Advances in Environmental Protection 环境保护前沿, 2011, 1, 20-23
http://dx.doi.org/10.12677/aep.2011.12004 Published Online December 2011 (http://www.hanspub.org/journal/aep)

Approximate Confidence Limits of the Lognormal Mean from Left-Censored Samples

Jiaqing Xu
College of Sciences, Ningbo University of Technology, Ningbo
Email: zb422@126.com
Received: Oct. 28th, 2011; revised: Nov. 25th, 2011; accepted: Nov. 29th, 2011.

Abstract: This paper discusses the maximum likelihood estimate of population parameters from left-censored samples and obtains the approximate confidence limits of mean as well as the test statictics of hypothesis for log-normal distributions. Some numerical results on implementing the proposed method via example and simulation studies are presented in this paper.

Keywords: Log-Normal Distributions; Left-Censored Samples; Maximum Likelihood Estimate; Approximate Confidence Limits

左删失监测数据下对数正态总体均值的近似置信限

许家清
宁波工程学院理学院，宁波
Email: zb422@126.com
收稿日期：2011 年 10 月 28 日；修回日期：2011 年 11 月 25 日；录用日期：2011 年 11 月 29 日

摘　要：文章讨论了在对数正态总体下，当环境监测数据为左删失时未知参数的极大似然估计，总体均值的渐近置信限和检验统计量，并通过计算机模拟验证了方法的可行性，同时给出了一个实际例子。

关键词：对数正态分布；左删失数据；极大似然估计；渐近置信限

1. 引言

在许多环境监测活动中，大量的总体可认为是服从对数正态分布的，其参数估计等统计推断问题往往将样本数据取对数后利用正态分布的统计推断方法来分析数据，同时经常碰到的问题是由于监测设备的原因，得到的是一组左删失样本数据，即监测设备无法监测到小于某一个值(称为删失点)的数据，一般采用的替代方法是用一组常数代替删失的数据，再用这一完全样本数据来估计均值 μ 和标准差 σ 的极大似然估计。如 Gilliom 和 Helsel(1986)[1]给出了二个简单的替代方法，即用 0 和删失点代替删失的数据。Helsel 和 Hirsch(1988)[2]等还讨论了一些其他的替代方法。另一个利用来自正态总体的删失样本获得未知参数极大似然估计的方法是 Dempster 等(1977)[3]提出的 *EM* 算法。这些算法在讨论估计量的统计性质时比较困难和繁琐，本文利用 Cohen(1956)[4]提出的算法给出均值 μ 和标准差 σ 的极大似然估计计算程序，同时利用大样本理论得到了总体均值的近似置信限和检验统计量。为比较不同地区的环境污染程度提供了统计检验方法。

2. 散失数据的统计推断

2.1. 参数的极大似然估计

假设总体 T 服从对数正态分布 $LN\left(\mu,\sigma^2\right)$，从中获得一样本容量为 N 的左删失样本，记为 $t_1,t_2,\cdots,t_{n_1},t_{n_1+1},\cdots t_{n_1+n_0}$。其中 $t_1,t_2,\cdots,t_{n_1}$ 为左删失的数据，即仅观测到 $t_i<t_0$，$i=1,2,\cdots,n_1$，t_0 为删失点，并无具体观测值，$t_{n_1+1},\cdots t_{n_1+n_0}$ 为实际观测到的样本观测值，且满足 $t_i\geq t_0$，$i=n_1+1,\cdots,n_1+n_0$。令 $x_i=\ln t_i$，则 $x_1,x_2,\cdots x_{n_1},x_{n_1+1},\cdots,x_{n_1+n_0}$ 可看成是来自正态总体

$X\sim N\left(\mu,\sigma^2\right)$ 的一个左删失样本，删失点为 $x_0=\ln t_0$。利用此左删失样本可得似然函数为：

$$\begin{aligned}&L\left(x_1,x_2,\cdots x_n;\mu,\sigma\right)\\&=L\left(\mu,\sigma\right)\\&=\frac{n!}{n_1!n_0!}\left(1-I\left(\theta\right)\right)^{n_i}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=n_1+1}^{n_1+n_0}\left(x_i-\mu\right)^2\right)\end{aligned}$$

