

数据的统计处理和解释

正态样本异常值的判断和处理

Statistical interpretation of data—Detection and handling
of outlying observations in normal sample

1 引言

1.1 本标准规定了判断和处理在正态样本中出现的异常值的一般原则和实施办法。

1.2 异常值(或异常观测值)是指样本中的个别值,其数值明显偏离它(或它们)所属样本的其余观测值。

异常值可能是总体固有的随机变异性的极端表现。这种异常值和样本中其余观测值属于同一总体。

异常值也可能是由于试验条件和试验方法的偶然偏离所产生的后果,或产生于观测、计算、记录中的失误。这种异常值和样本中其余观测值不属于同一总体。

1.3 本标准使用的其它统计学名词,见国家标准GB 3358—82《统计学名词及符号》。

1.4 应用条件:所考查样本中诸观测值(或经过一定的函数变换后得到的值),除了个别异常值外,其余大部分值(样本主体)来自同一正态总体或近似正态总体。

关于样本来自正态总体或近似正态总体的判断,可以根据物理上的、技术上的知识;也可通过与考查对象有同样性质的以往数据,进行正态性检验,其原理和方法见国家标准GB 4882—85《数据的统计处理和解释—正态性检验》。

2 判断异常值的统计学原则

2.1 本标准在下述不同情形下判断样本中的异常值:

上侧情形:根据以往经验,异常值都为高端值;

下侧情形:根据以往经验,异常值都为低端值;

双侧情形:异常值是在两端都可能出现的极端值。

注:上侧情形和下侧情形统称单侧情形。

2.2 执行本标准时,应规定在样本中检出异常值的个数的上限(占样本观测值个数的较小比例),当超过了这个上限,对此样本的代表性,应作慎重的研究和处理。

2.3 判断单个异常值的检验规则

根据实际情况,选定适宜的异常值检验规则(见4、5、6章);

指定为检出异常值的统计检验的显著性水平 α ,简称检出水平;

根据 α 和观测值个数 n 确定统计量的临界值;

将各观测值代入检验规则中给出的统计量,所得值若超过临界值,则判断事先确定待查的极端观测值为异常值;否则就判断“没有异常值”。

检出水平 α 的宜取值是5%,1%(或10%)。

2.4 判断多个异常值的检验规则

在允许检出异常值个数可大于1的情形,本标准规定的方法是重复使用同一种判断单个异常值的检验规则,即用指定的检出水平和符合2.3规定的规则首先检验全体观测值,若不能检出异常值,则整个检验停止;若检出了一个异常值,就再用相同的检出水平和相同的规则,对除去已检出的异常值后余下的观测值继续检验……直到不能检出异常值,或检出的异常值个数超过上限为止。

3 处理异常值的一般规则

3.1 对检出的异常值, 应尽可能寻找产生异常值的技术上的、物理上的原因, 作为处理异常值的依据。

3.2 处理异常值的方式有:

- 异常值保留在样本中参加其后的数据分析;
- 允许剔除异常值, 即把异常值从样本中排除;
- 允许剔除异常值, 并追加适宜的观测值计入样本;
- 在找到实际原因时修正异常值。

3.3 标准使用者应根据实际问题的性质, 权衡寻找产生异常值原因的花费, 正确判断异常值的得益及错误剔除正常观测值的风险, 确定实施下述三个规则中的一个。

- a. 对任何异常值, 若无充分的技术上的、物理上的说明其异常的理由, 则不得剔除或进行修正。
- b. 异常值中除有充分的技术上的、物理上的说明其异常的理由者外, 表现统计上高度异常的, 也允许剔除或进行修正, 其意义是:

指定为判断异常值是否高度异常的统计检验的显著性水平 α^* , 简称剔除水平, 其值小于检出水平 α ;

实施时, 按2.3规定进行检验后, 立即对检出的异常值, 再按2.3规定以剔除水平 α^* 代替检出水平 α 进行检验, 若在剔除水平下此检验是显著的, 则判此异常值表现高度异常。

在重复使用同一检验规则的情况下, 每次检出了异常值后都要再检验它在剔除水平下是否高度异常。若某次检验中检出的异常值为高度异常, 则这个异常值及在它前面检出的异常值都可被剔除或进行修正。

除特殊情况外, 剔除水平一般采用1%或更小, 而不宜采用大于5%的值。

在选用剔除水平的情况下, 检出水平可取5%或再大些。

- c. 检出的异常值都可被剔除或进行修正。

3.4 被检出的异常值, 被剔除或修正的观测值及其理由, 应予记录以备查询。

4 已知标准差情形下判断和处理异常值的规则

4.1 本章规定使用奈尔(Nair)检验法或奈尔检验法的重复使用。

4.1.1 上侧情形的检验法

- a. 对于按大小排列的观测值 $x_{(1)} < x_{(2)} < \dots < x_{(n)}$, 计算统计量

$$R_n = (x_{(n)} - \bar{x}) / \sigma$$

这里 σ 是已知的总体标准差, $\bar{x}$ 是样本均值。

- b. 确定检出水平 α , 在表A1查出对应 n 、 α 的临界值 $R_{1-\alpha}(n)$ 。
 - c. 当 $R_n > R_{1-\alpha}(n)$, 判断最大值 $x_{(n)}$ 为异常值, 否则, 判断“没有异常值”。
 - d. 在给出剔除水平 α^* 的情况下, 在表A1查出对应 n 、 α^* 的临界值 $R_{1-\alpha^*}(n)$ 。
- 当 $R_n > R_{1-\alpha^*}(n)$, 判断 $x_{(n)}$ 为高度异常; 否则, 判断“没有高度异常的异常值”。

4.1.2 下侧情形的检验法

与4.2.1规则相同, 但要使用统计量

$$R'_n = (\bar{x} - x_{(1)}) / \sigma$$

代替 R_n , 要判断的是最小值 $x_{(1)}$ 。

4.1.3 双侧情形的检验法

- a. 计算 R_n 与 R'_n 的值;
- b. 确定检出水平 α , 在表A1查出对应 n 、 $\alpha/2$ 的临界值 $R_{1-\alpha/2}(n)$;

c. 当 $R_n > R'_n$, 且 $R_n > R_{1-\alpha/2}(n)$, 判断最大值 $x_{(n)}$ 为异常值; 当 $R'_n > R_n$, 且 $R'_n > R_{1-\alpha/2}(n)$, 判断最小值 $x_{(1)}$ 为异常值; 否则, 判断“没有异常值”。

d. 在给出剔除水平 α^* 的情况下, 在表 A1 查出对应 n , $\alpha^*/2$ 的临界值 $R_{1-\alpha^*/2}(n)$ 。

当 $R_n > R'_n$, 且 $R_n > R_{1-\alpha^*/2}(n)$, 判断最大值 $x_{(n)}$ 为高度异常; 当 $R'_n > R_n$, 且 $R'_n > R_{1-\alpha^*/2}(n)$, 判断最小值 $x_{(1)}$ 为高度异常; 否则, 判断“没有高度异常的异常值”。

