Comparative Analyses of the Poisson-Exponential-Gamma to Selected Discrete Distributions with Applications

Abstract

In this paper, we compare results arising from implementing the new Poisson-Exponential-Gamma distribution (PEG) to other competing two-parameter distributions such as the generalized Poisson Lindley (GPL), the Poisson generalized Lindley (PGL), the Negative binomial (NB) and several others to several frequency data sets exhibiting different characteristics and data with covariates exhibiting over-or-under dispersion. In all, we show that the PEG does not perform better than most existing distributions. We also demonstrate the equivalence of the GPL and the PGL (the latter being a re-parameterized version of the former). We further show that the New Poisson generalized Lindley (NGPL) distribution is also equivalent to the two-parameter discrete Lindley (TDL) distribution and that in some cases, these two degenerate to the one-parameter geometric distribution (GD). Our results here indicate the limitations of the PEG especially to under-dispersed data. For very strong over-dispersed data, the NB, the New logarithmic distribution (NLD), the New Geometric Discrete Pareto Distribution (NGDP) or the discrete Weibull(DW) perform much better. SAS PROC NLMIXED is employed in our estimation. Adjusted group X^2 as well as Wald's test statistics were computed using estimated theoretical means and variances.

Keywords: *Generalized Poisson Lindley, Poissonn generalized Lindley, Poisson inverse Gaussian, GLM*

Mathematics Subject Classification: 60E05,62N01,62G05,62N05

1 Introduction

For count data exhibiting overdispersion as measured by the dispersion index (DI), which is a ratio of $s^2/\bar{Y}$ with overdispersion manifesting if DI > 1. Several two and three parameter distributions have been proposed for fitting such data. These include the Negative binomial, the Generalized Poisson, the Poisson inverse Gaussian, the DW and several Poisson-Lindley mixture distributions, as well as several three-parameter type distributions; which include the extended Com-Poisson, the quasi-negative binomial, the generalized negative, the Delaporte distribution, the Inverse-trinomial and the negative exponential generalized exponential distributions amongst several others. Umar & Yaya [36] have proposed a new Poisson-exponential Gamma distribution (PEG), which is a two parameter distribution. Its properties are discussed and its applications to three frequency data sets that exhibit over-dispersion are presented. The PEG is said to be better than (i) the generalized Poisson Lindley (GPL) [25] and the new generalized Poisson-Lindley (NGPL) [7] distribution The authors stated and i quote

> *It can be observed from the various results reported in Tables 1, 2, and 3 for all the three datasets analysed that the proposed PEG distribution with the smallest estimated values of all the models*

assessment criteria is most efficient than any of the other five existing similar distributions considered. In all the three tables of results, the new PEG distribution with the least values of 2logLik, AIC, AICC and BIC provided the best fit to the three data sets than any of the other five existing distributions considered.

The main goal of this paper therefore is to ascertain the veracity of this claim since the results from the PEG applications were exclusively based on the carefully selected data sets by extending the performance of the PEG model to other variety of count data sets and compare its performances to other distributions being employed in this paper.

The authors also compared the PEG to a host of one-parameter distributions, such as the, Poisson (P), Poisson-Lindley (PL), the Poisson-Shanker (PS), the Poisson-Exponential (PS) as well as the two-parameter Negative Binomial (NB). The PEG is presented as being better in performance to these other models based on the selected data on which the application of the PEG is based.

However, since count data come in several forms, we propose to examine the PEG relative to its two parameter counterparts GPL, NGPL, NB, NLD , DW, TDL and NGDP only, since it was established that none of the one-parameter models outperform the PEG based on the goodness-of-fit criteria (-2logLL, AIC and BIC) employed in the study for data exhibiting excess zeros and hyper over-dispersion [**?**]. We will also extend our results to the case where we have an over-dispersed count data having covariates. SAS PROC NLMIXED is employed for all our computations. The distributions employed in this study as well as their likelihood functions are presented in the following sections.

2 Methodology

In this section we describe the probability distributions employed in this study. because the Poisson is the underlying model for counter data, we also the Poisson in our discussion.

2.1 The New Poisson-Exponential Gamma Distribution-PEG

Yahya & Umar [36] proposed the new Poisson-Exponential Gamma (PEG) distribution with parameters θ, α, which is a mixture of Poisson paramater is assumed to follow the Exponential-Gamma distribution [33], that is,

$$Y|\lambda \sim P(\lambda), \quad \text{and} \quad \lambda|\theta, \alpha \sim \text{EG}(\theta, \alpha)$$

where, the Exponential -Gamma distribution [33] has the pmf:

$$f(y; \alpha, \theta) = \frac{\theta}{\theta + \Gamma(\alpha)}(\theta + \theta^{\alpha-1}y^{\alpha-1})e^{-\theta y}, \; y > 0, \alpha > 0, \theta > 0 \tag{1}$$

The resulting unconditional pmf of is the PEG given by (2).

$$f_{PEG}(y; \theta, \alpha) = \frac{\theta}{y![\theta + \Gamma(\alpha)]} \cdot \frac{\theta(\theta+1)^{\alpha}\Gamma(y+1) + \theta^{\alpha-1}(\theta+1)\Gamma(\alpha+y)}{(\theta+1)^{y+\alpha+1}} \tag{2}$$

$y = 1, 2, \ldots$ and $\alpha, \theta > 0$. The properties of the PEG are fully discussed in [36], however, we present below expressions for the mean and variance of PEG as established in Yahya & Umar, viz:

$$\begin{aligned} \mu &= \frac{\theta + \alpha\Gamma(\alpha)}{\theta[\theta + \Gamma(\alpha)]} \\ \sigma^2 &= \frac{(\theta + \Gamma(\alpha))[\theta^2 + \alpha\theta\Gamma(\alpha) + 2\theta + \Gamma(\alpha+2)] - [\theta + \alpha\Gamma(\alpha)]^2}{\theta^2[(\theta + \Gamma(\alpha)]^2} \end{aligned} \tag{3}$$

The probability that $\Pr(Y = 0)$ becomes,

$$f(0) = \frac{\theta}{\theta + \Gamma(\alpha)} \cdot \frac{\theta(\theta+1)^{\alpha} + \theta^{\alpha-1}(\theta+1)\Gamma(\alpha)}{(\theta+1)^{\alpha+1}}$$

2.2 The Generalized Poisson-Lindley -GPL Distribution

The generalized Poisson-Lindley (GPL) distribution having $Y \sim GPL(\alpha, \gamma = 1, \theta)$, proposed in [25] has the pmf given by:

$$f(y;\alpha,\gamma=1,\theta) = \frac{\Gamma(y+\alpha)\theta^{\alpha+1}\left(\alpha+\frac{y+\alpha}{1+\theta}\right)}{y!\Gamma(1+\alpha)(1+\theta)^{y+\alpha+1}} \tag{4}$$

It is a mixture of Poisson distribution and the generalized-Lindley (GL) distribution ([37]). Its moments are:

$$E(Y) = \frac{\alpha(1+\theta)+1}{\theta(1+\theta)} = \mu \tag{5a}$$

$$E(Y^2) = \mu + \frac{(\alpha+1)[\alpha(1+\theta)+2]}{\theta^2(1+\theta)} - \mu^2 \tag{5b}$$

Hence, variance is:

$$\sigma^2 = \frac{\alpha(\theta+1)^3+\theta^2+3\theta+1}{\theta^2(\theta+1)^2} \tag{6}$$

and the dispersion index is:

$$1 + \frac{\alpha(\theta+1)^2+2\theta+1}{\alpha\theta(\theta+1)^2+\theta(\theta+1)}$$

indicating over-dispersion for values of α and θ, with equi-dispersion occurring if

$$\frac{\alpha(\theta+1)^2+2\theta+1}{\alpha\theta(\theta+1)^2+\theta(\theta+1)} = 0$$

2.3 The New Generalized Poisson-Lindley Distribution-NGPL

The new generalized Poisson-Lindley (NGPL) distribution proposed in [7] is a mixture distribution of Poisson and the two parameter Lindley (TPLD) distribution such that,

$$Y|\lambda \sim P(\lambda), \quad \text{and} \quad \lambda|\theta,\alpha \sim \text{TPLD}(\theta,\alpha)$$

where the TPLD is defined in [31] as:

$$f(y;\theta,\alpha) = \frac{\theta^2}{(\theta+\alpha)}(1+\alpha y)e^{-\theta y}.$$

