Discriminating Between The Weibull and Log-Normal Distributions

Debasis Kundu[1] Anubhav Manglick[2]

Abstract

Log-Normal and Weibull distributions are the most popular distributions for modeling skewed data. In this paper, we consider the ratio of the maximized likelihood in choosing between the two distributions. The asymptotic distribution of the logarithm of the maximized likelihood ratio has been obtained. It is observed that the asymptotic distribution is independent of the unknown parameters. The asymptotic distribution has been used to determine the minimum sample size required to discriminate between two families of distributions for a user specified probability of correct selection. We perform some numerical experiments to observe how the asymptotic methods work for different sample sizes. It is observed that the asymptotic results work quite well even for small samples also. Two real data sets have been analyzed.

Key Words and Phrases Asymptotic distributions; Likelihood ratio tests; Probability of correct selection; Location Scale Family.

Address of Correspondence: Professor Debasis Kundu, Department of Mathematics, Indian Institute of Technology Kanpur, Pin 208016, INDIA. e-mail: kundu@iitk.ac.in.

[1] Department of Mathematics, Indian Institute of Technology Kanpur, Pin 208016, INDIA.

[2] Faculty of Mathematics and Informatics, University of Passau, GERMANY.

1 Introduction:

We address the following problem in this paper. Suppose an experimenter has observed n data points, say $x_1, \ldots, x_n$ and he wants to use either two-parameter log-normal model or two parameter Weibull model, which one is preferable?

It is well known that both the log-normal and Weibull models can be used quite effectively to analyze skewed data set. Although, these two models may provide similar data fit for moderate sample sizes, but still it is desirable to select the correct or more nearly correct model, since the inferences based on the model will often involve tail probabilities, where the affect of the model assumptions are very critical. Therefore, even if we have small or moderate samples, it is still very important to make the best possible decision based on whatever data are available.

The problem of testing whether some given observations follow one of the two probability distributions is quite old in the statistical literature. See for example the work of Atkinson (1969, 1970), Bain and Englehardt (1980), Chambers and Cox (1967), Chen (1980), Cox (1961, 1962), Dyer (1973), Fearn and Nebenzahl (1991), Gupta and Kundu (2003, 2004), Kundu, Gupta and Manglick (2004), Wiens (1999) and the references therein.

In this paper, we consider the problem of discriminating between Weibull and log-normal distributions. We use the ratio of the maximized likelihood (RML) in discriminating between the two distribution functions and using the approach of White (1982a,b), we obtain the asymptotic distribution of the logarithm of RML. It is observed that the asymptotic distribution is asymptotically normal and it is independent of the unknown parameters. The asymptotic distribution can be used to compute the probability of correct selection (PCS) and it is observed in the simulation study that the asymptotic distribution works quite well even for small sample sizes. Dumonceaux and Antle (1973) proposed a likelihood ratio test

in discriminating between the log-normal and Weibull distributions. The asymptotic results can be used to obtain the critical regions of the corresponding testing of hypotheses problem also.

We also find the minimum sample size required to discriminate between the two distribution functions for a given PCS. Using the asymptotic distribution of the logarithm of RML, we obtain the minimum sample size required to discriminate between the two distribution functions for a given user specified protection level, *i.e.* the PCS.

The rest of the paper is organized as follows. We describe the likelihood ratio method in section 2. Asymptotic distributions of the logarithm of RML statistics under null hypotheses are obtained in section 3. Sample size determination has been performed in section 4. Some numerical experiments are performed in section 5 and two real data sets are analyzed in section 6. Finally we conclude the paper in section 7.

