Enrollment No.: **2017IMSST008**

Convolution of Mixed Poisson distribution for modeling count data with equi-, under- and overdispersion, and its applications

Project Report

submitted in partial fulfillment of the requirements for the award of degree of

Integrated Master of Sciences

in

Statistics

Submitted By:
Shilpa Soni
(2017IMSST008)

Supervisor:
Dr. Deepesh Bhati
(Assistant Professor)

Department of Statistics
School of Statistics, Mathematics and Computational Sciences
Central University of Rajasthan, Ajmer - 305817

2017-2022

Dr. Deepesh Bhati
Assistant Professor
Department of Statistics

Email : deepesh.bhati@curaj.ac.in
Website : http://www.curaj.ac.in

राजस्थान केन्द्रीय विश्वविद्यालय
Central University of Rajasthan
(संसद के अधिनियम के तहत स्थापित केन्द्रीय विश्वविद्यालय)
(A Central University by an Act of Parliament)
NH-8, Bandarsindri, 305801
Kishangarh (Ajmer), Rajasthan, INDIA
Phone (Office): +91-1463-238755/260200
Telefax: +91-1463-238722

Certificate

This is to certify that the work embodied in the accompanying project report entitled ***"Convolution of Mixed Poisson distribution for modeling count data with equi-, under- and overdispersion, and its applications"*** has been successfully carried by ***Shilpa Soni (2017IMSST008)***, a X Semester student, of the Department of Statistics, Central University of Rajasthan, under my guidance and supervision.

To the best of my knowledge, the outcomes/findings incorporated in this thesis have not been presented to any other University/Institute for the award of any other Degree/Diploma.

She worked from January 2022 to May 2022 and her work carried out is satisfactory.

Place: Bandarsindri

Date:

Dr. Deepesh Bhati

(Supervisor)

TO WHOMSOEVER IT MAY CONCERN

This is to certify that ***Shilpa Soni***, Enrollment No. 2017IMSST008, Integrated M.Sc. Statistics student of Department of Statistics has done project work ***"Convolution of Mixed Poisson distribution for modeling count data with equi-, under- and overdispersion, and its applications"*** under the guidance of ***Dr. Deepesh Bhati***, Department of Statistics, Central University of Rajasthan towards the partial fulfillment of the award of ***"Master of Science in Statistics"*** during the period January 2022 to May 2022.

Head

Department of Statistics

Central University of Rajasthan

Declaration

I, **Shilpa Soni**, hereby declare that the project work entitled **"Convolution of Mixed Poisson distribution for modeling count data with equi-, under- and overdispersion, and its applications"** submitted to Department of Statistics, Central University of Rajasthan as a partial fulfillment of requirements of X-Semester examination, is a bona fide record of work under taken by me, under the supervision of **Dr. Deepesh Bhati**, Designation at Department of Statistics, Central University of Rajasthan and it has never been presented to any other university or institution for award of any other Degree/Associate-ship or Fellowship.

I certify that all sources and data are fully compliant in this sense and the similarities are within the allowed range.

Signature of Candidate

Shilpa Soni

Enroll. No.: 2017IMSST008

Place: Ajmer

Date:

Acknowledgement

It gives me immense pleasure to express my deep sense gratitude and indebtedness to my supervisor **Dr. Deepesh Bhati** for his valuable support and encouraging mentality throughout the project. I am highly thankful to him for providing me this opportunity to carry out the idea and work during my project and helping me to gain the successful completion of my project.

Besides my guide, I am also thankful to **Ishfaq Shah** and **Girish Aradhye**, for their encouraging and insightful comments on the challenging parts of my thesis.

Lastly, a special thanks to **Richa Kumari** for being my support system, who encouraged me throughout my hardships and pushed me towards excellence.

Shilpa Soni

Department of Statistics

Central University of Rajasthan

Dedicated to my sister, Richa.

Your passion inspires me.

Contents

Abstract

Mixed Poisson distribution as a model for count data has been presented. It can be used for modeling count data with equidispersion, underdispersion and overdispersion. A couple of exemplary sub-models have been proposed and their main properties are derived, such as probability density function, probability generating function, moment generating function and subsequently, moments, which are all obtained in elegant unconditional closed forms. The maximum likelihood estimation method is used for estimating the model parameters. The proposed model is then, fitted to two count data illustrating its capabilities in the challenging cases of overdispersed and underdispersed count data. The freedom of the choice of different underlying distributions has been entertained which will lead to further research and possibly the application of mpre complex techniques and even more applicability.

Chapter 1

Introduction

Modeling count data is of great research interest across all disciplines known to mankind. This problem as been addressed by multiple researchers over the years. One of the reasons for this is the ginormous amount of varied data available to us at-hand. This has resulted in the introduction and evolution of multiple models. In spite of this, as of today, there is unavailability of a a unifying approach to modelling all series data of counts.