Consequently, the pmf of the NGPL is given by:

$$f(y;\theta,\alpha) = \frac{\theta^2}{(\theta+\alpha)(1+\theta)^{y+1}}\left(1+\frac{\alpha(y+1)}{(1+\theta)}\right), y = 0,1,\ldots \tag{7}$$

where $\theta > 0$, $\alpha > 0$. Its moments are:

$$\begin{aligned} \mu_1 &= \frac{2\alpha+\theta}{\theta(\alpha+\theta)} \\ \sigma^2 &= \frac{2\alpha^2(1+\theta)+\theta^2(1+\theta)+\alpha\theta(4+3\theta)}{\theta^2(\alpha+\theta)^2} \end{aligned} \tag{8}$$

f(0) for the model in (11) becomes

$$f(0) = \frac{\theta^2(1+\alpha+\theta)}{(\alpha+\theta)(1+\theta)^2}$$

2.4 The New Poisson generalized Lindley-NPGL

The New Poisson generalized Lindley (NPGL) proposed in [3] ia a mixture of Poisson and generalized Lindley (GL) distributions. Here, the random variable Y having a Poisson distribution with parameter λ, that is,

$$Y|\lambda \sim P(\lambda), \quad \text{and} \quad \lambda|\theta,\alpha \sim \text{GL}(\theta,\alpha)$$

while the parameter λ is assumed distributed as the Abouammoh *et al.* (2015) [1] Generalized Lindley (GPL) distribution with pmf,

$$g(\lambda) = \frac{\theta^{\alpha}\lambda^{\alpha-2}}{(\theta+1)\Gamma(\alpha)}(\lambda+\alpha-1)e^{-\theta\lambda} \tag{9}$$

for $\lambda > 0, \alpha > 1$ and $\theta > 0$. Hence, employing the well established conditional formulation, then

$$f(y) = \int_0^{\infty} f(y|\lambda)g(\lambda)d\lambda \tag{10}$$

The integral in (10) when integrated out gives the unconditional NGPL distribution with parameters α, θ with pmf,

$$f(y;\alpha,\theta) = \frac{\theta^{\alpha}\Gamma(\alpha+y-1)[\alpha(\theta+2)-\theta+y-2]}{\Gamma(\alpha)\Gamma(y+1)(\theta+1)^{\alpha+y+1}};\ y = 0,1,\ldots \tag{11}$$

with $\alpha > 1$ and $\theta > 0$.
The mean and variance of the NGPL(α,θ) are given respectively as:

$$\mu = \frac{(\alpha-1)\theta+\alpha}{\theta(\theta+1)} \tag{12a}$$

$$\sigma^2 = \frac{(\alpha-1)\theta+\alpha}{\theta(\theta+1)} + \frac{(\alpha-1)\theta^2+2\alpha\theta+\alpha}{\theta^2(\theta+1)^2} \tag{12b}$$

The properties of both the GPL and NPGL distributions are fully discussed in [25]and [3] respectively and would not be further discussed in this paper. However, the expressions for the means and variances of both distributions are given respectively in (5) and (12).

2.5 The Negative Binomial-NB

The Negative binomial distribution (NB) has the probability mass function (pmf):

$$f(y;r,p) = \frac{\Gamma(r+y)}{y!\Gamma(r)}p^y(1-p)^r, \quad y = 0,1,\ldots \tag{13}$$

where $r \in (0,\infty) > p$ and $p \in (0,1)$. The mean and variance of the NB model with parameters r and p in (13) are given respectively in (14a) and (14b) respectively.

Hence,

$$\mu = rp/(1-p) \quad \Longrightarrow\ p\ = \frac{\mu}{r+\mu} \tag{14a}$$

$$\sigma^2 = rp/(1-p)^2 \quad \Longrightarrow \sigma^2 = \mu + \frac{\mu^2}{r} \tag{14b}$$

Of course the NB is a mixture of the Poisson-Gamma distributions.

2.6 The Two-parameter Discrete Lindley Distribution-TDL

Hussien *et al.* [19] proposed the two-parameter discrete Lindley distribution (TDL) which has the pmf:

$$f(y|p,\beta) = \frac{(1-p)^2(1+\beta y)p^y}{1+p(\beta-1)}, \quad y = 0,1,2,\ldots,\ 0<p<1,\ \beta \geq 0 \tag{15}$$

The mean and variance of the TDL are given respectively in (16a) and (16b),

$$\mu = \frac{[1-p+\beta(1+p)]p}{[1+p(\beta-1)](1-p)} \tag{16a}$$

$$\sigma^2 = \frac{(1-p)^2+(1-3p^2+2p)\beta+2(p\beta)^2}{[1+p(\beta-1)]^2(1-p)^2} \tag{16b}$$

2.7 The New Geometric Discrete Pareto Distribution-NGDP

The NGDP proposed in [8] has the pmf:

$$f(y|q,\alpha) = \frac{q^y}{(y+1)^\alpha} - \frac{q^{(y+1)}}{(y+2)^\alpha}, \quad y = 0,1,2,\ldots, \quad 0<q<1,\ \alpha \geq 0. \tag{17}$$

Its mean and variance can be computed from expressions in (18a) and 18b) respectively,

$$\mu_y = q\Phi(q,\alpha,2) \tag{18a}$$

$$\sigma_y^2 = 2q\Phi(q,\alpha-1,2) - q\Phi(q,\alpha,2)[3+q\Phi(q,\alpha,2)] \tag{18b}$$

where $\Phi(z,s,a) = \sum_{k=0}^{\infty} \frac{z^k}{(a+k)^s}$

2.8 The New logarithmic Distribution-NLD

Gómez-Déniz *et al.* [14] proposed the new logarithmic distribution (NLD) whose pmf has the form:

$$f(y|\alpha,\theta) = \frac{\log(1-\alpha\theta^y) - \log(1-\alpha\theta^{y+1})}{\log(1-\alpha)}; \quad y = 0,1,\ldots,\ 0<\theta<1;\ \alpha<1(\alpha \neq 0) \tag{19}$$

Its mean and variance can be computed from expressions in (20a) and 20b) respectively,

$$\mu_Y = \frac{1}{\log(1-\alpha)} \sum_{y=1}^{\infty} \log(1-\alpha\theta^y) \tag{20a}$$

$$\sigma_Y^2 = \frac{1}{\log(1-\alpha)} \sum_{y=1}^{\infty} (2y-1)\log(1-\alpha\theta^y) - \mu_Y^2 \tag{20b}$$

3 Parameter Estimation:

For a single observation i, the log-likelihood for the NB, GPL, NGPL, PGL, TDL, PEG, NGDP and NLD models are presented respectively in LL1 to LL8 in (21).

$$\text{LL1} = \log[\Gamma(r+y)] + y\log(p) + r\log(1-p) - \log y! - \log[\Gamma(r)] \tag{21a}$$

$$\text{LL2} = \log[\Gamma(y+\alpha)] + (\alpha+1)\log(\theta) + \log\left[\alpha + \frac{y+\alpha}{1+\theta}\right] - \log y! - \log[\Gamma(1+\alpha)] - (y+\alpha+1)\log(1+\theta) \tag{21b}$$

$$\text{LL3} = 2\log(\theta) - \log(\theta+\alpha) - (y+1)\log(1+\theta) + \log\left[1 + \frac{\alpha(y+1)}{(1+\theta)}\right] \tag{21c}$$

$$\text{LL4} = \alpha\log(\theta) + \log[\Gamma(\alpha+y-1)] + \log[\alpha(\theta+2) - \theta + y - 2] - \log[\Gamma(\alpha)] - \log[\Gamma(y+1)] - (\alpha+y+1)\log(\theta+1) \tag{21d}$$

$$\text{LL5} = 2\log(1-p) + \log(1+\beta y) + y\log(p) - \log[1+p(\beta-1)] \tag{21e}$$

$$\text{LL6} = \log\theta - \log(y!) - \log[\theta+\Gamma(\alpha)] - (y+\alpha+1)\log(\theta+1) + \log\left[\theta(\theta+1)^{\alpha}\Gamma(y+1) + \theta^{\alpha-1}(\theta+1)\Gamma(\alpha+y)\right] \tag{21f}$$

$$\text{LL7} = \log\left[\frac{q^{y}}{(y+1)^{\alpha}} - \frac{q^{(y+1)}}{(y+2)^{\alpha}}\right] \tag{21g}$$

$$\text{LL8} = \log\left[\frac{\log(1-\alpha\theta^{y}) - \log(1-\alpha\theta^{y+1})}{\log(1-\alpha)}\right] \tag{21h}$$

Maximum-likelihood estimation of (21) is carried out with PROC NLMIXED in SAS, which minimizes the function $-LL(y,\Theta)$ over the parameter space Θ numerically. The integral approximations in PROC NLMIXED is the Adaptive Gaussian Quadrature (Pinheiro & Bates, 1995) and our choice optimization algorithm here is the Newton-Raphson techniques.

