双参数指数分布异常数据的检验

李云飞，黄继伟，朱 宏

(电子科技大学应用数学学院 成都 610054)

【摘要】针对双参数指数分布的异常大数据，给出了一种新的检验方法。寻找到总体参数的具有较好稳健性的估计量，并在此基础上构造出检验统计量，求出了该检验统计量精确的概率密度函数和大样本情形下的近似分布，得到了检验临界值简洁的近似表达式。

关 键 词 双参数指数分布； 异常数据； 次序统计量； 分位数； 检验统计量

中图分类号 O212.1； O212.7 文献标识码 A

Detection of Outliers from the Two-Parameter Exponential Distribution

LI Yun-fei，HUANG Ji-wei，ZHU Hong

(School of Applied Mathematics, UEST of China Chengdu 610054)

Abstract In this paper, a method of examination for the upper outliers from the two－parameter exponential distribution is discussed. A new test statistic is given. We derive the accurate density function and the approximate distribution of this test statistic.

Key words two-parameter exponential distribution; outlier; order statistic; fractile; test statistic

双参数指数分布在可靠性理论中有着非常广泛的应用，它可以用来描述元件或系统的寿命。在做这些寿命试验时会得到大量的观测数据，但是由于受过失性误差的影响会产生异常数据，这些异常数据会对试验结果造成不良影响。因此，如何检测这些异常数据是一个重要的问题。

假设 $X_1, X_2, \cdots, X_n$ 是来自双参数指数分布总体X的独立同分布样本，总体X的密度函数为：

$$f(x,\mu,\sigma)=\begin{cases}\dfrac{1}{\sigma}\exp\left(-\dfrac{x-\mu}{\sigma}\right) & x\geqslant\mu\\ 0 & x<\mu\end{cases}$$

式中 $\mu\in R$ 为门限参数，$\sigma\in R^+$ 为尺度参数。设 $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$ 是以上样本的次序统计量，$x_{(1)}, x_{(2)}, \cdots, x_{(n)}$ 是相应的次序统计量的观测值。如果这些观测值中有异常大数据，一定出现在 $x_{(1)}, x_{(2)}, \cdots, x_{(n)}$ 的高端。针对 $G\left(\dfrac{x-\mu}{\sigma}\right)$ 型分布中的异常大数据，提出过一种检验方法[1]。但是，当样本中出现多个异常大数据时，该方法不能抵抗来自于这些异常数据的污染。因此，作为 $G\left(\dfrac{x-\mu}{\sigma}\right)$ 型分布的特例，当双参数指数分布样本中出现多个异常大数据时，上述方法不能检验出这些异常数据。本文针对以上问题，给出一种检验方法，可以

收稿日期：2004－02－24

作者简介：李云飞(1980－)，男，硕士生，主要从事数理统计方面的研究.

满足检验多个异常大数据的需要。

1 检验统计量的构造

由设 $X_{([np]+1)}$ 为样本 p 分位数 $(0<p<1)$。已经知道，样本分位数能够较好抵抗异常数据的干扰，尤其是越靠近样本中位数的样本分位数，其抵抗力越强[2]。而且对于一般常见的分布，样本 p 分位数是总体 p 分位数的渐近无偏，一致估计[3]。

定理 1 设 $X_1,X_2,\cdots,X_n$ 是来自双参数指数分布总体 $f(x,\mu,\sigma)$ 的独立同分布样本， $X_{(1)},X_{(2)},\cdots,X_{(n)}$ 是以上样本的次序统计量，则 $X_{(1)}$ 是门限参数 μ 的极大似然估计。

证明 设样本观测值为 $x_1,x_2,\cdots,x_n$，相应次序统计量的观测值为 $x_{(1)},x_{(2)},\cdots,x_{(n)}$，则似然函数为：

$$L(\theta)=\begin{cases}\dfrac{1}{\sigma^n}\exp\left(-\sum\limits_{i=1}^{n}\dfrac{x_i-\mu}{\sigma}\right) & \mu\leqslant x_{(1)}\leqslant x_{(2)}\leqslant\cdots\leqslant x_{(n)}\\ 0 & \text{其他}\end{cases}$$