We use the following notation for the rest of the paper. The density function of a log-normal random variable with scale parameter $\theta > 0$ and shape parameter $\sigma > 0$, will be denoted by

$$f_{LN}(x;\sigma,\theta) = \frac{1}{\sqrt{2\pi}x\sigma} e^{-\frac{(\ln x - \ln\theta)^2}{2\sigma^2}}; \quad \text{for} \;\; x > 0. \tag{1.1}$$

A log-normal distribution with the shape and scale parameters σ and θ will be denoted by $LN(\sigma,\theta)$. The density function of a Weibull distribution, with shape parameter $\beta > 0$ and scale parameter $\lambda > 0$, will be denoted by

$$f_{WE}(x;\beta,\lambda) = \beta\lambda^{\beta} x^{\beta-1} e^{-(x\lambda)^{\beta}}; \quad \text{for} \;\; x > 0. \tag{1.2}$$

A Weibull distribution with the shape and scale parameters β and λ respectively, will be denoted by $WE(\beta,\lambda)$. We denote $\psi(x) = \frac{d}{dx} ln(\Gamma(x))$ and $\psi'(x) = \frac{d}{dx}\psi(x)$, as the digamma and polygamma functions respectively.

2 Ratio of Maximized Likelihood

In this section, we assume that we have a sample $X_1, \ldots, X_n$, from one of the two distribution functions. The likelihood functions, assuming that the data follow $LN(\sigma, \theta)$ or $WE(\beta, \lambda)$, are

$$L_{LN}(\sigma, \theta) = \prod_{i=1}^{n} f_{LN}(X_i; \sigma, \theta)$$

and

$$L_{WE}(\beta, \lambda) = \prod_{i=1}^{n} f_{WE}(X_i; \beta, \lambda),$$

respectively. The logarithm of RML is defined as

$$T = \ln \left[\frac{L_{LN}(\hat{\sigma}, \hat{\theta})}{L_{WE}(\hat{\beta}, \hat{\lambda})} \right]. \tag{2.1}$$

Here $(\hat{\sigma}, \hat{\theta})$ and $(\hat{\beta}, \hat{\lambda})$ are maximum likelihood estimators (MLEs) of (σ, θ) and (β, λ) respectively based on the sample $X_1, \ldots, X_n$. The logarithm of RML can be written as follows;

$$T = n \left[\frac{1}{2} - \ln \left(\hat{\sigma} \hat{\beta} (\hat{\lambda} \hat{\theta})^{\hat{\beta}} \sqrt{2\pi} \right) \right]. \tag{2.2}$$

For log-normal distribution

$$\hat{\theta} = \left(\prod_{i=1}^{n} X_i \right)^{\frac{1}{n}} \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(\ln X_i - \ln \hat{\theta} \right)^2, \tag{2.3}$$

and for Weibull distribution $\hat{\beta}$ and $\hat{\lambda}$ satisfy the following relation

$$\hat{\lambda} = \left(\frac{n}{\sum_{i=1}^{n} X_i^{\hat{\beta}}} \right)^{\frac{1}{\hat{\beta}}}. \tag{2.4}$$

The following discrimination procedure can be used. Choose the log-normal distribution if the test statistic $T > 0$, otherwise choose the Weibull distribution as the preferred model. Note that if we transform the data as $Y_i = \ln X_i$ and consider the logarithm of RML of the corresponding transformed distributions, namely normal and extreme-value distributions, then the value of the test statistic will be same as (2.2). Therefore, using the results of

Dumonceaux, Antle and Haas (1973), it follows that the distribution of T is independent of (σ, θ) and (β, λ) if the data follow $LN(\sigma, \theta)$ and $WE(\beta, \lambda)$ respectively. The discrimination procedure is equivalent to the likelihood ratio method.

3 Asymptotic Properties of the Logarithm of RML

In this section we derive the asymptotic distribution of RML for two different cases. We use the following notation. Almost sure convergence will be denoted by $a.s.$ For any Borel measurable function, $h(.)$, $E_{LN}(h(U))$, and $V_{LN}(h(U))$ will denote the mean and variance of $h(U)$ under the assumption that U follows $LN(\sigma, \theta)$. Similarly, we define, $E_{WE}(h(U))$ and $V_{WE}(h(U))$ as mean and variance of $h(U)$ under the assumption that U follows $WE(\beta, \lambda)$. Moreover, if $g(.)$ and $h(.)$ are two Borel measurable function, we define $cov_{LN}(g(U), h(U)) = E_{LN}(g(U)h(U)) - E_{LN}(g(U))E_{LN}(h(U))$ and $cov_{WE}(g(U)h(U)) = E_{WE}(g(U)h(U))$ - $E_{WE}(g(U))E_{WE}(h(U))$, where U follows $LN(\sigma, \theta)$ and $WE(\beta, \lambda)$ respectively.