Many of the earlier works used the Poisson distribution as a integral part of the process. But, the Poisson distribution is unsuitable on real datasets many-a-times as a model due to the fact that the corresponding assumption of the relationship between the mean and variance is heavily violated. The observation of overdispersed or underdispersed counts is fairly common. One may use the more flexible mixed generalised family of distributions which has been illustrated in the present paper (Less generalised cases have been illustrated in Marcelo Bourguignon and Christian H. Weiß's paper of 2017 [2], Consul's 1989 paper [6], Lambert's 1992 paper [7] and Puig's paper of 2006 [8]). In a way, one is assuming that the parameter of the distribution in question is itself is a random variable with some distribution. Thus, the aim of this paper is to bring together results concerning the modeling of count data using the Mixed Poisson distribution as introduced by Karlis and Xekalaki, 2005 ([1])

1.1 Research Objectives of the Work

1. Generalization/modification of classical count distributions such as Poisson and Geometric.
2. To study the applicability of proposed count models generated by Mixed Poisson distribution on general datasets like Grainger and Reid and Chan, Riley *et al.*
3. To model count data with excess of zeros and possessing over dispersion as well as underdispersion.
4. Further study and inference of the class of count models generated by Mixed Poisson distribution has been done.

1.2 Proposed Problems

• While dealing with count data, there are some limitations to look for as: Over-dispersion and Under-dispersion, Excess of zeros/ones or zero-vertex modality..

• Mainly three standard distributions have been used to model count data namely Poisson distribution, Negative Binomial distribution and Geometric distribution. As the basic characteristic of Poisson distribution is its equi-dispersion property. The problem is that when modeling real data, the equi-dispersion criterion is rarely satisfied and grossly violated. Analysts usually must adjust their Poisson model in some way to account for any under- or over-dispersion that is in the data.

• Over-dispersion is, by far, one of the foremost problem facing analysts who use Poisson regression when modeling count data. Thus, one has to look for such an model which will accommodate over-dispersion.

• The Negative Binomial distribution allows more flexibility in modeling over-dispersed data than does a single-parameter Poisson model. The negative binomial is derived as a Poisson-gamma mixture model, with the

dispersion parameter being distributed as gamma shaped. The gamma PDF is pliable and allows for a wide variety of shapes. As a consequence, most overdispersed count data can be appropriately modeled using a negative binomial regression. The advantage of using the negative binomial rests with the fact that when the dispersion parameter is zero (0), the model is Poisson. Values of the dispersion parameter greater than zero indicate that the model has adjusted for correspondingly greater amounts of overdispersion. But, it is pertinent to mention that even though the negative binomial model adjusts for the Poisson over-dispersion; however it cannot be used to model under-dispersed Poisson data. Thus, we can see that if one problem of over-dispersion is dealt with with by a model, a new problem crops us.

- Because of under/over dispersion, the generalized Poisson distribution has been introduced. The generalized Poisson distribution has a second parameter, also referred to as the dispersion or scale parameter. Also like the previously introduced models, the generalized Poisson reduces to Poisson when the dispersion is zero. The nice feature of the generalized Poisson, however, is that the dispersion parameter can have negative values, which indicate an adjustment for Poisson under-dispersion.
- Keeping in view the applications of Count data and based on the limitations of standard models, the proposed work will focus on development of further extension of count models in order to overcome the shortcomings of classical models and make advancements in applications to solve complex real life problems.
- The proposed work will intensively look to overcome these limitations and make further contributions to count data models generated by Poisson distribution and its application in various fields like Medical, Insurance, Biometrics, Econometrics and much more.

As for preliminaries, a random variable X follows a mixed Poisson distribution with mixing distribution having probability density function g if its probability function is given by

$$P(X = x) = P(x) = \int_0^\infty \frac{e^{-\lambda}\lambda^x}{x!} g(\lambda) d\lambda; \quad x = 0, 1, \ldots \quad (1.1)$$

Here, λ can be assumed to follow any distribution. We are looking at cases where λ follows any of the following distributions:

- Exponential, and
- Lindley.

This paper introduces the BerMP distribution as a convolution of the Bernoulli and Mixed Poisson distributions, where λ is assumed to follow a distinct distribution.

The proposed models find application and solve the problem of any deviations from the equidispersed case really well. We can fine-tune model by choosing the underlying distribution of the parameter of the Mixed Poisson distribution to obtain the perfect fit for the data-set and also modify them to get the desired result.

The project report is organized as follows:

In chapter 2, we lay down the research objectives, introduce the preliminary (baseline) distribution the Mixed Poisson distribution and give essential results. After that we have introduce the actual convoluted Mixed Poisson model with a formal definition. We have also formally derived the PGF of the proposed distribution.

In chapter 3, we have formulated the subcases of using a different underlying distribution of the paramater of the Mixed Poisson distribution, *lambda* and derived the various statistical properties of each of the models.

In chapter 4, we also undertook our estimation procedure and found maximum likelihood estimates of the parameters.

After this, in chapter 5, we carry out the data analysis of two datasets, specifically Dental Caries and glucocorticoid and mineralocorticoid receptors (GR and MR) data and try to fit the model appropriately.

