

ON MODELING ZERO-TRUNCATED COUNT DATA

*Samuel Adewale Aderoju¹, Emmanuel Teju Jolayemi²

^{*1}Department of Statistics and Mathematical Sciences, Kwara State University, Malete,
P.M.B. 1530, Ilorin, Kwara State, Nigeria.
samuel.aderoju@kwasu.edu.ng¹

²Department of Statistics, University of Ilorin, P.M.B. 1515, Ilorin, Nigeria.
tejujola@unilorin.edu.ng²

Abstract : In this paper, a more generalized zero-truncated distribution that come from the mixture of distributions was developed and implemented for non-zero count data. The application of the distribution was demonstrated and its performance assessed against some existing ones using real life datasets. Following the idea of mixed distributions, the Zero-Truncated Com-Binomial (ZTCB) is from the mixture of Conway-Maxwell-Poisson type generalization to the Binomial distribution. The first two moments via probability generating function were also derived. The Maximum Likelihood Estimations of the parameters were also obtained by direct maximization of the log-likelihood function using “optim” routine in R software. The findings of the study showed that: the ZTCB distribution is more robust to handle all levels of dispersion than Zero-Truncated Multiplicative Binomial distribution. The statistics (chi square goodness-of-fit) as well as Deviance (in example four) and the AIC show that the proposed ZTCB distribution yields best result among the models under consideration. This paper therefore provides useful alternative to the existing zero-truncated distributions.

Keywords - Zero-Truncated, Com-binomial, Maximum likelihood, Multiplicative Binomial, structurally non-zero.

1. INTRODUCTION

The Binomial and Poisson distributions are the popular base distributions to describe a binary or non-binary count, data type. The Poisson and Binomial models have been applied in many disciplines, including medicine, Economy, epidemiology, Biology and Demography.

Compoisson distribution was originally developed in 1962 to model Queuing process by [1] the distribution was later revisited by [2]. [2] derived many of the basic properties of the distribution. [3] compared the efficiency of quasi-likelihood and Maximum Likelihood Estimate estimation approaches for estimating the parameters of a single-link Compoisson based on simulated data sets. [4] developed a Maximum Likelihood Estimate (MLE) for a single link GLM based on the original Compoisson distribution. The major advantage of the Compoisson distribution is its ability to handle both under-dispersion and over-dispersion within a single conditional distribution. This is an alternative to the restricted generalized Poisson (RGP) distribution by [5].

Nevertheless, when the data of interest is structurally zero-truncated, the distributions must be adjusted to account for the missing zeros. A typical example is a study of length of hospital stay (for patients admitted into the hospital). Length of hospital stay is recorded as a minimum of at least one day [6].

The rest of this paper is organized as follows. In section two, we present the ZTCB distribution and describes its properties and the maximum likelihood method for estimating the parameters in section three. In section four, we apply the ZTCB distribution to two real datasets and show that this model provides an excellent fit to these datasets. Finally, some concluding remarks are made in section five.

2. ZERO-TRUNCATED COM-BINOMIAL (ZTCB) DISTRIBUTION

The Compoisson-binomial (CMPB) Distribution: The probability mass function (pmf) of the Compoisson-binomial (CMPB) is defined as, [2];

$$f(Y = y; n, \pi, v) = \frac{\binom{n}{y}^v \pi^y (1 - \pi)^{n-y}}{\sum_{i=0}^n \binom{n}{i}^v \pi^i (1 - \pi)^{n-i}}, \quad y = 0, 1, \dots, n \quad 1$$

for $n \in \mathbb{Z}^+$, $\pi \in (0,1)$ and $v \in \mathbb{R}$

where, Z^+ is a set of unknown non – negative integers

NOTE:

when $v = 1$, we have the Binomial distribution,

$v > 1$, we have under – dispersion,

$v < 1$, we have over – dispersion with respect to the Binomial distribution

Note that (according to [7]):

- The CMPB distribution can be interpreted as a sum of equi-correlated Bernoulli variables.
- The Compoisson distribution ([1]) is approximation to the CMPB distribution when n is large. (for details, see [7])

