# SAS/STAT® 9.2 User's Guide
# The SURVEYMEANS Procedure
## (Book Excerpt)

# Chapter 85

# The SURVEYMEANS Procedure

## Contents

# Overview: SURVEYMEANS Procedure

The SURVEYMEANS procedure estimates characteristics of a survey population by using statistics computed from a survey sample. You can estimate statistics such as means, totals, proportions, quantiles, and ratios. PROC SURVEYMEANS also provides domain analysis, which computes estimates for subpopulations or domains. The procedure also estimates variances and confidence limits and performs *t* tests for these statistics. PROC SURVEYMEANS uses either the Taylor series (linearization) method or replication (subsampling) methods to estimate sampling errors of estimators based on complex sample designs. The sample design can be a complex survey sample design with stratification, clustering, and unequal weighting. See Lohr (1999), Särndal, Swensson, and Wretman (1992), and Wolter (1985) for more details.

# Getting Started: SURVEYMEANS Procedure

This section demonstrates how you can use the SURVEYMEANS procedure to produce descriptive statistics from sample survey data. For a complete description of PROC SURVEYMEANS, see the section "Syntax: SURVEYMEANS Procedure" on page 6465. The section "Examples: SURVEYMEANS Procedure" on page 6510 provides more complicated examples to illustrate the applications of PROC SURVEYMEANS.

## Simple Random Sampling

This example illustrates how you can use PROC SURVEYMEANS to estimate population means and proportions from sample survey data. The study population is a junior high school with a total of 4,000 students in grades 7, 8, and 9. Researchers want to know how much these students spend weekly for ice cream, on average, and what percentage of students spend at least $10 weekly for ice cream.

To answer these questions, 40 students were selected from the entire student population by using simple random sampling (SRS). Selection by simple random sampling means that all students have an equal chance of being selected and no student can be selected more than once. Each student selected for the sample was asked how much he or she spends for ice cream per week, on average. The SAS data set IceCream saves the responses of the 40 students:

```
data IceCream;
   input Grade Spending @@;
   if (Spending < 10) then Group='less';
   else Group='more';
   datalines;
7 7   7   7   8 12   9 10   7   1   7 10   7   3   8 20   8 19   7 2
7 2   9 15   8 16   7   6   7   6   7   6   9 15   8 17   8 14   9 8
9 8   9   7   7   3   7 12   7   4   9 14   8 18   9   9   7   2   7 1
7 4   7 11   9   8   8 10   8 13   7   2   9   6   9 11   7   2   7 9
;
```

The variable Grade contains a student's grade. The variable Spending contains a student's response regarding how much he spends per week for ice cream, in dollars. The variable Group is created to indicate whether a student spends at least $10 weekly for ice cream: Group='more' if a student spends at least $10, or Group='less' if a student spends less than $10.

You can use PROC SURVEYMEANS to produce estimates for the entire student population, based on this random sample of 40 students:

```
title1 'Analysis of Ice Cream Spending';
title2 'Simple Random Sample Design';
proc surveymeans data=IceCream total=4000;
   var Spending Group;
   run;
```

The PROC SURVEYMEANS statement invokes the procedure. The TOTAL=4000 option specifies the total number of students in the study population, or school. The procedure uses this total to adjust variance estimates for the effects of sampling from a finite population. The VAR statement names the variables to analyze, Spending and Group.

Figure 85.1 displays the results from this analysis. There are a total of 40 observations used in the analysis. The "Class Level Information" table lists the two levels of the variable Group. This variable is a character variable, and so PROC SURVEYMEANS provides a categorical analysis for it, estimating the relative frequency or proportion for each level. If you want a categorical analysis for a numeric variable, you can name that variable in the CLASS statement.

**Figure 85.1** Analysis of Ice Cream Spending

```
                        Analysis of Ice Cream Spending
                          Simple Random Sample Design

                            The SURVEYMEANS Procedure

                                  Data Summary

                     Number of Observations              40

                             Class Level Information

                     Class
                     Variable       Levels    Values

                     Group               2    less more

                                  Statistics

                                              Std Error
Variable   Level              N         Mean    of Mean      95% CL for Mean
------------------------------------------------------------------------------
Spending                     40     8.750000   0.845139   7.04054539 10.4594546
Group      less              23     0.575000   0.078761   0.41568994  0.7343101
           more              17     0.425000   0.078761   0.26568994  0.5843101
------------------------------------------------------------------------------
```

The "Statistics" table displays the estimates for each analysis variable. By default, PROC SURVEYMEANS displays the number of observations, the estimate of the mean, its standard error, and the 95% confidence limits for the mean. You can obtain other statistics by specifying the corresponding *statistic-keywords* in the PROC SURVEYMEANS statement.

The estimate of the average weekly ice cream expense is $8.75 for students at this school. The standard error of this estimate if $0.85, and the 95% confidence interval for weekly ice cream expense is from $7.04 to $10.46. The analysis variable Group is a character variable, and so PROC SURVEYMEANS analyzes it as categorical, estimating the relative frequency or proportion for each level or category. These estimates are displayed in the Mean column of the "Statistics" table. It is estimated that 57.5% of all students spend less than $10 weekly on ice cream, while 42.5% of the students spend at least $10 weekly. The standard error of each estimate is 7.9%.

# Stratified Sampling

Suppose that the sample of students described in the previous section was actually selected by using stratified random sampling. In stratified sampling, the study population is divided into nonoverlapping strata, and samples are selected from each stratum independently.

The list of students in this junior high school was stratified by grade, yielding three strata: grades 7, 8, and 9. A simple random sample of students was selected from each grade. Table 85.1 shows the total number of students in each grade.

**Table 85.1** Number of Students by Grade

| Grade | Number of Students |
|-------|--------------------|
| 7 | 1,824 |
| 8 | 1,025 |
| 9 | 1,151 |
| Total | 4,000 |

To analyze this stratified sample, you need to provide the population totals for each stratum to PROC SURVEYMEANS. The SAS data set StudentTotals contains the information from Table 85.1:

```
data StudentTotals;
   input Grade _total_; datalines;
7 1824
8 1025
9 1151
;
```

The variable Grade is the stratum identification variable, and the variable _TOTAL_ contains the total number of students for each stratum. PROC SURVEYMEANS requires you to use the variable name _TOTAL_ for the stratum population totals.

The procedure uses the stratum population totals to adjust variance estimates for the effects of sampling from a finite population. If you do not provide population totals or sampling rates, then the procedure assumes that the proportion of the population in the sample is very small, and the computation does not involve a finite population correction.

In a stratified sample design, when the sampling rates in the strata are unequal, you need to use sampling weights to reflect this information in order to produce an unbiased mean estimator. In this example, the appropriate sampling weights are reciprocals of the probabilities of selection. You can use the following DATA step to create the sampling weights:

```
data IceCream;
   set IceCream;
   if Grade=7 then Prob=20/1824;
   if Grade=8 then Prob=9/1025;
   if Grade=9 then Prob=11/1151;
   Weight=1/Prob;
```

When you use PROC SURVEYSELECT to select your sample, the procedure creates these sampling weights for you.

The following SAS statements perform the stratified analysis of the survey data:

```
title1 'Analysis of Ice Cream Spending';
title2 'Stratified Simple Random Sample Design';
proc surveymeans data=IceCream total=StudentTotals;
   stratum Grade / list;
   var Spending Group;
   weight Weight;
run;
```

The PROC SURVEYMEANS statement invokes the procedure. The DATA= option names the SAS data set IceCream as the input data set to be analyzed. The TOTAL= option names the data set StudentTotals as the input data set that contains the stratum population totals. Comparing this to the analysis in the section "Simple Random Sampling" on page 6459, notice that the TOTAL=StudentTotals option is used here instead of the TOTAL=4000 option. In this stratified sample design, the population totals are different for different strata, and so you need to provide them to PROC SURVEYMEANS in a SAS data set.

The STRATA statement identifies the stratification variable Grade. The LIST option in the STRATA statement requests that the procedure display stratum information. The WEIGHT statement tells the procedure that the variable Weight contains the sampling weights.

Figure 85.2 displays information about the input data set. There are three strata in the design and 40 observations in the sample. The categorical variable Group has two levels, 'less' and 'more.'

**Figure 85.2** Data Summary

```
                 Analysis of Ice Cream Spending
              Stratified Simple Random Sample Design

                   The SURVEYMEANS Procedure

                         Data Summary

           Number of Strata                      3
           Number of Observations              40
           Sum of Weights                    4000


                   Class Level Information

           Class
           Variable      Levels    Values

           Group             2     less more
```

Figure 85.3 displays information for each stratum. The table displays a stratum index and the values of the STRATA variable. The stratum index identifies each stratum by a sequentially assigned number. For each stratum, the table gives the population total (total number of students), the sampling rate, and the sample size. The stratum sampling rate is the ratio of the number of students in the

sample to the number of students in the population for that stratum. The table also lists each analysis variable and the number of stratum observations for that variable. For categorical variables, the table lists each level and the number of sample observations in that level.

**Figure 85.3** Stratum Information

```
                        Stratum Information

    Stratum              Population  Sampling
     Index      Grade         Total      Rate      N Obs  Variable  Level
    -----------------------------------------------------------------------
       1          7           1824     1.10%         20   Spending
                                                          Group       less
                                                                      more
       2          8           1025     0.88%          9   Spending
                                                          Group       less
                                                                      more
       3          9           1151     0.96%         11   Spending
                                                          Group       less
                                                                      more
    -----------------------------------------------------------------------

                        Stratum Information

  Stratum                Population  Sampling
   Index       Grade          Total      Rate      N Obs  Variable            N
  -------------------------------------------------------------------------------
     1           7            1824     1.10%         20   Spending           20
                                                          Group              17
                                                                              3
     2           8            1025     0.88%          9   Spending            9
                                                          Group               0
                                                                              9
     3           9            1151     0.96%         11   Spending           11
                                                          Group               6
                                                                              5
  -------------------------------------------------------------------------------
```

Figure 85.4 shows the following:

- The estimate of average weekly ice cream expense is $9.14 for students in this school, with a standard error of $0.53, and a 95% confidence interval from $8.06 to $10.22.

- An estimate of 54.5% of all students spend less than $10 weekly on ice cream, and 45.5% spend more, with a standard error of 5.8%.

**Figure 85.4** Analysis of Ice Cream Spending

```
                           Statistics

                                       Std Error
Variable  Level          N        Mean    of Mean     95% CL for Mean
------------------------------------------------------------------------
Spending                40    9.141298    0.531799  8.06377052 10.2188254
Group      less         23    0.544555    0.058424  0.42617678  0.6629323
           more         17    0.455445    0.058424  0.33706769  0.5738232
------------------------------------------------------------------------
```

## Output Data Sets

PROC SURVEYMEANS uses the Output Delivery System (ODS) to create output data sets. This is a departure from older SAS procedures that provide OUTPUT statements for similar functionality. For more information about ODS, see Chapter 20, "Using the Output Delivery System."

For example, to save the "Statistics" table shown in Figure 85.4 in the previous section in an output data set, you use the ODS OUTPUT statement as follows:

```
title1 'Analysis of Ice Cream Spending';
title2 'Stratified Simple Random Sample Design';
proc surveymeans data=IceCream total=StudentTotals;
   stratum Grade / list;
   var Spending Group;
   weight Weight;
   ods output Statistics=MyStat;
run;
```

The statement

```
ods output Statistics=MyStat;
```

requests that the "Statistics" table that appears in Figure 85.4 be placed in a SAS data set MyStat.

The PRINT procedure displays observations of the data set MyStat:

```
proc print data=MyStat;
run;
```

Figure 85.5 displays the data set MyStat.

**Figure 85.5** The Data Set MyStat

```
                    Analysis of Ice Cream Spending
                 Stratified Simple Random Sample Design

              Var                                      Lower      Upper
Obs VarName  Level        N        Mean      StdErr    CLMean     CLMean

 1  Spending              40     9.141298            0.531799 8.06377052 10.2188254
 2  Group     less        23     0.544555            0.058424 0.42617678  0.6629323
 3  Group     more        17     0.455445            0.058424 0.33706769  0.5738232
```

The section "ODS Table Names" on page 6509 gives the complete list of the tables produced by PROC SURVEYMEANS.

# Syntax: SURVEYMEANS Procedure

The following statements are available in PROC SURVEYMEANS:

**PROC SURVEYMEANS** < *options* > < *statistic-keywords* > **;**
    **BY** *variables* **;**
    **CLASS** *variables* **;**
    **CLUSTER** *variables* **;**
    **DOMAIN** *variables* < *variable*∗*variable variable*∗*variable*∗*variable ...* > < / *option* > **;**
    **RATIO** < *'label'* > *variables* / *variables* **;**
    **REPWEIGHTS** *variables* < / *options* > **;**
    **STRATA** *variables* < / *option* > **;**
    **VAR** *variables* **;**
    **WEIGHT** *variable* **;**

The PROC SURVEYMEANS statement invokes the procedure. It optionally names the input data sets, specifies statistics for the procedure to compute, and specifies the variance estimation method. The PROC SURVEYMEANS statement is required.

The VAR statement identifies the variables to be analyzed. The CLASS statement identifies those numeric variables that are to be analyzed as categorical variables. The STRATA statement lists the variables that form the strata in a stratified sample design. The CLUSTER statement specifies cluster identification variables in a clustered sample design. The DOMAIN statement lists the variables that define domains for subpopulation analysis. The RATIO statement requests ratio analysis for means or proportions of analysis variables. The WEIGHT statement names the sampling weight variable. The REPWEIGHTS statement names replicate weight variables for BRR or jackknife variance estimation. You can use a BY statement with PROC SURVEYMEANS to obtain separate analyses for groups defined by the BY variables.

