ARTICLE IN PRESS

Available at
www.ElsevierMathematics.com
POWERED BY SCIENCE DIRECT®
Journal of Statistical Planning and Inference ▮▮▮ (▮▮▮▮) ▮▮▮–▮▮▮

ELSEVIER

journal of statistical planning and inference

www.elsevier.com/locate/jspi

Discriminating between the log-normal and generalized exponential distributions

Debasis Kundu[a,*], Rameshwar D. Gupta[b], Anubhav Manglick[a]

[a] *Department of Mathematics, Indian Institute of Technology Kanpur 208016, India*
[b] *Department of Computer Science and Applied Statistics, University of New Brunswick, Saint Jonh, Canada E2L 4L5*

Received 5 November 2002; accepted 10 August 2003

Abstract

The two-parameter generalized exponential distribution was recently introduced by Gupta and Kundu (Austral. New Zealand J. Statist. 40 (1999) 173). It is observed that the Generalized Exponential distribution can be used quite effectively to analyze skewed data set as an alternative to the more popular log-normal distribution. In this paper, we use the ratio of the maximized likelihoods in choosing between the log-normal and generalized exponential distributions. We obtain asymptotic distributions of the logarithm of the ratio of the maximized likelihoods and use them to determine the required sample size to discriminate between the two distributions for a user specified probability of correct selection and tolerance limit.

Keywords: Asymptotic distributions; Generalized exponential distribution; Kolmogorov–Smirnov distances; Likelihood ratio test statistic

1. Introduction

Recently Gupta and Kundu (1999) introduced the Generalized Exponential (GE) distribution and studied quite extensively several properties of the GE distribution, see for example (Gupta and Kundu, 1999, 2001a,b, 2002). The readers may be referred to some of the related literature on GE distribution by Raqab (2002), Raqab and Ahsanullah (2001) and Zheng (2002). The two-parameter GE family has the

* Corresponding author. Tel.: +91-512-2597141; fax: +91-512-2597500.
E-mail address: kundu@iitk.ac.in (D. Kundu).

0378-3758/$ - see front matter © 2003 Published by Elsevier B.V.
doi:10.1016/j.jspi.2003.08.017

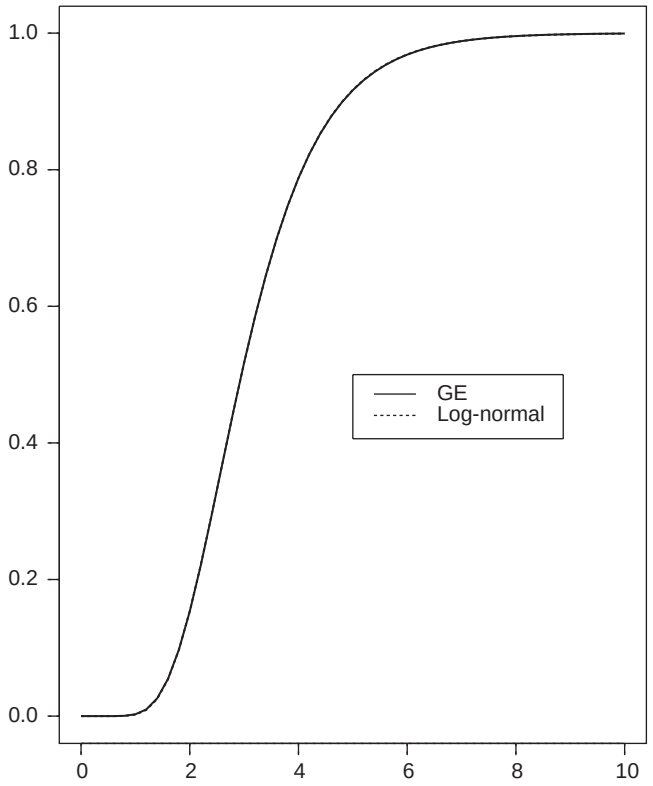

Fig. 1. The distribution functions of GE (12.9, 1) and LN (0.3807482, 2.9508672).

distribution function

$$F_{\mathrm{GE}}(x;\alpha,\lambda) = (1-\mathrm{e}^{-\lambda x})^{\alpha}, \quad x > 0. \tag{1}$$

The corresponding density function is

$$f_{\mathrm{GE}}(x;\alpha,\lambda) = \alpha\lambda(1-\mathrm{e}^{-\lambda x})^{\alpha-1}\mathrm{e}^{-\lambda x}, \quad x > 0. \tag{2}$$

Here $\alpha > 0$ and $\lambda > 0$ are the shape and scale parameters, respectively. When $\alpha = 1$, it coincides with the exponential distribution with mean $1/\lambda$. When $\alpha \leqslant 1$, the density function is strictly decreasing and for $\alpha > 1$ it has a unimodal shape. These densities are illustrated in Gupta and Kundu (2001a). It is clear that the GE density functions are always right skewed and it is observed that GE distributions can be used quite effectively to analyze skewed data sets. Among several other distributions, the two-parameter log-normal distribution is also used quite effectively to analyze skewed data sets. Log-normal density function is always unimodal in nature. Shapes of the different log-normal density functions can be found in Johnson et al. (1995). It is clear that the shapes of these two density functions are quite similar at least for certain ranges of the parameters. See for example Fig. 1, where the two distribution

functions are almost identical. Although these two models may provide similar data fit for moderate sample sizes, it is still desirable to select the correct or more nearly correct model, since inferences based on the model will often involve tail probabilities, where the effect of the model assumption will be more crucial.

