Discriminating Between the Generalized Rayleigh and Log-Normal Distribution

Debasis Kundu[1] and Mohammad Z. Raqab[2]

Abstract

Surles and Padgett recently introduced two-parameter Burr Type X distribution, which can also be described as the generalized Rayleigh distribution. It is observed that the generalized Rayleigh and log-normal distributions have many common properties and both the distributions can be used quite effectively to analyze skewed data set. For a given data set the problem of selecting either generalized Rayleigh or log-normal distribution is discussed in this paper. The ratio of maximized likelihood is used in discriminating between the two distributing functions. Asymptotic distributions of the ratio of the maximized likelihoods under null hypotheses are obtained and they are used to determine the minimum sample size required in discriminating between these two families of distributions for a used specified probability of correct selection and the tolerance limit.

Key Words and Phrases: Burr Type X distribution; Maximum likelihood estimator; Hazard function; Asymptotic distributions; Likelihood Ratio Test; Probability of Correct Selection; Tolerance limit.

Address of correspondence: Debasis Kundu, e-mail: kundu@iitk.ac.in, Phone no. 91-512-2597141, Fax no. 91-512-2597500.

[1] Department of Mathematics and Statistics, Indian Institute of Technology Kanpur, Pin 208016, India.

[2] Department of Mathematics, University of Jordon Amman 11942, JORDON.

1 Introduction

Recently, Surles and Padgett [6] introduced the two-parameter Burr Type X distribution and correctly named as the generalized Rayleigh (GR) distribution. The two-parameter GR distribution for $\alpha > 0$ and $\lambda > 0$ has the distribution function

$$F_{GR}(x;\alpha,\lambda) = \left(1 - e^{-(\lambda x)^2}\right)^{\alpha}; \quad \text{for} \quad x > 0, \tag{1}$$

and the density function

$$f_{GR}(x;\alpha,\lambda) = 2\alpha\lambda^2 x e^{-(\lambda x)^2}\left(1 - e^{-(\lambda x)^2}\right)^{\alpha-1}; \quad \text{for} \quad x > 0. \tag{2}$$

Here α and λ are the shape and scale parameters respectively. From now on the the generalized Rayleigh distribution with the shape parameter α and scale parameter λ will be denoted by $GR(\alpha,\lambda)$. The shapes of the density functions or hazard functions do not depend on λ, they only depend on α. It is known that for $\alpha \le \frac{1}{2}$, the density function is strictly decreasing and for $\alpha > \frac{1}{2}$, it is unimodal. It is also observed that the hazard function of a GR distribution can be either bathtub type or an increasing function depending on the shape parameter α. For $\alpha \le \frac{1}{2}$, the hazard function of $GR(\alpha,\lambda)$ is inverted bathtub type and for $\alpha > \frac{1}{2}$, it has an increasing hazard function.

It is known that the GR density functions are always right skewed and they can be used quite effectively to analyze skewed data set. Among several other distributions two parameter log-normal distribution is also used quite effectively to analyze skewed data set. Log-normal density function is always unimodal and it has the inverted bathtub type hazard function. Shapes of the different log-normal density functions can be found in Johnson, Kotz and Balakrishnan [3]. At least for certain ranges of the parameters, shapes of these two density functions are quite similar. See for example in the Figures 1 and 2 the density and distribution functions of $GR(15,1)$ and $LN(0.1822,1.7620)$ (defined in Section 2), where they are almost indistinguishable.

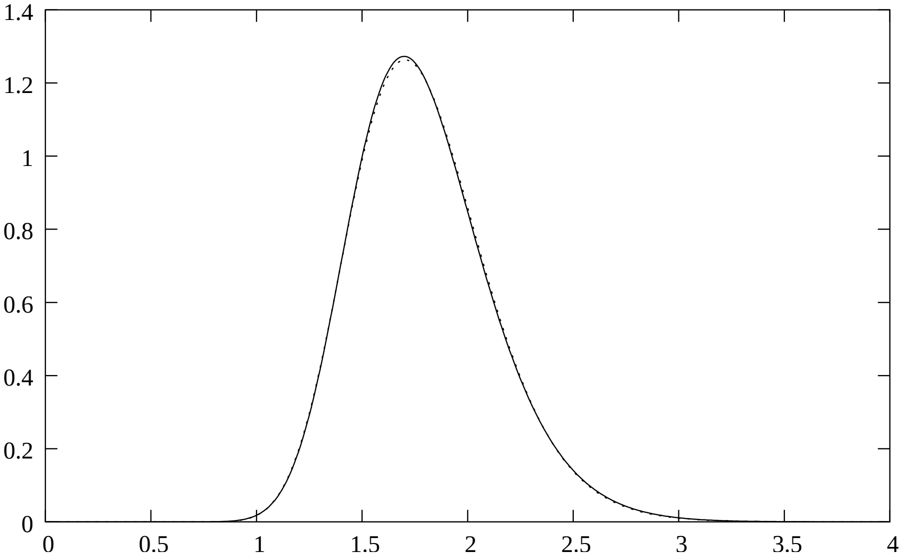

Figure 1: The density functions of $GR(15, 1)$ and $LN(0.1822, 1.7620)$.

