2020

The Zero Inflated Negative Binomial - Shanker distribution and its application to HIV exposed infant data.

Kibika, Stella Andia
Strathmore Institute of Mathematical Sciences
Strathmore University

The Zero Inflated Negative Binomial - Shanker Distribution and its Application to HIV Exposed Infant Data

Kibika, Stella Andia

Master of Science, Statistical Sciences

2020

The Zero Inflated Negative Binomial - Shanker Distribution and its Application to HIV Exposed Infant Data

Kibika, Stella Andia

Submitted in partial fulfillment of the requirements for the Degree of Master of Science, Statistical Sciences at Strathmore University

Strathmore Institute of Mathematical Sciences
Strathmore University
Nairobi,Kenya

September, 2020

Declaration

I declare that this work has not been previously submitted and approved for the award of a degree by this or any other University. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made in the thesis itself.

Stella Andia Kibika

17/09/2020

Approval

The thesis of Stella Andia Kibika was reviewed and approved by the following:

Dr. Collins Odhiambo,
Lecturer, Strathmore Institute of Mathematical Sciences,
Strathmore University

Dr. Elphas Okango,
Lecturer, Strathmore Institute of Mathematical Sciences,
Strathmore University

Dr. Godfrey Madigu,
Dean, Strathmore Institute of Mathematical Sciences,
Strathmore University

Dr. Bernard Shibwabo,
Director of Graduate Studies,
Strathmore University

Abstract

Motivated by HIV exposed infants (HEI) sero-conversion data, we provide an extension of Zero-inflated Negative Binomial (ZINB) distribution to Zero-Inflated Negative Binomial – Shanker (ZINB-SH) distribution. We review the classical Poisson, and negative binomial distribution when applying count data and there zero-inflated versions. After reviewing the conceptual and computational features of these methods, we generate a new extension which is intrinsically a combination of Zero-inflated Negative Binomial and Shanker distribution. In this setting the ZINB-SH, distribution provides an alternative to the Poisson-Shanker distribution in particular, when data exhibits over dispersion brought by excess zeros. The HIV Exposed infant data is characterized by both structured and non-structured zeroes which makes the feature ideal in this context. We describe the properties of ZINB-SH distribution and estimate its parameters. Extensive simulations were conducted and the results in terms of goodness-of-fit, compared to the standard Negative Binomial, Shanker, Zero-Inated Negative Binomial and Negative Binomial-Shanker distributions. The ZINB-SH distribution is competitive under different settings of simulation and does well as sample size increases. To validate the distribution we apply real typical HIV Exposed Infant data.

Acknowledgement

The authors would like to thank the Strathmore Institute of Mathematical Sciences and the faculty who have supported the research devoting their time and intellectual resources.
I would also like to thank my supervisors, Dr. Collins Odhiambo and Dr. Elphas Okango for their contribution and support throughout the period of research.

Table of Contents

List of Figures

List of Tables

Abbreviations

MTCT – Mother to Child Transmission
HEI - HIV Exposed Infant
EID – Early Infant Diagnosis
HIV – Human Immuno Deficiency Virus
ZINB – Zero Inflated Negative Binomial
SH - Shanker
ZINB-SH - Zero Inflated Negative Binomial Shanker
ZINB-CR - Zero Inflated Crack
ZINB-GE - Zero Inflated Generalized Exponential
NB – Negative Binomial
MLE-Maximum Likelihood Estimation
GLM - Generalized Linear Model.

1 Introduction

1.1 Background

Zero inflated models have been developed to analyze count data that exhibits many zeros. The skewed nature of the resulting distribution makes it difficult to transform the data to a normal distribution. Count models such as Poisson and negative binomial are preferred. In cases where these models are not able to handle the number of zeros, they are extended to zero-inflated models. The zero inflated Poisson (ZIP) and zero inflated negative binomial (ZINB) are the most popular models. Hurdle models are an alternative way of modelling zero inflated data. The Poisson hurdle model was introduced in 1986 by John Mullahy. The hurdle models also model the zeros and the non-zero values (zero-truncated) separately. The difference between the zero inflated models and the hurdle models is that the hurdle models do not distinguish between the structured and random zeros. All the zeros are assumed to be structured (Mullahy, 1986). Zero-inflated models are not simply two-part models that can be estimated separately, as are hurdle models. Rather, zero-inflated models are mixture models. They use logistic or probit regression for the binary component, but both components – the binary and count – include the same zero counts when being estimated (Hilbe, 2014).
General form of the hurdle model:

$$g(X) = \begin{cases} \pi & x = 0 \\ h(X) & x > 0 \end{cases} \tag{1}$$

π=proportion of structured zeros
h(X) is the probability distribution function of the count data.
Failure to take into account both structured and unstructured zeros may introduce bias into the final results if there exists random zeros in the data (Blasco-Moreno, Pérez-Casany, Puig, Morante, & Castells, 2019). In cases where the data exhibits over dispersion, the ZINB is preferred to the Poisson. The zero inflated models introduce a second link function, the logit link, which allows the zeros to be distinguished as either structured or unstructured. There also exist hurdle models that assume all the zeros are structured. It truncates the non-zero part of the data. The hurdle models do not distinguish the zeros and may be applicable when the researcher is sure that all zeros arising from the data are structured for example when all the observations exist in a controlled space. An example would be looking at the occurrences of a disease when the entire population under investigation has been vaccinated for the particular disease. The zero inflated models will outperform hurdle models whenever the zeros in the data arise both structurally and randomly (Yang, 2014). In order to create more flexible models, there are some mixed models such as zero inflated negative binomial-Crack (Saengthong, Bodhisuwan, & Thongteeraparp, 2015) (ZINB-CR), zero inflated negative binomial-Sushila (Yamrubboon, Thongteeraparp, Bodhisuwan, & Jampachaisri, 2017) (ZINB-S) and zero inflated negative binomial-Generalized exponential (ZINB-GE)(Aryuyuen, Bodhisuwan, & Supapakorn, 2014) that have been developed. These models have been shown to perform better than the ZINB with fewer parameters. They are all parametric models that involve estimation of parameters. The disadvantage of the extended models is the complexity that is introduced with mixing two distributions. Estimation of parameters becomes more complex and requires algorithms such as the Newton Raphson and Expectation-Maximization (EM) algorithms. To achieve convergence, statistical software is used to make the process easier. There is also no standard procedure to tell when a dataset is exhibiting zero inflation (Warton, 2005). The ZINB model is a model that is used to address the issue of over dispersion in count data with excess zeros. We seek to extend the ZINB probability distribution function (PDF) to allow for greater flexibility by introducing more randomness. Structured zeros occur due to the presence of a group that is not at risk of exhibiting the phenomena under study. Structured zeros are inevitable while unstructured zeros occur by chance. (Ridout, Demétrio, & Hinde, 1998) gives an example of the distinction between the zeros using horticulture data where, in counting disease lesions on plants, a plant may have no lesions either because it is resistant to the disease(structured), or simply because no disease spores have landed on it(unstructured). In this study the focus will be on mothers who undergo prenatal and post-natal care in facilities that are equipped to prevent transmission (structured) and mothers who do not receive pre-natal and postnatal care or visit facilities that are not equipped to prevent transmission

(unstructured). Distinguishing structured and unstructured zeros is important here because the two groups have different probabilities of mother to child transmission. The zero inflated models are applied to such kind of data, with more zeros than the probability distribution expects, to account for the excess zeros. Creating more flexible models is important for the bias-variance trade off and creating models that have better predictive ability. The mixed models mentioned in the introduction have been proven to perform better than the original form of ZINB. Majority of sero-conversion among HIV Exposed Infants (HEI) occurs during the pregnancy, delivery and breastfeeding process (Nekesa, Odhiambo, & Chaba, 2019). The government and World Health Organization have put measures in place for Prevention of Mother to Child Transmission (PMTCT). These interventions have been introduced as a result of the high infant mortality recorded due to HIV in the period 1970-1990 where it also emerged that Mother to Child Transmission was a key factor (Shapiro & Lockman, 2010). The interventions put in place have resulted in decrease number of MTCT from 29.7% in 2015 to 11.5% in 2017 with PMTCT coverage of 77% according to the Ministry of Health HIV Estimates report of 2018. The difference in the quality of health services offered across the country leads to suboptimal procedures for PMTCT in certain facilities leading to random zeros in data collected for HIV sero-conversion(Nekesa et al., 2019). In many studies relying on count data where the counts may exhibit more zeros than the common count models can handle, zero inflated models can be used to model the outcomes. This is useful especially when some of the zeros recorded are not occurring randomly. For example, if an intervention is put in place to prevent a phenomenon from occurring, the data collected will contain zeros occurring randomly and those that are a result of the intervention. The negative binomial distribution is useful since it accounts for over dispersion that may be present in the data. However, it is not able to cater for excess zero therefore the ZINB distribution is applied. The ZINB applies weights to the structured and random zeros. It gives a weight π to the structured zeros and $(1-\pi)$ to the random zeros and other count values greater than zero. Given a random variable X,

$$P(X=x)=\begin{cases}\pi+(1-\pi)(1+\frac{\theta}{r})^{-r} & x_i=0\\ (1-\pi)\frac{\Gamma(x_i+r)}{x_i!\Gamma(r)}(1+\frac{\theta}{r})^{-1(1+\frac{r}{\theta})^{-x_i}} & x_i>0\end{cases} \tag{2}$$

where,
π = proportion of structured zeros
θ = probability of success
r = dispersion parameter.
This paper seeks to extend the ZINB to allow more flexibility to the model. The ZINB-SH will allow a parameter of the NB to be random and follow its own distribution. Mixture models have been used to increase the flexibility and robustness of probability distributions. This work aims to determine whether a mixture of the NB and Shanker distributions will provide greater flexibility and robustness when fitting zero inflated models. Fitting more than one model to a given dataset is common to establish the best model for a given situation.

