Discriminating Between the Log-Normal and Log-Logistic Distributions

Arabin Kumar Dey[†] & Debasis Kundu[†]

Abstract

Log-normal and log-logistic distributions are often used to analyze lifetime data. For certain ranges of the parameters, the shape of the probability density functions or the hazard functions can be very similar in nature. It might be very difficult to discriminate between the two distribution functions. In this paper, we consider the discrimination procedure between the two distribution functions. We use the ratio of maximized likelihood for discrimination purposes. The asymptotic properties of the proposed criterion have been investigated. It is observed that the asymptotic distributions are independent of the unknown parameters. The asymptotic distributions have been used to determine the minimum sample size needed to discriminate between these two distribution functions for a user specified probability of correct selection. We have performed some simulation experiments to see how the asymptotic results work for small sizes. For illustrative purpose two data sets have been analyzed.

Keywords: Asymptotic distributions; likelihood ratio test; probability of correct selection; log-location-scale family; model selection; Kolmogorov-Smirnov distance.

[†] Department of Mathematics and Statistics, Indian Institute of Technology Kanpur, Kanpur, Pin 208016, INDIA.

Corresponding Author: Debasis Kundu, Phone no. 91-512-2597141, Fax No. 91-512-2597500, e-mail: kundu@iitk.ac.in.

1 Introduction

Log-normal and log-logistic distributions are often used for analyzing skewed data. The two distributions have several interesting properties and their probability density functions (PDFs) can take different shapes. For example, the log-normal can have unimodal PDFs and they are always log-concave. Similarly, the PDF of the log-logistic distribution is either reversed J shaped or unimodal. It is known that the log-normal distribution can have inverted bath-tub hazard function, whereas the hazard function of the log-logistic distribution is either decreasing or inverted bath-tub, see Johnson, Kotz and Balakrishnan [19]. For certain ranges of the parameters, the shape of their probability density functions (PDFs) or the hazard functions can be very similar in nature. Therefore, if it is known or apparent from the histogram of the data that the data are coming from a right skewed distribution, then any one of them may be used for analyzing the data. Therefore, to analyze a skewed data an experimenter can choose any one of the two models. Although, these two distributions may provide similar data fit for moderate sample sizes but it is still desirable to choose the correct or more nearly correct model, since the inferences based on the tail probabilities where the affect of the model assumption will be very crucial.

In this paper we address the following problem. Suppose one observes a random sample of size n, say $x_1, \ldots, x_n$ and the person has a choice to use one of the distributions, *viz*. log-normal or log-logistic, which one is preferable. The problem of testing whether some given observations follow a particular distribution is a classical problem. Cox [8, 9] first considered this problem and developed a testing procedure for two non-nested families. Since then, extensive work has been done in discriminating between two families of distributions. See for example Atkinson [2, 3] Wiens [36], Dumonceaux and Antle [11], Dumonceaux *et al.* [10], Quesenberry and Kent [33], Balasooriya and Abeysinghe [5], Kim *et al.* [21] and Kundu, Gupta and Manglick [22] for some of the recent references.

In this paper we consider the problem of discriminating between the log-normal and log-logistic distributions. We use the ratio of the maximized likelihood (RML) in discriminating between the two distribution functions. We also obtain the asymptotic distributions of the RMLs using the idea of White [34, 35]. Using the asymptotic distributions of the RMLs, we compute the probability of correct selection (PCS) under each model. It is observed that the asymptotic PCS works quite well even for moderate sample sizes. We suggest some small sample corrections based on the numerical simulations. The asymptotic distributions of the RMLs are independent of the unknown parameters. Therefore, these asymptotic distributions can be used to determine the minimum sample size needed to discriminate between these two distributions for a given user specified PCS.

The rest of the paper is organized as follows. In section 2, we provide the notation and discrimination procedure. The asymptotic results are provided in section 3. Determination of sample size is obtained in section 4. Numerical results and data analysis are presented in section 5 and section 6 respectively. Finally the conclusions appear in section 7.

2 Notation and Discrimination Procedure

We will use the following notation in this paper.

Log-normal Distribution: The log-normal distribution with the shape parameter $\eta > 0$ and the scale parameter $\theta > 0$ will be denoted by $LN(\eta, \theta)$ and the corresponding PDF is

$$f_{LN}(x;\eta,\theta) = \frac{1}{\sqrt{2\pi}x\eta} e^{-\frac{1}{2}\left(\frac{(\ln x - \ln\theta)^2}{\eta^2}\right)}; \quad x > 0. \tag{1}$$

Log-logistic Distribution: The log-logistic distribution with the shape parameter $\sigma > 0$ and the scale parameter $\xi > 0$ has the following PDF;

$$f_{LL}(x;\sigma,\xi) = \frac{1}{\sigma x} \times \frac{e^{\left(\frac{\ln x - \ln\xi}{\sigma}\right)}}{\left(1 + e^{\frac{\ln x - \ln\xi}{\sigma}}\right)^2}; \quad x > 0. \tag{2}$$

From now on it will be denoted as $LL(\sigma, \xi)$.

