A new statistical approach to model the counts of novel coronavirus cases

M. El-Morshedy

Department of Mathematics, College of Sciences and Humanities Studies in Al-Kharj, Prince Sattam Bin Abdulaziz University, Al-Kharj 11942, Saudi Arabia.

Emrah Altun (✉ emrahaltun@bartin.edu.tr)

Department of Mathematics, Bartin University, Bartin 74100, Turkey

M. S. Eliwa

Department of Mathematics, Faculty of Science, Mansoura University, Mansoura 35516, Egypt.

Research Article

Keywords: COVID-19, Discrete distribution, Gamma Lindley distribution, Maximum likelihood estimation, Regression, Simulation.

Posted Date: May 23rd, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-31163/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Version of Record: A version of this preprint was published at Mathematical Sciences on March 16th, 2021. See the published version at <https://doi.org/10.1007/s40096-021-00390-9>.

A new statistical approach to model the counts of novel coronavirus cases

M. El-Morshedy^{1,2}, Emrah Altun^{3*} and M. S. Eliwa^{2,4}

¹Department of Mathematics, College of Sciences and Humanities Studies in Al-Kharj, Prince Sattam Bin Abdulaziz University, Al-Kharj 11942, Saudi Arabia.

²Department of Mathematics, Faculty of Science, Mansoura University, Mansoura 35516, Egypt.

³Department of Mathematics, Bartın University, Bartın 74100, Turkey.

⁴Department of Mathematics, College of Science in Al-Zulfi, Majmaah University, Al-Majmaah 11952, Saudi Arabia

Abstract

This study proposes new statistical tools to analyze the counts of the daily coronavirus cases and deaths. Since the daily new cases and deaths exhibit highly overdispersion, we introduce a new two-parameter discrete distribution, called *discrete generalized Lindley*, which enables us to model all kinds of dispersion such as under, equi and overdispersion. Additionally, we introduce a new count regression model based on the proposed distribution to investigate the effects of the important risk factors on the counts of deaths for OECD countries. Three data sets are analyzed with proposed models and competitive models. Empirical findings show that air pollution, the proportion of obesity, and smokers in a population do not affect the counts of deaths for OECD countries. The interesting empirical result is that the countries with having larger alcohol consumption have lower counts of deaths.

Key words: COVID-19; Discrete distribution; Gamma Lindley distribution; Maximum likelihood estimation; Regression; Simulation.

MSC: 60E05; 62E10; 62F10; 62N05.

1 Introduction

The first case of the COVID-19 (coronavirus disease, 2019) was reported in Wuhan, China, in December 2019. The World Health Organization (WHO) has declared that the COVID-19 is a pandemic on the date of 11 March 2020. After this date, all countries have increased their measures to decrease the spread rate of the COVID-19 by closing schools, shopping centers, airlines, and also their borders. As of date 17 May 2020, the counts of COVID-19 cases are over 4.72 million and the counts of deaths are 313,221. This number may be of little importance to anyone, but almost half the population of Luxembourg.

The researchers and academicians have spent their time finding medical solutions such as drugs and vaccines to return our normal life. Besides these medical researches, the researchers have also focused on the mathematical and statistical modeling of the COVID-19 outbreak.

*Corresponding author: emrahaltun@bartin.edu.tr

For instance, Remuzzi and Remuzzi (2020) predicted the needed hospital beds and personnel for Italy under the exponential trend. Maleki et al. (2020) used autoregressive time series models based on the two-piece scale mixture normal distributions to forecast the recovered and confirmed COVID-19 cases. Nesteruk (2020) predicted the daily new COVID-19 cases in China by using the mathematical model, called SIR. As in Nesteruk (2020), Batista (2020) also used the SIR model to predict COVID-19 cases. Caccavo (2020) introduced the SIRD compartmental model to predict COVID-19 cases in China and Italy. Ayyoubzadeh et al. (2020) predicted COVID-19 cases in Iran by using long short-term memory (LSTM) which is a deep-learning method.

This study aims to model the daily new cases and deaths of the COVID-19 employing a new statistical tool. To achieve this aim, we introduce a new flexible two-parameter discrete model, called as *discrete analogous of the generalized Lindley*, shortly DsGLi, distribution. The generalized Lindley distribution was introduced by Nedjar and Zeghdoudi (2016) and modified by Messaadia and Zeghdoudi (2017). We introduce the DsGLi distribution by applying the survival discretization method of Roy (2003). The question may come to mind of anyone: why do we need this distribution? In primary data analysis, we have realized that most of the existing probability distributions do not provide acceptable results for modeling the COVID-19 cases. The reason for that the counts of deaths or daily new cases exhibit excessive overdispersion (see, Section 2 for the definition of the overdispersion). At this point, we see the advantage of DsGLi distribution over other distributions. Because, the proposed distribution, DsGLi, provides a new opportunity to model all kinds of dispersed count data sets. Some important properties of the DsGLi distribution can be summarized as follows: (i) it has closed forms expressions for its statistical properties, (ii) it has increasing hazard rate function (hrf), (iii) it can be used to model both skewed and leptokurtic count data sets. Additionally, a new count regression model is defined based on the DsGLi distribution to analyze the effects of explanatory variables such as smoking, obesity, air pollution on the daily cases, and deaths of the COVID-19 outbreak.

The remaining parts of the presented study are organized as follows: Section 2 deals with the statistical properties of the DsGLi distribution. In Section 3, we discuss the parameter estimation process of the DsGLi distributions with maximum likelihood estimation method, and the performance of the estimation method is investigated by a simulation study for its finite sample size behavior. In Section 4, we introduce the DsGLi regression model and clarify its parameter estimation process and residual analysis. Section 5 contains three applications to COVID-19 data sets. The empirical results obtained in Section 5 are discussed in Section 6 in detail. Section 7 contains some important remarks about the presented study.