All statements can appear multiple times except the PROC SURVEYMEANS statement and the WEIGHT statement, which can appear only once.

The rest of this section gives detailed syntax information for the BY, CLASS, CLUSTER, DO-MAIN, RATIO, REPWEIGHTS, STRATA, VAR, and WEIGHT statements in alphabetical order after the description of the PROC SURVEYMEANS statement.

## 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, specify the variance estimation method, and provide sample design information. The DATA= option names the input data set to be analyzed. The VARMETHOD= option specifies the variance estimation method, which is the Taylor series method by default. For Taylor series variance estimation, you can include a finite population correction factor in the analysis by providing either the sampling rate or population total with 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 that contains 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=**$\alpha$

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 $\alpha$ produces $100(1-\alpha)\%$ confidence limits. The default of ALPHA=0.05 produces 95% confidence limits.

**DATA=**ized*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**

treats missing values as a valid (nonmissing) category for all categorical variables, which include CLASS, STRATA, CLUSTER, and DOMAIN variables.

By default, if you do not specify the MISSING option, an observation is excluded from the analysis if it has a missing value. For more information, see the section "Missing Values" on page 6483.

**NOMCAR**

requests that the procedure treat missing values in the variance computation as *not missing completely at random* (NOMCAR) for Taylor series variance estimation. When you specify the NOMCAR option, PROC SURVEYMEANS computes variance estimates by analyzing the nonmissing values as a domain or subpopulation, where the entire population includes both nonmissing and missing domains. See the section "Missing Values" on page 6483 for more details.

By default, PROC SURVEYMEANS completely excludes an observation from analysis if that observation has a missing value, unless you specify the MISSING option. Note that the NOMCAR option has no effect on a categorical variable when you specify the MISSING option, which treats missing values as a valid nonmissing level.

The NOMCAR option applies only to Taylor series variance estimation. The replication methods, which you request with the VARMETHOD=BRR and VARMETHOD=JACKKNIFE options, do not use the NOMCAR option.

**NONSYMCL**

requests nonsymmetric confidence limits for the quantiles when you request quantiles with PERCENTILE= or QUANTILE= option. See the section "Confidence Limits" on page 6497 for more details.

**NOSPARSE**

suppresses the display of analysis variables with zero frequency in the output. By default, the procedure displays all continuous variable and all levels of categorical variables.

**ORDER=DATA | FORMATTED | FREQ | 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. |
| FREQ | order values by descending frequency count; the most observed value comes first in the order. |
| 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.

**PERCENTILE=(***values***)**

specifies percentiles you want the procedure to compute. You can separate values with blanks or commas. Each value must be between 0 and 100. You can also use the *statistic-keywords* DECILES, MEDIAN, Q1, Q3, and QUARTILES to request common percentiles.

PROC SURVEYMEANS uses Woodruff method (Dorfman and Valliant 1993; Särndal, Swensson, and Wretman 1992; Francisco and Fuller 1991) to estimate the variances of quantiles. BRR and jackknife variance estimation do not apply to the variance estimation for percentiles. See the section "Quantiles" on page 6496 for more details.

**QUANTILE=(***values***)**

specifies quantiles you want the procedure to compute. You can separate values with blanks

or commas. Each value must be between 0 and 1. You can also use the *statistic-keywords* DECILES, MEDIAN, Q1, Q3, and QUARTILES to request common quantiles.

PROC SURVEYMEANS uses Woodruff method (Dorfman and Valliant 1993; Särndal, Swensson, and Wretman 1992; Francisco and Fuller 1991) to estimate the variances of quantiles. BRR and jackknife variance estimation do not apply to the variance estimation for quantiles. See the section "Quantiles" on page 6496 for more details.

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

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

specifies the sampling rate as a nonnegative *value*, or specifies an input data set that contains the stratum sampling rates. The procedure uses this information to compute a finite population correction for Taylor series variance estimation. The procedure does not use the RATE= option for BRR or jackknife variance estimation, which you request with the VARMETHOD=BRR or VARMETHOD=JACKKNIFE option.

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" on page 6484 for more details.

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 converts 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= or RATE= option, then the Taylor series variance estimation does not include a finite population correction. You cannot specify both the TOTAL= and RATE= options.

**STACKING**

requests that the procedure produce the output data sets by using a stacking table structure, which was the default before SAS 9. The new default is to produce a rectangular table structure in the output data sets.

A rectangular structure creates one observation for each analysis variable in the data set. A stacking structure creates only one observation in the output data set for all analysis variables.

The STACKING option affects the following tables:

- Domain
- Ratio
- Statistics
- StrataInfo

See the section "Rectangular and Stacking Structures in an Output Data Set" on page 6502 for more details.

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

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

specifies the total number of primary sampling units in the study population as a positive *value*, or specifies an input data set that contains the stratum population totals. The procedure uses this information to compute a finite population correction for Taylor series variance estimation. The procedure does not use the TOTAL= option for BRR or jackknife variance estimation, which you request with the VARMETHOD=BRR or VARMETHOD=JACKKNIFE option.

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" on page 6484 for more details.

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

*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" on page 6476 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 might not be the same for all analysis variables. See the section "Missing Values" on page 6483 for more information.

When you use a RATIO statement, available statistics computed for a ratio include N, NCLU, SUMWGT, RATIO, STDERR, DF, T, PROBT, and CLM, as shown in the following list. If no statistics are requested, the procedure computes 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 $\alpha$ is determined by the ALPHA= option; the default is $\alpha = 0.05$ |
| CLSUM | $100(1 - \alpha)\%$ two-sided confidence limits for the SUM, where $\alpha$ is determined by the ALPHA= option; the default is $\alpha = 0.05$ |
| CV | coefficient of variation for MEAN |
| CVSUM | coefficient of variation for SUM |
| DECILES | the 10th, the 20th, …, and the 90th percentiles including their standard errors and confidence limits |
| DF | degrees of freedom for the *t* test |
| LCLM | $100(1 - \alpha)\%$ one-sided lower confidence limit of the MEAN, where $\alpha$ is determined by the ALPHA= option; the default is $\alpha = 0.05$ |
| LCLSUM | $100(1 - \alpha)\%$ one-sided lower confidence limit of the SUM, where $\alpha$ is determined by the ALPHA= option; 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 |
| MEDIAN | median for a numeric variable |
| MIN | minimum value |
| NCLUSTER | number of clusters |
| NMISS | number of missing observations |
| NOBS | number of nonmissing observations |
| Q1 | lower quartile |
| Q3 | upper quartile |
| QUARTILES | lower quartile (25%th percentile), median (50%th percentile), and upper quartile (75%th percentile), including their standard errors and confidence limits |
| RANGE | range, MAX−MIN |
| 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 $\theta$ is a requested statistic

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

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

VAR          variance of the MEAN or RATIO

VARSUM          variance of the SUM

See the section "Statistical Computations" on page 6485 for details about how PROC SUR-VEYMEANS computes these statistics.

**VARMETHOD=BRR <(***method-options***)>**

**VARMETHOD=JACKKNIFE | JK <(***method-options***)>**

**VARMETHOD=TAYLOR**

specifies the variance estimation method. VARMETHOD=TAYLOR requests the Taylor series method, which is the default if you do not specify the VARMETHOD= option or the REPWEIGHTS statement. VARMETHOD=BRR requests variance estimation by balanced repeated replication (BRR), and VARMETHOD=JACKKNIFE requests variance estimation by the delete-1 jackknife method.

For VARMETHOD=BRR and VARMETHOD=JACKKNIFE you can specify *method-options* in parentheses. Table 85.2 summarizes the available *method-options*.

**Table 85.2**   Variance Estimation Method-Options

| **Keyword** | **Method** | **(*Method-Options*)** |
|---|---|---|
| BRR | Balanced repeated replication | DFADJ<br>FAY <=*value*><br>HADAMARD \| H=*SAS-data-set*<br>OUTWEIGHTS=*SAS-data-set*<br>PRINTH<br>REPS=*number* |
| JACKKNIFE \| JK | Jackknife | DFADJ<br>OUTJKCOEFS=*SAS-data-set*<br>OUTWEIGHTS=*SAS-data-set* |
| TAYLOR | Taylor series | |

*Method-options* must be enclosed in parentheses following the method keyword. For example:

```
varmethod=BRR(reps=60 outweights=myReplicateWeights)
```

The following values are available for the VARMETHOD= option:

BRR < (*method-options*) >    requests balanced repeated replication (BRR) variance estima-
tion. The BRR method requires a stratified sample design with two pri-
mary sampling units (PSUs) per stratum. See the section "Balanced Re-
peated Replication (BRR) Method" on page 6498 for more information.

You can specify the following *method-options* in parentheses following
VARMETHOD=BRR.

**DFADJ**

computes the degrees of freedom as the number of nonmiss-
ing strata for an analysis variable. The degrees of freedom for
VARMETHOD=BRR equal the number of strata, which by default
is based on all valid observations in the data set. But if you spec-
ify the DFADJ method-option, PROC SURVEYMEANS does not
count any empty strata due to all observations containing missing
values for an analysis variable.

See the section "Degrees of Freedom" on page 6490 for more in-
formation. See the section "Data and Sample Design Summary" on
page 6504 for details about valid observations.

The DFADJ method-option has no effect on categorical variables
when you specify the MISSING option, which treats missing values
as a valid nonmissing level.

The DFADJ method-option cannot be used when you provide repli-
cate weights with a REPWEIGHTS statement. When you use a
REPWEIGHTS statement, the degrees of freedom equal the num-
ber of REPWEIGHTS variables (or replicates), unless you specify
an alternative value in the DF= option in the REPWEIGHTS state-
ment.

**FAY** < =*value* >

requests Fay's method, a modification of the BRR method, for vari-
ance estimation. See the section "Fay's BRR Method" on page 6499
for more information.

You can specify the *value* of the Fay coefficient, which is used in
converting the original sampling weights to replicate weights. The
Fay coefficient must be a nonnegative number less than 1. By de-
fault, the value of the Fay coefficient equals 0.5.

**HADAMARD=***SAS-data-set*

**H=***SAS-data-set*

names a SAS data set that contains the Hadamard matrix for BRR
replicate construction. If you do not provide a Hadamard matrix
with the HADAMARD= method-option, PROC SURVEYMEANS
generates an appropriate Hadamard matrix for replicate construc-
tion. See the sections "Balanced Repeated Replication (BRR)

Method" on page 6498 and "Hadamard Matrix" on page 6501 for details.

If a Hadamard matrix of a given dimension exists, it is not necessarily unique. Therefore, if you want to use a specific Hadamard matrix, you must provide the matrix as a SAS data set in the HADAMARD=*SAS-data-set* method-option.

In the HADAMARD= input data set, each variable corresponds to a column of the Hadamard matrix, and each observation corresponds to a row of the matrix. You can use any variable names in the HADAMARD= data set. All values in the data set must equal either 1 or −1. You must ensure that the matrix you provide is indeed a Hadamard matrix—that is, $\mathbf{A}'\mathbf{A} = R\mathbf{I}$, where $\mathbf{A}$ is the Hadamard matrix of dimension $R$ and $\mathbf{I}$ is an identity matrix. PROC SURVEYMEANS does not check the validity of the Hadamard matrix that you provide.

The HADAMARD= input data set must contain at least $H$ variables, where $H$ denotes the number of first-stage strata in your design. If the data set contains more than $H$ variables, the procedure uses only the first $H$ variables. Similarly, the HADAMARD= input data set must contain at least $H$ observations.

If you do not specify the REPS= method-option, then the number of replicates is taken to be the number of observations in the HADAMARD= input data set. If you specify the number of replicates—for example, REPS=*nreps*—then the first *nreps* observations in the HADAMARD= data set are used to construct the replicates.

You can specify the PRINTH option to display the Hadamard matrix that the procedure uses to construct replicates for BRR.

**OUTWEIGHTS=**derived*SAS-data-set*

names a SAS data set that contains replicate weights. See the section "Balanced Repeated Replication (BRR) Method" on page 6498 for information about replicate weights. See the section "Replicate Weights Output Data Set" on page 6501 for more details about the contents of the OUTWEIGHTS= data set.

The OUTWEIGHTS= method-option is not available when you provide replicate weights with the REPWEIGHTS statement.

**PRINTH**

displays the Hadamard matrix.

When you provide your own Hadamard matrix with the HADAMARD= method-option, only the rows and columns of the Hadamard matrix that are used by the procedure are displayed.

See the sections "Balanced Repeated Replication (BRR) Method" on page 6498 and "Hadamard Matrix" on page 6501 for details.

The PRINTH method-option is not available when you provide replicate weights with the REPWEIGHTS statement because the procedure does not use a Hadamard matrix in this case.

**REPS=***number*

specifies the number of replicates for BRR variance estimation. The value of *number* must be an integer greater than 1.

If you do not provide a Hadamard matrix with the HADAMARD= method-option, the number of replicates should be greater than the number of strata and should be a multiple of 4. See the section "Balanced Repeated Replication (BRR) Method" on page 6498 for more information. If a Hadamard matrix cannot be constructed for the REPS= value that you specify, the value is increased until a Hadamard matrix of that dimension can be constructed. Therefore, it is possible for the actual number of replicates used to be larger than the REPS= value that you specify.