1.2 Problem Statement

Analysis of count data is important in social sciences, epidemiology, public health and agriculture among many others.
In some cases the count of zeros in a dataset is more than what the common count models expect. In other cases the count of zero occurs due to a part of the population that does not exhibit the phenomenon under study(Mwalili, Lesaffre, & Declerck, 2008). In these cases, standard models such as the poisson and negative binomial may not be optimal.
Models have been developed to deal with these scenarios. The zero inflated poisson model for simulating zero inflated data for manufacturing developed by (Lambert, 1992) is one of the most popular methods. The other models developed include ZINB, ZINB-CR and ZINB-GE.
Excess zeros introduce overdispersion in the data.According to (Yang, 2014) when the data include excessive zeros (even as low as 20%) and over-dispersion, zero inflated models (i.e. ZIP, ZINB, ZAP, and ZANB) perform better than the standard count models.
Datasets exhibit different variations depending on the goals and method of collection. Datasets will exhibit variation in terms of shape. This may render standard models unfit for the count data. When standard distributions fail to fit a dataset accurately,new distributions are formulated to fit the shape of the dataset. In this work we develop one of the alternative distributions

that can be used when standard count distributions do not give a good fit.

1.3 Objectives

1.3.1 Main Objective

To create the ZINB-Shanker mixed distribution , find its properties and estimate the parameters.

1.3.2 Specific Objectives

i) To conduct simulations to compare the performance of standard count models, ZINB and ZINB-SH.

ii) To apply real data i.e. HIV exposed infants data, to the ZINB-SH distribution.

Counts datasets have been used in many areas i.e. number of vehicles crossing an intersection per hour, number of ER visits happening each month, number of motor vehicle insurance claims filed per year, number of defects found in a mass produced printed circuit board etc. Many real life phenomena produce counts that have zeros. For example number of times a machine fails each month, number of exoplanets discovered each year, the number of billionaires living in every single city in the world. Because such datasets are difficult to deal with using conventional count models i.e. the Poisson, the Binomial or the Negative Binomial regression models because they contain more number of zero valued counts than what one would expect to observe using the conventional probability distribution models. Standard Poisson and NB models on such data sets may generate poor quality predictions, no matter how much you tweak its parameters. A modify a standard counts model such as Poisson or Negative Binomial to account for the presence of the extra zeroes is therefore naturally implemented. One technique known as the Hurdle model can also be used though it is not a focus of this study. In this study, we'll look at the zero-inflated model. Specifically, we'll focus on the Zero Inflated Negative Binomial model, and then extend the work to Shanker Model.

1.4 Research Questions

i) Does increasing the randomness in the zero inflated negative binomial model improve the performance of the model?

ii) Does accounting for zero inflation in a dataset using zero inflated models improve the results of the study?

1.5 Scope of the study

This study will majorly focus on developing the ZINB-SH and its properties including mean and variance.The study will also cover estimation of parameters of the new distribution using maximum likelihood estimation. We will also test the performance of ZINB-SH compared to ZINB. We will begin by simulating zero inflated data and using a chi-square goodness of fit compare the performance of the two without covariates. The simulation will also cover the case where covariates are present and a regression is conducted. Here we use AIC and rootograms to compare performance. The new distribution will also be used to analyze the HEI dataset. The density plots and histograms will be used to compare the fit of ZINB and ZINB-SH to the HEI dataset.

1.6 Significance of the study

This study will develop a model that can be used as an alternative to model zero inflated datasets. The study will develop the probability distribution function and provide the properties of the distribution. The final results will give precise mathematical properties and formulas that can be directly applied to a dataset. The ZINB-SH will provide a model that can be used and compared with existing zero inflated models. The model can be applied to datasets without covariates and datasets containing covariates.

2 Literature Review

Poisson and negative binomial are the most popular count distributions used for count data. The Poisson regression model has been widely used to analyze count data. For example (Osgood, 2000) uses the Poisson regression to analyze crime rates, (Zou, 2004) proposes a modified Poisson regression model for measuring relative risk in epidemiological and medical studies. However, it is limited by its characteristic of equality of the mean and variance. When the mean and variance are not equal, the data is considered overdispersed. According to (Jiang & House, 2017) there is a special case of overdispersion occurs when there are excessive zeros and ignoring the abundant zeros results in poor parameter estimates and poor model fit. The negative binomial model is capable of taking into account overdispersion. (Hadi, Aruldhas, Chow, & Wattleworth, 1995) used the negative binomial regression model to estimate safety effects of cross-sections on highways. However, several studies show that the two models fail when there are excess zeros. (Desjardins, 2013) showed that the negative binomial does not work well in the presence of excess zeros using suspended school days data.

Most data exhibiting many zero counts is often due to a part of the population that are not at risk of exhibiting certain behavior during the period of the study (Hua, Wan, Wenjuan, & Paul, 2014).Count variables have structured zeros whose interpretation may be quite different from the random zeros (Hua et al., 2014). (Cohen, 1963) developed zero inflated models without the use of covariates. (Lambert, 1992) introduced the zero inflated Poisson (ZIP) regression model to deal with count data that exhibits an abundance of zeros. The model allows the zeros in the data to be classified as either structured or unstructured. ZIP uses a logistic regression component to distinguish the structured from the unstructured. However the ZIP model does not work well when there is overdispersion, the ZINB model is used in the presence of overdispersion. Failing to make the distinction results in highly biased results and significant loss of explanatory power (Hua et al., 2014).There is no research that conclusively shows one of these models to perform better than the other. The performance of the models depends on the particular scenario being studied.(Yang, 2014) shows the different scenarios in which each of the models is better. He also notes that the ZINB produces more consistent results when compared to ZIP and hurdle models at different levels of overdispersion and zero inflation using the AIC as the measure of accuracy.
Zero inflation is exhibited in many areas such as healthcare. (Loquiha et al., 2018) apply zero inflated models to model maternal mortality, (Hua et al., 2014) show the impact of failing to distinguish the random and non-random zeros in alcohol studies, (Lambert, 1992) creates a ZIP model with covariates for application in defects in manufacturing and finds that the ZIP models are better at predicting count data than Poisson. (Hall, 2000) developed a ZINB model used in horticulture and concluded that these models are important in modelling sources of heterogeneity in the data. Lambert shows how the zero inflated can be used in regression with covariates. The regression allows the factors affecting the outcome to be modelled. The covariates are estimated using the ordinary least squares method. The regression of zero inflated models can be extended to non-parametric models.There has been a study that has fitted a spline model to the zero inflated models. (Opitz, Tramini, & Molinari, 2013) fitted a model that would be used in cases when the GLM was too rigid. The spline model fitted was based on B-splines. The study brings out the importance of identifying the non-linear relationships in assessing the impact of risk factors. Using a dental data set, the classical GLM and ZI-GLM failed to bring out the significance of some of the predictors on the response.
Mixture models provide an alternative to non-parametric while being less restrictive than the usual distributions(Diebolt & Robert, 1994). There are several examples of mixture distributions that have been developed to improve existing distributions. The Sujatha distribution (Shanker et al., 2016) mixes the exponential and two gamma distributions to generate a lifetime distribution with an increasing hazard as opposed to making an assumption of constant mortality. The Sujatha can also be mixed with Poisson to create a better count model for frequency events. The parameter for the Poisson is assumed to follow a sujatha distribution. The negative binomial model is itself a mixture model of Poisson assuming the Poisson parameter follows a gamma distribution. Mixture models are important and Batang 2018 attributes the revival of the negative binomial distribution to the mixture models available. Mixture models had early applications in atmospheric data analysis.(Cohen, 1963) developed various mixture

models, from two normals, exponentials and poisson-negative binomial mixtures.
The Shanker is an important lifetime distribution that improves on the Lindley distribution. It combines the exponential and gamma distributions. It has been shown to provide better results compared to exponential and Lindley distributions. Given the problems that arise with excess zeros in the standard distributions such as overdispersion, this study will aim develop a zero inflated model for the Shanker using the negative binomial as the count distribution. The final model will give a mixture model with Shanker that can handle overdispersion and improve on the Poisson Shanker for scenarios where overdispersion is observed. The new model will gives the option for comparison with other models and a greater pool of choices for modelling overdispersed data with excess zeros. It will also provide greater flexibility that is brought about by mixing distributions for zero inflated models. The study will show whether the flexibility makes the model better by performing a comparison with the original ZINB.