4.2 使用奈尔检验法的示例:

考查某种化纤纤维干收缩率, 得25个独立观测值: 3.13, 3.49, 4.01, 4.48, 4.61, 4.76, 4.98, 5.25, 5.32, 5.39, 5.42, 5.57, 5.59, 5.59, 5.63, 5.63, 5.65, 5.66, 5.67, 5.69, 5.71, 6.00, 6.03, 6.12, 6.76, (单位%)。已知在正常条件下, 测试量服从正态分布, $\sigma = 0.65$, 现考查下侧的异常值。

规定至多检出三个异常值, 采用3.3中b的处理方式。取检出水平 $\alpha = 5\%$, 剔除水平 $\alpha^* = 1\%$ 。

对 $n = 25$, 得 $\bar{x} = 5.2856$, $R'_{25} = (\bar{x} - x_{(1)})/\sigma = (5.2856 - 3.13)/0.65 = 3.316$ 。而 $R_{0.95}(25) = 2.815$, $R_{0.99}(25) = 3.282$, $R'_{25} > R_{0.99}(25)$, 故判断3.13是高度异常的异常值。

取出3.13后在余下的24个观测值中计算均值 $\bar{x} = 5.375$, 这时最小值为3.49, $R'_{24} = (5.375 - 3.49)/0.65 = 2.90$ 。对 $n = 24$, $R_{0.95}(24) = 2.800$, $R_{0.99}(24) = 3.269$, 因 $R'_{24} > R_{0.95}(24)$, 判断3.49是异常值。

取出3.13、3.49后, 余下23个观测值的样本均值为5.457, 这时最小值为4.01, $R'_{23} = (5.457 - 4.01)/0.65 = 2.227$ 。对 $n = 23$, $R_{0.95}(23) = 2.784$, 因 $R'_{23} < R_{0.95}(23)$, 故判断“没有异常值”。

检出3.13和3.49是异常值, 其中3.13高度异常, 可考虑剔除。

5 未知标准差情形下判断和处理异常值的规则 (I)

——检出异常值的个数不超过1

5.1 本章给出格拉布斯 (Grubbs) 检验法和狄克逊 (Dixon) 检验法, 标准使用者可根据实际要求选定实施其中一种检验法 (参考附录B)。

5.2 格拉布斯检验法

5.2.1 上侧情形的检验法

a. 对于观测值 $x_1, \dots, x_n$, 计算统计量

$$G_n = (x_{(n)} - \bar{x})/s$$

的值, 这里 $x_{(n)}$ 是最大观测值, $\bar{x}$ 和 s 是样本均值和样本标准差, 即 $\bar{x} = (x_1 + \dots + x_n)/n$,

$$s = \left[\frac{1}{n-1} \left(\sum_{i=1}^n x_i^2 - n\bar{x}^2 \right) \right]^{1/2};$$

b. 确定检出水平 α , 在表 A2 查出对应 n , α 的临界值 $G_{1-\alpha}(n)$;

c. 当 $G_n > G_{1-\alpha}(n)$, 判最大值 $x_{(n)}$ 为异常值; 否则, 判断“没有异常值”;

d. 在给出剔除水平 α^* 的情况下, 在表 A2 查出对应 n , α^* 的临界值 $G_{1-\alpha^*}(n)$ 。

当 $G_n > G_{1-\alpha^*}(n)$, 判 $x_{(n)}$ 高度异常; 否则, 判断“没有高度异常的异常值”。

5.2.2 下侧情形的检验法

与5.2.1规则相同, 但要使用统计量

$$G'_n = (\bar{x} - x_{(1)})/s$$

代替 G_n , 要判断的是最小观测值 $x_{(1)}$ 。

5.2.3 双侧情形的检验法

a. 计算 G_n 和 G'_n 的值;

b. 确定检出水平 α , 在表 A2 查出对应 n , $\alpha/2$ 的临界值 $G_{1-\alpha/2}(n)$;

c. 当 $G_n > G'_n$, 且 $G_n > G_{1-\alpha/2}(n)$, 判断 $x_{(n)}$ 为异常值; 当 $G'_n > G_n$, 且 $G'_n > G_{1-\alpha/2}(n)$, 判断 $x_{(1)}$ 为异常值; 否则, 判断“没有异常值”;

d. 在给出剔除水平 α^* 的情况下, 在表 A2 查出对应 n , $\alpha^*/2$ 的临界值 $G_{1-\alpha^*/2}(n)$ 。

当 $G_n > G'_n$, 且 $G_n > G_{1-\alpha^*/2}(n)$, 判断 $x_{(n)}$ 为高度异常; 当 $G'_n > G_n$, 且 $G'_n > G_{1-\alpha^*/2}(n)$, 判断 $x_{(1)}$ 为高度异常; 否则, 判断“没有高度异常的异常值”。

5.2.4 使用格拉布斯检验法的示例

检验某种砖的一个交付批的10个样品的抗压强度数据 (自小而大排列) 4.7, 5.4, 6.0, 6.5, 7.3, 7.7, 8.2, 9.0, 10.1, 14.0 (单位: MPa)。

检验最大值是否异常值, 取检出水平 $\alpha = 5\%$ 。

计算 $\bar{x} = (4.7 + 5.4 + 6.0 + 6.5 + 7.3 + 7.7 + 8.2 + 9.0 + 10.1 + 14.0) / 10 = 7.89$

$$s^2 = [(4.7 - 8)^2 + (5.4 - 8)^2 + (6.0 - 8)^2 + (6.5 - 8)^2 + (7.3 - 8)^2 + (7.7 - 8)^2 + (8.2 - 8)^2 + (9.0 - 8)^2 + (10.1 - 8)^2 + (14.0 - 8)^2 - 10(8 - 7.89)^2] / 9 = 7.312$$

$$s = 2.704$$

(计算 s 时把各观测值减 8 是为了简化计算)。

$$G_{10} = (x_{(10)} - \bar{x}) / s = (14 - 7.89) / 2.704 = 2.260$$

对 $n = 10$, $G_{0.95}(10) = 2.176$, 因 $G_{10} > G_{0.95}(10)$, 判断 $x_{(10)} = 14$ 为异常值。

