第 18卷 第 1期
1997年 3 月
郑州工业大学学报
Journal of Zhengzhou University of Technology
Vo 1.18 No .1.
Mar . 1997

拉依达 (PauTa) 准则与异常值剔除

张 敏 (郑州工业大学数力系)
袁 辉 (河南省计算中心)

摘 要: 针对用拉依达准则剔除异常值时要求有足够多的测量数据,且剔除方法较繁等问题,本文着重讨论了拉依达准则的适用条件,并予以证明;推导出变化的贝塞尔公式,为循环剔除异常值提供了简便方法。

关键词: 拉依达准则 异常值 残余误差 标准偏差
中图分类号: O 241.1

在工程、计量以及实验等测量过程中,由于测量者的一时疏忽,读错、记错测量数据或者操作不当、仪器突然失常等主客观因素,都会使测量数据含有粗大误差,从而严重影响测量精度,歪曲实验结果,因此,在处理测量数据时,应该把含有粗大误差的数值从中剔除。但是,判断一个数据是否含有粗大误差而决定取舍时,一定要特别慎重,要有充分的理论依据,否则会将反映客观事实的数据错误剔除,或将真正含有粗大误差的数据保留,造成人为地破坏测量数据的可靠性。下面针对判断测量数据是否含有粗大误差的拉依达准则进行较全面的讨论。

1 异常值和拉依达准则

用拉依达准则判断粗大误差的基本思想是以给定的置信概率 99.7% 为标准,以三倍测量列的标准偏差限为依据,凡超过此界限的误差,就认为它不属于随机误差的范畴,而是粗大误差。含有粗大误差的测量值称为异常值,异常值是不可取的[1],应该从测量数据中剔除。

用拉依达准则判断和剔除含有粗大误差的异常值时,应先算出等精度独立测量列 $X_i(i = 1, 2, \cdots, n)$ 的平均值 $\bar{x}$ 及残余误差 $\nu = x_i - \bar{x}$,并按贝塞尔公式算出该测量列的标准偏差 S,如果某测量值 X_d 的残余误差 $\nu = x_d - \bar{x}(1 \leqslant d \leqslant n)$ 满足下式

$$|\nu| > 3S \tag{1}$$

则认为 X_d 是含有粗大误差的异常值,须剔除不要[2]。该判别式即为拉依达准则。

2 拉依达准则的使用条件

用拉依达准则剔除异常值虽在科研、工程和教学上普遍采用,但对其使用范围和如何简便快捷地剔除异常值却很少涉及。为此,首先需要明确,该准则只有在测量次数 n 较大时才适用,至少应使 $n > 10$次才行,否则使用该准则无效。因为如果 $n \leqslant 10$时,即使测量列中存在含有粗大误差的异常值,也不能判断出来予以剔除。对此现证明如下:

已知对某物理量进行等精度独立测量,测得值为 $x_i(i = 1, 2, \cdots, n)$,其平均值为 $\bar{x}$,残余误

收稿日期: 1996-04-03

差为 $v_i = x_i - \bar{x}$，根据贝塞尔公式，可得测量列的标准偏差

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} v_i^2}$$

由此可得 $(n-1)S^2 = v_1^2 + v_2^2 + \cdots + v_n^2$

当 $n \leqslant 10$ 时，$n-1 \leqslant 9$，故上式中任何一个 v_i^2 均满足 $v_i^2 \leqslant 9S^2$

即 $$|v_i| \leqslant 3S \qquad (2)$$

由此可知，当测量次数 $n \leqslant 10$ 时，残余误差 $v_i (i = 1, 2, \cdots, n)$ 的绝对值均小于 $3S$，所以这时用拉依达准则是不能判断出含有粗大误差的异常值的。

