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

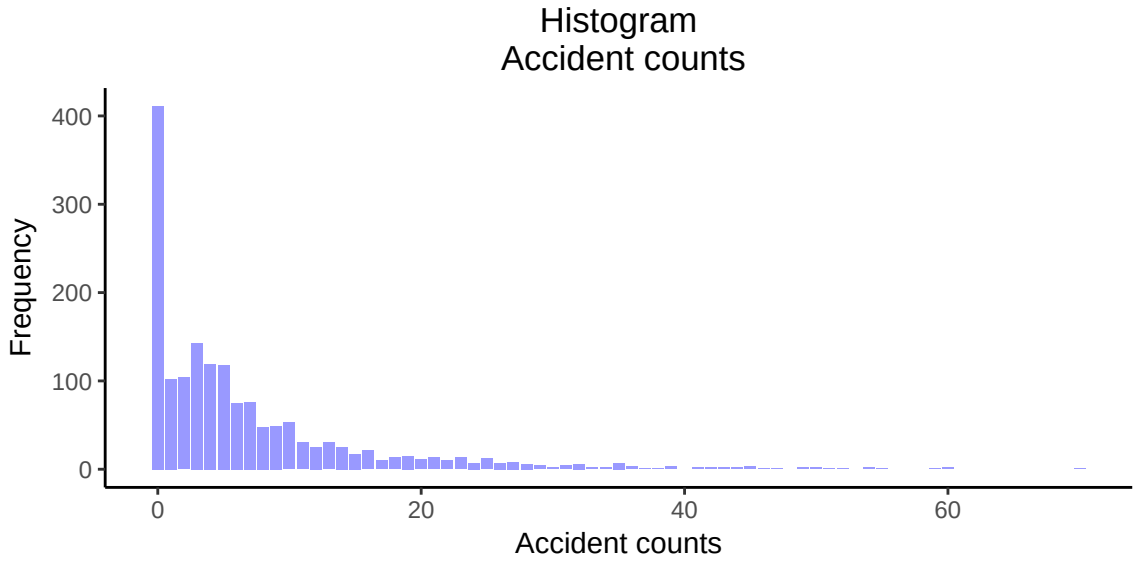
1. Introduction

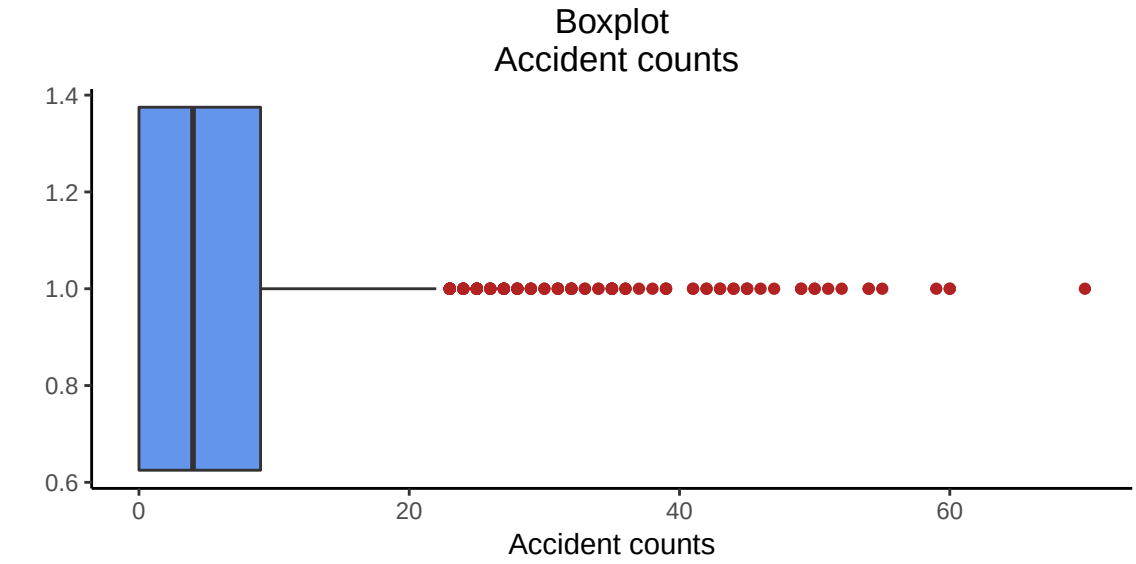
This assignment is an explorative report on a dataset containing accident counts. The aim of the report is to find and fit a model which accurately describes the accident dataset. This report will first explore the data then fit different adequate distributions and choose the most appropriate one. Once 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. The sample variance far exceeds the sample mean. This indicates overdispersion if the Poisson distribution were to be used. This is when the observations are more variable than what would be expected. This suggests that alternative count models and mixture distributions should be used.