2.1 The ZTCB Distribution

The pmf of the CMPB distribution as defined above is

$$f(Y = y; n, \lambda, v) = \frac{\binom{n}{y}^v \pi^y (1 - \pi)^{n-y}}{\sum_{i=0}^n \binom{n}{i}^v \pi^i (1 - \pi)^{n-i}} \quad y = 0, 1, 2, 3 \dots, n \quad 2$$

However, the Zero-truncated version of the distribution, which we here referred to as Zero-truncated Com-binomial (ZTCB) can be derived as, ([8]);

$$P(y; n, \pi, v) = \frac{f(Y = y; n, \pi, v)}{1 - f(Y = 0; n, \pi, v)} \quad y = 1, 2, 3 \dots, n \quad 3$$

Where, $f(Y = 0; n, \pi, v) = \frac{(1-\pi)^n}{\sum_{i=0}^n \binom{n}{i}^v \pi^i (1-\pi)^{n-i}}$ and $f(Y = y; n, \pi, v)$ is the CMPB defined above.

Therefore, the pmf of the ZTCB distribution is derived as

$$\begin{aligned} P(y; n, \pi, v) &= \frac{\binom{n}{y}^v \pi^y (1 - \pi)^{n-y}}{\sum_{i=0}^n \binom{n}{i}^v \pi^i (1 - \pi)^{n-i}} \div \left\{ 1 - \frac{(1 - \pi)^n}{\sum_{i=0}^n \binom{n}{i}^v \pi^i (1 - \pi)^{n-i}} \right\} \\ &= \frac{\binom{n}{y}^v \pi^y (1 - \pi)^{n-y}}{\sum_{i=0}^n \binom{n}{i}^v \pi^i (1 - \pi)^{n-i}} * \left[\frac{\sum_{i=0}^n \binom{n}{i}^v \pi^i (1 - \pi)^{n-i}}{\sum_{i=0}^n \binom{n}{i}^v \pi^i (1 - \pi)^{n-i} - (1 - \pi)^n} \right] \\ &= \frac{\binom{n}{y}^v \pi^y (1 - \pi)^{n-y}}{\sum_{i=0}^n \binom{n}{i}^v \pi^i (1 - \pi)^{n-i} - (1 - \pi)^n}, \quad y = 1, 2, 3 \dots, n \quad 4 \end{aligned}$$

Hence, the pmf of the ZTCB distribution can be re-written as

$$P(y; n, \pi, v) = \frac{\binom{n}{y}^v \pi^y (1 - \pi)^{n-y}}{\sum_{i=1}^n \binom{n}{i}^v \pi^i (1 - \pi)^{n-i}}, \quad y = 1, 2, 3 \dots, n \quad 5$$

3. PROPERTIES OF ZTCB DISTRIBUTION

From the pmf of the ZTCB distribution defined above, when $v = 1$, the expected value of the random variable Y , $E(Y) = \frac{n\pi}{1-(1-\pi)^n}$.

The probability generating function of the ZTCB can be expressed as:

$$E(t^y) = \sum_{y=1}^n t^y \frac{\binom{n}{y}^v \pi^y (1-\pi)^{n-y}}{\sum_{j=1}^n \binom{n}{j}^v \pi^j (1-\pi)^{n-j}} \quad 6$$

Hence the first and second moment can be expressed respectively as:

$$\mu'_1 = \sum_{y=1}^n \frac{y \binom{n}{y}^v \pi^y (1-\pi)^{n-y}}{\sum_{j=1}^n \binom{n}{j}^v \pi^j (1-\pi)^{n-j}} \quad 7$$

and

$$\mu'_2 = \sum_{y=1}^n \frac{y(y-1) \binom{n}{y}^v \pi^y (1-\pi)^{n-y}}{\sum_{j=1}^n \binom{n}{j}^v \pi^j (1-\pi)^{n-j}} \quad 8$$

