指数分布参数基于不完全数据的区间估计

罗 艳*
(重庆教育学院数学系 重庆 400067)

【摘要】 对不完全样本观测数据,讨论了指数分布总体参数的区间估计;给出了构造置信区间的一种方法并推导出了相应的分布密度函数表达式;并说明了该方法在样本中可能存在异常值时的应用。

关 键 词 指数分布; 参数估计; 置信区间; 不完全数据; 异常值; 稳健性
中图分类号 O212.1; O212.7

在产品的寿命试验以及一些特殊的试验中,通常得到不完全的样本观测数据,关于这一类样本的参数估计是一个比较重要的问题。不少文献对此问题的讨论大都只限于点估计。由于不完全样本的特殊性,基于其构造的统计量往往很难求出相应的概率分布函数表达式,因此总体参数基于不完全数据的区间估计就有一定的难度。近年来,对 Weibull 分布、对数正态分布已求出了近似的置信区间[1~3]。

指数分布是讨论电子产品、元器件等使用寿命的重要分布。本文针对指数分布总体讨论总体参数基于不完全数据的区间估计问题,给出了构造置信区间的一个方法并推导出了构造置信区间所需枢轴量的概率密度函数表达式。该方法也适用于样本数据中可能存在异常值的情形。

1 枢轴量的构造

设 $X_1, X_2, \cdots, X_n$ 是来自指数分布总体 X 的独立同分布(iid)样本,则

$$f_Z(x) = \begin{cases} \lambda e^{-\lambda x} & x > 0 \\ 0 & x \leqslant 0 \end{cases} \tag{1}$$

其中 $\lambda > 0$ 为总体参数。记 $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$ 为 $X_1, X_2, \cdots, X_n$ 的顺序统计量,定义

$$Z_k = \frac{X_k}{\lambda X_{(k+1)}^2} \qquad 1 \leqslant k < n \tag{2}$$

下面证明在对参数 λ 进行区间估计时,Z_k 是枢轴量。

定理 1 由式(2)定义的 Z_k 的概率分布与参数 λ 的取值无关。

证明 设 $X_1, X_2, \cdots, X_n$ 是总体(1)的样本,$X_{(1)}, X_{(2)}, \cdots, X_{(n)}$ 是相应的顺序统计量,则不难证明 $\lambda X_1, \lambda X_2, \cdots, \lambda X_{(n)}$ 是总体(1)当 $\lambda = 1$ 时的样本,且由于 $\lambda > 0$,相应的顺序统计量是 $\lambda X_{(1)}, \lambda X_{(2)}, \cdots, \lambda X_{(n)}$,而

$$Z_k = \frac{X_{(k)}}{\lambda X_{(k+1)}^2} = \frac{\lambda X_{(k)}}{[\lambda X_{(k+1)}]^2}$$

因此,λ 取大于 0 的任何值,Z_k 的概率分布均与 $\lambda = 1$ 时相同,即 Z_k 的概率分布与 λ 的取值无关。

证毕

1998 年 1 月 30 日收稿,1998 年 7 月 8 日修改定稿
* 女 35 岁 大学 讲师

定理1说明了在对参数 λ 进行区间估计时，Z_k 是枢轴量。

2 枢轴量的概率分布

本节讨论 Z_k 的概率分布，沿用上节记号，有

引理1[4] $(X_{(k)}, X_{(k+1)})$的联合密度函数为

$$f_{k,k+1}(x_1,x_2)=\begin{cases}n(n-1)C_{n-2}^{k-1}[F_X(x_1)]^{k-1}[1-F_X(x_2)]^{n-k-1}f_X(x_1)f_X(x_2) & x_1<x_2\\ 0 & x_1\geqslant x_2\end{cases}$$

式中 $F_X(x)$、$f_X(x)$分别是总体分布函数和密度函数。

由定理1，我们只考虑 $\lambda=1$ 的情形。

引理2 当总体(1)参数 $\lambda=1$ 时，$(X_{(k)}, X_{(k+1)})$的联合密度函数为

$$f_{k,k+1}(x_1,x_2)=\begin{cases}n(n-1)C_{n-2}^{k-1}[1-\exp(-x_1)]^{k-1}\exp(-x_1)\exp[-(n-k)x_2] & 0<x_1<x_2\\ 0 & \text{其他}\end{cases}$$

