Statistics and sampling strategy

With the latest revision of ISO/IEC 17025:2017 coming to effect by the end of this year, more attention is now being given to the component of the uncertainty arising from the sampling steps of an analysis, as it is reckoned that the result of a sample is as good as the sample that is based upon. Indeed, in many cases, sampling uncertainty contribution might be the largest single component of the overall uncertainty of the analysis. This is particularly the case when the material under testing is heterogeneous in nature. This occurs in the field of geology, environmental and food sciences. The importance of good sampling practice therefore cannot be over emphasized.

Statistically speaking, if we make one analysis on each of the h samples drawn from a population, we obtain the confidence limits of the sample mean as below:

$$\mu = \bar{x} \pm t_{(n-1)} s / \sqrt{h} \qquad \ldots \text{ Eq [1]}$$

where $\bar{x}$ is the mean of the h sample measurements and s^2 is the variance of the measurements.

Now, the total variance of the population is given by σ^2, which is estimated by s^2 and is the sum of measurement and sampling variances, i.e. σ^2_{meas} and $\sigma^2_{sampling}$, i.e. $\sigma^2 = \sigma^2_{meas} + \sigma^2_{sampling}$. That also means that σ^2/h (estimated by s^2/h) is the variance of the mean $\bar{x}$.

The argument follows that if there is only a single measurement made on each of the h samples, we are not in the position to examine the variance of measurement. Hence, the confidence limit of the sample mean of the h samples does not cover the uncertainty of analysis (measurement).

If the value of each sample is the mean of n replicate measurements, then the variance of the mean is to be:

$$\left(\frac{\sigma^2_{meas}}{n} + \sigma^2_{sampling} \right) / h \quad \text{or} \quad \left(\frac{\sigma^2_{meas}}{nh} + \frac{\sigma^2_{sampling}}{h} \right) \qquad \ldots \text{ Eq [2]}$$

For maximum precision of reported result, it is obvious that we must reduce the variance of the mean as much as possible. In the part of analysis, we can improve the measurement precision either by using a more precise method of analysis or by increasing n replicate measurements. For the sampling precision, we can simply increase the number of samples, h.

By simple calculation, it can be shown that there is a limitation to reduce the measurement variance because a point will be reached where any further reduction of measurement variance does not significantly improve the total variance. Statistically, it can also be proved that if the measurement variance is less than 1/3 of the sampling variance, the omission of measurement variance in the overall variance estimation will only cause a 5% error and so can be neglected.

A possible sampling strategy with bulk material suggested is to take h samples and blend them before carrying out n replicate measurements. By doing so, the variance of the mean of these replicate measurements is then:

$$\sigma^2_{meas}/n + \sigma^2_{sampling}/h \qquad \ldots \text{Eq [3]}$$

When this equation is compared with Equation [3], it is obvious that the variance of mean is smaller when we analyze h samples individually for n replicate measurements. However, the costs of sampling and analysis (nh against n measurements) and economic consideration are significantly different and must be seriously noted.

Eurachem, for example, has made a recommendation on good sampling practice, involving taking several samples from the target material and each of these samples is divided into two, and duplicate measurements are made on each of the two sub-samples in repeatability conditions. ANOVA can then be used to separate the contributions that the sample and measurement variations make to the uncertainty.

In the next blog article, we shall demonstrate the step-by-step ANOVA principle in tackling this challenge.