Sains Malaysiana 54(1)(2025): 313-323
http://doi.org/10.17576/jsm-2025-5401-25

A Ratio-Type Weighted Geometric Distribution for Modelling Overdispersed Count Data

(Taburan Geometri Berpemberat Jenis Nisbah untuk Memodelkan Data Bilangan Terlebih Serakan)

SHIN ZHU SIM[1], HASSAN S. BAKOUCH[2], RAZIK RIDZUAN MOHD TAJUDDIN[3,*] & ULYA ABDUL RAHIM[4]

[1]*School of Mathematical Sciences, University of Nottingham Malaysia, 43500 Semenyih, Selangor, Malaysia*
[2]*Department of Mathematics, College of Science, Qassim University, Buraydah, Saudi Arabia*
[3]*Department of Mathematical Sciences, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor, Malaysia*
[4]*Deanery Office, Faculty of Medicine, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor, Malaysia*

Received: 11 June 2024/Accepted: 7 October 2024

ABSTRACT

Weighted distributions have always been a popular approach in developing flexible distributions for data modelling. In this paper, we introduce a flexible ratio-type weighted geometric distribution by adopting the geometric distribution as a basic standard distribution and opting for weights, represented as $w(x) = (x + 1)/(x + 2)$. The proposed distribution is overdispersed and is capable of accommodating data with small mode values such as 0, 1 and 2. The proposed distribution has the following properties – unimodal, log-concave and has increasing failure rates. The moment estimator is obtained, and the resulting estimated parameter is utilized as the initial point in finding the estimators based on the maximum likelihood technique and probability generating function. A probability comparison between the typical geometric distribution and the proposed distribution is discussed as well. A collection of insurance claim datasets is utilized for model fitting, and it was found out that generally, the proposed distribution can adequately fit the datasets as opposed to other contending distributions.

Keywords: Discrete distributions; estimation geometric; simulation; weights

ABSTRAK

Taburan berpemberat selalu menjadi pendekatan yang popular dalam mengembangkan taburan fleksibel untuk pemodelan data. Dalam kajian ini, kami memperkenalkan taburan geometrik berpemberat jenis nisbah yang fleksibel dengan mengambil kira taburan geometrik sebagai taburan piawai asas dan memilih berat, yang diwakili sebagai $w(x) = (x + 1)/(x + 2)$. Taburan yang dicadangkan adalah terlebih serak dan mampu mengendalikan data dengan nilai mod kecil seperti 0, 1 dan 2. Taburan yang dicadangkan mempunyai sifat berikut – unimodal, log-cekung dan mempunyai kadar kegagalan yang meningkat. Penganggar momen diperoleh, dan parameter yang dianggarkan digunakan sebagai titik permulaan dalam mencari penganggar berdasarkan teknik kebolehjadian maksimum dan fungsi penjana kebarangkalian. Perbandingan kebarangkalian antara taburan geometrik biasa dan taburan yang dicadangkan turut dibincangkan. Koleksi set data tuntutan insurans digunakan untuk pemodelan dan didapati bahawa secara umum, taburan yang dicadangkan dapat memadankan set data dengan baik berbanding taburan lain yang dipertimbangkan.

Kata kunci: Geometrik; pemberat; penganggaran; simulasi; taburan diskret

INTRODUCTION

When dealing with positive count data modelling, it is advisable to explore a mixed distribution that involves truncation or alternatives with similar effects, such as weighted distributions, to address the variability in the data. Various weighted distributions, including one termed 'probability proportional to the size', where the weight is proportionate to the observation size, have been introduced. Biased data are prevalent across various scientific disciplines, prompting statisticians and researchers to diligently seek solutions for addressing these biases. To remove biases and achieve an appropriate distribution, researchers commonly employ the weighted concept of biased observation. This approach contributes to the formulation of a weighted distribution. This paper effectively aims to assimilate weighted approach for a well-known discrete distribution and thus, enriching the literatures in this field.

Assuming a random variable X follows a probability function $f(x)$, which can be a probability mass in the case of discrete X or probability density for continuous X, and the likelihood of recording the observation x is $0 < w(x) < 1$, then the $f(x)$ of the recorded observation, X^w, denoted as $f^w(x)$, , can be expressed as follows:

$$f^w(x) = \frac{w(x) f(x)}{\omega},$$

where the normalizing factor, denoted as ω is acquired to ensure that the overall probability equals unity. Consequently, ω can be termed as the visibility factor. It is important to observe that f^w equals f only when $w(x)$ remains constant. When $w(x) = x$, the resulting f^w is the pmf for a size-biased distribution, a special case of weighted distribution.

The weighted distribution theory has emerged as a comprehensive approach to model biased data, with Fisher (1934) initially exploring it, followed by Rao (1965) and Patil (1991) who further examined it in a unified manner. These pioneers identified scenarios in which recorded observations cannot be regarded as a random sample from the original distribution, encompassing non-experimental, non-replicated, and non-random categories. Such deviations may arise from factors like the non-observability of occurrences, partial destruction of observations, and sampling with uneven probabilities of observations (Rao 1965).

Typically, when the events are unobserved, it leads to data truncation. Additionally, certain data, particularly those originating from natural sources, might face destruction. Moreover, focusing on a specific event and retracing its actual occurrences in the population may not afford an equitable chance for the event to manifest within the population. These factors introduce distortions in the collected sample data. Consequently, opting for size-biased distributions becomes a logical choice for modelling this type of data. Examples of size-biased distributions encompass the size-biased Poisson, size-biased binomial, and size-biased negative binomial distributions, along with the size-biased hypergeometric distribution (Patil & Rao 1978).

