Известия НАН Армении, Математика, том 59, н. 2, 2024, стр. 16 – 34.

A NEW REGULARLY VARYING DISCRETE DISTRIBUTION GENERATED BY WARING-TYPE PROBABILITY

D. FARBOD

Department of Mathematics, Faculty of Engineering Science,
Quchan University of Technology, Quchan, Iran
E-mail: *d.farbod@qiet.ac.ir*

Abstract. In this paper, based on the discretization method, we construct a new 2-parameter regularly varying discrete distribution generated by Waring-type probability (2-RDWP). Some useful plots are displayed for the model. From the mathematical point of view, to suggest 2-RDWP as a new discrete probability distribution in bioinformatics, some statistical facts such as unimodality, skewness to the right, upward/downward convexity, regular variation at infinity and asymptotically constant slowly varying component are established for the model. We provide the conditions of coincidence of solution for the system of likelihood equations with the maximum likelihood estimators for the unknown parameters. Simulation studies are performed using the Monte Carlo method and Nelder-Mead optimization algorithm to obtain maximum likelihood estimations of the unknown parameters. Asymptotic expansion of the probability function with two terms is considered, and then the moment's existence of integer orders is investigated. Finally, a real count data set is used to show the applicability of the new model compared to other models in bioinformatics.

MSC2020 numbers: 60E05, 62E10, 62F10, 62P10.

Keywords: asymptotic expansion; discretization; maximum likelihood; Monte Carlo method; statistical facts; Waring-type probability.

1. Introduction

Probability distributions are commonly applied to describe phenomena in biomolecular systems, bioinformatics, etc. Due to the usefulness of probability distributions in bioinformatics, their mathematical theory is widely studied, and new discrete distributions (frequency distributions) are developed. According to the variety, diversity and complexity of real data sets in bioinformatics and biomolecular systems, it is impossible to figure out and suggest a universal model suitable for all situations. Hence, the interest in developing discrete distributions in bioinformatics and biomolecular systems remain strong in probability and statistics.

Many discrete probability distributions have been introduced based on different methods for the needs of bioinformatics systems. For a review of different methods, see, for example, [3]. Let us point out two of the known producers as follows.

Using the method of birth-death process, we refer the readers to, for example, [2, 5, 13, 14, 15]. Besides the method of birth-death process, there are other methods, in particular by *discretization method* which we refer to, for example, [6, 7, 8, 10].

The advantage of constructing new probability distribution is proposed to parametric ones because by changing the parameters, one hopes to find the best approximation for the unknown model. Because of the wide variety of phenomena in bioinformatics, we shall attempt to introduce new parametric distribution (based on discretization method).

A continuous analog of the 2-parameter regularly varying Waring probability was given by dediscretization method [1, 2, 9]. Its probability density function is stated as

$$f_x(\alpha) = \frac{1}{c(\alpha)} \times \frac{(r+x-1)^{(r+x-1)}}{(q+x)^{(q+x)}}, \quad x \in (0,\infty) \tag{1.1}$$

where $\alpha = (r, q)$ is the unkown parameter such that $r > 0, \quad q > 0$. r is called numerator parameter and q denominator parameter and $q - r > 0$. Also, $c(\alpha)$ is the normalization factor and $c(\alpha) = \int_0^\infty \frac{(r+t-1)^{(r+t-1)}}{(q+t)^{(q+t)}} dy$.

We note that the continuous analog of the 2-parameter regularly varying Waring probability (1.1) is a continuous probability distribution. Here, let us call the model (1.1) as *Waring-type probability*.

The novelty and the motivation to write this paper is to construct a new skewed discrete probability model (frequency distribution) for the needs of biosystems using (1.1). We use discretization method, and then study mathematical properties, statistical inferences and applications.

2. The 2-RDWP distribution

The desired discrete probability distribution is possible to obtain using the discretization method. We use a type of discretization of densities used by, for example, Farbod [6, 8] and Farbod and Gasparian [10]. Let us consider the numerator of (1.1) as follows:

$$p_x(\alpha) = \frac{(r+x-1)^{r+x-1}}{(q+x)^{q+x}}, \qquad x > 0. \tag{2.1}$$

To have $p_x(\alpha)$ (2.1) as a probability mass function (pmf), we use discretization method [6, 7, 8, 10] to get a new *discrete probability distribution*, denoted by $g_x(\alpha)$, with the following pmf:

$$g_x(\alpha) = \big(d(\alpha)\big)^{-1} \times \frac{(r+x-1)^{r+x-1}}{(q+x)^{q+x}}, \tag{2.2}$$

where $x = 1, 2, ...,$ and $d(\alpha)$ is the *normalization factor (normalization constant)* given by

$$d(\alpha) = \sum_{y=1}^{\infty} \frac{(r+y-1)^{r+y-1}}{(q+y)^{q+y}} \tag{2.3}$$

and $\alpha = (r, q)$ is the unknown parameter such that $r > 0$ and $q > r$.

Remark 2.1. *It is obvious that* $g_x(\alpha) \geq 0$ *and also* $\sum_{x=1}^{\infty} g_x(\alpha) = 1$. *Thus, function (2.2) is a probability function and can be considered as a new pmf on the set of positive integers* $x \in \mathbb{N}_+ = \{1, 2, 3, ...\}$.

A probability measure (distribution function of random variable X) is given by

$$F_x(\alpha) = P(X \leq x) = \big(d(\alpha)\big)^{-1} \sum_{m=1}^{x} g_m(\alpha) = \big(d(\alpha)\big)^{-1} \sum_{m=1}^{x} \frac{(r+m-1)^{r+m-1}}{(q+m)^{q+m}}. \tag{2.4}$$

We call model (2.4) a "*2-parameter regularly varying discrete distribution generated by Waring-type probability*"(in short, 2-RDWP). The pmf of 2-RDWP is given by Eq.(2.2). This paper investigates some mathematical properties, statistical inferences and applications for the model (2.2).

The remaining sections of the paper can be summarized as follows. Section 3 presents some plots of pmf and log-log plot of the 2-RDWP model for different values of parameters. *Statistical facts*, for our model, are verified for the mathematical needs of bioinformatics in Section 4. In Section 5, we propose maximum likelihood (ML) estimators of the 2-RDWP's parameters, which are coincided with some moment estimators. Section 6 uses the Monte Carlo method and Nelder-Mead optimization algorithm to simulate for obtaining the ML estimations of parameters. Section 7 gives an asymptotic expansion with two terms for the pmf, tail behavior of distribution function, and also the moment's existence of integer orders is investigated. Section 8 presents application of the proposed model and compares it with other rival models. The study is concluded in Section 9. Section 10 considers an Appendix containing the pmfs of some rival models arising in bioinformatics.

3. Figures

This section presents two types of figures for the 2-RDWP model (2.2). To depict figures, we need to consider the model's pmf as truncated. First, some pmfs for different possible values of parameters r and q are plotted in Figures 1(A-J). Second, some log-log plots ($\ln g_x(\alpha)$ versus $\ln x$) are displayed in Figures 2(A-J). Figures 1(A-J) show skewness to the right and also unimodality of the pmf and Figures 2(A-J) show the deviations of $\ln g_x(\alpha)$ versus $\ln x$ from the straight line,

which is discussed in Section 4.