2 Discrete analogue of GLi distribution

Roy (2003) introduced a new method to generate a new discrete distribution based on the survival function of any continuous probability distribution. This method is called a survival discretization method. Let the continuous random variable X has the survival function (sf) $S(x; \xi) = \Pr(X > x)$, then the probability mass function (pmf), corresponding to $S(x; \xi) = \Pr(X > x)$, is

$$\Pr(X = x) = S(x; \xi) - S(x + 1; \xi); \quad x = 0, 1, 2, 3, \dots \quad (1)$$

This approach has been received considerable attention over the recent years. For instance, Gómez-Déniz and Calderín-Ojeda (2011), Bebbington et al. (2012), Nekoukhou et al. (2013), Bakouch et al. (2014), Alamatsaz et al. (2016), El-Morshedy et al. (2020a, 2020b), Eliwa et al. (2020a, 2020b), and references cited therein. Nedjar and Zeghdoudi (2016) proposed a new probability distribution for modeling data, in the so-called gamma Lindley (GLi) distribution.

It is a mixture of a $\text{gamma}(2, \theta)$ and $\text{Lindley}(\alpha)$ distribution. The pdf of GLi distribution can be expressed as

$$f(x; \alpha, \theta) = \frac{\theta^2}{\alpha(1 + \theta)} ([\theta\alpha + \alpha - \theta]x + 1) e^{-\theta x}; \quad x > 0, \alpha > 0, \theta > 0. \quad (2)$$

Unfortunately, Equation (2) is not a proper PDF, because it can be negative for some values of the parameters $\alpha > 0$ and $\theta > 0$. Messaadia and Zeghdoudi (2017) modified the parameter space to be $\alpha \geq \frac{\theta}{1+\theta}$ and $\theta > 0$, and consequently the proper pdf of GLi model can be written as

$$f(x; \alpha, \theta) = \frac{\theta^2}{\alpha(1 + \theta)} ([\theta\alpha + \alpha - \theta]x + 1) e^{-\theta x}; \quad x > 0, \alpha \geq \frac{\theta}{1 + \theta}, \theta > 0. \quad (3)$$

The sf corresponding to Equation (3) is

$$R(x; \alpha, \theta) = \frac{(\theta x + 1)(\theta\alpha + \alpha - \theta) + \theta}{\alpha(1 + \theta)} e^{-\theta x}; \quad x > 0, \quad (4)$$

where $\alpha \geq \frac{\theta}{1+\theta}$ and $\theta > 0$. Using the survival discretization method and sf of GLi distribution, we define the rf of the DsGLi as given below

$$S(x; \alpha, \eta) = \frac{(1 - \ln \eta^{x+1})(\alpha - \alpha \ln \eta + \ln \eta) - \ln \eta}{\alpha(1 - \ln \eta)} \eta^{x+1}; \quad x \in \mathbb{N}_0, \quad (5)$$

where $\alpha \geq \frac{-\ln \eta}{1 - \ln \eta}$, $0 < \eta < 1$ and $\mathbb{N}_0 = \{0, 1, 2, 3, \dots, k\}$ for $0 < k < \infty$. The pmf and cumulative distribution function (cdf) of the DsGLi distribution are given, respectively, by

$$P_x(x; \alpha, \eta) = \frac{\eta^x}{1 - \ln \eta} \left\{ 1 - \eta - \ln \eta [1 + x - \eta(x + 2)] + (1 - \frac{1}{\alpha})(\ln \eta)^2 [x - \eta(x + 1)] \right\}, \quad (6)$$

$$F(x; \alpha, \eta) = 1 - \frac{(1 - \ln \eta^{x+1})(\alpha - \alpha \ln \eta + \ln \eta) - \ln \eta}{\alpha(1 - \ln \eta)} \eta^{x+1}, \quad (7)$$

where $x \in \mathbb{N}_0$. The pmf in (6) is log-concave, where $\frac{P_x(x+1; \alpha, \eta)}{P_x(x; \alpha, \eta)}$ is a decreasing function in x for all values of the model parameters. Figure 1 shows the pmf plots for different values of the model parameters. From Figure 1, the pmf of the DsGLi distribution is unimodal and right-skewed.

Figure 1: The pmf plots of the DsGLi distribution.

The hrf of the DsGLi is be expressed as

$$h(x; \alpha, \eta) = 1 - \frac{[(1 - \ln \eta^{x+1})(\alpha - \alpha \ln \eta + \ln \eta) - \ln \eta] \eta}{(1 - \ln \eta^x)(\alpha - \alpha \ln \eta + \ln \eta) - \ln \eta}; \quad x \in \mathbb{N}_0, \quad (8)$$

where $h(x; \alpha, \eta) = \frac{P_x(x; \alpha, \eta)}{S(x-1; \alpha, \eta)}$. Figure 2 shows the hrf plots of the DsGLi distribution. It is noted that the shape of the hrf is increasing. Further, the value of failure rate decreases with $\eta \rightarrow 1$ for fixed value of α .

Figure 2: The hrf plots of the DsGLi distribution.

