Austrian Journal of Statistics
2024, Volume 53, 43–63.
http://www.ajs.or.at/
doi:10.17713/ajs.v53i4.1818

On Positive Inflated Geometric Distribution: Properties and Applications

Zehra Skinder
Dept. of Mathematical
Sciences, IUST

Peer Bilal Ahmad*
Dept. of Mathematical
Sciences, IUST

Muneeb Ahmad Wani
Dept. of Statistics,
IIT KANPUR

Abstract

One-inflation in zero-truncated count data has recently found considerable attention. In this regard, zero-truncated Geometric distribution and distribution to a point mass at one are used to create a one-inflated model, namely, one-inflated zero-truncated Geometric distribution. Its reliability characteristics, generating functions, and distributional properties are investigated in detail, which includes survival function, hazard rate function, reverse hazard rate function, probability generating function, characteristic function, variance, skewness, and kurtosis. Monte Carlo simulation have been undertaken to evaluate the effectiveness of the maximum likelihood estimators. To test the compatibility of our proposed model, the baseline model and the proposed model are distinguished by using the two different test procedures. The adaptability of the suggested model is demonstrated using two real-life datasets from separate domains by taking various performance measures into consideration.

Keywords: zero-truncation, one-inflation, goodness of fit, simulation, hypothesis testing, geometric distribution.

1. Introduction

In every discipline of knowledge, including epidemiology, engineering, sociology, biological research, insurance, agriculture, and public health, the statistical analysis and modelling of count data is very important. We fit a valid probability model to count data in order to build up decision-making while dealing with count data.

When dealing with positive count data possessing variability, one can model positive data by truncating the distribution at zero, resulting in a zero-truncated distribution. When a specific range of values for the variables is ignored or cannot be seen, the resulting model is said to be truncated. Truncation of probability distributions is an essential statistical feature with several applications in different areas. It is preferable to use a zero-truncated probability distribution instead of any other discrete distribution, when data is to be represented or produced without zeros. Many datasets exclude zero counts, such as the number of siblings in a family, the number of passengers in a car including the driver, the number of articles published in different journals from various disciplines, the number of disturbing events reported by patients, the number of flowers bloomed, and the number of times a voter has cast a ballot in a general election, etc. Zero-truncated probability models behave well when modelling such types of situations, and the results drawn from them seem quite sound.

Positive count data modelling can be traced back to the mid-twentieth century, when the first truncated model, known as the zero-truncated Poisson distribution (ZTPD), was put forward by David and Johnson (1952) to model such data. Later, several models like the Negative Binomial distribution (NBD), were modified to zero-truncated Negative binomial distribution by Sampford (1955) as an alternative to the zero-truncated version of the Poisson distribution. Ghitney *et al.* (Ghitany, Al-Mutairi, and Nadarajah (2008)) introduced the zero-truncated Poisson–Lindley distribution (ZTPLD) and developed the estimation methods based on the moment method and the maximum likelihood method. Phang and Loh (2013) discussed the applications of ZTNBD in analyzing the abundance of rare species and hospital stays. Shanker and Fesshaye (2016) studied the nature and behaviour of other truncated distributions like ZTPD, ZTPLD, and zero-truncated Poisson-Sujhata distribution (ZTPSD) by drawing different inferences. Shibu *et al.* (Shibu, Chesneau, Monisha, Maya, and Irshad (2023)) introduced the novel Lagrangian zero-truncated Katz distribution and investigated various structural properties of the model and showed that the model is both over-dispersed as well as under-dispersed. Elah *et al.* (Elah, Ahmad, and Wani (2023)) introduced a new truncated model called zero-truncated New discrete distribution to analyze the various applications in different fields. Kiani (2020) introduced a simple structural model called zero-truncated discrete Lindly distribution. Based on the values of parameters, the distribution can be thought of as a two-model mixture of a zero-truncated Geometric distribution (ZTGD) and ZTNBD. Park *et al.* (Park, Gou, and Wang (2022)) studied the estimation of the parameters of truncated Geometric distribution. The baseline model used in this article is the zero-truncated Geometric distribution. A random variable with non-negative integer support is said to possess Geometric distribution from Gómez-Déniz (2010), if its probability mass function (pmf) is of form

$$f(t) = p(1-p)^t; t = 0, 1, 2..., 0 < p < 1 \tag{1}$$

where t denotes the number of failures until the first success. One of the essential properties of this distribution, among all other distributions, is the lack of memory property. which plays an important role in the branch of applied probability.

The pmf of zero-truncated Geometric distribution from Devi *et al.* (Devi, Gupta, and Kumar (2017)) is given by

$$g(t;p) = p(1-p)^{t-1}; t = 1, 2, 3..., 0 < p < 1 \tag{2}$$

The mean and variance of the zero-truncated Geometric distribution are as follows

$$Mean = \frac{1}{p}$$

$$Variance = \frac{1-p}{p^2}$$

To account for large number of ones in the dataset, inflated models based on the zero-truncated distributions have been investigated. One-inflation in zero-truncation is when there are excessive number of ones than predicted in the observed data. Accommodating one inflation in zero-truncation is crucial in situations where the event occurrence is limited, and there are several reasons, such as avoidance or behavioral changes, stigma, and heterogeneity, that lead to only a single instance of the event. Modelling one-inflation in such cases allows us to capture scenarios where an event happens only once, and subsequent events become less likely due to these reasons. Accommodating one inflation acknowledges the possibility that the individuals may be willing to report an event only once, and subsequent events might go unreported. As is the case related to the drunk driving dataset, the majority of the arrests happen as a result of police stopping and questioning every passing motorist. However, because drunk driving carries a stigma, there might be a strong behavioral reaction after the initial arrest, leading to some people leaving the population of drunk drivers. One-inflation is justified by the possibility that the subject will learn to avoid being observed again, which we term avoidance ability.

Several zero-truncated models, inflated at "1", have been recently attracted by several researchers like Godwin and Böhning (2017) proposed the one-inflated positive Poisson model to deal with phenomena of excess 1's. Godwin (2017, 2019) proposed a one-inflated zero-truncated Negative Binomial (OIZTNB) model and a positive Poisson mixture model, respectively, and used them as truncated distributions in the Horvitz–Thompson estimation of unknown population size. They also analyzed the various applications in different fields to check the model's adaptability. Tajuddin *et al.* (Tajuddin, Ismail, and Ibrahim (2021)) investigated the parameter estimation techniques of one-inflated positive Poisson distribution and compared different estimation methods in terms of unbiasedness, consistency, efficiency, and deficiency and found that all the estimators are consistent and asymptotically normal. They also developed a one-inflation index and analyzed the presence of excess ones in the dataset. Some of the recent works regarding the inflation aspect in count data are by Skinder *et al.* (Skinder, Ahmad, and Elah (2023)) and Wani and Ahmad (2023). Tajuddin *et al.* (Tajuddin, Ismail, and Ibrahim (2022)) introduced the count model called one-inflated positive Poisson Lindly distribution to estimate the population size of criminals. Kaskasamkul and Bohning (2017) proposed the inflated count model based on the Geometric distribution to estimate population size.

For positive count data, most of the data comes from the count "1" because "0" counts remain unobserved and it is believed that excessive number of "1" counts can contribute to the dispersion (over or under) in the data. Although statistical modelling in this subject has come a long way, new models are still required from time to time. These new models were inspired by the emerging trends that frequently occur in our count data. As a result, we have developed a simple and flexible two-parametric probability model that can handle the statistical dispersion (over and under-dispersion) and the excess of "1" counts in the data. The model developed may be used as an alternative to some other distributions mentioned in the application section as it provides better fitting as compared to them.

The remainder of the paper is organized as follows. In section 2, the proposed model, and its probability mass function (pmf) and cumulative distribution function (cdf) are introduced. Further, reliability characteristics and generating functions are also presented in this section. In section 3, we have obtained the various structural properties including mean, variance, dispersion index, skewness, and kurtosis. In section 4, we discuss the estimation of the parameters of the proposed model by the maximum likelihood method. A rigorous simulation study is also discussed in this section. In section 5, different test procedures are applied for examination to check the significance of the inflation parameter. Certain real-life applications are considered in section 6 from various domains for highlighting the functionality of the model. Lastly, the conclusion is discussed in section 7 itself.