Case 1: The Data Follow Log-Normal Distribution

In this case we have the following main result.

Theorem 1: Under the assumption that the data follow $LN(\sigma, \theta)$, the distribution of T as defined in (2.2) is approximately normally distributed with mean $E_{LN}(T)$ and $V_{LN}(T)$.

To prove theorem 1, we need the following lemma.

Lemma 1: Suppose the data follow $LN(\sigma, \theta)$, then as $n \to \infty$, we have

(1) $\hat{\sigma} \to \sigma$ a.s., $\hat{\theta} \to \theta$ a.s., where

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

(2) $\hat{\beta} \to \tilde{\beta}$ a.s., $\hat{\lambda} \to \tilde{\lambda}$ a.s., where

$$E_{LN}\left(\ln(f_{WE}(X;\tilde{\beta},\tilde{\lambda}))\right) = \max_{\beta,\lambda} E_{LN}\left(\ln(f_{WE}(X;\beta,\lambda))\right).$$

Note that $\tilde{\beta}$ and $\tilde{\lambda}$ may depend on σ and θ, but we do not make it explicit for brevity. Let us denote;

$$T^* = \ln\left[\frac{L_{LN}(\sigma,\theta)}{L_{WE}(\tilde{\beta},\tilde{\lambda})}\right], \tag{3.1}$$

(3) $n^{-\frac{1}{2}}\left(T - E_{LN}(T)\right)$ is asymptotically equivalent to $n^{-\frac{1}{2}}\left(T^* - E_{LN}(T^*)\right)$.

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

Proof of Theorem 1: Using the Central Limit Theorem (CLT), it can be easily seen that $n^{-\frac{1}{2}}\left(T^* - E_{LN}(T^*)\right)$ is asymptotically normally distributed. Therefore, the proof immediately follows from part (3) of Lemma 1 and the CLT.

Now we discuss how to obtain $\tilde{\beta}$ and $\tilde{\lambda}$, $E_{LN}(T)$ and $V_{LN}(T)$. Let us define;

$$g(\beta,\lambda) = E_{LN}\left(\ln(f_{WE}(X;\beta,\lambda))\right) = E_{LN}\left[\ln\beta + \beta\ln\lambda + (\beta-1)\ln X - (\lambda X)^{\beta}\right]$$

$$= \ln\beta + \beta\ln\lambda + (\beta-1)\ln\theta - (\lambda\theta)^{\beta}e^{\frac{\beta^2\sigma^2}{2}}. \tag{3.2}$$

By differentiating $g(\beta,\lambda)$ with respect to β and λ and equating them to zero, we obtain

$$\tilde{\beta} = \frac{1}{\sigma} \quad \text{and} \quad \tilde{\lambda} = \frac{1}{\theta}e^{-\frac{\sigma}{2}}. \tag{3.3}$$

Now we provide the expression for $E_{LN}(T)$ and $V_{LN}(T)$. Note that $\lim_{n\to\infty}\frac{E_{LN}(T)}{n}$ and $\lim_{n\to\infty}\frac{V_{LN}(T)}{n}$ exist. Suppose, we denote $\lim_{n\to\infty}\frac{E_{LN}(T)}{n} = AM_{LN}$ and $\lim_{n\to\infty}\frac{V_{LN}(T)}{n} = AV_{LN}$. It is already observed that the distribution of T is independent of σ and θ, therefore to compute $E_{LN}(T)$ and $V_{LN}(T)$, without loss of generality, we consider $\sigma = \theta = 1$. For large n,

$$\begin{aligned}
\frac{E_{LN}(T)}{n} &\approx AM_{LN} = E_{LN}\left[\ln f_{LN}(X;1,1) - \ln f_{WE}(X;\tilde{\beta},\tilde{\lambda})\right] \\
&= E_{LN}\left[-\frac{1}{2}\ln(2\pi) - \tilde{\beta}\ln X - \frac{1}{2}(\ln X)^2 - \ln\tilde{\beta} - \tilde{\beta}\ln\tilde{\lambda} + (X\tilde{\lambda})^{\tilde{\beta}}\right] \\
&= -\frac{1}{2}\ln(2\pi) - \frac{1}{2} - \ln\tilde{\beta} - \tilde{\beta}\ln\tilde{\lambda} + \tilde{\lambda}^{\tilde{\beta}}e^{\frac{1}{2}\tilde{\beta}^2} \\
&= -\frac{1}{2}\ln(2\pi) + 1 = 0.0810614 \qquad (3.4)
\end{aligned}$$

