General discussion on Sampling Statistics

Sampling theory is the field of statistics that is involved with the collection, analysis and interpretation of data gathered from random samples of a population under study. Some test laboratory personnel may get themselves involved in field sampling before carrying out analyses, such as those working in factory QC laboratories and cargo inspection laboratories.

It may be noted that the application of sampling theory is concerned not only with the proper selection of observations from the population that will constitute the random sample, but also involves the use of probability theory, along with prior knowledge about the population parameters, to analyse the data from the random samples and develop inferences and conclusions from the analysis results.

The normal distribution, along with related probability distributions, is most heavily utilized in developing the theoretical background for sampling theory.

Sampling Plans

Development of a sampling plan is one of the most important parts of the overall planning process for providing reliable samples for laboratory analysis. The sample selection process itself is critical because we should avoid bias in the course of sampling.

There are different sampling procedures in choosing the observations that will constitute their random samples of the population. The objective of these procedures is to select samples that will be representative of the population from where they originate. An important point to note is that these samples, also known as random samples, will have the property that each sample has the *same* probability of being drawn from the population as another sample.

Simple Random Sampling

Simple random sampling is the process of selecting a random sample from a finite or infinite population. It has two properties that make it the standard against which we measure all other methods:

- *Unbiased*: each unit has the same chance of being chosen

- *Independence*: selection of one unit has no influence on the selection of other units.

If there is a finite population of size N and we want to take n size sample each time, then we can draw a total of ${}^{N}C_{n}$ different samples of such size n, and each of these random samples have an equal $\frac{1}{{}^{N}C_{n}}$ probability of being selected. If for example, there are 100 drums of chemicals for a shipment, how many ways that 4 different drums can be randomly selected to form a sample from them for testing?

The answer is: we can have ${}^{100}C_{4}$ or $\frac{100\times 99\times 98\times 97}{4\times 3\times 2\times 1}$ or 3,921,225 ways!

If we can devise a procedure for selecting a sample of 4 drums such that each of these almost 4 millions samples has an equal probability (i.e. equal to 1/3,921,225) of being selected, then the sample selected would be a random sample.

The main problem of random sampling is that the experience of the person who is doing the sampling is very important and can strongly influence any subjective distortions such as :

- preferred sampling at easily accessible locations
- intuitive selection of either obviously darker or lighter colour of the population
- tendency towards an intuitive regular distribution of sampling points
- etc.
-

Systematic sampling

Systematic sampling is the sampling procedure wherein the k^{th} element of the population under study is selected for the sample, with the starting point randomly determined from the first k elements. The value of k is often dependent on the structure and objectives of the sampling experiment, as well as the population under study.

In systematic sampling, the sample values are spread more evenly across the population, thus, many systematic samples are highly representative of the population from which they were selected.

Yet, one must be careful that the value of *k* does not result in a sampling interval whose periodicity would compromise the randomness of the observations.

However, systematic sampling is easier, quicker, and cheaper than random sampling. The precision of the results obtained is mainly influenced by the 'gap' between the sampling points.

The primary disadvantage of systematic sampling is that no valid estimate of sampling error can be calculated from a *single* sample taken.

For example, in inspecting a shipment of 5,000 bags of rice, we can choose to inspect every 50th bag in the batch. The bags inspected are therefore the 50th, 100th, 150th, and so on until the 5000th bag.

In doing so, we ensure that each 50th bag is not specifically picked for inspection. Otherwise, the samples selected for inspection will not be representative of the entire batch of 5,000 bags of rice.

Stratified Random Sampling

Stratified random sampling is the sampling procedure that divides the population under study into mutually exclusive sub-populations, and then selects random samples from each of these sub-populations. These sub-groups (the so-called *stratas*) are determined in such a way that the parameter of interest is fairly homogenous within a sub-population.

By doing so, the variability of the population parameter within each sub-population should be considered less than its variability for the entire population. Oftentimes, there is a relationship between the characteristics of a certain population and the population parameter.

Example :

In an Environmental Impact Assessment (EIA) study of a designated area for industrial development, one can classify the total area to be investigated into more homogenous sub-units, such as:

- climatic conditions
- human activities
- ecological conditions
- river qualities
- etc. etc.

before randomly select samples of the sub-units for study.

How representative is a sample?

We assume the analytical value of the parent population under consideration is *normally distributed*, with a population mean μ and the standard deviation σ. When a limited number of *n* samples is analyzed, the mean result would be $\bar{x}$ and standard deviation *s*.

It is expected that the bulk sample taken from the parent population represents this parent population for the analyte of interest. This representativeness of a sample may be quantified by the degree and the reliability of approximation of the obtained mean value x as compared to the true mean of μ of the parent population.

