ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 4, Issue 3 , March 2017

Zero-Truncated Compoisson-Binomial Distribution and its Application

Samuel Adewale Aderoju

Department of Statistics and Mathematical Sciences, Malete, P.M.B. 1530, Ilorin, Kwara State, Nigeria

ABSTRACT:In this paper, Zero-truncated Com-binomial distribution was derived and investigated its behavior in modeling structurally non-zero data. The proposed distribution is characterized by two parameters, which make it flexible. The maximum likelihood method is used to obtain the estimators of the parameters through R-software. Two real-life datasets were used to evaluate its performance. The statistic (chi square goodness-of-fit) with the p-value shows that the proposed Zero-truncated Com-binomial distribution yields "a good fit".

KEYWORDS: Zero-Truncated, Com-Binomial, maximum likelihood, structurally non-zero, goodness-of-fit.

I. INTRODUCTION

When the data to be modelled is a count, binary or non-binary, data type it is important to consider distributions that handle such data well. In this case the Binomial and Poisson models are the popular distributions to describe such data. The Poisson and Binomial models have been applied in many disciplines, including medicine, Economy, epidemiology, Biology and Demography.

Compoission distribution was originally developed in 1962 to model Queuing process by [1], the distribution was later revisited by [2] after a period in which it was not been widely used. [2]derived many of the basic properties of the distribution. The Compoission distribution belongs to the exponential family as well as to the two-parameter power series family of distributions. It introduces an extra parameter, v, which governs the rate of decay of successive ratios of probabilities. It built-in the usual Poisson (when $v= 1$), geometric (when $v= 0$) and Bernoulli (when $v= \infty$) distributions and it allows for both thicker and thinner tails than the Poisson distribution ([2], [3]).

The distribution has recently become much more widely known and applied, including studies such as, birth process models ([4]), internet search engine visits ([5]), analyzing word length ([2]), prediction of purchase timing and quantity decisions ([6]), quarterly sales of clothing ([2]), the timing of bid placement and extent of multiple bidding ([7]), modeling electric power system reliability ([8]), developing cure rate survival models ([9]), modeling the number of car breakdowns ([10]), and modeling motor vehicle crashes ([11]; [12]).

[13]compared the efficiency of quasi-likelihood and Maximum Likelihood Estimate estimation approaches for estimating the parameters of a single-link Compoission based on simulated data sets. [14]developed a Maximum Likelihood Estimate (MLE) for a single link GLM based on the original Compoission distribution.

The major advantage of the Compoission 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 [15]. [14]demonstrated that Compoission model formulation assumes a constant dispersion level across all observation, such an extension still maintains the structure of an exponential family, unlike that of the generalized Poisson model of [15].

However, 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. Another example is a study by the county traffic court on the number of tickets received by teenagers ([16]). Only individuals who have received at least one citation are in the traffic court files.

The rest of this paper is organized as follows: in the next section we proposed a new distribution, properties and the parameters estimation of ZTCMB distribution are discussed in section III. In section IV, the application of the distribution is illustrated. The results were discussed and conclusion made in section V.

.

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 4, Issue 3 , March 2017

II. THE PROPOSED DISTRIBUTION

The probability mass function (pmf) of the Com-Binomial (CMB) is defined as, ([2]);

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

for $n\in Z^{+}$, $\pi(0,1)$ *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 [17]):

i. The CMB distribution can be interpreted as a sum of equi-correlated Bernoulli variables.

ii. The Compoission distribution ([1]) is approximation to the CMB distribution when n is getting large. (for details, see [17])

Zero-Truncated Com-Binomial (ZTCMB) Distribution

The pmf of the CMB distribution as defined above is

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

However, the Zero-truncated version of the distribution, which we refer to as Zero-truncated Com-Binomial (ZTCMB) can be derived as

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

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 CMB defined above.

