Thailand Statistician
January 2025; 23(1): 64-71
http://statassoc.or.th
Contributed paper

Further Statistical Properties of Size-Biased Geometric Distribution

Nuri Celik [a]* and Iklim Gedik Balay [b]
[a] Department of Mathematics/Gebze Technical University, Kocaeli, Turkey
[b] Department of Business and Finance, Ankara Yildirim Beyazit University, Ankara, Turkey
*Corresponding author; e-mail: nuricelik@gtu.edu.tr

Received: 19 January 2022
Revised: 29 July 2022
Accepted: 4 September 2022

Abstract

In this article, size biased geometric distribution is introduced. Some of its statistical and reliability properties are discussed like probability function, cumulative distribution and reliability function, hazard rate function and moments. The unknown distribution parameter is obtained by the moments and the maximum likelihood method in a closed form. Monte Carlo simulation study is performed in order to investigate the performance of maximum likelihood estimators. Additionally, this distribution is evaluated for the new count regression model. Real life data which comes from the statistical literature are considered for the illustration and the results are compared with Poisson, geometric, discrete Weibull, discrete Gamma and discrete Lindley distributions.

Keywords: Discrete random variables, count regression, discrete Gamma, discrete Weibull.

1. Introduction

In statistics literature, the plenty of discrete distribution have been introduced in order to model some random experiments. In recent years, it has been seen that, these traditional distributions have limited application areas especially in reliability, failure times and count models. For this reason, some authors have proposed certain methods for generating new discrete distributions from the well-known continuous distributions. Roy (2003) first proposed the method for the discretization of a given continuous random variable. This method is based on the survival function of the continuous random variable X and stich with the same functional form of the survival function as that of the continuous one. Therefore, many characterization of these new discrete distributions remain unchanged. Chakraborty and Chakravarty (2012) summarized the discretization of many well-known distributions by using this method, such as discrete Weibull distribution (Nakagawa and Osaki, 1975, Stein and Dattero, 1984, Khan et al., 1989 and Kulasekera, 1994) , discrete normal distribution (Roy, 2003), discrete Rayleigh distribution (Roy, 2004), discrete Burr (Krishna and Pundir, 2009; Khorashadizadeh et al., 2013), discrete Pareto distributions (Krishna and Pundir, 2009) and discrete inverse Weibull distribution (Jazi et al., 2009).

On the other hand, some authors proposed new discrete distributions by using other methods. Baker (2000) introduced a partial probability function and applied this new distribution to the epidemiology. Deniz et al. (2011) proposed another probability mass function as an alternative to the negative binomial, Poisson-inverse Gaussian, hyper-Poisson and generalized Poisson distributions

and used in actuarial application area. Zheng (2010) studied a new model based on the limiting distribution induced by Bartlett's formulation of the LuriaDelbrck mutation model. El-Desouky et al. (2021) presented a new class of distribution obtained by compounding generalized HermiteGenocchi polynomials given by Gould and Hopper with the powers series distribution.

It can be easily noticed that, there have been a great interest to new discrete distributions. However, many of these new introduced distributions have limited application areas and have the same characterization with the related continuous distributions. Some of them, on the other hand, have some mathematical restrictions and difficulties especially in moments and cumulative distribution functions (cdf). Most of these distributions have two or more unknown parameters which causes estimation and fitting problems. In this paper, we deal with size-biased geometric distribution and some further statistical properties of it. The probability function is

$$P(X = x) = (1-\theta)^2 x\theta^{x-1}; \;\; x = 1, 2, 3, ... \tag{1}$$

where θ is unknown parameter with $0 < \theta < 1$. Since

$$\sum_{i=1}^{\infty} x_i \theta^{x_i - 1} = \frac{1}{(1-\theta)^2}, \quad \text{for } \; 0 < \theta < 1.$$

The rest of the paper organizes as follow. In Section 2, we give some mathematical and statistical properties of the size biased geometric distribution such as the probability function, cdf, moments, survival and hazard function and maximum likelihood estimation. The regression model based on size biased gometric distribution is studied in the next section. In Section 4, a Monte Carlo simulation study is conducted and in Section 5, we give some real life data set in order to show the performance of this distribution corresponding to the other traditional distributions. Conclusion is given at the end of this paper.

2. Statistical Properties

In this section, we give the properties of the size biased gometric distribution. Firstly, Figure 1 shows some possible shapes of Equation (1). We can see that the probability mass function is always unimodal.

