Previous Page | Next Page

Definitions and Notation

For a stratified clustered sample design, together with the sampling weights, the sample can be represented by an $n \times (P+1)$ matrix

$$\begin{aligned}(\mathbf{w},\mathbf{Y}) &= (w_{hij},\mathbf{y}_{hij}) \\ &= \left(w_{hij},y_{hij}^{(1)},y_{hij}^{(2)},\ldots,y_{hij}^{(P)}\right)\end{aligned}$$

where

- $h=1,2,\ldots,H$ is the stratum index
- $i=1,2,\ldots,n_h$ is the cluster index within stratum h
- $j=1,2,\ldots,m_{hi}$ is the unit index within cluster i of stratum h
- $p=1,2,\ldots,P$ is the analysis variable number, with a total of P variables
- $n=\sum_{h=1}^{H}\sum_{i=1}^{n_h} m_{hi}$ is the total number of observations in the sample
- w_{hij} denotes the sampling weight for unit j in cluster i of stratum h
- $\mathbf{y}_{hij}=\left(y_{hij}^{(1)},y_{hij}^{(2)},\ldots,y_{hij}^{(P)}\right)$ are the observed values of the analysis variables for unit j in cluster i of stratum h, including both the values of numerical variables and the values of indicator variables for levels of categorical variables.

For a categorical variable C, let l denote the number of levels of C, and denote the level values as $c_1,c_2,\ldots,c_l$. Let $y^{(q)}\,(q\in\{1,2,\ldots,P\})$ be an indicator variable for the category $C=c_k\,(k=1,2,\ldots,l)$ with the observed value in unit j in cluster i of stratum h:

$$y_{hij}^{(q)}=I_{\{C=c_k\}}(h,i,j)=\begin{cases}1 & \text{if } C_{hij}=c_k\\ 0 & \text{otherwise}\end{cases}$$

Note that the indicator variable $y_{hij}^{(q)}$ is set to missing when C_{hij} is missing. Therefore, the total number of analysis variables, P, is the total number of numerical variables plus the total number of levels of all categorical variables.

The sampling rate f_h for stratum h, which is used in Taylor series variance estimation, is the fraction of first-stage units (PSUs) selected for the sample. You can use the TOTAL= or RATE= option to input population totals or sampling rates. See the section Specification of Population Totals and Sampling Rates for details. If you input stratum totals, PROC SURVEYMEANS computes f_h as the ratio of the stratum sample size to the stratum total. If you input stratum sampling rates, PROC SURVEYMEANS uses these values directly for f_h. If you do not specify the TOTAL= or RATE= option, then the procedure assumes that the stratum sampling rates f_h are negligible, and a finite population correction is not used when computing variances. Replication methods specified by the VARMETHOD=BRR or the VARMETHOD=JACKKNIFE option do not use this finite population correction f_h.

Previous Page | Next Page

Mean

When you specify the keyword MEAN, the procedure computes the estimate of the mean (mean per element) from the survey data. Also, the procedure computes the mean by default if you do not specify any ***statistic-keywords*** in the PROC SURVEYMEANS statement.

PROC SURVEYMEANS computes the estimate of the mean as

$$\widehat{\bar{Y}} = \left(\sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij}\, y_{hij} \right) / w_{...}$$

where

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

is the sum of the weights over all observations in the sample.

Previous Page | Next Page | Top of Page

Previous Page | Next Page

Variance and Standard Error of the Mean

When you specify the keyword STDERR, the procedure computes the standard error of the mean. Also, the procedure computes the standard error by default if you specify the keyword MEAN, or if you do not specify any ***statistic-keywords*** in the PROC SURVEYMEANS statement. The keyword VAR requests the variance of the mean.