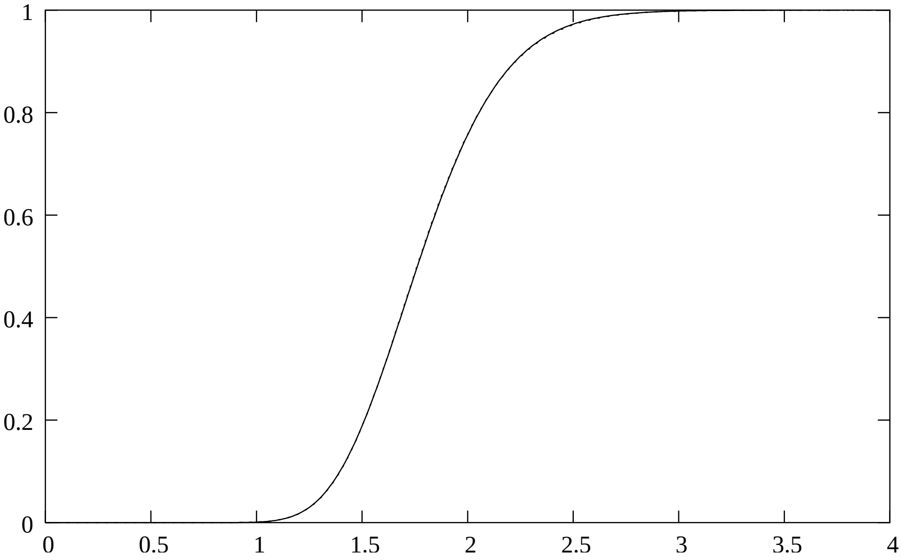

Figure 2: The distribution functions of $GR(15, 1)$ and $LN(0.1822, 1.7620)$.

Although, these two models may provide similar data fit for moderate sample sizes, it is still desirable to choose the correct or nearly correct order model, since the inferences based on a particular model will often involve tail probabilities, where effect of model assumptions will be more crucial. Therefore, even if large sample sizes are not available, it is still very important to make the best possible decision based on the given observation.

In this paper we use the ratio of the maximized likelihoods (RML) in discriminating between the GR and log-normal distributions. Using the basic ideas of White [7, 8] and following similar steps as of Gupta and Kundu [2], the asymptotic distribution of RML has been obtained. It is observed by extensive Monte Carlo simulations that the asymptotic distributions work quite well in discriminating between the two distribution functions, even when the sample size is not very large. Based on the asymptotic distributions and the distance between the two distribution functions, we obtain the minimum sample size required to discriminate between the two distribution functions as a user specified protection level.

The rest of the paper is organized as follows. In Section 2, we obtain the RML of the two distribution functions. The asymptotic distributions of the logarithm of RML are provided in Section 3. The minimum sample size required to discriminate between the two distribution functions at a user specified probability of correct selection and a tolerance level are presented in Section 4. Numerical results based on Monte Carlo simulations are presented in Section 5. For illustrative purposes, one data analysis is performed and presented in Section 6 and finally the conclusions appear in Section 7.

2 Ratio of the Maximized Likelihoods

Suppose $X_1, \ldots, X_n$ are independent and identically distributed (i.i.d.) random variables from any of the two distribution functions. The density function of a GR random variable with shape parameter α and scale parameter λ is provided in (2). The density function of a log-normal random variable with scale parameter β and shape parameter $\sigma > 0$ is as follows;

$$f_{LN}(x;\beta,\sigma) = \frac{1}{\sqrt{2\pi}x\sigma} e^{-\frac{(\ln x - \ln\beta)^2}{2\sigma^2}}; \qquad x > 0. \tag{3}$$

A GR distribution with shape parameter α and scale parameter λ will be denoted by $GR(\alpha,\lambda)$. Similarly, a log-normal distribution with shape parameter σ and scale parameter β will be denoted by $LN(\sigma,\beta)$.