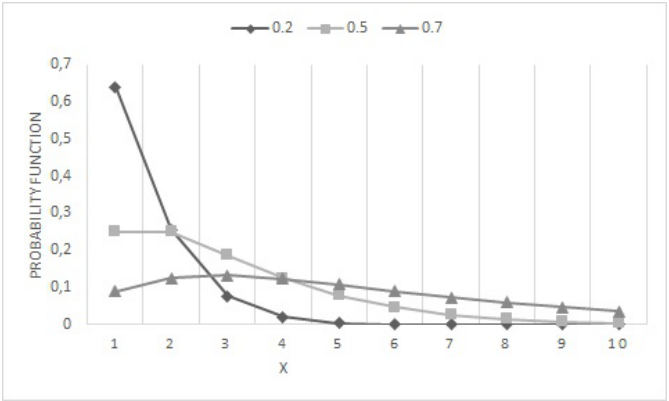


Figure 1 Probability mass function for θ=0.2, 0.5 and 0.7

2.1. Moments

Let X be a new discrete random variable. Then, the probability generating function of X can be expressed as

$$M(t) = E(e^{tx}) = \sum_{i=1}^{n} e^{tx_i} x_i \theta^{x_i - 1}(1-\theta)^2. \tag{2}$$

Equation (2) can be solved as

$$M(t) = \frac{(1-\theta)^2}{\theta} \sum_{i=0}^{n} x_i \Big(\theta e^t\Big)^{x_i}.$$

Setting $y = \theta e^t$ then,

$$M(t) = \frac{(1-\theta)^2}{\theta} \frac{\theta e^t}{(1-\theta e^t)^2}.$$

Therefore the first four moment of the distribution are

$$E(X) = \frac{1+\theta}{1-\theta}, \quad E(X^2) = \frac{\theta^2 + 4\theta + 1}{(1-\theta)^2},$$
$$E(X^3) = \frac{\theta^3 + 10\theta^2 + 12\theta + 1}{(1-\theta)^3},$$
$$E(X^4) = \frac{\theta^4 + 24\theta^3 + 66\theta^2 + 28\theta + 1}{(1-\theta)^4},$$

respectively. Then the variance can be obtained as

$$V(X) = \frac{2\theta}{(1-\theta)^2}.$$

To better understand the shape of the distribution, see the coefficients of expected values, variances and skewness (γ_1) and the kurtosis (γ_2) for some representative values of θ given in Table 1. It is clear from Table 1 that the skewness and the kurtosis of the distribution decreases as the θ increases.

Table 1 The expected values, the variances, the skewness and the kurtosis values

θ	0.1	0.2	0.5	0.7	0.9
$E(X)$	1.222	1.500	3.000	5.666	19.00
$V(X)$	0.247	0.650	4.000	15.55	180.0
γ_1	5.663	1.852	1.208	1.009	0.976
γ_2	37.59	3.262	1.215	0.666	0.489

2.2. Cumulative distribution function

The cdf of the proposed distribution can be found as

$$F(b) = P(X \le b) = \sum_{x=1}^{b} (1-\theta)^2 x \theta^{x-1}; \;\; x = 1, 2, 3, ...$$

In order to find the cdf the simple fact which is commonly used in serial representation can be used. Let $S_n = \sum_{x=1}^{n} (1-\theta)^2 x \theta^{x-1}$. Therefore,

$$S_n - \theta S_n = (1-\theta)\Big(\frac{1-\theta^n}{1-\theta} - n\theta^n\Big). \tag{3}$$

Hence, the cdf can be written as

$$F(x) = P(X \le x) = (1-\theta^x) - x(1-\theta)\theta^x; \;\; x = 1, 2, 3, ...$$

2.3. Survival function and hazard rate

The survival function (sf) and the hazard rate function of the distribution can be obtained

$$S(x) = P(X > x) = 1 - \Big((1-\theta^x) - x(1-\theta)\theta^x\Big); \quad x = 1, 2, 3, ...$$

$$h(x) = \frac{(1-\theta)^2 x\theta^{x-1}}{1 - \Big((1-\theta^x) - x(1-\theta)\theta^x\Big)}; \quad x = 1, 2, 3, ...,$$

respectively. Figure 3 shows the shapes of the hazard rate function for some representative values of θ. As can be noticed from the Figure 2, the failure rate function has always increasing form for all θ values.

