The SURVEYMEANS Procedure

Overview

The SURVEYMEANS procedure produces estimates of survey population means and totals from sample survey data. The procedure also produces variance estimates, confidence limits, and other descriptive statistics. When computing these estimates, the procedure takes into account the sample design used to select the survey sample. The sample design can be a complex survey sample design with stratification, clustering, and unequal weighting.

PROC SURVEYMEANS uses the Taylor expansion method to estimate sampling errors of estimators based on complex sample designs. This method obtains a linear approximation for the estimator and then uses the variance estimate for this approximation to estimate the variance of the estimate itself (Woodruff 1971, Fuller 1975). When there are clusters, or primary sampling units (PSUs), in the sample design, the procedure estimates variance from the variation among PSUs. When the design is stratified, the procedure pools stratum variance estimates to compute the overall variance estimate.

PROC SURVEYMEANS uses the Output Delivery System (ODS) to place results in output data sets. This is a departure from older SAS procedures that provide OUTPUT statements for similar functionality.

Previous | Next | Top of Page

The SURVEYMEANS Procedure

PROC SURVEYMEANS Statement

PROC SURVEYMEANS < *options* > < *statistic-keywords* > **;**

The PROC SURVEYMEANS statement invokes the procedure. In this statement, you identify the data set to be analyzed and specify sample design information. The DATA= option names the input data set to be analyzed. If your analysis includes a finite population correction factor, you can input either the sampling rate or the population total using the RATE= or TOTAL= option. If your design is stratified, with different sampling rates or totals for different strata, then you can input these stratum rates or totals in a SAS data set containing the stratification variables.

In the PROC SURVEYMEANS statement, you also can use statistic-keywords to specify statistics for the procedure to compute. Available statistics include the population mean and population total, together with their variance estimates and confidence limits. You can also request data set summary information and sample design information.

You can specify the following options in the PROC SURVEYMEANS statement:

ALPHA=α
sets the confidence level for confidence limits. The value of the ALPHA= option must be between 0 and 1, and the default value is 0.05. A confidence level of α produces $100(1-\alpha)$ % confidence limits. The default of ALPHA=0.05 produces 95% confidence limits.

DATA=*SAS-data-set*
specifies the SAS data set to be analyzed by PROC SURVEYMEANS. If you omit the DATA= option, the procedure uses the most recently created SAS data set.

MISSING
requests that the procedure treat missing values as a valid category for all categorical variables, which include categorical analysis variables, strata variables, cluster variables, and domain variables.

ORDER=DATA | FORMATTED | INTERNAL
specifies the order in which the values of the categorical variables are to be reported. The following shows how PROC SURVEYMEANS interprets values of the ORDER= option:

DATA
orders values according to their order in the input data set.

FORMATTED
orders values by their formatted values. This order is operating environment dependent. By default, the order is ascending.

INTERNAL
orders values by their unformatted values, which yields the same order that the SORT procedure does. This order is operating environment dependent.

By default, ORDER=FORMATTED.

The ORDER= option applies to all the categorical variables. When the default ORDER=FORMATTED is in effect for numeric variables for which you have supplied no explicit format, the levels are ordered by their internal values.

RATE=*value* | *SAS-data-set*
R=*value* | *SAS-data-set*
specifies the sampling rate as a nonnegative *value*, or names an input data set that contains the stratum sampling rates. The procedure uses this information to compute a finite population correction for variance estimation. If your sample design has multiple stages, you should specify the *first-stage sampling rate*, which is the ratio of the number of PSUs selected to the total number of PSUs in the population.

For a nonstratified sample design, or for a stratified sample design with the same sampling rate in all strata, you should specify a nonnegative *value* for the RATE= option. If your design is stratified with different sampling rates in the strata, then you should name a SAS data set that contains the stratification variables and the sampling rates. See the section "Specification of Population Totals and Sampling Rates" for more details.

The sampling rate *value* must be a nonnegative number. You can specify *value* as a number between 0 and 1. Or

you can specify *value* in percentage form as a number between 1 and 100, and PROC SURVEYMEANS will convert that number to a proportion. The procedure treats the value 1 as 100%, and not the percentage form 1%.

