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

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 and health practices and determine the applicability of regulatory limitations prior to use.*

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.

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.

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 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 Compute test statistic *T* for each observation in the initial starting data set (DTS_0) as follows:

$$T = |x - \bar{x}| / s \qquad (1)$$

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.

4.3 Remove the observation in the data set with the largest absolute magnitude of the test statistic *T* and form a reduced data set (DTS_i), where i = number of observations removed from the initial data set.

4.4 Re-calculate *T* for all observations in the reduced data set from 4.3.

4.5 Repeat steps 4.3 to 4.4 until *r* number of observations have been removed from the initial data set. That is, until

calculation of all *T*'s for all observations in the reduced data set DTS$_r$ has been completed.

4.6 Compare the maximum *T* computed in each data set (DTS$_0$ to DTS$_r$) to a critical value $\lambda_{critical}$ associated the data set DTS$_i$, where λ is chosen based on a false identification probability of 0.01. See Table A1.1 in Annex A1 for λ values applicable to different data set sizes.

4.7 Identify the data set DTS$_m$ for which the maximum *T* exceeds $\lambda_{critical}$, and *m* (number of observations removed from the initial data set DTS$_0$) is the largest value ($0 < m \leq r$).

4.8 All observations removed from data sets DTS$_0$ to form DTS$_m$, along with the observation associated with the maximum *T* in data set DTS$_m$, are declared as outliers.

NOTE 1—In the rare occasion where *m=r*, the total number of outliers identified will actually be (*r* + 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 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$_0$ is the initial data set.

5.2.2 The observation 24.6, corresponding to the maximum *T* value in DTS$_0$, is removed to from a reduced data set DTS$_1$.

5.2.3 The above is repeated up to DTS$_6$.

5.2.4 From 4.7, the largest *m* value for which the maximum *T* value of the data set DTS$_m$ exceeds $\lambda_{critical}$ is 2 (see data set column labeled DTS$_2$).

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 have the most extreme T values (T$_{max}$) in the data set i. For the convenience of readers T$_{max}$ is shown in the third last row from the bottom of this table. For instance, in DTS$_5$, the value 33.5 has a corresponding T$_5$ value of 1.65, which is the largest T value (T$_{max}$) for DTS$_5$. 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 iteration. For example, in DTS$_5$, the value 33.2 identified as the most extreme from DTS$_4$ is removed, and the removed cell is shown as a blank entry in DTS$_5$.

The decision on which of these highlighted cells are outliers is made only after completion of the required iterations (in this case, up to DTS$_6$).

To make the outlier decision, start from DTS$_6$. Compare the T$_{max}$ value to the critical value (lambda), both are listed at the bottom of this table for readers' convenience. If T$_{max}$ is smaller than critical value below it, move to the previous DTS (DTS$_5$), and if it's smaller, move to DTS$_4$ ... and so forth. Stop at the first DTS$_i$ where the T$_{max}$ exceeds the critical value, which is DTS$_2$ 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$_2$ (which is 26.0), and all the extreme values identified in all the DTS's before DTS$_2$, which are: 25.3 in DTS$_1$, and 24.6 in DTS$_0$. The total number of outliers identified in this example is 3.

