THE UNIT GENERALIZED HALF NORMAL DISTRIBUTION: A NEW BOUNDED DISTRIBUTION WITH INFERENCE AND APPLICATION

Mustafa Ç. Korkmaz[1]

In this paper, a new distribution defined on (0,1) bounded interval has been introduced. We define the distribution by transformation of a positive random variable (rv) with respect to exponential function. Basic distributional properties of newly defined distribution are studied. We pointed out the different estimation methods for its parameters' estimations. We assess the performance of the estimators of these estimation methods by the simulation study. Application of the proposed distribution to a real data set shows better fit than many known distributions on the (0,1) interval.

Keywords: unit distributions, unit generalized half normal distribution, different estimation methods, beta distribution

MSC2010: 62E05, 62F10

1. Introduction

The unit distributions defined on the unit interval are applied to model the behavior of rvs limited to intervals of (0,1) length and they have found applications in fields like health, biology, meteorology, hydrology, financial modeling and other sciences. The beta distribution is the well-known statistical distribution to model data sets on the (0,1) interval and is a convenient and useful model in many areas of statistics. However, its data modeling ability may be insufficient to explain the data. So, alternative distributions to the beta distribution have been defined and applied in the literature such as Johnson S_B [4], Topp-Leone [12], Kumaraswamy (Kw) [6], standart two-sided power (STSP) [11], exponentiated Topp Leone (ETL) [8], unit inverse Gaussian (UIG) [3] and logit slash [5] distributions. The goal of this paper is to introduce the new alternative distribution defined on the (0,1) interval based on the transformation of the generalized half normal (GHN) distribution [2]. We are also motivated to introduce the new distribution because (i) it is capable of modeling increasing, modified-bathtub (N-shaped) and then bathtub shaped hazard rate; and (ii) the proposed distribution has distinguished properties of the shapes and provides better fits than some well known unit distributions. Its some basic properties have been obtained. We consider the different estimations procedures for its the model parameters. An application of the model to a real data set is presented and is compared to the fit attained by some other well-known distributions on the (0,1) interval. The paper is ended with conclusion remarks.

2. Unit Generalized Half Normal (UGHN) distribution

Let the rv Y follow a GHN distribution with probability density function (pdf) and cumulative distribution function (cdf)

$$f_{GHN}(y,\alpha,\beta) = 2\alpha y^{\alpha-1}\beta^{-\alpha}\phi\left[\left(\frac{y}{\beta}\right)^{\alpha}\right]$$

[1] Assoc. Professor, Artvin Çoruh University, Turkey, e-mail: mcagatay@artvin.edu.tr

and

$$F_{GHN}(y,\alpha,\beta)=2\Phi\left[\left(\frac{y}{\beta}\right)^{\alpha}\right]-1=1-2\Phi\left[-\left(\frac{y}{\beta}\right)^{\alpha}\right]$$

respectively, where $y>0$, $\alpha,\beta>0$, $\phi[\cdot]$ is the pdf of standard normal distribution and $\Phi[\cdot]$ is the cdf of standard normal distribution. Most of the statistical properties of the GHN distribution were obtained by [2]. For example, [2] obtained the rth moments of this distribution as

$$E(Y^r)=\frac{\beta^r}{\sqrt{\pi}2^{r/\alpha}}\Gamma\left(\frac{\alpha-r}{2\alpha}\right),$$

where $\Gamma(\cdot)$ is the gamma function. A rv X has a unit GHN distribution with shape parameter α and scale parameter β, if its pdf is given by for $x\in(0,1)$ and $\alpha,\beta>0$

$$f(x,\alpha,\beta)=\sqrt{\frac{2}{\pi}}\left(\frac{\alpha}{-x\log x}\right)\left(\frac{-\log x}{\beta}\right)^{\alpha}e^{-\frac{1}{2}\left(\frac{-\log x}{\beta}\right)^{2\alpha}}=\frac{2\alpha(-\log x)^{\alpha-1}}{x\,\beta^{\alpha}}\phi\left[\left(\frac{-\log x}{\beta}\right)^{\alpha}\right]. \tag{1}$$