4 Applications

4.1 Data Set I

This dataset in Table 1 gives the distribution of the number of Haemocytometer yeast cell counts per square presented in [30] and recently re-analyzed in [36]. The data has observed mean of 0.4599 and variance 0.6046, hence the dispersion index (DI) is $1.3146 > 1$, indicating mild over-dispersion of the data. The results of applications of the models described above are presented in Table 1. The following fit criteria are employed to access the performances of the various models

- Pearson's goodness-of-fit test,
$$X^2 = \sum_{i=0}^{K} \frac{(y_i - \hat{m}_i)^2}{\hat{m}_i}$$
- -2LL: Twice the log-likelihood
- AIC: Akaike Information criterion
- BIC: Bayesian Information Criterion
- Root mean squared error (RMSE) defined as:
$$\sqrt{\frac{1}{K+1}\sum_{i=0}^{K}(y_i - \hat{m}_i)^2}, \quad \text{such that,} \quad \sum_{i=0}^{K}\hat{m}_i = n.$$

- Wald's test Statistic,

$$X^2 = \sum_{j=1}^{n} \frac{(y_j - \hat{m}_j)^2}{\hat{\sigma}_j^2}$$

where n is the number of observations in the data.

Y	Count	P	NB	GPL	NPGL	PEG	TDL	NGPL	NGDP
0	128	118.0627	126.7259	126.9315	126.9314	126.3275	126.5635	126.5636	126.724
1	37	54.2962	42.0847	41.7448	41.7448	42.1846	42.1919	42.1917	42.080
2	18	12.4852	12.8355	12.8874	12.8874	13.3999	12.9668	12.9668	12.849
3	3	1.9140	3.7988	3.8539	3.8539	3.8052	3.7950	3.7950	3.799
4	1	0.2201	1.1071	1.1283	1.1283	0.9809	1.0750	1.0750	1.104
Total	187	186.9781	186.5520	186.5459	186.5459	186.6981	185.5172	186.5922	186.556
		(0.2420)	(1.5551)	(1.5825)	(1.5825)	(1.2828)	(1.4828)	(1.4828)	(1.5487)
y_a		8	15	15	15	13	16	15	16
MLE		$\hat{\lambda}$=0.4599	$\hat{p}$=0.2779	$\hat{\theta}$=2.9037	$\hat{\theta}$=2.9037	$\hat{\alpha}$=4.0146	$\hat{p}$=0.2402	$\hat{\theta}$=3.1631	$\hat{\alpha}$=-0.2278
			$\hat{r}$=1.1950	$\hat{\alpha}$=1.0799	$\hat{\alpha}$=2.0799	$\hat{\theta}$=5.6001	$\hat{\beta}$=0.3879	$\hat{\alpha}$=2.6374	$\hat{q}$=0.2752
μ	0.4599								
σ^2	0.6046								
$\bar{y}$		0.4599	0.4599	0.4601	0.4601	0.4595	0.4599	0.4599	0.4598
s^2		0.4599	0.6369	0.6412	0.6412	0.6138	0.6301	0.6301	0.6363
-2LL		347.7	340.0	340.0	340.0	339.4	339.9	339.9	340.0
AIC		349.7	344.0	344.0	344.0	**343.4**	343.9	343.9	344.0
BIC		352.9	350.5	350.4	350.4	349.8	350.3	350.3	350.5
X^2_W		244.5117	176.5640	175.3745	175.3745	183.2001	178.4725	178.4716	176.7257
d.f.		185	184	184	184	184	184	184	184
pvalue		0.0022	0.6399	**0.6635**	**0.6635**	0.5028	0.6011	0.6011	0.6366
X^2_g		11.7722	3.0711	2.9801	2.9801	2.4712	2.9326	2.9326	2.8658
d.f.		3	2	2	2	2	2	2	2
p-value		0.0082	0.2153	0.2254	0.2254	**0.2907**	0.2308	0.2308	0.2386
RMSE		9.2744	3.3195	**3.1894**	**3.1894**	3.2114	3.3232	3.3231	3.3138

Table 1: Parameter Estimates and Expected Values under three Models

Results:

We observe the following:

- For results in Table 1, all the models have their expected values not summing to $n = 187$ the sample size, within the range of Y. For example, the PEG for instance has $\sum_{i=0}^{4} \hat{m}_i = 186.6981 < 187$. This has been shown to be the case for all discrete distributions ([23], [24]). The $\hat{m}_4 = 0.9809$ reported is not the expected value employed in computing the grouped Pearson's X^2_g, but the 1.2828 in the parentheses, to ensure that $\sum \hat{m}_i = 187$ within $0 \leq Y \leq 4$. Clearly, the sum of expected frequencies 187.6 reported for the NB model in Table 1 of [36] can therefore not be correct.

- The parameter estimates presented in Yahya & Umar for the NB and PEG, as well as the -2LL, AIC and BIC are not reproducible in SAS PROC NLMIXED or in R *optim* application.

- Again for the PEG, the expected values do not sum to $n = 187$ until $y_a = 13$. At $Y = 13$ (which is outside the range of observed Y), the mean and variance of PEG are equal those of the expected theoretical values of 0.4595 and 0.6138 respectively. Indeed, all the models converge to their theoretical means and variances at the values of y_a presented in the Table. All the two-parameter models converge at about y_a equal 15 or 16.

- We observe that the Generalized Poisson-Lindley (GPL) of [25] and the new Poisson-Generalized Lindley (NPGL) presented in [3] are equivalent and the latter is just a re-parameterized version of the GPL. They both give the same estimates of the θ parameter, same -2LL, AIC and BIC values but different α parameter estimates with the NPGL estimate being $1 + \hat{\alpha}_{GPL}$.

- Similarly, the New Generalized Poisson-Lindley (NGPL) presented in [8] behaves very much like the two-parameter Discrete Lindley (TDL) distribution presented in [19].

- The PEG is the most parsimonious model for this data set, amongst the models considered here when the Pearson's X^2 and AIC criteria are employed
- With the Wald's GOF and RMSE criteria however, the GPL and its equivalent NPGL are the most parsimonious.
- All the models considered here with the exception of the Poisson fit this data set well. Both the NGDP and NLD exponentiated models behave similarly and only the results for the former are presented here.

4.2 Example Data Set II

The data set in in this example presented in Table 2 is the number of mistakes in copying groups of random digits [20]. The data has also been re-analyzed in [29] and presented in Table 2 of [36]. The sample size here is $n = 60$, $\mu = 0.7833$ and $\sigma^2 = 1.2573$ leading to a dispersion index DI $= 1.605 > 1$. Thus, the data is over-dispersed.

Y	Count	P	NB	GPL	NPGL	PEG	TDL	NGPL
0	35	27.4128	33.9494	34.3865	34.3866	34.1181	33.2677	33.2677
1	11	21.4734	14.4920	13.9056	13.9054	14.3309	15.0807	15.0807
2	8	8.4104	6.3905	6.3513	6.3512	6.3395	6.6569	6.6569
3	4	2.1961	2.8480	2.9222	2.9222	2.8474	2.8814	2.8814
4	2	0.4301	1.2759	1.3361	1.3362	1.2882	1.2285	1.2285
Total	60	59.9227	58.9558	58.9017	58.9016	58.9241	59.1151	59.1151
		(0.5073)	(2.3201)	(2.4344)	(2.4345)	(2.3641)	(2.1134)	(2.1134)
y_a		10	25	31	23	26	23	21
MLE		$\hat{\lambda}$=0.7833	$\hat{p}$=0.4551	$\hat{\theta}$=1.3876	$\hat{\theta}$=1.3875	$\hat{\alpha}$=0.8288	$\hat{p}$=0.3799	$\hat{\theta}$=1.6324
			$\hat{r}$=0.9381	$\hat{\alpha}$=0.6703	$\hat{\alpha}$=1.6703	$\hat{\theta}$=1.1693	$\hat{\beta}$=0.1933	$\hat{\alpha}$=0.6308
μ	0.7833							
σ^2	1.2573							
$\bar{y}$		0.7833	0.7833	0.7849	0.7849	0.7831	0.7833	0.7833
s^2		0.7833	1.4375	1.4771	1.4771	1.4583	1.3386	1.3386
-2LL		155.1	146.7	146.5	146.5	146.7	146.7	146.7
AIC		157.1	150.7	150.5	150.5	150.7	150.7	150.7
BIC		159.2	154.9	154.6	154.6	154.7	154.9	154.9
X^2_W		94.7022	51.6074	50.2239	50.2228	50.8709	55.4170	55.4171
d.f.		58	57	57	57	57	57	57
pvalue		0.0017	0.6769	**0.7253**	**0.7253**	0.7030	0.5347	0.5347
X^2_g		8.9142	1.7895	1.5211	1.5210	1.7556	1.9058	1.9058
d.f.		3	2	2	2	2	2	2
p-value		0.0305	0.4087	**0.4674**	**0.4674**	0.4157	0.3856	0.3856
RMSE		5.8806	1.8611	**1.6055**	**1.6054**	1.7939	2.1317	2.1317

Table 2: Parameter Estimates and Expected Values under three Models

4.3 Results:

Again the results for this data set are identical to all the observations made for the results in Table 1. Here too, the parameter estimates presented in Table 2 of [36]for the NB and PEG models can not be reproduced here, the exception being the Poisson. For the PEG, θ is correctly estimated but the parameter α is incorrectly estimated. Further, the -2LL, AIC and BIC presented in [36] are un-realized in this study. Again we see here the equivalences that GPL $\equiv$ NPGL and that similarly, NGPL $\equiv$ TDL. Further, the GPL and NPGL are the most parsimonious models for this data set. They both behave better than the PEG.

