Statistical consideration for strategic sampling – Part III

Following the discussions on random sampling methods after simple random sampling and stratified random sampling, we will now look into the other techniques.

3. **Systematic sampling**

As the term 'systematic' dictates, we have to select the first test sample at random and then pick up the subsequent samples at a fixed interval, such as every fifth item or at a fixed time during the streaming of items on a conveyor belt upon production.

For example, consider 1000 sequentially numbered items in a batch of a product of which 20 need to be selected for testing. What we have to do is to select the first at random from items marked 1–50, which every 50th item is chosen. If item number 9 was selected at random as the first test sample, the subsequent samples chosen would be 59, 109, 159, …, 959.

The main advantage for systematic sampling is due to its simplicity to implement over a period of time or a stream of material. In this type of 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. It also reduces the occurrence of long or very short intervals that may occur in simple random sampling.

However, systematic sampling can only be considered equivalent to random sampling if the variation in the material is reasonably independent of the item numbering or location. An extreme case can be that all items selected at intervals turn out to be of good quality whilst there might be bad ones in between! Thus, the precision of the test result obtained is mainly influenced by the 'gap' between the sampling points, and no valid estimate of sampling error can be calculated from a *single* sample taken.

Systematic samples are treated as if random, so the statistical treatment is as for simple random sampling as discussed in the Part I article previously.

4. **Cluster sampling**

Cluster sampling can be used to overcome time and cost constraints when a relatively small number of samples are required from a large and widely dispersed population which would be costly to examine in full using simple random sampling.

The first stage involves dividing the population into groups or 'clusters', usually on the basis of location. A number of these clusters are then selected at random for further evaluation. This differs from stratified random sampling in that in stratified random sampling, all strata are sampled; in cluster sampling, only selected sub-groups are sampled

A further important difference is that cluster sampling is most effective when groups are internally as inhomogeneous as possible, while stratified sampling works best on highly homogeneous strata.

Cluster sampling is often used in geographical surveys because of the large costs of travelling between very different locations. It is also applied in conducting census of population of a multi-racial society. Cluster sampling may also be useful in situations such as product QC sampling, where packaged products are boxed for distribution, and taking several units from a small number of distribution packages is less costly than opening a fresh distribution package for every unit taken.

Advantages

Where the cost of sampling distinct locations or groups is large, cluster sampling provides a random sample at lower cost than stratified random or simple random sampling.

Disadvantages

We can expect larger (that is poorer) sampling variance than simple random sampling with the same total number of individual test samples, because it selects a small number of clusters which may be significantly different.

Within a laboratory, it is unusual to incur disproportionate costs for sampling from different groups, so cluster sampling is unlikely to be useful in a laboratory setting.

Statistical treatment

If the clusters are of equal size and all units within a cluster are tested, the cluster means can be treated as a simple random sample, so that the estimated population mean is the mean of cluster means and the standard error is based on the standard deviation of cluster means.

For more complex situations, the calculations for cluster sampling are considerably more intricate, requiring weighting by cluster proportions and nested ANOVA to provide within- and between-cluster variances.

5. **Sequential sampling**

Sequential sampling is used when the number of samples required cannot be estimated before the start of the study. Samples are selected one at a time in sequence. After each sample has been tested, a decision is made as to whether the cumulative result is sufficient to make a decision (e.g. whether to accept or reject a batch of material) or whether further sampling is required. The sample size is not fixed in advance nor is the timeframe for data collection.

Sequential sampling technique is initially developed as a tool for product quality control. It has, provided that a suitable upper limit is set, the advantage of minimizing the number of samples to be taken to reach a conclusion. However, because each decision on further sampling depends on measurement results on previous samples, it can be very slow.

Statistical treatment

Generally speaking, in all sampling techniques, we will all come to a conclusion to accept the results under the null hypothesis or rejecting the null hypothesis while accepting the alternative hypothesis. In sequential sampling, we have a third option, i.e. to accept the null hypothesis, accept the alternative hypothesis or select another pool of samples and conduct the experiment once again. By so doing, we can obtain limitless number of samples before finally making a decision whether to accept the null or alternative hypothesis.

In conclusion, we note that sampling introduces the majority of the uncertainty in measurements on bulk materials (population). Proper planning and implementation of sampling are a crucial part of the overall measurement process. Hence, the desire to have representative sample for analysis cannot be over emphasized.

Within a laboratory, analysts are often faced with the problem of sub-sampling from a larger laboratory sample. Two strategies described in this series of notes, simple random sampling and stratified random sampling, are often sufficient for this purpose.