定理2 记

$$h_i(z)=z_1^3\{\exp(z_1^2)\sqrt{\pi}(1+2z_1^2)[\Phi(\sqrt{2}z_2)-\Phi(\sqrt{2}z_1)]+(2z_1-z_2)\exp(-z_2^2+z_1)-z_1\}$$

其中 $z_1=\dfrac{n-k}{2\sqrt{(k-i)z}}$，$z_2=z_1+\sqrt{\dfrac{k-i}{z}}$，$n$、$k$ 意义同前，$i=0,1,\cdots,k-1$；$\Phi(x)$是标准正态分布函数。则有枢轴量 Z_k 的分布密度函数$f_k(z)$为

$$f_k(z)=\begin{cases}\dfrac{4n(n-1)}{(n-k)^3}C_{n-2}^{k-1}\displaystyle\sum_{i=0}^{k-1}(-1)^{k-i-1}C_{k-1}^{i}h_i(z) & z>0\\ 0 & z\leqslant 0\end{cases}\tag{3}$$

证明 作变换

$$\begin{cases}Z_k=\dfrac{X_{(k)}}{X_{(k+1)}^2}\\ T=X_{k+1}\end{cases}\tag{4}$$

则由

$$\begin{cases}z=\dfrac{x_1}{x_2^2}\\ t=x_2\end{cases}$$

解出

$$\begin{cases}x_1=zt^2\\ x_2=t\end{cases}$$

于是式(4)的雅可比$|J|$为

$$|J|=\begin{vmatrix}t^2 & 2tz\\ 0 & 1\end{vmatrix}=t^2$$

由引理2可得到(Z_k,T)的联合密度为

$$g(z,t)=t^2f_{k,k+1}(zt^2,t)$$

当 $z=0$ 时，$zt^2\leqslant 0$，故

$$g(z,t)=0$$

当 $z>0$ 且 $z\geqslant 1/t$ 时，$zt^2\geqslant t$，故

$$g(z,t)=0$$

当 $0<z<1/t$ 时

$$g(z,t)=n(n-1)C_{n-2}^{k-1}t^2[1-\exp(-zt^2)]^{k-1}\exp[-zt^2-(n-k)t]$$

而 $f_k(z)=\int_{-\infty}^{+\infty}g(z,t)\mathrm{d}t$，于是有

当 $z\leqslant 0$ 时

$$f_k(z)=0$$

当 $z>0$ 时

$$f_k(z)=\int_0^{\frac{1}{z}}n(n-1)C_{n-2}^{k-1}t^2[1-\exp(-zt^2)]^{k-1}\exp[-zt^2-(n-k)t]\mathrm{d}t=$$

$$n(n-1)C_{n-2}^{k-1}\sum_{i=0}^{k-1}(-1)^{k-i-1}C_{k-1}^{i}\int_0^{\frac{1}{z}}t^2\exp[-(k-i)zt^2-(n-k)t]\mathrm{d}t=$$

$$n(n-1)C_{n-2}^{k-1}\sum_{i=1}^{k-1}(-1)^{k-i-1}C_{k-1}^{i}\int_0^{\frac{1}{z}}t^2\exp[-(\sqrt{(k-i)z}\,t+z_1)^2+z_1^2]\mathrm{d}t=$$

$$\frac{8n(n-1)}{(n-k)^3}C_{n-2}^{k-1}\sum_{i=0}^{k-1}(-1)^{k-i-1}C_{k-1}^{i}z_1^3\exp(z_1^2)\int_{z_1}^{z_2}(u-z_1)^2\exp(-u^2)\mathrm{d}u \tag{5}$$

不难求出

$$\int_{z_1}^{z_2}-2u\exp(-u^2)\mathrm{d}u=\exp(-z_2^2)-\exp(-z_1^2) \tag{6}$$

$$\int_{z_1}^{z_2}\exp(-u^2)\mathrm{d}u=\sqrt{\pi}[\Phi(\sqrt{2}z_2)-\Phi(\sqrt{2}z_1)] \tag{7}$$

$$\int_{z_1}^{z_2}u^2\exp(-u^2)\mathrm{d}u=\frac{\sqrt{\pi}}{2}[\Phi(\sqrt{2}z_2)-\Phi(\sqrt{2}z_1)]-\frac{1}{2}[z_2\exp(-z_2^2)-z_1\exp(-z_1^2)] \tag{8}$$

