Discriminating Between The Log-normal and Gamma Distributions

Debasis Kundu[*]& Anubhav Manglick[†]

Abstract

For a given data set the problem of selecting either log-normal or gamma distribution with unknown shape and scale parameters is discussed. It is well known that both these distributions can be used quite effectively for analyzing skewed non-negative data sets. In this paper, we use the ratio of the maximized likelihoods in choosing between the log-normal and gamma distributions. We obtain asymptotic distributions of the ratio of the maximized likelihoods and use them to determine the minimum sample size required to discriminate between these two distributions for user specified probability of correct selection and tolerance limit.

Key Words and Phrases: Asymptotic distribution; Kolmogorov-Smirnov distances; probability of correct selection; tolerance level.

1 Introduction

It is a quite important problem in statistics to test whether some given observations follow one of the two possible probability distributions. In this paper we consider the problem of selecting either log-normal or gamma distribution with unknown shape and scale parameters for a given data set. It is well known (Johnson, Kotz and Balakrishnan [13]) that both log-normal and gamma distributions can be used quite effectively in analyzing skewed

[*] Department of Mathematics, Indian Institute of Technology Kanpur, Pin 208016, INDIA. E-mail: kundu@iitk.ac.in, corresponding author

[†] Faculty of Mathematics and Informatics, University of Passau, GERMANY

positive data set . Sources in the literature indicate that these two distributions are often interchangeable (Wiens [18]). Therefore, to analyze a skewed positive data set an experimenter might wish to select one of them. Although these two models may provide similar data fit for moderate sample sizes but it is still desirable to choose the correct or nearly correct order model, since the inference based on a particular model will often involve tail probabilities where the affect of the model assumption will be more crucial. Therefore, even if large sample sizes are not available, it is very important to make the best possible decision based on the given observations.

The problem of testing whether some give observations follow one of the two probability distributions, is quite old in the statistical literature. Atkinson [1, 2], Chen [5], Chambers and Cox [4], Cox [6, 7], Jackson [12] and Dyer [9] considered this problem in general for discriminating between two arbitrary distribution functions. Due to the increasing applications of the lifetime distributions, special attention has been given to discriminate some specific lifetime distribution functions. Pereira [15] developed two tests to discriminate between log-normal and Weibull distributions. Dumonceaux and Antle [8] also considered the same problem of discriminating between log-normal and Weibull distributions. They proposed a test and provided its critical values in that paper. Fearn and Nebenzahl [10] used the maximum likelihood ratio method in discriminating between the Weibull and gamma distributions. Bain and Englehardt [3] provided the probability of correct selection (PCS) of Weibull versus gamma distributions based on extensive computer simulations. Firth [11] and Wiens [18] discussed the problem of discriminating between the log-normal and gamma distributions.

In this paper we consider the problem of discriminating between the log-normal and gamma distribution functions. We use the ratio of maximized likelihoods (RML) in discriminating between these two distributions, which was originally proposed by Cox [6, 7] in discriminating between two separate models. We obtain the asymptotic distributions of the RML. It is observed by extensive simulations study that these asymptotic distributions work quite well to compute the PCS, even if the sample size is not very high. Using these asymptotic distributions and the distance between these two distribution functions, we compute the minimum sample size required to discriminate the two distribution functions at a user specified protection level and a tolerance limit.

The rest of the paper is organized as follows. We briefly discuss the RML in section 2. We obtain the asymptotic distributions of RML in section 3. In section 4, we compute the minimum sample size required to discriminate between the two distribution functions. Some numerical experiments are performed to observe how the asymptotic results behave for finite sample in section 5. Data analysis are performed in section 6 and finally we conclude the paper 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 a gamma or from a log-normal distribution function. The density function of a log-normal random variable with scale parameter θ and shape parameter σ is denoted by

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

The density function of a gamma distribution with shape parameter α and scale parameter λ will be denoted by

$$f_{GA}(x;\alpha,\lambda) = \frac{1}{\lambda\Gamma(\alpha)} \left(\frac{x}{\lambda}\right)^{\alpha-1} e^{-\left(\frac{x}{\lambda}\right)}; \qquad x, \alpha, \lambda > 0. \tag{2}$$

A log-normal distribution with shape parameter σ and scale parameter θ will be denoted by $LN(\sigma,\theta)$ and similarly a gamma distribution with shape parameter α and scale parameter λ will be denoted as $GA(\alpha,\lambda)$.