Suppose, the data come from $GR(\alpha,\lambda)$, then the log-likelihood function of the observed data is

$$L_{GR}(\alpha,\lambda) = \prod_{i=1}^{n} f_{GR}(x_i;\alpha,\lambda) = 2^n\alpha^n\lambda^{2n}\prod_{i=1}^{n} x_i e^{-\sum_{i=1}^{n}(\lambda x_i)^2}\prod_{i=1}^{n}\left(1 - e^{-(\lambda x_i)^2}\right)^{\alpha-1}. \tag{4}$$

Similarly, if the data are coming from the log-normal distribution, the likelihood function becomes;

$$L_{LN}(\sigma,\beta) = \prod_{i=1}^{n} f_{LN}(x_i;\sigma,\beta) = (2\pi)^{-n/2}\prod_{i=1}^{n} x_i^{-1}\sigma^{-n}e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(\ln x_i - \ln\beta)^2}. \tag{5}$$

In this case RML can be defined as follows;

$$RML = \frac{L_{GR}(\hat\alpha,\hat\lambda)}{L_{LN}(\hat\sigma,\hat\beta)}, \tag{6}$$

where $(\hat\alpha,\hat\lambda)$ and $(\hat\sigma,\hat\beta)$ are the maximum likelihood estimators of (α,λ) and (σ,β) respectively. Suppose $T = \ln(RML)$, then

$$T = n\left[\ln(\hat\alpha\hat\sigma(\tilde{X}\hat\lambda)^2) - 2\left(\frac{\hat\alpha - 1}{\hat\alpha}\right) - \hat\lambda^2\bar{X^2} + \frac{1}{2}\left(1 + \ln(2\pi)\right)\right], \tag{7}$$

where $\bar{X^2} = \frac{1}{n}\sum_{i=1}^{n} X_i^2$ and $\tilde{X} = \left(\prod_{i=1}^{n} X_i\right)^{\frac{1}{n}}$. Moreover, $\hat\alpha$ has the following expression in terms of $\hat\lambda$;

$$\hat\alpha = -\frac{2n}{\sum_{i=1}^{n}\ln\left(1 - e^{-(\hat\lambda X_i)^2}\right)}. \tag{8}$$

For log-normal distribution, $\hat{\sigma}$ and $\hat{\beta}$ have the explicit expressions as follows;

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(\ln X_i - \ln \beta \right)^2, \qquad \text{and} \qquad \hat{\beta} = \tilde{X}. \tag{9}$$

The following simple discrimination procedure can be used to discriminate between the two distribution functions. For a given data set if $T > 0$, then choose GR distribution as the preferred model, otherwise choose the log-normal distribution. Note that this discrimination procedure is intuitively very appealing provided we do not have any preference over any model.

Comments: Suppose the data come from $GR(\alpha, \lambda)$ distribution. In this case the distribution of λX_i is independent of λ. It is also known (see Bain and Engelhardt [1]) that since λ is the scale parameter, $\frac{\hat{\lambda}}{\lambda}$ is independent of λ. Since the distribution of $\frac{X_i}{\tilde{X}}$ is independent of λ, therefore, the distribution of $\hat{\sigma}$ is independent of λ. It implies, the distribution of T is independent of λ. Similarly, when the data come from $LN(\sigma, \beta)$, then the distribution of T is independent of β.

3 Asymptotic Properties of the Logarithm of RML

In this section, we derive the asymptotic distribution of the logarithm of RML when the data come from (a) GR distribution or (b) Log-normal distribution. We will be using the following notations/abbreviations for the rest of the paper. For any Borel measurable functions $g(\cdot)$ and $h(\cdot)$, $E_{GR}(g(U))$, $V_{GR}(g(U))$ denote the mean, variance of $g(U)$ respectively and $cov_{GR}(g(U), h(U))$ denotes the covariance between $g(U)$ and $h(U)$ under the assumption that U follows $GR(\alpha, \lambda)$. Similarly, when U follows $LN(\sigma, \beta)$ then $E_{LN}(h(U))$, $V_{LN}(h(U))$ and $cov_{LN}(g(U), h(U))$ are also defined. We also abbreviate almost surely by $a.s.$.

Suppose the data come from $GR(\alpha, \lambda)$ then along the same line as Gupta and Kundu [2], it can be shown that as $n \to \infty$,

(i) $\hat{\alpha} \to \alpha \quad a.s., \quad \hat{\lambda} \to \lambda \quad a.s.$, where

$$E_{GR}(\ln f_{GR}(X;\alpha,\lambda)) = \max_{\bar{\alpha},\bar{\lambda}} E_{GR}\left[\ln(f_{GR}(X;\bar{\alpha},\bar{\lambda})\right].$$

(ii) $\hat{\sigma}^2 \to \tilde{\sigma}^2, \quad a.s., \quad \hat{\beta} \to \tilde{\beta} \quad a.s.$, where

$$\tilde{\sigma}^2 = Var(\ln Z), \qquad \tilde{\beta} = \frac{1}{\lambda}e^{E(\ln Z)} \text{ and } Z \sim GR(\alpha,1).$$

(iii) T is asymptotically normally distributed with mean $E_{GR}(T)$ and variance $V_{GR}(T)$.