2. Methodology

By combining the ZTG distribution with a point mass π, a distribution is obtained that accounts for the inflated frequency at one. Take an experiment into consideration that led to the following two processes:

- The first process generates only one count with probability π, $0 < \pi < 1$.
- The second process is governed by ZTGD with probability $(1 - \pi)$.

Also, suppose that the experiment is repeated a number of times independently. Assume that the first process occurs with probability π and the second process occurs with probability (1-π). Now, take a count variable, say Z, into consideration that takes some distribution which allows for frequent one-valued observations. When the first process occurs, Z is set at Z=1; when the second process occurs i.e., the counts are generated according to ZTGD random variable.

Thus, for Z=1, which could be from the occurrence of either process first with probability π or process second with probability (1-π). We could have

$$P(Z=1)=\pi+(1-\pi)p \qquad ;z=1$$

For $Z>1$, the pmf of Z follows the ZTGD written as

$$P(Z=z)=(1-\pi)p(1-p)^{z-1} \qquad ;z=2,3,4,...,$$

Hence, the pmf of count variable Z is obtained by combining the above two equations and is given as

$$P(Z=z)=\begin{cases} \pi+(1-\pi)p; & z=1 \\ (1-\pi)p(1-p)^{z-1}; & z=2,3,4,..., \\ 0; & \text{otherwise,} \end{cases} \tag{3}$$

where 0<π<1 and $0<p<1$. Equation (3) is known as one-inflated zero-truncated Geometric distribution. To prove that equation (3) is a proper pmf, consider

$$\sum_{z=1}^{\infty} P(Z=z)=\pi+(1-\pi)p+(1-\pi)\sum_{z=2}^{\infty} p(1-p)^{z-1}$$

$$=\pi+(1-\pi)\sum_{z=1}^{\infty} p(1-p)^{z-1}$$

$$=\pi+(1-\pi)\sum_{z=1}^{\infty} g(z),$$

where $g(z)$=$p(1-p)^{z-1}$ is the pmf of the ZTGD in terms of "t" in equation (2). Thus, $\sum_{z=1}^{\infty} g(z)=1$.

Clearly, when $\pi=0$, the distribution reduces to ZTGD with pmf given in (2).

The pmf plots in figure (1) for different combinations of parametric values indicate that the OIZTGD (3) is uni-modal. Further, the mode is at one for different combinations of parameters. Moreover, the tail shows a rapid decrease as the value increases for different combinations of parameters.

2.1. Cumulative distribution function (cdf)

Theorem 1. *If Z $\sim$ OIZTGD (π, p), then its cdf is given as*

$$F(Z)=[1-(1-p)^{z}(1-\pi)] \tag{4}$$

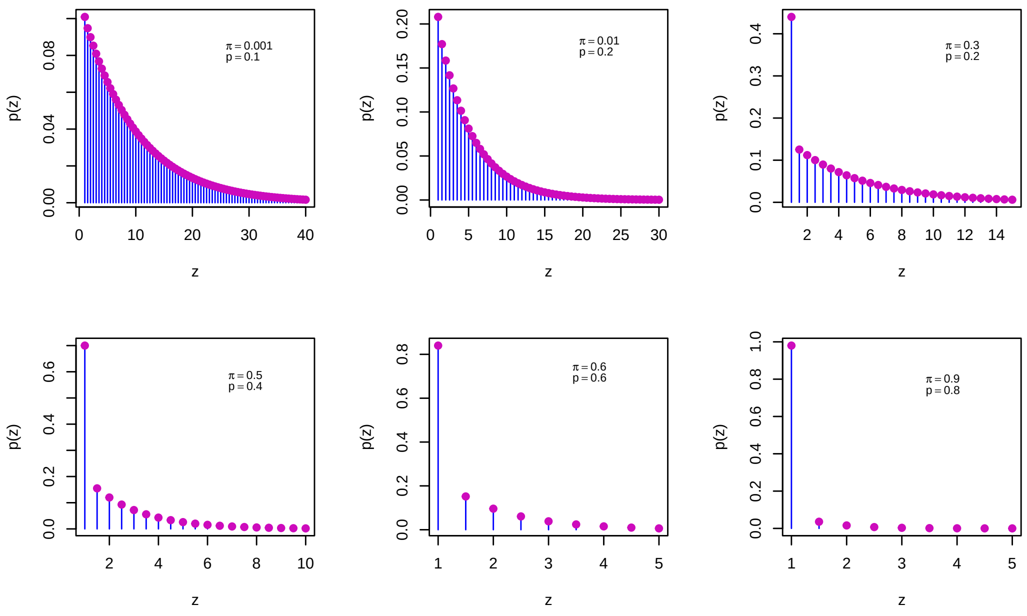

Figure 1: The PMF Plots of OIZTGD

Proof. If Z∼ OIZTGD (π, p), then its cdf is as follows

$$F(Z) = P(Z \leq z)$$

$$= \sum_{t=1}^{z} P(Z = t)$$

$$= \pi + (1-\pi)p \sum_{t=1}^{z} (1-p)^{t-1}$$

$$= \pi + (1-\pi)p \left[\frac{1-(1-p)^{z}}{p} \right]$$

$$= \pi + (1-\pi)[1-(1-p)^{z}]$$

Hence proved. □

The cdf plots of OIZTGD (π, p) with different combinations of parameters of π and p are provided in figure (2).

2.2. Reliability characteristics along with generating functions

In this part, various reliability characteristics like survival analysis, hazard function, and reverse hazard function are discussed along with generating functions. Further, the distributional properties are also discussed.

Survival function (SF)

The probability that a system will survive beyond a certain time period is called survival function. It is also called reliability or survivor function and is denoted by “S". Further, the

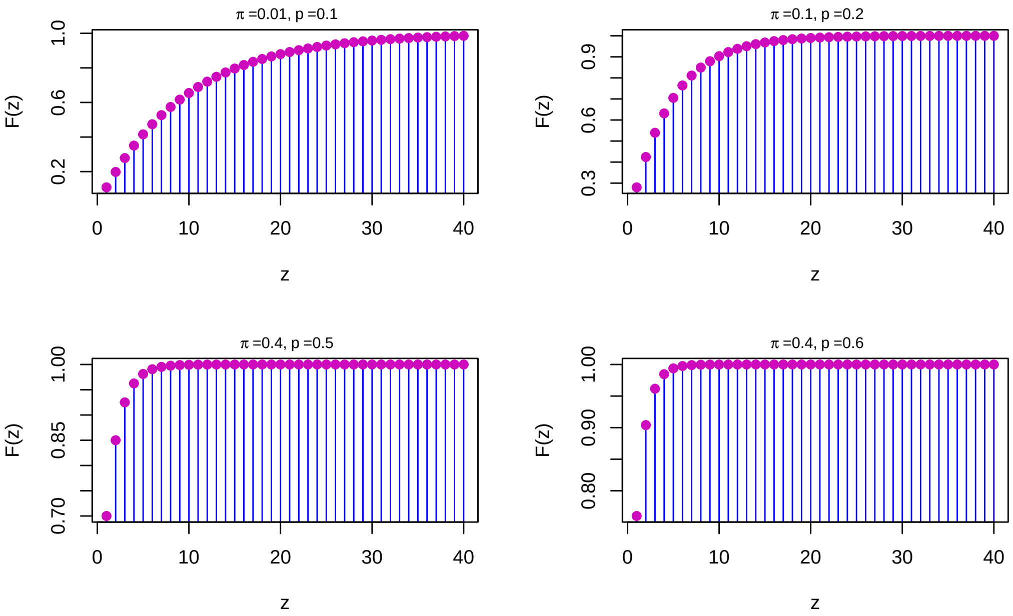

Figure 2: The CDF plots of OIZTGD

survival plots are also shown in figure (3).
If Z ∼ OIZTGD (π, p), then its survival function is as follows:

$$S(Z) = 1 - F(Z) = 1 - [\pi + (1-\pi)[1-(1-p)^z]] = [(1-\pi)(1-p)^z] \tag{5}$$

Hazard rate function (HRF)

