Journal of Modern Applied Statistical Methods

Volume 11 | Issue 1 Article 22

5-1-2012

The Length-Biased Lognormal Distribution and Its Application in the Analysis of Data from Oil Field Exploration Studies

Makarand V. Ratnaparkhi
Wright State University, makarand.ratnaparkhi@wright.edu

Uttara V. Naik-Nimbalkar
Pune University, Pune, India

Journal of Modern Applied Statistical Methods
May 2012, Vol. 11, No. 1, 255-260

1538 – 9472/12/$95.00

The Length-Biased Lognormal Distribution and Its Application in the Analysis of Data from Oil Field Exploration Studies

Makarand V. Ratnaparkhi
Wright State University,
Dayton, OH

Uttara V. Naik-Nimbalkar
Pune University,
Pune, India

The length-biased version of the lognormal distribution and related estimation problems are considered and sized-biased data arising in the exploration of oil fields is analyzed. The properties of the estimators are studied using simulations and the use of sample mode as an estimate of the lognormal parameter is discussed.

Key words: Length-biased lognormal distribution, estimation, simulations, sample mode, bootstrap.

Introduction

The term length-biased data refers to sample data where the probability of recording an observation depends on the magnitude, for example x, of the observation. In particular, the larger the observation, the higher the probability of observing the related event and, hence, including the corresponding observation in the sample. Length-biased data occur in many research areas and in fields of application, such as, medical science, ecology and geological sciences. Further, the term size-biased data is used to describe the situation where the probability of inclusion of an observation depends on a certain function: $w(x) > 0$ of x.

The length-biased version of the original probability density function (pdf.) that is of interest as a model is considered for modeling length-biased data. The lognormal distribution (LN) with parameters (μ,σ) is known to be a useful model in many applications. Therefore, it is natural to expect the applications of length-biased lognormal distributions (LBLN) in some data analysis problems. For example, the lognormal distribution is commonly used in the analysis of data in geological studies, and Meisner and Demirmen (1981) observed that size-biased data occur in oil-field exploration studies. Yan (2004) considered the presence of length-biasedness in data on incubation periods arising in the SARS epidemic. Among many probability models that are considered for the analysis of these data, the length-biased lognormal distribution is one such model. Quin, et al. (2002) considered such a distribution for data on Breslow thickness in cancer research.

With respect to the properties of the length-biased lognormal distribution, in general, if $f(x;\theta)$ is the original pdf of a non-negative random variable X with $E(X)<\infty$, then its length-biased version is given by

$$g(x;\theta^*)=x\,f(x;\theta)/E(X),\ x>0,$$

where $\theta \in \Omega$ is a scalar or a vector of the parameters of the original distribution of X and $\theta^* \in \Omega^*$ denotes a scalar or a vector of the parameters of the corresponding length-biased version. In some cases θ^* is the same as θ. In practical situations, the interest is in estimating θ, the parameter(s) of the original distribution using length-biased data. However, due to the nature of the available data (length-biased data)

Makarand V. Ratnaparkhi is a Professor of Statistics at Wright State University in Dayton, Ohio, USA. Email him at: makarand.ratnaparkhi@wright.edu. Uttara V. Naik-Nimbalkar is a Professor of Statistics at University of Pune, Pune, India. Email her at: uvnaik@stats.unipune.ac.in.

the experimenter has no other choice but to use the length-biased version of the original distribution. Thus, there is a need to study the properties of the θ estimator with respect to $g(x; \theta*)$, but such an estimation problem is not straightforward for the lognormal distribution.

Methodology

The random variable X is said to have a lognormal distribution with parameters (μ, σ) (denoted by LN (μ, σ)), if its pdf is

$$f(x;\mu,\sigma)=\frac{1}{x\sigma\sqrt{2\pi}}\exp[-\frac{1}{2\sigma^2}(\log x-\mu)^2], \tag{1}$$

where

$$x>0,\ -\infty<\mu<\infty,\ \sigma>0.$$

Length-Biased Lognormal Distribution-Definition and Properties

Using the definition of length-biased distribution, the pdf of the length-biased lognormal distribution (denoted by LBLN (μ, σ)) is given by

$$g\left(x:\mu,\sigma\right)=\frac{1}{x\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2\sigma^2}\left(\log x-\left(\mu+\sigma^2\right)\right)^2\right] \tag{2}$$

where

$$x>0,\ -\infty<\mu<\infty,\ \sigma>\alpha.$$

For convenience (2) will be expressed as

$$g(x;\mu,\sigma)=\frac{1}{x\sigma\sqrt{2\pi}}\exp[-\frac{1}{2\sigma^2}(\log x-\mu^*)^2], \tag{2a}$$

where

$$x>0,\ \mu^*=\mu+\sigma^2,-\infty<\mu<\infty \text{ and } \sigma>0.$$

The properties of the pdfs (1) and (2) are presented in Table 1. The mode of LBLN shown in Table 1 depends only on μ and not on σ^2 as for LN. From Table 2, it is clear that the structure of the Fisher information for LBLN is not the same; hence the related results will not be the same when LBLN is used instead of LN in data analysis.

