NIST Internal Report
NISTIR 8526

Statistical Detection of Outliers in the Certification of NIST Reference Charpy Lots

Enrico Lucon

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8526

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

NIST Internal Report
NISTIR 8526

Statistical Detection of Outliers in the Certification of NIST Reference Charpy Lots

Enrico Lucon
Applied Chemicals and Materials Division
Material Measurement Laboratory

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8526

June 2024

U.S. Department of Commerce
Gina M. Raimondo, Secretary

National Institute of Standards and Technology
Laurie E. Locascio, NIST Director and Under Secretary of Commerce for Standards and Technology

NIST Internal Report 8526
June 2024

Certain commercial equipment, instruments, software, or materials, commercial or non-commercial, are identified in this paper in order to specify the experimental procedure adequately. Such identification does not imply recommendation or endorsement of any product or service by NIST, nor does it imply that the materials or equipment identified are necessarily the best available for the purpose.

Publication History
Approved by the NIST Editorial Review Board on 2024-05-31

How to Cite this NIST Technical Series Publication
E. Lucon (2024) Detection and Handling of Statistical Outliers in the Certification of NIST Reference Charpy Lots. (National Institute of Standards and Technology, Boulder, CO), NIST Series (NISTIR) 8526.
https://doi.org/10.6028/NIST.IR.8526

NIST Author ORCID iDs
Enrico Lucon: 0000-0002-3021-4785

Contact Information
enrico.lucon@nist.gov

Public Comment Period
June 1, 2024 – May 31, 2025

Submit Comments
enrico.lucon@nist.gov

Abstract

This report describes the procedures used by the NIST Charpy Machine Verification Program to detect statistically significant outliers within the test results obtained from the certification of Charpy reference lots. The evaluation of outliers is based on an array of rigorous statistical procedures. The process begins with the screening of potential/suspected outlier data points for datasets assuming an approximately normal distribution, followed by the establishment of their statistical significance using multiple statistical tests. When this examination identifies a data point as a significant outlier, the corresponding Charpy test and specimen are subjected to in-depth examination, in order to establish whether a test- or material-related problem has occurred, and could justify discarding the test result before calculating the certified absorbed energy of the lot and its associated uncertainties, and establishing whether the lot should be accepted or rejected. The same strategy can also be applied to suspicious values reported by customers. Practical examples are illustrated in appendices.

Keywords

Certification tests, Charpy reference lot, NIST Charpy Program, statistical outlier, statistical significance.

Table of Contents

List of Tables

List of Figures

1. Introduction

Since 1989 [1], the Charpy Machine Verification Program at NIST in Boulder, Colorado, has provided thousands of companies in more than 60 countries worldwide with certified reference Charpy specimens for the indirect verification of their machines according to ASTM E23 [2] and ISO 148-2 [3].

Reference Charpy specimens are available at three absorbed energy levels: low (15 J to 20 J), high (80 J to 120 J), and super-high (190 J to 230 J). At the low- and high-energy level, specimens have been certified at -40 °C since the inception of the program at NIST; however, NIST recently introduced low-energy and high-energy specimens to be tested at room temperature (21 °C) [4].

During the certification process of a lot (batch) of NIST reference specimens, 150 randomly selected samples are tested in order to establish the certified value of absorbed energy of the lot (KV_{ref}), and its associated uncertainties. Tests are performed on the three Charpy machines located in Boulder, which are identified as "reference machines" in section A2.4.1.2 of ASTM E23-23a. The acceptance or rejection of a new lot of Charpy specimens is based on the outcome of the dimensional controls of 60 randomly selected specimens [5], as well as the scatter of absorbed energy values obtained from the 150 impact tests mentioned above. More details on the NIST certification process are provided in the following section.

Sporadically, one or more values of absorbed energy may appear significantly different from the remaining test results. This report describes a rigorous statistical process to be followed in order to confirm those values as statistical outliers and determine whether specific reasons can be identified that would explain the anomalous behavior of the specimen(s). Such causes can be test-related (experimental outliers: errors in test execution, such as incorrect positioning of the specimen on the machine anvils, jamming between a specimen half and a stationary or moving part of the machine, communication error occurred between machine encoder and acquisition system, etc.) or material-related (physical outliers: presence of macroscopic inhomogeneities on the fracture surface, such as inclusions, pores, or delaminations). If a specific explanation is found, it can be legitimate to remove the result from the data set, and establish the acceptability of the lot based on the remaining tests. If, however, a cause could not be determined, the data point should not be rejected, and the acceptance or rejection of the lot must be determined considering all the results obtained. As stated by John Mandel in 1991: "*We do not recommend rejection on the basis of purely statistical considerations. Our main reason is that while such rejection procedures always improve the appearance of the data, for example, by drastically reducing the standard deviations, they do nothing in terms of avoiding future instances of outlying results. They have simply sharply reduced the field to which the inferences from the study apply.*" [6]

In other words, **an absorbed energy value cannot be excluded from a Charpy certification dataset based only on statistical reasons**. If an experimental or physical reason cannot be determined, the datapoint must be retained.

2. Certification of a lot of Charpy indirect verification specimens

The certification of a lot of indirect verification Charpy specimens is a two-step process:

(1) Dimensional inspection, hardness measurement, and impact testing of a preliminary group of 100 specimens (*pilot lot*).

(2) Dimensional inspection, hardness measurement, and impact testing of the remaining 1,800 specimens (*production lot*).

Dimensional inspections and hardness measurements are performed on 30 randomly selected specimens, while impact tests are conducted on the same 30 specimens plus 45 additional samples, also randomly selected.

2.1. Dimensional inspection

The dimensional control of NIST indirect verification specimens consists in measuring the following dimensions by means of a digital optical comparator [5]: length, notch centering, width, thickness, ligament, notch radius, and notch angle. Additionally, the perpendicularity between adjacent sides is verified using a specific gage. Each measured dimension is validated against the tolerances prescribed by the NIST dimensional specifications, which are in most cases stricter than those of ASTM E23 and ISO 148-1 [7]. Depending on the number of dimensionally non-compliant specimens and the magnitude of the deviations from the nominal values, the pilot or production lot (or both) can be rejected even before conducting the impact tests [5].

2.2. Hardness measurements

The Rockwell C hardness (HRC) of 30 specimens from the pilot or production lot is measured by performing measurements at the two specimen ends on the surface opposite the notch. The two HRC values are averaged to obtain the hardness value of the individual sample, while the mean hardness of the 30 specimens is associated to the whole lot.

The following conditions are checked:

(a) The hardness measured on the two specimen ends must not differ by more than 1.5 HRC.

(b) The standard deviation from the HRC values measured on 30 Charpy specimens must not exceed 2 % of the mean value (*i.e.*, coefficient of variation, $CV \leq 2$ %).

Charpy lots cannot be rejected solely on the basis of hardness measurements.

2.3. Impact tests

The crucial step for the acceptance or rejection of an indirect verification Charpy lot is the performance of 75 Charpy tests, 25 on each of the three NIST reference machines (labeled SI3, TO2, and TK). Tests are conducted at -40 °C or 21 °C, in accordance with ASTM E23 and ISO 148-1.

The **average absorbed energy** of the lot is defined as

$$\overline{KV} = \frac{\sum_{i=1}^{n} KV_i}{n} \quad , \tag{1}$$

with KV_i = absorbed energy from the i-th test ($i = 1,\ldots,n$). The standard deviation is defined as

$$s = \sqrt{\frac{\sum_{i=1}^{n} (KV_i - \overline{KV})^2}{n-1}} \quad . \tag{2}$$

In both Eq.(1) and Eq.(2), n is the number of specimens tested (25 for an individual machine, 75 for the overall lot).

The **pooled standard deviation** for the lot is defined as

$$s_p = \sqrt{\frac{\sum_{j=1}^{P}(n_j-1)s_j^2}{\sum_{j=1}^{P} n_j-1}} \quad , \tag{3}$$

where n_j and s_j are the number of tests and the standard deviation for the specific reference machine, and P is the number of reference machines (normally 3). When the number of tests is 25 for each of 3 reference machines, eq. (3) simplifies to:

$$s_p = \sqrt{\frac{s_1^2+s_2^2+s_3^2}{3}} \quad , \tag{4}$$

where s_1, s_2, and s_3 are the standard deviations of the 3 reference machines.

The **sample size** of the lot, which represents the minimum number of specimens from a given lot that should be tested in a verification test, is defined as

$$n_{SS} = \left(\frac{3s_p}{E}\right)^2 \quad , \tag{5}$$

where E is the greater between 1.4 J (1 ft-lbs.) or 5 % of $\overline{KV}$.

The most important acceptance criterion for a pilot or production lot of indirect verification Charpy specimens is:

$$\boldsymbol{n_{SS} \leq 5.0} \quad .$$

The ratio between the standard deviation of a particular machine and the pooled standard deviation from Eq.(4) is labeled k_{SI3}, k_{TO2}, or k_{TK}. If any of these factors is greater than 1.25, the largest of the machine standard deviations shall be used to calculate the sample size instead of s_p.

3. Occurrence of statistical outliers in the certification of Charpy lots

In statistics, an outlier is defined as a value that is abnormally distant from other values in a random sample from a population [8]. In mechanical testing, an outlier can also have an experimental or physical explanation:

- A test performed in an erroneous way or not in conformance with the standard is an experimental outlier; for example, a Charpy specimen tested at the wrong temperature (room temperature instead of -40 °C), or placed on the machine supports rotated by 90°.
- A specimen containing a large inhomogeneity or impurity, such as a big carbide or discontinuity (pore, delamination, …), represents an example of physical outlier.

Experimental or physical outliers, when properly identified, can be confidently excluded from certification datasets before calculating the sample size. However, if no experimental or physical cause can be found, an outlier should never be discarded purely on the basis of its extreme (low or high) value with respect to the other results in the dataset.

The first step is the detection of possible outliers, which allows identifying data points that should be subjected to further analysis. Once potential/suspected outliers are identified, their statistical significance is established, so that specific investigations concerning their experimental or physical nature can be eventually performed.

The NIST Charpy Machine Verification Program policy on the treatment of statistical outliers, described in the present report, consists in the following three steps:

(1) Preliminary screening of a Charpy dataset to identify potential outliers.

(2) Evaluation of the statistical significance of potential outliers through the application of multiple statistical tests.

(3) Identification of possible experimental and/or physical causes, and, if these can be identified, exclusion of the corresponding values and recalculation of the relevant parameters (average absorbed energy, standard deviations, pooled standard deviation, and sample size).

Obviously, if step (1) does not reveal any potential outliers, or if step (2) does not identify any statistically significant outlier, no additional investigations are warranted, and the original calculated parameters for the whole Charpy dataset are retained.

4. Screening of potential/suspected outliers

4.1. General strategy

Two approaches (box-and-whisker plots and modified z-score) are discussed in subsections 4.3 and 4.4 below. **Test results identified as potential/suspected outliers by either approach are subject to further analysis to assess their statistical significance in accordance with the methods described in section 5 below**. As a preliminary step, a check of the dataset must be performed to assess whether absorbed energy data exhibit a normal (or approximately normal) distribution. If the dataset does not appear to be normally distributed, and the departure from normality is not due to outlying measurements, a power transformation is suggested.

4.2. Normality check of certification test results

Most of the statistical tests mentioned in this report, and in general the majority of the criteria for outliers, are based on an assumed underlying normal (Gaussian) population or distribution. When the test results are not normally (or approximately normally) distributed, the probabilities associated with the tests will be different.

Therefore, before proceeding with the identification of potential/suspected outliers, it is good practice to examine the distribution of the absorbed energy values and verify the hypothesis that they are normally, or approximately normally, distributed.

Visual inspection of the distribution may be used to assess normality, although this approach does not guarantee that the distribution is normal. Various graphical representations of the dataset can be used to this purpose, such as a histogram, stem-and-leaf plot, boxplot, probability (P-P) plot, and quantile-quantile (Q-Q) plot. The latter plots the quantiles (values that split a dataset into equal portions) versus the numerical values, and it's easier to interpret in case of large sample size (typically, $n > 30$). The Q-Q plot is the recommended graphical representation here, and has a straightforward qualitative interpretation: if the data values fall along a roughly straight line, then data appear to be normally distributed.

An example of a Q-Q plot normally distributed dataset is shown in Figure 1: data points fall relatively close to a straight line.

