# Chapter 4
## The SURVEYMEANS Procedure

## Chapter Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>OVERVIEW</td>
<td>151</td>
</tr>
<tr>
<td>GETTING STARTED</td>
<td>151</td>
</tr>
<tr>
<td>Simple Random Sampling</td>
<td>151</td>
</tr>
<tr>
<td>Stratified Sampling</td>
<td>154</td>
</tr>
<tr>
<td>Create Output Data Set</td>
<td>156</td>
</tr>
<tr>
<td>SYNTAX</td>
<td>157</td>
</tr>
<tr>
<td>PROC SURVEYMEANS Statement</td>
<td>158</td>
</tr>
<tr>
<td>BY Statement</td>
<td>161</td>
</tr>
<tr>
<td>CLASS Statement</td>
<td>162</td>
</tr>
<tr>
<td>CLUSTER Statement</td>
<td>163</td>
</tr>
<tr>
<td>STRATA Statement</td>
<td>163</td>
</tr>
<tr>
<td>VAR Statement</td>
<td>164</td>
</tr>
<tr>
<td>WEIGHT Statement</td>
<td>164</td>
</tr>
<tr>
<td>DETAILS</td>
<td>164</td>
</tr>
<tr>
<td>Specification of Population Totals and Sampling Rates</td>
<td>164</td>
</tr>
<tr>
<td>Missing Values</td>
<td>165</td>
</tr>
<tr>
<td>Primary Sampling Units (PSUs)</td>
<td>166</td>
</tr>
<tr>
<td>Statistical Computation</td>
<td>166</td>
</tr>
<tr>
<td>Output Data Sets</td>
<td>171</td>
</tr>
<tr>
<td>Displayed Output</td>
<td>171</td>
</tr>
<tr>
<td>ODS Table Names</td>
<td>173</td>
</tr>
<tr>
<td>EXAMPLES</td>
<td>174</td>
</tr>
<tr>
<td>Example 4.1 Stratified Cluster Sample Design</td>
<td>174</td>
</tr>
<tr>
<td>Example 4.2 Unequal Weighting</td>
<td>177</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>181</td>
</tr>
</tbody>
</table>
Chapter 4
The SURVEYMEANS Procedure

Overview

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

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

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

Getting Started

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, please refer to the “Syntax” section on page 157. The “Examples” section on page 174 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 the 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 using simple random sampling (SRS). Selection by simple random sampling means that all students have an equal chance or being selected, and no student can be selected more than once. Each student selected for the sample was asked how much he spends for ice cream per week, on the average. The SAS data set named IceCream saves the responses of the 40 students.

```sas
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 on 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.

```sas
title1 'Analysis of Ice Cream Spending';
title2 'Simple Random Sampling 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 4.1. Analysis of Ice Cream Spending, Simple Random Sampling Design

Figure 4.1 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.

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 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 in this school. The standard error of this estimate if $0.84, 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 using stratified random sampling. In stratified sampling, the study population is divided into nonoverlapping strata, and samples are selected independently from each stratum.

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 4.1 shows the total number of students in each grade.

Table 4.1. Number of Students by Grade

<table>
<thead>
<tr>
<th>Grade</th>
<th>Number of Students</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>1,824</td>
</tr>
<tr>
<td>8</td>
<td>1,025</td>
</tr>
<tr>
<td>9</td>
<td>1,151</td>
</tr>
<tr>
<td>Total</td>
<td>4,000</td>
</tr>
</tbody>
</table>

To analyze this stratified sample from a finite population, you need to provide the population totals for each stratum to PROC SURVEYMEANS. The SAS data set named StudentTotal contains the information from Table 4.1.

```sas
data StudentTotal;
  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 store the stratum population totals in a variable named `_TOTAL_`.

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 it does not include a finite population correction in the computations.

The following SAS statements perform the analysis of the survey data.

```sas
  title1 'Analysis of Ice Cream Spending';
  title2 'Stratified Simple Random Sampling Design';
  proc surveymeans data=IceCream total=StudentTotal;
    stratum Grade / list;
    var Spending Group;
  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 StudentTotal as the input data set containing the stratum population totals. Comparing this to the analysis in the “Simple Random Sampling” section on page 183, notice that the TOTAL=StudentTotal 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 they are provided 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.

![Data Summary](image1)

**Figure 4.2.** Data Summary

Figure 4.2 provides information on 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'.

![Stratum Information](image2)

**Figure 4.3.** Stratum Information
Figure 4.3 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.

### Analysis of Ice Cream Spending
#### Stratified Simple Random Sampling Design

The SURVEYMEANS Procedure

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>Std Error Mean</th>
<th>Lower 95% CL for Mean</th>
<th>Upper 95% CL for Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spending</td>
<td>40</td>
<td>8.75</td>
<td>0.53</td>
<td>7.67</td>
<td>9.82</td>
</tr>
<tr>
<td>Group = less</td>
<td>23</td>
<td>0.57</td>
<td>0.06</td>
<td>0.45</td>
<td>0.69</td>
</tr>
<tr>
<td>Group = more</td>
<td>17</td>
<td>0.43</td>
<td>0.06</td>
<td>0.31</td>
<td>0.54</td>
</tr>
</tbody>
</table>

**Figure 4.4.** Ice Cream Spending Analysis, Stratified SRS Design

Figure 4.4 shows that the estimate of average weekly ice cream expense is $8.75 for students in this school, with a standard error of $0.53, and a 95% confidence interval from $7.68 to $9.82. The mean estimate of $8.75 is the same as the one shown in Figure 4.1, which was computed under the assumption of a simple random sampling design without stratification. However, the standard error computed for the stratified design is $0.53, which is less than the standard error of $0.84 shown in Figure 4.1.

