Applicable Analysis and Discrete Mathematics
available online at http://pefmath.etf.rs

Appl. Anal. Discrete Math. **18** (2024), 431–451.
https://doi.org/10.2298/AADM230705015K

ON THE RADEMACHER PERTURBED GEOMETRIC DISTRIBUTION

*Maher Kachour and Christophe Chesneau**

This paper introduces a new discrete distribution that depends on two parameters. It can be described as a perturbed version of the standard geometric distribution by adding a "random noise" following the Rademacher distribution. Depending on the values of its parameters, this new distribution can be unimodal or bimodal, as well as underdispersed or overdispersed. After reviewing its main properties, the problem of parameter estimation is examined and an application to a practical data set is given.

1. INTRODUCTION

Discrete models are very important in handling count data encountered in several theoretical and practical fields. In particular, earthquakes, traffic accidents, counts of landslides, and the number of people dying from the disease can be modeled by discrete (probability) distributions. Further, the reliability of a switching device is a function of the number of times the switch is operated, or the reliability of a computer is a function of the number of times the computer has broken down. According to [**8**], almost all observed values are actually discrete because they are measured to only a finite number of decimal places and cannot really constitute all points in a continuum.

Many research papers have been published on the study and applications of distributions with support included in $\mathbb{N} = \{0, 1, 2, 3, 4, \ldots\}$. A large number

*Corresponding author. Christophe Chesneau
2020 Mathematics Subject Classification. 60E05, 62E15, 62F10.
Keywords and Phrases. Discrete distributions, Geometric distribution, Rademacher distribution, Underdispersion, Overdispersion, Unimodality, Bimodality, Inflated and deflated values.

of these distributions can be found in [**19**] and [**12**]. The geometric distribution (originally interpreted as the number of independent and identical trials to get the first success, i.e., its support is $\mathbb{N}\backslash\{0\} = \{1, 2, 3, 4, \ldots\}$) and the second-type geometric distribution (originally interpreted as the number of failures of independent and identical trials before getting the first success, i.e., its support is $\mathbb{N}$) are well-known discrete distributions. They have been studied by many researchers due to their empirical applications. In recent decades, there has been an increased interest in constructing new flexible distributions defined on $\mathbb{N}$. In particular, many generalizations of the geometric distribution were attempted via different methods. A brief state of the art on this topic is given below. A generalization of the right-truncated geometric distribution was considered in [**24**]. The author in [**14**] introduced another generalization of the geometric distribution by discretizing the generalized exponential distribution, originally proposed in [**26**]. A new generalization of the geometric distribution was examined in [**9**] by employing the techniques in [**28**]. A new generalization of the geometric distribution based on the quadratic transmutation techniques in [**34**] was introduced in [**7**]. A generalized geometric distribution using a discrete analog of the weighted exponential distribution established in [**15**] was studied in [**6**]. The authors in [**30**] used the skewing mechanism in [**4**] for continuous distributions to derive a new generalization of the geometric distribution. A new parametric extension of the geometric distribution using also Azzalini's method was investigated in [**10**]. The authors in [**2**] applied the transmuted record-type geometric method to construct an extended form of the geometric distribution. A sequence of independent and identical trinomial experiments, and a typical stopping rule to generate a generalized geometric distribution were examined in [**32**].

In parallel to these contributions, many recent research papers have been interested in an "inflated" version of the geometric distribution. Indeed, the geometric and the second-type geometric distributions may be inadequate for dealing with overdispersed count data. This case occurs for instance in the abundance of zero counts in the data. To overcome this over-dispersion problem, the zero-inflated geometric (ZIG) distribution was introduced (for details, see [**19**], [**33**] and the references therein). Explicitly, for a random variable X following the $ZIG(p, \pi)$ distribution, its probability mass function (pmf) is specified by

$$\mathbb{P}(X = k) = \begin{cases} \pi + p(1-\pi) & \text{if } k = 0, \\ (1-\pi)(1-p)^k p & \text{if } k \geq 1, \end{cases}$$

where p is probability of success in the geometric distribution and π is the mixture additional weight, which belongs to $\left(-\frac{p}{1-p}, 1\right)$. Negative values of π, i.e., when $\pi \in \left(-\frac{p}{1-p}, 0\right)$, have a natural interpretation in terms of zero-deflation, relative to a second-type geometric model. Correspondingly, $\pi \in (0, 1)$ can be regarded as zero inflation (see [**19**]) Inspired by the ZIG distribution described above, [**20**] introduced the generalized inflated geometric (GIG) distribution, which permits inflation and deflation at several specific values. Indeed, the inflated geometric

distribution models were recently considered and studied due to their empirical needs and applications. For example, the authors in [**31**] used an inflated geometric model to fit the total number of migrants in the household cohort (including international migrants) of the rural areas of Comilla District, Bangladesh. Furthermore, the author in [**3**] proposed an extension of the geometric distribution zero-one inflated to estimate the frequencies of the number of major derogatory reports in the credit history of individual credit card applicants. More applications based on these models are also presented in [**35**], [**25**], and [**36**].

For the purpose of this paper, the geometric distribution with parameter $p \in (0,1)$, denoted by $\mathrm{G}(p)$, needs to be defined. For a random variable X following the $\mathrm{G}(p)$ distribution, its pmf is given by

$$\mathbb{P}(X = k) = p(1-p)^{k-1}, \ k \in \mathbb{N}\backslash\{0\}.$$

Thus, we consider a " $-1/+1$" perturbed version of the $\mathrm{G}(p)$ distribution. The considered perturbed scheme is as follows: Let X and Y be independent random variables with X following the $\mathrm{G}(p)$ distribution and $Y = 2U - 1$, where U follows the Bernoulli distribution with parameter $\alpha \in (0,1)$. Indeed, in this case, one can say that Y follows a Rademacher (α) distribution. Therefore, we have

$$Y(\Omega) = \{-1, 1\}, \text{ with } \mathbb{P}(Y = 1) = 1 - \mathbb{P}(Y = -1) = \alpha.$$

Thus, Y can model a moderate discrete noise with intensity modulated by α. Hence, the version of X perturbed with an additional "random noise" modeled by Y can be defined as

$$Z = X + Y. \tag{1}$$

Remark 1. *Data perturbation is a popular technique, for example, in the privacy-preserving data mining field. The noise additive perturbation (used here to generate our new geometric distribution) is one of the most basic methods for data perturbation. This type of technique relies on the fact that data owners may not want to equally protect all values in a record; thus, a column-based value distortion can be applied to perturb some sensitive columns (see, e.g., [***1***], [***13***], and [***11***]).*

Remark 2. *Most of the inflated geometric distribution models proposed in the literature are based on the second-type geometric distribution, which already has* $\mathbb{N}$ *as support. Our new distribution is founded on the standard geometric distribution, which is defined on* $\mathbb{N}\backslash\{0\}$*. Thus, noise additive perturbation with* -1 *as minimal modality (such as the Rademacher distribution used here) allows the standard geometric perturbed distribution to obtain* $\mathbb{N}$ *as support.*

The definition of the corresponding distribution is presented below.