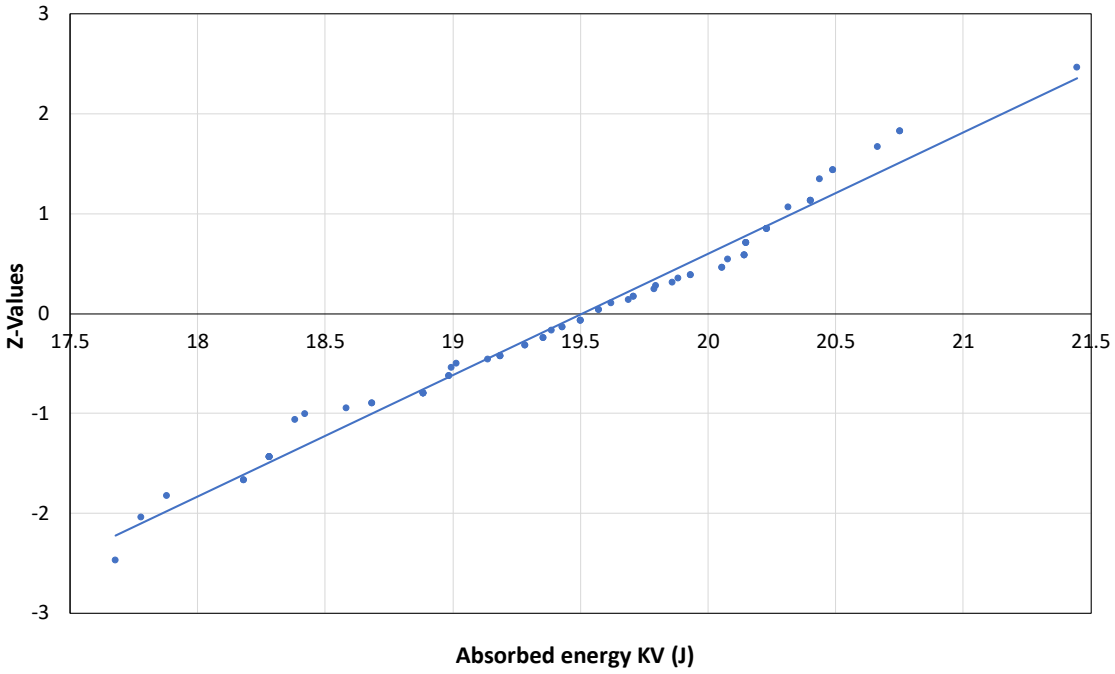

Figure 1 - Q-Q plot for a normally distributed Charpy low-energy dataset.

An example of another low-energy dataset that does not follow a normal distribution is shown in Figure 2: data points corresponding to the low and high "tails" of the distribution clearly deviate from the linear regression line. Note in Figure 2 the structure and deviation in the tails from the straight line, which clearly indicates that the data likely does not come from a normal distribution.

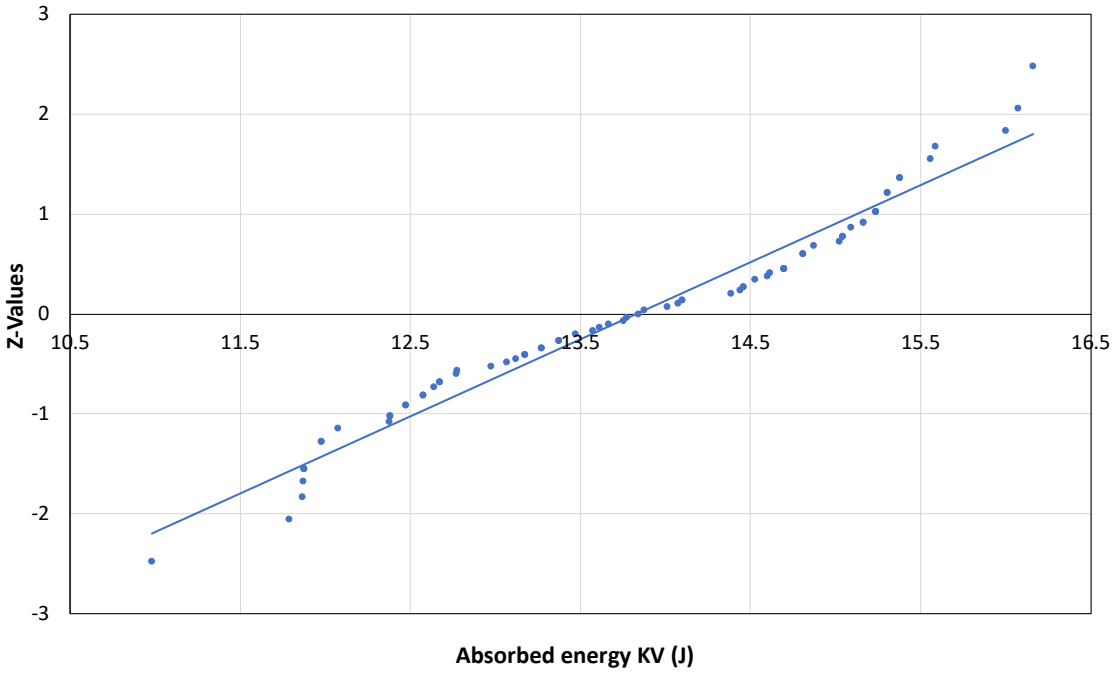

Figure 2 - Q-Q plot for a non-normally distributed Charpy low-energy dataset.

In Figure 1 and Figure 2, the *Z*-values plotted along the ordinate axis are determined as follows.

(a) Absorbed energy values are sorted in ascending order, *i.e.*, from smallest to largest. The rank of the KV_i value (with $i = 1,\ldots,n$) is denoted r_i.

(b) The percentile of the *i*-th value is given by:

$$p_i = \frac{r_i - 0.5}{n} \quad . \tag{6}$$

(c) The *Z*-value of the *i*-th data point in the Q-Q plot corresponds to the **inverse normal cumulative distribution for the probability p_i**.

Numerous statistical tests are available to test data analytically for normal distribution. In all these tests, one is testing the null hypothesis (H_0) that data are normally distributed. If the ***p*-value** (probability value) returned by the normality tests is **smaller than 0.05, then a normal distribution is not assumed** (*i.e.*, H_0 is rejected).

Among the many available tests, the following four have been selected herein:

- Anderson-Darling (A-D) test [9],
- Shapiro-Wilk (S-W) test [10],
- d'Agostino-Pearson (dA-P) test [11], and
- Jarque-Bera (J-B) test [12].

The last two tests (dA-P and J-B) are based on the sample skewness (measure of symmetry, or lack of symmetry) and kurtosis (measure of the "heaviness" of the distribution tails). More about skewness and kurtosis will be provided in section 5.2.2 below.

The probabilities *p* calculated for the four tests listed above, corresponding to the two datasets in Figure 1 (normal) and Figure 2 (non-normal), are listed in Table 1. For both datasets, all tests return the same outcome.

Table 1 - Outcome of normality tests on the datasets shown in Figure 1 and Figure 2.

Dataset in Figure 1		
Test	*p*	Outcome
A-D	0.096	**Normal**
S-W	0.183	**Normal**
dA-P	0.360	**Normal**
J-B	0.452	**Normal**
Dataset in Figure 2		
A-D	0.044	**Non-normal**
S-W	0.011	**Non-normal**
dA-P	0.042	**Non-normal**
J-B	0.044	**Non-normal**

In some cases, the distribution of a certification dataset may be non-normal in the presence of one or more outliers (Figure 3). Once the suspected outlier(s) is/are removed, the distribution becomes normal (Figure 4). The corresponding probabilities obtained from the normality tests are presented in Table 2.

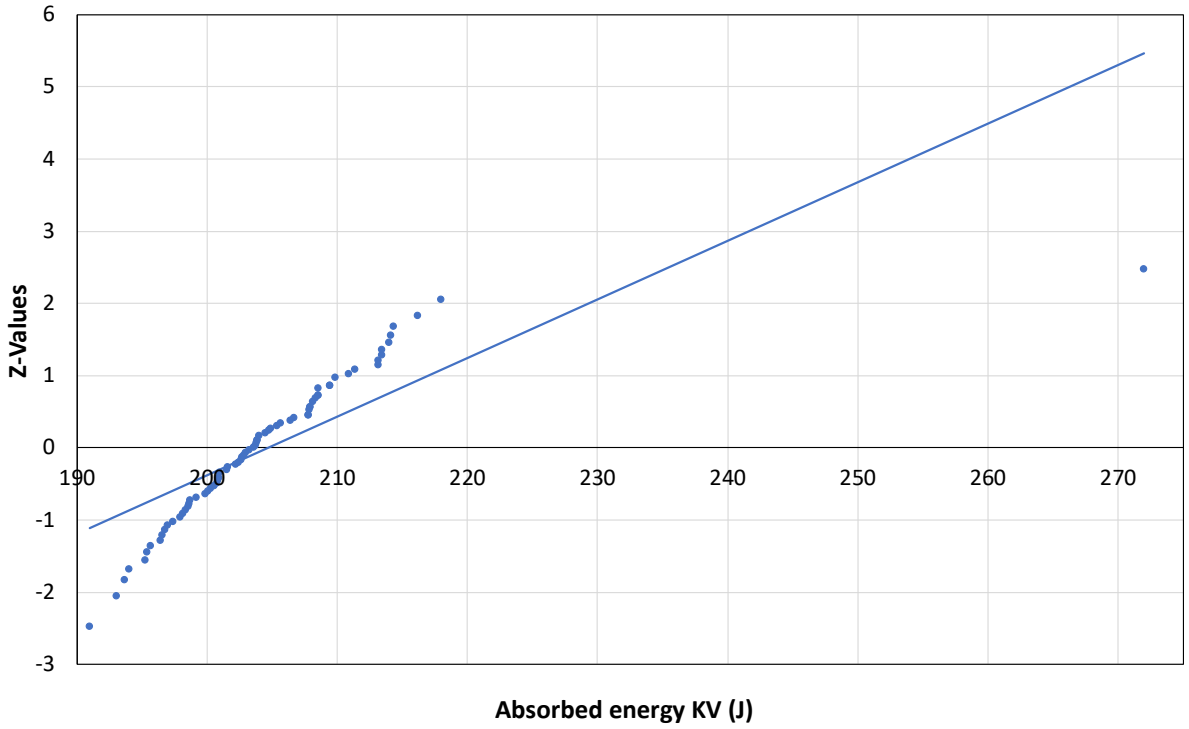

Figure 3 - Q-Q plot for a super-high-energy dataset presenting a very high outlier.

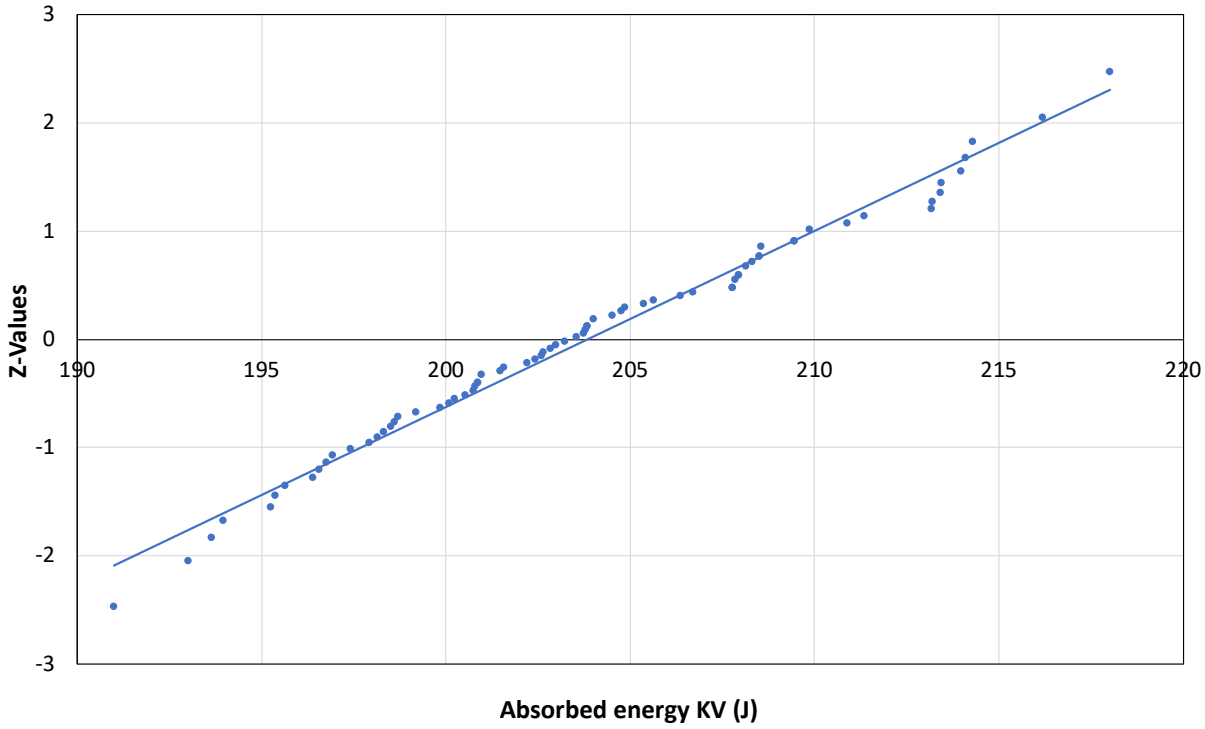

Figure 4 - Q-Q plot for the dataset shown in Figure 3 after removing the outlier.

