Asian European Journal of Probability and Statistics
Vol. 1, No-2, (2024), pp. 137-151

A Generalization to Size Biased Negative Binomial Distribution and Its Applications in Covid-19 Fertility Rates

C. Satheesh Kumar[a] and Prince Sathyan[b]

[a] *Department of Statistics, University of Kerala, India* [b] *Department of Statistics, University of Kerala, India*

ARTICLE HISTORY
Compiled April 12, 2025

Received 10 February 2024; Accepted 02 September 2024

ABSTRACT
In this study, we introduce a size biased version of the negative binomial distribution named as generalized size biased negative binomial distribution and demonstrate its applicability by fitting it to COVID-19 data sets. We derive several key properties of the distribution, including the probability generating function, cumulative distribution function, survival and hazard rate functions, along with recurrence relations for probabilities. Additionally, we explore parameter estimation methods and develop statistical tests to assess the significance of the distribution's parameters. Furthermore, a simulation study is conducted to evaluate the performance of the parameter estimators obtained using the maximum likelihood method.

KEYWORDS
count data modeling; maximum likelihood estimation; MCMC simulation; model selection; negative binomial distribution; size biased model; survival function

1. Introduction

The Negative Binomial Distribution (NBD) is a commonly utilized discrete probability distribution for modeling count data, particularly in cases where overdispersion occurs, meaning that the variance exceeds the mean. It provides a more flexible alternative to the Poisson distribution by incorporating a dispersion parameter, allowing for greater variability in observed data. Due to this adaptability, the NBD is widely applied in fields such as epidemiology, finance, insurance, and ecology to analyze count-based phenomena, including disease incidence, insurance claims, and species abundance Hilbe [3]; Cameron and Trivedi [1]. The distribution is often conceptualized as a gamma-Poisson mixture, where the Poisson rate parameter follows a gamma distribution, effectively capturing heterogeneity in real-world data. It has certain limitations, one major drawback is its inability to model underdispersed data, where the variance is lower than the mean, restricting its applicability in some cases Ridout et. al[9]. Furthermore, the assumption that the underlying heterogeneity follows a gamma distribution may not always be suitable for empirical data, potentially leading to model misfit. To overcome these challenges, researchers have proposed

CONTACT C. Satheesh Kumar[a]. Email: drcsatheeshkumar@gmail.com

modifications such as weighted, zero-inflated, or generalized variations of the NBD, which offer improved flexibility in modeling complex data patterns Zhang and Peleato [10]. Considering these factors, further exploration of alternative extensions of the NBD is necessary to enhance its effectiveness across diverse statistical modeling applications.
Size-biased distributions are class of probability models where the probability of observing a particular outcome is proportional to its size or magnitude, leading to a bias toward larger values. This approach is particularly useful in fields where larger units are more likely to be sampled or observed, such as ecology, actuarial science, and industrial reliability. For instance, in forestry, larger trees are more likely to be included in a sample because they occupy more space, making a size-biased model more representative of the actual sampling process. Similarly, in insurance, claims with higher amounts are more likely to be reported or noticed, necessitating size-biased modeling. These distributions are especially relevant when analyzing overdispersed or skewed count data, as they can account for unequal probabilities of selection due to size or weight. However, care must be taken when interpreting results, as size bias can distort parameter estimates if not properly accounted for during analysis or inference Patil and Rao [7].

In medical sciences, weighted distributions help analyze disease incidence rates and survival data, particularly in cases where patients are selected based on pre-existing conditions Rao [8]. In ecology and environmental studies, these distributions are applied to model species abundance, where larger or more prominent species have a higher probability of being sampled Patil [6]. Additionally, in insurance and actuarial science, size-biased and length-biased distributions are used to assess claim sizes and risk factors, ensuring more accurate premium calculations Cox [2]. Furthermore, in linguistics and bibliometrics, weighted models assist in studying word frequency distributions and citation analysis, where certain words or articles are disproportionately represented due to underlying selection mechanisms , Johnson et al. [4]. These applications highlight the versatility and necessity of discrete weighted distributions in handling biased or preferentially sampled data across diverse disciplines

In this article, we propose a weighted version of the Negative Binomial Distribution (NBD), referred to as the "generalized size biased negative binomial distribution (GSNBD)", and investigate its key statistical properties. Section 2 introduces the definition of the GSNBD and derives essential functions, including the probability generating function (p.g.f.), cumulative distribution function (c.d.f.), survival function, and hazard rate function. Additionally, we obtain expressions for recurrence relations for its probabilities. In Section 3, we focus on the estimation of GSNBD parameters using the method of maximum likelihood. Moreover, Section 4 presents statistical test procedures for evaluating the significance of the distribution's parameters. To demonstrate the practical utility of the GSNBD, Section 5 applies the model to Covid-19 mortality rate data, illustrating both the parameter estimation and hypothesis testing methods discussed in Section 4. Finally, Section 6 includes a simulation study to assess the performance of the maximum likelihood estimators (MLEs), offering insights into their accuracy and efficiency.