data set=>	DTS$_0$	T$_0$	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	35.0	1.16
	36.6	0.05	36.6	0.04	36.6	0.17	36.6	0.37	36.6	0.32	36.6	0.40	36.6	0.49
	34.7	0.37	34.7	0.52	34.7	0.73	34.7	1.08	34.7	1.06	34.7	1.17	34.7	1.29
	36.2	0.04	36.2	0.14	36.2	0.29	36.2	0.52	36.2	0.48	36.2	0.56	36.2	0.66
	37.0	0.14	37.0	0.06	37.0	0.05	37.0	0.22	37.0	0.17	37.0	0.24	37.0	0.32
	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	37.2	0.24
	41.3	1.09	41.3	1.12	41.3	1.20	41.3	1.38	41.3	1.50	41.3	1.49	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	35.5	0.95
	35.4	0.21	35.4	0.34	35.4	0.52	35.4	0.82	35.4	0.79	35.4	0.89	35.4	1.00
	39.9	0.78	39.9	0.78	39.9	0.79	39.9	0.86	39.9	0.96	39.9	0.93	39.9	0.90
	39.2	0.62	39.2	0.60	39.2	0.59	39.2	0.60	39.2	0.69	39.2	0.65	39.2	0.60
	36.6	0.05	36.6	0.04	36.6	0.17	36.6	0.37	36.6	0.32	36.6	0.40	36.6	0.49
	37.2	0.18	37.2	0.11	37.2	0.00	37.2	0.15	37.2	0.09	37.2	0.16	37.2	0.24
	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	34.0	1.59
	35.7	0.15	35.7	0.27	35.7	0.43	35.7	0.71	35.7	0.67	35.7	0.77	35.7	0.87
	39.2	0.62	39.2	0.60	39.2	0.59	39.2	0.60	39.2	0.69	39.2	0.65	39.2	0.60
	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	35.7	0.87
	40.2	0.84	40.2	0.85	40.2	0.88	40.2	0.97	40.2	1.08	40.2	1.05	40.2	1.02
	36.6	0.05	36.6	0.04	36.6	0.17	36.6	0.37	36.6	0.32	36.6	0.40	36.6	0.49
	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	1.40
	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	1.40
	39.1	0.60	39.1	0.58	39.1	0.56	39.1	0.56	39.1	0.65	39.1	0.61	39.1	0.56
	40.6	0.93	40.6	0.95	40.6	1.00	40.6	1.12	40.6	1.23	40.6	1.21	40.6	1.19
	41.3	1.09	41.3	1.12	41.3	1.20	41.3	1.38	41.3	1.50	41.3	1.49	41.3	1.49
average	36.37		36.78		37.19		37.60		37.43		37.60		37.77	
std dev	4.54		4.02		3.42		2.68		2.58		2.48		2.38	
T$_{max}$		2.60		2.85		3.27		1.68		1.64		1.65		1.59
$\lambda_{critical}$		3.24		3.22		3.20		3.18		3.16		3.14		3.11
	m=0		m=1		m=2		m=3		m=4		m=5		m=6	

5.2.5 From 4.8, observations 24.6 from DTS_0, 25.3 from DTS_1, and 26.0 from DTS_2 are declared outliers.

6. Keywords

6.1 GESD; outliers

ANNEX

(Mandatory Information)

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

TABLE A1.1 $\lambda_{critical}$ for Various Data Set Sizes (0.01 significant)