5.3 狄克逊检验法

5.3.1 单侧情形的检验法

a. 对于按大小排列的观测值 $x_{(1)} < x_{(2)} < \dots < x_{(n)}$, 计算统计量

样本大小	检验高端异常值	检验低端异常值
$n: 3 \sim 7$	$D = r_{10} = \frac{x_{(n)} - x_{(n-1)}}{x_{(n)} - x_{(1)}}$	$D' = r'_{10} = \frac{x_{(2)} - x_{(1)}}{x_{(n)} - x_{(1)}}$
$n: 8 \sim 10$	$D = r_{11} = \frac{x_{(n)} - x_{(n-1)}}{x_{(n)} - x_{(2)}}$	$D' = r'_{11} = \frac{x_{(2)} - x_{(1)}}{x_{(n-1)} - x_{(1)}}$
$n: 11 \sim 13$	$D = r_{21} = \frac{x_{(n)} - x_{(n-2)}}{x_{(n)} - x_{(2)}}$	$D' = r'_{21} = \frac{x_{(3)} - x_{(1)}}{x_{(n-1)} - x_{(1)}}$
$n: 14 \sim 30$	$D = r_{22} = \frac{x_{(n)} - x_{(n-2)}}{x_{(n)} - x_{(3)}}$	$D' = r'_{22} = \frac{x_{(3)} - x_{(1)}}{x_{(n-2)} - x_{(1)}}$

b. 确定检出水平 α , 在表 A3 查出对应 n , α 的临界值 $D_{1-\alpha}(n)$;

c. 检验高端值时, 当 $D > D_{1-\alpha}(n)$, 判断 $x_{(n)}$ 为异常值; 检验低端值时, 当 $D' > D_{1-\alpha}(n)$, 判断 $x_{(1)}$ 为异常值; 否则, 判断“没有异常值”;

d. 在给出剔除水平 α^* 的情况下, 在表 A3 查出对应 n , α^* 的临界值 $D_{1-\alpha^*}(n)$ 。

检验高端值时, 当 $D > D_{1-\alpha^*}(n)$, 判断 $x_{(n)}$ 为高度异常; 检验低端值时, 当 $D' > D_{1-\alpha^*}(n)$, 判断 $x_{(1)}$ 为高度异常; 否则, 判断“没有高度异常的异常值”。

5.3.2 双侧情形的检验法

a. 计算 D 与 D' 的值, 这里 D 与 D' 由 5.3.1 的 a 给出;

b. 确定检出水平 α , 在表 A3' 查出对应 n , α 的临界值 $\tilde{D}_{1-\alpha}(n)$;

c. 当 $D > D'$, $D > \tilde{D}_{1-\alpha}(n)$, 判断 $x_{(n)}$ 为异常值; 当 $D' > D$, $D' > \tilde{D}_{1-\alpha}(n)$, 判断 $x_{(1)}$ 为异常值; 否则, 判断“没有异常值”。

d. 在给出剔除水平 α^* 的情况下, 在表 A3' 查出对应 n , α^* 的临界值 $\tilde{D}_{1-\alpha^*}(n)$ 。

当 $D > D'$, $D > \tilde{D}_{1-\alpha^*}(n)$, 判断 $x_{(n)}$ 为高度异常; 当 $D' > D$, $D' > \tilde{D}_{1-\alpha^*}(n)$, 判断 $x_{(1)}$ 为高度异常; 否则, 判断“没有高度异常的异常值”。

5.3.3 使用狄克逊检验法的示例

射击16发子弹, 射程(自小而大排列)分别为1125, 1248, 1250, 1259, 1273, 1279, 1285, 1285, 1293, 1300, 1305, 1312, 1315, 1324, 1325, 1350 (单位: m)。

a. 检验低端值是否异常值。指定 $\alpha = 1\%$

对 $n = 16$, 使用

$$D' = r'_{22} = \frac{x_{(3)} - x_{(1)}}{x_{(14)} - x_{(1)}} = \frac{1250 - 1125}{1324 - 1125} = \frac{125}{189} = 0.6614$$

因 $D_{0.99}(16) = 0.595$, $D' > D_{0.99}(16)$, 故判断最小值1125为异常值;

b. 双侧情形

对 $n = 16$, 计算 $D' = 0.6614$ 和

$$D = r_{22} = \frac{x_{(16)} - x_{(14)}}{x_{(16)} - x_{(3)}} = \frac{1350 - 1324}{1350 - 1250} = \frac{26}{100} = 0.26$$

查表 A3' 得 $\tilde{D}_{0.99}(16) = 0.627$ 。

因 $r'_{22} > r_{22}$, $r'_{22} > \tilde{D}_{0.99}(16)$, 故判断最小值1125为异常值。

6 未知标准差情形下判断和处理异常值的规则 (II)

——检出异常值的个数上限大于1。

6.1 本章给出偏度—峰度检验法和狄克逊检验法的重复使用方法, 标准使用者可根据实际要求选定实施其中一种检验法 (参考附录 B)。

6.2 偏度—峰度检验法

6.2.1 使用条件: 考查样本诸观测值, 确认它们的样本主体来自正态总体, 而极端值应较明显的偏离样本主体。

6.2.2 单侧情形——偏度检验法

a. 对于观测值 $x_1, x_2, \dots, x_n$, 计算偏度统计量

$$b_s = \frac{\sqrt{n} \sum_{i=1}^n (x_i - \bar{x})^3}{\left[\sum_{i=1}^n (x_i - \bar{x})^2 \right]^{3/2}} = \frac{\sqrt{n} \left[\sum_{i=1}^n x_i^3 - 3\bar{x} \sum_{i=1}^n x_i^2 + 2n(\bar{x})^3 \right]}{\left[\sum_{i=1}^n x_i^2 - n\bar{x}^2 \right]^{3/2}}$$

b. 确定检出水平 α , 在表 A4 查出对应 n , α 的临界值 $b'_{1-\alpha}(n)$;

c. 对上侧情形, 当 $b_s > b'_{1-\alpha}(n)$, 判断最大值 $x_{(n)}$ 为异常值; 否则, 判断“没有异常值”。对下侧情形, 当 $-b_s > b'_{1-\alpha}(n)$, 判断最小值 $x_{(1)}$ 为异常值; 否则, 判断“没有异常值”。

d. 在给出剔除水平 α^* 的情况下, 在表 A4 查出对应 n , α^* 的临界值 $b_{1-\alpha^*}(n)$;

对上侧情形, 当 $b_s > b_{1-\alpha^*}(n)$, 判断 $x_{(n)}$ 为高度异常; 对下侧情形, 当 $-b_s > b_{1-\alpha^*}(n)$, 判断 $x_{(1)}$ 为高度异常; 否则, 判断“没有高度异常的异常值”。