For any real numbers a, b and z such that $z \neq 0, -1, -2, \ldots$, the Gauss hypergeometric function (as well as confluent hypergeometric function are respectively) is defined

as in the following, for $|t| < 1$.

$$ {}_2F_1(a, b, z; t) = \sum_{r=0}^{\infty} \frac{(a)_r (b)_r t^r}{(z)_r r!} \tag{1} $$

and

$$ {}_1F_0(a; -; t) = \sum_{r=0}^{\infty} \frac{(a)_r t^r}{r!} \tag{2} $$

in which

$$ (a)_r = a(a+1)(a+2)\dots(a+r-1) = \frac{\Gamma(a+r)}{\Gamma(a)} \tag{3} $$

for $r = 1, 2, \dots$ and $(a)_0 = 1$, is the Pochhammer's symbol. For further details regarding the hypergeometric function see Mathai and Haubold [5].

2. Definition and Properties of GSNBD

Here, first we present the definition of the GSNBD.

Definition 2.1. A non-negative integer-valued random variable Y is said to follow a generalized size biased negative binomial distribution (GSNBD) if its probability mass function (p.m.f), denoted by $f_Y(.)$, is defined for $y = m,\ m+1,\ m+2, ...$; with parameters $r > 0$, $0 < p < 1$ and $q = 1 - p$.

$$ f_Y(y) = P(Y = y) = \frac{(y + r - 1)!\ p^{r+m}\ q^{y-m}}{(r + m - 1)!\ (y - m)!} \tag{4} $$

A graphical representation for various shapes of the p.m.f of the GSNBD for parameters values of its parameter are given in Figure 1, 2 and 3. These figures show that the p.m.f of the GSNBD can be right-skewed, symmetric, or decreasing curves. Next we present certain properties of the GSNBD through the following results.

Proposition 2.2. *The probability generating function (p.g.f) $G(t)$ of the G.,GSNBD with p.m.f (4) is given by*

$$ G(t) = p^{r+m}\ t^m {}_1F_0(m + r; -; qt). \tag{5} $$

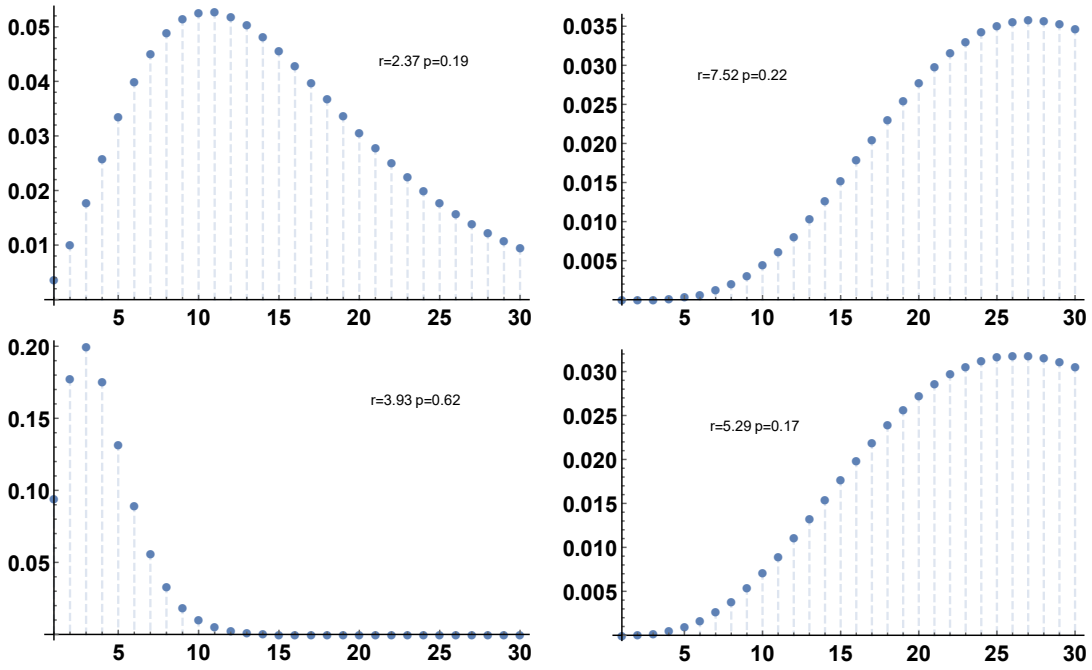

Figure 1. Plots of probability mass functions of GSNBD, when m=1 and different values of parameters