Let z_1, z_2, z_3, ..., z_n be a random sample from OIZTGD (π, p) as given by equation (3). Suppose Y is the number of z_i's taking the value one. Then equation (3) can be written as follows:

$$P(Z = z_i) = [\pi + (1-\pi)p]^Y [(1-\pi)p(1-p)^{z-1}]^{1-Y}$$

Now, using S(z) from equation (5). The hazard rate function of OIZTGD $(\pi,\, p)$ is given as

$$H(Z) = \frac{P(z)}{S(z)}$$

$$= \frac{[\pi + (1-\pi)p]^Y [(1-\pi)p(1-p)^{z-1}]^{1-Y}}{[(1-\pi)(1-p)^z]}$$

Reverse hazard rate (RHR)

The reverse hazard rate is defined as the ratio of the probability mass function to the distribution function and is denoted as

$$R(Z) = \frac{P(z)}{F(z)}$$

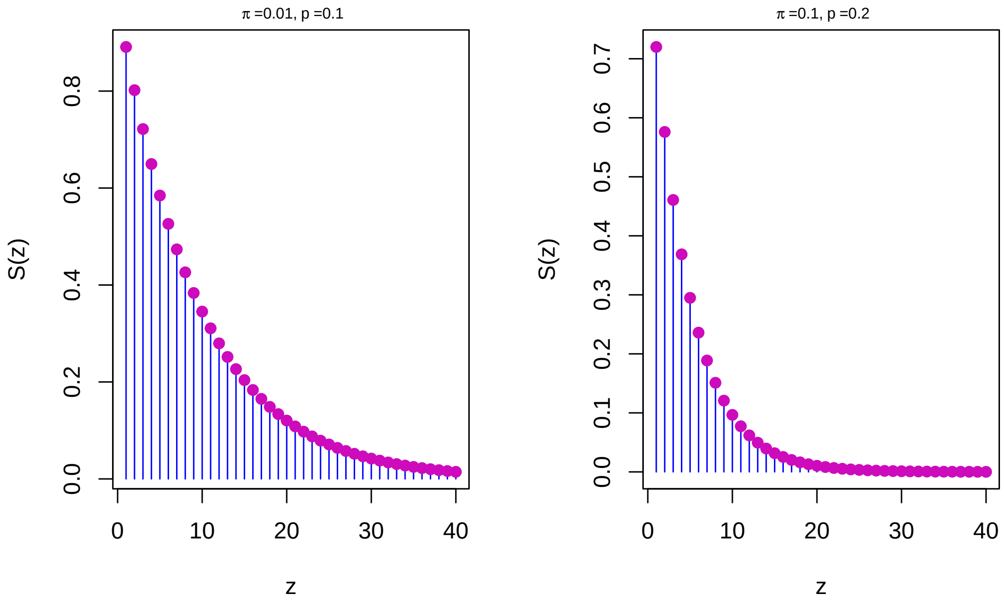

Figure 3: The survival plots of OIZTGD

$$= \frac{[\pi + (1-\pi)p]^Y [(1-\pi)p(1-p)^{z-1}]^{1-Y}}{\pi + (1-\pi)[1-(1-p)^z]}.$$

Mill's ratio

Mill's ratio is defined as the ratio of the survival function to the probability mass function and is denoted as

$$m(z) = \frac{1}{H(z)} = \frac{S(z)}{P(z)}$$

$$= \frac{[(1-\pi)(1-p)^z]}{[\pi + (1-\pi)p]^Y [(1-\pi)p(1-p)^{z-1}]^{1-Y}}$$

2.3. Probability generating function (pgf)

Theorem 2. *If $Z \sim$ OIZTGD $(\pi,\, p)$, then its pgf, $P_z(t)$ is given as*

$$P_z(t) = t\pi + (1-\pi)tp\frac{1}{(1-t-tp)} \tag{6}$$

Proof. If Z $\sim$ OIZTGD $(\pi,\, p)$, then its pgf is as follows

$$P_z(t) = \sum_{z=1}^{\infty} t^z P(Z=z)$$

$$P_z(t) = \sum_{z=1}^{\infty} t^z \left[(\pi + (1-\pi)p) + (1-\pi)p(1-p)^{z-1}\right]$$

$$P_z(t) = t\pi + \frac{(1-\pi)p}{(1-p)} \sum_{z=1}^{\infty} t^z (1-p)^{z-1}$$

$$P_z(t) = t\pi + (1-\pi)tp\frac{1}{(1-t-tp)}$$

Hence proved. □

Remark 1. *Putting $t = e^t$ in equation (6), the moment generating function, $M_z(t)$ of OIZTGD(π, p) is defined as*

$$M_z(t) = e^t\pi + (1-\pi)e^tp\frac{1}{(1-e^t-e^tp)}$$

Remark 2. *Putting $t = e^{it}$ in equation (6), the characteristic function, $\psi_z(t)$ of OIZTGD(π, p) is defined as*

$$\psi_z(t) = e^{it}\pi + (1-\pi)e^{it}p\frac{1}{(1-e^{it}-e^{it}p)}$$

3. Distributional properties

3.1. Moments

Theorem 3. *If Z ∼ OIZTGD (π, p), Then its r^{th} order moment about zero is as follows:*

$$\mu_r' = E(z^r) = \pi + (1-\pi)\frac{p}{(1-p)}Li_{-r}(1-p) \tag{7}$$

Proof. If Z ∼ OIZTGD (π, p), then the r^{th} order moment about zero is

$$\begin{aligned}\mu_r' &= E(Z^r)\\ &= \sum_z z^r p(\pi, p)\\ &= \pi + (1-\pi)\frac{p}{(1-p)}\sum_{z=1}^{\infty} z^r(1-p)^z\\ &= \pi + (1-\pi)\frac{p}{(1-p)}Li_{-r}(1-p) \qquad \left[\because \sum_{z=1}^{\infty} z^r(1-p)^z = Li_{-r}(1-p)\right],\end{aligned}$$

where $Li_n(x)$ is the polylogarithm function of order n and argument x.
Hence proved. □

In particular, the first four raw moments of the proposed model are obtained by putting r=1, 2, 3, 4.

$$\mu_1' = \pi + (1-\pi)\left[\frac{1}{p}\right]$$

$$\mu_2' = \pi + (1-\pi)\left[\frac{(2-p)}{p^2}\right]$$

$$\mu_3' = \pi + (1-\pi)\left[\frac{(p^2-6p+6)}{p^3}\right]$$

$$\mu_4' = \pi + (1-\pi)\left[\frac{(14p^2 - p^3 - 36p + 24)}{p^4}\right].$$

Therefore, Variance(σ^2) of the proposed model is given as

$$\sigma^2 = \mu_2' - (\mu_1')^2 = \frac{[1 + p^2\pi + 2p\pi^2 - \pi^2 - p(1 + \pi + p\pi^2)]}{p^2}.$$

Index of dispersion

If Z $\sim$ OIZTGD (π, p), then its index of dispersion (γ) is as follows:

$$\gamma = \frac{Variance}{Mean} = \frac{1 + 2p\pi^2 + p^2\pi - \pi^2 - p(1 + \pi + p\pi^2)}{p(p\pi + 1 - \pi)}$$

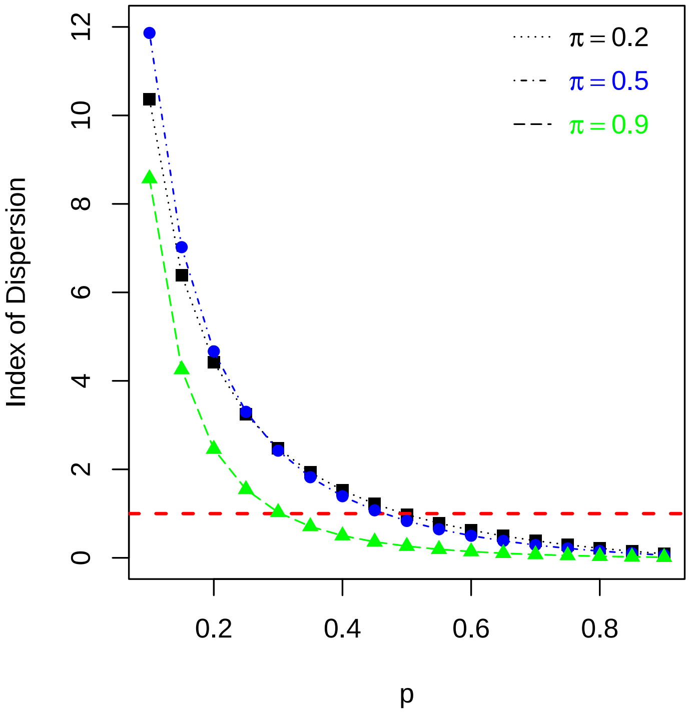

Figure 4: Index of dispersion plot of OIZTGD

As can be seen from figure (4), the proposed model is both over-dispersed as well as under-dispersed. At lower values of π and p, the model is over-dispersed, and for higher values of π and p, the model is under-dispersed. So, we can say that the model shows over and under-dispersion for different combinations of parameters. This is also shown in table 1.