We also have

$$\begin{aligned}
\frac{V_{LN}(T)}{n} &\approx AV_{LN}(\sigma) = V_{LN}\left(\ln f_{LN}(X;1,1) - \ln f_{WE}(X;\tilde{\beta},\tilde{\lambda})\right) \\
&= V\left(\tilde{\beta}\ln X + \frac{1}{2}(\ln X)^2 - (\tilde{\lambda}X)^{\tilde{\beta}}\right) \\
&= \tilde{\beta}^2 + \frac{1}{2} + \tilde{\lambda}^{2\tilde{\beta}}\left(e^{2\tilde{\beta}^2} - e^{\tilde{\beta}^2}\right) - 3\tilde{\beta}^2\tilde{\lambda}^{\tilde{\beta}}e^{\frac{1}{2}\tilde{\beta}^2} = e - \frac{5}{2} = 0.2182818. \qquad (3.5)
\end{aligned}$$

Case 2: The Data Follow Weibull Distribution

Theorem 2: Under the assumption that the data follow $WE(\beta,\lambda)$, the distribution of T as defined in (2.2), is asymptotically normally distributed with mean $E_{WE}(T)$ and variance $V_{WE}(T)$.

To prove Theorem 2, we need Lemma 2, similar to Lemma 1.

Lemma 2: Suppose the data follow $WE(\beta,\lambda)$, then as $n\to\infty$, we have

(1) $\hat{\beta}\to\beta$ a.s., $\hat{\lambda}\to\lambda$ a.s. where

$$E_{WE}\left(\ln(f_{WE}(X;\beta,\lambda))\right) = \max_{\bar{\beta},\bar{\lambda}} E_{WE}\left(\ln(f_{WE}(X;\bar{\beta},\bar{\lambda}))\right).$$

(2) $\hat{\sigma}\to\tilde{\sigma}$ a.s., $\hat{\theta}\to\tilde{\theta}$ a.s, where

$$E_{WE}\left(\ln(f_{LN}(X:\tilde{\sigma},\tilde{\theta}))\right) = \max_{\sigma,\theta} E_{WE}\left(\ln(f_{LN}(X;\sigma,\theta))\right).$$

Note that here also $\tilde{\sigma}$ and $\hat{\theta}$ may depend on β, but we do not make it explicit for brevity.

Let us denote here;

$$T_* = \ln\left[\frac{L_{LN}(\tilde{\sigma},\tilde{\theta})}{L_{WE}(\beta,\lambda)}\right] = \ln\left[\frac{L_{LN}(\tilde{\sigma},\tilde{\theta})}{L_{WE}(\beta,1)}\right]. \tag{3.6}$$

(3) $n^{-\frac{1}{2}}\left[T - E_{WE}(T)\right]$ is asymptotically equivalent to $n^{-\frac{1}{2}}\left[T_* - E_{WE}(T_*)\right]$.

Proof of Lemma 2 It also follows from Theorem 1 of White (1982b).

Proof of Theorem 2: It follows similarly as Theorem 1.

Now we discuss how to obtain $\tilde{\sigma}$, $\tilde{\theta}$, $E_{WE}(T)$ and $V_{WE}(T)$. Let us define;

$$\begin{aligned} h(\sigma,\theta) &= E_{WE}\left(\ln(f_{LN}(X;\sigma,\theta))\right) = E_{WE}\left[-\frac{1}{2}\ln(2\pi) - \ln X - \ln\sigma - \frac{(\ln X - \ln\theta)^2}{2\sigma^2}\right] \\ &= -\frac{1}{2}\ln(2\pi) - \frac{1}{\beta}\psi(1) + \ln\lambda - \ln\sigma \\ &\quad -\frac{1}{2\sigma^2}\left[\frac{1}{\beta^2}\Gamma''(1) - \frac{2}{\beta}\psi(1)\ln\lambda + (\ln\lambda)^2 - 2\ln\theta\left(\frac{1}{\beta}\psi(1) - \ln\lambda\right) + (\ln\theta)^2\right]. \end{aligned} \tag{3.7}$$

