Title: Over-Dispersed and Under-Dispersed Generalized Poisson Distribution

Author(s): M.Y. Hassan

Department: Statistics

UAEU-CBE-Working Paper Series

No. 2019-01
Over-Dispersed and Under-Dispersed Generalized Poisson Distribution

M.Y. Hassan ¹
Department of Statistics
United Arab Emirates University
Al Ain, UAE

Abstract

We propose an over-dispersed, and under-dispersed Generalized Poisson Distribution. This distribution will overcome some of the potential limitations suffered by the existing dominant distributions including lack of modeling over-dispersion, under-dispersion. This distribution is flexible and could be applied to a variety of problems from different disciplines like business, finance, medicine and reliability in engineering.

Keywords: Generalized Poisson Distribution, Heterogeneity, Link function Functions.

1 Introduction and Motivation

Many attempts have been made to find new families of probability distributions by extending well-known families of distributions to provide more flexibility in modeling various data sets, see for example, Arnold and Beaver (2000a, 2000b). Models based on the Poisson distribution are frequently used to analyze Count data in different domains. The traditional approach for these models has been to link their means to relevant covariates using the canonical (log) link function. Count data generally involve unobserved heterogeneity caused by excess zeros or the covariates. The unobserved heterogeneity cause over-dispersion in which the Poisson based models cannot provide natural representations but result misspecification of those models. Alternative methods for the analysis of this problem have been proposed to deal with the over-dispersion issue of the count data. (Budhavarapu, 2016) recently modeled unobserved heterogeneity using finite mixture random parameters for spatially correlated discrete count data. (Lambert, 1992) combined the Poisson regression model with a logit model which resulted the zero-inflated Poisson (ZIP) to cover the zero inflation that causes over-dispersion in the count variable of the dependent variable. (Lim, 2014 ) also used the zero-inflated Poisson (ZIP) to fit count data with excess zeros whereas (Zhou, 2015) employed the zero inflated Negative Binomial (ZINB) models to cater for different dispersion parameters and heterogeneity aspects. To address the issue of over-dispersion (Minkova et al 2013) employed weighted Poisson (WPD) of (Fisher, 1934; Rao, 1965), Patil (2002). (Dalrymple , 2003) used finite mixture of zero-inflated Poisson and hurdle models with applications to count data to capture over-dispersion. (Shmueli et al., 2005) used the Conway-Maxwell-Poisson (CMP) distribution to fit both the under-dispersion and over-dispersion patterns of the count data. Most of these models capture over-dispersion but not able to model data involving underdispersion except the Conway-Maxwell-Poisson (CMP) model which can represent both over-dispersion and under-dispersion. The CMP distribution has been used in a variety of count-data applications and has been extended methodologically in various directions (see a survey of CMP-based methods and applications in Sellers et al., 2012). However the CMP model has two weaknesses, first, it can not probably capture bimodality and secondly, its parameter estimation has convergence problems. For more research in resolving problems arising from count data, see for example (Balakrishnan, N., et al. 2008), (Bockenholt, 1998), (Karlis et al. 2007), (Wang, P., et al. 1998 ), (Land et al 1996), Consul (1992), and (Frees, 2011).

¹Corresponding author. Email: myusuf@uaeu.ac.ae
2 Equi-Dispersed, Over-Dispersed and Under-Dispersed G-Poisson

Definition 2.1. A random variable $X$ has G-Poisson$(\lambda, \delta)$ distribution with parameters $\lambda$ and $\delta$ if its probability mass function is given by

$$
\phi(x, \lambda, \delta) = \left(\frac{e^{-\lambda} \lambda^x}{x!(\lambda^2 + \lambda + \delta)}\right)[\delta + x^2] \quad \lambda > 0, \delta \geq 0 \quad x = 0, 1, ...
$$

(1)

2.1 Moments, Moment Generating Function and Fisher’s Index of Dispersion

If $X$ is distributed as G-Poisson$(\lambda, \delta)$ presented in (21), the moment generating function of $X$ is given by

$$
M_X(t) = \sum_{x=0}^{\infty} \left(\frac{e^{tx}(e^{-\lambda} \lambda^x / x!)}{x!(\lambda^2 + \lambda + \delta)}\right)[\delta + x^2] = \sum_{x=0}^{\infty} \left(\frac{e^{tx}(e^{-\lambda} \lambda^x / x!)}{\lambda^2 + \lambda + \delta}\right)\delta + \sum_{x=0}^{\infty} \left(\frac{e^{tx}(e^{-\lambda} \lambda^x / x!)}{\lambda^2 + \lambda + \delta}\right)x^2
$$

$$
= \frac{\delta(e^{-\lambda+\lambda e^t}) + \lambda e^{-\lambda+\lambda e^t}(1 + \lambda e^t)}{\lambda^2 + \lambda + \delta} = \frac{e^{-\lambda+\lambda e^t}(\delta + \lambda e^t + \lambda^2 e^{2t})}{\lambda^2 + \lambda + \delta}
$$