From the pmf of the ZTCB distribution defined above, when $v = 1$, the expected value of the random variable Y ,

$$E(Y) = \frac{n\pi}{1-(1-\pi)^n}$$

However, when $v > 1$, suppose $p = v$ and $q = v - 1$, then the first two moment are derived respectively as:

$$\mu'_1 = \frac{n^p \pi T}{(\pi - 1) \left(1 - pFq \left[\{-n_1, -n_2, \dots, -n_p\}, \{1_1, 1_2, \dots, 1_q\}, \frac{\pi}{\pi - 1} \right] \right)}$$

and

$$\mu'_2 = \frac{n^p \pi (T - S)}{(\pi - 1) \left(1 - pFq \left[\{-n_1, -n_2, \dots, -n_p\}, \{1_1, 1_2, \dots, 1_q\}, \frac{\pi}{\pi - 1} \right] \right)}$$

where

$$T = pFq \left[\{(1-n)_1, (1-n)_2, \dots, (1-n)_p\}, \{2_1, 2_2, \dots, 2_q\}, \frac{\pi}{\pi - 1} \right]$$

$$S = pFq \left[\{(1-n)_1, (1-n)_2, \dots, (1-n)_p\}, \{1, 2_1, 2_2, \dots, 2_{q-1}\}, \frac{\pi}{\pi - 1} \right]$$

These is another contribution of this paper.

Figures below present the pmf of the ZTCB distribution for $n=7$ at different values of π (as p) and v . For $v \rightarrow \infty$, the pmf is peak around the point $n\pi$ and for $v \rightarrow -\infty$, it peaks around 1 or n .

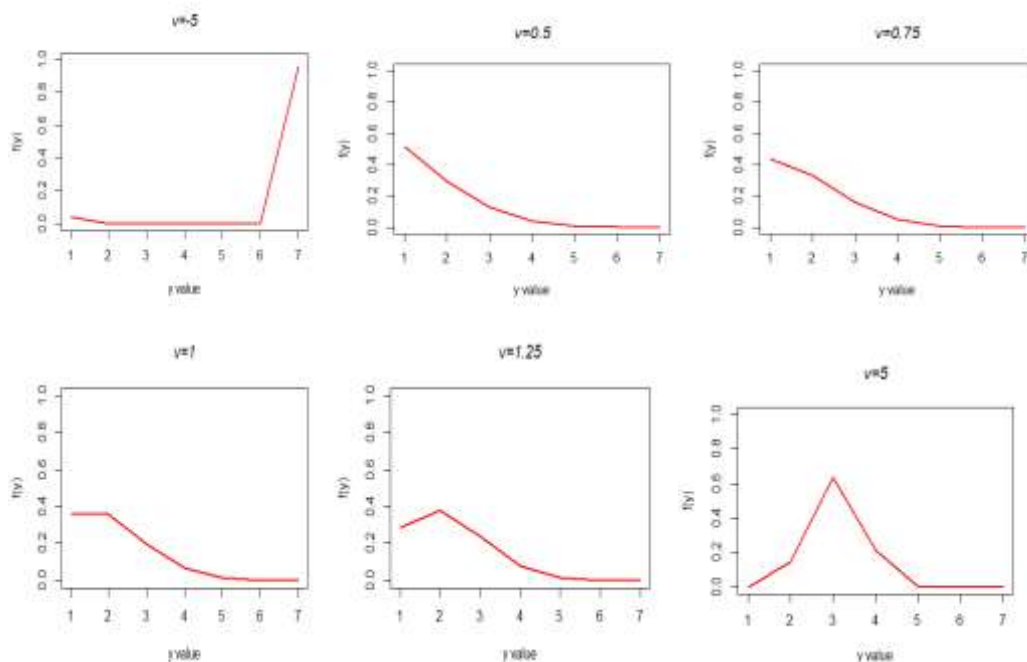


Fig.1: The pmf of ZTCB distribution with $\pi = p = 0.25$ and at various values of v .