Now we provide the expressions for $E_{GR}(T)$ and $V_{GR}(T)$, as they will be useful. Note that $\lim_{n\to\infty}\frac{E_{GR}(T)}{n}$ and $\lim_{n\to\infty}\frac{V_{GR}(T)}{n}$ exist and we denote them as $AM_{GR}(\alpha)$ and $AV_{GR}(\alpha)$ respectively. Therefore, for large n

$$\begin{aligned}\frac{E_{GR}(T)}{n} \approx AM_{GR}(\alpha) &= E_{GR}\left[\ln f_{GR}(\alpha,\lambda) - \ln f_{LN}(\tilde{\sigma},\tilde{\beta})\right] \\ &= \ln 2 + \ln\alpha + \frac{1}{2}\ln(2\pi) + \frac{1}{2} - E(Z^2) + (\alpha-1)E\ln(1-e^{-Z^2}) \\ &\quad +\frac{1}{2}\ln(Var(\ln Z)).\end{aligned}$$

Also

$$\begin{aligned}\frac{V_{GR}(T)}{n} \approx AV_{GR}(\alpha) &= V_{GR}\left(\ln(f_{GR}(\alpha,\lambda) - \ln(f_{LN}(\tilde{\sigma},\tilde{\beta}))\right) \\ &= V_{GR}\left(2\ln Z - Z^2 + (\alpha-1)\ln(1-e^{-Z^2}) + \frac{(\ln Z - E\ln Z)^2}{2\tilde{\sigma}^2}\right) \\ &= 4V_{GR}(\ln Z) + V(Z^2) + (\alpha-1)^2 V(\ln(1-e^{-Z^2})) + \frac{1}{4\tilde{\sigma}^4}V(\ln Z) \\ &- 4Cov(\ln Z, Z^2) + (\alpha-1)Cov(\ln(1-e^{-Z^2}), (\ln Z - E\ln Z)^2) \\ &+ \frac{2}{\tilde{\sigma}^2}Cov(\ln Z, (\ln Z - E\ln Z)^2) - 2(\alpha-1)Cov(Z^2, \ln(1-e^{-Z^2})) \\ &- \frac{2}{\tilde{\sigma}^2}Cov(Z^2, (\ln Z - E\ln Z)^2) + 4(\alpha-1)Cov(\ln Z, \ln(1-e^{-Z^2})).\end{aligned}$$

Now let us consider the case when the data come from $LN(\sigma,\beta)$. In this case again along the same line as Gupta and Kundu [2], it can be shown that as $n \to \infty$.

(i) $\hat{\sigma} \to \sigma \quad a.s, \quad \hat{\beta} \to \beta \quad a.s.$, where

$$E_{LN}(\ln f_{LN}(X;\sigma,\beta)) = \max_{\bar{\sigma},\bar{\beta}} E_{LN}(\ln f_{LN}(X;\bar{\sigma},\bar{\beta}).$$

(ii) $\hat{\alpha} \to \alpha \quad a.s, \quad \hat{\lambda} \to \lambda \quad a.s.$, where

$$E_{LN}(\ln f_{GR}(X;\tilde{\alpha},\tilde{\lambda})) = \max_{\alpha,\lambda} E_{GR}(\ln f_{LN}(X;\alpha,\lambda).$$

(iii) T is asymptotically normally distributed with mean $E_{LN}(T)$ and variance $V_{LN}(T)$.

Now we discuss how to obtain $\tilde{\alpha}$ and $\tilde{\lambda}$. Let us define

$$\begin{aligned} g(\alpha,\lambda) &= E_{LN}(\ln f_{GR}(X;\alpha,\lambda)) \\ &= E_{LN}\left(\ln 2 + \ln\alpha 2\ln\lambda + \ln X - (\lambda X)^2 + (\alpha-1)\ln(1-e^{-(\lambda X)^2})\right) \\ &= \ln 2 + \ln\alpha + 2\ln\lambda + \ln\beta - \lambda^2\beta^2 e^{2\sigma^2} + (\alpha-1)U(\sigma,\beta\lambda), \end{aligned}$$

where

$$U(a,b) = \int_0^\infty \frac{1}{\sqrt{2\pi}az} e^{-\frac{1}{2a^2}(\ln Z)^2} \ln(1-e^{-(bZ)^2})dz.$$

Therefore, $\tilde{\alpha}$ and $\tilde{\lambda}$ can be obtained as solutions of

$$\frac{1}{\tilde{\alpha}} + U(\sigma,\beta\tilde{\lambda}) = 0 \qquad (10)$$

$$2 - 2(\tilde{\lambda}\beta)^2 e^{2\sigma^2} + (\tilde{\alpha}-1)\tilde{\lambda}\beta U_2(\sigma,\beta\tilde{\lambda}) = 0, \qquad (11)$$

here $U_2(a,b)$ is the derivative $U(a,b)$ with respect to b. From (10), it is clear $(\beta\tilde{\lambda})$ is a function of $\tilde{\alpha}$ and σ only. From (11) it follows that $\tilde{\alpha}$ is a function σ only, therefore, $(\beta\tilde{\lambda})$ is a function of σ only.

