THE POWER TO KNOW.

SAS/STAT 15.2® User's Guide

The SURVEYSELECT Procedure

SAS® Documentation
November 06, 2020

This document is an individual chapter from *SAS/STAT® 15.2 User's Guide*.

The correct bibliographic citation for this manual is as follows: SAS Institute Inc. 2020. *SAS/STAT® 15.2 User's Guide*. Cary, NC: SAS Institute Inc.

SAS/STAT® 15.2 User's Guide

SAS Institute Inc., SAS Campus Drive, Cary, NC 27513-2414

November 2020

Chapter 123
The SURVEYSELECT Procedure

Contents

Overview: SURVEYSELECT Procedure **10178**
Getting Started: SURVEYSELECT Procedure **10180**
Simple Random Sampling 10181
Stratified Sampling 10183
Stratified Sampling with Control Sorting 10186
Syntax: SURVEYSELECT Procedure **10187**
PROC SURVEYSELECT Statement 10188
CONTROL Statement 10211
FREQ Statement 10211
ID Statement 10212
SAMPLINGUNIT | CLUSTER Statement 10212
SIZE Statement 10213
STRATA Statement 10214
Details: SURVEYSELECT Procedure **10219**
Missing Values 10219
Sorting by CONTROL Variables 10220
Random Number Generation 10221
Sample Selection Methods 10222
Simple Random Sampling 10222
Unrestricted Random Sampling 10223
Balanced Bootstrap Sampling 10223
Systematic Random Sampling 10223
Sequential Random Sampling 10224
Bernoulli Sampling 10225
Poisson Sampling 10226
Sequential Poisson Sampling 10226
PPS Sampling without Replacement 10227
PPS Sampling with Replacement 10228
PPS Systematic Sampling 10229
PPS Sequential Sampling 10229
Brewer’s PPS Method 10231
Murthy’s PPS Method 10231
Sampford’s PPS Method 10232
Sample Size Allocation 10232
Proportional Allocation 10233
Optimal Allocation 10233

Neyman Allocation 10234
Specifying the Margin of Error 10235
Secondary Input Data Set 10236
Sample Output Data Set 10237
Allocation Output Data Set 10241
Random Assignment Output Data Set 10242
Displayed Output 10243
ODS Table Names 10247
Examples: SURVEYSELECT Procedure **10247**
Example 123.1: Replicated Sampling 10247
Example 123.2: PPS Selection of Two Units per Stratum 10250
Example 123.3: PPS (Dollar-Unit) Sampling 10253
Example 123.4: Proportional Allocation 10256
References **10258**

Overview: SURVEYSELECT Procedure

The SURVEYSELECT procedure provides a variety of methods for selecting probability-based random samples. The procedure can select a simple random sample or can sample according to a complex multistage design that includes stratification, clustering, and unequal probabilities of selection. When you use probability sampling, each unit in the survey population has a known, positive probability of selection. This property of probability sampling avoids selection bias and enables you to use statistical theory to make valid inferences from the sample to the survey population.

To select a sample by using PROC SURVEYSELECT, you provide a SAS data set that contains the sampling frame (the list of units from which the sample is to be selected). The sampling units can be individual observations or groups of observations (clusters). You can also specify the selection method, the sample size or sampling rate, and other selection parameters. PROC SURVEYSELECT selects the sample and produces an output data set that contains the selected units, their selection probabilities, and their sampling weights. To select a sample in multiple stages, you can invoke the procedure separately for each stage of selection by providing the sampling frame and selection parameters for each stage.

PROC SURVEYSELECT provides methods for both equal probability sampling and probability proportional to size (PPS) sampling. In equal probability sampling, each unit in the sampling frame (or stratum) has the same probability of selection. In PPS sampling, each unit's selection probability is proportional to its size measure. For information about probability sampling methods, see Lohr (2010), Kish (1965), Kish (1987), Kalton (1983), and Cochran (1977).

PROC SURVEYSELECT provides the following equal probability sampling methods:

- simple random sampling (without replacement)
- unrestricted random sampling (with replacement)
- systematic random sampling

- sequential random sampling
- balanced bootstrap sampling
- Bernoulli sampling

The procedure also provides Poisson sampling and the following probability proportional to size (PPS) sampling methods:

- PPS sampling without replacement
- PPS sampling with replacement
- PPS systematic sampling
- PPS algorithms for selecting two units per stratum
- sequential PPS sampling with minimum replacement

PROC SURVEYSELECT uses fast, efficient algorithms for sample selection. Thus, it performs well even for large input data sets (sampling frames).

PROC SURVEYSELECT can perform stratified sampling by selecting samples independently within strata, which are nonoverlapping subgroups of the survey population. Stratification controls the distribution of the sample size in the strata. It is widely used in practice toward meeting a variety of survey objectives. For example, you can use stratification to ensure adequate sample sizes for subgroups of interest (including small subgroups), or you can use stratification to improve the precision of overall estimates. When you use a systematic or sequential selection method, PROC SURVEYSELECT can sort by control variables within strata for the additional control of implicit stratification.

For stratified sampling, PROC SURVEYSELECT provides survey design methods to allocate the total sample size among the strata. Available allocation methods include proportional, Neyman, and optimal allocation. Optimal allocation maximizes the estimation precision within the available resources by taking into account stratum sizes, costs, and variances.

PROC SURVEYSELECT provides replicated sampling, where the total sample is composed of a set of replicates, and each replicate is selected in the same way. You can use replicated sampling to study variable nonsampling errors, such as variability in the results obtained by different interviewers. You can also use replicated sampling to estimate standard errors for combined sample estimates and to perform a variety of other resampling and simulation tasks.

Getting Started: SURVEYSELECT Procedure

The following examples show how to use PROC SURVEYSELECT to select probability-based random samples. These examples use simulated data for a customer satisfaction survey.

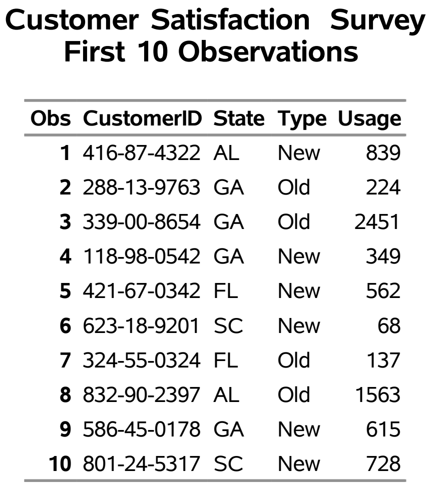
Suppose an internet service provider plans to conduct a customer satisfaction survey by selecting a random sample of customers from all current customers (the survey population). The company plans to interview the selected customers and make inferences about the survey population from the sample data.

The SAS data set Customers contains the sampling frame, which is the list of units in the survey population. The sample of customers will be selected from this sampling frame. The data set Customers is constructed from the company's customer database and contains 13,471 observations (one observation for each customer).

The following PROC PRINT statements display the first 10 observations of the data set Customers and produce the table shown in Figure 123.1:

```
title1 'Customer Satisfaction Survey';
title2 'First 10 Observations';
proc print data=Customers(obs=10);
run;
```

Figure 123.1 Customers Data Set (First 10 Observations)



Customer Satisfaction Survey
First 10 Observations

Obs	CustomerID	State	Type	Usage
1	416-87-4322	AL	New	839
2	288-13-9763	GA	Old	224
3	339-00-8654	GA	Old	2451
4	118-98-0542	GA	New	349
5	421-67-0342	FL	New	562
6	623-18-9201	SC	New	68
7	324-55-0324	FL	Old	137
8	832-90-2397	AL	Old	1563
9	586-45-0178	GA	New	615
10	801-24-5317	SC	New	728

The variable CustomerID contains the unique customer identification number. The variable State contains the state where the customer is located. The value of the variable Type is 'Old' if the customer has subscribed to the service for more than one year; otherwise, the value of Type is 'New'. The variable Usage contains the customer's average monthly service usage in minutes.

The following examples show how to use PROC SURVEYSELECT to implement three different probability sample designs. The first design is simple random sampling without stratification. The second design is a stratified design in which the list of customers is stratified by state and type; the sample is then selected by simple random sampling within strata. The third design is a stratified design that uses control sorting; customers are ordered by service usage within strata, and the sample is selected by systematic random sampling.

