Unveiling The Versatility: Applications And Advancements Of The Negative Binomial Distribution In Statistical Modeling

[1] Flowery Francis

[1] Assistant Professor of Statistics,
[1] Department of Statistics,
[1] Sri. C. Achutha Menon Govt. College, Thrissur, Kerala, India.

Abstract: This study encapsulates the diverse applications and advancements of the negative binomial distribution, highlighting its pivotal role in statistical modeling across various disciplines. From ecology to seismology, transportation to epidemiology, seminal studies by Power and Moser, Rao and Kaila, and Zou et al. demonstrate its efficacy in capturing variance-to-mean relationships, enabling robust hypothesis testing and parameter estimation. Innovations by White and Bennetts, Aeberhard et al., and others refine modeling techniques to accommodate skewed data and mitigate sensitivity to misspecifications, ensuring reliable inference in complex datasets. Furthermore, developments by Hoffman, Vangala, Preisser et al., and others extend the utility of the negative binomial distribution to scenarios marked by excess zeros and serial dependence, enriching statistical methodologies in areas such as crash data analysis and dental health studies. The exploration of alternative distributions, including the generalized negative binomial and quasi-negative binomial distributions, further broadens the statistical toolbox, addressing challenges posed by highly skewed data and moment nonexistence. Collectively, these contributions underscore the enduring relevance and ongoing evolution of the negative binomial distribution in facilitating data-driven discoveries across interdisciplinary domains.

***Index Terms* -** Crash data analysis, Dental health studies, Epidemiology, Negative binomial distribution, Reliable inference, Transportation.

I. Introduction

In the realm of statistical analysis, the negative binomial distribution has emerged as a powerful tool with versatile applications across diverse fields ranging from ecology and seismology to transportation and epidemiology. This distribution offers a robust framework for modeling count data characterized by variance-to-mean relationships, providing invaluable insights into phenomena where traditional models fall short. From addressing the clustering behavior of seismic events in seismology to identifying hotspots in transportation safety, researchers have leveraged the negative binomial distribution to tackle complex challenges such as skewed data distributions, excess zeros, and serial dependence. Furthermore, advancements in statistical methodologies, including the introduction of the generalized negative binomial distribution and the exploration of alternative distributions like the quasi-negative binomial, have further broadened the analytical toolkit, paving the way for more nuanced analyses in fields such as crash data analysis and dental health studies. This paper explores the applications and advancements of the negative binomial distribution, highlighting its pivotal role in facilitating reliable inference and driving data-driven discoveries across interdisciplinary domains.

The article is organized as follows: Section 2 explores the Negative Binomial Distribution, investigating its properties and applications. Following this, Section 3 offers a comprehensive review of significant research endeavors focused on exponential aspects of the distribution. This includes discussions on extensions, techniques for parameter estimation, and real-world applications. Finally, Section 4 provides a concluding summary of the findings presented in the article.

II. Exploring the Negative Binomial Distribution

The negative binomial distribution is a probability distribution used to model the number of successes in a sequence of independent Bernoulli trials before a specified number of failures occurs. It's particularly useful for count data where the number of trials needed to achieve a certain number of successes varies. The distribution is characterized by parameters: the probability of success in each trial (p) and the number of successes required (r). Its probability mass function calculates the likelihood of observing a specific number of successes (k) before r failures, taking into account the variable number of trials. The mean and variance of the distribution provide measures of central tendency and spread, respectively, enabling statistical analysis and inference in various fields, including biology, economics, and quality control.