r	N	m=0 $\lambda_{critical}$	m=1 $\lambda_{critical}$	m=2 $\lambda_{critical}$	m=3 $\lambda_{critical}$	m=4 $\lambda_{critical}$	m=5 $\lambda_{critical}$	m=6 $\lambda_{critical}$	m=7 $\lambda_{critical}$	m=8 $\lambda_{critical}$	m=9 $\lambda_{critical}$	m=10 $\lambda_{critical}$
2	6	1.97	1.76	1.50								
2	7	2.14	1.97	1.76								
2	8	2.27	2.14	1.97								
2	9	2.39	2.27	2.14								
2	10	2.48	2.39	2.27								
2	11	2.56	2.48	2.39								
2	12	2.64	2.56	2.48								
3	13	2.70	2.64	2.56	2.48							
3	14	2.76	2.70	2.64	2.56							
3	15	2.81	2.76	2.70	2.64							
3	16	2.85	2.81	2.76	2.70							
3	17	2.89	2.85	2.81	2.76							
4	18	2.93	2.89	2.85	2.81	2.76						
4	19	2.97	2.93	2.89	2.85	2.81						
4	20	3.00	2.97	2.93	2.89	2.85						
4	21	3.03	3.00	2.97	2.93	2.89						
4	22	3.06	3.03	3.00	2.97	2.93						
5	23	3.09	3.06	3.03	3.00	2.97	2.93					
5	24	3.11	3.09	3.06	3.03	3.00	2.97					
5	25	3.14	3.11	3.09	3.06	3.03	3.00					
5	26	3.16	3.14	3.11	3.09	3.06	3.03					
6	27	3.18	3.16	3.14	3.11	3.09	3.06					
6	28	3.20	3.18	3.16	3.14	3.11	3.09	3.06				
6	29	3.22	3.20	3.18	3.16	3.14	3.11	3.09				
6	30	3.24	3.22	3.20	3.18	3.16	3.14	3.11				
6	31	3.25	3.24	3.22	3.20	3.18	3.16	3.14				
6	32	3.27	3.25	3.24	3.22	3.20	3.18	3.16				
7	33	3.29	3.27	3.25	3.24	3.22	3.20	3.18	3.16			
7	34	3.30	3.29	3.27	3.25	3.24	3.22	3.20	3.18			
7	35	3.32	3.30	3.29	3.27	3.25	3.24	3.22	3.20			
7	36	3.33	3.32	3.30	3.29	3.27	3.25	3.24	3.22			
7	37	3.34	3.33	3.32	3.30	3.29	3.27	3.25	3.24			
8	38	3.36	3.34	3.33	3.32	3.30	3.29	3.27	3.25	3.24		
8	39	3.37	3.36	3.34	3.33	3.32	3.30	3.29	3.27	3.25		
8	40	3.38	3.37	3.36	3.34	3.33	3.32	3.30	3.29	3.27		
8	41	3.39	3.38	3.37	3.36	3.34	3.33	3.32	3.30	3.29		
8	42	3.40	3.39	3.38	3.37	3.36	3.34	3.33	3.32	3.30		
9	43	3.41	3.40	3.39	3.38	3.37	3.36	3.34	3.33	3.32	3.30	
9	44	3.43	3.41	3.40	3.39	3.38	3.37	3.36	3.34	3.33	3.32	
9	45	3.44	3.43	3.41	3.40	3.39	3.38	3.37	3.36	3.34	3.33	
9	46	3.45	3.44	3.43	3.41	3.40	3.39	3.38	3.37	3.36	3.34	
9	47	3.46	3.45	3.44	3.43	3.41	3.40	3.39	3.38	3.37	3.36	
10	48	3.46	3.46	3.45	3.44	3.43	3.41	3.40	3.39	3.38	3.37	3.36
10	49	3.47	3.46	3.46	3.45	3.44	3.43	3.41	3.40	3.39	3.38	3.37
10	50	3.48	3.47	3.46	3.46	3.45	3.44	3.43	3.41	3.40	3.39	3.38
10	51	3.49	3.48	3.47	3.46	3.46	3.45	3.44	3.43	3.41	3.40	3.39
10	52	3.50	3.49	3.48	3.47	3.46	3.46	3.45	3.44	3.43	3.41	3.40
10	53	3.51	3.50	3.49	3.48	3.47	3.46	3.46	3.45	3.44	3.43	3.41
10	54	3.52	3.51	3.50	3.49	3.48	3.47	3.46	3.46	3.45	3.44	3.43
10	55	3.52	3.52	3.51	3.50	3.49	3.48	3.47	3.46	3.46	3.45	3.44
10	56	3.53	3.52	3.52	3.51	3.50	3.49	3.48	3.47	3.46	3.46	3.45
10	57	3.54	3.53	3.52	3.52	3.51	3.50	3.49	3.48	3.47	3.46	3.46
10	58	3.55	3.54	3.53	3.52	3.52	3.51	3.50	3.49	3.48	3.47	3.46
10	59	3.55	3.55	3.54	3.53	3.52	3.52	3.51	3.50	3.49	3.48	3.47
10	60	3.56	3.55	3.55	3.54	3.53	3.52	3.52	3.51	3.50	3.49	3.48
10	61	3.57	3.56	3.55	3.55	3.54	3.53	3.52	3.52	3.51	3.50	3.49
10	62	3.57	3.57	3.56	3.55	3.55	3.54	3.53	3.52	3.52	3.51	3.50
10	63	3.58	3.57	3.57	3.56	3.55	3.55	3.54	3.53	3.52	3.52	3.51
10	64	3.59	3.58	3.57	3.57	3.56	3.55	3.55	3.54	3.53	3.52	3.52
10	65	3.59	3.59	3.58	3.57	3.57	3.56	3.55	3.55	3.54	3.53	3.52
10	66	3.60	3.59	3.59	3.58	3.57	3.57	3.56	3.55	3.55	3.54	3.53
10	67	3.60	3.60	3.59	3.59	3.58	3.57	3.57	3.56	3.55	3.55	3.54
10	68	3.61	3.60	3.60	3.59	3.59	3.58	3.57	3.57	3.56	3.55	3.55
10	69	3.62	3.61	3.60	3.60	3.59	3.59	3.58	3.57	3.57	3.56	3.55
10	70	3.62	3.62	3.61	3.60	3.60	3.59	3.59	3.58	3.57	3.57	3.56
10	71	3.63	3.62	3.62	3.61	3.60	3.60	3.59	3.59	3.58	3.57	3.57
10	72	3.63	3.63	3.62	3.62	3.61	3.60	3.60	3.59	3.59	3.58	3.57
10	73	3.64	3.63	3.63	3.62	3.62	3.61	3.60	3.60	3.59	3.59	3.58
10	74	3.64	3.64	3.63	3.63	3.62	3.62	3.61	3.60	3.60	3.59	3.59
10	75	3.65	3.64	3.64	3.63	3.63	3.62	3.62	3.61	3.60	3.60	3.59
10	76	3.65	3.65	3.64	3.64	3.63	3.63	3.62	3.62	3.61	3.60	3.60
10	77	3.66	3.65	3.65	3.64	3.64	3.63	3.63	3.62	3.62	3.61	3.60
10	78	3.66	3.66	3.65	3.65	3.64	3.64	3.63	3.63	3.62	3.62	3.61
10	79	3.67	3.66	3.66	3.65	3.65	3.64	3.64	3.63	3.63	3.62	3.62