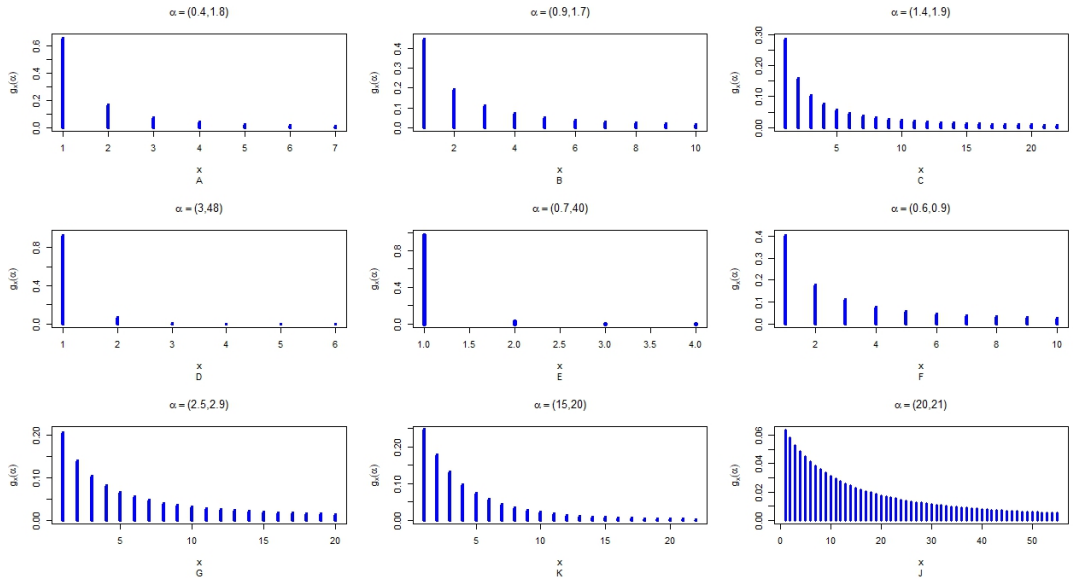

Рис. 1. Illustrations of the pmf of 2-RDWP model (2.2) for possible values of two parameters r and q.

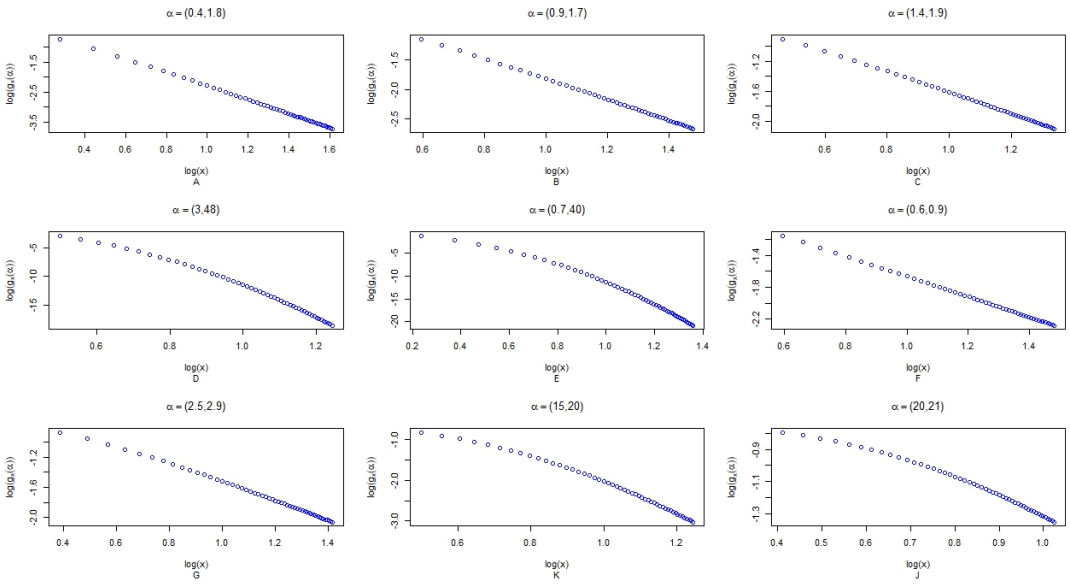

Рис. 2. Illustrations of the log-log plot of 2-RDWP model (2.2) for possible values of two parameters r and q.

4. Statistical facts

From the mathematical point of view, to suggest a discrete probability distribution as a new model in bioinformatics, we need to verify some common *statistical facts (empirical facts)* such as unimodality, skewness to the right, upward/downward convexity, regularly varying at infinity, and slowly varying at infinity. In other words, it was established that if a pmf (probability law) holds these *statistical facts*, then the corresponding pmf could be a mathematical framework for bioinformatics applications [2, 3, 5, 15]. So, to apply the 2-RDWP model (2.2) as a new probability model in bioinformatics, we need to check out the validity of such known *statistical facts*, mathematically, numerically and intuitively.

We notice that *statistical facts* (*empirical facts*) are common *mathematical properties* of the empirical frequency distributions (with complex forms and long right-side tails) observed in bioinformatics data sets and are systematically reproducible in biomolecular systems [2, 3, 5, 15].

4.1. **Log-log plot.** Biologists prefer to deal with log-log plot of distribution instead of its shape [2]. One of the statistical facts is that log-log plot of discrete distributions arising in bioinformatics systematically deviated from the straight line and shows

upward/downward convexity [2, 3, 15]. It means that the deviations of log-log plot of $g_x(\alpha)$ from the straight line must be *not too large*.

Let us investigate the log-log plot of our model. Namely, we deal with $\ln g_x(\alpha)$ versus $\ln x$. We write the log-log plot of the model ($\ln g_x(\alpha)$ versus $\ln x$) as follows:

$$\frac{\ln g_x(\alpha)}{\ln x} = \frac{(x+r-1)\ln(x+r-1) - (x+q)\ln(x+q) - \ln(d(\alpha))}{\ln x} \tag{4.1}$$

It is obvious that, for sufficiently large x (sufficiently large x means $x \longrightarrow \infty$), $\ln(x+a) \approx \ln x$ (a is some real constant). Therefore from Eq.(4.1), for sufficiently large x, we have

$$\frac{\ln g_x(\alpha)}{\ln x} \approx r - q - 1 - \frac{\ln d(\alpha)}{\ln x}. \tag{4.2}$$

Due to Eq.(4.2) we conclude that the deviations of $\ln g_x(\alpha)$ versus $\ln x$ from the straight line *constant* $= (r-q-1)$ are small, at least for large values of x which it turns out upward/downward convexity.