Almost sure convergence will be denoted by $a.s.$. For any Borel measurable functions $g(\cdot)$ and $h(\cdot)$, $E_{LN}(g(U))$ and $V_{LN}(g(U))$ and $Cov_{LN}(g(U), h(U))$ denote the mean of $g(U)$, variance of $g(U)$ and covariance between $g(U)$ and $h(U)$ respectively, when U has a $LN(\eta, \theta)$ distribution. Similarly, if U follows $LL(\sigma, \xi)$, the respective quantities are defined and they should be clear from the context.

Now we describe the discrimination procedure based on a random sample $X = \{x_1, \ldots, x_n\}$. It is assumed that the data have been generated from one of the two distributions, namely; $LN(\eta, \theta)$ or $LL(\sigma, \xi)$. Based on the observed sample, the corresponding likelihood functions are; $L_{LN}(\eta, \theta|X) = \prod_{i=1}^{n} f_{LN}(x_i; \eta, \theta)$ and $L_{LL}(\sigma, \xi|X) = \prod_{i=1}^{n} f_{LL}(x; \sigma, \xi)$ respectively. If $\widehat{\eta}$, $\widehat{\theta}$, $\widehat{\sigma}$ and $\widehat{\xi}$ are the maximum likelihood estimators of the corresponding parameters, then the discrimination procedure is as follows. Choose

(a) Log-normal distribution if

$$L_{LN}(\widehat{\eta}, \widehat{\theta}) > L_{LL}(\widehat{\sigma}, \widehat{\xi}). \tag{3}$$

(b) Log-logistic distribution if

$$L_{LL}(\widehat{\sigma}, \widehat{\xi}) > L_{LN}(\widehat{\eta}, \widehat{\theta}). \tag{4}$$

From (3) and (4) it is clear that given the data $\{x_1, \ldots, x_n\}$, one will be able to choose one particular distribution $a.s.$. Now the natural question is, what is the PCS in this discrimination procedure. It may be noted that the probability of correct selection will depend on the parent distribution, *i.e.*, the original distribution of $\{X_1, \ldots, X_n\}$. We will consider the two cases separately.

If the data were originally coming from $LN(\eta, \theta)$, the probability of correct selection

(PCS_{LN}) can be written as follows;

$$PCS_{LN} = P(T > 0| \text{ data follow log-normal distribution}), \tag{5}$$

where

$$T = \ln\left[\frac{L_{LN}(\widehat{\eta},\widehat{\theta})}{L_{LL}(\widehat{\sigma},\widehat{\xi})}\right].$$

Similarly,

$$PCS_{LL} = P(T < 0| \text{ data follow log-logistic distribution}). \tag{6}$$

Now to compute (5) and (6) one needs to compute the exact distributions of T under the respective parent distributions. Since they are difficult to compute, we rely on their asymptotic distributions. Based on the asymptotic distributions, we can obtain the approximate values of PCS_{LN} and PCS_{LL} for large n. In the next section we obtain the asymptotic distribution of T, under the respective parent distributions, but first let us look at the expressions of T's in terms of the corresponding MLEs. Note that;

$$T = n\left[-\frac{1}{2}(\ln\pi + \ln 2 + 1) - \frac{1}{\widehat{\sigma}}\ln\widetilde{X} - \ln\widehat{\eta} + \ln\widehat{\sigma} + \frac{1}{\widehat{\sigma}}\ln\widehat{\xi} + \frac{2}{n}\sum_{i=1}^{n}\ln\left(1 + e^{\frac{\ln X_i - \ln\widehat{\xi}}{\widehat{\sigma}}}\right)\right], \tag{7}$$

and

$$\widetilde{X} = \left(\prod_{i=1}^{n} X_i\right)^{\frac{1}{n}} = \widehat{\theta}, \quad \widehat{\eta} = \frac{1}{n}\sum_{i=1}^{n}\left(\ln X_i - \ln\widehat{\theta}\right)^2.$$

COMMENTS: Note that if we transform the data $Y_i = \ln X_i$ and consider the logarithm of RMLs of the corresponding transformed distributions, namely normal, extreme value and logistic distributions, then the values of the test statistics will be unchanged. Using the results of Dumonceaux *et al.* [10] it follows that the distributions of T's are independent of the parameters.

COMMENTS: Interestingly it is also observed that the Kolmogorov-Smirnov distance between the log-normal (log-logistic) and the best fitted asymptotic log-logistic (log-normal) is constant, see Appendix.

3 Asymptotic Distributions

It is not possible to obtain the exact distributions of T under the respective parent distributions and therefore we mainly rely on the asymptotic distributions. We consider the two cases separately.

First let us consider the case when the data are coming from log-normal distribution. We have the following result.

Theorem 1: If the data are from $LN(\eta, \theta)$, then T is asymptotically normally distributed with mean $E_{LN}(T)$ and $V_{LN}(T)$.

