Pakistan Journal of Statistics and Operation Research

A zero truncated discrete distribution: Theory and applications to count data

Tassaddaq Hussain

Department of Mathematics, Mirpur University of Science and Technology (MUST) Mirpur,10250 (AJK),Pakistan, tafkho2000@gmail.com

Abstract

The analysis and modeling of zero truncated count data is of primary interest in many fields such as engineering, public health, sociology, psychology, epidemiology. Therefore, in this article we have proposed a new and simple structure model, named a zero truncated discrete Lindley distribution. The distribution contains some sub models and represents a two-component mixture of a zero truncated geometric distribution and a zero truncated negative binomial distribution with certain parameters. Several properties of the distribution are obtained such as mean residual life function, probability generating function, moments of residual life function, raw moments, estimation of parameters, Shannon and Rényi entropies, a characterization, and stress-strength parameter. Moreover, the collective risk model is discussed by considering the proposed distribution as primary distribution and exponential and Erlang distributions as secondary ones. Test and evaluation statistics as well as three real-life data applications are considered to assess the performance of the distribution among the most frequently zero truncated discrete probability models.

Key Words: Characterization; Zero truncated generalized Poisson distribution; Mean residual life; Estimation.

Introduction

An important statistical property which is used in many fields, such as reliability theory, queueing systems, medicine, industry and many others, is the truncation of probability distributions. The truncation of probability models occurs when a range of possible values of the variables either is ignored or is impossible to be observed. The zero truncation of the models is also a truncation phenomena in which one intends to model count data without zero. For example: length of hospital stay is recorded as a minimum of at least one day, number of times a voter has voted (among the people who have voted) during the general election, number of journal articles published in different disciplines, number of tickets received by teenagers as predicted by school performance, number of children ever born to sample of mothers over 40 years of age, number of eggs and gall-cells (at different dates) were counted in flowers heads in two years, number of occupants in passenger cars, number of stressful events reported by patients, number of houses replying to a postal survey etc. For modeling the above situations usually the zero truncated Poisson (ZTP),\ zero truncated negative binomial (ZTNB), zero truncated generalized negative binomial (ZTGNB), zero truncated Poisson Lindley (ZTPL),Zero Truncated Discrete Shanker (ZTDS), zero-truncated Poisson-Amarendra (ZTPA), zero-truncated Poisson-Garima (ZTPGr) and zero truncated Poisson Gamma (ZTPG) distributions. Applications of the distributions above as follows. Finney and Varley (1955) used the ZTP for the analysis of number of eggs and gall-cells counts in flowers heads, Brass (1959) modeled the number of children ever born to sample of mothers over 40 years of age by ZTNB distribution , Creel and Loomis (1990) applied the ZTP distribution on deer hunting in California, Lindsay (1995) used the ZTP for the analysis of postal survey data, Lee et al. (2003) applied the ZTNB distribution in a regression model for analyzing the over dispersed positive count data of ischemic stroke hospitalizations, Kennedy (2005) used the ZTP, ZTNB and ZTGNB distributions for the analysis

of stroke count data, Phange and Loh (2013) and the reference therein discussed the application of ZTNB in the analysis of abundance of rare species and hospital stay. Groggan and Carson (1991) applied the truncated model to a number of recreational fishing trips taken from a sample of Alaskan fishermen. Moreover, Ghitany et al., (2008), Borah and Saikia (2017), Shanker (2017) and Shanker and Shukla (2017) discussed the ZTPL, ZTDS, ZTPA and ZTPGr distributions respectively and applied to real count data. It is observed in the above discussed zero truncated models that neither have high peakedness nor possess heavy tail. In addition, these models do not work very well in both over and under dispersion phenomena. Moreover, none of the above mentioned models exhibit skewness and kurtosis in the domain of interval $(2, \infty)$ and $(6, \infty)$ respectively, also these models do not possess closed form and algebraically manipulate able survival functions. The mentioned drawbacks above were the motivation to propose a simple structure zero truncated model with two parameters under the name zero truncated discrete Lindley distribution (ZTDL). The proposed ZTDL has closed form structure of its various statistical properties along with heavy right tail than the above competing zero truncated models. Moreover, the peakedness and skewness of the proposed ZTDL model approach to 6 and 2, respectively, which is not observed in the above mentioned models except the ZTPL distribution. Further, the proposed model may be helpful in developing hurdle models by acting as a baseline distribution and in regression modeling. Also, the distribution is over- and under-dispersed. The flexibility of the two parameter ZTDL over the one parameter ZTDL and ZTG (zero truncated geometric) distributions shall be investigated later by using stochastic ordering. Further, the flexibility of the model is investigated by its consonance with some real data sets as revealed through techniques such as empirical distribution function plots, empirical hazard function plots, Anderson-Darling, Kolmogrov-Simnorov (KS), Akaike information criterion (AIC), Bayesian information criterion (BIC) and Akaike information criterion corrected (AICc) statistics. For this purpose three real data sets are analyzed and it is noted that the proposed model is the most suitable distribution among others by exhibiting minimum values for those statistics. In addition, the proposed model only exhibits the increasing failure rate phenomena which is the desired requirement of the life testing experiment. Other motivations of the model in terms of its hazard rate and mean residual life functions are discussed in the coming sections. A Lindley distribution is used extensively to describe the lifetime of a process or a device in a wide variety of fields, including reliability, engineering, biology, and medicine. The first version of Lindley was proposed by Lindley (1958). After that a number of forms of the distribution has been described in academic literature, such as a generalized Lindley by Nadaragha et al. (2011), an extended Lindley (EL) distribution by Bakouch et al. (2012), a power Lindley by Ghitany et al. (2013), a new weighted Lindley distribution by Asgharzadeh et al. (2016). Similarly, Shanker et al. (2013) introduced the two parameter continuous Lindley (TPL) distribution defined by

$$f(x) = \frac{\theta^2}{\theta+\beta}(1 + \beta x)exp(-\theta x), x > 0; \beta \geq 0, \theta > 0. \qquad (1.1)$$

The proposed discrete distribution here is the zero truncated discrete version of the distribution specified by (1.1) and its probability mass function is obtained via the discretization approach by incorporating (1.1) in the following identity

$$P(Y = x) = f(x)\left(\sum_{k=1}^{\infty} f(k)\right)^{-1},$$

where *f(x)* is the probability density function of a continuous random variable. Therefore in the light of equations above we define the zero truncated discrete Lindley distribution denoted by ZTDL(p β) with parameters p and β as

$$P(Y = x) = p_x = \frac{(1-p)^2}{1+\beta-p}(1 + \beta x)p^{x-1}, \qquad x = 1,2,3,.... \qquad (1.2)$$

where $0 < p < 1$ and $\beta \geq 0$. For $\beta = 0$ the distribution reduces to zero truncated geometric distribution and for $\beta = 1$ it becomes one parameter zero truncated dis- crete Lindley distribution, respectively, noting that the case for $\beta = 1$ is not discussed before. Also, when $\beta = 2$, the distribution reduces to the zero truncated negative binomial distribution with parameters $(2, p)$. Moreover, the probability mass function in (1.2) shows that the ZTDL(p β) distribution is a two-component mixture of a zero truncated geometric distribution with parameter p and a zero truncated negative binomial distribution with parameters $(2, p)$ and mixing proportions $\frac{1-p}{1+\beta-p}$

and $\frac{\beta}{1+\beta-p}$ respectively. Figure 1 indicates how the parameters p and β affect the ZTDL probability mass function and shows right skewness and uni-modality. Moreover, it can be noted that, the ZTDL distribution can be obtained by using the expression

$$P(Y = x) = \frac{P(W = x)}{1 - P(W = 0)},$$

where $P(W = x)$ is the two parameter discrete Lindley distribution introduced by Hussain et al. (2016) and given by

$$P(Y = x) = \frac{(1 - p)^2}{1 + p(\beta - 1)}(1 + \beta x)p^x, \qquad x = 0, 1, 2, 3, \dots . \quad 0 < p < 1, \ \beta \geq 0 .$$

The recursive relation between ZTDL probabilities is given by

$$(1 + \beta x)P(Y = x + 1) = p\left(1 + \beta(x + 1)\right)P(Y = x).$$

Since ZTDL is based on two parameters namely shape parameter p and scale parameter β, so Figure 1 under the influence of theses parameters indicates that as $p \to 0$ the mode of the function moves towards left that is probability graph of ZTDL makes a reverse J shape for all values of β. However, as $p \to 1$ the mode of the distribution moves towards right for all values of β. Moreover, the distribution does not achieve perfect symmetry for all values of p and β. In addition to, we have also theoretically observe that $p_x = 0$ as $p \to 1$ and $p_x = x(1 - p)^2 p^{x-1}$ as $\beta \to \infty$. Thus the survival and hazard functions of the ZTDL distribution are

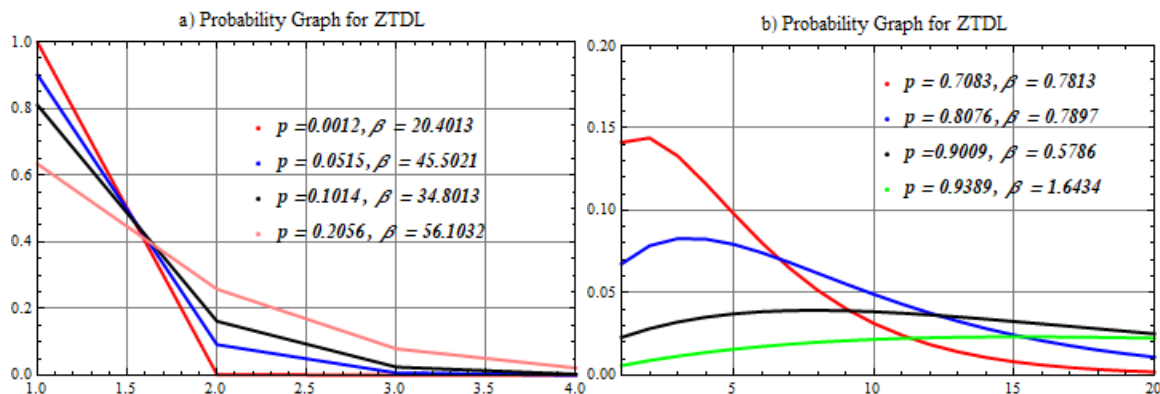Figure 1: Plots of the ZTDL Probasbility mass function for some parameter values.

$$S_x = \frac{\left(1 + \beta x + p(\beta - (1 + \beta x))\right)p^{x-1}}{1 + \beta - p}, \qquad x = 1,2,3, \dots . \quad 0 < p < 1, \ \beta \geq 0 , \qquad (1.3)$$

and

$$h_x = \frac{(1 + \beta x)(1 - p)^2}{(1 + \beta x + p(\beta - (1 + \beta x)))}, \qquad x = 1,2,3, \dots . \quad 0 < p < 1, \ \beta \geq 0 , \qquad (1.4)$$

It is clear from equation (1.4) that $h_x = 1$ as $p \to 0$ and $h_x = 0$ as $p \to 1$. Moreover, $h_x = 1 - p$ as $\beta \to 0$ and $h_x = \frac{x(1-p)^2}{x(1-p)+p}$ as $\beta \to \infty$. However, $h_x = 1 - p$ as $x \to \infty$, therefore, the hazard rate function of the ZTDL distribution is bounded above, which is an important property for the lifetime models. Figure 2 gives some plots of the hazard rate function for some selected values of the parameters. It is observed that the ZTDL distribution exhibits an increasing failure rate for all values x, p anf β. However, if p is considered as successfully working of the product then as p

$\rightarrow 1$ the intensity of failure rate decreases for all values of β. Moreover, if $p \rightarrow 0$ the distribution has a constant failure rate. The mean residual life (MRL) function of the ZTDL distribution is defined as

$$K_x = E(Y - x | Y \geq x) = \frac{\sum_{k=x}^{\infty} S_k}{S_x},$$

and for ZTDL is expressed as

$$K_x = \frac{p \, h_x}{(1-p)^2} + \frac{p \, \beta \, (1+p)}{(1-p)(1+\beta x + p(\beta - (1+\beta x)))}, x = 1,2,3, \dots. 0 < p < 1, \ \beta \geq 0 \,.$$

We may interpret the second term of equation (1.5) as a correction term for the MRL of the ZTDL distribution than the zero truncated geometric that for which $\beta = 0$. Also, it is noted that the MRL of ZTDL distribution is greater (equal) than (to) the MRL of zero truncated geometric and one parameter zero truncated discrete Lindley distribution based on value of the parameter β. Further discussion to MRL will be given later.