Table 1: Statistical measures of dispersion

π	Over-dispersion	Equi-dispersion	Under-dispersion
0.2	p<0.4948974	p=0.4948974	p>0.4948974
0.5	p<0.4691016	p=0.4691016	p>0.4691016
0.9	p<0.3035678	p=0.3035678	p>0.3035678

Coefficient of skewness

If Z ∼ OIZTGD (π, p), then the Pearson's coefficient of skewness is as follows:

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{[\mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3]^2}{[\mu_2' - \mu_1'^2]^3}$$

$$= \frac{[p^3\pi + (1-\pi)(p^2 - 6p + 6) - 3(p^2\pi + (1-\pi)(2-p))(p\pi + 1 - \pi) + 2(p\pi + 1 - \pi)^3]^2}{[1 + 2p\pi^2 + p^2\pi - \pi^2 - p(1 + \pi + p\pi^2)]^3}$$

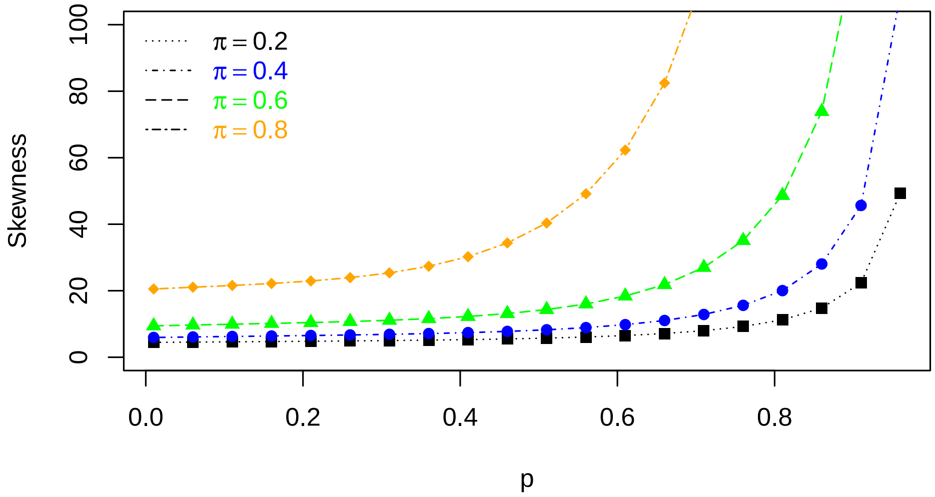

Figure 5: Skewness plot of OIZTGD

From figure (5), it can be observed that frequencies are lowest at the lower values and they rapidly increase as the value increases, which means that the model is negatively skewed and has an inverse j-shaped curve.

Coefficient of kurtosis

If Z ∼ OIZTGD (π, p), then the Pearson's coefficient of kurtosis is as follows:

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{[\mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4]}{[\mu_2' - \mu_1'^2]^2}$$

$$= \frac{\begin{bmatrix} p^4\pi + (1-\pi)(14p^2 - p^3 - 36p + 24) - 4\left((p^3\pi + (1-\pi)(p^2 - 6p + 6)(p\pi + 1 - p))\right) \\ +6\left((p^2\pi + (1-\pi)(2-p))(p\pi + 1 - \pi)^2\right) - 3(p\pi + 1 - \pi)^4 \end{bmatrix}}{[1 + 2p\pi^2 + p^2\pi - \pi^2 - p(1 + \pi + p\pi^2)]^2}$$

Furthermore, from figure (6), it can be observed that the OIZTGD is platykurtic i.e., the proposed model has a flat peak.

4. Parametric estimation

The parameters π and p of equation (3) can be obtained by using the maximum likelihood method of estimation as follows:

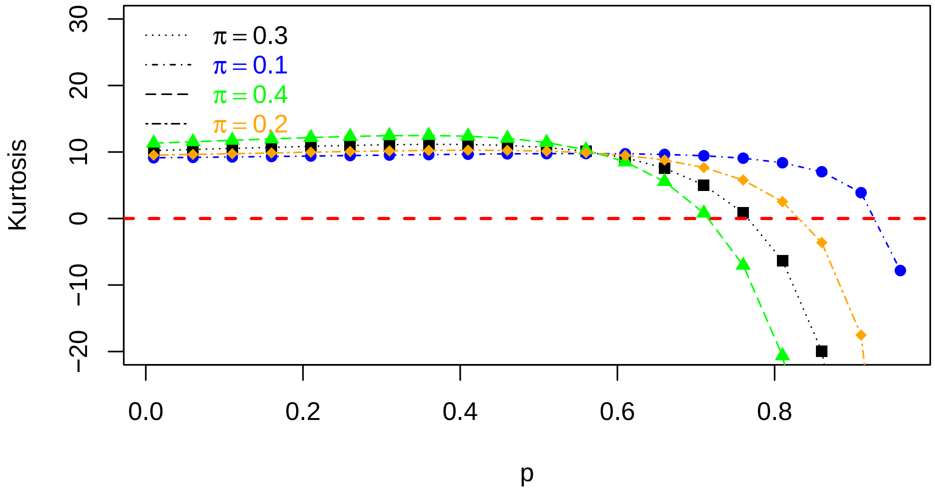

Figure 6: Kurtosis plot of OIZTGD

Let $z_1, z_2, z_3, ..., z_n$ be a random sample from OIZTGD, as given in equation no.(3) and let i=1, 2, ..., n

$$\omega_i = \begin{cases} 1 \quad ; \quad if \quad z = 1 \\ \\ 0 \quad ; \quad otherwise \end{cases}$$

Then, for i=1, 2, 3, .., n, equation no.(3) can be written in the following form

$$P(Z = z_i) = [\pi + (1-\pi)p]^{\omega_i}[(1-\pi)p(1-p)^{z_i-1}]^{1-\omega_i}$$

Hence the likelihood function will be L=L($\pi, p; z_1, z_2, ..., z_n$)

$$L = \prod_{i=1}^{n} [\pi + (1-\pi)p]^{\omega_i} \left[(1-\pi)p(1-p)^{z_i-1}\right]^{1-\omega_i}$$

$$= [\pi + (1-\pi)p]^{n_1} \prod_{\substack{i=1 \\ z_i \neq 1}}^{n} \left[(1-\pi)p(1-p)^{z_i-1}\right]^{c_i},$$

where $c_i = 1 - \omega_i$, $n_1 = \sum_{i=1}^{n} \omega_i$. Note that n_1 represents the number of ones (1s) in the sample. Therefore,

$$lnL = n_1 ln[\pi + (1-\pi)p] + (n-n_1)ln(1-\pi) + (n-n_1)lnp + \sum_{\substack{i=1 \\ z_i \neq 1}}^{n} c_i z_i ln(1-p) - (n-n_1)ln(1-p)$$

$$\frac{\partial lnL}{\partial \pi} = \frac{n_1(1-p)}{[\pi + (1-\pi)p]} - \frac{(n-n_1)}{(1-\pi)} \tag{8}$$

$$\frac{\partial lnL}{\partial p} = \frac{n_1(1-\pi)}{[\pi + (1-\pi)p]} + \frac{(n-n_1)}{p} - \frac{\sum_{\substack{i=1 \\ z_i \neq 1}}^{n} c_i z_i}{(1-p)} + \frac{(n-n_1)}{(1-p)}. \tag{9}$$