Table 2 - Outcome of normality tests on the datasets shown in Figure 3 and Figure 4.

Original dataset in Figure 3		
Test	*p*	Outcome
A-D	1.044×10^{-8}	**Non-normal**
S-W	9.741×10^{-12}	**Non-normal**
dA-P	0	**Non-normal**
J-B	0	**Non-normal**
Modified dataset in Figure 4		
A-D	0.444	**Normal**
S-W	0.516	**Normal**
dA-P	0.370	**Normal**
J-B	0.469	**Normal**

As a general rule, if **at least two of the four tests indicate normality of the dataset under investigation** (before and/or after the removal of suspected outliers), the analyses described in the sections below should be performed, *i.e.*, the dataset is deemed normally, or approximately normally, distributed.

If three, or all, of the four normality tests do not recognize the distribution as normal, even after removing potential outliers, a transformation of the data might be helpful. There are many transformations available to make non-normal data resemble normal data, and therefore render them amenable to the outlier analyses described in this document. More details on potential data transformations can be found in [13].

4.3. Box-and-whisker plots

One of the simplest statistical approaches used to identify possible outliers in a large dataset is the so-called box-and-whisker plot (or simply boxplot) [14].

With reference to Figure 5, the "bulk" of the points is represented by the box, which is centered around the median value of absorbed energy and whose width is given by the Interquartile Range (*IQR*), where $IQR = Q3 - Q1$, and *Q1* and *Q3* correspond to the 25th and 75th percentiles[1] of the dataset, respectively. The upper and lower whiskers extend to $Q3 + k \cdot IQR$ and $Q1 - k \cdot IQR$, with $k = 1.5$ or 3.

[1] Definition of *percentile*: each of the 100 equal groups into which a population can be divided according to the distribution of values of a particular variable. *Q1*, which corresponds to the 25th percentile, is the *first quartile*. *Q3*, which corresponds to the 75th percentile, is the *third quartile*.

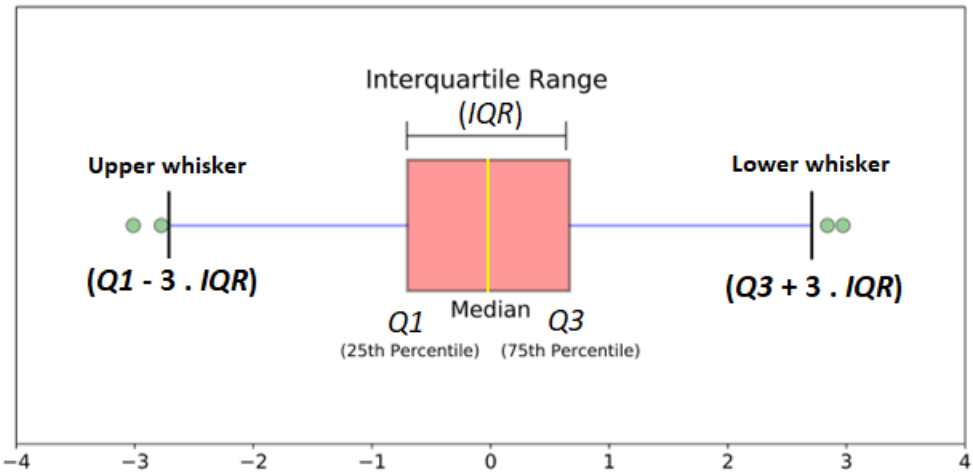

Figure 5 - Example of box-and-whiskers plot (k = 3).

Any individual data point falling above the upper whisker or below the lower whisker can be considered a potential outlier and is represented in Figure 5 by an individual symbol.

In the traditional analysis of indirect verification Charpy lots [1], a simple outlier analysis was performed using, for every machine, box-and-whiskers plot with k = 1.5. This basic analysis, however, was not used to remove individual outlier data points, but only to possibly[2] reject a lot that exhibited more than 5 % of outliers. Currently, this basic analysis does not lead to any consequence and is superseded by the step described below.

An additional level has recently been introduced in the box-and-whiskers analysis as a **potential/suspected outlier screening procedure**[3]: in this case, the box-and-whiskers procedure is applied to all test results combined (from all machines), with the lower whisker corresponding to $Q1 - 3 \cdot IQR$ and the upper whisker to $Q3 + 3 \cdot IQR$ (*i.e.*, k = 3).

If any of the values of absorbed energy obtained from the Charpy tests falls outside these extended whiskers, it is classified as a potential/suspected outlier and may be subject to further analysis in order to establish its statistical significance.

4.4. Modified *z*-score method

The *z-score*[4] is a numerical value that indicates how far a data point is from the mean value. Specifically, the z-score is expressed as the number of standard deviations that the data point is away from the mean, and is simply calculated as:

$$z\text{-score} = \frac{KV_i - \overline{KV}}{s} \quad . \tag{7}$$

Typically, observations with a z-score greater than 3 or smaller than -3 are classified as outliers.

[2] The original language in [1] is: "*If a lot has more than 5 % outliers, it may be rejected.*" The final decision rested on the Charpy Program Coordinator.
[3] This procedure is similar to the *Tukey boxplot rule* [15] cited in ASTM E178-21 [16], which classifies as *potential outliers* data points outside 1.5 of the interquartile range, and *probable outliers* data points outside 3 of the interquartile range.
[4] The z-score is not the same as the Z-value mentioned in section 4.2.

The main drawback in using *z*-scores to detect outliers is that, as the mean is sensitive to the presence of outliers in the data, the *z*-scores will be also.

The *modified z-score* [17] addresses this issue by considering the distance of each data point from the median, which is less sensitive to the presence of outliers, rather than the mean. The formula used to calculate a modified *z*-score, M_i, is:

$$M_i = \frac{0.6745(KV_i - \overline{KV})}{MAD} \quad , \tag{8}$$

where *MAD* is the median absolute deviation of the dataset:

$$MAD = median\left|KV_i - \widetilde{KV}\right| \ , \tag{9}$$

with $\widetilde{KV}$ = median value of the dataset. **According to [17], if the absolute value of M_i is larger than 3.5, the data point can be considered a potential outlier.**

5. Statistical analyses used to establish the significance of potential outliers

The ASTM E178-21 Standard Practice [16] covers outlying observations in samples and how to test the statistical significance of outliers. Four specific tests that are recommended in this document, two for single outliers and two for multiple outliers, are considered in this report, and will be detailed in the following subsections.

5.1. Criteria for single outliers

5.1.1. *T* criterion (section 7.1 of ASTM E178-21)

The *n* values of absorbed energy, *KV*, obtained during the certification process (normally, $n = 75$) must first be sorted in order of increasing magnitude. Suppose the largest value, KV_n, is the potential outlier. The test criterion, T_n, is given by:

$$T_n = \frac{KV_n - \overline{KV}}{s} \quad , \tag{10}$$

where *s* is the estimate of the population standard deviation calculated with Eq.(2). Obviously, if the suspected outlier is the lowest value, KV_1, the numerator in Eq. (5) becomes $\overline{KV} - KV_1$. The critical values, T_{crit}, corresponding to levels of significance $\alpha = 0.01$, 0.05, and 0.1 are given in Table 3 for $n = 3$ to 75. The values are taken from [18], and have been adjusted for division by $n - 1$ instead of n in calculating *s*.

Table 3 - Critical values of *T* (one-sided test) [18].

n	Level of significance, α			*n*	Level of significance, α		
	0.01	0.05	0.1		0.01	0.05	0.1
3	1.155	1.155	1.155	**40**	3.673	3.381	3.240
4	1.499	1.496	1.492	**41**	3.687	3.393	3.251
5	1.780	1.764	1.749	**42**	3.700	3.404	3.261
6	2.011	1.973	1.944	**43**	3.712	3.415	3.271
7	2.201	2.139	2.097	**44**	3.724	3.425	3.282
8	2.358	2.274	2.221	**45**	3.736	3.435	3.292
9	2.492	2.387	2.323	**46**	3.747	3.445	3.302
10	2.606	2.482	2.410	**47**	3.757	3.455	3.310
11	2.705	2.564	2.485	**48**	3.768	3.464	3.319
12	2.791	2.636	2.550	**49**	3.779	3.474	3.329
13	2.867	2.699	2.607	**50**	3.789	3.483	3.336

n	Level of significance, α			n	Level of significance, α		
	0.01	0.05	0.1		0.01	0.05	0.1
14	2.935	2.755	2.659	51	3.798	3.491	3.345
15	2.997	2.806	2.705	52	3.808	3.500	3.353
16	3.052	2.852	2.747	53	3.816	3.507	3.361
17	3.103	2.894	2.785	54	3.825	3.516	3.368
18	3.149	2.932	2.821	55	3.834	3.524	3.376
19	3.191	2.968	2.854	56	3.842	3.531	3.383
20	3.230	3.001	2.884	57	3.851	3.539	3.391
21	3.266	3.031	2.912	58	3.858	3.546	3.397
22	3.300	3.060	2.939	59	3.867	3.553	3.405
23	3.332	3.087	2.963	60	3.874	3.560	3.411
24	3.362	3.112	2.987	61	3.882	3.566	3.418
25	3.389	3.135	3.009	62	3.889	3.573	3.424
26	3.415	3.157	3.029	63	3.896	3.579	3.430
27	3.440	3.178	3.049	64	3.903	3.586	3.437
28	3.464	3.199	3.068	65	3.910	3.592	3.442
29	3.486	3.218	3.085	66	3.917	3.598	3.449
30	3.507	3.236	3.103	67	3.923	3.605	3.454
31	3.528	3.253	3.119	68	3.930	3.610	3.460
32	3.546	3.270	3.135	69	3.936	3.617	3.466
33	3.565	3.286	3.150	70	3.942	3.622	3.471
34	3.582	3.301	3.164	71	3.948	3.627	3.476
35	3.599	3.316	3.178	72	3.954	3.633	3.482
36	3.616	3.330	3.191	73	3.960	3.638	3.487
37	3.631	3.343	3.204	74	3.965	3.643	3.492
38	3.646	3.356	3.216	75	3.971	3.648	3.496
39	3.660	3.369	3.228				

In the case of a suspected outlier in the certification of an indirect verification Charpy lot, the data point is considered a statistically significant outlier if T_n or T_1 are larger than the critical value T_{crit}, with n corresponding to the number of Charpy tests performed (normally 75) and α = 0.05, as given in Table 3.

5.1.2. Dixon criteria for a single outlier (section 7.2 of ASTM E178-21)

The Dixon test [19] is based entirely on ratios of differences between the observations. The relevant statistic changes with the size of the dataset, n. For $n > 13$, the Dixon statistic is given by:

$$r_{22} = \frac{KV_3 - KV_1}{KV_{n-2} - KV_1} \quad \text{if the smallest value } (KV_1) \text{ is the suspected outlier, or} \tag{11}$$

$$r_{22} = \frac{KV_n - KV_{n-2}}{KV_n - KV_3} \quad \text{if the largest value } (KV_n) \text{ is the suspected outlier.} \tag{12}$$

In Eqs. (11) and (12), it is assumed that values are sorted in ascending order, so that x_3 is the third smallest value and x_{n-2} is the third largest value. Critical values of the Dixon statistic are shown in Table 4 for α = 0.1, 0.05, and 0.01, and n between 3 and 100. Values up to n = 50 were originally provided by Dixon [9], but were updated and extended up to n = 100 in [20] and [21].