If you provide a Hadamard matrix with the HADAMARD= method-option, the value of REPS= must not be less than the number of rows in the Hadamard matrix. If you provide a Hadamard matrix and do not specify the REPS= method-option, the number of replicates equals the number of rows in the Hadamard matrix.

If you do not specify the REPS= or HADAMARD= method-option and do not include a REPWEIGHTS statement, the number of replicates equals the smallest multiple of 4 that is greater than the number of strata.

If you provide replicate weights with the REPWEIGHTS statement, the procedure does not use the REPS= method-option. With a REP-WEIGHTS statement, the number of replicates equals the number of REPWEIGHTS variables.

JACKKNIFE | JK <(*method-options*)> requests variance estimation by the delete-1 jack-knife method. See the section "Jackknife Method" on page 6500 for de-tails. If you provide replicate weights with a REPWEIGHTS statement, VARMETHOD=JACKKNIFE is the default variance estimation method.

You can specify the following *method-options* in parentheses following VARMETHOD=JACKKNIFE:

**DFADJ**

computes the degrees of freedom as the number of nonmiss-ing strata for an analysis variable. The degrees of freedom for VARMETHOD=BRR equal the number of strata, which by default is based on all valid observations in the data set. But if you spec-ify the DFADJ method-option, PROC SURVEYMEANS does not

count any empty strata due to all observations containing missing values for an analysis variable.

See the section "Degrees of Freedom" on page 6490 for more information. See the section "Data and Sample Design Summary" on page 6504 for details about valid observations.

The DFADJ method-option has no effect on categorical variables when you specify the MISSING option, which treats missing values as a valid nonmissing level.

The DFADJ method-option cannot be used when you provide replicate weights with a REPWEIGHTS statement. When you use a REPWEIGHTS statement, the degrees of freedom equal the number of REPWEIGHTS variables (or replicates), unless you specify an alternative value in the DF= option in the REPWEIGHTS statement.

**OUTWEIGHTS=***SAS-data-set*

names a SAS data set that contains replicate weights. See the section "Jackknife Method" on page 6500 for information about replicate weights. See the section "Replicate Weights Output Data Set" on page 6501 for more details about the contents of the OUT-WEIGHTS= data set.

The OUTWEIGHTS= method-option is not available when you provide replicate weights with the REPWEIGHTS statement.

**OUTJKCOEFS=***SAS-data-set*

names a SAS data set that contains jackknife coefficients. See the section "Jackknife Method" on page 6500 for information about jackknife coefficients. See the section "Jackknife Coefficients Output Data Set" on page 6502 for more details about the contents of the OUTJKCOEFS= data set.

TAYLOR    requests Taylor series variance estimation. This is the default method if you do not specify the VARMETHOD= option and if there is no REPWEIGHTS statement. See the section "Statistical Computations" on page 6485 for more information.

## BY Statement

**BY** *variables* **;**

You can specify a BY statement with PROC SURVEYMEANS to obtain separate analyses on observations in groups defined by the BY variables.

Note that by 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. You should use the DOMAIN statement to obtain domain analysis.

When a BY statement appears, the procedure expects the input data sets to be sorted in order of the BY variables. The *variables* are one or more variables in the input data set.

If you specify more than one BY statement, the procedure uses only the latest BY statement and ignores any previous ones.

If your input data set is not sorted in ascending order, use one of the following alternatives:

- Sort the data by using the SORT procedure with a similar BY statement.

- Use the BY statement option NOTSORTED or DESCENDING. The NOTSORTED option does not mean that the data are unsorted but rather that the data are arranged in groups (according to values of the BY variables) and that these groups are not necessarily in alphabetical or increasing numeric order.

- Create an index on the BY variables by using the DATASETS procedure.

For more information about the BY statement, see *SAS Language Reference: Concepts*. For more information about the DATASETS procedure, see the *Base SAS Procedures Guide*.

## CLASS Statement

> **CLASS** *variables* ;

The CLASS statement names variables to be analyzed as categorical variables. For categorical variables, PROC SURVEYMEANS estimates the proportion in each category or level, instead of the overall mean. PROC SURVEYMEANS always analyzes character variables as categorical. If you want categorical analysis for a numeric variable, you must include that variable in the CLASS statement.

The CLASS *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 CLASS variables determine the categorical variable levels. Thus, you can use formats to group values into levels. See the FORMAT procedure in the *Base SAS Procedures Guide* and the FORMAT statement and SAS formats in *SAS Language Reference: Dictionary* for more information.

You can use multiple CLASS statements to specify categorical variables.

When you specify classification variables, you can use the SAS system option SUMSIZE= to limit (or to specify) the amount of memory that is available for data analysis. See the chapter on SAS system options in *SAS Language Reference: Dictionary* for a description of the SUMSIZE= option.

# CLUSTER Statement

> **CLUSTER** *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 you provide replicate weights for BRR or jackknife variance estimation with the REPWEIGHTS statement, you do not need to specify a CLUSTER statement.

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)" on page 6485 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. See the FORMAT procedure in the *Base SAS Procedures Guide* and the FORMAT statement and SAS formats in *SAS Language Reference: Dictionary* for more information.

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

# DOMAIN Statement

> **DOMAIN** *variables* < *variable*∗*variable variable*∗*variable*∗*variable* ... > < / *option* > ;

The DOMAIN statement requests analysis for subpopulations, or domains, in addition to analysis for the entire study population. The DOMAIN statement names the variables that identify domains, which are called domain variables.

It is common practice to compute statistics for domains. The formation of these domains might be unrelated to the sample design. Therefore, the sample sizes for the domains are random variables. In order to incorporate this variability into the variance estimation, you should use a DOMAIN statement.

Note that a DOMAIN statement is different from a BY statement. In a BY statement, you treat the sample sizes as fixed in each subpopulation, and you perform analysis within each BY group independently. See the section "Domain Statistics" on page 6494 for more details.

You should use the DOMAIN statement on the entire data set to perform a domain analysis. Creating a new data set from a single domain and analyzing that with SURVEYMEANS yields inappropriate estimates of variance.

A domain variable can be either character or numeric. The procedure treats domain variables as categorical variables. If a variable appears by itself in a DOMAIN statement, each level of this variable determines a domain in the study population. If two or more variables are joined by asterisks (*),

then every possible combination of levels of these variables determines a domain. The procedure performs a descriptive analysis within each domain defined by the domain variables.

The formatted values of the domain variables determine the categorical variable levels. Thus, you can use formats to group values into levels. See the FORMAT procedure in the *Base SAS Procedures Guide* and the FORMAT statement and SAS formats in *SAS Language Reference: Dictionary* for more information.

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

**DFADJ**

> computes the degrees of freedom by using the number of nonempty strata for an analysis variable in a domain.

> Given a specific domain, it is possible that there is no sampling unit in the sample falling into a domain for a stratum or there is no valid observations from a stratum in a domain due to missing values. Therefore, a domain might contain empty strata. By default, the procedure counts these empty strata when calculating the degrees of freedom.

> However, if you specify the DFADJ option, the procedure makes an adjustment by excluding these empty strata when compute the degrees of freedom. Note that prior to SAS 9.2, the procedure made this adjustment by default.

> See the section "Degrees of Freedom" on page 6490 for more information. See the section "Data and Sample Design Summary" on page 6504 for details about valid observations.

> The DFADJ option has no effect on categorical variables when you specify the MISSING option, which treats missing values as a valid nonmissing level.

## RATIO Statement

> **RATIO** < *'label'* > *variables* / *variables* ;

The RATIO statement requests ratio analysis for means or proportions of analysis variables. A ratio statement names the variables whose means are used as numerators or denominators in a ratio. Variables appearing before the slash (/), called *numerator variables*, are used as numerators. Variables appearing after the slash (/), called *denominator variables*, are used as denominators. These *variables* can be any number of analysis variables, either continuous or categorical in the input data set, except those named in the BY, CLUSTER, REPWEIGHT, STRATA, and WEIGHT statements.

You can optionally specify a label for each RATIO statement to identify the ratios in the output. Labels must be enclosed in single quotes.

If a RATIO statement does not have any numerator variable or denominator variable specified, the RATIO statement is ignored.

The computation of ratios depends on whether the numerator and denominator variables are continuous or categorical.

For continuous variables, ratios are calculated with the mean of the variables. For example, for continuous variables X, Y, Z, and T, the following RATIO statement requests that the procedure analyze the ratios $\bar{x}/\bar{z}$, $\bar{x}/\bar{t}$, $\bar{y}/\bar{z}$, and $\bar{y}/\bar{t}$:

```
ratio x y / z t;
```

If a continuous variable appears as both a numerator and a denominator variable, the ratio of this variable itself is ignored.

For categorical variables, ratios are calculated with the proportions for the categories of a categorical variable. For example, if the categorical variable Gender has the values 'Male' and 'Female,' with the proportions $p_m = \Pr(\text{Gender}=\text{'Male'})$ and $p_f = \Pr(\text{Gender}=\text{'Female'})$, and Y is a continuous variable, then the following RATIO statement requests that the procedure analyze the ratios $p_m/p_f$, $p_f/p_m$, $\bar{y}/p_m$, and $\bar{y}/p_f$:

```
ratio Gender y / Gender;
```

If a categorical variable appears as both a numerator and denominator variable, then the ratios of the proportions for all categories are computed, except the ratio of each category with itself.

You can have more than one RATIO statement. Each RATIO statement produces ratios independently by using its own numerator and denominator variables. Each RATIO statement also produces its own ratio analysis table.

Available statistics for a ratio are as follows:

- N, number of observations used to compute the ratio

- NCLU, number of clusters

- SUMWGT, sum of weights

- RATIO, ratio

- STDERR, standard error of ratio

- VAR, variance of ratio

- T, *t*-value of ratio

- PROBT, *p*-value of *t*

- DF, degrees of freedom of *t*

- CLM, two-sided confidence limits of ratio

- UCLM, one-sided upper confidence limit of ratio

- LCLM, one-sided lower confidence limit of ratio

The procedure calculates these statistics based on the *statistic-keywords* that you specified in the PROC SURVEYMEANS statement. If a *statistic-keyword* is not appropriate for a RATIO statement, that *statistic-keyword* is ignored. If no valid statistics are requested for a RATIO statement, the procedure computes the ratio and its standard error by default.

Note that ratios within a domain are currently not available.

When the means or proportions for the numerator and denominator variables in a ratio are calculated, an observation is excluded if it has a missing value in either the continuous numerator variable or the denominator variable. An observation with missing values is also excluded for the categorical numerator or denominator variables, unless the MISSING option is used.

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.

## REPWEIGHTS Statement

>   **REPWEIGHTS** *variables* < / *options* > ;

The REPWEIGHTS statement names variables that provide replicate weights for BRR or jackknife variance estimation, which you request with the VARMETHOD=BRR or VARMETHOD=JACKKNIFE option in the PROC SURVEYMEANS statement. If you do not provide replicate weights for these methods by using a REPWEIGHTS statement, then the procedure constructs replicate weights for the analysis. See the sections "Balanced Repeated Replication (BRR) Method" on page 6498 and "Jackknife Method" on page 6500 for information about replicate weights.

Each REPWEIGHTS variable should contain the weights for a single replicate, and the number of replicates equals the number of REPWEIGHTS variables. The REPWEIGHTS variables must be numeric, and the variable values must be nonnegative numbers.

If you provide replicate weights with a REPWEIGHTS statement, you do not need to specify a CLUSTER or STRATA statement. If you use a REPWEIGHTS statement and do not specify the VARMETHOD= option in the PROC SURVEYMEANS statement, the procedure uses VARMETHOD=JACKKNIFE by default.

If you specify a REPWEIGHTS statement but do not include a WEIGHT statement, the procedure uses the average of each observation's replicate weights as the observation's weight.

You can specify the following options in the REPWEIGHTS statement after a slash (/):

**DF=**df

>   specifies the degrees of freedom for the analysis. The value of *df* must be a positive number. By default, the degrees of freedom equals the number of REPWEIGHTS variables. See the section "Degrees of Freedom" on page 6490 for details.

>   PROC SURVEYMEANS uses the DF= value in computing confidence limits for proportions, totals, and other statistics. See the section "Statistical Computations" on page 6485 for details.

**JKCOEFS=**value

>   specifies a jackknife coefficient for VARMETHOD=JACKKNIFE. The coefficient *value* must

be a nonnegative number less than one. See the section "Jackknife Method" on page 6500 for details about jackknife coefficients.

You can use this option to specify a single value of the jackknife coefficient, which the procedure uses for all replicates. To specify different coefficients for different replicates, use the JKCOEFS=*values* or JKCOEFS=*SAS-data-set* option.

**JKCOEFS=***values*

specifies jackknife coefficients for VARMETHOD=JACKKNIFE, where each coefficient corresponds to an individual replicate identified by a REPWEIGHTS variable. You can separate *values* with blanks or commas. The coefficient *values* must be nonnegative numbers less than one. The number of *values* must equal the number of replicate weight variables named in the REPWEIGHTS statement. List these values in the same order in which you list the corresponding replicate weight variables in the REPWEIGHTS statement.

See the section "Jackknife Method" on page 6500 for details about jackknife coefficients.

To specify different coefficients for different replicates, you can also use the JKCOEFS=*SAS-data-set* option. To specify a single jackknife coefficient for all replicates, use the JKCOEFS=*value* option.

**JKCOEFS=***SAS-data-set*

names a SAS data set that contains the jackknife coefficients for VARMETHOD=JACKKNIFE. You provide the jackknife coefficients in the JKCOEFS= data set variable JKCoefficient. Each coefficient value must be a nonnegative number less than one. The observations in the JKCOEFS= data set should correspond to the replicates that are identified by the REPWEIGHTS variables. Arrange the coefficients or observations in the JKCOEFS= data set in the same order in which you list the corresponding replicate weight variables in the REPWEIGHTS statement. The number of observations in the JKCOEFS= data set must not be less than the number of REPWEIGHTS variables.