4.4 Example III: Insurance Data

This example data set is from [2] (p.379) and gives the distribution of the number of accidents in the age-group 26-30 years during the first year of service for a group of railyard shunters and was previously

analyzed in [**?**]. Here $n = 227, \bar{y} = 0.5815$ and $s^2 = 0.5719$. Thus the dispersion index here is $0.9834 < 1$. The data is therefore under-dispersed. The results of applying the PEG model to this data set is presented in Table 3, where the Y's are the number of insurance claims and the counts are the frequencies of each claim for a given number of claims. This example allows us examine the behavior of the PEG model when the data is under-dispersed. In the last two columns are the results from NGDP and NLD models. These models, because of their variance functions allow for under and over-dispersed data.

Y	Count	PEG	GPL	TDL	NGPL	NGDP	NLD
0	121	141.7501	126.9549	121.1050	136.2706	121.7622	125.0343
1	85	54.2834	73.7241	84.1072	61.3464	82.4026	78.4176
2	19	19.9380	21.4578	18.1254	20.7432	19.2040	20.4530
3	1	7.1532	4.1736	3.1021	6.2369	3.1442	2.7443
4	0	2.5269	0.6103	0.4792	1.7583	0.4283	0.3118
5	0	0.8829	0.0716	0.0698	0.4759	0.0521	0.0347
6	1	0.3058	0.0070	0.0098	0.1252	0.0059	0.0038
Total	227	226.8402	226.9994	226.9985	226.9568	226.9993	226.9995
y_a		18	10	10	15	10	10
MLE		$\hat{\alpha}$=1.7454	$\hat{\alpha}$=414.57	$\hat{p}$=0.1177	$\hat{\alpha}$=1000.0	$\hat{\alpha}$=-2.6390	$\hat{\alpha}$=-37.4399
		$\hat{\theta}$=2.0920	$\hat{\theta}$=712.90	$\hat{\beta}$=4.8990	$\hat{\theta}$=3.4329	$\hat{q}$=0.0744	$\hat{\theta}$=0.1103
μ	0.5815						
σ^2	0.5719						
$\bar{y}$		0.5867	0.5815	0.5815	0.5816	0.5826	0.5683
s^2		0.8940	0.5823	0.5181	0.7513	0.5236	0.5182
-2LL		469.7	450.2	445.6	459.0	446.5	447.3
AIC		473.7	454.2	449.6	463.0	450.5	451.3
X^2_g		> 29.1652	4.8385	1.8210	> 18.0	1.9758	2.3065
d.f.		3	3	2	2	3	3
p-value		0.0000	0.1840	**0.4023**	0.0000	0.5774	0.5113
X^2_W		144.5680	221.9356	249.4413	172.0204	246.8261	249.4825
d.f.		224	224	224	224	224	224
p-value		1.0000^a	**0.5264**	0.1169	0.9959^a	0.1412	0.1166
RMSE		14.2442	5.0721	**1.0144**	10.8704	1.3701	3.0671

Table 3: Parameter Estimates, Expected Frequencies and GOFs

Results

We have employed the GPL, the TDL, NGPL, NGDP and the NLD here. The parameter estimates under the NGPL are such that they generated highly inflated standard errors. However, both the PEG and NGPL give over estimated observed variances and consequently fail to fit the data. On the other hand the GPL and TDL behave well and both fit the data well with the GPL being the most parsimonious. Clearly, the PEG and NGPL do not handle under-dispersed data well. The NGDP and NLD also fit the data well and either would be a suitable candidate for under-dispersed frequency count data.

5 Excess Zero Data Examples

5.1 Example IV: Accident Data

The data for this example is presented in [15]. The data is in Table 4 and provides the frequency distribution of number of accidents among 647 machine operators in a fixed period. The observed mean and variance are 0.4652 and 0.6919 respectively, with the dispersion index (DI) being $1.49 > 1$. The data is therefore over-dispersed. The percentage of zeros in the observed data is 69.1% while the corresponding percentage under the Poisson model is $100e^{-0.4652} = 62.8\%$. Thus, this data set exhibits excess zeros. Ignoring the zero outcome in modeling such data usually lead to biasness etc. [18]. We therefore explore the zero-inflated model application to the PEG and subsequently compared with corresponding models for other distributions previously presented.

The probability mass function of a zero-inflated distribution (ZI) is a two-part process manifested by the structural zeros part and the process that generates random counts and can be written in the form proposed in [21]:

$$\Pr(Y=y|\phi)=\begin{cases}\phi+(1-\phi)\ \Pr(Y=0) & \text{if } y_i=0\\ (1-\phi)\ \Pr(Y=y_i)) & \text{if } y_i=1,2,\cdots\end{cases} \tag{22}$$

where $\phi \in (0,1)$ is the extra proportion of zeros and Y is the count random variable with specified parameters. A constant inflation parameter ϕ is modeled here in the logit form. That is, $\phi = 1/[1+\exp(-a0)]$.
In Table 4 are the results of applications of zero-inflated models to this data set.

		Regular Models				Zero-Inflated Models			
Y	Count	NB	GPL	NGPL	PEG	ZINB	ZIGPL	ZINGPL	ZIPEG
0	447	445.8864	446.3985	441.5707	446.3011	447.0000	447.0000	447.0000	447.0000
1	132	134.8957	133.7239	140.2033	134.3026	130.1862	130.2621	130.6287	130.3477
2	42	43.9920	44.4548	44.5160	44.0433	47.7803	47.6738	47.2299	47.4009
3	21	14.6924	14.9354	14.1343	14.7665	15.4765	15.4760	15.4174	15.6667
4	3	4.9647	5.0017	4.4878	4.9997	4.6795	4.6960	4.7485	4.7521
5	2	1.6893	1.6652	1.4249	1.7021	1.3544	1.3638	1.4086	1.3468
Total	647	646.1205	646.1795	646.3371	646.1154	646.4768	646.4717	646.4332	646.5141
y_a		19	18	18	21	18	21	16	18
		$\hat{p}$=0.3497	$\hat{\alpha}$=0.7364	$\hat{\theta}$=2.1495	$\hat{\alpha}$=0.7013	$\hat{p}$=0.2377	$\hat{\alpha}$=1.9492	$\hat{\alpha}$=19.2937*	$\hat{\alpha}$=3.7745
		$\hat{r}$=0.8651	$\hat{\theta}$=2.2446	$\hat{\alpha} \approx 0.000$	$\hat{\theta}$=1.8884	$\hat{r}$=2.0880	$\hat{\theta}$=33799	$\hat{\theta}$=3.0183	$\hat{\theta}$=4.3792
						$\hat{\phi}$=0.2855	$\hat{\phi}$=0.2779	$\hat{\phi}$=0.2470	$\hat{\phi}$=0.1562
μ	0.4652								
σ^2	0.6919								
$\bar{y}$		0.4652	0.4654	0.4652	0.4652	0.4652	0.4652	0.4652	0.4652
s^2		0.7154	0.7150	0.6817	0.7175	0.6968	0.6973	0.7000	0.6961
X^2_g		3.9091	3.5172	4.5209	3.6996	3.3066	3.2889	3.2592	3.1132
d.f.		3	3	3	3	2	2	2	2
p-value		0.2714	**0.3185**	0.2104	0.2958	0.1914	0.1931	0.1960	0.2109
X^2_W		624.8	625.11	655.7069	622.9144	641.4920	640.9940	638.5115	642.0974
d.f.		644	644	644	644	643	643	643	643
-2LL		1184.5	**1184.3**	1185.0	1184.5	1183.8	1183.8	1183.8	1183.6
AIC		1188.5	1188.3	1189.0	1188.5	1189.8	1189.8	1189.8	1189.6
BIC		1197.5	1197.2	1197.9	1197.4	1203.2	1203.2	1203.2	1203.0
RMSE		3.0974	**2.8978**	5.0409	2.9767				

Table 4: Distribution of Number of accidents among machine operators

Results:

The results from Table 4 indicate the following:

- For all the models, the sum of expected values did not sum to the sample size $n = 647$ until y_a which is outside the range of the data, $0 \le Y \le 5$. Thus the theoretical means and variances of these models are not realized until y_a.
- For the regular models, the most parsimonious model is the Generalized Poisson-Lindley (GPL). The New generalized Poisson-Lindley give an approximate parameter estimate for $\alpha \approx 0.0000$ indicating that for this data set, the GPL converges to the Geometric distribution with parameter,

$$\left(\frac{\hat{\theta}}{1+\hat{\theta}}\right)=\frac{2.1495}{3.1495}=0.6825.$$

  That is, a one-parameter geometric distribution GD (0.6825).
- All the regular models estimate the observed mean of 0.4652 well, but overestimate the observed variance of 0.6919, the exception being the NGPL.
- Because of strong deviations of these estimated variances from the observed variance, we need to fit zero-inflated corresponding models to this data set which exhibits mild excess zeros to ameliorate these deviations.