Parameter Spaces of LN (μ, σ) and LBLN (μ, σ)

In practical situations, for the analysis of length-biased data, the LN (μ, σ) is replaced by LBLN (μ, σ) . Further, examination of the pdfs for (1), (2) and (2a) shows that, although the listed pdfs seem to have the same form, there exists an in-built relationship between the parameters (μ, σ) of LBLN (μ, σ). Thus, studying the implications of this relationship is necessary for the interpretation and estimation of parameters. A brief discussion related to the parameter spaces of (μ, σ) for these two distributions is useful for identifying the underlying problems in data analysis.

Let Ω_1 and Ω_2 denote the respective parameters of LN (μ, σ) and LBLN (μ, σ). Then,

$$\Omega_1=\{(\mu,\sigma)|-\infty<\mu<\infty,\sigma>0\} \tag{3}$$

and

$$\Omega_2=\{(\mu,\sigma)|-\infty<\mu^*=\mu+\sigma^2<\infty,\sigma>0\}. \tag{4}$$

From (3) and (4) it is clear that if the LBLN (μ, σ) is used as a model with $\mu*$ = 0, then it will represent only those members of the original LN (μ, σ) model for which $\mu^*=\mu+\sigma^2=0$, i.e. $\mu=-\sigma^2$. A similar restriction will arise for other values, $\mu*$ = c for example, of μ^*, where c is some constant. Thus, there is a built-in restriction on the choice of the LBLN distribution with respect to the selection of the appropriate model for representing the original LN.

Maximum Likelihood Estimation (MLE)

Let $(X_1,X_2,\ldots X_n)$ be a random sample from a LBLN (μ, σ) distribution. The log-likelihood function $l(\mu, \sigma)$ is then given as

Table 1: Properties of LN (μ, σ) and LBLN (μ, σ)

Property	LN (μ, σ)	LBLN(μ, σ)
Mean	$\exp(\mu+\sigma^2/2)$	$\exp(\mu^*+\sigma^2/2)$
Median	$\exp(\mu)$	$\exp(\mu^*)$
Mode	$\exp(\mu-\sigma^2)$	$\exp(\mu^*-\sigma^2)=\exp(\mu)$
Variance	$\exp(2\mu+\sigma^2)\{\exp(-\sigma^2)-1\}$	$\exp(2\mu^*+\sigma^2)\{\exp(-\sigma^2)-1\}$

Table 2: The Fisher Information Matrix of LN (μ, σ) and LBLN (μ, σ)

Fisher Information Matrix	LN(μ, σ)	LBLN(μ, σ)
	$I_1=\begin{bmatrix}1/\sigma^2 & 0\\ 0 & 2/\sigma^2\end{bmatrix}$	$I_2=\begin{bmatrix}1/\sigma^2 & 2/\sigma\\ 2/\sigma & 4+2/\sigma^2\end{bmatrix}$

$$l(\mu,\sigma)=-n\sum\log x_i-n\log\sigma - \log\sqrt{2\pi}\,\frac{1}{\sigma^2}\sum(\log x_i-\mu-\sigma^2)^2. \tag{5}$$

The solutions to likelihood equations $\frac{\partial l}{\partial\mu}=0$ and $\frac{\partial l}{\partial\sigma}=0$ give the MLEs as:

$$\hat{\mu}=(\sum\log x_i/n)-\hat{\sigma}^2$$

and

$$\hat{\sigma}^2=\frac{1}{n}[\sum(\log x_i)^2-(\sum\log x_i)^2/n]. \tag{6}$$

To study the properties of $(\hat{\mu},\hat{\sigma}^2)$ the transformation $Y=\log X$ is considered in (2). It is known that $Y\sim$ Normal $(\mu+\sigma^2, \sigma)$. This leads to the following estimates corresponding to (6):