2.1 Review of Count Data and Count Models

A count can be described as a positive value between the range of zero to infinity.Parametric statistical models rely on an underlying distribution. This distribution allows a researcher to provide inference and predict values based on a sample obtained from a larger population. This is the frequentist approach to statistical modelling, Bayesian approaches can also be used to model count data. Methods such as the maximum likelihood estimation and least squares regression aide in estimation of parameters that fit a particular dataset based on the underlying distribution.
Many count models exist that are used for count data. A count model has the response variable as count data. The most common models are the poisson and negative binomial count models.The most common problem that exists when modelling count data is overdispersion. Overdispersion can be described as a scenario where the variance of a dataset is greater than the mean. It can also be described as a scenario where the variance of the observed count data is greater than the variance of predicted counts. A model that fails to properly adjust for overdispersed data is called an overdispersed model. Its standard errors are biased and cannot be trusted, the standard errors associated with model predictors may appear from the model to significantly contribute to the understanding of the response, but in fact they may not (Hilbe, 2014). The negative binomial model is used to solve the problem of overdispersion. A dataset that structurally excludes zeros is modelled using zero truncated models while a dataset that has excess zeros uses zero inflated or hurdle models.
The negative binomial model will be the base model in this work.
Akaike Information Criterion(AIC) and Bayesian Information Criterion(BIC) can be used to choose the optimal count model to use for a given dataset.

2.2 Review of zero inflated datasets and Models

In this section we provide ways of assessing whether a dataset is exhibiting zero inflation.
According to (Warton, 2005) abundance of zeros does not necessarily imply zero inflation. (Warton, 2005) found that the negative binomial model was the best fitting of the count distributions, without zero-inflation. The high frequency of zeros was well described by the systematic component of the model and so it was rarely necessary to modify the random component of the model.
First, a dataset needs to be modelled using zero inflated models if the frequency of the zeros is not necessarily random.In a case where part of the population of interest never exhibits the phenomenon under study, part of the zeros in the data need to be treated as non-random.
Second, we can use model prediction to show the point at which a distribution fails to handle the number of zeros, that is, a dataset has excess zeros. We can predict how many zeros, a distribution expected and compare to the number of zeros there are in the dataset.
Negative Binomial Generalized Linear Model
This is the generalized linear model with the response variable following a negative binomial distribution.
Generalized linear models comprise three components:

1. A random component, specifying the conditional distribution of the random variable given the values of the explanatory variables in the model.

2. A linear predictor η_i
   $\eta_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik}$.
3. A link function which transforms the expectation of the response variable to the linear predictor.
   $g(\mu_i) = \eta_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik}$

where μ_i= expectation of the response variable to the linear predictor.
We use the negative binomial GLM to fit and predict. The maximum likelihood estimation can be used to determine the parameters of the negative binomial model. The estimated parameters can be used to fit the density of the negative binomial. The sum of the densities with a vector of zeros is used to determine the number of zeros that the distribution expected. A comparison of the expected zeros and the zeros present determines whether the dataset is zero inflated.
Data with an excessive number of zero counts is a problem for many count models (Hilbe, 2014). For example, given a specific mean value for a Poisson distribution of counts, the probability of having zero counts is defined by the PDF. When the mean is low, the probability of having zero counts is quite good. But for a mean is high, the Poisson PDF specifies that the probability of a zero count is very near zero. For a dataset following a Poisson distribution with high mean and high proportion of zeros, the Poisson distribution may be unable to handle the zeros. The zero inflated models were developed to model datasets with excessive zeros,that is, more zeros than the standard count models can handle.
In this work we introduce more randomness to the ZINB model and compare the performance of the new model to the ZINB. The goal is to assess whether extra randomness improves the performance of the model. It will also provide an alternative model to ZINB.

2.3 Zero-Inflated Poisson

The Poisson hurdle model was introduced in 1986 by John Mullahy. The hurdle models also model the zeros and the non-zero values (zero-truncated) separately. The difference between the zero inflated models and the hurdle models is that the hurdle models do not distinguish between the structured and random zeros. All the zeros are assumed to be structured (Mullahy, 1986)
The ZIP distribution allows some zeros to be non-random and other zeros follow the Poisson distribution.The ZIP without covariates was firs studied by (Cohen, 1963) Although Poisson models are popular when analyzing count data, the model might not be the best fit due to the characteristic of the Poisson count model of the mean-variance equation, which is specified by $\mu_i = E(Y_i) = Var(Y_i)$. The assumption is very restrictive and is easily violated (Jiang & House, 2017).

2.3.1 MLE for Zero-Inflated Poisson

In a zero-inflated model, we group all zeros into two parts. One part comes from the perfect modules, i.e., modules with zero faults. The other part comes from the non-perfect modules where the number of faults follows some standard distribution. In a ZIP regression model, we introduce a parameter π, which represents the probability of a module being perfect. Hence, the probability of the module being non-perfect is $1 - \pi$. Also, we assume that in non-perfect modules, the number of faults follows a Poisson distribution (Khoshgoftaar, Gao, & Szabo, 2001). The probability density function of the zero inflated poisson model is therefore,

$$P(X = x) = \begin{cases} \pi + (1-\pi)\exp^{-\mu_i} & x = 0 \\ \\ (1-\pi)\frac{\exp^{-\mu_i}\,\mu_i^{x_i}}{x_i!} & x > 0. \end{cases}$$

The parameters can be estimated using maximum likelihood and numeric optimization methods. The likelihood function:

$$\begin{aligned} L(\mu, \pi; \mathbf{x}) = \prod_{i=1}^{n} [[I_{(x_i=0)}(\pi_i + (1-\pi_i)\exp^{-\mu_i}] \\ + [I_{(x_i>0)}(1-\pi_i)\frac{\exp^{-\mu_i}\,\mu_i^{x_i}}{x_i!}]] \end{aligned} \tag{3}$$

The log-likelihood function:

$$\begin{aligned} logL = \sum_{i=1}^{x_i}[[I_{x_i=0}(log\pi_i + (1-\pi_i) - \mu_i] \\ +[I_{x_i>0}log(1-\pi_i) - \mu_i + x_i log\mu_i - x_i!]] \end{aligned} \tag{4}$$

Partial derivatives:

$$\frac{\partial logL}{\partial \pi} = \sum_{i=1}^{n}[[I_{x_i=0}\frac{-\mu_i}{\pi}] + [I_{x_i>0} - \frac{-\mu_i}{1-\pi}]] = 0 \tag{5}$$

$$\frac{\partial logL}{\partial \mu} = \sum_{i=1}^{n}[[I_{x_i=0} - (1-\pi)] + [I_{x_i>0} - (1-\pi) * \frac{x_i}{\mu}]] = 0 \tag{6}$$

2.4 Zero-Inflated Negative Binomal

In cases where the data exhibits over dispersion, the ZINB is preferred to the ZIP.
The ZINB uses the negative binomial as its base distribution. The negative binomial distribution involves a parameter for dispersion. The ZINB distribution allows some zeros to be non-random and other zeros follow the negative binomial distribution.

2.4.1 MLE for Zero-Inflated Negative Binomial

The probability density function of the zero inflated negative model is,

$$P(X=x) = \begin{cases} \pi + (1-\pi)g(x) & x = 0 \\ (1-\pi)g(x) & x > 0. \end{cases} \quad g(x) = \frac{\Gamma(x_i+\alpha^{-1})}{\Gamma(\alpha^{-1})\Gamma(x_i+1)}(\frac{1}{1+\alpha\mu_i})^{\alpha^{-1}}(\frac{\alpha\mu_i}{1+\alpha\mu_i})^{x_i}$$

$$\begin{aligned} L(\alpha,\mu,\pi;\mathbf{x}) = \prod_{i=1}^{n}[[I_{(x_i=0)}\pi_i + (1-\pi_i)(\frac{1}{1+\alpha\mu_i})^{\alpha^{-1}}] \\ +[I_{(x_i>0)}(1-\pi_i)\frac{\Gamma(x_i+\alpha^{-1})}{\Gamma(\alpha^{-1})\Gamma(x_i+1)}(\frac{1}{1+\alpha\mu_i})^{\alpha^{-1}}(\frac{\alpha\mu_i}{1+\alpha\mu_i})^{x_i}] \end{aligned} \tag{7}$$

$$\begin{aligned} lnL(\alpha,\mu,\pi;\mathbf{x}) = \prod_{i=1}^{n}[\log[I_{(x_i=0)}\pi_i + (1-\pi_i)(\frac{1}{1+\alpha\mu_i})^{\alpha^{-1}}] \\ +[I_{(x_i>0)}log(1-\pi_i) + \log\Gamma(x_i+\alpha^{-1}) - \log\Gamma(\alpha^{-1}) - \log\Gamma(x_i+1) + \log(\frac{1}{1+\alpha\mu_i})^{\alpha^{-1}} + \log(\frac{\alpha\mu_i}{1+\alpha\mu_i})^{x_i}] \end{aligned} \tag{8}$$

$$\begin{aligned} \frac{\partial logL}{\partial \alpha} = \sum_{i=1}^{n}[[I_{x_i=0}\frac{1}{\pi_i + (1-\pi_i)(\frac{1}{1+\alpha\mu_i^{\alpha-1}})}(1-\pi_i)(\alpha-1)\frac{1}{1+\alpha\mu_i}^{\alpha-2}\mu_i] \\ +[I_{x_i>0}\frac{1}{\Gamma(x_i+\alpha^{-1})}\Gamma'(x_i+\alpha^{-1}) - \frac{\Gamma'(\alpha^{-1})}{\Gamma(\alpha^{-1})}]] = 0 \end{aligned} \tag{9}$$

$$\begin{aligned} \frac{\partial logL}{\partial \mu} = \sum_{i=1}^{n}[[I_{x_i=0}(1-\pi_i)(1-\alpha)(\frac{1}{1+\alpha\mu_i})^{\alpha-2}(\frac{1+\alpha}{(1+\alpha\mu_i)^2})] \\ +[I_{x_i>0}(\alpha-1)(\alpha-1)(\frac{1}{1+\alpha\mu_i})\frac{1+\alpha}{(1+\alpha\mu_i)^2} + x_i(\frac{\alpha\mu_i}{1+\alpha\mu_i})^{x_i-1}\frac{(\alpha\mu_i)(1+\alpha)-\mu_i^2}{(1+\alpha\mu_i)^2}]] = 0 \end{aligned} \tag{10}$$

$$\frac{\partial logL}{\partial \pi} = \sum_{i=1}^{n}[[I_{x_i=0}1 - (\frac{1}{1+\alpha\mu_i})^{\alpha-1}] + [I_{x_i>0}\frac{-1}{1-\pi_i}]] = 0 \tag{11}$$

