文章编号:1007 – 6492(2001)01 – 0049 – 03

指数分布参数极大似然估计检验的样本崩溃点

侯紫燕,严广松,原新凤

(河南纺织高等专科学校基础部,河南 郑州 450007)

摘 要:如何量化一种统计方法对异常值的不敏感性一直是稳健统计研究的一个重要课题.检验的样本崩溃点是样本中能逆转判决的离群值的最小比例.在研究相关文献的基础上,计算出指数分布参数极大似然估计检验的样本崩溃点,并分析了样本崩溃点的渐近正态性,为量化统计方法的稳健性提供了一种新的途径.

关键词:指数分布;极大似然估计;样本崩溃点;渐近正态性

中图分类号:O 212.7 文献标识码:A

0 引言

众所周知,在运用统计方法解决实际问题时,要先收集数据,然后再按照一定的统计方法由这些数据计算统计量的值,并进行统计推断,因此数据是统计分析的基础.然而,在获得数据的过程中,往往会因种种原因或多或少地夹杂进一些反常数值,通常称之为异常值,这些异常值将会不同程度地影响到统计推断的结果.有些统计方法有较强的抗干挠能力,少量的异常值不会对统计结果产生任何影响;有些统计方法则对异常值相当敏感,个别异常值就会使统计量的取值和统计推断结果发生较大的变化,以致导出不合理的甚至完全错误的结论.如果要在某种意义下对这些统计方法之优劣做出评价,用稳健统计的语言则称前者是稳健的,后者是不稳健的.

近半个世纪以来,各国统计学家为寻求稳健的统计方法及合理描述稳健方法的统计指标作了多方面的努力,众多成果相继问世.其中 F. R. Hampel[1] 1971 年给出了稳健性的一个严格定义,并提出了描述稳健性的两个重要概念:崩溃点和影响曲线,就崩溃点而言,它的含义是在不造成破坏影响的前提下,统计量所能容许的实际总体与假定模型之间的最大偏差. D. L. Donoho[2] 1983 年引进了样本崩溃点,此概念从大范围上描述统计量能承受多大比例的离群值.张健[3] 1996 年引进简化替换型样本崩溃点,由于直接与当前样本相联系,可将其看作是随机变量.本文将在文献[3]定义的基础上,计算出指数分布参数极大似然估计检验的样本崩溃点,并分析该样本崩溃点的渐近特性.

1 样本崩溃点的定义

简化替换型样本崩溃点的直观解释是:设数据集 $X=(X_1,X_2,\cdots,X_n)$ 的样本容量为 n,对该数据集中的最后 m 个观测值用任意的 m 个数值作替换,存在一个替换,使统计量的取值变得能逆转当前判决,且比值 $\frac{m}{n}$ 为最小,该比值即为简化替换型样本崩溃点.它的严格数学定义如下表述.

定义[3] 检验 $\phi(X)$ 在样本 X 处接受判决的简化替换型样本崩溃点为

$$\varepsilon_{SA}(X)=\frac{1}{n}\min\{m:0\leqslant m<n,\sup_{Y\in R^m}\phi(X_{n-m}\cup Y)=1\}.$$

检验 $\phi(X)$ 在样本 X 处拒绝判决的简化替换型样本崩溃点为

$$\varepsilon_{SR}(X)=\frac{1}{n}\min\{m:0\leqslant m<n,\inf_{Y\in R^m}\phi(X_{n-m}\cup Y)=0\},$$

其中, $X_{n-m}=(X_1,X_2,\cdots,X_{n-m})$,

$$\phi(X)=\begin{cases}0 & \text{(接受判决)},\\ 1 & \text{(拒绝判决)}.\end{cases}$$

收稿日期:2000 – 10 – 25;修订日期:2000 – 12 – 02

基金项目:河南省教委自然科学基金资助项目(1999110022)

作者简介:侯紫燕(1954 –),女,山西省清徐县人,河南纺织高等专科学校副教授,主要从事数理统计的理论与应用方面的研究.

2 样本崩溃点及其渐近正态性

设 $X_1, X_2, \cdots, X_n$, iid $\sim F_\lambda(x) = 1 - e^{-\lambda x}$, $(\lambda > 0, x \geq 0)$, 则 $\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} X_i}$ 为 λ 的极大似然估计, 容易证明

$$\sqrt{n}(\hat{\lambda} - \lambda) \xrightarrow{L} N(0, \lambda^2), \tag{1}$$

于是, 对于给定的检验水平 α $(0 < \alpha < 1)$, 存在 $\phi^{-1}(1-\alpha)$, 当 $\lambda = \lambda_0$ (λ_0 已知), 且样本容量 n 充分大时, 有

$$P\left(\sum_{i=1}^{n} X_i \geq \frac{n}{\lambda_0(1 + \phi^{-1}(1-\alpha)/\sqrt{n})} \right) = 1 - \alpha \tag{2}$$

成立.