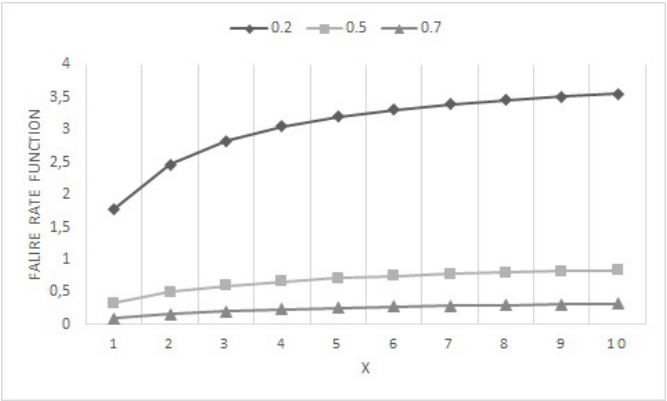


Figure 2 Failure rate function for θ=0.2, 0.5 and 0.7

2.4. Estimation

In this section we obtain the estimators of the unknown parameter θ by using moments and maximum likelihood method. The estimator of θ by using moments method can be found as

$$\frac{1+\theta}{1-\theta} = \frac{\sum_{i=1}^{n} X_i}{n}, \quad \hat{\theta} = \frac{\bar{X}-1}{\bar{X}+1}. \tag{4}$$

In order to obtain the ML estimator of the unknown parameter, the log-likelihood function of θ is

$$\log L(\theta) = \sum_{i=1}^{n} \ln X_i + \ln(\theta) \sum_{i=1}^{n} (X_i - 1) + 2n \ln(1-\theta).$$

Taking the derivative of the log-likelihood function with respect to the unknown parameter θ and equating it to the zero, we obtain the following equation

$$\frac{\partial \ln L}{\partial \theta} = \frac{\sum_{i=1}^{n} (X_i - 1)}{\theta} - \frac{2n}{1-\theta}.$$

The solution is called maximum likelihood estimator (MLE) of θ and obtained as

$$\hat{\theta} = \frac{\sum_{i=1}^{n}(X_i - 1)}{2n + \sum_{i=1}^{n}(X_i - 1)}.$$

It can be seen that the likelihood equation can be solved and ML estimator is obtained in an open form which brings us some advantages for modelling and calculations.

3. Count Regression Based on Size Biased Geometric Distribution

In many economic, medicine or engineering context the dependent variable of interest is a non-negative integer. This dependent variable can be explained in terms of a set of covariates. The classical regression analysis models the continous response and covariates. Therefore, regression models for count data has been studied in statistical analysis. The Poisson regression model assumes that the observed outcome variable follows a Poisson distribution. It has been used for modelling count data and contingency table. On the other hand, logistic regression is very suitable for binary data. However, both model have some restriction in application. As can be noticed, logistic distribution can be only used for binary data. Poisson regression is also suitable for the count data whose mean and variance are approximately equal. In this paper, we introduced a new regression model in order to model the count data. It brings us flexibility for any count data. The Poisson regression model is

$$\log(E(Y = y|x)) = \alpha + \beta x = \beta^{'} x.$$

Taking consideration the Equation (3), we proposed the following linear model based on size biased geometric distribution.

$$\frac{\theta + 1}{1 - \theta} = e^{\beta^{'} x},$$

where $\frac{\theta+1}{1-\theta}$ is the expected value of size biased geometric distribution. Solving this equation we reach the following model

$$\theta = \frac{e^{\beta^{'} x} - 1}{1 + e^{\beta^{'} x}}. \tag{5}$$

Further, the likelihood function is

$$L = \prod y_i \theta^{y_i - 1}(1 - \theta)^2.$$

Taking the logarithm of the likelihood function and putting Equation (4) instead of θ, we obtain the log-likelihood function

$$\log L = \sum_{i=1}^{n} \ln y_i + \frac{e^{\beta^{'} x} - 1}{1 + e^{\beta^{'} x}} \sum_{i=1}^{n}(y_i - 1) + 2n \ln\left(1 - \frac{e^{\beta^{'} x} - 1}{1 + e^{\beta^{'} x}}\right).$$

Taking derivatives of the log-likelihood function with respect to the unknown parameters and setting the derivatives to zero, we obtain the maximum likelihood estimation of $\beta^{'}$.