Thus, the moment generating function of the distribution is given by

$$
M_X(t) = \frac{e^{-\lambda+\lambda e^t}(\delta + \lambda e^t + \lambda^2 e^{2t})}{\lambda^2 + \lambda + \delta}
$$

(2)

The first and the second derivatives of the moment generating are as follows:

$$
M'(t) = \frac{\lambda e^{-\lambda+\lambda e^t}(\delta + \lambda e^t + \lambda^2 e^{2t}) + e^{-\lambda+\lambda e^t}(\lambda e^t + 2\lambda^2 e^{2t})}{\lambda^2 + \lambda + \delta} = \frac{\lambda e^{\lambda(e^t-1)+t}(1 + \delta + 3\lambda e^t + \lambda^2 e^{2t})}{\lambda^2 + \lambda + \delta}
$$

$$
M''(t) = \frac{\lambda e^{\lambda(e^t-1)+t}(1 + \lambda e^t)(1 + \delta + 3\lambda e^t + \lambda^2 e^{2t}) + \lambda e^{\lambda(e^t-1)+t}(3\lambda e^t + 2\lambda^2 e^{2t})}{\lambda^2 + \lambda + \delta}
$$

$$
= \frac{\lambda e^{\lambda(e^t-1)+t}(1 + \lambda + 7\lambda e^t + \lambda^3 e^{3t})}{\lambda^2 + \lambda + \delta}
$$
It follows that, the mean and the variance are given by

\[ M'(0) = \mu = \frac{\lambda(1 + \delta + 3\lambda + \lambda^2)}{\lambda^2 + \lambda + \delta}, \quad M''(0) = \text{E}(x^2) = \frac{\lambda(1 + \lambda + 7\lambda + \lambda\delta + 6\lambda^2 + \lambda^3)}{\lambda^2 + \lambda + \delta} \]

and

\[ V(x) = \frac{\lambda[2\lambda^3 + \lambda^4 + 6\lambda\delta + 2\lambda^2(1 + \delta) + \delta(1 + \delta)]}{(\lambda^2 + \lambda + \delta)^2} \]  

Consequently the Fisher’s index of dispersion is given by

\[ \Phi(x) = \frac{V(x)}{\text{E}(x)} = \frac{[2\lambda^3 + \lambda^4 + 6\lambda\delta + 2\lambda^2(1 + \delta) + \delta(1 + \delta)]}{(\lambda^2 + \lambda + \delta)(1 + \delta + 3\lambda + \lambda^2)} \]  

**Theorem 2.1.** Let \( X^* \) be a \( G - \text{Poisson}(\lambda, \delta) \) random variable with the probability mass function in (22) then \( X^* \) is

(i) Equi-dispersed if and only if \( \delta = \lambda + \lambda^2 + 1/2 \)

(ii) Over-dispersed if and only if \( \delta > \lambda + \lambda^2 + 1/2 \)

(iii) Under-dispersed if and only if \( \delta < \lambda + \lambda^2 + 1/2 \)

**Proof**

(i) The prove of this part is trivial.

(ii) Suppose \( \Phi(x) > 1 \), that is to say \( [2\lambda^3 + \lambda^4 + 6\lambda\delta + 2\lambda^2(1 + \delta) + \delta(1 + \delta)]/((\lambda^2 + \lambda + \delta)(1 + \delta + 3\lambda + \lambda^2)) > 1 \) which implies \( [2\lambda^3 + \lambda^4 + 6\lambda\delta + 2\lambda^2(1 + \delta) + \delta(1 + \delta)] - (\lambda^2 + \lambda + \delta)(1 + \delta + 3\lambda + \lambda^2) > 0 \) simplifying the above equation gives \( -\lambda(1 + 2\lambda + 2\lambda^2 - 2\delta) > 0 \) which finally gives \( \delta > \lambda + \lambda^2 + 1/2 \).