- The results of the ZI models are presented in the last four columns of Table 4. With exception of the NGPL, all the estimated variances in the regular models are greater than the observed variance. The zero-inflated models on the other hand tempered these estimates by lowering their estimates, thus bring them closer to the observed variance of 0.6919. Here, however, the ZIPEG is the most parsimonious model.
- The ZINGPL provides estimate for α whose estimated standard error is extremely large-thus casting doubt on the adequacy of this model for zero-inflated data.

5.2 Example V

This example is again taken from [**?**] and relate to claim counts of third party liability vehicle insurance in a Zaire insurance company [34]. The data in Table 5 are therefore the distribution of claims from 4000 vehicle polices.

		Regular Models				Zero-Inflated Models			
Y	Count	NB	GPL	NGPL	PEG	ZINB	ZIGPL	ZINGPL	ZIPEG
0	3719	3719.2220	3718.7790	3681.5495	3719.0678	3719.2243	3719.0000	3718.9997	3719.0760
1	232	229.9009	229.5918	293.0978	228.5440	229.8990	229.2586	228.2113	228.5374
2	38	39.9106	41.3967	23.3343	42.0282	39.9103	41.3928	42.8720	42.0270
3	7	8.4156	8.1604	1.8577	8.2743	8.4155	8.2272	8.0540	8.2741
4	3	1.9313	1.6484	0.1479	1.6624	1.9313	1.6823	1.5130	1.6624
5	1	0.4648	0.3361	0.0118	0.3370	0.4648	0.3478	0.2842	0.3370
Total	4000	3999.8453 (0.6195)	3999.9135	3999.9990	3999.9138	3999.8453	3999.9088	3999.9343	3999.9138
y_a		-	13	10	14	15	14	14	14
		$\hat{p}$=0.2854	$\hat{\alpha}$=0.1332	$\hat{\alpha}$=0.0014	$\hat{\alpha}$=0.0906	$\hat{p}$=0.2854	$\hat{\alpha}$=0.2264	$\hat{\alpha} \approx 0.000$	$\hat{\alpha}$=0.0906
		$\hat{r}$=0.2166	$\hat{\theta}$=3.9018	$\hat{\theta}$=11.5622	$\hat{\theta}$=3.8597	$\hat{r}$=0.2166	$\hat{\theta}$=3.8565	$\hat{\theta}$=4.3231	$\hat{\theta}$=3.8597
						$\hat{\phi} \approx 0.000$	$\hat{\phi}$= 0.2289	$\hat{\phi}$= 0.6261	$\hat{\phi}$= 0.0001
μ	0.0865								
σ^2	0.1225								
$\bar{y}$		0.0865	0.0864	0.0865	0.0866	0.0865	0.0864	0.0865	0.0866
s^2		0.1210	0.1192	0.0940	0.1199	0.1210	0.1196	0.1190	0.1199
X_g^2		1.1738	2.3660	> 36.000	2.4969	1.1736	2.2429	3.4229	2.4970
d.f.		3	3	2	3	2	2	2	2
p-value		0.7593	0.5000	0.0000	0.4759	0.5561	0.3258	0.1806	0.2869
X_W^2		4048.6975	4110.1005	5214.5678	4087.5685	4048.7284	4099.2354	4117.0158	4087.6800
d.f.		3997	3997	3997	3997	3996	3996	3996	3996
-2LL		2367.1	2367.9	2414.8	2368.0	2367.1	2367.8	2368.6	2368.0
AIC		2371.1	2371.9	2418.8	2372.0	2373.1	2373.8	2374.6	2374.0
BIC		2383.7	2384.4	2431.4	2384.6	2392.0	2392.7	2393.5	2392.9

Table 5: Distribution of claims from an Insurance Company

The observed data has a mean of 0.0865 and thus under the Poisson model the percentage of expected zeros would be exp(−0.0865)=91.72%. However the observed data has about 93.98% zeros. Because the percentage of observed zeros is not too far from that expected under the Poisson, the data therefore exhibits moderate excess zeros.

Results:

Results from Table 5 indicate the following observations:

- All the regular models (with the exception of the NGPL) fit the data very well, with the NB being the most parsimonious. The NGPL does not fit well at all.
- The zero-inflated corresponding models do not show much improvemts but the estimated variances are adjusted upwards to match better the observed variance in both ZIGPL and ZINGPL.
- The ZINGPL gives a parameter estimate $\hat{\alpha} \approx 0.0000$ indicating again convergence to the zero-inflated geometric model with estimated parameter $\hat{r} = 0.8121$ which is equivalent to $\hat{r} = \dfrac{\hat{\theta}}{1+\hat{\theta}} = \dfrac{4.3231}{5.3231} = 0.8121$. Further, the ZINGPL now fits the data with a p-value of 0.1806.

- The NB is the most parsimonious among both the regular and zero-inflated models. It performs much better than the PEG. So too is the regular GPL model and its corresponding zero-inflated ZIGPL model.

6 GLM Applications

Our example data here is the U.S. Medical Expenditure Panel Survey (MEPS) data set relating to the number of doctor visits (Y=docvis) in 2003 for a number of elderly patients as well as several other covariates relating to patients' characteristics (Hilbe, [17]. The covariates are:

- private insurance coverage (supplemental to Medicare) (0,1)
- medicaid-eligibility for low income Medicaid coverage (0,1)
- female-gender of patients (1 if female, 0 if male)
- actlim-limitation of activity (0,1)
- totchr-number of chronic conditions
- phylim-physical limitation (0,1)
- educyr-number of years of educational attainment.

We present the first and last five observations for this data set ($n = 3677$).

Obs	docvis	female	phylim	private	medicaid	educyr	actlim	totchr
1	4	1	0	1	0	15	0	3
2	6	1	1	0	0	8	1	2
3	2	1	1	0	1	11	0	2
4	11	0	0	1	0	13	0	3
5	3	1	0	1	0	14	0	1
3671	5	1	1	1	0	16	0	1
3672	2	0	0	0	0	6	1	2
3673	15	1	1	0	1	12	1	3
3674	8	1	1	1	0	9	1	6
3675	6	1	0	1	0	13	0	2
3676	14	1	1	0	0	3	1	2
3677	10	0	1	0	0	4	1	1

We also created the interaction term fem_{edu} of **female** and **educyr**.
For data having covariates $x_1, x_2, \ldots, x_p$, the linear predictor therefore becomes,

$$\mathbf{x}'\boldsymbol{\beta} = \beta_0 + \beta_1\text{female} + \beta_2\text{phylim} + \beta_3\text{private} + \beta_4\text{medicaid} + \beta_5\text{educyr} + \beta_6\text{actlim} + \beta_7\text{totchr} + \beta_8\text{fem}_{\text{edu}}$$

the parameters μ_i in NB and PIG; θ_i in GPL,NPGL and PEG; GP; and p_i in TDL are modeled respectively as follows:

$$\mu_i = \exp(\mathbf{x}'\boldsymbol{\beta});\ \theta_i = \exp(\mathbf{x}'\boldsymbol{\beta});\ \theta_i = \exp(\mathbf{x}'\boldsymbol{\beta} + \text{offset});\ p_i = 1/[1 + \exp(-\mathbf{x}'\boldsymbol{\beta})];\ i = 1, 2, \ldots, 3677.$$

Here, the p_i, are modeled in the logit form. We may first note here that the data is grossly over-dispersed with a dispersion parameter DI=6.5178 under the Poisson model. The results therefore in Table 6 gives the performances of the distributions considered as alternatives to the Poisson. For GLM analyses, we shall employ the fit criteria to access the performances of the various models, -2LL, AIC, BIC, and Wald's X^2.