Simple Random Sampling

The following PROC SURVEYSELECT statements select a probability sample of customers from the Customers data set by using simple random sampling:

```
title1 'Customer Satisfaction Survey';
title2 'Simple Random Sampling';
proc surveyselect data=Customers method=srs n=100
                  out=SampleSRS;
run;
```

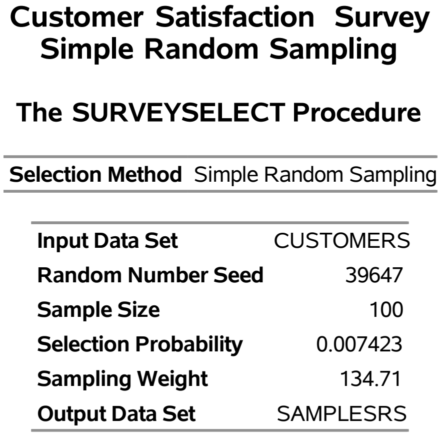
The PROC SURVEYSELECT statement invokes the procedure. The DATA= option names the SAS data set Customers as the input data set from which to select the sample. The METHOD=SRS option specifies simple random sampling as the sample selection method. In simple random sampling, each sampling unit (observation) has an equal probability of selection, and sampling is performed without replacement. (Without-replacement sampling means that a unit cannot be selected more than once.) The N= option specifies a sample size of 100 customers. The OUT= option stores the sample in the SAS data set named SampleSRS.

Figure 123.2 displays the output from PROC SURVEYSELECT, which summarizes the sample selection. The “Sample Selection Method” table shows that the selection method is simple random sampling. The “Sample Selection Summary” table shows that a sample of 100 customers is selected from the input data set Customers.

When you use simple random sampling (without stratification), all sampling units have the same selection probability. In this example, the selection probability for each customer is 0.007423, which is the sample size (100) divided by the population size (13,471). The sampling weight for each customer in the sample is 134.71, which is the inverse of the selection probability. If you specify the STATS option, PROC SURVEYSELECT includes the selection probabilities and sampling weights in the output data set. For more complex designs, PROC SURVEYSELECT includes this information in the output data set by default.

The “Sample Selection Summary” table also displays the initial seed that PROC SURVEYSELECT uses for random number generation (39647). When you do not specify the SEED= option, PROC SURVEYSELECT uses the time of day from the computer’s clock to obtain an initial seed. To reproduce this same sample, you can specify SEED=39647 (for the same input data set and sample selection method).

Figure 123.2 Sample Selection Summary



```
 Customer Satisfaction  Survey
    Simple Random Sampling

   The SURVEYSELECT Procedure

Selection Method  Simple Random Sampling

    Input Data Set          CUSTOMERS
    Random Number Seed          39647
    Sample Size                   100
    Selection Probability    0.007423
    Sampling Weight            134.71
    Output Data Set         SAMPLESRS
```

The sample of 100 customers is stored in the SAS data set SampleSRS. PROC SURVEYSELECT does not display this output data set. The following PROC PRINT statements display the first 20 observations of SampleSRS:

```
title1 'Customer Satisfaction Survey';
title2 'Sample of 100 Customers, Selected by SRS';
title3 '(First 20 Observations)';
proc print data=SampleSRS(obs=20);
run;
```

Figure 123.3 displays the first 20 observations of the output data set SampleSRS, which contains the sample of customers. This data set includes all variables in the input data set Customers. If you do not want to include all variables, you can use the ID statement to specify which variables to copy from the input data set to the output (sample) data set.

Figure 123.3 Customer Sample (First 20 Observations)

Customer Satisfaction Survey
Sample of 100 Customers, Selected by SRS
(First 20 Observations)

Obs	CustomerID	State	Type	Usage
1	017-27-4096	GA	New	168
2	026-37-3895	AL	New	59
3	038-54-9276	GA	New	785
4	046-40-3131	FL	New	60
5	070-37-6924	GA	New	524
6	100-58-3342	FL	New	302
7	107-61-9029	AL	New	235
8	110-95-0432	FL	New	12
9	112-81-9251	SC	New	347
10	137-33-0478	GA	New	551
11	143-83-4677	AL	New	203
12	147-19-9164	GA	New	172
13	159-51-0606	FL	New	102
14	164-14-7799	GA	Old	388
15	165-05-7323	SC	New	606
16	174-69-3566	AL	Old	111
17	177-69-6934	FL	New	202
18	181-58-3508	AL	Old	261
19	207-41-8446	AL	Old	183
20	207-64-7308	GA	New	193

Stratified Sampling

This example shows how to use PROC SURVEYSELECT to select a stratified random sample. The sampling frame (list of customers) is stratified by the variables State and Type. This stratification divides the sampling frame into nonoverlapping subgroups that are determined by the values of State and Type. Samples are then selected independently within the strata.

PROC SURVEYSELECT requires that the input data set be sorted by the STRATA variables. The following PROC SORT statements sort the Customers data set by the stratification variables State and Type:

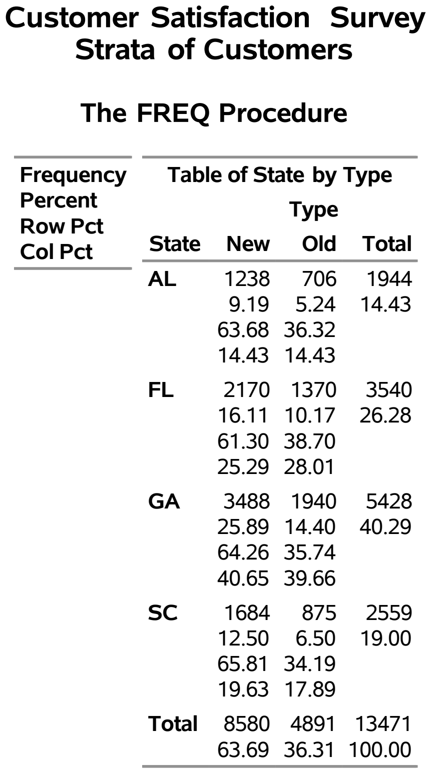
```
proc sort data=Customers;
   by State Type;
run;
```

The following PROC FREQ statements display the crosstabulation of the Customers data set by State and Type:

```
title1 'Customer Satisfaction Survey';
title2 'Strata of Customers';
proc freq data=Customers;
   tables State*Type;
run;
```

Figure 123.4 shows the table of State by Type for the 13,471 customers. There are four states and two levels of Type, which form eight strata.

Figure 123.4 Stratification of Customers by State and Type



Customer Satisfaction Survey
Strata of Customers

The FREQ Procedure

Frequency Percent Row Pct Col Pct	Table of State by Type			
		Type		
	State	New	Old	Total
	AL	1238 9.19 63.68 14.43	706 5.24 36.32 14.43	1944 14.43
	FL	2170 16.11 61.30 25.29	1370 10.17 38.70 28.01	3540 26.28
	GA	3488 25.89 64.26 40.65	1940 14.40 35.74 39.66	5428 40.29
	SC	1684 12.50 65.81 19.63	875 6.50 34.19 17.89	2559 19.00
	Total	8580 63.69	4891 36.31	13471 100.00

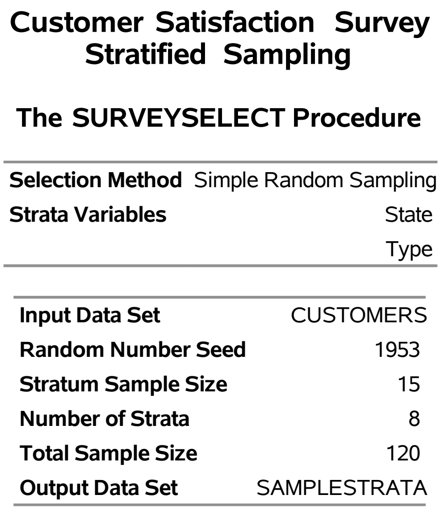
The following PROC SURVEYSELECT statements select a probability sample of customers from the Customers data set according to the stratified sample design:

```
title1 'Customer Satisfaction Survey';
title2 'Stratified Sampling';
proc surveyselect data=Customers method=srs n=15
                  seed=1953 out=SampleStrata;
   strata State Type;
run;
```

The STRATA statement names the stratification variables State and Type. In the PROC SURVEYSELECT statement, the METHOD=SRS option specifies simple random sampling, and the N= option specifies a sample size of 15 customers in each stratum. If you want to specify different sample sizes for different strata, you can use the N=*SAS-data-set* option to name a secondary data set that contains the stratum sample sizes. The SEED= option specifies 1953 as the initial seed for random number generation.