现考虑原假设 $H_0: \lambda \leq \lambda_0$ (λ_0 已知), 对立假设 $H_1: \lambda > \lambda_0$, 注意到 $\hat{\lambda}$ 是 λ 的相合估计, 因此当 n 充分大时, 在 H_0 成立下, 应有

$$P\left(\sum_{i=1}^{n} X_i \geq \frac{n}{\lambda_0(1 + \phi^{-1}(1-\alpha)/\sqrt{n})} \right) \geq 1 - \alpha \tag{3}$$

成立.

于是可取 $C_n \cong \frac{n}{\lambda_0(1 + \phi^{-1}(1-\alpha)/\sqrt{n})}$, 则检验 H_0 是否成立可化为如下形式的检验问题

$$\phi(X) = \begin{cases} 0 & \sum_{i=1}^{n} X_i \geq C_n \\ 1 & \sum_{i=1}^{n} X_i < C_n \end{cases}, \tag{4}$$

这里, $\xrightarrow{L}$ 表示以分布收敛, $N(0,1)$ 为标准正态分布, $\phi^{-1}(x)$ 是标准正态分布函数 $\phi(x)$ 的反函数.

定理 设 $X = (X_1, X_2, \cdots, X_n)$ 是来自指数分布 $F_\lambda(x)$ 的简单随机样本, 由式(4)定义的检验 $\phi(X)$ 在样本 X 处的接受判决和拒绝判决的简化替换型样本崩溃点分别为

$$\varepsilon_{SA}(X) = \frac{1}{n}\min\left\{ m : 0 \leq m < n, \frac{m}{n} > 1 - \frac{1}{\lambda_0(1 + \phi^{-1}(1-\alpha)/\sqrt{n})} \left[\frac{1}{n-m} \sum_{i=1}^{n-m} X_i \right]^{-1} \right\}, \tag{5}$$

$$\varepsilon_{SR}(X) = \frac{1}{n}, \tag{6}$$

且当 $n \to \infty$ 时,

$$\varepsilon_{SA}(X) \to \frac{\lambda}{\lambda_0} = \varepsilon_{SA}^{*}(F_\lambda) \text{ a.s } (\lambda \leq \lambda_0), \tag{7}$$

$$\varepsilon_{SA}(X) \to 0 = \varepsilon_{SR}^{*}(F_\lambda) \quad \text{a.s} \quad (\lambda > \lambda_0), \tag{8}$$

在 H_0 成立的情况下,

$$\sqrt{n}[\varepsilon_{SA}(X) - \varepsilon_{SA}^{*}(F_\lambda)] \xrightarrow{L} N\left(0, \frac{\lambda}{\lambda_0}\right) \quad (0 < \lambda \leq \lambda_0). \tag{9}$$

3 主要结论的证明

式(5)和式(6)的思路相近, 就式(5)说明. 对于当前样本和给定的检验水平 α, 若 H_0 实际为真, 在正常情况下, 统计量 $\sum_{i=1}^{n} X_i$ 的取值会较大, 一般表现为 $\sum_{i=1}^{n} X_i \geq C_n$ 成立. 但是, 当样本因故遭到某种污染, 致使 m 个过小的数值混入其中(不妨设第 $n-m+1 \sim n$ 个观测值被过小数值代替), 从而统计量的值表现为 $\sum_{i=1}^{n-m} X_i + \sum_{i=n-m+1}^{n} X_i < C_n$, 从而拒绝接受 H_0. 我们称导致逆转当前判决的最少的异常值的个数 m 与样本容量 n 之比 $\frac{m}{n}$ 为 $\phi(X)$ 在 X 处的接受的样本崩溃点. 注意到 X_i 的可能取值范围, 从极端的情况考虑, 让样本中的后 m 个数据都用 0 代替, 可得结论. 又注意到

$$\sum_{i=1}^{n-m} X_i < \frac{n}{\lambda_0(1 + \phi^{-1}(1-\alpha)/\sqrt{n})} \tag{10}$$

等价于

$$\frac{m}{n} > 1 - \frac{1}{\lambda_0(1 + \phi^{-1}(1-\alpha)/\sqrt{n})} \cdot \left[\frac{1}{n-m} \sum_{i=1}^{n-m} X_i \right]^{-1}, \tag{11}$$

从而得到式(5).