Parameter	Distributions							
	NB	GPL	NGPL[b]	TDL	NPGL[a]	PEG	GP	PIG
Intercept	0.6412	-0.1840	-1.1701	0.1549	-0.1840	-0.1950	0.7806	0.7707
female	0.2870*	-0.2603*	-0.2606*	0.2632*	-0.2603*	-0.2619*	0.2939*	0.3036*
phylim	0.1975*	-0.1800*	-0.1802*	0.1826*	-0.1800*	-0.1810*	0.1866*	0.1878*
private	0.1445*	-0.1313*	-0.1314*	0.1326*	-0.1313*	-0.1321*	0.1428*	0.1450*
medicaid	0.0835	-0.0753	-0.0754	0.0777	-0.0753	-0.0756	0.0302	0.0290
educyr	0.0445*	-0.0407*	-0.0408*	0.0415*	-0.0407*	-0.0409*	0.0390*	0.0390*
actlim	0.0761	-0.0694	-0.0695	0.0719	-0.0694	-0.0694	0.0245	0.0176
totchr	0.2682*	-0.2458*	-0.2460*	0.2482*	-0.2458*	-0.2471*	0.2450*	0.2479*
fem_{edu}	-0.0278*	0.0254*	0.0254*	-0.0258*	0.0254*	0.0255*	-0.0268*	-0.0272*
ML	$\hat{r}$=1.5677*	$\hat{\alpha}$=0.9772*	$\hat{\alpha}$=1.0124*	$\hat{\beta}$=0.4399*	$\hat{\alpha}$=1.9772*	$\hat{\alpha}$=1.9360*	$\hat{\delta}$=0.5839*	$\hat{\beta}$=5.3636*
-2LL	21166	21141	21141	21154	21141	21140	21020	21065
AIC	21186	21161	21161	21174	21161	21160	21040	21085
BIC	21248	21223	21223	21236	21223	21222	21103	21147
X^2_W	4569.7399	4598.2752	4649.5680	4651,6360	4600.0054	4576.8044	4015.1579	3651.8763
d.f.	3667	3667	3667	3667	3667	3667	3667	3667
p-value	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00004	0.56712

Table 6: Parameter estimates & GOF statistics for the Distributions

[b] Bhatti *et al (2019)*, [a] Atikankul(2023), [a] Significant at $\alpha = 0.01$.

A Poisson model applied to the response variables has Wald's $X^2_W = 29,463.8387$ on 3676 d.f., giving a dispersion parameter of 8.0152, which clearly indicates very strong over-dispersion, considering the size of the data. In Table 6 are the estimated parameters, together with their computed Wald's GOFs and corresponding -2LL, AIC and BIC values.
In the last two columns of Table 6 are the results of applying both the generalized Poisson of the second kind (GP), [12] and the Poisson-Inverse Gaussian (PIG), Wilmot [34] models to the data. Results in Table 6 indicate the following:

- Based on the -2LL, AIC and BIC, the Generalized Poisson has the smallest values and would be considered the most parsimonious on these criteria. This is followed by the PIG. The PEG gives lower AIC and BIC than the other models and may be preferable over the NB, GPL, TDL and the NPGL.
- Based on the Wald's test statistic, X^2_W however, the PIG is the only model that fits the data with a p-value of (0.56712). All the other models fail to fit the data. Although,
- the GP has the lowest AIC criterion, but the PIG gives the only significant or better p-values based on Wald's GOF. This is why it is important that we should employ as many criteria as possible in selecting the most parsimonious model.
- We observe again that the GPL and NPGL are equivalent with the latter giving a +1 parameter estimate for $\hat{\alpha}$.
- The Models NB, GP and PIG are presented here because they are all parameterized in terms of their means μ. We observe that both the GP and PIG provide better fits with the former being more parsimonious in terms of AIC and BIC, while the latter performs better based on the Wald's goodness-of-fit test statistic $X^2_W = \sum_{i=1}^{N} \frac{(y_i - \hat{m}_i)^2}{\hat{\sigma}^2_i}$.

6.1 Under-dispersed GLM

The asthma inhaler data presented in Greenwald *et al.* [16] and [11], which comprises of 5209 daily count observations from 48 children suffering from asthma during the school day, for a certain period in Denver, Colorado. The students are aged 6 to 13 years. The covariates here are: (i) the percentage of humidity (humidity), (ii) the barometric pressure (in mmHG/1000-pressure), (iii) the average daily temperature (in Fahrenheit degree/100- temperature), and (iv) the morning levels of PM25, which are small air particles

less than 25mm in diameter (particles). The linear predictor here is $\mathbf{x}'\boldsymbol{\beta} = \beta_0 + \beta_1 \text{humidity} + \beta_2 \text{pressure} + \beta_3 \text{temperature} + \beta_4 \text{particle}$. The response variable (Y), which is the inhaler use count, has a sample mean of 1.2705, sample variance of 0.9183, and therefore, a dispersion index DI = 0.6637 < 1. Thus, the data is under-dispersed. In Table 7 are the results of the applications of some of the models discussed above to the inhaler data.

	Distributions						
Parameter	P	GPL	NPGL*	TDL	NGDP	NLD	DW
Intercept	-2.2132	9.0891	9.6351	-5.2305	-4.9008	-3.7526	1.6193
humidity	-0.1125	0.1126	0.1123	-0.1974	-0.0064	-0.3741**	0.2477**
pressure	4.0950	-4.0987	-4.2543	6.6845	4.9708**	2.7998	-5.1593
temperature	-0.2035	0.2036	0.2023	-0.3553	-0.0471	-0.5706**	0.4150**
particles	0.0225	-0.0225	-0.0226	0.0380	0.0217**	0.0346	-0.0317
ML		$\hat{\alpha}$=962.71	$\hat{\alpha}$=1513.98	$\hat{\beta}$=11.0305	$\hat{\alpha}$=-2.5050*	$\hat{\alpha}$=-1790.77	$\hat{\beta}$=2.1282*
-2LL	13905.47	13908	13907	13500	13448	14,134	13,472
AIC	13915.47	13920	13919	13512	13460	14,146	13,484
BIC	13948.26	13959	13958	13551	13500	14,186	13,524
X^2_W	3448.93	3444.34	3446.25	4570.63	4898.21	4332.80	5162.92
d.f.	5204	5203	5203	5203	5203	5203	5203
p-value	1.0000	1.0000	1.0000	1.0000	0.9988	1.0000	0.6508

Table 7: Parameter estimates & GOF statistics for the Distributions

* Atikankul (2023) Model; * significant at the 5% point

Results:

Results from Table 7 indicate that the Poisson, the Generalized Poisson-Lindley (GPL) and the New Poisson generalized Lindley (NPGL) did not improve on the under-dispersion of the data. The GPE and the NGPL ([8] did not converge and as earlier stated, they both are not suitable for strong under-dispersed count data. However, in the last two columns are results from applying (a) the New Geometric Discrete Pareto Distribution (NGDP), NLD and the Discrete Type I Weibull (DW) distribution with pmf [26]:

$$f(y|q,\beta) = q^{y^\beta} - q^{(y+1)^\beta} \quad y = 0, 1, \ldots \tag{23}$$

The parameter q is modeled in the logit form, viz: $q = 1/(1 + \exp(-\mathbf{x}'\boldsymbol{\beta}))$. The NGDP and DW perform much better than the others and both seem capable of handling under-dispersed count data well.

6.2 Case of Mixed-Level of Dispersion

The bids data taken from [10] is a data set with mixed level of dispersion, where the the conditional distribution is over-dispersed relative for some covariate pattern but is under-dispersed for another covariate pattern. The response variable (Y) here is the number of bids received by 126 US firms that were targets of tender offers during a certain period of time. The response variable Y, has $\bar{y} = 1.7381$ and $s^2 = 2.0509$ with a DI=1.180, which indicates moderate over-dispersion and $0 \leq Y \leq 10$. The following covariates are employed in our analysis. In Table 8 are the results of applications of our models to this data set.

- bid price-taken as the price at a particular week divided by the price 14 working days before the bid,
- size-that is, the total book value of assets measured in billions of dollars,
- regulator, a dummy variable that is equal to 1 if there was an intervention by federal regulators and 0 otherwise.