To prove Theorem 1, we need the following Lemma.:

Lemma 1: Under the assumption that the data are from $LN(\eta, \theta)$, we have the following results as $n \to \infty$

(a) $\widehat{\eta} \longrightarrow \eta \;\; a.s$, and $\widehat{\theta} \longrightarrow \theta, \;\; a.s.$, where

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

(b) $\widehat{\sigma} \longrightarrow \widetilde{\sigma}$, $a.s.$, and $\widehat{\xi} \longrightarrow \widetilde{\xi}, \;\; a.s, \;\;$ where

$$E_{LN}(\ln f_{LL}(X; \widetilde{\sigma}, \widetilde{\xi}) = \max_{\sigma, \xi} E_{LN}(\ln f_{LL}(X; \sigma, \xi).$$

Let us denote

$$T^* = \ln \left[\frac{L_{LN}(\eta, \theta)}{L_{LL}(\widetilde{\sigma}, \widetilde{\xi})} \right].$$

(c)

$$\frac{1}{\sqrt{n}} \left[T - E_{LN}(T) \right] \text{ asymptotically equivalent to } \frac{1}{\sqrt{n}} \left[T^* - E_{LN}(T^*) \right].$$

Proof of Lemma 1: The proof follows using similar argument as of White (1982, Theorem 1) and therefore, it is omitted.

PROOF OF THEOREM 1: Using the Central limit theorem (CLT), it immediately follows that $\frac{1}{\sqrt{n}}\left[T^* - E_{LN}(T^*)\right]$ is asymptotically normally distributed. Therefore the result follows using (c) of Lemma 1. ∎

It should be mentioned that $\widetilde{\sigma}$ and $\widetilde{\xi}$ depend on η and θ, but we do not make it explicit for brevity. For further development we need to compute $\widetilde{\sigma}$ and $\widetilde{\xi}$ for $\eta = 1$ and $\theta = 1$ and we will denote them $\widetilde{\sigma}_1$ and $\widetilde{\xi}_1$. Let us define

$$\begin{aligned} h(\sigma,\xi) &= E_{LN}\left(f_{LL}(X:\sigma,\xi)\right) \\ &= E_{LN}\left[\frac{\ln X - \ln\xi}{\sigma} - 2\ln\left(1+e^{\frac{\ln X-\ln\xi}{\sigma}}\right) - \ln\pi - \ln\sigma\right] \\ &= -\frac{1}{\sigma}\ln\xi - \frac{2}{\sqrt{2\pi}}\int_0^\infty \frac{1}{x}\ln\left(1+e^{\frac{\ln X-\ln\xi}{\sigma}}\right)e^{-\frac{1}{2}(\ln x)^2}dx. \end{aligned} \tag{8}$$

then $\widetilde{\sigma}_1, \widetilde{\xi}_1$ can be obtained by maximizing $h(\sigma,\xi)$ with respect to σ and ξ. Unfortunately, it is not possible to obtain the analytical solutions. Numerically it is observed that $h(\sigma,\xi)$ is maximized when $\widetilde{\sigma}_1 = 0.5718$ and $\widetilde{\xi}_1 = 1.0000$.

Now we will compute $E_{LN}(T)$ and $V_{LN}(T)$. If we denote $AM_{LN} = \lim_{n\to\infty}\frac{1}{n}E_{LN}(T)$ and $AV_{LN} = \lim_{n\to\infty}\frac{1}{n}V_{LN}(T)$, then

$$\begin{aligned} \lim_{n\to\infty}\frac{1}{n}E_{LN}(T) = AM_{LN} &= E_{LN}\left[\ln f_{LN}(X;1,1) - \ln f_{LL}(X;\widetilde{\sigma}_1,\widetilde{\xi}_1)\right] \\ &= -\frac{1}{2}\ln 2 - \frac{1}{2}\ln\pi - \frac{1}{2} + \ln\widetilde{\sigma}_1 + \frac{\ln\widetilde{\xi}_1}{\widetilde{\sigma}_1} + 2E\ln\left(1+\widetilde{\xi}_1^{-1/\widetilde{\sigma}_1}e^{Z/\widetilde{\sigma}_1}\right), \\ &\approx 0.0095. \end{aligned}$$

here Z follows $N(0,1)$. Moreover,

$$\begin{aligned} \lim_{n\to\infty}\frac{1}{n}V_{LN}(T) = AV_{LN} &= V_{LN}\left[\ln f_{LN}(X;1,1) - \ln f_{LL}(X;\widetilde{\sigma}_1,\widetilde{\xi}_1)\right] \\ &= V_{LN}\left[-\frac{1}{2}Z^2 - Z\left(\frac{1}{\widetilde{\sigma}_1}\right) + 2\ln\left(1+e^{Z/\widetilde{\sigma}_1}\xi_1^{-1/\widetilde{\sigma}_1}\right)\right], \\ &\approx 0.0137 \end{aligned}$$

Now we consider the case when the data are coming from the log-logistic distribution. We have the following result.

THEOREM 2: If the data are from $LL(\sigma, \xi)$, then T is asymptotically normally distributed with mean $E_{LL}(T)$ and $V_{LL}(T)$.