其中 $\theta=\dfrac{x_0-\mu}{\sigma}$，$I\left(\theta\right)=\dfrac{1}{\sqrt{2\pi}}\int_{\theta}^{+\infty}\exp\left(-\dfrac{t^2}{2}\right)=1-\Phi\left(\theta\right)$

$\Phi\left(\theta\right)$ 是标准正态分布的分布函数。

为求未知参数 μ 和 σ 的极大似然估计，作变换[1] $y_i=x_{n_1+i}-x_0$ 则对数似然函数可写为

$$\begin{aligned}l\left(\theta,\sigma\right)&=\ln L\left(\mu,\sigma\right)\\&=n_1\ln\left(1-I\left(\theta\right)\right)-n_0\ln\sigma-\frac{1}{2}\sum_{i=1}^{n_0}\left(\theta+\frac{y_i}{\sigma}\right)+c\end{aligned}$$

其中 c 为常数。

似然方程 $\dfrac{\partial l}{\partial\theta}=0$ 和 $\dfrac{\partial l}{\partial\sigma}=0$ 经过整理后有

$$\sigma\left(Y\left(\theta\right)-\theta\right)=a_1\text{ 和 }\sigma^2\left[1-\theta\left(Y\left(\theta\right)-\theta\right)\right]=a_2$$

其中 $a_k=\dfrac{1}{n_0}\sum_{i=1}^{n_0}y_i^k$，$Y\left(\theta\right)=\dfrac{n_1}{n_0}\dfrac{\phi\left(\theta\right)}{\Phi\left(\theta\right)}$，$\phi\left(\theta\right)=\Phi'\left(\theta\right)$ 为标准正态分布的密度函数，在上述似然方程中消去 σ 后得：

$$\frac{a_2}{a_1^2}=\frac{1}{Y\left(\theta\right)-\theta}\left(\frac{1}{Y\left(\theta\right)-\theta}-\theta\right)$$

利用计算机可求的上述方程的根作为未知参数 θ 的极大似然估计，记为 θ，从而可得 μ 和 σ 的极大似然估计

$$\hat{\sigma}=\frac{a_1}{Y\left(\hat{\theta}\right)-\hat{\theta}},\ \hat{\mu}=x_0-\hat{\theta}\hat{\sigma},$$

而总体 T 的均值 $\delta=ET$ 的极大似然估计为

$$\hat{\delta}=\exp\left(\hat{\mu}+\frac{\hat{\sigma}^2}{2}\right)$$

2.2. 近似置信区间

为求得 δ 的近似置信区间，需计算对数似然函数关于 μ 和 σ 的 Hessian 矩阵，记 $\beta=(\mu,\sigma)'$ $\hat{\beta}=(\hat{\mu},\hat{\sigma})'$。由对数似然函数的表达式得：

$$\frac{\partial l}{\partial\mu}=\frac{\partial l}{\partial\theta}\frac{\partial l}{\partial\mu}=-\frac{1}{\sigma}\frac{\partial l}{\partial\theta}=\frac{n_0}{\sigma^2}\left[\bar{x}-\mu-\sigma Y\left(\theta\right)\right]$$

$$\begin{aligned}\frac{\partial l}{\partial\sigma}&=-n_1\frac{\phi\left(\theta\right)\theta}{1-I\left(\theta\right)}\frac{1}{\sigma}-\frac{n_0}{\sigma}+\frac{n_0}{\sigma}\theta^2+\frac{n_0}{\sigma^3}a_2+\frac{2\theta n_0a_1}{\sigma^2}\\&=\frac{n_0}{\sigma^2}\left[\frac{s^2+\left(\bar{x}-\mu\right)^2}{\sigma}-\sigma-\sigma\theta Y\left(\theta\right)\right]\end{aligned}$$

