应用概率统计　第二十四卷
第四期　2008年8月

Chinese Journal of Applied Probability
and Statistics Vol.24 No.4 Aug. 2008

基于数据分组与右删失情形下对数正态分布的参数估计 *

刘　欣　　陈　惠　　费鹤良

(上海师范大学数理信息学院, 上海, 200234)

摘　要

本文研究了对数正态分布数据在分组与删失情形下参数的估计问题. 一是给出未知参数的极大似然估计存在且唯一的充要条件. 二是利用EM算法对参数值进行了估计.

关键词: 对数正态分布, 分组右删失, 极大似然估计, EM算法.

学科分类号: O213.2.

§1. 引　　言

对数正态分布是可靠性和寿命试验中广泛应用的模型之一, 已有很多著作讨论了对数正态分布在完全样本、Type-I和Type-II截尾样本下的参数估计, 可见Laweless (1983), Bain和Engelhardt (1991)等. 对于对数正态分布在分组数据下的MLE, 王静(2003)给出了MLE存在且唯一的充要条件, 以及证明了MLE具有相合性与收敛速度服从重对数律. 现在我们讨论比分组数据更为一般的情况, 就是在每个分组的右端再去掉一些未失效的产品, 这就使数据成了分组与右删失的情形. 对于数据分组和右删失情形下Weibull分布参数的MLE已有了一定的研究. Liu (2001)讨论了Weibull分布极大似然估计存在且唯一的充要条件. 本文我们讨论对数正态分布在分组与右删失情形下未知参数的MLE存在且唯一的充要条件, 并且用EM算法给出了参数的估计值.

设某产品的寿命服从对数正态分布, 其分布函数为

$$F(x;\mu,\sigma)=\int_0^x \frac{1}{\sqrt{2\pi}\sigma t}\exp\Big[-\frac{(\ln t-\mu)^2}{2\sigma^2}\Big]\mathrm{d}t,\qquad x>0. \tag{1.1}$$

这里μ,σ都未知且$\sigma>0$, $\mu\in(-\infty,+\infty)$.

现取n个产品进行寿命试验, 获得数据如下: 将$[0,+\infty)$, 分成$N+1$个区间, 前N个区间记作$[T_{i-1},T_i)$, 其中$i=1,\cdots,N$, $0=T_0<T_1<\cdots<T_N<\infty$. d_i为落入到第i个区间$[T_{i-1},T_i)$中的失效产品数, 并记在T_i时刻截尾产品数为λ_i. 这样我们有$\sum\limits_{i=1}^{N}(d_i+\lambda_i)=n$.

*国家自然科学基金(10571057, 10671129)和教育部高校博士点专项科研基金项目(20060270002).
本文2005年4月29日收到, 2005年10月13日收到修改稿.

似然函数为

$$\begin{aligned} & L(\mu,\sigma;d,\lambda) \\ = & \prod_{i=1}^{N}\Big(\int_{T_{i-1}}^{T_i}\frac{1}{\sqrt{2\pi}\sigma t}\exp\Big[-\frac{(\ln t-\mu)^2}{2\sigma^2}\Big]\mathrm{d}t\Big)^{d_i}\Big(1-\int_0^{T_i}\frac{1}{\sqrt{2\pi}\sigma t}\exp\Big[-\frac{(\ln t-\mu)^2}{2\sigma^2}\Big]\mathrm{d}t\Big)^{\lambda_i}, \end{aligned}$$

其中$d=(d_1,d_2,\cdots,d_N)$, $\lambda=(\lambda_1,\lambda_2,\cdots,\lambda_N)$.

作变换$x=(\ln t-\mu)/\sigma$, 令$\alpha=1/\sigma$, $\beta=\mu/\sigma$, 则$L(\mu,\sigma)=l(\alpha,\beta)$,

$$l(\alpha,\beta)=\prod_{i=1}^{N}\Big[\Big(\int_{\alpha\ln T_{i-1}-\beta}^{\alpha\ln T_i-\beta}g(x)\mathrm{d}x\Big)^{d_i}\Big(1-\int_0^{\alpha\ln T_i-\beta}g(x)\mathrm{d}x\Big)^{\lambda_i}\Big]. \tag{1.2}$$

对数似然函数为

$$h(\alpha,\beta)=\ln l(\alpha,\beta)=\sum_{i=1}^{N}\Big[d_i\ln\int_{\alpha\ln T_{i-1}-\beta}^{\alpha\ln T_i-\beta}g(x)\mathrm{d}x+\lambda_i\ln\Big(1-\int_0^{\alpha\ln T_i-\beta}g(x)\mathrm{d}x\Big)\Big], \tag{1.3}$$

其中$g(x)=(1/\sqrt{2\pi})\cdot\exp\{-x^2/2\}$. 这样我们将求$L(\mu,\sigma)$的最大值点存在且唯一的充要条件转化为了求$h(\alpha,\beta)$的最大值点存在且唯一的充要条件, 两者是一致的.

