Pakistan Journal of Statistics and Operation Research

A New Unit Distribution Based On The Unbounded Johnson Distribution Rule: The Unit Johnson S_U Distribution

Selim Gündüz [1], Mustafa Ç. Korkmaz [2*]

* Corresponding Author

1. Department of Business, Adana Alparslan Türkeş Science and Technology University, Adana, Turkey, sgunduz@atu.edu.tr
2. Department of Measurement and Evaluation, Artvin Çoruh University, Artvin, Turkey, mustafacagataykorkmaz@gmail.com

Abstract

This paper proposes a new probability distribution, which belongs a member of the exponential family, defined on (0,1) unit interval. The new unit model has been defined by relation of a random variable defined on unbounded interval with respect to standard logistic function. Some basic statistical properties of newly defined distribution are derived and studied. The different estimation methods and some inferences for the model parameters have been derived. We assess the performance of the estimators of these estimation methods based on the three different simulation scenarios. The analysis of three real data examples which one is related to the coronavirus data, show better fit of proposed distribution than many known distributions on the unit interval under some comparing criteria.

Key Words: Unit Distribution; Johnson's Distributions; Point Estimation; Beta Distribution; Exponential Family; Coronavirus Data; COVID-19; Recovery Rate; Burr Modeling.

Mathematical Subject Classification: 60E05, 62F10

1 Introduction

Johnson (1949) has pioneered a general system of the continuous distribution based on some transformations of the standard normal random variable (rv). The Johnson system distributions have been developed as follows. Let Z be a standard normal rv and define the transformation:

$$Y = g^{-1}\left(\frac{Z-\mu}{\sigma}\right),$$

where $g^{-1}(\cdot)$ is the any suitable function and $\mu \in \Re$ and $\sigma > 0$ are the parameters. Three important systems such as log-normal, unbounded (unbounded support) and bounded (unit support) have been defined by the author. The support of the system depends on the selection of the $g^{-1}(\cdot)$ function. Following g functions have been used $g(y) = \log y$, $g(y) = \log[y/(1-y)]$ and $g(y) = \sinh^{-1}[y] = \log[y + \sqrt{1+y^2}]$ for the log-normal (Johnson S_L), bounded (Johnson S_B) and unbounded (Johnson S_U) systems respectively by the author, where $\sinh^{-1}[t], t \in \Re$ is the inverse of the hyperbolic sine function. Johnson has also studied these systems in detail.

The probability density function (pdf) of the unbounded Johnson distribution system, Johnson S_U, is given by

$$f_{S_U}(y, \mu, \sigma) = \frac{\sigma}{\sqrt{1+y^2}} \phi(\mu + \sigma \sinh^{-1}[y]), -\infty < y < \infty, \tag{1}$$

where, the $\phi(\cdot)$ is the pdf of the standard normal distribution, $\sinh^{-1}[t] = \log\left(t + \sqrt{1 + t^2}\right), t \in \Re$ is the inverse of the hyperbolic sine function, and $\mu \in \Re$ and $\sigma > 0$ are the shapes parameters. The corresponding cumulative distribution function (cdf) is written as

$$F_{S_U}(y, \mu, \sigma) = \Phi(\mu + \sigma \sinh^{-1}[y]), \tag{2}$$

where $\Phi(\cdot)$ is the cdf of the standard normal distribution. The median value of this model is equal to $-\sinh\left(\frac{\mu}{\sigma}\right)$, where $\sinh(z) = 0.5(e^x - e^{-x})$ for $z \in \Re$ is the hyperbolic sine function. The model is unimodal and the mode value is between median and zero value. When μ is positive the mode is greater than the median, implying negative skewness, and vice versa. Many others characteristics of the Johnson S_U distribution can be found in Johnson (1949, 1955). For Johnson S_U distribution, a estimation procedure of the moments method based on the Fortran algorithm has been studied by Hill et al. (1976). Slifker and Shapiro (1980) introduced a selection rule to give estimates of the Johnson parameters. Wheeler (1980) has compared the quantile estimators with the moment estimators for the Johnson system distributions. Castagliola (1998) has presented a methodology which suggests to fit the normal sample median distribution with a symmetrical Johnson S_U distribution having approximatively the same second and fourth moments. Chen and Kamburowska (2001) have concerned with selection of a distribution from the Johnson system by matching percentiles as well as a modified selection procedure has been proposed by the authors. Tuenter (2001) have described an algorithm to fit a Johnson S_U system by moment matching. A analytic proof of the completeness and uniqueness of the skewness and kurtosis of the Johnson S_U distribution has been presented by Dudewicz et al. (2004). A new approach to estimate the four parameters of the Johnson systems has been introduced by George and Ramachandran (2011). Venkataraman and Rao (2016) have proposed Johnson S_U distribution in conjunction with ARMA-GARCH model based on superior out-sample value-at-risk prediction. Van Dorp and Jones (2020) have pointed out the behavior of Johnson's density functions, and their skewness and kurtosis with graphical and limiting perspectives in detail.

On the other hand, the unit distributions, defined on (0,1) support, is applied to model the behavior of random variables limited to intervals of (0,1) length. They are useful to model the percentages and proportions. No doubt, the beta distribution is one the first distribution that comes to mind to model the percentages and proportions since it is a convenient and useful model in many areas of statistics. However, its data modeling ability may be insufficient to explain the data. So, there are some alternative distributions, defined on unit interval, to beta distribution in the literature such as Johnson S_B, Topp-Leone (Topp and Leone, 1955), Johnson S_B' (Johnson, 1955), unit gamma (Consul and Jain, 1971), Kumaraswamy (Kumaraswamy, 1980), arcsine (Arnold and Groeneveld, 1980), unit logistic (Tadikamalla and Johnson, 1982), generalized beta type I distribution (McDonald, 1984), simplex (Barndorff-Nielsen and Jorgensen, 1991), standart two-sided power distribution (van Dorp and Kotz 2002), Mc arcsine (Cordeiro and Lemonte, 2014), log-Lindley (Gomez-Deniz et al., 2019), two-sided generalized Kumaraswamy, (Korkmaz and Genç, 2017), log-xgamma (Altun and Hamedani, 2018), unit Birnbaum-Saunders (Mazucheli et al., 2018), unit Lindley (Mazucheli et al., 2019b), unit inverse Gaussian (Ghitany et al., 2019), unit Gompertz (Mazucheli at al., 2019), unit improved second degree Lindley (Altun and Cordeiro, 2020), log-weighted exponential (Altun, 2020), logit slash (Korkmaz, 2020a) and unit generalized half normal (Korkmaz, 2020b) distributions. Many of above distributions have been obtained via transforming of the base distribution and they have given better results than beta distribution in terms of data modeling.

The aim of this study is to propose a new alternative unit distribution, which belongs to the exponential family, to model of the percentages and proportions. Since the proposed distribution belongs to exponential family the sufficient statistics are obtained for its model parameters. The new distribution is derived from a transformation of the Johnson S_U distribution in (1). The proposed distribution has distinguished properties of the density and hazard rate shapes among many unit distributions and provides better fits than some well known unit distributions such as beta, Kumaraswamy and Johnson S_B.

The rest of the paper is organized as follows. We define the proposed distribution and discuss its some basic distributional properties in Section 2. Section 3 is devoted to procedures of the different estimation methods to estimate its unknown parameters and the expected Fisher information matrix has been obtained. In Section 4, the simulation studies based on the three different scenarios are given to see the performance of the different estimators of the model parameters. The analysis of the three real data examples, which one is related to the coronavirus data, have been illustrated in Section 5. Finally, the paper is ended with concluding.

2 The unit Johnson S_U distribution

Based on (1), (2) and the transformation $X = e^Y/(1 + e^Y)$, we define the new unit distribution with the following cdf and pdf

$$F(x, \mu, \sigma) = \Phi\big(\mu + \sigma \sinh^{-1}(\log[x/(1-x)])\big) \tag{3}$$

and

$$f(x, \mu, \sigma) = \frac{\sigma}{x(1-x)\sqrt{1+[\log(x/(1-x))]^2}} \phi(\mu + \sigma \sinh^{-1}[\log(x/(1-x))]) \tag{4}$$

respectively, where $0 < x < 1$, $\sigma > 0$ and $\mu \in \Re$ are model parameters, and $\Phi(\cdot)$ is the cdf of standard normal distribution. From (3), it is clear that $F(x, \mu, \sigma) = F_{S_U}(\log[x/(1-x)], \mu, \sigma)$. We denote the unit Johnson S_U distribution with $UJS_U(\mu, \sigma)$. The UJS_U rv is expressed by the following transformations of the rvs

$$Z \sim N(0,1)$$
$$Y = \sinh((Z - \mu)/\sigma) \sim \text{Johnson } S_U$$
$$X = e^Y(1 + e^Y)^{-1} \sim UJS_U.$$

Hence, the newly defined distribution can be seen as the logit transformation of the unbounded Johnson distribution. The first derivation of the pdf obtained as

$$\frac{\partial f(x,\mu,\sigma)}{\partial x} = \frac{(2x-1)f(x,\mu,\sigma)}{x(1-x)} - \frac{2\pi f(x,\mu,\sigma)\log(x/(1-x))}{x(1-x)(1+2\pi\log(x/(1-x)))} \times \frac{\sigma f(x,\mu,\sigma)(\mu+\sigma\sinh^{-1}[\log(x/(1-x))])}{x(x-1)\sqrt{1+\log^2(x/(1-x))}}.$$

The signs of this equation can not be determined analytically. So, the plots of pdf and shape regions of the pdf are drawn for its density shapes in Figure 1. From Figure 1, we see that various pdf shapes such as w-shaped, U-shaped, uni-modal shaped, N-shaped, inverse N-shaped, decreasing and increasing for the UJS_U distribution. At the same time, these pdf shapes can be seen with plot of the pdf regions. As a result, we can say that the both μ and σ parameters are the shape parameters of the model. Due to these two parameters of the distribution, which ensure different shapes to model, we obtain many unit distributions on the unit interval with very flexible form.
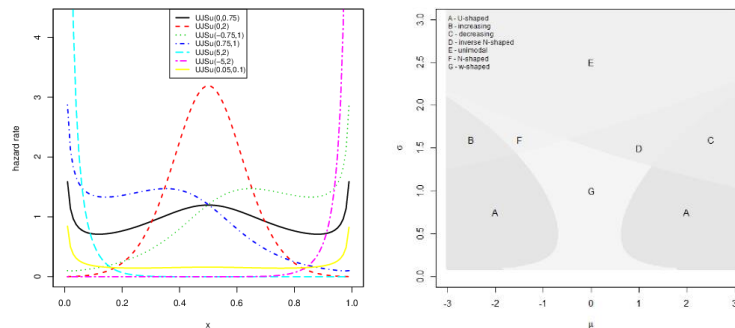