Now, let $\frac{\partial lnL}{\partial \pi}$=0. Then from equation (8)

$$\Rightarrow [(n_1 - n_1 p)(1-\pi) - (n-n_1)(\pi + p - \pi p)] = 0$$

$$\Rightarrow \hat{\pi} = \frac{(np - n_1)}{n(p-1)}.$$

Now, let $\frac{\partial lnL}{\partial p}$=0. Then from equation (9)

$$\frac{\partial lnL}{\partial p} = \frac{n_1(1-\pi)}{[\pi + (1-\pi)p]} + \frac{(n-n_1)}{p} - \frac{(\sum_{\substack{i=1 \\ z_i \neq 1}}^{n} c_i z_i - n + n_1)}{(1-p)} = 0$$

$$p^2 \left(n_1\pi - n_1 - \sum_{\substack{i=1 \\ z_i \neq 1}}^{n} c_i z_i + \pi \sum_{\substack{i=1 \\ z_i \neq 1}}^{n} c_i z_i \right) + p \left(n - \pi \sum_{\substack{i=1 \\ z_i \neq 1}}^{n} c_i z_i - n\pi \right) + \pi(n-n_1) = 0. \quad (10)$$

Now, substituting the value of $\hat{\pi}$ in equation (10). Equation (10) reduces to

$$p^2 \left(nn_1 - n_1^2 - n_1 \sum_{\substack{i=1 \\ z_i \neq 1}}^{n} c_i z_i \right) + p \left(n_1 \sum_{\substack{i=1 \\ z_i \neq 1}}^{n} c_i z_i + n_1 - nn_1 \right) + n_1(n_1 - n) = 0 \quad (11)$$

Since equation (11) is in quadratic form. Therefore, the estimated value of the parameter p can be obtained by solving the above quadratic equation, i.e.

$$\hat{p} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a},$$

where

$$b = n_1 \sum_{\substack{i=1 \\ z_i \neq 1}}^{n} c_i z_i + n_1 - nn_1, \quad a = nn_1 - n_1^2 - n_1 \sum_{\substack{i=1 \\ z_i \neq 1}}^{n} c_i z_i, \quad c = n_1(n_1 - n)$$

Therefore,

$$\hat{p} = \frac{-\left(n_1 \sum_{\substack{i=1 \\ z_i \neq 1}}^{n} c_i z_i + n_1 - nn_1 \right) \pm \sqrt{(n_1 \sum_{\substack{i=1 \\ z_i \neq 1}}^{n} c_i z_i + n_1 - nn_1)^2 - 4(nn_1 - n_1^2 - n_1 \sum_{\substack{i=1 \\ z_i \neq 1}}^{n} c_i z_i) n_1(n_1 - n)}}{2\left(nn_1 - n_1^2 - n_1 \sum_{\substack{i=1 \\ z_i \neq 1}}^{n} c_i z_i \right)}$$

$$\hat{p} = \frac{\left(\sum_{\substack{i=1 \\ z_i \neq 1}}^{n} c_i z_i + 1 - n \right) + \sqrt{(\sum_{\substack{i=1 \\ z_i \neq 1}}^{n} c_i z_i)^2 + 1 + 5n^2 - 2n + (n_1 - 3n + 2) \sum_{\substack{i=1 \\ z_i \neq 1}}^{n} c_i z_i - 4n_1(2n + n_1)}}{2\left(n - n_1 - \sum_{\substack{i=1 \\ z_i \neq 1}}^{n} c_i z_i \right)}$$

After solving the above quadratic equation, we get the estimated value of $\hat{p}$.
Therefore, the second-order differentials with respect to π and p can be obtained as

$$\frac{\partial^2 logL}{\partial \pi^2} = -\frac{n_1(1-p)^2}{[\pi + (1-\pi)p]^2} - \frac{(n-n_1)}{(1-\pi)^2}$$

Similarly,

$$\frac{\partial^2 logL}{\partial p^2} = -\frac{n_1(1-\pi)^2}{[\pi+(1-\pi)p]^2} - \frac{(n-n_1)}{p}^2 - \frac{\sum_{\substack{i=1\\ z_i\neq 1}}^{n} c_i z_i}{(1-p)^2} + \frac{(n-n_1)}{(1-p)^2}$$

Similarly,

$$\frac{\partial^2 logL}{\partial \pi \partial p} = -\frac{n_1}{[\pi+(1-\pi)p]} - \frac{n_1(1-\pi)(1-p)}{[\pi+(1-\pi)p]^2}$$

4.1. Simulation

In this section, we carry out a simulation study to investigate the finite sample behaviour of the maximum likelihood estimators for different sample sizes (n=25,75,100,300,600) on various parameter settings through the use of discrete variant of the inverse CDF technique. The procedure was repeated 1000 times in R-software (RCore (2016)) for calculation of Bias, Variance, Mean Square Error (MSE), and Mean Relative Estimate (MRE) and the results are given in Table (2).

$$Bias = \frac{1}{N}\sum_{i=1}^{N}\hat{p} - p \qquad Variance = \frac{1}{N}\sum_{i=1}^{N}\hat{p}^2 - \left(\frac{1}{N}\sum_{i=1}^{N}\hat{p}\right)^2$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(\hat{p}-p)^2 \qquad MRE = \frac{1}{N}\sum_{i=1}^{N}\frac{\hat{p}}{p}$$

here, $\hat{p}$ is the estimate of p and N=1000, is the number of replications. It can be seen from the table 2, that as the sample size increases, the Variance and MSE decreases and are close to zero for large sample sizes. Also, MRE tends to be 1 as the sample size increases. These results suggest that maximum likelihood estimates are consistent and therefore can be used in estimating the unknown parameters of the proposed model.

5. Hypothesis testing

In this part, we have checked the significance of the inflation parameter (π) by likelihood ratio test and Wald's test.

5.1. Likelihood ratio test

In order to test the significance of the inflation parameter π of the OIZTGD, the likelihood ratio test is carried out to distinguish between ZTGD (p) and OIZTGD (π, p). Here, the null hypothesis is

$$H_o : \pi = 0 \; Vs \; the \; alternative \; hypothesis \; H_1 \pi \neq 0$$

In case of likelihood ratio test, the test statistic is given by

$$-2ln\xi = 2(l_1 - l_2), \tag{12}$$

where, $l_1 = lnL(\hat{\omega}; z)$, Where $\hat{\omega}$ is the maximum likelihood estimator for $\omega = (\pi, p)$ without limitation, and $l_2 = lnL(\hat{\omega}^*, z)$, in which $\hat{\omega}^*$ is the maximum likelihood estimator for ω under the null hypothesis H_o.

