Sparse matrix representations for the Horvitz–Thompson estimator

Thomas Lumley

May 17, 2009

Let R_i be the sampling indicator, $E[R_i] = \pi_i$, $E[R_i R_j] = \pi_{ij}$, $\check{x}_i = x_i/\pi_i$, and write the covariance of the sampling indicators as

$$\Delta_{ij} = \mathrm{cov}[R_i, R_j] = \pi_{ij} - \pi_i\pi_j$$

$$\check{\Delta}_{ij} = \frac{\Delta_{ij}}{\pi_{ij}} = 1 - \frac{\pi_i\pi_j}{\pi_{ij}}.$$

The Horvitz–Thompson estimator is

$$\hat{\sigma}^2 = \check{x}\check{\Delta}\check{x}^T$$

with computation time proportional to the number of entries of $\check{\Delta}$. After calibration the formula is

$$\hat{\sigma}^2 = (g\check{r})\check{\Delta}(g\check{r})^T,$$

where g is the calibration weight and r the calibration residual.

For cluster sampling from a stratum with n clusters, $\check{\Delta}_{ij} = (1 - \pi_i)$ for (i, j) in the same cluster and $\check{\Delta}_{ij} = -(1 - \pi_i)/(n_h - 1)$ for (i, j) in different clusters. Under sampling with replacement we simply set the between-cluster terms to zero. For stratified sampling $\check{\Delta}$ is zero for observations in different strata.

Under multistage sampling

$$1 - \check{\Delta} = (1 - \check{\Delta}_{(1)})(1 - \check{\Delta}_{2|1})$$

so

$$\check{\Delta} = \check{\Delta}_{(1)} + \check{\Delta}_{(2|1)} - \check{\Delta}_{(1)}\check{\Delta}_{(2|1)}.$$

It is easy to construct $\check{\Delta}$ recursively, using a similar algorithm to that used in computing the Horvitz–Thompson estimator.

This is also true for multi-phase sampling if we work with the observable probabilities $\pi^*_i = \pi_{1i} \times \pi_{2|\text{phase 1},i}$

$$1 - \check{\check{\Delta}}^* = (1 - \check{\Delta}_{(1)})(1 - \check{\Delta}_{2|\text{phase 1}})$$

so

$$\check{\check{\Delta}}^{*} = \check{\Delta}_{(1)} + \check{\Delta}_{(2|\text{phase 1})} - \check{\Delta}_{(1)}\check{\Delta}_{(2|\text{phase 1})}$$

The entries of $\check{\Delta}_{(1)}$ corresponding to observations that are not in phase 2 do not need to be computed.

The reason for choosing this representation of the joint sampling probabilities is that $\check{\Delta}_{ij} = 0$ when R_i and R_j are independent. In particular, the matrix $\check{\Delta}$ will be sparse when the first stage or phase of sampling is either with-replacement or highly stratified. With a sparse-matrix representation such as those provided by the Matrix package, storage and computation time will both be proportional to the number of non-zero entries of $\check{\Delta}$. Under general PPS or adaptive sampling there will be no saving in space or time, but there is still a gain in ease of programming.

In version 3.15, this algorithm is used for two-phase designs, allowing multistage sampling at each phase. The function `Dcheck_multi` computes $\check{\Delta}_2$ for phase 2 using the recursive formulation, and `Dcheck_multi_subset` computes the subset of $\check{\Delta}_1$ for the phase 1 design, for observations that end up in phase 2. On an ordinary laptop computer the computations are feasible for designs where the second phase has a few thousand observations.

In future versions this approach may be used to allow exact computation of the Horvitz–Thompson and Yates–Grundy estimators for PPS designs, and various approximations, and perhaps to support adaptive sampling.