Remark 4.1. *We note that there are not any specific definitions for upward/downward convexity concept in bioinformatics and it is issue of the mathematical disciplines. In other words, some of the peculiarities of the shapes of empirical frequency distributions in bioinformatics are: upward/downward convexity, the only point where the frequency distribution achieves it's maximal value, etc. For more details about mathematical and applied concepts of upward/downward convexity, we refer the readers to Astola and Danielian* [2, Sec.1.4, Sec.2.5].

Additionally, Figures 2(A-J) show the log-log plot of 2-RDWP with different values of parameters. Figures 2(A-J) provide that the deviations of $\ln g_x(\alpha)$ versus $\ln x$ from the straight line may be small, at least for some large values of x. From Figures 2(A-J), we observed a significant shift and variation of the power law-like right-side tail of the pmfs.

4.2. **Regular variation.** This subsection shows that the model $g_x(\alpha)$ varies regularly at infinity and also we present an asymptotically constant slowly varying component for it. Compared to Astola and Danielian [2], let us state two definitions for our model (2.2).

Definition 4.1. *The frequency distribution $g_x(\alpha)$ varies regularly at infinity with exponent $(-\rho)$ if it may be presented in the form*

$$g_x(\alpha) = x^{-\rho} \cdot R(x)(1+o(1)), \quad x \longrightarrow \infty, \quad \rho \in (-\infty, \infty),$$

where $R(x) > 0$ for $x = 1, 2, ...,$ and for $\kappa = 2, 3, ...,$ $\lim_{x\longrightarrow\infty} \frac{R(\kappa x)}{R(x)} = \kappa^{-\rho}$.

Definition 4.2. *Let, for $\kappa = 2, 3, ...$, the limit exists*

$$\lim_{x \longrightarrow \infty} \frac{R(\kappa x)}{R(x)} = 1$$

then $g_x(\alpha)$ exhibits the asymptotically constant slowly varying component L if we have

$$\lim_{x \longrightarrow \infty} R(x) = L \in (0, \infty).$$

Remark 4.2. *It is clear that Definition 4.2 is a particular case of Definition 4.1. Thus, a function varying regularly at infinity with exponent $\rho = 0$ varies slowly at infinity* [2].

Let us establish the function $g_x(\alpha)$ varies regularly at infinity with exponent $(-\rho)$ having $-\rho = -(q + 1 - r)$. We propose theorem, remark and numerical example as follows.

Theorem 4.1. *The model $g_x(\alpha)$ (2.2) varies regularly at infinity with exponent $(-\rho)$ and*

$$-\rho = -(q + 1 - r) < -1. \tag{4.3}$$

Proof. From (2.2) and (2.3), for sufficiently large x, we get

$$g_x(\alpha) \approx (d(\alpha))^{-1} \cdot e^{-(q+1-r)} x^{-(q+1-r)} \approx x^{-(q+1-r)}. \tag{4.4}$$

It follows from (4.4) that $g_x(\alpha)$ (2.2) varies regularly at infinity if $\rho = q + 1 - r > 1$. Theorem 4.1 is proved. □

Remark 4.3. *From (4.4) and based on Remark 3, we observe that $L = (d(\alpha))^{-1}$ is an asymptotically constant slowly varying component for the model $g_x(\alpha)$ (2.2). In other words, $g_x(\alpha)$ exhibits the asymptotically constant slowly varying component given by $L = (d(\alpha))^{-1}$.*

Let us give a numerical example as follows.

Example 4.1. *Let us compute the value of ρ corresponding to selected two parameters r and q (used in Figures 1 and 2) by:*

$$\begin{array}{ll}
\alpha = (0.4, 1.8), & -\rho = -2.4 < -1 \\
\alpha = (0.9, 1.7), & -\rho = -1.8 < -1 \\
\alpha = (1.4, 1.9), & -\rho = -1.5 < -1 \\
\alpha = (3, 48), & -\rho = -46 < -1 \\
\alpha = (0.7, 40), & -\rho = -40.3 < -1 \\
\alpha = (0.6, 0.9), & -\rho = -1.3 < -1 \\
\alpha = (2.5, 2.9), & -\rho = -1.4 < -1 \\
\alpha = (15, 20), & -\rho = -6 < -1 \\
\alpha = (20, 21), & -\rho = -2 < -1
\end{array}$$

We see that our numerical values are agreed with the variation of the value of regular variation exponent $(-\rho)$ and are met in the condition (4.3).

4.3. **Unimodality.** Unimodality is an essential feature for discrete distributions arising in bioinformatics. For details about this, we refer the readers to, for example, [2, 13, 14, 15]. In this subsection, we study such feature for the 2-RDWP model. Compared to Bhati and Bakouch [4], let us give a proposition as follows.

Proposition 4.1. *The pmf (2.2) is unimodal with mode value at* $x = 1$.

Proof. Let us consider pmf (2.2) for the positive integer value of x. Then for $x \geq 1$, we get

(4.5)

$$\frac{dg_x(\alpha)}{dx} = \frac{d}{dx}\left(\frac{1}{d(\alpha)} \cdot \frac{(r+x-1)^{r+x-1}}{(q+x)^{q+x}}\right)$$

$$= \frac{1}{d(\alpha)} \cdot \frac{1}{\left((q+x)^{q+x}\right)^2} \cdot (r+x-1)^{r+x-1} \cdot (q+x)^{q+x} \cdot \big[\ln(r+x-1) - \ln(q+x)\big].$$

It is obvious that for $0 < r < q$

$$\ln(r+x-1) - \ln(q+x) < 0.$$

So, we conclude that $\frac{dg_x(\alpha)}{dx}$ given by (4.5) is always negative. It implies that $g_x(\alpha)$ decreases and takes its mode at $x = 1$. The proof is completed. □

In addition to Proposition 4.1, let us investigate unimodality as numerical. We have a recursive formula given by

$$\frac{g_{x+1}(\alpha)}{g_x(\alpha)} = \frac{(r+x)^{r+x}(q+x)^{q+x}}{(r+x-1)^{r+x-1}(q+x+1)^{q+x+1}}, \qquad x = 1, 2, ... \tag{4.6}$$

Numerically, it can be shown that $\frac{g_{x+1}(\alpha)}{g_x(\alpha)} < 1$. Let us have the following example.

Example 4.2. *Let us consider some values of parameters* $(r = 0.7, q = 1)$ *and* $(r = 1.5, q = 2.5)$. *From (4.6), we calculate* $\frac{g_{x+1}(\alpha)}{g_x(\alpha)}$, *for* $x = 1, 2, 3, 4, 5, 6$, *in Table 1 as follows:*

Таблица 1. The behavior of $\frac{g_{x+1}(\alpha)}{g_x(\alpha)}$ (4.6) for different values of parameters r and q

$\alpha = (r, q)$	$\frac{g_2(\alpha)}{g_1(\alpha)}$	$\frac{g_3(\alpha)}{g_2(\alpha)}$	$\frac{g_4(\alpha)}{g_3(\alpha)}$	$\frac{g_5(\alpha)}{g_4(\alpha)}$	$\frac{g_6(\alpha)}{g_5(\alpha)}$	$\frac{g_7(\alpha)}{g_6(\alpha)}$
$\alpha = (0.7, 1)$	*0.46869*	*0.62523*	*0.70967*	*0.76288*	*0.79955*	*0.82638*
$\alpha = (1.5, 2.5)$	*0.49602*	*0.59820*	*0.66569*	*0.71370*	*0.74962*	*0.77752*

From Table 1, we see that the expression, as in (4.6), increases when x increases and also for $x = 1, 2, 3, 4, 5, 6,$ the values $\frac{g_{x+1}(\alpha)}{g_x(\alpha)} < 1$. Numerically, it seems that $g_x(\alpha)$ defined by (2.2) decreases and is downward convex. Automatically, the unimodality of $g_x(\alpha)$ is received.

