测量数据中异常数据的检验比较*

宋宜容
(青海大学 水电系, 青海 西宁 810016)

[摘 要] 在简要介绍和分析一般测量数据检验处理的基础上, 对几种数据检验结果进行了比较, 介绍了一种具有抗差能力的样本分位数统计检验方法, 并举例进行了粗差检验说明。

[关键词] 测量数据; 异常数据; 粗差定位; 抗差估计
[中图分类号] O212.1 [文献标识码] A [文章编号]1000-2782(2002)03-0123-04

测量数据的异常数据检验是测量数据处理的重要组成部分。通常粗差是指比正常值大得多或小得多的异常数据。对这类异常数据, 必须设法在资料检查过程中加以删除, 以防给计算分析造成不良影响。尽管资料检查过程能发现和剔除部分粗差, 但仍需作一定的数据处理。处理方法归纳起来分为两大类, 一类是基于经典统计理论的误差统计检验法, 如文献[1, 2]介绍的各种异常值判别准则以及文献[3, 4]介绍的数据探测技术; 另一类是基于函数或统计推值的比较判别法。为了判别测量数据中的异常数据, 一般需要事先构造一些检验统计量, 这些统计量通常来源于经典的最小二乘法, 当观测值中存在粗差时, 最小二乘法的估计值是有偏的, 利用不可靠的估值来构造统计量并进行统计检验很难保证检验结果的可靠性。抗差能力的高低直接决定检验结果的可靠性水平, 也就可能出现所谓的异常值"遮蔽"现象[2]。由于最小二乘法估计不具有抗差性, 因此其估计值一般不宜用于构造检验统计量, 而必须用到抗差估计, 抗差估计中L 估计具有计算简单、直观明了的特点, 其中样本分位数是一类最简单、最基本的L 估计[5]。本研究简单介绍基于样本分位数检测异常值的基本原理, 将样本分位数法结果与文献[6]中讨论过的狄克松(Dixon)判别法、格拉布斯(Grubbs)判别法、莱以特判别法的检验结果进行比较, 为测量中异常数据的检验处理提供参考。

1 样本分位数统计检验法

设$V_1, V_2, \cdots, V_n$是取自正态总体$N(\mu, \sigma^2)$的一个样本, μ 表示总体均值, σ^2 表示总体方差。把样本观测的残差按大小顺序排列, 构成顺序统计量$V_{(1)} \leq V_{(2)} \leq \cdots \leq V_{(n)}$, 称$V_p$ 为样本 P 的分位, 即指

$$V_p = \begin{cases} V_{(np)}, & \text{当 } np \text{ 是整数时} \\ V_{[(np)+1]}。 & \text{当 } np \text{ 是非整数时} \end{cases} \quad (1)$$

式中, 符号[.]表示取不超过其变量本身的最大整数。如利用 1/4 样本分位数检测异常值, 是指取上四分位数和下四分位数的均值作为总体位置参数的估计, 取四分位数的离散度即上四分位数与下四分位数之差作为尺度参数的估计, 然后构造如下检验统计量。

$$S_n = \frac{V_{(n)} - \frac{1}{2}\{V_{(n_3)} + V_{(n_4)}\}}{V_{(n_3)} - V_{(n_4)}}, \quad (2)$$

$$S_1 = \frac{\frac{1}{2}\{V_{(n_3)} + V_{(n_4)}\} - V_{(1)}}{V_{(n_3)} - V_{(n_4)}}, \quad (3)$$

$$MRS = \max_{i=1,\cdots,n} \left| \frac{V_i - \frac{1}{2}\{V_{(n_3)} + V_{(n_4)}\}}{V_{(n_3)} - V_{(n_4)}} \right|。 \quad (4)$$

