

Rogues and Suspects: How to Tackle Outliers

Analytical scientists, and indeed everyone involved in quantitative studies, know that it is usually necessary to make replicate measurements in order to characterise the random variations that inevitably occur. They also know that when such replicates are performed, it is not uncommon to find that one (or more) of the results seems not to agree with the rest. Proper treatment of such suspect values is an area of great importance.

Some Simple Examples

Suppose four titrations are performed, using the same equipment, reagents, and manipulative techniques to ensure that the results are as repeatable as possible. Quite often we shall find that three of the results are in close agreement, but the fourth is rather different (Figure 1a). If we cannot simply explain the discrepancy (because of faulty equipment, technique, or data recording) we must decide whether it is permissible to reject the fourth value as an *outlier* before calculating the mean and standard deviation of the results, using them in significance tests etc, or whether we must utilise all the values, including the suspect one. In a separate, but very common, type of experiment we might use eight reference materials to yield a calibration graph for an instrumental analysis. We may then find that seven of the resulting measurements lie close to a straight line or smooth curve, while the eighth value is inexplicably out of line. The same question arises: can we reject this eighth value as an outlier before calculating the best straight line or curve, or must we include it?

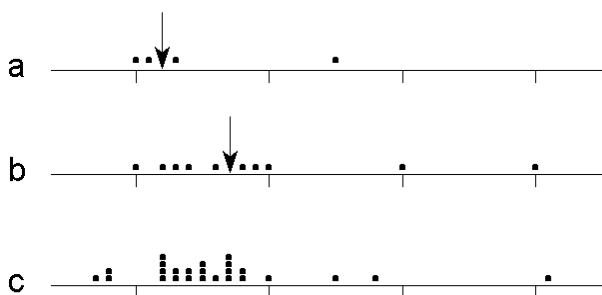


Figure 1. Examples of dot-plots of replicated results. Medians, where relevant, are denoted by arrows.

This problem of suspect values can arise in any area of analytical measurement, such as the results of proficiency testing schemes and of method performance studies (collaborative trials). Suspect values may also be seen in multivariate measurements. The problems they pose in theory and in practice are clear if we re-examine the titration example. As every effort was made to ensure that the titrations were repeatable, we would expect the results to form a sample from a (hypothetically infinite) population with a normal (Gaussian) error distribution. There is thus a finite chance that an entirely valid single measurement might be substantially different from the sample mean. Does this imply that we can never reject suspect values? Probably not, though it is certainly true that, for small samples, one measurement must be very different indeed from the rest before it can be considered as an outlier. If suspect values are needlessly retained, measurement quality suffers, while if they are wrongly rejected we may be guilty of adjusting the data, perhaps subconsciously, to fit our preconceived idea of what the results should be - a misdemeanour of which scientific giants such as Isaac Newton, Gregor Mendel and Robert Millikan, amongst many others, have been accused!

Another aspect of handling suspect values is that the dictum “if in doubt, take more measurements” may not help. If we perform our titration ten times, it is possible that eight of the results will agree closely, the other two being suspect. Both the latter may be higher than the rest of the values (Figure 1b), both lower, or one higher and one lower! So any satisfactory method of handling suspect values must also be able to deal with multiple instances.

Initial Data Analysis (IDA)

Simple IDA methods can provide much commonsense guidance in this area. The human eye-brain combination detects patterns and trends much more easily in graphical form than when the data are presented as a list of numbers. So calibration and other graphs should always be inspected visually, and dot-plots such as those in Figure 1, in which each measurement is represented by a dot on a linear scale, provide simple graphical summaries of replicate data. Very simple situations (Figure 1a) may not gain much from graphical presentation, but for larger data sets, such as those in Figure 1b (two outliers?) and Figure 1c (possibly a skewed error distribution, so the highest value may not be an outlier) dot-plots often highlight features that would not otherwise be so apparent.

