Complex Samples: Covariance Matrix of Total

Introduction

This document describes the algorithms used in the complex sampling module procedures for estimation of covariance matrix of population total estimates. It contains a more general formulation of the algorithms given in CSDESCRIPTIVE and CSTABULATE.

Complex sample data must contain both the values of the variables to be analyzed and the information on the current sampling design. Sampling design includes the sampling method, strata and clustering information, inclusion probabilities and the overall sampling weights.

Sampling design specification may include up to three stages of sampling. Any of the following general sampling methods may be assumed in the first stage: random sampling with replacement, random sampling without replacement and equal probabilities and random sampling without replacement and unequal probabilities. The first two sampling methods can also be specified for the second and the third sampling stage.

Notations

H	Number of strata.
n_h	Sampled number of primary sampling units (PSU) per stratum.
f_h	Sampling rate per stratum.
m_{hi}	Number of elements in the i^{th} sampled unit in stratum h, $i = 1,\ldots,n_h$.
w_{hij}	Overall sampling weight for j^{th} element in the i^{th} sampled unit in stratum h.
$\mathbf{y}_{hij}$	Values of vector $\mathbf{y}$ for the j^{th} element in the i^{th} sampled unit in stratum h.
$\mathbf{y}_T$	Population total sum for vector of variables $\mathbf{y}$.
n	Total number of elements in the sample.
N	Total number of elements in the population.

Weights

Overall weights specified for each ultimate element are processed as given. They can be obtained as a product of weights for corresponding units computed in each sampling stage.

When sampling without replacement in a given stage, substitution $w_{hi} = 1/\pi_{hi}$ for unit i in stratum h will result in application of the estimator for the population totals due to Horvitz and Thompson (1952). Corresponding variance estimator (2) or (3) will also be unbiased. π_{hi} is the probability of unit i from stratum h being selected in the given stage.

If sampling with replacement in a given stage, substitution $w_{hi} = 1/(n_h p_{hi})$ yields the estimator for the population totals due to Hansen and Hurwitz (1943). Repeatedly selected units should be replicated in the data. Corresponding variance estimator (1) will be unbiased. p_{hi} is the probability of selecting unit i in a single draw from stratum h in the given stage.

Weights obtained in each sampling stage need to be multiplied when processing multi-stage samples. The resulting overall weights for the elements in the final stage are used in all expressions and formulas below.

Z expressions

$$\mathbf{z}_{hij} = w_{hij}\mathbf{y}_{hij}$$

$$\mathbf{z}_{hi} = \sum_{j=1}^{m_{hi}} \mathbf{z}_{hij}$$

$$\bar{\mathbf{z}}_h = \frac{1}{n_h}\sum_{i=1}^{n_h} \mathbf{z}_{hi}$$

$$\mathbf{S}_h^2(\mathbf{y}) = \frac{1}{n_h - 1}\sum_{i=1}^{n_h} (\mathbf{z}_{hi} - \bar{\mathbf{z}}_h)(\mathbf{z}_{hi} - \bar{\mathbf{z}}_h)'$$

For multi-stage samples, index h denotes a stratum in the given stage, and i stands for unit from h in the same stage. Index j runs over all final stage elements contained in unit hi.

Total estimation

An estimate for the population total of vector of variables $\mathbf{y}$ in a single-stage sample is the weighted sum over all the strata and all the clusters:

$$\hat{\mathbf{y}}_T = \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij}\mathbf{y}_{hij}$$

Alternatively, we compute the weighted sum over all the elements in the sample:

$$\hat{\mathbf{y}}_T = \sum_{i=1}^{n} w_i \mathbf{y}_i$$

The latter expression is more general as it also applies to multi-stages samples.

Total covariances

For a multi-stage sample containing a with replacement sampling stage, all specifications other than weights are ignored for the subsequent stages. They make no contribution to the variance estimates.

Single stage sample

Covariance of the total for vector $\mathbf{y}$ in a single-stage sample is estimated by the following:

$$\hat{\mathbf{V}}(\hat{\mathbf{y}}_T) = \hat{\mathbf{V}}_1(\hat{\mathbf{y}}_T) = \sum_{h=1}^{H} \mathbf{U}_h(\hat{\mathbf{y}}_T)$$

where $\mathbf{U}_h(\hat{\mathbf{y}}_T)$ is an estimate contribution from stratum $h = 1, \ldots, H$ and depends on the sampling method as follows:

For sampling with replacement

$$\mathbf{U}_h(\hat{\mathbf{y}}_T) = n_h \mathbf{S}_h^2(\mathbf{y}) \qquad \textbf{(1)}$$

For simple random sampling

$$\mathbf{U}_h(\hat{\mathbf{y}}_T) = (1 - f_h) n_h \mathbf{S}_h^2(\mathbf{y}) \qquad \textbf{(2)}$$

For sampling without replacement and unequal probabilities

$$\mathbf{U}_h(\hat{\mathbf{y}}_T) = \sum_{i=1}^{n_h} \sum_{i>j}^{n_h} \left(\frac{\pi_{hi}\pi_{hj}}{\pi_{hij}} - 1\right)(\mathbf{z}_{hi} - \mathbf{z}_{hj})(\mathbf{z}_{hi} - \mathbf{z}_{hj})' \qquad \textbf{(3)}$$

In the variance estimator (3), π_{hi} and π_{hj} are the inclusion probability for units i and j in stratum h, and π_{hij} is the joint inclusion probability for the same units. This estimator is due to Yates and Grundy (1953) and Sen (1953). In some situations it may yield a negative estimate and is treated as undefined.