GE has an exponential tail while log-normal has heavier tail than exponential. Therefore, even if large sample sizes are not available it is still very important to make a best possible decision based on whatever data are available.

The problem of testing whether some given observations follow one of the two probability distribution functions is quite old in the statistical literature. Atkinson (1969, 1970), Chen (1980), Chambers and Cox (1967), Cox (1961,1962), Jackson (1968) and Dyer (1973) considered this problem in general for discriminating between two models. Between the models, the effect of choosing a wrong model was originally discussed by Cox (1961) in general and recently Wiens (1999) demonstrated it nicely by a real data example. Due to increasing applications of the lifetime distributions, special attention is given to the discrimination between the log-normal and Weibull distributions (Dumonceaux and Antle, 1973; Pereira, 1978; Chen, 1980; Quesenberry and Kent, 1982), the log-normal and gamma distributions (Jackson, 1969; Quesenberry and Kent, 1982; Wiens, 1999), the gamma and Weibull distributions by Bain and Englehardt (1980) and Fearn and Nebenzahl (1991), the Weibull and generalized exponential distributions by Gupta and Kundu (2003a) and the gamma and generalized exponential distributions by Gupta and Kundu (2003b).

In this paper, we consider the problem of discriminating between the log-normal and GE distributions. We use the ratio of the maximized likelihood (RML) in discriminating between the two distribution functions. We obtain the asymptotic distributions of the logarithm of RML and under each model observe by extensive Monte Carlo simulations that the asymptotic distributions work quite well in discriminating between the two distribution functions even when the sample size is not too large. Using these asymptotic distributions and the distance between the two distribution functions, we determine the minimum sample size needed to discriminate between the two models at a user specified protection level. It is observed experimentally that the distance between the two distribution functions can be quite small for certain ranges of the parameter values.

The rest of the paper is organized as follows. The ratio of the maximized likelihoods is described in Section 2. Asymptotic distributions of the logarithm of RMLs are developed in Section 3. In Section 4, the asymptotic distributions are used to compute the minimum sample size required to discriminate two distribution functions at a user specified probability of correct selection and a tolerance level. Numerical results are presented in Section 5 and finally we conclude the paper in Section 6.

2. Ratio of the maximized likelihoods

Suppose $X_1,\ldots,X_n$ are independent and identically distributed (i.i.d.) random variables from any one of the two distribution functions. The density function of a GE random variable with shape parameter α and scale parameter λ is given in (2). The

density function of a log-normal random variable with scale parameter $\theta > 0$ and shape parameter $\sigma > 0$ is denoted by

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

A GE distribution with shape parameter α and scale parameter λ will be denoted by $\mathrm{GE}(\alpha,\lambda)$. Similarly, a log-normal distribution with shape parameter σ and scale parameter θ will be denoted by $\mathrm{LN}(\sigma,\theta)$. The likelihood functions assuming that the data are coming from $\mathrm{GE}(\alpha,\lambda)$ or $\mathrm{LN}(\sigma,\theta)$ are

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

respectively. The RML is defined as

$$L = \frac{L_{\mathrm{GE}}(\hat{\alpha},\hat{\lambda})}{L_{\mathrm{LN}}(\hat{\sigma},\hat{\theta})}. \tag{4}$$

Here $(\hat{\alpha},\hat{\lambda})$ and $(\hat{\sigma},\hat{\theta})$ are maximum likelihood estimators of (α,λ) and (σ,θ), respectively. The logarithm of RML can be written as

$$T = n\left[\ln(\hat{\alpha}\hat{\lambda}\tilde{X}\hat{\sigma}) - \frac{\hat{\alpha}-1}{\hat{\alpha}} - \hat{\lambda}\bar{X} + \frac{1}{2}(1+\ln(2\pi))\right], \tag{5}$$

where $\bar{X} = 1/n\sum_{i=1}^{n} X_i$ and $\tilde{X} = \left(\prod_{i=1}^{n} X_i\right)^{1/n}$. Moreover, $\hat{\alpha}$ and $\hat{\lambda}$ have the following relation (Gupta and Kundu, 2001a):

$$\hat{\alpha} = -\frac{n}{\sum_{i=1}^{n}\ln(1-\mathrm{e}^{-\hat{\lambda}X_i})}. \tag{6}$$

In case of the log-normal distribution, $\hat{\theta}$ and $\hat{\sigma}$ have the following form:

$$\hat{\theta} = \tilde{X} \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left[\ln\left(\frac{X_i}{\hat{\theta}}\right)\right]^2. \tag{7}$$

Now we propose the following discrimination procedure. Choose the GE distribution if $T > 0$, otherwise choose the log-normal distribution as the preferred model.

Now consider the case when the data come from the $\mathrm{GE}(\alpha,\lambda)$ distribution. In this case the distribution of λX_i is clearly independent of λ and from Bain and Englehardt (1991) it easily follows that the distribution $\hat{\lambda}/\lambda$ is independent of λ. From the expression of $\hat{\sigma}^2$, it is immediate that $\hat{\sigma}^2$ is independent of λ. It shows that the distribution of T is independent of λ and depends only on α. Similarly it can be shown that when the data come from $\mathrm{LN}(\sigma,\theta)$ then the distribution of T depends only on σ and is independent of θ.