Taylor Series Method

When you use VARMETHOD=TAYLOR, or by default if you do not specify the VARMETHOD= option, PROC SURVEYMEANS uses the Taylor series method to estimate the variance of the mean $\bar{\hat{Y}}$. The procedure computes the estimated variance as

$$\hat{V}(\bar{\hat{Y}}) = \sum_{h=1}^{H} \hat{V}_h(\bar{\hat{Y}})$$

where if $n_h > 1$,

$$\hat{V}_h(\bar{\hat{Y}}) = \frac{n_h(1-f_h)}{n_h-1} \sum_{i=1}^{n_h} (e_{hi\cdot} - \bar{e}_{h\cdot\cdot})^2$$

$$e_{hi\cdot} = \left(\sum_{j=1}^{m_{hi}} w_{hij} \, (y_{hij} - \bar{\hat{Y}}) \right) / \, w_{\cdots}$$

$$\bar{e}_{h\cdot\cdot} = \left(\sum_{j=1}^{n_h} e_{hi\cdot} \right) / \, n_h$$

and if $n_h = 1$,

$$\hat{V}_h(\bar{\hat{Y}}) = \begin{cases} \text{missing} & \text{if } n_{h'} = 1 \text{ for } h' = 1,2,\ldots,H \\ 0 & \text{if } n_{h'} > 1 \text{ for some } 1 \le h' \le H \end{cases}$$

Replication Methods

When you specify VARMETHOD=BRR or VARMETHOD=JACKKNIFE, the procedure computes the variance $\hat{V}(\bar{\hat{Y}})$ with replication methods by using the variability among replicate estimates to estimate the overall variance. See the section Replication Methods for Variance Estimation for more details.

Standard Error

The standard error of the mean is the square root of the estimated variance.

$$\text{StdErr}(\bar{\hat{Y}}) = \sqrt{\hat{V}(\bar{\hat{Y}})}$$

Previous Page | Next Page | Top of Page

Previous Page | Next Page

Ratio

When you use a RATIO statement, the procedure produces statistics requested by the ***statistic-keywords*** in the PROC SURVEYMEANS statement.

Suppose that you want to calculate the ratio of variable Y over variable X. Let x_{hij} be the value of variable X for the jth member in cluster i in the hth stratum.

The ratio of Y over X is

$$\widehat{R} = \frac{\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij}\, y_{hij}}{\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij}\, x_{hij}}$$

PROC SURVEYMEANS uses the Taylor series method to estimate the variance of the ratio $\widehat{R}$ as

$$\widehat{V}(\widehat{R}) = \sum_{h=1}^{H} \widehat{V}_h(\widehat{R})$$

where if $n_h > 1$,

$$\widehat{V}_h(\widehat{R}) = \frac{n_h(1-f_h)}{n_h - 1}\sum_{i=1}^{n_h}(g_{hi\cdot} - \bar{g}_{h\cdot\cdot})^2$$

$$g_{hi\cdot} = \frac{\sum_{j=1}^{m_{hi}} w_{hij}\,(y_{hij} - x_{hij}\widehat{R})}{\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij}\, x_{hij}}$$

$$\bar{g}_{h\cdot\cdot} = \left(\sum_{j=1}^{n_h} g_{hi\cdot}\right) / n_h$$

and if $n_h = 1$,

$$\widehat{V}_h(\widehat{R}) = \begin{cases} \text{missing} & \text{if } n_{h'} = 1 \text{ for } h' = 1, 2, \ldots, H \\ 0 & \text{if } n_{h'} > 1 \text{ for some } 1 \le h' \le H \end{cases}$$

The standard error of the ratio is the square root of the estimated variance:

$$\text{StdErr}(\widehat{R}) = \sqrt{\widehat{V}(\widehat{R})}$$

When the denominator for a ratio is zero, then the value of the ratio is displayed as '–Infty', 'Infty', or a missing value, depending on whether the numerator is negative, positive, or zero, respectively; and the corresponding internal value is the special missing value '.M', the special missing value '.I', or the usual missing value, respectively.

Previous Page | Next Page | Top of Page

Previous Page | Next Page

Replication Methods for Variance Estimation

Recently replication methods have gained popularity for estimating variances in complex survey data analysis. One reason for this popularity is the relative simplicity of replication-based estimates, especially for nonlinear estimators; another is that modern computational capacity has made replication methods feasible for practical survey analysis. For details see Lohr (1999); Wolter (1985); Rust (1985); Dippo, Fay, and Morganstein (1984); Rao and Shao (1999); Rao, Wu, and Yue (1992); and Rao and Shao (1996).