The new pdf can be obtained with transformation of the $X=e^{-Y}$ rv, where Y has GHN rv. On the other word, a rv X is distributed unit GHN distribution on the interval (0,1) if its log transformation, $-\log x$, is distributed $GHN(\alpha,\beta)$. We denote it with $UGHN(\alpha,\beta)$. For $\alpha=1$, unit half normal distribution is obtained. The corresponding cdf is given by:

$$F(x,\alpha,\beta)=2-2\Phi\left[\left(\frac{-\log x}{\beta}\right)^{\alpha}\right]=2\Phi\left[-\left(\frac{-\log x}{\beta}\right)^{\alpha}\right]=1-\operatorname{erf}\left(\left[\frac{-\log(x)}{\beta}\right]^{\alpha}/\sqrt{2}\right), \tag{2}$$

where $0<x<1$ and

$$\operatorname{erf}(z)=\frac{2}{\sqrt{\pi}}\int_0^z\exp\left(-t^2\right)dt$$

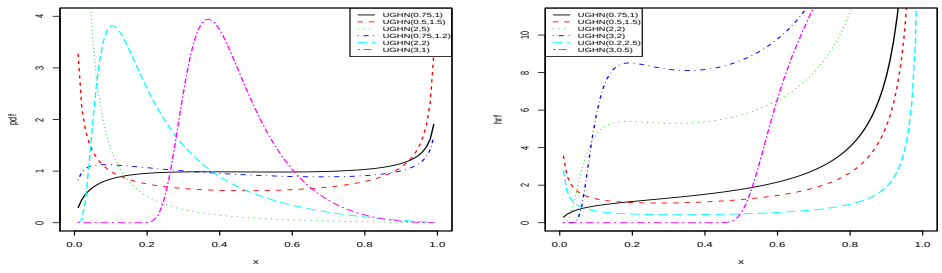
is the error function. It is clearly seen that

$$F(x,\alpha,\beta)=1-F_{GHN}(-\log x,\alpha,\beta).$$

Using (1) and (2), the hazard rate function (hrf) of UGHN distribution is given by

$$h(x,\alpha,\beta)=f(x,\alpha,\beta)/[1-F(x,\alpha,\beta)].$$

Analytically, it may be difficult to identify signs of $\partial f(x,\alpha,\beta)/\partial x$ and $\partial h(x,\alpha,\beta)/\partial x$. So, we sketched the plots of the pdf and hrf to see their possible shapes. Figure 1 indicates the possible pdf and hrf shapes and their shape regions of UGHN distribution for changed α and β parameters. As seen from Figure 1, the shapes of the pdf can be decreasing, increasing, uni-modal, U-shaped and N-shaped as well as the shapes of the hrf can be increasing, N-shaped and bathtub shaped.



(A) pdf shapes (B) hrf shapes

FIGURE 1. The possible pdf and hrf shapes $UGHN$ distribution for selected parameters values

2.1. Moments

The rth raw moment of the UGHN distribution is given by

$$\mu'_r = E\left(X^r\right) = E\left(e^{-rY}\right) = \sum_{i=0}^{\infty} \frac{(-r)^i}{i!} E\left(Y^i\right) = \sum_{i=0}^{\infty} \frac{(-r)^i}{i!} \beta^i \sqrt{\frac{2^{i/\alpha}}{\pi}} \Gamma\left(\frac{i+\alpha}{2\alpha}\right).$$

As it can be seen, μ'_r's can not be expressed in a closed form. The numerical integration can be applied to obtain the mean and other important related measures. The j^{th} order central moment can be obtained by the following relationship

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

With this formula, the skewness and kurtosis coefficients are respectively given by

$$\sqrt{\beta_1} = \mu_3 \mu_2^{-3/2}$$

and

$$\sqrt{\beta_2} = \mu_4 \mu_2^{-2}.$$

However, these above calculations can be easily computed using many packet programs such as R, Matlab, Maple and Wolfram.