§2. MLE存在且唯一的充要条件

如果在整个寿命试验过程中没有一个产品失效, 那么未知参数(μ,σ)的MLE是不存在的. 所以我们假设至少有一个产品失效. 类似于文献[4]的方法可将数据的所有情形分成9种情况讨论:

情形1: $d_1>0$, $d_2=\cdots=d_N=0$, $d_1+\lambda_1<n$.

情形2: $d_1>0$, $d_2=\cdots=d_N=0$, $d_1+\lambda_1=n$.

情形3: $d_1>0$, $d_2>0$, $d_3=\cdots=d_N=\lambda_2=\cdots=\lambda_N=0$.

情形4: $d_1>0$, $d_2>0$, $d_3=\cdots=d_N=0$, 并且存在一个整数$j\geqslant 2$, 使得$\lambda_j>0$.

情形5: $d_1>0$, 并且存在一个整数$j>2$, 使得$d_j>0$.

情形6: $d_1=\cdots=d_{N-1}=0$, $d_N>0$.

情形7: $d_1=0$, 存在整数i_1,i_2, $(1<i_1<i_2\leqslant N)$, 使得$d_{i_1}>0$, $\lambda_{i_2}>0$.

情形8: $d_1=0$, 存在整数i_1,i_2, $(1<i_1<i_1+1<i_2\leqslant N)$, 使得$d_{i_1}>0$, $d_{i_2}>0$, 但是$\lambda_{i_1+1}=\cdots=\lambda_N=0$.

情形9: $d_1=0$存在整数i_1, $(1<i_1<N)$, 使得$d_{i_1}>0$, $d_{i_1+1}>0$, 而且$d_{i_1+2}=\cdots=d_N=\lambda_{i_1+1}=\cdots=\lambda_N=0$.

2.1 若干引理的介绍

引理 2.1[5] 设$g(x)=(1/\sqrt{2\pi})\cdot\exp\{-x^2/2\}$, $G\triangleq G(u,v)=\int_u^v g(x)\mathrm{d}x$, $(-\infty<u<v<+\infty)$, 则$\ln G$的Hessian矩阵

$$H(u,v)=\begin{pmatrix}\dfrac{\partial^2\ln G}{\partial u^2} & \dfrac{\partial^2\ln G}{\partial u\partial v}\\ \dfrac{\partial^2\ln G}{\partial u\partial v} & \dfrac{\partial^2\ln G}{\partial v^2}\end{pmatrix}$$

是负定的.

引理 2.2[5] 设$g(x)=(1/\sqrt{2\pi})\cdot\exp\{-x^2/2\}$, $-\infty<c<d<+\infty$, $\alpha\in(0,+\infty)$, $\beta\in(-\infty,+\infty)$,

$$A_1=A_1(\alpha,\beta)=\ln\int_{\alpha c-\beta}^{\alpha d-\beta}g(x)\mathrm{d}x,$$
$$A_2=A_2(\alpha,\beta)=\ln\int_{-\infty}^{\alpha d-\beta}g(x)\mathrm{d}x,$$
$$A_3=A_3(\alpha,\beta)=\ln\int_{\alpha c-\beta}^{+\infty}g(x)\mathrm{d}x,$$

则A_1的Hessian矩阵$H_1(\alpha,\beta)$是负定的, A_2和A_3的Hessian矩阵$H_2(\alpha,\beta)$和$H_3(\alpha,\beta)$是半负定的.

引理 2.3 设$h(\alpha,\beta)$由(1.3)确定, d_i,λ_i, $1\leqslant i\leqslant N$如前定义, 若下列条件:

条件1: $d_1>0$, $d_2>0$, $d_3=\cdots=d_N=0$, 并且存在一个整数$j\geqslant 2$, 使得$\lambda_j>0$.

条件2: $d_1>0$, 并且存在一个整数$j>2$, 使得$d_j>0$.

条件3: $d_1=0$, 存在整数i_1,i_2, $(1<i_1<i_2\leqslant N)$, 使得$d_{i_1}>0$, $\lambda_{i_2}>0$.

条件4: $d_1=0$, 存在整数i_1,i_2, $(1<i_1<i_1+1<i_2\leqslant N)$, 使得$d_{i_1}>0$, $d_{i_2}>0$, 但是$\lambda_{i_1+1}=\cdots=\lambda_N=0$.

任何一条满足, 则$h(\alpha,\beta)$的Hessian矩阵$\widetilde{H}(\alpha,\beta)$在$(0,+\infty)\times(-\infty,+\infty)$上处处负定.