To prove Theorem 2, we need the following Lemma 2, similar to Lemma 1. The proof of Lemma 2 also follows along the same line as of Lemma 1.

LEMMA 2: Under the assumption that the data are from $LL(\sigma, \xi)$, we have the following results as $n \to \infty$

(a) $\widehat{\sigma} \longrightarrow \sigma \;\; a.s$, and $\widehat{\xi} \longrightarrow \xi, \;\; a.s.$, where

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

(b) $\widehat{\eta} \longrightarrow \widetilde{\eta}$, $a.s.$, and $\widehat{\theta} \longrightarrow \widetilde{\theta}, \;\; a.s, \;\;$ where

$$E_{LL}(\ln f_{LN}(X; \widetilde{\eta}, \widetilde{\theta}) = \max_{\eta, \theta} E_{LL}(\ln f_{LN}(X; \eta, \theta).$$

Let us denote

$$T_* = \ln \left[\frac{L_{LN}(\widetilde{\eta}, \widetilde{\theta})}{L_{LL}(\sigma, \xi)} \right].$$

(c)

$$\frac{1}{\sqrt{n}} \left[T - E_{LL}(T) \right] \text{ asymptotically equivalent to } \frac{1}{\sqrt{n}} \left[T_* - E_{LL}(T_*) \right].$$

PROOF OF THEOREM 2: Follows along the same line as of Theorem 1. ∎

In this case also $\widetilde{\eta}$ and $\widetilde{\theta}$ depend on σ and ξ and we are not making it explicit for brevity. We need $\widetilde{\eta}$ and $\widetilde{\theta}$ when $\sigma = 1$ and $\xi = 1$ and we will denote them $\widetilde{\eta}_1$ and $\widetilde{\theta}_1$ respectively. Now we will discuss how to compute $\widetilde{\eta}_1, \widetilde{\theta}_1$ and also $E_{LL}(T), V_{LL}(T)$ for further development. Observe that $\widetilde{\eta}_1, \widetilde{\theta}_1$ can be obtained by maximizing $g(\eta, \theta)$, where

$$g(\eta, \theta) = E_{LL} \left[\ln f_{LN}(X; \eta, \theta) \right] = -E_{LL} \left[\frac{(\ln X - \ln \theta)^2}{2\eta^2} + \ln X + \ln \eta + \frac{1}{2} \ln(2\pi) \right].$$

with respect to η and θ respectively. By simple calculation it can be easily observed that the maximization occurs at $\widetilde{\eta}_1 = \sqrt{3}$ and $\widetilde{\theta}_1 = 1$. Moreover,

$$\begin{aligned}
\lim_{n\to\infty} \frac{1}{n} E_{LL}(T) = AM_{LL} &= E_{LL}\left[\ln f_{LL}(X;1,1) - \ln f_{LN}(X;\widetilde{\eta}_1,\widetilde{\theta}_1)\right] \\
&= E_{LL}\left[-2\ln(1+X) + \frac{3}{2}(\ln X)^2 + \frac{1}{2}\ln 2\pi - \frac{1}{2}\ln 3\right] \\
&= -2 + \frac{1}{3}\pi^2 + \frac{1}{2}\ln 2\pi - \frac{1}{2}\ln 3 \approx 0.0144,
\end{aligned}$$

$$\begin{aligned}
\lim_{n\to\infty} \frac{1}{n} V_{LL}(T) = AV_{LL} &= V_{LL}\left[\ln f_{LL}(X;1,1) - \ln f_{LN}(X;\widetilde{\eta}_1,\widetilde{\theta}_1)\right] \\
&= V_{LL}\left[-2\ln(1+X) + \frac{3}{2}(\ln X)^2 + \ln X\right] \approx 0.0486.
\end{aligned}$$

4 Numerical Results

In this section we perform some numerical experiments to observe how the asymptotic results, derived in section 3, perform for different sample sizes and for different parent distributions. All the computations are performed at the Indian Institute of Technology Kanpur, using a Pentium IV processor. All the computations are performed using S-PLUS or R, and they can be obtained from the authors on request. We compute the probability of correct selection based on simulations and also based on the asymptotic distributions derived in section 3. Since the distributions are independent of the scale and the shape parameters in each case, we have considered them to be one, in all cases. We have reported the results for $n = 20$, 40, 60, 80, 100, 500.

First we consider the case when the parent distribution is log-normal. In this case we generate a random sample of size n from $LN(1,1)$. We compute T and check whether $T > 0$ or not. We replicate the process 1000 times and obtain an estimate of PCS. We also compute the PCS based on Theorem 1. The results are reported in Table 1. Similarly, we report the results when the parent distribution is log-logistic. The results are reported in Tables 2. In

each table, the first and second rows represent the results based on Monte Carlo simulations and asymptotic distributions respectively.

Table 1: The probability of correct selection based on Monte Carlo simulations and also based on asymptotic results when the data are from log-normal distribution and the alternative is log-logistic.

