Multiple-Frame Sampling

Most survey samples are selected from a single sampling frame presumably covering all of the units in the target population. Multiple-frame sampling refers to surveys where two or more frames are used and independent samples are taken respectively from each of the frames. Interences about the target population are based on the combined sample data. The method is referred to as Dual-Frame Sampling when the survey uses two frames.

Sampling designs are often dictated by several key factors, including the target population and parameters of interest, the population frame or frames for sampling selection of units, the mode of data collection, inference tools available for analyzing data under the chosen design, and the total cost. There are two major motivations behind the use of multiple-frame sampling method: (i) to achieve a desired level of accuracy with reduced cost and (ii) to have a better coverage of the target population and hence to reduce biases due to coverage errors. Even if a complete frame such as a household address list is available, it is often more cost effective to take a sample of reduced size from the complete frame and supplement the sample by additional data taken from other frames such as telephone directories or institutional lists which might be incomplete but less expensive to sample from. For surveys of human populations where the goal is to study special characteristics of individuals, such as persons with certain rare diseases, a sample taken from the frame for general population health surveys is usually not very informative. Other frames, such as lists of general hospitals and/or special treatment centers, often provide more informed data as well as extended coverage of the target population.

There are, however, unique features, issues and problems with inferences under
multiple-frame sampling, which require unique treatments and special techniques. Let 
\[ Y = \sum_{i=1}^{N} y_i \] be the population total of a study variable \( y \), where \( N \) is the overall population size. Suppose there are three frames \( A \), \( B \) and \( C \). Each of them may be incomplete but together they cover the entire target population. Let \( s_A \), \( s_B \) and \( s_C \) be the three independent samples taken respectively from frames \( A \), \( B \) and \( C \). The basic question is how to estimate \( Y \) using all three samples. It turns out that none of the samples can directly be used if the frames are incomplete. The most general picture is that the three frames divide the target population into seven disjoint domains: \( A \), \( AB \), \( ABC \), \( AC \), \( B \), \( C \) and \( BC \), where \( A \) contains population units from frame \( A \) but not covered by \( B \) or \( C \), \( AB \) includes all units from both \( A \) and \( B \) but not \( C \), \( ABC \) represents the set of units covered by all three frames, etc. If for instance frames \( B \) and \( C \) are non-overlapping, then the domain \( BC \) vanishes. We can re-write the overall population total as 
\[ Y = Y_A + Y_B + Y_C + Y_{AB} + Y_{AC} + Y_{BC} + Y_{ABC} \], where, for instance, \( Y_A \) is the population total for domain \( A \). Each of the three samples can also be partitioned according to the involved population domains: 
\[ s_A = s_a \cup s_{ab} \cup s_{ac} \cup s_{abc} \], \( s_B = s_b \cup s_{ba} \cup s_{bc} \cup s_{bac} \) and 
\[ s_C = s_c \cup s_{ca} \cup s_{cb} \cup s_{cab} \], where for instance units in both \( s_{ab} \) and \( s_{ba} \) are selected from the domain \( AB \) but \( s_{ab} \) is from frame \( A \) but \( s_{ba} \) is from frame \( B \), indicated by the first letter in the subscript. Estimation of \( Y \) is typically carried out through the estimation of domain totals using relevant sample data.

Major issues and problems with estimation under multiple-frame sampling include but are not restricted to the following: (i) Frame membership identification for all sampled units. This is required in order to post-stratify samples from different frames into
appropriate population domains. Additional questions regarding frame memberships need to be included for data collection; (ii) Estimation of domain totals using multiple samples. For instance, both \( s_{ab} \) and \( s_{ba} \) are selected from the domain \( AB \) and need to be combined to estimate \( Y_{AB} \). This may not be as straightforward as it appears to be, since the sampling designs used for frames \( A \) and \( B \) could be different. To obtain efficient domain estimates can be challenging, especially for domains such as \( ABC \) where all three samples \( s_{abc} \), \( s_{bac} \) and \( s_{cub} \) need to be combined; (iii) Lack of information on the domain population sizes. Under certain designs one may, for instance, have an estimator readily available for the domain mean \( \bar{Y}_{AB} \) and estimation of the total \( Y_{AB} \) requires that the domain size \( N_{AB} \) be either known or easily estimable, which is not always the case; (iv) Identifying and removing duplicated units from multiple-frame samples. This is required by some methods based on pooled samples where no single unit is allowed to be used more than once; (v) Handling the extra variation induced by the random sample sizes. Even if all the initial sample sizes are fixed, the sizes of the post-stratified samples are still random. This creates difficulties for variance estimation; (vi) Use of auxiliary information for estimation. The known auxiliary population information could be for the entire target population or for specific frames or for both. Incorporating such information with multiple-frame samples requires approaches which differ from the conventional single-frame methodologies.

When all frames are complete, multiple-frame sampling becomes the so-called multiple surveys where several independent samples are taken from the same target population.
Changbao Wu

See also Coverage Error; Dual-frame sampling; Frame; Mode of data collection; Target population

Further Readings