3 Research Methodology

3.1 Research Design

This study will create a mixture distribution for the ZINB and Shanker distributions. The model will then be applied on a real dataset for HIV exposed infants. The new distribution is made of the negative binomial and Shanker distributions and then adjusted for zero inflation.

The negative binomial Distribution

Suppose that X is a random variable from the negative binomial distribution, the PDF is given by:

$$P(X=x)=\binom{m+x_i-1}{x_i}p^m(1-p)^{x_i} \quad x_i=0,1,2,... \quad m>0, \quad 0<p<1, \tag{12}$$

then the expected value and variance of X are
$E(X)=\frac{m(1-p)}{p}$,
and
$E(X^2)=\frac{m(1-p)[1+m(1-p)]}{p^2}$ respectively.
$p=\exp(-\lambda)$
where λ follow a Shanker(θ) distribution.

Shanker Distribution

The Shanker distribution is proposed by (Shanker, 2015) for modelling lifedata. It is a one parameter distribution which is a mixture of $exponential(\theta)$ and $gamma(2,\theta)$ distributions. Suppose that X is a random variable from the Shanker distribution, the PDF is given by:

$$P(X=x)=\frac{\theta^2}{\theta^2+1}(\theta+x_i)e^{(-\theta x_i)}; \quad x_i>0, \quad \theta>0, \tag{13}$$

then the expected value and variance of X are
$E(X)=\frac{\theta^2+2}{\theta(\theta^2+1)}$,
and
$Var(X)=\frac{\theta^4+4\theta^2+2}{\theta^2(\theta^2+1)^2}$ respectively.

The Negative binomial-Shanker (NB-SH) Distribution

This is a compound distribution developed by (Tlhaloganyang, Mooketsi, Leinanyane, & Sakia, n.d.). It is a mixture of Negative Binomial and Shanker. Suppose that X is a random variable from the Negative Binomial-Shanker distribution, the PDF is given by:

$$P(X=x)=\binom{m+x_i-1}{x_i}\sum_{k=0}^{x_i}\binom{x_i}{k}(-1)^k(1+\frac{\theta(m+k)}{\theta^2+1})(1+\frac{m+k}{\theta})^{-2}, \tag{14}$$

then the expected value and variance of X are
$E(X)=\frac{m\theta^2}{\theta^2+1}(\omega-\delta)$,
and
$Var(X)=\frac{m(m+1)\theta^2}{\theta^2+1}(\rho-2\omega+\delta)$
where,
$\delta=\frac{\theta^2+1}{\theta^2}$, $\omega=\frac{\theta(\theta-1)+1}{(\theta-1)^2}$, $\rho=(\frac{\theta-1}{\theta-2})^2$.
This distribution is a special case of the generalized negative binomial-Shanker(GNB-SH) with the parameter $\beta=1$.

The first objective: To create a ZINB-Shanker distribution will be achieved by using the method used by (Lambert, 1992) which distinguishes the structured and random zeros. The model is a mixture of Bernoulli and Negative binomial-Shanker. The random zeros will follow the ZINB-Shanker pdf in equation 3. The model for zero inflation is as below:

$$P(X=x)=\begin{cases}\pi+(1-\pi)((1+\frac{\theta m}{\theta^2+1})(1+\frac{m}{\theta})^{-2} & x=0\\ (1-\pi)\binom{m+x-1}{x}\sum_{k=0}^{x}\binom{x}{k}(-1)^k(1+\frac{\theta(m+k)}{\theta^2+1})(1+\frac{m+k}{\theta})^{-2} & x>0.\end{cases}$$ where,

π=proportion of structured zeros

θ=probability of success
r=dispersion parameter.
To get the properties of the ZINB-SH, some general rules on finding mean and variance of zero inflated models are used.
Suppose we knew which zeros come from the random NB state that is, suppose we could observe $Z_i = 1$ when X_i is from the structured zero and $Z_i = 0$ when Y_i is from the random zero.
Then Z is a Bernoulli distribution defined by,

$$Z = \begin{cases} 1 & with\, probability\, \pi \\ 0 & with\, probability\, 1-\pi \end{cases} \tag{15}$$

$E[X] = E[E[X|Z]]$
$E[X|Z = 0]P[Z = 0] + E[X|Z = 1]P[Z = 1]$,since $E[X|Z = 1] = 0$
$E[X] = (1 - \pi)E(X|Z = 0)$,
and
$Var[X] = Var[E[X|Z]] + E[Var[X|Z]]$, since $Var[X|Z = 1] = 0$ and $E[X|Z = 1] = 0$,
$Var[X] = (1 - \pi)var[X|Z = 0] + \pi(1 - \pi)E[X|Z = 0]^2$.
Suppose that X is a random variable from the zero inflated negative binomial shanker distribution. Then the expected value and variance of X are respectively,
$E(X) = (1 - \pi)\frac{m\theta^2}{\theta^2+1}(\omega - \delta)$,
$Var(X|Z = 0) = \frac{m(m+1)\theta^2}{\theta^2+1}(\rho - 2\omega + \delta) - [\frac{m\theta^2}{\theta^2+1}(\omega - \delta)]^2$,
and
$Var(X) = (1 - \pi)(\frac{m(m+1)\theta^2}{\theta^2+1}(\rho - 2\omega + \delta) - [\frac{m\theta^2}{\theta^2+1}(\omega - \delta)]^2) + \pi(1 - \pi)(\frac{m(m+1)\theta^2}{\theta^2+1}(\rho - 2\omega + \delta))^2$.
The parameters of the distribution were estimated using the maximum likelihood estimation method. The maximum likelihood method differentiates the product of the probability distribution function with respect to each of the parameters. Numerical methods are used to solve the final equations.
The likelihood function:

$$\begin{aligned} L(m, \theta, \pi; \mathbf{x}) = \prod_{i=1}^{n} [[I_{(x_i=0)}(\pi_i + (1 - \pi_i)(1 + \frac{\theta m}{\theta^2 + 1})(1 + \frac{m}{\theta})^{-2})] \\ +[I_{(x_i>0)}(1 - \pi_i)\binom{m + x_i - 1}{x_i}\sum_{k=0}^{x_i}\binom{x_i}{k}(-1)^k(1 + \frac{\theta(m + k)}{\theta^2 + 1})(1 + \frac{m + k}{\theta})^{-2}]] \end{aligned} \tag{16}$$

The log-likelihood function:

$$\binom{m + x_i - 1}{x_i} = \frac{\Gamma(m + x_i)}{\Gamma(m + 1)\Gamma(x_i)} \tag{17}$$

$$\begin{aligned} logL = \sum_{i=1}^{x_i}[log[I_{x_i=0}(\pi_i + (1 - \pi_i)(1 + \frac{\theta m}{\theta^2 + 1})(1 + \frac{m}{\theta})^{-2})] \\ +[I_{x_i>0}log(1 - \pi_i) + log\Gamma(m + x_i) - log\Gamma(m + 1) - log\Gamma(x_i) \\ +log\sum_{r=0}^{x_i}\binom{x_i}{r}(-1)^r + log(1 + \frac{\theta(m + r)}{\theta^2 + 1}) + log(1 + \frac{m + r}{\theta})^{-2}]] \end{aligned} \tag{18}$$