4. Simulation Study

In this section, we perform a simulation study for the performance of the maximum-likelihood estimate given by Equation (18). First, we generate 10,000 samples of size n using $Q(p) = \inf\{x \in N, p <= F(x)\}$ and we compute the maximum-likelihood estimates for the 10,000 samples, say $\hat{\theta}$ for $i = 1, 2, ...$ Then, we compute the estimates, biases, standard errors (SE), mean-squared errors (MSE) and confidence intervals as

1. $E(\hat{\theta}) = \dfrac{1}{n}\sum_{i=1}^{n}\hat{\theta}_i = \bar{\theta}$

2. $Bias(\hat{\theta}) = \dfrac{1}{n}\sum_{i=1}^{n}(\hat{\theta}_i - \theta)$

3. $E(SE(\hat{\theta})) = \sqrt{\dfrac{1}{n}\sum_{i=1}^{n}\left(\dfrac{-\partial^2 \ln L}{\partial \hat{\theta_i}^2}\right)}$

4. $MSE(\hat{\theta}) = \dfrac{1}{n}\sum_{i=1}^{n}(\hat{\theta}_i - \theta)^2$

5. $CI_{95\%} = \hat{\theta} \pm 1.96\ SE(\hat{\theta})$

We repeat these steps for $n = 10, 20$ and 50 with $\theta = 0.1, 0.2, 0.5$ and 0.9. Table 2 shows the result of the simulation.

Table 2 Simulation results for MLE

True Parameter	Estimates	Sample Size		
		10	25	50
$\theta = 0.1$	$E(\hat{\theta})$	0.037	0.069	0.081
	$Bias(\hat{\theta})$	0.006	0.002	0.001
	$E(SE(\hat{\theta}))$	0.006	0.002	0.001
	$MSE(\hat{\theta})$	0.005	0.001	0.001
	$\%p$ in $CI_{95\%}$	%91.8	%92.4	%93.1
$\theta = 0.2$	$E(\hat{\theta})$	0.167	0.177	0.184
	$Bias(\hat{\theta})$	0.003	0.001	0.001
	$E(SE(\hat{\theta}))$	0.004	0.003	0.003
	$MSE(\hat{\theta})$	0.002	0.001	0.001
	$\%p$ in $CI_{95\%}$	%92.7	%93.4	%94.2
$\theta = 0.5$	$E(\hat{\theta})$	0.473	0.489	0.495
	$Bias(\hat{\theta})$	0.002	0.002	0.001
	$E(SE(\hat{\theta}))$	0.004	0.003	0.003
	$MSE(\hat{\theta})$	0.002	0.001	0.001
	$\%p$ in $CI_{95\%}$	%92.8	%93.5	%94.4
$\theta = 0.7$	$E(\hat{\theta})$	0.675	0.689	0.696
	$Bias(\hat{\theta})$	0.003	0.001	0.001
	$E(SE(\hat{\theta}))$	0.004	0.002	0.002
	$MSE(\hat{\theta})$	0.002	0.001	0.001
	$\%p$ in $CI_{95\%}$	%93.0	%93.8	%94.5

From Table 2, it can be seen that as the sample size increases, the bias, the mean square error, and the standard error of the estimates decreases. Further, the estimators satisfies the normality from the confidence intervals.

5. Numerical Example

In this section we study two data set to illustrate the performance of the size biased geometric distribution. The first data set is about 2003 final examination marks of 48 slow space students in mathematics in the Indian Institute of Technology at Kanpur and studied by Gupta and Kundu (2009). The data is given in Table 3.

Table 3 Data Set 1

29	25	50	15	13	27	15	18	7	7	8	19	12	18	5	21	15	86	21
15	14	39	15	14	70	44	6	23	58	19	50	23	11	6	34	18	28	34
12	37	4	60	20	23	40	65	19	31	12								

The second data set is about of remission times in weeks for 20 leukaemia patients randomly assigned to a certain treatment and studied by Lawless (1982), Damien and Walker (2002), Kottas (2006) and Bakouch et al. (2014). The data is given in Table 4.

Table 4 Data Set 2

1	3	3	6	7	7	10	12	14	15	18	19
22	26	28	29	34	40	48	49				

The six models (size biased geometric distribution, geometric, Poisson, discrete Weibull, discrete gamma and discrete Lindley) were fitted to each of the data sets The parameter estimates, p-values and the Kolmogorov-Smirnov statistics are shown in Tables 5 and 6.

