第10卷 第32期 2010年11月
1671—1815(2010)32-7993-04

科 学 技 术 与 工 程
Science Technology and Engineering

Vol. 10 No. 32 Nov. 2010
ⓒ 2010 Sci. Tech. Engng.

不完全数据情形下指数分布参数的极大似然估计

李雪芳
(暨南大学经济学院统计系,广州 510632)

摘 要 处理生存分析观测数据使用的参数估计方法有很多,极大似然估计法是最常见的一种估计方法。当寿命分布为指数分布时,给出了定时截尾数据、定数截尾数据情形的极大似然估计,以及随机右删失下参数极大似然估计的一般表达式。此外,还提出了分组数据场合参数极大似然估计的图解求法。

关键词 指数分布 定时截尾数据 定数截尾数据 分组数据 极大似然估计

中图法分类号 O213.2; **文献标志码** A

0 引言

随着科学技术的发展,产品的可靠性愈来愈受到人们的重视。为了弄清被测试产品的寿命分布,求出各项可靠性指标,研究产品失效机理以便对提高产品可靠性提出建议,常常需要进行寿命试验。寿命试验按样品的失效情况分为完全寿命试验和截尾寿命试验,后者运用最广泛。

譬如在产品寿命试验中,由于试验设备、观测手段或有其他方面的困难造成某些试验数据丢失或未观测到的现象等。这样我们得到各种"删失数据"。

如何对删失数据进行分析,是一个重要的统计问题。很多学者都对删失数据做了研究,例如:王启华在文献[1]中给出了随机删失下指数分布的参数极大似然估计的一些结果;刘力平在文献[2]中研究了威布尔情形下删失数据的一些统计推断问题;田霆,刘次华在文献[3]中着重研究了定时结尾数据的删失数据,文献[4]着重研究了定数截尾删失数据。

由于删失机制多种多样,不能用同一种方法进行数据处理,因此,我们有必要进行分类研究,以期能解决各种不同的实际问题。本文针对[3]中的定时截尾试验模型和文献[4]中的定数截尾试验模型,利用文献[5]中的方法对定时和定数试验模型给出参数的极大似然估计。

2010年8月4日收到,8月20日修改

1 定时截尾删失数据的极大似然估计

假设总体服从参数为 θ 的指数分布。若取 n 个产品同时参加定时截尾试验,试验进行到 $\tau(\tau>0)$ 时刻停止。设在 τ 时刻以前有 r 个产品失效,记相应的失效时间为 $t_1 \leqslant t_2 \leqslant \cdots \leqslant t_r \leqslant \tau$,则总试验时间 $T = \sum_{i=1}^{r} t_i + (n-r)\tau$。记第 i 个个体得到的观测值是 $X_i \wedge \tau$,令 $t_1 = X_i \wedge \tau, \delta_i = I_{X_i \leqslant \tau}$,则得到数据 $(t_i,\delta_i)i = (1,\cdots,n)$,$\delta_i = 1$ 表示是 t_i 寿终数据,$\delta_i = 0$ 表示 t_i 是删失数据。$(t_1,\delta_1),\cdots,(t_n,\delta_n)$ 的似然函数为:

$$L(\theta) = \prod_{i=1}^{n} f(t_i,\theta)^{\delta_i}[1 - F(t_i,\theta)]^{1-\delta_i} \quad (1)$$

产品的失效概率为:

$$p = P(t < \tau) = 1 - e^{-\frac{\tau}{\theta}} \quad (2)$$

已知一个产品在 $[t_i, t_i + \mathrm{d}t_i]$ 内失效的概率为 $f(t_i)\mathrm{d}t_i$,其余 $n-r$ 个产品的寿命超过 τ 的概率为 $(e^{-\frac{\tau}{\theta}})^{n-r}$。所以上述观察结果出现的概率近似为:

$$k[(1/\theta)e^{-\frac{t_1}{\theta}}\mathrm{d}t_1]\cdots[(1/\theta)e^{-\frac{t_r}{\theta}}\mathrm{d}t_r](e^{-\frac{\tau}{\theta}})^{n-r}。$$

其中 k 是某一个常数。常数因子对求极大似然估计无影响,故由式(1)可得指数分布下定时截尾删失

数据的似然函数：

$$L(\theta) = \frac{1}{\theta^r}\exp\left\{-\frac{[t_1 + \cdots + t_r + (n-r)\tau]}{\theta}\right\} \tag{3}$$