Moreover, in Section 3, we plotted the pmf of 2-RDWP (2.2) for different values of parameters. In other words, intuitively and from the graphical approach in Figures 1(A-J), it is readily seen that the pmf of 2-RDWP is unimodal. The modes are observed for all plots in Figures 1(A-J) at $x = 1$.

4.4. **Skewness to the right.** One of the essential properties of discrete distributions (frequency distributions) arising in biomolecular systems is the skewness to the right of the pmf. This property has been discovered by experimental methods based on the observation of various data sets of such systems. The conception of skewness for biologists is based on intuition and the shapes of graphs of discrete distributions [2, 3]. Section 3 displayed the plots of the pmf of 2-RDWP (2.2) for different possible parameter values. Intuitively and from the graphical approach in Figures 1(A-J), it can be observed that the pmf of 2-RDWP (2.2) is skewed to the right. Here, let us have a numerical example.

Example 4.3. *Let us have some real data that includes the number of proteins assigned to Panther families or subfamilies as follows* [18]*:*
1, 17, 11, 22, 16, 10, 61, 10, 12, 15, 22, 10, 5, 1, 33, 6, 11, 1, 5, 3, 2, 9, 22, 10, 3, 86, 1, 1, 15, 5, 8, 26, 2, 14, 2, 9, 62, 7, 114, 113, 20, 22, 14, 12, 13, 6, 24, 26, 22, 51, 56, 106, 59, 55, 29, 1, 141, 168, 607, 395, 616, 1, 7, 19, 3, 29, 59, 4, 4, 1, 3, 18, 60, 46, 11, 56, 269, 812.

The value of skewness for these data is 3.960*.*

The following mathematical result is received from Subsections 4.1 – 4.4.

Corollary 4.1. *The common statistical facts (unimodality, skewness to the right, upward/downward convexity, regular variation at infinity, asymptotically constant slowly varying component) hold for the model $g_x(\alpha)$ (2.2). Therefore, from the mathematical point of view, the 2-RDWP model (2.2) may be considered as a new regularly varying frequency distribution for the needs of large-scale biomolecular systems, bioinformatics, etc. For details about this, see* [2, 3, 5, 15].

5. On the ML estimators

This section gives ML estimators for the model (2.2). We get the conditions of coincidence of solution for the system of likelihood equations with the ML estimators

for the unknown parameters. Let us define the functions $h(x;\alpha)$ and $t(x;\alpha)$ by

$$h(x;\alpha) = \ln(r+x-1)+1, \quad t(x;\alpha) = -\big(\ln(q+x)+1\big),$$

and $\overline{h_n(\alpha)} = \frac{1}{n}\sum_{i=1}^{n} h(x_i;\alpha)$, $\overline{t_n(\alpha)} = \frac{1}{n}\sum_{i=1}^{n} t(x_i;\alpha)$. We state a lemma for the model (2.2).

Lemma 5.1. *For model (2.2), we have the following*

$$E\big[h(\xi;\alpha)\big] < \infty, \qquad E\big[t(\xi;\alpha)\big] < \infty,$$

where $E[\cdot]$ is the mathematical expectation.

Proof. Based on the definition of mathematical expectation, the proof is satisfied, obviously.

From Lemma 5.1, and compared to Farbod and Gasparian [11], let us present a theorem.

Theorem 5.1. *The likelihood equations for obtaining the ML estimators of parameter α with the model (2.2) have the following moments equations*

$$\begin{cases} E\big[h(\xi;\alpha)\big] = \overline{h_n(\alpha)} \\ \\ E\big[t(\xi;\alpha)\big] = \overline{t_n(\alpha)} \end{cases} \tag{5.1}$$

Proof. We consider the likelihood function $L(X^n;\alpha) = \prod_{i=1}^{n} g_{x_i}(\alpha)$. The logarithm of the likelihood function is given by

$$l(X^n;\alpha) = \ln L(X^n;\alpha) = \sum_{i=1}^{n} \ln \frac{(r+x_i-1)^{r+x_i-1}}{(q+x_i)^{q+x_i}} - n\ln d(\alpha) \tag{5.2}$$

If the following conditions hold

$$\frac{\partial l(X^n;\alpha)}{\partial r} = 0, \quad \frac{\partial l(X^n;\alpha)}{\partial q} = 0,$$

then the ML estimators of the parameters $\alpha = (r,q)$ exist.

Let us obtain derivatives by parameters r and q. We have

$$\frac{\partial l(X^n;\alpha)}{\partial r} = \sum_{i=1}^{n} \Big(\ln(r+x_i-1)+1\Big) - n\frac{1}{d(\alpha)}\frac{\partial d(\alpha)}{\partial r}$$

where $\frac{1}{d(\alpha)}\frac{\partial d(\alpha)}{\partial r} = E\big[h(\xi;\alpha)\big]$. From $\frac{\partial l(X^n;\alpha)}{\partial r} = 0$, we get $E\big[h(\xi;\alpha)\big] = \overline{h_n(\alpha)}$.

Meanwhile, we have

$$\frac{\partial l(X^n;\alpha)}{\partial q} = \sum_{i=1}^{n} -\Big(\ln(q+x_i)+1\Big) - n\frac{1}{d(\alpha)}\frac{\partial d(\alpha)}{\partial q}$$

where $\frac{1}{d(\alpha)}\frac{\partial d(\alpha)}{\partial q} = E\big[t(\xi;\alpha)\big]$. From $\frac{\partial l(X^n;\alpha)}{\partial q} = 0$, we obtain $E\big[t(\xi;\alpha)\big] = \overline{t_n(\alpha)}$.

The Theorem 5.1 is proved. □

We aim to show that the solution $\hat{\alpha}$ of the system (5.1) is the ML estimator of the parameter α. It is sufficient to establish that the matrix $\hat{M}_n = (\hat{M}_{ij}^n)_{i,j=1}^2$ with $\hat{M}_{ij}^n = \hat{M}_{ij}^n(\hat{\alpha})$, $\hat{M}_{ij}^n(\hat{\alpha}) = \frac{\partial^2 l(X^n;\alpha)}{\partial r \partial q}|_{\alpha=\hat{\alpha}}$ is negative definite. Let us state two lemmas.