2.2. Stochastic ordering

Let us denote the pdf, cdf, hrf and mean residual life function (mrl) of a positive continuous rv X by $f_X(\cdot), F_X(\cdot), h_X(\cdot)$ and $m_X(\cdot)$, respectively, and those of another positive continuous rv Y by $f_Y(\cdot), F_Y(\cdot), h_Y(\cdot)$ and $m_Y(\cdot)$, respectively. We recall some basic definitions.

A rv X is said to be smaller than a rv Y in the

(i) The stochastic order $(X \leq_{(sto)} Y)$ if $F_X(x) \geq F_Y(x), \forall\ x$.
(ii) The hazard rate order $(X \leq_{(hro)} Y)$ if $h_X(x) \geq h_Y(x), \forall\ x$.
(iii) The mean residual life order $(X \leq_{(mrlo)} Y)$ if $m_X(x) \leq m_Y(x), \forall\ x$.
(iv) The likelihood ratio order $(X \leq_{(lro)} Y)$ if $\frac{f_X(x)}{f_Y(x)}$ decreases in x.

The below given implications (see [9]) are well justified:

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

The following Proposition shows that the UGHN distributions are ordered with respect to different stochastic orderings.

Proposition 2.1. *Let $X \sim UGHN(\alpha, \beta_1)$ and $Y \sim UGHN(\alpha, \beta_2)$. If $\beta_2 < \beta_1$, then $\left[X \leq_{(lro)} Y\right]$ and $\left[X \leq_{(hro)} Y\right]$, $\left[X \leq_{(mrlo)} Y\right]$, $\left[X \leq_{(sto)} Y\right]$.*

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

$$g(x) = \frac{f_X(x)}{f_Y(x)} = \left(\frac{\beta_1}{\beta_2}\right)^{\alpha} \exp\left[-\frac{(-\log x)^{2\alpha}}{2}\left(\frac{1}{\beta_1^{2\alpha}} - \frac{1}{\beta_2^{2\alpha}}\right)\right].$$

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

$$\frac{\partial \log g(x)}{\partial x} = \frac{\alpha}{x(-\log x)}\left[\left(\frac{-\log x}{\beta_1}\right)^{\alpha} - \left(\frac{-\log x}{\beta_2}\right)^{\alpha}\right]\left[\left(\frac{-\log x}{\beta_1}\right)^{\alpha} + \left(\frac{-\log x}{\beta_2}\right)^{\alpha}\right].$$

If $\beta_2 < \beta_1$, then $\frac{\partial \log g(x)}{\partial x} \leq 0$ which purposes that $\left[X \leq_{(lro)} Y\right]$ and $\left[X \leq_{(hro)} Y\right]$, $\left[X \leq_{(mrlo)} Y\right]$, $\left[X \leq_{(sto)} Y\right]$. Hence, the proof is completed. □

2.3. Order statistics

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

$$\begin{aligned} f\left(x_{(r)}, \alpha, \beta\right) &= \frac{\alpha(-\log x)^{\alpha-1} n!}{x\beta^{\alpha}(r-1)!(n-r)!}\phi\left[\left(\frac{-\log x}{\beta}\right)^{\alpha}\right] \\ &\quad \sum_{k=0}^{n-r}(-1)^k \binom{n-r}{k} 2^{r+k}\left[\left(\Phi\left[-\left(\frac{-\log x}{\beta}\right)^{\alpha}\right]\right)\right]^{r+k-1} \end{aligned}$$

respectively, where $r = 1, 2, \ldots, n$. For $r = 1$ and $r = n$, we have the pdf of the $X_{(1)} = min\{X_1, X_2, ..., X_n\}$ and $X_{(n)} = max\{X_1, X_2, ..., X_n\}$, respectively.

3. Different Estimation Methods

In this section, we point out various estimators for estimating the unknown parameters of the $UGHN$ distribution.