记：

$$\begin{aligned}t_{11}&=\frac{\partial^2l}{\partial\mu^2}=\frac{\partial}{\partial\mu}\left(\frac{\partial l}{\partial\mu}\right)=\frac{1}{\sigma^2}\frac{\partial^2l}{\partial\theta^2}\\&=-\frac{1}{\sigma^2}\left[n_0+n_1\left(Z\left(-\theta\right)\theta+Z^2\left(-\theta\right)\right)\right]\\&=-\frac{1}{\sigma^2}\left[n_0+n_1Z'\left(-\theta\right)\right]\end{aligned}$$

$$\begin{aligned}t_{22}&=\frac{\partial^2l}{\partial\sigma^2}\\&=-\frac{1}{\sigma^2}\left[\frac{3n_0\left(s^2+\left(\bar{x}-\mu\right)^2\right)}{\sigma^2}-n_0\right.\\&\qquad\left.+n_1\theta\left(-2Z\left(-\theta\right)+\theta Z'\left(-\theta\right)\right)\right]\\&=-\frac{1}{\sigma^2}\left[\frac{3n_0\left(s^2+\left(\bar{x}-\mu\right)^2\right)}{\sigma^2}-n_0+n_1\varsigma\left(-\theta\right)\right]\end{aligned}$$

$$\begin{aligned}t_{12}&=\frac{\partial^2l}{\partial\mu\partial\sigma}=-\frac{1}{\sigma^2}\left[\frac{2n_0\left(\bar{x}-\mu\right)}{\sigma}\right.\\&\qquad\left.-n_1\left(Z\left(-\theta\right)-\theta Z'\left(-\theta\right)\right)\right]\\&=-\frac{1}{\sigma^2}\left[\frac{2n_0\left(\bar{x}-\mu\right)}{\sigma}-n_1\lambda\left(-\theta\right)\right]\end{aligned}$$

其中

$$\bar{x}=\frac{1}{n_0}\sum_{i=n_1+1}^{n_1+n_0}x_i=x_0+a_1$$

$$s^2=\frac{1}{n_0}\sum_{i=n_1+1}^{n_1+n_0}\left(x_i-\bar{x}\right)^2=a_2-a_1^2$$

$$Z\left(\theta\right)=\frac{\phi\left(\theta\right)}{I\left(\theta\right)}$$

$$Z'\left(\theta\right)=\frac{\partial Z\left(\theta\right)}{\partial\theta}=Z\left(\theta\right)\left[Z\left(\theta\right)-\theta\right]$$

$$\varsigma\left(\theta\right)=\theta\left(ZZ\left(\theta\right)-\theta+\lambda Z\left(\theta\right)-\theta\right)。$$

得：$T(\beta)=\begin{pmatrix}\dfrac{\partial^2l}{\partial\mu^2} & \dfrac{\partial^2l}{\partial\mu\partial\sigma}\\ \dfrac{\partial^2l}{\partial\mu\partial\sigma} & \dfrac{\partial^2l}{\partial\sigma^2}\end{pmatrix}=\begin{pmatrix}t_{11} & t_{12}\\ t_{12} & t_{22}\end{pmatrix}$

由此得 Fisher 信息阵 $I(\beta)$的一个相合估计为

$$\hat{I}(\beta)=-\frac{1}{n}T(\beta)\Big|_{\beta=\hat{\beta}}=\hat{\Sigma}$$

利用极大似然估计的渐近正态性有：

$$\sqrt{n}\left(\hat{\beta}-\beta\right)\xrightarrow{a.s}N\left(0,\Sigma^{-1}\right)$$

记 $\delta=\exp\left(\mu+\frac{\sigma^2}{2}\right)$ $\hat{\delta}=\exp\left(\hat{\mu}+\frac{\hat{\sigma}^2}{2}\right)$,

$$u_1(\mu,\sigma)=\frac{\partial\delta}{\partial\mu}=\exp\left(\mu+\frac{\sigma^2}{2}\right)$$

$$u_2(\mu,\sigma)=\frac{\partial\delta}{\partial\sigma}=\sigma\exp\left(\mu+\frac{\sigma^2}{2}\right)$$

$$U=\left(u_1(\mu,\sigma),u_2(\mu,\sigma)\right),$$

可以证明[5]：$\sqrt{n}\left(\hat{\delta}-\delta\right)\xrightarrow{a.s}N\left(0,U\Sigma^{-1}U'\right)$