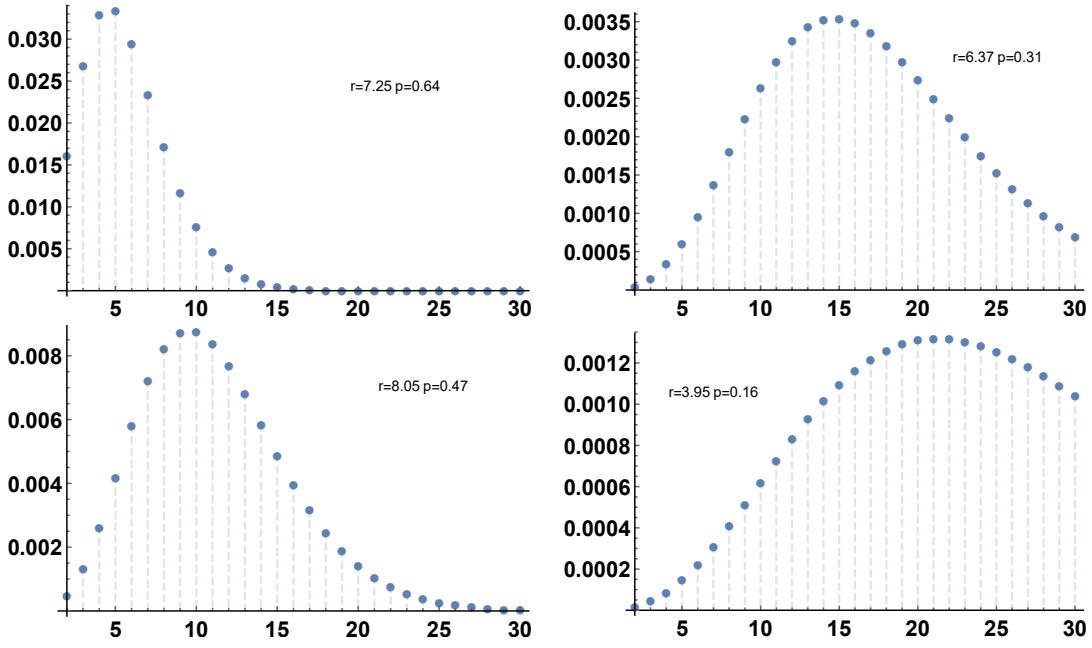

Figure 2. Plots of probability mass functions of GSNBD, when m=2 and different values of parameters

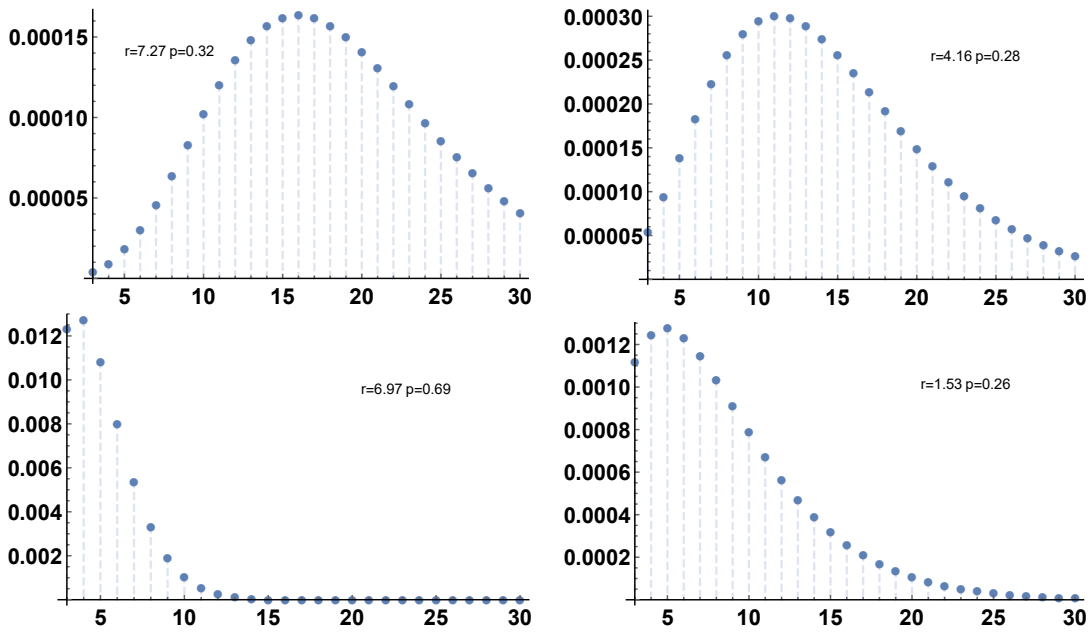

Figure 3. Plots of probability mass functions of GNBD, when m=3 and different values of parameters

Proof. By definition, the p.g.f of the GSNBD with p.m.f (4) is given by

$$\begin{aligned} G(t) &= \sum_{y=m}^{\infty} f_Y(y)t^y \\ &= \sum_{y=m}^{\infty} \frac{(y+r-1)!\ p^{r+m}\ q^{y-m}}{(r+m-1)!\ (y-m)!}t^y \qquad (6) \\ &= \frac{p^{r+m}}{(r+m-1)!}\sum_{y=0}^{\infty} \frac{(y+m+r-1)!\ q^y}{y!}t^{y+m}, \qquad (7) \end{aligned}$$

