ISSN 2079-7141

UAEU College of Business and Economics

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

Author(s): M.Y. Hassan

Department: Statistics

UAEU-CBE-Working Paper Series

No. 2019-01

Series Founding and Acting Editor: Prof. Dr. Abdulnasser Hatemi-J

The views and conclusions expressed in this working paper are strictly those of the author(s) and do not necessarily represent, and should not be reported as, those of the FBE/UAEU. The FBE and the editor take no responsibility for any errors, omissions in, or for the correctness of, the information contained in this working paper.

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 under-dispersion 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).

[1] Corresponding 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 λ and δ if its probability mass function is given by

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

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

$$\begin{aligned}\mathbb{M}_X(t) &= \sum_{x=0}^{\infty}(\frac{e^{tx}(e^{-\lambda}\lambda^x/x!)}{\lambda^2+\lambda+\delta})[\delta+x^2] = \sum_{x=0}^{\infty}\frac{e^{tx}(e^{-\lambda}\lambda^x/x!)}{\lambda^2+\lambda+\delta}\delta + \sum_{x=0}^{\infty}(\frac{e^{tx}(e^{-\lambda}\lambda^x/x!)}{\lambda^2+\lambda+\delta})x^2 \\ &= \frac{\delta(e^{-\lambda+\lambda e^t}) + \lambda e^{-\lambda+\lambda e^t+t}(1+\lambda e^t)}{\lambda^2+\lambda+\delta} = \frac{e^{-\lambda+\lambda e^t}(\delta+\lambda e^t+\lambda^2 e^{2t})}{\lambda^2+\lambda+\delta}\end{aligned}$$

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

$$\mathbb{M}_X(t) = \frac{e^{-\lambda+\lambda e^t}(\delta+\lambda e^t+\lambda^2 e^{2t})}{\lambda^2+\lambda+\delta} \tag{2}$$

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

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

$$\begin{aligned}\mathbb{M}^{''}(t) &= \frac{\lambda e^{\lambda(e^t-1)+t}(1+\lambda e^t)(1+\delta+3\lambda e^t+\lambda^t e^{2t}) + \lambda e^{\lambda(e^t-1)+t}(3\lambda e^t+2\lambda^2 e^{2t})}{\lambda^2+\lambda+\delta} \\ &= \frac{\lambda e^{\lambda(e^t-1)+t}(1+\lambda+7\lambda e^t+\lambda\delta e^t+6\lambda^2 e^{2t}+\lambda^3 e^{3t})}{\lambda^2+\lambda+\delta}\end{aligned}$$

It follows that, the mean and the variance are given by

$$\mathbb{M}^{'}(0) = \mu = \frac{\lambda(1+\delta+3\lambda+\lambda^2)}{\lambda^2+\lambda+\delta}, \qquad \mathbb{M}^{''}(0) = \mathbb{E}(x^2) = \frac{\lambda(1+\lambda+7\lambda+\lambda\delta+6\lambda^2+\lambda^3)}{\lambda^2+\lambda+\delta}$$

and

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

Consequently the Fisher's index of dispersion is given by

$$\mathbb{I}(x) = \mathbb{V}(x)/\mathbb{E}(x) = \frac{[2\lambda^3+\lambda^4+6\lambda\delta+2\lambda^2(1+\delta)+\delta(1+\delta)]}{(\lambda^2+\lambda+\delta)(1+\delta+3\lambda+\lambda^2)} \tag{4}$$

Theorem 2.1. *Let X^* be a $G-Poisson(\lambda,\delta)$ 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 $\mathbb{I}(x) > 1$, 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 δ by $\lambda+\lambda^2+1/2$ gives $\lambda+\lambda^4+4\lambda^3+4\lambda^2+2\lambda^2\delta+4\lambda\delta+\delta+\delta^2 = 2\lambda(\lambda+\lambda^2+1/2)+\lambda^4+2\lambda^3+2\lambda^2+2\lambda^2\delta+4\lambda\delta+\delta+\delta^2 < 2\lambda\delta+\lambda^4+2\lambda^3+2\lambda^2+2\lambda^2\delta+4\lambda\delta+\delta+\delta^2$ since $\delta > (\lambda+\lambda^2+1/2)$. Substituting the simplifying terms in equation (24) gives $\mathbb{I}(x) = [2\lambda^3+\lambda^4+6\lambda\delta+2\lambda^2(1+\delta)+\delta(1+\delta)]/(\lambda^2+\lambda+\delta)(1+\delta+3\lambda+\lambda^2) > \lambda+\lambda^4+4\lambda^3+4\lambda^2+2\lambda^2\delta+4\lambda\delta+\delta+\delta^2/(2\lambda\delta+\lambda^4+2\lambda^3+2\lambda^2+2\lambda^2\delta+4\lambda\delta+\delta+\delta^2) = 1$ 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 θ, we get the following probability mass function.