Definition 1. *Let* Z *be a random variable as defined by the random sum in Equation* (1)*. According to this structure, we have* $Z(\Omega) = \mathbb{N}$ *and the associated pmf*

can be described as follows:

$$\mathbb{P}(Z=k)=\begin{cases} p(1-\alpha) & \text{if } k=0,\\ p(1-\alpha)(1-p) & \text{if } k=1,\\ p(1-p)^{k-2}\left(\alpha+(1-\alpha)(1-p)^2\right) & \text{if } k\geq 2. \end{cases}$$

Thus, Z is said to follow the Rademacher perturbed geometric distribution with parameters α and p, denoted by R-$G(\alpha,p)$.

Remark 3. *At first glance, based on the expression of its pmf, it will be natural to wonder if there is a link between our new distribution and the already existing inflated geometric distributions. Specifically, it is entirely appropriate to ask whether the Rademacher perturbed geometric distribution represents a special case of the GIG distribution. To address these questions, we must mention two important elements:*

- *The GIG distribution is based on the second-type geometric distribution (which, by definition, has $\mathbb{N}$ as support).*
- *One can consider the pmf of the GIG distribution as "ad-hoc" constructed. Indeed, to obtain an inflated\deflated structure at specific values (for example, at zero in the ZIG distribution, see the pmf described above), an additional weight is used to define the probability of these specific values (indeed, the pmf of a considered specific value is a linear combination between the additional weight and the second-type geometric pmf of this value). The realization probability of the other values is similar to that of the pmf of the second-type geometric distribution (for the same values) with a multiplicative standardization constant, which depends on the additional weights used for the specific values considered.*

Indeed, the noise additive perturbation technique (used to define our new distribution) offers a pmf structure different from that presented above. Assume that W is a second-type geometric random variable (i.e., $W(\Omega)=\mathbb{N}$), with p as a parameter and E a random variable following the Bernoulli(π) distribution, where E and W are independent. Thus, one can easily see that the pmf of $W+E$ is not equal to that of the ZIG distribution.

The pmf of the R-G(α,p) distribution is plotted in Figure 1 for some chosen values of parameters. It reveals the following information:

- A clear geometric decrease from lag 2.
- Depending on parameter values, the R-G(α,p) distribution can zero deflate or inflate (indeed, the pmf at 0 can be written as a bivariate function that depends on p and α, i.e., $\mathbb{P}(Z=0)=f(p,\alpha)=p(1-\alpha)$. This function will increase with large values of p and weak values of α. However, $\mathbb{P}(Z=0)$ will decrease with large values of α).

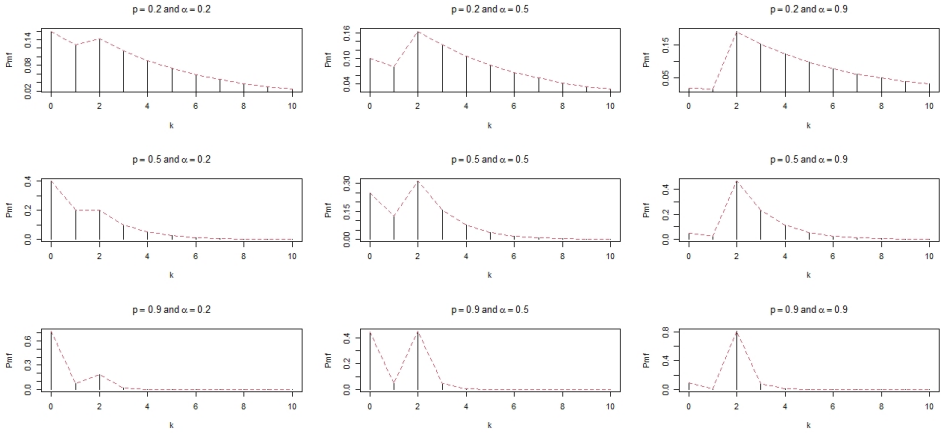

Figure 1: Plots of the pmf of the R-G (α, p) distribution for different parameter value combinations.

- Depending on parameter values, the R-G (α, p) distribution can inflate twice.

The characteristics and flexibility of the distribution can be useful in many fields of application. For example, the R-G (α, p) distribution can model complete female fertility, which accounts for a relative excess of both zero and two children (see [**27**] and [**25**]). Moreover, it can fit data from a dental epidemiological study. Explicitly, the R-G (α, p) distribution can be used to model the DMFT index (which is a count number standing for the number of decayed, missing, and filled teeth) measured on children from a specific age, where data have a surplus amount of zero, one, or two (see [**5**]).

The contents of the paper are arranged as follows: In Section 2, various properties of this distribution, including the mode, the failure rate function and the probability generating function (pgf), are studied. In Section 3, the method of moments and the method of maximum likelihood estimation are used for parameter estimation. Also, a simulation study is carried out to study the performance of the obtained estimates. Application of this distribution in real-world data modeling is illustrated in Section 4, and conclusions are presented in Section 5.

2. MAIN PROPERTIES

This section is devoted to the main theoretical properties of the R-G (α, p) distribution. Hereafter, we consider a random variable Z that follows the R-G distribution.

2.1. Mode(s). First, let us investigate the mode of the R-G (α, p) distribution. Recall that the mode associated with Z, denoted by Mode (Z), it is the value (or

values) k^* such that

$$\mathbb{P}(Z = k^*) \geq \mathbb{P}(Z = k), \ \forall k \in \mathbb{N}.$$

On the other hand, one can see that

- $\mathbb{P}(Z = 2) \geq \mathbb{P}(Z = k), \ \forall k \geq 2,$
- $\mathbb{P}(Z = 0) > \mathbb{P}(Z = 1).$

Therefore, we can conclude that 0 and 2 are the unique "candidates" to be the mode of Z. The precise result is described in the next proposition.

Proposition 1. *Depending on the parameter values, we have*

- *Mode* $(Z) = 2$, *if* $\frac{\alpha}{1-\alpha} > p(2-p)$.
- *Mode* $(Z) = 0$, *if* $\frac{\alpha}{1-\alpha} < p(2-p)$.
- *Mode* $(Z) = \{0, 2\}$, *if* $\frac{\alpha}{1-\alpha} = p(2-p)$.

Proof. Let

$$\Delta = \mathbb{P}(Z = 2) - \mathbb{P}(Z = 0) = p(\alpha - (1-\alpha) p(2-p)).$$

Thus, we have

- $\Delta > 0$ if $\frac{\alpha}{1-\alpha} > p(2-p)$.
- $\Delta < 0$ if $\frac{\alpha}{1-\alpha} < p(2-p)$.
- $\Delta = 0$ if $\frac{\alpha}{1-\alpha} = p(2-p)$.

This ends the proof. □

Remark 4. *This result is consistent with the construction of the distribution (see Equation* (1)*). In fact,* 1 *is the mode of a geometric distribution, and the perturbation resulting from the Rademacher distribution can explain why* 0 *and/or* 2 *are the modes of the R-G*(α, p) *distribution. On the other hand, this result is interesting because it justifies the use of this distribution to fit data with excess* 0 *and/or* 2.

2.2. Cumulative distribution function. The following proposition determines the cumulative distribution function of the R-G (α, p) distribution.