Results:

The results in Table 8 indicate the followings:

Parameter	Distributions						
	P	NB	GPL	NPGL*	TDL	PEG	DW
Intercept	1.5318	1.5276	1.9857	1.9857	0.9212	1.9788	-3.5165
price	-0.7849	-0.7824	0.7814	0.7814	-1.4240	0.7824	1.4767
size	0.0362	0.0369	-0.0368	-0.0368	0.0687	-0.0369	-0.1126
regulator	0.0547	0.0544	-0.0543	-0.0543	0.0919	-0.0544	-0.0569
ML		$\hat{r}$=33.3289	$\hat{\alpha}$=33.4711	$\hat{\alpha}$=34.4710	$\hat{\beta}$=23.0210	$\hat{\alpha}$=33.3289	$\hat{\beta}$=1.9409
-2LL	394.3	393.9	393.9	393.9	361.3	393.9	385.1
AIC	402.3	403.9	403.9	403.9	371.3	403.9	395.1
BIC	419.6	418.1	418.1	418.1	385.5	418.1	409.3
X^2_W	123.2578	116.3302	116.3641	116.3641	144.2486	116.3302	126.5202
d.f.	122	121	121	121	121	121	121
p-value	0.4511	0.6030	0.6021	0.6021	0.6552	0.6030	0.3474

Table 8: Parameter estimates & GOF statistics for the Distributions

* Atikankul (2023) Model

- The PEG, NB, GPL and NPGL (Atikankul, 2023) all behave alike for this special data set. The parameters of the NB are the negatives of those of the other three because of parameterization. The parameter α for the NPGL is 34.4710 as against that of GPL of 33.4711 as expected since $\boldsymbol{\alpha}_{ngpl} = 1 + \boldsymbol{\alpha}_{gpl}$ theoretically.
- All the four produce the same -2LL, AIC, BIC as well as equivalent Wald's goodness-of-fit statistic X^2_W.
- The dispersion parameters $\hat{\sigma}^2_1/\hat{\mu}_i,\ i = 1, 2, \ldots, 126$ for the Poisson, NB, GPL, NPGL and PEG did not indicate the mixed level of the data as reflected in Table 9. For these five distributions, the minimum DI and maximum DI are all greater than 1 reflecting over-dispersion as expressed as a variance-function for these models. The exception of course being the Poisson, with DI being 1 across the entire data.
- However, both TDL and DW models exhibit the inherent mixed level of dispersion in the data, with the DI being less than 1.00 for some observations-indicating under-dispersion and being greater than 1.00 for other observations-indicating over-dispersion. These two distributions would therefore be suitable for this data set with the TDL being the most parsimonious of the two in terms of AIC and BIC, while the DW will be the most parsimonious of the two if the Wald's GOF criterion is employed.

Models	$\hat{\mu}_i$		$\hat{\sigma}^2_i$		DI	
	Min	Max	Min	Max	Min	Max
P	0.9241	4.0316	0.9241	4.0316	1.0000	1.0000
NB	0.9253	4.0777	0.9510	4.5766	1.0278	1.1223
GPL	0.9254	4.0773	0.9510	4.5738	1.0276	1.1218
NPGL	0.9254	4.0773	0.9510	4.5738	1.0276	1.1218
TDL	0.9948	5.0656	0.5070	12.7112	0.5097	2.5093**
PEG	0.9253	4.0777	0.9510	4.5766	1.0278	1.1223
DW	0.8037	7.0230	0.5608	16.4058	0.6978	2.3360 **

Table 9: Minimum and Maximum Values of Estimated moments with corresponding DI

** Exhibit Under and over-dispersions

6.3 Zero-Inflated GLM Application:

In this section, we would compare the PEG model to the other models considered in this study to a data set exhibiting excess zeros.an example Our example data here is the very well analyzed German National Health Registry (GNHR) [17] data set which comprises of 3874 respondents. The data has 3874

observations (it is a subset of the main data set, which has 27,300 observations) and 16 variables including the response variable **docvis**, number of doctor's visits. We have however presented below the first and last five observations from this subset data and the five explanatory variables of interest, namely, gender (1=female, 0 if male), the age of the individual (age); kids-number of children, educ-years of education and marital status (1 if married; 0 if not married).
The response variable Y has a range [0,121] with mean 3.162881 and variance 39.387611. Thus, the dispersion index (ID) here is 12.4531, indicating a very strong over-dispersion. In addition, 41.58% of the data have zero counts which indicate excess zeros in the data.

Our model formulation here is based on the five explanatory variables employed in Saffari *et al.* (2019), namely, **sex, age, kids, educ** and **marital status**. If we let the linear predictor $\mathbf{x}\boldsymbol{\beta}$ be defined as in (24),

$$\mathbf{x}\boldsymbol{\beta} = b_0 + b_1\text{female} + b_2\text{age} + b_3\text{children} + b_4\text{educ} + b_5\text{married} \tag{24}$$

Then, the mean $\mu = rp/(1-p)$ for the Negative binomial, the parameter θ_i in GPL, NPGL and PEG are modeled as $\exp(\mathbf{x}\boldsymbol{\beta})$, while the parameter p in the TDL is modeled in the logit form $p = 1/(1+\exp(-\mathbf{x}\boldsymbol{\beta}))$. The zero-component of the zero-inflated models is also modeled in the logit form, viz:

$$\log\left(\frac{\phi_i}{1-\phi_i}\right) = a_0 + a_1\text{female} + a_2\text{age} + a_3\text{children} + a_4\text{educ} + a_5\text{married} \tag{25}$$

The results of implementing these models are presented in Table 10.

	ZINB	ZIGPL	ZIPGL*	ZITDL	ZINGPL**	ZIPEG	ZIGD	ZINGDP	ZINLD
Log link Parms									
Intercept	0.6112	-0.9108	-0.9108	1.0005	1.0005	-1.1741	1.0005	1.3580	1.1769
female	0.2142	-0.2122	-0.2122	0.1839	0.1839	-0.2179	0.1839	0.2274	0.2013
age	0.0205	-0.0176	-0.0176	0.0173	0.0173	-0.0216	0.0173	0.0234	0.0199
kids	-0.0219	0.0203	0.0203	-0.0088	-0.0088	0.0180	-0.0088	0.0042^a	-0.0119^a
educ	-0.0233	0.0211	0.0211	-0.0270	-0.0270	0.0269	-0.0270	-0.0391	-0.0214^a
married	-0.1724	0.1448	0.1448	-0.1718	-0.1718	0.1885	-0.1718	-0.2636	-0.1756
	$\hat{r}$= 0.5435	$\hat{\alpha} \approx 0.000$	$\hat{\alpha}$=1.000	$\hat{\beta} \approx 0.000$	$\hat{\alpha} \approx 0.000$	$\hat{\alpha}$=0.5292	na	$\hat{\alpha}$=0.3455	$\hat{\alpha}$=0.8336
Logit link Parms									
Intercept	-1.6058	-1.6866	-1.6855	-0.1399	-0.1399	-1.4677	-0.1389	-0.9557	-1.1490
female	-2.2433	-1.2951	-1.2951	-0.8232	-0.8232	-2.1893	-0.8232	-1.7995	-1.3830
age	-0.0333	-0.0203	-0.0203	-0.0235	-0.0235	-0.0348	-0.0235	-0.0379	-0.0242
kids	0.8220	0.6802	0.6802	0.3731	0.3731	0.7974	0.3731	0.7113	0.6496
educ	0.1291	0.0928	0.0928	0.0437	0.0437	0.1252	0.0437	0.1059	0.0870
married^a	-0.4016	-0.4311	0.4311	-0.1733	-0.1733	-0.3682	-0.1733	-0.2875	-0.3823
-2LL	16565	16668	16668	16662	16662	16,570	16,662	16,563	16,553
AIC	16591	16694	16694	16688	16688	16,596	16,686	16,589	16,577
BIC	16672	16776	16776	16769	16769	16,678	16,761	16,670	16,660
X^2	5799.637	6967.963	6968.041	7010.514	7010.816	5803.973	7010.816	5441.204	5478.641
d.f.	3861	3861	3861	3861	3861	3861	3862	3861	3861
p-value	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
$p_{textzero}$	0.4192	0.4309	0.4309	0.4157	0.4157	0.4195	0.4157	0.4205	0.4157

Table 10: Parameter Estimates
* Atitunku; ** Bhatti *et al.*, ^{a} not significant.

6.4 Results

Results from Table 10 indicate the followings:

- Zero-inflated models ZITDL and ZINGPL do not converge readily and are most intractable to implement.
- None of the models fit the data. However, the ZINLD (the zero-inflated New logarithmic Distribution,[14]) ZINB outperforms others in terms of having lowest -2LL, AIC and BIC. This is followed by the ZINGDP (the zero-inflated New Geometric Discrete Pareto Distribution, [8]). These are closely followed by the ZINB (zero-inflated Negative Binomial).

- In terms of the Wald's GOF criterion, the ZINGD has the lowest value of the test statistic, 5,441.204 on 3861 d.f.
- Covariate parameter estimates of ZIGPL and ZIPGL are identical as previously observed. The $\hat{\alpha}$ parameter estimated under the ZIGPL is approximately zero and the corresponding estimate for the ZINPGL is as expected 1.00.
- For the ZITDL and ZINGPL, they similarly provide equivalent covariates' parameter estimates. However, the parameter estimate $\hat{\beta}$ and $\hat{\alpha}$ are approximately zeros. Thus, the two distributions reduce to zero-inflated geometric distribution. The last column in Table 10 gives the estimated parameters and goodness-of-fit test statistics under the ZIGM. We observe that these are identical to those of ZITDL and ZINGPL.
- The estimated proportion of zeros from our models are presented as p_{zero} for all the models, compared to the observed (1611/3874)=0.4158. In this regard the ZINLD, ZITDL, its equivalent ZINGPL and the ZIGD provide close estimates while the ZINB is also reasonably close.
- All the models did not fit this data because the data is strongly skewed to the right. The range of the response variable is [0,121] but 97.91% of the data are in the range [0,20]. Thus, right truncated model would hopefully be appropriate for this data set.

