# CSDESCRIPTIVES

This document describes the algorithms used in the complex sampling estimation procedure CSDESCRIPTIVES. The data do not have to be sorted.

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, and inclusion probabilities for all units at every sampling stage. The overall sampling weight must be specified for each observation.

The sampling design specification for CSDESCRIPTIVES 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.

## Notation

The following notation is used throughout this chapter unless otherwise stated:

| | |
|---|---|
| $H$ | 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 $h$, $i = 1,\ldots,n_h$. |
| $y_{hij}$ | Value of variable $y$ for the $j^{th}$ element in the $i^{th}$ sampled unit in stratum $h$. |
| $w_{hij}$ | Overall sampling weight for $j^{th}$ element in the $i^{th}$ sampled unit in stratum $h$. |
| $n$ | Total number of elements in the sample. |
| $N$ | Total number of elements in the population. |
| $Y$ | Population total sum for variable $y$. |

### 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, substituting $w_{hi} = 1/\pi_{hi}$ for unit $i$ in stratum $h$ results in the application of the estimator for the population totals due to Horvitz and Thompson (1952). The corresponding variance estimator (2) or (3) will also be unbiased. $\pi_{hi}$ is the probability of unit $i$ from stratum $h$ being selected in the given stage.

If sampling with replacement in a given stage, substituting $w_{hi} = 1/(n_h p_{hi})$ yields the estimator for the population totals due to Hansen and Hurwitz (1943). Repeatedly selected

units should be replicated in the data. The 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

$$z_{hij} = w_{hij} y_{hij}$$

$$z_{hi} = \sum_{j=1}^{m_{hi}} z_{hij}$$

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

$$S_h^2 = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (z_{hi} - \bar{z}_h)^2$$

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

## Variable Total

An estimate for the population total of variable $y$ in a single-stage sample is the weighted sum over all the strata and all the clusters:

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

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

$$\hat{Y} = \sum_{i=1}^{n} w_i y_i$$

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

# Variable Total Variance

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

The variance of the total for variable $y$ in a single-stage sampling is estimated by the following:

$$\hat{V}(\hat{Y}) = \hat{V}_1(\hat{Y}) = \sum_{h=1}^{H} U_h$$

where $U_h$ is an estimated contribution from stratum $h = 1, \ldots, H$ and depends on the sampling method as follows:

- For sampling with replacement

$$U_h = n_h S_h^2 \quad \textbf{(1)}$$

- For simple random sampling

$$U_h = (1 - f_h) n_h S_h^2 \quad \textbf{(2)}$$

- For sampling without replacement and unequal probabilities

$$U_h = \sum_{i=1}^{n_h} \sum_{i>j}^{n_h} \left(\frac{\pi_{hi}\pi_{hj}}{\pi_{hij}} - 1\right)(z_{hi} - z_{hj})^2 \quad \textbf{(3)}$$

In the variance estimator (3), $\pi_{hi}$ and $\pi_{hj}$ are the inclusion probabilities for units $i$ and $j$ in stratum $h$, and $\pi_{hij}$ is the joint inclusion probability for the same units. This estimator is due to Yates and Grundy (1953) and Sen (1953).

For each stratum $h$ containing a single element, the variance contribution $U_h$ 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 variance of the total for variable $y$:

$$\hat{V}(\hat{Y}) = \hat{V}_2(\hat{Y}) = \hat{V}_1(\hat{Y}) + \sum_{h=1}^{H}\sum_{i=1}^{n_h}\pi_{hi}\sum_{k=1}^{K_{hi}}U_{hik}$$

where

- $\pi_{hi}$ is the first stage inclusion probability for the primary sampling unit $i$ in stratum $h$. In the 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*.
- $U_{hik}$ is a variance contribution from the second stage stratum $k$ from the primary sampling unit $hi$. It depends on the second stage sampling method. The 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 variance of the total for variable $y$:

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

where

- $f_{hik}$ is the sampling rate for the secondary sampling units in the second stage stratum $hik$.
- $L_{hikj}$ is the number of third stage strata in the secondary sampling unit $hikj$.
- $U_{hikjl}$ is a variance 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.