Recently, the use of weighted distributions has consistently been a favoured method in the development of adaptable probability distributions with different number of parameters. Some of the examples include the works by Bhati and Joshi (2018) as well as Gupta and Kundu (2009). Bhati and Joshi (2018) derived geometric distribution through a power-type weight function, $w(x) = 1 - p^{\alpha(x+1)}$ which can also be viewed as a discrete analogue of the weighted exponential distribution introduced by Gupta and Kundu (2009), resulting in a two-parameter weighted geometric distribution. This paper does consider geometric distribution as the underlying base distribution but with a relatively simpler weight function, which results in a one-parameter weighted geometric distribution.

Additionally, several researchers have incorporated variations of this weighted distribution in discrete data. Recently, Almuhayfith et al. (2023) have considered a ratio weight and Poisson distribution in developing a novel discrete distribution, named as Semi-Poisson distribution. Bakouch (2018) introduced a flexible discrete distribution, which involves the negative binomial and size-biased negative binomial distributions as sub-models among others, and it is a weighted version of the two-parameter discrete Lindley distribution. The credibility of the proposed distribution by Bakouch (2018), is recommended for several types of over- and under-dispersed count data. These findings align with the results of the earlier study conducted by Del Castillo and Pérez-Casany (1998) where they introduced new exponential families derived from the concept of weighted distribution, encompassing, and extending the Poisson distribution. This feature renders them suitable for accommodating discrete data in situations of overdispersion or underdispersion.

In contrast, Ridout and Besbeas (2004) proposed the weighted Poisson distribution as a model for counting data exhibiting underdispersion. The credibility of the proposed distribution is recommended for modelling strong underdispersion. Tajuddin & Ismail (2023) have modelled underdispersed count data by proposing a one parameter size-biased Poisson distribution. However, the flexibility of the distribution is limited as it relies exclusively on a single parameter Tajuddin & Ismail (2023). Recently, Puig, Valero and Fernández-Fontelo (2024) studied the mechanisms leading to underdispersion in count data by using weighted Poisson and other well-known distributions.

The paper is structured as follows. We introduce a ratio-type weighted geometric (RWG) distribution and provide some of its distributional properties in the next section. After that, we addresses the issue of estimating the proposed model using three estimation methods – moment, maximum likelihood and probability generating function-based estimators, supported by an accompanying simulation analysis. Subsequently, we employ the proposed distribution to undertake fitting of selected automobile claim data. In an effort to exhibit the practicality of the proposed model, we compare its performance with two other single-parameter models. The conclusions are presented in the final section of the paper.

DEFINITION AND DISTRIBUTIONAL PROPERTIES

Let X be a geometric random variable defined by $\Pr(X = x) = (1-p)p^x$ where $0 < p < 1$ for $x = 0, 1, 2, \cdots$. Assume that the probability of ascertaining the event X has a weighting factor $w(x) = \frac{x+1}{x+2}$. It is important to note that $0 < w(x) < 1$ because

$x+1 < x+2$. This will ensure that the weighting factor to decrease slowly at a gradient, $w'(x) = (x+2)^{-2}$. The weighted distribution corresponds to the selected weight $w(x)$ has probability mass function (pmf) as

$$\Pr\left(X^{w}=x\right)=\frac{w(x)\Pr\left(X=x\right)}{\sum_{x=0}^{\infty} w(x)\Pr\left(X=x\right)};x=0,1,2,...$$

where

$$\sum_{x=0}^{\infty} w(x)\Pr\left(X=x\right)=\frac{p+(1-p)\ln(1-p)}{p^{2}(1-p)}.$$

We obtain a ratio-type weighted geometric (RWG) distribution given by

$$\Pr\left(X^{w}=x\right)=C_{p}\left(\frac{x+1}{x+2}\right)p^{x+2}, \quad (1)$$

where $C_p = \frac{1-p}{p+(1-p)\ln(1-p)}$. The RWG distribution can also be obtained if $X_2 \sim \Pr(X_2=x)=\frac{(1-p)^2(1+\beta x)}{1+p(\beta-1)}p^x$ and $\beta = 1$, such that $\Pr(X_2=x)=(1-p)^2(1+x)p^x$ and $w_2(x)=\frac{1}{x+2}$, then $\Pr(X^*=x)=\frac{w_2(x)}{E[w_2(x)]}\Pr(X_2=x)$. The $\Pr(X_2=x)$ refers to the probability mass function for the two-parameter discrete Lindley distribution with $\beta = 1$ (8). A similar weight function was employed by considering Poisson distribution as the baseline distribution to develop a Semi-Poisson distribution (Almuhayfith et al. 2023). To examine the shapes of the proposed RWG distribution, the probabilities are computed for different values of parameter p and presented in Figure 1.

Critical patterns within the data have been observed in Figure 1. With a small p value, an overly represented peak

