GESD – A Robust and Effective Technique for Dealing with Multiple Outliers

By Alex T.C. Lau

Q I suspect that my data contain more than one outlier. Is there a preferred technique to use to isolate them?

A Two techniques have been discussed in Data Points for testing if a single observation "with a value that does not appear to belong with the rest of the values in a data set" can be declared as an outlier. In the Data Points column, "Dealing with Outliers" (*SN*, Nov./Dec. 2008), the problem associated with one outlier masking another outlier in a single outlier test was mentioned, and a reference to the generalized extreme studentized deviate (GESD) was provided as a robust and comprehensive technique to effectively identify multiple outliers. This column provides a simple example of outlier masking and how to apply GESD to identify multiple outliers.

ILLUSTRATION OF MASKING

Reproducing the data from the earlier article: [5.3, 3.1, 4.9, 3.9, 7.8, 4.7, 4.3], the value 7.8 was visually identified as being discordant with the rest of the data, and its outlier status was confirmed using the Grubbs technique by comparing its T statistic against a critical value as follows:

$\bar{X}$ = average = 4.86; s = standard deviation = 1.48; $T_{7.8}$ = | 7.8 - 4.86 | / 1.48 = 1.99

For n = 7 and a 5 percent false declaration risk, the critical value is 1.938. Since T = 1.99 exceeds this critical value, the value 7.8 is confirmed to be an outlier.

Suppose we now have the following data set: [5.3, 3.1, 4.9, 3.9, 7.8, 4.7, 4.3, 8.0, 4.5, 5.1, 3.5] (see dot plot in Figure 1). We wish to test if the value 8.0 is an outlier.

Figure 1 - Dot Plot of Data

Following the example from the earlier article using the Grubbs technique, we have:

$\bar{X}$ = average = 5.01; s = standard deviation = 1.58; $T_{8.0}$ = | 8.0 - 5.01 | / 1.58 = 1.89.

For n = 11 and a 5 percent risk, the critical value is 2.234. Since $T_{8.0}$ = 1.89 is less than this critical value, we cannot declare the value 8.0 as an outlier. So, *what happened*?

The problem illustrated above is a phenomenon know as *masking*. In this dataset, the two visually obvious discordant values inflated the standard deviation, hence making the T statistic small relative to the critical value.

USE OF GESD TO IDENTIFY MULTIPLE OUTLIERS

Common outlier detection techniques such as Dixon and Grubbs require a *priori* examination of the dataset to determine the number of potential outliers and where they reside (large or small) among the dataset. If multiple outliers exist, depending on the size of these outliers, the test statistics could be erroneously small and hence result in an insignificant test for rejection.

Rosner (1983)[1] proposed a technique that he named "Generalized Extreme Studentized Deviation (GESD) Many Outlier Procedure" for the effective identification of multiple outliers that does not require a priori examination of the dataset to decide on how many and where to test for outliers. Detailed comparison of this procedure against four other popular techniques were critically examined in a 1993 publication by the American Society of Quality Control, with GESD being the recommended technique among those examined.[2] And, in 2014, ASTM D7915, Practice for Application of Generalized Extreme Studentized Deviate (GESD) Technique to Simultaneously Identify Multiple Outliers in a Data Set, was completed by D02.94 Subcommittee on Quality Assurance and Statistics, part of Committee D02 on Petroleum Products, Liquid Fuels and Lubricants. A simple description of this procedure is described below.

GESD Procedure:

1. Decide a *priori* the maximum number of outliers to test for. Let's call this number ***r***. (A general recommendation is to set $r = 20$ percent of n.)
2. Set current cycle index $i = 1$.
3. Calculate the quantity $T = |\text{observation} - \text{average}| \div s$ for every member of the dataset in the current cycle.
4. Identify the observation with the largest T. Designate this as $T_{1\,max}$ (i.e., maximum T for the first cycle.)
5. Remove the observation identified in 4) from the dataset.
6. Increase current cycle index i by 1 : i.e. $i = i + 1$
7. Repeat steps 3 to 6 using remaining data up to and including $i = r$.
8. On completion of step 7, beginning with $T_{r\,max}$, maximum T value in cycle ***r***, and working backwards ($T_{r\text{-}1\,max}$, $T_{r\text{-}2\,max}$ .... and so on), compare this maximum value versus the critical value for the specific cycle (λ_i) obtained the ASQC publication.
9. Identify the highest cycle for which $T_{i\,max}$ exceeds its limit value. The observation associated with the $T_{i\,max}$ for that cycle and all observations associated with the $T_{i\,max}$'s for all previous cycles up to and including cycle 1 are considered to be outliers.

To illustrate the GESD procedure using the example above, we have $n = 11$ data points. So normally we would set $r = 2$ (20 percent of n). However, for the purpose of methodology illustration, as well as showing that this technique is robust to over-specification of ***r*** (i.e., more than necessary), we will set $r = 3$ for this exercise.

The GESD calculations are listed in Table 1 below. Since T_{2max} at cycle 2 is the highest cycle that exceeds its corresponding critical value (λ_2), observation 7.8 for Cycle 2, and observation 8.0 for Cycle 1 are both identified as outliers.

Table 1 — GESD Calculations

Average for Cycle =	5.01		4.71		4.37
Standard Deviation for Cycle =	1.58		1.29		0.74
Cyle 1	*T*	Cycle 2	*T*	Cycle 3	*T*
5.3	0.18	5.3	0.46	5.3	1.26
3.1	1.21	3.1	1.25	3.1	1.71 =$T_{3\,max}$
4.9	0.07	4.9	0.15	4.9	0.72
3.9	0.70	3.9	0.63	3.9	0.63
7.8	1.77	7.8	0.239* =$T_{2\,max}$		
4.7	0.20	4.7	0.01	4.7	0.45
4.3	0.45	4.3	0.32	4.3	0.09
8.0	1.90 =$T_{1\,max}$				
4.5	.32	4.5	0.16	4.5	0.18
5.1	.06	5.1	0.30	5.1	0.99
3.5	.96	3.5	0.94	3.5	1.17
Critical Value λ at 5 Percent Risk	2.36		2.29		2.22

* Denotes the highest cycle for which T_{imax} exceeds its limit value

References

1. Rosner, Bernard, "Percentage Points for a Generalized ESD Many-Outlier Procedure," Technometrics, Vol. 25, No. 2, May 1983, pp. 165-172.
2. Iglewicz, Boris, and Hoaglin, D.C., The ASQC Basic References in Quality Control: Statistical Techniques, Volume 16: How to Detect and Handle Outliers, American Society for Quality Control Quality Press, Milwaukee.

Alex T.C. Lau, TCL Consulting, Whitby, Ontario, Canada, is chairman of the Coordinating Subcommittee on Quality Assurance and Statistics (D02.94) and D02.01.0B on Precision, which are part of ASTM Committee D02 on Petroleum Products, Liquid Fuels and Lubricants. An ASTM International fellow, Lau is also a member of Committees E11 on Quality and Statistics, E36 on Accreditation and Certification, and F08 on Sports Equipment and Facilities.

Dean V. Neubauer, Corning Inc., Corning, New York, is an ASTM International fellow, is chairman of E11.90.03 on Publications and coordinator of the Data Points column; he is immediate past chairman of Committee E11 on Quality and Statistics.