Mikailalsys Journal of
Mathematics and Statistics

e-ISSN : 3030-816X
p-ISSN : 3030-8399

Index : Harvard, Boston, Sydney University, Dimensions, Lens, Scilit, Semantic, Google, etc

https://doi.org/10.58578/MJMS.v3i2.5171

Generic Count Distributions and Their Zero-Inflated Forms: A Simulation Study

Adetunji A. A. & Sabri S. R. M.
Federal Polytechnic, Ile-Oluji, Nigeria; Universiti Sains Malaysia, Penang, Malaysia
adeadetunji@fedpolel.edu.ng; rijal@usm.my

Article Info:

Submitted:	Revised:	Accepted:	Published:
Feb 15, 2025	Feb 27, 2025	Mar 12, 2025	Mar 17, 2025

Abstract

The percentage of zero observations necessitating zero-inflated distributions in count data modelling has been a major issue. The challenge in such a situation is determining when to shift from parent distributions to their zero-inflated versions. In most studies, the performances of parent distributions are assessed with those of their zero-inflated forms. This study conducts simulation studies for the Poisson and the negative binomial distributions and their respective zero-inflated forms. Count data [0, 4] with different percentages of zero counts are simulated using different sample sizes. Both negative log-likelihood and Bayesian information criterion (which considers the number of estimated parameters) are used to assess performance. Results show that the zero-inflated Poisson distribution best suits modelling all forms of data when the negative log-likelihood value is used to assess performance. When the BIC is used, the Poisson distribution gives the best performance for both 10% and 20% zeros, while the ZIP distribution is the best for both 50% and 90% zeros. The NB distribution outperforms the ZINB distribution in all situations. Also, in all cases, the negative binomial performs better than the zero-inflated negative binomial distributions. To further assess the distributions, four count data sets with varying percentages of zero are examined. Both the ZIP and the NB distributions perform better than others.

Keywords: Count Data, Zero-Inflation, Poisson Distribution, Negative Binomial Distribution

Introduction

The nature of some count data makes the frequency of zero counts (higher) inflated or (lower) deflated. Zero-deflation is when the observed frequency of zero in a count data is much lower and inconsistent with the expected frequency. This scenario is typically less reported in the literature (Angers & Biswas, 2003; Dietz & Böhning, 2000) than in instances where zero frequency is substantially larger (Bauer *et al.*, 2007; Conceição *et al.*, 2017). Higher zero frequencies are found in diverse fields of study, including the number of defects in manufacturing (Lambert, 1992), the number of claims in actuarial science (Adetunji & Sabri, 2021), the number of antenatal care visits (Bekalo & Kebede, 2021), studies involving health service utilization (Feng, 2021), and in several others. A simple question then is what percentage of zero counts in observation is needed to assume zero inflation? Since there is no particular rule of thumb on this, researchers often compare some parent discrete distributions (like Poisson, and negative binomial distributions) with their zero-inflated forms (Bekalo & Kebede, 2021; Loeys *et al.*, 2011; Tajuddin & Ismail, 2022). This is done by assuming that the process of generating some count data with excess zero may not be correctly modelled using common parent distributions (Feng, 2021; Loeys *et al.*, 2011).

Generally, a zero-inflated distributions have both the structural and sampling parts. The structural part generates 0, while the sampling part generates the counts. Assuming ω as the zero-inflation parameter, the zero0inflation form of a discrete distribution with PMF P_x is denoted by:

$$P_x^z = \begin{cases} \omega + (1-\omega)P_0, & x = 0 \\ (1-\omega)P_x, & x = 1,2,3,\dots \end{cases}$$

The form of the distribution P_x^z is determined by the value of ω. If $\omega = 0$, $P_x^z = P_x$, when $\omega < 0$, P_x^z becomes a zero-deflated distribution and when $\omega > 0$, P_x^z becomes a zero-inflated distribution.

In this study, count data [0, 4] of sizes 50, 250, 2000, 10000, and 100000, each with 10%, 20%, 50%, and 90% zeros are simulated. The simulated data are assessed with the Poisson and the negative binomial (NB) distributions along with their respective zero-inflated forms (zero-inflated Poisson – ZIP, zero-inflated negative binomial – ZINB). Comparisons are made using both negative log-likelihood values and the Bayesian information criterion.