Rest of the article is organized as follows. Section 2 gives several properties of the ZTDL distribution such as reliability function, mean residual life function, probability generating function, moments, moments of residual life function, estimation of parameters via maximum likelihood method and associated Fisher information matrix, Shannon and Rènyi entropies. A characterization, convolution and stress-strength parameter are investigated in Section 3, besides the collective risk model where the proposed distribution is considered as primary distribution and exponential and Erlang distributions as secondary ones. Test and evaluation statistics as well as three real data application for the ZTDL distribution are discussed in Section 4. Finally, Section 5 deals with conclusions.

2.   Properties of the ZTDL distribution

In this section, some statistical and reliability properties of the ZTDL distribution are investigated by the following theorems, propositions and corollaries.

Log-concaveity and its implications: A probability distribution is said to be log-concave if its probability mass function (pmf) satisfies the inequality $p_x^2 \geq p_{x-1} p_{x+1}, \forall \ x$. The log-concavity of a distribution has important implications on the characteristic of its reliability function, failure rate function, tail probabilities and moments. Therefore, the log-concavity produces an increasing failure rate and a monotonically decreasing mean residual life time function (see Chakraborty and Ong, 2016). The next proposition shows the log-concavity of the ZTDL distribution. **Proposition 2.1:** If $Y \sim ZTDL(p, \beta)$ then the probability mass function (pmf), of the random variable Y is log-concave for all choices of β and independent of p.

Proof: In order to show that the ZTDL distribution defined by (1.2) is log-concave, it is sufficient to show that $p_x^2 \geq p_{x-1} p_{x+1}, for \ \beta \geq$ and $\ x = 2,3,4, \dots$ which follows by noting that $(1 + \beta x)^2 \geq (1 + \beta x)^2 - \beta^2$.

Corollary 2.1: The following results are the direct consequence of log-concavity (see Chakraborty and Ong, 2016):

i) ZTDL has an increasing failure rate function. ii) Convolution of ZTDL with any other discrete distribution will also result in a log concave distribution. iii) For any integer $\frac{p_{x+k}}{p_x} \geq \frac{p_{y+k}}{p_y}$ for $\ x < y$ that is $\frac{p_{x+k}}{p_x}$ is non increasing in x. iv) ZTDL has at most an exponential tail, $\lim_{x \to \infty} e^{bx} P(Y = x) = 0 \Rightarrow P(Y = x) = Oe^{-bx}$ for some $\ b > 0$ as $x \to \infty$. v) ZTDL has a monotonically decreasing mean residual life time function.

Now, recall that the log-concave probability mass functions are strongly unimodal, see Nekoukhou et al. (2012), and have an increasing failure rate. Therefore, we have the following corollary

Corollary 2.2: If Y~ZTDL(p, β) then the mode of the random variable Y is located at m and m satisfies, $\frac{p\,\beta-(1-p)}{(1-p)\beta} \leq m \leq \frac{\beta-(1-p)}{(1-p)\beta}$.

Proof: According to Abouammoh and Mashhour (1981) and Nekoukhou et al., (2012) the unimodality condition is defined as $p_x \geq p_{x-1}$ for $x \leq m$ and $p_{x+1} \leq p_x$ for $x \geq m$, so for ZTDL $p_x \geq p_{x-1} \Rightarrow \frac{\beta\,p-(1-p)}{(1-p)\beta} \leq x$ and $p_{x+1} \leq p_x \Rightarrow \frac{\beta-(1-p)}{(1-p)\beta}$. This completes the proof.

The following proposition confirms again the increasing failure rate of the ZTDL distribution.

Proposition 2.2: If Y~ZTDL(p, β) then hazard function of ZTDL i.e. $0 \leq h_x \leq 1$ is an increasing failure rate (IFR).

Proof: For a discrete random variable Y with pmf $p_x = P(Y = x)$ and failure rate function $h_x = \frac{p_x}{S_x}$ and let $\eta_x = \frac{p_x - p_{x+1}}{p_x}$ then $\Delta\eta_x = \eta_{x+1} - \eta_x = \frac{p_{x+1}}{p_x} - \frac{p_{x+2}}{p_{x+1}}$. Hence, we get the three results (Gupta et al., 1997)

    i. $\Delta\eta_x > 0$, h_x is an increasing failure rate (IFR)

    ii. $\Delta\eta_x < 0$, h_x is an decreasing failure rate (DFR)

    iii. $\Delta\eta_x = 0$, h_x is a constant failure rate (CFR)

Therefore, for the ZTDL distribution we find that $\Delta\eta_x = \frac{p\beta^2}{(1+\beta x)(1+\beta(x+1))} > 0$, which implies the hazard function of ZTDL is an increasing failure rate (IFR).

For infinitely divisibility, self-decomposability and stability property of ZTDL(p,β), it is noted that according to Steutel and Harn (1979) and Nekoukhou et al., (2012), a necessary conditions for infinite divisibility of discrete distribution p_x is that $p_0 > 0 > 0$, which in the ZTDL distribution case is not fulfilled as $p_0 = 0$. Moreover, classes of self-decomposable and stable distributions are subclasses of infinitely divisible distributions, so non infinitely divisible distribution is neither self-decomposable nor stable, see Nekoukhou et al., (2012).

The recognition of any discrete probability distribution is usually based on its probability generating function (pgf). The following theorem gives the pgf of ZTDL distribution.

Proposition 2.3: If Y~ZTDL(p, β) then the probability generating function (pmf), of the random variable Y is expressed as

$$G_Y(t) = \frac{t(1-p)^2(1-pt+\beta)}{(1-pt)^2(1-p+\beta)},$$

where $\beta \geq 0, 0 < p < 1$ for $0 < pt < 1$.

Proof: The proof can be obtained by routine calculations.

Corollary 2.3: If t is replaced by e^t, we get the moment generating function (mgf) of the ZTDL distribution as

$$M_Y(t) = \frac{e^t(1-p)^2(1-pe^t+\beta)}{(1-pe^t)^2(1-p+\beta)}, \tag{2.1}$$

where $0 < pe^t < 1$.

Corollary 2.4: The rth. derivatives of equation (2.1) with respect to t to $t = 0$. yield the rth moment

$$\mu_r' = \frac{(1-p)^2}{p(1-p+\beta)}\left(Li_{-r}(p) + \beta Li_{-r-1}(p)\right) \qquad (2.2)$$

Hence, from above corollary we have the first four moments about origin as

$$\mu_1' = \frac{(1-p)^2}{p(1-p+\beta)}\left(Li_{-1}(p) + \beta Li_{-2}(p)\right), \quad \mu_2' = = \frac{(1-p)^2}{p(1-p+\beta)}\left(Li_{-2}(p) + \beta Li_{-3}(p)\right),$$

$$\mu_{31}' = \frac{(1-p)^2}{p(1-p+\beta)}\left(Li_{-3}(p) + \beta Li_{-4}(p)\right), \quad \mu_4' = = \frac{(1-p)^2}{p(1-p+\beta)}\left(Li_{-4}(p) + \beta Li_{-5}(p)\right),$$

where $Li_{-r}(p) = \sum_{k=1}^{\infty} k^r p^k$ (see Gradshteyn, and Ryzhik, 2008) is the poly logarithm function. Therefore, by using these expressions, the mean and variance of the distribution are, respectively, given by

$$\text{Mean} = \frac{(1+\beta-p+p\beta)}{(1-p)(1-p+\beta)}, \text{Variance} = \frac{p\left(2\beta+(1-p)^2+\beta(1-p)(3+p)\right)}{(1-p)^2(1-p+\beta)^2}.$$

In statistics, the phenomena of over- and under-dispersion relative to the Poisson distribution is generally observed in count data. There are various causes of such phenomenon, like heterogeneity and aggregation for over-dispersion and repulsion for under-dispersion although less frequent, see Kokonendji and Mizre (2005). As the index of dispersion (ID) is a measure of such phenomena and defined as variance to mean ratio so for the ZTDL distribution it can be expressed as

$$\text{ID} = \frac{1}{(1-p)} + \frac{\beta+1}{1+\beta-p} - \frac{\beta+1}{1+\beta-p+p\beta} - 1.$$

From the above expression it is obvious that as $p \rightarrow 0$ the ID becomes equal to zero, i.e., for smaller values of p the distribution is under dispersed. Similarly as $p \rightarrow 1$ the ID approaches to infinity. Moreover, for $\beta \rightarrow 0$ the ID becomes $\frac{p}{1-p}$ and for $\beta \rightarrow \infty$ the ID becomes $\frac{1}{1-p} - \frac{1}{2}$. So the ID depends solely upon the value of p.
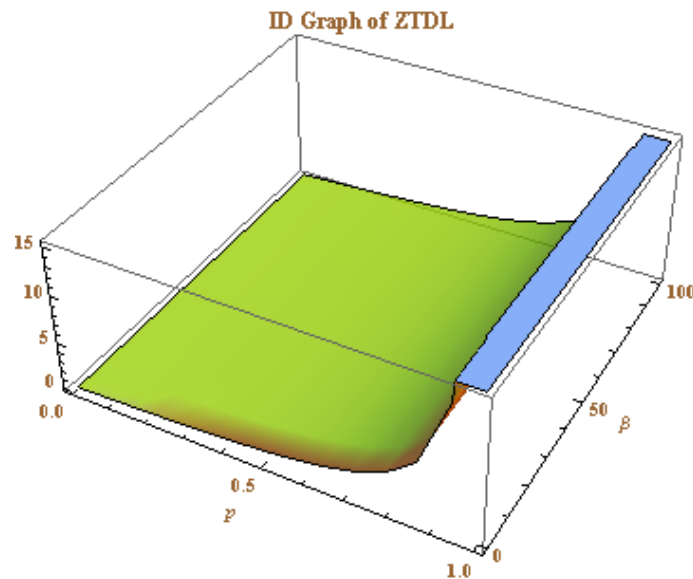