See the section "Jackknife Method" on page 6500 for details about jackknife coefficients.

To specify different coefficients for different replicates, you can also use the JKCOEFS=*values* option. To specify a single jackknife coefficient for all replicates, use the JKCOEFS=*value* option.

## STRATA Statement

**STRATA** *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 "Survey Data Analysis" on page 6484 for more information.

If you provide replicate weights for BRR or jackknife variance estimation with the REPWEIGHTS statement, you do not need to specify a STRATA statement.

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 FORMAT procedure in the *Base SAS Procedures Guide* and the FORMAT statement and SAS formats in *SAS Language Reference: Dictionary* for more information.

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" on page 6504 for more details.

## 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" on page 6476 for more information.

When you specify a variable in a RATIO statement, but not in a VAR statement, the procedure includes this variable as an analysis variable.

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, REPWEIGHT, STRATA, and WEIGHT statements.

## WEIGHT Statement

> **WEIGHT** *variable* ;

The WEIGHT statement names the variable that contains the sampling weights. This variable must be numeric, and the 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. See the section "Missing Values" on page 6483 for more information. If you specify more than one WEIGHT statement, the procedure uses only the first WEIGHT statement and ignores the rest.

If you do not specify a WEIGHT statement but provide replicate weights with a REPWEIGHTS statement, PROC SURVEYMEANS uses the average of each observation's replicate weights as the observation's weight.

If you do not specify a WEIGHT statement or a REPWEIGHTS statement, PROC SUR-VEYMEANS assigns all observations a weight of one.

# Details: SURVEYMEANS Procedure

## Missing Values

If you have missing values in your survey data for any reason, such as nonresponse, this can compromise the quality of your survey results. If the respondents are different from the nonrespondents with regard to a survey effect or outcome, then survey estimates might be biased and cannot accurately represent the survey population. There are a variety of techniques in sample design and survey operations that can reduce nonresponse. After data collection is complete, you can use imputation to replace missing values with acceptable values, and/or you can use sampling weight adjustments to compensate for nonresponse. You should complete this data preparation and adjustment before you analyze your data with PROC SURVEYMEANS. See Cochran (1977); Kalton and Kaspyzyk (1986); and Brick and Kalton (1996) for more information.

If an observation has a missing value or a nonpositive value for the WEIGHT variable, then that observation is excluded from the analysis.

An observation is also excluded if it has a missing value for design variables such as STRATA variables, CLUSTER variables, and DOMAIN variables, unless missing values are regarded as a legitimate categorical level for these variables, as specified by the MISSING option.

If you specify the MISSING option in the PROC SURVEYMEANS statement, the procedure treats missing values of a categorical variable as a valid category.

By default, when computing statistics for an analysis variable, PROC SURVEYMEANS omits observations with missing values for that variable. The procedure computes statistics for each variable based only on observations that have nonmissing values for that variable. This treatment is based on the assumption that the missing values are missing completely at random (MCAR). However, this assumption is sometimes not true. For example, evidence from other surveys might suggest that observations with missing values are systematically different from observations without missing values. If you believe that missing values are not missing completely at random, then you can specify the NOMCAR option to let variance estimation include these observations with missing values in the analysis variables.

Whether or not the NOMCAR option is used, observations with missing or invalid values for WEIGHT, STRATA, CLUSTER, or DOMAIN variables are always excluded, unless the MISSING option is also specified.

When you specify the NOMCAR option, the procedure treats observations with and without missing values for analysis variables as two different domains, and it performs a domain analysis in the domain of nonmissing observations.

The procedure performs univariate analysis and analyzes each VAR variable separately. Thus, the number of missing observations might be different for different variables. You can specify the keyword NMISS in the PROC SURVEYMEANS statement to display the number of missing values for each analysis variable in the "Statistics" table.

If you use a REPWEIGHTS statement, all REPWEIGHTS variables must contain nonmissing values.

## Survey Data Analysis

### Specification of Population Totals and Sampling Rates

To include a finite population correction (*fpc*) in Taylor series variance estimation, you can input either the sampling rate or the population total by using the RATE= or TOTAL= option in the PROC SURVEYMEANS statement. (You cannot specify both of these options in the same PROC SURVEYMEANS statement.) The RATE= and TOTAL= options apply only to Taylor series variance estimation. The procedure does not use a finite population correction for BRR or jackknife variance estimation.

If you do not specify the RATE= or TOTAL= option, the Taylor series variance estimation does not include a finite population correction. For fairly small sampling fractions, it is appropriate to ignore this correction. See Cochran (1977) and Kish (1965) for more information.

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)" on page 6485 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 can use the RATE=*value* or TOTAL=*value* option. If your sample design is stratified with different sampling rates or population totals in different strata, use the RATE=*SAS-data-set* or 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 converts 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 when the Taylor series variance method is used. See the section "Variance and Standard Error of the Mean" on page 6488 and the section "Variance and Standard Deviation of the Total" on page 6493 for more information.

BRR or jackknife variance estimation methods draw multiple replicates (or subsamples) from the full sample by following a specific resampling scheme. These subsamples are constructed by deleting PSUs from the full sample.

If you use a REPWEIGHTS statement to provide replicate weights for BRR or jackknife variance estimation, you do not need to specify a CLUSTER statement. Otherwise, you should specify a CLUSTER statement whenever your design includes clustering at the first stage of sampling. If you do not specify a CLUSTER statement, then PROC SURVEYMEANS 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 that uses the entire sample is called *domain analysis* (subgroup analysis, subpopulation analysis, subdomain analysis). The formation of these subpopulations of interest might be unrelated to the sample design. Therefore, the sample sizes for the subpopulations might 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, see Kish (1965).

## Statistical Computations

The SURVEYMEANS procedure uses the Taylor series (linearization) method or replication (resampling) methods to estimate sampling errors of estimators based on complex sample designs. For details see Wolter (1985); Lohr (1999); Kalton (1983); Hidiroglou, Fuller, and Hickman (1980); Fuller et al. (1989); Lee, Forthoffer, and Lorimor (1989); Cochran (1977); Kish (1965); Hansen,

Hurwitz, and Madow (1953); Rust (1985); Dippo, Fay, and Morganstein (1984); Rao and Shao (1999); Rao, Wu, and Yue (1992); and Rao and Shao (1996). You can use the VARMETHOD= option to specify a variance estimation method to use. By default, the Taylor series method is used.

The Taylor series 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 equal the number of clusters minus the number of strata in the sample design.

For a multistage sample design, the Taylor series estimation depends only on the first stage of the sample design. Therefore, 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 that 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.

However, replication methods have recently 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.

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 population parameters 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 a REPWEIGHTS statement to provide your own replicate weights for variance estimation. For more information about using replication methods to analyze sample survey data, see the section "Replication Methods for Variance Estimation" on page 6498.

## 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

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

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" on page 6484 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 $f_h$.

## 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

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

where

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

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.

### 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 $\widehat{\bar{Y}}$. The procedure computes the estimated variance as

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

where if $n_h > 1$,

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

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

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

and if $n_h = 1$,

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

### Replication Methods

When you specify VARMETHOD=BRR or VARMETHOD=JACKKNIFE, the procedure computes the variance $\widehat{V}(\widehat{\bar{Y}})$ with replication methods by using the variability among replicate estimates to estimate the overall variance. See the section "Replication Methods for Variance Estimation" on page 6498 for more details.

### Standard Error

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 *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 $j$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 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 $n_h = 1$,

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

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.

## *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.

## Degrees of Freedom

PROC SURVEYMEANS computes degrees of freedom $df$ to obtain the $100(1 - \alpha)\%$ confidence limits for means, proportions, totals, ratios, and other statistics. The degrees of freedom computation depends on the variance estimation method that you request. Missing values can affect the degrees of freedom computation. See the section "Missing Values" on page 6483 for details.

### *Taylor Series Variance Estimation*

For the Taylor series method, 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 the degrees of freedom equal the number of observations minus the number of strata. If the design is not stratified, then the degrees of freedom equal the number of PSUs minus one.

If in a stratum, all observations are excluded from the analysis due to missing values, then that stratum is called an *empty stratum*. Empty strata are not counted in the total number of strata for the table. Similarly, empty clusters and missing observations are not included in the totals counts of cluster and observations used to compute the degrees of freedom for the analysis.

If you specify the MISSING option, missing values are treated as valid nonmissing levels for a categorical variable and are included in computing degrees of freedom. If you specify the NOM-CAR option for Taylor series variance estimation, observations with missing values for an analysis variable are included in computing degrees of freedom.

### *Replicate-Based Variance Estimation*

When there is a REPWEIGHTS statement, the degrees of freedom equal the number of REP-WEIGHTS variables, unless you specify an alternative in the DF= option in a REPWEIGHTS statement.

For BRR or jackknife variance estimation without a REPWEIGHT statement, by default PROC SURVEYMEANS computes the degrees of freedom by using all valid observations in the input data set. A valid observation is an observation that has a positive value of the WEIGHT variable and nonmissing values of the STRATA and CLUSTER variables unless you specify the MISSING option. See the section "Data and Sample Design Summary" on page 6504 for details about valid observations.

For BRR variance estimation (including Fay's method) without a REPWEIGHTS statement, PROC SURVEYMEANS calculates the degrees of freedom as the number of strata. PROC SURVEYMEANS bases the number of strata on all valid observations in the data set, unless you specify the DFADJ method-option for VARMETHOD=BRR. When you specify the DFADJ option, the procedure computes the degrees of freedom as the number of nonmissing strata for an analysis variable. This excludes any empty strata that occur when observations with missing values of that analysis variable are removed.

For jackknife variance estimation without a REPWEIGHTS statement, PROC SURVEYMEANS calculates the degrees of freedom 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. For jackknife variance estimation, PROC SURVEYMEANS bases the number of strata and clusters on all valid observations in the data set, unless you specify the DFADJ method-option for VARMETHOD=JK. When you specify the DFADJ option, the procedure computes the degrees of freedom from the number of nonmissing strata and clusters for an analysis variable. This excludes any empty strata or clusters that occur when observations with missing values of an analysis variable are removed.

The procedure displays the degrees of freedom for the $t$ test if you specify the keyword DF in the PROC SURVEYMEANS statement.

## 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

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

where $\widehat{\overline{Y}}$ is the estimate of the mean, $\text{StdErr}(\widehat{\overline{Y}})$ is the standard error of the mean, and $t_{df,\,\alpha/2}$ is the $100(1 - \alpha/2)$th percentile of the $t$ distribution with $df$ calculated as described in the section "$t$ Test for the Mean" on page 6490.

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

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

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

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

## 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:

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

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:

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

## 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 "Definitions and Notation" on page 6486, 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" on page 6488. Similarly, the procedure computes confidence limits for proportions as described in the section "Confidence Limits for the Mean" on page 6491.

## 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.

### *Taylor Series Method*

When you use VARMETHOD=TAYLOR, or by default, PROC SURVEYMEANS uses the Taylor series method to estimate 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 $n_h = 1$,

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

### *Replication Methods*

When you specify VARMETHOD=BRR or VARMETHOD=JACKKNIFE option, the procedure computes the variance $\widehat{V}(\widehat{Y})$ with replication methods by measuring the variability among the estimates derived from these replicates. See the section "Replication Methods for Variance Estimation" on page 6498 for more details.

### *Standard Deviation*

The standard deviation of the total equals

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

## Confidence Limits for the 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

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

where $\widehat{Y}$ is the estimate of the total, $\mathrm{Std}(\widehat{Y})$ is the estimated standard deviation, and $t_{df, \alpha/2}$ is the $100(1 - \alpha/2)$th percentile of the $t$ distribution with $df$ calculated as described in the section "*t* Test for the Mean" on page 6490.

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

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

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

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

## 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

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

The requested statistics for variable $y$ in domain $D$ are computed by using the new weights $v$.

### Domain Mean

The estimated mean of $y$ in the domain $D$ is

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

where

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

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 $n_h > 1$,

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

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

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

and if $n_h = 1$,

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

### Domain Total

The estimated total in domain $D$ is

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

and its estimated variance is

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

where if $n_h > 1$,

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

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

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

and if $n_h = 1$,

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

## Quantiles

Let Y be the variable of interest in a complex survey. Denote $F(t) = Pr(Y \leq t)$ as the cumulative distribution for $Y$.

For $0 < p < 1$, the $p$th quantile of the population cumulative distribution function is

$$Y_p = \inf\{y : F(y) \geq p\}$$

### Estimate of Quantile

Let $\{y_{hij}, w_{hij}\}$ be the observed values for variable $Y$ associated with sampling weights, where $(h, i, j)$ are the stratum index, cluster index, and member index, respectively, as shown in the section "Definitions and Notation" on page 6486. Let $y_{(1)} < y_{(2)} < ... < y_{(d)}$ denote the sample order statistics for variable $Y$.

An estimate of quantile $Y_p$ is

$$\hat{Y}_p = \begin{cases} y_{(1)} & \text{if } p < \hat{F}(y_{(1)}) \\ y_{(k)} + \dfrac{p - \hat{F}(y_{(k)})}{\hat{F}(y_{(k+1)}) - \hat{F}(y_{(k)})}(y_{(k+1)} - y_{(k)}) & \text{if } \hat{F}(y_{(k)}) \leq p < \hat{F}(y_{(k+1)}) \\ y_{(d)} & \text{if } p = 1 \end{cases}$$

where $\hat{F}(t)$ is the estimated cumulative distribution for $Y$:

$$\hat{F}(t) = \frac{\sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij} I(y_{hij} \leq t)}{\sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij}}$$

and $I(\cdot)$ is the indicator function.