$n \longrightarrow$	20	40	60	80	100	500
MC	0.77	0.78	0.79	0.81	0.82	0.96
AS	0.64	0.70	0.73	0.77	0.79	0.97
ASC	0.73	0.74	0.77	0.79	0.81	0.96

Table 2: The probability of correct selection based on Monte Carlo simulations and also based on asymptotic results when the data are from log-logistic distribution and the alternative is log-normal.

$n \longrightarrow$	20	40	60	80	100	500
MC	0.42	0.53	0.61	0.66	0.69	0.94
AS	0.62	0.66	0.69	0.72	0.74	0.93
ASC	0.51	0.60	0.65	0.69	0.72	0.94

It is clear from the Tables 1 and 2 that as the sample size increases, the PCS increases as expected. Interestingly, when log-normal is the parent distribution, then the PCS based on Monte Carlo simulation is found to be significantly higher than the other case particularly for small sample sizes. For example, when the sample size is 20, and the parent distribution is log-normal, the PCS is 0.77. But when the parent distribution is log-logistic, for the same sample size the PCS is only 0.42. It seems if the log-logistic is the parent distribution then its approximation by the log-normal distribution is better than the other way. We have verified this by comparing the Kolmogorov-Smirnov distance and it is observed that the K-S distances between (i) log-logistic (parent) and best fitted log-normal and (ii) log-normal (parent) and the best fitted log-logistic are 0.023 and 0.015 respectively.

Finally, based on the numerical results we have suggested the following simple correction

factors which may be used mainly for small sizes:

$$\begin{aligned} AM_{LN} &= 0.0095 + \frac{0.0655}{n} + \frac{0.7270}{n^2} - \frac{3.0292}{n^3}, \\ AV_{LN,LL} &= 0.0906 - \frac{0.0414}{n} - \frac{0.3945}{n^2} + \frac{1.8426}{n^3}, \\ AM_{LL} &= 0.0144 - \frac{0.2681}{n} - \frac{0.4322}{n^2} + \frac{4.8427}{n^3}, \\ AV_{LL,LN} &= 0.0486 - \frac{0.7521}{n} + \frac{5.6154}{n^2} - \frac{18.5252}{n^3}. \end{aligned}$$

We have recalculated the PCS based on the correction factors and the results are reported in Tables 1 and 2 respectively.

It should be also mentioned that the Tables 1 and 2 can be easily used to compute the minimum sample size needed to discriminate between log-normal and log-logistic distributions for a user specified PCS. For example if the PCS = 0.70, then at least 100 sample size is needed to discriminate between the two distribution functions.

5 Data Analysis

In this section we analyze two data sets for illustrative purposes and use our method to discriminate between the distribution functions.

Data Set 1: This data set (from Linhart and Zuchini [27]) represents the failure times of the air conditioning system of an air plane (in hours) : 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.

We fit the two distribution functions, the MLEs of the different parameters for different distribution functions and the corresponding log-likelihood values are as follows. Log-Normal: $\widehat{\eta} = 1.3192$, $\widehat{\theta} = 28.7343$, $LL_{LN}(\widehat{\eta}, \widehat{\theta})$ = -151.706, Log-Logistic: $\widehat{\sigma} = 0.7259$, $\widehat{\xi} = 26.5007$, $LL_{LL}(\widehat{\sigma}, \widehat{\xi})$ = - 152.3468.

Therefore, based on the log-likelihood values clearly, log-normal distribution is the preferred one. The Kolmogorov-Smirnov distances and the corresponding p values (within brackets) between the empirical distribution function and the fitted distribution functions for the two cases are as follows: Log-Normal: 0.1047 (0.88), Log-logistic: 0.1300 (0.69). Therefore, it is clear that the fitted log-normal is much closer to the empirical distribution function than the log-logistic distribution. The non-parametric survival function and the fitted survival functions are plotted in Figure 1. The K-S distance between the two fitted survival functions is 0.043. We also present the observed, expected frequencies for differ-

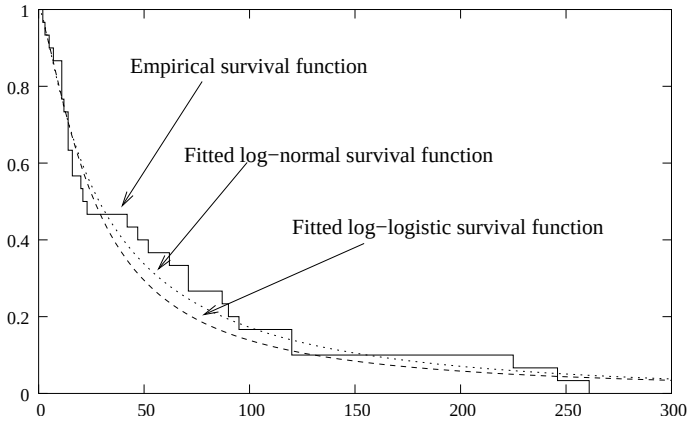

Figure 1: The empirical survival function and the fitted survival functions for data set 1.

ent groups and the corresponding χ^2 statistics for both the distributions to the fitted data. The results are presented in Table 4. The χ^2 values are 3.562 and 5.172 for the log-normal and log-logistic distributions respectively. In this case in terms of the log-likelihood values, K-S distances and also from the χ^2 values, between the two distribution functions, clearly log-normal is the better choice.