Figure 3: Index of Dispersion plots of ZTDL

Figure 3 displays plots of the ID for different values of the distribution parameters. It is noted that, for $\beta < 8$ and $p < 0.4145$, the distribution is under-dispersed and as $p \to 1$ even at smaller values of β the distribution is over dispersed, i.e., $\beta \to 0$ and $p \to 1$, the ID > 1.Another important statistical properties is the skewness which usually refer to the departure from symmetry which is the ratio of square of the third mean moments to the cube of the second moment about mean, for ZTDL distribution it is expressed as

$$
\text{Skewness} = \frac{1}{p\big(2\beta^2 + (1-p)^2 + \beta(1-p)(3+p)\big)^3} \Big\{ \big(2\beta^3(1+p) + (1-p)^3 + \beta(1-p)^2\big(4 + p(7+p)\big) + \beta^2\big(5 + p(3 + (p-9)p))\big)\big)^2 \Big\}.
$$

While, kurtosis measures the degree of peakedness or flatness of a unimodal frequency curve and defined as the ratio of fourth mean moment to the square of the second moment about mean, for the ZTDL distribution it is expressed as

$$
\text{Kurtosis} = \frac{1}{p\big(2\beta^2 + (1-p)^2 + \beta(1-p)(3+p)\big)^2} \{(1-p)^4(1 + p(7+p)) + 2\beta^4(1 + p(10+p))
$$
$$
+ \beta^3(1-p)(7+p)(1 + p(10+p)) + \beta^2(1-p)^2 (9 + p(89 + (35-p)p))
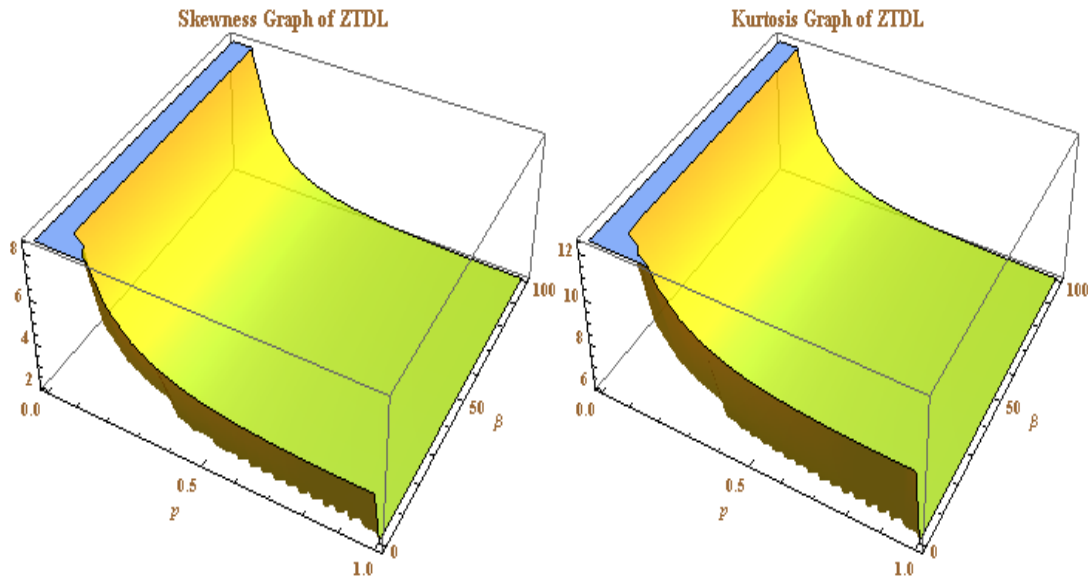$$
$$
+ \beta(-1+p)^3(5 + p(45 + p(21+p)))\}
$$

Figure 4: Skewness and Kurtosis plots for ZTDL

Figure 4 gives plots of the skewness and kurtosis for different values of the distribution parameters. It is observed that the distribution is positively skewed and leptokurtic in nature. Skewness becomes equal to 2 as p and β increases and peakedness of the distribution gives higher values at smaller p and β, however as p and β increases it settled down to 6. From the discussion above it is obvious that the ordinary moments are generally helpful in estimating the unknown parameters and characterization of the shape of the distributions.

2.1   Residual life function with some properties

Residual life random variable is used much in risk analysis and so we investigate some of its properties, like survival function, mean, variance and rth moment for the ZTDL distribution. The residual life is defined by the conditional

random variable, $R(t) = X - t|X > t, t \geq 0$, and described as the the period from time t until the time of failure. The survival function of the residual lifetime R(t), for the ZTDL distribution is

$$S_{R(t)}(x) = \frac{S_{x+t}}{S_t} = \frac{p^x(1 + \beta(x + t) + p(\beta - (1 + (x + t))))}{1 + \beta t + p(\beta - (1 + \beta t))}, x > 0,$$

and the corresponding probability mass function can be obtained by $P(R(t) = x) = S_{R(t)}(x) - S_{R(t)}(x + 1)$. The mean residual life is defined as

$$K_t = E(R(t)) = \frac{1}{S_t}\sum_{x=t}^{\infty} xp_x - t,$$

$$K_t = \frac{p(1 + \beta - p + p\beta) + (1-p)(1 + p(2\beta - 1))t + (1-p)^2t^2}{(1-p)(1 + \beta t + p(\beta - (1 + \beta t)))} - t,$$

where p_x and S_t are given in (1.2) and (1.3), respectively. Based on the behaviour of the failure function hx of the ZTDL distribution, which is increasing, we conclude that K_t is decreasing. The variance residual life has considerable interest in the recent years and for the ZTDL distribution it is given as

$$V_t = Var(R(t)) = -t^2 + \frac{1}{S_t}\sum_{x=t}^{\infty} x^2p_x - 2t K_t - K_t{}^2,$$

$$V_t = \frac{1}{(1-p)^2(1 + \beta t + p(\beta - (1 + \beta t)))}\{p(1 + \beta + 4\beta p + p^2(\beta - 1)) + \beta(1-p)^3t^3$$
$$+ p(1-p)(2(1-p) + 3\beta(1 + p))t + (1-p)^2(1 - p + 3\beta p)t2\}$$

$$-\frac{(p(1 + \beta - p + p\beta) + (1-p)(1 + p(2\beta - 1))t + (1-p)^2t^2)^2}{(1-p)2(1 + \beta t + p(\beta - (1 + \beta t)))^2}.$$

Moreover, the rth moment of the residual life of ZTDL distribution is

$$M_t = E[(Y - t)^r|Y > t] = \frac{1}{S_t}\sum_{x=t}^{\infty}(x - t)^r p_x = \frac{(Li_{-r}(p)(1 + \beta t) + \beta Li_{-r-1}(p))(1 - p)^2}{(1 + \beta t + p(\beta - (1 + \beta t)))}.$$

2.2   Entropies

In statistical mechanics, the entropy may be interpreted as measure of disorder in the distribution. The most frequently discussed entropies are Shannon and Renyi entropy. The derived expressions for these entropies when Y is a discrete ZTDL random variable are given below. The Shannon entropy is defined by $E[-\log(P(Y = x))]$ and expressed by using the Taylor series expansion for $ln(1 + z)$ as

$$H(x) = ln(1 + \beta - p) - 2ln(1 - p) - lnp(\mu - 1) - \sum_{m=1}^{\infty}\frac{(-1)^{m+1}\beta^m\mu'_m}{m},$$

where μ'_m and μ are the mth order moment about origin and mean of the distribution. The Renyi entropy is defined as

$$I_R(\lambda) = \frac{1}{1 - \lambda}ln\left(\sum_{x=1}^{\infty}(P(Y = x))^2\right),$$

hence

$$\mathrm{I}_R(\lambda) = \frac{\lambda}{1-\lambda}\{2ln(1-p) - ln(1 + \beta - p)\} + \frac{p^{-\lambda}}{1-\lambda} ln\{\textstyle\sum_{m=0}^{\infty} \binom{\lambda}{m}\beta^m Li_{-m}(p^\lambda)\},$$

where $\lambda > 0$ and $\lambda \neq 1$.

2.3   Parameter Estimation with Inference

Let $Y_1, Y_2, \ldots, Y_n$ be a random sample drawn identically independently from the ZTDL distribution with observed values $x_1, x_2, \ldots, x_n$, then their joint probability function as a log-likelihood function is

$$ln\big(L(p;\beta)\big) = 2nln(1 - p) - nln\big((1 + \beta - p)\big) + \sum_{i=1}^{n} x_i ln\, p - n\, l\, n\, p + \sum_{i=1}^{n} l\, n(1 + \beta\, x_i). \qquad (2.3)$$

Now on partially differentiating both sides of equation (2.3) with respect to p and β and equating them to zero we get MLEs of p and β, respectively, as

$$\frac{2p}{1-p} - \frac{p}{1-p+\beta} = \bar{x} - 1,$$

and

$$\sum_{i=1}^{n} \frac{x_i}{1 + \beta\, x_i} = \frac{n}{1-p+\beta}.$$

The MLEs are computed using computational package such as Mathematica software.

Proposition 2.4: The MLEs $\hat{p}$ and $\hat{\beta}$ of the ZTDL distribution has a bivariate normal distribution with mean (p, β) and variance-covariance matrix $(\mathrm{I}\big((p, \beta)\big))^{-1}$, where

$$\mathrm{I}\left((p,\beta)\big|p = \hat{p}, \beta = \hat{\beta}\right) = \begin{bmatrix} \mathbf{E}\left(\dfrac{-\partial^2 lnL(p,\beta)}{\partial p^2}\right) & \mathbf{E}\left(\dfrac{-\partial^2 lnL(p,\beta)}{\partial p\,\partial\beta}\right) \\[2ex] \mathbf{E}\left(\dfrac{-\partial^2 lnL(p,\beta)}{\partial p\,\partial\beta}\right) & \mathbf{E}\left(\dfrac{-\partial^2 lnL(p,\beta)}{\partial\beta^2}\right) \end{bmatrix},$$

and the expected information matrix $\mathrm{I}(\hat{p}, \hat{\beta})$ is given by

$$\mathrm{I}\left((p,\beta)\big|p = \hat{p}, \beta = \hat{\beta}\right) =$$

$$\begin{bmatrix} \dfrac{n(2\hat{\beta}^2 + (1 - \hat{p})^2 + \hat{}\,\beta(1 - \hat{p})(3 + \hat{p}))}{\big(1 + \hat{\beta} - \hat{p}\big)^2(1 - \hat{p})^2\,\hat{p}} & \dfrac{n}{\big(1 + \hat{\beta} - \hat{p}\big)^2} \\[4ex] \dfrac{n}{\big(1 + \hat{\beta} - \hat{p}\big)^2} & \dfrac{n(1 - \hat{p})^2\,{}_3F_2\left(2,2,1 + \frac{1}{\hat{\beta}}; 2 + \frac{1}{\hat{\beta}}; \hat{p}\right)}{\big(1 + \hat{\beta} - \hat{p}\big)(1 + \hat{\beta})} - \dfrac{n}{\big(1 + \hat{\beta} - \hat{p}\big)^2} \end{bmatrix}$$

Proof of Proposition 2.4 is investigated by the next corollaries.