2.1 Properties

The statistical properties of the DsGLi distribution are obtained and reported in this section. Let X be random variable having a pmf in (6). The probability generating function (pgf) of X is

$$\begin{aligned} G_x(z) &= \sum_{x=0}^{\infty} z^x G_x(x; \alpha, \eta) \\ &= \frac{-2\eta(\alpha - 1)(z - 1) \ln \eta + \alpha(\eta^2 z - 2\eta + 1) \ln \eta - \alpha(\eta - 1)(\eta z - 1)}{\alpha(\ln \eta - 1)(\eta z - 1)^2}. \end{aligned} \quad (9)$$

By replacing z by e^z in (9), one can obtain the moment generating function (mgf) of X which is given by

$$M_X(z) = \frac{-2\eta(\alpha - 1)(e^z - 1) \ln(\eta) + \alpha(\eta^2 e^z - 2\eta + 1) \ln \eta - \alpha(\eta - 1)(\eta e^z - 1)}{\alpha(\ln \eta - 1)(\eta e^z - 1)^2}. \quad (10)$$

The similar relation is also valid between the mgf and characteristic function (cf). The cf function of X is obtained by replacing e^z by e^{iz} . Then, we have

$$\varphi_X(z) = \frac{-2\eta(\alpha - 1)(e^{iz} - 1) \ln(\eta) + \alpha(\eta^2 e^{iz} - 2\eta + 1) \ln \eta - \alpha(\eta - 1)(\eta e^{iz} - 1)}{\alpha(\ln \eta - 1)(\eta e^{iz} - 1)^2}. \quad (11)$$

The partial derivatives of (10) according to z at $z = 0$ gives raw moments of X . Using this property, the first two moments of the DsGLi model are given, respectively, by

$$E(X) = -\eta \frac{(\alpha - 1) \ln \eta^2 + \alpha(\eta - 2) \ln \eta - \alpha(\eta - 1)}{\alpha(\ln \eta - 1)(\eta - 1)^2}, \quad (12)$$

$$E(X^2) = \eta \frac{(3\alpha\eta + \alpha - 3\eta - 1) \ln \eta^2 + \alpha(\eta^2 - 3\eta - 2) \ln \eta - \alpha\eta^2 + \alpha}{\alpha(\ln \eta - 1)(\eta - 1)^3}. \quad (13)$$

The variance of the DsGLi distribution can be calculated by using $\text{Var}(X) = E(X^2) - E(X)^2$. The skewness and kurtosis measures of the DsGLi can be also easily calculated by using well-known relations. The other important measure of any discrete distribution is dispersion index (DI) which is defined as $\text{DI} = \text{Var}(X) / E(X)$. The flexibility of the DI measure is very important to model different types of data sets such as over-dispersed ($\text{DI} > 1$), equi-dispersed ($\text{DI} = 1$) and under-dispersed ($\text{DI} < 1$). The statistical measures of the DsGLi distribution are computed and reported in Tables 1 and 2. To interpret the individual effects of the parameters α and η , the results are calculated for fixed $\alpha = 0.9$ and $\eta = 0.01$. As seen from Table 1, the mean, variance, DI are the increasing function of the parameter η for fixed $\alpha = 0.90$ whereas the skewness and kurtosis decrease when the parameter η increases. As seen in Table 2, the mean and variance is the increasing function of α . The DI, skewness, and kurtosis decrease when the parameter α increases. As seen from these results, the DsGLi distribution has flexible DI which can be over or under one. So, the DsGLi distribution can be an appropriate choice in modeling all types of count data.

Table 1: The numeric values of the statistical measures of the DsGLi distribution for $\alpha = 0.9$.

	η										
Measure	0.01	0.05	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Mean	0.0141	0.0803	0.1751	0.4082	0.7182	1.1438	1.7557	2.6959	4.2971	7.5582	17.478
Variance	0.0143	0.0852	0.1976	0.5248	1.0663	1.9968	3.6960	7.1035	15.049	39.363	179.20
DI	1.0118	1.0612	1.1283	1.2855	1.4845	1.7457	2.1051	2.6348	3.5022	5.2079	10.253
Skewness	8.5415	3.8420	2.8210	2.1538	1.8816	1.7274	1.6251	1.5508	1.4947	1.4528	1.4249
Kurtosis	77.619	19.256	12.369	8.9408	7.7616	7.1397	6.7437	6.4663	6.2648	6.1217	6.0321

Table 2: The numeric values of the statistical measures of the DsGLi distribution for $\eta = 0.01$.

	α										
Measure	1.5	3.0	4.5	6.0	7.5	9.0	10.5	12.5	15.0	17.5	20.0
Mean	0.0313	0.0442	0.0485	0.0506	0.0519	0.0527	0.0534	0.0539	0.0545	0.0548	0.0551
Variance	0.0314	0.0438	0.0479	0.0499	0.0511	0.0519	0.0525	0.5301	0.5355	0.0538	0.0541
DI	1.0025	0.9915	0.9876	0.9857	0.9845	0.9837	0.9831	0.9826	0.9821	0.9818	0.9815
Skewness	5.6703	4.6978	4.4593	4.351	4.289	4.2489	4.2208	4.1943	4.1714	4.1552	4.1432
Kurtosis	35.341	24.741	22.438	21.431	20.866	20.504	20.253	20.017	19.815	19.6729	19.567

3 Estimation

The unknown parameters of the DsGLi distribution are obtained by the maximum likelihood estimation (MLE) method. This method is based on the maximization of the log-likelihood function for a given data set. Let assume that we have a sample comes from the DsGLi distribution, denoted as $X_1, X_2, \dots, X_n$. Then, we have the following log-likelihood function for the DsGLi distribution

$$\begin{aligned} \ell(x; \alpha, \eta) &= \ln \eta \sum_{i=1}^n x_i - n \ln(1 - \ln \eta) \\ &+ \sum_{i=1}^n \ln \left(1 - \eta - \ln \eta [1 + x_i - \eta(x_i + 2)] + \left(1 - \frac{1}{\alpha}\right) (\ln \eta)^2 [x_i - \eta(x_i + 1)] \right) \end{aligned} \quad (14)$$