Figure 4.4 shows that an estimate of 57.5% of all students spend less than $10 weekly on ice cream, and 42.5% spend more, with a standard error of 0.06%.

### Create Output Data Set

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 on ODS, see the chapter titled “Using the Output Delivery System” in the SAS/STAT Procedures Guide.

For example, to place the “Statistics” table shown in Figure 4.4 in the previous section in an output data set, you use the ODS OUTPUT statement as follows:

```sas
title1 'Analysis of Ice Cream Spending';
title2 'Stratified Simple Random Sampling Design';
proc surveymeans data=IceCream total=StudentTotal;
   stratum Grade / list;
   var Spending Group;
   ods output Statistics=MyStat;
run;
```
The statement

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

requests that the “Statistics” table that appears in Figure 4.4 be placed in a SAS data set named MyStat.

The PRINT procedure displays observations of the data set MyStat:

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

Figure 4.4 displays the observations in the data set MyStat.

![Figure 4.4](image)

Syntax

The following statements are available in PROC SURVEYMEANS.

```
PROC SURVEYMEANS < options > < statistic-keywords > ;
VAR variables ;
CLASS variables ;
STRATA variables < / option > ;
CLUSTER variables ;
WEIGHT variable ;
BY variables ;
```

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

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

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

**ALPHA=α**

sets the confidence level for confidence limits. The value of the ALPHA= option must be between 0.0001 and 0.9999, 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. If $\alpha$ is between 0 and 1 but outside the range of 0.0001 to 0.9999, the procedure uses the closest range endpoint. For example, if you specify ALPHA=0.000001, the procedure uses 0.0001 to determine confidence limits.

**DATA=SAS-data-set**

specifies the SAS data set to be analyzed by PROC SURVEYMEANS. If you omit the DATA= option, the procedure uses the most recently created SAS data set.
MISSING
requests that the procedure treat missing values as a valid category for categorical variables.

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

- DATA orders values according to their order in the input data set.
- FORMATTED orders values by their formatted values. This order is operating environment dependent. By default, the order is ascending.
- INTERNAL orders values by their unformatted values, which yields the same order that the SORT procedure does. This order is operating environment dependent.

By default, ORDER=FORMATTED.

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

For a nonstratified sample design, or for a stratified sample design with the same sampling rate in all strata, you should specify a positive 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 164 for details.

The sampling rate value must be a positive number. You can specify value as a number between 0 and 1. Or you can specify value in percentage form as a number between 1 and 100, and PROC SURVEYMEANS will convert that number to a proportion. The procedure treats the value 1 as 100%, and not the percentage form 1%.

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

TOTAL=value | SAS-data-set
N=value | SAS-data-set
specifies the total number of primary sampling units (PSUs) in the study population as a positive value, or names an input data set that contains the stratum population totals. The procedure uses this information to compute a finite population correction for variance estimation.
Chapter 4. The SURVEYMEANS Procedure

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 164 for details.

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

statistic-keywords

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

PROC SURVEYMEANS performs univariate analysis, analyzing each variable separately. Thus the number of nonmissing and missing observations may not be the same for all analysis variables. See the section “Missing Values” on page 165 for more information.

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 162 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 the same for categorical and numeric variables, as the sum of the weights over all nonmissing observations.

The valid statistic-keywords are as follows:

<table>
<thead>
<tr>
<th>Keyword</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL</td>
<td>all statistics listed</td>
</tr>
<tr>
<td>CLM</td>
<td>100(1 - (\alpha))% confidence limits for the MEAN, where (\alpha) is determined by the ALPHA= option described on page 191, and the default is (\alpha = 0.05)</td>
</tr>
<tr>
<td>CV</td>
<td>coefficient of variation</td>
</tr>
<tr>
<td>DF</td>
<td>degrees of freedom for the (t) test</td>
</tr>
<tr>
<td>MAX</td>
<td>maximum value</td>
</tr>
<tr>
<td>MEAN</td>
<td>mean for a numeric variable, or the proportion in each category for a categorical variable</td>
</tr>
<tr>
<td>MIN</td>
<td>minimum value</td>
</tr>
</tbody>
</table>
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 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 information on estimating the variance of subpopulation means and totals from sample survey data, refer to Cochran (1977).

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 using the SORT procedure with a similar BY statement.
- Use the BY statement options NOTSORTED or DESCENDING in the BY statement. 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 using the DATASETS procedure.