其中方差$U\Sigma^{-1}U'$的相合估计为$\hat{U}\hat{\Sigma}^{-1}\hat{U}'$，$\hat{U}=\left(u_1(\hat{\mu},\hat{\sigma}),u_2(\hat{\mu},\hat{\sigma})\right)$。由此得到$\delta$的$100(1-\alpha)\%$近似双侧置信限，置信下限，置信上限分别为：

$$\hat{\delta}-z_{\frac{\alpha}{2}}\sqrt{\left(\hat{U}\hat{\Sigma}^{-1}\hat{U}'\right)/n}\ ,\quad \hat{\delta}+z_{\frac{\alpha}{2}}\sqrt{\left(\hat{U}\hat{\Sigma}^{-1}\hat{U}'\right)/n}$$

$$\hat{\delta}-z_{\alpha}\sqrt{\left(\hat{U}\hat{\Sigma}^{-1}\hat{U}'\right)/n}\ ,\quad \hat{\delta}+z_{\alpha}\sqrt{\left(\hat{U}\hat{\Sigma}^{-1}\hat{U}'\right)/n}$$

其中z_α为$N(0,1)$的α上侧分位数。

考虑单边假设检验问题$H_0:\hat{\delta}<\delta_0$，由上述的近似分布不难得到检验统计量为：

$$Z=\frac{\hat{\delta}-\delta}{\sqrt{\left(\hat{U}\hat{\Sigma}^{-1}\hat{U}'\right)/n}}\sim N(0,1)$$

由此可计算检验的P值。

Table 1. Simulation results
表 1. 模拟结果

t_0	n	emu	emsigma	emdeta	emdetal	emdetau	rcov	n1
5.6	30	1.999	0.196	7.534	6.999	8.07	0.936	4
	50	1.999	0.198	7.536	7.117	7.955	0.942	7
	80	2	0.198	7.537	7.205	7.869	0.942	6
	100	2	0.199	7.537	7.24	7.834	0.942	8
5.8	30	1.999	0.196	7.533	6.996	8.07	0.94	4
	50	2	0.198	7.538	7.118	7.958	0.947	6
	80	2	0.199	7.537	7.204	7.87	0.943	7
	100	1.999	0.199	7.535	7.237	7.834	0.942	10
6	30	1.999	0.196	7.531	6.991	8.071	0.936	5
	50	2	0.198	7.537	7.116	7.957	0.947	11
	80	1.999	0.199	7.532	7.198	7.866	0.944	13
	100	1.999	0.199	7.533	7.234	7.831	0.946	18
6.2	30	1.999	0.197	7.526	6.983	8.069	0.935	6
	50	1.999	0.198	7.532	7.109	7.954	0.939	10
	80	2	0.199	7.539	7.203	7.875	0.948	11
	100	2	0.2	7.538	7.237	7.838	0.951	19
6.6	30	1.999	0.195	7.538	6.987	8.089	0.937	9
	50	1.999	0.197	7.536	7.106	7.966	0.938	16
	80	1.999	0.199	7.532	7.19	7.875	0.945	21
	100	1.999	0.199	7.536	7.23	7.842	0.947	29
6.8	30	2.004	0.19	7.566	7.021	8.111	0.927	9
	50	2.001	0.196	7.547	7.113	7.981	0.939	16
	80	1.999	0.198	7.536	7.19	7.883	0.946	24
	100	2	0.199	7.537	7.226	7.847	0.949	33
7	30	2.019	0.172	7.673	7.162	8.184	0.865	16
	50	2.001	0.184	7.607	7.188	8.027	0.91	14
	80	2.004	0.193	7.566	7.22	7.912	0.938	27
	100	2.002	0.196	7.553	7.24	7.866	0.948	39

