

ST 432 Dual and Multiple Frame Sampling.

Class Notes for your Information

April 2009

Introduction

List Frames

I have tried to emphasize the crucial importance of having a good sampling frame for sound application of probability sampling. In many examples of human surveys, mentioned in this class, I have implicitly emphasized lists of names (e.g., telephone directory; list of licensed anglers etc). There are several problems that sometimes arise with list frames.

One problem is obviously that a list frame for the population of interest may not exist. Then there are two reasonable solutions: firstly to construct one which could be very costly and not easily done and secondly is to use an area frame which I discuss below. For a crucially important survey that is of national significance a list frame may be constructed but this is seldom done. I totally discount another approach which is not to use probability-based sampling. This is commonly done in market surveys, but I think it is really illogical and potentially very dangerous as totally unsound inferences may result and this could potentially discredit surveys in general when in fact only poorly designed surveys are at fault.

Sometimes list frames may suffer from having duplicate names on the list and usually those duplications would be removed before using the frame. Another, often more serious problem, is that the list frame used may be very incomplete which can cause negative bias in estimates if not accounted for before final estimates are presented. Recall that

$$\hat{\tau} = N\bar{y},$$

is the standard estimator of the population total for simple random sampling without replacement and can be used to illustrate the problem we face. If we have N' from the list which is less than the now unknown N then:

$$\hat{\tau} = N'\bar{y}',$$

and clearly that will cause a negative bias for the population total. Of course the incompleteness is likely “nonrandom” as well so that \bar{y}' is also likely to be a biased

estimate of the population mean. This bias could be positive or negative depending on the nature of the incompleteness. Therefore, for a valid survey design, we have to deal with the incompleteness and we use a dual frame approach described below but first I discuss area frames because they are part of the solution as well.

Area Frames

To make things more intelligible let us think of a household survey in a poor country where there are no lists of households available. Suppose your population of interest is defined by some region or total area (county, state, country) and you define your *primary sampling units* as areas A_1, A_2, \dots, A_N . Then you take a sample A_1, A_2, \dots, A_n probably using probability proportional to some measure of size of the areas as they are likely unequal.

Within each A_i there are M_i households which are the *secondary sampling units*, but this number is unknown except for the sampled areas where part of the survey is to establish how many households are present. Once M_i has been enumerated by searching the area then a sample of m_i of these households would be sampled in some probabilistic way (perhaps srs). In some surveys any one member of the household might be chosen to be interviewed. In others a sample of household members might be taken making the members of the household the *tertiary sampling units*.

Estimation would involve approaches from multi-stage cluster sampling where special account would probably need to be taken of size of the areas (psu's) and size of the households (ssu's). Estimators could not assume that the M_i are all known for the population (only for the sampled psu's).

Area Frame sampling involves complete coverage of the population of households (all households have a probability of being in the sample) whereas an incomplete list frame does not (households left off the list frame have 0 probability of inclusion). However, the complete enumeration of households in each sampled area means that the cost of doing the survey could be high. To quote Kott and Vogel (1995) "Area frame sampling ensures completeness but at a greater cost per completed interview, while list frame sampling is less complete but also less costly and more effective for targeting large and/or rare items".

Now we consider a combined approach of sampling the list and area frames which is called dual (or more generally multiple) frame sampling. The gain is that one can exploit the strengths and offset the weaknesses of each type of frame.

Dual Frame Sampling

Multiple frame (typically dual frame) estimation methods have been utilized for more than 50 years (Hansen et al. 1953; Hartley 1962; Cochran 1965; Hartley 1974). In a typical application such as a business or farm survey there is an *incomplete* list frame and a *complete* area frame (Kott and Vogel 1995 is a good reference for an overview of the topic). Haines (1997) and Haines and Pollock (1998) applied and extended these methods to some wildlife surveys. By taking a large sample from the incomplete but inexpensive-to-sample list frame and a smaller sample from the complete but expensive-to-sample area frame one can obtain a much more efficient estimator for a fixed cost than if the area frame alone were used. (If the list frame alone were used very serious bias would result due to its substantial incompleteness).

To illustrate the general approach one can use the simplest dual frame estimator called the screening estimator derived by Hartley (1962). The general idea is that the list frame is incomplete whereas the area frame is complete and thus, therefore, the overlap domain is the list frame whereas the non-overlap domain consists of those members of the area frame which are not on the list frame. Therefore, assuming simple random sampling in each frame, an estimate of the population total would be the sum of the population estimates for the two domains:

$$\hat{\tau} = \hat{\tau}_{OL} + \hat{\tau}_{NOL}.$$

For the overlap domain one uses the usual population total estimator for the list frame:

$$\hat{\tau}_{OL} = N_L \bar{y}_{OL},$$

where N_L is the number of elements on the list frame, and \bar{y}_{OL} is the sample mean based on the list frame. All units are screened out from the area frame that are on the list frame and only the remaining units are used so that:

$$\hat{\tau}_{NOL} = N_A \bar{y}_{NOL},$$

where here N_A is the number of area units in the area frame, and \bar{y}_{NOL} is the sample mean for a sample of n_A units from that area frame where all elements on the list frame are screened out.