6.2.3 双侧情形——峰度检验法

a. 对于观测值 $x_1, x_2, \dots, x_n$, 计算峰度统计量

$$b_k = \frac{n \sum_{i=1}^n (x_i - \bar{x})^4}{\left[\sum_{i=1}^n (x_i - \bar{x})^2 \right]^2} = \frac{n \left[\sum_{i=1}^n x_i^4 - 4\bar{x} \sum_{i=1}^n x_i^3 + 6\bar{x}^2 \sum_{i=1}^n x_i^2 - 3n\bar{x}^4 \right]}{\left[\sum_{i=1}^n x_i^2 - n\bar{x}^2 \right]^2}$$

- b. 确定检出水平 α , 在表 5 查出对应 n , α 的临界值 $b'_{1-\alpha}(n)$;
- c. 当 $b_k > b'_{1-\alpha}(n)$, 判断离均值 $\bar{x}$ 最远的观测值为异常值; 当 $b_k \leq b'_{1-\alpha}(n)$, 判断“没有异常值”;
- d. 在给出剔除水平 α^* 的情况下, 在表 A5 查出对应 n , α^* 的临界值 $b'_{1-\alpha^*}(n)$;
- 当 $b_k > b'_{1-\alpha^*}(n)$, 判断离均值 $\bar{x}$ 最远的观测值为高度异常, 否则, 判断“没有高度异常的异常值”。

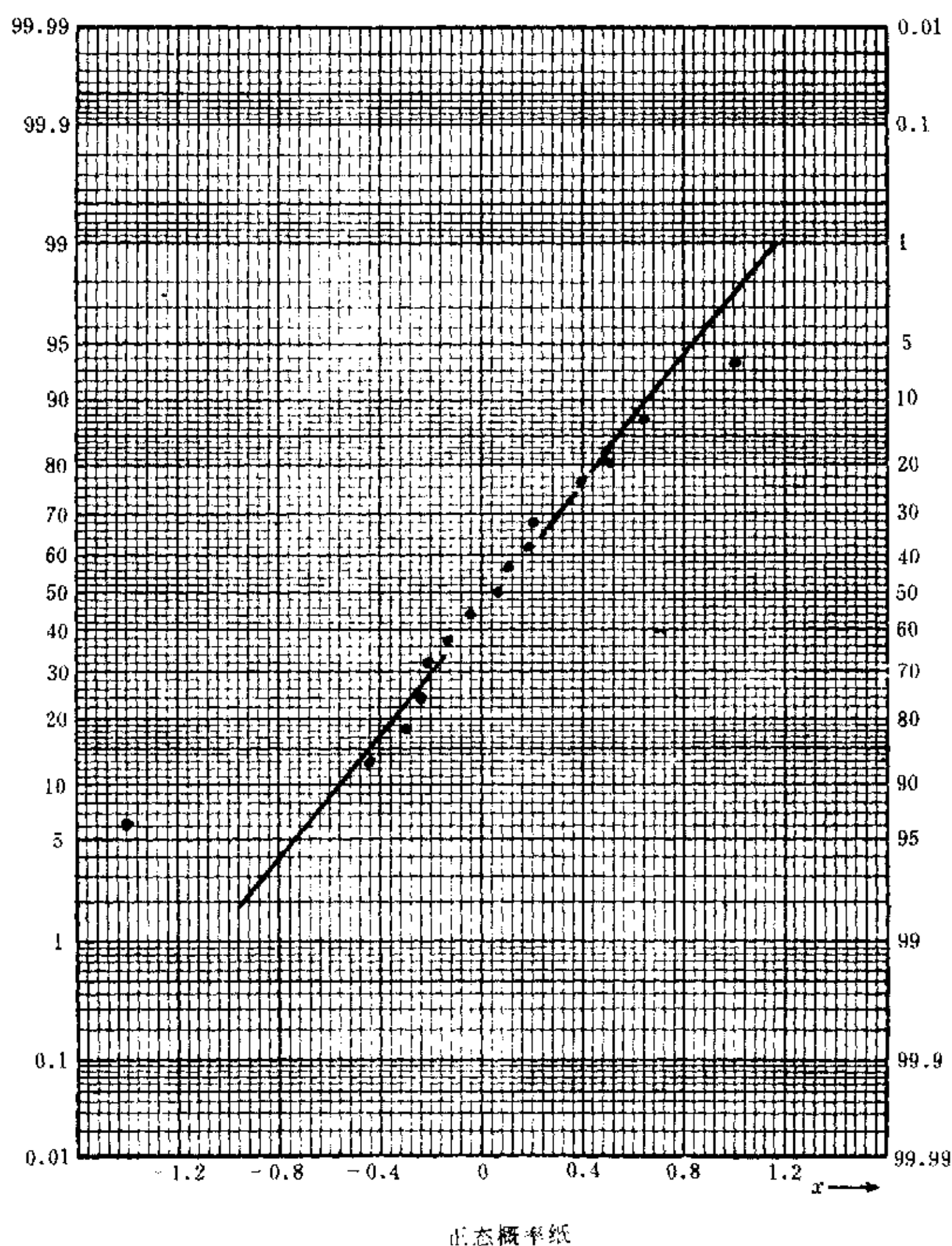
6.2.4 重复使用峰度检验法的示例。

异常值问题早期研究中的著名实例 (1883 年), 对观测金星垂直半径的 15 个观测数据的残差: (单位: 秒)。

-1.40, -0.44, -0.30, -0.24, -0.22, -0.13, -0.05, 0.06, 0.10, 0.18, 0.20, 0.39, 0.48, 0.63, 1.01。

要判断 -1.40 和 1.01 是否异常。

首先考查使用条件, 用正态概率纸 (用法见国家标准 GB 4882—85 《正态性检验》)



点在正态概率纸上的诸点, 样本主体在图上近似在一条直线近旁, 当画出适宜的直线后, 样本一端或两端的个别点明显向外偏离, 故可用偏度—峰度检验法。

计算得

$$\sum_{i=1}^{15} x_i \quad \sum_{i=1}^{15} x_i^2 \quad \sum_{i=1}^{15} x_i^3 \quad \sum_{i=1}^{15} x_i^4$$

$$0.27 \quad 4.2545 \quad -1.417671 \quad 5.17024805$$

$$\bar{x} = 0.27/15 = 0.018$$

$$b_k = 15 [5.17024805 + 4 \times 0.018 \times (-1.417671) + 6 \times (0.018)^2 \times 4.2545 - 45 (0.018)^4] / [4.2545 - 15 (0.018)^2]^2 = 79.20879579/18.05944013 = 4.3860$$