We have two choices to obtain the MLEs of the parameters α and η . The first way, we can use (14) to direct maximization of the log-likelihood to get the MLEs of the parameters, say $\hat{\alpha}$, and $\hat{\eta}$. The second way, the score vectors, given below, can be simultaneously solved for zero.

$$\frac{\partial \ell}{\partial \alpha} = \sum_{i=1}^n \frac{\left(\frac{\ln \eta}{\alpha}\right)^2 [x_i - \eta(x_i + 1)]}{1 - \eta - \ln \eta [1 + x_i - \eta(x_i + 2)] + \left(1 - \frac{1}{\alpha}\right) (\ln \eta)^2 [x_i - \eta(x_i + 1)]}, \quad (15)$$

$$\begin{aligned} \frac{\partial \ell}{\partial \eta} &= \frac{1}{\eta} \sum_{i=1}^n x_i + \frac{n}{\eta(1 - \ln \eta)} \\ &+ \sum_{i=1}^n \frac{(x_i + 2)(\ln \eta + 1) - \frac{1+x_i}{\eta} - 1 + \ln \eta \left(1 - \frac{1}{\alpha}\right) \left(\frac{2x_i - 2\eta(x_i + 1)}{\eta} - (x_i + 1) \ln \eta\right)}{1 - \eta - \ln \eta [1 + x_i - \eta(x_i + 2)] + \left(1 - \frac{1}{\alpha}\right) (\ln \eta)^2 [x_i - \eta(x_i + 1)]}. \end{aligned} \quad (16)$$

In this study, we prefer the first choice, direct maximization of (14) by means of **constrOptim** function of **R**. To obtain the asymptotic standard errors and confidence intervals, the observed information matrix is used evaluated at $\hat{\alpha}$ and $\hat{\eta}$. The observed information matrix can be numerically calculated by **hessian** function of **R** software.

3.1 Simulation

We assess the finite sample performance of the MLE method in estimating the unknown parameters of the DsGLi distribution. Therefore, we conduct a simulation study. The simulation replication number, N , is taken as 1,000. The true parameter values are used as $\alpha = 0.5$ and $\eta = 0.5$. We generate random samples from the DsGLi distribution with $n = 50, 55, 60, \dots, 300$ sample sizes. The simulation results are interpreted based on the estimated biases, mean square

errors (MSEs), and mean relative errors (MREs). The required mathematical formulations of these metrics can be found in the works of Altun (2019, 2020). We expect to see that biases and MSEs are near the zero and MREs are near the one for sufficiently large sample sizes. The simulation results are graphically summarized and displayed in Figure 3. These results confirm our expectation that the estimated biases and MSEs are very near the zero for nearly all samples of sizes. Also, the estimated MREs are near the one, as expected. These results also show that the MLEs of the parameters of the DsGLi distribution are asymptotically unbiased and consistent. The similar results are obtained for different parameter settings, but not reported here for the sake of simplicity.

Figure 3: The simulation results of the DsGLi distribution.

4 DsGLi regression model

When modeling the discrete response variable with associated covariates, the Poisson regression model is the first thing come to mind. However, it is well known that the Poisson regression produces inaccurate results when the response variable is over-dispersed or under-dispersed, except equi-dispersion. In this study, we propose an alternative model to provide new opportunities in predicting the over-dispersed counts. Let Y be a random variable following a DsGLi density. Using the below re-parametrization

$$\alpha = \frac{\eta \ln(\eta)^2}{(\ln(\eta) - 1)((\eta - 1)(\eta + \mu(\eta - 1)) + \eta \ln(\eta))}, \quad (17)$$

we have

$$P(y_i; \eta, \mu) = \frac{\eta^{y_i}}{1 - \ln(\eta)} \left\{ \begin{aligned} &1 - \eta - \ln(\eta) [1 + y - \eta(y + 2)] \\ &+ \left(\frac{\gamma(\eta, \mu) - 1}{\gamma(\eta, \mu)} \right) \ln(\eta)^2 [y_i - \eta(y_i + 1)] \end{aligned} \right\}, \quad (18)$$

where

$$\gamma(\eta, \mu_i) = \frac{\eta \ln(\eta)^2}{(\ln(\eta) - 1)((\eta - 1)(\eta + \mu_i(\eta - 1)) + \eta \ln(\eta))}. \quad (19)$$

The density in (18) is denoted as $Y \sim \text{DsGLi}(\alpha, \mu)$. After this re-parametrization, the mean of the random variable is $E(Y) = \mu$ and its variance is

$$\text{Var}(Y) = \frac{\mu(3\eta + 1)}{1 - \eta} - \frac{2\eta^2}{(1 - \eta)^2} - \mu^2 \quad (20)$$

The parameter η is a dispersion parameter of the re-parametrized DsGLi distribution. Now, using the density in (18), we propose a new count regression model. Let the response variable Y follows the density in (18) and consider the regression structure given below

$$g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}, \quad (21)$$

where $\mathbf{x}_i^T = (x_{i1}, x_{i2}, \dots, x_{ik})$ is the explanatory variable vector and $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_k)$ is the vector of regression parameters. The function in (21) is known as link function. The link functions play an important role in generalized linear models to construct a bridge between predictors and the mean of the response variable. The suitable choice of the link function depends on the domain the response variable. Since the response variable is defined on $\mathbb{Z}_+$, the log-link function is used.