Figure 1: The possible pdf shapes (left) and pdf shape regions (right) plots of the UJS_U distribution for the selected parameters values

On the other hand, it is interesting that when $\mu = 0$, the distribution is symmetric about $1/2$ and the expected value of the distribution is equal to $1/2$. To see this, we can write following equations

$$f(1/2 - x, \mu, \sigma) = \frac{\sigma\big(1 + [\log((0.5-x)/(0.5+x))]^2\big)^{-0.5}}{(0.5-x)(0.5+x)\sqrt{2\pi}} e^{-1/2(\mu+\sigma\sinh^{-1}[\log((0.5-x)/(0.5+x))])^2}$$

and

$$f(1/2 + x, \mu, \sigma) = \frac{\sigma\big(1 + [-\log((0.5-x)/(0.5+x))]^2\big)^{-0.5}}{(0.5+x)(0.5-x)\sqrt{2\pi}} e^{-1/2(\mu+\sigma\sinh^{-1}[-\log((0.5-x)/(0.5+x))])^2}.$$

It can be easily seen that $f(1/2 - x, \mu, \sigma) = f(1/2 + x, -\mu, \sigma)$ from above equations using the $\sinh^{-1}(z) = -\sinh^{-1}(-z)$ for $z \in \Re$. From above equations, it is also noticed that $f(x, \mu, \sigma) = f(1 - x, -\mu, \sigma)$. Moreover, if the $UJS_U(\mu, \sigma)$ may be right skewed, then the $UJS_U(-\mu, \sigma)$ is left skewed (see moments subsection).

Consequently, we can say that the shapes of this distribution, especially w-shaped, N-shaped and inverse N-shaped pdfs, can be distinguishing feature on data modeling. Its some distribution properties are given by the following subsections.

2.1 Hazard rate function

The hazard rate function (hrf), of the UJS_U distribution is given by

$$h(x, \mu, \sigma) = \frac{\sigma}{x(1-x)\sqrt{1+[\log(x/(1-x))]^2}} \frac{\phi(\mu + \sigma \sinh^{-1}[\log(x/(1-x))])}{(1 - \Phi(\mu + \sigma \sinh^{-1}[\log(x/(1-x))]))}, \quad 0 < x < 1. \tag{5}$$

To define hrf shapes, we obtain the first derivation of the $\log h(x, \mu, \sigma)$ as

$$\partial \log h(x, \mu, \sigma) \, \partial x = -\frac{1}{x(1-x)} - \frac{\log\left(\frac{x}{1-x}\right)}{x(1-x)[\log^2(x1-x)+1]} - \frac{\sigma\left(\mu + \sigma \sinh^{-1}\left[\log\left(\frac{x}{1-x}\right)\right]\right)}{x(1-x)\sqrt{1+[\log(x1-x)]^2}} + h(x, \mu, \sigma).$$

The signs of this equation can not be determined analytically. So, the plots of hrf and shape regions of the hrf are drawn for shapes of the hrf in Figure 2. From this Figure, its hrf shapes can be w-shaped, bathtub shaped and increasing. At the same time, these hrf shapes can be seen with plot of the hrf regions. It is distinctive feature that the UJS_U distribution has w-shaped hrf on the unit interval. So, the w-shaped hrf shape of this distribution can be striking property on data modeling as well as in its density.
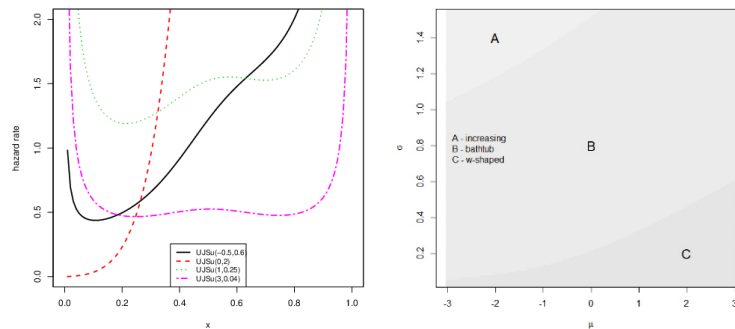

Figure 2: The possible hrf shapes (left) and hrf shape regions (right) plots of the UJS_U distribution for the selected parameters values

2.2 Moments

The rth moment of the $UJS_U(\mu, \sigma)$ distribution is given by

$$\mu'_r = \int_0^1 x^r f(x, \mu, \sigma) dx, \quad z = \mu + \sigma \sinh^{-1}\left[\log\left(\frac{x}{1-x}\right)\right]$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \left(1 - \frac{1}{1 + e^{\sinh\left(\frac{(z-\mu)}{\sigma}\right)}}\right)^r e^{-z^2/2} dz$$

$$= \frac{1}{\sqrt{2\pi}} \sum_{i=0}^{r} (-1)^i \binom{r}{i} \int_{-\infty}^{\infty} \left(1 + e^{\sinh\left(\frac{z-\mu}{\sigma}\right)}\right)^{-i} e^{-z^2/2} dz$$

$$= 1 + \frac{1}{\sqrt{2\pi}} \sum_{i=1}^{r} (-1)^i \binom{r}{i} \int_{-\infty}^{\infty} \left(1 + e^{\sinh\left(\frac{z-\mu}{\sigma}\right)}\right)^{-i} e^{-z^2/2} dz.$$

The first four moments of the distribution are given by

$$\mu'_1 = 1 - \int_{-\infty}^{\infty} \left(1 + e^{\sinh((z-\mu)/\sigma)}\right)^{-1} \phi(z) dz,$$

$$\mu'_2 = 2\mu'_1 - 1 + \int_{-\infty}^{\infty} \left(1 + e^{\sinh((z-\mu)/\sigma)}\right)^{-2} \phi(z)dz,$$

$$\mu'_3 = 1 + 3(\mu'_2 - \mu'_1) - \int_{-\infty}^{\infty} \left(1 + e^{\sinh((z-\mu)/\sigma)}\right)^{-3} \phi(z)dz$$

and

$$\mu'_4 = 4\mu'_3 - 6\mu'_2 + 4\mu'_1 - 1 + \int_{-\infty}^{\infty} \left(1 + e^{\sinh((z-\mu)/\sigma)}\right)^{-4} \phi(z)dz.$$

Alternatively, the following equation can be given for μ'_r

$$\mu'_r = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \left(1 + e^{-\sinh((z-\mu)/\sigma)}\right)^{-r} e^{-z^2/2} dz.$$

From above equations, we see that the rth raw moments can not be put forth analytically however, numerical integration can be applied to calculate the mean and other important related measures.
The jth central moment can be obtained by the following equation

$$\mu_j = E[(X - \mu'_1)^j] = \sum_{r=0}^{j} \binom{j}{r} \mu'_r (-\mu'_1)^{j-r}, \; j = 2,3,\dots$$

Then, the skewness and kurtosis coefficients are respectively given by $\sqrt{\frac{\mu_3^2}{\mu_2^3}}$ and $\frac{\mu_4}{\mu_2^2}$. These above calculations can be easily computed using many packet programs such as R, S-Plus, SAS and Wolfram.
The plots of the skewness and kurtosis coefficients for selected values of the μ and σ parameters are shown in Figure 3. From this Figure, we see that both two parameters affect skewness and kurtosis of the distribution. The distribution can be left skewed, right skewed and symmetrical. For $\mu = 0$, the skewness of distribution is equal to zero as we expect. For fixed μ, when σ increases, the skewness goes to zero as well as the kurtosis goes to 3. These plots indicate that the this distribution can model various data types on unit interval in terms of skewness and kurtosis.
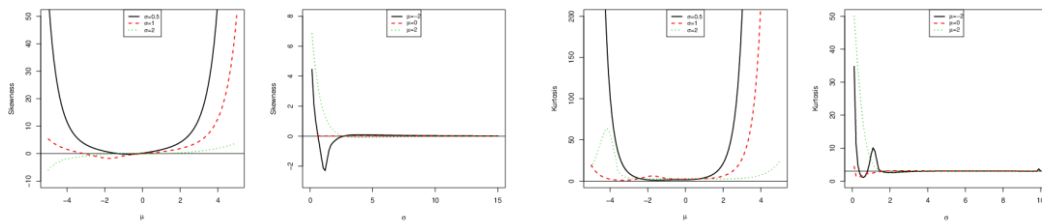

Figure 3: The skewness and kurtosis plots of the UJS_U distribution for selected parameters values

2.3 Exponential family
A distribution belongs to the exponential family if its pdf (or probability mass function) can be expressed as

$$f(x, \Theta) = h(x)c(\Theta)\exp\{t(x)w(\Theta)\},$$

where $\Theta = (\theta_1, \theta_2, \dots, \theta_k)^T$ is the parameter vector. The pdf of the UJS_U distribution can be rewritten as follows

$$f(x, \mu, \sigma) = \frac{\sigma e^{-\mu^2/2}}{x(1-x)\sqrt{1+\left[\log\left(\frac{x}{1-x}\right)\right]^2}\sqrt{2\pi}} e^{-\mu\sigma \sinh^{-1}\left[\log\left(\frac{x}{1-x}\right)\right] - \frac{\sigma^2}{2}\left(\sinh^{-1}\left[\log\left(\frac{x}{1-x}\right)\right]\right)^2}$$

$$= \frac{\overbrace{\sigma e^{-\mu^2/2}}^{c(\mu,\sigma)} e^{\overbrace{\left((\sinh^{-1}[\log(x/(1-x))])^2, \sinh^{-1}[\log(x/(1-x))]\right)}^{t(x)}{}^T \overbrace{\binom{-\sigma^2/2}{-\mu\sigma}}^{w(\mu,\sigma)}}}{\underbrace{x(1-x)\sqrt{1+[\log(x/(1-x))]^2}\sqrt{2\pi}}_{h(x)}}.$$

Therefore, the UJS_U distribution belongs to the exponential family and according to factorization theorem, the $t(X) = (\sum_{i=1}^{n} \sinh^{-1}[\log(X_i/(1-X_i))], \sum_{i=1}^{n} (\sinh^{-1}[\log(X_i/(1-X_i))])^2)$ is the sufficient statistics for the (μ, σ), where n is the sample size.