1.1.2. Symmetry

This property is visually seen in the histogram and boxplot. All counts are greater than zero with a median value of 4 accidents. The largest accident observed is 70 accidents. THe histogram shows that the data are non-symetrical and positively skewed which is usually expected of count data.

1.1.3. Outliers

From the boxplot it clear that many outliers exist. One common method of classifying a point as an extreme value or outlier is 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%.

2. Methods

2.1. Model Formulation

The data is discrete, asymmetric, positive definite, contains many positive outliers and many zeros. This would suggest distributions such as Poisson, Negative Binomial, mixture distribution of 2 Poissons and a zero inflated Poisson.

2.2. Akiake Information Coefficient (AIC)

should we look at biC? The AIC metric can be used to compare models from different families of distributions. They can be used to compare relative goodness of fit between models. A lower AIC value indicates a better fitting model.

$\text{AIC} = -2l(\hat{\theta}) + 2\text{p}$
$p =$ Number of estimated parameters

2.2.1. Poisson

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

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

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

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

The Poisson is characterised by the λ parameter which denotes the population average rate of event occurence. In this context it would be the average number of accidents per unit time frame.

Lambda Estimate	AIC
6.917892	20263.64

2.2.2. Negative Binomial

$$p(x) = \frac{\Gamma(r+x)}{x!\,\Gamma(r)} \left(\frac{m}{r+m}\right)^x \left(\frac{r}{r+m}\right)^r \quad \text{for } x = 0, 1, 2, \ldots$$

$$L(m, r|x) = \prod_{i=1}^{n} p(x_i)$$

$$L(m, r|x) = [\frac{1}{\Gamma(r)}]^n \prod_{i=1}^{n} \frac{\Gamma(r+x_i)}{x_i!} (\frac{m}{r+m})^{\Sigma_{i=1}^{n} x_i} (\frac{r}{r+m})^{nr}$$

$$l(m, r|x) = -n \ln[\Gamma(r)] + \sum_{i=1}^{n} \ln(\Gamma(r+x_i)) - \sum_{i=1}^{n} \ln x_i! + \sum_{i=1}^{n} x_i \ln(\frac{m}{r+m}) + nr \ln(\frac{r}{r+m})$$

This parameterisation of the negative binomial is characterised by the mean parameter m and the shape parameter r. It is important to note that the variance of a Negative Binomial under this parameterisationg is $m + \frac{m^2}{r}$. Shape parameters are often regarded as nuisance parameters and do not play a meaningful role in maximising likelihood. Therefore, since we desire no under or over dispersion, we can express the shape parameter as a function of the mean parameter to be esimated and the sample variance.

CHECK THIS! LOOK AT PAPER CALLED ESTIMATING SHAPE. I HAVE HIGHLIGHTED SOME STUFF ON SECOND PAGE

$$Var(x) = m + \frac{m^2}{r}$$
$$r = \frac{m^2}{Var(x) - m}$$
$$\hat{r} = \frac{\hat{m^2}}{S^2 - \hat{m}}$$

M Mean Estimate	r Shape Estimate	AIC
6.810803	0.592616	9626.917

2.2.3. Mixture of 2 Poissons

A finite mixture distributon of two Poisson variables will now be explored. A possibly reason for the overdispersion is that the data are from two separate Poisson distributions. Since it is not known from which distribution that any given data point is from, presuming that the two distribution mixture is appropriate, an additional mixing parameter p needs to be estimated.

$$p(x|\lambda_1, \lambda_2, p) = p\frac{e^{-\lambda_1}\lambda_1^x}{x!} + (1-p)\frac{e^{-\lambda_2}\lambda_2^x}{x!}, \quad x \in \{0, 1, \ldots, \infty\}, \lambda_1, \lambda_2, p > 0$$

$$L(\lambda_1, \lambda_2, p|x) = \prod_{i=1}^{n} p(x_i)$$

$$L(\lambda_1, \lambda_2, p|x) = \prod_{i=1}^{n} p\frac{e^{-\lambda_1}\lambda_1^{x_i}}{x_i!} + (1-p)\frac{e^{-\lambda_2}\lambda_2^{x_i}}{x_i!}$$

$$l(\lambda_1, \lambda_2, p|x) \sum_{i=1}^{n} \ln[p\frac{e^{-\lambda_1}\lambda_1^{x_i}}{x_i!} + (1-p)\frac{e^{-\lambda_2}\lambda_2^{x_i}}{x_i!}]$$

The parameters λ_1 and λ_2 refer to the average number of road accidents per road stretch for the first and second distribution respectively. The parameter p is the mixing parameter. This represents the probability that a given observation belongs to distribution 1. Therefore, the probability that an observation belongs to distribution 2 is the $1 - p$ quantity.

