

基于狄克逊准则的异常值自动化检测方法及应用

□成杰 彭明 罗检民 易平安 李锋 贺敏娟

永州市计量测试检定所

【摘要】异常值的处理一直是测试人员备受关注的问题，因为异常值的存在不仅对测试结果的精准度产生显著影响，严重扭曲数据的统计特征，而且易造成结论出现偏差，影响测试工作的科学性与严谨性。对于不服从正态分布或存在多个异常值的中小规模数据样本（ $n=3 \sim 30$ ），本文提出一种基于狄克逊准则，依托 Matlab 软件自动识别与剔除异常值的检测方法，对测量数据中的异常值实现精准的识别，提升了测量数据预处理的效率。

【关键词】异常值；狄克逊准则；Matlab；自动化检测；统计量分析

文献标识码：A 文章编号：1003-1870（2026）06-0037-05

Automated Outlier Detection Method Based on Dixon's Criterion and Its Application

【Abstract】 The handling of outliers has long been a major concern for testers, as their presence not only significantly impacts the accuracy of test results and severely distorts the statistical characteristics of data, but also likely leads to biased conclusions, undermining the scientific rigor and accuracy of testing work. For small to medium-scale data samples ($n = 3 - 30$) that do not follow a normal distribution or contain multiple outliers, an automated detection method based on Dixon's criterion using Matlab software is proposed in this paper to precisely identify and remove outliers from measurement data, thereby improving the efficiency of data preprocessing.

【Keywords】 outlier; Dixon's criterion; Matlab; automatic detection; statistical analysis

引言

异常值又称离群值或粗大误差，是指对一个被重复测量时得到的有限测量数据结果中，出现了与其他值偏离较远且不符合统计规律的个别值^[1]。特别是户外测试情况，人为操作偏差、系统噪声、环境干扰、仪器计量误差等因素，都可能导致异常值的数据产生。异常值的存在会干扰计量误差分析与不确定度评定，影响测量结论的准确性与权威性^[2]。因

此，科学地对异常值进行处理，是保障计量工作规范性与精准性的关键。

1 异常值判别方法及核心理论

3σ 准则、格拉布斯准则、狄克逊准则等方法都是在计量领域对异常值判别的常用手段^[3]。因三种方法对异常值判别原理不同，其使用场合也明显的有差异。

1.1 3σ 准则（拉依达准则）

3 σ 准则是基于正态分布的统计特性的一种异常值判别手段, 其核心判定逻辑为: 对于测试数据: $x_1, x_2, \dots, x_n$ 计算其总体均值 $\bar{x}$ 与实验标准差 s , 若某一测量值 x_i 满足

$$|x_i - \bar{x}| > 3s$$

则可判定该测量值为异常值。该准则具备计算简洁、易于实现的优势, 但只适用于测量数据服从正态分布且样本量满足 $n > 30$ 的场合。

1.2 格拉布斯准则

格拉布斯准则对异常值判断依据为: 计算可疑数据点 (通常是最大值或最小值) x_d 与样本均值 $\bar{x}$ 之间残余误差与标准差 s 的比值, 如果其比值超过了一个由样本量和显著性水平决定的临界值时则判定该可疑值为异常值^[4], 计算公式如下:

$$\frac{|x_d - \bar{x}|}{s} > G(\alpha, n);$$

格拉布斯准则对单一异常值具有较高的识别精度, 但对于不服从正态分布且存在多个异常值的样本数据, 则不适用于该准则判别异常值。

1.3 狄克逊准则

狄克逊准则的异常值判断依据是计算可疑值与邻近观测值的极差比值, 通过双侧检测并结合相应临界值完成异常值的判别。具体判定步骤如下: ①对测量数据首先按从小到大进行排序 $x_1, x_2, \dots, x_n$ 。②再根据样本量的取值范围, 选取对应的计算公式, 具体执行步骤如下:

(1) 当 $n=3 \sim 7$ 时

$$r_{10} = \frac{x_{(n)} - x_{(n-1)}}{x_{(n)} - x_{(1)}}, r'_{10} = \frac{x_{(2)} - x_{(1)}}{x_{(n)} - x_{(1)}}$$