If you do not specify the TOTAL= option or the RATE= option, then the variance estimation does not include a finite population correction. You cannot specify both the TOTAL= option and the RATE= option.

STACKING
requests the procedure to produce the output data sets using a stacking table structure, which was the default in releases prior to Version 9. The new default is to produce a rectangular table structure in the output data sets.

The STACKING option affects the following tables:

- Domain
- Ratio
- Statistics
- StrataInfo

When you use the ODS statement to create SAS data sets for these tables in the output, the data set structure can be either stacking or rectangular. A rectangular structure creates one observation for each analysis variable in the data set. However, if you use the STACKING option in Version 9, the procedure creates only one observation in the output data set for all analysis variables. The following example shows these two structures in output data sets.

```
data new;
  input sex$ x;
  datalines;
M 12
F 5
M 13
F 23
F 11
;

proc surveymeans data=new mean;
  ods output statistics=rectangle;
  run;

proc print data=rectangle;
  run;

proc surveymeans data=new mean stacking;
  ods output statistics=stacking;
  run;

proc print data=stacking;
  run;
```

Figure 70.6 shows the rectangular structure of the output data set for the statistics table.

rectangle structure in the output data set

OBS	VarName	VarLevel	Mean	StdErr
1	x		12.800000	2.905168
2	sex	F	0.600000	0.244949
3	sex	M	0.400000	0.244949

Figure 70.6: Rectangular Structure in the Output Data Set

Figure 70.7 shows the stacking structure of the output data set for the statistics table.

stacking structure in the output data set

OBS	x	x_Mean	x_StdErr	sex_F	sex_F_Mean	sex_F_StdErr	sex_M	sex_M_Mean	sex_M_StdErr
1	x	12.800000	2.905168	sex=F	0.600000	0.244949	sex=M	0.400000	0.244949

Figure 70.7: Stacking Structure in the Output Data Set

TOTAL=*value*| *SAS-data-set*

N=*value*| *SAS-data-set*

specifies the total number of primary sampling units (PSUs) in the study population as a positive *value*, or names an input data set that contains the stratum population totals. The procedure uses this information to compute a finite population correction for variance estimation.

For a nonstratified sample design, or for a stratified sample design with the same population total in all strata, you should specify a positive *value* for the TOTAL= option. If your sample design is stratified with different population totals in the strata, then you should name a SAS data set that contains the stratification variables and the population totals. See the section "Specification of Population Totals and Sampling Rates" for more details.

If you do not specify the TOTAL= option or the RATE= option, then the variance estimation does not include a finite population correction. You cannot specify both the TOTAL= option and the RATE= option.

statistic-keywords

specifies the statistics for the procedure to compute. If you do not specify any statistic-keywords, PROC SURVEYMEANS computes the NOBS, MEAN, STDERR, and CLM statistics by default.

The statistics produced depend on the type of the analysis variable. If you name a numeric variable in the CLASS statement, then the procedure analyzes that variable as a categorical variable. The procedure always analyzes character variables as categorical. See the section "CLASS Statement" for more information.

PROC SURVEYMEANS computes MIN, MAX, and RANGE for numeric variables but not for categorical variables. For numeric variables, the keyword MEAN produces the mean, but for categorical variables it produces the proportion in each category or level. Also for categorical variables, the keyword NOBS produces the number of observations for each variable level, and the keyword NMISS produces the number of missing observations for each level. If you request the keyword NCLUSTER for a categorical variable, PROC SURVEYMEANS displays for each level the number of clusters with observations in that level. PROC SURVEYMEANS computes SUMWGT in the same way for both categorical and numeric variables, as the sum of the weights over all nonmissing observations.

PROC SURVEYMEANS performs univariate analysis, analyzing each variable separately. Thus the number of nonmissing and missing observations may not be the same for all analysis variables. See the section "Missing Values" for more information.

If you use the keyword RATIO without the keyword MEAN, the keyword MEAN is implied.

Other available statistics computed for a ratio are N, NCLU, SUMWGT, RATIO, STDERR, DF, T, PROBT, and CLM, as listed below. If no statistics are requested, the procedure will compute the ratio and its standard error by default for a RATIO statement.