For more information on the BY statement, refer to the discussion in *SAS Language Reference: Concepts*. For more information on the DATASETS procedure, refer to the discussion in the *SAS Procedures Guide*.

### CLASS Statement

```sas
CLASS | CLASSES 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. Refer to the discussion of the FORMAT procedure in the *SAS Procedures Guide* and to the discussions of the FORMAT statement and SAS formats in *SAS Language Reference: Dictionary*.

You can use multiple CLASS statements to specify categorical variables.

When you specify class variables, the procedure uses the SAS system option SUMSIZE= for the amount of memory that is available for data analysis. Refer to the chapter on SAS System options in *SAS Language Reference: Dictionary* for a description of the SUMSIZE= option.
**STRA T A Statement**

```
CLUSTER | CLUSTERS variables ;
```

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

If your sample design has clustering at multiple stages, you should identify only the first-stage clusters, or primary sampling units (PSUs), in the CLUSTER statement. See the section “Primary Sampling Units (PSUs)” on page 201 for more information.

The CLUSTER variables are one or more variables in the DATA= input data set described on page 158. These variables can be either character or numeric. The formatted values of the CLUSTER variables determine the CLUSTER variable levels. Thus, you can use formats to group values into levels. Refer to the discussion of the FORMAT procedure in the *SAS Procedures Guide* and to the discussions of the FORMAT statement and SAS formats in *SAS Language Reference: Dictionary*.

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

**STRATA Statement**

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

The STRATA statement names variables that form the strata in a stratified sample design. The combinations of categories of STRATA variables define the strata in the sample. If your sample design has stratification at multiple stages, you should identify only the first-stage strata in the STRATA statement. See the section “Statistical Computation” on page 166 for more information.

The STRATA variables are one or more variables in the DATA= input data set described on page 158. These variables can be either character or numeric. The formatted values of the STRATA variables determine the levels. Thus, you can use formats to group values into levels. See the discussion of the FORMAT procedure in the *SAS Procedures Guide*.

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

**LIST**

displays a “Stratum Information” table, which includes values of the STRATA variables and sampling rates for each stratum. This table also provides the number of observations and number of clusters for each stratum and analysis variable. See the section “Displayed Output” on page 171 for more details.
Chapter 4. The `SURVEYMEANS` Procedure

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 162 for more information.

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

WEIGHT Statement