The likelihood functions assuming that the data are coming from $GA(\alpha, \lambda)$ or $LN(\theta, \sigma)$ are

$$L_{GA}(\alpha, \lambda) = \prod_{i=1}^{n} f_{GA}(x; \alpha, \lambda) \quad \text{and} \quad L_{LN}(\theta, \sigma) = \prod_{i=1}^{n} f_{LN}(x; \theta, \sigma)$$

respectively. The RML is defined as

$$L = \frac{L_{LN}(\hat{\sigma}, \hat{\theta})}{L_{GA}(\hat{\alpha}, \hat{\lambda})}, \tag{3}$$

where $(\hat{\alpha}, \hat{\lambda})$ and $(\hat{\theta}, \hat{\sigma})$ are maximum likelihood estimators of (α, λ) and (θ, σ) respectively based on the sample $\{X_1, \ldots, X_n\}$. The natural logarithm of RML can be written as

$$T = n\left[\ln\left(\frac{\Gamma(\hat{\alpha})}{\hat{\sigma}}\right) - \hat{\alpha}\ln\left(\frac{\tilde{X}}{\hat{\lambda}}\right) + \frac{\bar{X}}{\hat{\lambda}} - \frac{1}{2\hat{\sigma}^2 n}\sum_{i=1}^{n}\left(\ln\left(\frac{X_i}{\theta}\right)\right)^2 - \frac{1}{2}\ln(2\pi)\right], \tag{4}$$

here $\bar{X}$ and $\tilde{X}$ are arithmetic and geometric means of $\{X_1, \ldots X_n\}$ respectively, *i.e.*

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad \tilde{X} = \left(\prod_{i=1}^{n} X_i\right)^{\frac{1}{n}}. \tag{5}$$

Note that in case of log-normal distribution, $\hat{\theta}$ and $\hat{\sigma}$ have the following forms;

$$\hat{\theta} = \tilde{X} \quad \text{and} \quad \hat{\sigma} = \left(\frac{1}{n}\sum_{i=1}^{n}\ln\left(\frac{X_i}{\hat{\theta}}\right)^2\right)^{\frac{1}{2}}. \tag{6}$$

Also $\hat{\alpha}$ and $\hat{\lambda}$ satisfy the following relation

$$\hat{\alpha} = \frac{\bar{X}}{\hat{\lambda}}. \tag{7}$$

The following procedure can be used to discriminate between gamma and log-normal distributions. Choose the log-normal distribution if $T > 0$, otherwise choose the gamma distribution as the preferred one. From the expression of T as given in (4), it is clear that if the data come from a log-normal distribution, then the distribution of T is independent of θ and depends only on σ. Similarly, if the data come from a gamma distribution, then its distribution depends only on α and it is independent of λ.

We estimate the PCS by using extensive computer simulations for different sample sizes and for different shape parameters. First we generate a sample of size n from a $LN(\sigma, 1)$ and we compute $\hat{\sigma}$, $\hat{\theta}$, $\hat{\alpha}$ and $\hat{\lambda}$ from that sample. Based on that sample we compute T and verify whether $T > 0$ or $T < 0$. We replicate the process 10,000 times and obtain the percentage of times it is positive. It provides an estimate of the PCS when the data come from a log-normal distribution. Exactly the same way we estimate the PCS when the data come from a gamma distribution. The results are reported in Tables 5 and 6 respectively.

Some of the points are quite clear from Tables 5 and 6. In both cases for fixed shape parameter as sample size increases the PCS increases as expected. When the data come from a log-normal distribution, for a fixed sample size, the PCS increases as the shape parameter decreases. Interestingly, when the data come from a gamma distribution the PCS increases as the shape parameter increases. From these simulation experiments, it is clear that the two distribution functions become closer if the shape parameter of the log-normal distribution decreases and the corresponding shape parameter of the gamma distribution increases.

3 Asymptotic Properties Of The RML

In this section we obtain the asymptotic distributions of RML for two different cases. From now on we denote the almost sure convergence by $a.s.$.

Case 1: The data are coming from a log-normal distribution.

We assume that n data points $\{X_1, \ldots, X_n\}$, are from a $LN(\sigma, \theta)$ and $\hat{\alpha}$, $\hat{\lambda}$, $\hat{\theta}$ and $\hat{\sigma}$ are same as defined before. We use following notations. For any Borel measurable function $h(.)$, $E_{LN}(h(U))$ and $V_{LN}(h(U))$ denote mean and variance of $h(U)$ under the assumption that U follows $LN(\sigma, \theta)$. Similarly we define $E_{GA}(h(U))$ and $V_{GA}(h(U))$ as mean and variance of $h(U)$ under the assumption that U follows $GA(\alpha, \lambda)$. If $g(.)$ and $h(.)$ are two Borel measurable functions, we define along the same line that $cov_{LN}(g(U), h(U)) = E_{LN}(g(U)h(U))$ - $E_{LN}(g(U))E_{LN}(h(U))$ and similarly $cov_{GA}(g(U), h(U))$ also, where U follows $LN(\theta, \sigma)$ and $GA(\alpha, \lambda)$ respectively. The following lemma is needed to prove the main result.

