar STRSEA001[a], Fabio Fehr FHRFAB001[b]

[a] *UCT Statistics Honours, Cape Town, South Africa*

[b] *UCT Statistics Honours, Cape Town, South Africa*

Abstract

This project will explore the Accidents dataset and try fit a Poisson, Negative Binomial, Mixture of 2 Poissons and zero inflated Poisson models to the data. The model with the strongest support will be chosen and discussed. Profile likelihoods and confidence intervals for the parameters will be found and displayed of the chosen model.

Keywords: Likelihood, Overdispersion, Soek

JEL classification

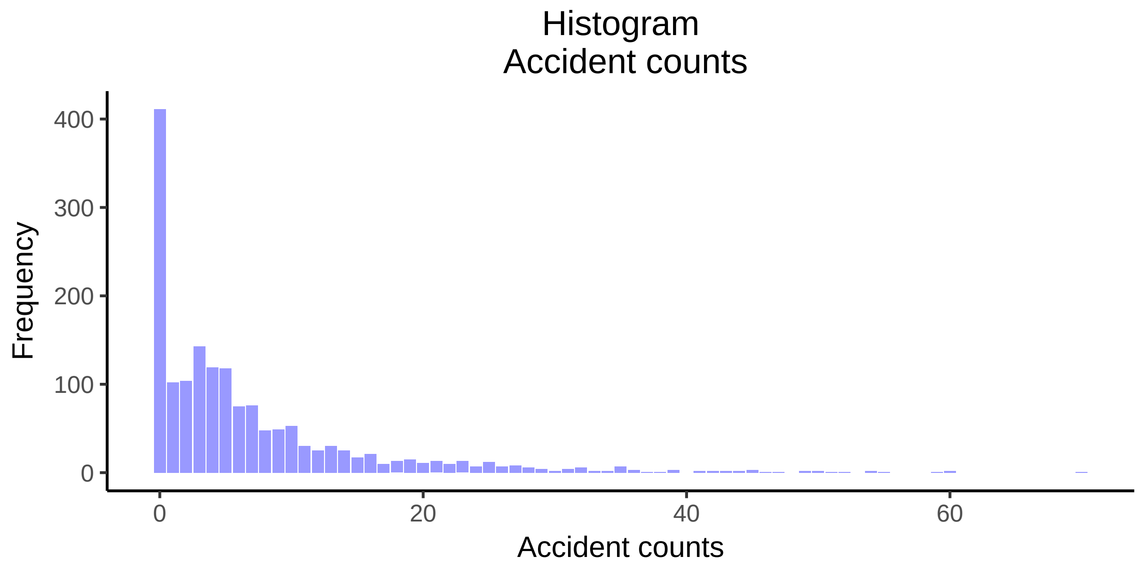
1. Introduction

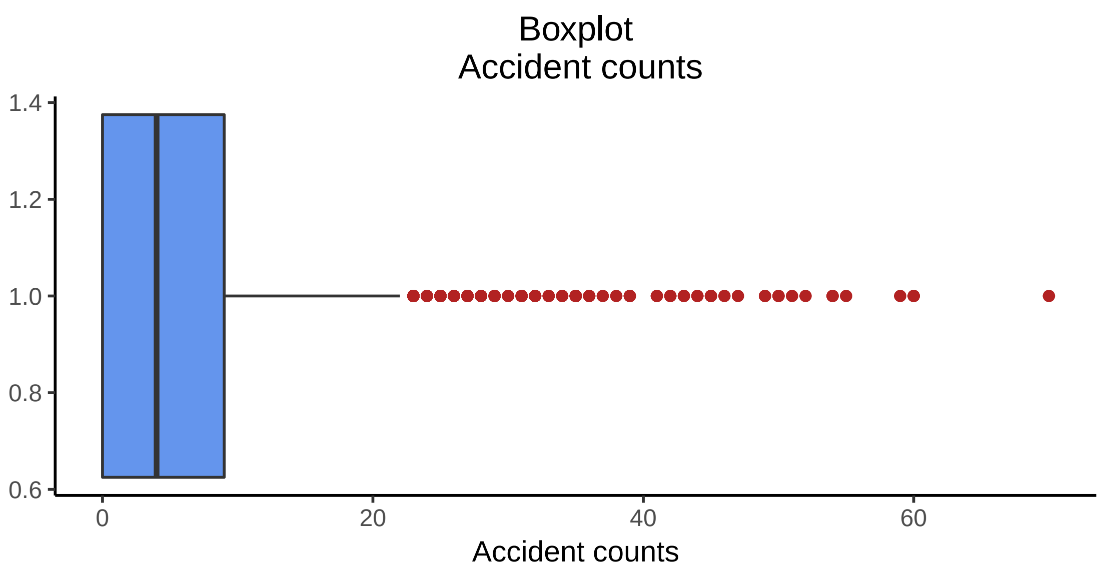
This assignment is an explorative report on a dataset containing accident counts. The aim of the report is to find and fit a model which accurately describes the accident dataset. This report will first explore the data then fit different adequate distributions and choose the most appropriate one. Once a model has been selected the profile likelihood and confidence intervals will be programmed and calculated from from first principles. The results will then be analysed critically and conclusions will be made and consider further considerations in the study.