3. Asymptotic properties of the logarithm of RML

In this section, we obtain the asymptotic distributions of the logarithm 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* $\mathrm{GE}(\alpha,\lambda)$. We assume that n data points are from $\mathrm{GE}(\alpha,\lambda)$ and $\hat{\alpha}$, $\hat{\lambda}$, $\hat{\theta}$ and $\hat{\sigma}$ are as defined before. We use the following notation. For any Borel measurable function $h(.)$, $E_{\mathrm{GE}}(h(U))$ and $V_{\mathrm{GE}}(h(U))$ denote the mean and variance of $h(U)$ under the assumption that U follows $\mathrm{GE}(\alpha,\lambda)$. Similarly, we define $E_{\mathrm{LN}}(h(U))$ and $V_{\mathrm{LN}}(h(U))$ as the mean and variance of $h(U)$ under the assumption that U follows $\mathrm{LN}(\sigma,\theta)$. If $g(.)$ and $h(.)$ are two Borel measurable functions, we define along the same line that $\mathrm{cov}_{\mathrm{GE}}(g(U),h(U))=E_{\mathrm{GE}}(g(U)h(U))-E_{\mathrm{GE}}(g(U))E_{\mathrm{GE}}(h(U))$ and $\mathrm{cov}_{\mathrm{LN}}(g(U),h(U))=E_{\mathrm{LN}}(g(U)h(U))-E_{\mathrm{LN}}(g(U))E_{\mathrm{LN}}(h(U))$, where U follows $\mathrm{GE}(\alpha,\lambda)$ and $\mathrm{LN}(\sigma,\theta)$, respectively. The following lemma is needed to prove the main result.

Lemma 1. *Under the assumption that the data are from* $\mathrm{GE}(\alpha,\lambda)$ *we have as* $n\to\infty$,

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

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

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

$$E_{\mathrm{GE}}[ln(f_{\mathrm{LN}}(X;\tilde{\sigma},\tilde{\theta}))]=\max_{\sigma,\theta}E_{\mathrm{GE}}[ln(f_{\mathrm{LN}}(X;\sigma,\theta))].$$

It may be noted that $\tilde{\theta}$ *and* $\tilde{\sigma}$ *may depend on* α *and* λ *but we do not make it explicit for brevity. Let us denote*

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

(iii) $n^{-1/2}[T-E_{\mathrm{GE}}(T)]$ *is asymptotically equivalent to* $n^{-1/2}[T^*-E_{\mathrm{GE}}(T^*)]$

Proof. The proof follows a similar argument as in White (1982, Theorem 1) and is therefore omitted.

Now we can state the main result:

Theorem 1. *Under the assumption that the data are from* $\mathrm{GE}(\alpha,\lambda)$, T *is asymptotically normally distributed with mean* $E_{\mathrm{GE}}(T)$ *and variance* $V_{\mathrm{GE}}(T^*)=V_{\mathrm{GE}}(T)$.

Proof. Using the central limit theorem and part (ii) of Lemma 1, it follows that $n^{-1/2}[T^*-E_{\mathrm{GE}}(T^*)]$ is asymptotically normally distributed with mean zero and variance $V_{\mathrm{GE}}(T^*)$. Therefore using part (iii) of Lemma 1, the result immediately follows.

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

$$\begin{aligned} g(\sigma,\theta) &= E_{\mathrm{GE}}[\ln(f_{\mathrm{LN}}(X;\sigma,\theta))] \\ &= -\frac{1}{2}\ln 2\pi - \ln\sigma - E(\ln(Z)) + \ln\lambda - \frac{1}{2\sigma^2}E(\ln(Z))^2 \\ &\quad - \frac{1}{2\sigma^2}(\ln\lambda)^2 - \frac{1}{2\sigma^2}(\ln\theta)^2 + \frac{\ln\theta E(\ln Z)}{\sigma^2} + \frac{\ln\lambda E(\ln Z)}{\sigma^2} \\ &\quad - \frac{\ln\theta\ln\lambda}{\sigma^2}, \end{aligned} \tag{8}$$

where Z follows $\mathrm{GE}(\alpha,1)$. Therefore, $\tilde{\theta}$ and $\tilde{\sigma}$ can be obtained as

$$\tilde{\theta} = \frac{1}{\lambda}\mathrm{e}^{E(\ln Z)}, \tag{9}$$

$$\begin{aligned} \tilde{\sigma}^2 &= E(\ln Z)^2 + (\ln(\lambda\tilde{\theta}))^2 - 2E(\ln Z)\ln(\lambda\tilde{\theta}) \\ &= E(\ln(Z))^2 - (E(\ln Z))^2. \end{aligned} \tag{10}$$

From (9) and (10) it is clear that $\lambda\tilde{\theta}$ and $\tilde{\sigma}$ are functions of α only. Note that $E(\ln Z)^2$ and $(E(\ln Z))^2$ can be easily obtained using the results of Gupta and Kundu (1999). Now we provide the expressions for $E_{\mathrm{GE}}(T)$ and $V_{\mathrm{GE}}(T)$. Observe that $\lim_{n\to\infty} E_{\mathrm{GE}}(T)/n$ and $\lim_{n\to\infty} V_{\mathrm{GE}}(T)/n$ exist and we denote them as $\mathrm{AM}_{\mathrm{GE}}(\alpha)$ and $\mathrm{AV}_{\mathrm{GE}}(\alpha)$, respectively. Therefore, for large n

$$\begin{aligned} \frac{E_{\mathrm{GE}}(T)}{n} &\approx AM_{\mathrm{GE}}(\alpha) = E_{\mathrm{GE}}[\ln(f_{\mathrm{GE}}(\alpha,\lambda)) - \ln(f_{\mathrm{LN}}(\tilde{\sigma},\tilde{\theta}))] \\ &= \frac{1}{2}\ln 2\pi + E_{\mathrm{GE}}(\ln Z)\left(1 - \frac{\ln\tilde{\theta}}{\tilde{\sigma}^2}\right) + \ln\tilde{\sigma} + \frac{(\ln\tilde{\theta})^2}{2\tilde{\sigma}^2} + \frac{(E_{\mathrm{GE}}(\ln Z))^2}{2\tilde{\sigma}^2} \\ &= \frac{1}{2}\ln 2\pi + E_{\mathrm{GE}}(\ln Z) + \ln\tilde{\sigma} + \frac{1}{2}. \end{aligned} \tag{11}$$