$$\hat{\mu}=\bar{Y}+cS^2,$$

where $c=-(n-1)/n$, $\hat{\sigma}^2=\frac{(n-1)}{n}S^2$, and for the sample

$$Y_i=\log X_i, i=1,2,...,n, \bar{Y}=\sum\frac{Y_i}{n}$$

and

$$S^2=\sum\frac{(Y_i-\bar{Y})^2}{(n-1)}. \tag{7}$$

From (7), it is clear that $\hat{\mu}$ is a biased estimator of μ. Further, the distribution of $\hat{\mu}$ cannot be expressed in closed form. Thus, the distributional properties of $\hat{\mu}$, unlike in the case of the original LN distribution, are not readily

available for statistical inference. Therefore, simulations are considered in order to understand the properties of the above defined estimator of μ. In particular, finding the confidence interval (C.I.) for μ is not straightforward. Hence, the bootstrap method for constructing the confidence interval for μ was considered to obtain the results.

Results

To illustrate the use of the methodology introduced above, data from oil field explorations (Meisner & Demirmen, 1981) was analyzed.

Table 3: Sizes of Oil Fields Data
(X = Field Size, Oil (10^6 BBLS), n = 58)

28	26	775	114	31
337	41	113	1328	21
13	455	89	482	70
215	62	58	6.9	154
177	43	33	178	15
22	11	8.1	35	25
170	19	56	42	335
21	50	181	93	75
8.8	29	450	5.9	8.8
49	100	10	8.8	17
12	125	20	8.8	8.8
6.9	25	100		

MLE of the Parameters (μ, σ) of LBLN

The MLE's of μ and σ were obtained (see Table 4) using the formulas in (7). The amount of bias in the estimate of μ can also be estimated using (7).

Table 4: Estimates of μ and σ

Parameter	Estimate	Standard Error
μ	2.0748	0.3729
σ	1.3317	0.1236

Simulations for Studying the Properties of $\hat{\mu}$

As noted, the distribution of $\hat{\mu}$ is not available in closed form (see equation (7)); therefore, to understand the properties of $\hat{\mu}$ simulations were conducted. In particular, the amount of bias in the estimates of values of μ is of interest. For these simulations, different values of n and (μ, σ) were used. The simulation results for n = 20 and certain values of (μ, σ) are shown in Table 5. Results obtained were expected from (7), and show that the absolute value of the bias in the estimate $\hat{\mu}$ of μ increases as σ increases. Results from the simulations for other values of n were not different from those recorded above and therefore for brevity are not included.

Estimation of μ Based on the Mode of LBLN (μ, σ)

The mode M of the LBLN (μ, σ) distribution from Table 1 is given by M = $\exp(\mu)$ which is free of σ^2 and, hence, leads to the formula $\mu = \ln(M)$. This expression can be employed to estimate μ using the sample mode. Note that such estimate of μ, unlike the MLE of μ, does not depend on the estimate of σ^2.

For data presented in Table 3, the estimate of μ using the sample mode is 2.1747. This estimate is comparable with the MLE estimate of 2.0748 (see Table 4); however, because the sample mode is not known to be an efficient estimator of the location parameter it is not considered further.

Bootstrap Estimation of μ

As noted previously, because the distribution of $\hat{\mu}$ is not available in closed form, the nonparametric bootstrap method was used to estimate μ and its related confidence interval; results are shown in Table 6. Based on these results, the 95% and 90% C.I.s can be constructed.

The purpose of the above computations is for illustration, not for comparison of the results obtained herein with those obtained by Meisner and Demirmen (1981). However, the definition of the size-biased (also known as the weighted distribution) version of the LN (μ, σ)

Table 5: Simulation Results for the Properties of $\hat{\mu}$ (# of simulations = 5,000)

μ	σ	Mean of $\hat{\mu}$	MSE	μ	σ	Mean of $\hat{\mu}$	MSE	μ	σ	Mean of $\hat{\mu}$	MSE
-2	0.5	-1.98	0.0186	0	0.5	0.0122	0.0186	1	0.5	1.0111	0.0189
-2	1.0	-1.94	0.1416	0	1.0	0.0548	0.1486	1	1.0	1.0559	0.1417
-2	1.5	-1.89	0.6199	0	1.5	0.1187	0.6086	1	1.5	1.1146	0.5893
-2	2.0	-1.79	1.7518	0	2.0	0.2132	1.7597	1	2.0	1.2008	1.7003

Table 6: Bootstrap Estimate of μ for the Oil Fields Data
(Number of Replications: 3,000)

Summary Statistics	Observed	Bias	Mean	SE
Parameters	2.075	0.03512	2.11	0.265
BCa Percentiles	2.5%	5%	95%	97.5%
Parameters	1.5028	1.5857	2.4702	2.5518

given, and also considered by Meisner and Demirmen, may be useful to some readers.