Table 4 - Critical values of Dixon statistic for testing of extreme observations.

n	Level of significance, α			n	Level of significance, α		
	0.01	0.05	0.1		0.01	0.05	0.1
3	0.885	0.941	0.988	**52**	0.181	0.219	0.293
4	0.679	0.765	0.889	**53**	0.180	0.218	0.292
5	0.558	0.642	0.782	**54**	0.179	0.217	0.290
6	0.484	0.562	0.699	**55**	0.178	0.216	0.288
7	0.434	0.508	0.637	**56**	0.177	0.215	0.287
8	0.398	0.467	0.591	**57**	0.176	0.214	0.286
9	0.370	0.436	0.555	**58**	0.175	0.213	0.285
10	0.349	0.412	0.526	**59**	0.174	0.212	0.283
11	0.331	0.392	0.503	**60**	0.173	0.211	0.282
12	0.317	0.376	0.483	**61**	0.173	0.210	0.281
13	0.305	0.362	0.466	**62**	0.172	0.209	0.279
14	0.294	0.350	0.452	**63**	0.171	0.208	0.278
15	0.285	0.339	0.439	**64**	0.170	0.207	0.278
16	0.277	0.329	0.427	**65**	0.169	0.206	0.277
17	0.269	0.321	0.417	**66**	0.169	0.205	0.275
18	0.263	0.314	0.408	**67**	0.168	0.205	0.274
19	0.256	0.307	0.400	**68**	0.167	0.204	0.274
20	0.251	0.301	0.392	**69**	0.167	0.203	0.272
21	0.246	0.295	0.385	**70**	0.166	0.202	0.271
22	0.242	0.290	0.379	**71**	0.165	0.201	0.271
23	0.238	0.285	0.374	**72**	0.165	0.201	0.270
24	0.234	0.280	0.367	**73**	0.164	0.200	0.268
25	0.230	0.276	0.363	**74**	0.164	0.199	0.268
26	0.227	0.273	0.358	**75**	0.163	0.198	0.267
27	0.224	0.269	0.354	**76**	0.163	0.198	0.266
28	0.221	0.266	0.350	**77**	0.162	0.197	0.266
29	0.218	0.262	0.346	**78**	0.161	0.196	0.265
30	0.216	0.259	0.343	**79**	0.161	0.196	0.264
31	0.213	0.257	0.339	**80**	0.160	0.195	0.263
32	0.211	0.254	0.336	**81**	0.160	0.194	0.262
33	0.209	0.251	0.332	**82**	0.159	0.194	0.261
34	0.207	0.249	0.329	**83**	0.159	0.193	0.261
35	0.205	0.247	0.327	**84**	0.158	0.193	0.260
36	0.203	0.245	0.324	**85**	0.158	0.192	0.259
37	0.201	0.242	0.321	**86**	0.157	0.192	0.258
38	0.199	0.241	0.319	**87**	0.157	0.191	0.257
39	0.197	0.238	0.316	**88**	0.156	0.191	0.257
40	0.196	0.237	0.314	**89**	0.156	0.190	0.257
41	0.194	0.235	0.312	**90**	0.155	0.190	0.256
42	0.193	0.233	0.310	**91**	0.155	0.189	0.255
43	0.192	0.232	0.308	**92**	0.154	0.189	0.254
44	0.190	0.230	0.306	**93**	0.154	0.188	0.254
45	0.189	0.229	0.305	**94**	0.154	0.188	0.254
46	0.188	0.227	0.303	**95**	0.153	0.187	0.252
47	0.187	0.226	0.301	**96**	0.153	0.187	0.252
48	0.185	0.224	0.299	**97**	0.152	0.186	0.251
49	0.184	0.223	0.297	**98**	0.152	0.186	0.251
50	0.183	0.222	0.296	**99**	0.152	0.185	0.250
51	0.182	0.221	0.294	**100**	0.151	0.185	0.250

ASTM E178-21 [16] remarks that "*in most situations, the Dixon criteria is less powerful at detecting an outlier than the criterion given in 7.1*" (*T* criterion, herein described in section 5.1.1).

In the case of a suspected outlier in the certification of an indirect verification Charpy lot, the smallest or the largest data point is considered a statistically significant outlier if r_{22}, given by Eq. (11) or (12), is larger than the critical value corresponding to the number of tests performed n and a level of confidence $\alpha = 0.05$, as provided in Table 4.

5.1.3. Grubbs' test

Grubbs test [22], also known as the *maximum normed residual test*, is used to detect a single outlier in a dataset that follows an approximately normal distribution.

The test statistic corresponds to the largest absolute deviation from the mean, normalized by the standard deviation, and is defined as:

$$G = \frac{max|KV_i - \overline{KV}|}{s} \quad . \tag{13}$$

For a one-sided test, the data point corresponding to the statistic G is considered a statistically significant outlier at a significance level α, if:

$$G > \frac{n-1}{\sqrt{n}} \sqrt{\frac{\left(t_{\alpha/n,n-2}\right)^2}{n-2+\left(t_{\alpha/n,n-2}\right)^2}} \quad , \tag{14}$$

with $t_{\alpha/n,n-2}$ denoting the critical value of the t distribution with n-2 degrees of freedom and a significance level of α/n.

5.2. Criteria for multiple outliers

According to ASTM E178-21 [16], "*recursive application of a test for a single outlier may be used.*" This entails that, if a suspected outlier (lowest or highest value in the dataset) is identified as statistically significant using the tests in section 5.1.1 (T criterion) and/or section 5.1.2 (Dixon criteria), the same tests can be run again on the modified dataset after removing the significant outlier. However, the standard also warns: "*The performance of most tests for single outliers is affected by masking, where the probability of detecting an outlier using a test for a single outlier is reduced when there are two or more outliers. Therefore, the recommended procedure is to use a criterion designed to test for multiple outliers, using recursive testing to investigate after the initial criterion is significant.*"

5.2.1. David, Hartley, and Pearson criterion for two outliers on opposite sides of a sample (section 7.4 of ASTM E178-21)

If both the lowest value and the largest value in a dataset are suspected outliers, the test proposed by David, Hartley, and Pearson [23] can be used. This test uses the ratio between the sample width, w, and its standard deviation, s:

$$\frac{w}{s} = \frac{x_n - x_1}{s} \tag{15}$$

The critical values of the $\frac{w}{s}$ statistic corresponding to significance levels of 0.1, 0.05, and 0.01 are provided in Table 5. The approximate critical value corresponding to n = 75 and α = 0.05, obtained by linear interpolation, is 5.59.

Table 5 - Critical values of the $\frac{w}{s}$ statistic. Each entry was calculated by 50,000,000 simulations.

n	Level of significance, α		
	0.01	0.05	0.1
3	1.997	1.999	2.000
4	2.409	2.429	2.445
5	2.712	2.755	2.803
6	2.949	3.012	3.095
7	3.143	3.222	3.338
8	3.308	3.399	3.543
9	3.449	3.552	3.720
10	3.574	3.685	3.875
11	3.684	3.803	4.011
12	3.782	3.909	4.133
13	3.871	4.005	4.244
14	3.952	4.092	4.344
15	4.025	4.171	4.435
16	4.093	4.244	4.519
17	4.156	4.311	4.597
18	4.214	4.374	4.669
19	4.269	4.433	4.736
20	4.320	4.487	4.799
21	4.368	4.539	4.858
22	4.413	4.587	4.913
23	4.456	4.633	4.965
24	4.497	4.676	5.015
25	4.535	4.717	5.061
26	4.572	4.756	5.106
27	4.607	4.793	5.148
28	4.641	4.829	5.188
29	4.673	4.863	5.226
30	4.704	4.895	5.263
35	4.841	5.040	5.426
40	4.957	5.162	5.561
45	5.057	5.265	5.674
50	5.144	5.356	5.773
60	5.29	5.50	5.93
80	5.51	5.73	6.18
100	5.68	5.90	6.36

In the case of a suspected outlier in the certification of an indirect verification Charpy lot, the smallest or the largest data point is considered a statistically significant outlier if $\frac{w}{s}$, calculated by means of Eq. (15), is larger than the critical value corresponding to the number of tests performed *n* and a level of confidence $\alpha = 0.05$.

5.2.2. Skewness and kurtosis criteria (section 7.8 of E178-21)

ASTM Standard Practice E2586 [24] provides the definition of skewness and kurtosis. The *skewness* of a population or sample is a measure of symmetry of the distribution, while the *kurtosis* of a population or sample is a measure of the weight of the tails of a distribution relative to the center.

The coefficient of skewness for a Charpy dataset is given by:

$$g_1 = \frac{n\sum(KV_i - \overline{KV})^3}{(n-1)(n-2)s^3} \quad , \tag{16}$$

and in the presence of several suspected outliers, should be used to test against change in level of several observations in the same direction.

The coefficient of kurtosis for a Charpy dataset is expressed as:

$$g_2 = \frac{n(n+1)\sum(KV_i-\overline{KV})^4}{(n-1)(n-2)(n-3)s^4} - \frac{3(n-1)^2}{(n-2)(n-3)} \ , \tag{17}$$

and is recommended to test against change in level to both higher and lower values, and also for changes in scale (variance). After computing g_1 and g_2, if their values exceed the significance levels given in Table 6 (skewness) and Table 7 (kurtosis), then the observation that is farthest from the mean is rejected, and the procedure is repeated until no more outliers are identified.

Table 6 – Significance levels for skewness g_1. Each entry was calculated by 50,000,000 simulations.

n	Level of significance, α		
	0.1	0.05	0.01
3	1.647	1.711	1.731
4	1.439	1.709	1.940
5	1.224	1.564	1.994
6	1.090	1.428	1.959
7	1.014	1.320	1.886
8	0.956	1.246	1.813
9	0.903	1.183	1.735
10	0.862	1.131	1.668
11	0.828	1.086	1.610
12	0.798	1.049	1.556
13	0.770	1.011	1.504
14	0.744	0.977	1.461
15	0.722	0.950	1.418
16	0.702	0.922	1.379
17	0.684	0.899	1.345
18	0.667	0.875	1.310
19	0.651	0.856	1.281
20	0.636	0.836	1.252
21	0.624	0.818	1.225
22	0.610	0.800	1.196
23	0.599	0.786	1.175
24	0.587	0.770	1.150
25	0.578	0.757	1.132
26	0.567	0.743	1.108
27	0.558	0.731	1.091
28	0.549	0.718	1.070
29	0.541	0.708	1.056
30	0.532	0.695	1.036
35	0.497	0.649	0.965
40	0.467	0.610	0.904
45	0.442	0.578	0.853
50	0.422	0.551	0.812

Table 7 – Significance levels for kurtosis g_2. Each entry was calculated by 50,000,000 simulations.

n	Level of significance, α		
	0.1	0.05	0.01
4	3.075	3.518	3.900
5	2.772	3.506	4.454
6	2.482	3.319	4.685
7	2.257	3.110	4.735
8	2.067	2.935	4.687
9	1.904	2.772	4.586
10	1.778	2.627	4.467
11	1.678	2.505	4.350
12	1.597	2.399	4.234
13	1.529	2.300	4.106

n	Level of significance, α		
	0.1	0.05	0.01
14	1.471	2.217	4.000
15	1.422	2.145	3.887
16	1.378	2.081	3.784
17	1.340	2.021	3.702
18	1.303	1.966	3.605
19	1.271	1.921	3.524
20	1.243	1.873	3.450
21	1.214	1.831	3.370
22	1.188	1.788	3.298
23	1.167	1.757	3.233
24	1.143	1.719	3.169
25	1.123	1.690	3.116
26	1.102	1.658	3.051
27	1.085	1.630	2.995
28	1.066	1.601	2.943
29	1.052	1.578	2.903
30	1.035	1.550	2.845
35	0.969	1.446	2.642
40	0.913	1.358	2.470
45	0.867	1.285	2.322
50	0.830	1.223	2.210

Significance levels in Table 6 and Table 7 are provided in [16] up to $n = 50$. The approximate values corresponding to $n = 75$ can be estimated by extrapolating power law regressions of the tabulated data[5], and correspond to 0.471 for skewness and 1.025 for kurtosis.

5.2.3. Tietjen-Moore Test for multiple outliers

The Tietjen-Moore test [25] is a generalization of the Grubbs' test in case of multiple outliers in a dataset that follows an approximately normal distribution. It is important to note that this test requires the exact number of suspected outliers, *m*, to be known.

After sorting the absorbed energy values in ascending order, the test statistics for the *m* largest *KV* values is

$$L_m = \frac{\sum_{i=1}^{n-m}(KV_i - \overline{KV_m})^2}{\sum_{i=1}^{n}(KV_i - \overline{KV})^2}, \tag{18}$$

while for the smallest *m* points it is given by:

$$L_m = \frac{\sum_{i=m+1}^{n}(KV_i - \overline{KV_m})^2}{\sum_{i=1}^{n}(KV_i - \overline{KV})^2}, \tag{19}$$

where $\overline{KV_m}$ indicates the mean absorbed energy of the dataset after removing the largest *m* values.