Partial derivatives:

$$\frac{\partial logL}{\partial \pi} = \sum_{i=1}^{n}[[I_{x_i=0}\frac{1}{(\pi_i) + (1 - \pi_i)(1 + \frac{\theta m}{\theta^2+1})(1 + \frac{m}{\theta})^{-2}}] + [I_{x_i>0}\frac{1}{1 - \pi_i}]] = 0 \tag{19}$$

$$\begin{aligned} \frac{\partial logL}{\partial \theta} = \sum_{i=1}^{n}[[I_{x_i=0}(1 + \frac{1 + \theta m}{\theta^2 + 1})(\frac{\theta m(2\theta)^{-1} - (\theta^2 + 1)m}{\theta m(\theta^2 + 1)} - 2(1 + \frac{m}{\theta})^2\theta)] \\ +[I_{x_i>0}((1 + \frac{\theta(m + r)}{\theta^2 + 1})^{-1})(\frac{(m + r)}{\theta(m + r)(\theta^2 + 1)}) - 2(1 + \frac{m + r}{\theta})^2(-\frac{1}{\theta})^{-1}]] = 0 \end{aligned} \tag{20}$$

$$\frac{\partial logL}{\partial r} = \sum_{i=1}^{n}[I_{x_i>0}(1 + \frac{\theta(m + r)^{-1}}{\theta^2 + 1})\theta + (1 + \frac{m + r}{\theta})^2] \tag{21}$$

$$\frac{\partial logL}{\partial m} = \sum_{i=1}^{n}[[I_{x_i=0}(1+\frac{\theta m}{\theta^2+1})^{-1}\theta - 2(1+\frac{m}{\theta})^2]$$
$$+[I_{x_i>0}\frac{1}{\Gamma(m+x_1)}\frac{\partial}{\partial m}\Gamma(m+x_i) - \frac{1}{\Gamma(m+1)}\frac{\partial}{\partial m}\Gamma(m+1) \tag{22}$$
$$+(1+\frac{\theta(m+r)}{\theta^2+1})^{-1}\theta - 2(1+\frac{m+r}{\theta})^2]] = 0$$

3.2 Data

The data used will be secondary data from publicly available information. The data used will be from three high burden areas, Kisumu, Nairobi and Mombasa. This data exhibits zero inflation because of the measures that have been put in place to reduce the rate of Mother to Child Transmission (MTCT). This increases the chance of children being born HIV-free and the chance of recording a zero. Due to the intervention the results will contain structured zeros. All HEI sero-conversion who were registered in the EID programme in Kisumu, Nairobi and Mombasa between January 2014 and December 2018 were included in the study. From study sampling frame, a total of 494 samples were collected from HEI visiting 60 health facilities across the three cities in Kenya and obtained PCR testing together with the results. HEI with missing age or greater than 2 years old were excluded from analysis.

Statistical analysis

Data were transferred from the Microsoft excel windows 12 to R Studio15 where data were analyzed. The analysis involves generating new random variables and getting density curves. The density curves will show the manner in which a given distribution fits the data, especially the zeros. We obtain a chi-square goodness of fit test to compare the distributions.

Chi-square test for goodness of fit test:

$$\chi^2_{\alpha,d} = \frac{(O-E)^2}{E}$$

where,
O=observed Value
E=expected Value
α = significance level
d = degrees of freedom.

Ethics

Data is secondary and readily available from National AIDS and STI Control Programme (NASCOP) website. No patient identification information is included in NASCOP database.

3.3 Simulation

Simulation will be achieved by use of common simulation methods such as the acceptance rejection region method to fit a new probability distribution. The method is an iterative method and statistical software will be used to run the algorithm. The goal will be to generate random numbers following the ZINB-SH(m,θ). The results will be used to generate plots to visualize the shape of the distribution. The steps used will be:

i) Generate U from the Uniform(0,1) distribution.

ii) Let λ come from the shanker distribution(θ).

iii) Generate Y from the NB(m,p) distribution.

iv) Generate U^* fron the Uniform(0,1) distribution.

v) if $U^* > \omega$ set X=Y, otherwise X=0

4 Results

A simulation was carried out based on four distributions. Negative Binomial (NB), Negative Binomial–Shanker (NB-SH), Zero Inflated Negative Binomial (ZINB) and Zero Inflated Negative Binomial Shanker (ZINB-SH). The simulation is based on the following parameters estimated using maximum likelihood method in R statistical software:

Table 4.1: Parameter Estimation.

	NB	SH	NB-SH	ZINB	ZINB-SH
θ	0.2432	-	3.5572	0.8396	0.7647
exp rate	-	2.3429	-	-	-
m	-	-	-	479	479
π	-	-	-	0.4141	0.4141
r(dispersion parameter)	-	-	-	0.5574	0.5574

The simulations produce the density plots below:

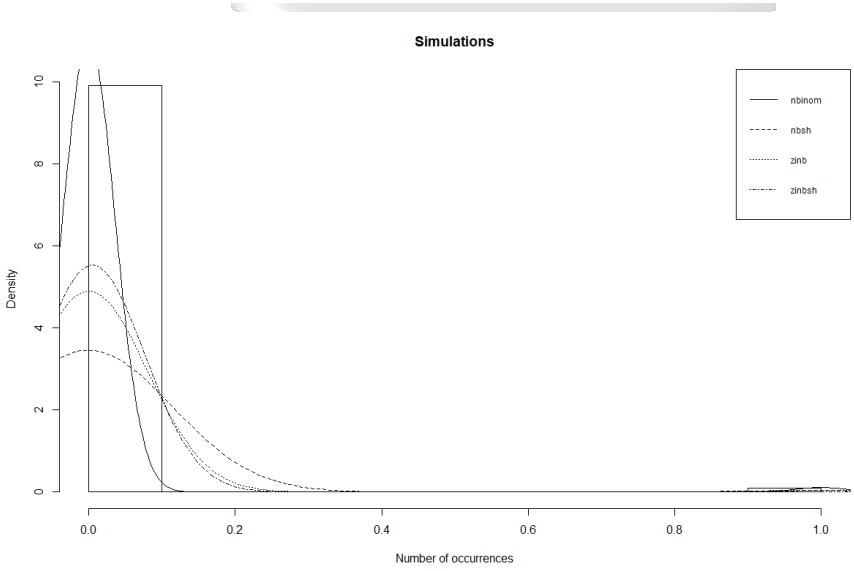


Figure 4.1: Density plots from simulations

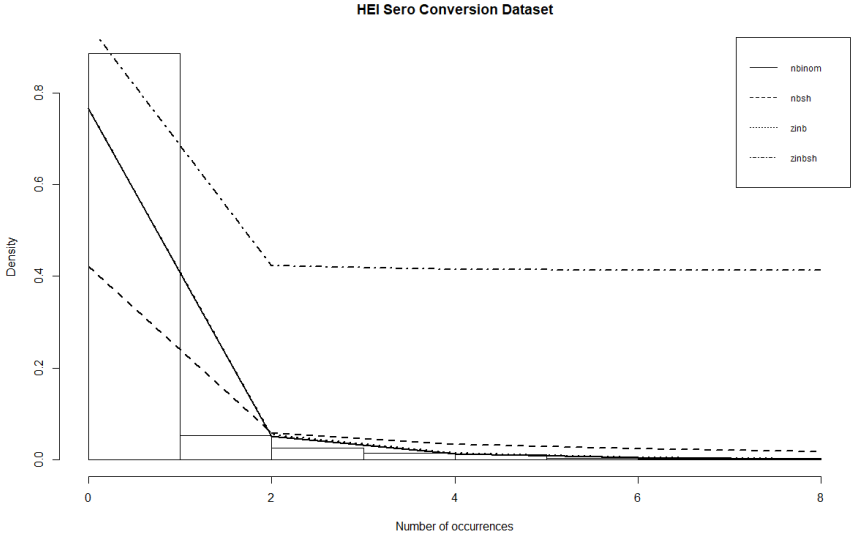


Figure 4.2: Density plots from HEI sero conversion dataset

The density plots show how the different models fit the data with zero inflation, the ZINB, ZINB-SH and NB-SH distributions partition the zeros. A portion of the zero values that may be considered random appear on the density plots. The NB distribution attempts to fit all the zero under the density curve since all the zeros are considered random.

Table 4.2: Observed and expected count values for ZINB and ZINB-SH.

Number of counts	Observed	ZINB	ZINB-SH	NB	NB-SH
0	378	380	381	381	454
1	59	61	47	47	29
2	26	27	23	23	8
3	13	10	20	20	3
4	7	7	10	10	0
>4	11	9	13	13	0
χ^2goodness of fit test(p-value)		0.09	0.78	0.78	0.05

The goodness of fit statistics show that there is no significant difference between the observed and expected values of the distributions. This implies that the models can be implemented in different scenarios and the best model chosen based on criteria such as Akaike Information Criterion (AIC).

Performance of these models can also be assessed using rootograms. We simulate independent variables and conduct a regression in order to be able to plot the rootograms. A rootogram uses bar plots to show how the model fit each count value. Hanging bars, that is, bars plotted above or below the zero line show a poor fit while bars plotted exactly on the zero line indicate a good fit.If a bar doesn't reach the zero line then the model over predicts a particular count bin, and if the bar exceeds the zero line it under predicts.
The figures below show the rootograms for the simulated dataset.