Table 5 Fitted estimates for data set 1

Model	Estimates	$p-value$	$K-S$
Size-Biased Geometric	$\hat{\theta}=0.925$	0.7325	0.0833
Poisson	$\hat{\lambda}=25.89$	0.0000	0.3998
Geometric	$\hat{p}=0.037$	0.0150	0.2223
Discrete Weibull	$\hat{q}=0.648, \hat{\beta}=0.675$	0.0000	0.7419
Discrete Gamma	$\hat{\alpha}=0.809, \hat{\beta}=0.035$	0.0000	0.2993
Discrete Lindley	$\hat{\theta}=0.076$	0.4667	0.1191

Table 6 Fitted estimates for data set 2

Model	Estimates	$p-value$	$K-S$
Size-Biased Geometric	$\hat{\theta}=0.902$	0.8130	0.1190
Poisson	$\hat{\lambda}=19.55$	0.0100	0.3523
Geometric	$\hat{p}=0.0487$	0.7436	0.1447
Discrete Weibull	$\hat{q}=0.368, \hat{\beta}=0.0.360$	0.0000	0.7233
Discrete Gamma	$\hat{\alpha}=0.5623, \hat{\beta}=0.031$	0.0541	0.2909
Discrete Lindley	$\hat{\theta}=0.1$	0.0541	0.2909

We can say that the size biased geometric distribution provides better fits than other competing models for at least two data sets.

6. Conclusion

In this article, further statistical inferences about size biased geometric distribution is introduced. Its important probabilistic and mathematical properties and parameter estimation are studied. As can noticed from the previous section, the size biased geometric distribution is easy to interpret and have no mathematical difficulties. On the other hand, both moments and ML estimators of the unknown parameter can easily be obtained in a closed form. Additionally, the hazard rate function has an increasing form. Numerical examples show that the proposed distribution is suitable for the data comparing to the well-known distributions such as Poisson and Geometric and the other discrete distributions which have been recently introduced such as discrete Weibull, Gamma and Lindley distri-

butions. Therefore, we conclude that the size biased geometric distribution will be a good alternative to the other distributions with easy interpretation and mathematical advantages.

References

Baker RD. Application of a new discrete distribution. J Appl Stat. 2000; 27(1): 5-21.

Bakouch HS, Jazi MA, Nadarajah S. A new discrete distribution. Statistics. 2014; 48(1):200-240.

Chakraborty S, Chakravarty, D. Discrete gamma distribution: Properties and parameter estimation. Commun Stat Theory. 2012; 41(18): 3301-3324.

Damien P, S. Walker S. A Bayesian non-parametric comparison of two treatments. Scand J Stat. 2002; 29: 51-56.

Deniz EG, Sarabia JM, Ojeda EC. A new discrete distribution with actuarial applications. Insur Math Econ. 2006; B48(3): 406-412.

El-Desouky BS, Gomaa RS, Magar AM. New Discrete Lifetime Distribution with Applications to Count Data. J Stat Theory Appl. 2021; 20(2): 304-317.

Gupta RD, Kundu D. A new class of weighted exponential distributions. Statistics. 2009; 43: 621-634.

Jazi MA, Lai CD, Alamatsaz MH. A discrete inverse Weibull distribution and estimation of its parameters. Stat Method. 2009; 7(2): 121-132.

Khan MSA, Khalique A, Abouammoh AM. On estimating parameters in a discrete Weibull distribution. IEEE T Reliab. 1989; 38: 248-250.

Khorashadizadeh M, Rezaei Roknabadi AH, Mohtashami Borzadaran GR. Characterisation of life distributions using log-odds rate in discrete ageing. Commun Stat Theory. 2013; 42(1): 76-87.

Kottas A. Nonparametric Bayesian survival analysis using mixtures of Weibull distributions. J Stat Plan Infer. 2006; 136: 578-596.

Krishna H, Pundir PS. Discrete Burr and discrete Pareto distributions. Stat Method. 2009;6: 177-188.

Kulasekera KB. Approximate mles of the parameters of a discrete Weibull distribution with type-I censored-data. Microelectron Reliab. 1994; 34: 1185-1188.

Lawless JF. Statistical Models and Methods for Lifetime Data, 1st edition: John Wiley and Sons, New York; 1982.

Nakagawa T, Osaki S. Discrete Weibull distribution. IEEE T Reliab. 1975; 24: 300-301.

Roy D. Discrete Rayleigh distribution. IEEE T Reliab. 2004; 53: 255-260.

Roy D. The discrete normal distribution. Commun Stat Theory. 2003; 32(10): 1871-1883.

Stein WE, Dattero R. A new discrete Weibull distribution. IEEE T Reliab. 1984; 33: 196-197.

Zheng Q. A new discrete distribution induced by the LuriaDelbrck mutation model. Statistics. 2010; 44: 529-540.