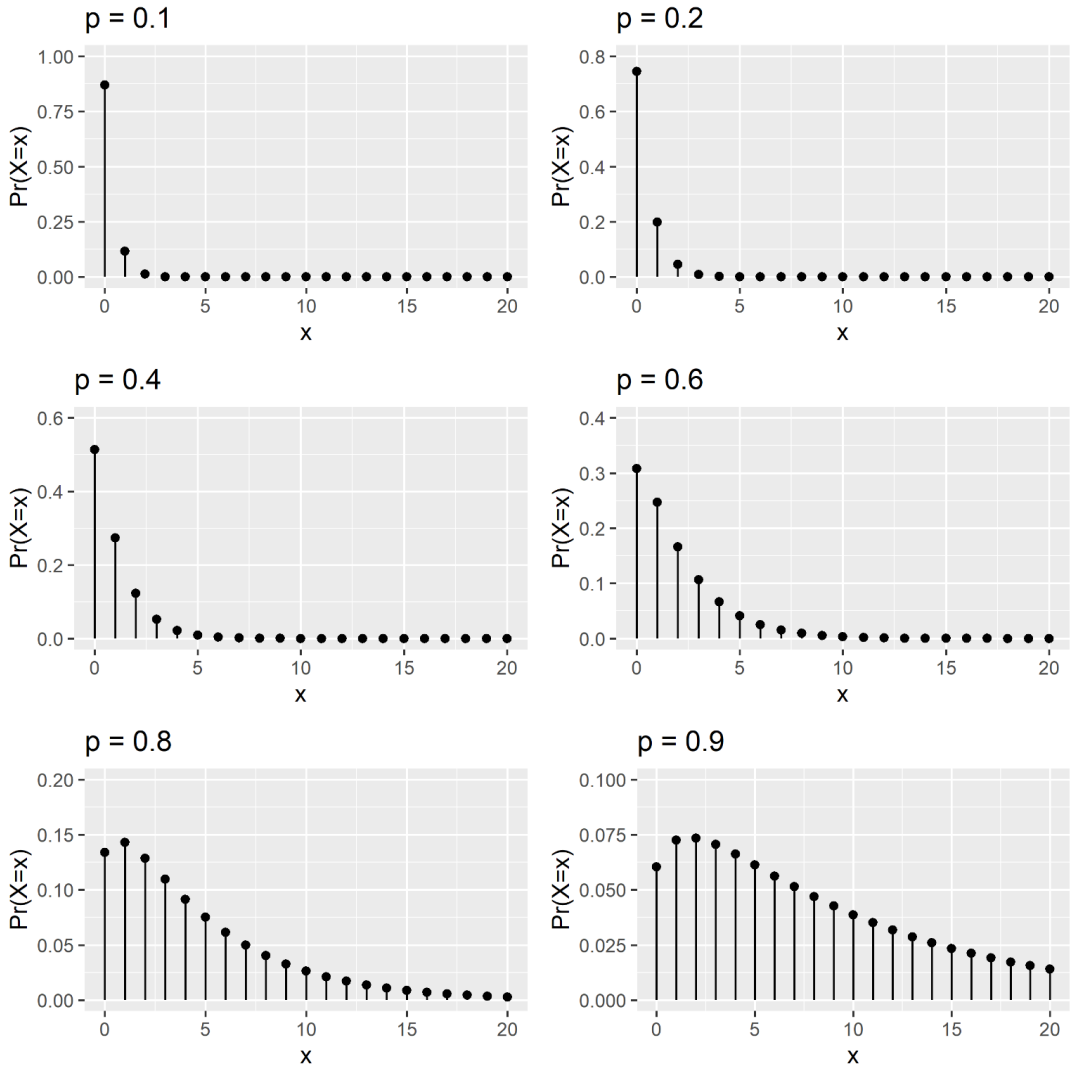

FIGURE 1. Graphs of probability mass functions for p = 0.1, 0.2, 0.4, 0.6, 0.8 and 0.9 respectively, for the proposed RWG distribution

at zero is evident, which suggests an excess of zeros in the data. However, an inverse relationship exists as the value of p increases; the prominence of the zero-mode diminishes, and the magnitude of other data points escalates. At $p = 0.8$, the mode transitions from zero, taking a position at one. This trend continues to the point where the mode approaches two, the largest value observed. This implies that with an increase in the value of p, the mode consistently shifts further from zero, making it plausible for the mode to reach as high as two.

SOME DISTRIBUTIONAL PROPERTIES

Without loss of generality, X^w is referred as X from hereon for simplicity. The n^{th} moment of RWG distribution is defined as

$$E\left(X^n\right) = C_p \sum_{x=0}^{\infty} x^n \left(\frac{x+1}{x+2}\right) p^{x+2} \tag{2}$$

Thus, the first four moments of the RWG distribution are:

$$E(X) = \frac{p(3p-2) - 2(1-p)^2 \ln(1-p)}{(1-p)\left[p + (1-p)\ln(1-p)\right]}, \tag{3}$$

$$E\left(X^2\right) = \frac{2p\left(4p^2 - 5p + 2\right) + 4(1-p)^3 \ln(1-p)}{(1-p)^2\left[p + (1-p)\ln(1-p)\right]}, \tag{4}$$

$$E\left(X^3\right) = \frac{2p(2p-1)\left(5p^2 - 6p + 4\right) - 8(1-p)^4 \ln(1-p)}{(1-p)^3\left[p + (1-p)\ln(1-p)\right]}, \tag{5}$$

$$E\left(X^4\right) = \frac{2p\left(24p^4 - 47p^3 + 63p^2 - 36p + 8\right) + 16(1-p)^5 \ln(1-p)}{(1-p)^4\left[p + (1-p)\ln(1-p)\right]}. \tag{6}$$

Using Equations (3) and (4), the variance of the RWG distribution is obtained as:

$$Var(X) = \frac{p^2\left[p(2-p) + (1-p)\ln(1-p)\right]}{(1-p)^2\left[p + (1-p)\ln(1-p)\right]^2}. \tag{7}$$

Hence, the index of dispersion is given as:

$$IOD = \frac{Var(X)}{E(X)} = \frac{p^2\left[-p(2-p) - 2(1-p)\ln(1-p)\right]}{(1-p)\left[p + (1-p)\ln(1-p)\right]\left[p(2-3p) + 2(1-p)^2 \ln(1-p)\right]}. \tag{8}$$

It is unclear whether the RWG distribution is overdispersed or underdispersed or both from Equation (8). This warrants further inspection visually. Equations (3-6) can be used determine the coefficient of variation, the skewness and the kurtosis using the well-known standard definitions. The mode for the distribution can be obtained by differentiating the log of probability mass function given by Equation (1) to get:

$$\ln \Pr\left(X = x\right) \propto \ln(x+1) + (x+2)\ln p - \ln(x+2).$$

Differentiating and setting the equation above to 0, yields a quadratic equation, which has the solution:

$$x_{mod} = \frac{-3\ln p - \sqrt{(\ln p)(\ln p - 4)}}{2\ln p}. \tag{9}$$

The integer part of the x_{mod}, $\lfloor x_{mod} \rfloor$ will be taken as the mode of the RWG distribution. To provide a clearer visual representation of the moment-based measures, Figure 2 displays the plots for the variance, the index of dispersion, the coefficient of variation, the skewness, the kurtosis, and the mode for the RWG distribution as p varies. From the figure, the RWG distribution is always overdispersed according to the index of dispersion $(IOD > 1)$. Besides that, as p increases, the RWG distribution will be concentrated around the mean. The RWG is also positively skewed so the right-tail of the data is long, which is supported by plots in Figure 1 as well. The kurtosis values are greater than 3, which means the excess kurtosis is always positive, hence the RWG distribution is described as leptokurtic. In other words, the RWG distribution will have fatter tail as supported by plots in Figure 1 as well. The RWG distribution can adequately explain data with modes 0, 1 and 2 as can be seen from mode plots of Figure 2.