1.1. Exploratory data analysis

To better understand our data this report shall explore the following properties; Firstly we examine the type of data within the accidents dataset and discuss whether our data is discrete ordinal or continuous. After the symmetry of the data and bounds will be discussed. This leads the exploration to outliers and extreme values.

Email addresses: sean.soutar@gmail.com (Sean Soutar STRSEA001), FHRFAB001@myuct.ac.za (Fabio Fehr FHRFAB001)





1.1.1. Data type

There are many instances where zero accidents were observed. This accounts for approximately 25.18% of the data. This suggests that the zero-inflated Poisson should be considered as this proportion is much higher than what would be expected of a regular Poisson distribution. The accident counts are discrete random variables. Specifically, they are discrete positive definite random variables on the interval $R \in \{0; +\infty\}$. Summary statistics of the data are shown below.

Mean	Variance	Median
6.917892	85.08584	4

In the Poisson distribution, the mean should equal the variance. It can be seen

1.1.2. Symmetry

This property is visually seen in the histogram and boxplot displaying the accident data. All count are greater than zero with the majority of count being below 20. The largest accident count being 70. This shows that the data is non symetrical and positively skewed.

1.1.3. Outliers

From the boxplot it clear that many outliers exist. An observation is termed an extreme value or outlier if it falls more than 1.5 times the inner-quartile range above the upper quartile. The proportion of outliers within our data set amount to 15.26% this give us reason to believe that population is also heavily skewed to the right. As aforementioned there are observations more extreme than what is displayed, which further reinforces our observation.

2. Methods

2.1. Model Formulation

Since our data is discrete, asymmetric, positive definite, contains many positive outliers and zeros this would suggest distributions such as Poisson, Negative Binomial and mixture distributions such as 2 poisson and a zero inflated Poisson.

2.1.1. Poisson

$$f(x) = e^{-\lambda}\frac{\lambda^x}{x!}, \quad x \in \{0, 1, \ldots, \infty\}, \lambda > 0$$

$$L(\lambda|x) = \prod_{i=1}^{n} f(x_i)$$

$$L(\lambda|x) = e^{-n\lambda}\frac{\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!}$$

$$l(\lambda|x) = -n\lambda + \left(\sum_{i=1}^{n} x_i\right)\ln\lambda + \ln(\prod_{i=1}^{n} x_i!).$$

-define all parameters -fit to the data

2.1.2. Negative Binomial

$$f(x) = \left(\begin{array}{c} x+j-1 \\ x \end{array}\right)(1-\pi)^x\pi^j, \quad x, \in \{0, 1, \ldots, \infty\}, 0 \leq \pi \leq 1$$

$$L(\lambda|x) = \prod_{i=1}^{n} f(x_i)$$

$$L(\lambda|x) = \prod_{i=1}^{n}\left(\begin{array}{c} x_i+j-1 \\ x_i \end{array}\right)(1-\pi)^{\sum_{i=1}^{n} x_i}\pi^{nj}$$

$$l(\lambda|x) = \sum_{i=1}^{n}\ln\left(\begin{array}{c} x_i+j-1 \\ x_i \end{array}\right) + \sum_{i=1}^{n} x_i\ln(1-\pi) + nj\ln(\pi)$$

-define all parameters -fit to the data

2.1.3. Mixture of 2 poissons

-Likelihood -define all parameters -loglikelihood

-fit to the data

-Here I am assuming the mixture will be poisson with rate = sample mean and the other poisson will have a rate of 0.1 to take into account the zero inflation ?

2.1.4. Zero inflated Poisson

-Likelihood -loglikelihood -define all parameters -fit to the data -We can use optimisers but we must program the likelihoods ourselves

2.2. Model Selection

-compare models and choose the best one -Illustrate how good the model is

-We need to reparameterize parameters so that they are unbounded

2.3. Profile Likelihood & Confidence Intervals

-Plot likelihood surface (two parameters at a time if necessary, fixing the other parameters at their MLEs).

-Must be program the profile likelihoods, CI's ourselves

3. Results

4. Conclusion

-What are the next steps and how can we improve the models

5. References