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

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

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

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

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

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

$$T^* = ln\left(\frac{L_{LN}(\sigma, \theta)}{L_{GA}(\tilde{\alpha}, \tilde{\lambda})}\right).$$

(iii) $n^{-\frac{1}{2}}\left[T - E_{LN}(T)\right]$ 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 [17, Theorem 1] and therefore it is omitted.

Now we can state the main result;

Theorem 1: Under the assumption that the data are from $LN(\sigma, \theta)$, T is asymptotically normally distributed with mean $E_{LN}(T)$ and variance $V_{LN}(T) = V_{LN}(T^*)$.

Proof of Theorem 1: Using the Central limit theorem and from part (ii) of lemma 1, it follows that $n^{-\frac{1}{2}}\left[T^* - E_{LN}(T^*)\right]$ is asymptotically normally distributed with mean zero and variance $V_{LN}(T^*)$. Therefore using part (iii) of lemma 1, the result immediately follows.

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

$$\begin{aligned} g(\alpha,\lambda) &= E_{LN}\left[\ln(f_{GA}(X;\alpha,\lambda))\right] \\ &= E_{LN}\left[(\alpha-1)\ln X - \frac{X}{\lambda} - \alpha\ln(\lambda) - \ln(\Gamma(\alpha))\right] \\ &= (\alpha-1)\ln\theta - \frac{\theta}{\lambda}e^{\frac{\sigma^2}{2}} + \alpha\ln\lambda + \ln(\Gamma(\alpha)). \end{aligned}$$

In this case, $\tilde{\alpha}$ and $\tilde{\lambda}$ have the following relations;

$$\tilde{\lambda} = \frac{\theta}{\tilde{\alpha}}e^{\frac{\sigma^2}{2}} \tag{8}$$

and

$$\psi(\tilde{\alpha}) = \ln\tilde{\alpha} - \frac{\sigma^2}{2}. \tag{9}$$

Here $\psi(x) = \frac{d}{dx}\ln\Gamma(x)$ is a psi function. Therefore, $\tilde{\alpha}$ can be obtained by solving the non-linear equation (9), and clearly it is a function of σ^2 only. Once $\tilde{\alpha}$ is obtained, $\tilde{\lambda}$ can be obtained from (8). It is immediate that $\left(\frac{\tilde{\lambda}}{\theta}\right)$ is also a function of σ^2 only.

Now we provide the expression for $E_{LN}(T)$ and $V_{LN}(T)$. Observe that $\lim_{n\to\infty}\frac{E_{LN}(T)}{n}$ and $\lim_{n\to\infty}\frac{V_{LN}(T)}{n}$ exist. We denote $\lim_{n\to\infty}\frac{E_{LN}(T)}{n} = AM_{LN}(\sigma^2)$ and $\lim_{n\to\infty}\frac{V_{LN}(T)}{n} = AV_{LN}(\sigma^2)$ respectively. Therefore for large n,

$$\begin{aligned} \frac{E_{LN}(T)}{n} &\approx AM_{LN}(\sigma) = E_{LN}\left[\ln(f_{LN}(X;\sigma,1)) - \ln(f_{GA}(X;\tilde{\alpha},\tilde{\lambda}))\right] \\ &= -\frac{1}{2}\ln(2\pi) - \ln\sigma - \frac{1}{2} - \frac{1}{\tilde{\lambda}}e^{\frac{\sigma^2}{2}} + \tilde{\alpha}\ln\tilde{\lambda} + \ln\Gamma(\tilde{\alpha}) \end{aligned} \tag{10}$$

We also have

$$\begin{aligned} \frac{V_{LN}(T)}{n} &\approx AV_{LN}(\sigma) = V_{LN}\left[\ln(f_{LN}(X;\sigma^2,1)) - \ln(f_{GA}(X;\tilde{\alpha},\tilde{\lambda}))\right] \\ &= V_{LN}\left[-\tilde{\alpha}\ln X + \frac{X}{\tilde{\lambda}} - \frac{1}{2\sigma^2}(\ln X)^2\right] \\ &= \tilde{\alpha}^2\sigma^2 + \frac{1}{\tilde{\lambda}^2}e^{\sigma^2}(e^{\sigma^2}-1) + \frac{1}{2} - 2\frac{\tilde{\alpha}}{\tilde{\lambda}}cov_{LN}(\ln X, X) \\ &\quad -\frac{1}{\tilde{\lambda}\sigma^2}cov_{LN}((\ln X)^2, X). \end{aligned} \tag{11}$$

Now we consider the other case.

Case 2: The data are from a gamma distribution.

Let us assume that a sample $\{X_1, \ldots, X_n\}$ of size n is obtained from GA(α, λ). In this case we have the following lemma.