3.1. Maximum likelihood estimation

In this subsection, we estimate the parameters of the $UGHN$ distribution by the method of maximum likelihood estimation (MLE). Let $X_1, X_2, \ldots, X_n$ be a random sample from the $UGHN$ distribution and

$$\boldsymbol{\Xi} = (\alpha, \beta)^T$$

be the vector of the model parameters. The log-likelihood function for $\boldsymbol{\Xi}$ may be expressed as

$$\begin{aligned} \ell(\boldsymbol{\Xi}) &= \tfrac{n}{2}\log\left(\tfrac{2}{\pi}\right) + n\log\alpha - n\alpha\log\beta - \sum_{i=1}^{n}\log x_i + (\alpha-1)\sum_{i=1}^{n}\log(-\log x_i) \\ &\quad -\tfrac{1}{2\beta^{2\alpha}}\sum_{i=1}^{n}(-\log x_i)^{2\alpha}. \end{aligned} \tag{3}$$

The MLEs, $\hat{\alpha}_{MLE}$ and $\hat{\beta}_{MLE}$, are obtained by maximizing $\ell(\boldsymbol{\Xi})$.

3.2. Maximum product spacing (MPS) estimation

The MPS method has been proposed by [1] as approximation to the Kullback-Leibler measure of information and is based on the idea that maximizing the geometric mean of the differences. The MPS estimators (MPSE), $\hat{\alpha}_{MPS}$ and $\hat{\beta}_{MPS}$,of the α and β are obtained by maximizing the geometric mean of the differences. The $\hat{\alpha}_{MPS}$ and $\hat{\beta}_{MPS}$ are obtained by maximizing

$$MPS\left(\boldsymbol{\Xi}\right)=\frac{1}{n+1}\sum_{i=1}^{n+1}\log\left[F(x_{(i)},\alpha,\beta)-F(x_{(i-1)},\alpha,\beta)\right]. \tag{4}$$

3.3. Least squares estimates

Let $X_{(1)},X_{(2)},\ldots,X_{(n)}$ be ordered sample of size n from UGHN distribution. Then, the expectation of the empirical cumulative distribution function is defined as

$$E\left[F\left(x_{(i)}\right)\right]=\frac{i}{n+1};i=1,2,\ldots,n.$$

The least square estimates (LSEs) say, $\hat{\alpha}_{LSE}$ and $\hat{\beta}_{LSE}$, of α and β are obtained by minimizing

$$QLSE\left(\boldsymbol{\Xi}\right)=\sum_{i=1}^{n}\left(F(x_{(i)},\alpha,\beta)-\frac{i}{n+1}\right)^{2}. \tag{5}$$

3.4. Weighted least squares estimates

Let $X_{(1)},X_{(2)},\ldots,X_{(n)}$ be ordered sample of size n from UGHN distribution. The variance of the empirical cumulative distribution function is defined as

$$V\left[F\left(x_{(i)}\right)\right]=\frac{i(n-i+1)}{(n+2)(n+1)^{2}};i=1,2,\ldots,n.$$

Then, the weighted least square estimates (WLSEs) say, $\hat{\alpha}_{WLSE}$ and $\hat{\beta}_{WLSE}$, of α and β are obtained obtained by minimizing

$$QWLSE\left(\boldsymbol{\Xi}\right)=\sum_{i=1}^{n}\frac{\left(F(x_{(i)},\alpha,\beta)-\frac{i}{n+1}\right)^{2}}{V\left[F\left(x_{(i)}\right)\right]}. \tag{6}$$

3.5. Anderson-Darling Estimation

This estimator is based on Anderson-Darling goodness-of-fits statistics. The Anderson-Darling (AD) minimum distance estimates, $\hat{\alpha}_{AD}$ and $\hat{\beta}_{AD}$, of α and β are obtained by minimizing

$$AD\left(\boldsymbol{\Xi}\right)=-n-\sum_{i=1}^{n}\tfrac{2i-1}{n}\left[\log F\left(x_{(i)},\alpha,\beta\right)+\log\left\{1-F\left(x_{(n+1-i),},\alpha,\beta\right)\right\}\right]. \tag{7}$$