Significance tests for outliers

In datasets where suspect values have arisen, several approaches to their treatment have been used. The most obvious and perhaps still the most popular is to apply a significance test. Several tests have been developed for this purpose. Each is used in the conventional way. First a null hypothesis (H_0) is established; in this case H_0 will be that the suspect value (here we assume there is just one) is not an outlier, that is, that the whole dataset resembles a random sample from a single population. Then the probability of obtaining the actual experimental results under H_0 is found, and if this probability is low (often $p < 0.05$) then H_0 is rejected, that is, the suspect value can be rejected as an outlier. The key probability is found using a simple equation to convert the experimental data into a *test statistic*. The equation depends on the test chosen, but it usually compares the proximity of the suspect value to the rest of the data with some measure of the overall spread of the data (the latter to include the suspect value because H_0 assumes it is *not* an outlier). For replicate measurements the best-known tests are those of Dixon and Grubbs: those and other tests will be described in more detail in future Technical Briefs. The main problem with outlier tests is that, to find the probability that the experimental data will occur, we have to make an assumption about the error distribution of the population from which the data sample comes. Usually it is assumed that this distribution is normal so, if this is not the case, misleading conclusions can result from the test. For example a value that is an apparent outlier on the assumption of a normal distribution may not be an outlier if the distribution is in fact log-normal. Outlier tests are also complicated (though not unusable) in the presence of two or more outliers (Figure 1b). Significance tests are also used to tackle outliers in other situations; for example Cochran's test is used to identify unusual group variances in Analysis of Variance (ANOVA) methods.

Median-based methods

A second approach to the problem of suspect values is to use median-based methods. The median of a data sample is found by putting the values in numerical order. If the number of measurements is odd the median is the middle value of the ordered sample; if there is an even number of values it is the mean of the two middle ones. The median is obviously unaffected by the values at the extremes of the data sample (see Figures 1a. and b.), so in some cases it may be a more sensible measure of central tendency than the mean. Equally, entirely valid data that contribute to the calculation of the mean do not affect the value of the median, so the mean is a better measure if outliers are not suspected. The median can also be used to provide a measure of the spread of the data through the *median absolute deviation* (MAD!). This is obtained by finding the magnitude (that is, without regard to + or - signs) of the deviation of each individual value in the sample from the sample median, and then the median of these deviations. The MAD forms the basis of another possible outlier test: if an individual result differs in magnitude from the sample median by more than 5 times the MAD it is a candidate for rejection. The MAD is also used in robust statistics (see below). Other median-based methods

have not been much used in the analytical sciences, though one provides the basis of a good method for calculating regression lines when outliers are suspected; this will also be described in a later Technical Brief.

Robust statistics

In recent years the favoured approach to the treatment of data containing suspect measurements has been the use of robust statistics (see Technical Brief 6), which work well when the data come from distributions that are not *greatly* different from the normal distribution. They are particularly suitable when there are suspect values, and also when the distribution of errors is *heavy-tailed*, that is, it is symmetrical, but with more values than expected that differ substantially from the mean. These two situations are clearly linked, though heavy-tailed data do not always contain outliers: they may arise from the overlaying of several normal distributions with similar means but with different variances, for example if measurements of a particular material are made using different techniques. The key feature of robust methods is that they avoid the question of accepting or rejecting doubtful values by *reducing their weights* (that is, the importance given to them in calculating statistics). Many robust methods are iterative, and their popularity reflects the fact that they are now easily implemented through simple software.

A simple example of the approach is provided by a process called *winsorisation*. This term covers a range of methods, but in a simple version the residuals of the measurements at the high and low ends of the data (that is, the differences between the values and the mean or median) are reduced so that they become the same as the *next largest* positive and negative ones. In effect the highest and lowest measurements have been shifted to become the same as the next highest and lowest ones! Obviously if there are no outliers at either end of the sample range, this adjustment will make only modest differences to the mean and standard deviation; but if such outliers do occur, their influence on the mean and standard deviation will be reduced. Such methods seem at first to involve unjustifiable interference with the raw data, but in practice robust methods work well in a range of measurement situations, and would probably be even more used if suitable software was widely available.

J N Miller

This Technical Brief was prepared for the Analytical Methods Committee by the Statistical Subcommittee under the chairmanship of M Thompson.

CPD Certification	I certify that I have studied this document as a contribution to Continuing Professional Development.
Name.....
Signature.....	Date.....
Name of supervisor.....
Signature.....	Date.....