In mathematical terms, the probability mass function (PMF) of the negative binomial distribution is:

$$P(X=k)=\binom{k+r-1}{k}p^k(1-p)^r$$

III. Unlocking Insights: Previous Year Studies on the Negative Binomial Distribution

- In their study "Application of the negative binomial to earthquake occurrences in the Alpide-Himalayan belt" (1986) [8] on earthquake occurrences in the Alpide-Himalayan belt, NM Rao and KL Kaila investigated shallow focus earthquakes (h ⩽ 60 km) with a magnitude range of M = 4.0–6.0 that occurred between 1954 and 1975. They assessed the applicability of the Poisson and negative binomial laws to describe the seismic activity in various high seismicity zones. They found that the clustering of events rendered the simple Poisson model inadequate for most zones, and instead, the negative binomial distribution provided an excellent fit for describing earthquake occurrences. The chi-square (X2) test was utilized to compare the actual observations with the theoretical distributions, confirming the suitability of the negative binomial model for capturing the clustering behavior of seismic events in the Alpide-Himalayan belt.

- In their study titled "Analysis of frequency count data using the negative binomial distribution" (1996) [13] GC White and RE Bennetts propose a likelihood-ratio testing framework based on the negative binomial distribution to analyze skewed count data commonly observed in organism counts. Unlike traditional methods that assume normality and constant variances across treatments, this approach allows for testing the goodness of fit of the negative binomial distribution to the observed counts and subsequently evaluates differences in means and/or aggregation among treatments. By incorporating information on dispersion, the proposed procedure offers insights beyond traditional ANOVA methods, with simulations demonstrating comparable statistical power. Using count data on Orange-crowned Warblers (Vermivora celata) as a case study, the authors illustrate the application and efficacy of their approach in ecological research contexts.

- In their work, "A generalized negative binomial and applications" (1998) [2] MS Bebbington and CD Lai introduce a novel distribution derived from the Markov Bernoulli sequence, termed the generalized negative binomial distribution. They explore the properties of this distribution, highlighting its versatility and applicability across various domains. Through examples drawn from discrete time queueing systems, they demonstrate how the distribution can effectively model real-world phenomena. Furthermore, the authors fit the distribution to two distinct datasets: the eruption record of Mt. Sangay and a record of computer disk failure accesses. Their analysis reveals that the generalized negative binomial distribution offers a suitable fit for data exhibiting strong serial dependence, such as the Mt. Sangay eruption record. However, in cases where the serial dependence is insufficient to justify the additional parameter of the distribution, as observed in the computer disk failure accesses dataset, alternative models may be more appropriate. The study concludes by showcasing the utility of the generalized negative binomial distribution in statistical quality control, underscoring its potential for addressing diverse analytical challenges across different fields.

- In their study titled "Linear model analysis of net catch data using the negative binomial distribution" (1999) [6] JH Power and EB Moser address the limitations of traditional linear model analyses when applied to count data from sampling with nets or trawls. Recognizing the suitability of the negative binomial distribution for describing count data with high variance-to-mean relationships, they propose an approach that allows for estimation of model parameters, including the negative binomial k parameter. Their method enables hypothesis testing of both continuous and discrete model effects and their interactions, incorporating adjustments for varying element sizes encountered during sampling. By utilizing bootstrap replication, they provide a robust framework for analyzing net catch data and evaluating relationships among organism counts and exogenous variables.

- In "Negative binomial control limits for count data with extra‐Poisson variation" (2003) [3] D Hoffman addresses the limitations of traditional control limit techniques based on the Poisson distribution when the assumption of Poisson distribution is violated due to over-dispersion. The negative binomial distribution, which accounts for over-dispersion, is proposed as a suitable alternative. Hoffman describes a straightforward method for calculating exact and approximate control limits for count data using the negative binomial distribution. The effectiveness of this approach is demonstrated through its application to water bacteria count data obtained from a water purification system, highlighting its utility in practical settings where Poisson assumptions may not hold.