Proposition 2. *Let* $F_Z(a) = \mathbb{P}(Z \leq a)$ *denote the cumulative distribution function associated with* Z. *Thus, we have*

$$F_Z(a) = \begin{cases} 0 & \text{if } a < 0, \\ p(1-\alpha) & \text{if } 0 \leq a < 1, \\ p(1-\alpha)(2-p) & \text{if } 1 \leq a < 2, \\ p(1-\alpha)(2-p) + \left(\alpha + (1-\alpha)(1-p)^2\right)\left(1 - (1-p)^{\lfloor a \rfloor - 1}\right) & \text{if } a \geq 2, \end{cases}$$

where $\lfloor \cdot \rfloor$ *is the floor function (which gives the greatest integer less than or equal to the input real value).*

Proof.

- If $0 \leq a < 1$, then $F_Z(a) = \mathbb{P}(Z = 0) = p(1-\alpha)$.
- If $1 \leq a < 2$, then $F_Z(a) = \mathbb{P}(Z = 0) + \mathbb{P}(Z = 1) = p(1-\alpha)(2-p)$.
- If $a \geq 2$, by using the geometric series formula, we get

$$\begin{aligned} F_Z(a) &= F_Z(1) + \sum_{k=2}^{\lfloor a \rfloor} \mathbb{P}(Z = k) \\ &= F_Z(1) + \left(\alpha + (1-\alpha)(1-p)^2\right)\left(p \sum_{j=0}^{\lfloor a \rfloor - 2} (1-p)^j\right) \\ &= p(1-\alpha)(2-p) + \left(\alpha + (1-\alpha)(1-p)^2\right)\left(1 - (1-p)^{\lfloor a \rfloor - 1}\right). \end{aligned}$$

This ends the proof. □

Remark 5. *Let $a \in \mathbb{N}$. Based on Proposition 2, one can easily define the corresponding failure rate function as follows:*

$$\kappa_Z(a) = \mathbb{P}(Z = a \mid Z \geq a) = \frac{\mathbb{P}(Z = a)}{\mathbb{P}(Z \geq a)} = \frac{\mathbb{P}(Z = a)}{1 - F_Z(a-1)}$$

$$= \begin{cases} p(1-\alpha) & \text{if } a = 0, \\ \dfrac{p(1-p)(1-\alpha)}{1-p(1-\alpha)} & \text{if } a = 1, \\ \dfrac{p(1-p)^{a-2}\left(\alpha + (1-\alpha)(1-p)^2\right)}{1 - \left(p(1-\alpha)(2-p) + \left(\alpha + (1-\alpha)(1-p)^2\right)\left(1 - (1-p)^{a-2}\right)\right)} & \text{if } a \geq 2. \end{cases}$$

This function is plotted in Figure 2 for some chosen parameter values. It reveals its increasing nature.

2.3. Moments. Moments play an important role in any statistical analysis. They allow for the measurement of crucial features of a distribution, such as the dispersion. Next, we provide the raw moments of the R-G (α, p) distribution. Let n be a non-zero positive integer. The n^{th} raw moment of Z can be expressed as follows:

$$\begin{aligned} \mathbb{E}(Z^n) &= \mathbb{E}\left((X+Y)^n\right) \\ &= \sum_{k=0}^{n} \binom{n}{k} \mathbb{E}\left(X^k\right) \mathbb{E}\left(Y^{n-k}\right) \\ &= \sum_{k=0}^{n} \binom{n}{k} \mathbb{E}\left(X^k\right) \left(\alpha + (-1)^{n-k}(1-\alpha)\right). \end{aligned}$$

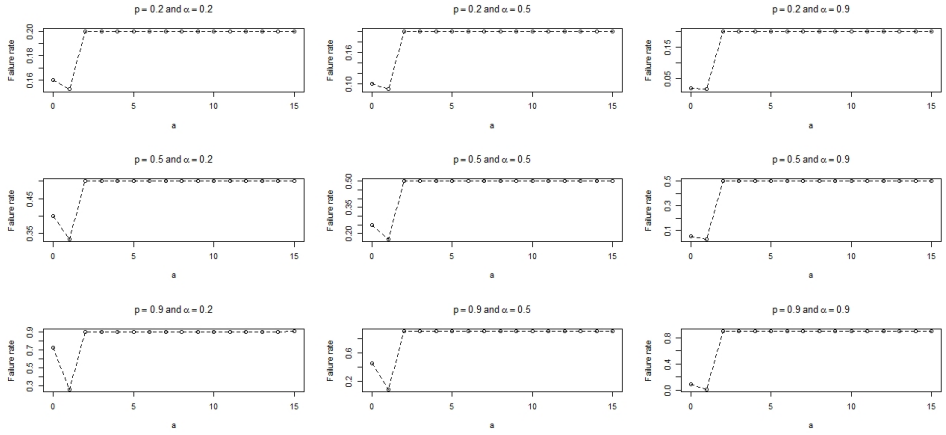

Figure 2: Plots of the failure rate function of the R-G (α, p) distribution for different parameter value combinations.

In particular, we have

$$\mathbb{E}(Z) = (2\alpha - 1) + \mathbb{E}(X) = (2\alpha - 1) + \frac{1}{p} \tag{2}$$

and

$$\begin{aligned}\mathbb{E}\left(Z^2\right) &= 1 + 2\mathbb{E}(X)(2\alpha - 1) + \mathbb{E}\left(X^2\right) \\ &= 1 + 2\frac{1}{p}(2\alpha - 1) + \frac{2-p}{p^2}.\end{aligned}$$

Thus, we deduce that

$$\begin{aligned}\mathbb{V}(Z) &= \mathbb{E}\left(Z^2\right) - (\mathbb{E}(Z))^2 \\ &= \frac{1-p}{p^2} + 4\alpha(1-\alpha).\end{aligned}$$

It also corresponds to $\mathbb{V}(Z) = \mathbb{V}(X) + \mathbb{V}(Y)$.

Remark 6. *The index of dispersion (ID) of Z is given by*

$$ID = \frac{\mathbb{V}(Z)}{\mathbb{E}(Z)} = \frac{\frac{1-p}{p^2} + 4\alpha(1-\alpha)}{(2\alpha - 1) + \frac{1}{p}}.$$

Suppose that $\alpha = \frac{1}{2}$. Thus, we obtain that $ID = p + \frac{1-p}{p} > 1$ (we can deduce that, in this case, the distribution is overdispersed). Table 1 contains the ID associated with the R-G(α, p) distribution for different value combinations of α and p.

Table 1: ID associated with the R-G (α, p) distribution for different parameter value combinations.