In the penultimate chapter 6, we conclude our paper by summarizing the various results obtained and also look for future prospective studies that can be carried out with the model.

Lastly, we lay out the references in chapter 7.

Chapter 2

Genesis of a new family of Count Models

2.1 Mixed Poisson Distribution with its properties

We have used the Mixed Poisson distribution as proposed by Karlis and Xekalaki, 2005 ([1]) as our second baseline distribution for the convolution., the first being Bernoulli. We are specifically using the mixed Poisson Distribution because of its credibility with dealing with over- and under-dispersed data. An excellent reference for mixed Poisson distributions is the book by Grandell (1997).

A random variable X follows a mixed Poisson distribution with mixing distribution having probability density function f if its probability function is given by:

$$P(X=x)=P(x)=\int_0^\infty \frac{e^{-\lambda}\lambda^x}{x!}f(\lambda)d\lambda, \quad x=0,1,\ldots, \quad (2.1)$$

Here, the interesting thing is that we have the freedom to choose the distribution of λ, $f(\lambda)$, as we like. As we will see later, we have chosen two exemplary distributions: Exponential and Lindley, which yield the corresponding Geometric Mixed Poisson Distribution and Lindley Mixed Poisson

Distributions, respectively.

What is also even more interesting is the sheer simplicity of the expressions that we will obtain. We could have used the stochastic representation to segregate the Poisson and the Exponential parts, after integration, but we don't need to, because of the unconditional closed form of the PMF. After which, the density that we obtained can be used to obtain the Log-Likelihood to estimate the unknown parameters. This process now becomes a straightforward one because of this reason and evalutaion of further statistical properties become easier.

We will also take two datasets to illustrate that our we can obtain beautifully simple models like the Geometric Mixed Poisson Distribution that perform a whole lot better than the conventional models. This is a very interesting result even more so because we have obtained it in an **unconditional closed form**, which is rare. This is miraculous because one of the greatest challenges we face when dealing with Mixed Distribution is that we don't get the PDF and/or PMF in a closed form most of the times.

In terms of the probability generating function, $G_X(s)$, of X, (equation 2.1) can be written in the form:

$$G_X(s) = \int_0^\infty e^{\lambda(s-1)} f(\lambda) d\lambda \tag{2.2}$$

It can be useful to note that the right hand side of this equation is $M_\lambda(s-1)$, the moment generating function of the mixing distribution evaluated at $(s-1)$.

Also λ is not necessarily a continuous random variable. It can be discrete or it can take a finite number of values. The latter case gives rise to finite Poisson mixtures.

The moments about the origin of the mixed Poisson distribution in terms

of those of the mixing distribution.

So,

$$\mathbb{E}(X) = \mathbb{E}(\lambda) \quad \text{and} \quad \mathbb{E}(X^2) \ = \mathbb{E}(\lambda^2) + \mathbb{E}(\lambda) \tag{2.3}$$

We have for the variance of the Poisson distribution that:

$$\begin{aligned} \mathbb{V}(X) &= \mathbb{E}(X^2) - [\mathbb{E}(X)]^2 \\ &= \mathbb{E}(\lambda^2) + \mathbb{E}(\lambda) - [\mathbb{E}(\lambda)]^2 \\ &= \mathbb{E}(\lambda) + \mathbb{V}(\lambda) \end{aligned} \tag{2.4}$$

It's fairly obvious that the variance of a mixed Poisson distribution is always greater than the variance of a simple Poisson distribution with the same mean.

In the next section, we see the convolution of Bernoulli and Mixed Poisson forming shape.

2.2 Convolution of Bernoulli and Mixed Poisson DIstribution

Let us consider a random variable Z such that $Z := X + Y$, where X and Y are two independent random variables and follow Bernoulli Distribution with parameter π such that $0 < \pi < 1$ and Mixed Poisson family of distribution (Karlis and Xekalaki (2005), [1]) with parameter λ such that $\lambda > 0$, respectively.

The probability mass function (PMF) of the random variable Z, is given by:

$$\begin{aligned} P(Z=z) &= \pi \int_0^\infty \frac{e^{-\lambda}\lambda^{z-1}}{(z-1)!} f(\lambda)d\lambda + (1-\pi)\int_0^\infty \frac{e^{-\lambda}\lambda^{z}}{z!} f(\lambda)d\lambda \\ &= \frac{1}{(z-1)!}\int_0^\infty e^{-\lambda}\lambda^{z-1}\left(\pi + \frac{(1-\pi)\lambda}{z}\right) f(\lambda)d\lambda, \end{aligned} \tag{2.5}$$

$$z = 0, 1, \ldots$$

Now from (2.5) we can make different appropriate choices of $f(\lambda)$ to get the new family of distributions. The notation for new family of distributions would be $Z \sim \mathscr{M}\mathscr{P}\mathscr{B}(\pi, \lambda)$ where $\pi \in (0,1)$ and mixing density will have the support of $\lambda > 0$.