取 $\alpha = 5\%$, 对应临界值为4.13, 因 $b_k = 4.3860 > 4.13$, 判断距离均值0.018最远的-1.40为异常值。去除了-1.40之后, 对余下14个值

$$\begin{array}{cccc} \sum_{i=1}^{14} x_i & \sum_{i=1}^{14} x_i^2 & \sum_{i=1}^{14} x_i^3 & \sum_{i=1}^{14} x_i^4 \\ 0.27 & 4.2545 & -1.417671 & 5.17024805 \\ +1.40 & -1.9600 & +2.744000 & -3.84160000 \\ \hline 1.67 & 2.2945 & 1.326329 & 1.32864805 \end{array}$$

$$\bar{x} = 1.67/14 = 0.1193, \text{再计算}$$

$$b_k = 14 [1.32864805 - 4 \times 0.1193 \times 1.326329 + 6 \times (0.1193)^2 \times 2.2945 - 3 \times 14 (0.1193)^4] / [2.2945 - 14 \times (0.1193)^2]^2 = 12.36462926/4.39025216 = 2.8164$$

对 $\alpha = 5\%$, $n = 14$ 。对应临界值约为4.11, 而 $b_k < 4.11$, 故不能再检出异常值。只检出-1.40为异常值。

6.3 狄克逊检验法

6.3.1 狄克逊检验法的规则见5.3。

6.3.2 重复使用狄克逊检验法的示例。

数据同于6.2.4 对 $n = 15$, 计算

$$r_{22} = \frac{x_{(15)} - x_{(13)}}{x_{(15)} - x_{(3)}} = \frac{1.01 - 0.48}{1.01 + 0.30} = \frac{0.53}{1.31} = 0.406$$

$$r'_{22} = \frac{x_{(3)} - x_{(1)}}{x_{(13)} - x_{(1)}} = \frac{-0.30 + 1.40}{0.48 + 1.40} = \frac{1.10}{1.88} = 0.585$$

取 $\alpha = 5\%$, 对双侧问题, 查出临界值 $\tilde{D}_{0.95}(15) = 0.565$, 由于 $r'_{22} > r_{22}$, 且 $r'_{22} > \tilde{D}_{0.95}(15)$, 故判断最小值-1.40为异常值。除去这个观测值以外的14个值 ($n = 14$), 使用

$$\begin{array}{l} r_{22} = \frac{x_{(14)} - x_{(12)}}{x_{(14)} - x_{(3)}} = \frac{1.01 - 0.48}{1.01 + 0.24} = \frac{0.53}{1.25} = 0.424 \\ r'_{22} = \frac{x_{(3)} - x_{(1)}}{x_{(12)} - x_{(1)}} = \frac{-0.24 + 0.44}{0.48 + 0.44} = \frac{0.20}{0.92} = 0.217 \end{array}$$

对 $\alpha = 5\%$, 临界值为 $\tilde{D}_{0.95}(14) = 0.586$, 故不能继续检出异常值, 只检出-1.40为异常值。

附 录 A
统 计 数 值 表
(参考件)

表 A1 奈尔检验法的临界值表

n	90%	95%	97.5%	99%	99.5%	n	90%	95%	97.5%	99%	99.5%
3	1.497	1.738	1.955	2.215	2.396	26	2.602	2.829	3.039	3.298	3.481
4	1.696	1.941	2.163	2.431	2.618	27	2.616	2.843	3.053	3.310	3.493
5	1.835	2.080	2.304	2.574	2.764	28	2.630	2.856	3.065	3.322	3.505
6	1.939	2.184	2.408	2.679	2.870	29	2.643	2.869	3.077	3.334	3.516
7	2.022	2.267	2.490	2.761	2.952	30	2.656	2.881	3.089	3.345	3.527
8	2.091	2.334	2.557	2.828	3.019	31	2.668	2.892	3.100	3.356	3.538
9	2.150	2.392	2.613	2.884	3.074	32	2.679	2.903	3.111	3.366	3.548
10	2.200	2.441	2.662	2.931	3.122	33	2.690	2.914	3.121	3.376	3.557
11	2.245	2.484	2.704	2.973	3.163	34	2.701	2.924	3.131	3.385	3.566
12	2.284	2.523	2.742	3.010	3.199	35	2.712	2.934	3.140	3.394	3.575
13	2.320	2.557	2.776	3.043	3.232	36	2.722	2.944	3.150	3.403	3.584
14	2.352	2.589	2.806	3.072	3.261	37	2.732	2.953	3.159	3.412	3.592
15	2.382	2.617	2.834	3.099	3.287	38	2.741	2.962	3.167	3.420	3.600
16	2.409	2.644	2.860	3.124	3.312	39	2.750	2.971	3.176	3.428	3.608
17	2.434	2.668	2.883	3.147	3.334	40	2.759	2.980	3.184	3.436	3.616
18	2.458	2.691	2.905	3.168	3.355	41	2.768	2.988	3.192	3.444	3.623
19	2.480	2.712	2.926	3.188	3.374	42	2.776	2.996	3.200	3.451	3.630
20	2.500	2.732	2.945	3.207	3.392	43	2.784	3.004	3.207	3.458	3.637
21	2.519	2.750	2.963	3.224	3.409	44	2.792	3.011	3.215	3.465	3.644
22	2.538	2.768	2.980	3.240	3.425	45	2.800	3.019	3.222	3.472	3.651
23	2.555	2.784	2.996	3.256	3.440	46	2.808	3.026	3.229	3.479	3.657
24	2.571	2.800	3.011	3.270	3.455	47	2.815	3.033	3.235	3.485	3.663
25	2.587	2.815	3.026	3.284	3.468	48	2.822	3.040	3.242	3.491	3.669
						49	2.829	3.047	3.249	3.498	3.675
						50	2.836	3.053	3.255	3.504	3.681