2.4 Stochastic ordering

Stochastic ordering of positive continuous random variables is an important tool to judge the comparative behavior of such variables. For this purpose, we shall recall some basic definitions. Let us denote the pdf, cdf and hrf of a positive continuous rv X by $f_X(\cdot), F_X(\cdot)$ and $h_X(\cdot)$ respectively, and those of another positive continuous rv Y by $f_Y(\cdot), F_Y(\cdot)$ and $h_Y(\cdot)$ respectively. We recall some basic definitions.
A rv X is said to be smaller than a rv Y in the

- The stochastic order ($X \leq_{(sto)} Y$) if $F_X(x) \geq F_Y(x), \forall x$

- The hazard rate order ($X \leq_{(hro)} Y$) if $h_X(x) \geq h_Y(x), \forall x$.

- The likelihood ratio order ($X \leq_{(lro)} Y$) if $g(x) = \frac{f_X(x)}{f_Y(x)}$ decreases in x.

The below given implications (see Shaked and Shanthikumar, 2007) are well justified:

$$\left[X \leq_{(lro)} Y\right] \Rightarrow \left[X \leq_{(hro)} Y\right] \Rightarrow \left[X \leq_{(sto)} Y\right]$$

The following Proposition shows that the UJS_U distributions are ordered with respect to different stochastic orderings. **Proposition 1.** Let $X \sim UJS_U(\mu_1, \sigma_1)$ and $Y \sim UJS_U(\mu_2, \sigma_2)$. If $\sigma_2 = \sigma_1$ and $\mu_1 > \mu_2$, then $\left[X \leq_{(lro)} Y\right]$ and $\left[X \leq_{(hro)} Y\right], \left[X \leq_{(sto)} Y\right]$.

Proof. For any $0 < x < 1$, the likelihood ratio is given by

$$g(x) = \frac{\sigma_1}{\sigma_2} e^{-\frac{1}{2}\left\{\left(\mu_1 + \mu_2 + (\sigma_1 + \sigma_2)\sinh^{-1}\left[\log\left(\frac{x}{1-x}\right)\right]\right)\left(\mu_1 - \mu_2 + (\sigma_1 - \sigma_2)\sinh^{-1}\left[\log\left(\frac{x}{1-x}\right)\right]\right)\right\}}$$

Thus, taking derivative with respect to x of $\log g(x)$, we have

$$\frac{\partial \log g(x)}{\partial x} = \frac{\mu_2\sigma_2 - \mu_1\sigma_1}{x(1-x)\sqrt{1 + \left[\log\left(\frac{x}{1-x}\right)\right]^2}} - \frac{(\sigma_1^2 - \sigma_2^2)\sinh^{-1}[\log(x1-x)]}{x(1-x)\sqrt{1 + \left[\log\left(\frac{x}{1-x}\right)\right]^2}}.$$

If $\sigma_1 = \sigma_2$ and $\mu_1 > \mu_2$, then $\frac{\partial \log g(x)}{\partial x} \leq 0$ for all x, which purposes that $\left[X \leq_{(lro)} Y\right]$ and $\left[X \leq_{(hro)} Y\right], \left[X \leq_{(sto)} Y\right]$. Hence, the proof is completed.

2.5 Relative Entropy

The relative entropy or Kullback-Leibler distance (also known as cross entropy or discriminant function) of a random variable with density function f with respect to another random variable with density g is defined as

$$E(f:g) = \int \log(f/g) dF$$

(Kullback, 1997). This is the discrimination function since $E(f:g) \geq 0$ and equality holds if and only if $f = g$ almost everywhere. The relative entropy is used as a measure for comparing information content of distributions in information theory. Since the symmetrical UJS_U distribution around 0.5 is sub-model of the UJS_U distribution, one may want to compare the information contents of the UJS_U distributions family to the information content of the symmetrical UJS_U distribution. Under this definition, the relative entropy is obtained as

$$E(f:g) = -\frac{\mu^2}{2} - \mu \int_0^1 \frac{\sigma^2 \sinh^{-1}\left[\log\left(\frac{x}{1-x}\right)\right] e^{-\frac{1}{2}\left(\mu+\sigma\sinh^{-1}\left[\log\left(\frac{x}{1-x}\right)\right]\right)^2}}{x(1-x)\sqrt{1+\left[\log\left(\frac{x}{1-x}\right)\right]^2}\sqrt{2\pi}} dx.$$

By using transformation $t = \sigma\sinh^{-1}\left[\log\left(\frac{x}{1-x}\right)\right]$ and after some calculations, we derive

$$E(f:g) = \mu^2/2.$$

As it can be seen that the relative entropy does not depend on σ, it is a function of only μ parameter. Therefore, no information is added to or subtracted from the information content of a UJS_U distribution by varying the parameter σ. Furthermore, this quantity attains its minimum value 0, as expected. When μ approaches to both ∞ and $-\infty$, the relative entropy approaches to ∞.

2.6 Order statistics

Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from the UJS_U distribution, and let $X_{(1)} \leq X_{(2)} \ldots \leq X_{(n)}$ denote the corresponding order statistics. It is well-known that the cdf and pdf of the rth order statistic, $X_{(r)}$, are defined by

$$F_{X(r)}(x) = \sum_{i=r}^n \binom{n}{i} [F(x)]^i [1-F(x)]^{n-i} = \sum_{i=r}^n \sum_{j=0}^{n-i} (-1)^j \binom{n}{i}\binom{n-i}{j} [F(x)]^{i+j}$$

and

$$f_{X(r)}(x) = \frac{n!}{(r-1)!(n-r)!}[F(x)]^{r-1}[1-F(x)]^{n-r}f(x) = \frac{1}{B(r,n-r+1)}\sum_{i=0}^{n-r} (-1)^i \binom{n-r}{i}[F(x)]^{r+i-1}f(x),$$

respectively, where $r = 1,2,\ldots,n$ and $B(\cdot,\cdot)$ is the beta function. Therefore, the cdf and pdf of the rth order statistic of the UJS_U distribution are given by

$$F_{X(r)}(x) = \sum_{i=r}^n \sum_{j=0}^{n-i} (-1)^j \binom{n}{i}\binom{n-r}{j}\left[\Phi\left(\mu + \sigma\sinh^{-1}\left(\log\left[\frac{x}{1-x}\right]\right)\right)\right]^{i+j}$$

and

$$f_{X(r)}(x) = \frac{\sigma\,\phi\left(\mu+\sigma\sinh^{-1}\left(\log\left[\frac{x}{1-x}\right]\right)\right)}{x(1-x)B(r,n-r+1)\sqrt{1+[\log(x1-x)]^2}}\sum_{i=0}^{n-r} (-1)^i \binom{n-r}{i}\left[\Phi\left(\mu + \sigma\sinh^{-1}\left(\log\left[\frac{x}{1-x}\right]\right)\right)\right]^{r+i-1}.$$

For $r = 1$ and $r = n$, we have the pdf of the $X_{(1)} = min\{X_1, X_2, \ldots, X_n\}$ and $X_{(n)} = max\{X_1, X_2, \ldots, X_n\}$, respectively.

2.7 Quantile function and random number generation

Let X be a UJS_U rv with cdf (3). Then, the quantile function, $x_u(\mu, \sigma) = F^{-1}(u, \mu, \sigma)$, of the UJS_U distribution can be given by

$$x_u(\mu,\sigma) = \exp\left[\sinh\left(\frac{\Phi^{-1}(u)-\mu}{\sigma}\right)\right]\left(1 + \exp\left[\sinh\left(\frac{\Phi^{-1}(u)-\mu}{\sigma}\right)\right]\right)^{-1},$$

where $0 < u < 1$ and $\Phi^{-1}(u)$ is the u th quantile of standart normal distribution. Hence, if U is a uniform rv on $(0,1)$ then, $X_U(\mu, \sigma)$ is the UJS_U rv. To generate random variables from the UJS_U distribution, we have the following algorithm.

Algorithm
- set μ and σ,
- simulate $U \sim Uniform(0,1)$,
- compute $Y = \sinh\left(\frac{\Phi^{-1}(U)-\mu}{\sigma}\right)$, then Y follows that $Johnson\ S_U(\mu, \sigma)$,
- compute $X = e^Y(1 + e^Y)^{-1}$, then X follows that $UJS_U(\mu, \sigma)$.

3 Different methods of the parameter estimation

In this section, we point out the six different estimators to estimate of the UJS_U model parameters. These methods are: the maximum likelihood (with inferences based on it), maximum product spacings, least squares, weighted least squares, Anderson-Darling and Cramer-von Mises estimators. The details are the following.

3.1 Maximum likelihood estimation

In this subsection, we estimate the parameters of the UJS_U distribution via the method of maximum likelihood estimation (MLE). Let $X_1, X_2, \ldots\ldots, X_n$ be a random sample from the UJS_U distribution with observed values $x_1, x_2, \ldots\ldots, x_n$, and $\Xi = (\mu, \sigma)^T$ be the vector of the model parameters. The log likelihood function for Ξ may be expressed as

$$\ell(\Xi) = n\log\sigma - \frac{n}{2}\log2\pi + \sum_{i=1}^{n} \log\left[\frac{\left(1+\left[\log\left(\frac{x_i}{1-x_i}\right)\right]^2\right)^{-\frac{1}{2}}}{x_i(1-x_i)}\right] - \frac{1}{2}\sum_{i=1}^{n}\left(\mu + \sigma\sinh^{-1}\left[\log\left(\frac{x_i}{1-x_i}\right)\right]\right)^2. \quad (6)$$

The MLEs, $\hat{\mu}_{MLE}$ and $\hat{\sigma}_{MLE}$, of the μ and σ parameters can be obtained as the simultaneous solution of the following non-linear equations:

$$\frac{\partial\ell(\Xi)}{\partial\mu} = -n\mu - \sigma\sum_{i=1}^{n}\sinh^{-1}\left[\log\left(\frac{x_i}{1-x_i}\right)\right] = 0 \quad (7)$$

and

$$\frac{\partial\ell(\Xi)}{\partial\sigma} = \frac{n}{\sigma} - \mu\sum_{i=1}^{n}\sinh^{-1}\left[\log\left(\frac{x_i}{1-x_i}\right)\right] - \sigma\sum_{i=1}^{n}\left(\sinh^{-1}\left[\log\left(\frac{x_i}{1-x_i}\right)\right]\right)^2 = 0.$$

From (7), the $\hat{\mu}_{MLE}$ is obtained as function of the σ and is given by

$$\hat{\mu}_{MLE} = -\frac{\sigma}{n}\sum_{i=1}^{n}\sinh^{-1}\left[\log\left(\frac{x_i}{1-x_i}\right)\right]. \quad (8)$$