3.6. The Cramer-von Mises Estimations

The Cramer-von Mises (CVM) minimum distance estimates, $\hat{\alpha}_{CVM}$ and $\hat{\beta}_{CVM}$, of α and β are obtained by minimizing

$$CVM\left(\boldsymbol{\Xi}\right)=\tfrac{1}{12n}+\sum_{i=1}^{n}\left[F\left(x_{(i)},\alpha,\beta\right)-\frac{2i-1}{2n}\right]^{2}. \tag{8}$$

To solve above equations, Equations (3), (4), (5), (6), (7) and (8) can be optimized either directly by using the `R` (`optim` and `maxLik` functions), `SAS` or `Ox` program or it can be used

nonlinear optimization methods such as the quasi-Newton algorithm to numerically optimize $\ell(\boldsymbol{\Xi})$, $MPS(\boldsymbol{\Xi})$, $QLSE(\boldsymbol{\Xi})$, $QWLSE(\boldsymbol{\Xi})$, $AD(\boldsymbol{\Xi})$ and $CVM(\boldsymbol{\Xi})$ functions.

4. Simulation experiments

In this Section, we perform a simulation study based on graphical results. We generate $N = 1000$ samples of size $n = 20, 25, \ldots, 1000$ from UGHN distribution with true parameter values $\alpha = 5$ and $\beta = 0.5$.Notice that, we can obtain random number from $UGHN$ distribution by the following algorithm. We can give this procedure as:

(i) set α and β,

(ii) simulate $U \sim Uniform(0,1)$,

(iii) compute $Y = -\beta\left[-\Phi^{-1}\left(\frac{U}{2}\right)\right]^{1/\alpha}$, then Y follows that $GHN(\alpha,\beta)$,

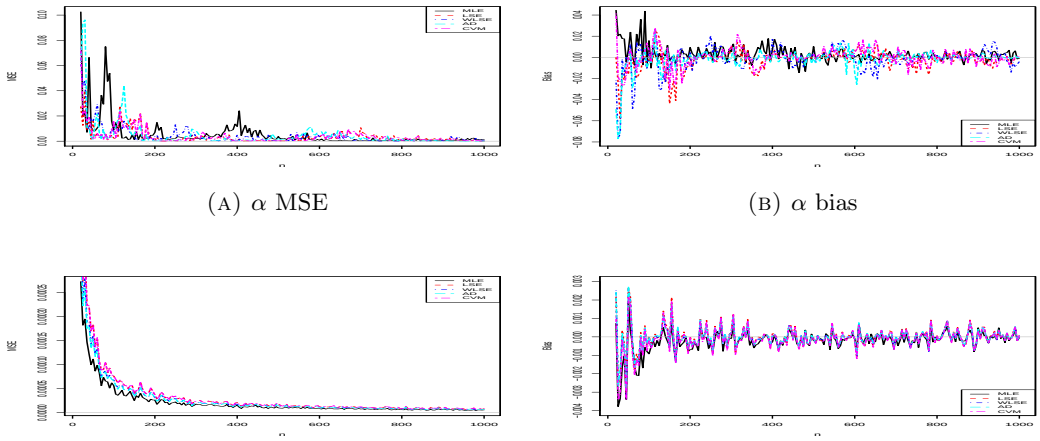
(iv) compute $X = e^{-Y}$, then X follows that $UGHN(\alpha,\beta)$. We compare above all estimators based on the empirical biases and mean square errors (MSEs) under varying sample size. The empirical bias and MSE are calculated by (for $h = \alpha, \beta$)

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

and

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

respectively. The results of this simulation study are shown in Figure 2. This Figure shows that all estimators are to be consistent since the MSE and biases decrease with increasing sample size. It is clear that the estimates of parameters are asymptotically unbiased. All empirical means are close true values. Hence, we can say that the performances of all estimators are close.