Figure 123.5 displays the output from PROC SURVEYSELECT, which summarizes the sample selection. A total of 120 customers are selected.

Figure 123.5 Sample Selection Summary



Customer Satisfaction Survey
Stratified Sampling

The SURVEYSELECT Procedure

Selection Method	Simple Random Sampling
Strata Variables	State Type

Input Data Set	CUSTOMERS
Random Number Seed	1953
Stratum Sample Size	15
Number of Strata	8
Total Sample Size	120
Output Data Set	SAMPLESTRATA

The following PROC PRINT statements display the first 30 observations of the output data set SampleStrata:

```
title1 'Customer Satisfaction Survey';
title2 'Sample Selected by Stratified Design';
title3 '(First 30 Observations)';
proc print data=SampleStrata(obs=30);
run;
```

Figure 123.6 displays the first 30 observations of the output data set SampleStrata, which contains the sample of 120 customers (15 customers from each of the eight strata). The variable SelectionProb contains the selection probability for each customer in the sample. Because customers are selected with equal probability within strata, the selection probability is the stratum sample size (15) divided by the stratum population size. The selection probabilities differ from stratum to stratum because the stratum population sizes differ. The selection probability for each customer in the first stratum (State=‘AL’ and Type=‘New’) is 0.012116, and the selection probability for customers in the second stratum (State=‘AL’ and Type=‘Old’) is 0.021246.

The variable SamplingWeight contains the sampling weights, which are computed as inverse selection probabilities.

Figure 123.6 Customer Sample (First 30 Observations)

Customer Satisfaction Survey
Sample Selected by Stratified Design
(First 30 Observations)

Obs	State	Type	CustomerID	Usage	SelectionProb	SamplingWeight
1	AL	New	015-57-9903	26	0.012116	82.5333
2	AL	New	052-18-5029	576	0.012116	82.5333
3	AL	New	064-72-0145	88	0.012116	82.5333
4	AL	New	291-22-2497	1221	0.012116	82.5333
5	AL	New	305-62-6833	187	0.012116	82.5333
6	AL	New	309-63-9722	534	0.012116	82.5333
7	AL	New	413-76-0209	435	0.012116	82.5333
8	AL	New	492-18-7867	70	0.012116	82.5333
9	AL	New	508-16-8324	189	0.012116	82.5333
10	AL	New	561-82-0366	392	0.012116	82.5333
11	AL	New	685-24-1718	74	0.012116	82.5333
12	AL	New	800-20-2155	21	0.012116	82.5333
13	AL	New	857-94-2672	77	0.012116	82.5333
14	AL	New	918-29-9618	540	0.012116	82.5333
15	AL	New	963-93-4916	33	0.012116	82.5333
16	AL	Old	182-45-1938	160	0.021246	47.0667
17	AL	Old	210-85-9046	184	0.021246	47.0667
18	AL	Old	211-14-1373	88	0.021246	47.0667
19	AL	Old	229-87-9527	362	0.021246	47.0667
20	AL	Old	239-16-9426	22	0.021246	47.0667
21	AL	Old	283-78-3723	595	0.021246	47.0667
22	AL	Old	293-90-2342	124	0.021246	47.0667
23	AL	Old	360-78-7048	375	0.021246	47.0667
24	AL	Old	432-96-1275	2283	0.021246	47.0667
25	AL	Old	534-79-2367	167	0.021246	47.0667
26	AL	Old	668-77-4832	30	0.021246	47.0667
27	AL	Old	681-88-8208	2133	0.021246	47.0667
28	AL	Old	794-79-7878	1274	0.021246	47.0667
29	AL	Old	954-40-0057	30	0.021246	47.0667
30	AL	Old	954-98-4646	1038	0.021246	47.0667

Stratified Sampling with Control Sorting

This example shows how to use PROC SURVEYSELECT to select a stratified random sample by implementing control sorting and systematic random sampling within strata. The sampling frame (list of customers) is stratified by the variable State, and the sampling units (customers) are sorted by the variables Type and Usage within each stratum (state). Customers are then selected by systematic random sampling within strata. Systematic sampling together with control sorting distributes the sample uniformly over the range of type and usage values within each state.

The following PROC SURVEYSELECT statements select a probability sample of customers from the Customers data set by using this design:

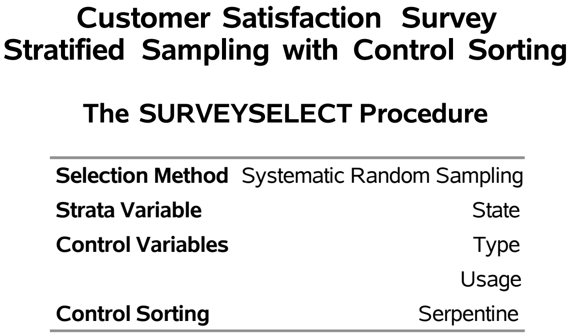
```
title1 'Customer Satisfaction Survey';
title2 'Stratified Sampling with Control Sorting';
proc surveyselect data=Customers method=sys rate=.02
                  seed=1234 out=SampleControl;
   strata State;
   control Type Usage;
run;
```

The STRATA statement names the stratification variable State. The CONTROL statement names the control variables Type and Usage. In the PROC SURVEYSELECT statement, the METHOD=SYS option requests systematic random sampling, and the RATE= option specifies a sampling rate of 2% for each stratum. The SEED= option specifies the initial seed for random number generation.

Figure 123.7 displays the output from PROC SURVEYSELECT, which summarizes the sample selection. A sample of 270 customers is selected by using systematic random sampling within strata that are determined by the variable State. The sampling frame Customers is sorted by the control variables Type and Usage within strata. Sorting is performed by using hierarchic serpentine sorting, which is the default type of sorting. For more information, see the section "Sorting by CONTROL Variables" on page 10220.

By default, the sorted data set replaces the input data set. To store the sorted input data in another data set, you can specify the OUTSORT= option. The output data set SampleControl contains the sample of customers.

Figure 123.7 Sample Selection Summary



```
                  Customer Satisfaction Survey
            Stratified Sampling with Control Sorting

                  The SURVEYSELECT Procedure

           Selection Method   Systematic Random Sampling
           Strata Variable                         State
           Control Variables                        Type
                                                   Usage
           Control Sorting                    Serpentine
```

Figure 123.7 *continued*

Input Data Set	CUSTOMERS
Random Number Seed	1234
Stratum Sampling Rate	0.02
Number of Strata	4
Total Sample Size	270
Output Data Set	SAMPLECONTROL

Syntax: SURVEYSELECT Procedure

The following statements are available in the SURVEYSELECT procedure:

```
PROC SURVEYSELECT options ;
    CONTROL variables ;
    FREQ variable ;
    ID variables ;
    SAMPLINGUNIT | CLUSTER variables < / options >  ;
    SIZE variable ;
    STRATA variables < / options >  ;
```

The PROC SURVEYSELECT statement invokes the SURVEYSELECT procedure. Optionally, it identifies input and output data sets. It also specifies the selection method, the sample size, and other sample design parameters. The PROC SURVEYSELECT statement is required.

The SIZE statement identifies the variable that contains the size measures of the sampling units. This statement is required for any probability proportional to size (PPS) selection method unless you specify the PPS option in the SAMPLINGUNIT statement.

The remaining statements are optional. The STRATA statement identifies a variable or set of variables that stratify the input data set. When you specify a STRATA statement, PROC SURVEYSELECT selects samples independently from the strata that are formed by the STRATA variables. The STRATA statement also provides options to allocate the total sample size among the strata.

The SAMPLINGUNIT statement identifies a variable or set of variables that group the input data set observations into sampling units (clusters). Sampling units are nested within strata. When you specify a SAMPLINGUNIT statement, PROC SURVEYSELECT selects clusters instead of individual observations.