Lemma 2: Under the assumption that the data are from a gamma distribution and as $n \to \infty$, we have

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

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

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

$$E_{GA}\left[ln(f_{LN}(X; \tilde{\sigma}, \tilde{\theta}))\right] = \max_{\sigma, \theta} E_{GA}\left[ln(f_{LN}(X; \sigma^2, \theta))\right].$$

Note that here also $\tilde{\sigma}$ and $\tilde{\theta}$ may depend on α and λ but we do not make it explicit for brevity. Let us denote $T_* = ln\left(\frac{L_{LN}(\tilde{\sigma}, \tilde{\theta})}{L_{GA}(\alpha, \lambda)}\right)$.

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

Theorem 2: Under the assumption that the data are from a gamma distribution, T is approximately normally distributed with mean $E_{GA}(T)$ and variance $V_{GA}(T) = V_{GA}(T_*)$.

Now to obtain $\tilde{\sigma}$ and $\tilde{\theta}$, let us define

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

Therefore, $\tilde{\sigma}$and $\tilde{\theta}$ can be obtained as

$$\tilde{\theta} = \lambda e^{\psi(\alpha)} \qquad \text{and} \qquad \tilde{\sigma} = (\psi'(\alpha))^{\frac{1}{2}}. \tag{12}$$

Here $\psi'(\alpha) = \frac{d}{d\alpha}\psi(\alpha)$. Now we provide the expressions for $E_{GA}(T)$ and $V_{GA}(T)$. Similarly as before, we observe that $\lim_{n\to\infty} \frac{E_{GA}(T)}{n}$ and $\lim_{n\to\infty} \frac{V_{GA}(T)}{n}$ exist. We denote $\lim_{n\to\infty} \frac{E_{GA}(T)}{n} = AM_{GA}(\alpha)$ and $\lim_{n\to\infty} \frac{V_{GA}(T)}{n} = AV_{GA}(\alpha)$ respectively, then for large n,

$$\begin{aligned}
\frac{E_{GA}(T)}{n} &\approx AM_{GA}(\alpha) = E_{GA}\left[\ln(f_{LN}(X;\tilde{\sigma},\tilde{\theta})) - \ln(f_{GA}(X;\alpha,1))\right] \\
&= -\frac{1}{2}\ln(2\pi) - \ln\tilde{\sigma} - \frac{1}{2\tilde{\sigma}^2}\left[\psi'(\alpha) + (\psi(\alpha) - \ln\tilde{\theta})^2\right] \\
&\quad + \ln(\Gamma(\alpha)) + \alpha(1 - \psi(\alpha))
\end{aligned}$$

$$\begin{aligned}
\frac{V_{GA}(T)}{n} &\approx AV_{GA}(\alpha) = V_{GA}\left[\ln(f_{LN}(X;\tilde{\sigma},\tilde{\theta})) - \ln(f_{GA}(X;\alpha,1))\right] \\
&= V_{GA}\left[X - \alpha\ln X - \frac{1}{2\tilde{\sigma}^2}(\ln X - \ln\tilde{\theta})^2\right] \\
&= \alpha + \alpha^2\psi'(\alpha) - 2\alpha(\psi(\alpha+1) - \psi(\alpha)) \\
&\quad -\frac{1}{\tilde{\sigma}^2}\left[\alpha(\alpha+1)(\psi'(\alpha+2) + \psi(\alpha+2))^2 - (\ln\tilde{\theta})\alpha\psi(\alpha+1)\right. \\
&\quad +\alpha(\ln\tilde{\theta})^2 - \alpha(\psi'(\alpha) + \psi(\alpha))^2 - 2(\ln\tilde{\theta})\psi(\alpha) - 2\alpha\psi(\alpha)\psi'(\alpha) \\
&\quad -\tilde{\alpha}\psi''(\alpha) + 2(\ln\theta)\psi'(\alpha)] \\
&\quad +\frac{1}{4\tilde{\sigma}^4}\left[\psi'''(\alpha) + 4\psi(\alpha)\psi''(\alpha) + 4\psi'(\alpha)(\psi(\alpha))^2 + 2(\psi'(\alpha))^2\right. \\
&\quad \left.-4(\ln\tilde{\theta})\psi''(\alpha) - 8\psi(\alpha)\psi'(\alpha)\ln\tilde{\theta} + 4\psi'(\alpha)(\ln\tilde{\theta})^2\right].
\end{aligned}$$

Note that $\tilde{\alpha}$, $\tilde{\lambda}$, $AM_{LN}(\sigma)$, $AV_{LN}(\sigma)$, $\tilde{\sigma}$, $\tilde{\theta}$, $AM_{GA}(\alpha)$ and $AV_{GA}(\alpha)$ are quite difficult to compute numerically. We present $\tilde{\alpha}$, $\tilde{\lambda}$ and also $AM_{LN}(\sigma)$ and $AV_{LN}(\sigma)$ for different values of σ in Table 1. We also present $\tilde{\sigma}$, $\tilde{\theta}$ and also $AM_{GA}(\alpha)$ and $AV_{GA}(\alpha)$ for different values of α in Table 2 for convenience.