Describing a distribution using recurrence relation may be useful in identifying the modality and the shape of the distribution. The recurrence relation for the RWG distribution can be represented as:

$$r(x) = \frac{\Pr(X = x+1)}{\Pr(X = x)} = \frac{p(x+2)^2}{(x+1)(x+3)} = p\left[1 + \frac{1}{(x+1)(x+3)}\right] > p < 1,$$

where

$$\Pr(X = 0) = \frac{p^2(1-p)}{2\left[p + (1-p)\ln(1-p)\right]}.$$

The inequality for $r(x) < 1$ shows that the RWG distribution is always decreasing. The difference in recurrence relation can be written as:

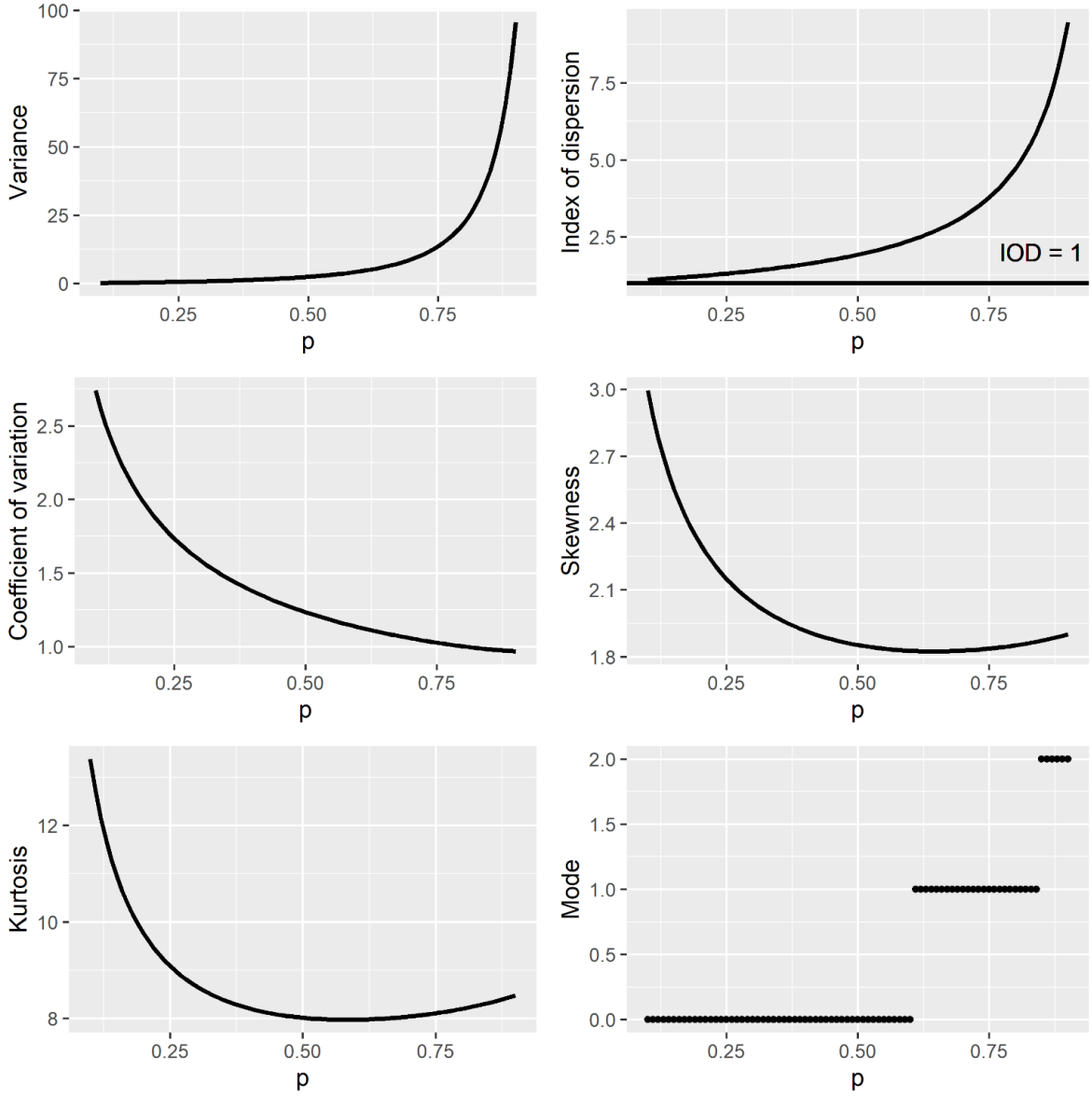

FIGURE 2. Variance, *IOD*, coefficient of variation, skewness, kurtosis, and mode for RWG distribution

$$\Delta r(x) = r(x) - r(x-1) = -\frac{p(2x+3)}{x(x+1)(x+2)(x+3)} < 0.$$

The negative $\Delta r(x)$ shows that the RWG is unimodal. The ratio of recurrence relation is given as:

$$\frac{r(x)}{r(x-1)} = -\frac{x(x+2)^3}{(x+1)^3(x+3)} < 1,$$

for $x \geq 1$, which implies that

$$\left[\Pr(X=x)\right]^2 > \left[\Pr(X=x+1)\right]\left[\Pr(X=x-1)\right] > 1,$$

which subsequently implies log-concavity. Since the RWG is log-concave, it also has increasing failure rates.