证明: 令$a_i=\ln T_i$, 则$-\infty\equiv a_0<a_1\cdots<a_N<+\infty$. 设

$$h_d^{(i)}(\alpha,\beta)=\ln\int_{\alpha\ln T_{i-1}-\beta}^{\alpha\ln T_i-\beta}g(x)\mathrm{d}x=\ln\int_{\alpha a_{i-1}-\beta}^{\alpha a_i-\beta}g(x)\mathrm{d}x,$$
$$h_\lambda^{(i)}(\alpha,\beta)=\ln\Big\{1-\int_{\alpha\ln T_{i-1}-\beta}^{\alpha\ln T_i-\beta}g(x)\mathrm{d}x\Big\}=\ln\Big\{1-\int_{\alpha a_{i-1}-\beta}^{\alpha a_i-\beta}g(x)\mathrm{d}x\Big\},$$

$i=1,2,\cdots,N$. 分别记$h_d^{(i)}(\alpha,\beta)$与$h_\lambda^{(i)}(\alpha,\beta)$的Hessian矩阵为$H_d^{(i)}(\alpha,\beta)$与$H_\lambda^{(i)}(\alpha,\beta)$. 由(1.3)式知

$$\widetilde{H}(\alpha,\beta)=\sum_{i=1}^{N}(d_iH_d^{(i)}(\alpha,\beta)+\lambda_iH_\lambda^{(i)}(\alpha,\beta)).$$

由引理2.2知, $H_d^{(i)}(\alpha,\beta)$, $i=2,\cdots,N$都是负定的; $H_d^{(1)}(\alpha,\beta)$与$H_\lambda^{(i)}(\alpha,\beta)$, $i=1,\cdots,N$都是半负定的, 从而只要至少存在一个i, $2\leqslant i\leqslant N$使$d_i>0$, $\widetilde{H}(\alpha,\beta)$就是负定的.

若条件1成立, 有$d_2>0$, 所以$\widetilde{H}(\alpha,\beta)$是负定的.

若条件2成立, 有$j>2$, 使$d_j>0$, 所以$\widetilde{H}(\alpha,\beta)$是负定的.

若条件3成立, 有$1<i_1<i_2\leqslant N$, 使$d_{i_1}>0$, $\lambda_{i_2}>0$, 所以$\widetilde{H}(\alpha,\beta)$是负定的.

若条件4成立, 有$1<i_1<i_2\leqslant N$, 使$d_{i_1}>0$, $d_{i_2}>0$, 所以$\widetilde{H}(\alpha,\beta)$是负定的.

从而上述条件中任何一条满足, $\widetilde{H}(\alpha,\beta)$就是负定的. □

引理 2.4 设$f(\theta)$是Θ上有二阶连续偏导数的实值函数, Θ是R^k中的凸开集. 又设梯度向量∇f在Θ中至少有一处为零, $f(\theta)$的Hessian矩阵$H(\theta)$在Θ中处处负定, 则$f(\theta)$在Θ上严格凹, 在Θ中有唯一的最大值点, 而且没有其它的极值点和稳定点.

2.2 MLE存在且唯一的充要条件

定理 2.1 在分组与右删失数据情形下, 为了对数正态分布参数的最大似然估计存在且唯一, 必须且只需下列条件之一满足.

条件1: $d_1>0$, $d_2>0$, $d_3=\cdots=d_N=0$, 并且存在一个整数$j\geqslant 2$, 使得$\lambda_j>0$.

条件2: $d_1>0$, 并且存在一个整数$j>2$, 使得$d_j>0$.

条件3: $d_1=0$, 存在整数i_1,i_2, $(1<i_1<i_2\leqslant N)$, 使得$d_{i_1}>0$, $\lambda_{i_2}>0$.

条件4: $d_1=0$, 存在整数i_1,i_2, $(1<i_1<i_1+1<i_2\leqslant N)$, 使得$d_{i_1}>0$, $d_{i_2}>0$, 但是$\lambda_{i_1+1}=\cdots=\lambda_N=0$.

条件1, 2, 3, 4分别是上述9种情形中的4, 5, 7, 8.

证明: 充分性: 由[3]知, 存在有界闭区域$[\alpha_1,\alpha_2]\times[\beta_1,\beta_2]$, 使$h(\alpha,\beta)$在其上存在最大值点$(\widehat{\alpha},\widehat{\beta})$且$(\widehat{\alpha},\widehat{\beta})$也是$h(\alpha,\beta)$在$D=(0,+\infty)\times(-\infty,+\infty)$上的最大值点. 由于$(\widehat{\alpha},\widehat{\beta})$是$D$之内点, 故梯度向量$\nabla h$在$(\widehat{\alpha},\widehat{\beta})$为零. 根据引理2.3, $h(\alpha,\beta)$的Hessian矩阵是负定的. 再利用引理2.4知$h(\alpha,\beta)$的最大值点存在且唯一. 充分性证毕.

下证必要性, 只要分别对情况1, 3, 6, 9, 2证明其MLE不存在或不唯一即可.