$\alpha \downarrow p \rightarrow$	0.1	0.2	0.3	0.4	0.5
0.1	9.821739	4.847619	3.212281	2.417647	1.9666667
0.2	9.642553	4.690909	3.079675	2.310526	1.8857143
0.3	9.462500	4.530435	2.937879	2.185714	1.7750000
0.4	9.281633	4.366667	2.788652	2.047826	1.6444444
0.6	8.917647	4.030769	2.472956	1.744444	1.3454545
0.7	8.734615	3.859259	2.308333	1.582759	1.1833333
0.8	8.550943	3.685714	2.140113	1.416129	1.0153846
0.9	8.366667	3.510345	1.968817	1.245455	0.8428571

$\alpha \downarrow p \rightarrow$	0.6	0.7	0.8	0.9
0.1	1.6974359	1.5467532	1.4944444	1.5539683
0.2	1.6416667	1.5113300	1.4653846	1.4937198
0.3	1.5403509	1.4119048	1.3558824	1.3548611
0.4	1.4121212	1.2797342	1.2119048	1.1891599
0.6	1.1095238	0.9654135	0.8775862	0.8263653
0.7	0.9440860	0.7941964	0.6984848	0.6375817
0.8	0.7725490	0.6173038	0.5148649	0.4461760
0.9	0.5963964	0.4362637	0.3280488	0.2529716

Hence, we can see that, for some sets of parameters, the ID value is strictly less than one (which implies that, for these sets of parameters, the distribution is underdispersed). Indeed, for a fixed p, when α approaches 1, the ID will decrease. In particular, for sets of parameters where $p > \frac{1}{2}$ and $\alpha > \frac{1}{2}$, we observe that $ID < 1$.

2.4. Incomplete moments. The incomplete moments find numerous applications in lifetime models. They allow for the definition of important quantities, such as the mean residual lifetime. Next, we provide expressions for the incomplete moments of the R-G (α, p) distribution.

Proposition 3. *Let $m \geq 2$ and n be a non-zero positive integer. The n^{th} incomplete moment of Z satisfies*

$$\mathbb{E}\left(Z^n \mathbb{1}_{\{Z \geq m\}}\right) = \frac{\left(\alpha + (1-\alpha)(1-p)^2\right)}{1-p} \mathbb{E}\left(X^n \mathbb{1}_{\{X \geq \lceil m \rceil\}}\right),$$

where $\lceil \cdot \rceil$ is the ceiling function (which gives the least integer that is greater than or equal to the input real value).

Proof.

Based on the definition of $\mathbb{P}(Z=k)$, we have

$$\begin{aligned}\mathbb{E}\left(Z^n \mathbb{1}_{\{Z\geq m\}}\right) &= \sum_{k=\lceil m\rceil}^{+\infty} k^n \mathbb{P}(Z=k)\\ &= \frac{\left(\alpha+(1-\alpha)(1-p)^2\right)}{1-p}\sum_{k=\lceil m\rceil}^{+\infty} k^n p(1-p)^{k-1}\\ &= \frac{\left(\alpha+(1-\alpha)(1-p)^2\right)}{1-p}\mathbb{E}\left(X^n \mathbb{1}_{\{X\geq\lceil m\rceil\}}\right).\end{aligned}$$

This ends the proof. □

2.5. Probability generating function. Several interesting characteristics of a distribution can be studied through its pgf. Next, we determine this function in the context of the R-G (α, p) distribution.

Proposition 4. *Let $G_Z(s)=\mathbb{E}\left(s^Z\right)$ denote the pgf associated with Z. Thus, we have*

$$G_Z(s)=p(1-\alpha)(1+s(1-p))+\left(\alpha+(1-\alpha)(1-p)^2\right)\left(\frac{ps^2}{1-(1-p)s}\right). \tag{3}$$

Proof. By using the geometric series formula, after some developments, we obtain

$$\begin{aligned}G_Z(s)=\mathbb{E}\left(s^Z\right) &= \sum_{k=0}^{+\infty} s^k \mathbb{P}(Z=k)\\ &= \mathbb{P}(Z=0)+s\mathbb{P}(Z=1)+\left(\alpha+(1-\alpha)(1-p)^2\right)s^2 p\sum_{k=2}^{+\infty}(s(1-p))^{k-2}\\ &= p(1-\alpha)(1+s(1-p))+\left(\alpha+(1-\alpha)(1-p)^2\right)\left(\frac{ps^2}{1-(1-p)s}\right).\end{aligned}$$

This finishes the proof. □

Remark 7. *One can derive the characteristic function associated with Z by simply replacing s with e^{it} in the pgf from Equation* (3)*. It thus takes the following form:*

$$\begin{aligned}\varphi_Z(t) &= \mathbb{E}\left(e^{itZ}\right)\\ &= p(1-\alpha)\left(1+e^{it}(1-p)\right)+\left(\alpha+(1-\alpha)(1-p)^2\right)\left(\frac{pe^{2it}}{1-(1-p)e^{it}}\right).\end{aligned}$$

2.6. Distribution of a sum. The sum of independent and identically distributed (iid) random variables that follow the R-G (α, p) distribution has a clear stochastic structure, as presented in the proposition below.

Proposition 5. *For any positive integer n, let $Z_1, \ldots, Z_n$ be iid random variables that follow the R-G(α, p) distribution. Then the sum random variable has the following stochastic representation:*

$$\sum_{i=1}^{n} Z_i = V + 2W - n,$$

where V is a random variable that follows the negative binomial NB(n, p) distribution and W is a random variable that follows the binomial B(n, α) distribution.

Proof. By the definition of the R-G(α, p) distribution, for any $i = 1, \ldots, n$, we can write $Z_i = X_i + 2U_i - 1$, where X_i is a random variable that follows the G(p) distribution and U_i is a random variable that follows the Bernoulli distribution with parameter α, all are independent with respect to the index i. We conclude with the well-known results that $V = \sum_{i=1}^{n} X_i$ follows the negative binomial NB(n, p) distribution and $W = \sum_{i=1}^{n} U_i$ follows the binomial B(n, α) distribution. This completes the proof. □

3. PARAMETER ESTIMATION

In this section, we consider the estimation of the parameters by the method of moments and the method of maximum likelihood. Also, simulation results on the behavior of the estimates are presented.

3.1. Method of moments. Let $Z_1, \ldots, Z_n$ denote a sample of iid random variables drawn from the R-G(α, p) distribution, and $z_1, \ldots, z_n$ some corresponding observations. Since

$$\frac{\mathbb{P}(Z = 1)}{\mathbb{P}(Z = 0)} = 1 - p,$$

a method of moment estimate of p can be defined as follows:

$$\bar{p}_n = 1 - \frac{\sum_{i=1}^{n} \mathbb{1}_{\{z_i = 1\}}}{\sum_{i=1}^{n} \mathbb{1}_{\{z_i = 0\}}}. \tag{4}$$

On the other hand, the method of moment estimate for α can be obtained using the first-order moment of Z given in Equation (2). This yields the following expression:

$$\bar{\alpha}_n = \frac{\bar{z}_n - \dfrac{\sum_{i=1}^{n} \mathbb{1}_{\{z_i = 0\}}}{\sum_{i=1}^{n} \mathbb{1}_{\{z_i = 0\}} - \sum_{i=1}^{n} \mathbb{1}_{\{z_i = 1\}}} + 1}{2}, \tag{5}$$

where

$$\bar{z}_n = \frac{\sum_{i=1}^{n} z_i}{n}.$$

Remark 8. *For theoretical results concerning the consistency and the asymptotic normality of the method of moment estimators, we refer to, e.g., [**16**] and [**23**].*

3.2. Behavior of the method of moment estimates via simulation study. The assessment of the performance of the method of moment estimates is based on a simulation study containing the following steps:

- Step 1: Fix $\alpha = \alpha_0$ and $p = p_0$.
- Step 2: Generate ten thousand samples of size n from the R-G (α_0, p_0) distribution. The representation in Equation (1) is used to generate samples.
- Step 3: Compute Equations (4) and (5) for the ten thousand samples, say $(\bar{\alpha}_{n,i}, \bar{p}_{n,i})$ for $i = 1, \ldots, 10000$.
- Step 4: Calculate the biases and mean squared errors (MSEs) given by

$$\mathrm{Bias}_1(n) = \frac{1}{10000}\sum_{i=1}^{10000}(\bar{\alpha}_{n,i} - \alpha_0), \quad \mathrm{Bias}_2(n) = \frac{1}{10000}\sum_{i=1}^{10000}(\bar{p}_{n,i} - p_0),$$

$$\mathrm{MSE}_1(n) = \frac{1}{10000}\sum_{i=1}^{10000}(\bar{\alpha}_{n,i} - \alpha_0)^2, \quad \mathrm{MSE}_2(n) = \frac{1}{10000}\sum_{i=1}^{10000}(\bar{p}_{n,i} - p_0)^2.$$

We repeated these steps for $n = 25, 50, 75, 100, 125, \ldots, 975, 1000$ and with $\alpha_0 = 0.2$ and $p_0 = 0.6$. Figure 3 shows how the biases vary with respect to n. Figure 4 shows how the MSEs vary with respect to n. Thus, one can see that the biases for each estimate increase to zero as $n \to +\infty$ and the MSEs for each estimate decrease to zero as $n \to +\infty$. Furthermore, the fit to a normal distribution is illustrated in Figure 5 for the method estimation (with $n = 10000$ and the parameters $\alpha_0 = 0.2$ and $p_0 = 0.6$) proposed in Subsection 3.1. Thus, this figure shows numerically the normal asymptotic distribution of the proposed estimates.

3.3. Method of maximum likelihood. Let $Z_1, \ldots, Z_n$ denote a sample of iid random variables drawn from the R-G (α, p) distribution, and $z_1, \ldots, z_n$ some corresponding observations. Let $Z_{obs} = \{z_i\}_{i=1}^n$ denote the set of observed data. Moreover, we set

- $I_0 = \{i \mid z_i = 0, 1 \leq i \leq n\}$ and $m_0 = \sum_{i=1}^n \mathbb{1}_{\{z_i=0\}}$ (= number of elements in I_0).
- $I_1 = \{i \mid z_i = 1, 1 \leq i \leq n\}$ and $m_1 = \sum_{i=1}^n \mathbb{1}_{\{z_i=1\}}$ (= number of elements in I_1).
- $I_{+2} = \{i \mid z_i \geq 2, 1 \leq i \leq n\}$. Note that the number of elements in I_{+2} equals $n - (m_0 + m_1)$.

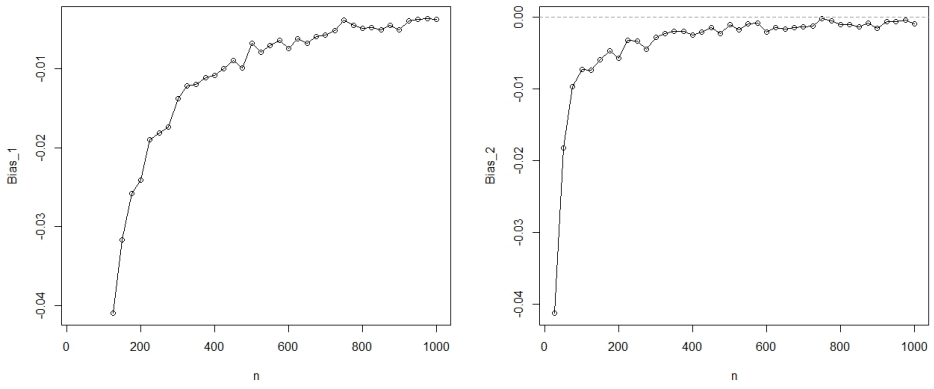

Figure 3: From left to right: Biases of Equations (4) and (5) for $n = 25, 50, 75, \ldots, 975, 1000$, when $\alpha_0 = 0.2$ and $p_0 = 0.6$.

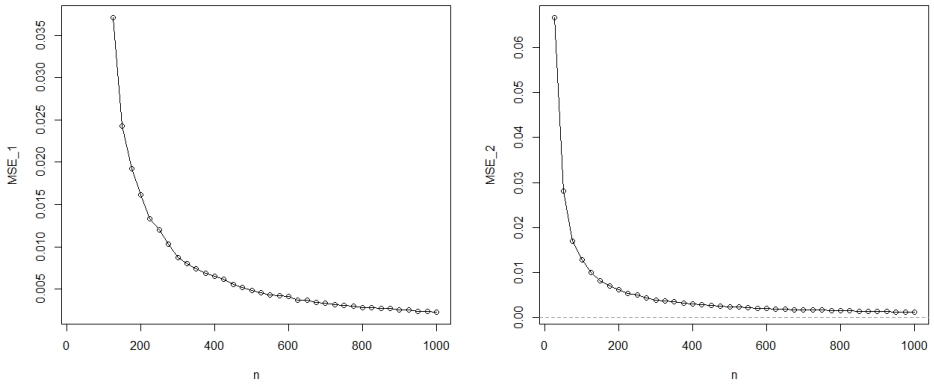

Figure 4: From left to right: MSEs of Equations (4) and (5) for $n = 25, 50, 75, \ldots, 975, 1000$, when $\alpha_0 = 0.2$ and $p_0 = 0.6$.

The observed-data likelihood function for $\theta = (\alpha, p)$ is then given by

$$\begin{aligned} L(\theta \mid Z_{obs}) &= (p(1-\alpha))^{m_0} \times (p(1-\alpha)(1-p))^{m_1} \\ &\quad \times \left(p\left(\alpha + (1-\alpha)(1-p)^2\right)\right)^{n-(m_0+m_1)} \times \prod_{i \in I_{+2}} (1-p)^{z_i - 2} \\ &= p^n (1-\alpha)^{m_0+m_1} \left(\alpha + (1-\alpha)(1-p)^2\right)^{n-(m_0+m_1)} \\ &\quad \times (1-p)^{\left(\sum_{i \in I_{+2}} z_i\right) - 2n + 2m_0 + 3m_1}. \end{aligned}$$

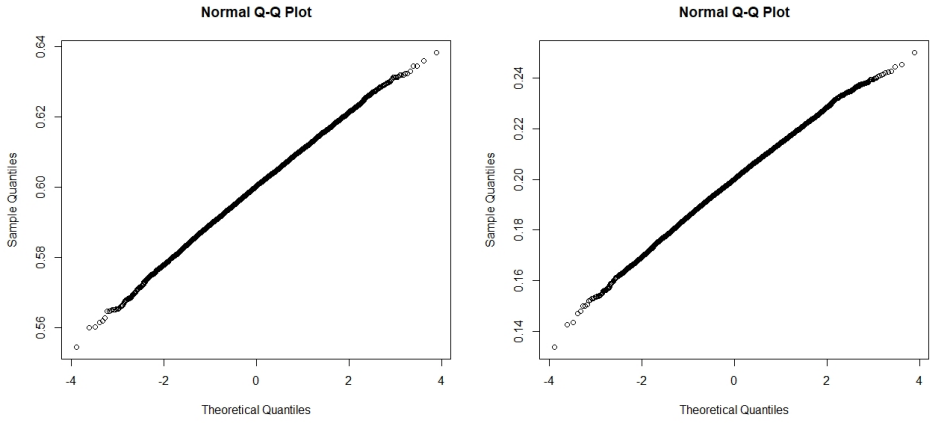

Figure 5: From left to right: Normal quantile-quantile (Q-Q) plots for the errors (from the method of the moment estimation, see Subsection 3.1), $(\bar{\alpha}_n - 0.2)$ and $(\bar{p}_n - 0.6)$, when the length series is $n = 1000$.