式中, $n_3 = \begin{cases} \frac{1}{4}n, & \text{当} \frac{1}{4}n \text{ 为整数时} \\ [\frac{1}{4}n] + 1; & \text{当} \frac{1}{4}n \text{ 不为整数时} \end{cases}$ (5)

$$n_4 = n - n_3 + 1。 \quad (6)$$

分别称S_n, S_1 和MRS 为总体方差 σ^2 未知时右侧、左侧和两侧异常值检验统计量。显然, 在这些统计量中用$\frac{1}{2}\{V_{(n_3)} + V_{(n_4)}\}$作为总体位置参数的估计, 用$V_{(n_3)} - V_{(n_4)}$作为尺度参数 σ的估计, 由于样本分位数具有较高的抗异常值污染的能力, 因此由其构成的检验统计量同样具有较强的抗差性, 也就是说, 这

* [收稿日期] 2001-11-06
[作者简介] 宋宜容(1964-), 女, 河南郑州人, 讲师, 主要从事水利工程测量的教学与研究。

些估计都具有抗异常值污染的能力。由(2)～(4)的定义可导出S_n, S_1和MRS变量的经验分布,并通过模拟方法求出在一定显著水平下检测异常值的临界值。文献[2]给出了$n=6$～35(步长为1)或$n=35$～100(步长为5),显著水平α分别取0.10,0.05和0.01时S_n, S_1和MRS的临界值。当由样本计算得到的S_n, S_1和MRS值大于其相应的临界值时,即可判定在给定的显著水平条件下被检测观测值为异常值。

2 数据检验示例

某GPS水准网的几何水准高程的拟合残差如表1。

表1 某GPS水准网几何水准高程的拟合残差值

Table 1 Fitting residual error value of geometrical level eleration for a GPS level net

Number	$V_{i(m)}$	$V_{(i)}$	$V^2_{(i)}$	$V'_{(i)}$	$V'^2_{(i)}$	$V''_{(i)}$	$V''^2_{(i)}$
1	0.270	-0.102	0.010 404	-0.040	0.001 600	-0.013	0.000 169
2	-0.002	-0.040	0.001 600	-0.013	0.000 169	-0.005	0.000 025
3	0.018	-0.013	0.000 169	-0.005	0.000 025	-0.002	0.000 004
4	0.008	-0.005	0.000 025	-0.002	0.000 004	-0.001	0.000 001
5	0.011	-0.002	0.000 004	-0.001	0.000 001	0.003	0.000 009
6	0.028	-0.001	0.000 001	0.003	0.000 009	0.004	0.000 016
7	0.012	0.003	0.000 009	0.004	0.000 016	0.008	0.000 064
8	-0.001	0.004	0.000 016	0.008	0.000 064	0.010	0.000 100
9	-0.102	0.008	0.000 064	0.010	0.000 100	0.011	0.000 121
10	0.003	0.010	0.000 100	0.011	0.000 121	0.012	0.000 144
11	0.018	0.011	0.000 121	0.012	0.000 144	0.018	0.000 324
12	0.004	0.012	0.000 144	0.018	0.000 324	0.018	0.000 324
13	0.010	0.018	0.000 324	0.018	0.000 324	0.027	0.000 729
14	-0.005	0.018	0.000 324	0.027	0.000 729	0.028	0.000 784
15	-0.013	0.027	0.000 729	0.028	0.000 784		
16	-0.040	0.028	0.000 784				