- In "Small-sample estimation of negative binomial dispersion, with applications to SAGE data" (2008) [9] MD Robinson and GK Smyth introduce a quantile-adjusted conditional maximum likelihood estimator for the dispersion parameter of the negative binomial distribution. They compare its performance, particularly in terms of bias, to various other methods, demonstrating its superiority in very small samples typical of those from serial analysis of gene expression (SAGE) studies, the motivating data for their research. Additionally, the authors investigate the impact of dispersion estimation on hypothesis testing, deriving an "exact" test that surpasses standard approximate asymptotic tests. This work contributes valuable insights into improving the accuracy of dispersion estimation and its implications for statistical inference in SAGE data analysis.

- In their paper, "Quasi-negative binomial distribution: Properties and applications" (2011) [5] S Li, F Yang, F Famoye, C Lee, and D Black explore a new distribution derived from the class of generalized Lagrangian probability distributions, termed the quasi-negative binomial distribution (QNBD). The QNBD encompasses the negative binomial distribution as a special case and exhibits various properties, including upper tail behavior and limiting distributions. The authors investigate scenarios where moments may not exist and demonstrate that the limiting distribution of QNBD can be the generalized Poisson distribution under certain conditions. Additionally, they introduce a zero-inflated version of QNBD. Through applications across diverse fields, the paper compares QNBD and its zero-inflated counterpart with other existing distributions such as Poisson, generalized Poisson, and negative binomial distributions, as well as their zero-inflated versions. The results indicate that QNBD or its zero-inflated version generally outperform other models, particularly in cases of highly skewed data, heavy-tailed distributions, or excessive numbers of zeros, as evidenced by evaluations using the chi-square statistic and the Akaike Information Criterion.

- In this paper, "One mixed negative binomial distribution with application" (2011) [12] Z Wang introduces the beta-negative binomial (BNB) distribution, a three-parameter distribution obtained as a beta mixture of the negative binomial (NB) distribution. The author derives the closed form and factorial moments of the BNB distribution and establishes recursion on the probability density function (pdf) of the BNB stopped-sum distribution. Stochastic comparisons between the BNB and NB distributions reveal that the BNB distribution exhibits a heavier tail than the NB distribution. Additionally, the paper applies the BNB distribution to insurance data, demonstrating its superior fit compared to the Poisson and NB distributions for count data.

- In their article "Robust inference in the negative binomial regression model with an application to falls data" (2014) [1] WH Aeberhard, E Cantoni, and S Heritier address the challenge of modeling overdispersed count data, such as the number of falls reported during intervention studies, using the negative binomial (NB) distribution. They propose two robust approaches for estimating regression parameters in generalized linear models with NB distribution to mitigate the sensitivity to model misspecifications. The first approach involves applying a bounded function on Pearson residuals, while the second approach bounds unscaled deviance components. Through simulations and application to a randomized controlled trial on reducing falls among Parkinson's disease patients, they demonstrate the effectiveness of their robust procedures in providing reliable inference despite model misspecifications.

- In their paper, "The negative binomial-Erlang distribution with applications" (2014) [4] S Kongrod and W Bodhisuwan introduce a novel three-parameter mixed distribution termed the negative binomial-Erlang distribution. This distribution is obtained by combining the negative binomial distribution with the Erlang distribution and is particularly suited for describing count data with a substantial number of zeros. The authors explore various properties of the negative binomial-Erlang distribution, including factorial moments, mean, variance, skewness, and kurtosis. They also provide parameter estimation techniques based on maximum likelihood estimation. Furthermore, the paper includes applications of the negative binomial-Erlang distribution to two real count datasets. The results demonstrate the superior fit of the negative binomial-Erlang distribution compared to the Poisson and negative binomial distributions for these datasets, highlighting its effectiveness in modeling count data with a large number of zeros.