The critical region for the Tietjen-Moore test is determined by simulation, performed by generating a standard normal random sample, typically of 10,000 samples. The critical values of L_m are given in Table 8 for *k* between 1 and 5, *n* up to 50, and significance levels $\alpha = 0.1$, 0.05, and 0.01.

[5] The equation of the power law regression for the values in Table 6 (skewness) is $g_1 = 3.0393\,n^{-0.432}$ (coefficient of determination $R^2 = 0.9883$). For kurtosis, the fitting equation is $g_2 = 7.3438n^{-0.456}$ ($R^2 = 0.9862$).

Table 8 - Critical values of L_m for the Tietjen-Moore test.

m	1			2			3			4			5		
n / α	0.1	0.05	0.01	0.1	0.05	0.01	0.1	0.05	0.01	0.1	0.05	0.01	0.1	0.05	0.01
3	0.011	0.003	0	…	…	…	…	…	…	…	…	…	…	…	…
4	0.098	0.049	0.01	0.003	0.001	0	…	…	…	…	…	…	…	…	…
5	0.199	0.127	0.044	0.038	0.018	0.004	…	…	…	…	…	…	…	…	…
6	0.283	0.203	0.093	0.092	0.056	0.019	0.02	0.01	0.002	…	…	…	…	…	…
7	0.35	0.27	0.145	0.148	0.102	0.044	0.056	0.032	0.01	…	…	…	…	…	…
8	0.405	0.326	0.195	0.199	0.148	0.075	0.095	0.064	0.028	0.038	0.022	0.008	…	…	…
9	0.45	0.374	0.241	0.245	0.191	0.108	0.134	0.099	0.048	0.068	0.045	0.018	…	…	…
10	0.488	0.415	0.283	0.286	0.23	0.141	0.17	0.129	0.07	0.098	0.07	0.032	0.051	0.034	0.012
11	0.52	0.451	0.321	0.323	0.267	0.174	0.208	0.162	0.098	0.128	0.098	0.052	0.074	0.054	0.026
12	0.548	0.482	0.355	0.355	0.3	0.204	0.24	0.196	0.12	0.159	0.125	0.07	0.103	0.076	0.038
13	0.573	0.51	0.386	0.384	0.33	0.233	0.27	0.224	0.147	0.186	0.15	0.094	0.126	0.098	0.056
14	0.594	0.534	0.414	0.411	0.357	0.261	0.298	0.25	0.172	0.212	0.174	0.113	0.15	0.122	0.072
15	0.613	0.556	0.44	0.435	0.382	0.286	0.322	0.276	0.194	0.236	0.197	0.132	0.172	0.14	0.09
16	0.631	0.576	0.463	0.456	0.405	0.31	0.342	0.3	0.219	0.26	0.219	0.151	0.194	0.159	0.108
17	0.646	0.593	0.485	0.476	0.426	0.332	0.364	0.322	0.237	0.282	0.24	0.171	0.216	0.181	0.126
18	0.66	0.61	0.504	0.494	0.446	0.353	0.384	0.337	0.26	0.302	0.259	0.192	0.236	0.2	0.14
19	0.673	0.624	0.522	0.511	0.464	0.373	0.398	0.354	0.272	0.316	0.277	0.211	0.251	0.217	0.154
20	0.685	0.638	0.539	0.527	0.48	0.391	0.42	0.377	0.3	0.339	0.299	0.231	0.273	0.238	0.175
25	0.732	0.692	0.607	0.591	0.55	0.468	0.489	0.45	0.377	0.412	0.374	0.308	0.35	0.312	0.246
30	0.766	0.73	0.65	0.637	0.601	0.527	0.523	0.506	0.434	0.472	0.434	0.369	0.411	0.376	0.312
35	0.792	0.762	0.69	0.674	0.641	0.573	0.586	0.554	0.484	0.516	0.482	0.418	0.458	0.424	0.364
40	0.812	0.784	0.722	0.702	0.673	0.61	0.622	0.588	0.522	0.554	0.523	0.46	0.499	0.468	0.408
45	0.826	0.802	0.745	0.726	0.698	0.641	0.648	0.618	0.558	0.586	0.556	0.498	0.533	0.502	0.444
50	0.84	0.82	0.768	0.746	0.72	0.667	0.673	0.646	0.592	0.614	0.588	0.531	0.562	0.535	0.483

Critical values for $n = 75$ can be obtained by fitting a line through L_m values between 25 and 50 in Table 8 and then extrapolating to $n = 75$. The following approximate critical values are calculated:

- For $m = 2$, $L_{m,crit} = 0.823$.
- For $m = 3$, $L_{m,crit} = 0.763$.

In the case of *m* suspected outliers in the certification of an indirect verification Charpy lot based on 75 impact tests, the *m* test results are considered statistically significant outliers at a significance level $\alpha = 0.05$, if L_m calculated by means of Eq. (18) or Eq. (19) is smaller than the corresponding critical values $L_{m,crit}$ provided above.

5.3. Single or multiple outliers: Generalized Extreme Studentized Deviate (ESD) test

The generalized ESD (Extreme Studentized Deviate) test [26] is used to detect one or more outliers in a dataset that follows an approximately normal distribution[6]. This test, unlike others (Grubbs and Tietjen-Moore), does not require the exact number of outliers to be known, but just an upper bound for the suspected number of outliers (*r*).

The statistic is:

$$R_i = \frac{max|KV_i - \overline{KV}|}{s} \quad . \tag{20}$$

[6] Datasets constituted by absorbed energy values from the certification of indirect verification Charpy lots are expected to follow, for the most part, a normal distribution.

After removing the value that maximizes $|KV_i - \overline{KV}|$, the statistic of Eq. (20) is recalculated with $n - 1$ observations. The process is repeated until r values have been removed.

For each of the r values, the following critical value is computed:

$$\lambda_i = \frac{(n-i)t_{p,n-i-1}}{\sqrt{\left(n-i-1+t_{p,n-i-1}^2\right)(n-i+1)}} \quad , \tag{21}$$

where i is between 1 and r, and $t_{p,v}$ is the value from the t distribution corresponding to $p = \alpha/n$ ($\alpha = 0.05$) and $v = n-i+1$, and

$$p = 1 - \frac{\alpha}{2(n-i+1)} \quad . \tag{22}$$

The number of significant outliers is determined by finding the largest i such that $R_i > \lambda_i$. In other words, if $R_i > \lambda_i$, the corresponding data point is a significant outlier.

6. Overview of the NIST strategy for handling outliers in the certification of indirect verification Charpy lots

(a) *Section 4.2* – Check the normality of the dataset using a Q-Q plot and four normality tests. If, after removing potential outliers, the dataset is not normally, or approximately normally, distributed according to at least 3 of the 4 normality tests, perform a power transformation on the values of absorbed energy.

(b) *Section 4.3* – Use a box-and-whisker plot (combining all tests performed on the three reference machines) to depict absorbed energy values, and identify data points falling outside the whiskers corresponding to $Q1 - 3 \cdot IQR$ and $Q3 + 3 \cdot IQR$.

(c) *Section 4.4* – Identify all data points for which the modified z-score is larger than 3.5.

(d) Select all test results that have been flagged as potential/suspected outliers in any of the above steps.

(e) If only one potential/suspected outlier has been identified in step (d), any one of the following statistical tests can be applied to confirm the significance of the outlier:

- *Section 5.1.1* – Evaluate the T criterion and compare with the critical value in Table 3.
- *Section 5.1.2* – Run the single-outlier Dixon test, by evaluating r_{22} and comparing it with the critical value in Table 4.
- *Section 5.1.3* – Run Grubbs' test, by evaluating the statistic G and comparing it with the critical value given by Eq. (14).
- *Section 5.3* – Apply the generalized ESD test, by computing the statistic R_i for the potential/suspected outlier and comparing it with the critical value λ_i.

If any of these tests confirms the statistical significance of the potential/suspected outlier, additional investigations are conducted to establish whether the data point can be considered an experimental or physical outlier. If this is the case, the corresponding value of absorbed

energy is removed from the dataset, and the sample size of the lot is recalculated, Eq. (5). If $n_{SS} \leq 5.0$, the lot is accepted[7]; if $n_{SS} > 5.0$, the lot is rejected.

(f) If multiple potential/suspected outliers have been identified in step (c), any one of the following statistical tests can be applied to confirm the significance of the outliers:

- *Section 5.2.1* – If both the lowest and the largest absorbed energy values are potential/suspected outliers, calculate the ratio *w/s* (David, Hartley, and Pearson test) and compare it with the critical value in Table 5. **Do not use if all potential/suspected outliers are on the same side of the dataset.**
- *Section 5.2.2* – Evaluate the coefficient of skewness, g_1, and the coefficient of kurtosis, g_2, for the dataset and compare them with the significance levels in Table 6 and Table 7, respectively.
- *Section 5.2.3* – Apply the Tietjen-Moore test on the m largest or smallest absorbed energy values, by calculating the test statistics L_m and comparing them with the critical values in Table 8.
- *Section 5.3* – Apply the generalized ESD test, by computing the statistic R_i for each potential/suspected outlier and comparing it with the corresponding critical value λ_i.

If any of these tests (as applicable) confirm the statistical significance of the potential/suspected outliers, the corresponding tests are subjected to further analyses to ascertain their potential experimental or physical outlier nature. For any test deemed to be an experimental or physical outlier, the corresponding value of absorbed energy is removed from the dataset, and the sample size of the lot is recalculated, Eq. (5). If $n_{SS} \leq 5.0$, the lot is accepted[7]; if $n_{SS} > 5.0$, the lot is rejected.

If any absorbed energy value is removed from the original indirect verification Charpy dataset, the reference (certified) absorbed energy of the lot, KV_{ref}, and its associated uncertainties (standard, u, and expanded, U) must be recalculated.

7. Treatment of potential/suspected outliers among customers' test results

From time to time, an individual value of absorbed energy KV reported by a customer of the NIST Charpy Program appears to be distant from the other values in the verification set, and significantly different from the certified absorbed energy, KV_{ref}.

To assess the significance of the potential outlier, a simple statistical procedure (described in ASTM E178-21 under *Alternative Outlier Procedures*, section 9.7), *Hampel's Rule*, can be applied. Also known as *Hampel Identifier* or *Hampel Filter*, it is a variation of the three-sigma rule of statistics, which is robust against outliers [27], and consists of the following steps [16]:

1. Establish the median value of absorbed energy, $\widetilde{KV}$, for the indirect verification Charpy lot, based on the certification results of the production lot or the combined pilot + production lot.[8]

[7] Assuming the dimensional inspection of the specimen is successful, see section 2.1 and [5].
[8] The certified absorbed energy and the expanded uncertainty of a Charpy lot are calculated after combining pilot and production lot results if there is no statistical difference between the means and standard deviations of the two lots. Conversely, if statistical differences are found between the means or the standard deviations, or both, only data from the production lot is used [1].

2. Calculate the median absolute deviation (MAD), or median of the absolute deviations from the data's median, as:

$$MAD = median\left(\left|KV_i - \widetilde{KV}\right|\right) \quad . \tag{25}$$

3. If the absolute difference between the customer's suspected outlier and the median absorbed energy of the indirect verification Charpy lot is larger than $3 \cdot MAD$, the following investigation steps shall be conducted:

    a. A visual examination of the tested specimen. If marks are clearly visible on the broken sample indicating bad positioning of the sample against the machine anvils, or jamming between the specimen and any part of the machine, the result should be excluded from further analyses. Too much or too little plastic deformation in comparison to the other specimens in the set, or with respect to the nominal energy level, might also indicate that the test has been performed at the wrong temperature.[9]

    b. If the visual inspection does not provide useful clues, then the customer should be directly contacted and asked whether anything unusual has occurred during the test or when reporting the test results (*e.g.*, typo). In the former case, the result will be discarded and the indirect verification of the customer's machine will be based on the remaining four tests; in the latter, the corrected value will be used in the subsequent analyses.

    Note that the above steps can always be performed based on engineering judgment, even if the statistical test is not significant.

4. If these circumstances affect more than one test value, the customer will be provided free of charge another set of five reference specimens, and the indirect verification will be repeated.

8. Practical examples of Charpy outlier detection and assessment

Application examples for some of the procedures described in this report are provided in Appendix A (one significant outlier), Appendix B (two significant outliers), and Appendix C (outlier customer's result).

9. Future work

From a scientific/statistical standpoint, the ideal approach to dealing with apparent outliers in the certification of Charpy lots would be to use statistical models that are robust (more robust than the normal distribution) to the presence of outliers in the data. Examples of these alternative distributions would be the Laplace model, also called double exponential distribution, or a rescaled and shifted Student's t model (with the number of degrees of freedom as a tunable parameter). A more general solution would be to consider a hierarchical Bayesian model with machine effects modeled as a sample from a normal distribution and within-machine effects modeled by a Laplace or rescaled Student's t distribution.