The confidence interval $(\bar{x} - CL) < \mu < (\bar{x} + CL)$ can be derived from the Student's *t*- distribution with:

$$CL = t_{p,n-1} \frac{s}{\sqrt{n}}$$

on the basis of the selected statistical certainty *P*. For *P* values of 90%, 95% or 99% are generally assumed. The factor $t_{p,n-1}$ whose values are tabulated in the Student's *t*– table becomes smaller with an increasing number of samples, *n* and with decreasing statistical certainty *P*, the confidence limit changes similarly.

Sample Uncertainties

The uncertainties in the data due to the sample always need to be evaluated. The variance of the population can be estimated easily, provided the samples are *randomly* selected and amenable to statistical analysis.

The sampling operation can introduce both systematic and random errors.
It may be impossible to quantify the individual components of sampling variance.

However, the overall sampling variance can be evaluated by taking a number (at least 7) of samples under conditions where the samples are expected to be essentially identical. The total variance consists of the sum of that due to the samples and to their measurement. Thus :

$$\left(\frac{s_{Total}}{\bar{x}_{Total}} \right)^2 = \left(\frac{s_{Sampling}}{\bar{x}_{Sampling}} \right)^2 + \left(\frac{s_{Analysis}}{\bar{x}_{Analysis}} \right)^2$$

The analysis variance is subtracted from the total variance to obtain sample variance.

It may be noted that in the case of solid samples, particularly solid waste where the distribution of pollutants of interest may not be homogeneously distributed, the sampling error usually exceeds the analytical error significantly. This applies equally well for sampling in the field and taking a sub-sample from a sample received in the laboratory for analysis.

The other contributing factors to sample uncertainties include:

a. *stratification,* which is an insidious source of error in some analytical samples. Samples that were initially well-mixed may separate, partially or fully, over a period of time. It may be difficult (perhaps impossible) to reconstitute them. A good example is the necessary melting of a margarine sample for analysis and due to the nature of high water content and emulsification of the product, it is almost impossible to reconstitute it back into its normal form should a repeated analysis is required.
b. *holding time*, which is defined as the maximum period of time that can elapse from sampling to measurement before significant deterioration can be expected to occur. This will apply to perishable food products which are sensitive to storage temperature and period of storage.

Measurement situations encountered by chemists

The following table demonstrates four possible scenario encountered by analytical chemists during the course of their analytical works:

Table 1 : Measurement Situations

Situation	Variance	Significant	Not Significant
A	Measurement variance Sample variance	X	 X
B	Measurement variance Sample variance	 X	X
C	Measurement variance Sample variance	X X	
D	Measurement variance Sample variance		X X

The following discussion on statistical sampling plan needs the following assumptions which are reasonable to be made:

- the sampling operation must be a randomization process to ensure the samples are randomly selected and independent of any other sample in the group;

- a Gaussian normal distribution is applicable

- an estimated or assumed standard deviation is treated as if it were the population standard deviation

- for statistical calculation, the sample standard deviation estimate, s, based on the analytical process, is used in the conventional manner

- in the analysis of variance from several independent sources, it is assumed that the total variance is equal to the sum of the various components as below:

$$s^2 = s_1^2 + s_2^2 + s_3^2 + \ldots\ldots + s_n^2$$

Note: In sampling, we will use the following notation:

σ_A = standard deviation of measurement
σ_s = standard deviation of samples (within a stratum)
σ_B = standard deviation between strata

The following equations describe the minimum number of samples and/or measurement necessary to limit the **total uncertainty**, to a value ***E***, with a stated level of confidence as indicated by the value of *z* that is selected. For the 95% confidence level, *z* =1.96 or approximately, 2.

Situation A – Measurement variance is *significant only*

Minimum number of measurements, n_A where sample σ_S is negligible:

$$n_A = \left(\frac{z\sigma_A}{E_A} \right)^2$$

Situation B – Sample variance is *significant only*

Minimum number of measurements, n_s, where measurement σ_A is negligible:

$$n_s = \left(\frac{z\sigma_s}{E_s} \right)^2$$

Note:

a. If number of measurements, n_A is more than being considered feasible, then:

we should improve the precision of the test method to decrease σ_A, or use a more precise method of measurement if available for smaller σ_A, or accept a larger uncertainty, E_A

- If more samples are required than is feasible, then use a larger sample (smaller σ_S), or use composites (smaller σ_S), or accept a larger uncertainty, E_s

Situation C : Measurement and sample variances are *both significant*

$$E_{Total} = \sqrt{\frac{\sigma_s^{\ 2}}{n_s} + \frac{\sigma_A^{\ 2}}{n_s n_A}}$$

where, n_A is the number of measurements per sample.