4 Determination Of Sample Size:

We are proposing a method to determine the minimum sample size required to discriminate between the log-normal and gamma distributions, for a given user specified PCS. Before discriminating between two fitted distribution functions it is important to know how close they are. There are several ways to measure the closeness or the distance between two distribution functions, for example, the Kolmogorov-Smirnov (K-S) distance or Hellinger distance etc.. It is very natural that if two distributions are very close then a very large sample size is needed to discriminate between them for a given PCS. On the other hand if the distance between two distribution functions is quite far, then one may not need very large sample size to discriminate between them. It is 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 using the Hellinger distance also, which is not pursued here.

We observed in section 3 that the RML statistics follow normal distribution approximately 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 $E_{LN}(T)$ and variance $V_{LN}(T)$, therefore the probability of correct selection (PCS) is

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

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

$$\Phi\left(\frac{-n \times AM_{LN}(\sigma)}{\sqrt{n \times AV_{LN}(\sigma)}}\right) = 1 - p^* \tag{14}$$

and solve for n. It provides

$$n = \frac{z_{p^*}^2 AV_{LN}(\sigma)}{(AM_{LN}(\sigma))^2}. \tag{15}$$

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

$$n = \frac{z_{p^*}^2 AV_{GA}(\alpha)}{(AM_{GA}(\alpha))^2}. \tag{16}$$

Here $AM_{GA}(\alpha)$ and $AV_{GA}(\alpha)$ are same as defined before. We report n, with the help of Table 2 for different values of α when $p^* = 0.9$ in Table 4. From Table 3, it is clear that as σ increases the required sample size decreases for a given PCS. Interestingly, from Table 4, it is immediate that as α increases the required sample size increases. Both the findings are quite intuitive in the sense one needs large sample sizes to discriminate between them if the two distribution functions are very close. It is clear that 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 in both the 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. The distance between two distribution functions is defined by the K-S distance. The K-S distance between two distribution functions, say $F(x)$ and $G(x)$ is defined as

$$\sup_x |F(x) - G(x)|. \tag{17}$$

We report K-S distance between $LN(\sigma, 1)$ and $GA(\tilde{\alpha}, \tilde{\lambda})$ for different values of σ in Table 3. Here $\tilde{\alpha}$ and $\tilde{\lambda}$ are same as defined in Lemma 1 and they have been reported in Table 1. Similarly, K-S distance between $GE(\alpha, 1)$ and $LN(\tilde{\sigma}, \tilde{\theta})$ for different values of α is reported in Table 4. Here $\tilde{\sigma}$ and $\tilde{\theta}$ are same as defined in Lemma 2 and they have been reported in Table 2. Now we explain how we can determine the minimum sample size required to discriminate between log-normal and gamma 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 a log-normal and gamma 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 $\sigma \geq 0.7$. Similarly from Table 4, it is clear that the K-S distance will be more than 0.05 if $\alpha \leq 2.0$. Therefore, if the data come from the log-normal distribution, then for the tolerance level $D^* = 0.05$, one needs at most $n = 96$ to meet the PCS, $p^* = 0.9$. Similarly if the data come from the gamma distribution then one needs at most $n = 95$ 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(95, 96) = 96 to meet the protection level $p^* = 0.9$ simultaneously for both the cases.

Table 1

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

σ	$AM_{LN}(\sigma)$	$AV_{LN}(\sigma)$	$\tilde{\alpha}$	$\tilde{\lambda}$
0.5	0.0207	0.0143	4.1594	0.2724
0.7	0.0389	0.0885	2.1930	0.5826
0.9	0.0608	0.1612	1.3774	1.0885
1.1	0.0861	0.2660	0.9588	1.8313
1.3	0.1131	0.4016	0.7133	3.2637
1.5	0.1409	0.5692	0.5556	5.5439

5 Numerical Experiments

In this section we perform some numerical experiments to observe how these asymptotic results derived in section 3 work for finite 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.* [16] and all the programs are written in C. They can be obtained from the authors on request. We compute the probability of correct selections based on simulations and we also compute it based on asymptotic results derived in section 3. We consider different sample sizes and also different shape parameters, the details are explained below.