2.3 Distributional Properties

Theorem 1. *Let $Z \sim \mathscr{MPB}(\pi, \lambda)$, then the probability generating function (PGF) is given by*

$$G_Z(s) = [1 + \pi(s-1)]\, M_\lambda(s-1), \quad |s| < 1, \tag{2.6}$$

where $M_\lambda(s-1)$ is the moment generating function (M.G.F.) of the mixing density function.

Proof. Since, by definition, we have $Z = X + Y$, where X and Y are two independent random variables following Bernouli with parameter π $(0 < \pi < 1)$, and Mixed Poisson family of distribution, respectively. Therefore, we can write:

$$\begin{aligned}
G_Z(s) &= G_{X+Y}(s) \\
&= G_X(s) \cdot G_Y(s) \\
&= [1 + \pi(s-1)]\; G_Y(s) \\
&= [1 + \pi(s-1)]\; \mathbb{E}(s^y) \\
&= [1 + \pi(s-1)]\; \sum_{y=0}^{\infty} s^y P_Y(y) \\
&= [1 + \pi(s-1)]\; \sum_{y=0}^{\infty} s^y \left[\int_0^\infty \frac{e^{-\lambda}\lambda^y}{y!} g(\lambda) d\lambda \right] \\
&= [1 + \pi(s-1)]\; \int_0^\infty e^{-\lambda} \left[\sum_{y=0}^{\infty} \frac{(s\lambda)^y}{y!} \right] g(\lambda) d\lambda \\
&= [1 + \pi(s-1)]\; \int_0^\infty e^{\lambda(s-1)} g(\lambda) d\lambda \\
&= [1 + \pi(s-1)]\; \mathbb{E}_\lambda(e^{-\lambda(s-1)})
\end{aligned}$$

$$= \ [1+\pi(s-1)]\ M_\lambda(s-1).$$

$$\implies G_Z(s) = [1+\pi(s-1)]\ M_\lambda(s-1)$$

Hence, proved.

□

Remark 1. *Replacing s by e^s in* (2.6)*, we will get the MGF of the random variable Z, as:*

$$M_Z(s) = [1+\pi(e^s-1)]\ M_\lambda(e^s-1). \tag{2.7}$$

This result can be used to find all the moments of the distribution.

Chapter 3

The Proposed Models

In this chapter, we will formally define our model from scratch and also establish some essential results.

3.1 Geometric Mixed Poisson Distribution

Choosing $f(\lambda)$ as exponential distribution with parameter $\theta > 0$ in (2.5), we will get a new distribution named as Geometric Mixed Poisson Distribution denoted by $\mathscr{MPB}(\pi,\theta)$ ([3]) with PMF given by

$$P(Z=z) = (\theta+\pi)\frac{\theta^{z-1}}{(1+\theta)^{z+1}}; \quad z = 0,1,\ldots, \tag{3.1}$$

such that $\theta > 0$ and $\pi \in (0,1)$. The PMF plot of $\mathscr{MPG}(\pi,\theta)$ for different choices of paramameters are displayed in figure 3.1

The MGF of exponential distribution with probability density function $\frac{1}{\theta}e^{-\frac{x}{\theta}}$ is given by:

$$M_X(s) = \frac{1}{1-\theta s}. \tag{3.2}$$

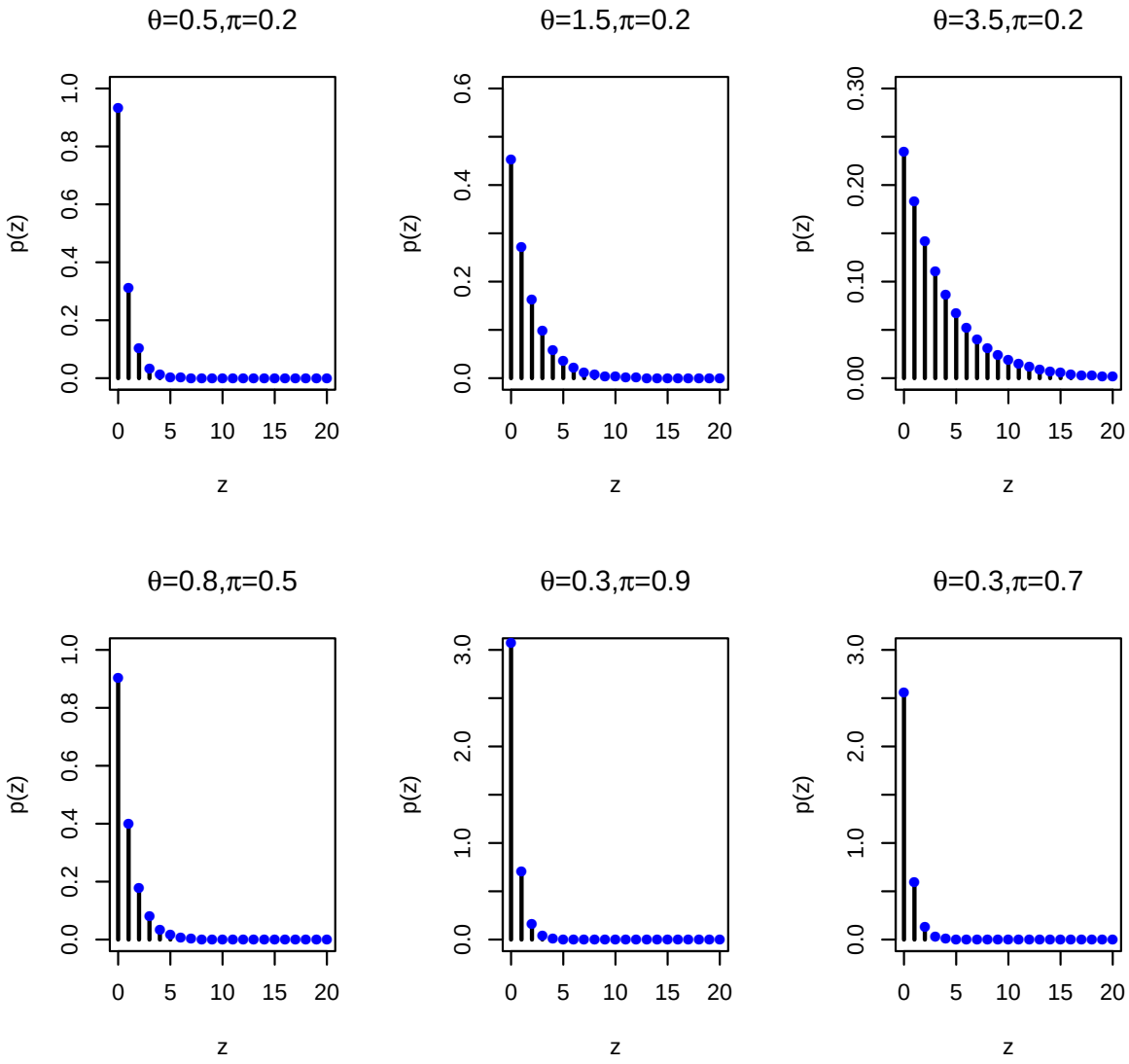

Figure 3.1: Plot of PMF of MPG for different values of parameters

From Theorem (2.6), the PGF of $\mathscr{MPG}(\pi,\theta)$ is

$$G_Z(s) = \frac{1+\pi(s-1)}{1-\theta(s-1)}. \tag{3.3}$$

Replacing s by e^s in (3.3), we will get the MGF of $\mathscr{MPG}(\pi,\theta)$ as

$$M_Z(s) = \frac{1+\pi(e^s-1)}{1-\theta(e^s-1)}. \tag{3.4}$$

Therefore, the mean and variance of $\mathscr{MPG}(\pi,\theta)$ can be obtained from (3.4) after doing some simple calculations, as under

$$\mathbb{E}(Z) = \pi+\theta \quad \textit{and} \quad \mathbb{V}(Z) = (\pi+\theta)\left[1+\theta-\pi\right]. \tag{3.5}$$

The index of dispersion $(\mathbb{ID})$ (the ratio of variance to mean) is given by

$$\mathbb{ID} = \frac{(\pi+\theta)\left[1+\theta-\pi\right]}{\pi+\theta} = 1+\theta-\pi. \tag{3.6}$$

From result (3.6), it is clear that this model can be

1. Equidispersed ($\mathbb{ID}=1$), when $\theta=\pi$.
2. Underdispersed ($\mathbb{ID}<1$), when $\theta<\pi$.
3. Overdispersed ($\mathbb{ID}>1$), when $\theta>\pi$.

3.2 Lindley Mixed Poisson Distribution

Choosing $f(\lambda)$ as Lindley distribution with parameter $\theta > 0$, we will get a new distribution named Lindley Mixed Poisson Distribution denoted by $\mathscr{MPL}(\pi,\theta)$ with PMF given by

$$P(Z=z) = \frac{\theta^2}{(\theta+1)^{z+3}}\,(1+\pi\theta); \quad z=0,1,\ldots \tag{3.7}$$

such that $\theta > 0$ and $\pi \in (0,1)$. The PMF plot of $\mathscr{MPG}(\pi,\theta)$ for different choices of paramameters are displayed in figure 3.1

The PGF of the Lindley distribution with probability density function $\frac{\theta^2}{\theta+1}\,(1+\lambda)e^{-\theta\lambda}$ is given by:

$$G_X(s) = \frac{\theta^2 e^{-\theta\lambda}}{(2-s)}\,\frac{(1+\lambda)}{(1+\theta)}. \tag{3.8}$$

From Theorem (2.6), the PGF of $\mathscr{MPL}(\pi,\theta)$ is

$$G_Z(s) = \frac{\theta^2 e^{-\theta\lambda}}{(2-s)}\,\frac{(1+\lambda)}{(1+\theta)}\,\big(\pi(s-1)+1\big). \tag{3.9}$$