Lemma 5.2. *Consider the model (2.2). Assuming the solution $\widehat{\alpha}$ of the system (5.1) (if it exists) holds in the following conditions*

$$\begin{cases} E\big[\psi(\xi;\alpha)\big] = \overline{\psi_n(\alpha)} \\ \\ E\big[\eta(\xi;\alpha)\big] = \overline{\eta_n(\alpha)} \end{cases} \tag{5.3}$$

where

$$\psi(\xi;\alpha) = \frac{1}{r+x-1},\ \overline{\psi_n(\alpha)} = \frac{1}{n}\sum_{i=1}^{n}\psi(x_i;\alpha);\quad \eta(\xi;\alpha) = -\frac{1}{q+x},\ \overline{\eta_n(\alpha)} = \frac{1}{n}\sum_{i=1}^{n}\eta(x_i;\alpha).$$

Then, the elements of the matrix $\hat{M}_n$ are as follows ($Var(\cdot)$ is the variance and $Cov(\cdot,\cdot)$ is the covariance):

$$\hat{M}_{11} = -n\ Var(h(\xi;\alpha)),$$

$$\hat{M}_{12} = \hat{M}_{21} = -n\ Cov\big(h(\xi;\alpha), t(\xi;\alpha)\big),$$

$$\hat{M}_{22} = -n\ Var(t(\xi;\alpha)).$$

Proof. We obtain second derivatives of the logarithm of likelihood functions by

$$\frac{\partial^2 l(X_n;\alpha)}{\partial r^2} = -n\Big(\frac{1}{d(\alpha)}\frac{\partial^2 d(\alpha)}{\partial r^2} - \big(\frac{1}{d(\alpha)}\frac{\partial d(\alpha)}{\partial r}\big)^2\Big) + n\overline{\psi_n(\alpha)}$$

$$\frac{\partial^2 l(X_n;\alpha)}{\partial r \partial q} = \frac{\partial^2 l(X_n;\alpha)}{\partial q \partial r} = -n\Big[\frac{1}{d(\alpha)}\frac{\partial^2 d(\alpha)}{\partial r \partial q} - \Big(\frac{1}{d(\alpha)}\frac{\partial d(\alpha)}{\partial r}\Big)\Big(\frac{1}{d(\alpha)}\frac{\partial d(\alpha)}{\partial q}\Big)\Big]$$

$$\frac{\partial^2 l(X_n;\alpha)}{\partial q^2} = -n\Big(\frac{1}{d(\alpha)}\frac{\partial^2 d(\alpha)}{\partial q^2} - \big(\frac{1}{d(\alpha)}\frac{\partial d(\alpha)}{\partial q}\big)^2\Big) + n\overline{\eta_n(\alpha)}$$

After some simplification, we have

$$M_{11} = -n\ Var\big(h(\xi;\alpha)\big) - n\big(E\big[\psi(\xi;\alpha)\big] - \overline{\psi_n(\alpha)}\big)$$

$$M_{12} = M_{21} = -n\ Cov\big(h(\xi;\alpha), t(\xi;\alpha)\big)$$

$$M_{22} = -n\ Var\big(t(\xi;\alpha)\big) - n\big(E\big[\eta(\xi;\alpha)\big] - \overline{\eta_n(\alpha)}\big)$$

With the help of (5.3) the proof of Lemma 5.2 is finished. □

Lemma 5.3. *Consider the model (2.2). Under the conditions (5.3), the matrix $\hat{M}_n$ is negative definite.*

Proof. It suffices to show $\hat{M}_{11}^n < 0$ and $\det(\hat{M}^n) > 0$. From Lemma 5.2, it can be concluded that $\hat{M}_{11}^n < 0$. To show that $\det(\hat{M}^n) > 0$ we give

$$\det(\hat{M}^n) = \hat{M}_{11}^n \hat{M}_{22}^n - (\hat{M}_{12}^n)^2$$

In accord with the value of $\hat{M}_{11}^n$, $\hat{M}_{22}^n$, $\hat{M}_{12}^n$ and based on Cauchy-Bunyakovski-Schwartz inequality the proof is completed. □

From Lemmas 5.2 and 5.3, the following result is given.

Corollary 5.1. *Suppose that the solution of the system (5.1) satisfies the conditions (5.3), then it coincides with the ML estimators of parameters.*

6. ML estimaton and simulation

Based on systems (5.1) and (5.3), it is not easy to derive closed forms for the solutions, analytically. So, we need to use a numerical method for the ML estimations of unknown parameters. Compared to Farbod [8], Nelder-Mead optimization algorithm (or *simplex search algorithm*) is suggested. Let us notice that the Nelder-Mead optimization algorithm is a free-derivative optimization method to nonlinear optimization problems and is suggested to apply for models with more than one parameter. This algorithm was introduced by Nelder and Mead [16]. See also [17].

For sampling, a simple stochastic sampling with replacement with the probability of variables is considered in which the probability of variables are probability functions. Simulation studies are proposed using the Monte Carlo method [17] with 1000 iterations to calculate ML estimations, biases and mean square errors (MSEs).

Remark 6.1. *First, we performed our simulation for the model (2.2). Based on (2.2), our simulation works well when $x = 1, 2, ..., xmax$ ($xmax = 100$). But we have some computational problems for large values x, such as $xmax = 150$ and bigger than 150. Let us notice that a type of function x^x exists in our pmf's form (2.2), and hence it raises problems for simulations and numerical calculations when x is large. For example, if $x = 500$, then using R statistical software (Version 4.2.2) $x^x = 500^{500} = \infty$. To solve this computational problem, without loss of generality and after some mathematical simplification, our pmf (2.2) can be written as follows:*

$$g_x^*(\alpha) = \left(\sum_{y=1}^{\infty} \frac{\left(1+\frac{r-1}{y}\right)^y \left(y+r-1\right)^{r-1}}{\left(1+\frac{q}{y}\right)^y (y+q)^q} \right)^{-1} \cdot \frac{\left(1+\frac{r-1}{x}\right)^x \left(x+r-1\right)^{r-1}}{\left(1+\frac{q}{x}\right)^x \left(x+q\right)^q}. \tag{6.1}$$

From Remark 6.1 and (6.1), we have the following corollary.

Corollary 6.1. *It is readily seen that the pmf (2.2) equals the pmf (6.1). So, for simulation studies, the pmf (6.1) is considered. In the formula (6.1), we need to have*

x as truncated. For simulation aims, let us set $x = 1$ to xmax ($xmax = 10000$). Namely, we have $x = 1, 2, ..., 10000$ and $y = 1, 2, ..., 10000$.

Let us consider the logarithm of the likelihood function (5.2). Based on (6.1) and $x = 1, 2, ..., 10000$, the ML estimations, biases, and MSEs are calculated. To simulation studies, we consider the values $(r = 0.4, q = 0.6), (r = 1, q = 2), (r = 2.5, q = 3.2)$ as true values, different sample sizes $n = 50, 100, 200, 500, 1000, 5000$, and using 1000 iterations.