Also,

$$\begin{aligned} \frac{V_{\mathrm{GE}}(T)}{n} &\approx AV_{\mathrm{GE}}(\alpha) = V_{\mathrm{GE}}[\ln(f_{\mathrm{GE}}(\alpha,\lambda)) - \ln(f_{\mathrm{LN}}(\tilde{\sigma},\tilde{\theta}))] \\ &= V_{\mathrm{GE}}\left[(\alpha-1)\ln(1-\mathrm{e}^{-Z}) - Z + \ln Z + \frac{1}{2\tilde{\sigma}^2}(\ln Z)^2 - \frac{1}{\tilde{\sigma}^2}\ln Z\,\ln(\lambda\tilde{\theta})\right] \\ &= \frac{(\alpha-1)^2}{\alpha^2} + (\psi'(1) - \psi'(\alpha+1)) + \left(1 - \frac{\ln\lambda\tilde{\theta}}{\tilde{\sigma}^2}\right)^2 V_{\mathrm{GE}}(\ln Z) + \frac{1}{4\tilde{\sigma}^4}V_{\mathrm{GE}}(\ln Z)^2 \\ &\quad - 2(\alpha-1)\,\mathrm{cov}_{\mathrm{GE}}(\ln(1-\mathrm{e}^{-Z}), Z) + 2(\alpha-1)\left(1 - \frac{\ln\lambda\tilde{\theta}}{\tilde{\sigma}^2}\right)\mathrm{cov}_{\mathrm{GE}}(\ln Z, \ln(1-\mathrm{e}^{Z})) \end{aligned}$$

$$+\frac{(\alpha-1)}{\tilde{\sigma}^2}\,\mathrm{cov}_{\mathrm{GE}}((\ln Z)^2,\ln(1-\mathrm{e}^{Z}))-2\left(1-\frac{\ln\lambda\tilde{\theta}}{\tilde{\sigma}^2}\right)\mathrm{cov}_{\mathrm{GE}}(Z,\ln Z)$$

$$-\frac{1}{\tilde{\sigma}^2}\,\mathrm{cov}_{\mathrm{GE}}(Z,(\ln Z)^2)+\frac{1}{\tilde{\sigma}^2}\left(1-\frac{\ln\lambda\tilde{\theta}}{\tilde{\sigma}^2}\right)\mathrm{cov}_{\mathrm{GE}}(\ln Z,(\ln Z)^2) \tag{12}$$

Case 2: The data are coming from a log-normal $\mathrm{LN}(\sigma,\theta)$.

Lemma 2. *Under the assumption that the data are from* $\mathrm{LN}(\sigma,\theta)$, *we have as* $n\to\infty$,

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

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

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

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

Here, $\tilde{\alpha}$ *and* $\tilde{\lambda}$ *also depend on* θ *and* σ *but for brevity we do not make it explicit. Let us denote*

$$T_*=\ln\left(\frac{L_{\mathrm{GE}}(\tilde{\alpha},\tilde{\lambda})}{L_{\mathrm{LN}}(\sigma,\theta)}\right).$$

(iii) $n^{-1/2}[T-E_{\mathrm{LN}}(T)]$ *is asymptotically equivalent to* $n^{-1/2}[T_*-E_{\mathrm{LN}}(T_*)]$.

The proof of Lemma 2 is omitted.

Theorem 2. *Under the assumption that the data are from* $\mathrm{LN}(\sigma,\theta)$, *the distribution of* T *is asymptotically normal with mean* $E_{\mathrm{LN}}(T)$ *and variance* $V_{\mathrm{LN}}(T_*)=V_{\mathrm{LN}}(T)$.

The proof of Theorem 2 follows along the same line as of Theorem 1.

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

$$\begin{aligned}h(\alpha,\lambda)&=E_{\mathrm{LN}}[\ln(f_{\mathrm{GE}}(X;\alpha,\lambda))]\\&=E_{\mathrm{LN}}[\ln\alpha+\ln\lambda-\lambda X+(\alpha-1)\ln(1-\mathrm{e}^{-\lambda X})]\\&=\ln\alpha+\ln\lambda-\lambda\theta\mathrm{e}^{\sigma^2/2}+(\alpha-1)u(\sigma,\lambda\theta),\end{aligned}$$

where

$$u(x,y)=\frac{1}{\sqrt{2\pi}x}\int_0^\infty\frac{1}{z}\ln(1-\mathrm{e}^{-yz})\mathrm{e}^{-(\ln z)^2/2x^2}\,\mathrm{d}z.$$

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

$$\frac{1}{\tilde{\alpha}}+u(\sigma,\tilde{\lambda}\theta)=0 \tag{13}$$

and

$$\frac{1}{\tilde{\lambda}} - \theta e^{\sigma^2/2} + (\tilde{\alpha} - 1)\theta u_2(\sigma, \tilde{\lambda}\theta) = 0. \tag{14}$$

Here $u_2(x, y)$ is the derivative of $u(x, y)$ with respect to y, *i.e.*,

$$u_2(x, y) = \frac{1}{\sqrt{2\pi}x} \int_0^\infty \frac{e^{-yz}}{(1 - e^{-yz})} e^{-(\ln z)^2/2x^2} \, dz. \tag{15}$$

From (13) it is clear that $(\tilde{\lambda}\theta)$ is a function of $\tilde{\alpha}$ and σ only. From (14) it follows that $\tilde{\alpha}$ is a function of σ only, therefore, $(\tilde{\lambda}\theta)$ is a function of σ only.