Data Set 2: The data is obtained from Lawless [26] and it represents the number of revolution before failure of each of 23 ball bearings in the life tests and they are as follows: 17.88, 28.92, 33.00, 41.52, 42.12, 45.60, 48.80, 51.84, 51.96, 54.12, 55.56, 67.80, 68.44, 68.64, 68.88, 84.12, 93.12, 98.64, 105.12, 105.84, 127.92, 128.04, 173.40.

Table 3: The observed, expected frequencies for different groups for both the distributions.

Intervals	Observed	Expected (LN)	Expected (LL)
0 - 15	11	9.33	9.39
15 - 30	5	6.01	6.88
30 - 60	3	5.91	6.37
60 - 100	6	3.57	3.21
100 -	5	5.18	4.15

In this case the two fitted distributions have the following MLEs: Log-Normal: $\widehat{\eta} = 0.5215$, $\widehat{\theta} = 63.4890$, $LL_{LN}(\widehat{\eta}, \widehat{\theta})$ = -113.1017, Log-Logistic: $\widehat{\sigma} = 0.3008$, $\widehat{\xi} = 64.0075$, $LL_{LL}(\widehat{\sigma}, \widehat{\xi})$ = - 113.3662. Therefore, based on the log-likelihood values in this case also, the log-normal distribution is the preferred one. The Kolmogorov-Smirnov (K-S) distance between the empirical distribution functions and the fitted distributions and the corresponding p values (in brackets) for log-normal and log-logistic distributions are 0.0901 (0.99) and 0.0937 (0.98) respectively. The non-parametric survival function and the fitted survival functions are plotted in Figure 2. The K-S distance between the two fitted distribution functions

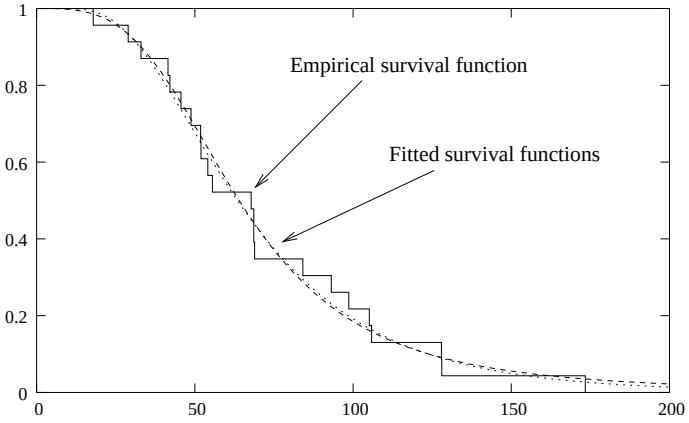

Figure 2: The empirical survival function and the fitted survival functions for data set 2.

is 0.018. Therefore, it is clear that based on the K-S distance the two distribution functions are almost equally close to the empirical distribution function. The observed and expected frequencies for log-normal and log-logistic distributions are provided in Table 4. The cor-

Table 4: The observed, expected frequencies for different groups for both the distributions.

Intervals	Observed	Expected (LN)	Expected (LL)
0 - 35	3	2.92	2.71
35 - 55	7	6.04	5.95
55-80	5	6.48	6.91
80 - 100	3	3.15	3.18
100 -	5	4.41	4.25

responding χ^2 values are 0.579 (log-normal) and 0.887 (log-logistic) respectively. It is clear from the log-likelihoods, K-S distances and also from the χ^2 values that the discrimination in this case is very difficult.

Since it is known that the log-normal has always unimodal hazard function where as the log-logistic distribution can have both unimodal as well as increasing hazard functions, we try to obtain an estimate of the shape of the hazard function from the observed data. A device called scaled TTT transform and its empirical version are relevant in this context. For a family with the survival function $S(y) = 1 - F(y)$, the scaled TTT transform, with $H_F^{-1}(u) = \int_0^{F^1(u)} S(y)dy$ defined for $0 < u < 1$ is $\phi_F(u) = H_F^{-1}(u)/H_F^{-1}(1)$. The empirical version of the scaled TTT transform is given by

$$\phi_n(r/n) = H_n^{-1}(r/n)/H_n^{-1}(1) = \left(\sum_{i=1}^{r} x_{i:n} + (n-r)x_{r:n}\right)/(\sum_{i=1}^{n} x_{i:n}),$$

here $r = 1, \ldots, n$ and $x_{i:n}$ for $i = 1, \ldots, n$ represent the order statistics of the sample. Aarset [1]showed that the scaled TTT transform is convex (concave) if the hazard rate is decreasing (increasing), and for bathtub (unimodal) hazard rates, the scaled TTT transform is first convex (concave) and then concave (convex). We have plotted the empirical version of the scaled TTT transform of the data set 2 in Figure 3. Since the empirical version of the scaled TTT transform is concave, it indicates that the hazard function is increasing. From the empirical version of the scaled TTT transfer we would prefer log-logistic than log-normal for data set 2.