Therefore, the pmf of the ZTCMB distribution is derived as

$$P(y;n,\lambda,v)=\frac{\binom{n}{y}^{v}\pi^{y}(1-\pi)^{n-y}}{\sum_{i=0}^{n}\binom{n}{i}^{v}\pi^{i}(1-\pi)^{n-i}}\div\left\{1-\frac{(1-\pi)^{n}}{\sum_{i=0}^{n}\binom{n}{i}^{v}\pi^{i}(1-\pi)^{n-i}}\right\}$$

$$=\frac{\binom{n}{y}^{v}\pi^{y}(1-\pi)^{n-y}}{\sum_{i=0}^{n}\binom{n}{i}^{v}\pi^{i}(1-\pi)^{n-i}}*\left[\frac{\sum_{i=0}^{n}\binom{n}{i}^{v}\pi^{i}(1-\pi)^{n-i}}{\sum_{i=0}^{n}\binom{n}{i}^{v}\pi^{i}(1-\pi)^{n-i}-(1-\pi)^{n}}\right]$$

$$=\frac{\binom{n}{y}^{v}\pi^{y}(1-\pi)^{n-y}}{\sum_{i=0}^{n}\binom{n}{i}^{v}\pi^{i}(1-\pi)^{n-i}-(1-\pi)^{n}}, \qquad y=1,2,3\ \dots,n \qquad 4$$

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

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 4, Issue 3 , March 2017

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

III.PROPERTIES OF ZTCMB DISTRIBUTION

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

Figures below present the pmf of the ZTCMB distribution for n= 7 at different values of π (as p)and v. For $v \to \infty$, the pmf is concentrated at the point nπ and for $v \to -\infty$ is concentrated at 1 or n.

