Ekonometri ve İstatistik Sayı:16 2012 44-54

İSTANBUL ÜNİVERSİTESİ
İKTİSAT FAKÜLTESİ
EKONOMETRİ VE İSTATİSTİK
DERGİSİ

DISCRIMINATING BETWEEN WEIBULL AND LOG-NORMAL DISTRIBUTIONS BASED ON KULLBACK-LEIBLER DIVERGENCE

Ali-Akbar BROMIDEH[*1]

Abstract

The Weibull and Log-Normal distributions are frequently used in reliability to analyze lifetime (or failure time) data. The ratio of maximized likelihood (RML) has been extensively used in choosing between the two distributions. The Kullback-Leibler information is a measure of uncertainty between two densities. We examine the use of Kullback-Leibler Divergence (KLD) in discriminating either the Weibull or Log-Normal distribution. An advantage of the KLD is that it incorporates entropy of each model. We explain the applicability of the KLD by a real data set and the consistency of the KLD with the RML is established.

Keywords: Model discrimination Weibull distribution Log-Normal distribution Kullback-Leibler divergence Ratio of maximized likelihood.
Jel Classification:

[*] Department of Statistics, Faculty of Mathematical Sciences, Shahid Beheshti University, Evin, Tehran, Iran.

[1] Strategic Planning Center, Iran Khodro Co., Tehran, E-mail: bromideh@gmail.com

1. INTRODUCTION

Weibull and Log-Normal distributions have been used in analyzing skewed positive data. Generally, positively skewed data play important roles in the reliability analysis. Both distributions are commonly used to model certain lifetimes in reliability and survival analysis. Although these two models may provide similar data fit for moderate sample sizes, but still it is desirable to select the correct model and make the best possible decision based on observed data. Often choosing a particular model is difficult and the relevant effect of model mis-selection can be quite severe.

Weibull distribution has been used frequently to describe the distribution of lifetime data. For example, see Cohen and Whitten (1988) and Abernethy (2002) for more details on this distribution and its characteristics. On the other hand, Log-Normal distribution is commonly used to model lifetimes in reliability and survival analysis, among several other distributions. Survival times of patients with certain types of cancer, failure times of semiconductor devices, insurance claim payments are a few of examples where can be well modeled by either distributions. See Meeker and Escobar (1998), Blishke and Murthy (2000) and Crow and Shimitzu (1988) and the references therein for an overview on this model.

In lifetime models, selection of the best model among several potential/candidate distributions is a problem of interest. It's observed that both distributions can be used quite effectively to analyze skewed data sets. The problem of discriminating between two distributions for testing whether some given data follow one of the two potential probability distributions is quite old in the statistical literature. The problem of testing whether some given observations follow one of the two possible distribution has been studied by many researchers. For instance see Dumonceaux and Antle (1973), Kundu and Manglick (2004) and Pascual (2005)). The idea has been extended to discriminate between Gamma and Weibull distributions (Bain and Englehardt (1980), Fearn and Nebenzahl (1991) and Mohd Saat et al. (2008)), between Gamma and Log-Normal distributions (Kundu and Manglick (2005)). Due to increasing applications of lifetime models, special attention is given to the discrimination between Weibull and Log-Normal distribution.

Although, the goodness-of-fit and significance testing is initially used in selection of two probability densities, but recently, the ratio of maximized likelihood (RML) test statistic is mostly used in the literature. To discriminate between Weibull and Log-Normal, RML is not a suitable selection criteria due to lack of inclusion of the mean parameter of Log-Normal, μ, and on the other hand, for small sample size it has low power (Dumonceaux and Antle (1973)). Further, this paper aims to introduce a new test statistic based on Kullback-Leibler information (distance) for model selection purposes. An advantage of this approach is that it incorporates information contained in both models. Second, all parameters play important role in the test statistic.In this paper, therefore, we will discuss on selection methodology between Weibull and Log-Normal distributions based on Kullback-Leibler divergence/distance (KLD). In fact, KLD indicates "how far away" a probability distribution **P** from another distribution **Q**.