Thus, the observed-data log-likelihood function is obtained as follows:

$$\ell = \log\left(L\left(\theta \mid Z_{obs}\right)\right) = n\log\left(p\right) + A\log\left(1-p\right) + \left(m_0 + m_1\right)\log\left(1-\alpha\right) + \left(n - \left(m_0 + m_1\right)\right)\log\left(\alpha + \left(1-\alpha\right)\left(1-p\right)^2\right),$$

with

$$A = \left(\sum_{i \in I_{+2}} z_i\right) - 2n + 2m_0 + 3m_1.$$

The maximum log-likelihood estimation of θ is defined as follows:

$$\hat{\theta}_n = \left(\hat{\alpha}_n, \hat{p}_n\right) = \underset{\theta \in (0,1)^2}{\operatorname{argmax}} \log\left(L\left(\theta \mid Z_{obs}\right)\right). \tag{6}$$

In other words, to obtain the maximum likelihood estimates, we maximize the log-likelihood function ℓ defined above with respect to the parameters α and p over the appropriate parameter space. Differentiating ℓ partially with respect to the parameters and setting them equal to zero yields the following system of likelihood equations or score equations:.

$$\frac{\partial \ell}{\partial p} = \frac{n}{p} - \frac{A}{1-p} - 2\left(n - \left(m_0 + m_1\right)\right)\frac{\left(1-p\right)\left(1-\alpha\right)}{\alpha + \left(1-\alpha\right)\left(1-p\right)^2} = 0$$

and

$$\frac{\partial \ell}{\partial \alpha} = \left(n - \left(m_0 + m_1\right)\right)\frac{1 - \left(1-p\right)^2}{\alpha + \left(1-\alpha\right)\left(1-p\right)^2} - \frac{m_0 + m_1}{1-\alpha} = 0.$$

Since equating the first-order log-likelihood derivatives to zero leads to a complicated system of equations, the maximum likelihood estimates are achieved using numerical methods. Explicitly, we use the "nlm" function (Non-Linear Minimization, from package "stats" of the software R) to find values that maximize the log-likelihood function defined above.

Remark 9. *Theoretical results concerning the asymptotic behavior of the maximum likelihood estimator have been widely discussed in the literature. For the consistency, see, e.g., [***37***], [***21***], and [***17***], and for the asymptotic normality, see the results of, e.g., [***22***], [***29***], and [***18***].*

3.4. Behavior of the maximum likelihood estimates via simulation study. In this section, we aim to test the efficiency of the parameter's estimation discussed in Subsection 3.3. Thus, using the R programming language, we simulate 1000 paths of length $n = 25$, 100, 250, 500 and 1000. These paths are simulated using the representation in Equation (1) with three sets of parameters: $(a)\,(\alpha_0, p_0) = (0.7, 0.4)$; $(b)\,(\alpha_0, p_0) = (0.5, 0.5)$; $(c)\,(\alpha_0, p_0) = (0.1, 0.6)$. Note that for each path, the maximum likelihood estimation, denoted by $(\hat{\alpha}_{n,i}, \hat{p}_{n,i})$ is calculated via the "nlm" function from package "stats" of software R. The mean values of maximum likelihood estimates for each set of parameters are given in Table 2. The standard deviations of the estimates are stated in brackets under the estimated values. Thus, one can see that the precision of these estimates increases when the size n increases. Explicitly, one can deduce that standard deviations decrease to zero when n increases.

Finally, the fit to the normal distribution is illustrated in Figure 6 for the maximum likelihood estimates (with the parameter set (c) and $n = 1000$). These figures show numerically the normal asymptotic distribution of the proposed estimates.

4. APPLICATION

In this section, the Rademacher perturbed geometric distribution is examined for a dataset arising from the business area. Indeed, "Secret d'Eve" is a small business (based in Lebanon) specializing in cosmetic products. The business model of "Secret d'Eve" is based on direct sales via social networks (in particular, Facebook and Instagram). The dataset used for this section gives the number of "coconut oil" (100 ml bottles), which is one of the products offered by "Secret d'Eve") sold per week in 2021:

$$0, 0, 2, 0, 2, 0, 3, 2, 3, 3, 5, 5, 3, 1, 0, 4, 5, 0, 5, 2, 4, 0, 2, 0, 2, 2,$$
$$2, 4, 3, 0, 1, 6, 2, 2, 0, 0, 4, 0, 2, 2, 0, 4, 1, 2, 2, 4, 3, 2, 2, 3, 0, 2.$$

Note that, during 2021, "Secret d'Eve" proposed a special offer with a 50% discount on the second bottle of "coconut oil" purchased. Thus, in total, we have 52 observations. Basic descriptive statistics concerning the observed data are presented in Table 3.

Table 2: Estimated parameters and the corresponding standard errors (in brackets) stated under the maximum likelihood method

a) $\alpha_0 = 0.7$ and $p_0 = 0.4$		
$n =$	$\hat{\alpha}$	$\hat{p}$
50	0.7009651	0.4047236
	(0.08765927)	(0.04716149)
100	0.6989499	0.4022161
	(0.06023791)	(0.03404406)
250	0.7008684	0.4025103
	(0.03970101)	(0.02143184)
500	0.6999166	0.4002412
	(0.02890456)	(0.01464088)
1000	0.6997472	0.3996503
	(0.02035622)	(0.0105077)
b) $\alpha_0 = 0.5$ and $p_0 = 0.5$		
$n =$	$\hat{\alpha}$	$\hat{p}$
50	0.4981015	0.5072683
	(0.103134)	(0.06299839)
100	0.4985988	0.5031262
	(0.0702593)	(0.04251333)
250	0.4986373	0.5017282
	(0.04425671)	(0.0278258)
500	0.4976185	0.5011049
	(0.03220724)	(0.01966106)
1000	0.499701	0.500419
	(0.02323199)	(0.01387747)
c) $\alpha_0 = 0.1$ and $p_0 = 0.6$		
$n =$	$\hat{\alpha}$	$\hat{p}$
50	0.092036	0.6072886
	(0.1073326)	(0.08610007)
100	0.09626411	0.6033917
	(0.0774742)	(0.06165065)
250	0.09907296	0.6021155
	(0.04552499)	(0.03828349)
500	0.0991955	0.6010666
	(0.03283201)	(0.02641448)
1000	0.09869812	0.6002303
	(0.0224928)	(0.01909936)

In Figure 7, we plot the observed data. Thus, we can see that 2 is the mode of the observed data, 0 is also well represented, and there is a decrease from the value of 2.