Competing Distributions

i. **Poisson Distribution:** A discrete random variable X has Poisson distribution if its PMF is defined as:

$$P_x = \frac{e^{-\theta}\theta^x}{x!}$$

ii. **Zero Inflated Poisson Distribution:** A discrete random variable X has a Zero-Inflated Poisson (ZIP) distribution with ω zero inflation parameter if its PMF is given as:

$$P_x^z = \begin{cases} \omega + (1-\omega)e^{-\theta}, & x = 0 \\ (1-\omega)\dfrac{e^{-\theta}\theta^x}{x!}, & x = 1,2,3,\dots \end{cases}$$

iii. Negative Binomial Distribution: A discrete random variable X has a negative binomial distribution if its PMF denoted by P_n is given:

$$P_x = \binom{x+r-1}{x}\left(\frac{\theta}{1+\theta}\right)^x\left(\frac{1}{1+\theta}\right)^r, \qquad x = 0,1,2,\dots, \qquad \theta, r > 0$$

iv. Zero-Inflated Negative Binomial Distribution: A discrete random variable X has a Zero-Inflated Negative Binomial (ZINB) distribution with ω zero inflation parameter if its PMF is given as:

$$P_x^z = \begin{cases} \omega + (1-\omega)\left(\dfrac{1}{1+\theta}\right)^r, & x = 0 \\ (1-\omega)\left[\dbinom{x+r-1}{x}\left(\dfrac{\theta}{1+\theta}\right)^x\left(\dfrac{1}{1+\theta}\right)^r\right], & x = 1,2,3,\dots \end{cases}$$

Simulations

Tables 1 to 10 show the results of the simulation studies for the four examined distributions along with each sample size and each percentage of zero counts.

Table 1. Negative log-likelihood values for various simulations when n = 50

Distribution	Percentage of 0			
	10%	20%	50%	90%
Poisson	76.0361	84.2655	67.4706	27.6112
Zero-Inflated Poisson	74.1379	83.2431	59.7509	21.9795
Negative Binomial	76.2293	84.2896	60.4916	22.7786
Zero-Inflated Negative Binomial	79.3877	94.6033	61.3227	22.911

Table 2. BIC values for various simulations when n = 50

Distribution	Percentage of 0			
	10%	20%	50%	90%
Poisson	155.9842	172.4430	138.8532	59.1344
Zero-Inflated Poisson	156.0998	174.3102	127.3258	51.7830
Negative Binomial	160.2826	176.4032	128.8072	53.3812
Zero-Inflated Negative Binomial	170.5115	200.9427	134.3815	57.5581

Table 3. Negative log-likelihood values for various simulations when n = 250

Distribution	Percentage of 0			
	10%	20%	50%	90%
Poisson	389.6850	430.0712	418.0942	116.1207
Zero-Inflated Poisson	386.8592	426.7672	378.1604	95.6742
Negative Binomial	390.8964	430.2475	388.4172	96.0647
Zero-Inflated Negative Binomial	428.7068	480.6581	401.0008	96.4681

Table 4. BIC values for various simulations when n = 250

Distribution	Percentage of 0			
	10%	20%	50%	90%
Poisson	784.8915	865.6639	841.7099	237.7629
Zero-Inflated Poisson	784.7613	864.5773	767.3637	202.3913
Negative Binomial	792.8357	871.5379	787.8773	203.1723
Zero-Inflated Negative Binomial	873.9780	977.8806	818.5660	209.5006

Table 5. Negative log-likelihood values for various simulations when n = 2000

Distribution	Percentage of 0			
	10%	20%	50%	90%
Poisson	3131.4090	3501.1929	3091.9810	1137.1980
Zero-Inflated Poisson	3104.9460	3447.2080	2812.7180	848.2981
Negative Binomial	3138.8310	3502.8570	2869.7460	871.4395
Zero-Inflated Negative Binomial	3420.0580	3889.3560	2942.7710	875.6283

Table 6. BIC values for various simulations when n = 2000

Distribution	Percentage of 0			
	10%	20%	50%	90%
Poisson	6270.4189	7013.3149	6191.5629	2281.9969
Zero-Inflated Poisson	6225.0938	6909.6178	5640.6378	1711.7980
Negative Binomial	6292.8638	7017.5876	5754.6938	1758.0808
Zero-Inflated Negative Binomial	6862.9187	7801.5147	5908.3447	1774.0593

Table 7. Negative log-likelihood values for various simulations when n = 10,000

Distribution	Percentage of 0			
	10%	20%	50%	90%
Poisson	15639.0500	17412.0300	15723.6500	5994.1690
Zero-Inflated Poisson	15532.0200	17188.3400	14241.1800	4561.5650
Negative Binomial	15667.9500	17416.7600	14589.1700	4620.4540
Zero-Inflated Negative Binomial	17202.2600	19562.8500	14992.9100	4642.3020