Lamda 1 estimate	Lambda 2 estimate	Mixing Parameter p	AIC
19.38094	2.84069	0.2465036	12075.12

2.2.4. Zero inflated Poisson

The Zero Inflated Poisson is also a finite mixture distribution. This model supposes that the data can come from two distributions. The one is a Zero Process and the other is a Poisson process that can only take on non-zero values. This model is useful if there are many zeroes in the data. This was seen to be the case as discussed in 1.1.1. The Zero Inflated Poisson is a piecewise defined distribution with different mass functions for predicting the probability that a given observation will be zero rather than non-zero.

$$p(x_i = 0) = \pi + (1 - \pi)e^{-\lambda}$$

$$p(x_i \neq 0) = (1 - \pi)\frac{\lambda^{x_i} e^{-\lambda}}{x_i!}, \qquad x_i \geq 1$$

$$L(\lambda, \pi|x) = L(\lambda, \pi|x = 0)L(\lambda, \pi|x \neq 0)$$

An indicator variable I is defined.

$$I = \begin{cases} 0 & x = 0 \\ 1 & x \neq 0 \end{cases}$$

$$L(\lambda, \pi|x) = \prod_{i=1}^{n} p(x_i = 0)^{1-I} p(x_i \neq 0)^I$$

$$L(\lambda, \pi|x) = \prod_{i=1}^{n} [\pi + (1 - \pi)e^{-\lambda}]^{1-I} [(1 - \pi)\frac{\lambda^{x_i} e^{-\lambda}}{x_i!}]^I$$

$$l(\lambda, \pi|x) = \sum_{i=1}^{n} \ln[(1 - I)[\pi + (1 - \pi)e^{-\lambda}] + I[(1 - \pi)\frac{\lambda^{x_i} e^{-\lambda}}{x_i!}]]$$

The parameter λ is the average rate of accidents per road stretch. The parameter π is the probability of extra zeroes in the data. MAYBE EXPLAIN THIS BETTER?