2.1 样本分位数统计法

以1/4样本分位值检验来计算对16个残差$V_{(i)}$的情况,由式(5)和(6)得

$$n_3=4,\quad n_4=14。$$

由式(3)得S_1,并从文献[2]附表中查得$S_1(16, 0.05)$的值

$$S_1=4.717>S_1(16,0.05)=2.051。$$

由式(2)得

$$S_{16}=0.652<S_1(16,0.05)=2.051。$$

故认为$V_{(1)}=0.102$是异常值。剔除$V_{(1)}$后,重新排序得$V'_{(1)}$,则由式(5)和(6)计算得

$$n_3=4,\quad n_4=12。$$

由式(3)得$S'_{(1)}=2.40$,查表得$S'_{(1)}(15, 0.05)=2.23$。

由式(2)得

$$S'_{(15)}=0.926<S'_{(1)}(15,0.05)=2.23。$$

故认为$V'_{(1)}=-0.040$也是异常值。剔除$V'_{(1)}$后,重新排序得$V''_{(1)}$,则由式(5)和(6)计算得

$$n_3=4,\quad n_4=11。$$

由式(2)和(3)计算得

$$S''_{(1)}=1.132<S''_{(1)}(14,0.05)=2.455,$$

$$S''_{(14)}=1.026<S''_{(14)}(14,0.05)=2.455,$$

故认为$V''_{(1)}$和$V''_{(14)}$都不是异常值。

2.2 狄克松(Dixon)判别法

首先判别最大残差(绝对值最大)$V_{(16)}$,因$n=16$,故取

$$r_{22}=\frac{V_{(16)}-V_{(14)}}{V_{(16)}-V_{(3)}}=\frac{0.028-0.018}{0.028+0.013}=0.244。$$

以$n=16$,$\alpha=0.05$为引数查表[2]可得$r_0(16, 0.05)=0.507$,因$r_{22}<r_0(16,0.05)$,故认为$V_{(16)}$相应的几何水准高程不是异常值。

其次判别最小残差$V_{(1)}=-0.102$,

$$r'_{22}=\frac{V_{(1)}-V_{(3)}}{V_{(1)}-V_{(14)}}=\frac{-0.102+0.013}{-0.102-0.018}=0.742,$$

则$r'_{22}>r_0(16,0.05)$,有理由认为$V_{(1)}$相应的几何水准高程是异常值。

舍去第一个残差$V_{(1)}$后,重新进行两端检验。仍然首先判别最大残差$V'_{(15)}$,

$$r_{22}=\frac{V'_{(15)}-V'_{(13)}}{V'_{(15)}-V'_{(3)}}=\frac{0.028-0.018}{0.028+0.005}=0.303,$$

因$r_{22}<r_0(15,0.05)=0.525$,故认为$V'_{(15)}$相应的几何水准高程不是异常值。同样再判别最小残差

$V'_{(1)} = -0.040$, 取

$$r'_{22} = \frac{V'_{(1)} - V'_{(3)}}{V'_{(1)} - V_{(13)}} = \frac{-0.040 + 0.005}{-0.040 - 0.018} = 0.603,$$

有 $r'_{22} > r_0(15, 0.05)$, 而 $r'_{22} < r_0(15, 0.01) = 0.616$, 这里可看出检验与显著水平有关。如剔除 $V'_{(1)}$ 后, 重新进行两端检验都不含有异常值。

2.3　格拉布斯(Grubbs)判别法

(1) 用格拉布斯判别法对 $V_{(i)}$ 计算得

$$\hat{\sigma} = \pm\sqrt{\frac{\sum_{i=1}^{16} V_{(i)}^2}{15}} = \pm 0.031\ 4,$$

$$\bar{V}_{(i)} = 0.001\ 5,$$

$$g_{(16)} = \frac{V_{(16)} - \bar{V}_{(i)}}{\hat{\sigma}} = 0.939。$$

以 $n = 16$, $\alpha = 0.05$ 为引数查表[2], 可得 $g_0(16, 0.05) = 2.44$, 即有 $g_{(16)} < g_0(16, 0.05)$, 故认为 $V_{(16)}$ 相应的几何水准高程不是异常值。

(2) 判别最小残差。

$$g_{(1)} = \frac{\bar{V}_{(i)} - V_{(1)}}{\hat{\sigma}} = -3.30,$$ 则 $|g_{(1)}| > g_0(16, 0.05)$, 故认为 $V_{(1)}$ 相应的几何水准高程是异常值。

(3) 舍去残差 $V_{(1)}$ 后, 按剩余的 15 个残差计算标准差得

$$\hat{\sigma}' = \pm\sqrt{\frac{\sum_{i=1}^{15} V_{(i)}^2}{14}} = \pm 0.017\ 8,$$

$$\bar{V}'_{(i)} = 0.005\ 3,$$

$$g_{(1)} = \frac{\bar{V}'_{(i)} - V'_{(1)}}{\hat{\sigma}'} = 2.54。$$

以 $n = 15$, $\alpha = 0.05$ 为引数查表[2], 可得 $g_0(15, 0.05) = 2.41$, 而 $g_{(15)} < g_0(15, 0.01) = 2.70$。