Therefore, by differentiating $h(\sigma,\theta)$ with respect to σ and θ and equating them to zero, we obtain $\tilde{\sigma}$ and $\tilde{\theta}$ as

$$\tilde{\theta} = \frac{1}{\lambda}e^{\frac{1}{\beta}\psi(1)}, \qquad \text{and} \qquad \tilde{\sigma} = \frac{\sqrt{\psi'(1)}}{\beta}. \tag{3.8}$$

Now we provide the expression for $E_{WE}(T)$ and $V_{WE}(T)$. Similarly as before, $\lim_{n\to\infty}\frac{E_{WE}(T)}{n}$ and $\lim_{n\to\infty}\frac{V_{WE}(T)}{n}$ exist and we denote $\lim_{n\to\infty}\frac{E_{WE}(T)}{n} = AM_{WE}$ and $\lim_{n\to\infty}\frac{V_{WE}(T)}{n} = AV_{WE}$. Note that as mentioned before, the distribution of T is independent of β and λ and we take them to be 1 for the calculations of AM_{WE} and AV_{WE}. Therefore for large n

$$\begin{aligned} \frac{E_{WE}(T)}{n} &\approx AM_{WE} = E_{WE}\left[\ln(f_{LN}(X;\tilde{\sigma},\tilde{\theta})) - \ln(f_{WE}(X;1,1))\right] \\ &= -\frac{1}{2}\ln(2\pi) - \ln\tilde{\sigma} - \frac{1}{2\tilde{\sigma}^2}\left[\frac{\pi^2}{6} + \left(\psi(1) - \ln\tilde{\theta}\right)^2\right] + 1 - \psi(1) \\ &= \frac{1}{2} - \frac{3}{2}\ln\pi + \frac{1}{2}\ln 3 - \psi(1) = -0.0905730. \end{aligned} \tag{3.9}$$

Note that the second equality follows by taking the expectations of the likelihood functions after putting the values of $\tilde{\sigma}$ and $\tilde{\theta}$ from (3.8). Moreover,

$$\frac{V_{WE}(T)}{n} \approx AV_{WE}(\beta) = V_{WE}\left[\ln(f_{LN}(X;\tilde{\sigma},\tilde{\theta})) - \ln(f_{WE}(X;1,1))\right]$$

$$
\begin{aligned}
&= V_{WE}\left[\ln X + \frac{(\ln X - \ln\tilde{\theta})^2}{2\tilde{\sigma}^2} - X\right] \\
&= \left(1 - \frac{\ln\tilde{\theta}}{\tilde{\sigma}^2}\right)^2 \left(\Gamma''(1) - (\Gamma'(1))^2\right) + \frac{1}{4\tilde{\sigma}^4}\left(\Gamma^{(4)}(1) - (\Gamma''(1))^2\right) + 1 - \frac{2}{\tilde{\sigma}^2}\Gamma'(1) \\
&\quad + \frac{1}{\tilde{\sigma}^2}\left(1 - \frac{\ln\theta}{\tilde{\sigma}^2}\right)\left(\Gamma^{(3)}(1) - \Gamma'(1)\Gamma''(1)\right) - 2\left(1 - \frac{\ln\tilde{\theta}}{\beta\tilde{\sigma}^2}\right) \\
&= \left(1 - \frac{\psi(1)}{\psi'(1)}\right)^2 \left(\Gamma''(1) - (\Gamma'(1))^2\right) + \frac{1}{4(\psi'(1))^2}\left(\Gamma^{(4)}(1) - (\Gamma''(1))^2\right) \\
&\quad + 1 + \frac{1}{\psi'(1)}\left(1 - \frac{\psi(1)}{\psi'(1)}\right)\left(\Gamma^{(3)}(1) - \Gamma'(1)\Gamma''(1)\right) - 2\left(1 - \frac{\psi(1)}{\psi'(1)}\right) - 2\frac{\psi(1)}{\psi'(1)} \\
&= \psi'(1)\left(1 - \frac{\psi(1)}{\psi'(1)}\right)^2 + \frac{1}{4(\psi'(1))^2}\left(\Gamma^{(4)}(1) - (\Gamma''(1))^2\right) - 1 \\
&\quad + \frac{1}{\psi'(1)}\left(1 - \frac{\psi(1)}{\psi'(1)}\right)\left(\Gamma^{(3)}(1) - \Gamma'(1)\Gamma''(1)\right) \\
&= 0.2834081. \qquad (3.10)
\end{aligned}
$$