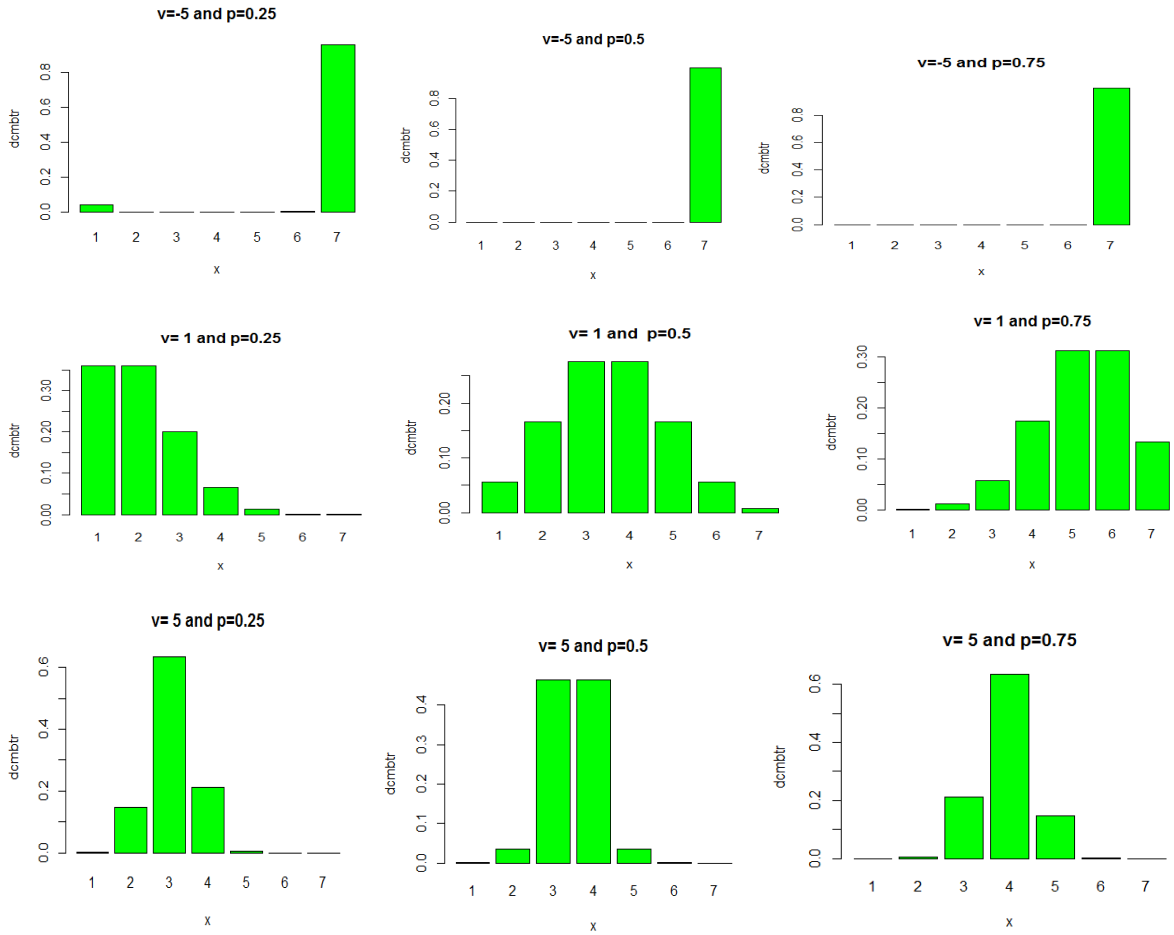


Figure1: *Graph of probability function of ZTCMB distribution for different values of the parameters (v = -5, 1, 5* **vs** *π = p = 0.25, 0.5, 0.75).*

Maximum Likelihood Estimation of the Parameters

The likelihood function of the ZTCMB is

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 4, Issue 3 , March 2017

$$L(\pi, v|y_i) = \prod_{i=1}^{n} \frac{\binom{n}{y_i}^{v} \pi^{y_i}(1-\pi)^{n-y_i}}{\sum_{j=1}^{n}\binom{n}{j}^{v} \pi^{j}(1-\pi)^{n-j}} \qquad 6$$

While the log-likelihood function is

$$\mathcal{L} = \sum_{i=1}^{n}\left\{v log\binom{n}{y_i} + y_i log\pi + (n-y_i)\log(1-\pi) - logQ\right\} \qquad 7$$

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

We are to find the first and second partial derivative of equation (7) with respect to each parameter and equate them to zero as:

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

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

i. Fisher Score Algorithm, or
ii. *Newton-Raphson* (NR) iterative method, as implemented by [18]
iii. *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 ([19]) with ***"L-BFGS-B"*** method. This can as well be done by using **PROC NLMIXED** in SAS.

IV. MODEL APPLICATION

Two datasets used by [20] are used here as example one and two.

Example 1: immunogold assay data

The data is taking from [21], 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 1: *observed and expected frequencies of immunogold assay data with MLE, Log-likelihood and chi-square statistic.*

X	1	2	3	4	*5*	*Total*	*MLE*	*Loglik*	x^2 *(P-value)*
Obs. Freq	122	50	18	4	4	**198**			
ZTCMB	123.49	47.09	17.96	6.85	2.61	**198**	$\beta = 0.2761$ $v = 0.0001$	203.3796	0.423 (0.94)

From table 1 above, the p-value is obtained as 0.94, which support the null hypothesis that the ZTCMB distribution fits the data. Obviously, the close agreement between the observed and expected frequencies indicates that the proposed ZTCMB distribution provides a good fit.

Example 2: flower heads data.

The data for example two is taking from [22], who gave counts of flower heads with 1, 2, . . . , 9 fly eggs. The sample mean and variance are 3.034 and 3.3056, respectively.

Table 2: *observed and expected frequencies of flower heads with MLE, Log-likelihood and chi-square statistic.*

X	**Obs.Freq**	**ZTCMB**
1	22	21.07
2	18	19.97
3	18	16.58
4	11	12.35
5	9	8.32
6	6	5.08

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 4, Issue 3 , March 2017

7	3	2.78
8	0	1.33
9	1	0.52
Total	**88**	**88**
MLE		$\pi = 0.4026$ $k = 0.2459$
Loglik		163.625
x^2 ***(P-value)***		0.568 (0.97)

From table 2 above, the p-value is obtained as 0.97, which support the null hypothesis that the ZTCMB distribution fits the data. Obviously, the close agreement between the observed and expected frequencies indicates that the proposed ZTCMB distribution provides a good fit.