TABLE A1.1 *Continued*

r	N	m=0 $\lambda_{critical}$	m=1 $\lambda_{critical}$	m=2 $\lambda_{critical}$	m=3 $\lambda_{critical}$	m=4 $\lambda_{critical}$	m=5 $\lambda_{critical}$	m=6 $\lambda_{critical}$	m=7 $\lambda_{critical}$	m=8 $\lambda_{critical}$	m=9 $\lambda_{critical}$	m=10 $\lambda_{critical}$
10	80	3.67	3.67	3.66	3.66	3.65	3.65	3.64	3.64	3.63	3.63	3.62
10	81	3.68	3.67	3.67	3.66	3.66	3.65	3.65	3.64	3.64	3.63	3.63
10	82	3.68	3.68	3.67	3.67	3.66	3.66	3.65	3.65	3.64	3.64	3.63
10	83	3.69	3.68	3.68	3.67	3.67	3.66	3.66	3.65	3.65	3.64	3.64
10	84	3.69	3.69	3.68	3.68	3.67	3.67	3.66	3.66	3.65	3.65	3.64
10	85	3.70	3.69	3.69	3.68	3.68	3.67	3.67	3.66	3.66	3.65	3.65
10	86	3.70	3.70	3.69	3.69	3.68	3.68	3.67	3.67	3.66	3.66	3.65
10	87	3.70	3.70	3.70	3.69	3.69	3.68	3.68	3.67	3.67	3.66	3.66
10	88	3.71	3.70	3.70	3.70	3.69	3.69	3.68	3.68	3.67	3.67	3.66
10	89	3.71	3.71	3.70	3.70	3.70	3.69	3.69	3.68	3.68	3.67	3.67
10	90	3.72	3.71	3.71	3.70	3.70	3.70	3.69	3.69	3.68	3.68	3.67
10	91	3.72	3.72	3.71	3.71	3.70	3.70	3.70	3.69	3.69	3.68	3.68
10	92	3.72	3.72	3.72	3.71	3.71	3.70	3.70	3.70	3.69	3.69	3.68
10	93	3.73	3.72	3.72	3.72	3.71	3.71	3.70	3.70	3.70	3.69	3.69
10	94	3.73	3.73	3.72	3.72	3.72	3.71	3.71	3.70	3.70	3.70	3.69
10	95	3.74	3.73	3.73	3.72	3.72	3.72	3.71	3.71	3.70	3.70	3.70
10	96	3.74	3.74	3.73	3.73	3.72	3.72	3.72	3.71	3.71	3.70	3.70
10	97	3.74	3.74	3.74	3.73	3.73	3.72	3.72	3.72	3.71	3.71	3.70
10	98	3.75	3.74	3.74	3.74	3.73	3.73	3.72	3.72	3.72	3.71	3.71
10	99	3.75	3.75	3.74	3.74	3.74	3.73	3.73	3.72	3.72	3.72	3.71
10	100	3.75	3.75	3.75	3.74	3.74	3.74	3.73	3.73	3.72	3.72	3.72

r = maximum number of outliers to be identified
m = number of observations removed from initial data set DTS_0

BIBLIOGRAPHY

(1) The ASQC Basic References in Quality Control: Statistical Techniques, Volume 16: "How to Detect and Handle Outliers," Boris Iglewicz and David C. Hoaglin

(2) Rosner, Bernard, "Percentage Points for a Generalized ESD Many-Outlier Procedure," Technometrics 25: 165-172

RELATED MATERIAL

ASTM Research Report D2-1481—Tutorial for Generalized Extreme Studentized Deviate Many Outlier Procedure