所以，由式(6)～式(8)得

$$\int_{z_1}^{z_2}(u-z_1)^2\exp(-u^2)\mathrm{d}u=\sqrt{\pi}[\Phi(\sqrt{2}z_2)-\Phi(\sqrt{2}z_1)](\frac{1}{2}+z_1^2)+$$

$$(z_1-\frac{1}{2}z_2)\exp(-z_2^2)-\frac{1}{2}z_1\exp(-z_1^2) \tag{9}$$

将式(9)代入式(5)，整理后，可得到式(3)。　　证毕

3　方法应用

首先说明利用上述结果，对于不完全的样本数据，如何构造总体参数的置信区间。

在寿命试验中常常遇到的截尾数据(Censored data)得到的是样本 $X_1,X_2,\cdots,X_n$ 中的前 r 个顺序统计量 $X_{(1)},X_{(2)},\cdots,X_{(r)}(r<n)$；一些特殊的试验只记录反应强度达到一定界限以上的数据，得到的是低端截尾的后几个顺序统计量；更一般的情形是只记录下顺序统计量中的任意一部分数据。本文构造的寻求总体参数置信区间的枢轴量 Z_k 只涉及顺序统计量 $X_{(k)}$ 及 $X_{(k+1)}$。因此，对于上述各种情形的不完全数据，大都能够找到样本数据许可的 k 值，计算出 $X_{(k)}/X_{(k+1)}^2$ 的数值，再根据式(3)计算出 Z_k 的分位数点(利用数值计算可实现)，从而得到总体参数 λ 的置信区间。有时可以得到样本数据许可的多个不同的 k 值，这时可以根据一定准则(比如在相同置信度下置信区间长度最短等)对 k 进行优选，确定一个较优的 k 值。

最后说明在怀疑样本数据中有异常值(Outlier)存在时，本文的方法仍然保持有效，具有稳健性(Robustness)。

异常值是指样本中的个别值,其数值明显地偏离其所在样本的其余观测值[5]。异常值可能是数据中随机性的极端表现,也可能是人为因素造成的后果。由于异常值对经典的统计方法影响甚大,如何尽量避免它的影响,已成为一个重要课题。在可能存在异常值的情况下,人们通常采用两种方式对待:1)先检验样本是否有异常值,如我国国家标准局公布的国家标准GB8056-87[5],然后再做相应处理;2)建立新的统计方法,使得即使存在异常值也不会对该方法有明显影响[6],亦即现在常说的稳健性或鲁棒性(Robustness)的含义之一。

当样本中存在异常值时,通常表现为样本观测值中最大或最小的若干个。很明显,本文的构造总体参数置信区间的方法,只要选取靠近 $n/2$(n 为样本容量)的 k 值,即可使得到的结果基本上不受异常值影响,因此,本方法对可能存在异常值的样本数据做总体参数的区间估计时,具有稳健性。

参 考 文 献

1 徐晓岭,费鹤良,陈振民.三参数 weibull 分布位置参数的置信限.数理统计与应用概率,1989,4(4):409~435

2 徐晓岭,费鹤良.三参数对数正态分布位置参数的置信限.数理统计与应用概率,1991,6(3):382~392

3 费鹤良,徐晓岭.三参数威布尔分布参数的联合置信域.应用概率统计,1992,8(4):398~402

4 David H A.Order Statistics,2nd ed.New York:John Wiley,1981

5 国家标准局.数据的统计处理和解释——指数样本异常值的判断和处理.中华人民共和国国家标准,1987,GB8056-87:1~11

6 Hubber P J.Robust Statistics.New York:John Wiley,1981

Confidence Interval Estimation for Parameter of Exponential Distribution from Incomplete Data

Luo Yan

(Dept.of Mathematics,Chongqing College of Education Chongqing 400067)

Abstract A method of confidence interval estimation for the parameter of exponential distribution from incomplete data is discussed.Making use of two order statistics to construct a new function of samples,the analytic expression of probability density function of the samples function is obtained.The result applies to the observed data when there are outliers.

Key words exponential distribution; parameter estimation; confidence interval; incomplete data; outlier; robustness

编辑 黄 辛