### Standard Error

PROC SURVEYMEANS uses Woodruff method (Dorfman and Valliant 1993; Särndal, Swensson, and Wretman 1992; Francisco and Fuller 1991) to estimate the variances of quantiles. This method first constructs a confidence interval on a quantile. Then it uses the width of the confidence interval to estimate the standard error of a quantile.

In order to estimate the variance for $\hat{Y}_p$, first the procedure estimates the variance of the estimated distribution function $\hat{F}(\hat{Y}_p)$ by

$$\hat{V}(\hat{F}(\hat{Y}_p)) = \sum_{h=1}^{H} \frac{n_h(1 - f_h)}{n_h - 1} \sum_{i=1}^{n_h} (d_{hi\cdot} - \bar{d}_{h\cdot\cdot})^2$$

where

$$d_{hi\cdot} = \left( \sum_{j=1}^{m_{hi}} w_{hij} \left( I(y_{hij} \leq \hat{Y}_p) - \hat{F}(\hat{Y}_p) \right) \right) / w_{\cdots}$$

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

Then $100(1 - \alpha)\%$ confidence limits of $\hat{F}(\hat{Y}_p)$ can be constructed by

$$(\hat{p}_L, \quad \hat{p}_U) = \left( \hat{F}(\hat{Y}_p) - t_{df, \alpha/2} \sqrt{\hat{V}(\hat{F}(\hat{Y}_p))}, \quad \hat{F}(\hat{Y}_p) + t_{df, \alpha/2} \sqrt{\hat{V}(\hat{F}(\hat{Y}_p))} \right)$$

where $t_{df, \alpha/2}$ is the $100(1 - \alpha/2)$th percentile of the $t$ distribution with $df$ degrees of freedom, described in the section "Degrees of Freedom" on page 6490.

When $(\hat{p}_L, \hat{p}_U)$ is out of the range of [0,1], the procedure does not compute the standard error.

The $\hat{p}_L$th quantile is defined as

$$\hat{Y}_{\hat{p}_L} = \begin{cases} y_{(1)} & \text{if } \hat{p}_L < \hat{F}(y_{(1)}) \\ y_{(k_L)} + \dfrac{\hat{p}_L - \hat{F}(y_{(k_L)})}{\hat{F}(y_{(k_L+1)}) - \hat{F}(y_{(k_L)})}(y_{(k_L+1)} - y_{(k_L)}) & \text{if } \hat{F}(y_{(k_L)}) \le \hat{p}_L < \hat{F}(y_{(k_L+1)}) \\ y_{(d)} & \text{if } \hat{p}_L = 1 \end{cases}$$

and the $\hat{p}_U$th quantile is defined as

$$\hat{Y}_{\hat{p}_U} = \begin{cases} y_{(1)} & \text{if } \hat{p}_U < \hat{F}(y_{(1)}) \\ y_{(k_U)} + \dfrac{\hat{p}_U - \hat{F}(y_{(k_U)})}{\hat{F}(y_{(k_U+1)}) - \hat{F}(y_{(k_U)})}(y_{(k_U+1)} - y_{(k_U)}) & \text{if } \hat{F}(y_{(k_U)}) \le \hat{p}_U < \hat{F}(y_{(k_U+1)}) \\ y_{(d)} & \text{if } \hat{p}_U = 1 \end{cases}$$

The standard error of $\hat{Y}_p$ then is estimated by

$$sd(\hat{Y}_p) = \frac{\hat{Y}_{\hat{p}_U} - \hat{Y}_{\hat{p}_L}}{2t_{df, \alpha/2}}$$

where $t_{df, \alpha/2}$ is the $100(1 - \alpha/2)$th percentile of the $t$ distribution with $df$ degrees of freedom.

### Confidence Limits

Symmetric $100(1 - \alpha)\%$ confidence limits are computed as

$$\left( \hat{Y}_p - sd(\hat{Y}_p) \cdot t_{df, \alpha/2}, \quad \hat{Y}_p + sd(\hat{Y}_p) \cdot t_{df, \alpha/2} \right)$$

If you specify the NONSYMCL option in the PROC statement, the procedure computes $100(1 - \alpha)\%$ nonsymmetric confidence limits:

$$\left( \hat{Y}_{\hat{p}_L}, \quad \hat{Y}_{\hat{p}_U} \right)$$

# 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

The balanced repeated replication (BRR) method requires that the full sample be drawn by using a stratified sample design with two primary sampling units (PSUs) per stratum. Let $H$ be the total number of strata. The total number of replicates $R$ is the smallest multiple of 4 that is greater than $H$. However, if you prefer a larger number of replicates, you can specify the REPS=*number* option. If a $number \times number$ Hadamard matrix cannot be constructed, the number of replicates is increased until a Hadamard matrix becomes available.

Each replicate is obtained by deleting one PSU per stratum according to the corresponding Hadamard matrix and adjusting the original weights for the remaining PSUs. The new weights are called replicate weights.

Replicates are constructed by using the first $H$ columns of the $R \times R$ Hadamard matrix. The $r$th $(r = 1, 2, ..., R)$ replicate is drawn from the full sample according to the $r$th row of the Hadamard matrix as follows:

- If the $(r, h)$th element of the Hadamard matrix is 1, then the first PSU of stratum $h$ is included in the $r$th replicate and the second PSU of stratum $h$ is excluded.

- If the $(r, h)$th element of the Hadamard matrix is $-1$, then the second PSU of stratum $h$ is included in the $r$th replicate and the first PSU of stratum $h$ is excluded.

The replicate weights of the remaining PSUs in each half sample are then doubled to their original weights. For more details about the BRR method, see Wolter (1985) and Lohr (1999).

By default, an appropriate Hadamard matrix is generated automatically to create the replicates. You can request that the Hadamard matrix be displayed by specifying the VARMETHOD=BRR(PRINTH) method-option. If you provide a Hadamard matrix by specifying the VARMETHOD=BRR(HADAMARD=) method-option, then the replicates are generated according to the provided Hadamard matrix.

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

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

with $H$ degrees of freedom, where $H$ is the number of strata.

You can use the VARMETHOD=BRR(OUTWEIGHTS=) method-option to save the replicate weights into a SAS data set.

## Fay's BRR Method

Fay's method is a modification of the BRR method, and it requires a stratified sample design with two primary sampling units (PSUs) per stratum. The total number of replicates $R$ is the smallest multiple of 4 that is greater than the total number of strata $H$. However, if you prefer a larger number of replicates, you can specify the REPS= method-option.

For each replicate, Fay's method uses a Fay coefficient $0 \leq \epsilon < 1$ to impose a perturbation of the original weights in the full sample that is gentler than using only half samples, as in the traditional BRR method. The Fay coefficient $0 \leq \epsilon < 1$ can be optionally set by the FAY $= \epsilon$ method-option. By default, $\epsilon = 0.5$ if only the FAY method-option is used without specifying a value for $\epsilon$ (Judkins 1990; Rao and Shao 1999). When $\epsilon = 0$, Fay's method becomes the traditional BRR method. For more details, see Dippo, Fay, and Morganstein (1984), Fay (1984), Fay (1989), and Judkins (1990).

Let $H$ be the number of strata. Replicates are constructed by using the first $H$ columns of the $R \times R$ Hadamard matrix, where $R$ is the number of replicates, $R > H$. The $r$th ($r = 1, 2, ..., R$) replicate is created from the full sample according to the $r$th row of the Hadamard matrix as follows:

- If the $(r, h)$th element of the Hadamard matrix is 1, then the full sample weight of the first PSU in stratum $h$ is multiplied by $\epsilon$ and that of the second PSU is multiplied by $2-\epsilon$ to obtain the $r$th replicate weights.

- If the $(r, h)$th element of the Hadamard matrix is $-1$, then the full sample weight of the first PSU in stratum $h$ is multiplied by $2-\epsilon$ and that of the second PSU is multiplied by $\epsilon$ to obtain the $r$th replicate weights.

You can use the VARMETHOD=BRR(OUTWEIGHTS=) method-option to save the replicate weights into a SAS data set.

By default, an appropriate Hadamard matrix is generated automatically to create the replicates. You can request that the Hadamard matrix be displayed by specifying the VARMETHOD=BRR(PRINTH) method-option. If you provide a Hadamard matrix by specifying the VARMETHOD=BRR(HADAMARD=) method-option, then the replicates are generated according to the provided Hadamard matrix.

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

SURVEYMEANS estimates the variance of $\hat{\theta}$ by

$$\widehat{V}(\hat{\theta}) = \frac{1}{R(1-\epsilon)^2} \sum_{r=1}^{R} \left(\hat{\theta}_r - \hat{\theta}\right)^2$$

with $H$ degrees of freedom, where $H$ is the number of strata.

## 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* $\alpha_r$. 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 $\alpha_r$ are the same for all replicates:

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

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

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

Denote the original weight in the full sample for the $j$th member of the $i$th PSU as $w_{ij}$. If the $i$th PSU is included in the $r$th replicate ($r = 1, 2, ..., R$), then the corresponding replicate weight for the $j$th member of the $i$th 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 jack-knife 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 $\theta$ is a population parameter of interest. Let $\hat{\theta}$ be the estimate from the full sample for $\theta$. Let $\hat{\theta}_r$ be the estimate from the $r$th replicate subsample by using replicate weights. PROC SURVEYMEANS estimates the variance of $\hat{\theta}$ by

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

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.

## Hadamard Matrix

A Hadamard matrix $\mathbf{H}$ is a square matrix whose elements are either 1 or $-1$ such that

$$\mathbf{H}\mathbf{H}' = k\mathbf{I}$$

where $k$ is the dimension of $\mathbf{H}$ and $\mathbf{I}$ is the identity matrix of order $k$. The order $k$ is necessarily 1, 2, or a positive integer that is a multiple of 4.

For example, the following matrix is a Hadamard matrix of dimension $k = 8$:

```
1    1    1    1    1    1    1    1
1   -1    1   -1    1   -1    1   -1
1    1   -1   -1    1    1   -1   -1
1   -1   -1    1    1   -1   -1    1
1    1    1    1   -1   -1   -1   -1
1   -1    1   -1   -1    1   -1    1
1    1   -1   -1   -1   -1    1    1
1   -1   -1    1   -1    1    1   -1
```

## Output Data Sets

You can use the Output Delivery System to create a SAS data set from any piece of PROC SURVEYMEANS output. See the section "ODS Table Names" on page 6509 for more information.

PROC SURVEYMEANS also provides an output data set that stores the replicate weights for BRR or jackknife variance estimation and an output data set that stores the jackknife coefficients for jackknife variance estimation.

### Replicate Weights Output Data Set

If you specify the OUTWEIGHTS= method-option for BRR or jackknife method in the VARMETHOD= option, PROC SURVEYMEANS stores the replicate weights in an output data

set. The OUTWEIGHTS= output data set contains all observations used in the analysis or all valid observations in the DATA= input data set. (A valid observation is an observation that has a positive value of the WEIGHT variable. Valid observations must also have nonmissing values of the STRATA and CLUSTER variables, unless you specify the MISSING option. See the section "Data and Sample Design Summary" on page 6504 for details about valid observations.)

The OUTWEIGHTS= data set contains the following variables:

- all variables in the DATA= input data set

- RepWt_1, RepWt_2, . . ., RepWt_n, which are the replicate weight variables

where $n$ is the total number of replicates in the analysis. Each replicate weight variable contains the replicate weights for the corresponding replicate. Replicate weights equal zero for those observations not included in the replicate.

After the procedure creates replicate weights for a particular input data set and survey design, you can use the OUTWEIGHTS= method-option to store these replicate weights and then use them again in subsequent analyses, either in PROC SURVEYMEANS or in the other survey procedures. You can use the REPWEIGHTS statement to provide replicate weights for the procedure.

## Jackknife Coefficients Output Data Set

If you specify the OUTJKCOEFS= method-option for VARMETHOD=JACKKNIFE, PROC SURVEYMEANS stores the jackknife coefficients in an output data set. The OUTJKCOEFS= output data set contains one observation for each replicate. The OUTJKCOEFS= data set contains the following variables:

- Replicate, which is the replicate number for the jackknife coefficient

- JKCoefficient, which is the jackknife coefficient

- DonorStratum, which is the stratum of the PSU that was deleted to construct the replicate, if you specify a STRATA statement

After the procedure creates jackknife coefficients for a particular input data set and survey design, you can use the OUTJKCOEFS= method-option to store these coefficients and then use them again in subsequent analyses, either in PROC SURVEYMEANS or in the other survey procedures. You can use the JKCOEFS= option in the REPWEIGHTS statement to provide jackknife coefficients for the procedure.

## Rectangular and Stacking Structures in an Output Data Set

When you use an ODS output statement to create SAS data sets for certain tables in PROC SURVEYMEANS, there are two types of table structure for a data set: *rectangular* and *stacking*. A rectangular structure creates one observation for each analysis variable in the data set. A stacking structure creates only one observation in the output data set for all analysis variables.

Before SAS 9, the stacking table structure, similar to the table structure in PROC MEANS, was the default in PROC SURVEYMEANS. Since SAS 9, the new default is to produce a rectangular table in the output data sets. You can use the STACKING option to request that the procedure produce the output data sets by using a stacking table structure.