From the way the estimate is constructed the two estimates are independent and, therefore, the variance of the overall estimate is just the sum of the two variances. In surveys like the National Farm Survey more complex sampling designs would need to be used but the general approach can easily be generalized.

Examples of Dual and Multiple Frame Sampling

Farm Survey

The dual frame approach finds one of its widest uses in agricultural surveys where the objective is to study farms as business enterprises and unfortunately a complete list of all farms is not available. Smaller farms tend to be the ones missed so that use of the incomplete list frame would be inadvisable (Kott and Vogel 1995).

Telephone Surveys

For telephone surveys a random digit dialing (RDD) frame has complete coverage in the sense that all people with phones (listed and unlisted numbers) are reached. However, there are better designs which use two frames. With a dual frame design for telephone surveys, an incomplete list frame sample is selected from a set of numbers based on telephone directories, and then it is combined with a sample selected from a complete random digit dialing (RDD) frame. Why use the incomplete list frame at all? Well it may be a lot cheaper to sample and also its use does allow a personalized approach on the call because the name is known. This more personalized call can result in higher response rates especially if combined with a pre call letter. Dual frame designs for telephone surveys still have complete coverage but in addition offer the prospect of reduced non response error and lower cost relative to using the RDD frame alone. One important operational point is that to use the dual frame estimator one needs to establish whether a number included on the RDD sample is on the list or not. This is to delineate the overlap and nonoverlap domains discussed in the previous section.

A good reference for this dual frame approach in telephone surveys is Traugott et al. (1986). There are probably other more recent ones as well.

This approach is being considered for national telephone surveys of anglers. Here there is an incomplete frame list of anglers based on a license file and there is basically a complete frame based on random digit dialing. Currently they just do just a RDD survey (NRC 2006).

Wildlife Nest Surveys

Here it is possible to get an incomplete list of bald eagle nests based on past observations. For completeness some areas need to be searched for nests. Therefore this is another example of dual frame sampling (Haines 1997, Haines and Pollock 1998).

Homeless Survey

Iachan and Dennis (1993) describe the use of multiple frames to sample the homeless population in Washington DC (See also Lohr 1999 p.402). There were three frames used which were (A) homeless shelters, (B) soup kitchens and (C) encampments and streets. Membership in more than one frame was estimated by asking all respondents irrespective of where sampled about their use of all three places. (One would need to assign each respondent into nonoverlapping domains a, b, c, ab, ac, bc, and abc. This is more complex than the estimator for two frames given earlier). This approach should reach most homeless people but especially when they are only in encampments or streets (C) the homeless may refuse to be interviewed or hide from interviewers.

Extension: Multiple Incomplete List Frames but No Area Frame

Another approach to solving the problem of incomplete list frames when an area frame cannot easily be used is to use several incomplete list frames in a capture-recapture approach. This is discussed in many references. See for example Haines (1997). The capture-recapture approach has also been used as one solution for the related problem of the US Census Undercount. In my opinion the capture-recapture approach is usually not nearly as satisfactory as the dual frame solution described earlier. Having one complete frame (usually the area frame) and one incomplete frame is the key to its wide applicability and simplicity!

References

Haines, D.E. (1997). Estimating population parameters using multiple frame and capture-recapture methodology. Ph D Thesis, North Carolina State University.

- Haines, D.E. and Pollock, K.H. (1998). Estimating the number of active and successful bald eagle nests: An application of the dual frame method. *Journal of Environmental and Ecological Statistics*. 5, 245-256.
- Hansen, M. H., Hurwitz, W. N., and Madow, W. G. (1953). *Sample Survey Methods and Theory Volume 1*. John Wiley and Sons, New York.
- Hartley, H. O. (1962). Multiple frame surveys. *Proceedings of the Social Statistics Section American Statistical Association* 203-206.
- Hartley, H. O. (1974). Multiple frame methodology and selected applications. *Sankhya C* 36, 99-118.
- Iachan, R. and M. L. Dennis. (1993). A multiple frame approach to sampling the homeless and transient population. *Journal of Official Statistics* 9, 747-764.
- Kott, P. S. and Vogel, F. A. (1995). Multiple Frame Business Surveys. *In Business Survey Methods*. Eds Cox, B. G., Binder, D. A., Chinappa, B. N., Christianson, A., Colledge, M. J., and Kott, P. S. John Wiley and Sons, New York.
- Lohr, S. L. (1999) *Sampling Design and Analysis*, Duxbury Press, Pacific Grove, California. (p. 401-402).
- Lohr, S. L. and Rao J. N. K. (2000). Inference from Dual Frame Surveys. *Journal of the American Statistical Association*, 95, 271-280.
- NRC Report 2006. *Review of Recreational Fisheries Survey Methods*. Ocean Studies Board. National Academy Press, Washington D. C.
- Traugott, M. W., Groves, R. M., and Lepkowski, J. (1986). Using dual frame designs to reduce nonresponse in telephone surveys. *Public Opinion Quarterly* 51, 522-539.