```
WEIGHT|WGT variable;
```

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

Details

Specification of Population Totals and Sampling Rates

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

If your design has multiple stages of selection and you are specifying the RATE= option, you should input the first-stage sampling rate, which is the ratio of the number of PSUs in the sample to the total number of PSUs in the study population. If you
are specifying the TOTAL= option for a multistage design, you should input the total number of PSUs in the study population. See the section “Primary Sampling Units (PSUs)” on page 201 for more details.

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

The secondary data set must contain all the stratification variables listed in the STRATA statement and all the variables in the BY statement. If there are formats associated with the STRATA variables and the BY variables, then the formats must be consistent in the primary and the secondary data sets. If you specify the TOTAL=SAS-data-set option, the secondary data set must have a variable named _TOTAL_ that contains the stratum population totals. Or if you specify the RATE=SAS-data-set option, the secondary data set must have a variable named _RATE_ that contains the stratum sampling rates. The secondary data set must contain all BY and STRATA groups that occur in the primary data set. 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 positive numbers. You can specify value as a number between 0 and 1. Or you can specify value in percentage form as a number between 1 and 100, and PROC SURVEYMEANS will convert that number to a proportion. The procedure treats the value 1 as 100%, and not the percentage form 1%.

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

---

**Missing Values**

If an observation has a missing value for the WEIGHT variable, then PROC SURVEYMEANS excludes that observation from the analysis. An observation is also excluded if it has a missing value for any STRATA or CLUSTER variable, unless the MISSING option described on page 159, is used.

When computing statistics for an analysis variable, by default PROC SURVEYMEANS omits observations with missing values for that variable. The procedure bases statistics for each variable only on observations that have nonmissing values for that variable. If you specify the MISSING option described on page 159, in the PROC SURVEYMEANS statement, the procedure treats missing values of a categorical variable as a valid category.
The procedure performs univariate analysis and analyzes each VAR variable separately. Thus, the number of missing observations may 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.

Primary Sampling Units (PSUs)

When you have clusters, or primary sampling units (PSUs), in your sample design, the procedure estimates variance from the variation among PSUs. See the section “Variance and Standard Error of the Mean” on page 168 and the section “Variance and Standard Deviation of the Total” on page 170. You can use the CLUSTER statement to identify the first stage clusters in your design. PROC SURVEYMEANS assumes that each cluster represents a PSU in the sample and that each observation is an element of a PSU. If you do not specify a CLUSTER statement, the procedure treats each observation as a PSU.

Statistical Computation

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

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

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

\[
(w, Y) = (w_{hij}, y_{hij}) = (w_{hij}, y_{hij}^{(1)}, y_{hij}^{(2)}, \ldots, y_{hij}^{(P)})
\]

where

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

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

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

Therefore, the total number of analysis variables, \( P \), is the total number of numerical variables plus the total number of levels of all categorical variables.

Also, \( f_h \) denotes the sampling rate for stratum \( h \). You can use the TOTAL= option or the RATE= option to input population totals or sampling rates. See the section “Specification of Population Totals and Sampling Rates” on page 164 for details. If you input the stratum totals, PROC SURVEYMEANS computes \( f_h \) as the ratio of the stratum sample size to the stratum total. If you input the stratum sampling rates, PROC SURVEYMEANS uses these values directly for \( f_h \). If you do not specify the TOTAL= option or the RATE= option, then the procedure assumes that the stratum sampling rates \( f_h \) are negligible, and a finite population correction is not used when computing variances.

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

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

PROC SURVEYMEANS computes the estimate of the mean as

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

where

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

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

Variance and Standard Error of the Mean

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

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

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

where

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

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

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

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

$$t(\bar{Y}) = \frac{\hat{Y}}{\text{StdErr}(\hat{Y})}$$

The two-sided p-value for this test is

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

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

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

**Confidence Limits for the Mean**

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

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

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

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

**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.

$$cv = \frac{\text{StdErr}(\hat{Y})}{\hat{Y}}$$

**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

$$\hat{p} = \frac{\sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hi} y_{[q]}^{[j]}}{\sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hi} j}$$
where \( y_{hij}^{(q)} \) is value of the indicator function for level \( C = c_k \), defined in the section “Definitions and Notation” on page 167, \( y_{hij}^{(q)} \) 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 168. Similarly, the procedure computes confidence limits for proportions as described in the section “Confidence Limits for the Mean” on page 169.

**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.

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

For a categorical variable level, \( \hat{Y} \) estimates its total frequency in the population.

**Variance and Standard Deviation of the Total**

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

PROC SURVEYMEANS estimates the variance of the total as

\[
\hat{V}(\hat{Y}) = \sum_{h=1}^{H} \frac{n_h (1 - f_h)}{n_h - 1} \sum_{i=1}^{n_h} (y_{hi.} - \bar{y}_{i.})^2
\]

where

\[
y_{hi.} = \sum_{j=1}^{m_{hi}} w_{hij} y_{hij}
\]

\[
\bar{y}_{i.} = \left( \sum_{i=1}^{n_h} y_{hi.} \right) / n_h
\]

The standard deviation of the total equals

\[
\text{Std}(\hat{Y}) = \sqrt{\hat{V}(\hat{Y})}
\]

**Confidence Limits of a Total**

If you specify the keyword SUMCLM, 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

\[
\hat{Y} \pm \text{Std}(\hat{Y}) \cdot t_{df, \alpha/2}
\]
where \( \hat{Y} \) is the estimate of the total, \( \text{Std}(\hat{Y}) \) is the estimated standard deviation, and 
\( t_{df, \alpha/2} \) is the 100(1 - \( \alpha/2 \)) percentile of the \( t \) distribution with \( df \) calculated as described in the section “t Test for the Mean” on page 169.

Output Data Sets

Output data sets from PROC SURVEYMEANS are produced using ODS (Output Delivery System). ODS encompasses more than just the production of output data sets. For a more detailed description on using ODS, see the chapter titled “Using the Output Delivery System” in the SAS/STAT Procedures Guide.

Displayed Output

By default PROC SURVEYMEANS displays a “Data Summary” table and a “Statistics” table. If you specify CLASS variables, or if you specify any character variables in the VAR statement, then the procedure displays a “Class Level Information” table. If you specify the LIST option in the STRATA statement, then the procedure displays a “Stratum Information” table.

Data and Sample Design Summary

The “Data Summary” table provides information on 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 may 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 165 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= option or the 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
- 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,

**Statistics**

The “Statistics” table displays all of the statistics that you request with statistic-keywords described on page 160, in the PROC SURVEYMEANS statement. 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 may contain the following information for each analysis variable, depending on which statistic-keywords you request:

- Variable name
- 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 \( H_0 : \text{population MEAN} = 0 \)
- \( \text{Pr} > |t| \), which is the two-sided \( p \)-value for the \( t \) test
- Lower and Upper 100(1 - \( \alpha \))% CL for Mean, which are confidence limits for the mean
- Coeff of Variation, which is the coefficient of variation
- Sum
- Std Dev, which is the standard deviation of the sum
- Var of Sum, which is the variance of the sum
- Lower and Upper 100(1 - \( \alpha \))% CL for Sum, which are confidence limits for the sum

## ODS Table Names

PROC SURVEYMEANS generates output using the Output Delivery System (ODS). For more information on ODS, see the chapter titled "Using the Output Delivery System" in the SAS/STAT Procedures Guide.

PROC SURVEYMEANS produces its output in a number of tables, and each table has a unique ODS name. You can use these names to select the tables and create output data sets using ODS. Table 4.2 lists the names of the tables produced by PROC SURVEYMEANS.

**Table 4.2. ODS Tables Produced in PROC SURVEYMEANS**

<table>
<thead>
<tr>
<th>ODS Table Name</th>
<th>Description</th>
<th>Statement</th>
<th>Option</th>
</tr>
</thead>
<tbody>
<tr>
<td>Summary</td>
<td>Data Summary</td>
<td>PROC</td>
<td>default</td>
</tr>
<tr>
<td>ClassVarInfo</td>
<td>Class Level Information</td>
<td>CLASS</td>
<td>default</td>
</tr>
<tr>
<td>StrataInfo</td>
<td>Stratum Information</td>
<td>STRATA</td>
<td>LIST</td>
</tr>
<tr>
<td>Statistics</td>
<td>Statistics</td>
<td>PROC</td>
<td>default</td>
</tr>
</tbody>
</table>

By referring to the names of such tables, you can use the ODS OUTPUT statement to place one or more of these tables in output data sets.

For example, the following statements create an output data set named MyStrata, which contains the “StrataInfo” table, and an output data set named MyStat, which contains the “Statistics” table for the ice cream study discussed in the section “Stratified Sampling” on page 186.
Examples

The “Getting Started” section on page 151 contains examples of analyzing data from simple random sampling and stratified simple random sampling designs. This section provides more examples that illustrate how to use PROC SURVEYMEANS.

Example 4.1. Stratified Cluster Sample Design

Consider the example in the section “Stratified Sampling” on page 186. 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 186 assumes that the sample of students was selected using a stratified simple random sampling 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 2 and 4 students. Table 4.3 shows the total number of study groups for each grade.

<table>
<thead>
<tr>
<th>Grade</th>
<th>Number of Study Groups</th>
<th>Number of Students</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>608</td>
<td>1,824</td>
</tr>
<tr>
<td>8</td>
<td>252</td>
<td>1,025</td>
</tr>
<tr>
<td>9</td>
<td>403</td>
<td>1,151</td>
</tr>
<tr>
<td>Total</td>
<td>617</td>
<td>4,000</td>
</tr>
</tbody>
</table>

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 8 study groups from the 7th grade, 3 groups from the 8th grade, and 5 groups from the 9th grade.

The SAS data set named IceCreamStudy saves the responses of the selected students.