## 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,

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

The variance of $\hat{N}$ is obtained by replacing $y_{hij}$ with 1; that is, by replacing $z_{hij}$ with $w_{hij}$ in the corresponding variance estimator formula for $\hat{V}(\hat{Y})$.

# Ratio and Mean Estimation

Let $R = Y/X$ be the ratio of the totals for variables $y$ and $x$. It is estimated by

$$\hat{R} = \hat{Y}/\hat{X}$$

where $\hat{Y}$ and $\hat{X}$ are the estimates for the corresponding variable totals.

The variance of $\hat{R}$ is approximated using the Taylor linearization formula following Woodruff (1971). The estimate for the approximate variance of the ratio estimate $\hat{V}(\hat{R})$ is obtained by replacing $z_{hij}$ with

$$z_{hij} = w_{hij}(y_{hij} - \hat{R}x_{hij})/\hat{X}$$

in the corresponding variance estimator $\hat{V}(\hat{Y})$.

## Mean Estimation

The mean $\bar{Y}$ for the variable $y$ is estimated by

$$\hat{\bar{Y}} = \hat{Y}/\hat{N}$$

where $\hat{Y}$ is the estimate for the total of $y$ and $\hat{N}$ is the population size estimate.

The variance of the mean is estimated using the ratio formulas, as the mean is a ratio of $\hat{Y}$ and $\hat{N}$. Accordingly, $\hat{V}(\hat{\bar{Y}})$ is obtained by substituting $z_{hij}$ with

$$z_{hij} = w_{hij}(y_{hij} - \hat{\bar{Y}})/\hat{N}$$

in the corresponding variance estimator $\hat{V}(\hat{Y})$.

# Domain Estimation

Let the population be divided into $D$ domains. For each domain $d = 1,\ldots,D$ define the following indicator variables:

$$\delta_{hij}(d) = \begin{cases} 1 & \text{if the sample unit } hij \text{ is in the domain } d \\ 0 & \text{otherwise} \end{cases}$$

To estimate a domain population total, domain variable total, ratios and means, substitute $y_i$ with $\delta_i(d)y_i$ in the corresponding formula for the whole population as follows:

- Domain variable total

$$\hat{Y}_d = \sum_{i=1}^{n} w_i \delta_i(d) y_i \ .$$

- Domain population total

$$\hat{N}_d = \sum_{i=1}^{n} w_i \delta_i(d) \ .$$

- Domain variable ratio

$$\hat{R}_d = \hat{Y}_d \big/ \hat{X}_d$$

- Domain variable mean

$$\hat{\bar{Y}}_d = \hat{Y}_d \big/ \hat{N}_d$$

Similarly, in order to estimate the variances of the above estimators, substitute $y_{hij}$ with $\delta_{hij}(d)y_{hij}$ in the corresponding formula for the whole population. The following substitution of $z_{ij}$ in the formulas for $\hat{V}(\hat{Y})$ are used for estimating the variance of:

- Domain variable total

$$z_{hij}(d) = \delta_{hij}(d) w_{hij} y_{hij}$$

- Domain population total

$$z_{hij}(d) = \delta_{hij}(d) w_{hij}$$

- Domain variable ratio

$$z_{hij} = \delta_{hij}(d) w_{hij} (y_{hij} - \hat{R}_d x_{hij}) \Big/ \hat{X}_d$$

- Domain mean

$$z_{hij} = \delta_{hij}(d) w_{hij} (y_{hij} - \hat{\bar{Y}}_d) \Big/ \hat{N}_d$$

# Standard Errors

Let $Z$ denote any of the population or subpopulation quantities defined above: variable total, population size, ratio or mean. Then the standard error of an estimator $\hat{Z}$ is the square root of its estimated variance:

$$StdError(\hat{Z}) = \sqrt{\hat{V}(\hat{Z})}\,.$$

# Coefficient of variation

The coefficient of variation of the estimator $\hat{Z}$ is the ratio of its standard error and its value:

$$CV(\hat{Z}) = \frac{SE(\hat{Z})}{\hat{Z}}.$$

The coefficient of variation is undefined when $\hat{Z} = 0$.