The valid statistic-keywords are as follows:

ALL
all statistics listed

CLM
$100(1-\alpha)$ % two-sided confidence limits for the MEAN, where α is determined by the ALPHA= option, and the default is $\alpha = 0.05$

CLSUM
$100(1-\alpha)$ % two-sided confidence limits for the SUM, where α is determined by the ALPHA= option, and the default is $\alpha = 0.05$

CV
coefficient of variation for MEAN

CVSUM
coefficient of variation for SUM

DF
degrees of freedom for the *t* test

LCLM
$100(1-\alpha)$ % one-sided lower confidence limit of the MEAN, where α is determined by the ALPHA= option, and the default is $\alpha = 0.05$

LCLMSUM
$100(1-\alpha)$ % one-sided lower confidence limit of the SUM, where α is determined by the ALPHA= option, and the default is $\alpha = 0.05$

MAX
maximum value

MEAN
mean for a numeric variable, or the proportion in each category for a categorical variable

MIN
minimum value

NCLUSTER
number of clusters

NMISS
number of missing observations

NOBS
number of nonmissing observations

RANGE
range, MAX-MIN

RATIO
ratio of means or proportions

STD
standard deviation of the SUM. When you request SUM, the procedure computes STD by default.

STDERR
standard error of the MEAN or RATIO. When you request MEAN or RATIO, the procedure computes STDERR by default.

SUM
weighted sum, $\sum w_i y_i$, or estimated population total when the appropriate sampling weights are used

SUMWGT
sum of the weights, $\sum w_i$

T
t-value and its corresponding *p*-value with DF degrees of freedom for
$H_0 : \theta = 0$
where θ is the population mean or the population ratio

UCLM
$100(1-\alpha)$ % one-sided upper confidence limit of the MEAN, where α is determined by the ALPHA= option, and the default is $\alpha = 0.05$

UCLMSUM
$100(1-\alpha)$ % one-sided upper confidence limit of the SUM, where α is determined by the ALPHA= option, and the default is $\alpha = 0.05$

VAR
variance of the MEAN or RATIO

VARSUM
variance of the SUM

See the section "Statistical Computations" for details on how PROC SURVEYMEANS computes these statistics.

Previous | Next | Top of Page

CLUSTER Statement

CLUSTER | CLUSTERS *variables* **;**

The CLUSTER statement names variables that identify the clusters in a clustered sample design. The combinations of categories of CLUSTER variables define the clusters in the sample. If there is a STRATA statement, clusters are nested within strata.

If your sample design has clustering at multiple stages, you should identify only the first-stage clusters, or primary sampling units (PSUs), in the CLUSTER statement. See the section "Primary Sampling Units (PSUs)" for more information.

The CLUSTER *variables* are one or more variables in the DATA= input data set. These variables can be either character or numeric. The formatted values of the CLUSTER variables determine the CLUSTER variable levels. Thus, you can use formats to group values into levels. Refer to the discussion of the FORMAT procedure in the *SAS Procedures Guide* and to the discussions of the FORMAT statement and SAS formats in *SAS Language Reference: Dictionary*.

You can use multiple CLUSTER statements to specify cluster variables. The procedure uses variables from all CLUSTER statements to create clusters.

Previous | Next | Top of Page

STRATA Statement

STRATA | **STRATUM** *variables* < / *option* > **;**

The STRATA statement names variables that form the strata in a stratified sample design. The combinations of categories of STRATA variables define the strata in the sample.

If your sample design has stratification at multiple stages, you should identify only the first-stage strata in the STRATA statement. See the section "Specification of Population Totals and Sampling Rates" for more information.

The STRATA *variables* are one or more variables in the DATA= input data set. These variables can be either character or numeric. The formatted values of the STRATA variables determine the levels. Thus, you can use formats to group values into levels. See the discussion of the FORMAT procedure in the *SAS Procedures Guide*.

You can specify the following option in the STRATA statement after a slash (/):

LIST
displays a "Stratum Information" table, which includes values of the STRATA variables and sampling rates for each stratum. This table also provides the number of observations and number of clusters for each stratum and analysis variable. See the section "Displayed Output" for more details.