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

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

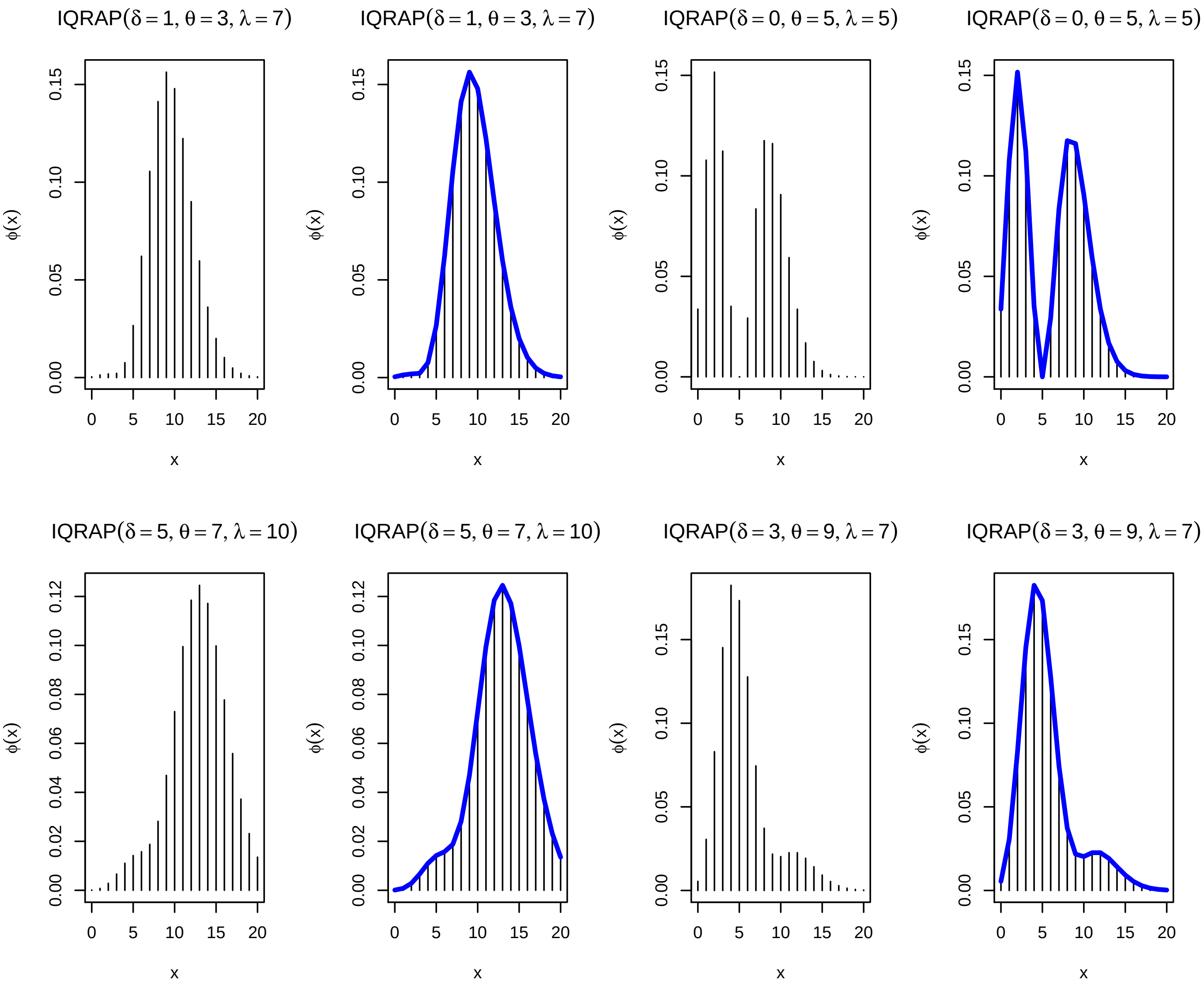

Figure 1: Line Plots of Probabilities and Probability mass Functions.

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.

Est	ν	λ	Z	BIC
ComPoisson	0.18	1.08	7.03	8673.5
IQRA-Poisson	δ	θ	λ	
Estimates	1.7	5.61	4.45	8665.64

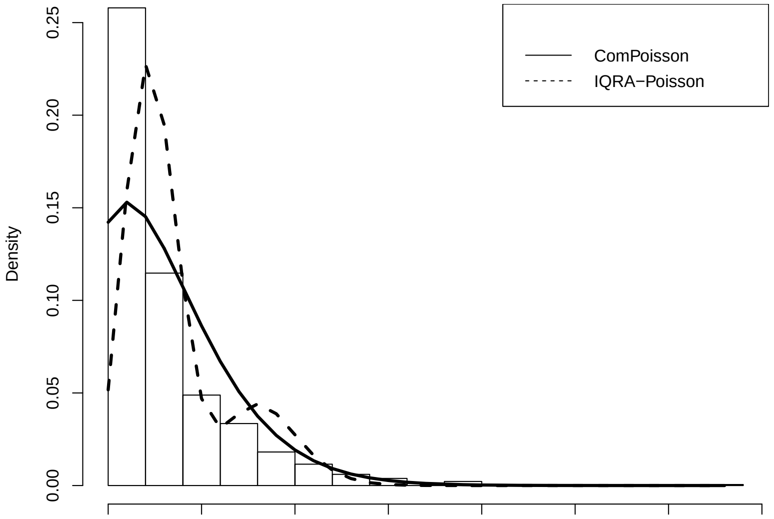

Figure 2: Fitted G-Poisson and COMPoisson for the Chemotherapy for Stage B/C Colon Cancer Data