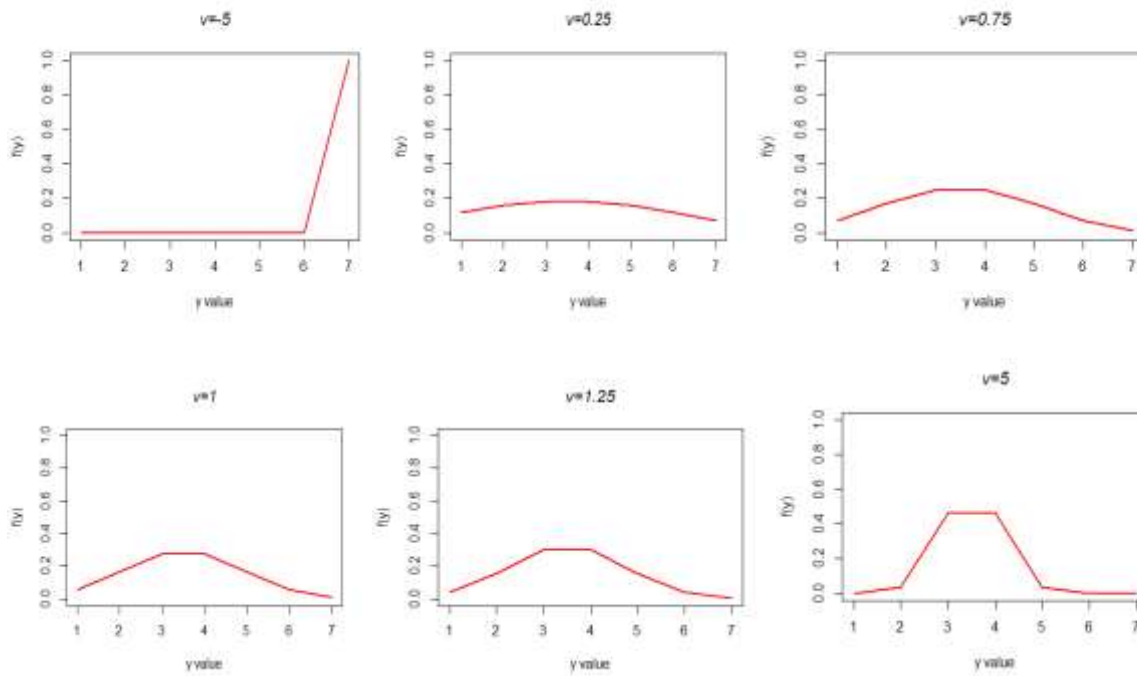


Fig.2: The pmf of ZTCB distribution with $\pi = p = 0.5$ and at various values of v .

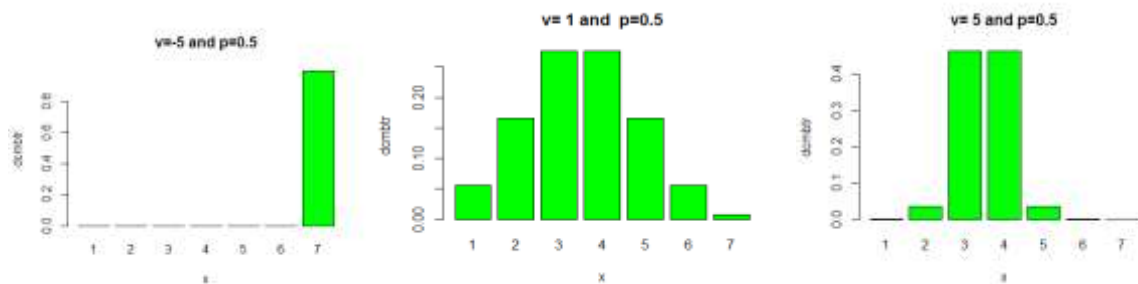


Fig.3: Graph of probability function of ZTCB distribution for different values of the parameters ($v = -5, 1, 5$ vs $\pi = p = 0.5$).

