Designation: ~~D7915 – 18~~ D7915 – 22

An American National Standard

Standard Practice for Application of Generalized Extreme Studentized Deviate (GESD) Technique to Simultaneously Identify Multiple Outliers in a Data Set[1]

This standard is issued under the fixed designation D7915; the number immediately following the designation indicates the year of original adoption or, in the case of revision, the year of last revision. A number in parentheses indicates the year of last reapproval. A superscript epsilon (ε) indicates an editorial change since the last revision or reapproval.

1. Scope*

1.1 This practice provides a step by step procedure for the application of the Generalized Extreme Studentized Deviate (GESD) Many-Outlier Procedure to simultaneously identify multiple outliers in a data set. (See Bibliography.)

1.2 This practice is applicable to a data set comprising observations that is represented on a continuous numerical scale.

1.3 This practice is applicable to a data set comprising a minimum of six observations.

1.4 This practice is applicable to a data set where the normal (Gaussian) model is reasonably adequate for the distributional representation of the observations in the data set.

1.5 The probability of false identification of outliers associated with the decision criteria set by this practice is 0.01.

1.6 It is recommended that the execution of this practice be conducted under the guidance of personnel familiar with the statistical principles and assumptions associated with the GESD technique.

1.7 *This standard does not purport to address all of the safety concerns, if any, associated with its use. It is the responsibility of the user of this standard to establish appropriate safety, health, and environmental practices and determine the applicability of regulatory limitations prior to use.*

1.8 *This international standard was developed in accordance with internationally recognized principles on standardization established in the Decision on Principles for the Development of International Standards, Guides and Recommendations issued by the World Trade Organization Technical Barriers to Trade (TBT) Committee.*

2. Terminology

2.1 *Definitions of Terms Specific to This Standard:*

2.1.1 *outlier, n*—an observation (or a subset of observations) which appears to be inconsistent with the remainder of the data set.

[1] This practice is under the jurisdiction of ASTM Committee D02 on Petroleum Products, Liquid Fuels, and Lubricants and is the direct responsibility of Subcommittee D02.94 on Coordinating Subcommittee on Quality Assurance and Statistics.

Current edition approved ~~July 1, 2018~~May 1, 2022. Published ~~August 2018~~May 2022. Originally approved in 1988. Last previous edition approved in ~~2014~~2018 as ~~D7915 – 14.~~D7915 – 18. DOI: ~~10.1520/D7915-18.~~10.1520/D7915-22.

***A Summary of Changes section appears at the end of this standard**

3. Significance and Use

3.1 The GESD procedure can be used to simultaneously identify up to a pre-determined number of outliers (r) in a data set, without having to pre-examine the data set and make *a priori* decisions as to the location and number of potential outliers.

3.2 The GESD procedure is robust to masking. Masking describes the phenomenon where the existence of multiple outliers can prevent an outlier identification procedure from declaring any of the observations in a data set to be outliers.

3.3 The GESD procedure is automation-friendly, and hence can easily be programmed as automated computer algorithms.

4. Procedure

4.1 Specify the maximum number of outliers (r) in a data set to be identified. This is the number of cycles required to be executed (see 4.2) for the identification of up to r outliers.

4.1.1 The recommended maximum number of outliers (r) by this practice is two (2) for data sets with six to twelve observations.

4.1.2 For data sets with more than twelve observations, the recommended maximum number of outliers (r) is the lesser of ten (10) or 20 %.

4.1.3 The recommended values for r in 4.1.1 and 4.1.2 are not intended to be mandatory. Users can specify other values based on their specific needs.

4.2 Set the current cycle number c to 1 ($c = 1$).

4.2.1 Assign the original data set to be assessed (in 4.1) as the data set for the current cycle 1 and label it as DTS_1.

4.3 Compute test statistic T for each observation in the data set assigned to the current cycle (DTS_c) as follows:

$$T = |x - \bar{x}|/s \tag{1}$$

where:

x = an observation in the data set,
$\bar{x}$ = average calculated using all observations in the data set, and
s = sample standard deviation calculated using all observations in the data set.

4.4 Identify the observation associated with the largest absolute magnitude of the test statistic T in the data set of the current cycle.

4.5 If current cycle c is less than r, execute 4.5.1 to 4.5.4; otherwise go to 4.6.

4.5.1 Remove the observation identified in 4.4 from the data set of the current cycle.

4.5.2 Increment the current cycle number by 1:
$c = c_{current} + 1$.

4.5.3 Assign the reduced data set in 4.5.1 (that is, data set with the observation identified in 4.4 removed) as the data set for the new cycle number and label it as DTS_c.