式中 $\ln L(\theta)=-n\ln\sigma-\sum\limits_{i=1}^{n}\dfrac{x_i-\mu}{\sigma}$，显然 $\dfrac{\partial\ln L(\theta)}{\partial\mu}>0$，故 $\ln L(\theta)$ μ 是严格单调增加函数，因此当 μ 取最大值时，似然函数 $L(\theta)$ 达到最大值，即：μ 的极大似然估计为 $\hat{\mu}=X_{(1)}$。 证毕

引理 1[4] 假设 $X_1,X_2,\cdots,X_n$ 是来自总体密度为 $f(x;\theta)$ 的一个样本，Θ 为参数空间，若 $\ln f(x;\theta)$ 在 Θ 上可微，且 $\forall\,\theta\neq\theta',\{x:f(x;\theta)\neq f(x;\theta')\}$ 不是零测集，则似然方程在 $n\to\infty$ 时以概率1有解，而且此解是 θ 的一致估计。

由定理1和引理1易知，$X_{(1)}$ 也是 μ 的一致估计，即：$X_{(1)}\xrightarrow{P}\mu$，由于双参数指数分布的分布函数为 $F(x)=1-\exp\{-\dfrac{x-\mu}{\sigma}\}$ $(x\geqslant\mu)$，且 $F(\mu+\sigma)=1-e^{-1}=0.632\,1$，所以 $\mu+\sigma$ 是总体的0.632 1分位数。记 $p=0.632\,1$，则 $X_{([np]+1)}$ 是 $\mu+\sigma$ 的一致估计，即 $X_{([np]+1)}\xrightarrow{P}\mu+\sigma$，又由 $X_{(1)}\xrightarrow{P}\mu$，可以得到 $X_{([np]+1)}-X_{(1)}\xrightarrow{P}\sigma$。

定理 2 定义统计量为：

$$T=\frac{X_{(n)}-X_{(1)}}{X_{([np]+1)}-X_{(1)}}$$

式中 T的分布与双参数指数分布的总体参数无关。

证明 如果 X 服从双参数指数分布 $f(x,\mu,\sigma)$，易证 $(X-\mu)/\sigma$ 服从参数为 $\mu=0,\ \sigma=1$ 的双参数指数分布。又由于 $\mu\in R,\sigma\in R^+$，因此 $(X_1-\mu)/\sigma,(X_2-\mu)/\sigma,\cdots,(X_n-\mu)/\sigma$ 就是来自于双参数指数分布 $f(x,0,1)$ 的样本，并且 $(X_{(1)}-\mu)/\sigma,(X_{(2)}-\mu)/\sigma,\cdots,(X_{(n)}-\mu)/\sigma$ 就是相应的次序统计量。将T变形为：

$$T=\frac{\dfrac{X_{(n)}-\mu}{\sigma}-\dfrac{X_{(1)}-\mu}{\sigma}}{\dfrac{X_{([np]+1)}-\mu}{\sigma}-\dfrac{X_{(1)}-\mu}{\sigma}}$$

由上式看出，不管参数 μ,σ 取什么值，T 的分布都与 $\mu=0,\ \sigma=1$ 时的分布相同。 证毕

显然当 $X_{(n)}$ 异常大时，T 的数值将偏大，因此统计量 T 可用来检验 $X_{(n)}$ 是否异常大，而且由定理2，以下可以就参数 $\mu=0,\ \sigma=1$ 的情形进行讨论。

2 检验统计量的分布

记 $n_1=[np]+1$，则：

$$T=\frac{X_{(n)}-X_{(1)}}{X_{(n_1)}-X_{(1)}}$$

此时，可以求出 $(X_{(1)}, X_{(n_1)}, X_{(n)})$ 的联合概率密度 $f(x_1,x_2,x_3)$ 为[4]：

1) 当 $0<x_1<x_2<x_3$ 成立时，

$$f(x_1,x_2,x_3)=\frac{n!}{(n_1-2)!(n-n_1-1)!}[F(x_2)-F(x_1)]^{n_1-2}[F(x_3)-F(x_2)]^{n-n_1-1}f(x_1)f(x_2)f(x_3)$$