Substituting (8) in (6), we obtain the profile log-likelihood of σ as

$$\ell(\sigma) = n\log\sigma - \frac{n}{2}\log2\pi + \sum_{i=1}^{n}\log\left[\frac{\left(1+\left[\log\left(\frac{x_i}{1-x_i}\right)\right]^2\right)^{-\frac{1}{2}}}{x_i(1-x_i)}\right] + \frac{\sigma^2}{2n}\left(\sum_{i=1}^{n}\sinh^{-1}\left[\log\left(\frac{x_i}{1-x_i}\right)\right]\right)^2$$

$$-\frac{\sigma^2}{2}\sum_{i=1}^{n}\left(\sinh^{-1}\left[\log\left(\frac{x_i}{1-x_i}\right)\right]\right)^2 \quad (9)$$

Therefore, the $\hat{\sigma}_{MLE}$ is obtained by maximizing (9) based on σ parameter. Following the normal routine of parameter estimation for the $\hat{\sigma}_{MLE}^2$, we have

$$\hat{\sigma}_{MLE}^2 = \frac{n}{\sum_{i=1}^{n}\left(\sinh^{-1}\left[\log\left(\frac{x_i}{1-x_i}\right)\right]\right)^2 - \frac{1}{n}\left(\sum_{i=1}^{n}\sinh^{-1}\left[\log\left(\frac{x_i}{1-x_i}\right)\right]\right)^2}.$$

Hence, $\hat{\sigma}_{MLE}^2 = \frac{1}{\mathrm{var}(h)}$ and $\hat{\mu}_{MLE} = -\frac{\bar{h}}{\sqrt{\mathrm{var}(h)}}$, where $\bar{h}$ is the mean and $var(h)$ is the variance of the values of the h, defined by $h_i = \sinh^{-1}\left[\log\left(\frac{x_i}{1-x_i}\right)\right]$, for $i = 1,2,\ldots n$.

For interval estimations of the μ and σ parameters, the Fisher information matrix is required. The Fisher information matrix is defined by

$$I(\Xi) = \begin{pmatrix} -E\left(\dfrac{\partial^2 \ell(\Xi)}{\partial \mu^2}\right) & -E\left(\dfrac{\partial^2 \ell(\Xi)}{\partial \mu\, \partial \sigma}\right) \\ -E\left(\dfrac{\partial^2 \ell(\Xi)}{\partial \sigma\, \partial \mu}\right) & -E\left(\dfrac{\partial^2 \ell(\Xi)}{\partial \sigma^2}\right) \end{pmatrix} = \begin{pmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{pmatrix}.$$

The elements of this information matrix are

$$I_{11} = n,$$

$$I_{12} = I_{21} = \sum_{i=1}^{n} E\left(\sinh^{-1}\left[\log\left(\frac{X_i}{1-X_i}\right)\right]\right)$$

and

$$I_{12} = \frac{n}{\sigma^2} + \sum_{i=1}^{n} E\left(\left(\sinh^{-1}\left[\log\left(\frac{X_i}{1-X_i}\right)\right]\right)^2\right).$$

Above expectations are calculated as

$$E\left(\sinh^{-1}\left[\log\left(\frac{x}{1-x}\right)\right]\right) = \frac{\int_0^1 \sigma \sinh^{-1}\left[\log\left(\frac{x}{1-x}\right)\right]\phi\left(\mu + \sigma\sinh^{-1}\left[\log\left(\frac{x}{1-x}\right)\right]\right)}{x(1-x)\sqrt{1+\left[\log\left(\frac{x}{1-x}\right)\right]^2}}\,dx$$

$$= \int_{-\infty}^{\infty} \left(\frac{u-\mu}{\sigma}\right)\phi(u)\,du = -\frac{\mu}{\sigma}$$

and

$$E\left(\left(\sinh^{-1}\left[\log\left(\frac{X_i}{1-X_i}\right)\right]\right)^2\right) = \frac{\int_0^1 \sigma\left(\sinh^{-1}\left[\log\left(\frac{X_i}{1-X_i}\right)\right]\right)^2 \phi\left(\mu+\sigma\sinh^{-1}\left[\log\left(\frac{x}{1-x}\right)\right]\right)}{x(1-x)\sqrt{1+\left[\log\left(\frac{x}{1-x}\right)\right]^2}}\,dx = \int_{-\infty}^{\infty}\left(\frac{u-\mu}{\sigma}\right)^2\phi(u)\,du = \frac{1+\mu^2}{\sigma^2}.$$

Hence, the Fisher information matrix is obtained as

$$I(\Xi) = \begin{pmatrix} n & -n\mu/\sigma \\ -n\mu/\sigma & n(\mu^2+2)/\sigma^2 \end{pmatrix}$$

Now, the variance-covariance matrix of $(\hat{\mu}_{MLE}, \hat{\sigma}_{MLE})$ is given by

$$\widehat{K}_n(\hat{\Xi}) = I^{-1}(\hat{\Xi})\bigg|_{\substack{\mu=\hat{\mu}_{MLE}\\\sigma=\hat{\sigma}_{MLE}}} = \frac{1}{I_{11}I_{22}-I_{12}^2}\begin{pmatrix} I_{22} & -I_{12} \\ -I_{21} & I_{11} \end{pmatrix} = \begin{pmatrix} Var(\mu) & Cov(\mu,\sigma) \\ Cov(\sigma,\mu) & Var(\sigma) \end{pmatrix} = \begin{pmatrix} \dfrac{\mu^2+2}{2n} & \dfrac{\mu\sigma}{2n} \\ \dfrac{\mu\sigma}{2n} & \dfrac{\sigma^2}{2n} \end{pmatrix}.$$

Under regularity conditions, the asymptotic distribution of $\sqrt{n}(\hat{\mu}_{MLE} - \mu, \hat{\sigma}_{MLE} - \sigma)$ as $n \to \infty$ is a bivariate normal with zero mean and variance covariance $\widehat{K}_n(\hat{\Xi})$. Finally, the asymptotic confidence intervals for the μ and σ parameters with a significance level δ are given by

$$\hat{\mu}_{MLE} - z_{\delta/2}\sqrt{\frac{\hat{\mu}_{MLE}^2+2}{2n}} < \mu < \hat{\mu}_{MLE} + z_{\delta/2}\sqrt{\frac{\hat{\mu}_{MLE}^2+2}{2n}}$$

and

$$\hat{\sigma}_{MLE} - z_{\delta/2}\frac{\hat{\sigma}_{MLE}}{\sqrt{2n}} < \sigma < \hat{\sigma}_{MLE} + z_{\delta/2}\frac{\hat{\sigma}_{MLE}}{\sqrt{2n}}$$

respectively, $z_{\delta/2}$ is the upper δth percentile of the standard normal model.

3.2 Maximum product spacing estimation

The maximum product spacing (MPS) method is an alternative method to MLE for parameter estimation. This method was proposed by Cheng and Amin (1979) and independently developed by Ranneby (1984) as approximation to the Kullback-Leibler measure of information. This method is based on the idea that differences (spacings) between the values of the cdf at consecutive data points should be identically distributed. The geometric mean of the differences is given as

$$GM = \sqrt[n+1]{\prod_{i=1}^{n+1} D_i},$$

where, the difference D_i is defined by

$$D_i = \int_{x_{(i-1)}}^{x_{(i)}} f(x, \mu, \sigma) dx; i = 1, 2, \dots, n+1, \tag{10}$$

where, $F(x_{(0)}, \mu, \sigma) = 0$ and $F(x_{(n+1)}, \mu, \sigma) = 1$. The maximum product spacing estimations (MPSEs), $\hat{\mu}_{MPS}$ and $\hat{\sigma}_{MPS}$, of the μ and σ parameters are obtained by maximizing the geometric mean (GM) of the differences. Substituting cdf of the UJS_U in (10) and taking logarithm of the above expression, we have

$$MPS(\Xi) = \frac{1}{n+1} \sum_{i=1}^{n+1} \log[F(x_{(i)}, \mu, \sigma) - F(x_{(i-1)}, \mu, \sigma)]. \tag{11}$$

The MPSEs can be obtained as the simultaneous solution of the following non-linear equations:

$$\frac{\partial MPS(\Xi)}{\partial \mu} = \frac{1}{n+1} \sum_{i=1}^{n+1} \left[\frac{F'_\mu(x_{(i)}, \mu, \sigma) - F'_\mu(x_{(i-1)}, \mu, \sigma)}{F(x_{(i)}, \mu, \sigma) - F(x_{(i-1)}, \mu, \sigma)} \right] = 0$$

and

$$\frac{\partial MPS(\Xi)}{\partial \sigma} = \frac{1}{n+1} \sum_{i=1}^{n+1} \left[\frac{F'_\sigma(x_{(i)}, \mu, \sigma) - F'_\sigma(x_{(i-1)}, \mu, \sigma)}{F(x_{(i)}, \mu, \sigma) - F(x_{(i-1)}, \mu, \sigma)} \right] = 0,$$

where $F'_\sigma(x, \mu, \sigma, \lambda) = \sinh^{-1}\left(\log\left[\frac{x}{1-x}\right]\right) \phi\left(\mu + \sigma \sinh^{-1}\left(\log\left[\frac{x}{1-x}\right]\right)\right)$ and
$F'_\mu(x, \mu, \sigma, \lambda) = \phi\left(\mu + \sigma \sinh^{-1}\left(\log\left[\frac{x}{1-x}\right]\right)\right)$.

3.3 Least squares estimation

Let $X_{(1)}, X_{(2)}, \dots, X_{(n)}$ be ordered statistics from the UJS_U distribution with sample size n. The least square estimations (LSEs), $\hat{\mu}_{LSE}$ and $\hat{\sigma}_{LSE}$, of the μ and σ parameters are obtained by minimizing the following equation

$$LSE(\Xi) = \sum_{i=1}^{n} \left(F\left(x_{(i)}, \mu, \sigma\right) - E\left[F\left(x_{(i)}\right)\right] \right)^2, \tag{12}$$

where $E\left[F\left(x_{(i)}\right)\right] = \frac{i}{n+1}$ is the expected value of the empirical cdf for $i = 1, 2, \dots, n$. Then, $\hat{\mu}_{LSE}$ and $\hat{\sigma}_{LSE}$ are solutions of the following equations:

$$\frac{\partial LSE(\Xi)}{\partial \mu} = \sum_{i=1}^{n} F'_\mu(x_{(i)}, \mu, \sigma) \left(F\left(x_{(i)}, \mu, \sigma\right) - \frac{i}{n+1} \right) = 0,$$

and

$$\frac{\partial LSE(\Xi)}{\partial \sigma} = \sum_{i=1}^{n} F'_\sigma(x_{(i)}, \mu, \sigma) \left(F\left(x_{(i)}, \mu, \sigma\right) - \frac{i}{n+1} \right) = 0,$$

respectively, where $F'_\mu(x_{(i)}; \mu, \sigma)$ and $F'_\sigma(x_{(i)}; \mu, \sigma)$ are provided before.