Previous | Next | Top of Page

The SURVEYMEANS Procedure

WEIGHT Statement

WEIGHT | **WGT** *variable* **;**

The WEIGHT statement names the variable that contains the sampling weights. This variable must be numeric. If you do not specify a WEIGHT statement, PROC SURVEYMEANS assigns all observations a weight of 1. Sampling weights must be positive numbers. If an observation has a weight that is nonpositive or missing, then the procedure omits that observation from the analysis. If you specify more than one WEIGHT statement, the procedure uses only the first WEIGHT statement and ignores the rest.

Previous | Next | Top of Page

VAR Statement

VAR *variables* **;**

The VAR statement names the variables to be analyzed.

If you want a categorical analysis for a numeric variable, you must also name that variable in the CLASS statement. For categorical variables, PROC SURVEYMEANS estimates the proportion in each category or level, instead of the overall mean. Character variables are always analyzed as categorical variables. See the section "CLASS Statement" for more information.

If you do not specify a VAR statement, then PROC SURVEYMEANS analyzes all variables in the DATA= input data set, except those named in the BY, CLUSTER, STRATA, and WEIGHT statements.

Previous | Next | Top of Page

The SURVEYMEANS Procedure

Survey Data Analysis

Specification of Population Totals and Sampling Rates

If your analysis should include a finite population correction (*fpc*), you can input either the sampling rate or the population total using the RATE= option or the TOTAL= option. (You cannot specify both of these options in the same PROC SURVEYMEANS statement.) If you do not specify one of these options, the procedure does not use the *fpc* when computing variance estimates. For fairly small sampling fractions, it is appropriate to ignore this correction. Refer to Cochran (1977) and Kish (1965).

If your design has multiple stages of selection and you are specifying the RATE= option, you should input the first-stage sampling rate, which is the ratio of the number of PSUs in the sample to the total number of PSUs in the study population. If you are specifying the TOTAL= option for a multistage design, you should input the total number of PSUs in the study population. See the section "Primary Sampling Units (PSUs)" for more details.

For a nonstratified sample design, or for a stratified sample design with the same sampling rate or the same population total in all strata, you should use the RATE=*value* option or the TOTAL=*value* option. If your sample design is stratified with different sampling rates or population totals in the strata, then you can use the RATE= *SAS-data-set* option or the TOTAL= *SAS-data-set* option to name a SAS data set that contains the stratum sampling rates or totals. This data set is called a *secondary data set*, as opposed to the *primary data set* that you specify with the DATA= option.

The secondary data set must contain all the stratification variables listed in the STRATA statement and all the variables in the BY statement. If there are formats associated with the STRATA variables and the BY variables, then the formats must be consistent in the primary and the secondary data sets. If you specify the TOTAL=*SAS-data-set* option, the secondary data set must have a variable named _TOTAL_ that contains the stratum population totals. Or if you specify the RATE=*SAS-data-set* option, the secondary data set must have a variable named _RATE_ that contains the stratum sampling rates. If the secondary data set contains more than one observation for any one stratum, then the procedure uses the first value of _TOTAL_ or _RATE_ for that stratum and ignores the rest.

The *value* in the RATE= option or the values of _RATE_ in the secondary data set must be nonnegative numbers. You can specify *value* as a number between 0 and 1. Or you can specify *value* in percentage form as a number between 1 and 100, and PROC SURVEYMEANS will convert that number to a proportion. The procedure treats the value 1 as 100%, and not the percentage form 1%.

If you specify the TOTAL=*value* option, *value* must not be less than the sample size. If you provide stratum population totals in a secondary data set, these values must not be less than the corresponding stratum sample sizes.

Primary Sampling Units (PSUs)

When you have clusters, or primary sampling units (PSUs), in your sample design, the procedure estimates variance from the variation among PSUs. See the section "Variance and Standard Error of the Mean" and the section "Variance and Standard Deviation of the Total." You can use the CLUSTER statement to identify the first stage clusters in your design. PROC SURVEYMEANS assumes that each cluster represents a PSU in the sample and that each observation is an element of a PSU. If you do not specify a CLUSTER statement, the procedure treats each observation as a PSU.

