Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.18576/amis/180418

Discrete Extension of the Inverse Weibull Distribution: Theory and Decision-Making for Count Data in Sustainability Analysis

Mahmoud El-Morshedy[1,2,*], *Mohamed S. Eliwa*[2], *Abhishek Tyagi*[3], *and Hend S. Shahen*[2]

[1]Department of Mathematics, College of Science and Humanities in Al-Kharj, Prince Sattam Bin Abdulaziz University, Al-Kharj 11942, Saudi Arabia
[2]Department of Mathematics, Faculty of Science, Mansoura University, Mansoura, 35516, Egypt
[3]Department of Statistics, Chaudhary Charan Singh University, Meerut, India

Received: 3 Mar. 2024, Revised: 15 May 2024, Accepted: 22 May 2024
Published online: 1 Jul. 2024

Abstract: In the realm of sustainability, lifetimes are often modeled with discrete measurements due to finite precision, lacking a continuous representation. Despite the inherent continuity in device or patient lifetimes, it is reasonable to consider their observations as stemming from a discretized distribution derived from a continuous model. This study introduces a discrete random probability model based on non-negative integers, formulated from the established Kumaraswamy family using recognized discretization methods while preserving the survival function's structure. The generated discrete model is called the Kumaraswamy discrete inverse Weibull. Various statistical properties, such as the hazard rate function, moments, dispersion index, skewness, kurtosis, quantile function, L-moments, and entropies, are explored. The new discrete model's parameters are estimated using maximum likelihood estimation, followed by a discussion on its performance in a simulation study. Additionally, three real-world sustainability applications using count data showcase the importance and versatility of this innovative discrete distribution.

Keywords: Survival discretization technique; Failure analysis; Dispersion index; Maximum likelihood approach; Simulation; Sustainability Count data; Goodness-of-fit test.

1 Introduction

Sustainability is fundamentally about balancing the needs of the present without compromising the ability of future generations to meet their own needs. Discrete probability distributions, a key concept in probability theory, play a crucial role in modeling and predicting various sustainability-related scenarios. These distributions enable the analysis of uncertain events with distinct outcomes and probabilities, helping decision-makers assess risks and make informed choices in areas such as sustainable resource management, environmental conservation, and energy planning. For example, discrete probability distributions can be used to evaluate the probability of different weather patterns affecting crop yields or to estimate the likelihood of specific outcomes in biodiversity conservation efforts. By applying these mathematical models, policymakers and researchers can quantify uncertainties, optimize resource allocation strategies, and design sustainable policies that account for various possible outcomes, contributing to a more resilient and environmentally conscious future.

Discretization is the process of converting a continuous random variable (RV) into a discrete RV by dividing it into equal intervals. This step is important in many data analysis and modeling applications for several reasons. Discretizing continuous RVs can simplify the data and make it easier to understand. Some statistical models, such as linear regression, assume continuous RVs, while others, such as logistic regression, assume categorical or discrete RVs. Discretization can capture non-linear relationships between RVs, reduce the dimensionality of data, and make it easier to visualize and analyze. Additionally, discretization can help handle

[*] Corresponding author e-mail: m.elmorshedy@psau.edu.sa

outliers by transforming them into a smaller number of intervals. Discrete probability distributions are foundational in statistics and probability theory, enabling data modeling, outcome analysis, and prediction. They are especially useful in fields like finance, business, marketing, medicine, engineering, and economics, where decisions often rely on limited data. For instance, in finance, these distributions model the probabilities of stock prices and market movements, helping investors make informed decisions. In business, they model customer behavior and the likelihood of purchases, guiding marketing strategies, inventory planning, and customer relationship management. In engineering, discrete probability distributions assess the reliability of designs over time, allowing engineers to model failure probabilities and enhance system safety. Due to the significance of these probabilistic models, numerous statisticians have developed various discrete distributions for data modeling. Several statistical methods exist for deriving a discrete probability model, including survival function discretization, Poisson mixing, binomial mixing, the T-geometric family of discrete distributions (see [1]), and utilizing the hazard rate function and its inverse.

In our case, we focus on the Kumaraswamy (Ku) and discrete inverse Weibull (DIW) models. The cumulative distribution functions (CDFs) of the Ku and DIW models can be formualted as

$$H(x;\mu,\sigma) = 1-(1-x^{\mu})^{\sigma};\ 0<x<1 \tag{1}$$

and

$$G(z;\alpha,\beta) = e^{-\left[\frac{\alpha}{z+1}\right]^{\beta}};\ z=0,1,2,3,.... \tag{2}$$

where $\mu>0$, $\sigma>0$, $\beta>0$ are the shape parameters whereas $\alpha>0$ is the scale parameter. For more information on the Ku and DIW models, see [2, 3], respectively. Recognizing the importance of discrete extensions of the Ku model, numerous researchers have developed and studied various extensions to model and analyze different data types across multiple fields. For instance, [4, 5] introduced the Kumaraswamy-geometric (KuGeo) distribution with integer support ranging from 0 to ∞. [6, 7] derived the discrete inverted Kumaraswamy (DIKu) distribution. [8] proposed a discrete Kumaraswamy Marshall-Olkin exponential distribution. [9] developed the discrete Kumaraswamy Erlang-truncated exponential distribution, also with integer support from 0 to ∞, among others. Discretization is a crucial step in data pre-processing, and it must be handled carefully to avoid omitting important information or introducing bias. The choice of the number of intervals and the discretization technique can significantly impact the results of the analysis. While many researchers have derived and studied numerous continuous probability distributions, they often do so by discretizing them. This approach has become a dominant trend in the literature, despite the lack of extensive works in this vital area of distribution theory. The importance of discretizing continuous models stems from the fact that much valuable data in fields such as medicine, engineering, and actuarial science cannot be effectively analyzed using continuous models alone. This necessity drives researchers to focus on developing discrete models.

In this context, a new discrete elastic extension of the Ku and DIW models has been developed and thoroughly discussed, resulting in what is known as the KuDIW distribution. Our motivations for introducing the KuDIW model can be summarized as follows: The KuDIW model can be applied to model asymmetric count data, analyze extreme and outlier observations, discuss dispersed data, explain different forms of kurtosis, assess various forms of failure and risk, and model heavy-tailed real data.

The remainder of the paper is structured as follows: Section 2 introduces the KuDIW model. In Section 3, various distribution statistics are derived. Section 4 details the parameter estimation of the KuDIW model using the maximum likelihood technique. A simulation study is presented in Section 5. Section 6 analyzes three real datasets to demonstrate the capability and applicability of the KuDIW model. Finally, Section 7 offers conclusions and suggestions for future research.

2 The KuDIW Model

The RV X is said to have the KuDIW distribution if its CDF can be listed as

$$F(z) = 1-\left(1-\left(e^{-\left[\frac{\alpha}{z+1}\right]^{\beta}}\right)^{\mu}\right)^{\sigma}, \tag{3}$$

where $z=1,2,3,...$ and $\mu,\sigma,\alpha,\beta>0$. The corresponding probability mass function (PMF) of Equation (3) can be formulated as

$$f(z) = \left(1-\left(e^{-\left[\frac{\alpha}{z}\right]^{\beta}}\right)^{\mu}\right)^{\sigma} - \left(1-\left(e^{-\left[\frac{\alpha}{z+1}\right]^{\beta}}\right)^{\mu}\right)^{\sigma}. \tag{4}$$

Conversely, in reliability theory, an essential statistical concept known as the hazard rate function (HRF) warrants study. This concept has demonstrated its significance in the maintenance process, particularly in the engineering field. The importance of the HRF extends beyond engineering, encompassing applications in insurance, medicine, and economics. The HRF can be expressed as follows

$$h(z) = 1-\left(1-\left(e^{-\left[\frac{\alpha}{z+1}\right]^{\beta}}\right)^{\mu}\right)^{\sigma}\left(1-\left(e^{-\left[\frac{\alpha}{z}\right]^{\beta}}\right)^{\mu}\right)^{-\sigma}. \tag{5}$$