Using R statistical software, the simulation results are given in Table 2. Our simulation studies work well and have satisfactory results for the model. The differences between real and estimated values of the parameters are small, in particular for large sample sizes.

Table 2 shows that when the sample size n increases, bias and MSE decrease. Moreover, from Table 2, we observe that when the true values of r and q are smaller (also a small value of $q-r$), the results are better, i.e. biases and MSEs are smaller. Let us notice that for the ML estimations, the conditions $\hat{q} - \hat{r} > 0$ and $\hat{\rho} > 1$ are satisfied.

7. Asymptotic expansion

Considering that our proposed model has no closed form for the pmf, obtaining some useful asymptotic expansion with two terms for the pmf is interesting. From (2.2) and (6.1), we get

$$g_x(\alpha) = (d(\alpha))^{-1} \cdot x^{r-q-1} \cdot \frac{\left(1+\frac{r-1}{x}\right)^x \left(1+\frac{r-1}{x}\right)^{r-1}}{\left(1+\frac{q}{x}\right)^x \left(1+\frac{q}{x}\right)^q} \tag{7.1}$$

We use two known asymptotic expansions as follows. For $x \longrightarrow \infty$, we have [5, 12]:

$$(1 + \tfrac{c}{x})^x = e^c \cdot \Big(1 - \tfrac{c^2}{2x} + O(\tfrac{1}{x^2})\Big). \tag{7.2}$$

Also, for $x \longrightarrow 0$ we have

$$(1 + x)^\alpha = 1 + \alpha x + O(x^2). \tag{7.3}$$

From (7.2) and (7.3), the formula (7.1) may be given by

(7.4)

$$\begin{aligned} g_x(\alpha) &= (d(\alpha))^{-1} \cdot x^{r-q-1} \cdot e^{r-q-1} \frac{\left(1+\frac{r-1}{x}\right)^x \left(1+\frac{r-1}{x}\right)^{r-1}}{\left(1+\frac{q}{x}\right)^x \left(1+\frac{q}{x}\right)^q} \\ &= (d(\alpha))^{-1} \cdot x^{r-q-1} \cdot e^{r-q-1} \cdot \frac{\left(1-\frac{(r-1)^2}{2x}+O(\frac{1}{x^2})\right)}{\left(1-\frac{q^2}{2x}+O(\frac{1}{x^2})\right)} \cdot \frac{\left(1+\frac{(r-1)^2}{x}+O(\frac{1}{x^2})\right)}{\left(1+\frac{q^2}{x}+O(\frac{1}{x^2})\right)} \end{aligned}$$

Таблица 2. Simulation results: The values of ML estimations $(\hat{r}, \hat{q})$, biases, and
MSEs for the 2-RDWP model (6.1)

	$(r = 0.4, q = 0.6); \quad x = 1, 2, ..., 10000$		
n	$(\hat{r}, \hat{q})$	Bias	MSE
50	(0.6156,0.8434)	(0.2156,0.2434)	(0.3234,0.4047)
100	(0.4844,0.6955)	(0.0844,0.0955)	(0.1013,0.1307)
200	(0.4318,0.6346)	(0.0318,0.0346)	(0.0420,0.0560)
500	(0.4108,0.6115)	(0.0108,0.0115)	(0.0150,0.0204)
1000	(0.4049,0.6051)	(0.0049,0.0051)	(0.0072,0.0098)
5000	(0.4002,0.6001)	(0.0002,0.0001)	(0.0014,0.0020)
	$(r = 1, q = 2); \quad x = 1, 2, ..., 10000$		
n	$(\hat{r}, \hat{q})$	Bias	MSE
50	(1.4683,2.6768)	(0.4683,0.6768)	(1.9440,3.8043)
100	(1.1716,2.2475)	(0.1716,0.2475)	(0.4026,0.7901)
200	(1.0673,2.0963)	(0.0673,0.0963)	(0.1254,0.2417)
500	(1.0236,2.0330)	(0.0236,0.0330)	(0.0461,0.0872)
1000	(1.0158,2.0229)	(0.0158,0.0229)	(0.0227,0.0430)
5000	(1.0039,2.0051)	(0.0039,0.0051)	(0.0041,0.0079)
	$(r = 2.5, q = 3.2); \quad x = 1, 2, ..., 10000$		
n	$(\hat{r}, \hat{q})$	Bias	MSE
50	(3.1808,3.9675)	(0.6808,0.7675)	(5.4404,6.8484)
100	(2.748,3.4778)	(0.2480,0.2778)	(1.3875,1.7214)
200	(2.6214,3.3357)	(0.1214,0.1357)	(0.5669,0.6986)
500	(2.5541,3.2609)	(0.0541,0.0609)	(0.2028,0.2494)
1000	(2.5367,3.2415)	(0.0367,0.0415)	(0.0992,0.1226)
5000	(2.5093,3.2102)	(0.0093,0.0102)	(0.0191,0.0238)

Let $\rho = q + 1 - r$. From (7.4), we get

(7.5)

$$g_x(\alpha) \approx \big(d(\alpha)\big)^{-1} \cdot x^{-\rho} \cdot e^{-\rho} \times \Big(1 + \tfrac{1}{2x}\big((r-1)^2 - q^2\big) + \big((r-1)^2 - q^2\big)O(\tfrac{1}{x^2})\Big)$$

$$\approx \big(d(\alpha)\big)^{-1} \cdot x^{-\rho} \cdot e^{-\rho} \cdot \Big(1 + \tfrac{1}{2x}\big((r-1)^2 - q^2\big) + O(\tfrac{1}{x^2})\Big).$$

7.1. **Tail behavior.** Using asymptotic expansions (7.2), (7.3) and based on (7.4), let us propose tail behavior of distribution function $F_x(\alpha)$ (2.4) when $x \longrightarrow \infty$.

From (2.2), we get

$$1 - F_x(\alpha) = P(X > x) = \sum_{m=x+1}^{\infty} g_x(\alpha) \tag{7.6}$$

By substituting (7.4) and (7.5) into (7.6), when $x \longrightarrow \infty$, we have

$$1 - F_x(\alpha) \approx \big(d(\alpha)\big)^{-1} \, e^{-(q+1-r)} \sum_{m=x+1}^{\infty} m^{-(q+1-r)}. \tag{7.7}$$

The following corollary is given.

Corollary 7.1. *It follows from (7.7) that the condition (4.3) must be met.*

7.2. **Moments.** It is known that some moments are undefined for every power law-like distribution. We investigate the moment's existence of the model (2.2). To do that, using asymptotic expansion (7.5), we shall propose the moment's existence of integer orders of the 2-RDWP model (2.2). Let $\rho = q + 1 - r$.