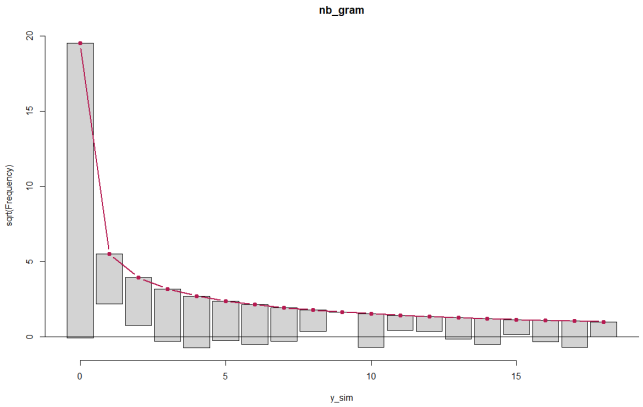


Figure 4.3: Rootogram plot for Negative Binomial Distribution

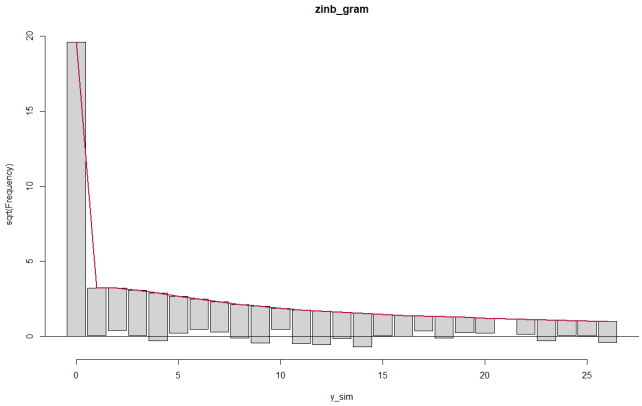


Figure 4.4: Rootogram plot for ZINB Distribution

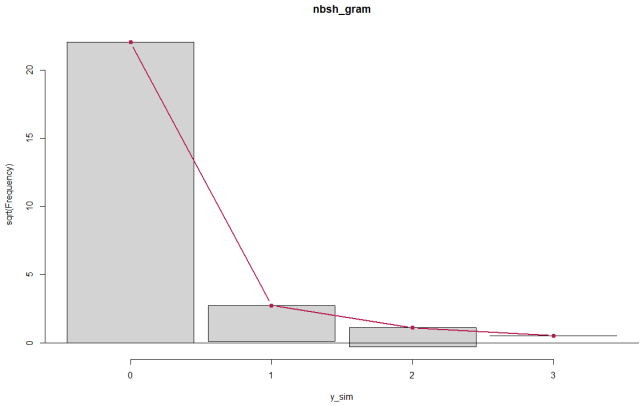


Figure 4.5: Rootogram plot for NB-SH Distribution

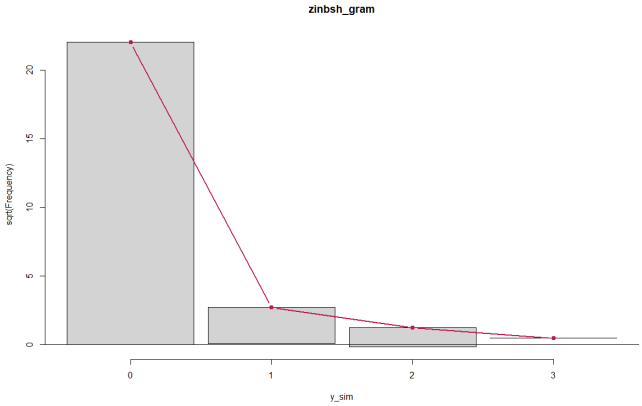


Figure 4.6: Rootogram plot for ZINB-SH Distribution

The simulations were also conducted on datasets with different proportions of non-random zeros. The performance based on AIC is shown in Table 3 below:
$AIC = -2logL(\theta) + 2k$
where $L(\theta)$ is the maximized likelihood function for the estimated model and k is the number of estimated parameters in the model.
AIC rewards goodness of fit (as assessed by the likelihood function), but it also includes a penalty that is an increasing function of the number of estimated parameters. The penalty discourages overfitting, which is desired because increasing the number of parameters in the model almost always improves the goodness of the fit.
The simulations below are based on a zero inflated dataset with 86% zeros.

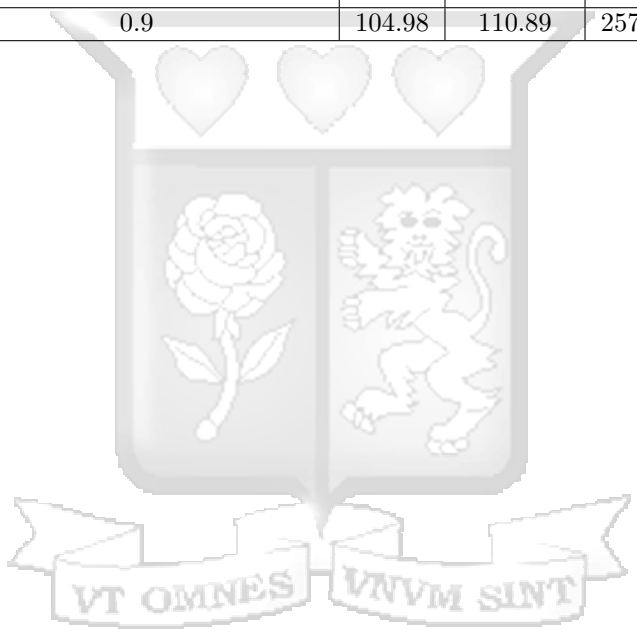
Table 4.3: Performance based on AIC values for dataset with 86% zeros.

Proportion of non-random zeros	ZINB	ZINB-SH	NB	NB-SH
0.0	2320	2377.60	598.18	1499.10
0.2	2050	2234.71	598.24	1500.01
0.4	1729.6	1998	1601.20	2553.42
0.8	665.4	705.34	1819.54	2897.81
0.9	322.6	350.04	2001.02	3228.52

We conduct similar AIC comparison on a zero inflated dataset with 14% zeros and obtain the results below:

Table 4.4: Performance based on AIC values for dataset with 14% zeros.

Proportion of non-random zeros	ZINB	ZINB-SH	NB	NB-SH
0.0	469.2	477.61	98.18	149.10
0.2	400.20	464.18	98.03	150.76
0.4	304.20	397.93	160.42	255.20
0.8	122.30	135.61	213.54	289.87
0.9	104.98	110.89	257.02	315.96



5 Discussion

The density plots using the simulated and real data show how well the distributions fit to the data. The zero inflated distributions fit the data better than the standard count models. Given a count dataset, we can plot histograms and density plots to determine the distribution of the dataset.

We further assess the goodness of fit by using the real dataset of HIV sero-conversion. We obtain expected values from each of the distributions and calculate the χ^2 goodness of fit. The results show that the distributions can be applied to the simulated dataset. This means that for the HIV sero-conversion dataset both standard and zero inflated models can be applied based on the χ^2 statistic.

The rootograms show the fit of the data at different counts for the simulated dataset. The negative binomial offers a poor fit of the data at count zero compared to the other models, it under predicts the zero counts. The zero inflated models give an accurate result for the zero counts. The rootograms give a visualization that shows the zero inflated models perform better in predicting a zero count when a dataset is exhibiting zero inflation.

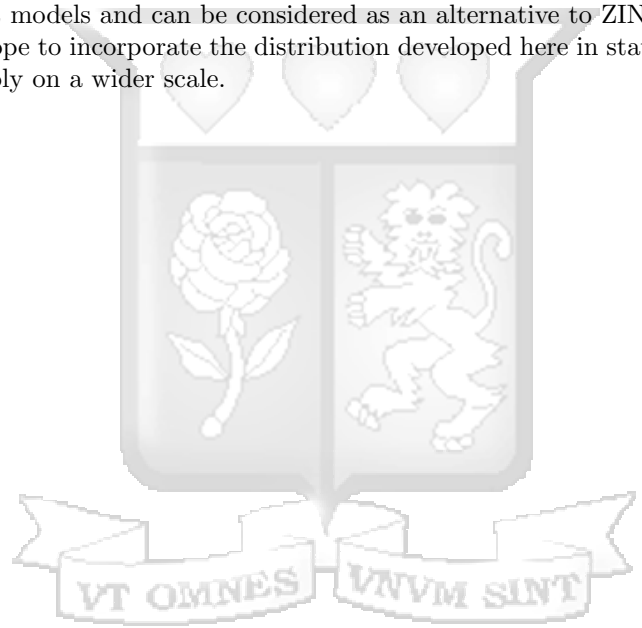
A dataset can exhibit zero inflation even when the overall percentage of zeros is small.It is important to test for zero inflation before using a zero inflated distribution for analysis. We use AIC values to show performance of a the distributions with simulated data exhibiting zero inflation with 86% of the overall values being zero and a second simulated dataset with 14% zeros. Since both datasets are zero inflated, there is a common trend.Overall, the ZINB distribution performs better than the ZINB-SH as more non-random zeros are introduced to the dataset for this simulated dataset. However,compared to standard count models, ZINB-SH performs better as the proportion of no random zeros increases. When all the values are zero, the standard count distributions have low AIC values and perform better than the ZINB-SH and ZINB. Non-random zeros result in an increase in AIC values for the standard models and a reduction in AIC for the zero inflated models. When the proportion of non-random zeros is 0.9, the ZINB and ZINB-SH have lower AIC values.