4.5.4 Repeat steps 4.3 to 4.5.

4.6 Beginning with $c = r$, compare the maximum T computed in the dataset DTS_c, to a critical value $\lambda_{critical}$ associated with the data set for cycle c, where $\lambda_{critical}$ is chosen based on a false identification probability of 0.01. See Table A1.1 in Annex A1 for $\lambda_{critical}$ values applicable to different data set sizes and cycle numbers (c).

4.7 If maximum T for data set DTS_c ~~does not exceed~~ exceeds $\lambda_{critical}$ for cycle ~~*c, reduce c* by 1 (that is, c= *c*~~*r*~~ − 1) and repeat the~~

~~comparison of maximum~~ the ~~T to~~ $\lambda_{critical}$ ~~until the first occurrence of maximum T exceeding~~ $\lambda_{critical}$ ~~is encountered. The~~ observation associated with maximum T ~~for this cycle (DTS~~in the dataset DTS_c~~)~~ and all observations associated with maximum T from DTS_c ~~–~~ ~~1~~ to DTS_1 are declared as outliers.

4.7.1 If maximum T for data set DTS_c does not exceed $\lambda_{critical}$ for cycle c, reduce c by 1 (that is, $c = c - 1$) and repeat the comparison of maximum T to $\lambda_{critical}$ until the first occurrence of maximum T exceeding $\lambda_{critical}$ is encountered. The observation associated with maximum T for this cycle (DTS_c) and all observations associated with maximum T from DTS_{c-1} to DTS_1 are declared as outliers.

4.8 The outlier identification procedure is declared complete at the first occurrence of maximum T exceeding $\lambda_{critical}$ for cycle c in 4.7, or completion of comparison for $c = 1$.

5. Worked Example

5.1 Listed below is a data set comprising 30 observations:

35.0	36.6	34.7	36.2	37.0	25.3	37.2	41.3	26.0	24.6
33.5	35.5	35.4	39.9	39.2	36.6	37.2	33.2	34.0	35.7
39.2	42.1	35.7	40.2	36.6	41.1	41.1	39.1	40.6	41.3

5.1.1 The total number of observations (N) = 30.

5.1.2 From 4.1.2, the maximum number of outliers r to be identified is six (20 % of 30), since six is less than ten.

5.2 Refer to Table 1 for the following discussions:

5.2.1 Data set labeled DTS_1 is the original data set, which is assigned to cycle (c) = 1.

5.2.2 The observation 24.6, corresponding to the maximum T value 2.60 in DTS_1, is identified and removed to form a reduced data set DTS_2.

5.2.3 The above is repeated up to DTS_6, where the observation 33.5 is identified as the having the maximum T value 1.65 but not removed since this is the last cycle for identifying up to $r = 6$ outliers.

5.2.4 In accordance with 4.7, working backwards from $c = 6$, the cycle number for which the first occurrence of maximum T value of the data set DTS_c exceeds $\lambda_{critical}$ is cycle number three (see data set column labeled DTS_3).

5.2.5 In accordance with 4.7, observations 24.6 from DTS_1, 25.3 from DTS_2, and 26.0 from DTS_3 are all declared as outliers by this practice.

6. Keywords

6.1 GESD; outliers

TABLE 1 Example Execution of the GESD Procedure for Worked Example in 5.1

NOTE 1—Explanation of Table 1:

The cell marked by a border for each DTS_i column is the observation with the most extreme T values (T_{max}) in the data set i. For the convenience of readers T_{max} is re-shown in the third last row from the bottom of this table. For instance, in DTS_1, the value 24.6 has a corresponding T_1 value of 2.60, which is the largest T value (T_{max}) for DTS_1. Marking of these cells with the border is only to help the readers. It does not mean these cells are outliers. What it means is that the marked cell is to be removed for the next required cycle. For example, in next required cycle DTS_2, the value 24.6 identified as the most extreme from the previous cycle DTS_1 is removed, and the removed cell is shown as a blank entry in DTS_2.

The decision on which of these highlighted cells are outliers is made only after completion of the required cycles (in this case, up to DTS_2, since $r = 6$).