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 ditribution 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 $\phi(x, \lambda, \delta) = C_1(e^{-\lambda}\lambda^x/x!)[\delta + x^\tau]$ and $\phi(x, \lambda, \delta) = C_2(e^{-\lambda}\lambda^x/x!)[\delta + e^{-\tau x}]$ respectively, where C_1 and C_2 are normalizing constants.

References

[1] Arnold, B. C., Beaver, R. J. (2000b). The skew Cauchy distribution. Statistics and Probability Letters, 49, 285-290

[2] Arnold, B.C.,Beaver, R.J. (2000a). Hidden truncation models. Sankhya, 62, pp. 23-35

[3] Balakrishnan, N., Kozubowski T.(2008) A class of weighted Poisson processes. Stat Probab Letters 78(15):2346-2352

[4] Bockenholt, U. “Mixed INAR(1998) Poisson regression models : Analyzing heterogeneity and serial dependencies in longitudinal count data” Journal of Econometrics Vol 89, 317-338.

[5] Budhavarapu, P, James, G. ,and Jorga A, (2016)“ Modeling unobserved heterogeneity using finate random parameters for spatially correlated counta data” Transportation Research Part B, Methodological Vol. 91 C, 492-510

[6] Consul, P.C. and F. Famoye (1992). Generalized Poisson Regression Model. Communications in Statistics: Theory and Method 21 (1) 89-109

[7] Dalrymple M., I. L. Hudson, and R. P. K. Ford, (2003)“Finite mixture, zero-inflated poisson and hurdle models with application to SIDS,” Computational Statistics and Data Analysis, vol. 41, no. 3-4, pp. 491–504

[8] Fisher, R. A. (1934) The effects of methods of ascertainment upon the estimation of frequencies. Ann. Eugenics, 6, 13-25

[9] Frees, E. W. (2011). Regression Modeling with Actuarial and Financial Applications Cambridge University Press

[10] Karlis, D., Rahmouni, M. (2007). Analysis of defaulters' behaviour using the Poisson mixture approach. IMA Journal of Management Mathematics, 18 (3): 297-311.

[11] Lambert, D. (1992) "Zero-inflated Poisson regression with an application to defects in manufacturing", Technometrics 34: 1–13.

[12] Land, K.C.; McCall, P.L.; Nagin, D.S. (1996) "A comparison of Poisson, negative

[13] Laurie, JA., CG Moertel, CG. et al. Surgical adjuvant therapy of large-bowel carcinoma: An evaluation of levamisole and the combination of levamisole and fluorouracil: The North Central Cancer Treatment Group and the Mayo Clinic. J Clinical Oncology, 7:1447-1456, 1989.

[14] Lim, H., Li, W., Yu, H., (2014) " Zero-inflated Poisson regression mixture model." Computational Statistics and Data Analysis, Vol 71, 2014 151–158

[15] Lin, DY. (1994) Cox regression analysis of multivariate failure time data: the marginal approach. Statistics in Medicine, 13: 2233-2247.

[16] Moertel, CG., Fleming,TR., et al, Levamisole and fluorouracil for adjuvant therapy of resected colon carcinoma. New England J of Medicine, 332:352-358, 1990.

[17] Minkova, L.D. and Balakrishnan, N.(2013). Compound weighted Poisson distributions, Metrika, 76, 543-558.

[18] Patil, G.P., 2002.Weighted distributions. In: El-Shaarawi, A.H., Piegorsch,W.W. (Eds.), Encyclopedia of Environmetrics, vol.4.Wiley, Chichester, pp. 2369–2377.

[19] Rao, C.R. (1965). On discrete distributions arising out of methods of ascertainment, in Classical and Contagious Discrete Distributions, G.P. Patil, ed., Pergamon Press and Statistical Publishing Society, Calcutta, pp. 320–332.

[20] Sellers,K., Borle, S., Shmueli,G. (2012) The CMP Model for Count Data: A Survey of Methods and Applications, Applied Stochastic Models in Business and Industry. 28, Issue 2, 104-116.

[21] Shmueli, G., Minka, T., Kadane, J., Borle, P. (2005) Boatwright, A Useful Distribution for Fitting Discrete Data: Revival of the Conway–Maxwell–Poisson Distribution, Journal of The Royal Statistical Society. Series C (Applied Statistics). 54, Issue 1, 127-142.

[22] Wang, P., Cockburn, I., Puterman, M. (1998) Analysis of patent data: a mixed Poisson regression model approach. J. Bus. Econom. Statist. 16 (1), 27–41.

[23] Zhou, M., Carin, L. (2015) "Negative binomial process count and mixture modeling," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 2, pp. 307–320.