Lamda estimate	Pi estimate	AIC
9.245627	0.2517669	15556.16

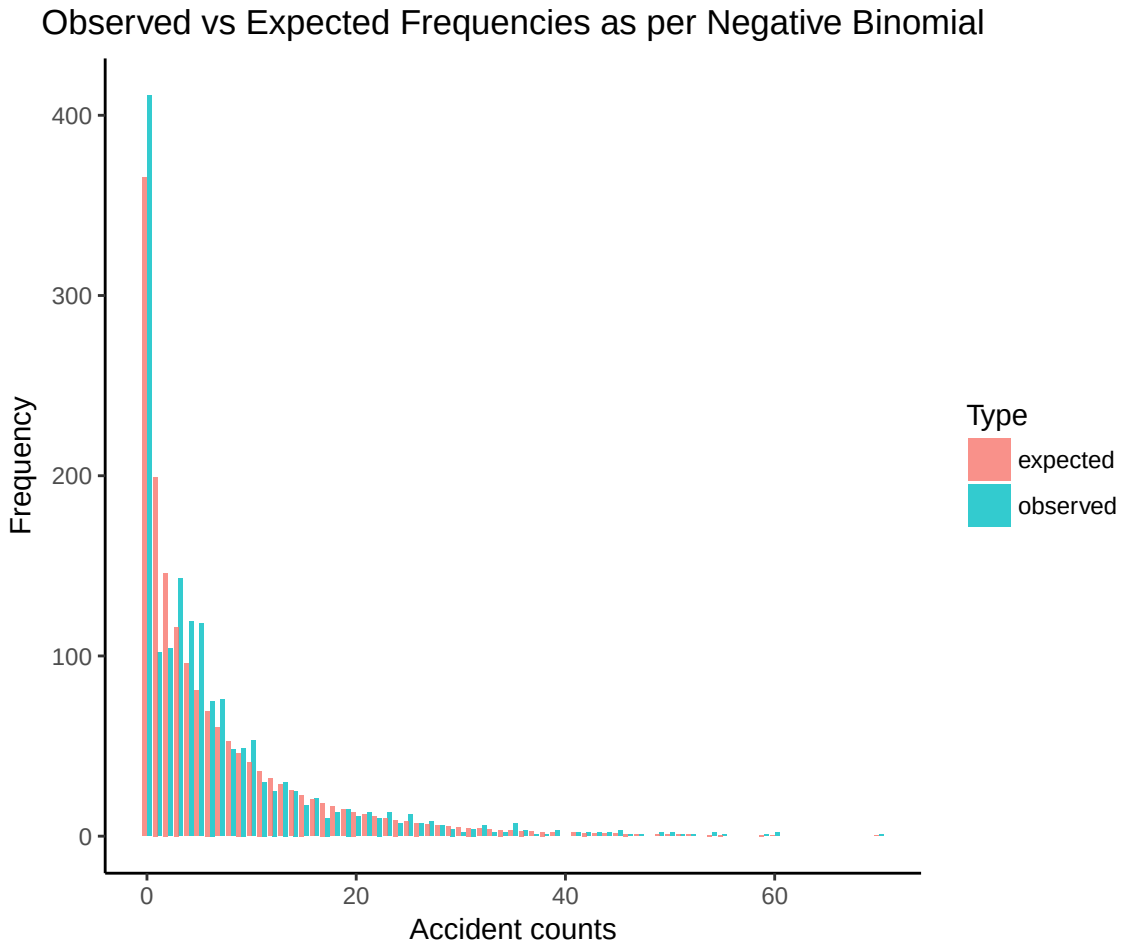
2.3. Model Selection

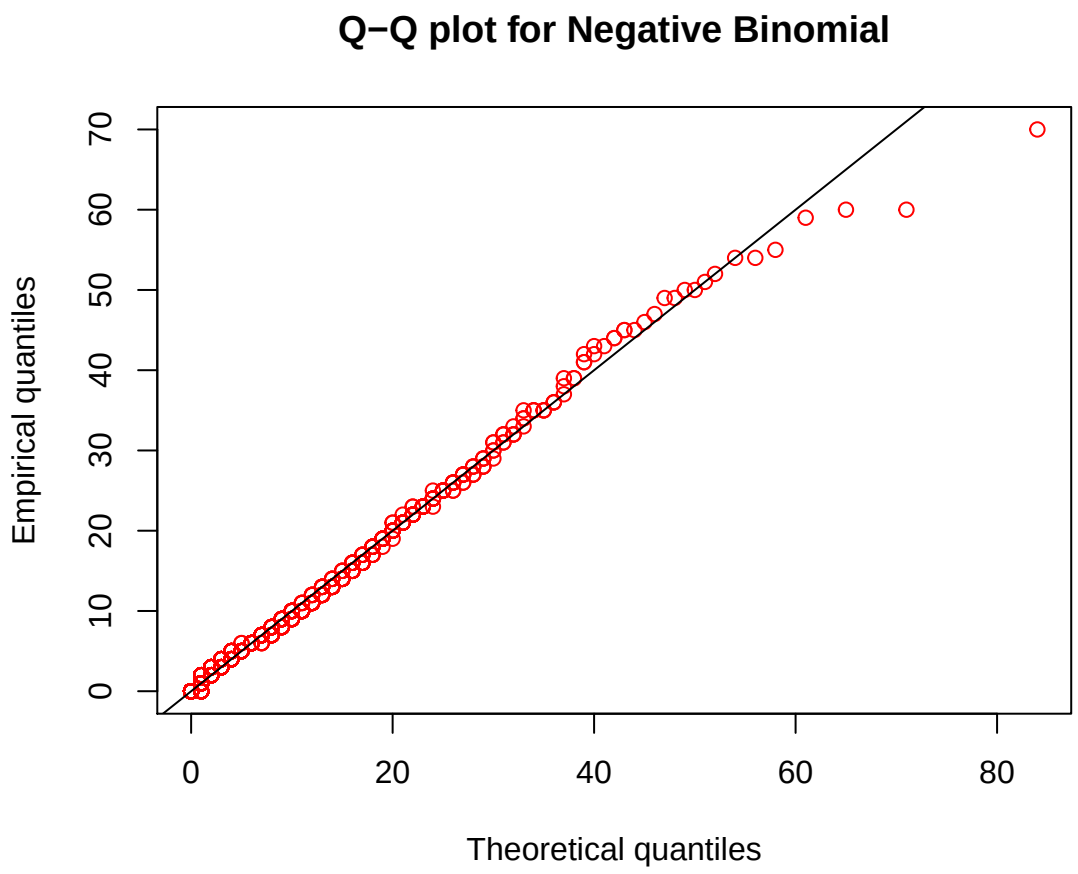
The AIC results for each model are summarised below.

Model	AIC
Poisson	20263.641
Negative Binomial	9626.917

Model	AIC
Poisson Mixture	12075.122
Zero Inflated Poisson	15556.157

The Negative Binomial has the lowest AIC value at 9626.92. This implies that is the best fitting model when compared to the other three models. The goodness of the Negative Binomial fit will now be assessed further.






```
##
##  Pearson's Chi-squared test
##
## data:  filter(comparison_df, Type == "observed")$Frequency and filter(comparison_df, Type !=
## X-squared = 1456, df = 1430, p-value = 0.3101
```

2.4. 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