Implementing an alternative, more robust to the presence of outliers, distribution will be the focus of future efforts within the NIST Charpy machine verification program. This will also entail modifying the criteria used for assessing the acceptability of a Charpy verification lot,

[9] Most likely, however, if the wrong test temperature has been set, this would affect all five Charpy tests in a set.

which cannot be based anymore on the concept of sample size as described in 2.3, which is strictly dependent on the assumption that certification data follow a normal distribution.

References

[1] McCowan CN, Siewert TA, and Vigliotti DP (2003) The NIST Charpy V-Notch Verification Program: Overview and Operating Procedures. (National Institute of Standards and Technology, Boulder, CO), NIST Interagency/Internal Report (NISTIR) - 1500-9 https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=851238 (Accessed January 9, 2024).

[2] ASTM International (2023) *ASTM E23-23a – Standard Test Methods for Notched Impact Bar Testing of Metallic Materials* (ASTM International, West Conshohocken, PA). https://doi.org/10.1520/E0023-23A

[3] ISO (2016) *ISO 148-2:2016 – Metallic materials – Charpy Pendulum Impact Test – Part 2: Verification of Testing Machines* (International Standards Organization, Geneva, Switzerland).

[4] Lucon E (2023) Test Temperature Range for NIST Certified Charpy Specimens for Testing at "Room Temperature". (National Institute of Standards and Technology, Boulder, CO), NIST Interagency/Internal Report (NISTIR) – 8470 https://doi.org/10.6028/NIST.IR.8470

[5] Lucon E, Eckhardt AC, and Santoyo RL (2024) Dimensional Inspection of Charpy Indirect Verification Specimens by Means of a Digital Optical Comparator. (National Institute of Standards and Technology, Boulder, CO), NIST Interagency/Internal Report (NISTIR) – in publication.

[6] Mandel J (1991) The validation of measurement through interlaboratory studies. *Chemometrics and Intelligent Laboratory Systems* 11(2), p.111. https://doi.org/10.1016/0169-7439(91)80058-X

[7] ISO (2016) *ISO 148-1:2016 – Metallic materials – Charpy Pendulum Impact Test – Part 1: Test Method* (International Standards Organization, Geneva, Switzerland).

[8] "What are outliers in the data?" (2012) *NIST/SEMATECH e-Handbook of Statistical Methods*, section 7.1.6, https://www.itl.nist.gov/div898/handbook/prc/section1/prc16.htm

[9] Stephens MA (1974) EDF Statistics for Goodness of Fit and Some Comparisons. *Journal of the American Statistical Association* 69(347): 730–737. https://doi.org/10.1080/01621459.1974.10480196

[10] Shapiro SS and Wilk MB (1965) An analysis of variance test for normality (complete samples). *Biometrika* 52(3-4): 591-611. https://doi.org/10.2307/2333709

[11] D'Agostino RB and Pearson ES (1973) Tests for Departure from Normality. Empirical Results for the Distributions of b2 and √b1. *Biometrika* 60(3): 613–622. https://doi.org/10.2307/2335012

[12] Jarque CM and Bera AK (1980) Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics Letters* 6(3): 255–259. https://doi.org/10.1016/0165-1765(81)90035-5

[13] Box GEP and Cox DR (1964) An Analysis of Transformations. *Journal of the Royal Statistical Society. Series B (Methodological)* 26(2): 211-252. https://doi.org/10.21236/ada110447

[14] du Toit SHC, Steyn AGW, and Stumpf RH (1986) *Graphical Exploratory Data Analysis* (Springer-Verlag Berlin and Heidelberg GmbH & Co. KG), 1st Ed.

[15] Tukey JW (1977) *Exploratory Data Analysis* (Addison-Wesley Publishing Company, Reading, MA).
[16] ASTM International (2021) *ASTM E178-21 – Standard Practice for Dealing with Outlying Observations* (ASTM International, West Conshohocken, PA). https://doi.org/10.1520/E0178-21
[17] Iglewitz B and Hoaglin D (1993) Volume 16: How to Detect and Handle Outliers. *The ASQC Basic References in Quality Control: Statistical Techniques* (ASQC Quality Press, Milwaukee, WI).
[18] Grubbs FE and Beck G (1972) Extension of Sample Sizes and Percentage Points for Significance Tests of Outlying Observations. *Technometrics*, TCMTA, 14(4): 847-854. https://doi.org/10.1080/00401706.1972.10488981
[19] Dixon WJ (1953) Processing Data for Outliers. *Biometrics*, BIOMA, 9(1): 74-89. https://doi.org/10.2307/3001634
[20] Verma SP and Quiroz-Ruiz A (2006) Critical Values for Six Dixon Tests for Outliers in Normal Samples up to Sizes 100, and Applications in Science and Engineering. *Revista Mexicana de Ciencias Geologicas* 23(2): 133-161.
[21] Bohrer A (2008) One-sided and Two-sided Critical Values for Dixon's Outlier Test for Sample Sizes up to n=30. *Economic Quality Control* 23(1): 5-13. https://doi.org/10.1515/EQC.2008.5
[22] Grubbs FE (1969) Procedures for Detecting Outlying Observations in Samples. *Technometrics*, TCMTA, 11(4): 1-12. https://doi.org/10.1080/00401706.1969.10490657
[23] David HA, Hartley HO, and Pearson ES (1954) The Distribution of the Ratio, in a Single Normal Sample, of Range to Standard Deviation. *Biometrika*, BIOKA, 41: 482–493.
[24] ASTM International (2019) *ASTM E2586-19$^{\varepsilon 1}$ – Standard Practice for Calculating and Using Basic Statistics* (ASTM International, West Conshohocken, PA). https://doi.org/10.1520/E2586-19E01
[25] Tietjen GL and Moore RH (1972) Some Grubbs-Type Statistics for the Detection of Several Outliers. *Technometrics*, TCMTA, 14(3): 583–597. https://doi.org/10.1080/00401706.1972.10488948
[26] Rosner B (1983) Percentage Points for a Generalized ESD Many-Outlier Procedure. *Technometrics*, TCMTA, 25(2): 165-172. https://doi.org/10.1080/00401706.1983.10487848
[27] Hampel FR (1974) The influence curve and its role in robust estimation. *Journal of the American Statistical Association* 69: 382-393. https://doi.org/10.1080/01621459.1974.10482962

Appendix A. Lot with one statistically significant outlier

A.1. SH-67 pilot lot certification results

The results of the Charpy tests for the certification of the super-high energy SH-67 pilot lot are shown in Table A.1. Seventy-five tests were performed at 21 °C, 25 on each of the three NIST reference machines. One of the absorbed energy values (272 J) is significantly higher than any other data point, and can be considered a potential/suspected outlier.

The sample size of the original dataset is $n_{SS} = 8.678$, and therefore unacceptable.

Table A.1 – Certification test results for SH-67 pilot lot. The suspected outlier is highlighted in **bold red**.

Machine	KV (J)	Machine	KV (J)	Machine	KV (J)
SI	195.37	TO	208.15	TK	202.83
	207.85		207.77		195.26
	201.48		199.86		207.94
	193.97		202.61		205.65
	198.52		209.45		207.94
	203.24		203.55		200.88
	195.64		202.42		197.94
	191.00		203.74		204.77
	196.57		199.19		200.97
	200.09		198.15		200.88
	198.70		200.52		213.17
	196.39		208.52		206.71
	208.31		209.45		213.43
	201.57		213.44		214.30
	206.38		207.77		216.20
	213.19		200.24		218.01
	200.74		203.83		193.02
	202.22		204.02		202.65
	209.88		214.09		211.35
	205.36		213.99		204.51
	272.00		202.98		203.80
	210.89		208.52		197.41
	196.94		203.83		200.79
	198.33		198.62		208.55
	196.76		193.65		204.86

A.2. Normality checks

The Q-Q plot of the original SH-67 dataset is provided in Figure A.1. All the tests indicate non-normality of the original dataset, with the following *p*-values: 1.04×10^{-8} (A-D), 9.74×10^{-12} (S-W), 0 (dA-P), and 0 (J-B).

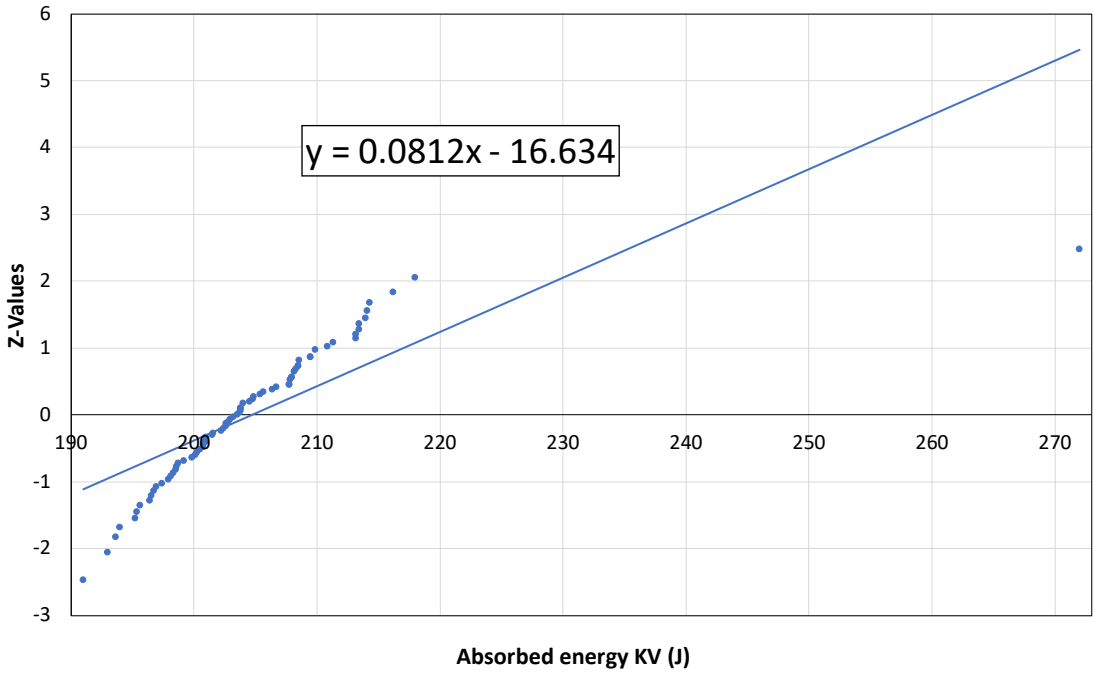

Figure A.1 – Q-Q plot of the original SH-67 dataset.

After removing the highest test result (Figure A.2), the dataset becomes normal according to all the normality tests ($p = 0.444$ for A-D, $p = 0.516$ for S-W, $p = 0.445$ for dA-P, and $p = 0.469$ for J-B). The analysis can therefore proceed to the next step.

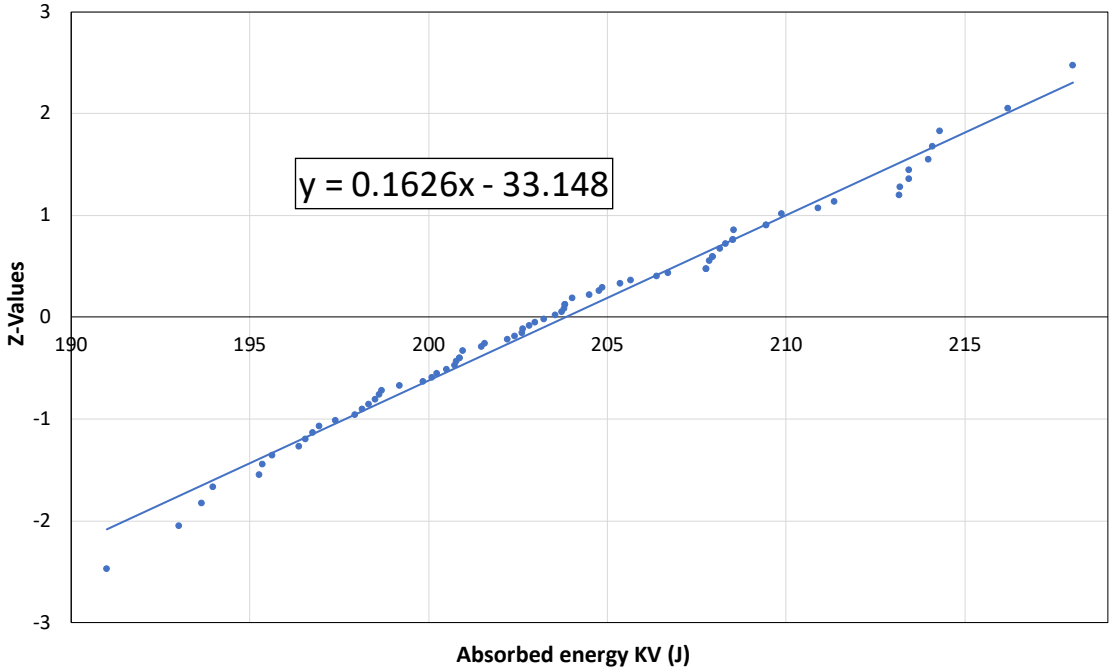

Figure A.2 – Q-Q plot for SH-67 after removing the suspected outlier.