```sas
data IceCreamStudy;
  input Grade StudyGroup Spending @@;
```
if (Spending < 10) then Group='less';
   else Group='more';
datalines;
7 34 7
7 34 2
9 230 8
7 403 4
8 143 12
8 143 10
9 312 10
7 321 3
7 78 1
7 78 6
7 34 7
7 34 7
9 230 7
7 501 3
8 59 20
8 59 13
8 59 18
9 235 9
9 235 6
8 156 19
8 156 14
7 489 2
7 489 2
7 315 6
7 315 2
7 489 2
7 489 2
7 489 2
7 156 1
7 156 2
9 301 8;

In the data set IceCreamStudy, the variable Grade contain a student’s grade. The variable StudyGroup 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 to 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 4.3.

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 164, 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.

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;
   stratum Grade / list;
   cluster StudyGroup;
   var Spending Group;
run;
Output 4.1.1. 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  

Class Level Information  
Class Variable Levels Values  
Group 2 less more  

Output 4.1.1 provides information on the sample design and the input data set. There are 3 strata in the sample design, and the sample contains 16 clusters and 40 observations. The variable Group has two levels, 'less' and 'more'.

Output 4.1.2. Stratum Information

Analysis of Ice Cream Spending  
Stratified Clustered Sample Design  
The SURVEYMEANS Procedure  

Stratum Information  
Stratum Population Sampling  
Index Grade Total Rate N Obs Variable N  
01 7 608 0.01 20 Spending 20  
Group = less 17  
Group = more 3  
02 8 252 0.01 9 Spending 9  
Group = less 0  
Group = more 9  
03 9 403 0.01 11 Spending 11  
Group = less 6  
Group = more 5  

Stratum Information  
Stratum Population Sampling  
Index Grade Total Rate N Obs Variable Clusters  
01 7 608 0.01 20 Spending 8  
Group = less 8  
Group = more 3  
02 8 252 0.01 9 Spending 3  
Group = less 0  
Group = more 3  
03 9 403 0.01 11 Spending 5  
Group = less 4  
Group = more 4
Example 4.2. Unequal Weighting

Output 4.1.2 displays information for each stratum. Since the primary sampling units in this design are study groups, the population totals shown in Output 4.1.2 are the total numbers of study groups for each stratum or grade. This differs from Figure 4.3 on page 155, which provides the population totals in terms of students since students were the primary sampling units for that design. Output 4.1.2 also displays the number of clusters for each stratum and analysis variable.

Output 4.1.3. Statistics

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean of Mean</th>
<th>Std Error</th>
<th>Lower 95%</th>
<th>Upper 95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spending</td>
<td>40</td>
<td>8.750000</td>
<td>0.634549</td>
<td>7.379140</td>
<td>10.120860</td>
</tr>
<tr>
<td>Group = less</td>
<td>23</td>
<td>0.575000</td>
<td>0.056274</td>
<td>0.453427</td>
<td>0.696573</td>
</tr>
<tr>
<td>Group = more</td>
<td>17</td>
<td>0.425000</td>
<td>0.056274</td>
<td>0.303427</td>
<td>0.546573</td>
</tr>
</tbody>
</table>

Output 4.1.3 displays the estimates of the average weekly ice cream expense and the percentage of students spending at least $10 weekly for ice cream. These estimates are the same as those shown in Figure 4.4 for the stratified SRS design. However, the variance estimates are different because of the different sample designs.

Example 4.2. Unequal Weighting

Quite often in complex surveys, respondents have unequal weights, which reflect unequal selection probabilities and adjustments for nonresponse and poststratification. In such surveys, the appropriate sampling weights must be used to obtain valid estimates for the study population. This example illustrates analysis of survey data with unequal sampling weights.

Suppose that economists want to study profiles of the 800 top-performing companies to provide information on their impact on the economy. A sample of 66 companies is selected with unequal probability.