4 Determination of Sample Size and Testing

4.1 Minimum Sample Size Determination

In this subsection section we propose a method to determine the minimum sample size needed to discriminate between the Weibull and log-normal distributions for a given user specified PCS. It is expected that the user specifies the PCS before hand.

First we consider Case 1, *i.e.* the data are assumed to follow $LN(\sigma, \theta)$. Since T is asymptotically normally distributed with mean $E_{LN}(T)$ and variance $V_{LN}(T)$, therefore, the PCS is

$$PCS = P[T > 0] \approx \Phi\left(\frac{E_{LN}(T)}{\sqrt{V_{LN}(T)}}\right) = \Phi\left(\frac{n \times AM_{LN}}{\sqrt{n \times AV_{LN}}}\right). \qquad (4.1)$$

Here Φ is the distribution function of the standard normal distribution. Now to determine

the minimum sample size required to achieve at least p^* protection level, we equate;

$$\Phi\left(\frac{n \times AM_{LN}}{\sqrt{n \times AV_{LN}}}\right) = p^* \tag{4.2}$$

and obtain

$$n = \frac{z_{p^*}^2 AV_{LN}}{(AM_{LN})^2}. \tag{4.3}$$

Here z_{p^*} is the $100p^*$ percentile points of a standard normal distribution. For Case 2, *i.e.* when the data follow $WE(\beta, \lambda)$, we obtain

$$n = \frac{z_{p^*}^2 AV_{WE}}{(AM_{WE})^2}. \tag{4.4}$$

Therefore, to achieve overall p^* protection level, we need at least

$$n = z_{p^*}^2 \max\{\frac{AV_{LN}}{(AM_{LN})^2}, \frac{AV_{WE}}{(AM_{WE})^2}\} = z_{p^*} \max\{33.2, 34.5\} = z_{p^*} 35. \tag{4.5}$$

4.2 Testing of Hypotheses

In this subsection we consider the discrimination problem as a testing of hypothesis problem as it was considered by Dumonceaux and Antle (1973). Dumonceaux and Antle (1973) considered the problem as the following two testing of hypotheses problems.

$$\text{Problem 1:} \quad H_0 : \text{Log-Normal} \quad vs.\ H_1 : \text{Weibull}. \tag{4.6}$$

$$\text{Problem 2:} \quad H_0 : \text{Weibull} \quad vs.\ H_1 : \text{Log-Normal}. \tag{4.7}$$

Dumonceaux and Antle (1973) provided the exact critical regions and the powers of the likelihood ratio tests based on Monte Carlo simulations. Our asymptotic results derived in the previous section can be used for testing the above two hypotheses as follows:

Test 1: For Problem 1: Reject the null hypothesis H_0 at the α % level of significance, if $T < n \times 0.0810614 - z_\alpha \times \sqrt{n \times 0.2182818}$, and accept otherwise.

Test 2: For Problem 2: Reject the null hypothesis H_0 at the α % level of significance, if $T > -n \times 0.0905730 + z_\alpha \times \sqrt{n \times 0.2834081}$, and accept otherwise.

5 Numerical Experiments:

In this section we perform some numerical experiments to observe how these asymptotic results derived in section 3, work for different sample sizes. All computations are performed at the Indian Institute of Technology Kanpur, using Pentium IV processor. We use the random deviate generator of Press *et al.* (1993) and all the programs are written in C. They can be obtained from the authors on request without any cost. We compute the PCS based on simulations and we also compute it based on the asymptotic normality results derived in section 3. Since the distribution (numerical value) of T is independent of the shape and scale parameters, we consider the shape and scale parameters to be one in all cases. We consider different sample sizes namely $n = 20$, 40, 60, 80 and 100.