- In "Negative binomial-generalized exponential distribution: generalized linear model and its applications" (2015) [11] P Vangala presents a novel approach to modeling crash data using the Negative Binomial Generalized Exponential (NB-GE) distribution. This distribution, combining elements of the Negative Binomial and Generalized Exponential distributions, is particularly suited for handling over-dispersed crash data characterized by a large number of zeros and long tails. Vangala develops a generalized linear model (GLM) for the NB-GE distribution and applies it to two over-dispersed crash datasets. Comparisons with Negative Binomial-Lindley (NB-L) and Negative Binomial (NB) models reveal that the NB-GE model performs comparably to NB-L and outperforms the NB model. The study investigates the conditions under which the NB-GE model is recommended over the NB model for ranking crash sites, considering factors such as the percentage of zeros and dispersion in the dataset. Simulated datasets under different scenarios confirm the superior performance of the NB-GE model in cases of high dispersion and a high percentage of zero counts, providing valuable insights for crash data analysis and site ranking decisions.

- In "Exploring the application of the negative binomial–generalized exponential model for analyzing traffic crash data with excess zeros" (2015) [10] P Vangala, D Lord, and SR Geedipally introduce the Negative Binomial–Generalized Exponential distribution (NB–GE) as a tool for analyzing crash data characterized by a large number of zero counts and over-dispersion. This three-parameter distribution combines elements of the Negative Binomial and Generalized Exponential distributions. The authors apply the NB–GE generalized linear model (GLM) to four datasets known for their large dispersion and/or a large number of zeros. Comparisons with the Poisson, Negative Binomial (NB), and Negative Binomial–Lindley (NB–L) models reveal that the NB–GE performs similarly to NB–L but outperforms the traditional NB model, particularly for datasets with significant over-dispersion and a large number of zeros. The study highlights the simplicity of the NB-GE model's modeling framework compared to NB–L, making it a straightforward choice for crash data analysis.

- In their paper, "Marginalized zero‐inflated negative binomial regression with application to dental caries" (2016) [7] JS Preisser, K Das, DL Long, and K Divaris introduce a marginalized zero‐inflated negative binomial regression (MZINB) model aimed at examining relationships between exposures and overdispersed count outcomes with many zeros, a common scenario in dentistry and various other fields. Unlike traditional zero-inflated negative binomial regression models, the MZINB model allows for straightforward inference of overall exposure effects by directly modeling the population marginal mean count through maximum likelihood estimation. Through simulation studies, the authors compare the performance of the MZINB model with other regression models such as marginalized zero-inflated Poisson, Poisson, and negative binomial regression. They then apply the MZINB model to evaluate the impact of a school-based fluoride mouthrinse program on dental caries in a cohort of 677 children. This research contributes to advancing statistical methods for analyzing count data with excess zeros and provides valuable insights into the effects of interventions on dental health.

- In their study, "Empirical Bayes estimates of finite mixture of negative binomial regression models and its application to highway safety" (2018) [14] Y Zou, JE Ash, BJ Park, D Lord, and L Wu address the need to improve safety analyses in transportation by developing a method to incorporate empirical Bayes (EB) estimates with a generalized finite mixture of negative binomial (GFMNB-K) models. While the traditional negative binomial (NB) model is commonly used in safety analyses due to its ability to handle overdispersion in crash data, recent research suggests that GFMNB-K models may offer better statistical performance. However, the EB method has not been applied to GFMNB-K models previously. The authors aim to fill this gap by developing GFMNB-K models with varying weight parameters to analyze crash data from Indiana and Texas. Their findings reveal differences in hotspot identification rankings between NB and GFMNB-2 models, particularly pronounced in the Texas dataset. They recommend future research to conduct a simulation study to determine which model formulation better identifies hotspots.

IV. Conclusion

The negative binomial distribution serves as a cornerstone in statistical analysis, providing a versatile and robust framework for modeling count data across various disciplines. Its capacity to capture variance-to-mean relationships and accommodate complex data structures, such as skewed distributions, excess zeros, and serial dependence, has transformed our approach to hypothesis testing, parameter estimation, and reliable inference. The discussed applications, ranging from seismology to transportation safety and dental health studies, highlight its impact beyond disciplinary confines, driving data-driven discoveries and informing evidence-based decision-making. Ongoing advancements, including the introduction of alternative distributions and refinements in statistical methodologies, promise to further enhance the efficacy and applicability of the negative binomial framework in addressing emerging challenges across diverse fields. As we continue to explore and leverage its potential, the negative binomial distribution remains an indispensable tool in the statistical toolkit, poised to shape the future of data analysis and scientific inquiry.