Corollary 2.5:

$$\mathbf{E}\left(\frac{-\partial^2 lnL(p,\beta)}{\partial p^2}\right) = \frac{n(2\beta^2 + (1-p)^2 + \beta(1-p)(3+p))}{(1+\beta-p)^2(1-p)^2 p}. \tag{2.4}$$

Proof: By twice differentiation of equation (2.3) with respect to p and then applying expectation on both sides, we get

$$-\mathbf{E}\left(\frac{\partial^2 lnL(p,\beta)}{\partial p^2}\right) = \frac{2n}{(1-p)^2} - \frac{n}{(1+\beta-p)^2} + \sum_{i=1}^{n}\frac{E(X_i-1)}{p^2},$$

using the $\mu = E(X)$ of ZTDL we get equation (2.4).

Corollary 2.6:

$$\mathbf{E}\left(\frac{-\partial^2 lnL(p,\beta)}{\partial \beta^2}\right) = \frac{(1-p)^2\,_3F_2\left(2,2,1+\frac{1}{\beta};2+\frac{1}{\beta};p\right)}{(1+\beta)(1+\beta-p)} - \frac{n}{(1+\beta-p)^2}. \tag{2.5}$$

Proof: By twice differentiation of equation (2.3) with respect to β and then applying expectation on both sides, we get

$$-\mathbf{E}\left(\frac{\partial^2 lnL(p,\beta)}{\partial \beta^2}\right) = -\frac{n}{(1+\beta-p)^2} + \sum_{i=1}^{n}\mathrm{E}\left(\frac{X_i^2}{(1+\beta X_i)^2}\right),$$

where

$$\mathrm{E}\left(\frac{X^2}{(1+\beta X)^2}\right) = \sum_{x=1}^{\infty}\frac{x^2}{(1+\beta x)^2}P(Y=x) = \frac{(1-p)^2\,_3F_2\left(2,2,1+\frac{1}{\beta};2+\frac{1}{\beta};p\right)}{(1+\beta)(1+\beta-p)}$$

and $_nF_m(a_1,a_2,\ldots,a_n;b_1,b_2,\ldots,b_n;z) = \sum_{k=0}^{\infty}\frac{(a_1)_k(a_2)_k\cdots(a_n)_k}{(b_1)_k(b_2)_k\cdots(b_m)_k}\frac{z^k}{k!}$,

on using the above expression and then simplifying it, we get (2.5).

Corollary 2.7:

$$\mathbf{E}\left(\frac{-\partial^2 lnL(p,\beta)}{\partial p\,\partial \beta}\right) = -\frac{n}{(1+\beta-p)^2}.$$

Proof: The proof is simple.

Corollary 2.8 : The minimum variance bound estimator of the ZTDL (p,β) when β is known is $\bar{x}$ with variance $\frac{p(2\beta^2 + 3\beta + 1 - 2p\beta - 2p + p^2 - \beta p^2)}{n(1-p)^2(1+\beta-p)^2}$.

3.   Characterization, convolution and collective risk model

3.1   Characterization

Characterization of probability function helps the researcher in resolving the problem of model identification.

Proposition 3.1: If X_1, X_2 and Y_n are independent random variables such that $X_1 \sim$ Bernoulli$\left(\frac{p}{1-p+\beta}\right)$, $X_2 \sim$ NB$(p,2)$ and $X_3 \sim$ a degenerate distribution, i.e., P(X = 1) = 1, then Y = $X_1 + X_2 + X_3$ has the ZTDL(p,β) distribution.

Proof: As $X_1 \sim$ Bernoulli$\left(\frac{p}{1-p+\beta}\right)$ then its moment generating function(mgf) is expressed as

$M_{X_1}(t) = \frac{1-pe^t+\beta}{1-p+\beta}$. Similarly X_2 and X_3 with their respective mgf are $M_{X_2}(t) = \frac{(1-p)^2}{(1-pe^t)^2}$ and $M_{X_3}(t) = e^t$, respectively. Since $Y = X_1 + X_2 + X_3$ then, the mgf of Y is given

$$M_Y(t) = \frac{(1 - pe^t + \beta)(1 - p)^2}{(1 - p + \beta)(1 - pe^t)^2} e^t,$$

$$M_Y(t) = M_{X_1}(t) M_{X_2}(t) M_{X_3}(t).$$

Proposition 3.2: If $Y_i \sim$ ZTDL($p, \ \beta$) for $i = 1,2,3,...n$, are identically independently distributed random variables, then $Z = Y_1 + Y_2 + Y_3 + \cdots + Y_n$ also has the ZTDL($n, p, \ \beta$) distribution.

Proof: As $Y_i \sim$ ZTDL($p, \ \beta$) for $i = 1,2,3, ..., n$, then the moment generating function of $Z = Y_1 + Y_2 + Y_3 + \cdots + Y_n$ is expressed as

$$M_Z(t) = \prod_{i=1}^{n} M_{Y_i}(t) = \left(\frac{1 + \beta}{1 + \beta - p} - \frac{p \, e^t}{1 + \beta - p} \right)^n \frac{(1 - p)^{2n}}{(1 - pe^t)^{2n}} \, e^{nt},$$

$$M_Z(t) = \left(1 - \frac{p \, (e^t + 1)}{1 + \beta - p} \right)^n \frac{(1 - p)^{2n}}{(1 - pe^t)^{2n}} \, e^{nt}.$$

It can also be rewritten as

$$M_Z(t) = M_{X_B}(t) M_{X_{NB}}(t) M_{X_{Degenerate}}(t),$$

$X_B \sim$ Binomial$\left(\frac{-p}{1-p+\beta}, n \right)$, $X_{NB} \sim$ NB($p, 2n$) and $X_{Degenerate} \sim$ a degenerate distribution

i.e., P($X = n$) = 1.

3.2   Convolution

Convolution is used in the mathematics of many fields, such as probability and statistics. In engineering, convolution is used to describe the relationship between three variables of interest: the input signal, the impulse response, and the output signal. The following theorem deals with the convolution of the ZTDL distribution.

Theorem 3.1: If $Y_i \sim$ ZTDL$_1$($p, \ \beta$) for $i = 1, 2$ are two i.i.d random variables, then probability function of sum of two random variables $Y_1 + Y_2 = Z$ is

$$P(Z = z) = \frac{(1 - p)^4 z p^{z-2}}{(1 + \beta - p)^2} \left\{ 1 + \beta \left(z + \beta \left(\frac{z^2 - 1}{6} \right) \right) \right\}, \tag{3.1}$$

where $\beta \geq 0, 0 < p < 1, z = 2,3,....$

Proof: By definition, probability function of sum of two discrete i.i.d random variables Y_1 and Y_2 can be expressed as $P(Z = z) = \sum_{y_1=1}^{z} P(Y_1 = y_1) P(Y_2 = z - y_1)$ where

$$P(Y_i = y_i) = \frac{(1 + \beta \, y_i)(1 - p)^2}{1 + \beta - p},$$

for $i = 1,2$ and $\beta \geq 0, 0 < p < 1, \ y_i = 1,2,3,...$ so we get

$$P(Z = z) = \frac{(1 - p)^4 z p^{z-2}}{(1 + \beta - p)^2} \sum_{y_1=1}^{z} \{(1 - (\beta \, y_1)^2) + \beta z(1 + \beta \, y_1)\},$$

which on simplification yields equation (3.1). This completes the proof. In the next corollaries, the abbreviations $ZTDL_1(.)$, $G_1(.)$, $G_0(.)$ and $DL_0(.)$ denote the ZTDL zero truncated geometric, geometric and one parameter zero truncated discrete Lindley distributions, respectively.

Corollary 3.1: If $Y_1 \sim ZTDL_1(p, \beta)$ and $Y_2 \sim G_1(1-p)$ are two i.i.d random variables, then the probability function of sum of the two random variables, i.e., $Y_1 + Y_2 = Z$ is

$$P(Z = z) = \frac{(1-p)^2 z p^{z-2}}{1 + \beta - p}\left\{1 + \frac{\beta(1+z)}{2}\right\},$$

where $\beta \geq 0, 0 < p < 1, z = 2,3,\ldots.$

Corollary 3.2: If $Y_1 \sim ZTDL_1(p, \beta)$ and $Y_2 \sim G_0(1-p)$ are two i.i.d random variables, then the probability function of sum of the two random variables, i.e., $Y_1 + Y_2 = Z$ is

$$P(Z = z) = \frac{(1-p)^3 z p^{z-1}}{1 + \beta - p}\left\{1 + \frac{\beta(1+z)}{2}\right\},$$

where $\beta \geq 0, 0 < p < 1, z = 2,3,\ldots.$

Corollary 3.3: If $Y_1 \sim ZTDL_1(p, \beta)$ and $Y_2 \sim DL_0(p, \beta)$ are two i.i.d random variables, then the probability function of sum of the two random variables, i.e., $Y_1 + Y_2 = Z$ is

$$P(Z = z) = \frac{(1-p)^4 z p^{z-1}}{(1+\beta-p)(1+p(\beta-1))}\left\{1 + \beta z + \frac{\beta^2(z^2-1)}{6}\right\},$$

where $\beta \geq 0, 0 < p < 1, z = 1,2,3,\ldots.$

3.3   Collective risk model

Generally, the random sum models are widely used in risk theory, queueing systems, reliability theory, economics, communications, and medicine. In insurance contexts, the compound distribution of random sum variables arises naturally as follows (see Bakouch and Severini (2009) and Bhati et al., (2015)). Let N be number of claims in a certain period which is a random variable and Xi be claim severity random variable, which is independent of N, is the size of ith claim. Therefore, aggregate loss is defined as $S = \sum_{i=1}^{N} X_i$. It is well known that the pdf of S is given as $f_S(x) = \sum_{n=1}^{\infty} p_n f^{n*}(x)$, where p_n denotes the probability of n claims (primary distribution) and $f^{n*}(x)$ is the nth fold convolution of $f(x)$ the claim amount (secondary distribution). Moreover, the distribution of the random variable S is known as a compound distribution of the random variables N and X_i. Now, we shall consider two such situations: In the first situation, the primary distribution is as defined in equation (1.2) and claim severity distribution as exponential distribution with parameter λ. In the second situation, we consider the Erlang distribution with parameters r and λ as the secondary distribution which usually arises in insurance settings when the individual claim amount also follow the gamma distribution.