Suppose C and D are two independent KuDIW RVs with parameters $(\mu_1,\sigma_1,\alpha_1,\beta_1)$ and $(\mu_2,\sigma_2,\alpha_2,\beta_2)$, respectively. Then, the HRF of $M=\min(C,D)$ can be expressed as

$$h_M(z;\Phi) = \frac{\Pr(\min(C,D)=z)}{\Pr(\min(C,D)\geq z)} = \frac{\Pr(\min(C,D)\geq z) - \Pr(\min(C,D)\geq z+1)}{\Pr(\min(C,D)\geq z)}$$

then

$$h_M(z;\Phi) = \Upsilon_1(z;\Phi) - \Upsilon_2(z;\Phi);\ z = 1,2,3,....,$$

where $\Phi = (\mu_1, \sigma_1, \alpha_1, \beta_1, \mu_2, \sigma_2, \alpha_2, \beta_2)$ and

$$\Upsilon_1(z;\Phi) = h(z;\mu_1,\sigma_1,\alpha_1,\beta_1) + h(z;\mu_2,\sigma_2,\alpha_2,\beta_2)$$
$$\Upsilon_2(z;\Phi) = h(z;\mu_1,\sigma_1,\alpha_1,\beta_1)\, h(z;\mu_2,\sigma_2,\alpha_2,\beta_2).$$

Similarly, the HRF of $K = \max(C,D)$ can be formulated as

$$h_K(z;\Phi) = 1 - \frac{1-\Theta_1(z;\Phi)}{1-\Theta_1(z-1;\Phi)}. \tag{6}$$

where

$$\Theta_1(z;\Phi) = F(z;\mu_1,\sigma_1,\alpha_1,\beta_1)\ F(z;\mu_2,\sigma_2,\alpha_2,\beta_2).$$

Figures 1 and 2 illustrate the PMF and HRF of the KuDIW distribution for specific parameter values.

As evident, the PMF can serve as a probabilistic model for analyzing and evaluating asymmetric data with a monomorphic form. Additionally, the associated failure function can be utilized to model a monotonically decreasing shape.

3 Distributional Properties

3.1 Moments

Assume the RV X have the KuDIW distribution, then the rth moment can be expressed as

$$\mu_r^{'} = \sum_{z=1}^{\infty} z^r f(z;\mu,\sigma,\alpha,\beta);\ r = 1,2,3,\ ...$$
$$= \sum_{z=1}^{\infty} [z^r - (z-1)^r] \left(1 - \left(e^{-\left[\frac{\alpha}{z}\right]^{\beta}}\right)^{\mu}\right)^{\sigma}. \tag{7}$$

The moment generating function (MGF) can be reported as

$$\mathrm{M}_Z(k) = \sum_{z=0}^{\infty}\sum_{i=0}^{\infty} \frac{(xk)^i}{i!} \begin{bmatrix} \left(1 - \left(e^{-\left[\frac{\alpha}{z}\right]^{\beta}}\right)^{\mu}\right)^{\sigma} \\ -\left(1 - \left(e^{-\left[\frac{\alpha}{z+1}\right]^{\beta}}\right)^{\mu}\right)^{\sigma} \end{bmatrix}, \tag{8}$$

where $k = 1,2,3....$ The rth moment can be derived from the MGF as $\mu_r^{'} = \frac{d^r}{dk^r}\mathrm{M}_Z(k)\,|_{k=0}$. Using Equation (7), the E(Z), Var(Z), Sk(Z), and Ku(Z) can be respectively given by

$$\mathrm{E}(Z) = \sum_{z=1}^{\infty} \left(1 - \left(e^{-\left[\frac{\alpha}{z+1}\right]^{\beta}}\right)^{\mu}\right)^{\sigma}, \tag{9}$$

$$\mathrm{Var}(Z) = \sum_{x=1}^{\infty} (2z-1) \left(1 - \left(e^{-\left[\frac{\alpha}{z}\right]^{\beta}}\right)^{\mu}\right)^{\sigma} - \mu_1^{'2}, \tag{10}$$

$$\mathrm{Sk}(Z) = \frac{\mu_3^{'} - 3\mu_2^{'}\mu_1^{'} + 2\mu_1^{'3}}{(Var(Z))^{3/2}}, \tag{11}$$

$$\mathrm{Ku}(Z) = \frac{\mu_4^{'} - 4\mu_1^{'}\mu_3^{'} + 6\mu_2^{'}\mu_1^{'2} - 3\mu_1^{'4}}{(Var(Z))^2}. \tag{12}$$

According to the first two moments, an important descriptive statistical concept is derived in the so-called index of dispersion (IoD) or variance-to-mean ratio can be effectively applied to the analysis and evaluation of actuarial data. The IoD of the KuDHLo distribution can be proposed as

$$\mathrm{IoD}(Z) = \frac{\sum_{z=1}^{\infty} (2z-1) \left(1 - \left(e^{-\left[\frac{\alpha}{z}\right]^{\beta}}\right)^{a}\right)^{b}}{\sum_{z=1}^{\infty} \left(1 - \left(e^{-\left[\frac{\alpha}{z+1}\right]^{\beta}}\right)^{a}\right)^{b}} - \sum_{z=1}^{\infty} \left(1 - \left(e^{-\left[\frac{\alpha}{z}\right]^{\beta}}\right)^{a}\right)^{b}. \tag{13}$$

After performing numerical computations, it has been observed that the model is suitable for analyzing asymmetric dispersion data across various forms of kurtosis.

3.2 Entropies

Entropy is a scientific concept in statistics and a measurable physical property often associated with states of chaos, randomness, or uncertainty. The term and concept are used across various fields, from classical thermodynamics where it was first recognized to the microscopic description of nature in statistical physics, and the principles of information theory. Entropy has extensive applications in weather science, physics, chemistry, economics, biological systems, cosmology, climate change, and information systems, including telecommunications. For more detailed properties and features of entropy, see [10]. This section discusses different types of entropy (En), such as Rényi entropy (REn), maximum entropy (XEn), Shannon entropy (SnEn), minimum entropy (MEn), and collision entropy (CoEn). All these types of entropy are applicable in information theory as they measure uncertain variability (see [11, 12]). For a random variable Z with the KuDIW

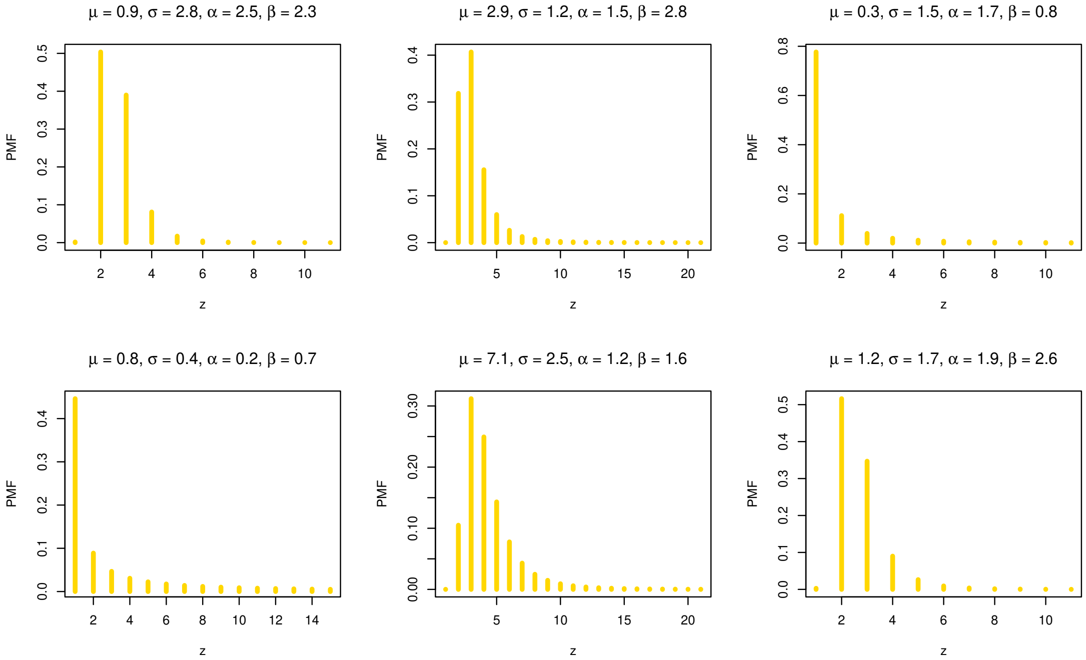

Fig 1. Various shapes for the PMF of the KuDIW distribution.