情形1:

$$\begin{aligned} l(\alpha,\beta) &= [\Phi(t_1)]^{d_1}\prod_{i=1}^{N}[1-\Phi(t_i)]^{\lambda_i} < [\Phi(t_1)]^{d_1}\prod_{i=1}^{N}[1-\Phi(t_1)]^{\lambda_i} \\ &= [\Phi(t_1)]^{d_1}[1-\Phi(t_1)]^{n-d_1} \leqslant \Big(\frac{d_1}{n}\Big)^{d_1}\Big(\frac{n-d_1}{n}\Big)^{n-d_1}. \end{aligned}$$

其中, $\Phi(t)$为标准正态分布的分布函数. $t_i=\alpha\ln T_i-\beta$. 取$\beta=\alpha\ln T_1-P_{d_1/n}$, $P_{d_1/n}$为标准正态分布的d_1/n分位数, 即$\Phi(P_{d_1/n})=d_1/n$. 则$\alpha\ln T_i-\beta=\alpha\ln T_i-(\alpha\ln T_1-P_{d_1/n})=$

$\alpha\ln(T_i/T_1)+P_{d_1/n}\longrightarrow P_{d_1/n}$, $(\alpha\to 0^+)$. 此时

$$\begin{aligned}\lim_{\alpha\to 0^+} l(\alpha,\beta) &= \lim_{\alpha\to 0^+}\Big\{[\Phi(t_1)]^{d_1}\prod_{i=1}^{N}[1-\Phi(t_i)]^{\lambda_i}\Big\}\\ &= [\Phi(P_{d_1/n})]^{d_1}\lim_{\alpha\to 0^+}\prod_{i=1}^{N}\Big[1-\Phi\Big(\alpha\ln\frac{T_i}{T_1}+P_{d_1/n}\Big)\Big]^{\lambda_i}\\ &= \Big(\frac{d_1}{n}\Big)^{d_1}\prod_{i=1}^{N}[1-\Phi(P_{d_1/n})]^{\lambda_i}=\Big(\frac{d_1}{n}\Big)^{d_1}\Big(\frac{n-d_1}{n}\Big)^{n-d_1}.\end{aligned}$$

所以MLE不存在.

情形3: 易知$d_1+d_2+\lambda_1=n$,

$$\begin{aligned}l(\alpha,\beta) &= [\Phi(t_1)]^{d_1}[\Phi(t_2)-\Phi(t_1)]^{d_2}[1-\Phi(t_1)]^{\lambda_1}\\ &< [\Phi(t_1)]^{d_1}[1-\Phi(t_1)]^{d_2+\lambda_1}\leqslant\Big(\frac{d_1}{n}\Big)^{d_1}\Big(\frac{n-d_1}{n}\Big)^{n-d_1}.\end{aligned}$$

取$\beta=\alpha\ln T_1-P_{d_1/n}$, 则$\alpha\ln T_1-\beta=P_{d_1/n}$, $\alpha\ln T_2-\beta=\alpha\ln(T_2/T_1)+P_{d_1/n}\longrightarrow+\infty$, $(\alpha\to+\infty)$. 此时

$$\begin{aligned}\lim_{\alpha\to+\infty} l(\alpha,\beta) &= \lim_{\alpha\to+\infty}\Big\{[\Phi(P_{d_1/n})]^{d_1}\Big[\Phi\Big(\alpha\ln\frac{T_2}{T_1}+P_{d_1/n}\Big)-\Phi(P_{d_1/n})\Big]^{d_2}[1-\Phi(P_{d_1/n})]^{\lambda_1}\Big\}\\ &= \Big[\frac{d_1}{n}\Big]^{d_1}\Big[1-\Phi(P_{d_1/n})\Big]^{d_1}=\Big(\frac{d_1}{n}\Big)^{d_1}\Big(\frac{n-d_1}{n}\Big)^{n-d_1}.\end{aligned}$$

所以MLE不存在.

情形6: 若$\lambda_N\neq 0$, 则

$$\begin{aligned}l(\alpha,\beta) &= [\Phi(t_N)-\Phi(t_{N-1})]^{d_N}\prod_{i=1}^{N}[1-\Phi(t_i)]^{\lambda_i}\\ &< [\Phi(t_N)]^{d_N}[1-\Phi(t_N)]^{\lambda_N}\\ &= \frac{d_N^{d_N}\lambda_N^{\lambda_N}}{(d_N+\lambda_N)^{d_N+\lambda_N}}.\end{aligned}$$

取$\beta=\alpha\ln T_N-P_{d_N/(d_N+\lambda_N)}$, 则$\alpha\ln T_N-\beta=P_{d_N/(d_N+\lambda_N)}$, $\alpha\ln T_i-\beta=\alpha\ln(T_i/T_N)+P_{d_N/(d_N+\lambda_N)}\longrightarrow-\infty$, $(\alpha\to+\infty)$. 此时

$$\begin{aligned}\lim_{\alpha\to+\infty} l(\alpha,\beta) &= \lim_{\alpha\to+\infty}\Big\{\Big[\Phi(P_{d_N/(d_N+\lambda_N)})-\Phi\Big(\alpha\ln\frac{T_{N-1}}{T_N}+P_{d_N/(d_N+\lambda_N)}\Big)\Big]^{d_N}\\ &\quad\times\prod_{i=1}^{N}\Big[1-\Phi\Big(\alpha\ln\frac{T_i}{T_N}+P_{d_N/(d_N+\lambda_N)}\Big)\Big]^{\lambda_i}\Big\}\\ &= [\Phi(P_{d_N/(d_N+\lambda_N)})]^{d_N}[1-\Phi(P_{d_N/(d_N+\lambda_N)})]^{\lambda_N}\\ &= \frac{d_N^{d_N}\lambda_N^{\lambda_N}}{(d_N+\lambda_N)^{d_N+\lambda_N}}.\end{aligned}$$