SOME GENERATING FUNCTIONS

The probability generating function (pgf) of a discrete random variable following RWG distribution can be expressed as

$$G_X(t)=\frac{p^2(1-p)}{2\left(p+(1-p)\ln(1-p)\right)}\left[\frac{1}{1-pt}-\Phi(pt,1,2)\right], \quad (10)$$

where

$$\Phi(z,s,a)=\sum_{k=0}^{\infty}\frac{z^k}{(k+a)^s}$$

for $|z|<1$ and $a\neq 0,-1,\cdots$, which is famously known as the Lerch transcendent function. The pgf in Equation (10) will be utilized to develop a pgf-based estimator for p.

Consequently, the moment generating function for a discrete random variable following RWG distribution can be expressed as:

$$M_X(t)=G_X(e^t)=\frac{p^2(1-p)}{2\left(p+(1-p)\ln(1-p)\right)}\left[\frac{1}{1-pe^t}-\Phi(pe^t,1,2)\right]. \quad (11)$$

The cumulant generating function for a discrete random variable following RWG distribution can be expressed as:

$$K_X(t)=\ln M_X(t)=\ln\left\{\frac{p^2(1-p)}{2\left(p+(1-p)\ln(1-p)\right)}\left[\frac{1}{1-pe^t}-\Phi(pe^t,1,2)\right]\right\}. \quad (12)$$

SURVIVAL AND HAZARD RATE FUNCTIONS

The survival function for RWG distribution can be obtained as:

$$S(k)=\Pr(X\geq k)=\sum_{x=k}^{\infty}\Pr(X=x)=\frac{p^{k+2}}{p+(1-p)\ln(1-p)}\left[1-(1-p)\Phi(p,1,k+2)\right], \quad (13)$$

where $\Phi(z,s,a)$ is the Lerch Transcendent Function. Using the survival function, the hazard rate function (hrf) can be obtained and given as:

$$h_x=\frac{\Pr(X=x)}{S(x)}=\frac{1-p}{1-(1-p)\Phi(p,1,x+2)}\left(\frac{x+1}{x+2}\right). \quad (14)$$

The limiting value of h_x as $x\to\infty$ is:

$$\lim_{x\to\infty}h_x=\lim_{x\to\infty}\frac{1-p}{1-(1-p)\Phi(p,1,x+2)}\left(\frac{x+1}{x+2}\right)=1-p, \quad (15)$$

because $\lim_{x\to\infty}\Phi(p,1,x+2)=0$ and $\lim_{x\to\infty}\left(\frac{x+1}{x+2}\right)=1$, which means the hrf is bounded above by $1-p$. The limiting value of h_x is also presented in Figure 3. From Figure 3, when p is small $(p\leq 0.2)$, the hrf is almost uniform whereas for larger values of p, $(p>0.2)$, the hrf is increasing in a logistic fashion or also known as 'increasing-constant'.

SOME COMPARISON OF PROBABILITY VALUES BETWEEN THE RATIO-TYPE WEIGHTED GEOMETRIC AND THE GEOMETRIC DISTRIBUTIONS

Since the RWG distribution is built on the geometric distribution, it is only reasonable to compare the probability values for the two distributions. Let $X_1\sim RWG(p)$ and $X_2\sim Geom(p)$ for $x=0,1,2,\ldots$ and $0<p<1$. So,

$$\Pr(X_1=x)=\frac{1-p}{p+(1-p)\ln(1-p)}\left(\frac{x+1}{x+2}\right)p^{x+2},$$

and

$$\Pr(X_2=x)=(1-p)\,p^x.$$

Let τ_x be the ratio of $\Pr(X_1=x)$ to $\Pr(X_2=x)$. So,

$$\tau_x=\frac{\Pr(X_1=x)}{\Pr(X_2=x)}=\frac{p^2}{p+(1-p)\ln(1-p)}\left(\frac{x+1}{x+2}\right)=\frac{p^2}{p+(1-p)\ln(1-p)}w(x).$$

It is easy to show that the lower and upper bound of the limits when $p\to 0^+$ and $p\to 1^-$ which is given as

$$w(x)<\tau_x<2w(x).$$

Generally, the ratio of probability values, τ_x is bounded between $w(x)$ and twice of $w(x)$. When $x=0$, $\frac{1}{2}<\tau_0<1$ for $0<p<1$. This distribution can only adequately explain data with number of zeroes less than from those estimated using geometric distribution.

STATISTICAL INFERENCES

In this section, we explore three methods of estimation of parameter p and a simulation study is carried out to study the properties of these estimators.

MOMENT ESTIMATION

The moment-based estimator $\tilde{p}$ can be obtained by equating the population moment to the corresponding sample moment. The equation is given as

$$\bar{x}=\frac{\tilde{p}(3\tilde{p}-2)-2(1-\tilde{p})^2\ln(1-\tilde{p})}{(1-\tilde{p})\left[\tilde{p}+(1-\tilde{p})\ln(1-\tilde{p})\right]}. \quad (16)$$

Although Equation (16) cannot be expressed as an explicit formula, it can be resolved quickly, even by using the 'FindRoot' feature in the Wolfram Mathematica program. The estimated $\tilde{p}$ will only serve as the initial value for maximum likelihood estimator (MLE) and pgf-based estimation methods, which are discussed in the following subsections.