V.CONCLUSION

The two datasets examples are used to illustrate the flexibility of the proposed distribution, introducing zero-truncated Com-Binomial (ZTCMB). It is characterized by two parameters. The maximum likelihood method is used to obtain the estimators of the parameters through R-software. The statistic (chi square goodness-of-fit) shows that the proposed ZTCMB yields 'good fit'.

Work is in progress to compare ZTCMB performance with the existing models using more real life datasets. Moreover, we will try as much as possible to obtain more mathematical properties of the new model as well.

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]Boatwright, P., Borle, S., and Kadane, J.B. (2003). A model of the joint distribution of purchase quantity and timing; Journal of the American Statistical Association, 2003;98:564-72.
[4]Ridout MS and Besbeas P. (2004). An empirical model for underdispersed count data. Statistical
[5 Telang R, Boatwright P, and Mukhopadhyay T. (2004). A mixture model for internet search Engine visits. Journal of Marketing ;41:206-14.
[6]Borle S, Boatwright P, Kadane JB, Nunes J, Shmueli G. (2005): Effect of Product Assortment Changes on Consumer Retention. *Marketing Science*; 24: 616-622.
[7]Borle, S., Boatwright, P. and Kadane, J. B. (2006). The timing of bid placement and extent of multiple bidding: An empirical investigation using eBay online auctions. Statistical Science, ;21(2):194-205.
[8]Guikema, S.D., Coffelt, J.P. (2008). A flexible count data regression model for risk analysis. Risk Analysis, 28(1):213-23.
[9]Rodrigues J, de Castro M, Cancho VG, and Balakrishnan N (2009). Compoission cure rate survival models and an application to a cutaneous melanoma data. Journal of Statistical Planning and Inference, 2009;139:3605-11.
[10]Khan NM, Jowaheer V (2010). A comparison of marginal and joint generalized Quasilikelihood estimating equations based on the Compoission GLM: Application to car breakdowns data. International Journal of Mathematical and Statistical Sciences, ;2(2).
[11]Lord D, Guikema SD, Geedipally SR (2008). Application of the Conway-Maxwell-Poisson generalized linear model for analyzing motor vehicle crashes. Accident Analysis & Prevention;40(3):1123-34.
[12]Lord D, Geedipally S, Guikema S. (2010). Extension of the application of Conway-Maxwell-Poisson models: Analyzing traffic crash data exhibiting under-dispersion. Risk Analysis,; 30(8):1268-76.
[13]Jowaheer V, Khan NAM (2009). Estimating regression effects in COM Poisson generalized linear model. World Academy of Science, Engineering, and Technology, ;53:213-23.
[14] Sellers K.F., and Shmueli G. (2010) A flexible regression model for count data. Annals of Applied Statistics, 4:943-61.
[15]Famoye, F. (1993). Restricted generalized Poisson regression model. Comm. Statist. Theory Methods 22 1335–1354. MR1225247
[16]Borge P, Rodrigues J and Balakrishnan N. (2014). A Compoission type generalization of the Binomial distribution and its properties and applications. Statistics and Probability Letters 87. 2014 Elsevier B.V. pages158–166.
[17] Jolayemi, E.T.(1990). Model selection for One-Dimensional Multinomials. Biom.J.32: 827 834.
[18]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/.
[19]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
[20]Matthews, J.N.S. and Appleton, D.R. (1993). An application of the truncated Poisson distribution to immunogold assay. *Biometrics*, 49, 617–621
[21]Finney, D.J. and Varley, G.C. (1955). An example of the truncated Poisson distribution. *Biometrics*, 11,387–394.