A.3. Potential/suspected outlier screening

The box-and-whiskers plot for the original dataset (Figure A.3) shows that the highest absorbed energy value (272 J) falls way above the upper whisker, corresponding to $Q3 + 3 \cdot IQR$, and can therefore be considered a potential/suspected outlier.

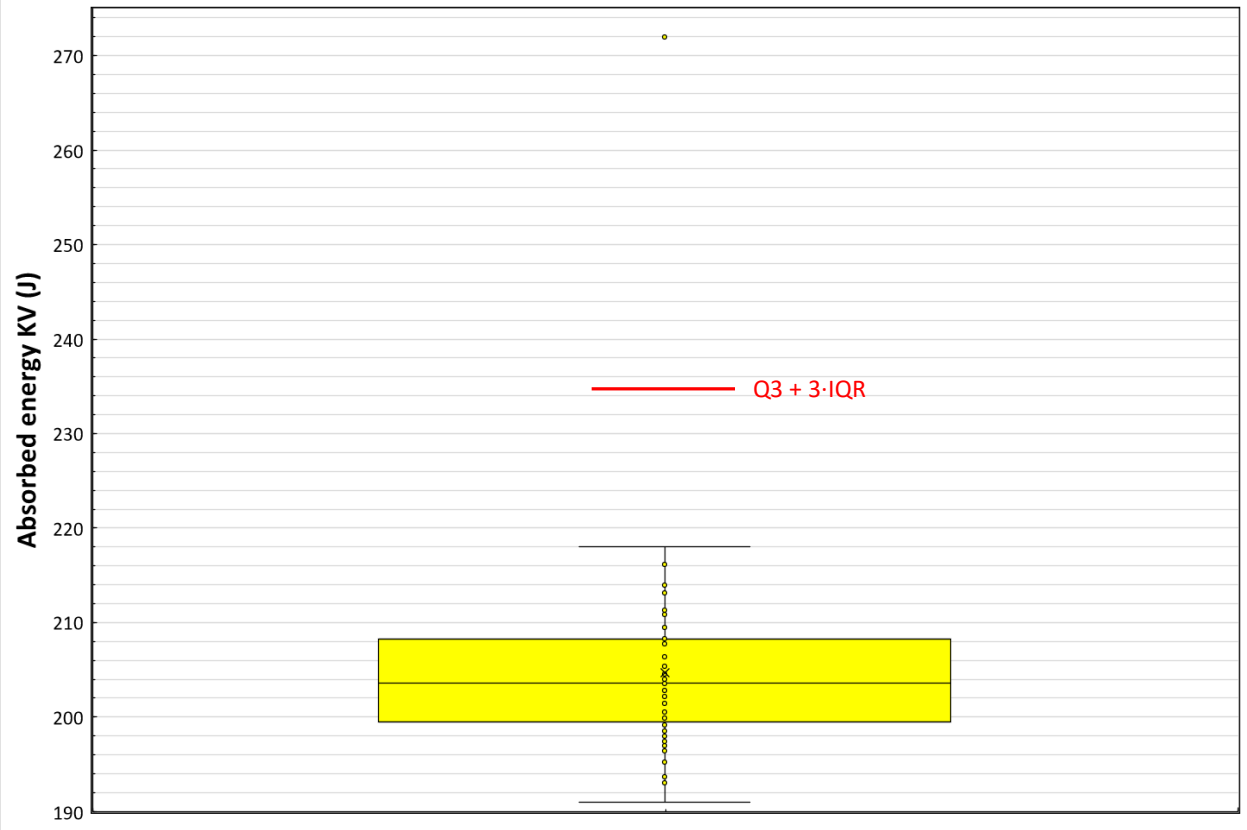

Figure A.3 – Box-and-whiskers plot showing one potential/suspected outlier.

Next, the modified z-score method (section 4.4) will be applied to confirm the potential/suspected outlier.

The absorbed energy values, sorted in ascending order, and the corresponding values of M_i calculated using eq. (8), are shown in Table A.2. For this dataset, $\widetilde{KV} = 203.55$ J and $MAD = 4.6$ J. The highest absorbed energy value has $M_i = 10.04 > 3.5$, and is therefore confirmed as a potential/suspected outlier.

In the following analysis steps aimed at assessing the statistical significance of the outlier, the criteria applicable to a single outlier (section 5.1) will be applied.

Table A.2 – Modified *z*-scores for the SH-67 pilot lot.

KV (J)	M_i	*KV* (J)	M_i	KV (J)	M_i
191.00	1.84	200.88	0.39	207.77	0.62
193.02	1.54	200.88	0.39	207.77	0.62
193.65	1.45	200.97	0.38	207.85	0.63
193.97	1.40	201.48	0.30	207.94	0.64
195.26	1.22	201.57	0.29	207.94	0.64
195.37	1.20	202.22	0.20	208.15	0.67
195.64	1.16	202.42	0.17	208.31	0.70
196.39	1.05	202.61	0.14	208.52	0.73
196.57	1.02	202.65	0.13	208.52	0.73
196.76	1.00	202.83	0.11	208.55	0.73
196.94	0.97	202.98	0.08	209.45	0.87
197.41	0.90	203.24	0.05	209.45	0.87
197.94	0.82	203.55	0.00	209.88	0.93
198.15	0.79	203.74	0.03	210.89	1.08
198.33	0.77	203.80	0.04	211.35	1.14
198.52	0.74	203.83	0.04	213.17	1.41
198.62	0.72	203.83	0.04	213.19	1.41
198.70	0.71	204.02	0.07	213.43	1.45
199.19	0.64	204.51	0.14	213.44	1.45
199.86	0.54	204.77	0.18	213.99	1.53
200.09	0.51	204.86	0.19	214.09	1.55
200.24	0.49	205.36	0.27	214.30	1.58
200.52	0.44	205.65	0.31	216.20	1.85
200.74	0.41	206.38	0.41	218.01	2.12
200.79	0.40	206.71	0.46	**272.00**	**10.04**

A.4. Assessment of statistical outlier's significance: *T* criterion

The outlying nature of the highest point (272.00 J) is very clear based on both the Q-Q plot and the box-and-whiskers plot. To confirm its statistical significance, the *T* criterion is selected.

For the potential/suspected outlier, the value of the test criterion is 6.777 according to eq. (10), which is higher than the critical value in Table 3 for $n = 75$ and $\alpha = 0.05$ ($T_{crit} = 3.648$).

The potential/suspected outlier is therefore **statistically significant** at a confidence level $\alpha = 0.05$ according to the *T* criterion.

A.5. Final decision

Since all the statistical tests agree that the data point with $KV = 272$ J is a significant outlier, additional analyses must be performed to identify the data point as either an experimental or a physical outlier. Should this be confirmed, the point must be removed from the dataset and the parameters for the lot are recalculated with $n = 74$. The comparison between the lot parameters before and after the removal of the outlier is shown in Table A.3. Note that removing the high outlier only marginally changes the mean absorbed energy (less than 1 J), but dramatically reduces standard deviation, range, coefficient of variation, and sample size.

Table A.3 - Statistical parameters of the SH-67 pilot lot before and after removing the outlier.

n	75	74
$\overline{KV}$ (J)	204.7	203.8
s	9.9	6.1
Range (J)	81.0	27.0
CV (%)	4.9	3.0
n_{SS}	8.678	3.010

LEGEND – Range = $KV_{max} - KV_{min}$; $CV = \frac{s}{\overline{KV}}$ (coefficient of variation).

The recalculated sample size (3.010) is less than 5.0, and therefore the SH-67 pilot lot can be considered acceptable.

Appendix B. Lot with two statistically significant outliers

B.1. HH-170 production lot certification results

The results of the Charpy tests for the certification of the high-energy HH-170 production lot are shown in Table B.4. Seventy-four tests were performed at 21 °C on the three NIST reference machines. The lowest (79.19 J) and the highest (120.00 J) absorbed energy values appear relatively distant from the remaining data points and appear to be potential outliers.

The sample size of the original dataset is n_{SS} = 10.445, and therefore unacceptable.

Table B.4 – Certification test results for the HH-170 production lot. The suspected outliers are highlighted in **bold red**.

Machine	KV (J)	Machine	KV (J)	Machine	KV (J)
SI	92.69	TO	107.2	TK	101.66
	105.5		97.932		100.62
	97.113		100.2		101.25
	90.401		101.9		104.79
	105.86		92.647		97.911
	101.92		97.837		102.81
	96.936		101.62		110.1
	102.55		105.02		99.162
	95.694		95.193		103.33
	100.32		97.082		98.015
	106.67		101.71		103.95
	96.492		**79.19**		104.06
	101.56		99.538		93.844
	96.669		99.916		101.66
	97.468		98.971		103.75
	93.925		102.75		102.6
	96.936		95.005		97.911
	98.535		93.212		93.948
	98.179		98.593		103.54
	98.446		98.782		100.83
	97.735		103.51		99.981
	90.753		105.31		95.304
	102.1		93.212		99.37
	110.91		98.593		93.323
			101.43		**120.00**

B.2. Normality checks

The Q-Q plot of the original dataset is provided in Figure B.4. Three of the normality tests indicate non-normality of the original dataset, with the following *p*-values: 0.0022 (S-W), 0.0022 (dA-P), and 2.31×10^{-12} (J-B). The exception is A-D, which returns $p = 0.051$ (barely above the normality threshold of 0.05).

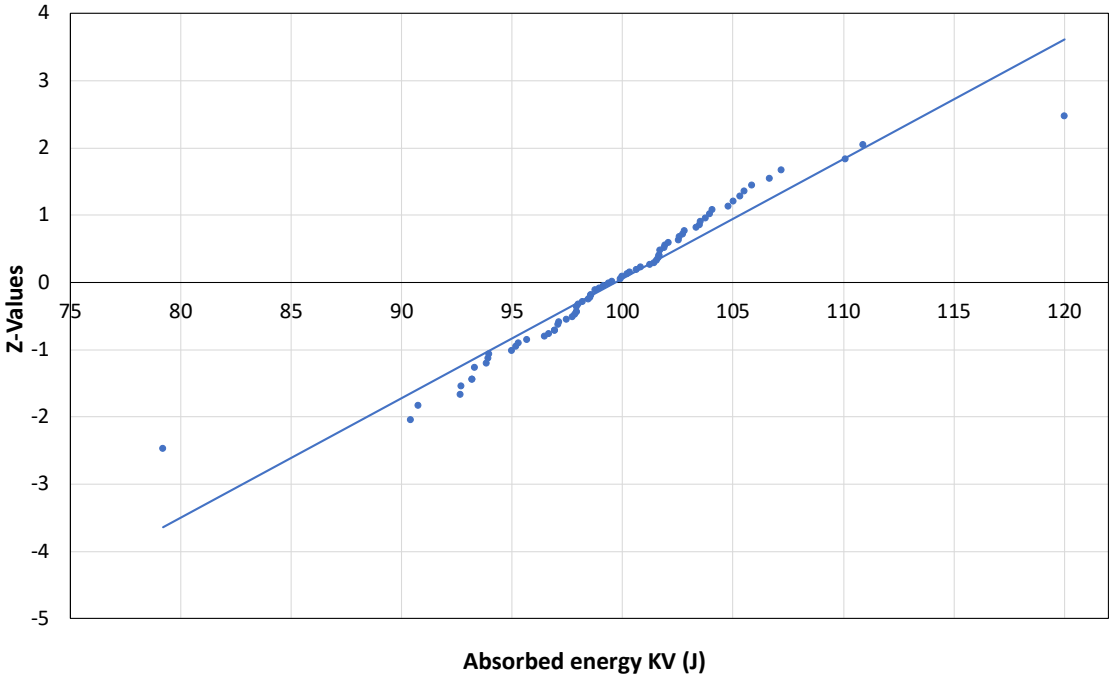

Figure B.4 – Q-Q plot of the original HH-170 dataset.