The CONTROL statement identifies variables for ordering units within strata. It can be used for systematic and sequential sampling methods. The ID statement identifies variables to copy from the input data set to the output data set of selected units.

The FREQ statement identifies a variable that contains the frequency of occurrence for each observation. The FREQ statement is available only for sample allocation when no sample is selected, which you can request by specifying the ALLOC= and NOSAMPLE options in the STRATA statement.

The following sections describe the PROC SURVEYSELECT statement and then describe the other statements in alphabetical order.

PROC SURVEYSELECT Statement

PROC SURVEYSELECT *options* **;**

The PROC SURVEYSELECT statement invokes the SURVEYSELECT procedure. Optionally, it identifies input and output data sets. If you do not name a DATA= input data set, the procedure selects the sample from the most recently created SAS data set. If you do not name an OUT= output data set to contain the sample of selected units, the procedure still creates an output data set and names it according to the DATA*n* convention.

The PROC SURVEYSELECT statement also specifies the sample selection method, the sample size, and other sample design parameters.

If you do not specify a selection method, PROC SURVEYSELECT uses simple random sampling (METHOD=SRS) by default unless you specify a SIZE statement or the PPS option in the SAMPLINGUNIT statement. If you specify a SIZE statement (or the PPS option), PROC SURVEYSELECT uses probability proportional to size selection without replacement (METHOD=PPS) by default. For more information, see the description of the METHOD= option.

You can use the SAMPSIZE=*n* option to specify the sample size, or you can use the SAMPSIZE=*SAS-data-set* option to name a secondary input data set that contains stratum sample sizes. You must specify a sample size or sampling rate except when you request one of the following: random assignment (GROUPS=); balanced bootstrap selection (METHOD=BALBOOTSTRAP); Poisson sampling (METHOD=POISSON); Brewer's method or Murthy's method, either of which selects two units from each stratum (METHOD=PPS_BREWER or METHOD=PPS_MURTHY); or sample allocation for a specified margin (MARGIN=).

You can provide stratum sample sizes, sampling rates, initial seeds, minimum size measures, maximum size measures, and certainty size measures in a secondary input data set. For more information, see the descriptions of the SAMPSIZE=, SAMPRATE=, SEED=, MINSIZE=, MAXSIZE=, CERTSIZE=, and CERTSIZE=P= options. You can name only one secondary input data set in each invocation of PROC SURVEYSELECT. For more information, see the section "Secondary Input Data Set" on page 10236.

Table 123.1 summarizes the options available in the PROC SURVEYSELECT statement. Descriptions of the options follow in alphabetical order.

Table 123.1 PROC SURVEYSELECT Statement Options

Option	Description
Input and Output Data Sets	
DATA=	Names the input SAS data set
OUT=	Names the output SAS data set that contains the sample
OUTORDER=RANDOM	Randomly orders the observations in the output data set
OUTSORT=	Names an output SAS data set that stores the sorted input data set
Selection Method	
METHOD=	Specifies the sample selection method
Sample Size	
SAMPSIZE=	Specifies the sample size
SELECTALL	Selects all stratum units when the sample size exceeds the total

Table 123.1 *continued*

Option	Description
Sampling Rate	
NMAX=	Specifies the maximum stratum sample size
NMIN=	Specifies the minimum stratum sample size
ROUND=	Specifies the rounding type
SAMPRATE=	Specifies the sampling rate
Replicated Sampling	
REPS=	Specifies the number of replicate samples
Size Measures	
CERTSIZE=	Specifies the certainty size measure
CERTSIZE=P=	Specifies the certainty proportion
MAXSIZE=	Specifies the maximum size measure
MINSIZE=	Specifies the minimum size measure
Control Sorting	
SORT=	Specifies the type of sorting
Random Number Generation	
RANUNI	Requests the RANUNI random number generator
SEED=	Specifies the initial seed
STRATUMSEED=	Controls the stratum initial seeds
Random Assignment	
GROUPS=	Requests random assignment
Displayed Output	
NOPRINT	Suppresses the display of all output
OUT= Data Set Contents	
CERTUNITS=	Includes number of certainty units
JTPROBS	Includes joint probabilities of selection
OUTALL	Includes all observations from the DATA= input data set
OUTHITS	Includes a distinct copy of each selected unit
OUTSEED	Includes the initial seed for each stratum
OUTSIZE	Includes additional design and sampling frame information
STATS	Includes selection probabilities and sampling weights

You can specify the following *options*:

CERTSIZE < **=***value* | *SAS-data-set* >

specifies the certainty size value. When you specify this option, PROC SURVEYSELECT selects with certainty all sampling units whose size measures are greater than or equal to the certainty size value. After removing these certainty units, the procedure selects the remainder of the sample by using the method that you specify in the METHOD= option. You provide the size measures by using the SIZE statement.

You can provide a single certainty *value* for the entire sample selection, or you can provide stratum-level certainty values by specifying a *SAS-data-set*. The certainty size values must be positive numbers.

When you specify this option, the OUT= output data set contains a variable named Certain that identifies units that are selected with certainty. The selection probability of each certainty unit is one.

The CERTSIZE= option is available for the following PPS selection methods: METHOD=PPS, METHOD=PPS_SAMPFORD, METHOD=PPS_SYS, METHOD=PPS_WR, and METHOD=SEQ_POISSON. The CERTSIZE= option is not available when you specify a SAMPLINGUNIT statement.

You can provide certainty size values by specifying one of the following forms:

CERTSIZE

indicates that certainty size values are provided in a secondary input data set that you name in another option (for example, the SAMPSIZE=*SAS-data-set* option). This data set should include a variable named _CERTSIZE_ that contains the certainty values. For more information, see the section "Secondary Input Data Set" on page 10236. You can name only one secondary input data set in each invocation of PROC SURVEYSELECT.

CERTSIZE=*value*

specifies a single certainty size *value*, which must be a positive number. If you request a stratified sample design by specifying the STRATA statement, PROC SURVEYSELECT uses the certainty value to determine certainty selections for all strata.

CERTSIZE=*SAS-data-set*

names a *SAS-data-set* that contains stratum-level certainty size values. You should provide the certainty values in the data set variable named _CERTSIZE_. Each observation in this data set should correspond to a stratum group, which is determined by the values of the STRATA variables.

This data set, which is a secondary input data set, must contain all stratification variables that you specify in the STRATA statement. The data set must also contain all stratum groups that appear in the DATA= data set. The order of the stratum groups in the CERTSIZE= data set must match the order of the groups in the DATA= data set. If formats are associated with the STRATA variables, the formats must be consistent in the two data sets. For more information, see the section "Secondary Input Data Set" on page 10236. You can name only one secondary input data set in each invocation of PROC SURVEYSELECT.

CERTSIZE=P < =*p* | *SAS-data-set* >

specifies the certainty proportion to use for iterative certainty selection. When you specify this option, PROC SURVEYSELECT selects with certainty the sampling units whose size measure proportions (of the total size) are greater than or equal to the certainty proportion *p*. After removing the selected certainty units, the procedure repeats certainty selection by using the remaining sampling units, the total size of the remaining units, and the certainty proportion *p*. This process continues until no more certainty units are selected. PROC SURVEYSELECT then selects the remainder of the sample by using the method that you specify in the METHOD= option.

You can provide a single certainty proportion *p* for the entire sample, or you can provide stratum-level certainty proportions by specifying a *SAS-data-set*.

The certainty proportions must be positive numbers. You can specify a certainty proportion as a number between 0 and 1. Or you can specify a proportion in percentage form as a number between 1 and 100, which PROC SURVEYSELECT converts to a proportion. The procedure treats the value 1 as 100% instead of 1%.

When you specify this option, the OUT= output data set contains a variable named Certain that identifies units that are selected with certainty. The selection probability of each certainty unit is one.

You use the SIZE statement to provide size measures for the sampling units. The CERTSIZE=P= option is available for METHOD=PPS, METHOD=PPS_SAMPFORD, and METHOD=SEQ_POISSON. This option is not available when you specify a SAMPLINGUNIT statement.

You can provide certainty size proportions by specifying one of the following forms:

CERTSIZE=P

indicates that certainty size proportions are provided in a secondary input data set that you name in another option (for example, the SAMPSIZE=*SAS-data-set* option). You should provide the certainty proportions in the data set variable named _CERTP_. For more information, see the section "Secondary Input Data Set" on page 10236. You can name only one secondary input data set in each invocation of PROC SURVEYSELECT.