```sas
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
```

SAS OnlineDoc™: Version 7-1
<table>
<thead>
<tr>
<th>Category</th>
<th>Mean 1</th>
<th>Mean 2</th>
<th>Mean 3</th>
<th>Mean 4</th>
<th>Mean 5</th>
<th>Mean 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Finance</td>
<td>5327.0</td>
<td>572.4</td>
<td>372.9</td>
<td>25.2</td>
<td>4.2</td>
<td>17.7</td>
</tr>
<tr>
<td>Energy</td>
<td>1602.7</td>
<td>678.4</td>
<td>653.0</td>
<td>75.6</td>
<td>2.8</td>
<td>6.0</td>
</tr>
<tr>
<td>Energy</td>
<td>5808.8</td>
<td>1288.4</td>
<td>2007.0</td>
<td>318.8</td>
<td>5.9</td>
<td>19.2</td>
</tr>
<tr>
<td>Medical</td>
<td>268.8</td>
<td>204.4</td>
<td>820.9</td>
<td>45.6</td>
<td>3.7</td>
<td>1.8</td>
</tr>
<tr>
<td>Transportation</td>
<td>5222.6</td>
<td>2627.8</td>
<td>1910.0</td>
<td>245.6</td>
<td>22.8</td>
<td>17.4</td>
</tr>
<tr>
<td>Other</td>
<td>872.7</td>
<td>1419.4</td>
<td>939.3</td>
<td>69.7</td>
<td>12.2</td>
<td>3.7</td>
</tr>
<tr>
<td>Retail</td>
<td>4461.7</td>
<td>9466.8</td>
<td>4662.7</td>
<td>289.0</td>
<td>132.1</td>
<td>15.8</td>
</tr>
<tr>
<td>HiTech</td>
<td>6719.2</td>
<td>6942.0</td>
<td>8240.2</td>
<td>381.3</td>
<td>85.8</td>
<td>22.0</td>
</tr>
<tr>
<td>Retail</td>
<td>833.4</td>
<td>1538.8</td>
<td>1090.3</td>
<td>64.9</td>
<td>15.4</td>
<td>3.5</td>
</tr>
<tr>
<td>Finance</td>
<td>415.9</td>
<td>167.3</td>
<td>1126.8</td>
<td>58.6</td>
<td>0.7</td>
<td>2.2</td>
</tr>
<tr>
<td>HiTech</td>
<td>442.4</td>
<td>1139.9</td>
<td>1039.9</td>
<td>57.6</td>
<td>22.7</td>
<td>2.3</td>
</tr>
<tr>
<td>Other</td>
<td>801.5</td>
<td>1157.0</td>
<td>664.2</td>
<td>56.9</td>
<td>15.5</td>
<td>3.4</td>
</tr>
<tr>
<td>Finance</td>
<td>8354.8</td>
<td>1538.8</td>
<td>1090.3</td>
<td>64.9</td>
<td>15.4</td>
<td>3.5</td>
</tr>
<tr>
<td>Finance</td>
<td>5345.8</td>
<td>530.1</td>
<td>337.4</td>
<td>36.4</td>
<td>4.3</td>
<td>17.8</td>
</tr>
<tr>
<td>Energy</td>
<td>3334.3</td>
<td>1644.7</td>
<td>1407.8</td>
<td>157.6</td>
<td>6.4</td>
<td>11.4</td>
</tr>
<tr>
<td>Manufacturing</td>
<td>1826.6</td>
<td>2671.7</td>
<td>483.2</td>
<td>71.3</td>
<td>25.3</td>
<td>6.7</td>
</tr>
<tr>
<td>Retail</td>
<td>618.8</td>
<td>2354.7</td>
<td>767.7</td>
<td>58.6</td>
<td>19.0</td>
<td>2.9</td>
</tr>
<tr>
<td>Retail</td>
<td>1529.1</td>
<td>6534.0</td>
<td>826.3</td>
<td>58.3</td>
<td>65.8</td>
<td>5.7</td>
</tr>
<tr>
<td>Manufacturing</td>
<td>4458.4</td>
<td>4824.5</td>
<td>3132.1</td>
<td>28.9</td>
<td>67.0</td>
<td>15.0</td>
</tr>
<tr>
<td>HiTech</td>
<td>5831.7</td>
<td>6611.1</td>
<td>9464.7</td>
<td>459.6</td>
<td>86.7</td>
<td>19.3</td>
</tr>
<tr>
<td>Medical</td>
<td>6468.3</td>
<td>4199.2</td>
<td>3170.4</td>
<td>270.1</td>
<td>59.5</td>
<td>21.3</td>
</tr>
<tr>
<td>Energy</td>
<td>1720.7</td>
<td>473.1</td>
<td>811.1</td>
<td>86.6</td>
<td>1.6</td>
<td>6.3</td>
</tr>
<tr>
<td>Energy</td>
<td>1679.7</td>
<td>1379.9</td>
<td>721.1</td>
<td>91.8</td>
<td>4.5</td>
<td>6.2</td>
</tr>
<tr>
<td>Retail</td>
<td>4018.2</td>
<td>16823.4</td>
<td>2038.3</td>
<td>178.1</td>
<td>162.0</td>
<td>13.6</td>
</tr>
<tr>
<td>Other</td>
<td>227.1</td>
<td>575.8</td>
<td>1083.8</td>
<td>62.6</td>
<td>1.9</td>
<td>1.6</td>
</tr>
<tr>
<td>Finance</td>
<td>3872.8</td>
<td>362.0</td>
<td>209.3</td>
<td>27.6</td>
<td>2.4</td>
<td>13.1</td>
</tr>
<tr>
<td>Manufacturing</td>
<td>3359.3</td>
<td>4844.7</td>
<td>2651.4</td>
<td>224.1</td>
<td>75.6</td>
<td>11.5</td>
</tr>
<tr>
<td>Energy</td>
<td>1295.6</td>
<td>356.9</td>
<td>180.8</td>
<td>162.3</td>
<td>0.6</td>
<td>5.0</td>
</tr>
<tr>
<td>Energy</td>
<td>1658.0</td>
<td>626.6</td>
<td>688.0</td>
<td>126.0</td>
<td>3.5</td>
<td>6.1</td>
</tr>
<tr>