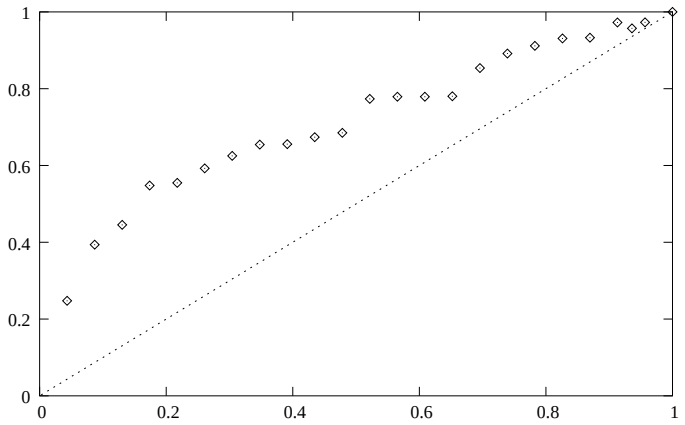

Figure 3: The empirical scaled TTT transform of the data set 2

6 Conclusions

In this paper we consider the problem of discriminating between the log-normal and log-logistic distribution functions. We have used the ratio of the maximized likelihood functions in discriminating between the two distribution functions. We have computed the PCS based on the Monte Carlo simulations for different sample sizes to see the performances of the method. Since both the distributions belong to the log-location-scale family, it is observed that the distributions of the ratio of the maximized likelihoods are independent of the unknown parameters. We compute the PCS based on the asymptotic distributions and suggested some small sample corrections to use the asymptotic results for small sizes. Tables 1 and 2 can be easily used to compute the minimum sample size needed for a given user specified PCS. We have analyzed two data sets for illustrative purposes. It is observed that for one data set the choice is quite clear but for the other data set the choice is not very clear from the ratio of maximized likelihoods, K-S distances and also from χ^2 values. We have used the empirical version of the scaled TTT in that case and we get some idea about the shape of the hazard function, which in turn helps us to choose the preferred distribution.

Appendix

THEOREM A.1 The K-S distance between the $LN(\eta,\theta)$ and $LL(\widetilde{\sigma},\widetilde{\xi})$ is independent of η and θ, where $\widetilde{\sigma}$ and $\widetilde{\xi}$ are obtained from part (b) of Lemma 1.

PROOF: Since θ and $\widetilde{\xi}$ are the respective scale parameters, it easily follows that $\widetilde{\xi}=c\theta$ for some constant c. Therefore, with out loss of generality we can take $\theta=1$. Furthermore, by simple observations on $E_{LN}(f_{LL}(X;\sigma,\xi))$, where $X\sim LN(\eta,1)$, it easily follows that

$$\widetilde{\sigma}=c_1\eta, \quad \text{and} \quad \ln\xi=c_2\eta,$$

where c_1 and c_2 are two constants independent of η. Now the K-S distance between $LN(\eta,1)$ and $LL(\widetilde{\sigma},\widetilde{\xi})$ can be written as

$$D=\sup_x\left|\Phi(\frac{\ln x}{\eta})-\frac{e^{\ln\left(\frac{x}{\widetilde{\xi}}\right)^{\frac{1}{\widetilde{\sigma}}}}}{1+e^{\ln\left(\frac{x}{\widetilde{\xi}}\right)^{\frac{1}{\widetilde{\sigma}}}}}\right|=\sup_z\left|\Phi(z)-\frac{c_3e^{c_1z}}{1+c_3e^{c_1z}}\right|, \tag{9}$$

where $c_3=e^{-\frac{c_2}{c_1}}$ is a constant. Since the (9) is independent of η, the result follows. ∎

THEOREM A.2 The K-S distance between the $LL(\sigma,\xi)$ and $LN(\widetilde{\eta},\widetilde{\theta})$ is independent of σ and ξ, where $\widetilde{\eta}$ and $\widetilde{\theta}$ are obtained from part (b) of Lemma 2.

PROOF: In this case since $\widetilde{\theta}=\xi$ and $\widetilde{\eta}=\sqrt{3}$, the result follows very easily by substituting those values in the K-S distance between $LL(\sigma,\xi)$ and $LN(\widetilde{\eta},\widetilde{\theta})$.

References

[1] Aarset, M. V. (1987), "How to identify a bathtub hazard rate", *IEEE Transactions on Reliability*, vol. 36, 106 - 108.

[2] Atkinson, A. (1969), "A test of discriminating between models", *Biometrika*, vol. 56, 337-341.

[3] Atkinson, A. (1970), "A method for discriminating between models" (with discussions), *Journal of the Royal Statistical Society*, Ser. B, vol. 32, 323-353.