Remark: The Zero-truncated Com-Poisson (ZTCP) distribution is an approximation of the ZTCB when n very large

$$\lim_{n \rightarrow \infty} P(y; n, \pi, v) = \lim_{n \rightarrow \infty} \frac{\binom{n}{y}^v \left(\frac{\lambda}{n}\right)^{y_i} \left(1 - \frac{\lambda}{n}\right)^{n-y_i}}{\sum_{j=1}^n \binom{n}{j}^v \frac{\lambda^j}{n} \left(1 - \frac{\lambda}{n}\right)^{n-j}}$$

$$= \frac{\lambda^y}{(y!)^v \sum_{k=1}^{\infty} \frac{\lambda^k}{(k!)^v}}, \quad y = 1, 2, \dots \quad 9$$

which is the pmf of the ZTCP distribution.

when $v = 1$, $\sum_{k=1}^{\infty} \frac{\lambda^k}{(k!)^v} = e^{\lambda} - 1$

Therefore, equation 8 above becomes

$$P(y | \lambda) = \frac{\lambda^y}{y! (e^{\lambda} - 1)}, \quad y = 1, 2, \dots \quad 10$$

which is the pmf of the Zero-truncated Poisson (ZTP) distribution.

Hence, the ZTCB distribution is the generalization of the ZTCP, ZTP and ZTB distributions.

3.1 MAXIMUM LIKELIHOOD ESTIMATION OF THE PARAMETERS

The likelihood function of the ZTCB is

$$L(\pi, v|y_i) = \prod_{i=1}^n \frac{\binom{n}{y_i}^v \pi^{y_i} (1-\pi)^{n-y_i}}{\sum_{j=1}^n \binom{n}{j}^v \pi^j (1-\pi)^{n-j}} \quad 11$$

While the log-likelihood function is

$$\mathcal{L} = \sum_{i=1}^n \left\{ v \log \binom{n}{y_i} + y_i \log \pi + (n - y_i) \log(1 - \pi) - \log Q \right\} \quad 12$$

where $Q = \sum_{j=1}^n \binom{n}{j}^v \pi^j (1 - \pi)^{n-j}$

We are to find the first and second partial derivative of equation (11) with respect to each parameter and equate them to zero as:

$$\frac{\partial \mathcal{L}}{\partial \pi} = 0, \quad \frac{\partial^2 \mathcal{L}}{\partial \pi^2} = 0 \quad \text{and} \quad \frac{\partial \mathcal{L}}{\partial v} = 0, \quad \frac{\partial^2 \mathcal{L}}{\partial v^2} = 0$$

However, the equations do not have closed form. Therefore, the maximum likelihood estimates (MLEs) of ZTCB cannot be solved analytically, an iterative methods such as

- Fisher Score Algorithm, or
- Newton-Raphson (NR) iterative method, as implemented by [9],
- Iteratively reweighted least squares (IRWLS) and so on, can be used.

We obtained the MLEs of the parameters by direct maximization of the log-likelihood function using “*optim*” routine of R software[10] with “*L-BFGS-B*” method. This can as well be done by using **PROC NLMIXED** in SAS.

4. Application of Zero-Truncated Com-Binomial Distribution to Real Life Datasets

Some examples were given below to explain the performance of the ZTCB distribution in comparison with some of the existing distributions.

Example 1: [Mortality data]

Data from [11]: counts of the number of mothers of the rural area having at least one live birth and one neonatal death. The sample mean and variance are 1.278 and 0.3546, respectively.

Table 1: The expected and observed number of mothers of the rural area having at least one live birth and one neonatal death

X	Obs.Freq	ZTP	ZTB	ZTMB	ZTCB
1	409	399.70	394.76	409.27	409.05
2	88	102.30	110.62	87.19	87.69
3	19	17.50	15.50	19.78	19.58
4	5	2.20	1.09	4.78	4.56
5	1	0.30	0.03	0.98	1.12
Total	522	522	522	522	522
MLE		$\theta = 0.5120$	$\pi = 0.1229$	$\beta = 0.1849$ $k = 0.6853$	$\beta = 0.1825$ $v = -0.0589$
Loglik		352.11	358.25	348.92	348.93
χ^2		3.4653	9.3643	0.0492	0.0363
(P-value)		(0.06)	(<0.01)	(0.82)	(0.85)
df		1	1	1	1
AIC		706.21	718.50	701.84	701.86

The data in example two, [11] applied Zero-truncated Poisson (ZTP) which they observed fit the data slightly. The ZTMB distribution performed better than ZTP distribution. However, our proposed ZTCB distribution outperforms ZTMB distribution.