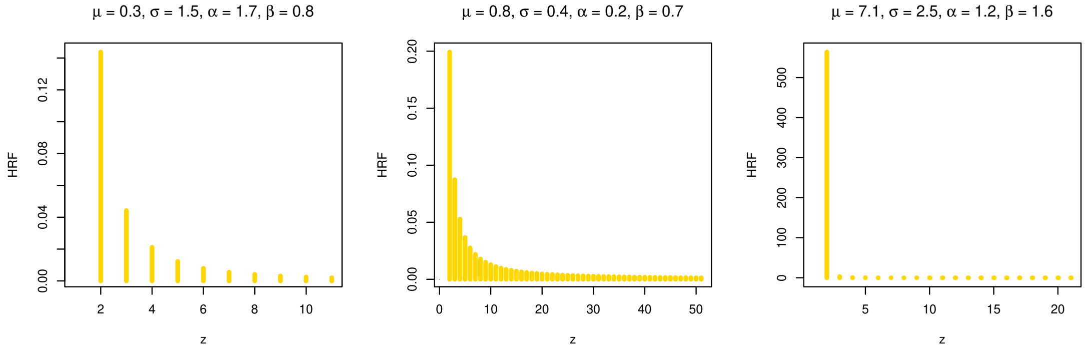

Fig 2. Various shapes for the HRF of the KuDIW model.

model, the REn is given as follows

$$I_{\sigma}(Z)=\frac{1}{1-\gamma}\log\sum_{z=0}^{\infty}\left[\begin{array}{c}\left(1-\left(e^{-\left[\frac{\alpha}{z}\right]^{\beta}}\right)^{\mu}\right)^{\sigma}\\ -\left(1-\left(e^{-\left[\frac{\alpha}{z+1}\right]^{\beta}}\right)^{\mu}\right)^{\sigma}\end{array}\right]^{\zeta}, \tag{14}$$

where $z\in\mathbb{N}_0$, $\gamma\in(0,1)$ and $\zeta\neq 1$. The REn generalizes the SnEn, CoEn, MEn, and XEn where the SnEn, CoEn, MEn, and XEn can be derived as a special case of the REn when $\gamma\longrightarrow 1$, $\gamma\longrightarrow 2$, $\gamma\longrightarrow\infty$, and $\gamma\longrightarrow 0$, respectively. Due to the complexity of obtaining a closed-form expression for En, numerical calculations are necessary to demonstrate the series' convergence and to explore its features. These computations reveal that the series converges and that En increases as the model parameters approach infinity.

3.3 Quantile function

In this section, we derive the quantile function (QF) of the KuDIW model. The QF has applications in various fields, particularly in hydrology for determining the levels of lakes and oceans. Additionally, it can be used to generate random samples for simulation purposes, among other applications. The uth QF, say z_u, of the KuDIW distribution is the solution of $F(z_u;\mu,\sigma,\alpha,\beta)-q=0$; $z_u>0$, then

$$z_u{=}\alpha\left(-\mu\ln\left[1-(1-u)^{\frac{1}{\sigma}}\right]\right)^{\frac{-1}{\beta}}-1, \tag{15}$$

where $u\in(0,1)$. Setting $u=0.5$, the median of the KuDIW distribution can be derived.

3.4 L-moment statistic

In statistics, the nth order statistic (OS) of a sample refers to its nth-smallest value, essentially its minimum. Order statistics, along with ranking statistics, are fundamental tools in nonparametric statistics and inference. For the RVs $Z_1,Z_{2,}...,Z_n$, the OS $Z_{1:n}\leq Z_{2:n}\leq...\leq Z_{n:n}$ are also RVs. Consider the RV Z have the KuDIW model, then the CDF of the ith OS can be listed as

$$\begin{aligned}F_{i:n}(z;\mu,\sigma,\alpha,\beta)&=\sum_{k=i}^{n}\binom{n}{k}\left[\begin{array}{c}[F_i(z;\mu,\sigma,\alpha,\beta)]^k\times\\ [1-F_i(z;\mu,\sigma,\alpha,\beta)]^{n-k}\end{array}\right]\\&=\sum_{k=i}^{n}\sum_{j=0}^{n-k}\sum_{l=0}^{k+j}\Phi_{(n,k)}^{(j,l)}S_i(z;\mu,\sigma l,\alpha,\beta),\end{aligned} \tag{16}$$

where

$$\Phi_{(n,k)}^{(j,l)}=(-1)^{j+l}\binom{n}{k}\binom{n-k}{j}\binom{k+j}{l}.$$

The corresponding PMF of the ith OS can be formulated as

$$\begin{aligned}f_{i:n}(z;.)&=F_{i:n}(z;.)-F_{i:n}(z-1;.)\\&=\sum_{k=i}^{n}\sum_{j=0}^{n-k}\sum_{l=0}^{k+j}\left[\Phi_{(n,k)}^{(j,l)}\left(\begin{array}{c}S_i(z;.)\\-S_i(z-1;.)\end{array}\right)\right],\end{aligned} \tag{17}$$

where $S_i(z;\mu,\sigma l,\alpha,\beta)$ is the survival function of the KuDIW model. Thus, the rth moment of $Z_{i:n}$ can be expressed as

$$\mathrm{E}(Z_{i:n}^r)=\sum_{x=0}^{\infty}\sum_{k=i}^{n}\sum_{j=0}^{n-k}\sum_{m=0}^{k+j}\left[\Phi_{(n,k)}^{(j,l)}z^r\left(\begin{array}{c}S_i(z;\mu,.)\\-S_i(z-1;.)\end{array}\right)\right]. \tag{18}$$

According to Equation (18), a significant descriptive statistic known as L-moments (L-M) is derived. L-moments can be used to summarize statistics for probability models (see [13] for more details). For the random variable Z, the L-moments are given by

$$\Omega_\alpha=\frac{1}{\lambda}\sum_{i-0}^{\lambda-1}(-1)^i\binom{\lambda-1}{i}\mathrm{E}(Z_{\lambda-i:\lambda}). \tag{19}$$

According to Equation (19), the mean, Sk, and Ku can be derived as mean = Ω_1, Sk = $\frac{\Omega_3}{\Omega_2}$, and Ku = $\frac{\Omega_4}{\Omega_2}$.

4 Maximum Likelihood Estimation

The maximum likelihood (ML) estimation technique is a widely used statistical method for estimating unknown parameters of a population. It is a robust approach for making inferences about data and has a variety of applications. ML estimation is based on the principle that the parameter to be estimated maximizes the likelihood of the observed data. This method is used to estimate population parameters such as the mean, standard deviation, and other characteristics. It is also commonly applied to determine the probability of specific events or outcomes given observed data. The technique works by maximizing the likelihood of the sample data, given assumed population parameters. This involves converting the sample data into a probability distribution, which is then used to calculate the probability of the observed data given the assumed parameters, ultimately identifying the most likely parameter values. In this segment, we derive the estimation of the KuDIW distribution parameters using the ML method based on a complete sample. Let $_1$, Z_2, ..., Z_n be a random sample from the KuDIW distribution. Then, the log-likelihood function, say $L(\mu,\sigma,\alpha,\beta|z_i)$, can be expressed as

$$L(\mu,\sigma,\alpha,\beta|z_i)=\sum_{i=1}^{n}\ln\left[\begin{array}{c}\left(1-\left(e^{-\left[\frac{\alpha}{z_i}\right]^\beta}\right)^\mu\right)^\sigma\\-\left(1-\left(e^{-\left[\frac{\alpha}{z_i+1}\right]^\beta}\right)^\mu\right)^\sigma\end{array}\right]. \tag{20}$$

To obtain the ML estimate of the model parameters, one approach is to maximize the likelihood function $L(\mu,\sigma,\alpha,\beta|z_i)$. Numerous numerical optimization methods are accessible in different programming platforms such as R and SAS, which can be employed to maximize$L(\mu,\sigma,\alpha,\beta|z_i)$. For instance, in R, the ‘optim‘ function and in SAS, the ‘PROC NLMIXED‘ procedure can be used for this optimization task.

5 Estimator Behaviour: Markov Chain Monte Carlo

Markov Chain Monte Carlo (MCMC) is a powerful simulation technique widely used for sampling from probability distributions in various contexts like Bayesian inference, optimization, and machine learning. Its key advantage lies in its ability to sample and estimate a distribution without requiring the calculation of the associated normalization constant, making it efficient and accurate for inference tasks. MCMC operates by constructing a Markov chain, which is a sequence of dependent random variables. Each sample in the chain depends on the previous one, ensuring that the samples follow the target distribution. The process begins with a random starting point and iteratively updates using a transition rule, specifying how the chain moves from one point to another. This iterative simulation continues until the chain reaches a steady state, converging to the target distribution. Simulation studies play a vital role in evaluating estimation techniques in statistics. They offer a comprehensive examination of different approaches across various scenarios. One common performance metric in simulation studies is the mean squared error (MSE), which measures the accuracy of a statistical model or estimator. MSE considers both bias and variance, unlike measures that solely focus on bias or variance alone, providing a more thorough assessment of estimation quality. The steps for conducting the simulation are as follows:

–Conduct a simulation by generating $N = 10000$ samples, each of size n, where n varies across 20, 80, 150, 300, and 500, from the KuDIW distribution using different sets of KuDIW parameters outlined in Table 1 and Table 2 where (Scheme I: $\mu = 0.3, \sigma = 0.6, \alpha = 0.8, \beta = 0.5$; Scheme II: $\mu = 0.9, \sigma = 1.2, \alpha = 0.3, \beta = 0.8$; Scheme III: $\mu = 1.1, \sigma = 1.2, \alpha = 0.1, \beta = 0.6$; Scheme IV: $\mu = 1.8, \sigma = 1.8, \alpha = 1.2, \beta = 1.6$).