7 Conclusions:

Results from this study have shown that the Poisson-Exponential-Gamma distribution proposed in [36] does not necessarily outperform the GPL [25], the NB or other similar two-parameter distributions. It does not fair well for under-dispersed count data but is a good alternative to count data exhibiting over-dispersion. We also established here that the New Poisson generalized Lindley distribution (NPGL) proposed in [3] is a re-parameterization of the generalized Poisson Lindley (GPL) distribution earlier proposed in [25], while the two Lindley-type distributions can provide adequate fits for some frequency data sets as considered in this study and some data having covariates, the two distributions are however not suitable for data exhibiting very strong over-dispersion as a result of excess zeros. The choice of which model is the most parsimonious depends on the data sets under consideration based on the variety of data sets considered in this study. However, the PEG distribution re-considered here adds to the literature and knowledge of discrete distributions that can be applied to over-dispersed count data as they provide alternatives for fitting any variety of over-dispersed count data. It should be noted here that the GPL converges faster than the NPGL. For mixed under and over-dispersed data, the NB, GPL and NPGL seem impervious of this as they return estimated dispersion indices that are greater than 1.0000.

Disclaimer (Artificial Intelligence)

Author(s) hereby declare that NO generative AI technologies such as Large Language Models (ChatGPT, COPILOT, etc) and text-to-image generators have been used during writing or editing of manuscripts.

Acknowledgments

The author is grateful to two anonymous referees for their comments that have improved this paper.

Competing Interests

There are no competing interests.

References

[1] Abouammoh, A.M, Alshangiti, A.M. & Ragab IE (2015). A new Generalized Lindley Distribution. *Journal of Statistical Computation and Simulation*, **85**(10),3662-3678.

[2] Adelstein,, A. M. (1949) Accident Proneness: A Criticism of the Concept Based Upon an Analysis of Shunters' Accidents. *Journal of the Royal Statistical Society*, A(115), 354–410

[3] Atikankul, Y. (2023). A New Generalized Lindley Regression Model. *Austrian Journal of Statistics*, 52, 39-50.

[4] Awad, Y.S., Bar-Lev,S.K. & Makov, U. (2016). A new class counting distributions embedded in the Lee-Carter model for mortality projections. *A Bayesian approach. Technical report*, No 146. Actuarial Research Center, University of Haifa. Israel.

[5] Barbiero, A. (2015). *Discrete Weibull Distributions (Type I and 3)*. R package version 1.0.1

[6] Barbiero, A. (2017). Discrete Weibull regression for modeling football outcomes. *Int. Jour. Business Intelligence & Data Mininig.*

[7] Bhati, D., Sastry, D.V., and Qadri, P.M. (2015). A new generalized Poisson-Lindley Distribution: Applications and Properties. *Austrian Journal of Statistics*, 11, 35–51.

[8] Bhati,D. and H.S. Bakouch (2019). A new infinitelydivisible discrete distribution with applications to count data modeling. *Communications in Statistics-Theory and Methods*, 48 (6): 140–16.

[9] Beall G (1940). The Fit and Significance of Contagious Distributions when Applied to Observations on Larval Insects. *Ecology*, **21** (4),460-474.

[10] Cameron, A.C. & Per Johansson (1997). Count Data Regression Models using Series Expansions: with Applications, *Journal of Applied Econometrics*, 12, 203-223.

[11] Canale, A. & Dunson, D.B. (2012). A Bayesian nonparametric model for count functional data. In *46th Scientific meeting of the Italian Statistical Society.*

[12] Consul, P.C. & Famoye, F. (1992). Generalized Poisson Regression-Model. *Communications in Statistics-Theory and Methods*, 21(1), 89-109.

[13] Deb, P., Trivedi, P.K. (1997). Demand for Medical Care by the Elderly: A Finite Mixture Approach. *Journal of Applied Econometrics*, 12, 313–336.

[14] Gómez-Déniz, E., Hernández-Bastida, A., and Fernández-Sánchez, M.P.A. (2016). A Suitable Discrete Distribution for Modelling Automobile Claim Frequencies. *Bulletin of the Malaysian Mathematical Sciences Society*, 39, 633–647.

[15] Greenwood, M., & Yule, U. (1920). An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents. *J. Roy. Statist. Soc. Ser. A* 83:255-279.

[16] Grunwald, G.K., Bruce,S.L., Jiang,L., Strand,G.K. & Rabinovitch, N.A. (2011). Statistical model for under-or-overdispersed clustered and longitudinal count data. *Biometric Journal*, 53(4), 578–594.

[17] Hilbe, J.M. (2007). German health data for 1984. *Negative Binomial Regression. Cambridge University Press.*

[18] Hilbe, J.M. (2014). *Negative Binomial Regression*. Cambridge University Press.

[19] Hussian, T., Aslam,M., & Ahmad, M. (2016). A two parameter Discrete Lindley Distribution.*Revista Colombiana de Estadistica*, 39 (1), 45–61.

[20] Kemp, C. D., & Kemp, A. W, (1965). Some properties of the Hermite Distribution. *Biometrika*, **52** (3-4), 381-394.

[21] Lambert, D. (1992). Zero-inflated Poisson Regression, with an application to Defects in Manufacturing. *Technometrics*, 34(1), 1–14.

[22] Lawal, H.B. (1980). Tables of percentage points of Pearson's goodness-of-fit statistic for use with small expectations. *Applied Statistics*, 29, 292–298.

[23] Lawal, H.B. (2018). Correcting for Non-Sum to 1 Estimated Probabilities in Applications of Discrete Probability Models to Count Data. *International Journal of Statistics and Probability*, 6, 119–131.

[24] Lawal, H.B. (2019). Quasi-Negative Binomial, Inverse Trinomial and Negative Binomial Generalized Exponential Distributions and their Applications. *Jour. Prob. Statist. Science*, 17(2), 205-221.

[25] Mahmoudi, E. & Zakerzadeh,H. (2010). Generalized Poisson-Lindley Distribution. *commun. Statist., Theory and Methods*, 39(10), 1785–1798.

[26] Nagakawa, T. and Osaki, S. (1975). The Discrete Weibull distribution. *IEEE transactions on Reliability*, 24(5), 300–301.

[27] Pinheiro, J.C. & Bates,D.M.(1995). Approximations to the Log-Likelihood Function in the Nonlinear Mixed-Effects Model *Journal of Computational and Graomegacal Statistics*,4: 12–35.

[28] Saffari,S.E., Allen J.C., Adnan, R., Ong, S.H., Sim S. Z., & Green, W. (2019). Frequency of visiting Doctor: A right truncated count regression model with excess zeros. *Biostatistics and Biometrics Open Access Journal*, 9(5):555773.

[29] Sankaran, M. (1970). The discrete Poisson-Lindley distribution, *Biometrics*, 26: 145–149.

[30] Shanker, R. and Hagos,F. (2015). On Poisson-Lindley Distribution and its Applications to Biological Sciences. it International Journal of Biometrics and Biostatistics, 2(4):00036

[31] Shanker, R., Sharma, S. and Shanker, R. (2013). A Two-Parameter Lindley Distribution for Modeling Waiting and Survival Time Data. *Applied Mathematics*, 4, 363–368.

[32] Umar, M. A. (2019). A zero-truncated Poisson-Exponential-Gamma Distribution and its Applications. M.Sc. Dissertation, University of Ilorin, Ilorin. Nigeria.

[33] Umar, M. A., Yahya, W. B.(2021). A New Exponential-Gamma Distribution with Applications. *Journal of Modern Applied Statistical Methods*. https://digitalcommons.wayne.edu/jmasm/about.html

[34] Willmot, G.E. (1987). The Poisson-inverse Gaussian distribution as an alternative to the negative binomial. *Scan Actura J.*, (3-4); 113–127.

[35] Wongrin W, Bodhisuwan W (2017). Generalized Poisson Lindley Linear Model for Count Data.ıJournal of Applied Statistics, **44**(15), 2659-2671.

[36] Yahya, W.B. & Umar, M.A. (2024). A New Poisson-Exponential-Gamma Distribution for modeling Count Data with Applications. *Quantity & Quality*

[37] Zakerzadeh H & Dolati, A. (2009). Generalized Lindley Distribution. *Journal of Mathematical Extension*, 3, 13–25.