续表 A1

n	90%	95%	97.5%	99%	99.5%	n	90%	95%	97.5%	99%	99.5%
51	2.843	3.060	3.261	3.509	3.687	76	2.974	3.185	3.381	3.624	3.798
52	2.849	3.066	3.267	3.515	3.692	77	2.978	3.189	3.385	3.628	3.801
53	2.856	3.072	3.273	3.521	3.698	78	2.983	3.193	3.389	3.631	3.805
54	2.862	3.078	3.279	3.526	3.703	79	2.987	3.197	3.393	3.635	3.808
55	2.868	3.084	3.284	3.532	3.708	80	2.991	3.201	3.396	3.638	3.812
56	2.874	3.090	3.290	3.537	3.713	81	2.995	3.205	3.400	3.642	3.815
57	2.880	3.095	3.295	3.542	3.718	82	2.999	3.208	3.403	3.645	3.818
58	2.886	3.101	3.300	3.547	3.723	83	3.002	3.212	3.407	3.648	3.821
59	2.892	3.106	3.306	3.552	3.728	84	3.006	3.216	3.410	3.652	3.825
60	2.897	3.112	3.311	3.557	3.733	85	3.010	3.219	3.414	3.655	3.828
61	2.903	3.117	3.316	3.562	3.737	86	3.014	3.223	3.417	3.658	3.831
62	2.908	3.122	3.321	3.566	3.742	87	3.017	3.226	3.421	3.661	3.834
63	2.913	3.127	3.326	3.571	3.746	88	3.021	3.230	3.424	3.665	3.837
64	2.919	3.132	3.330	3.575	3.751	89	3.024	3.233	3.427	3.668	3.840
65	2.924	3.137	3.335	3.580	3.755	90	3.028	3.236	3.430	3.671	3.843
66	2.929	3.142	3.339	3.584	3.759	91	3.031	3.240	3.433	3.674	3.846
67	2.934	3.146	3.344	3.588	3.763	92	3.035	3.243	3.437	3.677	3.849
68	2.938	3.151	3.348	3.593	3.767	93	3.038	3.246	3.440	3.680	3.852
69	2.943	3.155	3.353	3.597	3.771	94	3.042	3.249	3.443	3.683	3.854
70	2.948	3.160	3.357	3.601	3.775	95	3.045	3.253	3.446	3.685	3.857
71	2.952	3.164	3.361	3.605	3.779	96	3.048	3.256	3.449	3.688	3.860
72	2.957	3.169	3.365	3.609	3.783	97	3.052	3.259	3.452	3.691	3.863
73	2.961	3.173	3.369	3.613	3.787	98	3.055	3.262	3.455	3.694	3.865
74	2.966	3.177	3.373	3.617	3.791	99	3.058	3.265	3.458	3.697	3.868
75	2.970	3.181	3.377	3.620	3.794	100	3.061	3.268	3.460	3.699	3.871

表 A2 格拉布斯检验法的临界值表

n	90%	95%	97.5%	99%	99.5%	n	90%	95%	97.5%	99%	99.5%
3	1.148	1.153	1.155	1.155	1.155	6	1.729	1.822	1.887	1.944	1.973
4	1.425	1.463	1.481	1.492	1.496	7	1.828	1.938	2.020	2.097	2.139
5	1.602	1.672	1.715	1.749	1.764	8	1.909	2.032	2.126	2.221	2.274
						9	1.977	2.110	2.215	2.323	2.387
						10	2.036	2.176	2.290	2.410	2.482

续表 A2

n	90%	95%	97.5%	99%	99.5%	n	90%	95%	97.5%	99%	99.5%
11	2.088	2.234	2.355	2.485	2.564	56	2.811	3.000	3.172	3.383	3.531
12	2.134	2.285	2.412	2.550	2.636	57	2.818	3.006	3.180	3.391	3.539
13	2.175	2.331	2.462	2.607	2.699	58	2.824	3.013	3.186	3.397	3.546
14	2.213	2.371	2.507	2.659	2.755	59	2.831	3.019	3.193	3.405	3.553
15	2.247	2.409	2.549	2.705	2.806	60	2.837	3.025	3.199	3.411	3.560
16	2.279	2.443	2.585	2.747	2.852	61	2.842	3.032	3.205	3.418	3.566
17	2.309	2.475	2.620	2.785	2.894	62	2.849	3.037	3.212	3.424	3.573
18	2.335	2.504	2.651	2.821	2.932	63	2.854	3.044	3.218	3.430	3.579
19	2.361	2.532	2.681	2.854	2.968	64	2.860	3.049	3.224	3.437	3.586
20	2.385	2.557	2.709	2.884	3.001	65	2.866	3.055	3.230	3.442	3.592
21	2.408	2.580	2.733	2.912	3.031	66	2.871	3.061	3.235	3.449	3.598
22	2.429	2.603	2.758	2.939	3.060	67	2.877	3.066	3.241	3.454	3.605
23	2.448	2.624	2.781	2.963	3.087	68	2.883	3.071	3.246	3.460	3.610
24	2.467	2.644	2.802	2.987	3.112	69	2.888	3.076	3.252	3.466	3.617
25	2.486	2.663	2.822	3.009	3.135	70	2.893	3.082	3.257	3.471	3.622
26	2.502	2.681	2.841	3.029	3.157	71	2.897	3.087	3.262	3.476	3.627
27	2.519	2.698	2.859	3.049	3.178	72	2.903	3.092	3.267	3.482	3.633
28	2.534	2.714	2.876	3.068	3.199	73	2.908	3.098	3.272	3.487	3.638
29	2.549	2.730	2.893	3.085	3.218	74	2.912	3.102	3.278	3.492	3.643
30	2.563	2.745	2.908	3.103	3.236	75	2.917	3.107	3.282	3.496	3.648
31	2.577	2.759	2.924	3.119	3.253	76	2.922	3.111	3.287	3.502	3.654
32	2.591	2.773	2.938	3.135	3.270	77	2.927	3.117	3.291	3.507	3.658
33	2.604	2.786	2.952	3.150	3.286	78	2.931	3.121	3.297	3.511	3.663
34	2.616	2.799	2.965	3.164	3.301	79	2.935	3.125	3.301	3.516	3.669
35	2.628	2.811	2.979	3.178	3.316	80	2.940	3.130	3.305	3.521	3.673
36	2.639	2.823	2.991	3.191	3.330	81	2.945	3.134	3.309	3.525	3.677
37	2.650	2.835	3.003	3.204	3.343	82	2.949	3.139	3.315	3.529	3.682
38	2.661	2.846	3.014	3.216	3.356	83	2.953	3.143	3.319	3.534	3.687
39	2.671	2.857	3.025	3.228	3.369	84	2.957	3.147	3.323	3.539	3.691
40	2.682	2.866	3.036	3.240	3.381	85	2.961	3.151	3.327	3.543	3.695
41	2.692	2.877	3.046	3.251	3.393	86	2.966	3.155	3.331	3.547	3.699
42	2.700	2.887	3.057	3.261	3.404	87	2.970	3.160	3.335	3.551	3.704
43	2.710	2.896	3.067	3.271	3.415	88	2.973	3.163	3.339	3.555	3.708
44	2.719	2.905	3.075	3.282	3.425	89	2.977	3.167	3.343	3.559	3.712
45	2.727	2.914	3.085	3.292	3.435	90	2.981	3.171	3.347	3.563	3.716
46	2.736	2.923	3.094	3.302	3.445	91	2.984	3.174	3.350	3.567	3.720
47	2.744	2.931	3.103	3.310	3.455	92	2.989	3.179	3.355	3.570	3.725
48	2.753	2.940	3.111	3.319	3.464	93	2.993	3.182	3.358	3.575	3.728
49	2.760	2.948	3.120	3.329	3.474	94	2.996	3.186	3.362	3.579	3.732
50	2.768	2.956	3.128	3.336	3.483	95	3.000	3.189	3.365	3.582	3.736
51	2.775	2.964	3.136	3.345	3.491	96	3.003	3.193	3.369	3.586	3.739
52	2.783	2.971	3.143	3.353	3.500	97	3.006	3.196	3.372	3.589	3.744
53	2.790	2.978	3.151	3.361	3.507	98	3.011	3.201	3.377	3.593	3.747
54	2.798	2.986	3.158	3.368	3.516	99	3.014	3.204	3.380	3.597	3.750
55	2.804	2.992	3.166	3.376	3.524	100	3.017	3.207	3.383	3.600	3.754