Let $X \sim f(x;\theta)$. If $w(x) > 0$ is a function of x such that $E[w(X)] < \infty$, then the weighted distribution of X is defined by the pdf

$$g(x;\theta) = w(x) f(x;\theta) / E[w(X)], \tag{8}$$

where $w(x)$ is referred to as the weight function and θ is a scalar or a vector of parameters.

Meisner and Demirmen (1981) assumed that the original distribution of the size of the oil field, denoted by X, is LN (μ, σ). Further, in the exploration of the oil field, the probability of discovering an oil field depends on the size of the oil field. Therefore, for modeling the collected sample data of the oil fields Meisner and Demirmen considered the weighted lognormal (WLN) with weight function $w(x) = x^{\beta}$. Using (8) , the distribution of interest, the WLN (μ, σ) with the pdf is given by

$$g_2(x;\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp[-\frac{1}{2\sigma^2}(\log x - (\mu + \beta\sigma^2))^2], \tag{9}$$

where

$$x > 0,\ -\infty < \mu < \infty,\ \sigma > 0$$

and β may have a known or unknown value.

Meisner and Demirmen (1981) discussed the possible values of β. In particular, they noted that, because the sizes of oil fields change with the exploration period, the values of β could be in a two-sided neighborhood of the value 1. Therefore, this illustration, considering $\beta = 1$, that is assuming the sizes of the oil fields have the LBLN (μ, σ) distribution given by (2.2), is justified. Table 7 shows the estimates of μ for the other values of β; Meisner and Demirmen considered β as a random variable and developed a Bayesian approach for the analysis of these data. To construct the estimates

in Table 7, the modified version of (7) for accommodating β was used.

Table 7: Changes in the Estimate of μ for Values of β

β	0.9	1.0	1.1
μ Estimate	2.25	2.07	1.89

If the estimate 2.1747 of μ is acceptable, then considering that the mode of LBLN (μ, σ) is a function of μ alone, it can used for finding a value of β (a sort of ad hoc estimate of β) by extending Table 5 to include more values of β than may be necessary. In particular, using such a table it can be shown that if $\beta = 0.94$ then $\hat{\mu} = 2.17$ (approximately), which is close to the above estimate 2.1747.

Note that, in view of the unstable behavior of the sample mode, such an estimate should be carefully considered. However, in this analysis, observations show that the assumption of $\beta = 1$ (which is close to the value of $\beta = 0.94$) used for modeling the data from Table 4 has some relevance. Further, it should be noted that other more robust methods exist for locating the sample mode (Bickel & Fruthworth, 2006). The traditional method was used in this study for demonstration purposes.

Conclusion

The length-biased lognormal distribution was introduced along with an application in the analysis of data from oil field explorations. The maximum likelihood estimation of the parameters of the length-biased lognormal was discussed briefly. In particular, the properties of the estimator of μ are not tractable. Therefore, the related properties were studied using simulations. Results presented regarding the modal value of the length-biased lognormal show that the estimation of μ using the sample mode is straightforward, but the efficiency of such an estimator is doubtful. The concepts of weighted lognormal distribution as a generalization of the length-biased lognormal and related modeling problems were also briefly mentioned.

References

Bickel, D. R., & Fruthwirth, R. (2006). On a fast, robust estimator of the mode: Comparison to other robust estimators with applications. *Computational Statistics & Data Analysis*, 50, 3500-3530.

Meisner, J., & Demirmen, F. (1981).The creaming method: a Bayesian procedure to Forecast future oil and gas discoveries in mature exploration provinces. *Journal of the Royal Statistical Society*, *Series A*, *144*, 1-31.

Quin, I. J., Berwick, M., Ashbolt, R., & Dwyer, T. (2002). Quantifying the change of melanoma incidence by Breslow thickness. *Biometrics*, *58*, 665-670.

Yan, P. (2004). Estimation for the infection curves for the spread of severe acute respiratory syndrome (SARS) from a back-calculation approach. *Report of the Modeling and Projection Section*, *Centre for Infectious Disease Prevention & Control*. Population and Public Health Branch Health Canada, 1-17.