The STACKING option affects the following tables:

- Domain

- Ratio

- Statistics

- StrataInfo

Figure 85.6 and Figure 85.7 shows these two structures for analyzing the following data set:

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

The following statements request the default rectangular structure of the output data set for the statistics table:

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

Figure 85.6 shows the rectangular structure.

**Figure 85.6**  Rectangular Structure in the Output Data Set

```
         Rectangular Structure in the Output Data Set

            Var       Var
   Obs      Name      Level          Mean          StdErr

    1       x                     12.800000        2.905168
    2       sex        F           0.600000        0.244949
    3       sex        M           0.400000        0.244949
```

The following statements specify the STACKING option to request that the output data set have a stacking structure:

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

Figure 85.7 shows the stacking structure of the output data set for the statistics table requested by the STACKING option.

**Figure 85.7** Stacking Structure in the Output Data Set Requested by the STACKING option

```
              Stacking Structure in the Output Data Set

    Obs    x             x_Mean        x_StdErr     sex_F        sex_F_Mean

     1     x         12.800000        2.905168      sex=F         0.600000

    Obs    sex_F_StdErr    sex_M       sex_M_Mean    sex_M_StdErr

     1         0.244949    sex=M        0.400000        0.244949
```

# Displayed Output

The SURVEYMEANS procedure produces the following output.

## Data and Sample Design Summary

The "Data Summary" table provides information about the input data set and the sample design. This table displays the total number of valid observations, where an observation is considered *valid* if it has nonmissing values for all procedure variables other than the analysis variables—that is, for all specified STRATA, CLUSTER, and WEIGHT variables. This number might differ from the number of nonmissing observations for an individual analysis variable, which the procedure displays in the "Statistics" table. See the section "Missing Values" on page 6483 for more information.

PROC SURVEYMEANS displays the following information in the "Data Summary" table:

- Number of Strata, if you specify a STRATA statement

- Number of Clusters, if you specify a CLUSTER statement

- Number of Observations, which is the total number of valid observations

- Sum of Weights, which is the sum over all valid observations, if you specify a WEIGHT statement

## Class Level Information

If you use a CLASS statement to name classification variables for categorical analysis, or if you list any character variables in the VAR statement, then PROC SURVEYMEANS displays a "Class Level Information" table. This table contains the following information for each classification variable:

- Class Variable, which lists each CLASS variable name

- Levels, which is the number of values or levels of the classification variable

- Values, which lists the values of the classification variable. The values are separated by a white space character; therefore, to avoid confusion, you should not include a white space character within a classification variable value.

## Stratum Information

If you specify the LIST option in the STRATA statement, PROC SURVEYMEANS displays a "Stratum Information" table. This table displays the number of valid observations in each stratum, as well as the number of nonmissing stratum observations for each analysis variable. The "Stratum Information" table provides the following for each stratum:

- Stratum Index, which is a sequential stratum identification number

- STRATA variable(s), which lists the levels of STRATA variables for the stratum

- Population Total, if you specify the TOTAL= option

- Sampling Rate, if you specify the TOTAL= or RATE= option. If you specify the TOTAL= option, the sampling rate is based on the number of valid observations in the stratum.

- N Obs, which is the number of valid observations

- Variable, which lists each analysis variable name

- Levels, which identifies each level for categorical variables

- N, which is the number of nonmissing observations for the analysis variable

- Clusters, which is the number of clusters, if you specify a CLUSTER statement

## Variance Estimation

If the variance method is not Taylor series or if the NOMCAR option is used, by default, PROC SURVEYMEANS displays the following variance estimation specifications in the "Variance Estimation" table:

- Method, which is the variance estimation method

- Number of Replicates, which is the number of replicates if you specify the VARMETHOD=BRR or VARMETHOD=JACKKNIFE option

- Hadamard Data Set, which is the name of the SAS data set for the HADAMARD matrix if you specify the VARMETHOD=BRR(HADAMARD=) method-option

- Fay Coefficient, which is the value of the FAY coefficient if you specify the VARMETHOD=BRR(FAY) method-option

- Replicate Weights Data Set, which is the name of the SAS data set that contains the replicate weights

- Missing Levels Included (MISSING), if you specify the MISSING option

- Missing Levels Include (NOMCAR), if you specify the NOMCAR option

## Statistics

The "Statistics" table displays all of the statistics that you request with *statistic-keywords* in the PROC SURVEYMEANS statement, except DECILES, MEDIAN, Q1, Q3, and QUARTILES, which are displayed in the "Quantiles" table. If you do not specify any *statistic-keywords*, then by default this table displays the following information for each analysis variable: the sample size, the mean, the standard error of the mean, and the confidence limits for the mean. The "Statistics" table might contain the following information for each analysis variable, depending on which *statistic-keywords* you request:

- Variable name

- Variable Label

- Level, which identifies each level for categorical variables

- N, which is the number of nonmissing observations

- N Miss, which is the number of missing observations

- Minimum

- Maximum

- Range

- Number of Clusters

- Sum of Weights

- DF, which is the degrees of freedom for the $t$ test

- Mean

- Std Error of Mean, which is the standard error of the mean

- Var of Mean, which is the variance of the mean

- $t$ Value, for testing $H_0$ : population MEAN $= 0$

- Pr $> \mid t \mid$, which is the two-sided $p$-value for the $t$ test

- $100(1 - \alpha)\%$ CL for Mean, which are two-sided confidence limits for the mean

- $100(1 - \alpha)\%$ Upper CL for Mean, which is a one-sided upper confidence limit for the mean

- $100(1 - \alpha)\%$ Lower CL for Mean, which is a one-sided lower confidence limit for the mean

- Coeff of Variation, which is the coefficients of variation for the mean and the sum

- Sum

- Std Dev, which is the standard deviation of the sum

- Var of Sum, which is the variance of the sum

- $100(1 - \alpha)\%$ CL for Sum, which are two-sided confidence limits for the sum

- $100(1 - \alpha)\%$ Upper CL for Sum, which is a one-sided upper confidence limit for the sum

- $100(1 - \alpha)\%$ Lower CL for Sum, which is a one-sided lower confidence limit for the Sum

## Quantiles

The "Quantiles" table displays all the quantiles that you request with either *statistic-keywords* such as DECILES, MEDIAN, Q1, Q3, and QUARTILES, or the PERCENTILE= option, or the QUANTILE= option in the PROC SURVEYMEANS statement.

The "Quantiles" table contains the following information for each quantile:

- Variable name

- Variable Label

- Percentile, which is the requested quantile in the format of %

- Percentile Label, which is the corresponding common name for a percentile if it exists—for example, *Median* for 50th percentile

- Estimate, which is the estimate for a requested quantile with respect to the population distribution

- StdErr, which is the standard error of the quantile

- $100(1 - \alpha)\%$ Confidence Limits, which are two-sided confidence limits for the quantile

## Domain Analysis

If you use a DOMAIN statement, the procedure displays statistics in each domain in a "Domain Analysis" table. A "Domain Analysis" table contains all the columns in the "Statistics" table, plus columns of domain variable values.

Note that depending on how you define the domains with domain variables, the procedure might produce more than one "Domain Analysis" table. For example, in the DOMAIN statement

```
domain A B*C*D A*C C;
```

you use four definitions to define domains:

- A: all the levels of A

- C: all the levels of C

- A*C: all the interactive levels of A and C

- B*C*D: all the interactive levels of B, C, and D

The procedure displays four "Domain Analysis" tables, one for each domain definition. If you use an ODS OUTPUT statement to create an output data set for domain analysis, the output data set contains a variable Domain whose values are these domain definitions.

## Ratio Analysis

The "Ratio Analysis" table displays all of the statistics that you request with *statistic-keywords* in the PROC statement. If you do not specify any *statistic-keywords*, then by default this table displays the ratio and its standard error. The "Ratio Analysis" table might contain the following information for each ratio, depending on which *statistic-keywords* you request:

- Numerator, which identifies the numerator variable of the ratio

- Denominator, which identifies the denominator variable of the ratio

- N, which is the number of observations used in the ratio analysis

- number of Clusters

- Sum of Weights

- DF, which is the degrees of freedom for the $t$ test

- Ratio

- Std Error of Ratio, which is the standard error of the ratio

- Var of Ratio, which is the variance of the ratio

- $t$ Value, for testing $H_0$ : population RATIO $= 0$

- Pr $> | t |$, which is the two-sided $p$-value for the $t$ test

- $100(1 - \alpha)\%$ CL for Ratio, which are two-sided confidence limits for the Ratio

- Upper $100(1 - \alpha)\%$ CL for Ratio, which are one-sided upper confidence limits for the Ratio

- Lower $100(1 - \alpha)\%$ CL for Ratio, which are one-sided lower confidence limits for the Ratio

When you use the ODS OUTPUT statement to create an output data set, if you use labels for your RATIO statement, these labels are saved in the variable Ratio Statement in the output data set.

## Hadamard Matrix

If you specify the VARMETHOD=BRR(PRINTH) method-option in the PROC statement, PROC SURVEYMEANS displays the Hadamard matrix.

When you provide a Hadamard matrix with the VARMETHOD=BRR(HADAMARD=) method-option, the procedure displays only used rows and columns of the Hadamard matrix.

## ODS Table Names

PROC SURVEYMEANS assigns a name to each table it creates. You can use these names to reference the table when using the Output Delivery System (ODS) to select tables and create output data sets. These names are listed in Table 85.3. For more information about ODS, see Chapter 20, "Using the Output Delivery System."

**Table 85.3**  ODS Tables Produced by PROC SURVEYMEANS

| ODS Table Name | Description | Statement | Option |
|---|---|---|---|
| ClassVarInfo | Class level information | CLASS | default |
| Domain | Statistics in domains | DOMAIN | default |
| HadamardMatrix | Hadamard matrix | PROC | PRINTH |
| Ratio | Statistics for ratios | RATIO | default |
| Quantiles | Quantiles | PROC | default |
| Statistics | Statistics | PROC | default |
| StrataInfo | Stratum information | STRATA | LIST |
| Summary | Data summary | PROC | default |
| VarianceEstimation | Variance estimation | PROC | default |

For example, the following statements create an output data set MyStrata, which contains the

"StrataInfo" table, and an output data set MyStat, which contains the "Statistics" table for the ice cream study discussed in the section "Stratified Sampling" on page 6461:

```
title1 'Analysis of Ice Cream Spending';
title2 'Stratified Simple Random Sample Design';
proc surveymeans data=IceCream total=StudentTotals;
   strata Grade / list;
   var Spending Group;
   weight Weight;
   ods output StrataInfo = MyStrata
              Statistics = MyStat;
run;
```

# Examples: SURVEYMEANS Procedure

The section "Getting Started: SURVEYMEANS Procedure" on page 6459 contains examples of analyzing data from simple random sampling and stratified simple random sample designs. This section provides more examples that illustrate how to use PROC SURVEYMEANS.

## Example 85.1: Stratified Cluster Sample Design

Consider the example in the section "Stratified Sampling" on page 6461. The study population is a junior high school with a total of 4,000 students in grades 7, 8, and 9. Researchers want to know how much these students spend weekly for ice cream, on the average, and what percentage of students spend at least $10 weekly for ice cream.

The example in the section "Stratified Sampling" on page 6461 assumes that the sample of students was selected using a stratified simple random sample design. This example shows analysis based on a more complex sample design.

Suppose that every student belongs to a study group and that study groups are formed within each grade level. Each study group contains between two and four students. Table 85.4 shows the total number of study groups for each grade.

**Table 85.4**   Study Groups and Students by Grade

| Grade | Number of Study Groups | Number of Students |
|-------|------------------------|--------------------|
| 7     | 608                    | 1,824              |
| 8     | 252                    | 1,025              |
| 9     | 403                    | 1,151              |
| Total | 617                    | 4,000              |

It is quicker and more convenient to collect data from students in the same study group than to

collect data from students individually. Therefore, this study uses a stratified clustered sample design. The primary sampling units, or clusters, are study groups. The list of all study groups in the school is stratified by grade level. From each grade level, a sample of study groups is randomly selected, and all students in each selected study group are interviewed. The sample consists of eight study groups from the 7th grade, three groups from the 8th grade, and five groups from the 9th grade.

The SAS data set IceCreamStudy saves the responses of the selected students:

```
data IceCreamStudy;
   input Grade StudyGroup Spending @@;
   if (Spending < 10) then Group='less';
      else Group='more';
   datalines;
7   34   7      7   34   7      7 412   4      9   27 14
7   34   2      9 230 15      9   27 15      7 501   2
9 230   8      9 230   7      7 501   3      8   59 20
7 403   4      7 403 11      8   59 13      8   59 17
8 143 12      8 143 16      8   59 18      9 235   9
8 143 10      9 312   8      9 235   6      9 235 11
9 312 10      7 321   6      8 156 19      8 156 14
7 321   3      7 321 12      7 489   2      7 489   9
7   78   1      7   78 10      7 489   2      7 156   1
7   78   6      7 412   6      7 156   2      9 301   8
;
```

In the data set IceCreamStudy, the variable Grade contains a student's grade. The variable Study-Group identifies a student's study group. It is possible for students from different grades to have the same study group number because study groups are sequentially numbered within each grade. The variable Spending contains a student's response regarding how much he spends per week for ice cream, in dollars. The variable GROUP indicates whether a student spends at least $10 weekly for ice cream. It is not necessary to store the data in order of grade and study group.