Example 2: [Mortality data]

The data is taking from [11]: counts of the number of literate mothers with at least one live birth by the number of infant deaths. The sample mean and variance are 1.293 and 0.4185, respectively.

Table 2: The expected and observed number of literate mothers with at least one live birth by the number of infant deaths

X	Obs.Freq	ZTP	ZTB	ZTMB	ZTCB
1	683	658.97	649.90	686.15	685.61
2	145	177.40	192.44	136.74	138.05
3	29	31.84	28.49	34.89	34.40
4	11	4.29	2.11	11.40	10.61
5	5	0.50	0.06	3.82	4.33
Total	873	873	873	873	873

MLE	$\theta = 0.5384$	$\pi = 0.1289$	$\beta = 0.2033$ $k = 0.6249$	$\beta = 0.1994$ $v = -0.3073$
Loglik	615.397	634.09	601.55	601.28
χ^2	8.7063	20.0876	1.5477	1.2827
(P-value)	(<0.01)	(<0.01)	(0.21)	(0.26)
df	1	1	1	1
AIC	1232.79	1270.17	1207.11	1206.56

Similarly, for the data in example two, Zero-truncated Poisson (ZTP) and Zero-truncated Binomial (ZTB) distributions failed to fit the data. The ZTCB distribution performed better than ZTMB distribution, though both fit the data well.

Example 3: [immunogold assay data]

The data is taking from [12], who gave counts of sites with 1, 2, 3, 4 and 5 particles from immunogold assay data. The sample mean and variance are 1.576 and 0.7897, respectively.

Table 3: The results of the ZTRCMPB distribution and other distributions for example 1.

X	Obs.Freq	ZTP	ZTB	ZTMB	ZTCB
1	122	115.86	109.87	123.66	123.49
2	50	57.39	65.51	46.65	47.09
3	18	18.95	19.53	18.10	17.96
4	4	4.69	2.91	7.22	6.85
5	4	0.93	0.17	2.37	2.61
Total	198	198	198	198	198
MLE		$\lambda = 0.9906$	$p = 0.2297$	$\beta = 0.2795$ $k = 0.6972$	$\beta = 0.2761$ $v = 0.0001$
Loglik		205.9477	213.8822	203.7187	203.3796
χ^2		2.329	5.517667	0.529	0.423
(P-value)		(0.51)	(0.06)	(0.91)	(0.94)
df		3	3	2	2
AIC		413.89	429.76	411.44	410.76

The zero-tuncated Binomial introduced by [13] yields a fair fit to the data. [12] applied Zero-truncated Poisson (ZTP) and [14] applied Zero-truncated Multiplicative-Binomial (ZTMB) distributions, which they observed fit the data. The ZTMB distribution performed better than ZTP. Nevertheless, our proposed ZTCB distribution outperforms ZTMB distribution.

Example 4: [Mortality data]

The data is taking from NHS 2008: count of mothers of the complete fertility that experienced at least one child death. The sample mean and variance are 2.061 and 2.145, respectively.

Table 4: The expected and observed number of mothers of the complete fertility that experienced at least one child death

X	Obs.Freq	ZTP	ZTB	ZTMB	ZTCB
1	4770	3772.7	3603.7	4550.3	4724.5
2	2388	3159.5	3299.1	2664.3	2449.3
3	1237	1764.0	1869.7	1322.1	1264.6
4	677	738.6	733.6	621.8	650.9
5	350	247.4	211.1	293.6	334.1
6	177	69.1	46.0	143.8	171.1
7	97	16.5	7.7	74.3	87.4
8	37	3.5	1.0	40.9	44.6
9	20	0.6	0.1	24.0	22.7
10	12	0.1	0	14.9	11.5
11	1	0.0	0	9.5	5.8
12	3	0.0	0	6.2	2.9
13	2	0.0	0	3.8	1.5
14	0	0.0	0	1.9	0.7
15	1	0.0	0	0.6	0.4
Total	9772	9772	9772	9772	9772

