中华人民共和国国家标准

UDC 519.28

GB 4883—85

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方法。

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。

国家标准局1985-01-29发布　　1985-10-01实施

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$x_{(1)}<x_{(2)}<\cdots<x_{(n)}$，计算统计量

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

这里σ是已知的总体标准差，$\overline{x}$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

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

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$R_{0.95}(25) = 2.815$，$R_{0.99}(25) = 3.282$，$R'_n > R_{0.99}(25)$，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$x_1, \cdots x_n$，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$\bar{x} = (x_1 + \cdots + x_n)/n$，

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

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$G_{1-\alpha^*/2}(n)$。