CERTSIZE=P=*p*

specifies a single certainty size proportion *p*, which must be a positive number. If you request a stratified sample design by specifying the STRATA statement, PROC SURVEYSELECT uses the certainty proportion *p* to determine certainty selections for all strata.

CERTSIZE=P=*SAS-data-set*

names a *SAS-data-set* that contains stratum-level certainty size proportions. You should provide the certainty proportions in the data set variable named _CERTP_. Each observation in the data set should correspond to a stratum group, which is determined by the values of the STRATA variables.

This data set, which is a secondary input data set, must contain all stratification variables that you specify in the STRATA statement. The data set must also contain all stratum groups that appear in the DATA= input data set. The order of the stratum groups in the CERTSIZE=P= data set must match the order of the groups in the DATA= data set. If formats are associated with the STRATA variables, the formats must be consistent in the two data sets. For more information, see the section "Secondary Input Data Set" on page 10236. You can name only one secondary input data set in each invocation of PROC SURVEYSELECT.

CERTUNITS=*certain-option* | (*certain-options*)

controls the display and output of the number of certainty units. This option is available when you specify the CERTSIZE= or CERTSIZE=P= option.

You can specify one or both of the following *certain-options*. If you specify both *certain-options*, enclose the values in parentheses after CERTUNITS=.

NOPRINT

suppresses display of the number of certainty units in the "Sample Selection Summary" table. For more information, see the section "Displayed Output" on page 10243.

OUTPUT

includes the number of certainty units in the output data set. For more information, see the section "Sample Output Data Set" on page 10237.

DATA=*SAS-data-set*

names the *SAS-data-set* from which PROC SURVEYSELECT selects the sample. If you omit the DATA= option, the procedure uses the most recently created SAS data set. In sampling terminology, the input data set is the *sampling frame* (the list of units from which the sample is selected).

By default, the procedure uses input data set observations as sampling units and selects a sample of these units. Alternatively, you can use the SAMPLINGUNIT statement to define sampling units as groups of observations (clusters).

GROUPS=*n* | (*values*)

requests random assignment of the observations in the input data set to groups. You can specify the total number of groups as *n*, which must be a positive integer. Or you can provide a list of group size *values*, which are positive integers that specify the number of observations in the groups. When you use a STRATA statement, PROC SURVEYSELECT performs the specified random assignment independently in each stratum. Otherwise, the procedure performs the random assignment for the entire data set.

When you specify GROUPS=*n*, PROC SURVEYSELECT randomly assigns the observations in the data set (or stratum) to *n* groups. The number of observations in each group is equal, or as nearly equal as possible. For example, if the data set contains 100 observations and you specify GROUPS=3, PROC SURVEYSELECT creates three groups that contain 33, 33, and 34 observations. This is equivalent to specifying GROUPS=(33, 33, 34).

When you specify GROUPS=*values*, the number of groups is determined by the number of group size values that you list. You can separate the values with blanks or commas, and you must enclose the list of values in parentheses. The sum of the group size values must equal the total number of observations in the data set (or in the stratum, if you specify a STRATA statement).

The OUT= data set includes a variable named GroupID that identifies the group assignment of each observation. When you specify the OUTSIZE option, the output data set includes a variable named GroupSize that provides the number of units in the group; the output data set also includes the total number of units and the number of groups (in the data set, or in the stratum if you specify a STRATA statement). For more information, see the section "Random Assignment Output Data Set" on page 10242.

The following options are available when you specify the GROUPS= option: the SEED=, RANUNI, and OUTSEED options, which pertain to random number generation; the REPS= option, which provides independent replicates of the random assignment; the NOPRINT option, which suppresses display of the "Random Assignment" table; and the OUTSIZE option.

The GROUPS= option does not select a sample; you cannot specify sample selection options (for example, METHOD= or SAMPSIZE=) when you use the GROUPS= option. The SAMPLINGUNIT statement is not available when you use the GROUPS= option.

JTPROBS

includes joint probabilities of selection in the OUT= output data set. This option is available for the following probability proportional to size selection methods: METHOD=PPS, METHOD=PPS_SAMPFORD, and METHOD=PPS_WR. By default, PROC SURVEYSELECT outputs joint selection probabilities for METHOD=PPS_BREWER and METHOD=PPS_MURTHY, which select two units per stratum.

For information about joint selection probabilities for a particular sampling method, see the method description in the section "Sample Selection Methods" on page 10222. For more information about the contents of the output data set, see the section "Sample Output Data Set" on page 10237.

MAXSIZE **< =***value* **|** *SAS-data-set* **>**

specifies the maximum size value, which PROC SURVEYSELECT uses to adjust size measures before sample selection. When a size measure exceeds the maximum size value, PROC SURVEYSELECT replaces that size measure with the maximum size value.

You can provide a single maximum size *value* for the entire sample selection, or you can provide stratum-level maximum size values by specifying a *SAS-data-set*. The maximum size values must be positive numbers.

You provide size measures by specifying the SIZE statement or by specifying the PPS option in the SAMPLINGUNIT statement.

Unless you specify a SAMPLINGUNIT statement, the OUT= data set includes a variable named AdjustedSize that contains the adjusted size measures.

If you use a SAMPLINGUNIT statement to define sampling units (clusters), PROC SURVEYSELECT adjusts the sampling unit sizes (instead of the observation sizes). If you specify a SIZE statement, the size of a sampling unit is the sum of the size measures of the observations in the unit. If you specify the SAMPLINGUNIT PPS option, the size of a sampling unit is the number of observations in the unit.

When you use a SAMPLINGUNIT statement, the OUT= data set includes a variable named UnitSize that contains the adjusted sampling unit size measures.

You can provide maximum size values by specifying one of the following forms:

MAXSIZE

indicates that maximum size values are provided in a secondary input data set that you name in another option (for example, the SAMPSIZE=*SAS-data-set* option). You should provide the maximum size values in the data set variable named _MAXSIZE_. For more information, see the section "Secondary Input Data Set" on page 10236. You can specify only one secondary input data set in each invocation of PROC SURVEYSELECT.

MAXSIZE=*value*

specifies a single maximum size *value*, which must be a positive number. If you request a stratified sample design by specifying the STRATA statement, PROC SURVEYSELECT uses the value to adjust size measures in all strata.

MAXSIZE=*SAS-data-set*

names a *SAS-data-set* that contains stratum-level maximum size values. You should provide the maximum size values in the data set variable named _MAXSIZE_. Each observation in the data set should correspond to a stratum group, which is determined by the values of the STRATA variables.

This data set, which is a secondary input data set, must contain all stratification variables that you specify in the STRATA statement. The data set must also contain all stratum groups that appear in the DATA= data set. The order of the stratum groups in the MAXSIZE= data set must match the order of the groups in the DATA= data set. If formats are associated with the STRATA variables, the formats must be consistent in the two data sets. For more information, see the section "Secondary Input Data Set" on page 10236. You can name only one secondary input data set in each invocation of PROC SURVEYSELECT.

METHOD=*method*

M=*method*

specifies the sample selection *method*.

By default, METHOD=PPS if you specify a SIZE statement or the PPS option in the SAMPLINGUNIT statement; otherwise, METHOD=SRS by default.

You can specify one of the following *methods*:

BALBOOTSTRAP

BALBOOT

requests balanced bootstrap selection. This method selects r bootstrap samples of N units with equal probability and with replacement, where N is the total number of sampling units (in the stratum or in the data set) and r is the number of samples (replicates) that you specify in the REPS= option. The bootstrap selection is balanced so that the overall total number of selections is r for each sampling unit. For more information, see the section "Balanced Bootstrap Sampling" on page 10223.

When you request this method, you must specify the number of bootstrap samples r in the REPS= option. The sample size for each bootstrap replicate is fixed at N units; therefore, you cannot specify the SAMPSIZE=, SAMPRATE=, or ALLOC= option together with METHOD=BALBOOTSTRAP.

For balanced bootstrap sampling, the output data set contains an observation for each unit that is selected (in each replicate sample). If you specify the OUTHITS option, the output data set also includes the variable NumberHits, which provides the number of selections in the replicate for each selected unit.