Replacing s by e^s in (3.9), we will get the MGF of $\mathscr{MPL}(\pi,\alpha,\beta)$ as

$$M_Z(s) = \frac{\theta^2 e^{-\theta\lambda}}{(2-e^s)}\,\frac{(1+\lambda)}{(1+\theta)}\,\big(\pi(e^s-1)+1\big). \tag{3.10}$$

Therefore, the mean and variance of $\mathscr{MPL}(\pi,\theta)$ can be obtained from

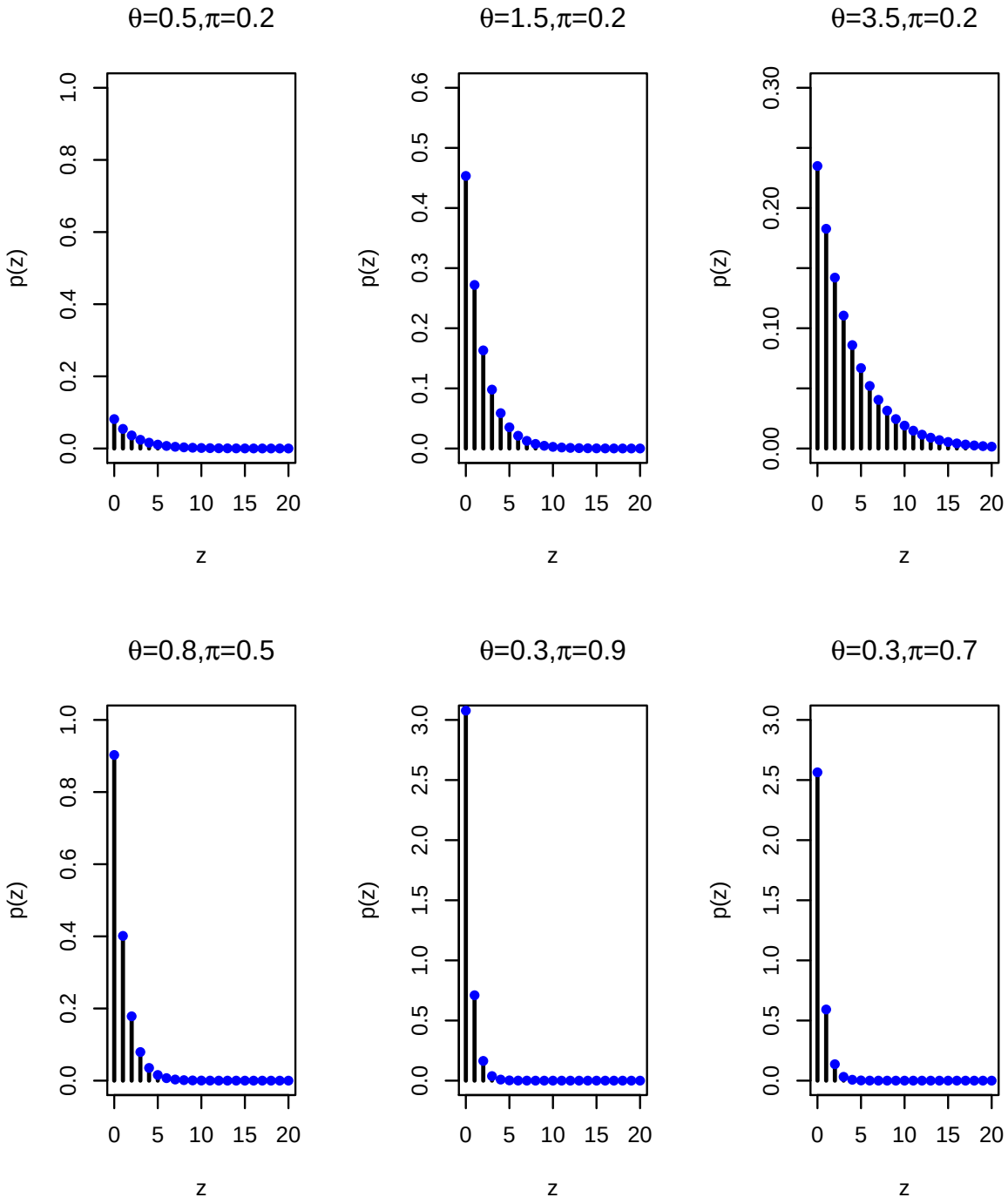

Figure 3.2: Plot of PMF of MPL for different values of parameters

(3.10) after doing some simple calculations, as under:

$$\mathbb{E}(Z) = \frac{\theta^2 e^{-\theta\lambda}}{(2-s)} \frac{(1+\lambda)}{(1+\theta)} (1+\pi) \quad \textit{and}$$
$$\mathbb{V}(Z) = \frac{(1+\pi)e^{-2\theta\lambda} \left(3(\theta+1)e^{\theta\lambda} - \pi - 1\right)}{(\theta+1)^2}. \tag{3.11}$$

The index of dispersion($\mathbb{ID}$) (ratio of variance to mean) is given by:

$$\mathbb{ID} = 3 - \frac{(1+\pi)}{(1+\theta)} e^{-\theta\lambda}. \tag{3.12}$$

From result (3.12), it is clear that this model can be

1. Equidispersed ($\mathbb{ID} = 1$), when $\frac{(1+\pi)}{(1+\theta)} e^{-\theta\lambda} = 2$.
2. Underdispersed ($\mathbb{ID} < 1$), when $\frac{(1+\pi)}{(1+\theta)} e^{-\theta\lambda} < 2$.
3. Overdispersed ($\mathbb{ID} > 1$), when $\frac{(1+\pi)}{(1+\theta)} e^{-\theta\lambda} > 2$.