First we consider the case when the data are coming from a log-normal distribution. In this case we consider n = 20, 40, 60, 80, 100 and σ = 0.5, 0.7, 0.9, 1.1, 1.3 and 1.5. For a fixed σ and n we generate a random sample of size n from LN(σ, 1), we finally compute T as defined in (4) and check whether T is positive or negative. We replicate the process 10,000 times and obtain an estimate of the PCS. We also compute the PCSs by using these asymptotic results as given in (13). The results are reported in Table 5. Similarly, we obtain the results when the data are generated from a gamma distribution. In this case we consider the same set of n and α = 2.0, 4.0, 6.0, 8.0, 10.0 and 12.0. The results are reported in Table 6. In each box the first row represents the results obtained by using Monte Carlo simulations and the second row represents the results obtained by using the asymptotic theory.

As sample size increases the PCS increases in both the cases. It is also clear that when the shape parameter increases for the log-normal distribution the PCS increases and when the shape parameter decreases for the gamma distribution the PCS increases. Even when the sample size is 20, asymptotic results work quite well for both the cases for all possible parameter ranges. From the simulation study it is recommended that asymptotic results can be used quite effectively even when the sample size is as small as 20 for all possible choices of the shape parameters.

Table 2
Different values of $AM_{GA}(\alpha)$, $AV_{GA}(\alpha)$, $\tilde{\sigma}$ and $\tilde{\theta}$ for different α.

α	$AM_{GA}(\alpha)$	$AV_{GA}(\alpha)$	$\tilde{\sigma}$	$\tilde{\theta}$
2.0	-0.0395	0.0904	0.8031	1.5262
4.0	-0.0207	0.0457	0.5328	3.5118
6.0	-0.0142	0.0305	0.4258	5.5075
8.0	-0.0109	0.0221	0.3649	7.5055
10.0	-0.0088	0.0180	0.3243	9.5044
12.0	-0.0074	0.0149	0.2948	11.5036

6 Data Analysis

In this section we analyze one data set and use our method to discriminate between two populations.

Data Set 1: The data set is from Lawless [14, 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 test 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 a log-normal distribution, the MLEs of the different parameters are $\hat{\sigma}$

= 0.5313, $\hat{\theta}$ = 0.63.5147 and $Ln(L_{LN}(\hat{\sigma}, \hat{\theta}))$ = -112.8552. The K-S distance between the fitted empirical distribution function and the fitted log-normal distribution function is 0.09. Similarly, if we use a gamma distribution, the MLEs of the different parameters are $\hat{\alpha}$ = 4.0196, $\hat{\lambda}$ = 17.9856 and $Ln(L_{GA}(\hat{\alpha}, \hat{\lambda}))$ = -113.0274. In this case, the K-S distance between the fitted empirical distribution function and the fitted gamma distribution function is 0.12. The K-S distance between the two fitted distributions is 0.034. They are quite close to each other. In terms of the K-S distance, log-normal distribution is closer to the empirical distribution function than a gamma distribution. Interestingly, $T = -112.8552 + 113.0274 = 0.1722 > 0$, also suggests to choose the log-normal distribution rather than the gamma distribution.

Table 3

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

$\sigma \rightarrow$	0.5	0.7	0.9	1.1	1.3	1.5
$n \rightarrow$	159	96	72	59	52	47
K-S	0.033	0.049	0.064	0.076	0.097	0.113

Table 4

The minimum sample size $n = \frac{z^2_{0.90} AV_{GA}(\alpha)}{(AM_{GA}(\alpha))^2}$, for $p^* = 0.9$ and when the data come from a gamma distribution is presented. The K-S distance between GA (α,1) and LN ($\tilde{\sigma}, \tilde{\theta}$) for different values of α is reported.

$\alpha \rightarrow$	2.0	4.0	6.0	8.0	10.0	12.0
$n \rightarrow$	95	175	249	306	382	447
K-S	0.049	0.034	0.025	0.023	0.021	0.013

Assuming that the original distribution was log-normal with σ = 0.5215 = $\hat{\sigma}$ and θ = 63.4784 = $\hat{\theta}$, we compute PCS by computer simulations (based on 10,000 replications) similarly as in section 5 and we obtain PCS = 0.6985. It implies that PCS $\approx$ 70%. On the other hand if the choice of log-normal distribution was wrong and the original distribution was gamma with shape parameter $\alpha = 4.0196 = \hat{\alpha}$ and scale parameter $\lambda = 17.9856 = \hat{\lambda}$, then similarly as before based on 10,000 replications we obtain PCS = 0.6788, yielding an estimated risk approximately 32% to choose the wrong model. Now we compute the PCSs based on large sample approximations. Assuming that the data are coming from the $LN(0.5313, 63.5147)$, we obtain $AM_{LN}(0.5215) = 0.0276$ and $AV_{LN}(0.5215) = 0.0578$, it implies $E_{LN}(T) \approx 0.6348$ and $V_{LN}(T) \approx = 1.3294$. Therefore, assuming that the data are from $LN(0.5313, 63.5147)$, T is approximately normally distributed with mean = 0.6348, variance = 1.3294 and PCS = 1 - $\Phi(-0.5505) = \Phi(0.5505) \approx 0.71$, which is almost equal to the above simulation result. Similarly, assuming that the data are coming from a gamma distribution, we compute $AM_{GA}(4.0196)$ = -0.0198 and $AV_{GA}(4.0196) = 0.0424$. and we have $E_{LN}(T) \approx$ -0.4554 and $V_{LN}(T) \approx = 0.9752$. Therefore, assuming that the data are from a gamma distribution the PCS = $\Phi(0.4612) \approx 0.68$, which is also very close to the simulated results. Therefore, based on K-S distances and also on the RML statistics, we would like to conclude that it is more likely that the data are coming from a log-normal distribution and the probability correct selection is $\approx$ 70 %.