From (7.5), it is readily seen that the first-order moment of X is finite if $q-r > 1$ (or equivalently $\rho > 2$). In other words, for model (2.2):

$$E(X) < \infty, \quad if\ \rho > 2.$$

For the second-order moment, it is easy to see that

$$E(X^2) < \infty, \quad if\ \ \rho > 3.$$

Hence, the variance for the model (2.2) is also finite if $\rho = q + 1 - r > 3$. In other words, we have

$$Var(X) = E(X^2) - E^2(X) < \infty, \quad if\ \rho > 3.$$

In the general case, if $q - r > j$ then

$$E(X^j) < \infty, \ \ j = 1, 2, ...; \quad if\ \ \rho > j + 1.$$

Corollary 7.2. *Assume that X is a regularly varying random variable with a distribution (2.2) and index ρ. Then the moment of order j is infinite if $\rho \leq j + 1$.*

Moreover, evaluating the mean and variance of the model (2.2) for practical needs is of interest. From the proposed asymptotic expansion (7.5), we can present an approximate form with two terms for the mean and variance. Let us obtain mean as a practical form for the truncated function with two terms as

(7.8)
$$E(X) = \sum_{x=1}^{\infty} g_x(\alpha) \approx (d(\alpha))^{-1} e^{-\rho} \Big[\sum_{x=1}^{\infty} x^{-\rho+1} + \tfrac{1}{2}((r-1)^2 - q^2) \sum_{x=1}^{\infty} x^{-\rho} \Big].$$

Compared to Astola and Danielian [2, p.29], we have

(7.9)
$$\begin{cases} \sum_{x=1}^{\infty} x^{-\rho} = \frac{1}{\Gamma(\rho-1)} \lim_{\lambda \longrightarrow 1} \int_0^1 \ln(1-\lambda t)(\ln \frac{1}{t})^{\rho-2} \frac{dt}{t} \\ \\ \sum_{x=1}^{\infty} x^{-\rho+1} = \frac{1}{\Gamma(\rho-2)} \lim_{\lambda \longrightarrow 1} \int_0^1 \ln(1-\lambda t)(\ln \frac{1}{t})^{\rho-3} \frac{dt}{t} \end{cases}$$

where $0 < \lambda < 1$ is some small constant and $\Gamma(\cdot)$ is the Gamma function. Substituting (7.9) into (7.8), an approximate form with two terms for the mean is given in the practical form and integral representation. Similarly, we can provide integral representations for the second order moment and also variance.

8. Application to data and comparison

As we pointed out in Section 2 and verified in Section 4, our new discrete distribution (2.2) may be considered in bioinformatics, biosystems, etc. Let us fit our model with a real count data set (Example 4.3) and then compare it with the other models in bioinformatics. Again, for simulation and fitting aims, we consider the pmf form (6.1).

The given real data set is the number of proteins in a biological system. In other words, we consider some real data that includes the number of proteins assigned to Panther families or subfamilies (see Subsection 4.4, Example 4.3). These data are collected from Venter et al. [18].

For these 78 data (data used in the Example 4.2), using (6.1) we obtain the ML estimations for two parameters r and q. ML estimations are given by $\hat{r} = 9.620101$, $\hat{q} = 10.378817$. It implies $-\hat{\rho} = -1.758716 < -1$. In addition, $\ln L = -357.063$ and p-value=0.713. Based on the Kolmogorov-Smirnov test and the 2-RDWP model, the p-value equals 0.713, which is a good fit for such real data. Additionally, based on some well-known statistical criteria such as:

Akaike information criterion (AIC) is given by $AIC = 2\ln L + 2k$ where k the number of parameters in the model; $-\ln L$ is the maximized value of the likelihood function for the estimated model; AIC with corrected (AICc) is given by $AICc = AIC + \frac{2k^2+2k}{n-k-1}$ where n is the sample size; and p-value, we compare the 2-RDWP with other discrete models arising in bioinformatics, such as the one-parameter skewed discrete Levy distribution (DLD) (10.1) [7], one-parameter skewed Power-Law (PL) model (10.2) [2], one-parameter truncated skewed discrete stable distribution (T-SDSD) (10.3) [8], one-parameter truncated skewed discrete stable distribution (T-DSD) (10.4) [8], and two-parameter truncated skewed discrete stable distribution (T-2SDSD) (10.5) [8], all having support on the set of positive integers, i.e. $x \in \mathbb{N}_+ = \{1, 2, 3, ...\}$.

Using R statistical software, our results are presented in Table 3. It can be observed from Table 3 that the 2-RDWP model has the smallest $-\ln L$, AIC, AICc, and the largest p-value. Accordingly, we can conclude that the 2-RDWP model provides the best fit among the compared models (DLD, PL, T-SDSD, T-DSD and T-2SDSD models). The pmfs of DLD, PL, T-SDSD, T-DSD and T-2SDSD are given in Section 10 (Appendix).

Таблица 3. Comparing results for 2-RDWP, DLD, PL, T-SDSD, T-DSD, and T-2SDSD models for data of Example 4.2

Model	$\ln L$	k	AIC	AICc	p-value
2-RDWP	-357.063	2	718.126	718.286	0.713
DLD	-360.51055	1	723.0211	723.07373	0.04948
PL	-366.2436	1	734.4872	734.53983	0.01773
T-SDSD	-362.7622	1	727.5244	727.57703	0.09018
T-DSD	-360.57675	1	723.1535	723.20613	0.1746
T-2SDSD	-360.55815	2	725.1163	725.2763	0.1723

9. Conclusions

In this paper, using the discretization methods, we formulated a new skewed regular varying discrete distribution, the so-called 2-RDWP, given by Eq.(2.2). Some plots for the pmf and log-log plots of the model have been illustrated for the different values of parameters satisfying the condition in Eq.(4.3). Figures 1(A-J) indicated the pmfs for the used parameters are skewed to the right and unimodal with mode value at $x = 1$. Significantly, Figures 1(A-J) showed that the length and shape of the right-side tails varied with parameter value changes. Figures 2(A-J) established the log-log plots of the 2-RDWP (2.2). The log-log plots of Figures 2(A-J) illustrated that the right-side tails could significantly deviate from the straight line, at least for large values of observed x.

The known common *statistical facts (empirical facts)*, including unimodality, skewness to the right, upward/downward convexity, stability by estimated parameters values, regular variation at infinity and asymptotically constant slowly varying component, have been proved for the 2-RDWP model. Hence, mathematically, we concluded that our model (2.2) could be used as a new probability distribution for the needs of bioinformatics and biomolecular systems.

ML estimators have been obtained based on some moment equations. The conditions of coincidence of solution for the system of likelihood equations with the ML estimators for the parameters have been proposed. Based on Monte Carlo method and Nelder-Mead optimization algorithm simulation studies have been given to get ML estimations, biases and MSEs. Simulation studies presented satisfactory results. We noted that for simulation aims, instead of Eq.(2.2), we considered the pmf in Eq.(6.1). The ML estimations $\hat{r}$ and $\hat{q}$ have met in the conditions, namely based on simulation studies $\hat{q} - \hat{r} > 0$ and $\hat{\rho} > 1$.

An asymptotic expansion with two terms for the pmf (2.2) has been given. Using asymptotic expansions, we proposed tail behavior of distribution function. Also, we

investigated the moment's existence of integer orders. Then, based on asymptotic expansion, useful formulas for the mean and variance in the truncated forms have been provided.