(2) 当 $n=8 \sim 10$ 时

$$r_{11} = \frac{x_{(n)} - x_{(n-1)}}{x_{(n)} - x_{(2)}}, r'_{11} = \frac{x_{(2)} - x_{(1)}}{x_{(n-1)} - x_{(1)}}$$

(3) 当 $n=11 \sim 13$ 时

$$r_{21} = \frac{x_{(n)} - x_{(n-2)}}{x_{(n)} - x_{(2)}}, r'_{21} = \frac{x_{(3)} - x_{(1)}}{x_{(n-1)} - x_{(1)}}$$

(4) 当 $n \geq 14$ 时

$$r_{22} = \frac{x_{(n)} - x_{(n-2)}}{x_{(n)} - x_{(3)}}, r'_{22} = \frac{x_{(3)} - x_{(1)}}{x_{(n-2)} - x_{(1)}}$$

令 r_{ij} 和 r'_{ij} 分别表示 r_{10} 、 r_{11} 、 r_{21} 、 r_{22} 和 r'_{10} 、 r'_{11} 、 r'_{21} 、 r'_{22} , 根据样本量 n 与查表的方式确定狄克逊检

验的临界 $G(\alpha, n)$; 4、当 $r_{ij} > r'_{11}$, $r_{ij} > G(\alpha, n)$ 时, x_n 为异常值, 当 $r_{ij} < r'_{ij}$, $r'_{ij} < G(\alpha, n)$ 时, x_1 为异常值, 否则没有异常值^[5]。

狄克逊准则能在对剔除原始数据异常值后对剩余数据进行重复检验, 直至数据样本中没有异常值的存在, 因狄克逊准则具有迭代检测, 可判定多个异常值特征, 在中小规模样本 ($3 \leq n \leq 30$) 日常测量中异常值判定, 该准则广泛使用。

2 异常值检测系统的设计与实现

本文介绍的异常值检测系统是基于狄克逊准则, 采用Matlab 软件中的模块化设计思路。该系统可实现异常值自动运算与剔除, 统计量结果可视化展示, 提升测量数据预处理的规范化与高效性。系统由主函数、狄克逊检测函数、统计量计算函数、可视化函数四个核心模块组成, 各模块具备独立的功能, 通过协同工作的方式, 实现异常值自动化运算与检测。

2.1 Matlab 简介与应用

Matlab 是美国 MathWorks 公司开发的高性能科学计算与工程仿真软件, 依托简洁的语法表述、优异的计算能力成为计量测试、数据分析、工程仿真领域的核心工具。该软件支持模块化函数编程, 便于开发者对设计的系统进行功能拆解、调试与优化。因其程序代码具备可读性与可移植性等特点, 测试人员可结合实际的设计场景灵活调整参数, 对不同类型、样本量的测量数据进行处理。Matlab 软件在计量领域中拥有广阔应用前景, 现阶段在高精度仪器测试、在线测量等新型场景中使用渗透率显著提升, 未来深度融合人工智能、神经网络、物联网等技术, 将成为高端计量领域中智能化测试的核心载体, 具有不可替代的核心作用。

2.2 系统整体架构

该系统设计遵循数据初始化、异常值检测、统计量计算、结果输出、可视化展示的逻辑顺序, 确保各环节衔接紧密、无逻辑断层, 整体运行流程如下: ①数据初始化: 通过主函数读取测量数据, 明确数据来源与用途, 完成参数初始化; ②异常值检测: 调用狄克逊准则检测函数, 对数据进行异常值识别, 输出异常值判定结果、具体数值及异常值在原始数据中的位置序号; ③统计量计算: 调用统计量

计算工具函数，分别计算原始数据与剔除异常值数据后的核心统计量（最大值、最小值、算术平均值、总体标准差、总体方差）；④结果输出：通过命令行窗口输出完整结果，包括原始样本量、异常值剔除后的样本数量、原始数据信息、异常值信息、核心统计量等；⑤可视化展示：调用可视化对比函数，绘制双栏对比图，对原始数据中的异常值进行识别标注，异常值剔除前后的数据分布差异与统计特征变化可以得到直观反映。检测系统整体架构如图1所示。