令 $\lambda = \frac{1}{\theta}$，并取对数对 λ 求导，再令 $\frac{\partial \ln(\lambda)}{\partial \lambda} = 0$，于是得平均寿命 θ 的极大似然估计为 $\theta = \frac{1}{\hat{\lambda}} = \frac{T}{r}$。

2 定数截尾删失数据的极大似然估计

假设将随机抽取 n 个产品在时刻时投入试验，试验进行到有 r 个（r 是事先规定的，$n \geqslant r$）产品失效时停止，r 个产品的失效时间分别为 $0 \leqslant t_1 \leqslant t_2 \leqslant \cdots \leqslant t_r$，这里 t_r 是第 r 个产品的失效时间，所得样本 $t_1, t_2, \cdots, t_n$ 称为定数截尾样本。

利用上述样本估计未知参数 θ（产品的平均寿命）。在时间区间 $[0, t_r]$ 内有 r 个产品失效，有 $n-r$ 个产品的寿命超过 t_r。

为了确定似然函数，观察上述结果出现的概率。产品在 $(t_i, \mathrm{d}t_i]$ 失效的概率近似的为：

$$f(t_i)\mathrm{d}t_i = \frac{1}{\theta}e^{-\frac{t_i}{\theta}}\mathrm{d}t_i, i = 1, 2, \cdots, r。$$

其余 $n-r$ 各产品寿命超过 t_r 的概率为：

$$\left(\int_{t_r}^{\infty}\frac{1}{\theta}e^{-\frac{t}{\theta}}\mathrm{d}t\right)^{n-r} = (e^{-\frac{t_m}{\theta}})^{n-r}。$$

上述观察结果出现的概率近似的为：

$$\frac{n!}{(n-r)!}\left(\frac{1}{\theta}e^{-\frac{t_1}{\theta}}\mathrm{d}t_1\right)\cdots\left(\frac{1}{\theta}e^{-\frac{t_r}{\theta}}\mathrm{d}t_r\right)\left(e^{-\frac{t_r}{\theta}}\right)^{(n-r)} = k\frac{1}{\theta^r}e^{-\frac{1}{\theta}[t_1+t_2+\cdots+t_r+(n-r)t_r]}。$$

因为忽略一个常数因子对求极大似然估计无影响，所以不考虑 $\frac{n!}{(n-r)!}\mathrm{d}t_1\cdots\mathrm{d}t_r$，

令 $T = t_1 + t_2 + \cdots + t_r + (n-r)t_r$。

似然函数可取如下形式 $L(\theta) = \frac{1}{\theta^r}e^{-\frac{T}{\theta}}$。

令 $\frac{\partial L(\theta)}{\partial \theta} = 0$，故可得 θ 的极大似然估计为

$$\hat{\theta} = \frac{T}{r}。$$

3 随机右删失数据参数的极大似然估计

定时截尾数据与定数截尾数据参数的最大似然估计有相似之处，原因定时截尾数据与定数截尾数据都是随机右删失的特殊情形。下面给出随机右删失数据的最大似然估计。

设截尾时间 $Y_1, Y_2, \cdots$ 是概率空间 (Ω, F^*, P_θ) $(\theta \in \Theta, \theta > 0)$ 上的相互独立、取正值的随机变量序列，Y_i 的分布函数为 $G_i(y)$，密度函数为 $g_i(y)$，$i = 1, 2, \cdots$，它们与参数 θ 无关。

假定 $\{X_i\}$ 与 $\{Y_i\}$ 相互独立。现在有 n 个受试样本。设观察到数据为 $\{Z_i\}$，$i = 1, 2, \cdots, n$。每个 Z_i 如下取值：

（1）当 $X_i < Y_i$ 时，产品在截尾之前失效，这时我们知道产品寿命的确切值，故可取 $X_i = Z_i$；

（2）当 $X_i \geqslant Y_i$ 时，产品寿命不小于截尾时间，这时我们仅知截尾时间而不知道产品寿命，故可取 $Z_i = Y$。

综上知：$Z_i = X_i \wedge Y_i = \min(X_i, Y_i)$。

再取：$\delta_i = \begin{cases} 1, \text{若 } X_i < Y \\ 0, \text{若 } X_i \geqslant Y_i \end{cases}$，$i = 1, 2, \cdots, n$。

