Statistical considerations for sampling strategies (Part I)

Sampling is the process of selecting a portion of material (i.e. *sample*) to represent or provide information about a larger body of material (i.e. *population*). The requirement of a properly defined sampling procedure whereby a part of a substance, material or product is taken to provide for testing or calibration as a representative of its whole body cannot be over emphasized.

Unfortunately, we chemists in most of the time have little or no involvement with the selection of the samples that are submitted to the laboratory for analysis. For example, the inspector of a testing-inspection-certification (TIC) company carries out sampling of a lot of cargo shipment in the client's warehouse and send the sample(s) back to the laboratory for testing. Notwithstanding this fact, it is important for us to remember that if an appropriate sampling strategy has not been used prior to analysis, the results produced in the laboratory will be of little value.

It is frequently found that sampling introduces the great majority of the uncertainty associated with measurements. Careful planning and execution of a sample protocol tailored for the specific application are therefore very important.

We have discussed in the last few notes on the subject of randomization which leads to procedures to collect random but representative samples from a population statistically.

Let's now discuss some of the strategies available for sampling and their strengths and weaknesses.

1. Simple random sampling

As we now know, in random sampling, each item in the population for laboratory analysis has an equal chance of being selected through the use of a random number table or a random number generation from MS Excel® spreadsheet or some statistical software like R programming.

This random sampling can also be applied in a laboratory setting for sub-sampling from a set of discrete items, such as ampoules or food packages. The usual method is to number the items sequentially, then use a table of random numbers or random number generated from software to select items randomly from the set.

For particulate materials such as resin or grainy beans, sub-samples can be obtained by repeated cone-and-quartering process or rotatory sampling, with due care to avoid bias from size or density selection and to prevent loss of moisture or volatile oil in some sensitive volatile materials such as black and white pepper, due to heat generated during the sub-sampling process.

The aim of all these procedures is to generate sub-samples in which every member of the population (laboratory sample) has an equal chance of appearing in a sub-sample.

Advantages:

- Simple to implement
- Often sufficient for sub-sampling within a laboratory

Disadvantages:

- having very variable intervals between successive test items, making it a poor choice for monitoring items from a continuous sequence
- for non-homogeneous materials or for sample consisting of identifiable sub-groups with substantially different properties, the variance in this simple random sampling method is among the highest of the strategies described here. A printed circuit board sample, for example, contains many electronic parts with different amounts of ROHS (Restriction of Hazardous Substances) under the EC Directives.

Statistical Treatment

Statistical treatment for random samples is comparatively straightforward.

Where the test samples are measured separately, the simple mean $\bar{x}$ of the measured values x_i for each test sample provides an unbiased estimate of the bulk material composition.

The uncertainty in $\bar{x}$ due to sampling variation (usually, for within-laboratory sampling, arising from inhomogeneity in the material) depends on the number of test samples n taken relative to the number of items N in the population. For example, in taking 10 tins of a product as representative of a laboratory sample of 100 tins, n = 10 and N = 100. The standard error $s(\bar{x})$ in $\bar{x}$ arising from sampling is given by:

$$s(\bar{x}) = s_{sam}\sqrt{\frac{1-f}{n}}$$

where $f = n/N$ and s_{sam} is the standard deviation found for the sampling process.

If the analytical uncertainty is small, s_{sam} is simply the standard deviation s of the observations x_i. Otherwise, if the test samples are measured with replication, s_{sam} can be obtained from the between-group component of variance from a one-way ANOVA. It can be proven that if the analytical uncertainty is less than 1/3 of the sampling uncertainty, it is 'safe' to ignore the contribution of analytical uncertainty in the analysis.

When N is very large, $s(\bar{x})$ converges to $s_{sam}/\sqrt{n}$ as usual; for $n < 0.1N$, the correction f can usually be ignored for analytical purposes. This is because n/N will be very small and so f is ignored.

As n approaches N, the uncertainty associated with sampling variability approaches zero; if we have sampled the entire population, there is no remaining possibility of variability due to a different choice of test samples. But, in practice, it is not likely for us to take the whole lot of items (i.e. population) for analysis!

The other sampling strategies will be discussed in the following papers.