–Evaluate the simulation based on bias and mean squared error (MSE), which serve as the performance criteria.

–Present the key findings numerically in Table 1 and Table 2, and visually in Figures 3-6.

–Observations reveal a decreasing trend in bias and MSE with increasing sample size n. This trend underscores the consistency property of the maximum likelihood (ML) approach across all parameter combinations for estimating the KuDIW parameters, highlighting its effective applicability.

6 Sustainability Data Analysis

This section delves into the implementations of the KuDIW distribution across three distinct real datasets. The competitive distributions can be listed as: Discrete inverse Weibull (DIW), discrete inverse Rayleigh (DIR), discrete Rayleigh (DR), three parameters discrete Lindley (DLi-III), t parameters discrete Lindley (DLi-II), one parameter discrete Lindley (DLi-I), negative binomial (NvBi), Poisson (Poi), discrete Pareto (Dpa), discrete log-logistic (DLogL), discrete Burr (DB), and discrete Burr-Hatke (DBH). The comparison among the fitted models is based on several criteria, including negative log likelihood ($-L$), Akaike information criterion (AIC), corrected Akaike information criterion (CAIC), Hannan-Quinn information criterion (HQIC), and Chi-square (χ^2). Information criteria are statistical tools used for model selection. They provide a means of balancing model fit with model complexity, thereby preventing overfitting. In essence, information criteria penalize the complexity of a model to ensure that simpler models are preferred if they explain the data adequately. The most common information criteria used in statistical modeling include the Akaike Information Criterion (AIC), the Corrected Akaike Information Criterion (CAIC), and the Hannan-Quinn Information Criterion (HQIC). Each of these criteria has its own specific formulation and use case, but they all serve the same fundamental purpose of model comparison and selection.

1. Akaike Information Criterion: The AIC is based on the concept of entropy, providing a relative measure of the information lost when a given model is used to represent the process that generated the data. The formula for AIC is:

$$\mathrm{AIC} = 2k - 2\ln(L)$$

where k is the number of parameters in the model, and L is the maximized value of the likelihood function for the model. A lower AIC value indicates a better-fitting model.

2. Corrected Akaike Information Criterion (CAIC): The CAIC is an extension of the AIC, adjusted for small sample sizes. This correction is important because the AIC can sometimes favor more complex models when the sample size is not large enough to justify them. The formula for CAIC is:

$$\mathrm{CAIC} = \mathrm{AIC} + \frac{2k(k+1)}{n-k-1}$$

where n is the sample size. The CAIC aims to provide a more accurate estimate of model quality for smaller samples.

3. Hannan-Quinn Information Criterion (HQIC): The HQIC is another criterion used for model selection, providing a more stringent penalty for model complexity compared to AIC and BIC. The formula for HQIC is:

$$\mathrm{HQIC} = -2\ln(L) + 2k\ln(\ln(n))$$

where n is the sample size, L is the maximized likelihood function, and k is the number of parameters. HQIC tends to favor simpler models more strongly, making it a useful criterion in certain contexts.

In practice, these criteria are often used together to compare models and select the one that offers the best balance of fit and simplicity. When different criteria suggest different models, analysts must consider the specific context and goals of their modeling task to make the final decision. We will juxtapose the KuDHLo distribution with other competitive models

6.1 Data set I

This dataset pertains to the count of European corn borer (ECB) larvae, also referred to as Pyrausta, observed in a field experiment detailed by [14]. The experiment entailed a random assessment of 8 hills, each with 15

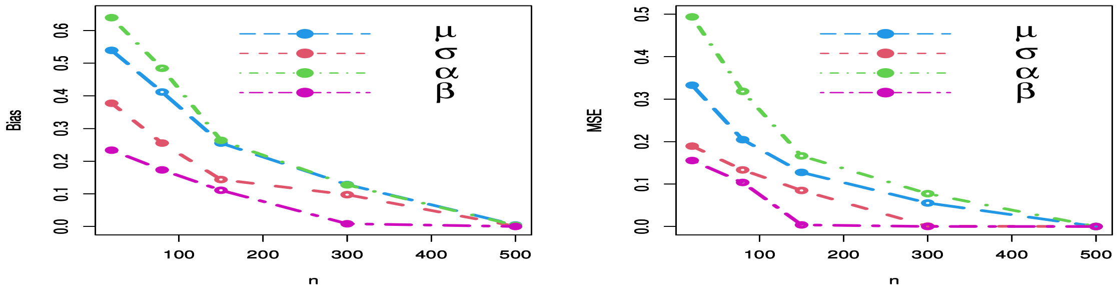

Fig 3. Simulation plots for scheme I.

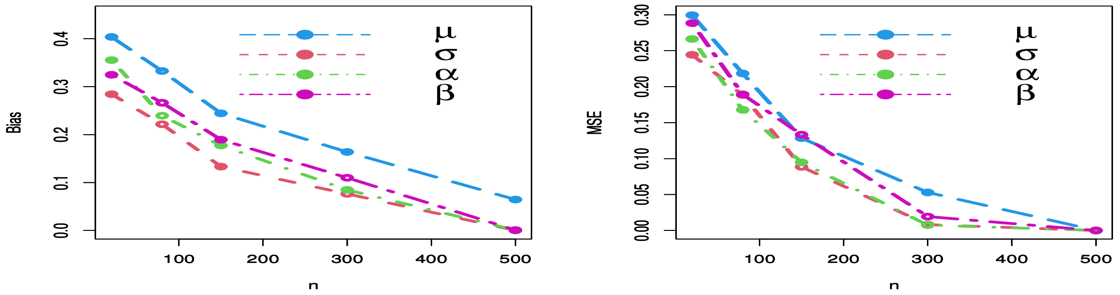

Fig 4. Simulation plots for scheme II.

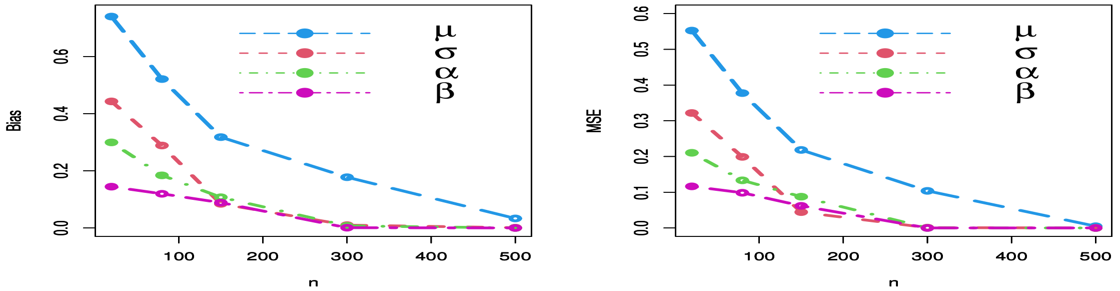

Fig 5. Simulation plots for scheme III.

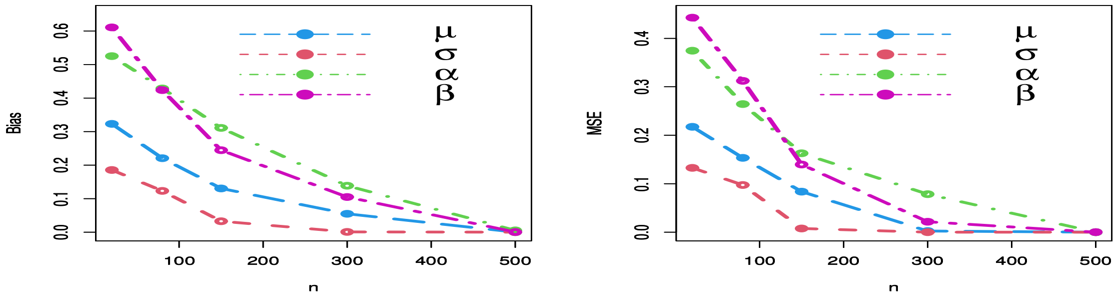

Fig 6. Simulation plots for scheme IV.