Now we provide the expression for $E_{LN}(T)$ and $V_{LN}(T)$. Since $\lim_{n\to\infty}\frac{E_{LN}(T)}{n}$ and $\lim_{n\to\infty}\frac{V_{LN}(T)}{n}$ exist, we denote them as $AM_{LN}(\sigma)$ and $AV_{LN}(\sigma)$ respectively. Therefore, for large n,

$$\begin{aligned} \frac{E_{LN}(T)}{n} \approx AM_{LN}(\sigma) &= E_{LN}\left[\ln f_{GR}(\tilde{\alpha},\tilde{\lambda}) - \ln f_{LN}(\sigma,1)\right] \\ &= \ln(2\tilde{\alpha}\tilde{\lambda}^2\sigma\sqrt{2\pi}) - \tilde{\lambda}^2 e^{2\sigma^2} + (\alpha-1)E_{LN}\ln(1-e^{-(\tilde{\lambda}X)^2}) + \frac{1}{2}. \end{aligned}$$

Table 1: Different values of $AM_{GR}(\alpha)$, $AV_{GR}(\alpha)$, $\tilde{\sigma}$ and $\tilde{\beta}$ for different α

$\alpha \rightarrow$	1.00	1.50	2.00	2.50	3.00	3.50	4.00
$AM_{GR}(\alpha)$	0.0898	0.0558	0.0378	0.0269	0.0200	0.0152	0.0117
$AV_{GR}(\alpha)$	0.2677	0.1561	0.1000	0.0685	0.0491	0.0363	0.0275
$\tilde{\sigma^2}$	0.4108	0.2397	0.1708	0.1344	0.1120	0.0969	0.0859
$\tilde{\beta}$	0.7494	0.9352	1.0597	1.1515	1.2236	1.2825	1.3320

Also

$$
\begin{aligned}
\frac{V_{LN}(T)}{n} \approx AV_{LN}(\sigma) &= V_{LN}\left[\ln f_{GR}(\tilde{\alpha}, \tilde{\lambda}) - \ln f_{LN}(\sigma, 1)\right] \\
&= V_{LN}\left[2\ln X - (\tilde{\lambda}X)^2 + (\alpha - 1)\ln(1 - e^{-(\tilde{\lambda}X)^2}) + \frac{(\ln X)^2}{2\sigma^2}\right] \\
&= 4\sigma^2 + \tilde{\lambda}^4(e^{8\sigma^2} - e^{4\sigma^2}) + (\alpha - 1)^2 V_{LN}(\ln(1 - e^{-(\tilde{\lambda}X)^2})) + \frac{1}{2} \\
&- 8(\sigma\tilde{\lambda})^2 e^{\sigma^4} - \tilde{\lambda}^2 e^{\sigma^4} + 4(\alpha - 1)Cov_{LN}(\ln X, \ln(1 - e^{-(\tilde{\lambda}X)^2})) \\
&- 2(\alpha - 1)Cov_{LN}((\tilde{\lambda}X)^2, \ln(1 - e^{-(\tilde{\lambda}X)^2})) \\
&+ \frac{(\alpha - 1)}{\sigma^2}Cov_{LN}((\ln X)^2, \ln(1 - e^{-(\tilde{\lambda}X)^2})).
\end{aligned}
$$

Note that $\tilde{\alpha}$, $\tilde{\lambda}$, $AM_{LN}(\sigma)$, $AV_{LN}(\sigma)$, $\tilde{\beta}$, $\tilde{\sigma}$, $AM_{GR}(\alpha)$ and $AV_{GR}(\alpha)$ are quite difficult to compute numerically. We present them in Tables 1 and 2 for convenience.

4 Determination Of Sample Size:

We propose a method to determine the minimum sample size required to discriminate between the log-normal and GR distributions, for a given user specified probability of correct selection (PCS) similarly as proposed in Gupta and Kundu [2]. To discriminate between two distribution functions it is important to know how close they are. There are several ways

Table 2: Different values of $AM_{LN}(\sigma)$, $AV_{LN}(\sigma)$, $\tilde{\alpha}$ and $\tilde{\lambda}$ for different σ

$\sigma^2 \rightarrow$	0.10	0.15	0.20	0.25	0.30	0.35
$AM_{LN}(\sigma)$	-0.0170	-0.0356	-0.0541	-0.0720	-0.0900	-0.1052
$AV_{LN}(\sigma)$	0.0450	0.1117	0.1982	0.3035	0.4271	0.5689
$\tilde{\alpha}$	3.2845	2.0649	1.5176	1.2071	1.0032	0.8614
$\tilde{\lambda}$	1.2554	1.0676	0.9331	0.8283	0.7418	0.6711

to measure the distance (inversely proportional to closeness) between them. Of course, the most popular one is the Kolmogorov-Smirnov (K-S) distance. The K-S distance between two distribution functions, say $F(x)$ and $G(x)$ is defined as

$$\sup_x |F(x) - G(x)|. \tag{12}$$

If two distributions are very close then it is natural that a very large sample size is needed to discriminate between them for a given PCS. It may be also true that if the distance between two distribution functions are small, then one may not need to differentiate the two distributions from any practical point of view. Therefore, it is expected that the user will specify before hand the PCS and also the tolerance limit in terms of the distance between two distribution functions. The tolerance limit simply indicates that the user does not want to make the distinction between two distribution functions if their distance is less than the tolerance limit. Based on the probability of correct selection and the tolerance limit, the required minimum sample size can be determined. Here we use the K-S distance to discriminate between two distribution functions but similar methodology can be developed for any other distance function also.