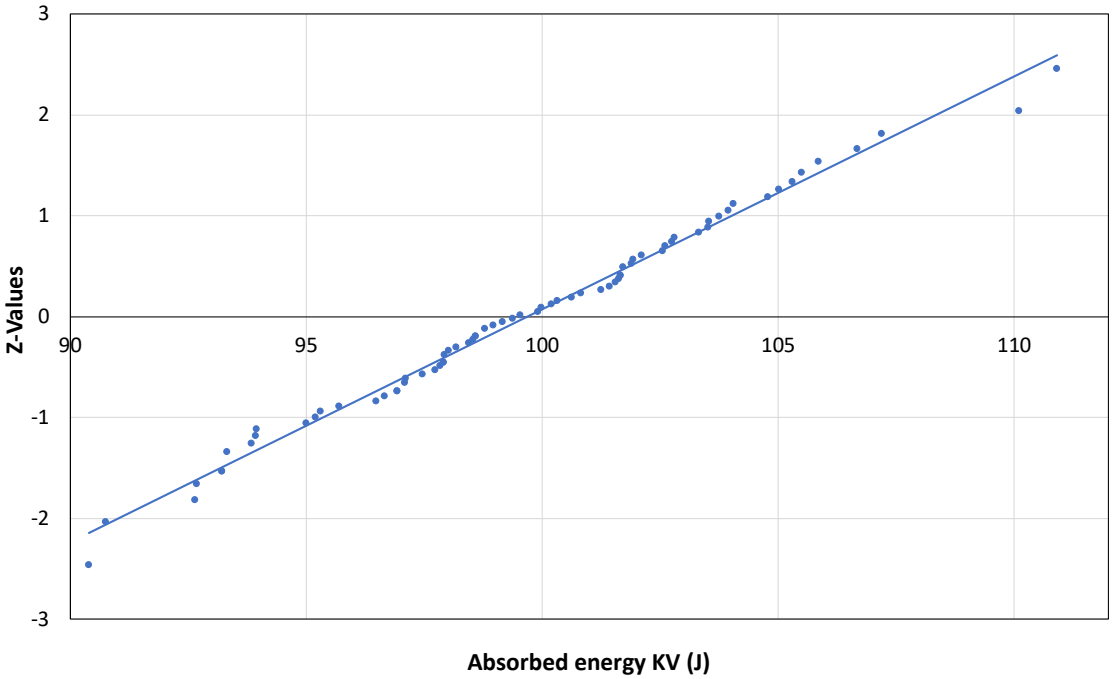

Figure B.5 – Q-Q plot for HH-170 after removing the two suspected outliers.

After removing the two extreme test results (Figure B.5), the dataset becomes normal according to all the normality tests ($p = 0.905$ for A-D, $p = 0.870$ for S-W, $p = 0.885$ for dA-P, and $p = 0.897$ for J-B). The analysis can therefore proceed to the next step.

B.3. Potential/suspected outlier screening

The box-and-whiskers plot for the original dataset (Figure B.6) shows that the lowest value (79.19 J) falls below the lower whisker, corresponding to $Q1 - 3 \cdot IQR$, and the highest value (120.00 J) falls above the upper whisker, corresponding to $Q3 + 3 \cdot IQR$. Both data points can therefore be considered potential/suspected outliers.

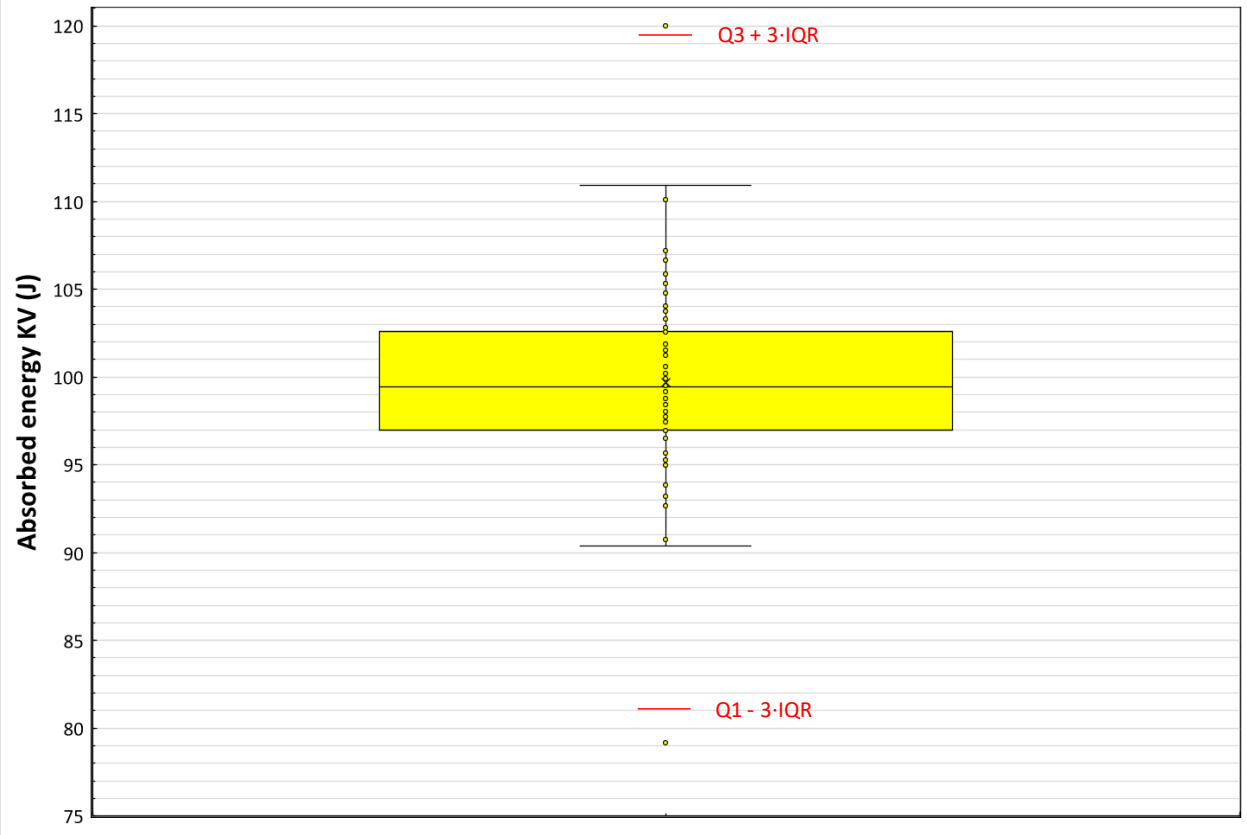

Figure B.6 – Box-and-whiskers plot for HH-170 showing two potential/suspected outliers.

Next, the modified z-score method (section 4.4) will be applied to confirm the potential/suspected outliers.

The absorbed energy values, sorted in ascending order, and the corresponding values of M_i calculated using eq. (8), are shown in Table B.5. For this dataset, $\widetilde{KV} = 99.45$ J and $MAD = 2.72$ J. The lowest absorbed energy value returns $M_i = 5.03$ and the highest returns $M_i = 5.10$. Both values are larger than 3.5, and therefore both test results are confirmed as potential/suspected outliers.

In the following analysis steps aimed at assessing the statistical significance of the two outliers, one of the available criteria applicable to multiple outliers will be applied.

Table B.5 – Modified *z*-scores for the HH-170 production lot. Potential/suspected outliers are shown in **bold red font**.

KV (J)	M_i	*KV* (J)	M_i	KV (J)	M_i
79.19	**5.03**	97.91	0.38	101.71	0.56
90.40	2.25	97.93	0.38	101.90	0.61
90.75	2.16	98.02	0.36	101.92	0.61
92.65	1.69	98.18	0.32	102.10	0.66
92.69	1.68	98.45	0.25	102.55	0.77
93.21	1.55	98.54	0.23	102.60	0.78
93.21	1.55	98.59	0.21	102.75	0.82
93.32	1.52	98.59	0.21	102.81	0.83
93.84	1.39	98.78	0.17	103.33	0.96
93.93	1.37	98.97	0.12	103.51	1.01
93.95	1.37	99.16	0.07	103.54	1.01
95.01	1.11	99.37	0.02	103.75	1.07
95.19	1.06	99.54	0.02	103.95	1.12
95.30	1.03	99.92	0.11	104.06	1.14
95.69	0.93	99.98	0.13	104.79	1.33
96.49	0.74	100.20	0.19	105.02	1.38
96.67	0.69	100.32	0.22	105.31	1.45
96.94	0.63	100.62	0.29	105.50	1.50
96.94	0.63	100.83	0.34	105.86	1.59
97.08	0.59	101.25	0.45	106.67	1.79
97.11	0.58	101.43	0.49	107.20	1.92
97.47	0.49	101.56	0.52	110.10	2.64
97.74	0.43	101.62	0.54	110.91	2.85
97.84	0.40	101.66	0.55	**120.00**	**5.10**
97.91	0.38	101.66	0.55		

B.4. Assessment of the statistical significance of the two outliers: Tietjen-Moore test

The preliminary analyses conducted by the use of the Q-Q plot and the box-and-whiskers plot indicated that the highest and the lowest data points are potential/suspected outliers. To confirm this, the Tietjen-Moore test is performed, as detailed below.

The value of m for the Tietjen-Moore test is 1, as the lowest and highest values in the HH-170 lot are potential/suspected outliers. The corresponding statistics L_m in accordance with eq. (18) and eq. (19) are 0.8054 for the largest point and 0.8026 for the smallest point, respectively. Both values are lower than the critical level corresponding to $n = 74$ and $\alpha = 0.05$ (0.8948), which is obtained by extrapolating to $n = 74$ the values in Table 8.

Both absorbed energy values are therefore **statistically significant** outliers according to the Tietjen-Moore test.

B.5. Final decision

Since the statistical tests described above agree that the lowest and the highest data points are statistically significant outliers, both tests and specimens need to be investigated, in order to determine whether they represent experimental or physical outliers. If this is the case for both tests, results are removed from the dataset and the parameters for the lot are recalculated with $n = 72$. The comparison between the lot parameters before and after removing the outliers is shown in Table B.6. As expected, removing the outliers significantly decreases standard deviation, range, coefficient of variation, and sample size. The certified absorbed energy value remains unchanged.

Table B.6 – Statistical parameters of the HH-170 production lot before and after removing the outliers.

n	74	72
$\overline{KV}$ (J)	99.67	99.67
s (J)	5.43	4.31
Range (J)	40.81	20.51
CV (%)	5.4	4.3
n_{SS}	10.347	6.770

Although the overall variability of the HH-170 lot is greatly improved by removing the outliers, the sample size (6.77) remains above the threshold (n_{SS} = 5.0), and the lot is therefore rejected.

This example shows that removing statistical outliers from a dataset doesn't always cause the lot to become acceptable.

Appendix C. Evaluation of customers' results containing a potential/suspected outlier

Suppose a set of five indirect verification results (absorbed energy values) from super-high-energy lot SH-67 is received from one of the NIST Charpy Program customers. One of the test results (test #2) is clearly higher than the remaining energy values, as seen in Table C.1.

We also assume that the Charpy results obtained by the customer at the low- and high-energy level are acceptable in accordance with ASTM E23.

Table C.1 - Hypothetical indirect verification test results from SH-67 lot, including a suspected high outlier (marked in **bold red font**).

Test #	KV (J)
1	205.04
2	**227.11**
3	198.97
4	200.09
5	210.01

Based on an engineering judgment by the NIST personnel, it is decided to apply the procedure outlined in section 7, before assessing the compliance of the customer's Charpy machine with ASTM E23 and/or ISO 148-2.

For the application of Hampel's Rule, we calculate the following:

- Median value of absorbed energy value for SH-67: $\widetilde{KV}$ = 203.4 J.
- Median absolute deviation for SH-67: MAD = 4.62 J.
- The difference between the customer's suspected outlier (227.11 J) and $\widetilde{KV}$ is 23.71 J, which is larger than $3 \cdot MAD$ = 13.86 J.

Hence, based on Hampel's Rule, the result of test #2 is a suspected/potential outlier, and the following steps are taken.

(a) A visual examination of the corresponding tested sample does not reveal any indication of experimental issues (jamming, wrong positioning, etc.) that could explain the high absorbed energy value.

(b) Nothing out of the ordinary can be observed on the fracture surface of the specimen.

(c) Upon direct questioning, the customer does not indicate anything anomalous about the test.

Ultimately, the indirect verification of the customer's machine at the super-high-energy level is based on all five reported absorbed energy values, and therefore the customer's machine fails the ASTM E23 indirect verification.