2) 当 $0<x_1<x_2<x_3$ 不成立时，

$$f(x_1,x_2,x_3)=0$$

做适当的变换，经过计算可以求出检验统计量 T 的概率密度函数为，

$$f_T(t)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(x_2-x_1)f[x_1,x_2,x_1+t(x_2-x_1)]\mathrm{d}x_1\mathrm{d}x_2$$

式中　当 $t\leqslant 1$ 时，

$$f_T(t)=0$$

当 $t>1$ 时，

$$f_T(t)=\int_0^{+\infty}\int_0^{x_2}\frac{n!(x_2-x_1)}{(n_1-2)!(n-n_1-1)!}[F(x_2)-F(x_1)]^{n_1-2}[F(x_1+t(x_2-x_1))-F(x_2)]^{n-n_1-1}\times$$
$$f(x_1)f(x_2)f[(x_1+t(x_2-x_1)]\mathrm{d}x_1\mathrm{d}x_2$$

式中　$F(x)$ 与 $f(x)$ 分别是双参数指数分布总体的分布函数和密度函数。

由于检验统计量 T 的概率密度函数表达式比较复杂，要通过它直接得到检验所需的临界值比较繁琐，但是可以通过计算机模拟得到近似的检验临界值。

下面讨论检验统计量 T 的大样本近似分布。

引理 2[2]　设 $\{\vec{\boldsymbol{\omega}}_n\},\{\vec{\boldsymbol{\varepsilon}}_n\}$ 是 k 维随机向量序列，$\{\vec{\boldsymbol{B}}_n\}$ 是 $k\times k$ 维随机矩阵序列，如果当 $n\to\infty$ 时，$\vec{\boldsymbol{\omega}}_n\xrightarrow{L}\vec{\boldsymbol{\omega}}$，$\vec{\boldsymbol{\varepsilon}}_n\xrightarrow{P}0$，$\vec{\boldsymbol{B}}_n\xrightarrow{P}\vec{\boldsymbol{B}}$，则有：

$$\boldsymbol{B}_n\boldsymbol{\omega}_n+\boldsymbol{\varepsilon}_n\xrightarrow{L}\vec{\boldsymbol{B}}\vec{\boldsymbol{\omega}}\qquad(n\to\infty)$$

式中　$\vec{\boldsymbol{\omega}}$ 是 k 维随机向量，$\vec{\boldsymbol{B}}$ 是 $k\times k$ 维常数矩阵，$\boldsymbol{0}$ 是 k 维零向量。

定理 3　记：

$$Z_j=\frac{X_j-X_{(1)}}{X_{(n_1)}-X_{(1)}}\qquad j=1,2,\cdots,n$$

设 N 是任意正整数，记：

$$\vec{\boldsymbol{Z}}_n=(Z_1,Z_2,\cdots,Z_N)^T$$

则当 $n\to\infty$ 时，随机向量序列 $\{\vec{\boldsymbol{Z}}_n\}$ 依分布收敛于 $\vec{\boldsymbol{Z}}$，其中 $\vec{\boldsymbol{Z}}$ 是 N 维随机向量，其联合分布函数为：

$$F_{\vec{\boldsymbol{Z}}}(z_1,z_2,\cdots,z_N)=\prod_{i=1}^{N}(1-\mathrm{e}^{-z_i})$$

证明　构造 N 维随机向量 $\vec{\boldsymbol{\varepsilon}}_n=\left(\frac{\mu-X_{(1)}}{X_{(n_1)}-X_{(1)}},\frac{\mu-X_{(1)}}{X_{(n_1)}-X_{(1)}},\cdots,\frac{\mu-X_{(1)}}{X_{(n_1)}-X_{(1)}}\right)^T$ 和 $\boldsymbol{N}\times\boldsymbol{N}$ 维随机矩阵序列：

$$\vec{\boldsymbol{B}}_n=\begin{pmatrix}\frac{\sigma}{X_{(n_1)}-X_{(1)}} & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & \frac{\sigma}{X_{(n_1)}-X_{(1)}}\end{pmatrix}$$