First we consider the case when the null distribution is Log-Normal and the alternative is Weibull. In this case we generate a random sample of size n from $LN(1,1)$ and compute T as defined in (2.2) and check whether T is positive or negative. We replicate the process 10,000 times and obtain an estimate of PCS. We also compute the PCS based on the asymptotic results derived in the Section 3. The results are reported in Table 1. Similarly, we obtain the results when the null distribution is Weibull and the alternative is Log-Normal, and the results are reported in Table 2.

It is quite clear from Tables 1 and 2 that as sample size increases the PCS increases as expected. Even when the sample size is 20, asymptotic results work quite well for both the cases. We performed normality tests on 10,000 T values and the null hypothesis can not be rejected even when the sample size is 20. From the simulation study, it is recommended that asymptotic results can be used quite effectively even when the sample size is small.

Now we consider the discrimination problem as a testing of hypothesis problem as defined in the previous section. Let us define the rejection regions as $\{T < 0\}$ and $\{T > 0\}$ for

problems 1 and 2 respectively. Therefore, it is immediate that P[Type I error] = 1 - PCS. From Tables 1 and 2, it also clear that the P[Type I error] varies between 0.22 and 0.04 as the sample size varies between 20 and 100 for both problems 1 and 2. Similarly, the power of the test varies between 0.78 and 0.96 as sample size varies between 20 and 100.

6 Data Analysis

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

Data Set 1: The first data set is as follows; (Lawless; 1986, page 228). The data given arose in tests on endurance of deep groove ball bearings. The data are the number of million revolutions before failure for each of the 23 ball bearings in the life tests and they are: 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.

When we use the Log-Normal model, the MLEs of the different parameters are $\hat{\theta} = 63.4890$, $\hat{\sigma} = 0.5215$ and the corresponding log-likelihood (LL) value is -113.1017. The Kolmogorv-Smirnov (K-S) distance between the data and the fitted log-normal distribution function is 0.0901 and the corresponding p value is 0.98. Similarly when we fit the Weibull model, the MLEs are $\hat{\beta} = 2.1050$, $\hat{\lambda} = 0.0122$ and the corresponding LL value is -113.6887. The non-parametric survival function, fitted Weibull survival function and fitted log-normal survival function are plotted in Figure 1.

The K-S distance between the data and the fitted Weibull distribution function is 0.1521 and the corresponding p value is 0.63. We also present the observed, expected frequencies for different groups and the corresponding χ^2 statistics for both the distributions to the fitted data. The results are presented below:

TABLE A

Intervals	Observed	LN	WE
0-35	3	2.92	3.01
35-55	7	6.04	5.31
55-80	5	6.48	6.52
80-100	3	3.15	3.62
100-	5	4.41	4.54

The χ^2 values are 0.579 and 1.045 for Log-Normal and Weibull distributions respectively. For data set 1, K-S distances, χ^2 values and Figure 1 indicate that both the distributions provide quite good fit to the data set.

The logarithm of RML *i.e.*, $T = 0.5870 > 0$. It indicates to choose the Log-Normal model. From (4.1), it is clear that if the data follow Log-Normal distribution then based on sample size 23, PCS = 0.80 and if the data follow Weibull then PCS = 0.79. Therefore, in this case the PCS is at least $\min\{0.79, 0.80\} = 0.79$. Based on the assumption that the data follow Log-Normal distribution, the p-value = 0.28. Similarly based on the assumption that the data follow Weibull distribution, the p-value = 0.15. Comparing the two p values also we would prefer to choose Log-Normal distribution over Weibull distribution. Therefore, in this case, the LL values, χ^2 values, K-S distances and our proposed method indicate to choose the Log-Normal model and the probability of correct selection is at least 79%. If we consider the two testing of hypotheses problems (4.6) and (4.7), then based on the data we can not reject the null hypotheses in both cases even with the 15% level of significance.

DATA SET 2: The second data set (Linhart and Zucchini; 1986, page 69) represents the failure times of the air conditioning system of an airplane: 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.