Table 1. Simulation results for the KuDIW parameters (Part I).

		Scheme I		Scheme II	
n	Parameter	Bias	MSE	Bias	MSE
μ	20	0.53934776	0.33273582	0.40376521	0.29947364
	80	0.41184635	0.20463381	0.33273583	0.21846394
	150	0.25539741	0.12747393	0.24437368	0.12846483
	300	0.12846879	0.05538621	0.16383644	0.05294631
	500	0.00467821	0.00008463	0.06453822	0.00008746
σ	20	0.37745088	0.18934653	0.28435238	0.24438459
	80	0.25548036	0.13329346	0.22162483	0.18849653
	150	0.14384622	0.08493613	0.13327469	0.08849632
	300	0.09746381	0.00083621	0.07638541	0.00814643
	500	0.00074552	0.00000083	0.00014774	0.00000876
α	20	0.63945232	0.49373897	0.35548682	0.26634832
	80	0.48452308	0.31836841	0.23947913	0.16789463
	150	0.26384268	0.16638936	0.17739464	0.09478585
	300	0.12745378	0.07736391	0.08465383	0.00748936
	500	0.00346357	0.00007376	0.00078467	0.00008645
β	20	0.23373950	0.15538362	0.32484625	0.28846391
	80	0.17332284	0.10373693	0.26649462	0.18936403
	150	0.11093878	0.00378931	0.18946452	0.13328469
	300	0.00836221	0.00007632	0.11001934	0.01936741
	500	0.00008682	0.00000008	0.00084538	0.00000347

Table 2. Simulation results for the KuDIW parameters (Part II).

		Scheme III		Scheme IV	
n	Parameter	Bias	MSE	Bias	MSE
μ	20	0.73964392	0.55243383	0.32287649	0.21745389
	80	0.52093478	0.37745392	0.22053972	0.15339643
	150	0.31777344	0.21835824	0.13038254	0.08364863
	300	0.17749362	0.10376498	0.05483532	0.00237648
	500	0.03324745	0.00483096	0.00008473	0.00000017
σ	20	0.44274782	0.32173584	0.18539638	0.13274834
	80	0.28845491	0.19894663	0.12304583	0.09735856
	150	0.17735394	0.08428469	0.03263894	0.00763674
	300	0.05539201	0.00984764	0.00083664	0.00000847
	500	0.00087482	0.00000437	0.00000746	0.00000029
α	20	0.29936481	0.21034734	0.52483695	0.37453883
	80	0.18396434	0.13327459	0.42846309	0.26438649
	150	0.10746539	0.08746381	0.31084664	0.16285834
	300	0.00735482	0.00007481	0.13846332	0.07846383
	500	0.00007367	0.00000054	0.0053864	0.00000464
β	20	0.14439469	0.11638455	0.61047654	0.44273538
	80	0.11947644	0.09847572	0.42349452	0.31228358
	150	0.08846372	0.06184638	0.24438455	0.13964784
	300	0.00083774	0.00000548	0.10485934	0.02150468
	500	0.00000084	0.00000006	0.00048763	0.00007914

replications, during which the observer recorded the number of borers per corn hill. The summary statistics for this dataset are as follows: the mean is 1.326, the variance is 3.669, the skewness is 1.976, and the kurtosis is 8.984. Figure 7 illustrates non-parametric infographics to analyze the behavior of dataset I. It is evident that the data exhibits right skewness, includes outliers, and displays a decreasing hazard rate function (HRF) shape. Table 3 lists the maximum likelihood estimates (MLEs) for the relevant parameter(s), accompanied by their respective standard errors and confidence intervals. Additionally, this table showcases the results of the goodness-of-fit test for dataset I. At a significance level of 0.05, it's evident that both the DIW and DB-XII distributions perform satisfactorily, alongside the KuDIW distribution. However, the KuDIW distribution emerges as the top-performing model among all tested options. Figure 8 displays the observed and expected PMFs for dataset I, affirming the superiority of the KuDIW model compared to other tested models.

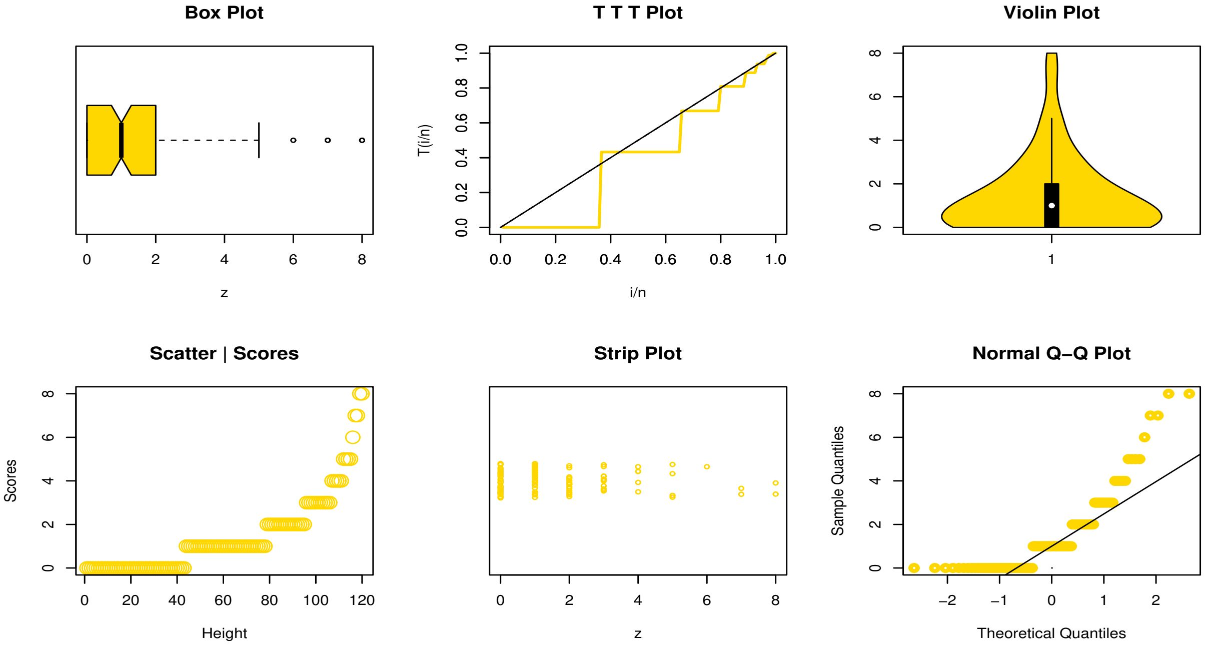

Fig 7. Non-parametric infographics for data set I.

6.2 Data set II

This dataset records the counts Kidney stones (a biochemical marker) in individuals based on steroids, as detailed by [15]. Before proceeding with an analysis of the CK data, it is essential to visually inspect the data to grasp its inherent characteristics. To achieve this, non-parametric plots were generated, as depicted in Figure 9. The visualization reveals a positively skewed distribution of the data, with some extreme observations present. This initial exploration through visualization plays a pivotal role in understanding the data's distribution, identifying outliers, and uncovering any anomalous patterns before delving into more extensive statistical analyses. The ML estimators of the tested models as well as goodness-of-fit test are listed in Table 4.

Both the KuDIW and DIW models demonstrate effective performance in modeling dataset II, with the KuDIW model standing out as the top-performing model among the competitors. The outcomes listed in Table 4 are depicted in Figure 10.

6.3 Data set III

The data retrieved from the Worldometer website presents the daily new deaths attributed to COVID-19 in South Korea, spanning from February 15, 2020, to December 12, 2020 (source: *https://www.worldometers.info/coronavirus/country/south-korea/*). To effectively analyze and interpret COVID-19 data, several key metrics and concepts are typically considered, including confirmed cases (total positive tests broken down by demographics), active cases (currently infected individuals), recoveries (individuals who have recovered based on regional criteria), deaths (fatalities attributed to the virus), testing rates (number of tests conducted), positivity rate (percentage of positive tests), hospitalizations (number of hospitalized individuals), vaccination rates (number of vaccinated individuals by dose), variants (circulating COVID-19 variants and their impact on transmission and vaccine efficacy), and public health measures (interventions like mask mandates and lockdowns), with data analysis often involving tracking these metrics over time to identify trends, assess the effectiveness of interventions, and make informed decisions about resource allocation and policy measures, using visualization tools like graphs and charts for clear presentation. In this analysis, we derived the ML estimators for the parameters of the KuDIW distribution based on the this data. To delve deeper into the characteristics of this dataset, non-parametric plots were created, as depicted in Figure 11. The visualization highlights a positively skewed distribution in the data, along with the presence of some extreme observations.