The SAS data set StudyGroup is created to provide PROC SURVEYMEANS with the sample design information shown in Table 85.4:

```
data StudyGroups;
   input Grade _total_; datalines;
7 608
8 252
9 403
;
```

The variable Grade identifies the strata, and the variable _TOTAL_ contains the total number of study groups in each stratum. As discussed in the section "Specification of Population Totals and Sampling Rates" on page 6484, the population totals stored in the variable _TOTAL_ should be expressed in terms of the primary sampling units (PSUs), which are study groups in this example. Therefore, the variable _TOTAL_ contains the total number of study groups for each grade, rather than the total number of students.

In order to obtain unbiased estimates, you create sampling weights by using the following SAS statements:

```
data IceCreamStudy;
   set IceCreamStudy;
   if Grade=7 then Prob=8/608;
   if Grade=8 then Prob=3/252;
   if Grade=9 then Prob=5/403;
   Weight=1/Prob;
```

The sampling weights are the reciprocals of the probabilities of selections. The variable Weight contains the sampling weights. Because the sampling design is clustered and all students from each selected cluster are interviewed, the sampling weights equal the inverse of the cluster (or study group) selection probabilities.

The following SAS statements perform the analysis for this sample design:

```
title1 'Analysis of Ice Cream Spending';
title2 'Stratified Clustered Sample Design';
proc surveymeans data=IceCreamStudy total=StudyGroups;
   strata Grade / list;
   cluster StudyGroup;
   var Spending Group;
   weight Weight;
run;
```

Output 85.1.1 provides information about the sample design and the input data set. There are three strata in the sample design, and the sample contains 16 clusters and 40 observations. The variable Group has two levels, 'less' and 'more.'

**Output 85.1.1** Data Summary and Class Information

```
                    Analysis of Ice Cream Spending
                   Stratified Clustered Sample Design

                       The SURVEYMEANS Procedure

                             Data Summary

                  Number of Strata                    3
                  Number of Clusters                 16
                  Number of Observations             40
                  Sum of Weights                 3162.6

                       Class Level Information

                  Class
                  Variable      Levels    Values

                  Group              2    less more
```

Output 85.1.2 displays information for each stratum. Since the primary sampling units in this design are study groups, the population totals shown in Output 85.1.2 are the total numbers of study groups for each stratum or grade. This differs from Output 85.3, which provides the population totals in

terms of students since students were the primary sampling units for that design. Output 85.1.2 also displays the number of clusters for each stratum and analysis variable.

**Output 85.1.2** Stratum Information

```
                        Stratum Information

   Stratum              Population  Sampling
   Index     Grade        Total       Rate         N Obs   Variable   Level
   ---------------------------------------------------------------------------
      1         7            608      1.32%           20    Spending
                                                           Group        less
                                                                        more
      2         8            252      1.19%            9    Spending
                                                           Group        less
                                                                        more
      3         9            403      1.24%           11    Spending
                                                           Group        less
                                                                        more
   ---------------------------------------------------------------------------

                        Stratum Information

Stratum                 Population  Sampling
   Index     Grade        Total       Rate         N Obs   Variable              N
   ---------------------------------------------------------------------------
      1         7            608      1.32%           20    Spending             20
                                                           Group                 17
                                                                                  3
      2         8            252      1.19%            9    Spending              9
                                                           Group                  0
                                                                                  9
      3         9            403      1.24%           11    Spending             11
                                                           Group                  6
                                                                                  5
   ---------------------------------------------------------------------------

                        Stratum Information

Stratum                 Population  Sampling
   Index     Grade        Total       Rate         N Obs   Variable       Clusters
   ---------------------------------------------------------------------------
      1         7            608      1.32%           20    Spending              8
                                                           Group                  8
                                                                                  3
      2         8            252      1.19%            9    Spending              3
                                                           Group                  0
                                                                                  3
      3         9            403      1.24%           11    Spending              5
                                                           Group                  4
                                                                                  4
   ---------------------------------------------------------------------------
```

Output 85.1.3 displays the estimates of the average weekly ice cream expenditure and the percentage of students spending at least $10 weekly for ice cream.

**Output 85.1.3** Statistics

```
                                  Statistics

                                            Std Error
Variable  Level              N          Mean     of Mean      95% CL for Mean
-----------------------------------------------------------------------------
Spending                    40      8.923860     0.650859   7.51776370 10.3299565
Group     less              23      0.561437     0.056368   0.43966057  0.6832130
          more              17      0.438563     0.056368   0.31678698  0.5603394
-----------------------------------------------------------------------------
```

# Example 85.2: Domain Analysis

Suppose that you are studying profiles of 800 top-performing companies to provide information about their impact on the economy. You are also interested in the company profiles within each market type. A sample of 66 companies is selected with unequal probability across market types. However, market type is not included in the sample design. Thus, the number of companies within each market type is a random variable in your sample. To obtain statistics within each market type, you should use domain analysis. The data of the 66 companies are saved in the following data set.

```
data Company;
   length Type $14;
   input Type$ Asset Sale Value Profit Employee Weight;
   datalines;
Other           2764.0   1828.0   1850.3    144.0    18.7    9.6
Energy         13246.2   4633.5   4387.7    462.9    24.3   42.6
Finance         3597.7    377.8     93.0     14.0     1.1   12.2
Transportation  6646.1   6414.2   2377.5    348.2    47.1   21.8
HiTech          1068.4   1689.8   1430.2     72.9     4.6    4.3
Manufacturing   1125.0   1719.4   1057.5     98.1    20.4    4.5
Other           1459.0   1241.4    452.7     24.5    20.1    5.5
Finance         2672.3    262.5    296.2     23.1     2.2    9.3
Finance          311.0    566.2    932.0     52.8     2.7    1.9
Energy          1148.6   1014.6    485.1     60.6     4.0    4.5
Finance         5327.0    572.4    372.9     25.2     4.2   17.7
Energy          1602.7    678.4    653.0     75.6     2.8    6.0
Energy          5808.8   1288.4   2007.0    318.8     5.9   19.2
Medical          268.8    204.4    820.9     45.6     3.7    1.8
Transportation  5222.6   2627.8   1910.0    245.6    22.8   17.4
Other            872.7   1419.4    939.3     69.7    12.2    3.7
Retail          4461.7   8946.8   4662.7    289.0   132.1   15.0
HiTech          6719.2   6942.0   8240.2    381.3    85.8   22.1
Retail           833.4   1538.8   1090.3     64.9    15.4    3.5
Finance          415.9    167.3   1126.8     56.8     0.7    2.2
HiTech           442.4   1139.9   1039.9     57.6    22.7    2.3
Other            801.5   1157.0    664.2     56.9    15.5    3.4
Finance         4954.8    468.8    366.4     41.7     3.0   16.5
Finance         2661.9    257.9    181.1     21.2     2.1    9.3
```

```
Finance          5345.8    530.1    337.4     36.4     4.3   17.8
Energy           3334.3   1644.7   1407.8    157.6     6.4   11.4
Manufacturing    1826.6   2671.7    483.2     71.3    25.3    6.7
Retail            618.8   2354.7    767.7     58.6    19.0    2.9
Retail           1529.1   6534.0    826.3     58.3    65.8    5.7
Manufacturing    4458.4   4824.5   3132.1     28.9    67.0   15.0
HiTech           5831.7   6611.1   9464.7    459.6    86.7   19.3
Medical          6468.3   4199.2   3170.4    270.1    59.5   21.3
Energy           1720.7    473.1    811.1     86.6     1.6    6.3
Energy           1679.7   1379.9    721.1     91.8     4.5    6.2
Retail           4018.2 16823.4   2038.3    178.1   162.0   13.6
Other             227.1    575.8   1083.8     62.6     1.9    1.6
Finance          3872.8    362.0    209.3     27.6     2.4   13.1
Retail           3359.3   4844.7   2651.4    224.1    75.6   11.5
Energy           1295.6    356.9    180.8    162.3     0.6    5.0
Energy           1658.0    626.6    688.0    126.0     3.5    6.1
Finance         12156.7   1345.5    680.7    106.6     9.4   39.2
HiTech           3982.6   4196.0   3946.8    313.9    64.3   13.5
Finance          8760.7    886.4   1006.9     90.0     7.5   28.5
Manufacturing    2362.2   3153.3   1080.0    137.0    25.2    8.4
Transportation   2499.9   3419.0    992.6     47.2    25.3    8.8
Energy           1430.4   1610.0    664.3     77.7     3.5    5.4
Energy          13666.5 15465.4   2736.7    411.4    26.6   43.9
Manufacturing    4069.3   4174.7   2907.6    289.2    38.2   13.7
Energy           2924.7    711.9   1067.8    146.7     3.4   10.1
Transportation   1262.1   1716.0    364.3     71.2    14.5    4.9
Medical           684.4    672.9    287.4     61.8     6.0    3.1
Energy           3069.3   1719.0   1439.0    196.4     4.9   10.6
Medical           246.5    318.8    924.1     43.8     3.1    1.7
Finance         11562.2   1128.5    580.4     64.2     6.7   37.3
Finance          9316.0   1059.4    816.5     95.9     8.0   30.2
Retail           1094.3   3848.0    563.3     29.4    44.7    4.4
Retail           1102.1   4878.3    932.4     65.2    47.3    4.4
HiTech            466.4    675.8    845.7     64.5     5.2    2.4
Manufacturing   10839.4   5468.7   1895.4    232.8    47.8   35.0
Manufacturing     733.5   2135.3     96.6     10.9     2.7    3.2
Manufacturing   10354.2 14477.4   5607.2    321.9   188.5   33.5
Energy           1902.1   2697.9    329.3     34.2     2.2    6.9
Other            2245.2   2132.2   2230.4    198.9     8.0    8.0
Transportation    949.4   1248.3    298.9     35.4    10.4    3.9
Retail           2834.4   2884.6    458.2     41.2    49.8    9.8
Retail           2621.1   6173.8   1992.7    183.7   115.1    9.2
;
```

For each company in your sample, the variables are defined as follows:

- Type identifies the type of market for the company.

- Asset contains the company's assets, in millions of dollars.

- Sale contains sales, in millions of dollars.

- Value contains the market value of the company, in millions of dollars.

- Profit contains the profit, in millions of dollars.

- Employee contains the number of employees, in thousands.

- Weight contains the sampling weight.

The following SAS statements use PROC SURVEYMEANS to perform the domain analysis, estimating means, and other statistics for the overall population and also for the subpopulations (or domain) defined by market type. The DOMAIN statement specifies Type as the domain variable:

```
title 'Top Companies Profile Study';
proc surveymeans data=Company total=800 mean sum;
   var Asset Sale Value Profit Employee;
   weight Weight;
   domain Type;
run;
```

Output 85.2.1 shows that there are 66 observations in the sample. The sum of the sampling weights equals 799.8, which is close to the total number of companies in the study population.

**Output 85.2.1** Company Profile Study

```
                     Top Companies Profile Study

                     The SURVEYMEANS Procedure

                           Data Summary

               Number of Observations              66
               Sum of Weights                   799.8

                            Statistics

                            Std Error
    Variable           Mean        of Mean            Sum          Std Dev
    -------------------------------------------------------------------------
    Asset        6523.488510     720.557075        5217486          1073829
    Sale         4215.995799     839.132506        3371953           847885
    Value        2145.935121     342.531720        1716319           359609
    Profit        188.788210      25.057876         150993            30144
    Employee       36.874869       7.787857          29493      7148.003298
    -------------------------------------------------------------------------
```

The "Statistics" table in Output 85.2.1 displays the estimates of the mean and total for all analysis variables for the entire set of 800 companies, while Output 85.2.2 shows the mean and total estimates for each company type.

**Output 85.2.2** Domain Analysis for Company Profile Study

```
                        Domain Analysis: Type

                                   Std Error
Type            Variable       Mean    of Mean          Sum       Std Dev
-------------------------------------------------------------------------
Energy          Asset    7868.302932  1941.699163      1449341        785962
                Sale     5419.679099  2416.214417       998305        673373
                Value    2249.297177   520.295162       414321        213580
                Profit    289.564658    52.512141        53338         25927
                Employee   14.151194     3.974697  2606.650000   1481.777769
Finance         Asset    7890.190264  1057.185336      1855773        704506
                Sale      829.210502   115.762531       195030         74436
                Value     565.068197    76.964547       132904         48156
                Profit     63.716837    10.099341        14986   5801.108513
                Employee    5.806293     0.811555  1365.640000    519.658410
HiTech          Asset    5031.959781   732.436967       321542        183302
                Sale     5464.292019   731.296997       349168        196013
                Value    6707.828482  1194.160584       428630        249154
                Profit    346.407042    42.299004        22135         12223
                Employee   70.766980     8.683595  4522.010000   2524.778281
Manufacturing   Asset    7403.004250  1454.921083       888361        492577
                Sale     7207.638833  2112.444703       864917        501679
                Value    2986.442750   799.121544       358373        196979
                Profit    211.933583    39.993255        25432         13322
                Employee   83.314333    31.089019  9997.720000   6294.309490
Medical         Asset    5046.570609  1218.444638       140799        131942
                Sale     3313.219713   758.216303        92439         85655
                Value    2561.614695   530.802245        71469         64663
                Profit    218.682796    44.051447  6101.250000   5509.560969
                Employee   46.518996    11.135955  1297.880000   1213.651734
Other           Asset    1850.250000   338.128984        58838         31375
                Sale     1620.784906   168.686773        51541         24593
                Value    1432.820755   297.869828        45564         24204
                Profit    115.089937    27.970560  3659.860000   2018.201371
                Employee   14.306604     2.313733   454.950000    216.327710
Retail          Asset    2939.845750   393.692369       235188         94605
                Sale     7395.453500  1746.187580       591636        263263
                Value    2103.863125   529.756409       168309         78304
                Profit    157.171875    31.734253        12574   5478.281027
                Employee   93.624000    15.726743  7489.920000   3093.832061
Transportation  Asset    4712.047359   888.954411       267644        163516
                Sale     4030.233275  1015.555708       228917        142669
                Value    1703.330282   313.841326        96749         58947
                Profit    224.762324    56.168925        12767   8287.585418
                Employee   30.946303     6.786270  1757.750000   1066.586615
-------------------------------------------------------------------------
```

## Example 85.3: Ratio Analysis

Suppose you are interested in the profit per employee and the sale per employee among the 800 top-performing companies in the data in the previous example. The following SAS statements illustrate

how you can use PROC SURVEYMEANS to estimate these ratios:

```
title 'Ratio Analysis in Top Companies Profile Study';
proc surveymeans data=Company total=800 ratio;
   var Profit Sale Employee;
   weight Weight;
   ratio Profit Sale / Employee;
run;
```

The RATIO statement requests the ratio of the profit and the sales to the number of employees.