BERNOULLI

requests Bernoulli sampling, which consists of N independent selection trials, each with constant inclusion probability π, where N is the total number of sampling units in the stratum or data set. The sample size is not fixed but is a random variable. For more information, see the section "Bernoulli Sampling" on page 10225.

When you specify this method, you must provide the sampling rate (inclusion probability π) in the SAMPRATE= option. For stratified sampling (which you request with the STRATA statement), you can specify the same sampling rate for each stratum in the SAMPRATE=*value* option. Or you can specify different sampling rates for different strata in the SAMPRATE=(*values*) or SAMPRATE=*SAS-data-set* option.

Because Bernoulli sampling is based on a specified inclusion probability instead of a fixed sample size, METHOD=BERNOULLI does not use the SAMPSIZE= option. Also, the ALLOC= option in the STRATA statement (which allocates the total sample size among strata) is not available with METHOD=BERNOULLI.

POISSON

requests Poisson sampling. A generalization of Bernoulli sampling, Poisson sampling consists of *N* independent selection trials with a separate inclusion probability specified for each unit, where *N* is the total number of sampling units in the stratum or data set. The sample size is not fixed but is a random variable. For more information, see the section "Poisson Sampling" on page 10226. For a fixed-sample-size modification of Poisson sampling, see METHOD=SEQ_POISSON.

You must provide inclusion probabilities for Poisson sampling in the SIZE variable. The probability values should be between 0 and 1. If a value of the SIZE variable is missing, nonpositive, or greater than 1, PROC SURVEYSELECT omits the observation from sample selection.

Because Poisson sampling is based on specified inclusion probabilities instead of on a fixed sample size, you cannot specify the SAMPSIZE= option together with METHOD=POISSON. Similarly, you cannot specify the ALLOC= option (which allocates the total sample size among strata) together with METHOD=POISSON.

The SAMPLINGUNIT statement is not available when you specify METHOD=POISSON.

When you specify the SAMPRATE= option for METHOD=POISSON but do not specify a SIZE statement, PROC SURVEYSELECT uses METHOD=BERNOULLI.

PPS

requests selection with probability proportional to size and without replacement. For more information, see the section "PPS Sampling without Replacement" on page 10227. When you specify this method, you must name a size measure variable in the SIZE statement or specify the PPS option in the SAMPLINGUNIT statement.

PPS_BREWER

BREWER

requests selection according to Brewer's method. Brewer's method selects two units from each stratum with probability proportional to size and without replacement. For more information, see the section "Brewer's PPS Method" on page 10231. When you specify this method, you must name a size measure variable in the SIZE statement or specify the PPS option in the SAMPLINGUNIT statement. You do not need to specify the sample size in the SAMPSIZE= option because Brewer's method selects two units from each stratum.

PPS_MURTHY

MURTHY

requests selection according to Murthy's method. Murthy's method selects two units from each stratum with probability proportional to size and without replacement. For more information, see the section "Murthy's PPS Method" on page 10231. When you specify this method, you must name a size measure variable in the SIZE statement or specify the PPS option in the SAMPLINGUNIT statement. You do not need to specify the sample size in the SAMPSIZE= option because Murthy's method selects two units from each stratum.

PPS_SAMPFORD

SAMPFORD

requests selection according to Sampford's method. Sampford's method selects units with probability proportional to size and without replacement. For more information, see the section "Sampford's PPS Method" on page 10232. When you specify this method, you must name a size measure variable in the SIZE statement or specify the PPS option in the SAMPLINGUNIT statement.

PPS_SEQ

CHROMY

requests sequential selection with probability proportional to size and with minimum replacement. This method is also known as Chromy's method. For more information, see the section "PPS Sequential Sampling" on page 10229. When you specify this method, you must name a size measure variable in the SIZE statement or specify the PPS option in the SAMPLINGUNIT statement.

PPS_SYS < (*method-options*) >

requests systematic selection with probability proportional to size. For more information, see the section "PPS Systematic Sampling" on page 10229. When you specify this method, you must provide size measures by specifying the SIZE statement or the PPS option in the SAMPLINGUNIT statement.

You can specify the following *method-options*:

DETAILS

displays the random start and the systematic interval in the "Sample Selection Summary" table when the design does not include strata or replicates. For more information, see the section "Displayed Output" on page 10243.

INTERVAL=*value*

specifies the interval *value* for PPS systematic selection. The interval value must be a positive number. It must not exceed the total of the size measures in the data set (or in each stratum if you specify a STRATA statement). By default, the systematic interval is the ratio of the size measure total to the sample size (which you provide in the SAMPSIZE= option). For more information, see the section "PPS Systematic Sampling" on page 10229.

You cannot use the INTERVAL= *method-option* when you specify a sample size in the SAMPSIZE= option or when you specify the ALLOC= option, which allocates the total sample size among strata.

START=*value*

specifies the starting *value* for PPS systematic selection. The starting value must be a positive number that is less than the systematic interval. By default, PROC SURVEYSELECT randomly determines a starting point in the systematic interval. For more information, see the section "PPS Systematic Sampling" on page 10229.

When you use this option to specify a systematic starting point (instead of allowing the procedure to randomly determine the starting point), the following options for random number generation have no effect: SEED=, RANUNI, and OUTSEED. You cannot use the REPS= option when you specify the START= *method-option*.

When the starting value that you provide is not randomly determined, the resulting selection is not a probability-based sample.

PPS_WR

requests selection with probability proportional to size and with replacement. For more information, see the section "PPS Sampling with Replacement" on page 10228. When you specify this method, you must name a size measure variable in the SIZE statement or specify the PPS option in the SAMPLINGUNIT statement.

SEQ

CHROMY

requests sequential selection according to Chromy's method. If you specify this method and do not specify a SIZE statement (or the PPS option in the SAMPLINGUNIT statement), PROC SURVEYSELECT uses sequential zoned selection with equal probability and without replacement. For more information, see the section "Sequential Random Sampling" on page 10224.

If you specify METHOD=SEQ and also specify a SIZE statement (or the PPS option in the SAMPLINGUNIT statement), PROC SURVEYSELECT uses METHOD=PPS_SEQ, which is sequential selection with probability proportional to size and with minimum replacement. For more information, see the section "PPS Sequential Sampling" on page 10229.

SEQ_POISSON

requests sequential Poisson sampling, which is a fixed-sample-size modification of Poisson sampling (METHOD=POISSON). For more information, see the section "Sequential Poisson Sampling" on page 10226.

When you request this method, you must provide size measures in the SIZE variable and you must specify the sample size in the SAMPSIZE= option.

If you request this method, you cannot specify a SAMPLINGUNIT statement.

SRS

requests simple random sampling, which is selection with equal probability and without replacement. For more information, see the section "Simple Random Sampling" on page 10222. METHOD=SRS is the default selection method if you do not specify the METHOD= option and also do not specify a SIZE statement (or the PPS option in the SAMPLINGUNIT statement).

SYS < (*method-options*) >

requests systematic random sampling. If you specify this method and do not specify a SIZE statement (or the PPS option in the SAMPLINGUNIT statement), PROC SURVEYSELECT uses systematic random sampling with equal probability. For more information, see the section "Systematic Random Sampling" on page 10223.

If you specify this method and also specify a SIZE statement (or the PPS option in the SAMPLINGUNIT statement), PROC SURVEYSELECT uses systematic random sampling with probability proportional to size (METHOD=PPS_SYS). For more information, see the section "PPS Systematic Sampling" on page 10229.

You can specify the following *method-options*:

DETAILS

displays the random start and the systematic interval in the "Sample Selection Summary" table when the design does not include strata or replicates. For more information, see the section "Displayed Output" on page 10243.

INTERVAL=*value*

specifies the interval for systematic random sampling. The interval *value* must be a positive number and must not exceed the number of sampling units in the data set (or the number of units in each stratum, if you specify a STRATA statement).

By default, PROC SURVEYSELECT determines the systematic interval from the sampling rate or sample size that you provide in the SAMPRATE= or SAMPSIZE= option, respectively. When you specify the sampling rate, PROC SURVEYSELECT computes the systematic interval as the inverse of the sampling rate. When you specify the sample size, the procedure computes the interval as the ratio of the number of sampling units to the sample size. For more information, see the section "Systematic Random Sampling" on page 10223.