Replication methods draw multiple replicates (also called subsamples) from a full sample according to a specific resampling scheme. The most commonly used resampling schemes are the ***balanced repeated replication*** (BRR) method and the ***jackknife*** method. For each replicate, the original weights are modified for the PSUs in the replicates to create replicate weights. The statistics of interest are estimated by using the replicate weights for each replicate. Then the variances of parameters of interest are estimated by the variability among the estimates derived from these replicates. You can use the REPWEIGHTS statement to provide your own replicate weights for variance estimation.

- **Balanced Repeated Replication (BRR) Method**
- **Fay's BRR Method**
- **Jackknife Method**
- **Hadamard Matrix**

Previous Page | Next Page | Top of Page

Previous Page | Next Page

Jackknife Method

The jackknife method of variance estimation deletes one PSU at a time from the full sample to create replicates. The total number of replicates R is the same as the total number of PSUs. In each replicate, the sample weights of the remaining PSUs are modified by the ***jackknife coefficient*** α_r. The modified weights are called replicate weights.

The jackknife coefficient and replicate weights are described as follows.
Without Stratification

If there is no stratification in the sample design (no STRATA statement), the jackknife coefficients α_r are the same for all replicates:

$$\alpha_r = \frac{R-1}{R} \text{ where } r = 1,2,...,R$$

Denote the original weight in the full sample for the jth member of the ith PSU as w_{ij}. If the ith PSU is included in the rth replicate ($r = 1,2,...,R$), then the corresponding replicate weight for the jth member of the ith PSU is defined as

$$w_{ij}^{(r)} = w_{ij}/\alpha_r$$

With Stratification

If the sample design involves stratification, each stratum must have at least two PSUs to use the jackknife method.

Let stratum $\tilde{h}_r$ be the stratum from which a PSU is deleted for the rth replicate. Stratum $\tilde{h}_r$ is called the ***donor stratum.*** Let $n_{\tilde{h}_r}$ be the total number of PSUs in the donor stratum $\tilde{h}_r$. The jackknife coefficients are defined as

$$\alpha_r = \frac{n_{\tilde{h}_r} - 1}{n_{\tilde{h}_r}} \text{ where } r = 1,2,...,R$$

Denote the original weight in the full sample for the jth member of the ith PSU as w_{ij}. If the ith PSU is included in the rth replicate ($r = 1,2,...,R$), then the corresponding replicate weight for the jth member of the ith PSU is defined as

$$w_{ij}^{(r)} = \begin{cases} w_{ij} & \text{if } i\text{th PSU is not in the donor stratum } \tilde{h}_r \\ w_{ij}/\alpha_r & \text{if } i\text{th PSU is in the donor stratum } \tilde{h}_r \end{cases}$$

You can use the VARMETHOD=JACKKNIFE(OUTJKCOEFS=) method-option to save the jackknife coefficients into a SAS data set and use the VARMETHOD=JACKKNIFE(OUTWEIGHTS=) method-option to save the replicate weights into a SAS data set.

If you provide your own replicate weights with a REPWEIGHTS statement, then you can also provide corresponding jackknife coefficients with the JKCOEFS= option.

Suppose that θ is a population parameter of interest. Let $\hat{\theta}$ be the estimate from the full sample for θ. Let $\hat{\theta}_r$ be the estimate from the rth replicate subsample by using replicate weights. PROC SURVEYMEANS estimates the variance of $\hat{\theta}$ by

$$\widehat{V}(\hat{\theta}) = \sum_{r=1}^{R} \alpha_r (\hat{\theta}_r - \hat{\theta})^2$$

with $R-H$ degrees of freedom, where R is the number of replicates and H is the number of strata, or $R-1$ when there is no stratification.

Previous Page | Next Page | Top of Page