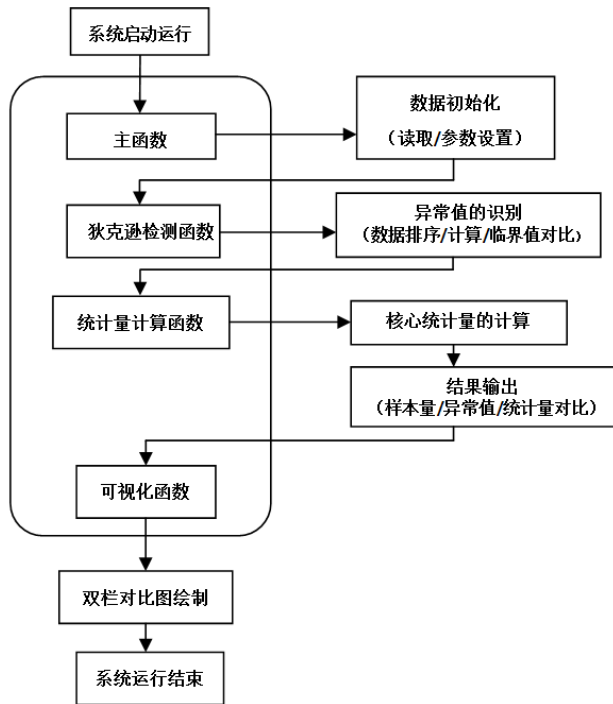


图1 检测系统整体架构图

2.3 核心函数实现

（1）主函数在系统中的作用是负责数据读取与初始化、功能函数调度、结果整合与输出，是保障系统稳定运行的关键中枢。其核心功能包括以下几个方面：①数据读取与初始化：读取测量数据序列，明确数据描述信息（如数据类型、测量场景），初始化样本量、异常值标志等参数；②函数调用：按逻辑顺序调用狄克逊检测函数、统计量计算函数及可视化函数，传递必要参数，确保各模块协同工作；③异常值的剔除与结果的整合：根据狄克逊检测函数返

回的异常值索引，剔除原始数据中的异常值得到净化后的数据，整合原始数据与净化后数据的统计量结果；④结果输出：通过控制台输出所有关键信息，确保输出内容清晰、规范，便于测试人员快速掌握关键的核心数据。

（2）狄克逊检测函数是实现异常值识别的核心模块，其输入为测试数据序列，输出为异常值判定标志、异常值具体数值及异常值在原始数据中的位置。核心功能是基于狄克逊准则的异常值判定原理，完成数据排序、统计量计算、临界值查询以及异常值的判定等工作。

（3）统计量计算函数在测量数据中极大值、极小值、算术平均值、总体标准差与总体方差的计算是通过统计量计算函数实现的，为数据对比分析提供了统计学依据，保证了测试结果的严谨性与可靠性。

（4）可视化函数用于绘制原始数据与剔除异常值数据后的双栏对比图，直观呈现剔除异常值前后的数据统计对比图，为检测结果分析提供可视化支撑。

3 试验验证与结果分析

为了验证本文提出的对异常值自动化检测方法的有效性、优越性及实用性，使用温湿度巡检仪对电热鼓风干燥箱进行温度测试，开展验证试验。

3.1 系统运行实例

测试实例使用的仪器：

便携式精密智能温湿度巡检仪1611A/HTDHG-9023A 型电热鼓风干燥箱。

测试实例使用的方案：

使用温湿度巡检仪对电热鼓风干燥箱内容某测试点进行温度数据采集。

将温湿度巡检仪的温湿度传感器放置在某干燥箱内；再将干燥箱设定温度为100.00℃后启动运行。等试验设备达到稳定状态后，开始记录测试点温度，数据采集间隔为2分钟，在30分钟内共采集16组有效测试数据。具体测量数据如表1所示。