若$\lambda_N=0$, 则$l(\alpha,\beta)=[\Phi(t_N)-\Phi(t_{N-1})]^{d_N}\prod\limits_{i=1}^{N}[1-\Phi(t_i)]^{\lambda_i}<1$. 取$\beta=\alpha\ln T^*$, $T_{N-1}<T^*<T_N$, 则$t_i=\alpha\ln T_i-\beta=\alpha\ln(T_i/T^*)\longrightarrow-\infty$ $(i\leq N-1)$或$\longrightarrow+\infty$ $(i=N)$ $(\alpha\to+\infty)$. 此时

$$\begin{aligned}\lim_{\alpha\to+\infty}l(\alpha,\beta) &= \lim_{\alpha\to+\infty}\left\{\left[\Phi\Big(\alpha\ln\frac{T_N}{T^*}\Big)-\Phi\Big(\alpha\ln\frac{T_{N-1}}{T^*}\Big)\right]^{d_N}\times\prod_{i=1}^{N-1}\left[1-\Phi\Big(\alpha\ln\frac{T_i}{T^*}\Big)\right]^{\lambda_i}\right\}\\ &= 1.\end{aligned}$$

所以MLE不存在.

情形9:

$$\begin{aligned}l(\alpha,\beta) &= \prod_{j=1}^{i_1}[1-\Phi(t_i)]^{\lambda_j}\prod_{j=i_1}^{i_1+1}[\Phi(t_i)-\Phi(t_{i-1})]^{d_j}\\ &< [1-\Phi(t_{i_1})]^{\lambda_{i_1}}[\Phi(t_{i_1})]^{d_{i_1}}[1-\Phi(t_{i_1})]^{d_{i_1+1}}\\ &= [\Phi(t_{i_1})]^{d_{i_1}}[1-\Phi(t_{i_1})]^{\lambda_{i_1}+d_{i_1+1}}\\ &\leqslant \frac{{d_{i_1}}^{d_{i_1}}(\lambda_{i_1}+d_{i_1+1})^{\lambda_{i_1}+d_{i_1+1}}}{(d_{i_1}+\lambda_{i_1}+d_{i_1+1})^{d_{i_1}+\lambda_{i_1}+d_{i_1+1}}}.\end{aligned}$$

取$\beta=\alpha\ln T_{i_1}-P_{(\lambda_{i_1}+d_{i_1+1})/(d_{i_1}+\lambda_{i_1}+d_{i_1+1})}$, 则$\alpha\ln(T_j/T_{i_1})+P_{(\lambda_{i_1}+d_{i_1+1})/(d_{i_1}+\lambda_{i_1}+d_{i_1+1})}\longrightarrow-\infty$ $(1\leqslant j\leqslant i_1-1)$或$\longrightarrow+\infty$ $(j=i_1+1)$ $(\alpha\to+\infty$, 此时

$$\begin{aligned}\lim_{\alpha\to+\infty}l(\alpha,\beta) &= \lim_{\alpha\to+\infty}\Bigg\{\prod_{j=1}^{i_1}\left[1-\Phi\Big(\alpha\ln\frac{T_i}{T_{i_1}}+P_{(\lambda_{i_1}+d_{i_1+1})/(d_{i_1}+\lambda_{i_1}+d_{i_1+1})}\Big)\right]^{\lambda_j}\\ &\quad\times\prod_{j=i_1}^{i_1+1}\left[\Phi\Big(\alpha\ln\frac{T_i}{T_{i_1}}+P_{(\lambda_{i_1}+d_{i_1+1})/(d_{i_1}+\lambda_{i_1}+d_{i_1+1})}\Big)\right.\\ &\quad\left.-\Phi\Big(\alpha\ln\frac{T_{i-1}}{T_{i_1}}+P_{(\lambda_{i_1}+d_{i_1+1})/(d_{i_1}+\lambda_{i_1}+d_{i_1+1})}\Big)\right]^{d_j}\Bigg\}\\ &= \lim_{\alpha\to+\infty}\Bigg\{\prod_{j=1}^{i_1-1}\left[1-\Phi\Big(\alpha\ln\frac{T_i}{T_{i_1}}+P_{(\lambda_{i_1}+d_{i_1+1})/(d_{i_1}+\lambda_{i_1}+d_{i_1+1})}\Big)\right]^{\lambda_j}\\ &\quad\times\left[1-\Phi\big(P_{(\lambda_{i_1}+d_{i_1+1})/(d_{i_1}+\lambda_{i_1}+d_{i_1+1})}\big)\right]^{\lambda_{i_1}}\left[\Phi\big(P_{(\lambda_{i_1}+d_{i_1+1})/(d_{i_1}+\lambda_{i_1}+d_{i_1+1})}\big)\right.\\ &\quad\left.-\Phi\Big(\alpha\ln\frac{T_{i-1}}{T_{i_1}}+P_{(\lambda_{i_1}+d_{i_1+1})/(d_{i_1}+\lambda_{i_1}+d_{i_1+1})}\Big)\right]^{d_{i_1}}\\ &\quad\times\left[\Phi\Big(\alpha\ln\frac{T_{i_1+1}}{T_{i_1}}+P_{(\lambda_{i_1}+d_{i_1+1})/(d_{i_1}+\lambda_{i_1}+d_{i_1+1})}\Big)\right.\\ &\quad\left.-\Phi\big(P_{(\lambda_{i_1}+d_{i_1+1})/(d_{i_1}+\lambda_{i_1}+d_{i_1+1})}\big)\right]^{d_{i_1+1}}\Bigg\}\\ &= \frac{{d_{i_1}}^{d_{i_1}}(\lambda_{i_1}+d_{i_1+1})^{\lambda_{i_1}+d_{i_1+1}}}{(d_{i_1}+\lambda_{i_1}+d_{i_1+1})^{d_{i_1}+\lambda_{i_1}+d_{i_1+1}}}.\end{aligned}$$