表 A3 狄克逊检验法的临界值表

n	统 计 量	90%	95%	99%	99.5%
3	$r_{10} = \frac{x_{(n)} - x_{(n-1)}}{x_{(n)} - x_{(1)}} \text{ 或 } r'_{10} = \frac{x_{(2)} - x_{(1)}}{x_{(n)} - x_{(1)}}$	0.886	0.941	0.988	0.994
4		0.679	0.765	0.889	0.926
5		0.557	0.642	0.780	0.821
6		0.482	0.560	0.698	0.740
7		0.434	0.507	0.637	0.680
8	$r_{11} = \frac{x_{(n)} - x_{(n-1)}}{x_{(n)} - x_{(2)}} \text{ 或 } r'_{11} = \frac{x_{(n)} - x_{(1)}}{x_{(n-1)} - x_{(1)}}$	0.479	0.554	0.683	0.725
9		0.441	0.512	0.635	0.677
10		0.409	0.477	0.597	0.639
11	$r_{21} = \frac{x_{(n)} - x_{(n-2)}}{x_{(n)} - x_{(2)}} \text{ 或 } r'_{21} = \frac{x_{(3)} - x_{(1)}}{x_{(n-1)} - x_{(1)}}$	0.517	0.576	0.679	0.713
12		0.490	0.546	0.642	0.675
13		0.467	0.521	0.615	0.649
14	$r_{22} = \frac{x_{(n)} - x_{(n-2)}}{x_{(n)} - x_{(3)}} \text{ 或 } r'_{22} = \frac{x_{(3)} - x_{(1)}}{x_{(n-2)} - x_{(1)}}$	0.492	0.546	0.641	0.674
15		0.472	0.525	0.616	0.647
16		0.454	0.507	0.595	0.624
17		0.438	0.490	0.577	0.605
18		0.424	0.475	0.561	0.589
19		0.412	0.462	0.547	0.575
20		0.401	0.450	0.535	0.562
21		0.391	0.440	0.524	0.551
22		0.382	0.430	0.514	0.541
23		0.374	0.421	0.505	0.532
24		0.367	0.413	0.497	0.524
25		0.360	0.406	0.489	0.516
26		0.354	0.399	0.486	0.508
27		0.348	0.393	0.475	0.501
28		0.342	0.387	0.469	0.495
29		0.337	0.381	0.463	0.489
30		0.332	0.376	0.457	0.483

表 A3' 双侧狄克逊检验法的临界值表

n	统 计 量	95%	99%	n	统 计 量	95%	99%
3	r_{10} 和 r'_{10} 中较大者	0.970	0.994	17	r_{22} 和 r'_{22} 中较大者	0.529	0.610
4		0.829	0.926	18		0.514	0.594
5		0.710	0.821	19		0.501	0.580
6		0.628	0.740	20		0.489	0.567
7		0.569	0.680	21		0.478	0.555
8	r_{11} 和 r'_{11} 中较大者	0.608	0.717	22		0.468	0.544
9		0.564	0.672	23		0.459	0.535
10		0.530	0.635	24		0.451	0.526
11	r_{21} 和 r'_{21} 中较大者	0.619	0.709	25		0.443	0.517
12		0.583	0.660	26		0.436	0.510
13		0.557	0.638	27		0.429	0.502
14	r_{22} 和 r'_{22} 中较大者	0.586	0.670	28		0.423	0.495
15		0.565	0.647	29		0.417	0.489
16		0.546	0.627	30		0.412	0.483

表 A4 偏度检验法的临界值表

n	95 %	99 %	n	95 %	99 %
8	0.99	1.42	40	0.59	0.87
9	0.97	1.41	45	0.56	0.82
10	0.95	1.39	50	0.53	0.79
12	0.91	1.34	60	0.49	0.72
15	0.85	1.26	70	0.46	0.67
20	0.77	1.15	80	0.43	0.63
25	0.71	1.06	90	0.41	0.60
30	0.66	0.98	100	0.39	0.57
35	0.62	0.92			

表 A5 峰度检验法的临界值表

n	95 %	99 %	n	95 %	99 %
8	3.70	4.53	40	4.05	5.02
9	3.86	4.82	45	4.02	4.94
10	3.95	5.00	50	3.99	4.87
12	4.05	5.20	60	3.93	4.73
15	4.13	5.30	70	3.88	4.62
20	4.17	5.38	80	3.84	4.52
25	4.14	5.29	90	3.80	4.45
30	4.11	5.20	100	3.77	4.37
35	4.08	5.11			

附录 B

判断和处理异常值的规则的选择指南 (参考件)

判断和处理异常值的各种方法适用于不同情况,不同要求,在选择时应注意下述几点:

B.1 判断和处理异常值的目的

B.1.1 三种不同的目的

B.1.1.1 识别与诊断,主要目的是找出异常值,从而进行生产诊断、新规律探索、技术考查等工作。

B.1.1.2 估计参数,主要目的在于估计总体的某个参数,寻找异常值的目的在于确定这些值是否计入样本,以估准参数。

B.1.1.3 检验假设,主要目的在于判断总体是否符合所考查的要求,寻找异常值的目的主要在于确定这些值是否计入样本,以使判断结果尽量准确。

B.1.2 判断异常值的不同目的引起的不同态度

B.1.2.1 对于以识别为目的,选择判断异常值的主要标准在于判断准确性,要根据所判断错误带来的风险不同,选择适宜的规则。

B.1.2.2 对于以估计和检验为目的,要判断异常值,就应把判断和处理异常值的方法和进一步作估计或检验的准确性统一起来考虑。如使用格拉布斯检验法作估计,实际是一种新估计量

$$\hat{\mu} = \begin{cases} (x_1 + \cdots + x_n)/n & \text{当 } G_n \leq G_{1-\alpha^*}(n) \\ (x_{(1)} + \cdots + x_{(n-1)})/(n-1) & \text{当 } G_n > G_{1-\alpha^*}(n) \end{cases}$$

B.1.2.3 若以估计和检验为目的,有时也可以不经过判断异常值的步骤,而采用稳健估计和稳健检验的办法。

例如:在塑料材料中,有时使用截割均值,把12个观测值的最大值与最小值舍去,以余下的10个观测值作算术平均以估计 μ 。(体操比赛评分时,也把诸裁判报出的最高分和最低分舍去,以余下的几个评分的平均值报出),并不需要追查舍去的一定是异常值,而这种估计也很好地预防了异常值的不利影响。