[4] Bain, L.J. and Englehardt, M. (1980), "Probability of correct selection of Weibull versus gamma based on likelihood ratio", *Communications in Statistics*, Ser. A., vol. 9, 375-381.

[5] Balasooriya, C.P. and Abeysinghe, T. (1994), "Selecting between gamma and Weibull distributions approach based on prediction of order statistics", *Journal of Applied Statistics*, vol. 21, 17 - 27.

[6] Chambers, E.A. and Cox, D.R. (1967), "Discriminating between alternative binary response models", *Biometrika*, 54, 573-578.

[7] Chen, W.W. (1980), "On the tests of separate families of hypotheses with small sample size", *Journal of Statistical Computations and Simulations*, vol. 2, 183-187.

[8] Cox, D.R. (1961), "Tests of separate families of hypotheses", *Proceedings of the Fourth Berkeley Symposium in Mathematical Statistics and Probability*, Berkeley, University of California Press, 105-123.

[9] Cox, D.R. (1962), "Further results on tests of separate families of hypotheses", *Journal of the Royal Statistical Society*, Ser. B, vol. 24, 406-424.

[10] Dumonceaux, R., Antle, C.E. and Haas, G. (1973), "Likelihood ratio test for discriminating between two models with unknown location and scale parameters", *Technometrics*, vol. 15, 19-31.

[11] Dumonceaux, R. and Antle, C.E. (1973), "Discriminating between the log-normal and Weibull distribution", *Technometrics*, vol. 15, 923-926.

[12] Dyer, A.R. (1973), "Discrimination procedure for separate families of hypotheses", *Journal of the American Statistical Association*, vol. 68, 970-974.

[13] Fearn, D.H. and Nebenzahl, E. (1991), "On the maximum likelihood ratio method of deciding between the Weibull and Gamma distributions", *Communications in Statistics - Theory and Methods*, vol. 20, 579-593.

[14] Firth, D. (1988), "Multiplicative errors: log-normal or gamma?", *Journal of the Royal Statistical Society, Ser. B*, 266-268.

[15] Gupta, R. D. and Kundu, D. (1999). "Generalized exponential distributions", *Australian and New Zealand Journal of Statistics*, vol. 41, 173 - 188.

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

[17] Gupta, R. D. and Kundu, D. (2003b), "Closeness of gamma and generalized exponential distribution", *Communications in Statistics - Theory and Methods*, vol. 32, no. 4, 705-721.

[18] Gupta, R. D. and Kundu, D. (2004), "Discriminating between gamma and generalized exponential distributions", *Journal of Statistical Computation and Simulation*, vol. 74, no. 2, 107-121.

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

[20] Kappeman, R.F. (1982), "On a method for selecting a distributional model", *Communications in Statistics - Theory and Methods*, vol.11, 663-672.

[21] Kim, H., Sun, D. and Tsutakawa, R. K. (2002), "Lognormal vs. Gamma: Extra Variations", *Biometrical Journal*, vol.44, 305-323.

[22] Kundu, D., Gupta, R.D. and Manglick, A. (2005), "Discriminating between the log-normal and generalized exponential distributions", *Journal of Statistical Planning and Inference*, vol. 127, 213 - 227.

[23] Kundu, D. and Manglick, A. (2004), "Discriminating between the Weibull and log-normal distributions", *Naval Research Logistics*, vol. 51, 893 - 905.

[24] Kundu, D. and Manglick, A. (2004), "Discriminating between the Weibull and Log-Normal distributions", *Naval Research Logistics*, vol. 51, 893-905, 2004.

[25] Kundu, D. and Manglick, A. (2005), "Discriminating between the Log-Normal and gamma distributions", *Journal of the Applied Statistical Sciences*, vol. 14, 175-187, 2005.

[26] Lawless, J.F. (1982), *Statistical Models and Methods for Lifetime Data*, John Wiley and Sons, New York.

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

[28] Marshall, A.W., Meza, J.C. and Olkin, I. (2001), "Can data recognize its parent distribution?", *Journal of Computational and Graphical Statistics*, vol. 10, 555 - 580.

[29] Pascual, F.G. (2005), "Maximum likelihood estimation under misspecified log-normal and Weibull distributions", *Communications in Statistics - Simulation and Computations* vol. 34, 503 - 524, 2005.

[30] Pereira, B, de (1978), "Empirical comparison of some tests of separate families of hypothesis", *Metrika*, vol. 25, 219 - 234.

[31] Press, W.H., Flannery, B.P., Teukolsky, S.A. and Vetterling, W.T. (1993), *Numerical Recipes; The Art of Scientific Computing*, Cambridge University Press, Cambridge, U.K.

[32] Taylor, J.A. and Jakeman, A.J. (1985), "Identification of a distributional model", *Communications in Statistics - Theory and Methods*, vol. 14, 497-508.

[33] Qusenberry, C.P. and Kent, J. (1982), "Selecting among probability distributions used in reliability", *Technometrics*, vol. 24, 59-65.

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

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

[36] Wiens, B.L. (1999), "When log-normal and gamma models give different results: a case study", *American Statistician*, vol. 53, 89-93.