4.1 Estimation process

The log-link is defined as $\mu_i = \exp(\mathbf{x}_i^T \boldsymbol{\beta})$. Using the log-link function, the log-likelihood function of the DsGLi regression model is given by

$$\begin{aligned} \ell(\eta, \boldsymbol{\beta}) = & \ln(\eta) \bar{y} - n \ln(1 - \ln(\eta)) \\ & + \sum_{i=1}^n \ln \left\{ \begin{aligned} & 1 - \eta - \ln(\eta) [1 + y - \eta(y + 2)] \\ & + \left(1 - \frac{(\ln(\eta) - 1)((\eta - 1)(\eta + \exp(\mathbf{x}_i^T \boldsymbol{\beta})(\eta - 1)) + \eta \ln(\eta))}{\eta \ln(\eta)^2} \right) \\ & \times \ln(\eta)^2 [y_i - \eta(y_i + 1)] \end{aligned} \right\} \end{aligned} \quad (22)$$

where n is the sample size, η is the unknown dispersion parameter and $\boldsymbol{\beta}$ is the unknown regression parameters. Let $\boldsymbol{\gamma} = (\eta, \boldsymbol{\beta})$ be a unknown parameter vector. Under the regularity conditions of the MLE, the asymptotic distribution of the $\boldsymbol{\gamma} - \hat{\boldsymbol{\gamma}}$ is the $(k + 2)$ -variate normal distribution with zero mean and variance-covariance matrix, $\Sigma_{(k+2) \times (k+2)}$ which is obtained by the inverse of observed information matrix, $I_{(k+2) \times (k+2)}$ calculated at the MLE of the $\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}$. To estimate the unknown parameter vector, $\boldsymbol{\gamma}$, the below estimation process is implemented.

1. First, the underlying data set is fitted by Poisson regression model to get the initial parameter vector of the $\boldsymbol{\beta}$ for DsGLi regression model.
2. Next, using the estimated regression parameters of the Poisson regression model as an initial parameter vector for the $\boldsymbol{\beta}$ and setting the initial value of $\eta = 0.5$, the minus log-likelihood function, $-\ell(\eta, \boldsymbol{\beta})$ in (22), is minimized by **constrOptim** function in **R** since the parameter $\eta \in (0, 1)$.
3. Then, using the **hessian** function in **R**, we obtain the observed information matrix, $I_{(k+2) \times (k+2)}$, calculated at $\hat{\boldsymbol{\gamma}}$.

4.2 Residual analysis

Residual analysis is carried out to be sure about the accuracy of the DsGLi regression model for the used data set. The randomized quantile residual (rqr) is used for this purpose.

Let the random variable Y having a cdf $F(y; \eta, \mu)$ which is the cdf of the re-parametrized DsGLi distribution. Then, the rqr is defined as

$$r_{q,i} = \Phi^{-1}(u_i), \quad (23)$$

where $u_i = F(y_i; \hat{\alpha}, \mu_i)$. Note that the rqr is distributed as $N(0, 1)$ when the model is acceptable for the used data.

5 Data analysis

The empirical importance of the proposed models are proved by three applications to COVID-19 data sets. The proposed models are compared with some competitive models to see its competitive power. The competitive models are listed below.

Table 3: The competitive models.

Distribution	Abbreviation	Reference
Poisson	Poi	-
Discrete Lindley	DsLi	Gómez-Déniz and Calderín-Ojeda (2011)
A new one-parameter discrete model	NDsIP	Eliwa and El-Morshedy (2020)
Discrete Rayleigh	DsR	Roy (2004)
Discrete Burr-XII	DsB-XII	Krishna and Pundir (2009)
Discrete Pareto	DsPa	Krishna and Pundir (2009)
Discrete Burr-Hatke	DsBH	El-Morshedy et al. (2020b)

In the first two empirical studies, we compare the fit of the DsGLi distribution with competitive models listed in Table (3). The goodness-of-fit test, Kolmogorov-Smirnov ($K - S$) are implemented to select a best-fitted model for COVID-19 data sets. The models having p-value larger than 0.05 are evaluated as possible accurate models, and the information criteria, listed below, are used to decide best fitted model in final stage. The data source is <https://www.worldometers.info/coronavirus/>. In the third application, we asses the performance of the DsGLi regression model by applying the model to the COVID-19 data set of the OECD countries.

- ✓ Akaike information criterion (AIC)
- ✓ Hannan-Quinn information criterion (HQIC)
- ✓ Bayesian information criterion (BIC)
- ✓ Corrected Akaike information criterion (CAIC)

5.1 Hong Kong

In the first application, we consider the daily new cases in Hong Kong, China. The data is available at <https://www.worldometers.info/coronavirus/country/china-hong-kong-sar/> and contains the daily new cases between 15 February to 23 March 2020. The data are: 0, 1, 3, 2, 3, 4, 0, 1, 4, 7, 4, 4, 3, 1, 2, 5, 0, 1, 2, 2, 3, 0, 7, 1, 5, 9, 1, 1, 10, 7, 6, 13, 25, 15, 48, 18, 44, 39.

This data set is modeled with DsGLi and other competitive models. Table 4 contains the estimated parameters and their corresponding standard errors (SEs) as well as confidence

intervals (CIs) for all fitted models. The results of the information criteria and goodness-of-fit test are given in Table 5. The best-fitted model should have the lowest values of these statistics. From Table 5, we conclude that the DsGLi model is the best-performed model among others since it has the lowest values of AIC, BIC, CAIC, HQIC, and K-S test statistic. The higher value of the p-value of the KS test shows the better-fitted model. If the p-value is less than 0.05, it means that the model cannot be used to predict the counts of COVID-19 cases. According to the modeled data set, DsGLi and NDsIP distributions have p-values larger than 0.05. However, the p-value of DsGLi distribution is higher than those of NDsIP distribution. Therefore, the proposed model is the best choice for modeling the data used.