For each stratum h containing a single element, the covariance contribution $\mathbf{U}_h(\hat{\mathbf{y}}_T)$ is always set to zero.

Two-stage sample

When the sample is obtained in two stages and sampling without replacement is applied in the first stage, we use the following estimate for the covariance of the total for vector $\mathbf{y}$:

$$\hat{\mathbf{V}}(\hat{\mathbf{y}}_T) = \hat{\mathbf{V}}_2(\hat{\mathbf{y}}_T) = \hat{\mathbf{V}}_1(\hat{\mathbf{y}}_T) + \sum_{h=1}^{H}\sum_{i=1}^{n_h}\pi_{hi}\sum_{k=1}^{K_{hi}}\mathbf{U}_{hik}(\hat{\mathbf{y}}_T).$$

π_{hi} is the first stage inclusion probability for the primary sampling unit i in stratum h. In case of simple random sampling, the inclusion probability is equal to the sampling rate f_h for stratum *h*.

K_{hi} is the number of second stage strata in the primary sampling unit i within the first stage stratum *h*.

$\mathbf{U}_{hik}(\hat{\mathbf{y}}_T)$ is a covariance contribution from the second stage stratum k from the primary sampling unit hi. It depends on the second stage sampling method. Corresponding formula (1) or (2) applies.

Three-stage sample

When the sample is obtained in three stages where sampling in the first stage is done without replacement and simple random sampling is applied in the second stage, we use the following estimate for the covariance of the total for vector $\mathbf{y}$:

$$\hat{\mathbf{V}}(\hat{\mathbf{y}}_T) = \hat{\mathbf{V}}_2(\hat{\mathbf{y}}_T) + \sum_{h=1}^{H}\sum_{i=1}^{n_h}\pi_{hi}\sum_{k=1}^{K_{hi}}f_{hik}\sum_{j=1}^{n_{hik}}\sum_{l=1}^{L_{hikj}}\mathbf{U}_{hikjl}(\hat{\mathbf{y}}_T)$$

f_{hik} is the sampling rate for the secondary sampling units in the second stage stratum hik.

L_{hikj} is the number of the third stage strata in the secondary sampling unit $hikj$.

$\mathbf{U}_{hikjl}(\hat{\mathbf{y}}_T)$ is a covariance contribution from the third stage stratum l contained in the secondary sampling unit $hikj$. It depends on the third stage sampling method. Corresponding formula (1) or (2) applies.

Total variance

Variance of the total estimate for the r^{th} element of the vector $\hat{\mathbf{y}}_T$, is estimated by the r^{th} diagonal element of the covariance matrix for $\hat{\mathbf{y}}_T$

$$\hat{V}((\hat{\mathbf{y}}_T)_r) = \hat{\mathbf{V}}(\hat{\mathbf{y}}_T)_{rr}.$$

Population Size Estimation

An estimate for the population size corresponds to the estimate for the variable total; it is sum of the sampling weights. We have the following estimate for the single-stage samples:

$$\hat{N} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij} \ .$$

More generally,

$$\hat{N} = \sum_{i=1}^{n} w_i \ .$$

Variance of $\hat{N}$ is obtained by replacing y_{hij} with 1, i.e. by replacing z_{hij} with w_{hij} in the corresponding variance estimator formula for $\hat{V}(\hat{y}_T)$.

References

Hansen, M. H., and Hurwitz, W. N. (1943), "On the theory of sampling from finite populations", Annals of Mathematical Statistics, volume 14, pages 333 - 362.

Horvitz, D. G., and Thompson, D. J. (1952), "A generalization of sampling without replacement from a finite universe", Journal of the American Statistical Association, volume 47, pages 663 - 685.

Särndal, C. E., Swenson, B., and Wretman, J. H. (1992), Model Assisted Survey Sampling, New York: Springer-Verlag.

Sen, A. R. (1953), "On the estimate of the variance in sampling with varying probabilities", Journal of the Indian Society of Agricultural Statistics, volume 5, pages 55 - 77.

Yates, F., and Grundy, P. M. (1953), "Selection without replacement from within strata with probability proportional to size", *Journal of the Royal Statistical Society Series B*, volume 15, pages 253 - 261.