3.4 Weighted least squares estimation

Let $X_{(1)}, X_{(2)}, \dots, X_{(n)}$ be ordered statistics from the UJS_U distribution with sample size n. The weighted least square estimations (WLSEs), $\hat{\mu}_{WLSE}$ and $\hat{\sigma}_{WLSE}$, of the μ and σ parameters are obtained by minimizing the following equation

$$WLSE(\Xi) = \sum_{i=1}^{n} \frac{(F(x_{(i)},\mu,\sigma) - E[F(x_{(i)})])^2}{V[F(x_{(i)})]},$$

(13)

where $V[F(x_{(i)})] = \frac{i(n-i+1)}{(n+2)(n+1)^2}$ is the variance of the empirical cdf for $i = 1,2,...,n$. Then, $\hat{\mu}_{WLSE}$ and $\hat{\sigma}_{WLSE}$ are solutions of the following equations:

$$\frac{\partial WLSE(\Xi)}{\partial \mu} = \sum_{i=1}^{n} \frac{F'_\mu(x_{(i)},\mu,\sigma)}{V[F(x_{(i)})]} \left(F(x_{(i)},\mu,\sigma) - \frac{i}{n+1} \right) = 0,$$

and

$$\frac{\partial WLSE(\Xi)}{\partial \sigma} = \sum_{i=1}^{n} \frac{F'_\sigma(x_{(i)},\mu,\sigma)}{V[F(x_{(i)})]} \left(F(x_{(i)},\mu,\sigma) - \frac{i}{n+1} \right) = 0,$$

respectively, where $F'_\mu(x_{(i)};\mu,\sigma)$ and $F'_\sigma(x_{(i)};\mu,\sigma)$ are provided before.

3.5 Anderson-Darling estimation

This estimator is based on Anderson-Darling goodness-of-fits statistics which is introduced by Anderson and Darling (1952). The Anderson-Darling (AD) minimum distance estimations, $\hat{\mu}_{AD}$, $\hat{\sigma}_{AD}$, of the μ, σ parameters are obtained by minimizing

$$AD(\Xi) = -n - \sum_{i=1}^{n} \frac{2i-1}{n} \left[\log F(x_{(i)},\mu,\sigma,\lambda) + \log\{1 - F(x_{(n+1-i)},\mu,\sigma,\lambda)\} \right].$$

(14)

Therefore, the $\hat{\mu}_{AD}$ and $\hat{\sigma}_{AD}$ can be obtained as the solution of the following system of equations:

$$\frac{\partial AD(\Xi)}{\partial \mu} = \sum_{i=1}^{n} (2i-1) \left[\frac{F'_\mu(x_{(i)},\mu,\sigma)}{F(x_{(i)}\mu,\sigma)} - \frac{F'_\mu(x_{(n+1-i)},\mu,\sigma)}{1 - F(x_{(n+1-i)},\mu,\sigma)} \right] = 0$$

and

$$\frac{\partial AD(\Xi)}{\partial \sigma} = \sum_{i=1}^{n} (2i-1) \left[\frac{F'_\sigma(x_{(i)},\mu,\sigma)}{F(x_{(i)},\mu,\sigma)} - \frac{F'_\sigma(x_{(n+1-i)},\mu,\sigma)}{1 - F(x_{(n+1-i)},\mu,\sigma)} \right] = 0.$$

3.6 The Cramer-von Mises estimation

The Cramer-von Mises (CVM) minimum distance estimations, $\hat{\mu}_{CVM}$ and $\hat{\sigma}_{CVM}$, of the μ and σ parameters are obtained by minimizing

$$CVM(\Xi) = \frac{1}{12n} + \sum_{i=1}^{n} \left[F(x_{(i)},\mu,\sigma) - \frac{2i-1}{2n} \right]^2.$$

(15)

Therefore, the $\hat{\mu}_{AD}$ and $\hat{\sigma}_{AD}$ can be obtained as the solution of the following system of equations:

$$\frac{\partial CVM(\Xi)}{\partial \mu} = \sum_{i=1}^{n} \left(F(x_{(i)},\mu,\sigma) - \frac{2i-1}{2n} \right) F'_\mu(x_{(i)},\mu,\sigma,\lambda) = 0,$$

and

$$\frac{\partial CVM(\Xi)}{\partial \sigma} = \sum_{i=1}^{n} \left(F(x_{(i)},\mu,\sigma) - \frac{2i-1}{2n} \right) F'_\sigma(x_{(i)},\mu,\sigma) = 0.$$

We note that one may see Chen and Balakrishnan (1995) for AD and CVM goodness-of-fits statistics in detail.
As it can be seen, all estimating equations except those of the MLE method contain non-linear functions, it is not possible to obtain explicit forms of all estimators directly. Therefore, they have to be solved by using numerical methods such as the Newton-Raphson and quasi-Newton algorithms.

4 Simulation experiments

In this Section, we perform the graphical simulation studies to see the performance of the above estimators of the UJS_U distribution with respect to varying sample size n. We generate $N = 1000$ samples of size $n = 20, 25, ...,1000$

from the UJS_U distribution based on the three different scenario. The actual parameter values used in the simulation studies are $\mu = 1$, $\sigma = 1$, $\mu = 2$, $\sigma = 1.5$ and $\mu = -2$, $\sigma = 2$ for the scenario-I, scenario-II and scenario-III respectively. The random numbers generation is obtained by random number generation algorithm. All estimations based on the estimation methods have been obtained by using constrOptim routine in the R program. Further, we calculate the empirical mean, bias and mean square error (MSE) of the estimations for comparisons between estimation methods. The bias and MSE are calculated by (for $h = \mu, \sigma$)

$$Bias_h(n) = \frac{1}{N}\sum_{i=1}^{N}\left(h_i - \hat{h}_i\right)$$

and

$$MSE_h(n) = \frac{1}{N}\sum_{i=1}^{N}\left(h_i - \hat{h}_i\right)^2$$

respectively. We expect that the empirical means are close to true values when the MSEs and biases are near zero. The results of this simulation study are shown in Figures 4-6. These Figures show that all estimators are consistent since the MSE and biasedness decrease to zero with increasing sample size as expected. All estimators are asymptotic unbiased also. For all scenarios, the amount of the biases and MSEs of the MLE, LSE, WLSE and AD methods are less than those of the CVM and MPS estimation methods and the CVM method has the maximum amount of the biases and MSEs . In all the cases, the performances of the other estimators are close and they are superior than the CVM and MPS methods. Therefore, all methods can be choosen as more reliable than another estimator of new distribution. The similar results can be also obtained for different parameter settings.
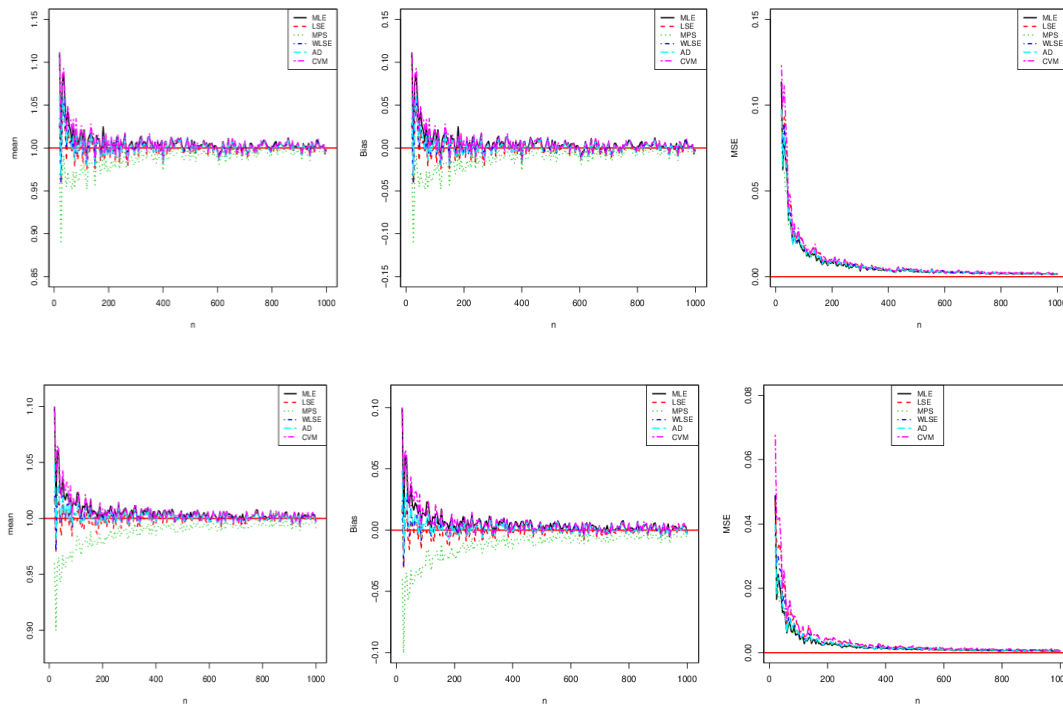
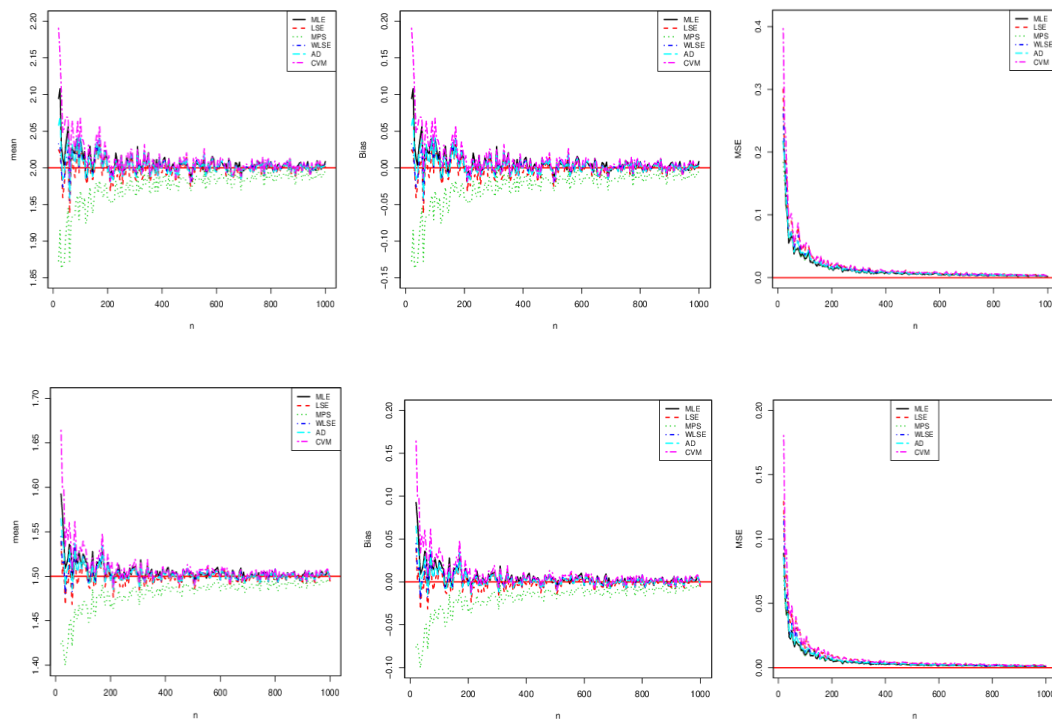Figure 4: The results of the μ (top) and σ (bottom) parameters for the scenario-I