You cannot use the INTERVAL= *method-option* when you specify the SAMPSIZE= option, the SAMPRATE= option, or the ALLOC= option (which allocates the total sample size among strata).

START=*value*

specifies the starting *value* for systematic selection. The starting value must be a positive number that is less than the systematic interval. By default, PROC SURVEYSELECT randomly determines a starting point in the systematic interval. For more information, see the section "Systematic Random Sampling" on page 10223.

When you use this option to specify a systematic starting point (instead of allowing the procedure to randomly determine the starting point), the following options for random number generation have no effect: SEED=, RANUNI, and OUTSEED. You cannot use the REPS= option when you specify the START= *method-option*.

When the starting value that you provide is not randomly determined, the resulting selection is not a probability-based sample.

URS

requests unrestricted random sampling, which is selection with equal probability and with replacement. For more information, see the section "Unrestricted Random Sampling" on page 10223.

MINSIZE < =*value* **|** *SAS-data-set* **>**

specifies the minimum size measure, which PROC SURVEYSELECT uses to adjust size measures before sample selection. When a size measure is less than the minimum size value, PROC SURVEYSELECT replaces that size measure with the minimum size value.

You can provide a single minimum size *value* for the entire sample selection, or you can provide stratum-level minimum size values by specifying a *SAS-data-set*. The minimum size values must be positive numbers.

You provide size measures by specifying the SIZE statement or by specifying the PPS option in the SAMPLINGUNIT statement.

Unless you specify a SAMPLINGUNIT statement, the OUT= data set includes a variable named AdjustedSize that contains the adjusted size measures.

If you use a SAMPLINGUNIT statement to define sampling units (clusters), PROC SURVEYSELECT adjusts the sampling unit sizes (not the observation sizes). If you specify a SIZE statement, the size of a sampling unit is the sum of the size measures of the observations in the unit. If you specify the SAMPLINGUNIT PPS option, the size of a sampling unit is the number of observations in the unit.

When you use a SAMPLINGUNIT statement, the OUT= data set includes a variable named UnitSize that contains the adjusted sampling unit size measures.

You can provide minimum size values by specifying one of the following forms:

MINSIZE

indicates that minimum size values are provided in a secondary input data set that you name in another option (for example, the SAMPSIZE=*SAS-data-set* option). You should provide the minimum size values in the data set variable named _MINSIZE_. For more information, see the section "Secondary Input Data Set" on page 10236. You can specify only one secondary input data set in each invocation of PROC SURVEYSELECT.

MINSIZE=*value*

specifies a single minimum size *value*, which must be a positive number. If you request a stratified sample design by specifying the STRATA statement, PROC SURVEYSELECT uses the minimum value to adjust size measures in all strata.

MINSIZE=*SAS-data-set*

names a *SAS-data-set* that contains stratum-level minimum size values. You should provide the minimum size values in the data set variable named _MINSIZE_. Each observation in the data set should correspond to a stratum group, which is determined by the values of the STRATA variables.

This data set, which is a secondary input data set, must contain all stratification variables that you specify in the STRATA statement. The data set must also contain all stratum groups that appear in the DATA= input data set. The order of the stratum groups in the MINSIZE= data set must match the order of the groups in the DATA= input data set. If formats are associated with the STRATA variables, the formats must be consistent in the two data sets. For more information, see the section "Secondary Input Data Set" on page 10236. You can name only one secondary input data set in each invocation of PROC SURVEYSELECT.

NMAX=*n*

specifies the maximum stratum sample size *n* for the SAMPRATE= option. When you specify the SAMPRATE= option, PROC SURVEYSELECT calculates the stratum sample size by multiplying the total number of units in the stratum by the specified sampling rate. If this sample size is greater than the value NMAX=*n*, PROC SURVEYSELECT selects only *n* units.

The maximum sample size *n* must be a positive integer. The NMAX= option is available only with the SAMPRATE= option, which you can specify for equal probability selection methods (METHOD=SRS, METHOD=URS, METHOD=SYS, and METHOD=SEQ). The NMAX= option is not available with METHOD=BERNOULLI, where the SAMPRATE= option specifies the constant inclusion probability.

NMIN=*n*

specifies the minimum stratum sample size *n* for the SAMPRATE= option. When you specify the SAMPRATE= option, PROC SURVEYSELECT calculates the stratum sample size by multiplying the total number of units in the stratum by the specified sampling rate. If this sample size is less than the value NMIN=*n*, PROC SURVEYSELECT selects *n* units.

The minimum sample size *n* must be a positive integer. The NMIN= option is available only with the SAMPRATE= option, which you can specify for equal probability selection methods (METHOD=SRS, METHOD=URS, METHOD=SYS, and METHOD=SEQ). The NMIN= option is not available with METHOD=BERNOULLI, where the SAMPRATE= option specifies the constant inclusion probability.

NOPRINT

suppresses the display of all output. You can use the NOPRINT option when you want only to create an output data set. This option temporarily disables the Output Delivery System (ODS). For more information, see Chapter 22, “Using the Output Delivery System.”

OUT=*SAS-data-set*

names the output data set. If you omit the OUT= option, the data set is named DATA*n*, where *n* is the smallest integer that makes the name unique. If you request sample selection by specifying the METHOD= option, the output data set contains the observations that are selected for the sample. If you request sample allocation without sample selection by specifying the ALLOC= and NOSAMPLE options in the STRATA statement, the output data set contains the allocated sample sizes. If you request random assignment by specifying the GROUPS= option, the output data set contains all observations in the input data set together with their assigned group identification.

When PROC SURVEYSELECT selects a sample, the output data set contains the units that are selected, sample design information, and selection statistics. You can specify options that control the information to include in the output data set. For more information, see the descriptions of the following options: JTPROBS, OUTALL, OUTHITS, OUTSEED, OUTSIZE, and STATS. For more information, see the section “Sample Output Data Set” on page 10237.

By default, the sample output data set contains only those units that are selected for the sample. To include all observations from the input data set in the output data set, use the OUTALL option.

By default, the sample output data set includes one copy of each selected unit, even when a unit is selected more than once, which can occur when you use with-replacement or with-minimum-replacement selection methods. For with-replacement or with-minimum-replacement selection methods, the output data set includes a variable NumberHits that records the number of hits (selections) for each unit. To include a distinct copy of each selection in the output data set when the same unit is selected more than once, use the OUTHITS option.

When you specify the ALLOC= and NOSAMPLE options in the STRATA statement, PROC SURVEYSELECT allocates the total sample size among the strata but does not select a sample. In this case, the OUT= data set contains the allocated sample sizes. For more information, see the section “Allocation Output Data Set” on page 10241.

When you specify the GROUPS= option, PROC SURVEYSELECT randomly assigns observations to groups; it does not select a sample. In this case, the OUT= data set contains all observations from the input data set and includes a variable named GroupID that identifies group assignments. For more information, see the section “Random Assignment Output Data Set” on page 10242.

OUTALL < (ZEROSTRATA) >

includes all observations from the sampling frame in the OUT= output data set. By default, the output data set includes only those units selected for the sample. When you specify the OUTALL option, the output data set includes all observations in the sampling frame along with a variable (Selected) that indicates each observation's selection status. The value of Selected is 1 for an observation that is selected or 0 for an observation that is not selected. For information about the contents of the output data set, see the section "Sample Output Data Set" on page 10237.

The OUTALL option is available for equal probability selection methods (METHOD=SRS, METHOD=URS, METHOD=SYS, METHOD=SEQ, and METHOD=BERNOULLI). and for METHOD=POISSON.

If you specify a sample size of 0 for a stratum, PROC SURVEYSELECT omits this stratum from the sampling frame. By default, PROC SURVEYSELECT also omits this stratum from the output data set when you specify the OUTALL option. You can specify the OUTALL(ZEROSTRATA) option to include strata that have sample sizes of 0 in the output data set. For more information, see the description of the SAMPSIZE= option.

OUTHITS

includes a distinct copy of each selected unit in the OUT= output data set when the same sampling unit is selected more than once. By default, the output data set contains a single copy of each unit selected, even when a unit is selected more than once, and the variable NumberHits records the number of hits (selections) for each unit. If you specify the OUTHITS option, the output data set contains *m* copies of a sampling unit for which NumberHits is *m*; for example, the output data set contains three copies of a unit that is selected three times (NumberHits is 3).