B.2 应用对于标准差的尽可能准确的信息

判断异常值(或判断其高度异常)的统计量都以标准差或其估计量作为尺度,因此要尽可能地利用已获得的标准差的信息。

B.2.1 可假设标准差已知的情形,判断准确性好,它适用于正常稳定的生产、试验、测试的数据。

B.2.2 存在对标准差的独立估计,此时有多个样本,可验证它们的标准差相等,从而把各个样本的标准差估计量合并起来,作为共同的标准差的估计。它对于单独一个样本来说,就存在对标准差的独立估计。本标准没有涉及这类问题,它适用于可认为有别的样本的标准差能代表本批样本的标准差。比如方差分析中就出现这种做法,试验室间试验中判断异常值的方法也有这种类型。

B.2.3 若无别的数据可利用,或不知别的样本的标准差能否代表本批样本,则使用本批样本来估计标准差。

B.3 对各种检验法的选择

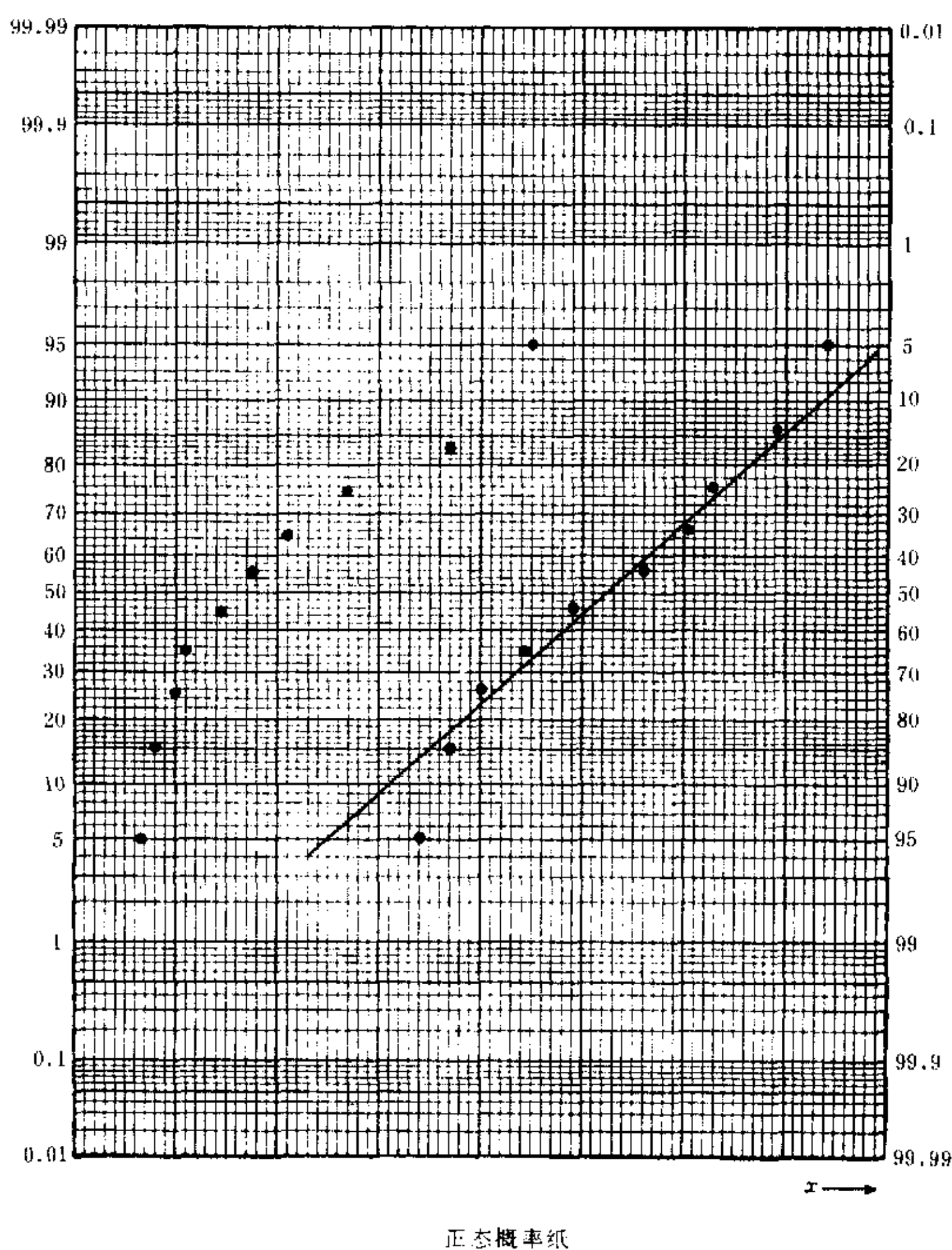
本标准第五章、第六章各给出了两种检验法,应该注意下述几点

B.3.1 在至多只有一个异常值时,格拉布斯检验法具有判断异常值的功效最优性,而狄克逊检验法正确判断异常值的功效与格拉布斯检验法相差甚微。

B.3.2 在出现多个异常值时,重复使用同一检验法可能犯判多为少(只检出一部分异常值)的错误,而不易犯判少为多(错到一部分观测值为异常值)的错误。这两类错误的概率以重复使用偏度—峰度检验法为少(可以证明,它也具有正确判断异常值的功效优良性)。但计算相对复杂得多;重复使

用狄克逊检验法的效果次之，而重复使用格拉布斯的功效则较差，故未采用。

B.3.3 偏度—峰度检验法又是正态性检验的优良检验法。不来自正态分布的样本都可能被它拒绝，但这不只是正态样本主体加异常值的模型，所以使用偏度—峰度检验法时，要满足规定的使用条件。比如，在正态概率纸上，若样本主体不是基本在一条直线的近旁，或是样本主体基本在一条直线近旁，而两端的值相对于这条直线而言不是向外偏离而是向内偏离（如下图），采用偏度—峰度检验法就可能把一部分观测值误判为异常值。



B.4 重视检出的异常值给出的信息

B.4.1 在一段时间后，考查检出的异常值的全体，往往能明显地发现其物理原因和系统倾向，如异常值出自某个测试者为多，说明此人的操作有系统偏差原因。

B.4.2 若各个样本中出现异常值较为经常，又常不能明确其物理原因，则应怀疑分布的正态性假设。此时可更细微的确定统计分布及选择适宜的统计量形式。

因此，推荐标准使用者完善判断和处理异常值的记录，并作定期分析。

附加说明：

本标准由全国统计方法应用标准化技术委员会提出。

本标准由全国统计方法应用标准化技术委员会数据的处理和解释分委员会负责起草。

本标准主要起草人吴传义。