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

$$\begin{aligned}
\frac{E_{\mathrm{LN}}(T)}{n} &\approx AM_{\mathrm{LN}}(\sigma) = E_{\mathrm{LN}}[\ln(f_{\mathrm{GE}}(\tilde{\alpha}, \tilde{\lambda})) - \ln(f_{\mathrm{LN}}(\sigma, \theta))] \\
&= E_{\mathrm{LN}}\left[\ln(\tilde{\alpha}\tilde{\lambda}) - \tilde{\lambda}\theta Y + (\tilde{\alpha} - 1)\ln(1 - e^{-\tilde{\lambda}\theta Y}) + \frac{1}{2}\ln(2\pi\sigma^2)\right. \\
&\quad \left. + \ln(\theta Y) + \frac{1}{2\sigma^2}(\ln Y)^2\right] \\
&= \ln(\tilde{\alpha}\tilde{\lambda}\theta) - \tilde{\lambda}\theta e^{\sigma^2/2} + (\tilde{\alpha} - 1)E_{\mathrm{LN}}[\ln(1 - e^{-\tilde{\lambda}\theta Y})] \\
&\quad + \frac{1}{2}\ln(2\pi\sigma^2) + \frac{1}{2}.
\end{aligned} \tag{16}$$

Also,

$$\begin{aligned}
\frac{V_{\mathrm{LN}}(T)}{n} &\approx AV_{\mathrm{LN}}(\sigma) = V_{\mathrm{LN}}[\ln(f_{\mathrm{GE}}(\tilde{\alpha}, \tilde{\lambda})) - \ln(f_{\mathrm{LN}}(\sigma, \theta))] \\
&= V_{\mathrm{LN}}\left[-\tilde{\lambda}\theta Y + (\tilde{\alpha} - 1)\ln(1 - e^{-\tilde{\lambda}\theta Y}) + \ln Y + \frac{1}{2\sigma^2}(\ln Y)^2\right] \\
&= \theta^2\tilde{\lambda}^2 e^{\sigma^2}(e^{\sigma^2} - 1) + (\tilde{\alpha} - 1)^2 V_{\mathrm{LN}}(\ln(1 - e^{-\tilde{\lambda}\theta Y})) + \sigma^2 + \frac{1}{2} \\
&\quad + 2(\tilde{\alpha} - 1)\,\mathrm{cov}_{\mathrm{LN}}(\ln(1 - e^{-\tilde{\lambda}\theta Y}), \ln Y) - 2\tilde{\lambda}\theta\,\mathrm{cov}_{\mathrm{LN}}(Y, \ln Y) \\
&\quad - \frac{\theta\tilde{\lambda}}{\sigma^2}\,\mathrm{cov}_{\mathrm{LN}}(Y, (\ln Y)^2) + \frac{\tilde{\alpha} - 1}{\sigma^2}\,\mathrm{cov}_{\mathrm{LN}}(\ln(1 - e^{-\tilde{\lambda}\theta Y}), (\ln Y)^2) \\
&\quad - 2\tilde{\lambda}\theta(\tilde{\alpha} - 1)\,\mathrm{cov}_{\mathrm{LN}}(Y, \ln(1 - e^{-\tilde{\lambda}\theta Y})) + \frac{1}{\sigma^2}\,\mathrm{cov}_{\mathrm{LN}}(\ln Y, (\ln Y)^2).
\end{aligned} \tag{17}$$

Note that $\tilde{\alpha}$, $\tilde{\lambda}$, $\mathrm{AM}_{\mathrm{LN}}(\sigma)$, $\mathrm{AV}_{\mathrm{LN}}(\sigma)$, $\tilde{\theta}$, $\tilde{\sigma}$, $\mathrm{AM}_{\mathrm{GE}}(\alpha)$ and $\mathrm{AV}_{\mathrm{GE}}(\alpha)$ are quite difficult to compute numerically. We present $\tilde{\alpha}$, $\tilde{\lambda}$, $\tilde{\theta}$, $\tilde{\sigma}$, $\mathrm{AM}_{\mathrm{LN}}(\sigma)$, $\mathrm{AV}_{\mathrm{LN}}(\sigma)$, $\mathrm{AM}_{\mathrm{GE}}(\alpha)$ and $\mathrm{AV}_{\mathrm{GE}}(\alpha)$ in Tables 1 and 2 for convenience.