所以MLE不存在.

情形2:

$$l(\alpha,\beta)=[\Phi(t_1)]^{d_1}[1-\Phi(t_1)]^{\lambda_1}\leqslant\Big(\frac{d_1}{d_1+\lambda_1}\Big)^{d_1}\Big(1-\frac{d_1}{d_1+\lambda_1}\Big)^{\lambda_1}=\frac{d_1^{d_1}\lambda_1^{\lambda_1}}{(d_1+\lambda_1)^{d_1+\lambda_1}}.$$

此时考虑选取α,β使$\Phi(t_1)=\int_{-\infty}^{\alpha\ln T_1}g(x)\mathrm{d}x=d_1/(d_1+\lambda_1)$, 就有

$$l(\alpha,\beta)=\frac{d_1^{d_1}\lambda_1^{\lambda_1}}{(d_1+\lambda_1)^{d_1+\lambda_1}}$$

成立, 而这样α,β的有无穷多组, 所以此时MLE存在但不唯一. □

§3. EM算法

我们在第2节中讨论了MLE存在且唯一的充要条件. 虽然似然方程有可能通过直接计算(如用牛顿法)求解, 但EM算法更方便高效, 所以本节介绍用EM算法求得参数的估计. 记X_j', $j=1,\cdots,n$为独立同分布随机变量, 其分布函数为(1.1). 作变换$X=\ln X'$, 记$a_i=\ln T_i$, $i=1,\cdots,N$. 则X_j, $i=1,\cdots,n$独立同分布于正态分布$N(\mu,\sigma^2)$. 它们分别落入区间$[a_{i-1},a_i)$或在a_i时刻被截尾, 即落入区间$[a_i,+\infty)$, 我们只能观测到落在区间$[a_{i-1},a_i)$的X_j的个数d_i及在a_i被截尾的个数λ_i, 其中$i=1,\cdots,N$, $-\infty=a_0<a_1<\cdots<a_N<+\infty$. 记随机变量$X_j$的全体为$X$, 其总数为$n$, 观测结果为$Y$, X_{ih},X_{il}分别为落入区间$[a_{i-1},a_i)$与在a_i被截尾的随机变量.

E步: 注意到这样一个事实, X实际上已经包含了Y的所有信息, 所以有$p(\mu,\sigma^2|X,Y)=p(\mu,\sigma^2|X)$, 由正态分布的密度函数可以得到

$$\begin{aligned}\log p(\mu,\sigma^2|X) &= \sum_{i=1}^{N}d_i\log\Big[\frac{1}{\sqrt{2\pi}\sigma}\exp\Big[-\frac{(X_{ih}-\mu)^2}{2\sigma^2}\Big]\Big]\\ &\quad+\sum_{i=1}^{N}\lambda_i\log\Big[\frac{1}{\sqrt{2\pi}\sigma}\exp\Big[-\frac{(X_{il}-\mu)^2}{2\sigma^2}\Big]\Big]\\ &= \sum_{i=1}^{N}(d_i+\lambda_i)(-\log\sqrt{2\pi}\sigma)-\sum_{i=1}^{N}d_i\frac{(X_{ih}-\mu)^2}{2\sigma^2}\\ &\quad-\sum_{i=1}^{N}\lambda_i\frac{(X_{il}-\mu)^2}{2\sigma^2},\\ Q(\mu,\sigma^2|\mu^{(m)},\sigma^{2(m)},Y) &\widehat{=} \mathsf{E}[\log p(\mu,\sigma^2|X)|\mu^{(m)},\sigma^{2(m)},Y]\\ &= -n\log\sqrt{2\pi}\sigma-\sum_{i=1}^{N}d_i\mathsf{E}\Big[\frac{(X_{ih}-\mu)^2}{2\sigma^2}\Big|\mu^{(m)},\sigma^{2(m)},Y\Big]\\ &\quad-\sum_{i=1}^{N}\lambda_i\mathsf{E}\Big[\frac{(X_{il}-\mu)^2}{2\sigma^2}\Big|\mu^{(m)},\sigma^{2(m)},Y\Big].\end{aligned}$$