Interventions that targets PMTCT remains important HIV management considerations and are intended to mitigate the risk of HIV transmission from HIV infected mothers to their child. Kenya, in particular, the country has to a large extent, made significant progress in reducing MTCT rate. Literature has shown, for effective reduction in transmission, interventions for HEI begins well before delivery (Nyamhanga, Frumence, & Simba, 2017). The HEI care and treatment, focuses on reducing the risk of infection with PreP, monitoring for signs and symptoms of HIV sero-conversion, and adhering to PCR testing schedule and starting treatment immediately. The UNAIDS report describes a 26% decline in incident HIV infections between 2009 and 2015 in the Global Plan priority countries in Sub-Saharan Africa. This study is motivated by typical HEI sero-conversion data. In this setting the ZINB-SH, distribution provides an alternative to the Poisson-Shanker distribution in particular, when data exhibits over dispersion brought by excess zeros. The HIV Exposed infant data is characterized by both structured and non-structured zeroes which makes the feature ideal in this context. HEI sero-conversion data is collected routinely by ministry of health (MoH) in Kenya. Naturally, implementation of PMTCT is heterogeneous and results to structured zeros among positive HEI (situation where PMTCT is implemented optimally) and random zero among positive HEI (situation where PMTCT is implemented sub-optimally). Failure to accommodate structured and random zero-inflation may result in false inference (Blasco-Moreno et al., 2019). With this backdrop, the conventional zero-inflated distributions that do not consider structured and random zeros in data may give misleading results. Several rigorous and non-rigorous count data analysis approaches with zero inflation have been proposed by different researchers. (Nekesa et al., 2019) did a comparison of four zero inflated models including the Zero Inflated Poisson (ZIP), Zero Altered Poisson (ZAP), Zero Inflated Negative Binomial (ZINB) and Zero Altered Negative Binomial (ZANB). (Nekesa et al., 2019) concluded that the ZAP was the best model based on AIC values for the different models. Covariates are used to run a regression based on the four models and HEI is shown to be twice likely to detect HIV compared to initial Polymerase chain reaction (PCR). Here we provide an extension of Zero-inflated Negative Binomial (ZINB) distribution to Zero Inflated Negative Binomial-Shanker (ZINB-SH) distribution. We have described the properties of ZINB-SH distribution and estimated its parameters. Extensive simulations were conducted and the results in terms of goodness-of-fit, compared to the

standard Negative Binomial, Zero-Inflated Negative Binomial and Negative Binomial–Shanker distributions. The ZINB-SH distribution is competitive under different settings of simulation and does well as sample size increases. To validate the distribution we apply real typical HIV Exposed Infant data.

6 Conclusion

In this work, the aim was to create a new distribution for zero inflated data and determine whether the new distribution performs better than the standard ZINB. The major difference between the two distributions is that ZINB-SH allows the parameters of the NB to be random and follow a distribution of their own, The parameters for the new distribution are determined using the maximum likelihood method. The new model is used in generating new random variables through simulations. It is also applied to a HEI Sero conversion dataset. In this case the ZINB-SH distribution has been shown to be competitive in performing analysis on zero inflated data. In the context of mixture distributions, a distribution with more parameters being random, with their own distribution, provides greater flexibility. ZINB-SH can be considered as a distribution when fitting models that exhibit excess zeros. The extra randomness in the model does not necessarily improve the ZINB, however, the ZINB-SH performs better than the standard count models and can be considered as an alternative to ZINB.
In future we hope to incorporate the distribution developed here in statistical software to make it easier to apply on a wider scale.



Appendices

R Codes

```
#FITTING THE ZERO INFLATED NEGATIVE BINOMIAL DISTRIBUTION TO THE HIV DATASET
library(fitdistrplus)
library(ggplot2)
library(distr)
library(actuar)
library(pscl)
library(ZIM)
library(pscl)
hiv<-read.csv("D:/Stella/Documents/MASTERS/Thesis/negative_binomial/Copy of HEI_Data.
#View(hiv)


rshanker<-function(n,exp_rate)
{iseed=101
w_exp=exp_rate/(exp_rate^2+1)
set.seed(iseed)
w_gam<-1-w_exp
r_exp <- rexp(n,exp_rate)
r_gam <- rgamma(n,shape=2,scale = exp_rate)
#flag <- rbinom(n,size=1,prob=cpct)
return(r_exp*w_exp + r_gam*w_gam)
}

qshanker<-function(p,exp_rate)
{iseed=101
w_exp=exp_rate/(exp_rate^2+1)
set.seed(iseed)
w_gam<-1-w_exp
q_exp <- qexp(p,exp_rate)
q_gam <- qgamma(p,shape=2,scale = exp_rate)
#flag <- rbinom(n,size=1,prob=cpct)
return(q_exp*w_exp + q_gam*w_gam)
}



pshanker<-function(q,exp_rate)
{iseed=101
w_exp=exp_rate/(exp_rate^2+1)
set.seed(iseed)
w_gam<-1-w_exp
p_exp <- pexp(q,exp_rate)
p_gam <- pgamma(q,shape=2,scale = exp_rate)
#flag <- rbinom(n,size=1,prob=cpct)
return(p_exp*w_exp + p_gam*w_gam)
}


dshanker <- function(x,exp_rate)
{iseed=101
w_exp=exp_rate/(exp_rate^2+1)
set.seed(iseed)
w_gam<-1-w_exp
d_exp <- dexp(x,exp_rate)
d_gam <- dgamma(x,shape=2,scale = exp_rate)
#flag <- rbinom(n,size=1,prob=cpct)
return(d_exp*w_exp + d_gam*w_gam)
}
```