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

$\alpha \rightarrow$	0.75	1.00	1.25	1.50	1.75	2.00	2.25	2.50
AM_{GE}	0.1100	0.0814	0.0625	0.0494	0.0399	0.0327	0.0272	0.0228
AV_{GE}	0.2599	0.1877	0.1407	0.1087	0.0862	0.0697	0.0572	0.0476
$\tilde{\sigma}$	1.592	1.287	1.106	0.985	0.898	0.832	0.780	0.738
$\tilde{\theta}$	0.369	0.554	0.720	0.867	0.998	1.115	1.221	1.318

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

$\sigma \rightarrow$	0.50	0.60	0.70	0.80	0.90	1.00	1.10	1.20
AM_{LN}	−0.0054	−0.0142	−0.0258	−0.0393	−0.0542	−0.0701	−0.0867	−0.1036
AV_{LN}	0.0090	0.0281	0.0579	0.0986	0.1507	0.2147	0.2905	0.3779
$\tilde{\alpha}$	6.181	3.850	2.664	1.976	1.537	1.239	1.026	0.868
$\tilde{\lambda}$	6.023	4.743	3.774	3. .018	2.416	1.932	1.541	1.225

4. Determination of sample size

We propose a method to determine the minimum sample size required to discriminate between the log-normal and GE distributions, for a given user specified probability of correct selection (PCS). It is very important to know the closeness between the two distribution functions before discriminating between them. There are several ways to measure the closeness or the distance between two distribution functions, but the most important one is the Kolmogorov–Smirnov (K–S) distance. If the distance between the two distributions is small, then a very large sample size is needed to discriminate between them for a given PCS and if the distance between two distribution functions is large one may not need very large sample size to discriminate between them. This is also true that if the distance between two distribution functions is small, one may not need to distinguish the two distributions from any practical point of view. This 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. The tolerance limit and PCS are equivalent to type I error and power in the corresponding testing of hypotheses problem. 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.

In Section 3 it was observed that the logarithm of the RML statistic follows approximately a normal distribution for large n. It can be used to determine the required sample size n such that the PCS achieves a certain protection level p^* for a given tolerance level D^*. This can be explained assuming case 1. Case 2 follows exactly along the same line.

Table 3
The minimum sample size $n = z^2_{0.70}\mathrm{AV}_{\mathrm{GE}}(\alpha)/(\mathrm{AM}_{\mathrm{GE}}(\alpha))^2$, using (4.5), for $p^* = 0.7$ and when the null distribution is GE is presented.

$\alpha \rightarrow$	0.75	1.00	1.25	1.50	1.75	2.00	2.25	2.50
$n \rightarrow$	6	8	10	13	15	18	22	26
K–S	0.311	0.117	0.096	0.081	0.070	0.062	0.055	0.049
Diff	8.65	5.24	3.71	2.83	2.27	1.87	1.57	1.35
Ratio	0.333	0.468	0.565	0.639	0.695	0.739	0.775	0.804

The K–S distance, the difference and ratio of the 99th percentile points of GE $(\alpha, 1)$ and LN$(\tilde{\sigma}, \tilde{\theta})$ for different values of α are reported.

Since T is asymptotically normally distributed with mean $E_{\mathrm{GE}}(T)$ and variance $V_{\mathrm{GE}}(T)$, therefore PCS is

$$\mathrm{PCS}(\alpha) = P[T > 0 \mid \alpha] \approx 1 - \Phi\left(\frac{-E_{\mathrm{GE}}(T)}{\sqrt{V_{\mathrm{GE}}(T)}}\right) = 1 - \Phi\left(\frac{-n \times \mathrm{AM}_{\mathrm{GE}}(\alpha)}{\sqrt{n \times \mathrm{AV}_{\mathrm{GE}}(\alpha)}}\right). \tag{18}$$

Here, Φ is the distribution function of the standard normal random variable. Now to determine the sample size needed to achieve at least a p^* protection level, equate

$$\Phi\left(\frac{-n \times \mathrm{AM}_{\mathrm{GE}}(\alpha)}{\sqrt{n \times \mathrm{AV}_{\mathrm{GE}}(\alpha)}}\right) = 1 - p^*, \tag{19}$$

and solve for n. It provides

$$n = \frac{z^2_{p^*}\mathrm{AV}_{\mathrm{GE}}(\alpha)}{(\mathrm{AM}_{\mathrm{GE}}(\alpha))^2}. \tag{20}$$

Here z_{p^*} is the $100p^*$ percentile point of a standard normal distribution. For $p^* = 0.7$ and for different α, the values of n are reported in Table 3. Similarly for case 2, we need

$$n = \frac{z^2_{p^*}\mathrm{AV}_{\mathrm{LN}}(\sigma)}{(\mathrm{AM}_{\mathrm{LN}}(\sigma))^2}. \tag{21}$$

We report n for different values of σ when $p^* = 0.7$ in Table 4. From Table 3, it is clear that as α increases the required sample size increases. Moreover, from Table 4, it is immediate that as σ increases the required sample size decreases. 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 (20) or (21) and using the fact that n is a monotone function of the shape parameters in both cases. But, unfortunately, in practice it may be completely unknown. Therefore, to have some idea of the shape parameter of the null distribution we make the following assumptions. It is assumed that the experimenter would like to choose the minimum sample size needed for a given protection level when the distance between two distribution functions is greater than a pre-specified tolerance level. The distance between two distribution functions is

Table 4
The minimum sample size $n = z^2_{0.70}\mathrm{AV}_{\mathrm{LN}}(\sigma)/(\mathrm{AM}_{\mathrm{LN}}(\sigma))^2$, using (4.6), for $p^* = 0.7$ and when the null distribution is log-normal is presented.