Table 2: Simulation table of MLE's for proposed model

Sample		$\pi = 0.01, p = 0.02$				$\pi = 0.02, p = 0.1$			
Size(n)	Parameter	Bias	Variance	MSE	MRE	Bias	Variance	MSE	MRE
25	$\hat{\pi}$	0.01456	0.00069	0.00090	2.45569	0.03299	0.00247	0.00356	2.64935
	$\hat{p}$	-0.00034	0.00052	0.00050	0.98277	0.01091	0.00070	0.00082	1.10909
75	$\hat{\pi}$	0.00437	0.00029	0.00031	1.43704	0.00050	0.00039	0.00039	1.02479
	$\hat{p}$	-0.00037	0.00049	0.00050	0.98168	0.00259	0.00011	0.00012	1.02589
100	$\hat{\pi}$	-0.00248	0.00006	0.00007	0.75180	-0.00338	0.00041	0.00043	0.83096
	$\hat{p}$	0.00035	0.00035	0.00041	1.01740	-0.00089	0.00004	0.00004	0.99115
300	$\hat{\pi}$	-0.00048	0.00004	0.00004	0.95192	-0.00645	0.00013	0.00017	0.67741
	$\hat{p}$	0.00009	0.00021	0.00032	1.00441	0.00059	0.00001	0.00001	1.00593
600	$\hat{\pi}$	-0.00024	0.00005	0.00005	0.97585	0.00023	0.00019	0.00019	1.01143
	$\hat{p}$	-0.00008	0.00011	0.00001	0.99614	0.00113	0.00002	0.00002	1.01130
Sample		$\pi = 0.03, p = 0.05$				$\pi = 0.04, p = 0.06$			
Size(n)	Parameter	Bias	Variance	MSE	MRE	Bias	Variance	MSE	MRE
25	$\hat{\pi}$	-0.00171	0.00231	0.00231	0.94312	-0.01318	0.00162	0.00180	0.67038
	$\hat{p}$	0.00131	0.00010	0.00010	1.02621	0.00096	0.00011	0.00011	1.01607
75	$\hat{\pi}$	-0.00650	0.00057	0.00061	0.78330	-0.00419	0.00123	0.00125	0.89530
	$\hat{p}$	0.00283	0.00003	0.00004	1.05661	0.00278	0.00003	0.00004	1.04638
100	$\hat{\pi}$	-0.00800	0.00025	0.00031	0.73349	-0.00796	0.00027	0.00033	0.80090
	$\hat{p}$	-0.00019	0.00002	0.00002	0.99620	0.00138	0.00002	0.00003	1.02296
300	$\hat{\pi}$	0.00140	0.00027	0.00028	1.04683	-0.00153	0.00030	0.00030	0.96174
	$\hat{p}$	0.00042	0.00001	0.00001	1.00848	0.00024	0.00001	0.00001	1.00400
600	$\hat{\pi}$	0.00052	0.00008	0.00008	1.01717	-0.00077	0.00026	0.00026	0.98084
	$\hat{p}$	0.00013	0.00001	0.00001	1.00265	0.00050	0.00001	0.00001	1.00834
Sample		$\pi = 0.04, p = 0.1$				$\pi = 0.4, p = 0.4$			
Size(n)	Parameter	Bias	Variance	MSE	MRE	Bias	Variance	MSE	MRE
25	$\hat{\pi}$	0.01541	0.00240	0.00264	1.38527	0.02866	0.06687	0.06769	1.07166
	$\hat{p}$	0.00061	0.00016	0.00016	1.00614	0.03359	0.01142	0.01255	1.08397
75	$\hat{\pi}$	-0.00240	0.00106	0.00106	0.93989	-0.05451	0.00961	0.01258	0.86372
	$\hat{p}$	0.00211	0.00015	0.00016	1.02113	0.00614	0.00250	0.00254	1.01535
100	$\hat{\pi}$	0.00570	0.00101	0.00104	1.14242	-0.01553	0.00529	0.00553	0.96116
	$\hat{p}$	-0.00179	0.00004	0.00004	0.98207	0.00392	0.00143	0.00145	1.00980
300	$\hat{\pi}$	-0.00811	0.00034	0.00040	0.79735	0.00842	0.00166	0.00173	1.02106
	$\hat{p}$	-0.00089	0.00002	0.00002	0.99112	-0.00535	0.00071	0.00074	0.98662
600	$\hat{\pi}$	0.00432	0.00022	0.00024	1.10797	0.00774	0.00092	0.00098	1.01936
	$\hat{p}$	-0.00056	0.00002	0.00002	0.99442	-0.00427	0.00030	0.00032	0.98930

Table 3: Calculated value of test statistic in case of likelihood ratio test

	Test statistic	Bootstrap-p value
Dataset 1	20.88	<0.0001
Dataset 2	27.84	<0.0001

5.2. Wald's test

Here for testing the significance of the inflation parameter π of OIZTGD, we assess Wald's test. To test the null hypothesis

$$H_0 : \pi = 0 \; Vs \; the \; alternative \; hypothesis \; H_1 : \pi \neq 0$$

In case of Wald's test, the test statistic is given by

$$W_\pi = \frac{\hat{\pi}^2}{Var(\hat{\pi})}, \tag{13}$$

Where $Var(\hat{\pi})$ represents the diagonal element of Fisher information matrix at $\pi = \hat{\pi}$ and $p = \hat{p}$.

Table 4: Calculated value of test statistic in case of wald's test

	Test statistic	Bootstrap-p value
Dataset 1	34.74	<0.0001
Dataset 2	72.38	<0.0001

5.3. Parametric bootstraping

In this part, we have written the parametric bootstrap procedure to obtain bootstrap p-value in case of likelihood ratio and wald's test, as both the tests are asymptotic tests. Following is the algorithm of parametric bootstrap procedure to evaluate the p-value of the test statistics:

(i) Based on the original data of size n, say $x_1, x_2, \cdots, x_n$, compute the value of the test statistic, say $S^{obs} = S(x_1, \cdots, x_n)$, and estimate $\hat{p}_n$.

(ii) Generate $j = 1, ..., B$ bootstrap samples $X_1^j, \cdots, X_n^j$ by independently sampling from a truncated estimated $TG(\hat{p}_n)$.

(iii) On the basis of each bootstrap sample compute the observed value of the test statistic, $S^{j,boot} = S(X_1^j, \cdots, X_n^j)$, $j = 1, ..., B$.

(iv) Then the $\hat{p}_{boot}$-value is given by

$$\hat{p}_{boot} = \frac{1}{B}\sum_{j=1}^{B} \mathbb{I}\left(S^{j,boot} > S^{obs}\right)$$

In the above algorithm, the test statistics S^{obs} considered here are $-2log\xi$ and W_π for the likelihood ratio test and Wald's test respectively. We compute test statistics values for both datasets and the parametric bootstrap p_{boot}-value is also given in the above tables. In both the datasets for both the tests, the p_{boot}- value is very low as compared to significance level 0.05. Hence we reject the null hypothesis that the data comes from Zero-truncated Geometric distribution.

6. Applications

In this part, we study the practical significance of one-inflated zero-truncated Geometric distribution. Two real-life datasets are taken to compare OIZTGD with few other distributions like zero-truncated Geometric distribution (ZTGD), zero-truncated Negative Binomial distribution (ZTNBD), zero-truncated Poisson Lindly distribution (ZTPLD), zero-truncated Discrete Lindly distribution (ZTDLD), zero-truncated two parameter discrete Lindly distribution (ZTTPDLD), zero-truncated Poisson distribution (ZTPD), one-inflated positive Poisson distribution (OIPPD), and one-inflated zero-truncated Negative Binomial distribution (OIZTNBD) to check the performance measures of the proposed model. R-software was used to perform all the computations, and the *fitdistr plus* command in R-software was used to estimate the parameters of the distribution.

Here, we consider two datasets, the first dataset shown in Table (6) has been taken from Williams (1943) and has been recently used by Hassan and Ahmad (2009). The dataset is related to the number of publications in the review of applied entomology. The second dataset shown in Table (8) has been taken from Heijden *et al.* (Van Der Heijden, Cruyff, and Van Houwelingen (2003)) and has been recently used by Tajuddin *et al.* (2021). The dataset is related to the frequency of a person being arrested for drunk driving and it reveals the details of five selected police regions in Dutch among 25 police regions for analyses. These two datasets are used to demonstrate and analyze the performance of various matrices. We have fitted OIZTGD, ZTGD, ZTNBD, ZTPLD, ZTDLD, ZTTPDLD, ZTPD, OIPPD, and OIZTNBD to both datasets for comparison. To check the suitability of the model, we have computed expected frequencies, the values of maximized log-likelihood, χ^2 statistic along with associated p-value, Akaike's Information Criteria (AIC), and Bayesian Information Criteria (BIC) shown in Tables (7) and (9).

From Table (7) and (9), it has been revealed that OIZTGD provides the best fit compared to existing models – ZTGD, ZTNBD, ZTPLD, ZTDLD, ZTTPDLD, ZTPD, OIPPD, and OIZTNBD as our proposed model has lowest AIC and BIC values, and also has the highest p-value in both datasets. Since OIZTNBD has no p-value because, all degrees of freedom got lost in the dataset related to the number of persons arrested for drunk driving.

Here, we have also adopted the likelihood ratio test and wald's test for testing the significance of the inflation parameter π. In the case of likelihood ratio test, the computed value for both datasets shown in Table (3) are 20.88 and 27.84. In the case of wald's test, the computed value for both datasets shown in Table (4) are 34.74 and 27.84. The null hypothesis is rejected in all cases of likelihood ratio test and wald's test as the value for both the datasets is greater than $\hat{p}_{boot}$-value. Hence, we conclude that π, the additional parameter in the model is significant, as discussed in Section 5.

Here, we have also plotted the estimated pmfs for both datasets. From the figures (7) and (8), it can be seen that the OIZTGD yields better fit to both datasets compared to all the existing models considered in the paper. This also supports the suitability of the proposed model to the given datasets.