The ML estimators of the competitive distributions as well as goodness-of-fit are discussed in Table 5. The KuDIW distribution stands out as the top-performing model among all the models tested. This conclusion is reinforced by Figure 12, which corroborates the findings listed in Table 5.

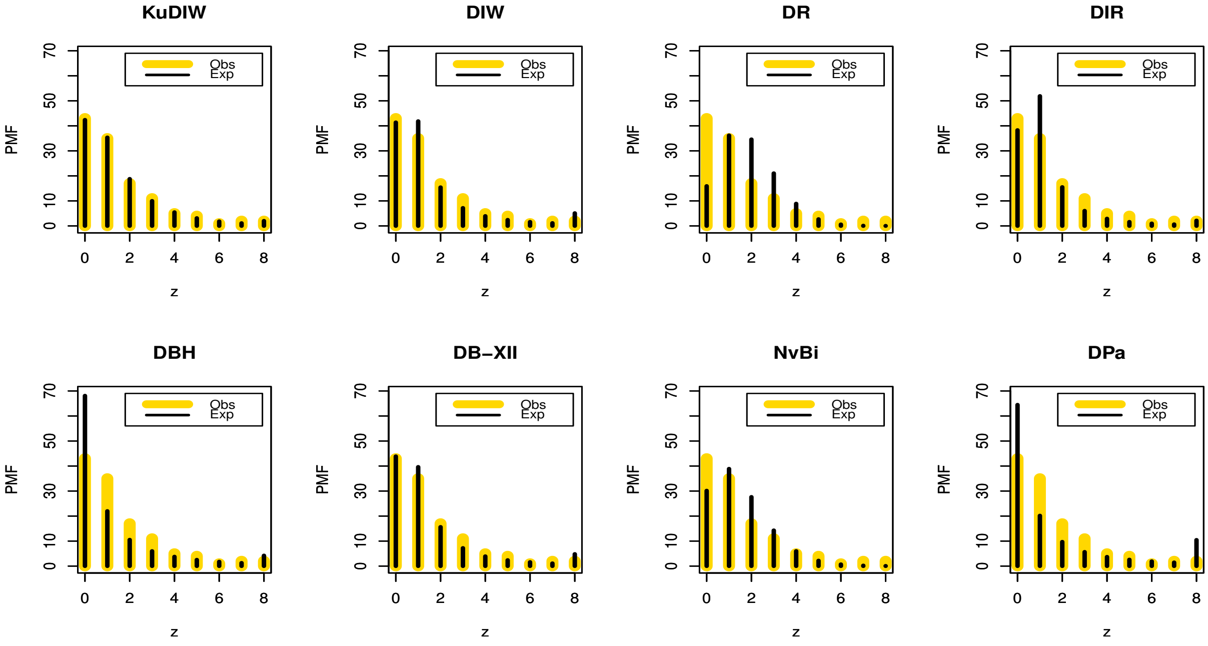

Fig 8. The observed and expected PMFs for data set I.

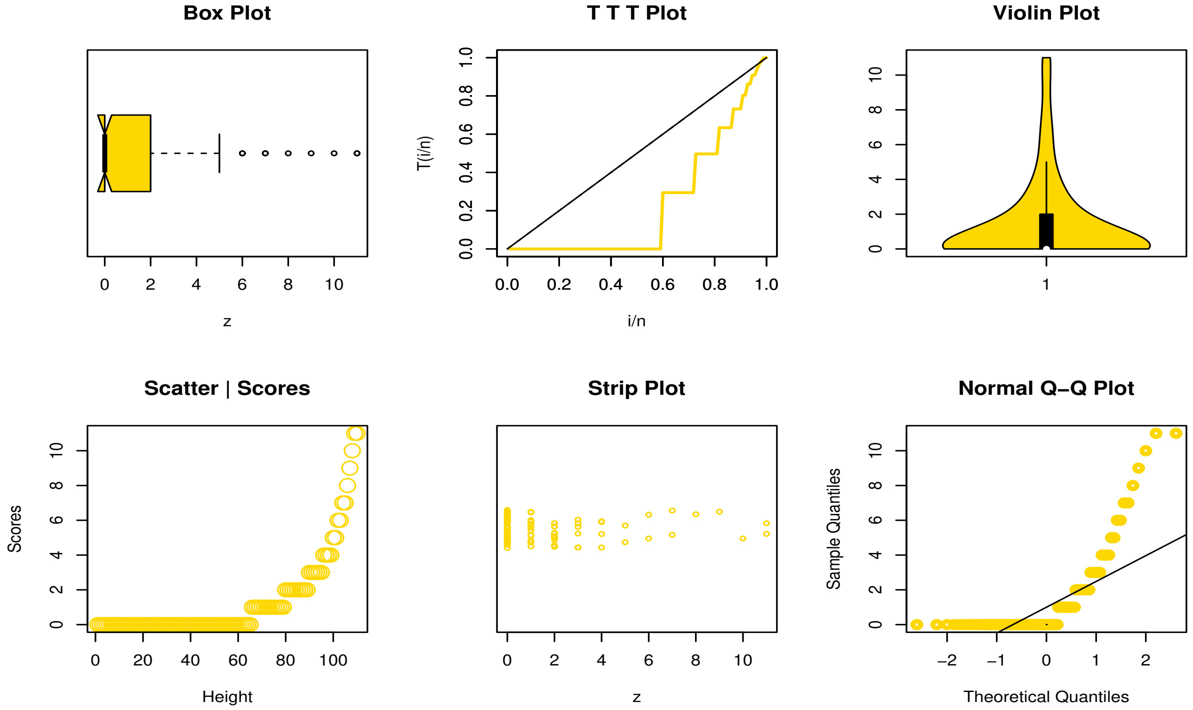

Figure 9. Non-parametric plots for data set II.

Table 3. The ML and goodness-of-fit test for data set I.

		ExFr							
X	ObFr	**KuDIW**	**DIW**	**DR**	**DIR**	**DBH**	**DB-XII**	**NvBi**	**DPa**
0	43	42.38	41.37	15.92	38.28	68.07	43.84	30.12	64.45
1	35	35.34	41.85	36.17	51.90	21.97	39.61	38.87	20.15
2	17	18.79	15.42	34.58	15.51	10.51	15.62	27.61	9.69
3	11	9.98	7.17	21.03	6.04	5.98	7.20	14.26	5.65
4	5	5.47	3.94	8.89	2.91	3.75	3.91	5.99	3.68
5	4	3.11	2.42	2.70	1.61	2.51	2.37	2.17	2.58
6	1	1.82	1.61	0.60	0.98	1.75	1.59	0.70	1.90
7	2	1.10	1.13	0.09	0.64	1.26	1.09	0.21	1.46
8	2	2.01	5.09	0.02	2.14	4.20	4.80	0.06	10.44
Total	**120**	**120**	**120**	**120**	**120**	**120**	**120**	**120**	**120**
$\widehat{\mu}$		25.356	–	–	–	–	–	0.870	–
$\widehat{\sigma}$		112.417	–	–	–	–	–	9.956	–
$\widehat{\alpha}$		0.002	0.345	0.867	0.319	0.865	0.519	–	0.329
$\widehat{\beta}$		0.246	1.541	–	–	–	2.358	–	–
$-L$		200.379	204.812	235.227	208.439	214.053	204.293	211.522	220.618
AIC		408.758	413.624	472.453	418.878	430.106	412.587	427.056	443.236
CAIC		409.106	413.727	472.487	418.912	430.139	412.689	427.143	443.270
HQIC		413.286	415.888	473.585	420.010	431.238	414.851	427.086	444.368
χ^2		0.442	5.511	60.059	14.274	27.05	4.664	20.367	32.462
DF		1	3	3	4	2	3	3	4
P-value		0.506	0.138	< 0.0001	< 0.0001	< 0.0001	0.198	0.0001	< 0.0001

Table 4. The ML and goodness-of-fit test for data set II.