Finally, we successfully applied 2-RDWP to a real data set. Based on well-known statistical criteria, we compared our results for the proposed model with other known models in biosystems. Our model gives better results than the other models for this real data set (Table 3).

The 2-RDWP model has a long right-side tail and power law-like behavior. It can be helpful in biomolecular systems, bioinformatics and other areas such as economics and physics.

10. Appendix

We present the pmfs of some rival models, used in Table 3. The pmf of the one-parameter DLD model is given by [7]

$$p_x(\gamma) = \frac{x^{-\frac{3}{2}}\exp(-\frac{\gamma}{2x})}{\sum_{y=1}^{\infty} y^{-\frac{3}{2}}\exp(-\frac{\gamma}{2y})}, \quad x = 1, 2, ...; \quad \gamma > 0. \tag{10.1}$$

The pmf of the one-parameter PL model is as [2]

$$p_x(\nu) = \frac{x^{-\nu}}{\sum_{y=1}^{\infty} y^{-\nu}}, \quad x = 1, 2, ...; \quad \nu > 1. \tag{10.2}$$

The pmf of T-SDSD when $0 < \theta < 1$, and $x = 1, 2, \ldots$, is given by [8]

$$p_x(\theta, 1) = \frac{\Gamma(\theta+1)x^{-\theta-1}\sin(\pi\theta) - \frac{1}{2}\Gamma(2\theta+1)x^{-2\theta-1}\sin(2\pi\theta)}{\sum_{y=1}^{\infty}\left(\Gamma(\theta+1)y^{-\theta-1}\sin(\pi\theta) - \frac{1}{2}\Gamma(2\theta+1)y^{-2\theta-1}\sin(2\pi\theta)\right)}. \tag{10.3}$$

The pmf of T-DSD when $0 < \theta < 2$, and $x = 1, 2, \ldots$, is given by [8]

$$p_x(\theta, 0) = \frac{\Gamma(\theta+1)x^{-\theta-1}\sin(\frac{\pi\theta}{2}) - \frac{1}{2}\Gamma(2\theta+1)x^{-2\theta-1}\sin(\pi\theta)}{\sum_{y=1}^{\infty}\left(\Gamma(\theta+1)y^{-\theta-1}\sin(\frac{\pi\theta}{2}) - \frac{1}{2}\Gamma(2\theta+1)y^{-2\theta-1}\sin(\pi\theta)\right)}. \tag{10.4}$$

The pmf of T-2SDSD when $0 < \theta < 2$, $0 < \beta < 1$, and $x = 1, 2, \ldots$, is given by [8]

$$p_x(\theta, \beta) = \frac{\Gamma(\theta+1)x^{-\theta-1}\sin(\frac{\pi\theta(1+\beta)}{2}) - \frac{1}{2}\Gamma(2\theta+1)x^{-2\theta-1}\sin(\pi\theta(1+\beta))}{\sum_{y=1}^{\infty}\left(\Gamma(\theta+1)y^{-\theta-1}\sin(\frac{\pi\theta(1+\beta)}{2}) - \frac{1}{2}\Gamma(2\theta+1)y^{-2\theta-1}\sin(\pi\theta(1+\beta))\right)}. \tag{10.5}$$

Code availability. All computational, fitting and simulation studies have been done using R statistical software. The R codes are available from the author upon request.

Acknowledgements. The author is grateful to the anonymous reviewers and the handling editor for their valuable comments and suggestions, which definitely improved the quality and presentation of the paper.

Список литературы

[1] J. Astola and E. Danielian, "Dediscretization of distributions arising in macroevolution models", Facta Universitatis, Series: Electronics and Energetics, **20**(2), 119 – 146 (2007).

[2] J. Astola and E. Danielian, Frequency Distributions in Biomolecular Systems and Growing Networks, Tampere International Center for Signal Processing (TICSP), Series no. 31, Tampere, Finland (2007).

[3] J. Astola, E. Danielian and S. Arzumanyan, "Frequency distributions in bioinformatics, a review", Proceedings of the Yerevan State University: Physical and Mathematical Sciences, **223**(3), 3 – 22 (2010).

[4] D. Bhati and H. S. Bakouch, "A new infinitely divisible discrete distribution with applications to count data modeling", Communications in Statistics - Theory and Methods, **48**(6), 1401 – 1416 (2019).

[5] E. Danielian, R. Chitchyan and D. Farbod, "On a new regularly varying generalized hypergeometric distribution of the second type", Mathematical Reports, **18(68)**(2), 217 – 232 (2016).

[6] D. Farbod, "M-estimators as GMM for stable laws discretizations", Journal of Statistical Research of Iran, **8**(1), 85 – 96 (2011).

[7] D. Farbod, "Some statistical inferences for two frequency distributions arising in bioinformatics", Applied Mathematics E-Notes, **14**, 151 – 160 (2014).

[8] D. Farbod, "Modeling and simulation studies for some truncated discrete distributions generated by stable densities", Mathematical Sciences, **16**(2), 105 – 114 (2022).

[9] D. Farbod, A. Iranmanesh and M. Basirat, "A continuous analog of the generalized hypergeometric distribution generated by dediscretization method", Journal of Mathematical Extension, **16**(10), 1 – 16 (2022).

[10] D. Farbod and K. Gasparian, "On the confidence intervals of parametric functions for distributions generated by symmetric stable laws", Statistica, **72**(4), 405 – 413 (2012).

[11] D. Farbod and K. Gasparian, "On the maximum likelihood estimators for some generalized Pareto-like frequency distribution", Journal of the Iranian Statistical Society (JIRSS), **12**(2), 211 – 233 (2013).

[12] I. S. Gradshteyn and I. M. Ryznik, Tables of Integrals, Series and Products, Translated from Russian by S. Technica, edited by A. Jeffrey and D. Zwillinger, Academic Press (2007).

[13] V. A. Kuznetsov, "Family of skewed distributions associated with the gene expression and proteome evolution", Signal Processing, **33**(4), 889 – 910 (2003).

[14] V. A. Kuznetsov, Mathematical modeling of avidity distribution and estimating general binding properties of transcription factors from genome-wide binding profiles, In: Tatarinova, T. V. and Nikolsky, Y. (eds), Biological Networks and Pathway Analysis, pp. 193 – 276, Springer, New York (2017).

[15] V. A. Kuznetsov, A. Grageda and D. Farbod, "Generalized hypergeometric distributions generated by birth-death process in bioinformatics", Markov Processes and Related Fields, **28**(2), 303 – 327 (2022).

[16] J. A. Nelder and R. Mead, "A simplex method for function minimization", Computer Journal, **7**(4), 308 – 313 (1965).

[17] M. L. Rizzo, Statistical Computing with R, 2nd edition, Chapmann and Hall/CRC, (2019).

[18] J. C. Venter, et al., "The sequence of the human genome", Science, **291**(5507), 1304 – 1351 (2001).

Поступила 19 мая 2023

После доработки 02 июня 2023

Принята к публикации 04 октября 2023