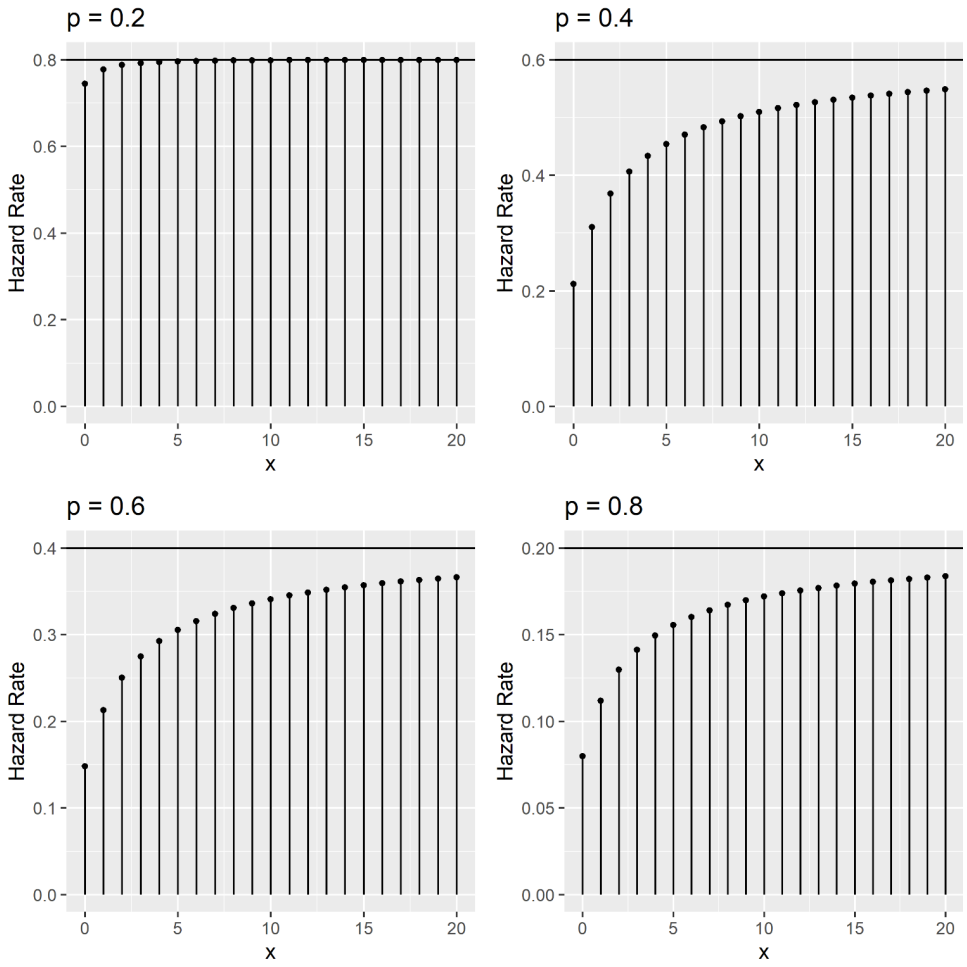

FIGURE 3. Hazard rate function for RWG distribution with different values of parameter p

MAXIMUM LIKELIHOOD ESTIMATION

The MLE is a widely favoured and efficient method for estimating unknown parameters. The log-likelihood function for the proposed distribution is

$$\ln L = \sum_{i=1}^{N} I_{\{X_i = x_i\}} \ln \Pr\left(X_i = x_i\right),$$

where $X_1, X_2, \cdots, X_N$ is a random sample. The first derivative of the log likelihood function is given by:

$$\frac{d \ln L}{dp} = \sum_{i=1}^{N} I_{\{X_i = x_i\}} \left[\ln(1-p) + (x+2)\ln p + \ln\left(\frac{x+1}{x+2}\right) - \ln\left[p + (1-p)\ln(1-p)\right] \right]. \quad (17)$$

The maximum likelihood estimate of p, denoted as $\hat{p}$, will be obtained by equating $d \ln L/dp = 0$. As this equation lacks a closed-form solution, numerical optimization with the *scipy.optimize* Python library is employed to compute the MLE.

PROBABILITY GENERATING FUNCTION-BASED ESTIMATION

The pgf-based estimator method is recommended due to its demonstrated consistency in estimating parameters for discrete distributions and its robustness in handling outliers, as demonstrated in prior research (Sim & Ong 2010). The pgf-based statistic considered here is

$$T = \int_0^1 \left[F_N(t) - F(t) \right]^2 dt,$$

where $F_N(t) = \frac{1}{N}\sum_{i=1}^{N} t^{x_i}$ and $F(t) = G_X(t) = E[t^X]$ are, respectively, the empirical pgf and the theoretical pgf of the distribution.

SIMULATION STUDY

This section reports the Monte Carlo simulation study on the performance of the proposed statistics for MLE and pgf-based estimation with respect to the mean square error (MSE) and bias. The estimated $\tilde{p}$ from the moment estimation technique will act as the initial value for finding the MLE and the pgf-based estimator. Sample sizes of n =50, 100, 500, 1000, and 2000 were examined, spanning small to large sample sizes, while different values of the parameter p were considered. Samples were generated using the inverse transform method.

It was found that after conducting 1000 simulation runs, results of sufficient accuracy were obtained. The bias and mean square error (MSE) of the simulated estimates are given as $\frac{1}{N}\sum_i^N(\hat{p}_i - p)$ and $\frac{1}{N}\sum_i^N(\hat{p}_i - p)^2$, respectively, where $\hat{p}_i$ is the ML or pgf-based estimator and p is the actual parameter.

Table 1 presents the average bias and MSE across various parameter values, comparing two different estimation approaches. The results demonstrate that the proposed estimation methods exhibit small average bias and MSE, especially when applied to sample sizes ranging from n =100 to n =2000. Furthermore, the pgf-based estimator consistently exhibits outperform in terms of average bias when compared to the MLE across different sample sizes and parameter values. On the other hand, regardless of sample size or parameter p, the MLE consistently yields lower average MSE values than the pgf-based estimating approach.