		ExFr							
X	ObFr	**KuDIW**	**DIW**	**DLi-I**	**Poi**	**DLi-II**	**DLi-III**	**DR**	**DIR**
0	65	63.39	63.91	40.29	27.39	46.03	46.01	10.89	60.89
1	14	18.94	20.69	29.83	38.08	26.79	26.77	26.62	33.99
2	10	9.13	8.05	18.36	26.47	15.57	15.57	29.45	8.12
3	6	5.25	4.23	10.34	12.26	9.05	9.06	22.29	3.00
4	4	3.33	2.56	5.52	4.26	5.27	5.27	12.63	1.42
5	2	2.25	1.75	2.85	1.19	3.06	3.07	5.54	0.78
6	2	1.59	1.26	1.44	0.27	1.78	1.78	1.91	0.47
7	2	1.17	0.95	0.71	0.05	1.04	1.04	0.53	0.31
8	1	0.88	0.74	0.35	0.00	0.60	0.60	0.12	0.21
9	1	0.68	0.59	0.17	0.00	0.35	0.35	0.02	0.15
10	1	0.53	0.48	0.08	0.00	0.20	0.20	0.00	0.11
11	2	2.86	4.79	0.06	0.03	0.26	0.28	0.00	0.55
Total	110	110	110	110	110	110	110	110	110
$\widehat{\mu}$		10.504	–	–	1.390	–	0.581	–	–
$\widehat{\sigma}$		72.839	–	–	–	–	358.728	–	–
$\widehat{\alpha}$		0.005	1.049	0.436	–	0.581	0.001	0.901	0.554
$\widehat{\beta}$		0.1615	0.581	–	–	0.001	–	–	–
$-L$		169.284	172.935	189.110	246.210	178.767	178.767	277.778	186.547
AIC		346.569	349.870	380.220	494.420	361.534	363.533	557.556	375.094
CAIC		346.949	349.982	380.257	494.457	361.646	363.759	557.593	375.131
HQIC		350.950	352.060	381.316	495.515	363.724	366.819	558.651	376.189
χ^2		1.767	6.445	34.635	89.277	19.091	19.096	306.515	40.456
DF		1	3	4	3	3	2	4	2
P-value		0.185	0.092	< 0.001	< 0.001	0.0003	< 0.0001	< 0.001	< 0.001

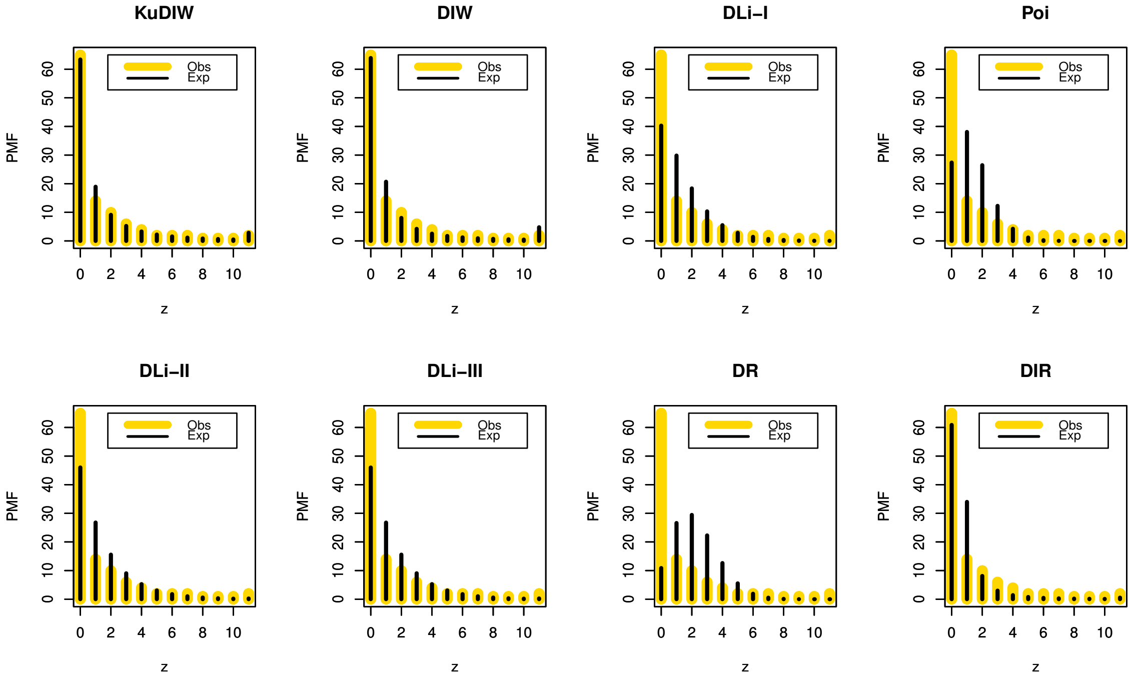

Figure 10. The observed and expected PMFs for data set II.

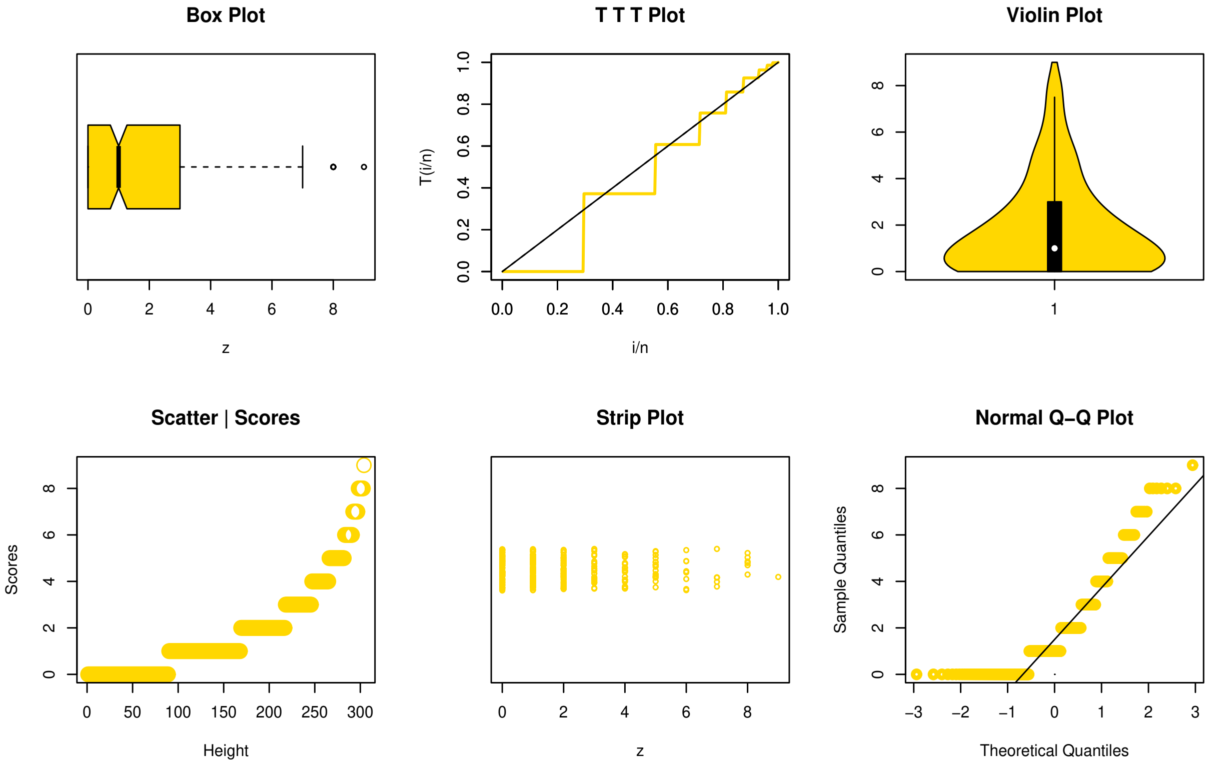

Fig 11. Non-parametric plots for data set III.

Table 5. The ML and goodness-of-fit test for data set III.