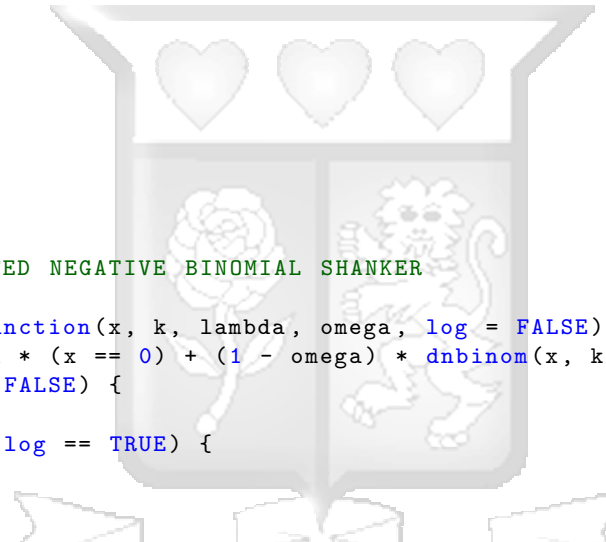
```r
#NEGATIVE BINOMIAL SHANKER DISTRIBUTION

rnbsh<-function(n,k,lambda){
  rnbinom(n,k,mu = lambda)
}

qnbsh<-function(p,k,m,lambda){
  q<-pnbinom(p, k, mu = lambda)
}

pnbsh<-function(q,k,m,lambda){
  p<-pnbinom(q, k, mu = lambda)
}


dnbsh<-function(x,k,lambda){
  d<-dnbinom(x,k,mu = lambda)
}
```



```r
#ZERO INFLATED NEGATIVE BINOMIAL SHANKER

dznbsh <- function(x, k, lambda, omega, log = FALSE) {
  d <- omega * (x == 0) + (1 - omega) * dnbinom(x, k, mu = lambda)
  if(log == FALSE) {
    d
  } else if(log == TRUE) {
    log(d)
  }
}

pznbsh <- function(q, k, lambda, omega, lower.tail = TRUE, log.p = FALSE) {
  p <- omega * (q >= 0) + (1 - omega) * pnbinom(q, k, mu = lambda)
  if(lower.tail == FALSE) {
    p <- 1 - p
  }
  if(log.p == TRUE) {
    p <- log(p)
  }
  p
}

qznbsh <- function(p, k, lambda, omega, lower.tail = TRUE, log.p = FALSE) {
  if(lower.tail == FALSE) {
    p <- 1 - p
  }
  if(log.p == TRUE) {
    p <- exp(p)
  }
  qnbinom(pmax(0, (p - omega) / (1 - omega)), k, mu = lambda)
}

rznbsh <- function(n, k, lambda, omega) {
  ifelse(rbinom(n, 1, omega), 0, rnbinom(n, k, mu = lambda))
```

```
}


fit1<-fitdist(data=hiv$Number.of.HEI.Sero.conversion,distr="nbinom",method="mle")
summary(fit1)
fit3<-fitdist(hiv$Number.of.HEI.Sero.conversion,"nbsh",start = list(k=1,lambda=1),lowe
summary(fit3)
fit4<-fitdist(data=hiv$Number.of.HEI.Sero.conversion,distr="zinb",
lower=c(0,0),start=list(k=1,lambda=1,omega=0.5))
summary(fit4)
fit5<-fitdist(hiv$Number.of.HEI.Sero.conversion, distr = "znbsh",
start = list(lambda=1,omega=0.5,k=1),method="mle")
summary(fit5)




a<-rznbsh(n=1000,omega=0.4141,lambda=rshanker(1,exp_rate=2.34291),k=1)
b<-rnbsh(n=1000,k=1,lambda=rshanker(1,exp_rate=2.34291))
c<-rnbinom(n=1000,mu=0.4919530,size=0.2432295)
d<-rzinb(n=1000, k= 0.5574 , lambda=0.8396,omega=0.4141)

hist(a,prob=T,main="Simulations",xlab="Number␣of␣occurrences")
lines(density(a),lty=1)
lines(density(b),lty=2)
lines(density(c),lty=3)
lines(density(d),lty=4)
legend('topright', legend=c("zinbsh", "nbinom","zinb","nbsh"),
       lty = c(1,2,3,4), cex=0.75)




#ggplot
temp <- as.data.frame(table(hiv$Number.of.HEI.Sero.conversion))
names(temp)<-c('no.','freq')
temp
temp$no.<-as.numeric(as.character(temp$no.))
temp$Nmod <- temp$freq / sum(temp$freq)
#temp$pois <- dpois(temp$no., lambda = mean(temp$no.))
temp$nbinom <- dnbinom(temp$no., size=0.2432295,mu=0.4919530)
temp$nbsh <-dnbsh(temp$no.,k=0.2433, lambda=0.49199)
temp$zinb<-dzinb(temp$no.,omega=0.4141,k= 0.5574 , lambda=0.8396)
temp$znbsh<-dznbsh(temp$no., lambda=0.8406,omega=0.4141,k=0.5584)
ggplot(temp, aes(x=no., y= Nmod)) +
  geom_histogram(stat="identity", binwidth = 2.5) +
  theme(panel.grid.minor.x=element_blank(),
        panel.grid.major.x=element_blank()) +
  geom_line(aes(y = nbinom, color = "znbsh"))+
  geom_line(aes(y = nbsh, color = "nbinom"))+
  geom_line(aes(y = zinb, color = "zinb"))+
  geom_line(aes(y = znbsh, color = "nbsh"))+
  scale_colour_manual("",values = c("red", "green", "blue","maroon"))+
  ggtitle("Comparison␣of␣Density␣Plots") +
  xlab("No␣of␣Infants") + ylab("Frequency")



##############################################################################
```

```r
#NB ROOTOGRAM
n <- 1000
male <- sample(c(0,1), size = n, replace = TRUE)
z <- rbinom(n = n, size = 1, prob = 0.1)
# mean(z == 0)
y_sim <- ifelse(z == 0, 0,
                rnbinom(n=n,mu=(0.4919530*male==1),size=0.2432295
                ))
table(y_sim, male)
barplot(table(y_sim))


#plot rootograms
library(pscl)
nb_gram <- glm.nb(y_sim ~ male)
summary(nb_gram)
countreg::rootogram(nb_gram)


#NBSH ROOTOGRAM
male <- sample(c(0,1), size = n, replace = TRUE)
z <- rbinom(n = n, size = 1, prob = 0.1)
# mean(z == 0)
y_sim <- ifelse(z == 0, 0,
                rnbsh(n=n,k=0.2433,lambda=rshanker(1,exp_rate=2.34291)
                ))
table(y_sim, male)
barplot(table(y_sim))


#plot rootograms
library(pscl)
nbsh_gram <- glm.nb(y_sim ~ male)
summary(nbsh_gram)
countreg::rootogram(nbsh_gram)


#ZINB ROOTOGRAM
male <- sample(c(0,1), size = n, replace = TRUE)
z <- rbinom(n = n, size = 1, prob = 0.1)
# mean(z == 0)
y_sim <- ifelse(z == 0, 0,
                rzinb(n=n, k= 0.5574 , lambda=0.8396,omega=0.0
                ))
table(y_sim, male)
barplot(table(y_sim))


#plot rootograms
library(pscl)
zinb_gram <- zeroinfl(y_sim ~ male | 1, dist = "negbin")
summary(zinb_gram)
countreg::rootogram(zinb_gram)


#ZINBSH ROOTOGRAM

male <- sample(c(0,1), size = n, replace = TRUE)
z <- rbinom(n = n, size = 1, prob = 0.1)
# mean(z == 0)
y_sim <- ifelse(z == 0, 0,
rznbsh(n=n,omega=0.9,lambda=(rshanker(1,exp_rate=2.34291)*male==1),k=0.5584
                ))
```

```r
table(y_sim, male)
barplot(table(y_sim))


#plot rootograms
library(pscl)
zinbsh_gram <- zeroinfl(y_sim ~ male | 1, dist = "negbin")
summary(zinbsh_gram)
countreg::rootogram(zinbsh_gram)
```

References

Aryuyuen, S., Bodhisuwan, W., & Supapakorn, T. (2014). Zero inflated negative binomial-generalized exponential distribution and its applications. *Songklanakarin J. Sci. Technol*, *36*(4), 483–491.

Blasco-Moreno, A., Pérez-Casany, M., Puig, P., Morante, M., & Castells, E. (2019). What does a zero mean? understanding false, random and structural zeros in ecology. *Methods in Ecology and Evolution*, *10*(7), 949–959.

Cohen, A. C. (1963). *Estimation in mixtures of discrete distributions*. Statistical Pub. Society.

Desjardins, C. D. (2013). Evaluating the performance of two competing models of school suspension under simulation-the zero-inflated negative binomial and the negative binomial hurdle.

Diebolt, J., & Robert, C. P. (1994). Estimation of finite mixture distributions through bayesian sampling. *Journal of the Royal Statistical Society: Series B (Methodological)*, *56*(2), 363–375.

Hadi, M. A., Aruldhas, J., Chow, L.-F., & Wattleworth, J. A. (1995). Estimating safety effects of cross-section design for various highway types using negative binomial regression. *Transportation Research Record*, *1500*, 169.

Hall, D. B. (2000). Zero-inflated poisson and binomial regression with random effects: a case study. *Biometrics*, *56*(4), 1030–1039.

Hilbe, J. M. (2014). *Modeling count data*. Cambridge University Press.

Hua, H., Wan, T., Wenjuan, W., & Paul, C.-C. (2014). Structural zeroes and zero-inflated models. *Shanghai archives of psychiatry*, *26*(4), 236.

Jiang, Y., & House, L. A. (2017). Comparison of the performance of count data models under different zero-inflation scenarios using simulation studies.

Khoshgoftaar, T. M., Gao, K., & Szabo, R. M. (2001). An application of zero-inflated poisson regression for software fault prediction. In *Proceedings 12th international symposium on software reliability engineering* (pp. 66–73).

Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, *34*(1), 1–14.

Loquiha, O., Hens, N., Chavane, L., Temmerman, M., Osman, N., Faes, C., & Aerts, M. (2018). Mapping maternal mortality rate via spatial zero-inflated models for count data: A case study of facility-based maternal deaths from mozambique. *PloS one*, *13*(11), e0202186.

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of econometrics*, *33*(3), 341–365.

Mwalili, S. M., Lesaffre, E., & Declerck, D. (2008). The zero-inflated negative binomial regression model with correction for misclassification: an example in caries research. *Statistical methods in medical research*, *17*(2), 123–139.

Nekesa, F., Odhiambo, C., & Chaba, L. (2019). Comparative assessment of zero-inflated models with application to hiv exposed infants data. *Open Journal of Statistics*, *9*(6), 664–685.

Nyamhanga, T., Frumence, G., & Simba, D. (2017). Prevention of mother to child transmission of hiv in tanzania: assessing gender mainstreaming on paper and in practice. *Health policy and planning*, *32*(suppl_5), v22–v30.

Opitz, T., Tramini, P., & Molinari, N. (2013). Spline regression for zero-inflated models. *arXiv preprint arXiv:1304.3347*.

Osgood, D. W. (2000). Poisson-based regression analysis of aggregate crime rates. *Journal of quantitative criminology*, *16*(1), 21–43.

Ridout, M., Demétrio, C. G., & Hinde, J. (1998). Models for count data with many zeros. In *Proceedings of the xixth international biometric conference* (Vol. 19, pp. 179–192).

Saengthong, P., Bodhisuwan, W., & Thongteeraparp, A. (2015). The zero inflated negative binomial–crack distribution: some properties and parameter estimation. *Songklanakarin Journal of Science & Technology*, *37*(6), 701–711.

Shanker, R. (2015). Shanker distribution and its applications. *International Journal of Statistics and Applications*, *5*(6), 338–348.

Shanker, R., et al. (2016). Sujatha distribution and its applications. *Statistics in Transition. New Series*, *17*(3), 391–410.

Shapiro, R. L., & Lockman, S. (2010). *Mortality among hiv-exposed infants: the first and final frontier*. The University of Chicago Press.

Tlhaloganyang, B. P., Mooketsi, D. R., Leinanyane, L., & Sakia, R. (n.d.). A compound of generalized negative binomial and shanker distribution.

Warton, D. I. (2005). Many zeros does not mean zero inflation: comparing the goodness-of-fit of parametric models to multivariate abundance data. *Environmetrics: The official journal of the International Environmetrics Society*, *16*(3), 275–289.

Yamrubboon, D., Thongteeraparp, A., Bodhisuwan, W., & Jampachaisri, K. (2017). Zero inflated negative binomial-sushila distribution and its application. In *Aip conference proceedings* (Vol. 1905, p. 050044).

Yang, S. (2014). A comparison of different methods of zero-inflated data analysis and its application in health surveys.

Zou, G. (2004). A modified poisson regression approach to prospective studies with binary data. *American journal of epidemiology*, *159*(7), 702–706.