Now applying (3) in (7) to obtain

$$G(t) = p^{r+m}t^m \sum_{y=0}^{\infty} \frac{(m+r)_y\ q^y}{y!}t^{y+m}, \qquad (8)$$

which implies (5) in the light of (2). □

Proposition 2.3. *The cumulative distribution function (c.d.f) $F_Y(n)$ of the GSNBD with p.m.f (4) is the following, for any $n \in \Re = (-\infty, \infty)$.*

$$F_Y(n) = F_Y(n) = 1 - \frac{q^{n+1-m}\ p^{r+m}\ (n+r)!}{(r+m-1)!\ (n-m+1)!}{}_2F_1(1, n+r; n+2-m; q) \qquad (9)$$

Proof. By definition, the c.d.f of the GSNBD with p.m.f (4) is given by

$$\begin{aligned} F_Y(n) &= P(Y \leq n) \\ &= \sum_{y=m}^{n} \frac{(y+r-1)!\ p^{r+m}\ q^{y-m}}{(r+m-1)!\ (y-m)!} \\ &= 1 - \sum_{y=n+1}^{\infty} \frac{(y+r-1)!\ p^{r+m}\ q^{y-m}}{(r+m-1)!\ (y-m)!} \\ &= 1 - \sum_{y=0}^{\infty} \frac{(y+n+r)!\ p^{r+m}\ q^{y+n+1-m}}{(r+m-1)!\ (y+n+1-m)!} \\ &= 1 - \frac{q^{n+1-m}\ p^{r+m}}{(r+m-1)!} \sum_{y=0}^{\infty} \frac{(y+n+r)!\ q^{y}}{(y+n+1-m)!}. \end{aligned} \tag{10}$$

In the light of (3) in (10) to obtain

$$F_Y(n) = 1 - \frac{q^{n+1-m}\ p^{r+m}\ (n+r)!}{(r+m-1)!\ (n-m+1)!} \sum_{y=0}^{\infty} \frac{(n+r)_y\ q^{y}}{(n+2-m)_y} \tag{11}$$

which leads to (9) by using (1). □

Proposition 2.4. *The survival function $S(.)$ and hazard rate function $h(.)$ of the GNBD are respectively, for any $t \in \Re$*

$$S(t) = \frac{q^{t+1-m}\ p^{r+m}\ (t+r)!}{(r+m-1)!\ (t-m+1)!} {}_2F_1(1, t+r; t+2-m; q) \tag{12}$$

and

$$h(t) = \frac{1}{{}_2F_1(1, t+r-1; t+1-m; q)}. \tag{13}$$

Proof following for the definition of $S(.)$ and $h(.)$ on

$$S(t) = 1 - P(Y > t)$$

and

$$h(t) = \frac{P(Y = t)}{P(Y > t-1)}.$$

By using Proposition 2, we have computed measures of skewness and kurtosis with the help of Mathematica software and plotted the values in Figure 4 and Figure 5. From the figures, it can be seen that the distribution enjoys positive and negative skewed behaviour as well as both platykurtic and leptokurtic nature. In the light of

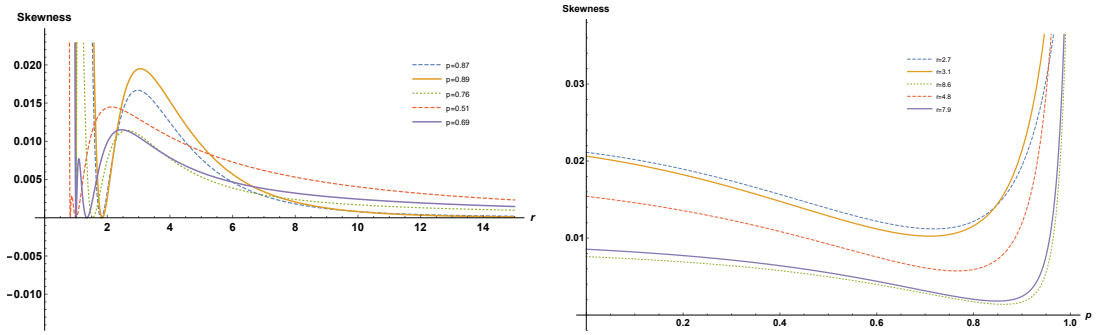

Figure 4. Plots of skewness of GNBD for particular values of its parameters.

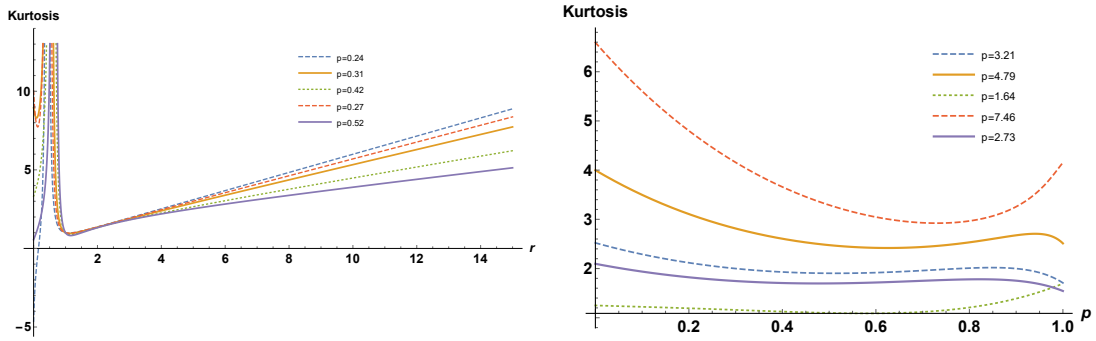

Figure 5. Plots of kurtosis of GSNBD for particular values of its parameters.