$\sigma \rightarrow$	0.50	0.60	0.70	0.80	0.90	1.00	1.10	1.20
$n \rightarrow$	85	39	24	18	15	13	11	10
K–S	0.022	0.045	0.072	0.105	0.144	0.191	0.244	0.304
Diff	1.99	2.57	3.30	4.23	5.40	6.86	8.70	10.99
Ratio	0.345	0.327	0.309	0.293	0.279	0.267	0.257	0.249

The K–S distance, the difference and ratio of the 99th percentile points of LN $(\sigma, 0.368)$ and GE $(\tilde{\alpha}, \tilde{\lambda})$ for different values of σ are reported.

defined by the K–S distance. The K–S distance between two distribution functions, say $F(x)$ and $G(x)$ is defined as

$$\sup_x |F(x) - G(x)|. \tag{22}$$

We report the K–S distance between the GE$(\alpha, 1)$ and LN$(\tilde{\sigma}, \tilde{\theta})$ for different values of α in Table 3. Here $\tilde{\sigma}$ and $\tilde{\theta}$ are as defined in Lemma 1 and are in Table 1. Similarly, the K–S distance between the LN$(\sigma, 0.368)$ (note that $\ln 0.368 = -1$ and we have taken the scale parameter of the log normal distribution as 0.368 for convenience) and GE$(\tilde{\alpha}, \tilde{\lambda})$ for different values of σ is reported in Table 4. Here $\tilde{\alpha}$ and $\tilde{\lambda}$ are as defined in Lemma 2 and are reported in Table 2. From Tables 3 and 4 it is observed that as the distance between the two distribution functions decreases the minimum sample size increases. The findings are quite intuitive in the sense that large sample sizes are needed to discriminate between the two distribution functions if they are very close and vice verse.

Now we discuss how we can determine the required sample size to discriminate between the log-normal and GE distribution functions for a user specified protection level and tolerance level. Suppose the protection level is $p^* = 0.7$ and the tolerance level is given in terms of K–S distance as $D^* = 0.07$. Here tolerance level $D^* = 0.07$ means that the practitioner wants to discriminate between a log-normal and GE distribution functions only when their K–S distance is more than 0.07. From Table 3, it is observed that the K–S distance will be more than 0.07 if $\alpha \leqslant 1.75$. Similarly from Table 4, it is clear that the K–S distance will be more than 0.07 if $\sigma \geqslant 0.70$. Therefore, if the data come from a GE distribution, then for the tolerance level $D^* = 0.07$, one needs at least $n = 15$ to meet the PCS, $p^* = 0.7$. Similarly if the data come from the log-normal distribution then one needs at least $n = 24$ to meet the above protection level $p^* = 0.7$ for the same tolerance level $D^* = 0.07$. Therefore, for the given tolerance level 0.07 one needs at least $\max(15, 24) = 24$ to meet the protection level $p^* = 0.7$ simultaneously for both cases.

Note that, two small tables are provided for the protection level 0.70 but for the other protection level the tables can be easily used as follows. For example if we need the protection level $p^* = 0.9$, then all the entries corresponding to the row of n, will be multiplied by $z^2_{0.9}/z^2_{0.7}$, because of (20) and (21). Therefore, Tables 3 and 4 can be used for any given protection level. Two of the referees pointed out that the K–S distance

may not be a good measure of distances between the two distribution functions. The better choice might be the difference or ratio of the upper percentile points of the two distribution functions. We report the difference and ratio of the 99th percentile points of the two distribution functions along with the K–S distances in Tables 3 and 4. They also can be used as the distance measure between the two distribution functions. They also provide similar results.

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 have been performed at the Indian Institute of Technology Kanpur, on a Pentium-IV processor and all the programs written in C, can be obtained from the authors on request. We use the random deviate generator of Press et al. (1993). We compute the probability of correct selections based on simulations and on the asymptotic results derived in

Table 5
The probability of correct selection based on Monte Carlo Simulations and also based on asymptotic results when the null distribution is GE

$\alpha \downarrow$	$n \rightarrow$				
	20	40	60	80	100
0.75	0.84 (0.83)	0.92 (0.91)	0.93 (0.95)	0.95 (0.97)	0.97 (0.98)
1.00	0.80 (0.80)	0.89 (0.88)	0.92 (0.93)	0.93 (0.95)	0.94 (0.96)
1.25	0.76 (0.77)	0.86 (0.85)	0.91 (0.90)	0.92 (0.93)	0.94 (0.95)
1.50	0.74 (0.75)	0.84 (0.83)	0.89 (0.88)	0.91 (0.91)	0.92 (0.93)
1.75	0.71 (0.73)	0.81 (0.81)	0.87 (0.85)	0.89 (0.89)	0.91 (0.91)
2.00	0.68 (0.71)	0.78 (0.78)	0.85 (0.83)	0.89 (0.87)	0.90 (0.89)
2.25	0.66 (0.69)	0.76 (0.76)	0.82 (0.81)	0.86 (0.85)	0.87 (0.87)
2.50	0.63 (0.68)	0.74 (0.75)	0.80 (0.79)	0.82 (0.82)	0.85 (0.85)

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.

Table 6
The probability of correct selection based on Monte Carlo Simulations and also based on asymptotic results when the null distribution is log-normal

$\sigma \downarrow$	$n \rightarrow$				
	20	40	60	80	100
0.50	0.62 (0.60)	0.65 (0.64)	0.68 (0.67)	0.70 (0.70)	0.72 (0.72)
0.60	0.65 (0.65)	0.70 (0.71)	0.75 (0.75)	0.77 (0.78)	0.81 (0.80)
0.70	0.68 (0.68)	0.75 (0.75)	0.80 (0.80)	0.84 (0.83)	0.86 (0.86)
0.80	0.70 (0.71)	0.79 (0.79)	0.85 (0.84)	0.88 (0.87)	0.90 (0.89)
0.90	0.72 (0.73)	0.82 (0.81)	0.88 (0.86)	0.91 (0.89)	0.92 (0.92)
1.00	0.75 (0.75)	0.85 (0.83)	0.90 (0.88)	0.92 (0.91)	0.93 (0.93)
1.10	0.76 (0.76)	0.87 (0.85)	0.90 (0.89)	0.94 (0.93)	0.95 (0.95)
1.20	0.78 (0.77)	0.88 (0.86)	0.90 (0.90)	0.93 (0.90)	0.98 (0.95)

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.