(A) α MSE (B) α bias

(C) β MSE (D) β bias

FIGURE 2. Simulation results

5. Data Analysis

In this section, we consider an application to real data set to show the modeling ability of the UGHN distribution. The data set consists of the failure times of 20 mechanical components given in [7]. Recently, this data was analyzed by [10]. Under MLE method, we fit the UGHN distribution to this data set and compare it with the beta, Johnson S_B, Kw,

Table 1.

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

Model	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\ell}$	AIC	BIC	A^*	W^*	KS
UGHN	6.3815	2.4225	**37.4527**	**-70.9055**	**-68.9140**	**0.5181**	**0.0619**	**0.1283**
	(1.2270)	(0.0636)						[**0.8971**]
Beta	3.1126	21.8245	27.8813	-51.7626	-49.7711	2.2611	0.3726	0.2537
	(1.0287)	(7.7997)						[0.1521]
Kw	1.5877	21.8673	25.6484	-47.2968	-45.3054	2.6889	0.4681	0.2626
	(0.3966)	17.9755						[0.1265]
ETL	1.7369	9.7109	26.1136	-48.2272	-46.2357	2.6147	0.4523	0.2640
	(0.3011)	(4.0910)						[0.1229]
STSP	0.0680	15.0169	35.5132	-67.0264	-65.0350	0.9698	0.1376	0.1758
	(0.0004)	(3.4249)						[0.5667]
Johnson S_B	3.8952	1.8605	31.3599	-58.7198	-56.7283	1.5531	0.2307	0.2039
	(0.6554)	(0.2942)						[0.3765]
UIG	0.5803	0.1215	33.0756	-62.1511	-60.1597	1.3777	0.2143	0.1991
	(0.1835)	(0.0125)						[0.4059]

STSP, ETL and UIG distributions based on 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. All computations are performed by the **maxLik** routine in the R program. We give the estimates and the values of goodnes-of-fits statistics in Table 1. When we see this Table, UGHN model can be chosen as the best model since it has the optimal values of all criteria. Consequently, the UGHN distribution provides better fit than its competitors.

6. Conclusion

We proposed a new distribution defined on (0,1) interval using a transformation of the GHN distribution. We investigated distributional properties of the new distribution. The model parameters were estimated by some estimation methods. A simulation study was performed to illustrate the performances of all estimators. Its usefulness on data modeling was shown via an application to the real data set. In summary, the proposed model can be an alternative to the classical bounded distributions available in the statistical literature to model rates and proportions.

REFERENCES

[1] Cheng, R. C. H. and Amin, N. A. K. (1979). Maximum product of spacings estimation with application to the lognormal distribution. Math Report, 791.

[2] Cooray, K. and Ananda, M. M. (2008). A generalization of the half–normal distribution with applications to lifetime data. Communications in Statistics–Theory and Methods, 37(9), 1323-1337.

[3] Ghitany, M. E., Mazucheli, J., Menezes, A. F. B. 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.

[4] Johnson, N. L. (1949). Systems of frequency curves generated by methods of translation. Biometrika, 36(1/2), 149-176.

[5] Korkmaz, M. Ç. (2020). A new heavy-tailed distribution defined on the bounded interval: the logit slash distribution and its application. Journal of Applied Statistics, DOI:10.1080/02664763.2019.1704701.

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

[7] Murthy, D. P., Xie, M. and Jiang, R. (2004). Weibull models (Vol. 505). John Wiley & Sons.

[8] Pourdarvish, A., S. M. T. K. Mirmostafaee and K. Naderi. (2015). The exponentiated Topp–Leone distribution: Properties and application. Journal of Applied Environmental and Biological Sciences 251–56.

[9] Shaked, M. and Shanthikumar, J. G. (2007). Stochastic orders. Springer Science & Business Media.

[10] Silva, R. B., Bourguignon, M., Dias, C. R. and Cordeiro, G. M. (2013). The compound class of extended Weibull power series distributions. Computational Statistics & Data Analysis, 58, 352-367.

[11] 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.

[12] 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.