The autocorrelation function (ACF) and the partial ACF (PACF) of the observed data are plotted in Figure 8. Clearly, this graph shows that the observed data can

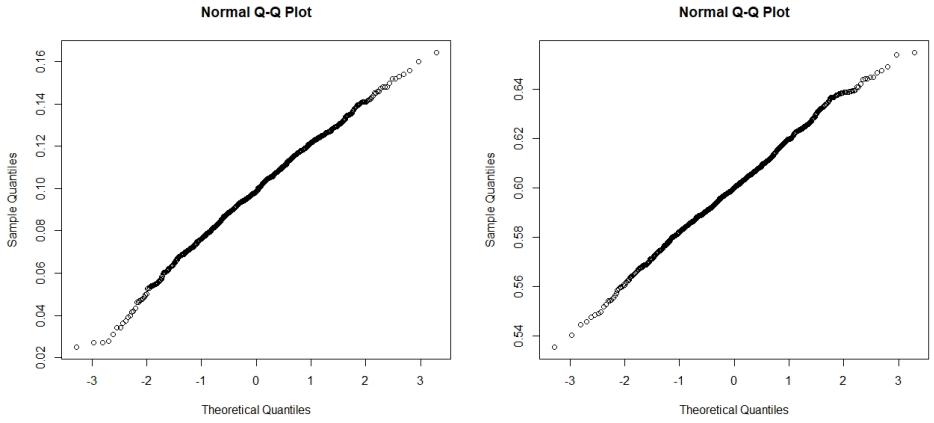

Figure 6: From left to right: Normal Q-Q plots for the errors (from the maximum likelihood estimation, see Subsection 3.3), $(\hat{\alpha}_n - 0.1)$ and $(\hat{p}_n - 0.6)$, when the length series is $n = 1000$.

Table 3: Basic descriptive statistics for the observed data associated with the number of "coconut oil" (100 ml bottles) sold per week in 2021.

Length	Mean	Min	Max	First quart.	Median	Third quart.	St. deviation
52	2.077	0	6	0	2	3	1.666817

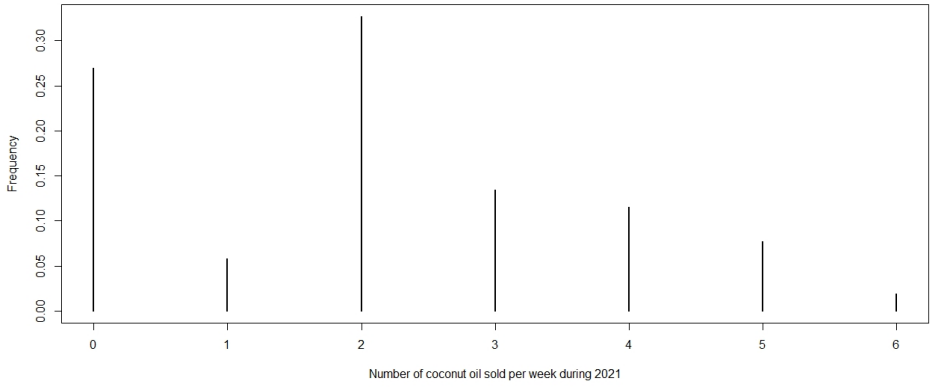

Figure 7: Plot of the number of "coconut oil" (100 ml bottles) sold per week in 2021.

be considered observations of white noise.

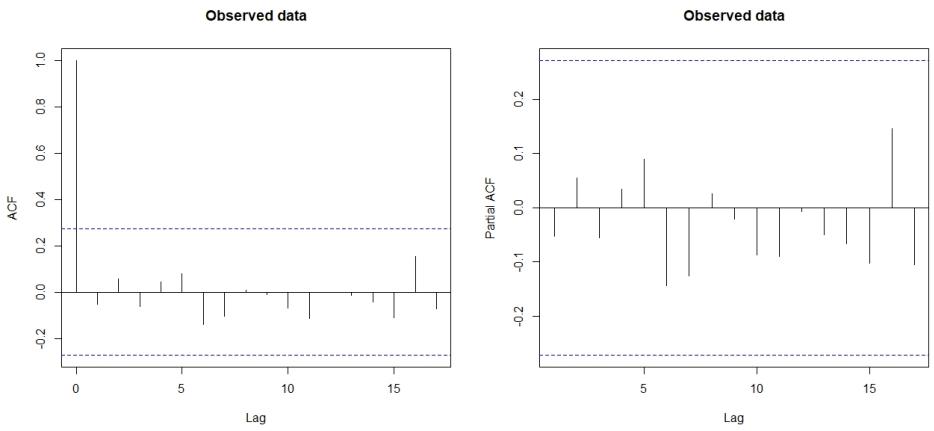

Figure 8: From left to right: ACF and PACF of the observed number of "coconut oil" (100 ml bottles) sold per week in 2021.

To fit these data, we propose four distributions: the Poisson (λ) distribution, the Geometric (a) distribution, the ZTIG (π_1, π_2, p) distribution (i.e., the zero-two inflated geometric distribution, which is a special case of the generalized inflated geometric (for details, see [**20**] and [**25**]), where π_1 and π_2 represent the weights associated with inflation of 0 and 2), and the R-G (α, p) distribution introduced in this paper. The maximum likelihood estimation of the involved parameters is presented in Table 4.

Table 4: Parameters estimation of the proposed distributions to fit the data

Distribution $\rightarrow$	R-G (α, p)	Poisson (λ)	Geometric (a)	ZTIG (π_1, π_2, p)
Estimations	$\hat{\alpha} = 0.574512$ $\hat{p} = 0.5186991$	$\hat{\lambda} = 2.076923$	$\hat{a} = 0.440642$	$\hat{\pi}_1 = 0.1175094$ $\hat{\pi}_2 = 0.2382797$ $\hat{p} = 0.2353197$

The p-values of the Chi-square test for the Poisson and geometric distributions (see Table 5) are very close to 0, suggesting that these distributions do not provide good models to fit the data.

Table 5: Goodness-of-fit statistics of the data.

Distribution	χ^2	DF	p-value
R-G	2.7583	2	0.25179
ZTIG	2.8111	1	0.09361
Poisson	21.9774	3	6.594125e-05
Geometric	38.6447	2	4.058916e-09

In Table 6, the classical Akaike information criterion (AIC), Bayesian information criterion (BIC), corrected AIC (AIC_C), Hannan-Quinn information criterion (HQIC), and "consistent" AIC (CAIC) are calculated for the ZTIG and R-G distributions in order to compare the performance of both distributions in the fitting of the considered data. Thus, one can see that the proposed R-G distribution yields minimum values of model fit statistics with respect to the mentioned information criteria.

Table 6: Comparison between model selection criteria.

Criterion	R-G	ZTIG
AIC	184.6112	185.2107
BIC	188.5137	191.0645
AIC_C	184.8561	185.7107
HQIC	186.1073	193.4549
CAIC	184.8561	191.7107

5. CONCLUSION

This paper offers a new two-parameter discrete distribution as a possible alternative to the standard geometric distribution. This new distribution can be seen as a perturbed version of the geometric distribution by adding "random noise" following the Rademacher distribution. We have studied several of its properties and explored the parameter estimation issue. The new distribution has proven to be very useful for modeling count data that presents inflated or deflated zero and/or inflated two and/or over or under dispersion and short- and long-tailed count data.