Table 5: Descriptive features of both datasets

	Mean	Variance	Coefficient of variation	Index of Dispersion
Dataset 1	1.5961	1.5740	0.7860	0.9862
Dataset 2	1.0650	0.0782	0.2626	0.0734

Table 6: Dataset related to the number of Publications in the Review of Applied Entomology by Williams (1943)

Claims	1	2	3	4	5	6	7	8	9	10
frequency	285	70	32	10	4	3	3	1	2	1

Table 7: Expected frequencies and χ^2 values for fitted models along with AIC and BIC

Claims	Observed Count	**OIZTGD**	ZTGD	ZTNBD	ZTPLD	ZTDLD	ZTTPDLD	ZTPD	OIPPD	OIZTNBD
1	285	285	258	257	256	251	258	236	285	285
2	70	65	96	97	98	103	96	121	88	65
3	32	31	36	36	36	38	36	41	29	32
4	10	15	13	13	13	13	13	10	7	15
5	4	7	5	5	5	4	5	2	1	7
6	3	4	2	2	2	1	2	0	0	4
7	3	2	1	1	1	0	1	0	0	2
8	1	1	0	0	0	0	0	0	0	1
9	2	1	0	0	0	0	0	0	0	0
10	1	0	0	0	0	0	0	0	0	0
Degrees of Freedom		**3**	3	2	2	3	3	2	1	2
ML Estimates		$\hat{\pi}$**=0.3688**	$\hat{\theta}$=0.6265	$\hat{p}$=0.3590	$\hat{\theta}$=2.1333	$\hat{\theta}$=1.2945	$\hat{p}$=0.3731	$\hat{\theta}$=1.0213	$\hat{\alpha}$=0.2660	$\hat{\alpha}$=0.3797
		$\hat{p}$**=0.5142**		$\hat{r}$=1.1110			$\hat{\beta}$=0.0010		$\hat{\theta}$=0.9990	$\hat{r}$=1.1112
										$\hat{p}$=0.4735
χ^2-value		**4.65**	15.50	17.13	16.92	33.01	11.00	54.11	35.99	6.05
$p-value$		**0.198**	0.001	<0.001	<0.001	<0.001	0.004	<0.001	<0.001	0.04
$-\hat{logl}$		**423.03**	433.47	434.50	435.71	441.36	433.47	475.83	465.10	423.21
AIC		**850.06**	868.94	873.00	873.42	884.73	870.94	953.67	934.20	852.43
BIC		**858.10**	872.96	881.04	877.44	884.73	878.98	957.69	942.23	864.48

Table 8: Dataset related to number of persons arrested for drunk driving by Van Der Heijden *et al.* (2003)

Claims	1	2	3	4	5
Observed	8877	481	51	8	1

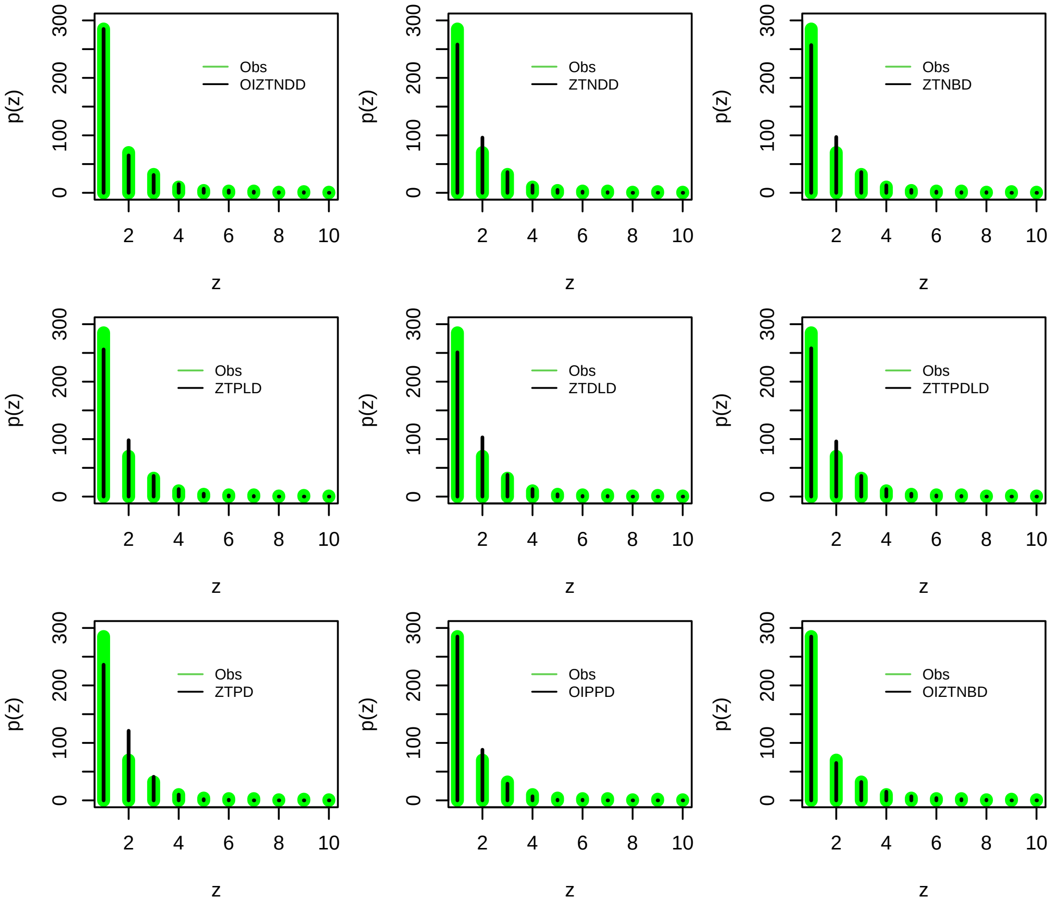

Figure 7: Fitted frequency plots of OIZTGD corresponding to dataset 1

Table 9: Expected frequencies and χ^2 values for fitted models along with AIC and BIC

Claims	Observed Count	**OIZTGD**	ZTGD	ZTNBD	ZTPLD	ZTDLD	ZTTPDLD	ZTPD	OIPPD	OIZTNBD
1	8877	8877	8843	8831	8843	8844	8840	8843	8877	8877
2	481	479	540	563	542	540	548	540	477	479
3	52	56	33	24	33	33	30	33	59	56
4	8	6	2	1	2	2	2	2	5	6
5	1	1	0	0	0	0	0	0	0	1
Degrees of freedom		**1**	1	0	1	1	0	1	1	-
ML Estimates		$\hat{\pi}$**=0.5032**	$\hat{\theta}$=0.9388	$\hat{p}$=0.0580	$\hat{\theta}$=16.2159	$\hat{\theta}$=3.18677	$\hat{p}$=0.0610	$\hat{\theta}$=0.1274	$\hat{\alpha}$=0.6682	$\hat{\alpha}$=0.5119
		$\hat{p}$**=0.8841**		$\hat{r}$=1.1110			$\hat{\beta}$= 0.0010		$\hat{\theta}$=0.3696	$\hat{r}$=1.1110
										$\hat{p}$=0.1119
χ^2-value		**0.86**	42.01	26.31	25.88	34.62	25.89	64.02	3.51	0.8654
$p - value$		**0.352**	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.060	-
$-\hat{logl}$		**2293.39**	2307.31	2307.99	2307.41	2312.05	2307.31	2326.18	2294.26	2293.42
AIC		**4590.79**	4616.63	4619.98	4616.83	4626.11	4618.63	4654.37	4592.52	4592.83
BIC		**4605.10**	4623.78	4634.28	4623.98	4633.26	4632.93	4661.52	4606.82	4614.28