例如对某物理量进行了十次等精度独立测量，所得数据如下

n	1	2	3	4	5	6	7	8	9	10
x_i	1.01	1.00	1.03	1.02	6.05	1.03	1.05	1.02	1.01	1.02

显然其中第五个数据 $x_5 = 6.05$ 是不合理的，应该属于异常值，现用拉依达准则予以判断。

$$\bar{x} = \frac{1}{10}\sum_{i=1}^{10} x_i = 1.52 \qquad S = \sqrt{\frac{1}{9}\sum_{i=1}^{10} v_i^2} = 1.59$$

$3S = 4.77$，而 x_5 的残余误差 $|v_5| = |6.05 - 1.52| = 4.53$，显然 $|v_5| < 3S$，故不能把含有粗大误差的异常值 x_5 判断出来，当然也就无法予以剔除。说明此时用拉依达准则无效。

当测量次数大于十时，这种方法还是比较好的，尤其是对测量数据中单个异常值的剔除，方法简单，且行之有效。比如对上例的物理量是等精度独立测量了十一次，数据为

n	1	2	3	4	5	6	7	8	9	10	11
x_i	1.01	1.00	1.03	1.02	6.05	1.03	1.05	1.02	1.01	1.02	1.04

现再来判断 $x_5 = 6.05$ 是否为异常值。

$$\bar{x} = \frac{1}{11}\sum_{i=1}^{11} x_i = 1.48 \qquad S = \sqrt{\frac{1}{10}\sum_{i=1}^{11} v_i^2} = 1.52$$

$3S = 4.56$，x_5 的残余误差 $|v_5| = |6.05 - 1.48| = 4.57$，$|v_5| > 3S$。满足拉依达准则，故 x_5 确为异常值，应当予以剔除。

在实际的测量数据中可能存在几个异常值，所以用拉依达准则时，测量次数要远大于十次，否则难以将异常值剔除尽。为此这种方法一般只有在 n 大于十三、四次时应用[3]。

3　判断与剔除异常值的简便方法

从上面的讨论中看出，每剔除一个异常值，就要重新计算一次 $n-1$ 个数据的平均值 $\bar{x}$、残余误差 v_i 及标准偏差 S，若测量列中有几个异常值需要剔除时，这样作显然是非常繁杂的。能否直接应用测量数据 x_i 来计算标准偏差 S，用某个固定数值 x_0 稍加修正来计算平均值 $\bar{x}$ 呢?回答是肯定的。只要将贝塞尔公式稍作变换，即可达到这一目的。

3.1 求平均值

通常可事先选取一个任意常数 x_0，该值可以是测量数据中的任意值，也可取与测量数据不同

的值，一般所选取的常数值 x_0 应与该测量列 x_i 的平均值相接近，并令

$$\delta = x_i - x_0$$

两边求和 $\sum_{i=1}^{n}\delta = \sum_{i=1}^{n}x_i - nx_0$

由平均值定义知

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n}x_i = x_0 + \frac{1}{n}\sum_{i=1}^{n}\delta \qquad (3)$$

由(3)式看出，求平均值时勿需直接应用测量数据，只要在原来所选取的常数值 x_0 上加一项 $\frac{1}{n}\sum_{i=1}^{n}\delta$ 即可。如果事先列出 $\delta = x_i - x_0$ 的数值，则该组数据的平均值就能很方便的算出。这正是循环剔除异常值时，计算各剩余数据的平均值所需要的。

3.2 推导变化的贝塞尔公式

若对某物理量进行等精度独立测量，测得数据 $x_i(i = 1, 2, \cdots, n)$，其平均值为 $\bar{x} = \frac{1}{n}\sum_{i=1}^{n}x_i$，各测量值的残余误差 $\upsilon = x_i - \bar{x}$。由贝塞尔公式可得该测量列的标准偏差