Theorem 3.2: Suppose that $ZTDL(p, \beta)$ acts as primary distribution and an exponential distribution with parameter λ acts as secondary a distribution, then the pdf of aggregate loss random variable $S = \sum_{i=1}^{N} X_i$ where $X \sim Gamma(n, \lambda)$ is given by

$$f_S(x) = \frac{(1-p)^2}{1+\beta-p}\left\{\lambda\, e^{-\lambda x\,(1-p)}\left(1 + \beta(1 + \lambda\, p\, x)\right)\right\}.\ for\ \ x > 0$$

Proof:

If the claim severity follows an exponential distribution with parameter $\lambda > 0$, since the nth fold convolution of exponential distribution is gamma distribution with parameter n and λ, the nth fold convolution is given by

$$f^{n*}(x) = \frac{\lambda^n}{(n-1)!} x^{n-1} e^{-\lambda x}, \quad n \in \aleph.$$

Hence, the pdf of the random variable S is given by

$$f_S(x) = \frac{(1-p)^2 e^{-\lambda x}}{px(1+\beta-p)} \sum_{n=1}^{\infty} \frac{(1+\beta n)(\lambda px)^n}{(n-1)!},$$

which after simplification yield

$$f_S(x) = \frac{(1-p)^2}{(1+\beta-p)} \{\lambda e^{-\lambda x(1-p)}(1+\beta(1+\lambda px))\},$$

which is mixture of $X_1 \sim \text{Exp}(\lambda(1-p))$ and $X_2 \sim \text{TPL}(p, \lambda(1-p))$ with weights $w_1 = \frac{1-p}{1+\beta-p}$ and $w_2 = \frac{\beta}{1+\beta-p}$. This completes the proof.

In the actuarial literature the mean of the aggregate loss random variable S can be obtained as $E(S) = E(X)E(N)$. Hence, if $X \sim \text{Exp}(\lambda)$ then

$$E(S) = \frac{1+\beta-p+p\beta}{(1-p)(1+\beta-p)\lambda}.$$

Theorem 3.3: Suppose that ZTDL(p, β) acts as primary distribution and an Erlang distribution with parameter $(2, \lambda)$ acts as secondary a distribution, then the pdf of aggregate loss random variable $S = \sum_{i=1}^{N} X_i$ is given by

$$f_S(x) = \frac{(1-p)^2 \lambda e^{-\lambda x}}{2(1+\beta-p)\sqrt{p}} \{(\beta+2)sinh(\lambda x\sqrt{p}) + \beta\lambda x\sqrt{p}cosh(\lambda x\sqrt{p}))\},$$

for $x > 0, 0 < p < 1, \lambda > 0$ and $\beta \geq 0$.

Proof:

By assuming that the claim severity follows an Erlang distribution with parameter $2, \lambda > 0$, the nth fold convolution is given by

$$f^{n*}(x) = \frac{\lambda^{2n}}{(2n-1)!} x^{2n-1} e^{-\lambda x}, \quad n \in \aleph.$$

hence the pdf of the random variable S is given by

$$f_S(x) = \frac{\lambda\sqrt{p}\,(1-p)^2 e^{-\lambda x}}{p(1+\beta-p)} \sum_{n=1}^{\infty} \frac{(1+\beta n)(\lambda\sqrt{p}x)^{2n-1}}{(2n-1)!},$$

Now using the series $\sum_{n=0}^{\infty} \frac{y^{2n-1}}{(2n+1)!} = sinh(y)$ and $\sum_{n=0}^{\infty} \frac{y^{2n}}{(2n!)} = cosh(y)$ (see Grad

shteyn and Ryzhik, 2008, pp.42) which after simplification yield

$$f_S(x) = \frac{(1-p)^2 \lambda e^{-\lambda x}}{2(1+\beta-p)\sqrt{p}} \{(\beta+2)sinh(\lambda x\sqrt{p}) + \beta\lambda x\sqrt{p}cosh(\lambda x\sqrt{p}))\}, x > 0.$$

It is also observed that the mean of the aggregate loss random variable S can be obtained as $E(S) = E(X)E(N)$. Hence, if $X \sim \text{Erlang}(2, \lambda)$ then

$$E(S) = \frac{2(1 + \beta - p + p\beta)}{(1 - p)(1 + \beta - p)\lambda}.$$

3.4 Stress-strength parameter

In the context of reliability, the stress-strength model describes the life of a component which has a random strength Y_1 that is subjected to a random stress Y_2. The component fails at the instant when the stress applied to it exceeds the strength, and the component will function satisfactorily whenever $Y_1 > Y_2$. Therefore, $\mathbf{R} = P(Y_1 > Y_2)$ is a measure of component reliability. It has many applications especially in engineering concepts such as structures, deterioration of rocket motors, fatigue failure of aircraft structures, and the ageing of concrete pressure vessels.

Proposition 3.3: Suppose $Y_1 \sim ZTDL(p_1, \beta_1)$ and $Y_2 \sim ZTDL(p_2, \beta_2)$ are two independent discrete random variables, then stress-strength parameter

$$\mathbf{R} = \frac{1}{(1 + \beta_2 - p_2)(1 + \beta_1 - p_1)(1 - p_1 p_2)^3} \times$$

$$\{((1 - p_2)^2((1 - p_1 p_2)(1 + \beta_1 - p_1 - p_1(1 + (\beta_1 - 1) p_1) p_2) + \beta_2(1 + \beta_1 - p_1 + p_1(\beta_1 - 1 + p_1 - 2\beta_1 p_1) p_2)))\}.$$

Proof: The proof obtained by the definition of $\mathbf{R}$ and simplification to

$$\mathbf{R} = \frac{(1 - p)^2}{(1 + \beta_2 - p_2)(1 + \beta_1 - p_1) p_1 p_2} \sum_{x=1}^{\infty} ((p_1 p_2)^x (1 + \beta_1 x)(1 + \beta_2 x)(1 - p_1) + \beta_1 p_1).$$
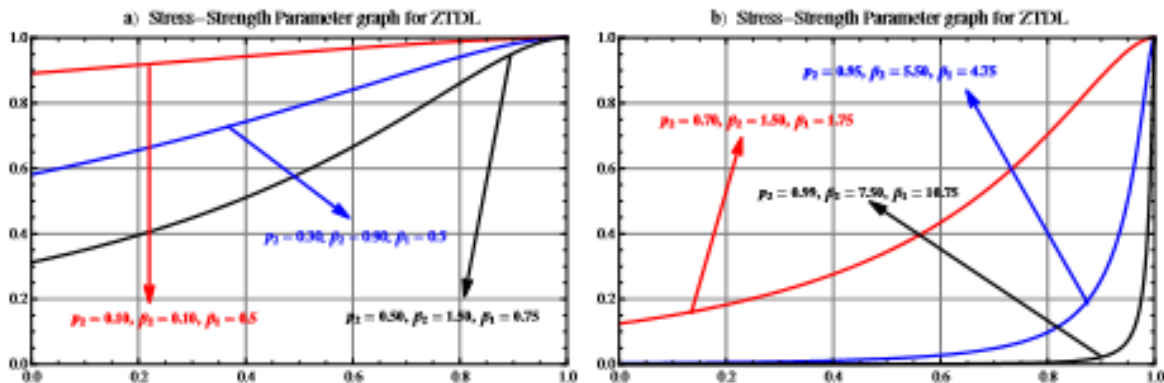

Figure 5: Stress Strength Parameter Graph for the indicated values of p and β

The plots in Figure 5 depict that as the stress parameter p_2 approaches to 1, the probability of the system approaches to zero for all values of β_1 and β_2.

4. Test and Evaluation Statistics with Data Applications

4.1 Test and evaluation statistics

In order to check the competence of the proposed model we have used the most recommended zero truncated models, such as Zero Truncated Poisson (ZTP), Zero Truncated Generalized Poisson (ZTGP), Zero Truncated Negative Binomial (ZTNB) and Zero Truncated Poisson Gamma (ZTPG), and probability functions of these models are given below

Distribution Formula	Domain
$P_{ZTGP}(Y = x) = \dfrac{\beta(\beta + p\,x)e^{-\beta - p\,x}}{(1 - e^{-\beta})x!}$	$x \in \aleph, p > -\dfrac{1}{\beta}$ and $\beta > 0$
$P_{ZTNB}(Y = x) = \begin{pmatrix} \beta + x - 1 \\ \beta - 1 \end{pmatrix}\dfrac{(1 - p)^x p^{\beta}}{(1 - p^{\beta})}$	$x \in \aleph, 0 < p < 1$ and $\beta > 0$
$P_{ZTPG}(Y = x) = \begin{pmatrix} \beta + x - 1 \\ \beta - 1 \end{pmatrix}\dfrac{p^x}{((1 + p)^{\beta} - 1)(1 + p)^x}$	$x \in \aleph, p > 0$ and $\beta > 0$
$P_{ZTP}(Y = x) = \dfrac{\beta^x}{(e^{\beta} - 1)x!}$	$x \in \aleph$ and $\beta > 0$

For this purpose we have used seven goodness of fit statistics. The computation of these statistics is based on MLEs which are computed by using Mathematica 7.0 computational package. These statistics include Log-Likelihood (l), Akaike information criterion (AIC), Bayesian information criterion (BIC) and Akaike information certerion corrected (AICc), the Anderson-Darling (A_0^*) Statistics and Kolmogrov Simnorov (KS) statistics with $p-value$ The formulae of the above mentioned statistics are

$$AIC = -2\,l(\hat{\theta}) + 2\,q, BIC = -2\,l(\hat{\theta}) + q\,ln(n), AICc = AIC + \frac{2q(q + 1)}{n - q - 1},$$

$$A_0^* = \left(\frac{2.25}{k^2} + \frac{0.75}{k} + 1\right)\{-k - \frac{1}{k}\sum_{i=1}^{k}(2i - 1)\,ln(z_i(1 - z_{k-i+1}))\},$$

$$KS = Max\left\{\frac{i}{k} - z_i, z_i - \frac{i - 1}{k}\right\},$$

where q denotes the number of parameters, $l(\hat{\theta})$ denotes the log-likelihood function evaluated at the maximum likelihood estimates, n denotes the number of observations in a sample, k denotes the number of classes and z_i denotes the cumulative distribution function (cdf) of the ith class.