APPLICATION

The analysis conducted by Gossiaux and Lemaire (1981) on six sets of over-dispersed automobile data is taken into account. These data sets display the frequency of automobile insurance claims per policy within a predetermined time frame. Joining the analysis, Poisson and the discrete Lindley (DLindley) distribution (Gómez-Déniz & Calderín-Ojeda 2011) are also fit to these six data sets, providing a point of comparison. Tables 2 exhibits the chi-square goodness of fit values for the MLE, and the pgf-based estimation methods. The chi-square is calculated using the following formula:

TABLE 1. Average bias and average MSE of the simulated estimates for MLE and pgf-based (in bracket)

Bias	Parameter p			
	0.2	0.4	0.6	0.8
n =50	-0.00394	-0.00606	-0.00612	-0.00435
	(-0.00051)	(-0.00335)	(-0.00302)	(-0.00341)
n =100	-0.00218	-0.00301	-0.00324	-0.00242
	(0.00025)	(-0.00084)	(-0.00075)	(-0.00126)
n =500	-0.00107	-0.00105	-0.00249	-0.00101
	(0.00039)	(0.000004)	(0.00014)	(-0.00027)
n =1000	-0.00133	-0.00128	-0.00150	-0.00110
	(0.00006)	(-0.00031)	(0.00004)	(-0.00021)
n =2000	-0.00110	-0.00109	-0.00129	-0.00122
	(0.00023)	(-0.00006)	(0.00023)	(0.00002)
MSE				
n =50	0.00183	0.00228	0.00164	0.00059
	(0.00211)	(0.00294)	(0.00266)	(0.00120)
n =100	0.00097	0.00119	0.00086	0.00031
	(0.00110)	(0.00151)	(0.00134)	(0.00060)
n =500	0.00020	0.00023	0.00017	0.00006
	(0.00023)	(0.00031)	(0.00025)	(0.00011)
n =1000	0.00010	0.00012	0.00008	0.00003
	(0.00011)	(0.00015)	(0.00012)	(0.00005)
n =2000	0.00005	0.00006	0.00004	0.00002
	(0.00006)	(0.00008)	(0.00007)	(0.00003)

TABLE 2. Fitting RWG to six automobile data set by Gossiaux and Lemaire (14) using MLE and pgf-based estimation

No of claims	0	1	2	3	4	5	6	7	χ^2	$\hat{p}$
Data1	96978	9240	704	43	9					
Poisson (ML)	96689.53	9773.44	493.95	16.64	0.43				332.18	0.101
Poisson (pgf)	96906.00	9578.62	473.40	15.60	0.39				360.91	0.099
DLindley (ML)	96981.05	9229.71	710.16	49.58	3.49				9.610	0.054
DLindley (pgf)	96979.76	9230.80	710.34	49.60	3.50				9.605	0.054
RWG (ML)	97015.98	9167.45	730.92	55.25	4.40				9.108	0.071
RWG (pgf)	96988.61	9190.44	734.79	55.70	4.45				9.112	0.071
ID = 1.06										
Data2	20592	2651	297	41	7	0	1			
Poisson (ML)	20420.94	2945.10	212.37	10.21	0.37	0.01	0.0003		4112.76	0.144
Poisson (pgf)	20551.34	2833.09	195.28	8.97	0.31	0.01	0.0002		5326.24	0.138
DLindley (ML)	20544.79	2720.36	292.41	28.55	2.64	0.24	0.02		57.67	0.077
DLindley (pgf)	20581.92	2691.18	285.61	27.53	2.51	0.22	0.02		62.27	0.074
RWG (ML)	20559.15	2696.55	298.42	31.31	3.21	0.32	0.04		34.28	0.0984
RWG (pgf)	20585.35	2676.13	293.54	30.53	3.10	0.31	0.03		36.23	0.0975
ID = 1.14										
Data 3	103704	14075	1766	255	45	6	2			
Poisson (ML)	102629.56	15921.95	1235.07	63.87	2.48	0.08	0.002		4176.31	0.16
Poisson (PGF)	103446.02	15228.91	1120.97	55.01	2.02	0.06	0.001		5364.18	0.15
DLindley (ML)	103347.35	14628.38	1682.27	175.79	17.38	1.66	0.17		137.05	0.08
DLindley (pgf)	103621.69	14416.80	1629.28	167.30	16.25	1.52	0.15		151.94	0.08
RWG (ML)	103430.35	14493.39	1713.59	192.10	21.03	2.27	0.27		79.36	0.11
RWG (pgf)	103641.08	14332.25	1672.29	185.00	19.99	2.13	0.25		86.81	0.10
ID = 1.16										
Data 4	370412	46545	3935	317	28	3				
Poisson (ML)	369246.88	48643.58	3204.09	140.70	4.63	0.12			665.91	0.13

continue to next page

continue from previous page

No of claims	0	1	2	3	4	5	6	7	χ^2	$\hat{p}$
Poisson (pgf)	370121.65	47883.01	3097.34	133.57	4.32	0.11			718.77	0.13
DLindley (ML)	371135.88	45202.24	4464.45	400.44	33.97	3.02			122.52	0.07
DLindley (pgf)	370610.00	45623.29	4555.07	413.03	35.43	3.19			127.03	0.07
RWG (ML)	371341.26	44842.45	4568.98	441.39	41.64	4.27			194.84	0.09
RWG (pgf)	370665.33	45379.47	4687.60	459.11	43.91	4.58			201.24	0.09
ID = 1.24										
Data 5	7840	1317	239	42	14	4	4	1		
Poisson (ML)	7635.62	1636.72	175.42	12.53	0.67	0.03	0.0010	0.00003	47412.31	0.21
Poisson (pgf)	7794.60	1510.18	146.30	9.45	0.46	0.02	0.0006	0.00002	90992.85	0.19
DLindley (ML)	7735.57	1463.22	225.71	31.67	4.21	0.54	0.07	0.01	399.08	0.11
DLindley (pgf)	7817.54	1405.97	206.07	27.47	3.47	0.42	0.05	0.01	542.24	0.10
RWG (ML)	7747.18	1446.21	227.79	34.02	4.96	0.71	0.10	0.02	252.56	0.14
RWG (pgf)	7819.93	1395.93	210.25	30.03	4.19	0.58	0.08	0.01	330.30	0.13
ID = 1.35										
Data 6	3719	232	38	7	3	1				
Poisson (MLE)	3668.54	317.33	13.72	0.40	0.01	0.0002			7878.05	0.09
Poisson (PGF)	3708.02	281.05	10.65	0.27	0.01	0.0001			14777.26	0.08
DLindley (ML)	3676.17	302.46	20.08	1.21	0.07	0.004			433.90	0.05
DLindley (pgf)	3709.62	273.24	16.22	0.87	0.04	0.002			710.42	0.04
RWG (ML)	3677.66	300.22	20.68	1.35	0.09	0.01			323.56	0.06
RWG (pgf)	3709.84	272.26	16.86	0.99	0.06	0.003			513.43	0.06
ID = 1.42										