We observed in Section 3 that the RML statistics follow normal distribution approxi-

Table 3: The minimum sample size $n = \frac{z_{0.90}^2 AV_{GR}(\alpha)}{(AM_{GR}(\alpha))^2}$, for $p^* = 0.9$ and when the data are coming from a GR distribution is presented. The K-S distance between $GR(\alpha, 1)$ and $LN(\tilde{\sigma}, \tilde{\beta})$ for different values of α is reported.

$\alpha \to$	1.0	1.5	2.0	2.5	3.0	3.5	4.0
$n \to$	55	82	115	156	202	258	330
K-S	0.071	0.055	0.045	0.038	0.032	0.028	0.025

mately for large n. Now it will be used with the help of K-S distance to determine the required sample size n such that the PCS achieves a certain protection level p^* for a given tolerance level D^*. We explain the procedure assuming case 1, case 2 follows exactly along the same line. Since T is asymptotically normally distributed with mean $E_{GR}(T)$ and variance $V_{GR}(T)$, therefore the probability of correct selection (PCS) is

$$PCS(\alpha) = P\left[T > 0|\alpha\right] \approx 1 - \Phi\left(\frac{-E_{GR}(T)}{\sqrt{V_{GR}(T)}}\right) = 1 - \Phi\left(\frac{-n \times AM_{GR}(\alpha)}{\sqrt{n \times AV_{GR}(\alpha)}}\right). \qquad (13)$$

Here Φ is the distribution function of the standard normal random variable. $AM_{GR}(\alpha)$ and $AV_{GR}(\alpha)$ are same as defined before. Now to determine the sample size needed to achieve at least a p^* protection level, equate

$$\Phi\left(\frac{-n \times AM_{GR}(\alpha)}{\sqrt{n \times AV_{GR}(\alpha)}}\right) = 1 - p^*, \qquad (14)$$

and solve for n. It provides

$$n = \frac{z_{p^*}^2 AV_{GR}(\alpha)}{(AM_{GR}(\alpha))^2}. \qquad (15)$$

Here z_{p^*} is the $100p^*$ percentile point of a standard normal distribution. For $p^* = 0.9$ and for different α, the values of n are reported in Table 3. Similarly for case 2, we need

$$n = \frac{z_{p^*}^2 AV_{LN}(\sigma)}{(AM_{LN}(\sigma))^2}. \qquad (16)$$

We report n, with the help of Table 2 for different values of α when $p^* = 0.9$ in Table 4. We report the K-S distance between $GR(\alpha, 1)$ and $GR(\tilde{\sigma}, \tilde{\beta})$ for different values of α in Table

Table 4: The minimum sample size $n = \frac{z_{0.90}^2 AV_{LN}(\sigma)}{(AM_{LN}(\sigma))^2}$, for $p^* = 0.9$ and when the data are coming from a log-normal distribution is presented. The K-S distance between $GR(\tilde{\alpha}, \tilde{\lambda})$ and $LN(\sigma, 1)$ for different values of σ is reported.

$\sigma^2 \to$	0.10	0.15	0.20	0.25	0.30	0.35
$n \to$	256	145	111	96	87	85
K-S	0.033	0.049	0.063	0.075	0.085	0.092

3 for different values α. Similarly, the K-S distance between $GR(\tilde{\alpha}, \tilde{\lambda})$ and $LN(\sigma, 1)$ for different values of σ is reported in Table 4. From Table 3, it is clear that as α increases the required sample size increases for a given PCS. It implies that the distance between the two distribution functions decreases. Similarly, in Table 4, it is observed that as σ increases the required sample size decreases, *i.e* the distance between them increases. Therefore, if one knows the ranges of the shape parameters of the two distribution functions, then the minimum sample size can be obtained using (15) or (16) and using the fact that n is a monotone function of the shape parameters for both cases. But unfortunately in practice it may be completely unknown. Therefore, to have some idea of the shape parameter of the null distribution we make the following assumptions. It is assumed that the experimenter would like to choose the minimum sample size needed for a given protection level when the distance between two distribution functions is greater than a pre-specified tolerance level.