4.2   Data Application

For comparison purposes, we have used three data sets which are reported by Mir and Ahmad (2009), Kundu and Gupta (2009), and Bakouch et al., (2014), respectively. The data sets along with summary statistics and index of dispersion are presented in Appendix: For data set-I see Table 1(a), Table 1(b) and Table 1(c), for data set-II, see Table 2(a), Table 2(b) and Table 2(c) and for data set III see Table 3(a), Table 3(b) and Table 3(c). Descriptive statistics of all data sets are summarized in Table 1(b), Table 2(b)and Table 3(b). Some theoretical statistics of the ZTDL distribution for the all data sets are given in Table 1(c), Table 2(c), and Table 3(c), where definitions of mean, variance, mean deviation about the mean and entropy are used for computations, noting that the parameters of the ZTDL are replaced by their corresponding ML estimates for each data set. From the six tables, it can be noted that the considered theoretical statistics of the ZTDL distribution are closed to the sample moments of all data sets. Count Data set: The count data set presented in Table 1(a) shows number of mites per leaf, the European red mite belongs to a group of plant-feeding mites, called spider mites. The summary descriptive statistics of this data set is presented in Table 1(b) also portrays that the data are under dispersed. Table 1(d) portrays the maximum likelihood estimates with estimated standard errors, $-l$, $A_0^*$and KS with $p-value$ and AIC, BIC and AICc values of competing models. From these tables it is evident that proposed model displays its appropriateness by showing smaller A_0^* and KS statistics with higher *p-values*. Moreover, the proposed model also shows smaller values of model selection statistics AIC, BIC and AICc. These smaller values of the model selection statistics advocate that the proposed model is the best probability model among the competing models for this under dispersed data set. Experimental Data sets: The second and third data sets are given in Tables 2(a) and 3(a) attached in appendix. The second data set consists of remission times in weeks for 20 leukemia patients randomly assigned to a certain treatment. Treatment for leukemia aims to achieve remission. Remission means that no leukemia cells can be found in the blood or bone marrow and the bone marrow is working normally again. In people treated for acute leukemia, remission may last many years, and then they are

considered cured. Whereas the third data set represents at least one goal scored directly from a penalty kick, foul kick or any other direct kick (all of them together will be called as kick goal) by any team. These data sets are taken from Bakouch et al., (2014) and Kundu and Gupta (2009) respectively. From Table 2 (b) and 3(b) attached in appendix, it is evident that these data sets are over dispersed data sets. The KS, A^*_0 , AIC, BIC and AICc statistics are provided in the Tables 2(d) and 3(d). From these results it is observed that the proposed model is also appropriate for over-dispersed data structure by showing minimum $-l$, A^*_0 and KS with higher p-values. Moreover, the corresponding distribution plots comparing the fitted and observed distribution functions of the three data sets are shown in Figures 6–8. Also, the hazard function plots comparing the fitted and empirical functions are displayed by Figures 9–11. A distribution plot is a plot of the empirical distribution function against fitted distribution function. For example, the zero truncated discrete Lindley distribution, $\sum_{j=1}^{n}\frac{I_{x_j \leq x-0.375}}{n+0.25}$, are plotted versus $1-\frac{\left(1+\hat{\beta}x_j+\hat{p}\left(\hat{\beta}-(1+\hat{\beta}x_j)\right)\right)\hat{p}^{x_j-1}}{1+\hat{\beta}-\hat{p}}$ as recommended by Chambers et al., (1983), where x(j) are the sorted values of the observed data in the ascending order, n is the number of observations and I ˆ A denotes the indicator function. An empirical hazard plot is a plot of the empirical hazard function against fitted hazard function. For example, the zero truncated discrete Lindley distribution:

For number of classes less than $10(1/2n - j - 5)$ are plotted versus $\frac{(1-\hat{p})^2(1+\hat{\beta}x_j)}{1+\hat{\beta}x_j+\hat{p}\left(\hat{\beta}-(1+\hat{\beta}x_j)\right)}$, for number of classes greater than 10 but less than 20 $20(1/2n - j - 10)$ are plotted versus $\frac{(1-\hat{p})^2(1+\hat{\beta}x_j)}{1+\hat{\beta}x_j+\hat{p}\left(\hat{\beta}-(1+\hat{\beta}x_j)\right)}$ and for number of classes greater than 20 $20(1/2n - j - 15)$ are plotted versus $\frac{(1-\hat{p})^2(1+\hat{\beta}x_j)}{1+\hat{\beta}x_j+\hat{p}\left(\hat{\beta}-(1+\hat{\beta}x_j)\right)}$. The distribution plots and empirical hazard function show that the proposed ZTDL distribution provides better fits than other competing distributions for at least three datasets
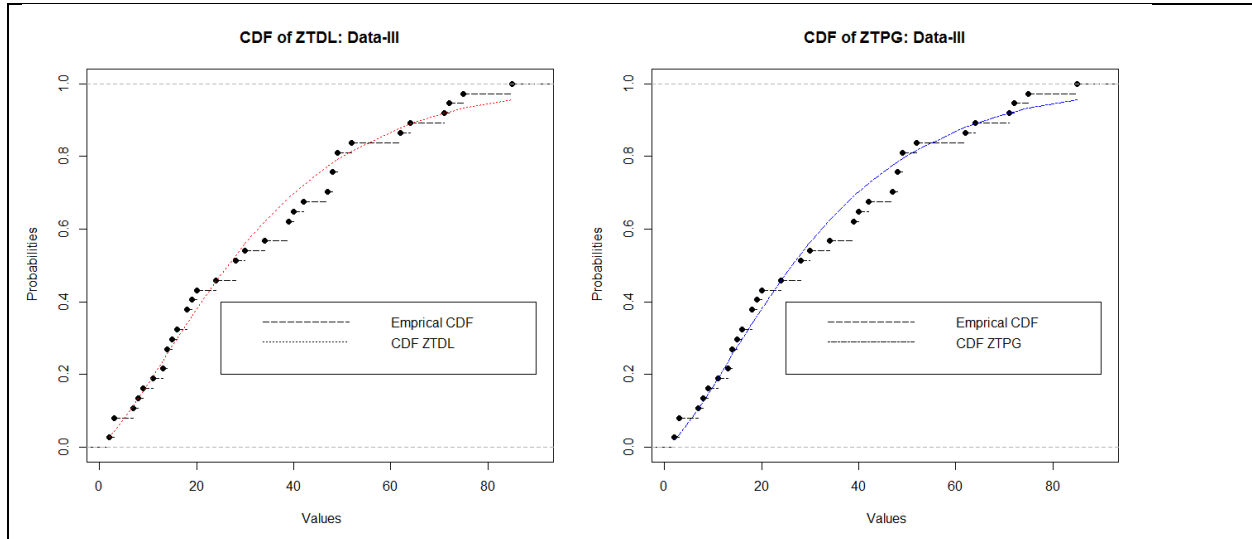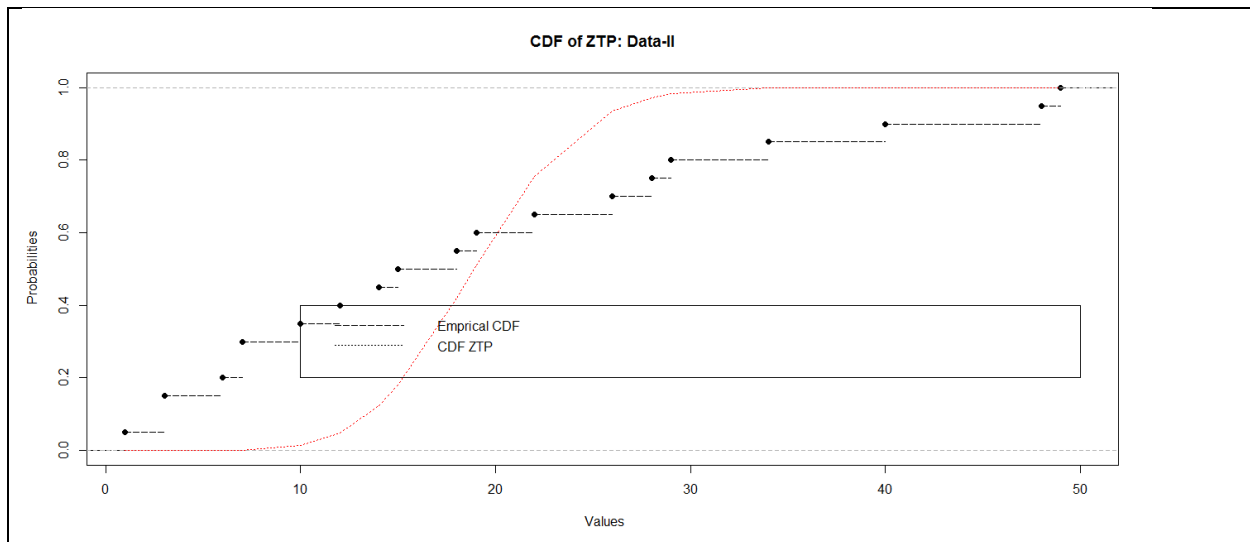
Figure 6: CDF plots Comparisons for data set –I

Figure 7: CDF plots Comparisons for data set -II

Figure 8: CDF plots Comparisons for data set -III

Pak.j.stat.oper.res. Vol.16 No. 1 2020 pp 167-190

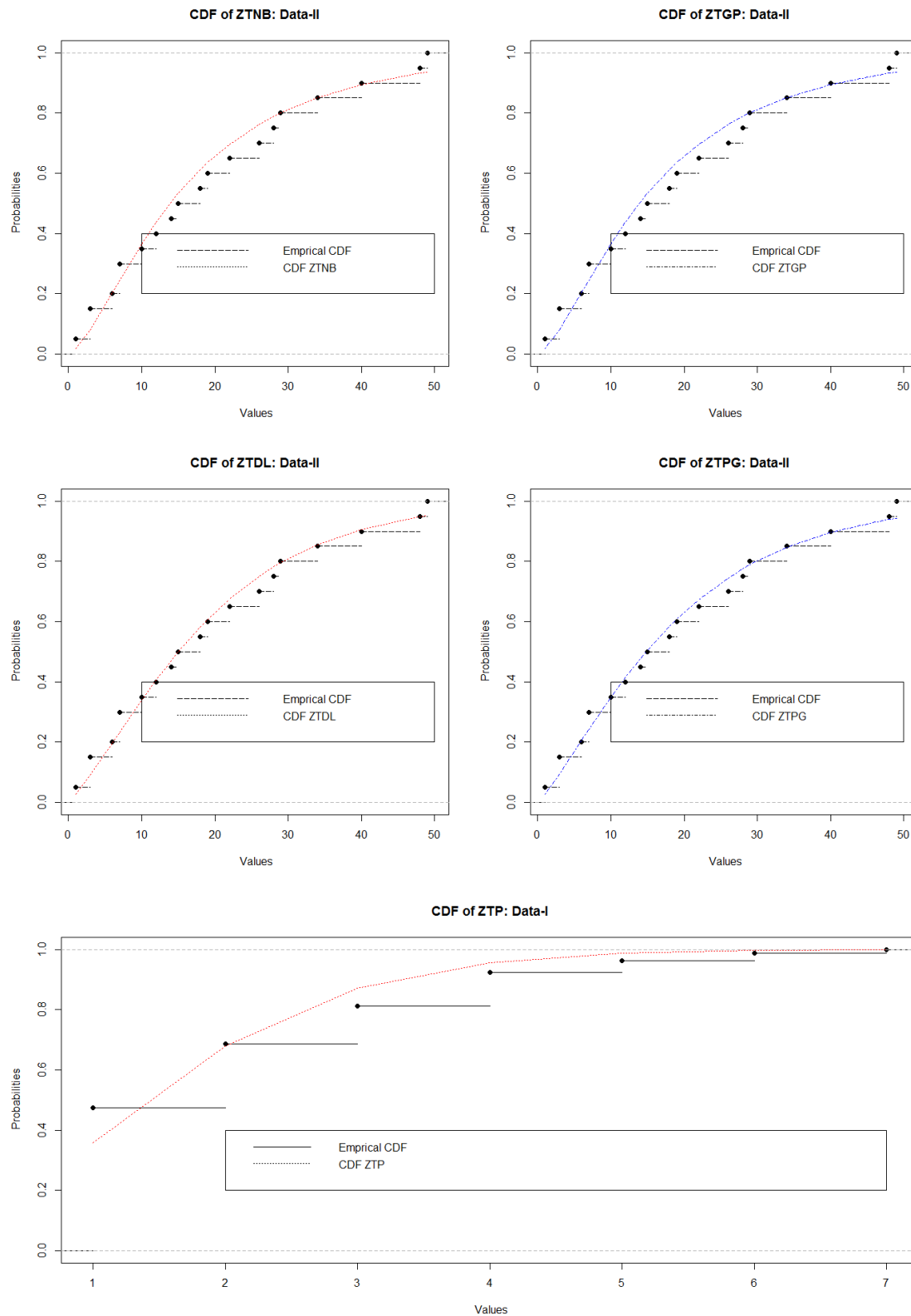DOI: http://dx.doi.org/10.18187/pjsor.v16i1.2133