		ExFr							
X	ObFr	**KuDIW**	**DIW**	**DIR**	**DLogL**	**DBH**	**DPa**	**DB**	**Poi**
0	89	84.98	82.35	69.89	80.93	166.60	149.39	92.89	45.41
1	79	83.99	103.70	140.61	92.78	54.59	50.50	97.79	86.34
2	49	52.32	44.39	47.69	51.43	26.67	25.48	42.68	82.07
3	29	31.35	22.32	19.13	27.34	15.54	15.38	21.17	52.01
4	19	18.89	12.93	9.33	15.56	10.02	10.31	12.17	24.73
5	17	11.57	8.25	5.19	9.52	6.89	7.39	7.75	9.40
6	9	7.21	5.64	3.16	6.19	4.95	5.53	5.31	2.98
7	6	4.58	4.05	2.09	4.25	3.68	4.41	3.83	0.81
8	6	2.96	3.03	1.43	3.06	2.81	3.47	2.88	0.19
9	1	6.15	17.34	5.48	12.94	12.25	32.14	17.53	0.06
Total	304	304	304	304	304	304	304	304	304
$\widehat{\mu}$		20.239	0.271	0.229	1.716	0.904	0.377	0.591	1.901
$\widehat{\sigma}$		258.176	1.411	–	1.878	–	–	2.466	–
$\widehat{\alpha}$		0.005	–	–	–	–	–	–	–
$\widehat{\beta}$		0.210	–	–	–	–	–	–	–
$-L$		567.207	586.855	606.870	577.011	620.466	633.531	587.652	621.098
AIC		1142.414	1177.711	1215.740	1158.023	1242.932	1269.061	1179.304	1244.195
CAIC		1142.548	1177.751	1215.754	1158.063	1242.945	1269.075	1179.344	1244.208
HQIC		1148.362	1180.684	1217.227	1160.997	1244.419	1270.548	1182.278	1245.682
χ^2		4.785	41.868	92.204	25.019	109.333	128.631	44.784	115.896
DF		3	6	6	6	6	6	6	4
P-value		0.188	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001

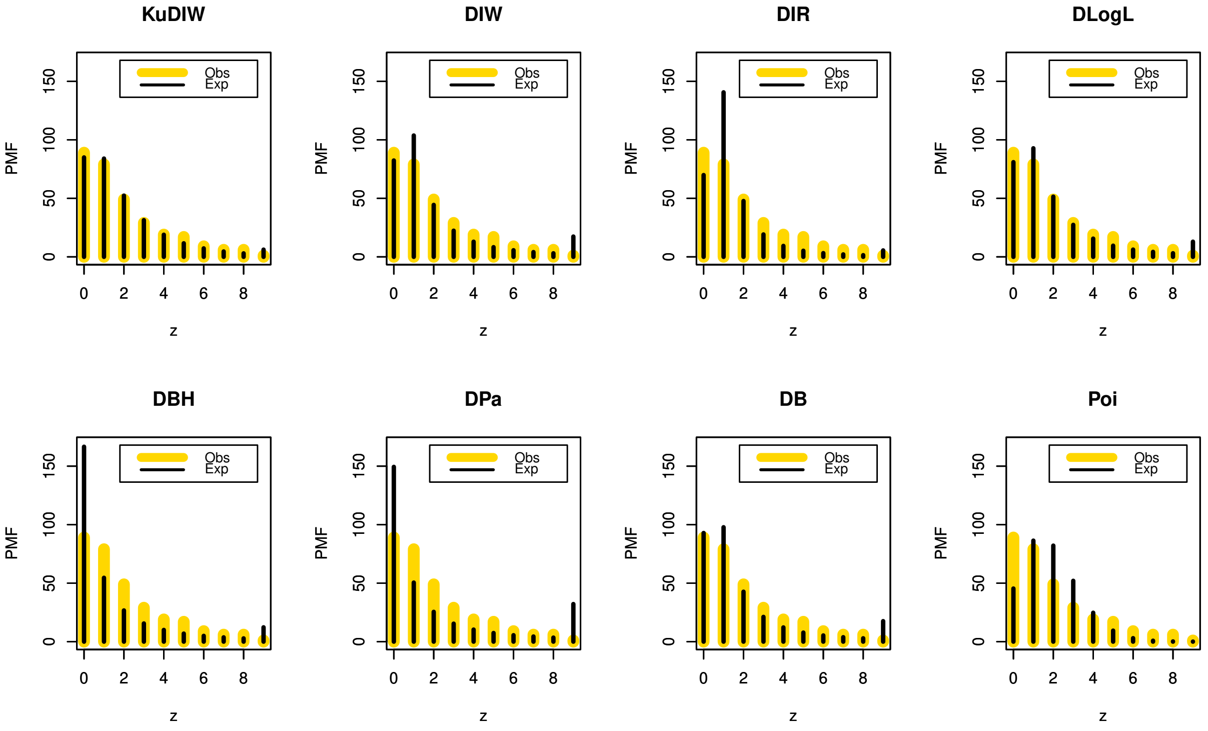

Fig 12. The observed and expected PMFs for data set III.

7 Concluding Observations and Future Work

In this article, a novel probabilistic mass function was introduced for analysis over the range $(0,\infty)$. Through a detailed examination of its statistical properties, it was observed that this discrete model was well-suited for evaluating monotonically decreasing hazard rate functions. Moreover, the proposed probabilistic mass function emerged as a robust choice for modeling positively skewed data across various kurtosis shapes. These capabilities, however, were just part of the model's strengths, as it also extended to enhancing the dispersion of real data. To complete the framework of this study, a maximum-likelihood approach was employed to derive optimal estimators for analyzing real-world data. Additionally, a simulation scheme was discussed to validate the efficacy of these estimators. Furthermore, the article explored three applications of real count sustainability data, showcasing the versatility and significance of this new discrete distribution. Looking ahead, future work would delve into univariate and bivariate extensions of time series using the proposed model, particularly for forecasting studies.

Acknowledgement: This project is sponsored by Prince Sattam Bin Abdulaziz University (PSAU) as part of funding for its SDG Roadmap Research Funding Programme project number PSAU-2023- SDG-24

References

[1] A. Alzaatreh, C. Lee, and F. Famoye. On the discrete analogues of continuous distributions, *Statistical Methodology*, **9(6)**, 589-603 (2012).

[2] P. Kumaraswamy. A generalized probability density function for double-bounded random processes, *Journal of Hydrology*, **46(1-2)**, 79-88 (1980).

[3] M. A. Jazi, C. D. Lai, and M. H. Alamatsaz. A discrete inverse Weibull distribution and estimation of its parameters, *Statistical Methodology*, **7(2)**, 121-132 (2010).

[4] Akinsete, F. Famoye, and C. Lee. The Kumaraswamy-geometric distribution, *Journal of Statistical Distributions and Applications*, **1***(1)*, 1-21 (2014).

[5] A. E. Abd EL-Hady, M. A. Hegazy, & A. A. EL-Helbawy. Discrete Exponentiated Generalized Family of Distributions, *Computational Journal of Mathematical and Statistical Sciences*, **2(2)**, 303-327 (2023).

[6] A. El-Helbawy, M. A. Hegazy, G. R. Al-Dayian, and R. E. Abd EL-Kader. A discrete analog of the inverted Kumaraswamy distribution: properties and estimation with application to COVID-19 data, *Pakistan Journal of Statistics and Operation Research*, **18(1)**, 297-328 (2022).

[7] M. Hegazy, R. Abd EL-Kader, A. D. Gannat, and A. A. A. EL-Helbawy. Discrete Inverted Kumaraswamy Distribution: Properties and Estimation, *Pakistan Journal of Statistics and Operation Research*, **18**, 297-328 (2022).

[8] J. Gillariose, L. Tomy, F. Jamal, and C. Chesneau. A Discrete Kumaraswamy Marshall-Olkin Exponential Distribution, *Journal of the Iranian Statistical Society*, **20***(2)*, 129-152 (2022).

[9] H. Eledum, and A. R. El-Alosey. Discrete Kumaraswamy Erlang-truncated exponential distribution with applications to count data, *Journal of Statistics Applications and Probability*, **12(2)**, 725-739 (2023).

[10] Wehrl. General properties of entropy, *Reviews of Modern Physics*, **50(2)**, 221 (1978).

[11] Rényi. On measures of entropy and information, *Mathematical Statistics and Probability*, **1**, 547-561 (1961).

[12] A. G. Abubakari, L. Anzagra, & S. Nasiru. Chen Burr-Hatke exponential distribution: Properties, regressions and biomedical applications, *Computational Journal of Mathematical and Statistical Sciences*, **2(1)**, 80-105 (2023).

[13] J. R. M. Hosking, and J. R. Wallis. Regional frequency analysis, **240** (1997).

[14] W. Bodhisuwan, and S. Sangpoom. The discrete weighted Lindley distribution, Proceedings of the International Conference on Mathematics, Statistics, and Their Applications, *Banda Aceh, Indonesia*, **5**, 4-6 (2016).

[15] S. K. Chan, P. R. Riley, K. L. Price, F. McElduff, P. J. Winyard, S. J. Welham, and D. A. Long. Corticosteroid-induced kidney dysmorphogenesis is associated with deregulated expression of known cystogenic molecules, as well as Indian hedgehog, *American Journal of Physiology-Renal Physiology*, **298(2)**, F346-F356 (2010).