In the next chapter, we will use the MLE technique to estimate the unknown parameters of the two proposed distributions. We haven't used the EM Algorithm for this estimation for two reasons. The PMF we have obtained is beautifully simple and is of a closed form which is rare. Employing the EM algorithm in such a case would have been a massive overkill. Although, we are open to using the EM Algorithm for other distributions which we will do in subsequent future research.

Secondly, MLE technique yielded very accurate estimates in our experience. Whichever initial seeds we chose, we arrived at the same results, which is a really good indication of the goodness of the model and the estimation technique as well. Hence, the regular maximisation algorithm or the Maximum Likelihood Estimation Technique served its purpose elegantly for our study and we didn't need to complicate our analysis any further by developing a complicated EM Algorithm.

We will see that in hindsight, our choice of the estimation technique will save us time as the estimates we obtain will **converge**. Everytime, we tried a new initial value, our likelihood converged to the same value. This would mean that our usual likelihood estimation procedure will converge, which does happen, as we will see later.

Chapter 4

Estimation

4.1 Maximum Likelihood Estimation

4.1.1 Geometric Mixed Poisson Distribution

Suppose $\underline{\mathbf{z}} = \{z_1, z_2, \ldots, z_n\}$ be a random sample of size n from the $\mathscr{MPG}(\pi, \theta)$ with pmf given in (3.1).

$$L(\theta, \pi|\underline{\mathbf{z}}) = \prod_{i=1}^{n}(\theta+\pi)\frac{\theta^{z_i-1}}{(1+\theta)^{z_i+1}} \tag{4.1}$$

The log-likelihood function is:

$$\begin{aligned}\log L(\theta, \pi, |\underline{\mathbf{z}}) = &\sum_{i=1}^{n}\log(\theta+\pi) + \sum_{i=1}^{n}\log\theta^{z_i-1} \\ &- \sum_{i=1}^{n}\log(1+\theta)^{z_i+1}\end{aligned} \tag{4.2}$$

The ML estimates $\hat{\theta}$ of θ and $\hat{\pi}$ of π, respectively, can be obtained by solving equations:

$$\frac{\partial \log L}{\partial \theta} = 0, \text{ and } \quad \frac{\partial \log L}{\partial \pi} = 0 \quad .$$

Since there are two parameters involved in the model, therefore, we will obtain two normal equations which are of implicit forms and are complex to be solved further numerically. So, we make use of the Mathematica Software 9.0 to find the estimates numerically by using the "NMaximize" function.

4.1.2 Lindley Mixed Poisson Distribution

Suppose $\underline{\mathbf{z}} = \{z_1, z_2, \ldots, z_n\}$ be a random sample of size n from the $\mathscr{MPL}(\pi, \theta)$ with pmf given in (3.7).

$$L(\theta, \pi|\underline{\mathbf{z}}) = \prod_{i=1}^{n} \frac{\theta^2}{(\theta+1)^{z_i+3}} (1+\pi\theta); \qquad z = 0, 1, \ldots \tag{4.3}$$

The log-likelihood function is:

$$\log L(\theta, \pi, |\underline{\mathbf{z}}) = \sum_{i=1}^{n} \log \theta^2 - \sum_{i=1}^{n} \log(\theta+1)^{z_i+3} + \sum_{i=1}^{n} \log(1+\pi\theta). \tag{4.4}$$

The ML estimates $\hat{\theta}$ of θ and $\hat{\pi}$ of π, respectively, can be obtained by solving equations:

$$\frac{\partial \log L}{\partial \theta} = 0, \text{ and } \quad \frac{\partial \log L}{\partial \pi} = 0 \quad .$$

Since there are two parameters involved in the model, therefore, we will obtain two normal equations which are of implicit forms and are complex to be solved further numerically. So, we make use of the Mathematica Software 9.0 to find the estimates numerically by using the "NMaximize" function.

Chapter 5

Data Analysis

To explore and check the practical potential of the proposed models, two data sets: *Grainger and Reid; 1954 ([9]) and Chan, Riley, Price, McElduff, Winyard, Welham and Long, 2009([10])*, have been taken into consideration.

Models like Poisson, $P(\lambda)$ and Negative Binomial, $NB(r,p)$ will be compared with the proposed models; Geometric Mixed Poisson, $MPG(\pi,\theta)$, Lindley Mixed Poisson, $MPL(\pi,\theta)$ distributions. These are being fitted to the given data sets and the parameters of each model have been estimated by Maximum Likelihood method(ML).