因为：

$$Z_i = \frac{X_i - X_{(1)}}{X_{(n_1)} - X_{(1)}} = \frac{X_i - \mu}{X_{(n_1)} - X_{(1)}} + \frac{\mu - X_{(1)}}{X_{(n_1)} - X_{(1)}} \qquad i = 1, 2, \cdots, N$$

所以：

$$\vec{\boldsymbol{Z}}_n = \vec{\boldsymbol{B}}_n \left(\frac{X_1 - \mu}{\sigma}, \frac{X_2 - \mu}{\sigma}, \cdots, \frac{X_N - \mu}{\sigma} \right)^T + \vec{\boldsymbol{\varepsilon}}_n$$

注意到 $X_{(n_1)} - X_{(1)} \xrightarrow{P} \sigma$， $X_{(1)} \xrightarrow{P} \mu$，从而有 $\vec{\boldsymbol{B}}_n \xrightarrow{P} \vec{\boldsymbol{I}}_{N\times N}$， $\vec{\boldsymbol{\varepsilon}}_n \xrightarrow{P} 0$ ($\vec{\boldsymbol{I}}_{N\times N}$ 为单位矩阵)，由引理2得 $\vec{\boldsymbol{Z}}_n \xrightarrow{L} \vec{\boldsymbol{I}}_{N\times N} \left(\frac{X_1 - \mu}{\sigma}, \frac{X_2 - \mu}{\sigma}, \cdots, \frac{X_N - \mu}{\sigma} \right)^T$。容易证明 $\left(\frac{X_1 - \mu}{\sigma}, \frac{X_2 - \mu}{\sigma}, \cdots, \frac{X_N - \mu}{\sigma} \right)^T$ 与该定理中 $\vec{\boldsymbol{Z}}$ 的分布函数相同，即：

$$\vec{\boldsymbol{Z}}_n \xrightarrow{L} \vec{\boldsymbol{Z}}$$ 证毕

通过该定理，实际上得到了检验统计量 T 的近似分布。因为 $T = \max\{Z_1, Z_2, \cdots, Z_n\}$，所以有 $P(T < z) = P(Z_1 < z, Z_2 < z, \cdots, Z_n < z)$。由定理3，当 n 足够大时，有：

$$P(T < z) \approx \prod_{j=1}^{n} (1 - \mathrm{e}^{-z}) = (1 - \mathrm{e}^{-z})^n$$

由此给出了检验统计量 T 的分布函数的近似表达式，该式可计算 T 相对显著性水平 α 的检验临界值 Z_α 的近似值为：

$$z_\alpha \approx -\ln[1 - (1 - \alpha)^{1/n}]$$

该表达式形式简洁，因此在进行异常数据检验时相关的计算比较简单。

3 结 束 语

检验统计量 T 能够抵抗来自于双参数指数分布异常大数据的干扰，具有较好稳健性。因此，利用该方法不仅能检验单个异常大数据，而且反复利用此方法也可以检验出多个异常大数据。

参 考 文 献

[1] 陈振民. $G(\frac{x-\mu}{\sigma})$ 型分布样本中异常值的统计检验[J]. 上海师范大学学报, 1987, 3: 13-17

[2] 成 平. 参数估计[M]. 上海: 上海科学技术出版社, 1985

[3] 陈希孺. 数理统计引论[M]. 北京: 科学出版社, 1997

[4] 茆诗松. 高等数理统计[M]. 北京: 高等教育出版社, 1998

[5] 朱 宏. Ⅰ型极值分布样本多个异常值的检验[J]. 电子科技大学学报, 1994, 23(3): 323-327

[6] 陈希孺. 非参数统计[M]. 上海: 上海科学技术出版社, 1989

编 辑 刘文珍