MLE	$\theta = 1.6749$	$\pi = 0.1157$	$\beta = 0.3551$ $k = 0.8547$	$\beta = 0.3376$ $v = 0.0085$
Loglik	14623.28	15026.18	13990.38	13950.54
X^2	2445.82	4375.83	86.90	8.91
(P-value)	(<0.01)	(<0.01)	(<0.01)	(0.35)
Deviance	1264.26	1862.12	89.82	9.21
df	6	5	8	8
AIC	29248.56	30054.35	27984.76	27905.09

From the example 1 through to 4 (in table 1 to 4), the proposed ZTCB distribution fits the datasets from different field well while the ZTMB distribution performs well in table 1 to 3 but fails to fit the data in example 4. Obviously, in the datasets where ZTMB perform well, the variances are less than means and $n < 10$. In example 4 where variance $\geq$ mean and $n > 10$, it failed woefully! The ZTB distribution fails to fit the examples here except in example 3 where it slightly fits the data. The examples illustrate the flexibility of the proposed ZTCB distribution, an edge on other existing distributions, which is one of the major contribution of this work.

5. CONCLUSION

The first three datasets have been studied previously with Zero-truncated Poisson (ZTP), zero-truncated Binomial (ZTB) and/or zero-truncated Multiplicative Binomial (ZTMB) distributions and now introducing zero-truncated Com-binomial (ZTCB) distribution. Our proposed model is more flexible than ZTMB and other distributions. It is characterized by two parameters. The maximum likelihood method is used to obtain the estimators of the parameters through R-software. The statistics (chi-square goodness-of-fit) as well Deviance (only in example 4), which are asymptotically the same and the AIC show that the proposed ZTCB yields best result among the models under consideration. Moreover, work is ongoing to obtain more mathematical properties of the ZTCB distribution.

REFERENCES

- [1] Conway, R. W. and Maxwell, W. L. (1962). A queuing model with state dependent service rates. *Journal of Industrial Engineering*, 12:132-6.
- [2] Shmueli G., Borle S., and Boatwright P. (2005). A useful distribution for fitting discrete data: Revival of the Conway-Maxwell-Poisson distribution. *Applied Statistics* ;54:127-42.
- [3] Jowaheer V, Khan NAM (2009). Estimating regression effects in COM Poisson generalized linear model. *World Academy of Science, Engineering, and Technology*, ;53:213-23.
- [4] Sellers K.F., and Shmueli G. (2010) A flexible regression model for count data. *Annals of Applied Statistics*, 4:943-61.
- [5] Famoye, F. (1993). Restricted generalized Poisson regression model. *Comm. Statist. Theory Methods* 22 1335–1354. MR1225247
- [6] Zuur A.F., Leno E.N., Walker N.J., Saveliev A. and Smith G.M. (2009). Zero-truncated and zero-inflated models for count data. In mixed effects models and extensions in ecology with R-Statistics for *Biology and Health*, 1-33.
- [7] Borges P, Rodrigues J and Balakrishnan N. (2014). A Compoisson type generalization of the Binomial distribution and its properties and applications. *Statistics and Probability Letters* 87.
- [8] Aderoju, S.A.,(2017). Zero-Truncated Compoisson-Binomial Distribution and its Application. *International Journal of Advanced Research in Science, Engineering and Technology, (IJARSET)*, Vol.4, Issue 3
- [9] Jolayemi, E.T.(1990). Model selection for One-Dimensional Multinomials. *Biom.J*.32: 827-834.
- [10] R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.
- [11] Shanker R, HagosFesshayee, Sujatha Selvaraj and AbreheYemane (2015), On Zero-Truncation of Poisson and Poisson-Lindley Distributions and Their Applications, *Biometrics & Biostatistics International Journal*, Vol 2, Issue 6.
- [12] Matthews, J.N.S. and Appleton, D.R. (1993). An application of the truncated Poisson distribution to immunogold assay. *Biometrics*, 49, 617–621
- [13] Thomas D.G. and Gart J.J. (1971). Small sample performance of some estimators of the truncated binomial distribution. *Journal of the American Statistical Association*, 66, 169-175.
- [14] Elamir E.A.H, (2013). Multiplicative-Binomial Distribution: Some Results on Characterization, Inference and Random Data Generation. *Journal of Statistical Theory and Applications*, Vol. 12, No. 1 (May 2013), 92-105