Table 8. BIC values for various simulations when n = 10,000

Distribution	Percentage of 0			
	10%	20%	50%	90%
Poisson	31287.3103	34833.2703	31456.5103	11997.5483
Zero-Inflated Poisson	31082.4607	34395.1007	28500.7807	9141.5507
Negative Binomial	31354.3207	34851.9407	29196.7607	9259.3287
Zero-Inflated Negative Binomial	34432.1510	39153.3310	30013.4510	9312.2350

Table 9. Negative log-likelihood values for various simulations when n = 100,000

Distribution	Percentage of 0			
	10%	**20%**	**50%**	**90%**
Poisson	156637.2000	174637.8000	156043.1000	58709.9300
Zero-Inflated Poisson	155792.8000	172220.0000	141900.7000	45457.5300
Negative Binomial	156961.0000	174664.4000	145052.5000	46014.3300
Zero-Inflated Negative Binomial	173339.7000	196348.5000	148949.2000	46220.2200

Table 10. BIC values for various simulations when n = 100,000

Distribution	Percentage of 0			
	10%	**20%**	**50%**	**90%**
Poisson	313285.9129	349287.1129	312097.7129	117431.3729
Zero-Inflated Poisson	311608.6259	344463.0259	283824.4259	90938.0859
Negative Binomial	313945.0259	349351.8259	290128.0259	92051.6859
Zero-Inflated Negative Binomial	346713.9388	392731.5388	297932.9388	92474.9788

Remark: The distribution with the lowest negative log-likelihood and Bayesian information criterion is considered the best.

Discussion of Simulation

When *n = 50*

The ZIP distribution outperforms all other competing distributions when the performance is assessed using the negative log-likelihood values. When BIC is used, the Poisson distribution gives the best performance for both 10% and 20% zeros while the ZIP distribution is the best for both 50% and 90% zeros. The NB distribution outperforms the ZINB distribution in all situations.

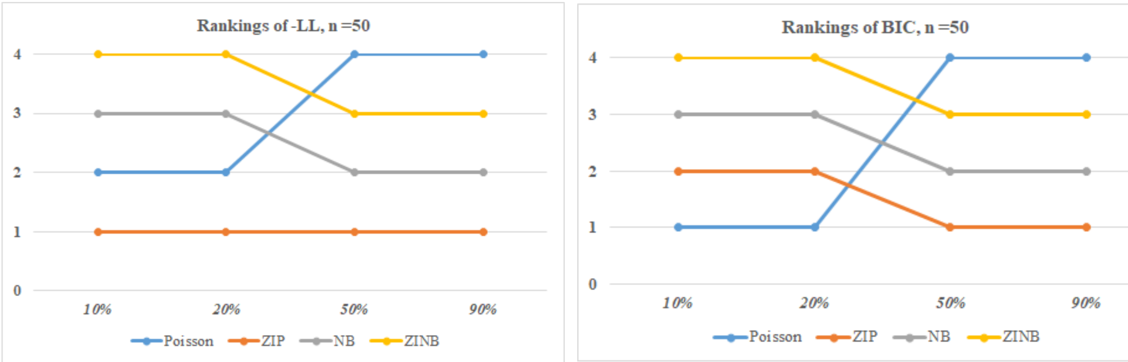

Figure 1: Rankings of –LL and BIC values when n = 50

When *n = 250; 2,000; 10,000; 100,000*

For the four sample sizes, both negative log-likelihood and BIC give the same conclusion. The ZIP distribution provides the best in all percentages of zeros. The NB distribution outperforms the ZINB distribution in all situations. The Poisson distribution performs better than NB distribution when both 10% and 20% of observations are zeros.

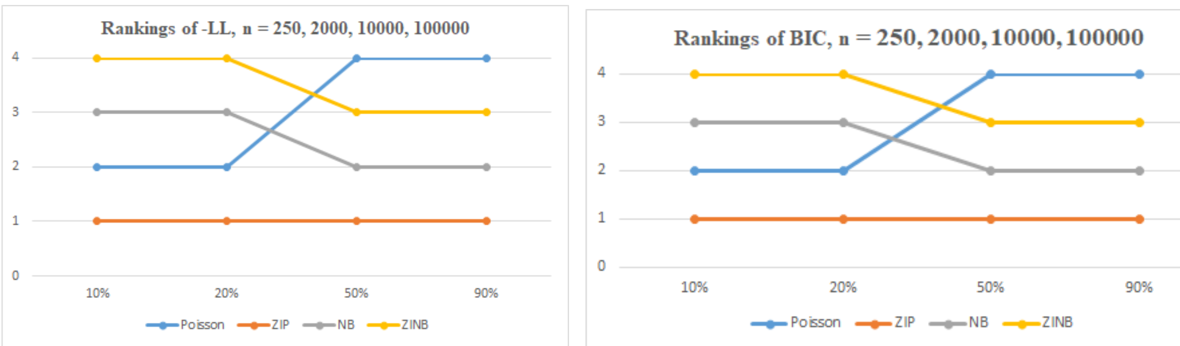

Figure 2: Rankings of –LL and BIC values when n = 250, 2000, 10000, 100000

Applications

Four real-life datasets with different percentages of zeros are used to assess observations from the above simulation. The summary of the data is presented in Table 11.