Now we use similar methodology as used in Gupta and Kundu [2] to determine the minimum sample size required to discriminate between GR and log-normal distribution functions for a user specified protection level and for a given tolerance level between them. Suppose the protection level is $p^* = 0.9$ and the tolerance level is given in terms of K-S distance as $D^* = 0.05$. Here tolerance level $D^* = 0.05$ means that the practitioner wants to discriminate between the GR and log-normal distribution functions only when their K-S distance is more than 0.05. From Table 3, it is observed that the K-S distance will be more than 0.05 if

$\alpha < 2.0$ (conservative). Similarly from Table 4, it is clear that the K-S distance will be more than 0.05 if $\sigma \leq 0.15$. Therefore, if the data come from the GR distribution, then for the tolerance level $D^* = 0.05$, one needs at most $n = 115$ to meet the PCS, $p^* = 0.9$. Similarly if the data come from the log-normal distribution then one needs at most $n = 145$ to meet the above protection level $p^* = 0.9$ for the same tolerance level $D^* = 0.05$. Therefore, for the given tolerance level 0.05 one needs max(115, 145) = 145 to meet the protection level $p^* = 0.9$ simultaneously for both cases. Although Tables 3 and 4 provide the required sample size for $p^* = 0.9$, but they can be used for any other p^* values also. For any other p^* value, the corresponding n will be multiplied by $z^2_{p^*}/z^2_{0.90}$.

5 Numerical Experiments

We perform some numerical experiments and present the results in this section. We mainly try to observe how the asymptotic results derived in Section 3 behave for different sizes. We perform all the computations at the Indian Institute of Technology Kanpur, using Pentium-IV processor. We use the random deviate generator of Press *et al.* [5] and all the programs are written in FORTRAN. We compute the PCS based on simulations and we also compute it based on asymptotic results derived in Section 3. We consider different sample sizes and different shape parameters for both the distributions. We consider both the cases separately, namely when the null distribution is (i) GR or (ii) log-normal. Since the PCS is independent of the scale parameter in both the cases, therefore, without loss of generality we consider them to be 1.

First we consider the case when the null distribution is GR. In this case we consider n = 20, 40, 60, 80, 100 and α = 1.0, 1.5, 2.0, 2.5, 3.0, 3.5 and 4.0. For a fixed α and n we generate a random sample of size n from $GR(\alpha, 1)$, and check whether T is positive or negative. We replicate the process 10,000 times and obtain an estimate of the PCS. We also

Table 5: The probability of correct selection based on Monte Carlo simulations and also based on asymptotic results when the null distribution is GR. The element in the first row in each box represents the results based on Monte Carlo simulations (10,000 replications) and the number in bracket represents the result obtained by using asymptotic results

$\alpha \downarrow n \rightarrow$	20	40	60	80	100
1.00	0.78 (0.78)	0.88 (0.86)	0.93 (0.91)	0.96(0.94)	0.97 (0.96)
1.50	0.73 (0.74)	0.82 (0.81)	0.88 (0.86)	0.91 (0.90)	0.93 (0.92)
2.00	0.69 (0.70)	0.78 (0.78)	0.83 (0.82)	0.87 (0.86)	0.89 (0.89)
2.50	0.66 (0.68)	0.74 (0.74)	0.79 (0.79)	0.83 (0.82)	0.85 (0.85)
3.00	0.63 (0.65)	0.71 (0.71)	0.76 (0.75)	0.80 (0.79)	0.82 (0.81)
3.50	0.61 (0.64)	0.68 (0.69)	0.73 (0.73)	0.76 (0.76)	0.78 (0.79)
4.00	0.59 (0.62)	0.66 (0.67)	0.70 (0.71)	0.73 (0.74)	0.76 (0.76)

compute the PCSs by using the asymptotic results derived in Section 4. The results are reported in Table 5. Similarly, we obtain the results when the data are generated from the log-normal distribution. In this case we consider the same set of n and $\sigma^2 = 0.10$, 0.15, 0.20, 0.25, 0.30 and 0.35. In this case the results are reported in Table 6. In each box the first element represents the result obtained by using Monte Carlo simulations and the element in the bracket represents the result obtained by using the asymptotic theory.

As sample size increases the PCS increases in both the cases. It is also clear that when α increases for the GR distribution the PCS decreases and when σ increases for the log-normal distribution the PCS increases. Moreover when the null distribution is GR, the asymptotic results work better than the case when the null distribution is log-normal. The asymptotic results work reasonably well even for small sample sizes.