It is also important to mention that the proposed perturbation technique of the geometric distribution can be used in a more generalized way by considering a different distribution associated with Y (the additive noise). Indeed, to find $\mathbb{N}$ as the support of the perturbed distribution, it is enough that the support of Y has -1 as the minimum modality (the other modalities are integer values). For example, one can consider Y as a random walk variable, i.e., $Y(\Omega) = \{-1, 0, 1\}$ (where the modality -1 represents a backstep, 0 stays put, and 1 is a step forward).

REFERENCES

1. R. Agrawal, R. Srikant: *Privacy-preserving data mining.* In Proceedings of the 2000 ACM SIGMOD international conference on Management of data (2000), 439-450.

2. M. M. A. Almazah, T. Erbayram, Y. Akdoğan, M. M. Al Sobhi, A. Z. Afify: *A new extended geometric distribution: Properties, regression model, and actuarial applications.* Mathematics, **9** (2021), 1336.

3. R.S. Alshkaki: *On zero-one inflated geometric distribution.* Internat. Res. J. Math. Eng. IT., **3** (2016), 10-21.

4. A. Azzalini: *A class of distributions which includes the normal ones.* Scand. J. Stat. Theory Appl., **12** (1985), 171-178.

5. M. F. Baksh, D. Böhning, R. Lerdsuwansri: *An extension of an over-dispersion test for count data.* Comput. Stat. Data Anal., **55** (2011), 466-474.

6. D. Bhati, S. Joshi: *Weighted geometric distribution with new characterizations of geometric distribution.* Commun. Stat. - Theory Methods, **47** (2018), 1510-1527.

7. S. Chakraborty, D. Bhati: *Transmuted geometric distribution with applications in modeling and regression analysis of count data.* SORT-Stat. Oper. Res. T., **40** (2016), 153-176.

8. S. Chakraborty, D. Chakravarty: *Discrete gamma distribution: properties and parameter estimation.* Commun. Stat. - Theory Methods, **41** (2012), 3301-3324.

9. S. Chakraborty, R. D. Gupta: *Exponentiated geometric distribution: Another generalization of geometric distribution.* Commun. Stat. - Theory Methods, **44** (2015), 1143-1157.

10. S. Chakraborty, S. H. Ong, A. Biswas: *An extension of the geometric distribution with properties and applications.* Austrian J. Stat., **52** (2023), 124-142.

11. K. Chen, L. Liu: *Geometric data perturbation for privacy preserving outsourced data mining.* Knowl. Inf. Syst., **29** (2011), 657-695.

12. M. Evans, N. Hastings, B. Peacock, C. Forbes: *Statistical distributions.* John Wiley & Sons, 2011.

13. A. Evfimievski, R. Srikant, R. Agrawal, J. Gehrke: *Privacy preserving mining of association rules.* In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, (2002), 217-228).

14. E. Gómez-Déniz: *Another generalization of the geometric distribution.* Test, **19** (2010), 399-415.

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

16. A. Hall: *Generalized method of moments.* Oxford, 2004.

17. P.J. Huber: *The behavior of maximum likelihood estimates under nonstandard conditions.* In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. 1, 1 (1967), 221-233.

18. J. Jiang: *REML estimation: asymptotic behavior and related topics.* Ann. Stat., **24** (1996), 255-286.

19. N. L. Johnson, A. W. Kemp, S. Kotz: *Univariate discrete distributions (Vol. 444).* John Wiley & Sons, 2005.

20. R. D. Joshi: *A generalized inflated geometric distribution (1015).* Master [Thesis]. Huntington: Marshall University.

21. G. Kulldorff: *On the conditions for consistency and asymptotic efficiency of maximum likelihood estimates.* Skand. Aktuartidskr., **40** (1957), 129-144

22. L. Le Cam: *On the asymptotic theory of estimation and testing hypotheses.* In Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics, (1956), (Vol. 3, pp. 129-157). University of California Press.

23. E. L. Lehmann, G. Casella: *Theory of point estimation.* Springer Science & Business Media, 2006.

24. J. Makcutek: *A generalization of the geometric distribution and its application in quantitative linguistics.* Rom. J. Phys., **60** (2008), 501-509.

25. A. Mallick, R. Joshi: *Parameter estimation and application of generalized inflated geometric distribution.* J. Stat. Theory Appl., **17** (2018), 491-519.

26. A. W. Marshall, I. Olkin: *A new method for adding a parameter to a family of distributions with application to the exponential and Weibull families.* Biometrika, 84(3) (1997), 641-652.

27. M. Melkersson, D. O. Rooth: *Modeling female fertility using inflated count data models.* J. Popul. Econ., **13** (2000), 189-203.

28. G. S. Mudholkar, D. K. Srivastava: *Exponentiated Weibull family for analyzing bathtub failure-rate data.* IEEE Trans. Reliab., **42** (1993), 299-302.

29. R. H. Norden: *A survey of maximum likelihood estimation.* Int. Stat. Rev., **40** (1972), 329-354.

30. S. H. Ong, S. Chakraborty, A. Biswas: *A new generalization of the geometric distribution using Azzalini's mechanism: properties and application.* (2020), arXiv preprint arXiv:2010.04507.

31. A. Pandey, H. Pandey, V. K. Shukla: *An Inflated Probability Model On Rural Out-Migration.* J. Math. Comput. Sci., **6** (2015), 702-711.

32. R. N. Rattihalli, S. R. Rattihalli: *A generalisation of geometric distribution.* Commun. Stat. - Theory Methods, **52** (2021), 1-13.

33. D. V. S. Sastry, D. Bhati, R. N. Rattihalli and E. Gómez-Déniz: *On zero-distorted generalized geometric distribution.* Commun. Stat. - Theory Methods, **45** (2016), 5427-5442.

34. W. T. Shaw, I. R. Buckley: *The alchemy of probability distributions: beyond Gram-Charlier expansions, and a skew-kurtotic-normal distribution from a rank transmutation map.* arXiv preprint, (2009), arXiv:0901.0434.

35. D. T. Shirke, S. R. Supanekar, D. Bhati: *On k-distorted generalized discrete family of distributions.* Commun. Stat. - Theory Methods, **46** (2017), 11591-11603.

36. P. Srisuradetchai, K. Dangsupa: *On interval estimation of the geometric parameter in a zero-inflated geometric distribution.* Thail. Stat., **21** (2023), 93-109.

37. A. Wald: *Note on the consistency of the maximum likelihood estimate.* Ann. Math. Stat., **20** (1949), 595-601.

Maher Kachour
ESSCA School of Management,
Lyon, France
E-mail: *kachour.maher@gmail.com*

Christophe Chesneau
Laboratoire de Mathématiques Nicolas Oresme (LMNO),
Université de Caen-Nomandie,
Caen, France
E-mail: *christophe.chesneau@gmail.com*

(Received 05. 07. 2023.)
(Revised 30. 06. 2024.)