References

1. Aeberhard, W. H., Cantoni, E., & Heritier, S. (2014). Robust inference in the negative binomial regression model with an application to falls data. Biometrics, 70(4), 920-931. https://doi.org/10.1111/biom.12212

2. Bebbington, M. S., & Lai, C. D. (1998). A generalized negative binomial and applications. Communications in Statistics - Theory and Methods, 27(10), 2515–2533. https://doi.org/10.1080/03610929808832240

3. Hoffman, D. (2003). Negative binomial control limits for count data with extra-Poisson variation. Pharmaceutical Statistics: The Journal of Applied Statistics in the Pharmaceutical Industry, 2(2), 127-132. https://doi.org/10.1002/pst.51

4. Kongrod, S., Bodhisuwan, W., & Payakkapong, P. (2014). The negative binomial-Erlang distribution with applications. International Journal of Pure and Applied Mathematics, 92(3), 389-401. http://dx.doi.org/10.12732/ijpam.v92i3.7

5. Li, S., Yang, F., Famoye, F., Lee, C., & Black, D. (2011). Quasi-negative binomial distribution: Properties and applications. Computational Statistics & Data Analysis, 55(7), 2363-2371. https://doi.org/10.1016/j.csda.2011.02.003)

6. Power, J. H., & Moser, E. B. (1999). Linear model analysis of net catch data using the negative binomial distribution. Canadian Journal of Fisheries and Aquatic Sciences, 56(2), 191-200. https://doi.org/10.1139/f98-150)

7. Preisser, J. S., Das, K., Long, D. L., & Divaris, K. (2016). Marginalized zero-inflated negative binomial regression with application to dental caries. Statistics in medicine, 35(10), 1722-1735. https://doi.org/10.1002/sim.6804

8. Rao, N. M., & Kaila, K. L. (1986). Application of the negative binomial to earthquake occurrences in the Alpide-Himalayan belt. Geophysical Journal International, 85(2), 283-290. https://doi.org/10.1111/j.1365-246X.1986.tb04513.x

9. Robinson, M. D., & Smyth, G. K. (2008). Small-sample estimation of negative binomial dispersion, with applications to SAGE data. Biostatistics, 9(2), 321-332. https://doi.org/10.1093/biostatistics/kxm030

10. Vangala, P., Lord, D., & Geedipally, S. R. (2015). Exploring the application of the negative binomial–generalized exponential model for analyzing traffic crash data with excess zeros. Analytic methods in accident research, 7, 29-36. https://doi.org/10.1016/j.amar.2015.06.001

11. Vangala, Prathyusha (2015). Negative Binomial-Generalized Exponential Distribution: Generalized Linear Model and its Applications. Master's thesis, Texas A & M University. https://hdl.handle.net/1969.1/155124

12. Wang, Z. (2011). One mixed negative binomial distribution with application. Journal of Statistical Planning and Inference, 141(3), 1153-1160. https://doi.org/10.1016/j.jspi.2010.09.020

13. White, G. C., & Bennetts, R. E. (1996). Analysis of frequency count data using the negative binomial distribution. Ecology, 77(8), 2549-2557. https://doi.org/10.2307/2265753

14. Zou, Y., Ash, J. E., Park, B. J., Lord, D., & Wu, L. (2018). Empirical Bayes estimates of finite mixture of negative binomial regression models and its application to highway safety. Journal of Applied Statistics, 45(9), 1652–1669. https://doi.org/10.1080/02664763.2017.1389863