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad (4)$$

式中由于 $\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n}(x_i - \frac{1}{n}\sum_{i=1}^{n}x_i)^2$

$$= \sum_{i=1}^{n}\left[x_i^2 - \frac{2x_i}{n}\sum_{i=1}^{n}x_i + \frac{1}{n^2}(\sum_{i=1}^{n}x_i)^2\right]$$

$$= \sum_{i=1}^{n}x_i^2 - \sum_{i=1}^{n}\frac{2x_i}{n}\sum_{i=1}^{n}x + \frac{1}{n}(\sum_{i=1}^{n}x_i)^2$$

$$= \sum_{i=1}^{n}x_i^2 - \frac{2}{n}(\sum_{i=1}^{n}x_i)^2 + \frac{1}{n}(\sum_{i=1}^{n}x_i)^2$$

$$= \sum_{i=1}^{n}x_i^2 - \frac{1}{n}(\sum_{i=1}^{n}x_i)^2$$

故(4)式可写成如下形式

$$S = \sqrt{\frac{1}{n-1}\left[\sum_{i=1}^{n}x_i^2 - \frac{1}{n}(\sum_{i=1}^{n}x_i)^2\right]} \qquad (5)$$

由(5)式可知，标准偏差 S 只和测量数据 x_i 及其平方 x_i^2 两项有关，而与平均值 $\bar{x}$ 及残余误差 υ 无关，从而在循环剔除异常值时，计算各组数据的标准偏差 S 就显得比较方便。尽管如此，该式还可进一步变换，使计算方法更为简便。即

由于 $\delta = x_i - x_0$ 或 $x_i = \delta + x_0$

$$x_i^2 = \delta^2 + 2\delta x_0 + x_0^2$$

两边求和 $\sum_{i=1}^{n}x_i^2 = \sum_{i=1}^{n}\delta^2 + 2x_0\sum_{i=1}^{n}\delta + nx_0^2 \qquad (6)$

由(3)式知 $\bar{x} = x_0 + \frac{1}{n}\sum_{i=1}^{n}\delta = \frac{1}{n}\sum_{i=1}^{n}x_i$

则 $x_0 = \frac{1}{n}\sum_{i=1}^{n}x_i - \frac{1}{n}\sum_{i=1}^{n}\delta$

将x_0代入(6)式,整理可得

$$\sum_{i=1}^{n}x_i^2-\frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)^2=\sum_{i=1}^{n}\delta_i^2-\frac{1}{n}\left(\sum_{i=1}^{n}\delta_i\right)^2$$

与(5)式比较可得

$$S=\sqrt{\frac{1}{n-1}\left[\sum_{i=1}^{n}\delta_i^2-\frac{1}{n}\left(\sum_{i=1}^{n}\delta_i\right)^2\right]}\quad[4] \tag{7}$$

该式即为变化的贝塞尔公式,其优点在于求标准偏差更为方便,并且在计算过程中不会因算平均值除不尽而产生舍入误差,数字本身的计算可以精确到任意位。

总之,只要给出所选定的任意值x_0和测量数据x_i,利用(3)、(7)两式就可很方便地算出测量列的平均值$\bar{x}$和标准偏差S,从而实现数据的及时处理和异常值的循环剔除。

4　例题分析

测某一长度15次,得l_i值如下:(单位:cm)

16.42	16.43	16.40	16.44	16.42	16.42	16.39	16.43
16.30	16.40	16.41	16.42	16.41	16.40	16.40	

用拉依达准则判断与剔除其中的异常值。

解:选取$l_0=16.38(cm)$,则$\delta_i=l_i-16.38(cm)$

为计算方便,将l_i、δ_i和δ_i^2各值列表如下

n	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$l_i(cm)$	16.42	16.43	16.40	16.44	16.42	16.42	16.39	16.43	16.30	16.40	16.41	16.42	16.41	16.40	16.40
$\delta_i(cm)$	0.04	0.05	0.02	0.06	0.04	0.04	0.01	0.05	−0.08	0.02	0.03	0.04	0.03	0.02	0.02
$\delta_i^2(cm)^2$	0.0016	0.0025	0.0004	0.0036	0.0016	0.0016	0.0001	0.0025	0.0064	0.0004	0.0009	0.0016	0.0009	0.0004	0.0004