X_{ih}的条件密度

$$f(t|\mu^{(m)},\sigma^{2(m)},Y)=\frac{\frac{1}{\sqrt{2\pi}\sigma^{(m)}}\exp\Big[-\frac{(t-\mu^{(m)})^2}{2\sigma^{2(m)}}\Big]}{\int_{a_{i-1}}^{a_i}\frac{1}{\sqrt{2\pi}\sigma^{(m)}}\exp\Big[-\frac{(t-\mu^{(m)})^2}{2\sigma^{2(m)}}\Big]\mathrm{d}x},$$
$$j=1,2\cdots,N,\ t\in[a_{i_1},a_i).$$

X_{il}的条件密度

$$f(t|\mu^{(m)},\sigma^{2(m)},Y)=\frac{\frac{1}{\sqrt{2\pi}\sigma^{(m)}}\exp\Big[-\frac{(t-\mu^{(m)})^2}{2\sigma^{2(m)}}\Big]}{1-\int_{-\infty}^{a_i}\frac{1}{\sqrt{2\pi}\sigma^{(m)}}\exp\Big[-\frac{(t-\mu^{(m)})^2}{2\sigma^{2(m)}}\Big]\mathrm{d}x},$$
$$j=1,2\cdots,N,\ t\in[a_i,+\infty).$$

为方便起见将$f(x_{ih})=f(t|\mu^{(m)},\sigma^{2(m)},Y)$, $f(x_{il})=f(t|\mu^{(m)},\sigma^{2(m)},Y)$, 分别记为$P_{ih}(t)$, $P_{il}(t)$. 则有

$$\begin{aligned}&Q(\mu,\sigma^2|\mu^{(m)},\sigma^{2(m)},Y)\\=\ &-\frac{n}{2}\log 2\pi\sigma^2-\sum_{i=1}^{N}d_i\int_{a_{i-1}}^{a_i}\frac{(x-\mu)^2}{2\sigma^2}P_{ih}(x)\mathrm{d}x-\sum_{i=1}^{N}\lambda_i\int_{a_i}^{+\infty}\frac{(x-\mu)^2}{2\sigma^2}P_{il}(x)\mathrm{d}x.\end{aligned}$$

M步: 将$Q(\mu,\sigma^2|\mu^{(m)},\sigma^{2(m)},Y)$分别对$\mu,\sigma^2$求导, 求出使$Q(\mu,\sigma^2|\mu^{(m)},\sigma^{2(m)},Y)$极大化的点$(\mu^{(m+1)},\sigma^{2(m+1)})$.

首先对μ求导

$$\frac{\partial Q}{\partial\mu}=\sum_{i=1}^{N}d_i\int_{a_{i-1}}^{a_i}\frac{x-\mu}{\sigma^2}P_{ih}(x)\mathrm{d}x+\sum_{i=1}^{N}\lambda_i\int_{a_i}^{+\infty}\frac{x-\mu}{\sigma^2}P_{il}(x)\mathrm{d}x.$$

令$\partial Q/\partial\mu=0$, 有

$$\mu=\frac{\sum_{i=1}^{N}d_i\int_{a_{i-1}}^{a_i}xP_{ih}(x)\mathrm{d}x+\sum_{i=1}^{N}\lambda_i\int_{a_i}^{+\infty}xP_{ih}(x)\mathrm{d}x}{\sum_{i=1}^{N}d_i\int_{a_{i-1}}^{a_i}P_{ih}(x)\mathrm{d}x+\sum_{i=1}^{N}\lambda_i\int_{a_i}^{+\infty}P_{ih}(x)\mathrm{d}x}.\tag{3.1}$$

对σ^2求导

$$\frac{\partial Q}{\partial\sigma^2}=-\frac{n}{2\sigma^2}+\sum_{i=1}^{N}d_i\int_{a_{i-1}}^{a_i}\frac{(x-\mu)^2}{2\sigma^4}P_{ih}(x)\mathrm{d}x+\sum_{i=1}^{N}\lambda_i\int_{a_i}^{+\infty}\frac{(x-\mu)^2}{2\sigma^4}P_{il}(x)\mathrm{d}x.$$

令$\partial Q/\partial\sigma^2=0$, 有

$$\sigma^2=\frac{2}{n}\Big\{\sum_{i=1}^{N}d_i\int_{a_{i-1}}^{a_i}\frac{(x-\mu)^2}{2}P_{ih}(x)\mathrm{d}x+\sum_{i=1}^{N}\lambda_i\int_{a_i}^{+\infty}\frac{(x-\mu)^2}{2}P_{il}(x)\mathrm{d}x\Big\}.\tag{3.2}$$

利用(3.1), (3.2)所得到的μ,σ^2就是我们所要寻找的$\mu^{(m+1)},\sigma^{2(m+1)}$. 这样就完成了一次迭代: $(\mu^{(m)},\sigma^{2(m)})\longrightarrow(\mu^{(m+1)},\sigma^{2(m+1)})$.