Domain Analysis

It is common practice to compute statistics for subpopulations, or domains, in addition to computing statistics for the entire study population. Analysis for domains using the entire sample is called *domain analysis* (subgroup analysis, subpopulation analysis, subdomain analysis). The formation of these subpopulations of interest may be unrelated to the sample design. Therefore, the sample sizes for the subpopulations may actually be random variables.

In order to incorporate this variability into the variance estimation, you should use a DOMAIN statement. Note that using a BY statement provides completely separate analyses of the BY groups. It does not provide a statistically valid subpopulation or domain analysis, where the total number of units in the subpopulation is not known with certainty. For more detailed information about domain analysis, refer to Kish (1965).

Previous | Next | Top of Page

The SURVEYMEANS Procedure

Statistical Computations

The SURVEYMEANS procedure uses the Taylor expansion method to estimate sampling errors of estimators based on complex sample designs. This method obtains a linear approximation for the estimator and then uses the variance estimate for this approximation to estimate the variance of the estimate itself (Woodruff 1971, Fuller 1975). When there are clusters, or PSUs, in the sample design, the procedure estimates variance from the variation among PSUs. When the design is stratified, the procedure pools stratum variance estimates to compute the overall variance estimate. For *t* tests of the estimates, the degrees of freedom equals the number of clusters minus the number of strata in the sample design.

For a multistage sample design, the variance estimation method depends only on the first stage of the sample design. So, the required input includes only first-stage cluster (PSU) and first-stage stratum identification. You do not need to input design information about any additional stages of sampling. This variance estimation method assumes that the first-stage sampling fraction is small, or the first-stage sample is drawn with replacement, as it often is in practice.

Quite often in complex surveys, respondents have unequal weights, which reflect unequal selection probabilities and adjustments for nonresponse. In such surveys, the appropriate sampling weights must be used to obtain valid estimates for the study population.

For more information on the analysis of sample survey data, refer to Lee, Forthoffer, and Lorimor (1989), Cochran (1977), Kish (1965), and Hansen, Hurwitz, and Madow (1953).

Definition 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 number, with a total of H strata
- $i = 1, 2, \ldots, n_h$ is the cluster number within stratum h, with a total of n_h clusters
- $j = 1, 2, \ldots, m_{hi}$ is the unit number within cluster i of stratum h, with a total of m_{hi} units
- $p = 1, 2, \ldots, P$ 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 observation j in cluster i of stratum h
- $\mathbf{y}_{hij} = (y_{hij}^{(1)}, y_{hij}^{(2)}, \ldots, y_{hij}^{(P)})$ are the observed values of the analysis variables for observation 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 I denote the number of levels of C, and denote the level

values as c_1, c_2, ... , c_l. Then there are l indicator variables associated with these levels. That is, for level $C=c_k$ (k = 1, 2, ... , l), a $y^{(q)}\,(q \in \{1, 2, \ldots, P\})$ contains the values of the indicator variable for the category $C=c_k$, with the value of observation j in cluster i of stratum h:

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

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.

Also, f_h denotes the sampling rate for stratum h. You can use the TOTAL= option or the 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= option or the 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.

This notation is also applicable to other sample designs. For example, for a sample design without stratification, you can let H=1; for a sample design without clusters, you can let m_{hi}=1 for every h and i.

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

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

where

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

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

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.

PROC SURVEYMEANS uses the Taylor series expansion theory to estimate the variance of the mean $\widehat{\bar{Y}}$. The procedure computes the estimated variance as

$$\widehat{V}(\widehat{\bar{Y}}) = \sum_{h=1}^{H} \widehat{V_h}(\widehat{\bar{Y}})$$

where if n_h>1,

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

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

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

and if n_h=1,

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

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

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

Ratio

When you use a RATIO statement, the procedure produces statistics requested by the statistics-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 *j*th member in cluster *i* in the *h*th 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 expansion 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,

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

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

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

and if n_h=1,

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

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

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

t Test for the Mean

If you specify the keyword T, PROC SURVEYMEANS computes the *t*-value for testing that the population mean equals zero, $H_0 : \bar{Y} = 0$. The test statistic equals

$$t(\widehat{\bar{Y}}) = \widehat{\bar{Y}} \,/\, \text{StdErr}(\widehat{\bar{Y}})$$

The two-sided *p*-value for this test is

$$\text{Prob}(\,|T| > |t(\widehat{\bar{Y}})|\,)$$

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

PROC SURVEYMEANS calculates the degrees of freedom for the *t* test as the number of clusters minus the number of strata. If there are no clusters, then *df* equals the number of observations minus the number of strata. If the design is not stratified, then *df* equals the number of clusters minus one. The procedure displays *df* for the *t* test if you specify the keyword DF in the PROC SURVEYMEANS statement.

