

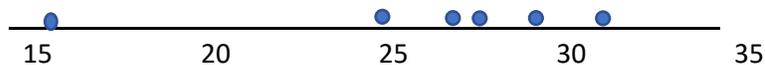
What is an outlier?

In repeated chemical analysis which assumes normal probability distribution, we may find some extreme (i.e. the biggest or smallest result) is a suspect which seems to be quite different from the rest of the data set. In other words, this result does not seem to belong to the distribution of the rest of the data. This suspect value is called an outlier.

Data given below and Figure 1 on the analysis of total aflatoxins in ppb level can be taken as a typical example:

15.1 24.9 26.7 27.1 28.4 31.1

Figure 1: Results of a determination repeated by two analysts



Visually, the result at 15.1 is suspected of being an outlier.

If we were to include it in calculating statistics such as the mean and standard deviation, the estimates would not be representative of the true population mean and standard deviation values which are used to evaluate and compare the precision and accuracy of analytical methods. Therefore, it is important that outliers are identified, confirmed and excluded from further calculations.

However, we must not simply discard any suspect measurements. To reject such a value wrongly might result in the value of the mean being shifted and the standard deviation being unrealistically small. Hence, there must be a basis for identifying data as outliers and a strategy for dealing with them.

The outlier statistic tests are abundant. They provide objective criteria for taking investigative or corrective action for any error. The important point to note is that a positive outcome from an outlier test is best considered as a signal to investigate the cause; usually, outliers must not be removed from the data set solely because of the result of a statistical test.

This is to avoid being biased. Of course, if there is good reason to believe that gross errors have incurred during the laboratory analysis, rejection of extreme outliers is necessary to prevent errors from unduly influencing results.

The general strategy to be taken include:

- Test at 95% and the 99% confidence level
- All outliers should be investigated and any error corrected
- Outliers significant at the 99% level (often termed stragglers) should be rejected only if there is an additional technical reason to do so
- Successive testing and rejection is permissible, but not to the extent of rejecting a large proportion of the data.

Remember that chemical analysis assumes underlying normality. If the data are Poisson distribution (discrete data such as microbiological colony counts), many valid high values might be incorrectly rejected because they appear inconsistent with a *normal* distribution. It is also crucial to consider whether outlying data might represent genuine features of the population.

Subsequent blog articles that follow will discuss various outlier statistic tests commonly adopted in chemical laboratories.