Based on the results of the Log-Likelihood (LL) values, the Akaike's information criterion (AIC) (*Akaike, 1973* [4]) and Bayesian information criterion (BIC) (*Schwartz, 1978* [5]), it can be said that there exists enough statistical evidenceto suggest that the Geomtric Mixed Poisson, $MPG(\pi,\theta)$ distribution fits the data really well and one can clearly see that the proposed model outperforms the competing models in terms of LL, AIC and BIC values and therefore establishes the superiority of the proposed model. (as shown in Table 5.1 and Table 5.2).

Table 5.1: Number of smooth surfaces affected by dental caries

Model	Estimates	LL	AIC	BIC
Poisson (λ)	$\hat{\lambda} = 1.97$	-342.456	686.912	687.076
NB(r, p)	$\hat{r} = 0.830, \hat{p} = 0.297$	-276.711	557.422	557.76
$\mathscr{M}\mathscr{P}\mathscr{G}(\pi, \theta)$	$\hat{\pi} = 0.893$, $\hat{\theta} = 0.973$	-199.599	403.198	403.527
$\mathscr{M}\mathscr{P}\mathscr{L}(\pi, \theta)$	$\hat{\pi} = 0.95$, $\hat{\theta} = 1.01$	-402.012	808.024	808.352

Table 5.2: Counts of cysts of kidneys using steroids

Model	Estimates	LL	AIC	BIC
Poisson (λ)	$\hat{\lambda} = 1.39$	-246.21	494.42	494.46
NB(r, p)	$\hat{r} = 0.321$, $\hat{p} = 0.188$	-167.544	339.088	339.17
$\mathscr{M}\mathscr{P}\mathscr{G}(\pi, \theta)$	$\hat{\pi} = 0.75$, $\hat{\theta} = 0.391$	-90.8726	185.75	185.83
$\mathscr{M}\mathscr{P}\mathscr{L}(\pi, \theta)$	$\hat{\pi} = 0.653$, $\hat{\theta} = 1.437$	-252.494	508.98	509.07

Chapter 6

Conclusion

This project presents a new convoluted mixed distribution function with two parameters, called the Ber-Mixed Poisson model. The BerMP or MPB model is constructed using the convolution of the Bernoulli and the Mixed Poisson distributions. The new distribution has two parameters, π and λ, the latter of which is assumed to follow a different distribution, either Exponential or Geometric. This accounts for count series composed of underdispersed and overdispersed data. We study various properties of BerMP distribution including the probability generating function and moments and ML estimates. Results of data analysis presented indicate good fit of the data as compared to the other conventional models. This is also a generalised distribution so we have the flexibility of choosing an appropriate underlying distribution of our liking for the parameter λ.

Some further characteristics of new model are:

- The proposed model has a closed form expression and presents more flexible for fitting than the conventional models.
- The moments have closed form expressions and are expressed easily for further analysis and computation.
- Three criteria namely LL, AIC and BIC were used in the paper for comparison of statistical models. Both criteria indicate that the convoluted

Mixed Poisson model produces a better fit with the data than the other conventional known models.

In the future, we hope to find more models like the ones we proposed and see if they perform better than the conventional models. We will also be eager to use other techniques to undertake estimation procedures if we don't obtain the PMF in a closed form.

We will also attempt to develop a **dedicated EM Algorithm**, especially for the Mixed Poisson Regression Model. So far, our needs were elegantly met by the conventional direct estimation procedure which yielded fairly accurate results.

Chapter 7

References

1. Karlis, D., and Xekalaki, E. (2005). Mixed Poisson Distributions. *International Statistical Review / Revue Internationale de Statistique,* 73(1), 35–58.

2. Bourguignon, M. and Weiß, C. H. (2017). An INAR(1) process for modeling count time series with equidispersion, underdispersion and overdispersion, *TEST* 26, 847–868, DOI: 10.1007/s11749-017-0536-4

3. Marcelo, B. and Rodrigo M.R. de M. (2022). A simple and useful regression model for fitting count data. *TEST*, 1–38.

4. Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control,* 19(6), 716–723.

5. Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics,* 6(2), 461–464.

6. Consul, P. C. (1989). *Generalized Poisson distributions:properties and applications,* Marcel Dekker, Inc., New York.

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

8. Puig, P. and Valero, J. (2006). Count data distributions: Some characterizations with applications. *Journal of the American Statistical Association*, 101(437), 332-340.

9. Grainger, R. M., and Reid, D. B. W. (1954). Distribution of Dental Caries in Children. *Journal of Dental Research,* 33(5), 613–623. DOI:10.1177/00220345540330050501.

10. Chan, S.-K., Riley, *et al.* (2010). Corticosteroid-induced kidney dysmorphogenesis is associated with deregulated expression of known cystogenic molecules, as well as indian hedgehog. *American Journal of Physiology-Renal Physiology* 298(2), F346–F356. DOI:10.1152/ajprenal.00574.2009