Table 5

The probability of correct selection based on Monte Carlo Simulations and also based on asymptotic results when the data are coming from log-normal. 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 immediately below represents the result obtained by using asymptotic results.

$\sigma \downarrow$ n $\rightarrow$	20	40	60	80	100
0.5	0.66 (0.68)	0.73 (0.74)	0.78 (0.79)	0.82 (0.82)	0.85 (0.85)
0.7	0.70 (0.72)	0.79 (0.80)	0.85 (0.84)	0.88 (0.88)	0.91 (0.91)
0.9	0.73 (0.75)	0.84 (0.83)	0.89 (0.88)	0.93 (0.92)	0.95 (0.94)
1.1	0.77 (0.76)	0.88 (0.86)	0.92 (0.91)	0.93 (0.93)	0.95 (0.95)
1.3	0.78 (0.79)	0.88 (0.87)	0.92 (0.92)	0.95 (0.95)	0.96 (0.96)
1.5	0.81 (0.80)	0.90 (0.89)	0.94 (0.93)	0.96 (0.96)	0.97 (0.97)

7 Conclusions

In this paper we consider the problem of discriminating the two families of distribution functions, namely the log-normal and gamma families. We consider the statistic based on the ratio of the maximized likelihoods and obtain the asymptotic distributions of the test statistics under null hypotheses. We compare the probability of correct selection using Monte Carlo simulations with the asymptotic results and it is observed that even when the sample size is very small the asymptotic results work quite well for a wide range of the parameter space. Therefore, the asymptotic results can be used to estimate the probability of correct selection. We use these asymptotic results to calculate the minimum sample size required for a user specified probability of correct selection. We use the concept of tolerance level based on the distance between the two distribution functions. For a particular D^* tolerance level the minimum sample size is obtained for a given user specified protection level. Two small tables are provided for the protection level 0.90 but for the other protection level the tables can be easily used as follows. For example if we need the protection level $p^* = 0.8$, then all the entries corresponding to the row of n, will be multiplied by $\frac{z_{0.8}^2}{z_{0.9}^2}$, because of (15) and (16). Therefore, Tables 3 and 4 can be used for any given protection level.

Table 6

The probability of correct selection based on Monte Carlo Simulations and also based on asymptotic results when the data are coming from a gamma distribution. 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 immediately below represents the result obtained by using asymptotic results.

$\alpha \downarrow$ n →	20	40	60	80	100
2.0	0.71 (0.72)	0.80 (0.80)	0.86 (0.85)	0.88 (0.88)	0.91 (0.91)
4.0	0.65 (0.67)	0.74 (0.73)	0.78 (0.77)	0.81 (0.81)	0.83 (0.83)
6.0	0.63 (0.64)	0.71 (0.70)	0.74 (0.74)	0.77 (0.77)	0.79 (0.79)
8.0	0.61 (0.63)	0.70 (0.69)	0.72 (0.72)	0.75 (0.75)	0.77 (0.77)
10.0	0.60 (0.61)	0.67 (0.66)	0.70 (0.70)	0.73 (0.72)	0.75 (0.75)
12.0	0.59 (0.61)	0.66 (0.65)	0.69 (0.68)	0.71 (0.71)	0.73 (0.73)

References

[1] Atkinson, A. (1969), "A test for discriminating between models", *Biometrika*, 56, 337-347.

[2] Atkinson, A. (1970), "A method for discriminating between models" (with discussions), *Jour. Royal Stat. Soc. Ser. B*, 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", *Jour. Stat. Comp. Simul.*, 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", *Jour. of Royal Statistical Society*, Ser. B, 24, 406-424.

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

[9] Dyer, A.R. (1973), “Discrimination procedure for separate families of hypotheses”, *Jour. Amer. Stat. Asso.*, 68, 970-974.

[10] Fearn, D.H. and Nebenzahl, E. (1991), “On the maximum likelihood ratio method of deciding between the Weibull and gamma distributions”, *Communications in Statistics, Ser. A*, 20, 2, 579=593.