The rest of the paper is organized as follows. In the following section, a brief presentation and MLE of parameters will be provided for both Weibull and Log-Normal distributions. In section 3, the proposed approach, i.e. Kullback-Leibler divergence, will be discussed. A real (lifetime) data set is analyzed in section 4, to illustrate how the proposed method works in practice. Finally, we will conclude the discussion in the last section.

2. PRELIMINARIES: WEIBULL AND LOG-NORMAL DISTRIBUTIONS

Weibull and Log-Normal distributions are among the possible models to analysis lifetime data in reliability and survival analysis. Both distributions can be found in any statistical literatures, but the notation is somehow different. To make it clear, the Weibull and Log-Normal distributions are recalled and the notations are introduced.

A positive random variable X is said to have a Weibull distribution, denoted by $We(\beta, \theta)$, when it has the probability density function (pdf) of

$$f(x|\beta,\theta) = \beta\theta^{\beta}x^{\beta-1}\exp\left(-(\theta x)^{\beta}\right) \tag{1}$$

where $x > 0$, and $\beta > 0, \theta > 0$ are shape and scale parameters, respectively. The MLE of β and θ can be calculated by:

$$\hat{\theta} = \left(\frac{n}{\sum_{i-1}^{n} x_i^{\hat{\beta}}} \right)^{-\frac{1}{\hat{\beta}}} \tag{2}$$

A numerical analysis is required to estimate the unknown parameters of β and θ in a Weibull distribution (see Thoman et. al. (1969)). The R code developed to estimate the parameters of a Weibull distribution can be shared upon request.

A random variable X is distributed as Log-Normal, denoted as $LN(\mu, \sigma^2)$, if $\ln(X)$ is normal, e.g. $\ln(X) \sim N(\mu, \sigma^2)$. The probability density of X is given by:

$$g(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma x}} \exp\left(- \frac{(ln(x) - \mu)^2}{2\sigma^2} \right) \tag{3}$$

where $x > 0, \mu > 0$ and $\sigma > 0$. The MLE of μ and σ^2 are given below, respectively:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} \ln(X_i) \;\; and \;\; \widehat{\sigma^2} = \frac{1}{n} \sum_{i=1}^{n} (\ln(X_i) - \hat{\mu})^2. \tag{4}$$

3. MODEL SELECTION: WEIBULL OR LOG-NORMAL?

Let $x_1, x_2, \dots, x_k$ be independent and identically distributed (iid) random variables from any one of the two probability distributions. Consider the problem of testing the null hypothesis H_0 that the distribution of the sample (data) is Weibull versus the alternative hypothesis that states it comes from a Log-Normal distribution. In other words, we are interested to test these hypotheses:

$$H_0 : X \sim LN(\mu, \sigma^2) \tag{5}$$

against

$$H_1 : X \sim We(\beta, \theta) \tag{6}$$

The main purpose of this paper is testing (5) against (6). Among the various testing methods, the most attention has been taken into consideration by goodness-of-fitness and RML. Recently, Kundu and Manglick (2004) and Pascual (2005) used the RML test statistic to choose Weibull and Log-Normal. They introduced the logarithm of RML as follows:

$$T = n\left[\frac{1}{2} - \ln\left(\hat{\sigma}\hat{\beta}(\hat{\mu}\hat{\theta})^{\hat{\beta}}\sqrt{2\pi}\right)\right]. \tag{7}$$

Finally, the following discrimination procedure was adopted: "Choose the Log-Normal distribution if the test statistic $T > 0$; otherwise choose the Weibull distribution as the preferred model." In the following section, the new test statistic based on Kullback-Leibler information/distance will be explained.

3.1. Kullback-Leibler Divergence based Test Statistic

In probability and information theory, the Kullback-Leibler divergence (also information discrepancy, information gain, relative entropy, or KLD) is a non-symmetric measure of the difference (dissimilarity) between two probability distributions **f** and **g**. Kullback-Leibler information between models f and g is defined for continues functions as:

$$KLD(f,g) = \int f(x) \ln\left(\frac{f(x)}{g(x)}\right) dx \tag{8}$$

It denotes the "information lost when **g** is used to approximate **f** or the distance from **g** to **f**." In other words, KLD is a measure of inefficiency of assuming that the distribution is **g** when the true distribution is **f**. Since the measure from **f** to **g** is not the same as the measure from **g** to **f**, then it can be conceptualized as a "*directed/oriented distance*" between the two models (Burnham and Anderson (2002)).