Proposition 2, we have the following important result, which depicts the nature of dispersion of the distribution.

Proposition 2.5. *The GSNBD over dispersed when*

$$\frac{7+(r-2)(1-p)(6+(r-3)(1-p))}{p^2} - \frac{p^2}{r(r+1)(1-p)^2}(\nu+\rho)^2 > \rho \tag{14}$$

where $\rho = \frac{r(1-p)(1+(r-1)(1-p))}{p^2}$ *and* $\nu = \frac{r(1-p)(1+(r-1)(1-p)(3+(r-2)(1-p)))}{p^3}$

Next we derive certain recursion formula for probabilities of the GNBD.

Proposition 2.6. *The following is a simple recursion formula for the probabilities* $f_Y(y)$ *of the GSNBD with p.m.f. (4) and is given by*

$$f_y(y+1) = \frac{(y+r)\ q}{(y+1-m)} f_Y(y) \tag{15}$$

Proof. From (4) we have the following:

$$f_Y(y) = \frac{(y+r-1)!\ p^{r+m}\ q^{y-m}}{(r+m-1)!\ (y-m)!} \tag{16}$$

$$f_Y(y+1) = \frac{(y+r)!\ p^{r+m}\ q^{y+1-m}}{(r+m-1)!\ (y+1-m)!} \tag{17}$$

Now, equation (16) and (17) together leads to

$$\frac{f_Y(y+1)}{f_Y(y)} = \frac{\frac{(y+r)!\ p^{r+m}\ q^{y+1-m}}{(r+m-1)!\ (y+1-m)!}}{\frac{(y+r-1)!\ p^{r+m}\ q^{y-m}}{(r+m-1)!\ (y-m)!}} \tag{18}$$

which implies (15). □

3. Estimation of the parameters

In this section, we discuss the estimation of the parameters of the GSNBD by the method of maximum likelihood.

3.1. Method of Maximum Likelihood

Here, we explore the maximum likelihood estimation method for determining the parameters r and p of the GSNBD.

Let $a(y)$ represent the observed frequency of y events for any $y = m, m+1, m+2\ldots$, and let z be the highest observed value of y. Then, the likelihood function corresponding to the given sample is:

$$L(\Theta; y) = \prod_{y=m}^{z} [f_Y(y)]^{a(y)}, \tag{19}$$

in which $f_Y(y)$ is the p.m.f of the GSNBD as given in (4). Now taking logarithm on both sides of (19), we have the following log-likelihood function.

$$l = \ln L(\Theta; y) = \sum_{y=m}^{z} a(y)[ln(y+r-1)! + (r+m)lnp + (y-m)ln(1-p) \\ -ln(r+m-1)! - ln(y-m)!] \tag{20}$$

Let $\hat{r}$ and $\widehat{p}$ denote the maximum likelihood estimators of the parameters r and p of the GSNBD. On differentiating the log-likelihood function (20) with respect to the parameters r and p and equating to zero, we obtain the following likelihood equations. In which $\varphi(\gamma + r) = \frac{\partial}{\partial r} \ln(\gamma + r)!$

$$\frac{\partial l}{\partial r} = 0$$

implies

$$\sum_{y=m}^{z} a(y) \{\varphi(y+r-1) + lnp - \varphi(m+r-1)\} = 0 \tag{21}$$

$$\frac{\partial l}{\partial p} = 0$$

implies

$$\sum_{y=m}^{z} a(y) \left\{ \frac{r+m}{p} + \frac{m-y}{1-p} \right\} = 0 \tag{22}$$

Obtaining explicit expressions for the parameters of the GSNBD through maximum likelihood estimation (MLE) is challenging. The likelihood equations (21)and (22) do not always have solutions, as the GSNBD is not a regular statistical model. In instances where these equations fail to produce a solution, the maximum of the likelihood function occurs at the boundary of the parameter space.

To tackle this issue, we calculated the second-order partial derivatives of $f_Y(y)$ concerning the parameters r and p. Using MATHEMATICA software, we confirmed that these derivatives are negative for all $r > 0$ and $0 < p < 1$. This finding demonstrates that the p.m.f. of the GSNBD is log-concave, ensuring that the maximum likelihood estimators $\hat{r}$ and $\widehat{p}$ remain unique within these parameter constraints. As a result, the MLEs for the GNBD parameters can be obtained by solving the system of equations (21) and (22) using computational tools like MATHEMATICA.