以相对显著水平 $\alpha = 0.05$ 或 $\alpha = 0.01$ 分别判断。假定再剔除 $V'_{(1)}$ 后按剩余的 14 个残差计算得

$$\hat{\sigma} = \pm\sqrt{\frac{\sum_{i=1}^{14} V_{(i)}^2}{n-1}} = \pm 0.014\ 7,$$

$$\bar{V}''_{(i)} = 0.008\ 6,$$

$$g_{(14)} = \frac{V''_{(14)} - \bar{V}''_{(i)}}{\sigma''} = 13.2 < g_0(14, 0.05),$$

$$g_{(14)} = \frac{\bar{V}''_{(i)} - V''_{(1)}}{\sigma''} = 2.37 < g_0(14, 0.05),$$

故认为 $V''_{(1)}$ 和 $V''_{(14)}$ 相应的几何水准高程不是异常值。

2.4　莱以特判别法

假定其残差落在 3 倍的标准差区间以外为粗差, 这时残差落在 3 倍的标准差区间以外的概率约为 0.27%, 对 16 个残差的情况, $\hat{\sigma} = \pm 0.031\ 4$, 首先判别绝对值最大的残差 $V_{(1)} = -0.102$, 而 $3\hat{\sigma} = 3(\pm 0.031\ 4) = \pm 0.094\ 2$, 因 $|V_{(1)}| > 3\hat{\sigma}$, 故认为 $V_{(1)}$ 相应的几何水准高程是异常值。舍去残差 $V_{(1)}$ 后, 按剩余的 15 个残差计算标准差得 $\hat{\sigma}' = \pm 0.017\ 8$, 再次判别绝对值最大的残差 $V'_{(1)} = -0.040$, 因 $|V'_{(1)}| < 3\hat{\sigma}'$, 故认为 $V'_{(1)}$ 相应的几何水准高程不是异常值。

$$\hat{\sigma} = \pm\sqrt{\frac{\sum_{i=1}^{16} V_{(i)}^2}{15}} = \pm 0.031\ 4,$$

$$\bar{V}_{(i)} = 0.001\ 5,$$

$$g_{(16)} = \frac{V_{(16)} - \bar{V}_{(i)}}{\sigma} = 0.939。$$

从这 4 种判别法可得出各自判别粗差的功效情况, 且都与显著水平有关。如 $V'_{(1)} = -0.040$ 的残差相对不同显著水平有不同的结果。在实际应用时最好能多种方法一齐使用, 综合分析作出结论。由于峰态和偏态统计量的检验效果较好, 但不能定位, 因此可先应用其判断数据中是否存在异常值, 然后再做检验。其统计量的计算公式可参考文献[2]。

3　结　论

从以上讨论可看出, 样本分位数统计检验具有计算简单、直观明了的特点; 同时具有一定的抗差能力。因为按最小二乘法求出的残差是统计相关, 上述讨论中没有使用剔除粗差后重新计算残差改正数来判断, 只作为一个例子来说明问题。实际应用时每一步应以重新平差的改正数来作检验, 同时还必须很好地研究引起残差的模型误差, 以及该模型误差与其他误差的可区分性。

[参考文献]

[1] 陈上及, 马继瑞. 海洋数据处理分析方法及其应用[M]. 北京: 海洋出版社, 1991. 74- 107.
[2] 张方仁, 张金通. 测量误差的统计分布和检验[M]. 北京: 中国计量出版社, 1992. 77- 144.
[3] 黄幼才. 数据探测与抗差估计[M]. 北京: 测绘出版社, 1990. 137- 190.

[4] 李德仁. 误差处理和可靠性理论[M]. 北京: 测绘出版社, 1988. 187- 203.
[5] 杨元喜. 抗差估计理论及其应用[M]. 北京: 八一出版社, 1993. 23- 31.
[6] 刘永明. GPS 水准的粗差检验[J]. 工程勘察, 1994, (6): 51- 54.