To make the outlier decision, start from DTS_6. Compare the T_{max} value to the critical value ($\lambda_{critical}$), both are listed at the bottom of this table for readers' convenience. If T_{max} does not exceed the critical value below it, move to the previous DTS ~~(DTS~~(DTS_5), and if it does not exceed the critical value below it, move to DTS_4 ... and so forth. Stop at the first DTS_i where the T_{max} exceeds the critical value, which is DTS_3 in the example, where T_{max} is 3.27, versus the critical value of 3.20. The outliers are then declared as the value associated with T_{max} at DTS_3 (which is 26.0), and all the extreme values identified in all the DTS's before DTS_3, which are: 25.3 in DTS_2, and 24.6 in DTS_1. The total number of outliers identified in this example is 3.

data set=>	DTS_1	T_1	DTS_2	T_2	DTS_3	T_3	DTS_4	T_4	DTS_5	T_5	DTS_6	T_6
	35.0	0.30	35.0	0.44	35.0	0.64	35.0	0.97	35.0	0.94	35.0	1.05
	36.6	0.05	36.6	0.04	36.6	0.17	36.6	0.37	36.6	0.32	36.6	0.40
	34.7	0.37	34.7	0.52	34.7	0.73	34.7	1.08	34.7	1.06	34.7	1.17
	36.2	0.04	36.2	0.14	36.2	0.29	36.2	0.52	36.2	0.48	36.2	0.56
	37.0	0.14	37.0	0.06	37.0	0.05	37.0	0.22	37.0	0.17	37.0	0.24
	25.3	2.44	25.3	2.85								
	37.2	0.18	37.2	0.11	37.2	0.00	37.2	0.15	37.2	0.09	37.2	0.16
	41.3	1.09	41.3	1.12	41.3	1.20	41.3	1.38	41.3	1.50	41.3	1.49
	26.0	2.29	26.0	2.68	26.0	3.27						
	24.6	2.60										
	33.5	0.63	33.5	0.81	33.5	1.08	33.5	1.53	33.5	1.52	33.5	1.65
	35.5	0.19	35.5	0.32	35.5	0.49	35.5	0.78	35.5	0.75	35.5	0.85
	35.4	0.21	35.4	0.34	35.4	0.52	35.4	0.82	35.4	0.79	35.4	0.89
	39.9	0.78	39.9	0.78	39.9	0.79	39.9	0.86	39.9	0.96	39.9	0.93
	39.2	0.62	39.2	0.60	39.2	0.59	39.2	0.60	39.2	0.69	39.2	0.65
	36.6	0.05	36.6	0.04	36.6	0.17	36.6	0.37	36.6	0.32	36.6	0.40
	37.2	0.18	37.2	0.11	37.2	0.00	37.2	0.15	37.2	0.09	37.2	0.16
	33.2	0.70	33.2	0.89	33.2	1.16	33.2	1.64	33.2	1.64		
	34.0	0.52	34.0	0.69	34.0	0.93	34.0	1.34	34.0	1.33	34.0	1.45
	35.7	0.15	35.7	0.27	35.7	0.43	35.7	0.71	35.7	0.67	35.7	0.77
	39.2	0.62	39.2	0.60	39.2	0.59	39.2	0.60	39.2	0.69	39.2	0.65
	42.1	1.26	42.1	1.32	42.1	1.43	42.1	1.68				
	35.7	0.15	35.7	0.27	35.7	0.43	35.7	0.71	35.7	0.67	35.7	0.77
	40.2	0.84	40.2	0.85	40.2	0.88	40.2	0.97	40.2	1.08	40.2	1.05
	36.6	0.05	36.6	0.04	36.6	0.17	36.6	0.37	36.6	0.32	36.6	0.40
	41.1	1.04	41.1	1.07	41.1	1.14	41.1	1.31	41.1	1.43	41.1	1.41
	41.1	1.04	41.1	1.07	41.1	1.14	41.1	1.31	41.1	1.43	41.1	1.41
	39.1	0.60	39.1	0.58	39.1	0.56	39.1	0.56	39.1	0.65	39.1	0.61
	40.6	0.93	40.6	0.95	40.6	1.00	40.6	1.12	40.6	1.23	40.6	1.21
	41.3	1.09	41.3	1.12	41.3	1.20	41.3	1.38	41.3	1.50	41.3	1.49
average	36.37		36.78		37.19		37.60		37.43		37.60	
std dev	4.54		4.02		3.42		2.68		2.58		2.48	
T_{max}		2.60		2.85		3.27		1.68		1.64		1.65
$\lambda_{critical}$		3.24		3.22		3.20		3.18		3.16		3.14
	$c = 1$		$c = 2$		$c = 3$		$c = 4$		$c = 5$		$c = 6$	

ANNEX

(Mandatory Information)

A1. $\lambda_{critical}$ FOR VARIOUS DATA SET SIZES