Output 85.3.1 shows the estimated ratios and their standard errors. Because the profit and the sales figures are in millions of dollars, and the employee numbers are in thousands, the profit per employee is estimated as \$5,120 with a standard error of \$1,059, and the sales per employee are \$114,332 with a standard error of \$20,503.

**Output 85.3.1** Estimate Ratios

```
           Ratio Analysis in Top Companies Profile Study

                      The SURVEYMEANS Procedure

                           Ratio Analysis

     Numerator Denominator          Ratio         Std Err
     ---------------------------------------------------------
     Sale      Employee        114.332497        20.502742
     Profit    Employee          5.119698         1.058939
     ---------------------------------------------------------
```

# Example 85.4: Analyzing Survey Data with Missing Values

As described in the section "Missing Values" on page 6483, the SURVEYMEANS procedure excludes an observation from the analysis if it has a missing value for the analysis variable or a nonpositive value for the WEIGHT variable.

However, if there is evidence indicating that the nonrespondents are different from the respondents for your study, you can use the NOMCAR option to compute descriptive statistics among respondents while still counting the number of nonrespondents.

We use the ice cream example in the section "Stratified Sampling" on page 6461 to illustrate how to perform similar analysis when there are missing values.

Suppose that some of the students failed to provide the amounts spent on ice cream, as shown in the following data set, IceCream:

```
data IceCream;
   input Grade Spending @@;
   if Grade=7 then Prob=20/1824;
   if Grade=8 then Prob=9/1025;
   if Grade=9 then Prob=11/1151;
   Weight=1/Prob;
   datalines;
7 7  7  7  8  .  9 10  7  .  7 10  7  3  8 20  8 19  7 2
7 .  9 15  8 16  7  6  7  6  7  6  9 15  8 17  8 14  9 .
9 8  9  7  7  3  7 12  7  4  9 14  8 18  9  9  7  2  7 1
7 4  7 11  9  8  8  .  8 13  7  .  9  .  9 11  7  2  7 9
;
data StudentTotals;
   input Grade _total_; datalines;
7 1824
8 1025
9 1151
;
```

Considering the possibility that those students who did not respond spend differently than those students who did respond, you can use the NOMCAR option to request the analysis to treat the respondents as a domain rather than exclude the nonrespondents.

The following SAS statements produce the desired analysis:

```
title 'Analysis of Ice Cream Spending';
proc surveymeans data=IceCream total=StudentTotals nomcar mean sum;
   strata Grade;
   var Spending;
   weight Weight;
run;
```

Output 85.4.1 summarize the analysis including the variance estimation method.

**Output 85.4.1** Analysis of Incomplete Ice Cream Data Excluding Observations with Missing Values

```
                    Analysis of Ice Cream Spending

                       The SURVEYMEANS Procedure

                             Data Summary

             Number of Strata                     3
             Number of Observations              40
             Sum of Weights                    4000


                         Variance Estimation

             Method                      Taylor Series
             Missing Values      Included (NOMCAR)
```

Output 85.4.2 shows the mean and total estimates when treating respondents as a domain in the student population. Although the point estimates are the same as the analysis without the NOMCAR option, for this particular example, the variance estimations are slightly higher when you assume that the missingness is not completely at random.

**Output 85.4.2** Analysis of Incomplete Ice Cream Data Excluding Observations with Missing Values

```
                          Statistics

                            Std Error
    Variable          Mean    of Mean           Sum        Std Dev
    ------------------------------------------------------------------
    Spending      9.770542   0.652347         32139    3515.126876
    ------------------------------------------------------------------
```

## Example 85.5: Variance Estimation Using Replication Methods

In order to improve service, the San Francisco Municipal Railway (MUNI) conducts a survey to estimated passenger's average waiting time for MUNI's subway system.

The study uses a stratified cluster sample design. Each MUNI subway line is a stratum. The subway lines included in the study are 'J-Church,' 'K-Ingleside,' 'L-Taraval,' 'M-Ocean View,' 'N-Judah,' and the street car 'F-Market & Wharves.' The MUNI vehicles in service for these lines during a day are primary sampling units. Within each stratum, two vehicles (PSUs) are randomly selected. Then the waiting times of passengers for a selected MUNI vehicle are collected.

Table 85.5 shows the number of passengers that are interviewed in each of the selected MUNI vehicles.

**Table 85.5** The Sample of the MUNI Waiting Time Study

| MUNI Line | Vehicle | Number of Passengers |
|---|---|---|
| F-Market & Wharves | 1 | 65 |
| | 2 | 102 |
| J-Church | 1 | 101 |
| | 2 | 142 |
| K-Ingleside | 1 | 145 |
| | 2 | 180 |
| L-Taraval | 1 | 135 |
| | 2 | 185 |
| M-Ocean View | 1 | 139 |
| | 2 | 203 |
| N-Judah | 1 | 306 |
| | 2 | 234 |

The collected data are saved in the SAS data set MUNIsurvey. The variable Line indicates which MUNI line a passenger is riding. The variable vehicle identifies the vehicle that a passenger is boarding. The variable Waittime is the time (in minutes) that a passenger waited. The variable weight contains the sampling weights, which are determined by selection probabilities within each stratum.

Output 85.5.1 displays the first 10 observations of the data set MUNIsurvey.

**Output 85.5.1** First 10 Observations in the Data Set from the MUNI Subway Survey

```
              MUNI subway passenger waiting time analysis


     Obs           line            vehicle    passenger    waittime    weight

      1     F-Market & Wharves        1           1           18         59
      2     F-Market & Wharves        1           2            0         59
      3     F-Market & Wharves        1           3           16         59
      4     F-Market & Wharves        1           4           13         59
      5     F-Market & Wharves        1           5            5         59
      6     F-Market & Wharves        1           6           13         59
      7     F-Market & Wharves        1           7            7         59
      8     F-Market & Wharves        1           8            5         59
      9     F-Market & Wharves        1           9           16         59
     10     F-Market & Wharves        1          10            5         59
```

Using the VARMETHOD=BRR option, the following SAS statements analyze the MUNI subway survey by using the BRR method to estimate the variance:

```
title 'MUNI passenger waiting time analysis using BRR';
proc surveymeans data=MUNIsurvey mean varmethod=brr mean clm;
   strata line;
   cluster vehicle;
   var waittime;
   weight weight;
run;
```

The STRATUM variable is line, which corresponds to MUNI lines. The two clusters within each stratum are identified by the variable vehicle. The sampling weights are stored in the variable weight. The mean and confident limits of passenger waiting time (in minutes) are requested statistics.

Output 85.5.2 summarizes the data and indicates that the variance estimation method is BRR with 8 replicates.

**Output 85.5.2** MUNI Passenger Waiting Time Analysis Using the BRR Method

```
            MUNI passenger waiting time analysis using BRR

                    The SURVEYMEANS Procedure

                         Data Summary

             Number of Strata                   6
             Number of Clusters                12
             Number of Observations          1937
             Sum of Weights                143040
```

**Output 85.5.2** *continued*

```
                         Variance Estimation

                 Method                           BRR
                 Number of Replicates              8
```

Output 85.5.3 reports that the average passenger waiting time for a MUNI vehicle is 7.33 minutes, with an estimated standard of 0.24 minutes, using the BRR method. The 95% confident limits for the mean are estimated as 6.75 to 7.91 minutes.

**Output 85.5.3** MUNI Passenger Waiting Time Analysis Using the BRR Method

```
                              Statistics

                                 Std Error
     Variable          Mean       of Mean      95% CL for Mean
     ------------------------------------------------------------------
     waittime        7.333012     0.237557    6.75172983 7.91429366
     ------------------------------------------------------------------
```

Alternatively, the variance can be estimated using the jackknife method if the VARMETHOD=JACKKNIFE option is used. The following SAS statements analyze the MUNI subway survey by using the jackknife method to estimate the variance:

```
title 'MUNI passenger waiting time analysis using jackknife';
proc surveymeans data=MUNIsurvey mean varmethod=jackknife mean clm;
   strata line;
   cluster vehicle;
   var waittime;
   weight weight;
run;
```

Output 85.5.4 summarizes the data and indicates that the variance estimation method is jackknife with 12 replicates.

**Output 85.5.4** MUNI Passenger Waiting Time Analysis Using the Jackknife Method

```
           MUNI passenger waiting time analysis using jackknife

                       The SURVEYMEANS Procedure

                            Data Summary

               Number of Strata                    6
               Number of Clusters                 12
               Number of Observations           1937
               Sum of Weights                 143040
```

**Output 85.5.4** *continued*

```
                        Variance Estimation

              Method                     Jackknife
              Number of Replicates              12
```

Output 85.5.5 reports the statistics computed using the jackknife method. Although the average passenger waiting time remains the same (7.33 minutes), the standard error is slightly smaller 0.23 minutes when the jackknife method is used, as opposed to 0.24 minutes when the BRR method is used. The 95% confident limits are between 6.76 and 7.90 minutes when the jackknife method is used.

**Output 85.5.5** MUNI Passenger Waiting Time Analysis Using the Jackknife Method

```
                           Statistics

                                 Std Error
      Variable          Mean       of Mean      95% CL for Mean
      -------------------------------------------------------------
      waittime        7.333012     0.232211   6.76481105 7.90121244
      -------------------------------------------------------------
```

# References

Brick, J. M. and Kalton, G. (1996), "Handling Missing Data in Survey Research," *Statistical Methods in Medical Research*, 5, 215–238.

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

Dippo, C. S., Fay, R. E., and Morganstein, D. H. (1984), "Computing Variances from Complex Samples with Replicate Weights," *Proceedings of the Survey Research Methods Section, ASA*, 489–494.

Dorfman, A. and Valliant, R. (1993), "Quantile Variance Estimators in Complex Surveys," *Proceedings of the Survey Research Methods Section, ASA*, 866–871.

Fay, R. E. (1984), "Some Properties of Estimators of Variance Based on Replication Methods," *Proceedings of the Survey Research Methods Section, ASA*, 495–500.

Fay, R. E. (1989), "Theory and Application of Replicate Weighting for Variance Calculations," *Proceedings of the Survey Research Methods Section, ASA*, 212–217.

Francisco, C. A. and Fuller, W. A. (1991), "Quantile Estimation with a Complex Survey Design," *Annals of Statistics*, 19, 454–469.

Fuller, W. A. (1975), "Regression Analysis for Sample Survey," *Sankhyā*, 37, Series C, Pt. 3, 117–

132.

Fuller, W. A., Kennedy, W., Schnell, D., Sullivan, G., and Park, H. J. (1989), *PC CARP*, Ames: Statistical Laboratory, Iowa State University.

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

Hidiroglou, M. A., Fuller, W. A., and Hickman, R. D. (1980), *SUPER CARP*, Ames, IA: Statistical Laboratory, Iowa State University.

Judkins, D. (1990), "Fay's Method for Variance Estimation," *Journal of Official Statistics*, 6, 223–239.

Kalton, G. (1983), *Introduction to Survey Sampling*, Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07-035, Beverly Hills, CA, and London: Sage Publications.

Kalton, G., and Kaspyzyk, D. (1986), "The Treatment of Missing Survey Data," *Survey Methodology*, 12, 1–16.

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

Lee, E. S., Forthoffer, R. N., and Lorimor, R. J. (1989), *Analyzing Complex Survey Data*, Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07-071, Beverly Hills, CA, and London: Sage Publications.

Lohr, S. L. (1999), *Sampling: Design and Analysis*, Pacific Grove, CA: Duxbury Press.

Rao, J. N. K. and Shao, J. (1996), "On Balanced Half Sample Variance Estimation in Stratified Sampling," *Journal of the American Statistical Association,* 91, 343–348.

Rao, J. N. K. and Shao, J. (1999), "Modified Balanced Repeated Replication for Complex Survey Data," *Biometrika,* 86, 403–415.

Rao, J. N. K., Wu, C. F. J., and Yue, K. (1992), " Some Recent Work on Resampling Methods for Complex Surveys," *Survey Methodology,* 18, 209–217.

Rust, K. (1985), "Variance Estimation for Complex Estimators in Sample Surveys," *Journal of Official Statistics,* 1, 381–397.

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

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

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

# Subject Index

# Syntax Index

# Your Turn

We welcome your feedback.

- If you have comments about this book, please send them to **yourturn@sas.com**. Include the full title and page numbers (if applicable).
- If you have comments about the software, please send them to **suggest@sas.com**.