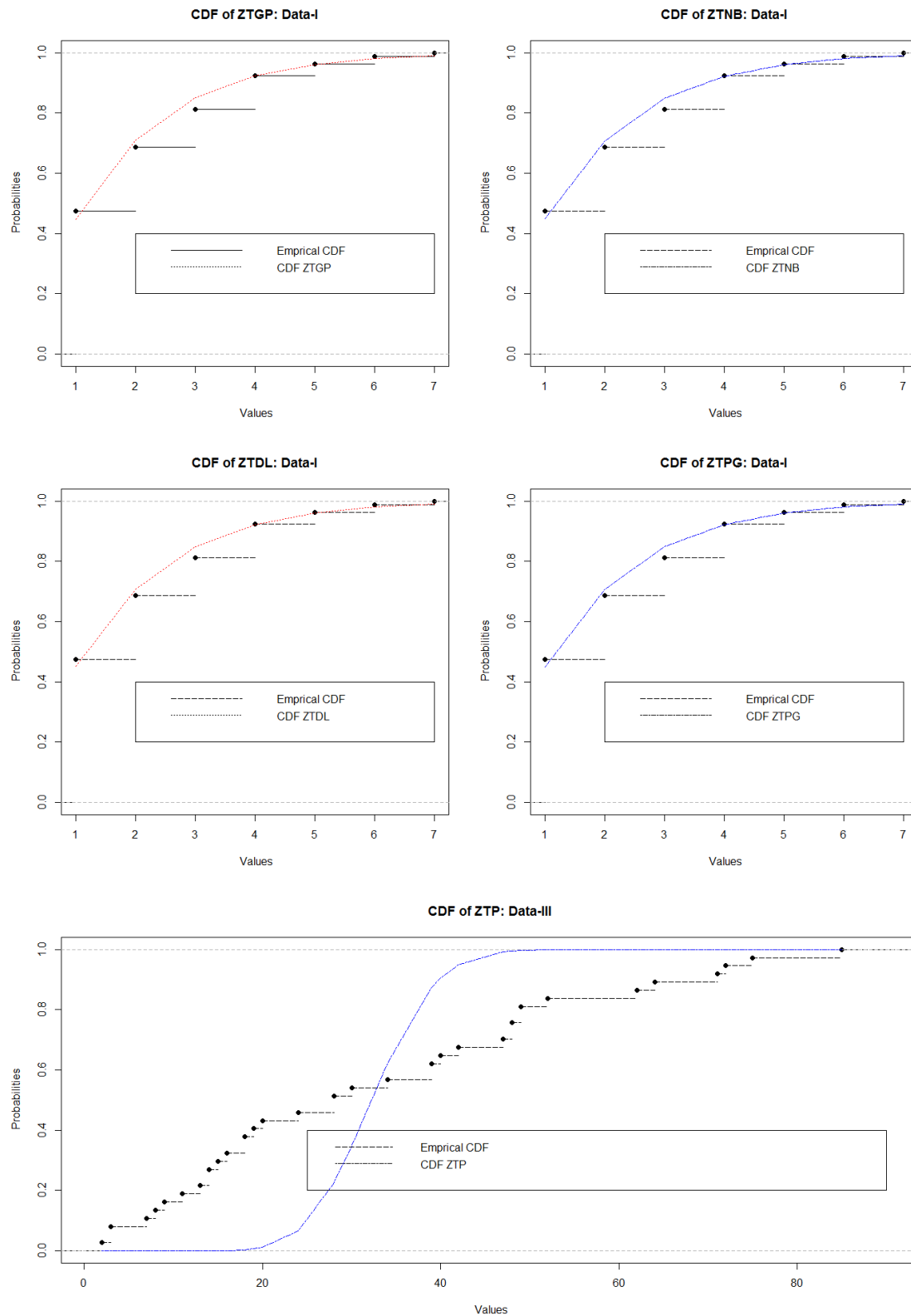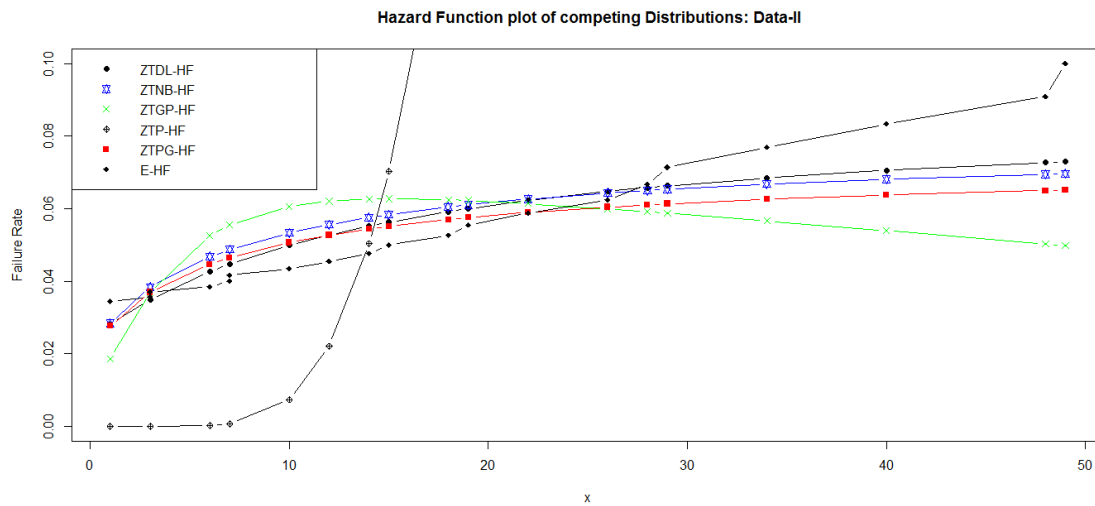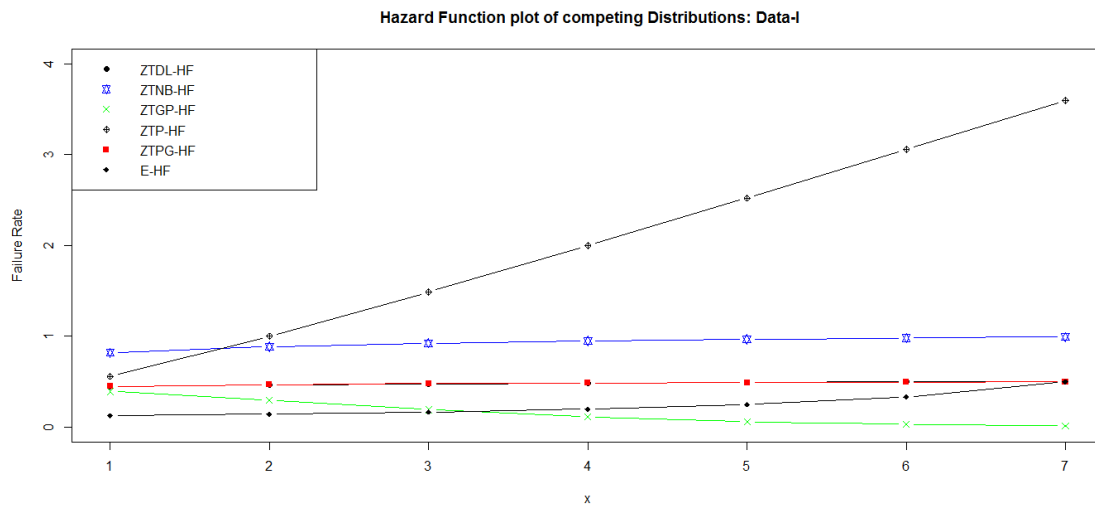Figure 9: Hazard Function plots Comparisons for data set-I

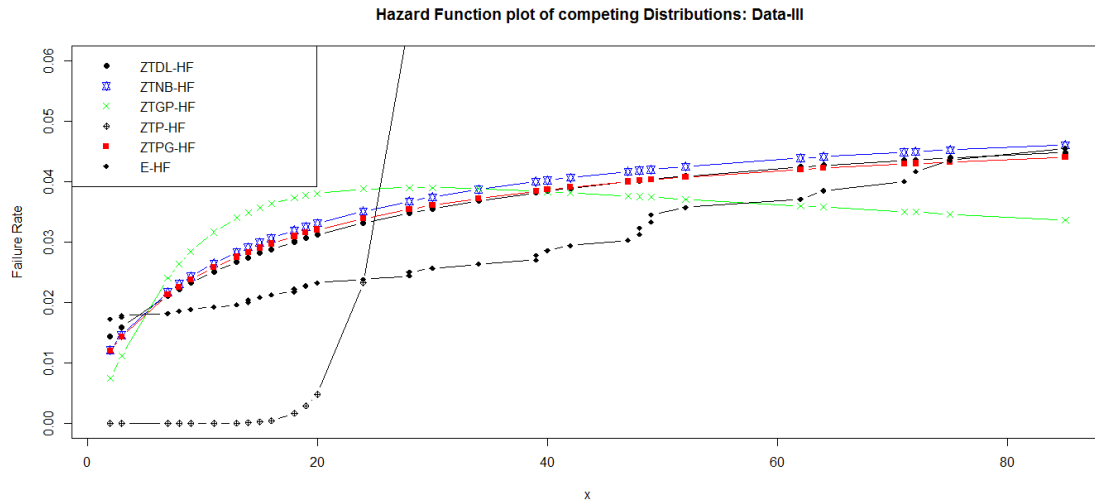Figure 10: Hazard Function plots Comparisons for data set-II

Figure 11: Hazard Function plots Comparisons for data set-III

Conclusions

A Zero truncated discrete Lindley with two parameters is proposed and it contains the zero truncated geometric and the one parameter zero truncated discrete Lindley distributions as submodels. Moreover, the distribution represents a two-component mixture of a zero truncated geometric distribution and a zero truncated negative binomial distribution with certain parameters. Various distributional properties, reliability characteristics and a characterization are studied. Also, the collective risk model is discussed by considering the proposed distribution as primary distribution and exponential and Erlang as secondary distributions. Moreover, it is found that the distribution has not only simple structure, mathematically amenable, more flexible and longer tail than other models but also more appropriate for modeling actuarial ecology, health, psychology, sociology and engineering data sets.