§4. 模　　拟

记X_i', $i=1,\cdots,n$为独立同分布于对数正态分布$LN(\mu,\sigma^2)$的随机变量, 取$T_0=0$, $T_1=0.3$, $T_2=3$, $T_3=30$, $T_4=300$, $T_5=3000$, $T_6=30000$, $T_7=300000$, 误差精度0.01. 对$j\leqslant 6$, 产品在T_j被截尾的概率取为$j/7$, 而到T_7所有未失效的产品都被截尾. 分别对$\mu=5$, $\sigma=2$与$\mu=8$, $\sigma=1.5$进行迭代, 模拟结果见下表. (其中μ',σ'为参数初值)

表1　取$\mu=5,\sigma=2,\mu'=6,\sigma'=1.5$的模拟结果

样本量	模拟次数	$\overline{\widehat{\mu}}$	$\overline{\widehat{\sigma}}$
100	10	4.8659	1.9675
	30	4.8732	1.9644
	50	5.0945	1.9634
300	10	5.0294	2.0967
	30	4.9711	2.0899
	50	5.0295	2.0866
800	10	4.9774	2.1063
	30	5.0238	2.1057
	50	5.0297	2.1049
2000	10	5.0338	2.1077
	30	5.0197	2.1067
	50	4.9826	2.1054

表2　取$\mu=8,\sigma=1.5,\mu'=6,\sigma'=2$的模拟结果

样本量	模拟次数	$\overline{\widehat{\mu}}$	$\overline{\widehat{\sigma}}$
100	10	7.9359	1.4733
	30	7.9488	1.4655
	50	8.1127	1.4653
1-4 300	10	8.0687	1.4890
	30	8.0846	1.4881
	50	7.9711	1.4879
1-4 800	10	8.0354	1.5237
	30	7.9822	1.5211
	50	7.9823	1.5196
1-4 2000	10	8.0346	1.5568
	30	8.0344	1.5509
	50	8.0288	1.5490

模拟结果表明随着样本量的增大, 估计值与参数真值的偏差有明显的减小. 当样本量为300时, 两个参数的平均偏差都在5%之内, 效果不错.

§5. 小　　结

本文针对数据分组与右删失情形下对数正态分布的参数估计问题进行了讨论, 主要给出未知参数的极大似然估计存在且唯一的充要条件; 并且介绍了如何利用EM算法对参数值进行估计. 从模拟结果中, 不难看出对于服从对数正态分布的分组右删失数据, 特别是样本较大的情况, 利用由EM算法得出的迭代公式可以得到对参数μ,σ的比较满意的估计$\widehat{\mu},\widehat{\sigma}$.

相应地, 也可以讨论这种情形下参数的区间估计问题, 如可以考虑极大似然估计是否具有渐近正态性, 从而研究区间估计或利用Bootstrap方法. 本文不做深入讨论.

参 考 文 献

[1] Laweless, J.F, *Statistical Models and Methods for Lifetime Data*, John wilay & sons, 1983.

[2] Bain, L.J, Engelhardt, M., *Statistical Analysis of Reliablity and Life Testing Models, Theory and Methods*, 2nd ed, New York: Marcel Dekker, 1991.

[3] Cheng, K.F., Chen, C.H., Estimation of the Weibull parameters with grouped data, *Commu. Statist.-Theor.*, **17(2)**(1988), 325–341.

[4] Liu, L.P., Existence of MLE for Weibull distribution with grouped and censored data, *Chinese Journal of Applied Probability and Statistics*, **17(2)**(2001), 133–138.

[5] 王静, 分组数据情形下对数正态分布参数的最大似然估计, 应用数学学报, **26(4)**(2003), 737–744.

[6] 郑明, 杨艺, 郑宇, 基于分组数据的Weibull分布的参数估计, 高校应用数学学报A辑, **18(3)**(2003), 303–310.

Estimation of the Parameters in the Lognormal Distribution with Grouped and Right-Censored Data

Liu Xin Chen Hui Fei Heliang

(*College of Mathematics and Sciences, Shanghai Normal University, Shanghai, 200234*)

In this paper we consider the estimation problem in the situation where the life of some products follows a lognormal distribution, and the observed data are grouped and right-censored. A necessary and sufficient condition for the existence and uniqueness of the Maximum Likelihood Estimate (MLE) of unknown parameters is given. Furthermore, we have estimated the unknown parameters by EM algorithm.

Keywords: Lognormal distribution, grouping and right-censoring, MLE, EM algorithm.

AMS Subject Classification: 62N01, 62N02.