Conversely, Suppose \( \delta > (\lambda + \lambda^2 + 1/2) \), expanding the numerator and the denominator of equation (4) gives \( \lambda^4 + 2\lambda^3 + 2\lambda^2 + 2\lambda^2\delta + 6\lambda\delta + \delta + \delta^2 \) and \( \lambda + \lambda^4 + 4\lambda^3 + 4\lambda^2 + 2\lambda^2\delta + 4\lambda\delta + \delta + \delta^2 \) respectively. Factoring the denominator again and substituting \( \delta \) by \( \lambda + \lambda^2 + 1/2 \) gives \( \lambda + \lambda^4 + 4\lambda^3 + 4\lambda^2 + 2\lambda^2\delta + 4\lambda\delta + \delta + \delta^2 = 2\lambda(\lambda + \lambda^2 + 1/2) + \lambda^4 + 2\lambda^3 + 2\lambda^2\delta + 4\lambda\delta + \delta + \delta^2 < 2\lambda\delta + \lambda^4 + 2\lambda^3 + 2\lambda^2\delta + 4\lambda\delta + \delta + \delta^2 \) since \( \delta > (\lambda + \lambda^2 + 1/2) \).

Substituting the simplifying terms in equation (24) gives \( \Phi(x) = [2\lambda^3 + \lambda^4 + 6\lambda\delta + 2\lambda^2(1 + \delta) + \delta(1 + \delta)]/((\lambda^2 + \lambda + \delta)(1 + \delta + 3\lambda + \lambda^2)) = 1 \) which gives the desired result.

(iii) The prove of this part follows the same lines as part (ii).
If the quadratic factor of the probability mass function in (1) is horizontally translated by $\theta$, we get the following probability mass function.

$$\phi(x, \lambda, \delta) = \left( \frac{e^{-\lambda x^2/x!}}{\lambda^2 + \lambda + \delta - 2\lambda \theta + \theta^2} \right) [\delta + (x - \theta)^2] \quad \lambda > 0, \quad \delta \geq 0, \quad x = 0, 1, ... \quad (5)$$

Clearly, the horizontal translation parameter $\theta$ controls the location of the concavity of the mass function whereas $\delta$ mainly controls the vertex of the concavity. Figure 1 displays the plots of the distribution with different values of $\theta$, $\delta$ and $\lambda$. Several different examples fitting this distribution with real count data will be presented.

![Figure 1: Line Plots of Probabilities and Probability mass Functions.](image-url)
3 Applications

3.1 G-Poisson VS Conway-Maxwell-Poisson Examples

Example 1: Chemotherapy for Stage B/C colon cancer

The data in this example contains the number of lymph nodes with detectable cancer. There are 1822 observations, with a sample mean and a sample variance of 3.66 and 12.76 respectively. It is one of the first successful trials of adjuvant chemotherapy for colon cancer. These data are originally collected and described in Laurie (1989). The main report is found in Moertel (1990) and a version with less follow-up time was used in the paper by Lin (1994).

Table 1: Fitted G-Poisson and COMPoisson.

<table>
<thead>
<tr>
<th>Est</th>
<th>ν</th>
<th>λ</th>
<th>Z</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>ComPoisson</td>
<td>0.18</td>
<td>1.08</td>
<td>7.03</td>
<td>8673.5</td>
</tr>
<tr>
<td>IQRA-Poisson</td>
<td>δ</td>
<td>θ</td>
<td>λ</td>
<td></td>
</tr>
<tr>
<td>Estimates</td>
<td>1.7</td>
<td>5.61</td>
<td>4.45</td>
<td>8665.64</td>
</tr>
</tbody>
</table>

![Figure 2: Fitted G-Poisson and COMPoisson for the Chemotherapy for Stage B/C Colon Cancer Data](chart.png)

Table 1 shows estimates of the maximum likelihood estimates and the BIC information criteria values of the two competing models the G-Poisson and the Com-Poisson. Both models are parsimonious and have
small number of parameters. Based on the BIC, G-Poisson performed better than the Com-Poisson. Besides that, for a number of other data sets the Comp-Poisson did not converge, so we do think that the convergence of the Comp-Poisson models need more investigation. Figure 2 displays qualitative assessment of the two fitted models and the plot demonstrated that the G-Poisson model better captured the modal class but both models did well on the right tail.

4 Conclusions and Possible Extensions

The general form of the G-Poisson distribution is presented and investigated. It has been shown that, this distribution is capable of capturing over-dispersed, under-dispersed and bimodal features of count data. The results of an example involving real data indicated the simplicity and potential superiority of the parsimonious G-Poisson model compared with other popular distributions such as the Com-Poisson distribution which has convergence problems. It is hoped that the parsimonious G-Poisson distribution will attract many researchers who can benefit the flexibility of this distribution. This distribution can be extended in many different ways based on the number of the parameters in the distribution. For Example, it could be extended in the following ways $\phi(x, \lambda, \delta) = C_1(e^{-\lambda x}/x)!\delta + x^\tau$ and $\phi(x, \lambda, \delta) = C_2(e^{-\lambda x}/x)!\delta + e^{-\tau x}$ respectively, where $C_1$ and $C_2$ are normalizing constants.

References