<td>Finance</td>
<td>12156.7</td>
<td>1345.5</td>
<td>680.7</td>
<td>106.6</td>
<td>9.4</td>
<td>39.2</td>
</tr>
<tr>
<td>HiTech</td>
<td>3982.6</td>
<td>4196.0</td>
<td>3946.8</td>
<td>313.9</td>
<td>64.3</td>
<td>13.5</td>
</tr>
<tr>
<td>Finance</td>
<td>8760.7</td>
<td>886.4</td>
<td>1006.9</td>
<td>90.0</td>
<td>7.5</td>
<td>28.5</td>
</tr>
<tr>
<td>Manufacturing</td>
<td>2362.2</td>
<td>3153.3</td>
<td>1080.0</td>
<td>137.0</td>
<td>25.2</td>
<td>8.4</td>
</tr>
<tr>
<td>Transportation</td>
<td>2499.9</td>
<td>3419.0</td>
<td>992.6</td>
<td>47.2</td>
<td>25.3</td>
<td>8.8</td>
</tr>
<tr>
<td>Energy</td>
<td>1430.4</td>
<td>1610.0</td>
<td>664.3</td>
<td>77.7</td>
<td>3.5</td>
<td>5.4</td>
</tr>
<tr>
<td>Energy</td>
<td>13666.5</td>
<td>15465.4</td>
<td>2736.7</td>
<td>411.4</td>
<td>26.6</td>
<td>43.9</td>
</tr>
<tr>
<td>Manufacturing</td>
<td>4069.3</td>
<td>4174.7</td>
<td>2907.6</td>
<td>289.2</td>
<td>38.2</td>
<td>13.7</td>
</tr>
<tr>
<td>Energy</td>
<td>2924.7</td>
<td>711.9</td>
<td>1067.8</td>
<td>146.7</td>
<td>3.4</td>
<td>10.1</td>
</tr>
<tr>
<td>Transportation</td>
<td>1262.1</td>
<td>1716.0</td>
<td>364.3</td>
<td>71.2</td>
<td>14.5</td>
<td>4.9</td>
</tr>
<tr>
<td>Medical</td>
<td>684.4</td>
<td>672.9</td>
<td>287.4</td>
<td>61.8</td>
<td>6.0</td>
<td>3.1</td>
</tr>
<tr>
<td>Energy</td>
<td>3069.3</td>
<td>1719.0</td>
<td>1439.0</td>
<td>196.4</td>
<td>4.9</td>
<td>10.6</td>
</tr>
<tr>
<td>Medical</td>
<td>246.5</td>
<td>318.8</td>
<td>924.1</td>
<td>43.8</td>
<td>3.1</td>
<td>1.7</td>
</tr>
<tr>
<td>Finance</td>
<td>11562.2</td>
<td>1128.5</td>
<td>580.4</td>
<td>64.2</td>
<td>6.7</td>
<td>37.3</td>
</tr>
<tr>
<td>Finance</td>
<td>9316.0</td>
<td>1059.4</td>
<td>816.5</td>
<td>95.9</td>
<td>8.0</td>
<td>30.2</td>
</tr>
<tr>
<td>Retail</td>
<td>1094.3</td>
<td>3848.0</td>
<td>563.3</td>
<td>29.4</td>
<td>44.7</td>
<td>4.4</td>
</tr>
<tr>
<td>Retail</td>
<td>1102.1</td>
<td>4878.3</td>
<td>932.4</td>
<td>65.2</td>
<td>47.3</td>
<td>4.4</td>
</tr>
<tr>
<td>HiTech</td>
<td>466.4</td>
<td>675.8</td>
<td>845.7</td>
<td>64.5</td>
<td>5.2</td>
<td>2.4</td>
</tr>
<tr>
<td>Manufacturing</td>
<td>10839.4</td>
<td>5468.7</td>
<td>1895.4</td>
<td>232.8</td>
<td>47.8</td>
<td>35.0</td>
</tr>
<tr>
<td>Manufacturing</td>
<td>733.5</td>
<td>2135.3</td>
<td>96.6</td>
<td>10.9</td>
<td>2.7</td>
<td>3.2</td>
</tr>
<tr>
<td>Manufacturing</td>
<td>10354.2</td>
<td>14477.4</td>
<td>5607.2</td>
<td>321.9</td>
<td>188.5</td>
<td>33.5</td>
</tr>
<tr>
<td>Energy</td>
<td>1902.1</td>
<td>2697.9</td>
<td>329.3</td>
<td>34.2</td>
<td>2.2</td>
<td>6.9</td>
</tr>
<tr>
<td>Other</td>
<td>2245.2</td>
<td>2132.2</td>
<td>2230.4</td>
<td>198.9</td>
<td>8.0</td>
<td>8.0</td>
</tr>
<tr>
<td>Transportation</td>
<td>949.4</td>
<td>1248.3</td>
<td>298.9</td>
<td>35.4</td>
<td>10.4</td>
<td>3.9</td>
</tr>
</tbody>
</table>
Example 4.2. Unequal Weighting

The variable *Type* identifies the type of market for the company. The variable *Asset* contains the company's assets in millions of dollars. The variable *Sale* contains sales in millions of dollars. The variable *Value* contains the market value of the company in millions of dollars. The variable *Profit* contains the profit in millions of dollars. The variable *Employee* stores the number of employees in thousands, and the variable *Weight* contains the sampling weight. In this example, the sampling weights are reciprocals of the selection probabilities.

Using a probability sample design and the appropriate sampling weights, you can obtain statistically valid estimates for the study population. The following SAS statements compute estimates for this study.

