Repeated Systematic Sampling, Topics in Systematic Random Sampling

Repeated systematic sampling: As we have seen, the problem with systematic random sampling is that the validity of SRS formulas for variance estimation depend on the assumption of a random population. One way to address this problem is to instead use repeated systematic sampling. Essentially, instead of taking one 1-in-k sample, we take n_s samples that are each 1-in-k', where $k' = kn_s$. For example, instead of taking one 1-in-5 sample to yield 60 measurements, we can take 10 1-in-50 samples to get the same sample size. We then use the variation between the repeated samples to obtain a more accurate variance estimate. Let $\overline{y}_i$ be the mean of the ith systematic sample. Then our estimator $\widehat{\mu}$ and estimated variance $\widehat{V}(\widehat{\mu})$ for $\widehat{\mu}$ are:

$$\widehat{\mu} = \frac{1}{n_s}\sum_{i=1}^{n_s}\overline{y}_i$$

and

$$\widehat{V}(\widehat{\mu}) = \left(\frac{N-n}{N}\right)\frac{s^2_{\overline{y}}}{n_s}, \quad \text{where } s^2_{\overline{y}} = \frac{\sum_{i=1}^{n_s}(\overline{y}_i - \widehat{\mu})^2}{n_s - 1}.$$

Understanding population structure via an ANOVA decomposition of clusters: Previously it was stated that

$$Var(\overline{y}_{sy}) = \frac{\sigma^2}{n}[1 + (n-1)\rho],$$

where ρ is the intraclass correlation between elements in the same systematic random sample. Now we will explain this concept a bit more. Imagine the population as consisting of k clusters, each of size n, where $N = nk$ is the total population size. As stated previously, when we draw a $1 - in - k$ systematic random sample, we are actually performing cluster sampling in which we select one of the k clusters. Think of the population as consisting of these k clusters, and consider performing an analysis of variance (ANOVA) for differences among the clusters. This would lead to familiar ANOVA concepts like the mean-square between clusters (MSB), the mean-square within clusters (MSW), and the total mean-square (MST). It is shown in the text that for large N, the intraclass correlation ρ is approximated by:

$$\rho \approx \frac{MSB - MST}{(n-1)MST}.$$

Upon examining this expression, it becomes easy to see how different population structures affect ρ, and hence $Var(\overline{y}_{sy})$. In a random population, MSB and MST will be almost equal, so $\rho \approx 0$ and $Var(\overline{y}_{sy}) \approx \sigma^2/n$. In an ordered population, MSB will be less than MST, so ρ will have a negative value and $Var(\overline{y}_{sy})$ will be lower than under SRS. In a periodic population, MSB will be more than MST, so ρ will have a positive value and $Var(\overline{y}_{sy})$ will be much higher than under SRS.

A variance estimator based on successive differences: If observations y_i are independent and identically distributed (iid) with variance σ^2, then $d_i = y_i - y_j$ satisfies $E(d_i) = 0$ and $V(d_i) = 2\sigma^2$. Since $E(d_i) = 0$, the usual sample variance of the $d_i's$, $\sum(d_i - \overline{d})^2/(n-1)$ is adjusted to give $\sum d_i^2/n$ as an estimator of $2\sigma^2$. We can use this idea to develop a variance estimator based on successive differences. Now let $d_i = y_{i+1} - y_i$ for $i = 1$ to $n-1$. Then our variance estimator is:

$$\widehat{V}_d(\overline{y}_{sy}) = \frac{(N-n)}{N}\frac{1}{n}\frac{\sum_{i=1}^{n-1} d_i^2}{2(n-1)}.$$

Some examples in the text are used to show that for ordered populations, this estimator is better than the variance estimator based on SRS. However, this estimator does not resolve the problems encountered by data from a periodic population; repeated systematic sampling is the only method we have studied that is adequate in that situation.