表1 温度采集测量数据

测试结果	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
单位: ℃	100.06	101.08	100.13	100.97	99.78	100.23	101.11	100.33	100.79	100.03	100.16	102.97	100.21	100.33	100.30	98.23

3.2 系统检测结果与分析

将16组温度测试数据输入到系统中，运行检测程序，检测结果如下：原始数据样本量为16组，系

统成功检测到2个异常值，分别为98.23℃（最小值）与102.97℃（最大值），剔除异常值后，样本数量为14组，结果输出如下（见图2，图3）：

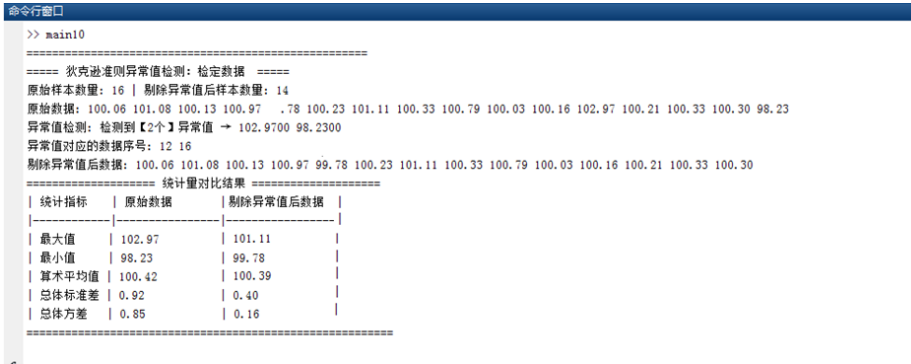


图2 系统输出的异常值检测结果

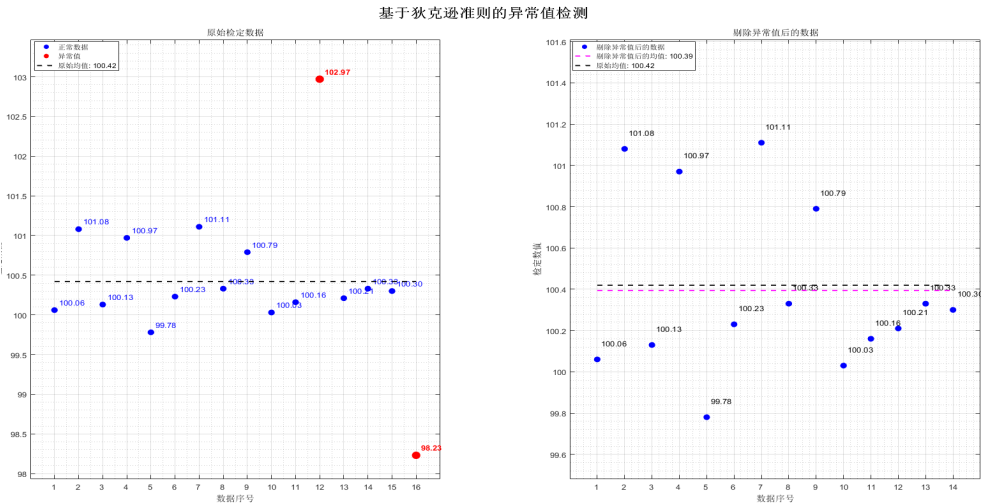


图3 剔除异常值数据前后的双栏对比图

(1) 检测结果可靠性分析：本次试验样本量=16时，使用人工计算的方式对系统检测计算结果进行验证，确定本次检测结果存在两个异常数值，分别是：102.97、98.23，与系统运算结果一致，该系统完全符合狄克逊准则的异常值判定标准，检测未出现漏判、误判现象，检测结果可追溯、可验证，说明本系统可精准识别中小样本中的多个异常值。

(2) 异常值合理性验证分析：对16组原始数据的数值分布特征分析可知，正常观测值的波动范围

仅为0.8 ~ 1.1℃，数据整体围绕100℃小幅波动。两处异常值与正常观测值的偏差介于1.4 ~ 2.7℃，显著超出正常波动范围，其数值特征明显偏离整体数据分布规律。结合实际工作情况分析，由于数据采集过程中环境温度的瞬时波动、测量设备受到瞬时干扰或人工操作引入的偶然误差，导致了异常值的产生，该结果也印证了异常值检测结果的真实性与合理性。