The KLD is a natural distance function between models and it is a fundamental quantity in science and information theory. It is usually used as a logical basis for model selection in conjunction with likelihood inference. Values of KLD are not based on only the mean and variance of the distributions; rather, the distributions in their entirety are the subject of comparison. The later is regarded an advantage of the KLD as a test statistic.

It is well known that $KLD(g,f) \neq KLD(f,g)$ and $KLD(f,g) \geq 0$ and the equality holds if and only if $f = g$ (Burnham and Anderson (2002)). The smaller $KLD(f,g)$ means that "**f**" is preferred and large values of KLD favor "**g**." To discriminate between the two distributions, in our case, to test (5) vs. (6) the KLD based test statistic is a ruler to measure the similarity between the two hypotheses/distributions. It is given by:

$$\begin{aligned} KLD(g_{ln}(x), f_{we}(x)) &= \int_0^\infty g_{ln}(x) \ln\left(\frac{g_{ln}(x)}{f_{we}(x)}\right) dx \\ &= H(g_{ln}(x)) - \int_0^\infty g_{ln}(x) \ln\,(f_{we}(\end{aligned} \qquad (9)$$

where $f_{we}(x)$ denotes on Weibull distribution and $g_{ln}(x)$ is the pdf of Log-Normal distribution and $H(g_{ln}(x))$ is the entropy of Log-Normal distribution.

Finally, the KLD test statistic for testing (5) vs. (6) is given by:

$$\begin{aligned} KLD(g_{LN}(x), f_{We}(x)) &= \hat{\beta} \ln(\hat{\theta}) + \hat{\theta}^{-\hat{\beta}} \exp\,[\hat{\beta}\hat{\mu} + \frac{1}{2}(\\ &\quad -\frac{1}{2} - \ln(\sqrt{2\pi}\hat{\sigma}) - \ln(\hat{\beta}) - \end{aligned} \qquad (10)$$

However, large values of KLD in (10), indicates that the data come from a Weibull distribution. In other words, we reject the null hypothesis, H_0, in favor of H_1, at the significance level α , if $KLD \geq C_{ln}^{n}(\alpha)$, where the critical point, $C_{ln}^{n}(\alpha)$, is determined by the $(1-\alpha) - quantile$ of the distribution KLD under the null hypothesis H_0. The detail on critical values computation is discussed in the following subsection. As we are considering the choice of a model as a test of hypothesis, it is important to allow either of the models to be the null hypothesis. We suppose that the researcher would assign to the null hypothesis the model which prefers to use. Unless there is convincing evidence that one should use the other. In order to allow the researcher this choice, we next provide two tables of critical values with Log-Normal and Weibull as the null hypothesis, respectively.

Hence, the KLD test statistic for testing $H_0: x \sim We(\beta,\theta)$ vs. $H_1: x \sim LN(\mu,\sigma^2)$ is given by:

$$KLD(f_{we}(x), g_{ln}(x)) = \ln(\sqrt{2\pi}\hat{\sigma}) + \ln(\hat{\beta}) - 1 - \hat{\gamma} + \frac{1}{2\hat{\sigma}^2}\left[(\hat{\mu} + \frac{\hat{\gamma}}{\hat{\beta}} - \ln(\hat{\theta}))^2 + \frac{1}{6}\right. \qquad (11)$$

where $\gamma = 0.5777$ is the Euler constant. Finally, the selection procedure is: "Choose the Log-Normal distribution if the test statistic $KLD > C_{we}^{n}(\alpha)$; otherwise choose the Weibull distribution as the best model." Since it is difficult to compute the exact distributions of $KLD(g_{ln}(x), f_{we}(x))$ and $KLD(f_{we}(x), g_{ln}(x))$, therefore, we use Mote Carlo simulations to compute the critical values for different sample sizes and models parameters. In the next section, we will analyze a real data to explain how the proposed test statistic works. Actually, we used the frequently discussed data in the literatures which has been analyzed by many researchers. For instance see Dumonceaux and Antle (1973), Kundu and Manglick (2004) and Pascual (2005).