For Data set 2, when we use the Log-Normal model, we have the following results: $\hat{\theta}$ = 28.7343, $\hat{\sigma}$ = 1.3192, LL = -151.706, χ^2 = 3.562, K-S = 0.1047 and the corresponding p

value = 0.88. Similarly, when we use the Weibull model they are $\hat{\beta} = 0.8554$, $\hat{\lambda} = 0.0183$, LL = -152.007, $\chi^2 = 3.053$, K-S = 0.1540 and the corresponding p value = 0.44. The non-parametric survival function, fitted log-normal survival function and fitted Weibull survival function are plotted in Figure 2. For data set 2 also K-S distances, χ^2 values and Figure 2 indicate that both the distributions provide quite reasonable fit to the data set.

The observed and the expected frequencies for different groups are presented below.

TABLE B

Intervals	Observed	LN	WE
0-15	11	9.33	8.45
15-30	5	6.01	5.06
30-60	3	5.91	6.33
60-100	6	3.57	4.55
100-	5	5.18	5.62

The logarithm of RML, $T = 0.3012$. In this case also since $T > 0$, we choose Log-Normal distribution as the preferred model. In this case the PCS = $\min\{0.83, 0.83\} = 0.83$. Based on the assumption that the data follow Log-Normal distribution, the p-value = 0.20 and similarly under the assumption that the data follow Weibull distribution, the corresponding p-value = 0.15. Based on the p values also we prefer to choose the Log-Normal model over Weibull model for data set 2. In this case also, if we consider the two testing of hypotheses problem, then based on the data we can not reject the null hypotheses in both cases even with the 15% level of significance.

7 CONCLUSIONS

In this paper we consider the problem of discriminating the two important families of distribution functions namely the log-normal and Weibull families. We consider the statistic

based on the ratio of maximized likelihoods and obtain the asymptotic distributions of the test statistics under null hypotheses. It is observed that the asymptotic distributions are asymptotically normal and they are independent of the parameters of the null distribution. We compare the probability of correct selection obtained using Monte Carlo simulations and the proposed asymptotic results, it is observed that the asymptotic results work very well even when the sample size is as small as 20. The normality tests on RML statistic suggests that T follows normal distribution, even when the sample size very small. Therefore the asymptotic results can be used quite effectively to calculate the PCS for a given data set. We use these asymptotic results to calculate the minimum sample size required to discriminate the two probability distributions for a given PCS. Our method can be used for discriminating any two members of the location and scale family. The exact mean and variance of the corresponding normal distribution needs to be derived in each case. Finally, we should mention that for a given data set it may happen that none of the two distribution functions provide good fit. It should be clear from the $K - S$ values and also from the χ^2 values. Those cases some other family members should be used.

ACKNOWLEDGMENTS: The authors would like to thank two referees and one associate editor for some very constructive suggestions.

References

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

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

[3] 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.

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

[5] 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.

[6] 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.

[7] 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.

[8] 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.

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

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

[11] 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.

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

[13] Gupta, R.D. and Kundu, D. (2004), "Discriminating between Gamma and Generalized exponential distributions", *Journal of Statistical Computation and Simulation*, vol. 74, 107-121.

[14] Kundu, D., Gupta, R.D and Manglick, A. (2004), "Discriminating between the log-normal and the generalized exponential distributions", *Journal of Statistical Planning and Inference*, (to appear).

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

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

[17] 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.

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

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

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

Table 1

The probability of correct selection based on Monte Carlo simulations (MC) with 10,000 replications and also based on the asymptotic results (AS) when the null distribution is Log-Normal and the alternative is Weibull.

n	30	60	90	120	150
MC	0.8401	0.9097	0.9481	0.9708	0.9796
AS	0.8405	0.9206	0.9579	0.9768	0.9871

Table 2

The probability of correct selection based on Monte Carlo simulations (MC) with 10,000 replications and also based on the asymptotic results (AS) when the null distribution is Weibull and the alternative is Log-Normal.

n	30	60	90	120	150
MC	0.7300	0.8592	0.9227	0.9499	0.9691
AS	0.7927	0.8689	0.9151	0.9436	0.9619