Acknowledgment

Appendix

Table 1(a): Observed number of Red Mites per leaf for Data Set-I								
Number of mites per leaf	1	2	3	4	5	6	7	Total
Observed Leaf	38	17	10	9	3	2	1	80

Table 1(b): Summary Statistics of number of Red Mites per leaf					
Mean	Variance	ID	Skewness	Kurtosis	Entropy
2.1500	2.1037	0.9785	1.2646	3.9613	1.4586

Table 1(c): Summary Statistics of number of Red Mites per leaf from ZTDL					
Mean	Variance	ID	Skewness	Kurtosis	Entropy
2.1499	2.2703	1.0559	3.6596	8.1462	2.7182

Table 1 (d): Test Summary for Data Set-I with standard error in parenthesis								
Distribution $\widehat{(p,\beta)}$	$\hat{p}$	$\hat{\beta}$	KS	p-value	A^*_0	$-l$	AIC	BIC	AICc
ZTDL	0.4542 (0.0331)	0.3371 (0.2685)	0.5648	0.107	7.1841	118.6661	241.3322	246.0963	238.4822
ZTNB	0.5190 (0.0356)	1.3843 (0.2731)	0.5655	0.106	7.1888	118.7285	241.4571	246.2211	238.6070
ZTGP	0.2646 (0.0548)	0.9986 (0.1805)	0.5662	0.106	7.2208	118.9245	241.849	246.6131	238.9990
ZTPG	0.9265 (0.1323)	1.3842 (0.2732)	0.5655	0.106	7.1888	118.7285	241.4571	246.2212	238.6071
ZTP	--------	1.7916 (0.1705)	0.5854	0.091	9.8636	122.7947	247.5894	249.9714	245.6394

Table 2(a): Remission times in week for 20 Leukomia patient Data Set-II
1,3,3,6,7,7,10,12,14, 15,18,19 22,26,28,29,34,40,48,49

Table 2(b): Summary Statistics Remission times in week for 20 Leukomia patient					
Mean	Variance	ID	Skewness	Kurtosis	Entropy
19.5500	216.05	11.0512	0.6521	2.3802	2.8571

Table 2(c): Summary Statistics Remission times in week for 20 Leukomia patient from ZTDL					
Mean	Variance	ID	Skewness	Kurtosis	Entropy
19.2148	229.557	11.9469	2.2216	6.2589	2. 7182

Table 2 (d): Test Summary for Data Set-II with standard error in parenthesis								
Distribution $\widehat{(p,\beta)}$	$\hat{p}$	$\hat{\beta}$	KS	p-value	A^*_0	$-l$	AIC	BIC	AICc
ZTDL	0.9127 (0.0135)	0.2514 (0.3318)	0.1392	0.706	0.5573	78.3798	160.7596	162.7511	158.3596
ZTNB	0.0727 (0.0128)	1.5026 (0.2687)	0.1432	0.691	0.5631	78.4828	160.9656	162.9571	158.5656
ZTGP	0.7513 (0.0488)	4.8236 (0.7983)	0.1703	0.593	0.7882	78.4828	160.9656	162.9571	158.5656
ZTPG	12.7577 (2.4268)	1.5026 (0.2686)	0.1432	0.691	0.5631	78.4828	160.9656	162.9571	158.5656
ZTP	--------	19.5499 (0.5606)	0.3254	0.3254	∞	152.7181	152.7181	308.4316	305.6360

Table 3(a): Number of Goal Scored by any team Data Set-III
20,18,19,85,40,49,8,71,39,48,72,62,9,3,75,18,14,42,52,34,
39,7,28,64,15,48,16,13,14,11,49,24,30,3,47,28,2

Table 3(b): Summary Statistics of Number of Goal Scored by any team					
Mean	Variance	ID	Skewness	Kurtosis	Entropy
32.8649	521.3423	15.8632	0.5444	2.2825	3.3487

Table 3(c): Summary Statistics of Number of Goal Scored by any team from ZTDL					
Mean	Variance	ID	Skewness	Kurtosis	Entropy
32.8059	642.3130	19.5435	2.1241	6.1401	2.7182

Table 3(d): Test Summary for Data Set-III with standard error in parenthesis									
Distribution $\widehat{(p,\beta)}$	$\hat{p}$	$\hat{\beta}$	KS	p-value	A^*_0	$-l$	AIC	BIC	AICc
ZTDL	0.9463 (0.0061)	0.2242 (0.2349)	0.1204	0.647	0.3859	163.5039	331.0078	334.2296	328.3321
ZTNB	0.0508 (0.0062)	1.7509 (0.2029)	0.1242	0.630	0.4046	163.6472	331.2944	334.5162	328.6187
ZTGP	0.7984 (0.0280)	6.6177 (0.7277)	0.1493	0.512	0.6147	165.2130	334.4260	337.6478	331.7503
ZTPG	18.6680 (2.3850)	1.7509 (0.2030)	0.1242	0.630	0.4047	163.6471	331.2942	334.5160	328.6185
ZTP	--------	32.8649 (0.9424)	0.4222	0.0050	∞	391.0929	784.1858	785.7967	782.2939

References

1. Abouammoh, A.M. and Mashhour, A.F. (1981). A Note on the Unimodality of Discrete Distributions. *Communication in Statistics Theory and Methods,* 10(13), 1345-1354.

2. Asgharzadeh, A., Bakouch, H.S., Nadarajah, S. and Sharafi, F. (2016). A new weighted Lindley distribution with application. *Brazilian Journal of Probability and Statistics,* 30(1), 1-27.

3. Bakouch, H. S., Al-Zahrani B. M., Al-Shomrani A. A., Marchi V. A., and Louzada F.,(2012). An extended Lindley Distribution. *Journal of the Korean Statistical Society,* 41, 75-85.

4. Bakouch, H. S.,Jazi, M.A. and Nadarajah, S. (2014). A new discrete distribution. A *Journal of Theoretical and Applied Statistics* 48, 200-240.

5. Bakouch, H.S. and Severini, T.A. (2009).Non Parametric Estimation in Random Sum Models. *Statistica* LXIX (1), 73-88.

6. Bhati, D., Satry, D.V.S. and Qadri, P. Z. M. (2015).A New Generalized Poisson Lindley Distribution: Applications and Properties. *Austrian Journal of Statistics*, 44, 35-51.

7. Borah, M. and Saikia, K.R. (2017). Zero- Truncated Discrete Shanker Distribution and Its Applications . *Biometrics Biostatistics International Journal*, 5(6): 00152. DOI: 10.15406/bbij.2017.05.00152.

8.   Brass, W.(1959). Simplified methods of fitting the truncated negative binomial distribution. *Biometrika* 45,59-68.

9.   Creel, M. D. and Loomis, J. B. (1990). Theoretical and Empirical Advantages of Truncated Count Data Estimators for Analysis of Deer Hunting in California. *American Journal of Agricultural Economics,* 72, 434-441.

10.   Chakraborty, S. and Ong, S. H. (2016). A COM-Poisson type generalization of the negative binomial distribution. *Communication in StatisticsTheory and Methods*, 45, 4117-4135.

11.   Chambers, J. M., Cleveland, W. S., Kleiner, B. and Tukey, P. A. (1983). *Graphical Methods for Data Analysis,* Chapman and Hall,London.

12.   Finney, D.J. and Varley, G.C. (1955). An example of the truncated Poisson Distribution. *Biometrics* 11,387-391.

13.   Ghitany, M.E., Al-Mutairi, D. K., Balakrishnan, N. and Al-Enezi, L.J. (2013). Power Lindley distribution and associated inference. *Computational Statistics and Data Analysis,* 64, 20-33.

14.   Ghitany, M.E., Al-Mutairi, D.K. and Nadarajah, S. (2008). Zero-truncated Poisson-Lindley distribution an its application. *Mathematics and Computers in Simulation* 79, 279-287.

15.   15. Gradshteyn, I.S. and Ryzhik, I.M. (2008).*Table of Integrals,series and products, 5th Edition.* Academic Press Limited,London.

16.   Groggan, J. T.and Carson,R. T. (1991). Models for truncated counts. *Journal of Applied Econometrics*, 6, 225-238.

17.   Hussain, T., Aslam, M. and Ahmad, M. (2016). A two parameter discrete Lindley distribution. *Revista Colombiana de Estadistica,* 39(1), 45-61.

18.   Kennedy, B.S. (2005). Does Race Predict Stroke Readmission? An Analysis Using the Truncated Negative Binomial Model. *Journal of the National Medical Association,* 97(5), 699-713.

19.   Kokonendji, C. C. and Mizre, D. (2005). Overdispersion and Underdispersion Characterization of Weighted Poisson Distribution (Technical Report No.0523). France: LMA.

20.   Kundu, D. and Gupta, R.D.(2009). Bivariate generalized exponential distribution. *Journal of Multivariate Analysis,* 100, 581-593.

21.   Lee, A. H., Wang, K., Yau, K. K. W and Somerford, P.G. (2003). Truncated Negative Binomial Mixed Regression Modelling of Ischaemic Stroke Hospitalizations. *Statistics in Medicine,* 22, 1129-1139.

22.   Lindley, D.V. (1958). Fiducial distributions and Bayes theorem. *Journal of the Royal Statistical Society, Series B*, 20, 102-107.

23.   Lindsay, J.K.(1995).*Modeling Frequency and Count Data,* Clarendon Press, Oxford University New York.

24.   Mir, K. A. and Ahmad, M.(2009). Size Biased Distributions and their Applications. *Pakistan Journal of Statistics,* 25(3), 283-294.

25.   Nadarajah, S., Bakouch, H. S. and Tahmasbi, R. (2011). A Generalized Lindley Distribution. *Sankhya B*, 73,331-359.

26.   Nekoukhou, V., Alamatsaz, M. H. and Bidram, H. (2012). A Discrete Analog of the Generalized Exponential Distribution. *Communication in Statistics Theory and Methods*, 41(11), 2000-2013.

27.   Phang, A. L. and Loh, E.F. (2013).Zero Truncated Strict Arcsine Model. *International Journal of Computer, Electrical, Automation, Control and Information Engineering, 7(7), 989-991.*

28.   Shanker, R. (2017). A Zero-Truncated Poisson-Amarendra Distribution and Its Application. *International Journal of Probability and Statistics,* 6(4),82-92

29.   Shanker, R. and Shukla, K.K. (2017). Zero-Truncated Poisson-Garima Distribution and its Applications. *Biostat Biometrics Open Access Journal,* 3(1): 555605. DOI: 10.19080/BBOAJ.2017.03.55560502.

30.   Shanker, R., Sharma, S. and Shanker, R. (2013). A two-parameter Lindley distribution for modeling waiting and survival times data. *Applied Mathematics,* 4, 363-368.

31.   Steutel, F. W. and Van Harn, K. (1979). Discrete Analogues of SelfDecomposability and Stability. *Ann. Prob.,* 7, 893-899.