Table 4: The estimated parameters of the fitted models for Hong Kong data set.

Model	α			η		
	MLE	SEs	CIs	MLE	SEs	CIs
DsGLi	0.888	0.024	[0.839, 0.936]	0.106	0.037	[0.035, 0.178]
DsLi	0.805	0.020	[0.766, 0.845]	—	—	—
NDsIP	0.911	0.015	[0.882, 0.939]	—	—	—
DsR	0.995	0.136	[0.728, 1]	—	—	—
DsB-XII	0.817	0.074	[0.672, 0.962]	3.298	1.427	[0.501, 6.096]
DsPa	0.559	0.053	[0.455, 0.663]	—	—	—
DsBH	0.995	0.013	[0.969, 1]	—	—	—
Poi	7.921	0.457	[7.026, 8.816]	—	—	—

Table 5: The goodness of fit test for Hong Kong data set.

Statistic	Model							
	DsGLi	DsLi	NDsIP	DsR	DsB-XII	DsPa	DsBH	Poi
$-\ell$	118.95	127.12	121.35	161.19	118.99	124.33	130.95	287.83
AIC	241.89	256.23	244.71	324.38	241.98	250.66	263.90	577.67
CAIC	242.23	256.34	244.82	324.49	242.32	250.77	264.01	577.78
BIC	245.17	257.87	246.35	326.02	245.26	252.30	265.54	579.31
HQIC	243.06	256.82	245.29	324.96	243.15	251.25	264.48	578.25
K-S	0.157	0.246	0.186	0.761	0.277	0.367	0.565	0.913
p-value	0.305	0.020	0.146	≤ 0.001	0.006	≤ 0.001	≤ 0.001	≤ 0.001

Figure 4 displays the estimated cdfs and probability-probability (PP) plots of the fitted models. These figures prove the suitability of the DsGLi distribution for modeling the counts of COVID-19 cases in Hong Kong.

Figure 4: The estimated cdfs (left) and PP plots (right) for Hong Kong data set.

Table 6 lists the theoretical values of the mean, variance, DI, skewness, and kurtosis measures of the DsGLi distribution obtained under the estimated parameters. The empirical mean and skewness of the used data set are 7.9210 and 2.2958, respectively. The theoretical mean and skewness are very closed to empirical ones.

Table 6: The numerical values for the statistical properties of the DsGLi distribution for Hong Kong data set.

Mean	Variance	DI	Skewness	Kurtosis
7.9203	70.6510	8.9203	2.0035	29.5274

5.2 Iran

In the second application, we consider the daily new deaths in Iran. Because, Iran is one of the countries affected very badly because of COVID-19 outbreak. The data is available at <https://www.worldometers.info/coronavirus/country/iran/> and contains the daily new cases between 15 February to 10 March 2020. The data are: 0, 0, 0, 0, 0, 0, 2, 2, 2, 4, 4, 3, 7, 8, 9, 11, 12, 11, 15, 16, 16, 21, 49, 43, 54.

The above data set is modeled with DsGLi and other competitive models and the estimated parameters and goodness-of-fit results are reported in Tables 7 and 8, respectively. According to the results in Table 8, we conclude that the DsGLi distribution is the best choice among other competitive models since it has the lowest values of the goodness-of-fit statistics. Additionally, three distributions, DsGLi, DsLi and NDsIP, have p-values larger than 0.05 for KS test results. However, as in previous application, the p-value of DsGLi distribution is higher than those of DsLi and NDsIP distributions. Therefore, the proposed model is the best choice for modeling the data used.

Table 7: The estimated parameters of the fitted models for Hong Kong data set.

Model	α			η		
	MLE	SEs	CI	MLE	SEs	CI
DsGLi	0.920	0.025	[0.871, 0.970]	0.077	0.042	[0, 0.158]
DsLi	0.857	0.019	[0.819, 0.894]	—	—	—
NDsIP	0.941	0.011	[0.919, 0.962]	—	—	—
DsR	0.997	0.048	[0.903, 1]	—	—	—
DsB-XII	0.650	0.102	[0.451, 0.849]	1.239	0.426	[0.405, 2.073]
DsPa	0.596	0.062	[0.475, 0.717]	—	—	—
DsBH	0.997	0.011	[0.975, 1]	—	—	—
Poi	11.560	0.680	[10.227, 12.893]	—	—	—

Table 8: The goodness of fit test for Iran data set.

Statistic	Model							
	DsGLi	DsLi	NDsIP	DsR	DsB-XII	DsPa	DsBH	Poi
$-\ell$	87.24	94.73	88.94	115.82	89.92	90.12	96.91	243.85
AIC	178.48	191.46	179.88	233.64	183.83	182.24	195.81	489.70
CAIC	179.03	191.64	180.06	233.81	184.38	182.42	195.98	489.88
BIC	180.92	192.68	181.10	234.86	186.27	183.46	197.03	490.92
HQIC	179.16	191.80	180.22	233.97	184.51	182.58	196.15	490.04
K-S	0.160	0.251	0.189	0.742	0.258	0.301	0.512	0.942
P-value	0.541	0.087	0.328	≤ 0.001	0.072	0.021	≤ 0.001	≤ 0.001

Figure 5 displays the estimated cdfs and PP plots of the fitted models. From these figures, it is concluded that the DsGLi model provides acceptable modeling performance for the used data.