```sas
title1 'Top Companies Profile Study';
title2 'Using Sampling Weights';
proc surveymeans data=Company total=800 mean sum;
var Asset Sale Value Profit Employee;
weight Weight;
run;
```

Output 4.2.1 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.
Chapter 4. The SURVEYMEANS Procedure

The “Statistics” table in Output 4.2.1 displays the estimates of the mean and total for all analysis variables.

If you do not use the appropriate sampling weights, then the results of the analysis may be biased. For example, the following statements analyze the data without the sampling weights that reflect the unequal probabilities of selection.

```
title1 'Top Companies Profile Study';
title2 'Without Using the Sampling Weights';
proc surveymeans data=Company total=800 mean sum;
   var Asset Sale Value Profit Employee;
run;
```

**Output 4.2.2.** Company Profile Study without Sampling Weights

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean of Mean</th>
<th>Std Error</th>
<th>Sum</th>
<th>Std Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asset</td>
<td>3557.753030</td>
<td>401.508963</td>
<td>234812</td>
<td>26500</td>
</tr>
<tr>
<td>Sale</td>
<td>2881.306061</td>
<td>407.864339</td>
<td>190166</td>
<td>26919</td>
</tr>
<tr>
<td>Value</td>
<td>1517.507576</td>
<td>206.197430</td>
<td>100156</td>
<td>13609</td>
</tr>
<tr>
<td>Profit</td>
<td>129.121212</td>
<td>13.824279</td>
<td>8522.000000</td>
<td>912.402420</td>
</tr>
<tr>
<td>Employee</td>
<td>27.704545</td>
<td>4.601392</td>
<td>1828.500000</td>
<td>303.691848</td>
</tr>
</tbody>
</table>

Output 4.2.2 shows the results of the analysis without using the sampling weights. These results just summarize the sample of 66 companies, and they are not statistically valid estimates for the study population of 800 companies. These statistics are substantially different from the estimates shown in Output 4.2.1. For example, the total assets computed without sampling weights is only $235 billion, compared to the estimate of $5,217 billion computed with the sampling weights. The estimated mean of the assets is $3.56 billion for the sample, but is estimated as $6.52 billion for the study population.
References


Statistical Laboratory (1989), *PC CARP*, Ames, IA: Statistical Laboratory, Iowa State University.