Figure 5: The results of the μ (top) and σ (bottom) parameters for the scenario-II
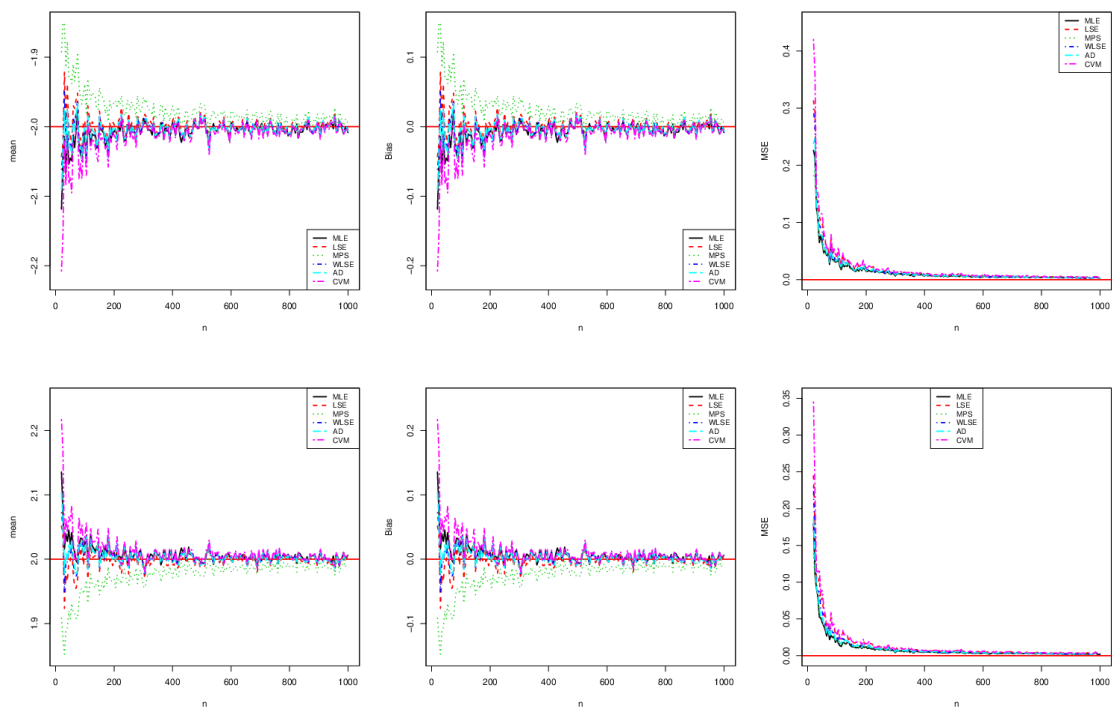

Figure 6: The results of the μ (top) and σ (bottom) parameters for the scenario-III

Furthermore, we also give simulation studies based on above results of the MLEs to see the performance of the 95% confidence intervals. The performance of the MLEs is evaluated based on the measurements of the coverage probability (CP) and coverage length (CL). The empirical CP and CL based on the MLEs are given by

$$CP_h(n) = \frac{1}{N}\sum_{i=1}^{N} I\left(\hat{h}_i - 1.9599 s_{\hat{h}_i} < h < \hat{h}_i + 1.9599 s_{\hat{h}_i}\right)$$

and

$$CL_h(n) = \frac{1}{N}\sum_{i=1}^{N} 3.9198 s_{\hat{h}_i}$$

respectively, where $I(\cdot)$ is the indicator function and the $s_{\hat{h}_i} = \left(s_{\hat{\mu}_i}, s_{\hat{\sigma}_i}\right)$ are the standard errors of the MLEs which are computed by inverting the Fisher information matrix. Figure 7 displays the simulation results for above measure. As seen from this Figure, For all scenarios, the CPs are very close to 0.95 and are around the nominal value as well as the CLs decrease when the sample size increases. These above results are as expected. It can be noticed that the CLs of the σ parameter are narrower than those of the μ parameter. The simulation results verify the consistency property of MLEs.
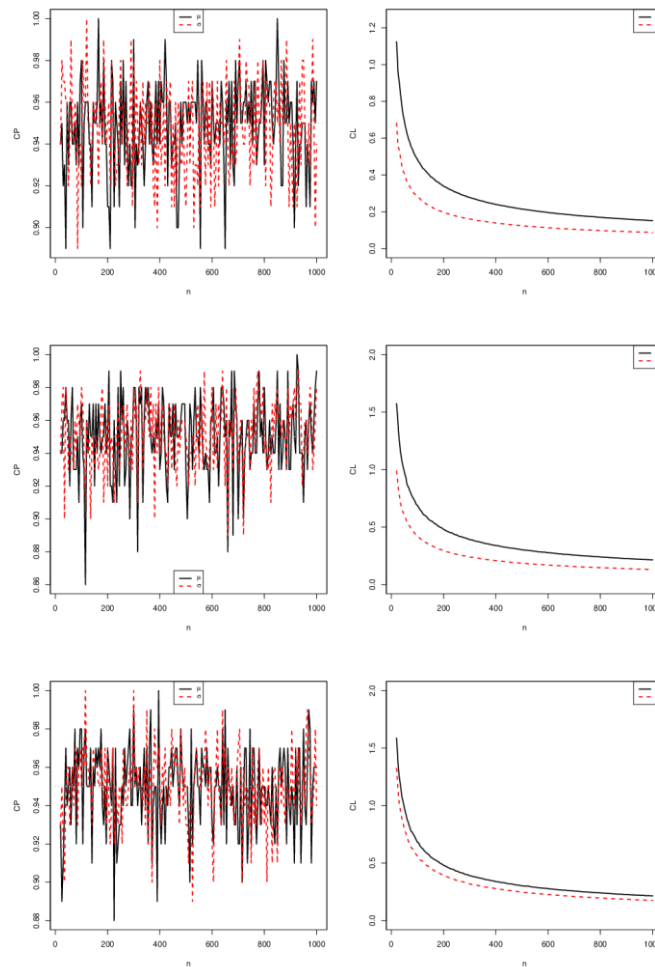
Figure 7: The plots of the empirical CPs and CLs of the first (top), second (center) and third (bottom) scenarios.

5 Data Analysis

In this section we analyze three data sets, which one is related to the coronavirus data and the others are related to burr measurements on the iron sheets, to compare the UJS_U distribution with well know unit distribution in the literature. These comparing models are (with their pdfs for $0 < x < 1$):

• Beta distribution:

$$f_{Beta}(x,\mu,\sigma) = \frac{1}{B(\mu,\sigma)} x^{\mu-1}(1-x)^{\mu-1}, \mu,\sigma > 0$$

where, $B(\mu,\sigma)$ is the beta function.
• Johnson S_B distribution:

$$f_{S_B}(x,\mu,\sigma) = \frac{\sigma}{x(1-x)}\phi\left(\sigma\log\left(\frac{x}{1-x}\right) + \mu\right) \ \ \mu \in \Re, \sigma > 0.$$

• Topp Leone (TL) distribution:

$$f_{TL}(x,\mu) = 2\mu x^{\mu-1}(1-x)(2-x)^{\mu-1}, \mu > 0.$$

• Kumaraswamy (Kw) distribution:

$$f_{Kw}(x,\mu,\sigma) = \mu\sigma x^{\mu-1}(1-x^{\mu})^{\sigma-1}, \mu,\sigma > 0.$$

• Unit inverse Gaussian (UIG) distribution:

$$f_{UIG}(x,\mu,\sigma) = \sqrt{\frac{\mu}{2\pi}}\frac{1}{x(-\log x)^{3/2}}\exp\left[\frac{\mu}{2\sigma^2\log x}(\log x + \sigma)^2\right], \ \mu,\sigma > 0.$$

• Unit Gompertz (UGom) distribution:

$$f_{UGom}(x,\mu,\sigma) = \mu\sigma x^{-\beta-1}\exp[-\mu(x^{-\sigma} - 1)], \mu,\sigma > 0.$$

• log-weighted exponential (LWE) distribution:

$$f_{LWE}(x,\mu,\sigma) = \frac{\mu+1}{\mu}\sigma x^{\sigma-1}(1-x^{\mu\sigma}), \mu,\sigma > 0.$$

To determine the best model, we also compute the estimated log-likelihood values $\hat{\ell}$, Akaike Information Criteria (AIC), Bayesian information criterion (BIC), Kolmogorov-Smirnov (KS), Cramer-von-Mises, (W^*) and Anderson-Darling (A^*) goodness of-fit statistics for all distribution models. In general, it can be chosen as the best model the one which has the smaller the values of the AIC, BIC, KS, W^* and A^* statistics and the larger the values of $\hat{\ell}$ and p-value of the goodness-of-statistics. It is notice that all results have been obtained under MLE method.

The first two data sets have been firstly introduced and studied by Dasgupta (2011) for burr measurements on the iron sheets. For the first data set of 50 observations on burr (in the unit of millimeters), the hole diameter is 12mm and the sheet thickness is 3.15mm. For the second data set of 50 observations, hole diameter and sheet thickness are 9 and 2mm, respectively. Hole diameter readings are taken on jobs with respect to one hole, selected and fixed as per a predetermined orientation. The two data sets relate to two different machines under comparison. One may see Dasgupta (2011) about the technical details of the data sets' measurements. These data sets have been analyzied by Korkmaz and Erişoğlu (2014), Dey et al., (2017), Dey et al., (2018) and ZeinEldin et al. (2019).