Figure 5: The estimated cdfs (left) and PP plots (right) for Iran data set.

Additionally, the theoretical values of the mean, variance, DI, skewness, and kurtosis mea-

sures of the DsGLi distribution are reported in Table 9 for Iran data set. The empirical mean and skewness of the used data set are 11.5600 and 1.7432, respectively. The theoretical mean and skewness are closed to empirical ones.

Table 9: The numerical values for the statistical properties of the DsGLi distribution for Iran data set.

Mean	Variance	IxD	Skewness	Kurtosis
11.5055	143.8836	12.5055	2.0017	9.0069

5.3 OECD

In the third application, the counts of deaths due to the COVID-19 are modeled for the OECD countries by the DsGLi regression model. The predictive performance of the DsGLi regression model is compared with Poisson regression. The response variable is the counts of deaths up to the date 16 May 2020. According to the *Health at a Glance* report (see OECD, 2019), the important risk factors on health are the use of smoking and alcohol, overweight (obese), and air pollution. These variables are available in OECD (2019) and measured in the year 2019. We try to explain the variability in the counts of deaths (y_i) due to the coronavirus with covariates, smoking (x_{i1} , % population aged 15+), alcohol (x_{i2} , % population aged 15+) and overweight (x_{i3} , % population with BMI ≥ 25) for population aged 15+). We also consider the population size for each country as an explanatory variable. The population size is transformed in three-level categoric variable and two dummy variables are created: the population size between 7-35 million (x_{i5} , 1=yes, 0=no) and the population size over 35 million x_{i6} , 1=yes, 0=no). The population size lower than 7 million is considered as a baseline category. To avoid the extreme outlier observations, we exclude the countries having less than 1 million population and over 100 million population sizes. The regression model in (24) is fitted by DsGLi and Poisson regression models.

$$\mu_i = \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i6} + \beta_5 x_{i6}). \quad (24)$$

The estimated regression parameters and their corresponding standard errors (SEs), as well as p-values, are listed in Table (10). The mean and variance of the response variable are 5391.4 and 104275974 which shows that it is highly over-dispersed. Because of the over-dispersed response variable, the Poisson regression model produces inaccurate results with higher AIC and BIC values than the DsGLi regression model. There are dramatic differences between the AIC and BIC values of two regression models. The DsGLi regression model enables to model over-dispersed response variable on the contrary to the Poisson regression model.

Table 10: The results of Poisson and DsGLi regression models

Parameters	Poisson			DsGLi		
	Estimates	SEs	p-values	Estimates	SEs	p-values
β_0	5.0493	0.0295	< 0.0001	7.0174	0.0033	< 0.0001
β_1	0.1108	0.0007	< 0.0001	0.1075	0.0734	0.14303
β_2	-0.0899	0.0017	< 0.0001	-0.2265	0.0763	0.0029
β_3	0.0070	0.0003	< 0.0001	-0.0097	0.0099	0.3265
β_4	-0.0206	0.0002	< 0.0001	-0.0472	0.0296	0.1108
β_5	3.6286	0.0190	< 0.0001	4.2760	0.0119	< 0.0001
β_6	1.5763	0.0202	< 0.0001	2.0978	0.0114	< 0.0001
η	-	-	-	0.9998	< 0.0001	-
$-\ell$	72891.2000			273.6595		
AIC	145796.4000			563.3190		
BIC	145806.2000			574.5286		

Figure 6 displays the results of the residual analysis of the DsGLi regression model. As seen from these figures, there is no observation to be evaluated as an outlier observation since all plotted points are in the envelopes.

Figure 6: The residual results of DsGLi regression model.

6 Discussion of empirical results

In this section, the empirical results are interpreted in detail. The first two applications are based on the modeling of the counts of daily new cases and deaths of Hong Kong and Iran, respectively. Using the estimated model parameters, some probabilities can be calculated. For instance, a researcher wants to know what is the probability that 10 or more deaths will occur in Iran in one day. Else, a researcher also wants to know that what is the probability

that 10 or more daily new cases will occur in Hong Kong. To answer these research questions, the estimated parameters of the DsGLi distribution and its cdf can be used. The probabilities related to these two research questions are calculated for different values of the deaths and daily new cases and reported in Table 11.

Table 11: The calculated probabilities for the daily new deaths and cases for Iran and Hong Kong.

Counts of deaths in Iran	Probability for Iran	Counts of new cases in Hong Kong	Probability for Hong Kong
Over 10	0.4015	Over 30	0.0251
Over 20	0.1751	Over 40	0.0076
Over 30	0.0764	Over 50	0.0023

The counts of deaths in OECD countries are modeled with some covariates by using the DsGLi and Poisson regression models in the third application. According to the estimated regression parameters, we conclude the following results.

- ✓ The proportion of smokers in the population does not affect the counts of deaths.
- ✓ In a kind of funny way, the countries having more alcohol consumption have lower counts of deaths.
- ✓ The proportion of obese individuals in the population does not affect the counts of deaths.
- ✓ The air pollution does not affect the counts of deaths.
- ✓ The countries with a population of over 35 million have $\exp(4.2760) = 71.9521$ times more counts of deaths than countries with a population below 7 million.
- ✓ The countries with a population of 7-35 million have $\exp(2.0978) = 8.1482$ times more counts of deaths than countries with a population below 7 million.