Section 3. We consider different sample sizes and also different shape parameters, as explained below.

First we consider the case when the data are coming from a GE distribution. In this case, we consider $n = 20, 40, 60, 80, 100$ and $\alpha = 0.75$, 1.00, 1.25, 1.50, 1.75, 2.00, 2.25 and 2.50. For a fixed α and n we generate a random sample of size n from GE$(\alpha, 1)$, compute T as defined in (6) 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 the asymptotic results as given in (18). The results are reported in Table 5. Similarly, we obtain the results when the data are generated from a log-normal distribution, for the same set of n and $\sigma = 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 1.1, 1.2$. 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 cases. It is also clear that when the shape parameter increases for the GE distribution the PCS decreases and when

the shape parameter increases for the log-normal 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.

6. Conclusions

In this paper, we consider the problem of discriminating between two families of distribution functions, the log-normal and GE families. We consider the statistic based on the logarithm of the ratio of the maximized likelihoods and obtain 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.

Acknowledgements

The authors thank the referees for their valuable comments. Part of the work was supported by a grant from the Natural Sciences and Engineering Research Council.

References

Atkinson, A., 1969. A test for discriminating between models. Biometrika 56, 337–347.

Atkinson, A., 1970. A method for discriminating between models (with discussions). J. Roy. Statist. Soc. Ser. B 32, 323–353.

Bain, L.J., Englehardt, M., 1980. Probability of correct selection of Weibull versus gamma based on likelihood ratio. Comm. Statist. Ser. A 9, 375–381.

Bain, L.J., Englehardt, M., 1991. Statistical Analysis of Reliability and Lifetime Model, 2nd Edition. Marcel Dekker, New York.

Chambers, E.A., Cox, D.R., 1967. Discriminating between alternative binary response models. Biometrika 54, 573–578.

Chen, W.W., 1980. On the tests of separate families of hypotheses with small sample size. J. Statist. Comput. Simulations 2, 183–187.

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, Berkley, CA, pp.105–123.

Cox, D.R., 1962. Further results on tests of separate families of hypotheses. J. Roy. Statist. Soc. Ser. B 24, 406–424.

Dyer, A.R., 1973. Discrimination procedure for separate families of hypotheses. J. Amer. Statist. Assoc. 68, 970–974.

Dumonceaux, R., Antle, C.E., 1973. Discriminating between the log-normal and Weibull distribution. Technometrics 15 (4), 923–926.

Fearn, D.H., Nebenzahl, E., 1991. On the maximum likelihood ratio method of deciding between the Weibull and Gamma distribution. Comm. Statist. Theory Methods 20 (2), 579–593.

Gupta, R.D., Kundu, D., 1999. Generalized exponential distributions. Austral. NZ J. Statist. 41 (2), 173–188.

Gupta, R.D., Kundu, D., 2001a. Exponentiated exponential distribution: an alternative to gamma and Weibull distributions. Biometrical J. 43 (1), 117–130.

Gupta, R.D., Kundu, D., 2001b. Generalized exponential distributions: different methods of estimations. J. Statist. Comput. Simulation 69 (4), 315–338.

Gupta, R.D., Kundu, D., 2002. Generalized exponential distributions: statistical inferences. J. Statist. Theory Appl. 1, 101–118.

Gupta, R.D., Kundu, D., 2003a. Discriminating between the Weibull and generalized exponential distributions. Comput. Statist. Data Anal. 43, 179–196.

Gupta, R.D., Kundu, D., 2003b. Discriminating between the gamma and generalized exponential distributions. J. Statist. Comput. Simulation, to appear.

Jackson, O.A.Y., 1968. Some results on tests of separate families of hypotheses. Biometrika 55, 355–363.

Jackson, O.A.Y., 1969. Fitting a gamma or log-normal distribution to fiber-diameter measurements on wool tops. Appl. Statist. 18, 70–75.

Johnson, N., Kotz, S., Balakrishnan, N., 1995. Continuous Univariate Distribution, Vol. 1. Wiley, New York.

Pereira, B. de, 1978. Empirical comparison of some tests of separate families of hypotheses. Metrika 25, 219–234.

Press et al., 1993. Numerical Recipes in C. Cambridge University Press, Cambridge.

Quesenberry, C.P., Kent, J., 1982. Selecting among probability distributions used in reliability. Technometrics 24 (1), 59–65.

Raqab, M.Z., 2002. Inference for generalized exponential distribution based on record statistics. J. Statist. Plann. Inference 104 (2), 339–350.

Raqab, M.Z., Ahsanullah, M., 2001. Estimation of the location and scale parameters of the generalized exponential distribution based on order statistics. J. Statist. Comput. Simulation 69 (2), 109–124.

White, H., 1982. Regularity conditions for Cox's test of non-nested hypotheses. J. Econometrics 19, 301–318.

Wiens, B.L., 1999. When log-normal and gamma models give different results: a case study. Amer. Statist. 53 (2), 89–93.

Zheng, G., 2002. On the Fisher information matrix in type-II censored data from the exponentiated exponential family. Biometrical J. 44 (3), 353–357.