这时，对第 i 个受试样本而言，随机向量 (Z_i, δ_i) 完全描述了试验是否截尾、以及试验时间的长短。这样，在试验终止时可得到 n 组观察值：$(Z_1, \delta_1), (Z_2, \delta_2), \cdots, (Z_n, \delta_n)$，这就是我们能获得的随机截尾试验数据。

若记 $\bar{F} = 1 - F, \bar{G} = 1 - G$，再假定 $\sum_{i=1}^{n}\delta_i > 0$，我们有[6]：

定理 1 Z_i 与 δ_i 的联合密度函数为：

$$L(\theta) = L_i(z_i, \delta_i; \theta) = [f(z_i;\theta)\bar{G}_i(z_i)]^{\delta_i}[g_i(z_i)\bar{F}(z_i;\theta)]^{1-\delta_i}。$$

由于截尾时间分布中不包含未知参数 θ，故若记 $A = \prod_{i=1}^{n}[g_i^{1-\delta_i}(z_i)\bar{G}_i^{\delta_i}(z_i)]$，则 $(Z_1, \delta_1), (Z_2, \delta_2), \cdots, (Z_n, \delta_n)$ 的联合密度函数为

$$L(\theta) = \prod_{i=1}^{n}L_i(z_i, \delta_i; \theta) =$$

$$\prod_{i=1}^{n}[f(z_i;\theta)\bar{G}_i(z_i)]^{\delta_i}[g_i(z_i)\bar{F}(z_i;\theta)]^{1-\delta_i} =$$

$$A\prod_{i=1}^{n}[f(z_i;\theta)]^{\delta_i}[\bar{F}(z_i;\theta)]^{1-\delta_i}。$$

据此,在指数分布情形下,$(Z_1,\delta_1),(Z_2,\delta_2),\cdots,(Z_n,\delta_n)$的联合密度函数为

$$L(\theta) = A\lambda^{\sum_{i=1}^{n}\delta_i}e^{-\lambda\sum_{i=1}^{n}z_i}。$$

取对数得

$$\ln L(\theta) = \ln A + \sum_{i=1}^{n}\delta_i\ln\lambda - \lambda\sum_{i=1}^{n}z_i,$$

令 $\frac{\partial\ln L(\theta)}{\partial\theta} = 0$,知参数 θ 的极大似然估计为

$$\hat{\theta} = \frac{1}{\hat{\lambda}} = \frac{\sum_{i=1}^{n}z_i}{\sum_{i=1}^{n}\delta_i}。$$

4　分组数据下的极大似然估计

现取 n 个个体的寿命进行观测,获得数据如下:将$(0,+\infty)$分成 $k+1$ 个区间,前 k 个区间记作$(t_{i-1},t_i](i=1,2,\cdots,k+1)$,这里 $0=t_0<t_1<\cdots<t_k<t_{k+1}=+\infty$。假定我们由于种种原因不能得到 n 个个体寿命的确切值,只能知道它们落入各个区间的个数:$n_1,n_2,\cdots,n_k,n_{k+1}$,这里 n_i 是 n 个个体中实际寿命值属于$(t_{i-1},t_i]$的个数。。

由于 n 个个体的寿命 $X_1,\cdots,X_n$ 是相互独立的,$n_i = \sum_{j=1}^{n}I(t_{i-1}<X_j\leqslant t_i), i=1,\cdots,k+1$,已知数据 $n_1,n_2,\cdots,n_k,n_{k+1}$对应的似然函数是:

$$L(\theta) = \prod_{i=1}^{k+1}[F(t_i,\theta)-F(t_{i-1},\theta)]^{n_i} \quad (4)$$

$L(\theta)$的最大值点 $\hat{\theta}$ 便是 θ 的极大似然估计。

首先给出指数分布下 MLE 存在且唯一的充要条件。则有如下定理:

定理 2　设 $X_1,\cdots,X_n$ 是相互独立同分布的正值随机变量,共同分布是 $F(t;\theta)=1-e^{-\frac{t}{\theta}}$,其中 θ 是未知的正数,则对于数据 $n_1,n_2,\cdots,n_k,n_{k+1},\theta$ 的极大似然估计存在且唯一的充要条件是 $n_1<n$ 且 $n_{k+1}<n$。

由式(4)知

$$\ln L(\theta) = \sum_{i=1}^{k+1}n_i\ln(F(t_i,\theta)-F(t_{i-1},\theta))。$$

将指数分布函数代入上式可得似然函数

$$\ln L(\theta) = \sum_{i=1}^{k+1}n_i\ln(e^{-\frac{t_{i-1}}{\theta}}-e^{-\frac{t_i}{\theta}})。$$