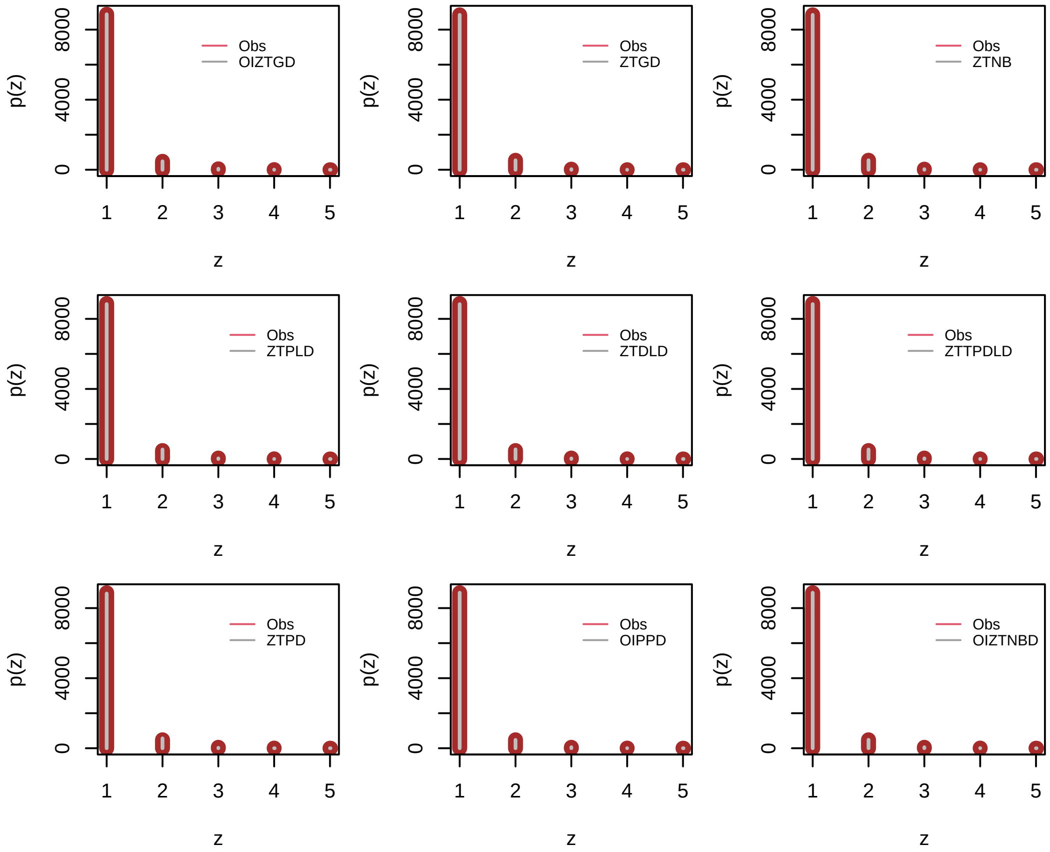

Figure 8: Fitted frequency plots of OIZTGD corresponding to dataset 2

7. Conclusions

A new one-inflated version of truncated distribution is proposed in this paper namely one-inflated zero-truncated Geometric distribution (OIZTGD). Key statistical properties of the distribution including generating functions, reliability characteristics, and moments have been derived. For parametric estimation purpose, the maximum likelihood method of estimation has been used. A simulation study has been done to evaluate the proficiency of the estimation measures considered in this paper. Further, the procedure of the likelihood ratio test and wald's test are carried out to test the significance of the inflation parameter. Two real-life datasets are reviewed to demonstrate the practicality of the proposed model juxtaposed to the existent models ZTGD, ZTNBD, ZTPLD, ZTDLD, ZTTPDLD, ZTPD, OIPPD, and OIZTNBD. We can see that OIZTGD in terms of Chi-square value and p-value gives the best fit as the existent models do not show the best fit. The information measures like Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC) have lowest values among all other competing distributions in terms of numerical value, revealing that OIZTGD can be considered as a suitable model in comparison to other models discussed in this paper.

References

David FN, Johnson NL (1952). "The Truncated Poisson." *Biometrics*, **8**(4), 275–285.

Devi B, Gupta R, Kumar P (2017). "Sequential Analysis of Zero Truncated Geometric Distribution." *Journal of Statistics Applications & Probability*, **6**(2), 385–389.

Elah N, Ahmad PB, Wani MA (2023). "A New Zero-truncated Distribution and Its Applications to Count Data." *Reliability: Theory & Applications*, **18**, 327–339.

Ghitany ME, Al-Mutairi DK, Nadarajah S (2008). "Zero-truncated Poisson–Lindley Distribution and Its Application." *Mathematics and Computers in Simulation*, **79**(3), 279–287.

Godwin RT (2017). "One-inflation and Unobserved Heterogeneity in Population Size Estimation." *Biometrical Journal*, **59**(1), 79–93.

Godwin RT (2019). "The One-inflated Positive Poisson Mixture Model for Use in Population Size Estimation." *Biometrical Journal*, **61**(6), 1541–1556.

Godwin RT, Böhning D (2017). "Estimation of the Population Size by Using the One-inflated Positive Poisson Model." *Journal of the Royal Statistical Society Series C: Applied Statistics*, **66**(2), 425–448.

Gómez-Déniz E (2010). "Another Generalization of the Geometric Distribution." *Test*, **19**, 399–415.

Hassan A, Ahmad PB (2009). "Misclassification in Size-biased Modified Power Series Distribution and Its Applications." *Journal of the Korean Society for Industrial and Applied Mathematics*, **13**(1), 55–72.

Kaskasamkul P, Bohning D (2017). "Population Size Estimation for One-inflated Count Data Based upon the Geometric Distribution." In *Capture-recapture Methods for the Social and Medical Sciences*, pp. 191–209.

Kiani TH (2020). "A Zero Truncated Discrete Distribution: Theory and Applications to Count Data." *Pakistan Journal of Statistics and Operation Research*, pp. 167–190.

Park C, Gou K, Wang M (2022). "A Study on Estimating the Parameter of the Truncated Geometric Distribution." *The American Statistician*, **76**(3), 257–261.

Phang YN, Loh EF (2013). "Zero Truncated Strict Arcsine Model." *International Journal of Computer and Information Engineering*, **7**(7), 989–991.

RCore T (2016). "R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria."

Sampford MR (1955). "The Truncated Negative Binomial Distribution." *Biometrika*, **42**(1/2), 58–69.

Shanker R, Fesshaye H (2016). "On Zero-truncation of Poisson, Poisson–Lindley and Poisson–Sujatha Distributions and Their Applications." *Biometrics & Biostatistics International Journal*, **3**(5), 1–13.

Shibu DS, Chesneau C, Monisha M, Maya R, Irshad MR (2023). "A Novel Zero-truncated Katz Distribution by the Lagrange Expansion of the Second Kind with Associated Inferences." *Analytics*, **2**(2), 463–484.

Skinder Z, Ahmad PB, Elah N (2023). "A New Zero-inflated Count Model with Applications in Medical Sciences." *Reliability: Theory & Applications*, **18**(3 (74)), 841–855.

Tajuddin RRM, Ismail N, Ibrahim K (2021). "Comparison of Estimation Methods for One-inflated Positive Poisson Distribution." *Journal of Taibah University for Science*, **15**(1), 869–881.

Tajuddin RRM, Ismail N, Ibrahim K (2022). "Estimating Population Size of Criminals: A New Horvitz–Thompson Estimator under One-Inflated Positive Poisson–Lindley Model." *Crime & Delinquency*, **68**(6-7), 1004–1034.

Van Der Heijden PGM, Cruyff M, Van Houwelingen HC (2003). "Estimating the Size of a Criminal Population from Police Records Using the Truncated Poisson Regression Model." *Statistica Neerlandica*, **57**(3), 289–304.

Wani MK, Ahmad PB (2023). "Zero-inflated Poisson-Akash Distribution for Count Data with Excessive Zeros." *Journal of the Korean Statistical Society*, pp. 1–29.

Williams CB (1943). "The Numbers of Publications Written by Biologists." *Annals of Eugenics*, **12**(1), 143–146.

Affiliation:

Peer Bilal Ahmad
Department of Mathematical Sciences
Islamic University of Science and Technology
Kashmir, India
E-mail: bilalahmadpz@gmail.com
URL: https://www.iust.ac.in/faculty-details.aspx?deptcode=DOM&empId=1345

Austrian Journal of Statistics http://www.ajs.or.at/
published by the Austrian Society of Statistics http://www.osg.or.at/
Volume 53 *Submitted:* 2023-10-20
2024 *Accepted:* 2024-04-03