关于式(7)的证明, 注意到结论式(5)中的 m 应同时满足

$$\sum_{i=1}^{n-m} X_i \leq \frac{n}{\lambda_0(1 + \phi^{-1}(1-\alpha)/\sqrt{n})} \tag{12}$$

和

$$\sum_{i=1}^{n-m+1} X_i \geq \frac{n}{\lambda_0(1 + \phi^{-1}(1-\alpha)/\sqrt{n})} \tag{13}$$

成立

利用 Kolmogorov 强大数定律和数学分析知识可知, 当 n 充分大时

$$\begin{cases} \frac{m}{n} \geq o(1) + 1 - \frac{1}{\lambda_0/\lambda(1 + \phi^{-1}(1-\alpha)/\sqrt{n})} & \text{a.s}; \\ \frac{m}{n} \leq o(1) + \left(1 + \frac{1}{n}\right) - \frac{1}{\lambda_0/\lambda(1 + \phi^{-1}(1-\alpha)/\sqrt{n})} & \text{a.s} \end{cases}$$

同时成立.

于是当 $n\to\infty$ 时,式(7)成立.

关于式(9)的证明,注意到式(10)和式(11)的等价性及式(12)和式(13)需同时成立,再根据如下引理

引理　(Anscombe[4]　1952)

设 $:X_1,X_2,\cdots,X_n$. iid ~ $N(0,1)$,再设存在另一正整数随机序列 $\{h_n\}$,使

$$P\left(\lim_{n\to\infty}\frac{h_n}{n}=\beta_0\right)=1\quad(\beta_0>0)$$

则 $\frac{1}{h_n}\sum_{i=1}^{h_n}X_i=\frac{1}{\sqrt{[n\beta_0]}}\sum_{i=1}^{[n\beta_0]}X_i+o_P(1)\xrightarrow{L}N(0,1)$,

令 $N(n-m)=\frac{1}{\sqrt{n-m}}\sum_{i=1}^{n-m}(X_i-E(X_i))$,

则 $N(n-m)\xrightarrow{L}N\left(0,\frac{1}{\lambda^2}\right)$.

因此,当 n 充分大时有

$$\frac{m}{n}\geq 1-\frac{1}{\lambda_0/\lambda(1+\phi^{-1}(1-\alpha)/\sqrt{n})}\cdot\left(1-\frac{\lambda}{\sqrt{n}\sqrt{1-\frac{m}{n}}}+O\left(\frac{1}{n}\right)\right)$$

和

$$\frac{m}{n}\leq 1-\frac{1}{\lambda_0/\lambda(1+\phi^{-1}(1-\alpha)/\sqrt{n})}\cdot\left(1-\frac{\lambda}{\sqrt{n}\sqrt{1-\frac{m}{n}+\frac{1}{n}}}+O\left(\frac{1}{n}\right)\right)$$

同时成立.

即当 $n\to\infty$ 时,式(9)成立.

参考文献:

[1] HAMPEL F R. A general qualitative definition of robustness[J]. Ann Math Statist,1971,42:1877 – 1896.

[2] DONOHO D L, HUBER P J. The notion of breakdown point[A]. Bickel P J. A Festschrift for Erich L Lehmann [C]. Belmont:Wadsworth,1983,157 – 184.

[3] ZHANG JIAN. The Sample breakdown points of teses[J]. Journal of Statistical Planning and Inference,1996,52:161 – 184.

[4] ANSCOMBE F. Large sample theory of sequential estimation[J]. Proc Combridge Phil Soc,1952,48:600 – 607

The Sample Breakdown Point of a Test for Maximum Likelihood Estimate of Exponential Distribution Parameter

HOU Zi – yan, YAN Guang – song, YUAN Xin – feng

(Department of Foundation, Henan Textile College, Zhengzhou 450007, China)

Abstract: The sample breakdown point of a test is defined as the smallest proportion of arbitrary outlier in the sample that reverses the test decision. In this paper, Wegive the sample breakdown point of a test for maximum likelihood estimate of exponential distribution parameter and analyze the asymptotically normal characteristic of the sample breakdown point.

Key words: exponential distribution; maximum likelihood estimate; sample breakdown point; asymptotically normal characteristic