Table 11. Summary for the four real-life dataset

	Data 1	Data 2	Data 3	Data 4
Mode	0.00	0.00	0.00	0.00
Skewness	10.46	4.27	2.56	1.31
Percentage of 0	98.96	93.49	79.01	58.33

The first data is the number of claims on motorcycle insurance from a Swedish company. The data has been previously used in modelling claim frequency (Adetunji & Sabri, 2023a; Omari *et al.*, 2018; Sabri & Adetunji, 2023). The second data represents the claim frequency of automobile injury in Singapore and has been used in count data modelling (Adetunji & Sabri, 2024; Wagh & Kamalja, 2017). Data 3 is the frequency of claims from policyholders as reported by a Turkish insurance company. The data has been previously used on zero-inflated modelling of claim frequency (Adetunji & Sabri, 2023c; Meytrianti et al., 2019; Sabri & Adetunji, 2023; Sarul & Sahin, 2015). The fourth data is the number of mistakes in copying random digits as previously found in (Adetunji & Sabri, 2023b, 2023c; Sah & Mishra, 2019; Samutwachirawong, 2021; Shanker & Mishra, 2016).

Table 12. Negative log-likelihood values for the real-life data

	Data 1	Data 2	Data 3	Data 4
Poisson	3871.9950	1941.1780	7153.1580	77.5456
Zero-Inflated Poisson	3840.8830	1933.1680	7038.9110	71.9210
Negative Binomial	3841.5000	1932.3830	7029.7140	73.3683
Zero-Inflated Negative Binomial	3857.3160	1938.2020	7057.0610	74.9561

Table 13. BIC values for the real-life data

	Data 1	Data 2	Data 3	Data 4
Poisson	7755.0652	3891.2764	14315.6046	7755.0652
Zero-Inflated Poisson	7703.9163	3884.1768	14096.3992	7703.9163
Negative Binomial	7705.1503	3882.6068	14078.0052	7705.1503
Zero-Inflated Negative Binomial	7747.8575	3903.1652	14141.9878	7747.8575

Discussion of Results

Tables 12 and 13 respectively show the negative log-likelihood and the BIC values of the examined four dataset. The ZIP distribution provides the overall best fits for Data 1 and Data 3 when both negative log-likelihood and BIC values are considered. The NB distribution gives the best fit for both Data 2 and 3. Both the Poisson and the ZINB distributions are outperformed in all the examined cases, although the ZINB distribution gives a relatively better fit.

Conclusion

When examining count data with excess zeros, focus in some studies is usually on different zero-inflated distributions. Comparing the frequentist and the Bayesian approach for zero-inflated regression, (Tajuddin & Ismail, 2022) reported that both ZIP and ZINB models from the Bayesian approach fit the data better than the frequentist models. Findings in this study suggest that incorporation of the classical Poisson and NB distributions to the study might have presented a different outcome.

Also, (Adetunji & Sabri, 2023c) found that NB distribution outperformed ZINB distribution since the negative binomial distribution is a form of the mixed Poisson distribution (when the gamma distribution is assumed for the parameter of the Poisson distribution). This finding is corroborated by the results obtained in this study.

Hence, when modelling count observation with a a higher frequency of zero counts than what any classical parent distribution would be able to model, it is essential to not focus on the likes of ZIP and ZINB distributions alone but also on their natural form.

References

Adetunji, A. A., & Sabri, S. R. M. (2021). Modelling Claim Frequency in Insurance Using Count Models. Asian Journal of Probability and Statistics, 14(4), 14–20. https://doi.org/10.9734/ajpas/2021/v14i430334

Adetunji, A. A., & Sabri, S. R. M. (2023a). A New Three-Parameter Mixed Poisson Transmuted Weighted Exponential Distribution with Applications to Insurance Data. Science and Technology Indonesia, 8(2), 235–244. https://doi.org/10.26554/sti.2023.8.2.235-244

Adetunji, A. A., & Sabri, S. R. M. (2023b). On the Poisson Transmuted Ailamujia Distribution with Applications to Dispersed and Skewed Count Data. Journal of

Statistics and Management Systems, 26(4), 929–943. https://doi.org/10.47974/JSMS-1023

Adetunji, A. A., & Sabri, S. R. M. (2023c). On zero-inflated mixed Poisson Transmuted Exponential Distribution: Properties and Applications to observation with excess zeros. Maejo International Journal of Science and Technology, 17(01), 68–80.