3.3 统计量对比分析

对比剔除异常值前后数据，测试结果数据的统计特性得到显著优化，如表2 中核心统计量对比分析所示，具体分析如下：

（1）波动范围优化。最大值从102.97℃降至101.11℃，最小值从98.23℃升至99.78℃，数据波动范围由4.74%收窄至1.33%，缩小3.41℃，极端值带来的偏差被有效消除，数据的整体稳定性显著提升。

（2）集中趋势优化。算术平均值从100.42℃调至100.39℃，数值更贴合干燥箱设定温度（100.00℃），核心统计量的参考价值得到提高，为后续计量误差分析提供了更可靠的数据支撑。

（3）离散程度优化。总体标准差从0.92℃降至0.40℃，下降幅度56.52%；总体方差从0.85降至0.16，下降幅度81.18%，数据的离散程度大幅降低，试验结果表明，本系统对异常值的剔除具有良好的数据净化效果，处理后的数据分布更为集中，能够真实反映实际测量中的状态。

表2 核心统计量对比分析

统计指标	原始数据 (℃)	剔除异常值后 数据(℃)	变化量
最大值	102.97	101.11	-1.86
最小值	98.23	99.78	1.55
算术平均值	100.42	100.39	-0.03
总体标准差	0.92	0.40	-0.52
总体方差	0.85	0.16	-0.69

4 结语

针对中小样本中计量测试数据中异常值检测计算量大、精度不足、适配性差等痛点，本文提出一种基于狄克逊准则，依托Matlab 软件编程实现的异常值自动化检测方法，通过理论分析、系统设计与实验验证，得出以下结论：

（1）本文方法可精准解决中小样本中多异常值检测难题，以狄克逊准则为核心，无需考虑数据分布类型，可有效适配中小规模样本量的数据，对测量数据中的异常值实现精准识别。

（2）通过 Matlab 软件搭建的系统，实现了测量数据异常值的自动识别与剔除、核心统计量计算及可视化展示，显著缩短计算耗时，规避了人工判别

带来的主观性与错误，提升了测量数据预处理的效率，保证了测量过程的规范性，降低了计量工作的人力成本。

（3）通过分析异常数据剔除后，测试数据的离散程度显著降低，集中趋势与稳定性大幅提升，统计特性更具代表性，为计量误差分析、不确定度评定提供可靠的数据支撑，方法操作便捷、实用性强，便于在现场测试工作中推广应用。

（4）对测量数据异常值自动化处理的只是数据预处理的一种手段，目的是提升工作效率，对于检定和校准结论不得随意剔除数据，出具检定结论时必须以原始测量数据为唯一依据。

参考文献

- [1] 叶德培. 一级注册计量师基础专业知识及专业实务[M]. 北京：中国质检出版社, 2017.
- [2] 范兴宇，杨阳. 计量视角下离群值识别法的研究综述与展望[J]. 赤峰学院学报（自然科学版）, 2023, 39(7):4-9.
- [3] 王小凯, 朱小文. 计量检定中3种判别和剔除异常值的统计方法[J]. 中国测试, 2018, 44(S1): 41-44.
- [4] 郝旭斌，沈大伟，张艳兵，黄福峰. 基于格拉布斯准则的膛压测试仪校准数据拟合方法[J]. 传感器与微系统, 2022, 41(9): 61-64.
- [5] 李常春, 常克昊，刘禹彬. 几种异常值判别准则在气象计量数据处理中的应用[J]. 电子测量技术, 2020, 43(23): 68-71.

作者简介

成杰（1986—）男，湖南永州人，硕士研究生，工程师。研究方向为：计量测试检定。

彭明（1975—）男，湖南省永州市人，本科学历，高级工程师。研究方向：计量检测技术应用及研究。

罗检民（1969—），男，正高级工程师，湖南省计量检测研究院衡器所所长。主要从事计量检测技术，非现场执法称重设备的研究。