# t Tests

Testing the hypothesis that a population quantity $Z$ equals $\theta_0$, i.e. $H_0 : Z = \theta_0$ is performed using the $t$ test statistic:

$$t(\hat{Z}) = \frac{\hat{Z} - \theta_0}{StdError(\hat{Z})}.$$

The *p*-value for the two-sided test is given by the probability

$$P(|T| > |t(\hat{Z})|)$$

where $T$ is a random variable form the $t$ distribution with $df$ degrees of freedom.

### Degrees of freedom

The number of the degrees of freedom $df$ for the $t$ distribution is calculated as the difference between the number of primary sampling units and the number of strata in the first stage of sampling.

## Confidence Limits

A level $1-\alpha$ confidence interval is constructed for a given $0 \le \alpha \le 1$. The confidence bounds are defined as

$$\hat{Z} \pm StdError(\hat{Z}) t_{df}(1-\alpha/2)$$

where $StdError(\hat{Z})$ is the estimated standard error of $\hat{Z}$, and $t_{df}(1-\alpha/2)$ is the $100(1-\alpha/2)$ percentile of the $t$ distribution with $df$ degrees of freedom.

## Design Effects

The design effect $Deff$ is estimated by

$$Deff = \frac{\hat{V}(\hat{Y})}{\hat{V}_{srs}(\hat{Y}_{srs})}$$

$\hat{V}(\hat{Y})$ is the estimate of variance of $\hat{Y}$ under the appropriate sampling design, while $\hat{V}_{srs}(\hat{Y}_{srs})$ is the estimate of variance of $\hat{Y}_{srs}$ under the simple random sampling assumption as follows:

$$\hat{V}_{srs}(\hat{Y}_{srs}) = (1-\frac{n}{\hat{N}})\frac{\hat{N}}{n-1}\sum_{i=1}^{n} w_i (y_i - \frac{\hat{Y}}{\hat{N}})^2$$

$Deff$ is undefined when $\frac{n}{\hat{N}} \ge 1$.

Whereas design effect is not relevant for estimates of the population size, we do compute the design effects for ratios and means in addition to the totals. The values of variable $y$ in $\hat{V}_{srs}$ are then replaced by the linearized values as follows:

- Ratio estimation

$$(y_i - \hat{R}x_i)/\hat{X}$$

- Mean estimation

$$(y_i - \hat{\bar{Y}}) \Big/ \hat{N}$$

When estimating design effects for domains we use the familiar substitution $\delta_i(d) y_i$ for $y_i$ in $\hat{V}_{srs}$ formula in addition to any ratio or mean substitutions.

We also provide the square root of design effect $Defft$ by computing

$$Defft = \sqrt{Deff} \ .$$

Design effects and their applications have been discussed by Kish (1965) and Kish (1995).

# References


Cochran, W. G. (1977), *Sampling Techniques*, Third Edition, New York: John Wiley & Sons.

Hansen, M. H., Hurwitz, W. N., and Madow, W. G. (1953), *Sample Survey Methods and Theory*, Volume II Theory, New York: John Wiley & Sons.

Hansen, M. H., Hurwitz, W. N., and Madow, W. G. (1953), *Sample Survey Methods and Theory*, Volume II Theory, New York: John Wiley & Sons.

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.

Kish, L. (1965), *Survey Sampling*, New York: John Wiley & Sons.

Kish, L. (1995), "Methods for Design Effects", *Journal of Official Statistics*, volume 11, pages 119 - 127.

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.

Wolter, K. M. (1985), *Introduction to Variance Estimation*, New York: Sringer-Verlag.

Woodruff, R. S. (1971), "A Simple Method for Approximating the Variance of a Complicated Estimate," *Journal of the American Statistical Association*, volume 66, pages 411 – 414.

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.