ML is MLE, pgf is pgf-based estimator and χ^2 is the chi-square goodness of fit values

$$\chi^2 = \sum_{x=0}^{\infty} \frac{\left(n_x - \hat{n}_x\right)^2}{\hat{n}_x},$$

where n_x is the observed data; $\hat{n}_x$ is the fitted data for $x = 0,1,2, \ldots$. To highlight the difference between the fitting achieved by MLE and pgf-based estimator, certain values are shown with increased decimal places.

The chi-square values presented in Table 2 indicate that the RWG distribution significantly outperforms the Poisson and Discrete Lindley distributions (either MLE method or the pgf-based estimation), except in the case of data set 4, where the RWG distribution exhibits a chi-square value comparable with that of the Discrete Lindley distribution. In general, the pgf-based estimator aligns well with the MLE, with the exception of data sets 2 and 5. This discrepancy can be attributed to the influence of small values within the final cell size.

CONCLUSION

This paper explores a versatile discrete RWG distribution. This unique approach has potential as an alternative to the Poisson distribution for modelling claim count data. Exhibiting several specific characteristics such as unimodality, overdispersion, log-concavity, increasing failure rates and a high peak at zero. The discrete ratio-type weighted geometric distribution provides intriguing opportunities for more nuanced analysis and understanding. Further improvements on the proposed RWG distribution can be done by taking a generalized weighting factor, $w(x)$

$$w(x)=\frac{x+a}{x+b} \quad \text{for} \quad a<b,$$

where a and b are parameters that can be further estimated. The inequality $a<b$ will ensure the weight, $w(x)$ decreases as x increases, in line with the proposed RWG distribution, where for the RWG distribution, $a=1$ and $b=2$. Besides that, an appropriate linear model with suitable link function may be developed. The resulting distribution using the generalized weight as well as the generalized linear model may have boundless usage.

ACKNOWLEDGMENTS

This study was funded by Geran Galakan Penyelidik Muda from Universiti Kebangsaan Malaysia (Grant no.: GGPM-2023-052). The authors declares that they have no competing interests.

REFERENCES

Almuhayfith, F.E., Bapat, S.R., Bakouch, H.S. & Alnaghmosh, A.M. 2023. A flexible semi-Poisson distribution with applications to insurance claims and biological data. *Mathematics* 11(5): 1-15.

Bakouch, H.S. 2018. A weighted negative binomial Lindley distribution with applications to dispersed data. *Anais da Academia Brasileira de Ciências* 90: 2617-2642.

Bhati, D. & Joshi, S. 2018. Weighted geometric distribution with new characterizations of geometric distribution. *Communications in Statistics-Theory and Methods* 47(6): 1510-1527.

Del Castillo, J. & Pérez-Casany, M. 1998. Weighted Poisson distributions for overdispersion and underdispersion situations. *Annals of the Institute of Statistical Mathematics* 50: 567-585.

Fisher, R.A. 1934. The effect of methods of ascertainment upon the estimation of frequencies. *Annals of Eugenics* 6(1): 13-25.

Gómez-Déniz, E. & Calderín-Ojeda, E. 2011. The discrete Lindley distribution: Properties and applications. *Journal of Statistical Computation and Simulation* 81(11): 1405-1416.

Gossiaux, A.M. & Lemaire, J. 1981. Méthodes d'ajustement de distributions de sinistres. *Bulletin of the Association of Swiss Actuaries* 81: 87-95.

Gupta, R.D. & Kundu, D. 2009. A new class of weighted exponential distributions. *Statistics* 43(6): 621-634.

Patil, G.P. 1991. Encountered data, statistical ecology, environmental statistics, and weighted distribution methods. *Environmetrics* 2(4): 377-423.

Patil, G.P. & Rao, C.R. 1978. Weighted distributions and size-biased sampling with applications to wildlife populations and human families. *Biometrics* 34(2): 179-189.

Puig, P., Valero, J. & Fernández-Fontelo, A. 2024. Some mechanisms leading to underdispersion: Old and new proposals. *Scandinavian Journal of Statistics* 51(1): 245-267.

Rao, C.R. 1965. On discrete distributions arising out of methods of ascertainment. *Sankhyā: The Indian Journal of Statistics, Series A* 27(2/4): 311-324.

Ridout, M.S. & Besbeas, P. 2004. An empirical model for underdispersed count data. *Statistical Modelling* 4(1): 77-89.

Sim, S.Z. & Ong, S.H. 2010. Parameter estimation for discrete distributions by generalized Hellinger-type divergence based on probability generating function. *Communications in Statistics - Simulation and Computation* 39(2): 305-314.

Tajuddin, R. R. M., & Ismail, N. 2023. A New One-Parameter Size-Biased Poisson Distribution for Modelling Underdispersed Count Data. *Malaysian Journal of Fundamental and Applied Sciences* 19(2)1: 64-172.

*Corresponding author; email: rrmt@ukm.edu.my