3. 实验结果

3.1. 仿真实验

假设总体 T服从对数正态分布 $LN(\mu,\sigma^2)$，删失点为 t_0，先由计算机模拟 n 个来自上述总体的样本数据, 根据删失点 t_0 得到一组左删失数据，利用 R 软件计算参数 μ,σ 的极大似然估计 $\hat{\mu},\hat{\sigma}$，以及相应的近似置信限。为验证置信限的精确度，对不同的 n 和 t_0 将上述过程重复 10,000 次，计算置信限覆盖真值的比例。过程如下：

取真值 $\mu=2,\sigma=0.2$。$\delta=7.538325$。在 $t_0=$ 5.6, 5.8, 6.0, 6.2,6.6, 6.8, 7.0 的每个值下样本容量 n 分别取 30, 50, 80, 100。有关结果见表 1，其中 t_0 为左删失点，n1 为删失数据个数，emu，emsigma，emdeta，emdetal，emdetau 分别为未知参数 μ,σ,δ 的点估计及 δ 的置信下限和上限在 10,000 次模拟中估计的平均值，$r\,\mathrm{cov}$ 为覆盖率。置信水平为 95%。

模拟的结论显示，当删失点不是很大，即与总体均值相比较小时，样本量 30 以上的置信限覆盖率非常接近预设的置信水平 95%，表明渐近置信限有较好的精度。同时置信区间的长度随着样本量的增加而减小。显然在实际的监测中，删失点不会离总体均值较近，所以方法有较强的实际可操作性。

3.2. 真实数据

以下是一组有关地下水环境污染的左删失监测数据[6]：

<10, <10, <10, 18, <10, 12, 10, 11, 11, 19, <10, <10, <10, 10, 10, 10, 10, 10, <10, 10, <10, 10, <10, 10, <10, 10, 10, 20, 20, <10, 20, 20, 20, <10, 10, 20, 620, 40, 50, 33, 10, 20, 10, 10, 10, 30, 20, 10, 20, 20, 20, <10, 20, 23, 17, 10, <10, 10, 20, 29, 20, <10, 10, <10, 10, <10, <10

根据长期的历史观察，可以认为数据是来自对数正态总体的，其中样本量为 67，删失点为 10，通过计算可得总体参数 μ 和 σ 的点估计分别为 2.48 和 0.79，置信水平为 95%的置信区间分别为[2.28, 2.68]和[0.63, 0.95]，总体均值的点估计为 16.358，置信水平为 95%的置信区间为[12.802, 19.913]。

4. 结论

在对数正态总体下，当监测数据为左删失时，参数的极大似然估计无法显式表示，本文利用 Cohen 的参数估计方法给出了参数的极大似然估计计算方法和程序，同时进一步给出了总体均值的渐近置信限，用同样方法可构造总体方差等其它数字特征的渐近置信限，对不同地区或不同时期的监测数据比较提供了统计检验方法，有关计算的程序可向作者索取.

5. 致谢

研究工作受到宁波市自然科学基金(编号 2011A610166)资助。

参考文献 (References)

[1] R. J. Gilliom, D. R. Helsel. Estimation of distributional parameters for censored trace level water quality data: 1. Estimation techniques. Water Resources Research, 1986, 22(2): 135-146.

[2] D. R. Helsel, R. M. Hirsch. Statistical methods in water resources. New York: Elsevier, 1988.

[3] A. P. Dempster, N. M. Laird and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. The Journal of Royal Statistical Society, 1977, 39(1): 1-38.

[4] C. Cohen Jr. On the solution of estimating equations for truncated and censored samples from normal populations. Biometrika, 1957, 44(1-2): 225-236.

[5] R. J. Serfling. Approximation theorems of mathematical statistics. New York: John Wiley & Sons, Inc., 1980.

[6] A. El-Makarim, A. Aboueissa. Maximum likelihood estimators of population parameters from multiply censored samples. Environmetrics, 2009, 20: 312-330.