Third data set is related to the new type of pandemic coronavirus in Turkey. On 12 January 2020, the World Health Organization (WHO) confirmed that a novel coronavirus was the cause of a respiratory illness in a cluster of people in Wuhan City, Hubei Province, China, which was reported to the WHO on 31 December 2019. The case fatality ratio for coronavirus disease 2019 (COVID-19) has been much lower than SARS of 2003 but the transmission has been significantly greater, with a significant total death toll. The ongoing pandemic of the COVID-19, a novel infectious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), spread to Turkey when the first case was officially confirmed on 12 March 2020. The first death due to COVID-19 in the Turkey has occurred on 17 March 2020 whereas the first official recoveries against to pandemic virus has been seen on 26 March 2020. On 20 April 2020, the Ministry of Health confirmed that the total number of cases increased to 90980 and that the death toll reached 2140. The total number of tests performed as of 20 April was 673980. This data set consists of 25 observations which are calculated as the ratio of daily total number of the recoveries to total number of confirmed cases in Turkey from 27 March to 20 April. This means that the daily percentage of people recovering in the total number of cases. The data are given as data set-III below. This data set and related descriptions are available in the Turkey's health ministry, WHO and Wikipedia with the following links respectively: https://covid19.saglik.gov.tr/, https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports   and https://en.wikipedia.org/wiki/2020_coronavirus_pandemic_in_Turkey.
All data sets are given as follows:

Data Set I

0.04, 0.02, 0.06, 0.12, 0.14, 0.08, 0.22, 0.12, 0.08, 0.26, 0.24, 0.04, 0.14, 0.16, 0.08, 0.26, 0.32, 0.28, 0.14, 0.16, 0.24, 0.22, 0.12, 0.18, 0.24, 0.32, 0.16, 0.14, 0.08, 0.16, 0.24, 0.16, 0.32, 0.18, 0.24, 0.22, 0.16, 0.12, 0.24, 0.06, 0.02, 0.18, 0.22, 0.14, 0.06, 0.04, 0.14, 0.26, 0.18, 0.16.

Data Set II

0.06, 0.12, 0.14, 0.04, 0.14, 0.16, 0.08, 0.26, 0.32, 0.22, 0.16, 0.12, 0.24, 0.06, 0.02, 0.18, 0.22, 0.14, 0.22, 0.16, 0.12, 0.24, 0.06, 0.02, 0.18, 0.22, 0.14, 0.02, 0.18, 0.22, 0.14, 0.06, 0.04, 0.14, 0.22, 0.14, 0.06, 0.04, 0.16, 0.24, 0.16, 0.32, 0.18, 0.24, 0.22, 0.04, 0.14, 0.26, 0.18, 0.16.

Data Set III

0.007371007, 0.009456904, 0.011325639, 0.014962594, 0.017958761, 0.021238599, 0.022883926, 0.023134649, 0.032840311, 0.038494218, 0.043882583, 0.046380721, 0.048291739, 0.050659855, 0.051521402, 0.056836698, 0.060502844, 0.064816787, 0.073704904, 0.081767351, 0.095548098, 0.109884654, 0.126966197, 0.138762079, 0.14761486.

We give the summary statistics of the data sets in Table 1. The three data sets are right skewed however the second data set is close to symmetrical distribution.

Table 1: Some summary statistics of the data sets

Data set	Mean	Median	St. Dev.	Skewness	Kurtosis
I	0.1632	0.1600	0.0810	0.0723	2.2166
II	0.1520	0.1600	0.0785	0.0061	2.3012
III	0.0559	0.0483	0.0408	0.8645	2.7781

Tables 2-4 show analyzing results based on the application models. When we see these Tables, UJS_U distribution model can be chosen as the best model since it has the smallest values of the AIC, BIC, W^* and A^* statistics. Further, Figures 8, 9 and 10 show all fitted densities and their cdfs for the data set I, data set II and data set III respectively. Hence, we observe that the UJS_U fitting has successfully captured shape of the empirical data, skewness and kurtosis for all data sets. Furthermore, we give parameter estimation results and goodness-of-fits statistics of the UJS_U distribution based on other estimation methods in Table 5. These results are better than the best results of the Tables 2-4 according to goodness-of-fits statistics. Consequently, the fitting results of the UJS_U distribution to three data sets is not ignored.

Table 2: MLEs, standard erros of the estimates (in parentheses), $\hat{\ell}$ and goodness-of-fits statistics for the data set I (p value is given in $[\cdot]$)

Model	$\hat{\mu}_{MLE}$	$\hat{\sigma}_{MLE}$	$\hat{\ell}$	AIC	BIC	A^*	W^*	KS
UJS_U	3.7875	2.9152	56.3943	-108.7886	-104.9645	0.5914	0.0970	0.1121
	(0.4046)	(0.2917)						[0.5555]
Beta	2.6824	13.8640	54.6066	-105.2133	-101.3893	0.9120	0.1538	0.1414
	(0.3025)	(1.3241)						[0.2699]
TL	0.7248	-	28.4078	-54.8156	-52.9036	8.2642	1.6471	0.3623
	(0.1025)							[0.000004]
UGom	0.0924	1.0747	40.6720	-77.3440	-73.5199	3.0810	0.5213	0.2046
	(0.0405)	(0.1413)						[0.0304]
Johnson S_B	2.3767	1.3175	51.3231	-98.6461	-94.8221	1.4647	0.2495	0.1740
	(0.2765)	(0.1317)						[0.0968]
Kw	2.0773	33.1338	56.0686	-108.1374	-104.3133	0.6777	0.1042	0.1103
	(0.1961)	10.2504						[0.5774]
UIG	0.2801	0.1632	46.2557	-88.5115	-84.6875	2.7699	0.5342	0.2333
	(0.0502)	(0.0176)						[0.00865]
LWE	0.0003	1.0065	31.8265	-59.6530	-55.8290	6.5681	1.2269	0.3182
	(0.3570)	(0.2082)						[0.00008]

Table 3: MLEs, standard erros of the estimates (in parentheses), $\hat{\ell}$ and goodness-of-fits statistics for the data set II (p value is given in [·])

Model	$\hat{\mu}_{MLE}$	$\hat{\sigma}_{MLE}$	$\hat{\ell}$	AIC	BIC	A^*	W^*	KS
UJS_U	3.8954	2.8876	57.9383	-111.8765	-108.0525	1.1141	0.1964	0.1698
	(0.4146)	(0.2889)						[0.1118]
Beta	2.4001	13.5204	55.9312	-107.8624	-104.0384	1.5649	0.2871	0.1981
	(0.2632)	(1.2381)						[0.0395]
TL	0.6805	-	30.4331	-58.8663	-56.9543	8.2380	1.6446	0.3771
	(0.0962)							[0.000001]
UGom	0.0916	1.0250	42.6099	-81.2198	-77.3957	3.4278	0.6091	0.2312
	(0.0411)	(0.1393)						[0.0095]
Johnson S_B	2.3682	1.2374	52.3050	-100.6100	-96.7860	2.2458	0.4145	0.2285
	(0.2760)	(0.1237)						[0.0108]
Kw	1.9606	31.3769	57.5214	-111.0429	-107.2189	1.2683	0.2093	0.1691
	(0.1797)	9.1912						[0.1145]
UIG	0.2206	0.1520	47.4751	-90.9501	-87.1261	3.6572	0.7335	0.2853
	(0.0441)	(0.0178)						[0.0006]
LWE	0.0003	0.9601	34.1277	-64.2554	-60.4314	6.4984	1.2165	0.3309
	(0.3739)	(0.2010)						[0.00003]

Table 4: MLEs, standard erros of the estimates (in parentheses), $\hat{\ell}$ and goodness-of-fits statistics for the data set III (p value is given in [·])

Model	$\hat{\mu}_{MLE}$	$\hat{\sigma}_{MLE}$	$\hat{\ell}$	AIC	BIC	A^*	W^*	KS
UJS_U	6.8239	3.7425	49.5364	-95.0728	-92.6351	0.2023	0.0267	0.1008
	(0.9936)	(0.5335)						[0.9400]
TL	0.3972	-	30.7211	-59.4423	-58.2235	5.3422	1.0836	0.4026
	(0.0794)							[0.00036]
Beta	1.7470	29.5386	49.4429	-94.8859	-92.4482	0.2042	0.0271	0.1025
	(0.3526)	(5.9653)						[0.9314]
UGom	0.0186	1.1145	46.5550	-89.1101	-86.6723	0.6500	0.1045	0.1549
	(0.0137)	(0.1774)						[0.5351]
Johnson S_B	3.6014	1.1529	49.0937	-94.1875	-91.7497	0.2876	0.0470	0.1194
	(0.5475)	(0.1631)						[0.8272]
Kw	1.4159	50.8893	49.3421	-94.6843	-92.2465	0.2126	0.0277	0.1016
	(0.1199)	13.4381						[0.9354]
UIG	0.0601	0.0559	48.6948	-93.3895	-90.9518	0.4877	0.0917	0.1584
	(0.0170)	(0.0108)						[0.5070]
LWE	0.00002	0.0628	34.4191	-64.8383	-62.1621	3.9935	0.7524	0.3382
	(0.3392)	(0.1574)						[0.0047]

Figure 8: Estimated densities and cdfs for the data set-I

Pak.j.stat.oper.res. Vol.16 No. 3 2020 pp 471-490
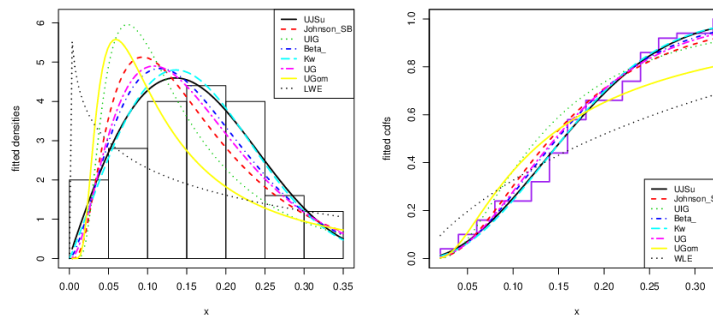
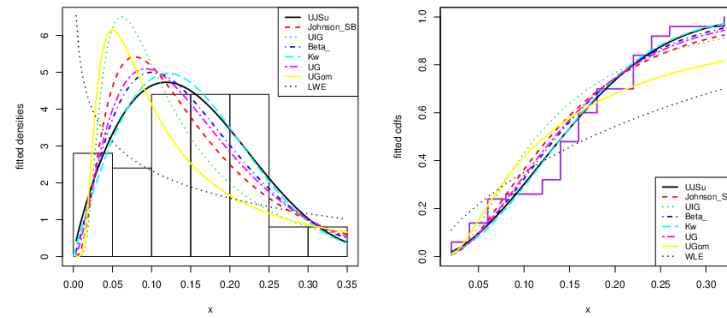DOI: http://dx.doi.org/10.18187/pjsor.v16i3.3421