If missing values or missing weights are present in your data, the number of strata, the number of observations, and the number of clusters are counted based on the observations in non-empty strata. See the section "Missing Values" for details. For degrees of freedom in domain analysis, see the section "Domain Statistics."

Confidence Limits for the Mean

If you specify the keyword CLM, the procedure computes two-sided confidence limits for the mean. Also, the procedure includes the confidence limits by default if you do not specify any statistic-keywords in the PROC SURVEYMEANS statement.

The confidence coefficient is determined by the value of the ALPHA= option, which by default equals 0.05 and produces 95% confidence limits. The confidence limits are computed as

$$\hat{\bar{Y}} \pm \mathrm{StdErr}(\hat{\bar{Y}}) \cdot t_{df,\,\alpha/2}$$

where $\hat{\bar{Y}}$ is the estimate of the mean, $\mathrm{StdErr}(\hat{\bar{Y}})$ is the standard error of the mean, and $t_{df,\,\alpha/2}$ is the $100(1-\alpha/2)$ percentile of the *t* distribution with *df* calculated as described in the section " t Test for the Mean."

If you specify the keyword UCLM, the procedure computes the one-sided upper $100(1-\alpha)$ confidence limit for the mean:

$$\hat{\bar{Y}} + \mathrm{StdErr}(\hat{\bar{Y}}) \cdot t_{df,\,\alpha}$$

If you specify the keyword LCLM, the procedure computes the one-sided lower $100(1-\alpha)$ confidence limit for the mean:

$$\hat{\bar{Y}} - \mathrm{StdErr}(\hat{\bar{Y}}) \cdot t_{df,\,\alpha}$$

Coefficient of Variation

If you specify the keyword CV, PROC SURVEYMEANS computes the coefficient of variation, which is the ratio of the standard error of the mean to the estimated mean.

$$cv(\bar{Y}) = \mathrm{StdErr}(\hat{\bar{Y}}) \,/\, \hat{\bar{Y}}$$

If you specify the keyword CVSUM, PROC SURVEYMEANS computes the coefficient of variation for the estimated total, which is the ratio of the standard deviation of the sum to the estimated total.

$$cv(Y) = \mathrm{Std}(\widehat{Y}) \,/\, \widehat{Y}$$

Proportions

If you specify the keyword MEAN for a categorical variable, PROC SURVEYMEANS estimates the proportion, or relative frequency, for each level of the categorical variable. If you do not specify any statistic-keywords in the PROC SURVEYMEANS statement, the procedure estimates the proportions for levels of the categorical variables, together with their standard errors and confidence limits.

The procedure estimates the proportion in level c_k for variable C as

$$\hat{p} = \frac{\sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij}\, y_{hij}^{(q)}}{\sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij}}$$

where $y_{hij}^{(q)}$ is the value of the indicator function for level $C=c_k$, defined in the section "Definition and Notation," and $y_{hij}^{(q)}$ equals 1 if the observed value of variable C equals c_k, and $y_{hij}^{(q)}$ equals 0 otherwise. Since the proportion estimator is actually an estimator of the mean for an indicator variable, the procedure computes its variance and standard error according to the method outlined in the section "Variance and Standard Error of the Mean." Similarly, the procedure computes confidence limits for proportions as described in the section "Confidence Limits for the Mean."

Total

If you specify the keyword SUM, the procedure computes the estimate of the population total from the survey data. The estimate of the total is the weighted sum over the sample.

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

For a categorical variable level, $\widehat{Y}$ estimates its total frequency in the population.