4. Testing of Hypothesis

In this section, we examine three test procedures designed to assess the significance of the parameter "r" in the GSNBD, whose probability mass function (p.m.f) is given by equation (4), as outlined below.

4.1. Generalized Likelihood Ratio Test

To evaluate the significance of the parameter r in the GSNBD, we employ the generalized likelihood ratio test (GLRT) procedure as described below.

Consider the null hypothesis $H_0 : r = 1$ against the alternative hypothesis $H_1 : r \neq 1$. In the case of the generalized likelihood ratio test (GLRT), the corresponding test statistic is

$$-2 \log \Delta = 2 \left(\log L(\hat{\Theta}; y) - \log L(\hat{\Theta}^*; y) \right), \tag{23}$$

Here, $\hat{\Theta}$ denotes the maximum likelihood estimator of $\Theta = (r, p)$ without any constraints, while $\hat{\Theta}^*$ represents the maximum likelihood estimator of Θ under the condition $r = 1$. The test statistic $-2 \log \Delta$, as stated in (23), asymptotically follows a chi-square distribution with one degree of freedom. For more details, see Rao (1947).

4.2. Rao's Efficient Score Test

Here, we investigate Rao's efficient score test (REST) to determine the significance of the parameter r in the GSNBD.

Let the null hypothesis be $H_0 : r = 1$ against the alternative hypothesis $H_1 : r \neq 1$.

In case of the Rao's efficient score test, the test statistic is

$$S = T'\Phi^{-1}T, \tag{24}$$

where

$$T' = \left(\frac{1}{\sqrt{n}}\frac{\partial \log L}{\partial q}, \frac{1}{\sqrt{n}}\frac{\partial \log L}{\partial r}, \frac{1}{\sqrt{n}}\frac{\partial \log L}{\partial \delta}\right),$$

Here, Φ represents the Fisher information matrix. The test statistic S, as given in (24), follows an asymptotic chi-square distribution with one degree of freedom (df). For further details on REST, refer to Rao (1965).

4.3. Wald's Test

Here, we employ Wald's test to assess the significance of the parameter r in the GSNBD. The null hypothesis is stated as

$H_0 : r = 1$ against the alternative hypothesis $H_1 : r \neq 1$.

The test statistic is given by

$$W_r = \frac{\hat{r}^2}{\hat{\text{Var}}(\hat{r})}, \tag{25}$$

Here, $\text{Var}(\hat{r})$ denotes the corresponding diagonal element of the Fisher information matrix, evaluated at $r = \hat{r}, p = \hat{p}$. The test statistic, as given in (25), asymptotically follows a chi-square distribution with one degree of freedom.

5. Applications

For numerical illustration, we have analyzed real-life data sets on COVID-19 mortality rates from various districts in Kerala. These data sets were sourced from the official website of the Directorate of Health Services, Kerala State, India (https://dhs.kerala.gov.in). Data Set-1 includes COVID-19 mortality counts from Trivandrum and Kollam districts during November 2020 – January 2021 and data Set-2 consists of mortality counts from Malappuram and Kozhikode districts during December 2020 – January 2021. We fitted the truncated Poisson distribution(TPD), truncated negative binomial distribution (TNPD), size- biased Poisson distribution(SPD), truncated alternative hyper-Poisson distribution (TAHPD) and generalized size biased negative binomial distribution (GSNBD) models to all these data sets. The results, including the expected frequencies, chi-square statistic, degrees of freedom (d.f.), P-value, AIC, BIC, AICc, and dispersion index for each model, are presented in Tables 1 and 2 respectively. Based on the computed values of the chi-square statistic, P-value, AIC, BIC, and AICc, it is evident that the GNBD model provides the best fit for all data sets, whereas the existing models, TPD, TNPD, WPD and TAHPD fail to do so.

We have also plotted the observed frequency curves of the data sets along with the fitted densities for the TPD, TNPD, WPD, TAHPD and GSNBD models. From Tables

Table 1. Distribution of Covid death of Trivandrum and Kollam (November 2020 - January 2021) and the expected frequencies computed using TPD, TNPD, SPD, TAHPD and GSNBD of Data Set-1.

X	Observed frequency	TPD	TNBD	SPD	TAHPD	GSNBD
2	13	8.05	26.02	3.83	18.88	18.35
3	14	13.87	22.95	12.27	19.55	17.21
4	12	16.85	16.81	19.35	17.54	10.95
5	13	16.87	11.17	20.51	13.87	9.86
6	17	14.03	6.87	16.31	9.56	9.07
7	8	10.03	3.97	10.37	6.03	8.69
8	6	6.27	2.18	5.49	3.46	7.22
9	4	3.49	1.15	2.51	1.82	5.02
10	2	1.75	0.59	0.99	0.89	3.01
11	3	0.79	0.29	0.37	0.40	2.62
Total	92	92	92	92	92	92
df		6	3	4	4	6
Estimates		$\lambda = 5.004$	p=0.64196 r=5.2929	$\lambda = 3.1798$	$\lambda = 4.2402$ $\theta = 6.4595$	$p = 5.822$ $r = 0.802$
χ^2-value		12.673	73.16	48.968	34.405	11.054
P-value		0.049	0	$5.928*10^{-10}$	$6.154*10^{-7}$	0.086
AIC		183.612	444.741	425.322	499.402	243.024
BIC		186.134	449.785	427.844	504.445	248.648
AICc		183.656	444.876	425.366	499.537	243.124

Table 2. Distribution of Covid death of Ernakulam and Thrissur (February 2021) and the expected frequencies computed using TPD, TNPD, SPD, TAHPD and GSNBD of Data Set-1.