[11] Firth, D. (1988), “Multiplicative errors: log-normal of gamma?”, *Journal of the Royal Statistical Society, Ser. B*, 2, 266-268.

[12] Jackson, O.A.Y. (1968), “Some results on tests separate families of hypotheses”, *Biometrika*, 55, 355-363.

[13] Johnson, N., Kotz, S. and Balakrishnan, N (1995), *Continuous Univariate Distributions*, 2nd Edition, Wiley, New York.

[14] Lawless, (1982), *Statistical Models and Methods for Lifetime Data*, New York, Wiley.

[15] Pereira, B. de B. (1977), “A note on the consistency and on the finite sample comparisons of some tests of separate families of hypotheses”, *Biometrika*, 64, 109-113.

[16] Press et al. (1993) *Numerical Recipes in FORTRAN*, Cambridge University Press, Cambridge.

[17] White, H. (1982), “Regularity conditions for Cox’s test of non-nested hypotheses”, *Journal of Econometrics*, vol. 19, 301-318.

[18] Wiens, B.L. (1999), “When log-normal and gamma models give different results: a case study”, *The American Statistician*, 53, 2, 89-93.

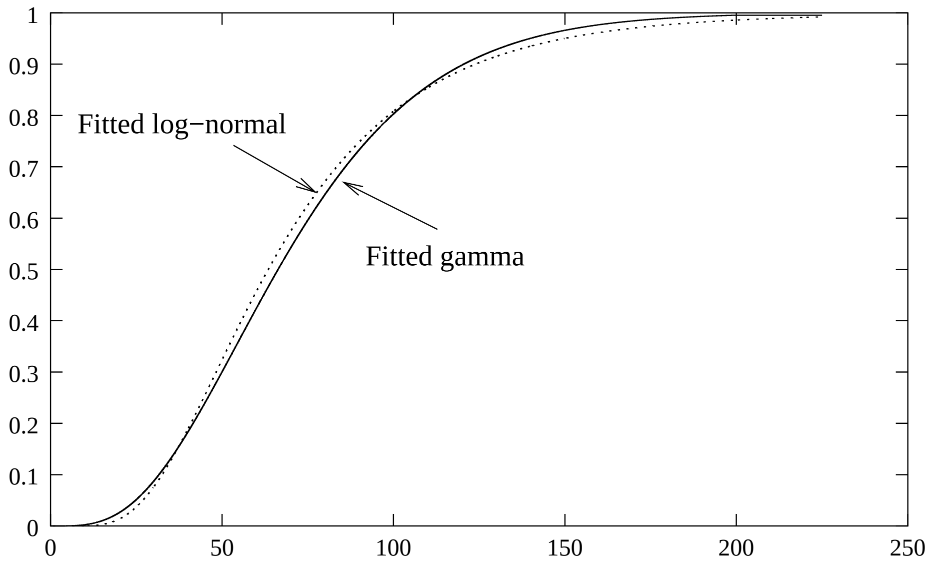

Figure 1: The two fitted distribution functions for the given data set

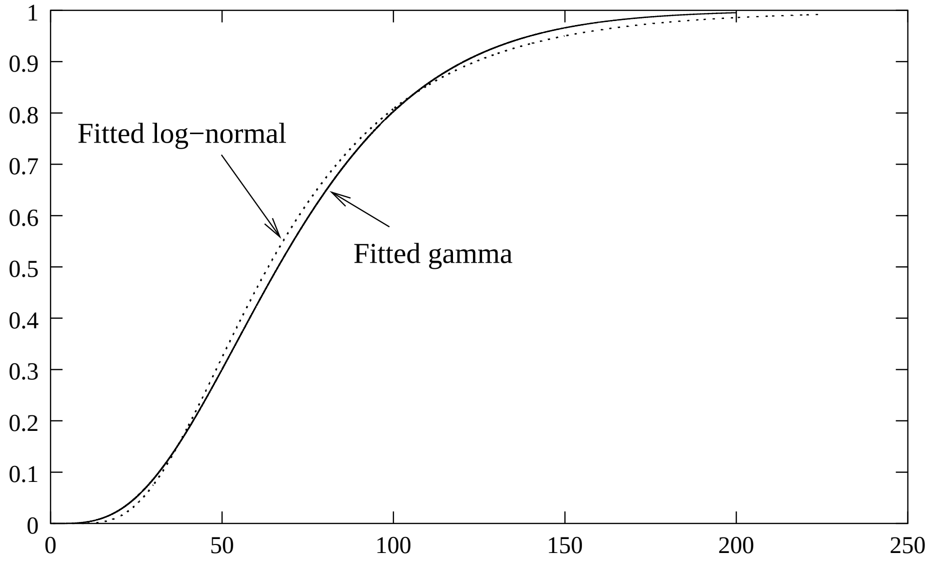

Figure 1: The two fitted distribution functions for the given data set