Variance and Standard Deviation of the Total

When you specify the keyword STD or the keyword SUM, the procedure estimates the standard deviation of the total. The keyword VARSUM requests the variance of the total.

PROC SURVEYMEANS estimates the variance of the total as

$$\widehat{V}(\widehat{Y}) = \sum_{h=1}^{H} \widehat{V_h}(\widehat{Y})$$

where if $n_h>1$,

$$\widehat{V_h}(\widehat{Y}) = \frac{n_h(1-f_h)}{n_h-1}\sum_{i=1}^{n_h}(y_{hi\cdot}-\bar{y}_{h\cdot\cdot})^2$$

$$y_{hi\cdot} = \sum_{j=1}^{m_{hi}} w_{hij}\, y_{hij}$$

$$\bar{y}_{h\cdot\cdot} = \left(\sum_{i=1}^{n_h} y_{hi\cdot}\right) / n_h$$

and if n_h=1,

$$\widehat{V_h}(\widehat{Y}) = \begin{cases} \text{missing} & \text{if } n_{h'}=1 \text{ for } h'=1,2,\ldots,H \\ 0 & \text{if } n_{h'}>1 \text{ for some } 1<h'<H \end{cases}$$

The standard deviation of the total equals

$$\text{Std}(\widehat{Y}) = \sqrt{\widehat{V}(\widehat{Y})}$$

Confidence Limits of a Total

If you specify the keyword CLSUM, the procedure computes confidence limits for the total. The confidence coefficient is determined by the value of the ALPHA= option, which by default equals 0.05 and produces 95% confidence limits. The confidence limits are computed as

$$\widehat{Y} \pm \text{Std}(\widehat{Y}) \cdot t_{df,\,\alpha/2}$$

where $\widehat{Y}$ is the estimate of the total, $\text{Std}(\widehat{Y})$ is the estimated standard deviation, and $t_{df,\,\alpha/2}$ is the $100(1-\alpha/2)$ percentile of the *t* distribution with *df* calculated as described in the section "t Test for the Mean."

If you specify the keyword UCLSUM, the procedure computes the one-sided upper $100(1-\alpha)$ confidence limit for the sum:

$$\widehat{Y} + \text{Std}(\widehat{Y}) \cdot t_{df,\,\alpha}$$

If you specify the keyword LCLSUM, the procedure computes the one-sided lower $100(1-\alpha)$ confidence limit for the sum:

$$\widehat{Y} - \text{Std}(\widehat{Y}) \cdot t_{df,\,\alpha}$$

Domain Statistics

When you use a DOMAIN statement to request a domain analysis, the procedure computes the requested statistics for each domain.

For a domain *D*, let I_D be the corresponding indicator variable:

$$I_D(h,i,j) = \begin{cases} 1 & \text{if observation } (h,i,j) \text{ belongs to } D \\ 0 & \text{otherwise} \end{cases}$$

Let

$$z_{hij} = y_{hij} I_D(h,i,j) = \begin{cases} y_{hij} & \text{if observation } (h,i,j) \text{ belongs to } D \\ 0 & \text{otherwise} \end{cases}$$

The requested statistics for variable *y* in domain *D* are computed based on the values of *z*.

Domain Mean The estimated mean of *y* in the domain *D* is

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

where

$$v_{hij} = w_{hij} I_D(h,i,j) = \begin{cases} w_{hij} & \text{if observation } (h,i,j) \text{ belongs to } D \\ 0 & \text{otherwise} \end{cases}$$

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

The variance of $\widehat{\bar{Y}_D}$ is estimated by

$$\widehat{V}(\widehat{\bar{Y}_D}) = \sum_{h=1}^{H} \widehat{V_h}(\widehat{\bar{Y}_D})$$

where if n_h>1,

$$\begin{aligned}
\widehat{V_h}(\widehat{\bar{Y}_D}) &= \frac{n_h(1-f_h)}{n_h-1}\sum_{i=1}^{n_h}(r_{hi\cdot}-\bar{r}_{h\cdot\cdot})^2 \\
r_{hi\cdot} &= \left(\sum_{j=1}^{m_{hi}} v_{hij}\,(z_{hij}-\widehat{\bar{Y}_D})\right) / v_{\cdots} \\
\bar{r}_{h\cdot\cdot} &= \left(\sum_{i=1}^{n_h} r_{hi\cdot}\right) / n_h
\end{aligned}$$

and if n_h=1,

$$\widehat{V_h}(\widehat{\bar{Y}_D}) = \begin{cases} \text{missing} & \text{if } n_{h'}=1 \text{ for } h'=1,2,\ldots,H \\ 0 & \text{if } n_{h'}>1 \text{ for some } 1<h'<H \end{cases}$$

