A short note on non-parametric and robust methods

We often assume that our analytical data collected follow the normal (Gaussian) distribution. This assumption is somehow supported by the central limit theorem which shows that the sampling distribution of the mean (average) results may be approximately normal even if the parent population has quite a different distribution. However, it must be stressed that the central limit theorem is not really valid for very small data sets (say, three or four readings) frequently used in analytical works.

Indeed, we cannot always assume the data collected are normally distributed. Some sets of data occurring in the analytical sciences can also have different distributions. For example, the antibody concentration in the blood sera of a group of different people can be expressed approximately as a log-normal distribution. Another example is the total bacterial colony count on a meat sample after 48-hr incubation at 37° C. The variation of the repeated counts tends to follow a Poisson distribution, instead, because the contributors to the result variation are many, including the colony form counting is on living microorganisms which can be of different growth strength.

There is growing evidence that even when repeated measurements are made on a single test material, the distribution of the results is sometimes symmetrical but not normal, e.g. the data may include more results than expected which are distant from the mean. The outlier testing fails to single out these outliers as there are results close to them. Such heavy-tailed distribution can be mistakenly regarded as normal distribution with the addition of grouped outliers arising from gross errors.

Heavy tailed distribution data may also arise from the superposition of two or more normal distributions with the similar mean value, but with significantly different standard deviations. This situation can happen if we use more than one piece of analytical instrument to do the measurements.

Hence, we can consider two groups of statistical tests for handling data that may not be normally distributed. Methods which make no assumption about the shape of the distribution from which the data are taken are called **non-parametric** or **distribution-free** methods. Examples of these methods include the sign test and the chi test. These calculations are relatively simple to be carried out.

In here, we do not talk about arithmetic mean or average as the "measure of central tendency" of a set of results. Instead, we use median, which is the value of the $\frac{n+1}{2}th$ observation if n is odd, and the average of the $\frac{n}{2}th$ and the $\frac{n+1}{2}th$ observations if n is even after arranging the set of data in ascending order. Median is not affected by outlier values.

In non-parametric statistics, the usual measure of dispersion (replacing standard deviation) is the interquartile range (IQR) as median divides the sample of measurements into two equal halves; if each of these halves is further divided into two, the points of division are called the upper and lower quartiles. However, this IQR concept is not widely practiced in analytical world. This is because we do not have many repeated measurements made and hence the differences in the calculated IQR values are large. But IQR has been extensively used in the laboratory proficiency testing programs with many participating laboratories where the chance to get groups of extreme results is high.

The median and the IQR of a measurement set are just two of the statistics which feature strongly in **initial data analysis** (IDA) or also called **exploratory data analysis** (EDA).

Robust methods are based on the belief that that the underlying population distribution may indeed be approximately normal, but with the addition of data such as outliers that may distort this distribution. These techniques in essence operate by *reducing* the weight given to suspicious results or outliers, so are appropriate in the cases of heavy-tailed distributions, and their acceptance and use have increased dramatically in recent years

The robust methods differ from non-parametric methods in that they often involve iterative calculations that would be lengthy or complex without a computer, but their rise in popularity certainly owes much to the universal availability of desk-top computers.

There are some very simple robust methods which do not require iterations because they arbitrarily eliminate suspicious data with some degree of confidence, rather than down-weighting a proportion of the data, such as **trimming** the outlier results.

Another less arbitrary robust approach is provided by **winsorisation**. In its simplest form, this process reduces the importance of the measurements giving the largest positive and negative deviations from the mean or median by moving the measurements so that these deviations become equal to the next largest or smallest ones (or perhaps the third largest ones). The advantage of this approach is that it is applied to a data set lacking suspect values or actual outliers, the effects on the calculated measures of location and spread are small, so no harm is done.

A robust variance estimate can also be derived from a statistic related to the unfortunately abbreviated **median absolute deviation** (MAD) which is calculated from:

$$\mathrm{MAD} = \mathrm{median}\ [\ |x_i - \mathrm{median}(x_i)|]$$

where median(x_i) is the median of all the x_i values, i.e. all the measurements.

I had written an article on MAD in one of my earlier blogs.