Comparasion of the anomaly data test in survey data

SONG Yi-rong
(*Department of Hydroelectcronical Qinghai University, Qinghai, Xining* 810016, *China*)

Abstract: Based on simple introduction and analysis of method of testing-processing for ordinary survey data, the paper compares several method of data testing, then it introduces a method of statistic-testing for the sample fraction which has the ability to resist error, and illuminates the method through an example's gross error testing.

Key words: survey data; data anomaly; gross error location; resisting error estimation

“西瓜一代杂种育种方法及‘西农 8 号’新品种选育”项目获国家科技进步二等奖

由西北农林科技大学王鸣教授带领课题组攻关的“西瓜一代杂种育种方法及‘西农 8 号’新品种选育”项目, 获 2001 年度国家科技进步二等奖。该项目标志着我国在为西瓜增产、增收提供新品种, 为消费者提供品质更优的商品西瓜科研工作达到了一个新的水平, 为我国西瓜产业迎接入世挑战作出了重大贡献。这是迄今我国西瓜研究中的最高奖项。

据联合国粮农组织(FAO)统计, 西瓜在世界 10 大果品中名列第五。西瓜在我国种植面积及产量均居世界第一, 年总产值达 150 亿元, 对于调整我国农村产业结构, 发展农村经济等方面具有重要作用。然而就科学技术而言, 我国并非“第一西瓜强国”, 特别是育种和品种水平低。20 世纪 80 年代, 进口品种长期主宰了我国西瓜产业, 其原因在于我国育种方法科技含量低, 缺乏创新的育种策略。因此, 要冲破西瓜育种的瓶颈, 取得突破性进展, 选育出真正超过进口西瓜的过硬品种, 必须在育种技术上有所创新和突破。

“西农 8 号”西瓜新品种于 1993 年通过陕西省品种审定, 其后又先后通过甘肃等多省的品种审定或认定, 1998 年通过全国品种审定, 1996 年荣获国家新产品证书, 1995 年在台湾举行的“海峡两岸西甜瓜育种研讨会”举办的两岸西瓜优良品种评比试验中被评为最优级。

该项目 1996 年获得国家发明专利, 1997 年获我国及世界知识产权组织(WIPO)联合签署颁发的中国专利发明创造金奖。2000 年又获中国专利十五周年最佳项目奖。其创新点在于, 综合采用常规育种(自交分离、杂交、抗病育种及杂种优势育种等)方法和多种现代育种(包括辐射育种、染色体工程及 RAPD 分子标记等)技术, 使其优点互相补充, 育成了高水平的西瓜新品种。同时, 该项目还采用了创新的或改进的抗病性鉴定和评价方法, 使抗病性鉴定和筛选的效果大为提高。在对育种种质资源和亲本材料在严格的人工控制条件下, 加大选择压力, 进行连续多代的苗期人工接种鉴定和筛选, 并首次创用西瓜炭疽病的离体叶人工接种鉴定技术和“AD 评价法”, 代替传统的“DI 评价法”, 省时、省工、省费用, 且结果准确可靠。此外, 在种质资源的鉴定、改造、创新方面, 亲本选择和选配等研究方面, 均有独到的创见。

查新报告结论表明, 该项育种技术成果在国内外均属于首创, 在我国西瓜育种历史上创造了新高度。“西农 8 号”是我国西瓜品种性状最全面, 适应范围最广, 推广面积最大, 经济、社会和生态效益最突出的名牌产品, 达到了国内领先水平。此外, 由于其抗病性强、耐重茬, 解决了西瓜连作障碍的难题, 且显著减少了农药用量, 降低了污染, 有利于保护环境和维护消费者身体健康, 创造出良好的生态效益、社会效益和经济效益。“西农 8 号”目前已在全国 20 多个省大面积推广, 累计创造经济效益 76.6 亿元, 基本上取代了进口西瓜品种, 节约了大量外汇开支, 为我国西瓜育种事业赢得了荣誉。