Figure 9: Estimated densities and cdfs for the data set-II
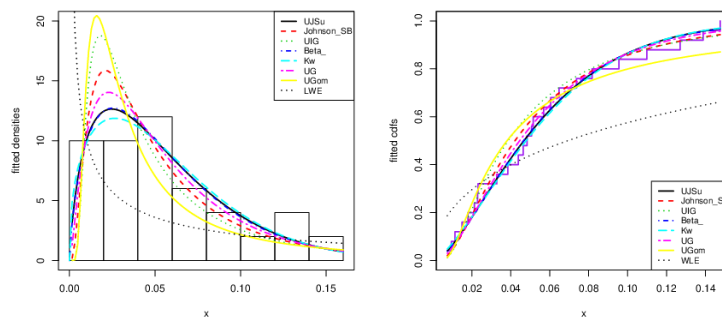

Figure 10: Estimated densities and cdfs for the data set-III

Table 5: The results of the UJS_U model parameters based on the different estimation methods for the data sets

	Data set-I					Data set-II					Data set III				
Method	$\hat{\mu}$	$\hat{\sigma}$	A^*	W^*	KS	$\hat{\mu}$	$\hat{\sigma}$	A^*	W^*	KS	$\hat{\mu}$	$\hat{\sigma}$	A*	W^*	KS
MPS	3.5752	2.7500	0.5502	0.0818	0.0949	3.6816	2.7287	1.0712	0.1529	0.1337	6.1302	3.3628	0.1993	0.0253	0.0781
LSE	3.5345	2.7615	0.5371	0.0834	0.1024	3.6645	2.7844	1.0258	0.1599	0.1492	5.9902	3.2785	0.1897	0.0244	0.0758
WLSE	3.6118	2.8040	0.5640	0.0963	0.1168	3.7422	2.8129	1.0456	0.1903	0.1706	6.0544	3.3164	0.1817	0.0239	0.0773
AD	3.5916	2.7879	0.5366	0.0838	0.1031	3.6080	2.7042	1.0028	0.1655	0.1546	6.3984	3.5043	0.1690	0.0226	0.0832
CVM	3.6477	2.8509	0.5576	0.0805	0.0916	3.8007	2.8911	1.1319	0.1515	0.1302	6.3668	3.4845	0.1694	0.0225	0.0806

6 Concluding remarks

A new alternative unit distribution belongs to exponential family is introduced. Statistical properties are studied in detail for newly defined distribution. For the model parameters, the six different estimators have been presented based on the methods of the maximum likelihood, maximum product spacing, least square, weighted least square, Anderson-Darling and Cramer-von-Mises estimation methods. Three simulation studies based on the different model scenarios are performed to illustrate the performances of above estimation methods. Three applications to the real data sets, which one is related to coronavirus data, show that the fitting results of the UJS_U distribution are not ignored. It is hoped that the UJS_U distribution will be a remarkable model in different disciplines as well as applied probability and statistics.

References

1. Altun, E. (2020). The log-weighted exponential regression model: alternative to the beta regression model. Communications in Statistics-Theory and Methods, DOI:10.1080/03610926.2019.1664586.
2. Altun, E. and Cordeiro, G. M. (2020). The unit-improved second-degree lindley distribution: inference and regression modeling. Computational Statistics, 35(1):259–279.
3. Altun, E. and Hamedani, G. (2018). The log-xgamma distribution with inference and application. Journal de la Société Française de Statistique, 159(3):40–55.
4. Anderson, T. W. and Darling, D. A. (1952). Asymptotic theory of certain" goodness of fit" criteria based on stochastic processes. The annals of mathematical statistics, pages 193–212.
5. Arnold, B. C. and Groeneveld, R. A. (1980). Some properties of the arcsine distribution. Journal of the American Statistical Association, 75(369):173–175.
6. Barndorff-Nielsen, O. E. and Jørgensen, B. (1991). Some parametric models on the simplex. Journal of multivariate analysis, 39(1):106–116.
7. Castagliola, P. (1998). Approximation of the normal sample median distribution using symmetrical Johnson su distributions: application to quality control. Communications in Statistics-Simulation and Computation, 27(2):289–301.
8. Chen, G. and Balakrishnan, N. (1995). A general purpose approximate goodness-of-fit test. Journal of Quality Technology, 27(2):154–161.
9. Chen, H. and Kamburowska, G. (2001). Fitting data to the Johnson system. Journal of Statistical Computation and Simulation, 70(1):21–32.
10. Cheng, R. and Amin, N. (1979). Maximum product of spacings estimation with application to the lognormal distribution (mathematical report 79-1). Cardiff: University of Wales IST.
11. Consul, P. C. and Jain, G. C. (1971). On the log-gamma distribution and its properties. Statistische Hefte, 12(2):100–106.
12. Cordeiro, G. M. and Lemonte, A. J. (2014). The mcdonald arcsine distribution: A new model to proportional data. Statistics, 48(1):182–199.
13. Dasgupta, R. (2011). On the distribution of burr with applications. Sankhya B, 73(1):1–19.
14. Dey, S., Mazucheli, J., and Anis, M. (2017). Estimation of reliability of multicomponent stress–strength for a kumaraswamy distribution. Communications in Statistics-Theory and Methods, 46(4):1560–1572.
15. Dey, S., Mazucheli, J., and Nadarajah, S. (2018). Kumaraswamy distribution: different methods of estimation. Computational and Applied Mathematics, 37(2):2094–2111.
16. Dudewicz, E. J., Zhang, C. X., and Karian, Z. A. (2004). The completeness and uniqueness of Johnson's system in skewness–kurtosis space. Communications in Statistics-Theory and Methods, 33(9):2097–2116.
17. George, F. and Ramachandran, K. (2011). Estimation of parameters of Johnson's system of distributions. Journal of Modern Applied Statistical Methods, 10(2):9.
18. Ghitany, M., Mazucheli, J., Menezes, A., and Alqallaf, F. (2019). The unit-inverse gaussian distribution: A new alternative to two-parameter distributions on the unit interval. Communications in Statistics-Theory and Methods, 48(14):3423–3438.
19. Gomez-Déniz, E., Sordo, M. A., and Calderʹın-Ojeda, E. (2014). The log–lindley distribution as an alternative to the beta regression model with applications in insurance. Insurance: Mathematics and Economics, 54:49–57.
20. Hill, I., Hill, R., and Holder, R. (1976). Algorithm as 99: Fitting Johnson curves by moments. Journal of the royal statistical society Series C (Applied statistics), 25(2):180–189.
21. Johnson, N. (1955). Systems of frequency curves derived from the first law of laplace. Trabajos de estadʹıstica, 5(3):283–291.
22. Johnson, N. L. (1949). Systems of frequency curves generated by methods of translation. Biometrika, 36(1/2):149–176.
23. Korkmaz, M. Ç . (2020a). A new heavy-tailed distribution defined on the bounded interval: the logit slash distribution and its application. Journal of Applied Statistics, 47(12):2097–2119.
24. Korkmaz, M. Ç. (2020b). The unit generalized half normal distribution: A new bounded distribution with inference and application. Unıversity Politehnica of Bucharest Scientific Bulletin-Series A-Applied Mathematics And Physics, 82(2):133–140.

25. Korkmaz, M. Ç. and Erişoğlu, M. (2014). The Burr XII-geometric distribution. Journal of Selçuk University Natural and Applied Science, 3(4):75–87.
26. Korkmaz, M. Ç. and Genç, A. İ. (2017). A new generalized two-sided class of distributions with an emphasis on two-sided generalized normal distribution. Communications in Statistics-Simulation and Computation, 46(2):1441–1460.
27. Kullback, S. (1997). Information theory and statistics. Courier Corporation.
28. Kumaraswamy, P. (1980). A generalized probability density function for double-bounded random processes. Journal of hydrology, 46(1-2):79–88.
29. Mazucheli, J., Menezes, A. F., and Dey, S. (2018). The unit-birnbaum-saunders distribution with applications. Chilean Journal of Statistics, 9(1):47–57.
30. Mazucheli, J., Menezes, A. F., and Dey, S. (2019a). Unit-Gompertz distribution with applications. Statistica, 79(1):25–43.
31. Mazucheli, J., Menezes, A. F. B., and Chakraborty, S. (2019b). On the one parameter unit-lindley distribution and its associated regression model for proportion data. Journal of Applied Statistics, 46(4):700–714.
32. McDonald, J. B. (1984). Some generalized functions for the size distribution of income. Econometrika, 52(3):647–663.
33. Ranneby, B. (1984). The maximum spacing method. an estimation method related to the maximum likelihood method. Scandinavian Journal of Statistics, pages 93–112.
34. Shaked, M. and Shanthikumar, J. G. (2007). Stochastic orders. Springer Science & Business Media.
35. Slifker, J. F. and Shapiro, S. S. (1980). The Johnson system: selection and parameter estimation. Technometrics, 22(2):239–246.
36. Tadikamalla, P. R. and Johnson, N. L. (1982). Systems of frequency curves generated by transformations of logistic variables. Biometrika, 69(2):461–465.
37. Topp, C. W. and Leone, F. C. (1955). A family of J-shaped frequency functions. Journal of the American Statistical Association, 50(269):209–219.
38. Tuenter, H. J. (2001). An algorithm to determine the parameters of S_U-curves in the Johnson system of probabillity distributions by moment matching. Journal of Statistical Computation and Simulation, 70(4):325–347.
39. van Dorp, J. R. and Jones, M. (2020). The Johnson system of frequency curvesˆahistorical, graphical, and limiting perspectives. The American Statistician, 74(1):37–52.
40. van Dorp, J. R. and Kotz, S. (2002). The standard two-sided power distribution and its properties: with applications in financial engineering. The American Statistician, 56(2):90–99.
41. Venkataraman, S. V. and Rao, S. N. (2016). Estimation of dynamic var using JS_U and PIV distributions. Risk Management, 18(2-3):111–134.
42. Wheeler, R. E. (1980). Quantile estimators of Johnson curve parameters. Biometrika, pages 725–728.
43. ZeinEldin, R. A., Chesneau, C., Jamal, F., and Elgarhy, M. (2019). Different estimation methods for type I half-logistic Topp–Leone distribution. Mathematics, 7(10):985.