Adetunji, A. A., & Sabri, S. R. M. (2024). A new two-parameter Poisson-transmuted exponential distribution: Properties and applications in count observations. Proceedings of the International Conference on Mathematical Sciences and Technology 2022 (MathTech 2022), 1–7. https://doi.org/10.1063/5.0192459

Angers, J., & Biswas, A. (2003). A Bayesian of Zero-Inflated Generalized Poisson Model. Computational Statistics and Data Analysis, 42, 37–46.

Bauer, T., Göhlmann, S., & Sinning, M. (2007). Gender Differences in Smoking Behavior. Health Economics, 16, 895–909.

Bekalo, D. B., & Kebede, D. T. (2021). Zero-Inflated Models for Count Data: An Application to Number of Antenatal Care Service Visits. Annals of Data Science, 8(4), 683–708. https://doi.org/10.1007/s40745-021-00328-x

Conceição, K. S., Louzada, F., Andrade, M. G., & Helou, E. S. (2017). Zero-Modified Power Series Distribution and its Hurdle Distribution Version. Journal of Statistical Computation and Simulation, 87(9), 1842–1862. https://doi.org/10.1080/00949655.2017.1289529

Dietz, E., & Böhning, D. (2000). On Estimation of the Poisson Parameter in Zero-Modified Poisson Models. Computational Statistics and Data Analysis, 34(4), 441–459. https://doi.org/10.1016/S0167-9473(99)00111-5

Feng, C. X. (2021). A Comparison of Zero-Inflated and Hurdle Models for Modeling Zero-Inflated Count Data. Journal of Statistical Distributions and Applications, 8(1), 1–19. https://doi.org/10.1186/s40488-021-00121-4

Lambert, D. (1992). Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing. Technometrics, 34, 1–14.

Loeys, T., Moerkerke, B., De-Smet, O., & Buysse, A. (2011). The analysis of zero-inflated count data: Beyond zero-inflated Poisson regression. British Journal of Mathematical and Statistical Psychology, 1–17. https://doi.org/10.1111/j.2044-8317.2011.02031.x

Meytrianti, A., Nurrohmah, S., & Novita, M. (2019) An Alternative Distribution for Modelling Overdispersion Count Data: Poisson Shanker Distribution. ICSA - International Conference on Statistics and Analytics 2019, 1, 108–120. https://doi.org/10.29244/icsa.2019.pp108-120

Omari, C. O., Nyambura, S. G., & Mwangi, J. M. W. (2018). Modeling the Frequency and Severity of Auto Insurance Claims Using Statistical Distributions. Journal of Mathematical Finance, 8(1), 137–160. https://doi.org/10.4236/jmf.2018.81012

Sabri, S. R. M. & Adetunji, A. A. (2023). Zero-Inflated Poisson Transmuted Weighted Exponential Distribution: Properties and Applications. The Borneo Journal of Science and Technology, 44(2), 1–16. https://doi.org/10.51200/bsj.v44i2

Sah, B. K, & Mishra, A. (2019). A Generalised Exponential-Lindley Mixture of Poisson Distribution. Nepalese Journal of Statistics, 3, 11–20. https://doi.org/10.4135/9781412952644.n350

Samutwachirawong, S. (2021). Poisson-Exponential and Gamma Distribution: Properties and Applications. Journal of Applied Statistics and Information Technology, 6(2), 17–24.

Sarul, L. S., & Sahin, S. (2015). An Application of Claim Frequency Data Using Zero Inflated and Hurdle Models in General Insurance. The Journal of Business, Economics and Finance, 4(4), 732–743.

Shanker, R., & Mishra, A. (2016). A Quasi Poisson-Lindley Distribution. Journal of the Indian Statistical Association, 54(1), 113–125.

Tajuddin, M. R. R., & Ismail, N. (2022). Frequentist and Bayesian Zero-Inflated Regression Models on Insurance Claim Frequency: A comparison study using Malaysia's Motor Insurance Data. Malaysian Journal of Science, 41(2), 16–29. https://doi.org/10.22452/mjs.vol41no2.2

Wagh, Y. S., & Kamalja, K. K. (2017). Modeling Auto Insurance Claims in Singapore. Sri Lankan Journal of Applied Statistics, 18, 105–118.