X	Observed frequency	TPD	TNBD	SPD	TAHPD	GSNBD
1	4	3.53	7.38	2.46	7.56	5.06
2	7	5.83	5.16	5.99	5.73	6.31
3	5	6.42	5.25	7.28	5.40	6.04
4	5	5.31	2.65	5.89	4.56	4.59
5	3	3.58	3.12	3.57	2.29	2.05
6	2	1.93	2.61	1.79	1.01	1.82
7	1	0.90	1.05	0.71	1.09	1.02
8	0	0.37	0.53	0.24	0.13	0.84
9	1	0.13	0.25	0.07	0.23	0.27
Total	28	28	28	28	28	28
df		2	1	2	1	1
Estimates		$\lambda = 3.302$	p=0.772 r=10.924	$\lambda = 2.428$	$\lambda = 2.792$ $\theta = 4.766$	$p = 3.279$ $\theta = 0.637$
χ^2-value		6.929	7.229	14.812	5.935	3.785
P-value		0.031	0.007	0.0001	0.014	0.051
AIC		112.56	117.804	121.804	111.804	109.804
BIC		113.893	120.469	123.137	113.137	111.137
AICc		112.714	118.284	121.958	111.958	109.958

1, 2 and Figures 6, 7, it is clear that none of these models provide the best fit to the data sets, except for the GSNBD. The GSNBD model emerges as the best fit based on the P-value and Chi-square statistic. Furthermore, the information criteria measures, including AIC, BIC, and AICc, further support the conclusion that GSNBD is a more appropriate model compared to the other models analyzed in this study.

Using the data sets provided in Table 1 and 2, we have computed the test statistic values for the GLRT, REST, and Wald Test, which are presented in Table 3. Since the critical value for the test at a 5% level of significance is 3.84 with one degree of freedom, the null hypothesis is rejected in all cases for the GLRT, REST, and Wald Test.

Table 3. Calculated values of the test statistic for GLRT, REST and Wald test for GNBD

	Calculated values of test statistic		
	GLRT	REST	Wald's Test
Data set 1	13.628	12.836	14.728
Data set 2	6.783	4.737	7.115

Table 4. Actual Bias and standard error(with in brackets)of parameters of GSNBD computed by MLE in case of the underdispersed situation corresponds to m= 1 and 2 using simulated datas for Parameter Set (r, p).

			MLE	
m	parameter set	Sample size	$\hat{r}$	$\hat{p}$
1	(r=3.29, p=0.81)	50	-0.7943 (0.63873)	0.83947 (0.73883)
		100	-0.62882 (0.28839)	0.437872 (0.37820)
		200	-0.08773 (0.07221)	0.092887 (0.09272)
		500	-0.04772 (0.018773)	0.062832 (0.027282)
		1000	-0.0074838 (0.0048729)	0.008383 (0.005823)
2	(r=1.17, p=0.62)	50	-0.73920 (0.38291)	0.82991 (0.28739)
		100	-0.28930 (0.083783)	0.39820 (0.072281)
		200	-0.073722 (0.036282)	0.06993 (0.028382)
		500	-0.01738 (0.0093883)	0.009738 (0.0083773)
		1000	-0.0083927 (0.0038392)	0.0028789 (0.00283921)

6. Simulation

In this section, we simulate random variates from the GSNBD and evaluate the bias and standard errors of the distribution's parameter estimators using the maximum likelihood method. Two sets of observations are simulated for $m = 1$ and 2 considering sample sizes of 50, 100, 200, 500, and 1000 under both overdispersed and underdispersed conditions. The results are presented in Tables 6 and 6.

Using these simulated observations, we estimated the parameters r and p of the GSNBD and subsequently computed the absolute bias and standard errors for each estimator. From Tables 6 and 6, it is evident that as the sample size increases, both the absolute bias and standard errors of the parameter estimators decrease.

Table 5. Actual Bias and standard error(with in brackets)of parameters of GSNBD computed by MLE in case of the overdispersed situation corresponds to m= 1 and 2 using simulated datas for Parameter Set (r, p).