Table 6: The probability of correct selection based on Monte Carlo simulations and also based on asymptotic results when the null distribution is LN. The element in the first row in each box represents the results based on Monte Carlo simulations (10,000 replications) and the number in bracket represents the result obtained by using asymptotic results

$\sigma^2 \downarrow n \rightarrow$	20	40	60	80	100
0.10	0.64 (0.64)	0.69 (0.70)	0.73 (0.73)	0.78 (0.77)	0.80 (0.79)
0.15	0.68 (0.68)	0.76 (0.75)	0.81 (0.80)	0.85 (0.83)	0.88 (0.86)
0.20	0.71 (0.71)	0.80 (0.78)	0.86 (0.83)	0.90 (0.86)	0.93 (0.89)
0.25	0.73 (0.72)	0.84 (0.80)	0.89 (0.85)	0.92 (0.89)	0.95 (0.91)
0.30	0.75 (0.73)	0.86 (0.81)	0.91 (0.87)	0.94 (0.89)	0.96 (0.92)
0.35	0.76 (0.74)	0.87 (0.82)	0.92 (0.87)	0.95 (0.90)	0.97 (0.92)

6 Data Analysis

In this section we consider one real life data set for illustrative purposes and verify how our methods work in practice. The data is obtained from Linhardt and Zucchini ([4], page 69). It represents the failure times (in hours) of the air conditioning system of 30 different airplanes and they are as follows: 23, 261, 87, 7, 120, 14, 62, 47, 225, 71, 246, 21, 42, 20, 5, 12, 120, 11, 3, 14, 71, 11, 14, 11, 16, 90, 1, 16, 52, 95.

When we use the GR distribution, the MLEs of the different parameters are $\hat{\alpha} = 0.3086$, $\hat{\lambda}$ = 0.0076 and the corresponding log-likelihood (LL) value is -154.102. Similarly, when we use the log-normal distribution, the MLEs are $\hat{\sigma}$ = 1.3192, $\hat{\beta}$ = 28.7343 and the corresponding LL value is -151.706. Therefore, $T = -154.102 + 151.706 < 0$ and it indicates we prefer the log-normal distribution than the GR distribution.

Table 7: The observed and expected frequencies for different groups based on the fitted GR and log-normal distributions

Intervals	Observed	GR	log-normal
0-15	11	7.84	8.45
15-30	5	4.11	5.06
30-60	3	5.95	6.33
60-100	6	5.37	4.55
100-	5	6.73	5.62

Now let us see how the two distribution functions fit the above data set. It is observed that the K-S distance between the data and the fitted GR distribution is 0.1941 and the corresponding p-value is 0.2082. The K-S distance between the data and the fitted log-normal distribution is 0.1047 and the corresponding p-value is 0.88. We also present the observed, expected frequencies for different groups and for both the distributions in Table 7. The χ^2 values are 4.632 and 3.562 and the corresponding p-values are 0.33 and 0.47 for the GR and log-normal distributions respectively. Therefore, from the K-S distance measures and also from the χ^2 values, it is clear that although both the distributions fit the data reasonably well, but log-normal is the preferred one.

7 Conclusions

In this paper we consider discriminating between two over-lapping distributions using the statistic based on RML. We also obtain the asymptotic distributions of the statistics under two different conditions. Note that the above problem can be put as a testing hypothesis problem also and the asymptotic distribution of the likelihood ratio test statistic can be easily obtained using the asymptotic results of T. Therefore, the power functions of the corresponding test statistics also can be obtained.

We have already observed that if two distribution functions are very close then the sample size must be very large to discriminate between the two. In Figure 2 it is observed that the two distribution functions are almost indistinguishable. Therefore, in this situation if the sample size is not very large and the data are generated from the GR distribution, then it will be very difficult to judge whether they have been generated from the GR or from the log-normal distribution. It implies that the GR distribution can be used to generate log-normal and in turn normal random numbers. Work is in progress and it will be reported elsewhere.

Acknowledgements

The authors would like to thank one referee for his/ her valuable comments.

References

[1] Bain, L.J. and Engelhardt, M. (1991), *Statistical Analysis of Reliability and Life-Testing Models*, 2nd. Edition, Marcel and Decker, New York.

[2] Gupta, R.D. and Kundu, D. (2003), "Discriminating between the Weibull and the GE distributions", *Computational Statistics and Data Analysis*, vol. 43, 179-196.

[3] Johnson, N.L., Kotz, S. and Balakrishnan, N. (1995), *Continuous Univariate Distribution* vol. 1, 2nd Edition, Wiley, New York.

[4] Linhardt, H. and Zucchini, W. (1986), *Model Selection*, Wiley, New York.

[5] Press, W.H., Teukolsky, S.A., Vellerling, W.T., Flannery, B.P. (1992), *Numerical Recipes in FORTRAN, The Art of Scientific Computing*, 2nd. Edition, Cambridge University Press, Cambridge.

[6] Surles, J.G. and Padgett, W.J. (2001), "Inference for reliability and stress-strength for a scaled Burr Type X distribution", *Lifetime Data Analysis*, vol. 7, 187-200.

[7] White, H. (1982a), "Maximum likelihood estimation of mis-specified models", *Econometrica*, vol. 50, 1-25.

[8] White, H. (1982b), "Regularity conditions for Cox's test of non-nested hypothesis", *Journal of Econometrics*, vol. 19, 301-318.