令 $\frac{\partial\ln L(\theta)}{\partial\theta} = 0$,有

$$\frac{\partial\ln L(\theta)}{\partial\theta} = \sum_{i=1}^{k+1}\alpha_i(\theta)n_i - t_kn_{k+1} = 0,$$

这里 $\alpha_i = \frac{t_ie^{-(t_i-t_{i-1})/\theta}-t_{i-1}}{1-e^{-(t_i-t_{i-1})/\theta}}$。

记 $H(\theta) = \sum_{i=1}^{k+1}\alpha_i(\theta)n_i$,则有:

$$\frac{\partial H(\theta)}{\partial\theta} = \frac{e^{-\frac{t_i-t_{i-1}}{\theta}}\frac{t_i-t_{i-1}}{\theta^2}[t_i-t_{i-1}]}{[1-e^{-\frac{t_i-t_{i-1}}{\theta}}]^2}。$$

因为 $t_i>t_{i-1}, i=(1,\cdots,n)$,故$\frac{\partial H(\theta)}{\partial\theta}>0$,所以$H(\theta)$是 θ 的增函数,且在定理假设条件下,$\frac{\partial^2\ln L(\theta)}{\partial\theta^2}<0$且 $\lim_{\theta\to0}\frac{\partial\ln L(\theta)}{\partial\theta} = \infty$, $\lim_{\theta\to\infty}\frac{\partial\ln L(\theta)}{\partial\theta}<0$,故方程 $\frac{\partial\ln L(\theta)}{\partial\theta} = 0\Leftrightarrow H(\theta) = t_kn_{k+1}$(右端是与 θ 无关的取正值的常数函数)。

由图解法,以 θ 的值为横坐标,左右两端的函数值为纵坐标画图,可知函数图像必与横轴相交于一点,又因为 $H(\theta)$是增函数,故存在唯一的交点,即在$(0,+\infty)$中恰有一个根 $\hat{\theta}$,且 $\hat{\theta}$ 是 $L(\theta)$的最大值点也就是所求的估计值。

5　有待继续研究的问题

本文仅对指数分布场合下三种常见删失数据模型的参数给出极大似然估计。而在实际中,威布尔分布、伽玛分布以及对数正态分布应用也很广泛。在许多类型的产品如:真空管、滚珠轴承、电器的绝缘材料,都广泛提倡用威布尔分布。伽玛分布虽不及威布尔分布那样常用,但伽马分布适合广泛

的寿命数据,并且一些失效过程模型还可导出伽玛分布。对数正态分布在分析电器绝缘体的失效时间,研究吸烟者中肺癌的出现时间等应广泛用。这些寿命分布下截尾删失数据的处理问题是进一步的研究方向。

参 考 文 献

1 *Wang Qihua*. Some Results on MLE of the Parameter of the Exponential Distribution from Randomly Censored Data. Acta Scientiarum Naturalium Universitatis Pekinensis,2000;36(5):583—590

2 刘力平. 威布尔分布组与删失数据下最大似然估计的存在性. 应用概率统计,2001;17(2):133—138

3 田 霆,刘次华. 定时截尾缺失数据下指数分布的统计推断. 华侨大学学报(自然科学版),2006;27(1):20—23

4 王乃生,王玲玲. 定数截尾数据缺失场合下指数分布参数的 Bayes 估计. 应用概率统计,2001;17(2):229—235

5 陈家鼎,孙山泽,李东风,等. 数理统计学讲义. 第 2 版. 北京:高等教育出版社,2006:12—16

6 刘海峰. 随机截尾指数寿命数据之参数 θ 的极大似然估计及其性质. 解放军理工大学学报. 2000;1(2):96—99

Maximum Likelihood Parameter Estimation of Exponential Distribution Based on Incomplete Sample

LI Xue-fang

(Department of Statistics, Economic Institute, Jinan University, Guangzhou 510632, P. R. China)

[**Abstract**] There are many methods to deal with measuring data in survival analysis. Maximum likelihood estimation is the most popular one. Under exponential distribution, maximum likelihood estimators of the parameter for Type-I censored data and Type-II censored data are obtained, and a general expression of maximum likelihood estimator for right randomly censoring data is also derived. Moreover, a graphical method is developed to solve maximum likelihood estimation of paremeter for Packet censored data.

[**Key words**] exponential distribution type-I censored data type-II censored data packet censored data maximum likelihood estimation