4. DATA ANALYSIS : IMPLEMENTATION OF THE KLD TEST

To illustrate the use of our proposed new test statistic, i.e. KLD, we analyze two real life data sets. In fact we use the KLD to discriminate between two distribution functions.

Suppose the following observations (as given by Lieblein and Zelen (1956) for the lifetime) are used to test whether the data come from a Weibull or a Log-Normal. 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 lifetime 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.

For these data we obtain $\hat{\mu} = 4.1506$ and and $\hat{\sigma} = 0.5215$ for Log-Normal distribution and $\hat{\beta} = 2.1026$ and $\hat{\theta} = 81.88$ in the Weibull model.

Case 1: $H_0: Log - Normal$ *vs.* $H_1: Weibull.$ we find $KLD(g_{ln}(x), f_{we}(x)) = 0.0923.$ We see from Table 1 that our calculated $KLD < C_{LN}^{23}(\alpha),$ and consequently we cannot reject the Log-Normality (i.e., H_0) in favor of the Weibull model(H_1).

Table 1: The critical values simulated by Monte Carlo when the null distribution is Log-Normal. The replication rate is 15,000.

n	$\alpha = 1\%$		$\alpha = 5\%$		$\alpha = 10\%$		$\alpha = 20\%$	
	C_{LN}^{n}	Power(%)	C_{LN}^{n}	Power(%)	C_{LN}^{n}	Power(%)	C_{LN}^{n}	Power(%)
20	0.269	(23)	0.173	(45)	0.141	(58)	0.116	(71)
30	0.193	(38)	0.142	(60)	0.123	(71)	0.106	(82)
40	0.168	(50)	0.128	(73)	0.113	(82)	0.101	(90)
60	0.126	(83)	0.106	(94)	0.098	(97)	0.092	(99)

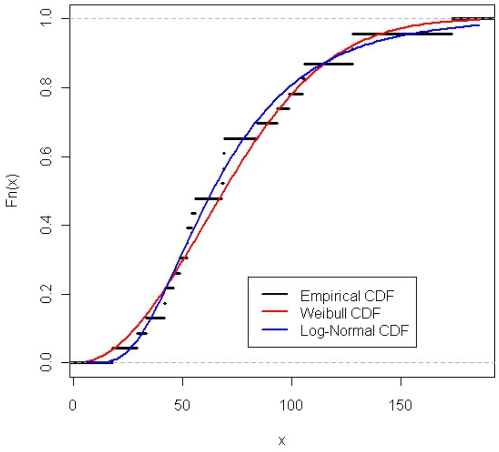
Case 2: $H_0: Weibull$ *vs.* $H_1: Log \quad Normal.$ we find $KLD(f_{we}(x), g_{ln}(x)) = 0.119.$ We see from Table 2 have $C_{We}^{23}(20\%),$ it is clear that one could not reject the Weibull model in favor of the Log-Normal model at the 0.20 level of significance.

Table 2: The critical values simulated by Monte Carlo when the null distribution is Weibull. The replication rate is 15,000 for each sample size.

n	$\alpha = 1\%$		$\alpha = 5\%$		$\alpha = 10\%$		$\alpha = 20\%$	
	C_{We}^{n}	Power(%)	C_{We}^{n}	Power(%)	C_{We}^{n}	Power(%)	C_{We}^{n}	Power(%)
20	0.215	(14)	0.159	(34)	0.138	(47)	0.119	(62)
30	0.183	(25)	0.140	(51)	0.126	(63)	0.111	(73)
40	0.162	(37)	0.130	(63)	0.118	(75)	0.107	(86)
60	0.130	(75)	0.112	(92)	0.106	(96)	0.100	(98)

Finally, we remind that the Log-Normal model may deserve some consideration for such data. The fitted Weibull and Log-Normal distributions are shown in Figure 1.

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



These data have been analyzed by many scholars to discriminate between the two distributions, so far. For example, by using RML, significance testing methods Dumonceaux and Antle (1973) and Kundu and Manglick (2004) could not reject the Log-Normal model in favor of the Weibull. It means that our introduced test statistic based on Kullback-Leibler information is consistent with other frequently testing statistic.