			MLE	
m	parameter set	Sample size	$\hat{r}$	$\hat{p}$
1	(r=7.43, p=0.47)	50	-0.36282 (0.92838)	0.73283 (0.489832)
		100	-0.19737 (0.59033)	0.28830 (0.29949)
		200	-0.09493 (0.29822)	0.084939 (0.19392)
		500	-0.038742 (0.0084633)	0.0188749 (0.0064846)
		1000	-0.009768 (0.002738)	0.007382 (0.0037473)
2	(r=3.85, p=0.61)	50	-0.47728 (0.738822)	0.287783 (0.673228)
		100	-0.18739 (0.281383)	0.09739 (0.28829)
		200	-0.08378 (0.078729)	0.063872 (0.057182)
		500	-0.017383 (0.03812)	0.026783 (0.0087628)
		1000	-0.0076228 (0.0046289)	0.001232 (0.0017382)

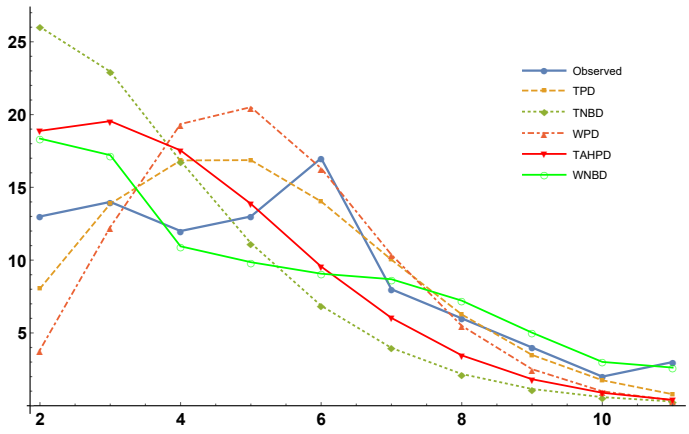

Figure 6. Frequency curves corresponding to various models based on data set 1

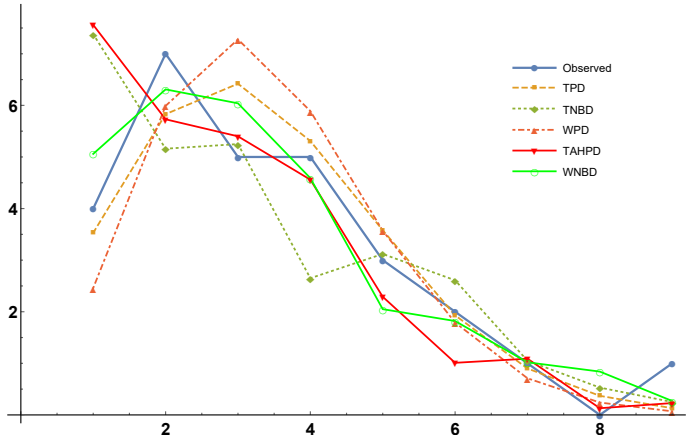

Figure 7. Frequency curves corresponding to various models based on data set 2

7. Summary and Conclusion

This study aims to develop an improved discrete statistical model for analyzing COVID-19 infection death rates in India during 2020–2021. To provide a more effective framework for modeling infection-related fatalities, we constructed an enhanced statistical model. The proposed model is a two-parameter extension formulated as a size biased version of the negative binomial distribution named as generalized size biased negative binomial distribution. We explored various statistical properties of the model, deriving expressions for its probability generating function, cumulative distribution function, survival and hazard rate functions, we established recursion formula for probabilities. Parameter estimation was performed using the maximum likelihood method. Furthermore, we developed specific test procedures to assess the significance of the additional parameter in the proposed distribution class.. The applicability of the generalized negative binomial was demonstrated by modeling COVID-19 mortality data from different districts in Kerala, India. Additionally, a simulation study was conducted to evaluate the performance of the maximum likelihood estimators for the model's parameters.

Acknowledgments

The authors are highly thankful to both the referees for their fruitful comments, which immensely helped to improve the quality and presentation of the article.

References

[1] Cameron, A. C. and Trivedi, P. K. (2013). *Regression analysis of count data.* Number 53. Cambridge university press.
[2] Cox, D. (2005). Some sampling problems in technology. *Selected Statistical Papers of Sir David Cox*, 1:81–92.
[3] Hilbe, J. M. (2011). *Negative binomial regression.* Cambridge University Press.
[4] Johnson, N. L., Kemp, A. W., and Kotz, S. (2005). *Univariate discrete distributions.* John Wiley & Sons.
[5] Mathai, A. M. and Haubold, H. J. (2008). Special functions for applied scientists.

[6] Patil, G. (2002). Weighted distributions,encyclopedia of environmetrics, vol. 4.
[7] Patil, G. P. and Rao, C. R. (1978). Weighted distributions and size-biased sampling with applications to wildlife populations and human families. *Biometrics*, pages 179–189.
[8] Rao, C. R. (1965). On discrete distributions arising out of methods of ascertainment. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 311–324.
[9] Ridout, M., Hinde, J., and Demétrio, C. G. (2001). A score test for testing a zero-inflated poisson regression model against zero-inflated negative binomial alternatives. *Biometrics*, 57(1):219–223.
[10] Zhang, Y. and Peleato, N. M. (2021). Predicting cyanobacteria abundance with bayesian zero-inflated negative binomial models. *Available at SSRN 3939421*.