7 Conclusion remarks

COVID-19 is still an unclear infectious disease. Each country's social and policy responsibility is affected by the COVID-19 outbreak. In this paper, our aim is to try to serve humanity in this difficult situation by modeling this outbreak by utilizing a new probability distribution, and therefore we have proposed a new two-parameter discrete gamma Lindley distribution for modeling such this count data. Some important statistical properties have been derived in closed forms which makes this model more flexible in practical fields, especially, in prediction sciences. The model parameters have been estimated by using the maximum likelihood approach which gives a unique estimator. The flexibility of the proposed model has been explained by utilizing three COVID-19 data sets in different countries. We hope that the empirical results of this study will be useful for policymakers, healthcare professionals, and researchers studying in this field.

References

- [1] Altun, E. (2020). A new generalization of geometric distribution with properties and applications. Communications in Statistics-Simulation and Computation, 49(3), 793-807.

- [2] Altun, E. (2019). A new model for over-dispersed count data: Poisson quasi-Lindley regression model. *Mathematical Sciences*, 13(3), 241-247.
- [3] Ayyoubzadeh, S. M., Ayyoubzadeh, S. M., Zahedi, H., Ahmadi, M. and Kalhori, S. R. N. (2020). Predicting COVID-19 Incidence Through Analysis of Google Trends Data in Iran: Data Mining and Deep Learning Pilot Study. *JMIR Public Health and Surveillance*, 6(2), e18828.
- [4] Alamatsaz, M., Dey, H., Dey, S., Harandi, T., and Shams, S., (2016). Discrete generalized Rayleigh distribution. *Pakistan journal of statistics*, 32(1), 1-20.
- [5] Bakouch, H. S., Aghababaei, M., and Nadarajah, S., (2014). A new discrete distribution. *Statistics*, 48(1), 200-240.
- [6] Batista, M. (2020). Estimation of the final size of the coronavirus epidemic by the SIR model. Online paper, ResearchGate.
- [7] Bebbington, M., Lai, C. D., Wellington, M., and Zitikis, R., (2012). The discrete additive Weibull distribution: a bathtub-shaped hazard for discontinuous failure data. *Reliability engineering and system safety*, 106, 37-44.
- [8] Caccavo, D. (2020). Chinese and Italian COVID-19 outbreaks can be correctly described by a modified SIRD model. medRxiv.
- [9] El-Morshedy, M., Eliwa, M. S., and Nagy, H., (2020a). A new two-parameter exponentiated discrete Lindley distribution: properties, estimation and applications. *Journal of Applied Statistics*, 47(2), 354-375.
- [10] El-Morshedy, M., Eliwa, M. S. and Altun, E. (2020b). Discrete Burr-Hatke Distribution With Properties, Estimation Methods and Regression Model. *IEEE Access*, 8, 74359-74370.
- [11] Eliwa, M. S., Alhussain, Z. A., and El-Morshedy, M. (2020a). Discrete Gompertz-G family of distributions for over-and under-dispersed data with properties, estimation, and applications. *Mathematics*, 8(3), 358.
- [12] Eliwa, M. S., Altun, E: El-Dawoody, M., El-Morshedy, M. (2020b). A new three-parameter discrete distribution with associated INAR(1) process and applications, *IEEE Access*, DOI=10.1109/ACCESS.2020.2993593.
- [13] Eliwa, M. S. and El-Morshedy, M. (2020). A one-parameter discrete distribution for over-dispersed data: Statistical and reliability properties with estimation approaches and applications, *Journal of Applied Statistics*, Forthcoming, to be published.
- [14] Gómez-Déniz, E., and Calderín-Ojeda, E., (2011). The discrete Lindley distribution: properties and applications. *Journal of statistical computation and simulation*, 81(11), 1405-1416.
- [15] Krishna, H. and Pundir, P. S. (2009). Discrete Burr and discrete Pareto distributions. *Statistical Methodology*, 6(2), 177-188.
- [16] Maleki, M., Mahmoudi, M. R., Wraith, D. and Pho, K. H. (2020). Time series modelling to forecast the confirmed and recovered cases of COVID-19. *Travel Medicine and Infectious Disease*, 101742.

- [17] Messaadia, H., and Zeghdoudi, H., (2017). Around gamma Lindley distribution. *Journal of modern applied statistical methods*, 16(2), 23.
- [18] Nesteruk, I. (2020). Statistics-based predictions of coronavirus epidemic spreading in mainland China, *Innovative Biosystems and Bioengineering*, 4(1), 13-18.
- [19] OECD (2019), *Health at a Glance 2019: OECD Indicators*, OECD Publishing, Paris, <https://doi.org/10.1787/4dd50c09-en>.
- [20] Nedjar, S., and Zeghdoudi, H., (2016). On gamma Lindley distribution: Properties and simulations. *Journal of computational and applied mathematics*, 298, 167-174.
- [21] Nekoukhou, V., Alamatsaz, M. H., and Bidram, H., (2013). Discrete generalized exponential distribution of a second type. *Statistics*, 47 (4), 876-887.
- [22] Remuzzi, A. and Remuzzi, G. (2020). COVID-19 and Italy: what next?. *The Lancet*.
- [23] Roy, D., (2003). The discrete normal distribution. *Communication in statistics: theory and methods*, 32(10), 1871-1883.
- [24] Roy, D. (2004). Discrete Rayleigh distribution. *IEEE Transactions on Reliability*, 53(2), 255-260.

Declarations

Competing interests: The authors declare no competing interests.

Figures

Figure 1

The pmf plots of the DsGLi distribution.

Figure 2

The hrf plots of the DsGLi distribution.

Figure 3

The simulation results of the DsGLi distribution.

Figure 4

The estimated cdfs (left) and PP plots (right) for Hong Kong data set.

Figure 5

The estimated cdfs (left) and PP plots (right) for Iran data set.

Figure 6

The residual results of DsGLi regression model.