5. CONCLUSION

In this paper we consider the problem of discriminating between two overlapping families of distribution functions, namely Log-Normal and Weibull. It is easy to realize the concept of Kullback-Leibler Divergence (information or distance) based test statistic and its usage in practice. The prime aim of this paper is to introduce another testing statistic. Notice that the comparison of test statistics is another story, which can cover in the future papers. It is observed that the proposed method is consistent with alternative testing statistic. Finally, it is suggested to interested research to test the approach for other similar distributions, such as (Generalized) Gamma, Inverse Gaussian, ... and compare the result with RML and as well as other testing approaches. Finding the exact and/or asymptotic distribution of the proposed test statistic can be an interesting research topic in this regard.

REFERENCES

Abernethy, R., (2006), *"The New Weibull Handbook: Reliability and Statistical Analysis for Predicting Life, Safety, Supportability, Risk, Cost and Warranty Claims,"* 5th Edition, Barringer & Associates.

Bain, L.J., Engelhardt, M. (1980), *"Probability of correct selection of Weibull versus gamma based on likelihood ratio,"* Comm. Statist. Ser. A. 9, 375-381.

Blishke, W.R. and Murthy, D.N.P. (2000) *"Reliability Modeling, prediction and optimization,"* Wiley Interscience Publications.

Burnham, K. P., and Anderson, D. R. (2002), *"Model selection and multimodel inference: a practical information-theoretic approach,"* 2nd eds., Springer, New York.

Cohen, A.C., Whitten, B.J. (1988), *"Parameter estimation in reliability and life span models,"* Marcel Dekker Inc.

Crow and Shimitzu, (1988), *"Log-Normal Distributions, Theory and Applications,"* Marcel Dekker Inc., New York.

Dey, A. K. and Kundu, D. (2009), *"Discriminating among the Log-Normal, Weibull and Generalized Exponential distributions,"* IEEE Transactions on Reliability , vol. 58, no. 3, 416-424.

Dumonceaux, R., and Antle, C.E., (1973), *"Discrimination between the log-normal and the Weibull distributions,* Technometrics 15 (4), 923-926.

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.

Gross, A.J., and Clark, V.A. (1975), *"Survival distribution: reliability application in the biomedical science,"* John Wiley & sons Inc.

Gupta, R.D., Kundu, D., (2003), *"Discriminating between Weibull and generalized exponential distributions,"* Computational Statistics & Data Analysis 43, 179-196.

Lieblein, J, and Zelen, M. (1956), *"Statistical investigation of the fatigue life of deep groove ball bearings,"* J. Res. Nat. Bar. Stand., 57, 273-316.

Lawless, J.F., (1982), *"Statistical Models and Methods for Lifetime Data,"* Wiley, New York.

Meeker and Escobar (1998), *"Statistical Methods for Reliability Data,* New York: John Wiley and Sons.

Mohd Saat, N. Z.; Jemain, A. A. and Al-Mashoor, S. H. (2008), *"A Comparison of Weibull and Gamma Distributions in Application of Sleep Spnea,"* Asian Journal of Mathematics and Statistics, 1 (3), 132-138.

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

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

Kundu D., Gupta, R. D. and Manglick, A. (2005), *"Discriminating between the Log-Normal and generalized exponential distributions,"* Journal of Statistical Planning and Inference. v127. 213-227.

Pascual, F.G. (2005), *"Maximum likelihood estimation under misspecied Log-Normal and Weibull distributions",* Communications in Statistics - Simulation and Computations vol. 34, No. 3, 503 - 524.

Pasha, G. R., Shuaib Khan, M. and Pasha, Ahmed Hesham, (2006), *"Discrimination Between Weibull and Log-Normal Distributions For Lifetime data",* Journal of Research (Science), Bahauddin Zakariya University, Multan, Pakistan. Vol. 17,No.2, April, pp. 103-114.

Thoman, D.R., Bain, L.J. and Antie, C.E. (1969), *"Inferences on the parameter of the weibull distribution,"* Thechnometrics, 11, 445-460.