datapoints
a statistics q&a

Dealing with Outliers

How to Evaluate a Single Straggler, Maverick, Aberrant Value

BY THOMAS MURPHY AND ALEX T. LAU

Q. How do you determine if a value is truly an outlier and how do you decide whether or not to proceed with the data analysis?

A. One of the prickly problems in data analysis is dealing with outliers in a set of data. An outlier is an observation with a value that does not appear to belong with the rest of the values in the data set. Outliers are also known by other names: maverick, flier, straggler or aberrant value. Two questions usually arise: 1) Is the value in question truly an outlier? 2) Can I eliminate the value and proceed with the data analysis?

Question 1 is one of outlier identification, and two essential tools are a graphical display of the data and a statistical test. An excellent graphic to look at the distribution of small data sets is the *dot plot*. For example, consider the data 5.3, 3.1, 4.9, 3.9, 7.8, 4.7 and 4.3, for which the dot plot is shown in Figure 1.

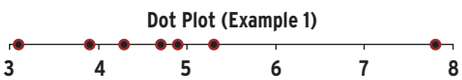

Figure 1 — Dot plot for data, 5.3, 3.1, 4.9, 3.9, 7.8, 4.7 and 4.3.

Here the value 7.8 appears to be an outlier because it falls well to the right of the others on the dot plot. In the plot, we are really looking at the gaps between the data values.

Two of the more commonly used statistical tests for a single outlier in a single set of data are the Dixon test and the Grubbs test. The Dixon test uses ratios of data gaps in different ways depending on the number of values in the data set. In the example above, the sample size is 7, and the ratio used is the gap between the outlier (7.8) and its nearest neighbor (5.3) divided by the gap between the largest and smallest values in the set. Thus, the Dixon ratio is:

$$(7.8 - 5.3)/(7.8 - 3.1) = 2.5/4.7 = 0.532$$

This value is compared with a critical value from a table, and the value is declared an outlier if it exceeds the critical value. The critical value depends on the sample size, n, and a chosen significance level, which is the risk of rejecting a valid observation. The table generally uses low risk significance levels like 1% or 5%. For $n = 7$ and a 5% risk, the critical value is 0.507. The Dixon ratio 0.532 exceeds this critical value, indicating that the value 7.8 is an outlier.

The Grubbs test uses a test statistic, T, that is the absolute difference between the outlier, X_O, and the sample average, $\bar{X}$, divided by the sample standard deviation, s. For the previous example, the sample average is $\bar{X} = 4.86$, and the sample standard deviation is $s = 1.48$. The calculated test statistic is:

$$T = |X_o - \bar{X}|/s = |7.8 - 4.86|/1.48 = 1.99$$

For $n = 7$ and a 5% risk, the critical value is 1.938, and $T = 1.99$ exceeds this critical value, again indicating that the value 7.8 is an outlier.

Getting to Question 2, it should be known that statistical tests are used to *identify* outliers, not to *reject* them from the data set. Technically, an observation should not be removed unless an investigation finds a probable cause to justify its removal. Some companies have defined procedures for such investigations, including retesting the material associated with the outlying observation, if possible.

In some cases, the physical situation may define the problem. For the three observations, 98.7, 90.0 and 99.7, the Dixon ratio is

$$8.7/9.7 = 0.897$$

The critical value for $n = 3$ and 5% risk is

0.941, so the value 90.0 cannot be identified as an outlier! Part of the reason might be the close proximity of the other two values. However, if the values recorded are human body temperatures in degrees Fahrenheit, then an outlier test is certainly not required to conclude that something is amiss. This example also illustrates that it is difficult to identify outliers in small data sets, such as $n < 5$. ASTM E691, Practice for Conducting an Interlaboratory Study to Determine the Precision of a Test Method, discourages such outlier tests for small groups of repeated test results within a single laboratory and suggests other methodologies for identifying aberrant data sets.

If an investigation does not find a probable cause, then what should be done? One approach would be to conduct the data analysis both with and without the outlier. If the conclusions are different, then the outlier is seen to be influential, and this should be noted in the report. Another option is to use robust estimators for characterizing the data set, such as the sample median rather than the sample average.

ASTM E178, Practice for Dealing with Outlying Observations, contains many statistical procedures for outlier testing. Other criteria are given in this standard for single outliers as well as tests for multiple outliers, and the standard also gives guidance on which test to use. A more comprehensive reference for outlier testing is the book, *Outliers in Statistical Data*, published by Wiley. Another useful and more practical reference is the American Society for Quality "Basic Reference in Quality Control, Statistical Techniques, Volume 16: *How to Detect and Handle Outliers*," ASQC Quality Press. Other references are listed in ASTM practice E178.

When there are multiple outliers in a data set, the investigation becomes more complicated, but test procedures are available for this situation. One problem is that one outlier may mask another outlier in a single outlier test. The Dixon test overcomes this by redefining the gaps to use as the sample size increases. This approach is well covered in E178 and other sources.

It is important to note that the first order of business is to look at the data graphically for potentially more than one outlier, either in the same or opposite direction, prior to using the Dixon or Grubbs technique. These techniques are designed to detect a single outlier in a dataset, and hence are not suitable for multiple outlier detection. One robust and comprehensive technique to effectively identify multiple outliers is the generalized extreme studentized deviate many-outlier procedure, described in the ASQ Basic Reference, Volume 16. While multiple outliers are beyond the intended scope of this article, interested readers are referred to the above literature for guidance, or you may choose to consult a statistician.

THOMAS MURPHY, *T.D. Murphy Statistical Consulting LLC, is chair of Subcommittee E11.30 on Statistical Quality Control, part of ASTM Committee E11 on Quality and Statistics.*

ALEX T. LAU, *Engineering Services Canada, is chair of Subcommittee D02.94 on Quality Assurance and Statistical Methods, part of ASTM Committee D2 on Petroleum Products and Lubricants, and a contributing member of E11.*

DEAN NEUBAUER *is the DataPoints column coordinator and E11.90.03 publications chair.*

Statistics play an important role in the ASTM International standards you write, such as the development of precision and bias statements for test methods, running interlaboratory studies, knowing how to round numbers properly and determining sample size. A panel of experts is ready to answer your questions about how to use statistical principles in ASTM standards. Please send your questions to SN ***Editor in Chief Maryann Gorman at mgorman@astm.org or ASTM International, 100 Barr Harbor Drive, P.O. Box C700, W. Conshohocken, PA 19428-2959.***