则$\sum_{i=1}^{15}\delta_i=0.39(cm)$,$\sum_{i=1}^{15}\delta_i^2=0.0249(cm)^2$

$$S=\sqrt{\frac{1}{n-1}\left[\sum_{i=1}^{15}\delta_i^2-\frac{1}{n}\left(\sum_{i=1}^{15}\delta_i\right)^2\right]}=0.033(cm)$$

$$3S=3\times0.033=0.099(cm)$$

$$l=l_0+\frac{1}{n}\sum_{i=1}^{15}\delta_i=16.406(cm)$$

在该组数据中,$l_9=16.30$偏小,应先判断它是否为异常值。

$$|v_9|=|l_9-l|=|16.30-16.406|=0.106(cm)$$

$|v_9|>3S$,故$l_9=16.30$为异常值,予以剔除。然后再对剩余的十四个数据进行异常值的判断,这时

$$\sum_{\substack{i=1\\ i\neq9}}^{15}\delta_i=\sum_{i=1}^{15}\delta_i-\delta_9=0.39-(-0.08)=0.47(cm)$$

$$\sum_{\substack{i=1\\ i\neq 9}}^{15}\delta_i^2=\sum_{i=1}^{15}\delta_i^2-\delta_9^2=0.0249-0.0064=0.0185(cm)^2$$

则 $$S'=\sqrt{\frac{1}{13}\left[\sum_{\substack{i=1\\ i\neq 9}}^{15}\delta_i^2-\frac{1}{14}\left(\sum_{\substack{i=1\\ i\neq 9}}^{15}\delta_i\right)^2\right]}=0.015(cm)$$

$$3S'=3\times 0.015=0.045(cm)$$

$$l'=l_0+\frac{1}{14}\sum_{\substack{i=1\\ i\neq 9}}^{15}\delta_i=16.414(cm)$$

剔除l_9后，剩余数据中l_4和l_7分别为最大值和最小值，应判断这两个数据是否为异常值。

$$|v_4|=|16.44-16.414|=0.026,\quad |v_4|<3S'$$

$$|v_7|=|16.39-16.414|=0.024,\quad |v_7|<3S'$$

v_4和v_7这两个数据都是正常值，故位于这两个极值之间的其它数据也都为正常值。

由此看出，当剔除某一数值x_d后，因为不影响其它测量数据$x_i(i\neq d)$及x_0值的大小，所以不用改变表中δ_i和δ_i^2的值，只要从原来求和的$\sum_{i=1}^{n}\delta_i$及$\sum_{i=1}^{n}\delta_i^2$两项中分别减去δ_d和δ_d^2值，即可进行$n-1$个数据的标准偏差S'及平均值$\bar{x}'$的计算。然后再利用拉依达准则判断剩余数据是否还有异常值，这样周而复始，直至把异常值循环剔除干净，非常简便。

参考文献

1 刘智敏 .误差与数据处理 .原子能出版社 .1981
2 肖明耀 .误差理论与应用 .计量出版社 .1985
3 孟尔熹(12)曹尔第 .实验误差与数据处理 .上海科学技术出版社 .1988
4 龚镇雄 .普通物理实验中的数据处理 .西北电讯工程学院出版社 .1985

The PauTa Criterion and Rejecting the Abnormal Value

Zhang Min Yuan Hui
(Zhengzhou University of Technology) (Henan Calculation Center)

Abstract As it is troublesome that using the PauTa criterion to reject the abnormal value and much enough measured data is required, the condition of application about the PauTa criterion is discussed and verified in this paper. The reformed Bessel formula have been developed. The paper provides a simple method for rejecting the abnormal value with round-Robin.

Keywords PauTa criterion Abnormal value Retained error Standard deviation