Domain Total The estimated total in domain *D* is

$$\widehat{Y}_D = \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} v_{hij}\,z_{hij}$$

and its estimated variance is

$$\widehat{V}(\widehat{Y}_D) = \sum_{h=1}^{H}\widehat{V_h}(\widehat{Y}_D)$$

where if n_h>1,

$$\begin{aligned}
\widehat{V_h}(\widehat{Y}_D) &= \frac{n_h(1-f_h)}{n_h-1}\sum_{i=1}^{n_h}(z_{hi\cdot}-\bar{z}_{h\cdot\cdot})^2 \\
z_{hi\cdot} &= \sum_{j=1}^{m_{hi}} v_{hij}\,z_{hij} \\
\bar{z}_{h\cdot\cdot} &= \left(\sum_{i=1}^{n_h} z_{hi\cdot}\right) / n_h
\end{aligned}$$

and if n_h=1,

$$\widehat{V_h}(\widehat{Y}_D) = \begin{cases} \text{missing} & \text{if } n_{h'}=1 \text{ for } h'=1,2,\ldots,H \\ 0 & \text{if } n_{h'}>1 \text{ for some } 1<h'<H \end{cases}$$

Degrees of Freedom For domain analysis, PROC SURVEYMEANS computes the degrees of freedom for *t* tests as the number of clusters in the non-empty strata minus the number of non-empty strata. When the sample design has no clusters, the degrees of freedom equals the number of observations in non-empty strata minus the number of non-empty strata. As discussed in the section "Missing Values," missing values and missing weights can result in empty strata. In domain analysis, an empty stratum can also occur when the stratum contains no observations in the specified domain. If no observations in a whole stratum belong to a domain, then this stratum is called an empty stratum for that domain.

For example,

```
data new;
   input str clu y w d;
   datalines;
1 1 . 40 9
1 2 2  . 9
1 3 . 25 9
2 4 5 20 9
2 5 8 15 9
3 6 5 30 7
3 7 9 89 7
3 8 6 23 7
;
proc surveymeans df nobs nclu nmiss;
   strata str;
   cluster clu;
   var y;
   weight w;
   domain d;
run;
```

Table 70.2: Calculations of *df* for Y

	Domain D=7	Domain D=9
Non Empty Strata	STR=3	STR=2
Clusters Used in the Analysis	CLU=6, CLU=7, and CLU=8	CLU=4 and CLU=5
df	3-1=2	2-1=1

Although there are three strata in the data set, STR=1 is an empty stratum for variable Y because of missing values and missing weights. In addition, no observations in stratum STR=3 belong to domain D=9. Therefore, STR=3 becomes an empty stratum as well for variable Y in domain D=9. As a result, the total number of non-empty strata for domain D=9 is one. The non-empty stratum for domain D=9 and variable Y is stratum STR=2. The total number of clusters for domain D=9 is two, which belong to stratum STR=2. Thus, for variable Y in domain D=9, the degrees of freedom for the *t* tests of the domain mean is *df*=2-1=1. Similarly, for domain D=7, strata STR=1 and STR=2 are both empty strata, so the total number of strata is one (STR=3), and the total number of clusters is three (CLU=6, CLU=7, and CLU=8). Table 70.2 illustrates how domains affect the total number of clusters and total number of strata in the *df* calculation. Figure 70.8 shows the *df* computed by the procedure.

The SURVEYMEANS Procedure

Domain Analysis: d

d	Variable	N	N Miss	Clusters	DF
7	y	3	0	3	6
9	y	2	2	2	4

Figure 70.8: Degrees of Freedoms in Domain Analysis

Previous | Next | Top of Page