A sampling unit can be selected more than once by with-replacement and with-minimum-replacement selection methods, which include METHOD=URS, METHOD=BALBOOTSTRAP, METHOD=PPS_WR, METHOD=PPS_SYS, and METHOD=PPS_SEQ. The OUTHITS option is available for these selection methods.

For information about the contents of the output data set, see the section "Sample Output Data Set" on page 10237.

OUTORDER=RANDOM < (*option*) >
OUTRANDOM < (*option*) >

randomly orders the sample observations in the OUT= output data set. If you specify a STRATA statement, the sample observations are randomly ordered within stratum groups. If you specify the REPS= option, the sample observations are randomly ordered within sample replicates (nested within stratum groups). This option does not affect the selection of the sample; it randomly orders the selected observations after sample selection is complete.

By default for most sample selection methods, the order of the sample observations in the OUT= data set corresponds to the order in the DATA= input data set. If you specify a CONTROL statement to sort the input data set for systematic selection methods, the order in the output data set is the same as the sorted order of the input data set. If you specify a sequential selection method, the order in the output data set is the selection order (which begins at a randomly chosen sampling unit). By default for METHOD=PPS, the order of the sample observations corresponds to the ascending order of the size measures.

The OUTORDER=RANDOM option is not available when you specify the GROUPS= option, a SAMPLINGUNIT statement, or the NOSAMPLE option in the STRATA statement.

You can specify the following *option*:

SEED=*value*

specifies the initial random number seed for randomly ordering the sample observations. To initialize random number generation, *value* must be a positive integer. If you do not specify this option or if you specify a *value* that is negative or 0, PROC SURVEYSELECT uses the existing random number stream that is used for sample selection. For more information, see the SEED= option and the section "Random Number Generation" on page 10221.

OUTSEED

includes the initial seed for each stratum in the OUT= output data set. The variable InitialSeed contains the stratum initial seeds. For information about the contents of the output data set, see the section "Sample Output Data Set" on page 10237. The OUTSEED option is not available when you specify the STRATUMSEED=NONE option for a stratified sample.

To reproduce the same sample for any stratum in a subsequent execution of PROC SURVEYSELECT, you can specify the same stratum initial seed in the SEED=*SAS-data-set* option together with the same sample selection parameters. For more information, see the section "Random Number Generation" on page 10221.

The "Sample Selection Summary" table displays the initial random number seed for the entire sample selection, which is the initial seed for the first stratum when the design is stratified. To reproduce the entire sample, you can specify this same seed value in the SEED= option, along with the same sample selection parameters.

OUTSIZE

includes additional design and sampling frame information in the OUT= output data set.

If you use a STRATA statement, the OUTSIZE option provides stratum-level values in the output data set. Otherwise, the OUTSIZE option provides overall values.

The OUTSIZE option includes the sample size or sampling rate in the output data set, depending on whether you specify the SAMPSIZE= option or the SAMPRATE= option, respectively. For PPS selection methods, the OUTSIZE option includes the total size measure in the output data set. If you do not provide size measures, or if you specify a SAMPLINGUNIT statement, the OUTSIZE option includes the total number of sampling units in the output data set.

If you request size measure adjustment or certainty selection, the OUTSIZE option includes the following information in the output data set: the minimum size measure if you specify the MINSIZE= option, the maximum size measure if you specify the MAXSIZE= option, the certainty size measure if you specify the CERTSIZE= option, and the certainty proportion if you specify the CERTSIZE=P= option.

For METHOD=BERNOULLI, the OUTSIZE option includes the following information in the output data set: total number of sampling units, selection probability (sampling rate), expected sample size, and actual sample size. See the section "Bernoulli Sampling" on page 10225 for descriptions of these statistics.

For more information about the contents of the output data set, see the section "Sample Output Data Set" on page 10237.

If you specify the GROUPS= option for random assignment, the OUTSIZE option adds the following information to the output data set: total number of units, number of groups, and number of units in the group. For more information, see the section "Random Assignment Output Data Set" on page 10242.

OUTSORT=*SAS-data-set*

names an output data set to store the sorted input data set. This option is available when you specify a CONTROL statement to sort the DATA= input data set for systematic or sequential selection methods (METHOD=SYS, METHOD=PPS_SYS, METHOD=SEQ, and METHOD=PPS_SEQ).

If you specify CONTROL variables but do not name an output data set in the OUTSORT= option, the sorted data set replaces the input data set.

RANUNI

requests uniform random number generation by the method of Fishman and Moore (1982), which is the random number generator that the RANUNI function provides. For more information, see the section "Random Number Generation" on page 10221. For information about the RANUNI function, see *SAS Functions and CALL Routines: Reference*.

By default, PROC SURVEYSELECT uses the Mersenne twister random number generator (Matsumoto and Nishimura 1998). The Mersenne twister random number generator has a very long period and good statistical properties. This is the random number generator that the RAND function provides for the uniform distribution. For information about the RAND function, see *SAS Functions and CALL Routines: Reference*.

If you use the RANUNI option to select a sample, you can reproduce this same sample by specifying the same SEED= value together with the RANUNI option (for the same input data set and sample selection parameters).

In releases before SAS/STAT 12.1, PROC SURVEYSELECT uses the RANUNI random number generator by default. To reproduce samples from releases before SAS/STAT 12.1, you can specify the same SEED= value together with the RANUNI option (for the same input data set and sample selection parameters).

REPS=*nreps* **<** **(***option***)** **>**

requests replicated sampling and specifies the number of replicate samples. The value of *nreps* must be a positive integer.

When you specify this option, PROC SURVEYSELECT selects *nreps* independent replicate samples. Each replicate sample is selected by using the same sample size (or sampling rate) and design parameters that you specify.

By default, the variable named Replicate in the OUT= data set contains replicate identification numbers for the sample observations. You can specify a different name for the replicate identification variable in the REPNAME= suboption.

You can use replicated sampling to provide a simple method of variance estimation for any form of statistic and to evaluate variable nonsampling errors such as interviewer differences. For more information, see Lohr (2010), Wolter (2007), Kish (1965), Kish (1987), and Kalton (1983). You can also use the REPS= option to perform a variety of other resampling and simulation tasks. For more information, see Cassell (2007).

You can specify the following *option*:

REPNAME=*name*

specifies the *name* of the replicate identification variable in the OUT= data set. By default, the variable name is Replicate. For more information, see the section "Sample Output Data Set" on page 10237.

ROUND=*type*

specifies the type of rounding to use when the sampling rate is converted to a positive, integer-valued sample size. This option is available when you specify the SAMPRATE= option for one of the following equal probability selection methods: METHOD=SRS, METHOD=URS, METHOD=SYS, and METHOD=SEQ.

For these selection methods, PROC SURVEYSELECT converts the sampling rate that you specify in the SAMPRATE= option to an integer-valued sample size before selecting the sample. PROC SURVEYSELECT multiplies the total number of units (in the stratum or in the data set) by the sampling rate to obtain the target sample size. If the target sample size is not an integer, the procedure rounds this number to a positive integer value. By default, PROC SURVEYSELECT always rounds the target sample size up to the nearest integer (ROUND=UP).

To provide positive selection probabilities for all sampling units, all stratum sample sizes must be greater than 0. Therefore, when a target sample size is less than 1, it is always rounded up to 1 (instead of down to 0), regardless of the ROUND=*type*.

You can specify one of the following *types*:

ALTERNATE

alternates the rounding direction (up or down) by strata. This option rounds up for the first stratum, down for the second stratum, up for the third stratum, and so on, where the strata are processed in the order in which they appear in the input data set. The alternating sequence skips strata for which the target sample size is an integer (and therefore does not require rounding). This option has no effect unless you request stratified sampling by specifying a STRATA statement.

DOWN

FLOOR

rounds down to the largest integer that does not exceed the target sample size.

NEAREST < (HALF=DOWN) >

rounds the target sample size to the nearest integer. This option rounds up when the fractional part of the target sample size is greater than 0.5 and rounds down when the fractional part is less than 0.5. By default, the sample size is rounded up when the fractional part is exactly 0.5. When you specify ROUND=NEAREST(HALF=DOWN), the sample size is rounded down when the fractional part is 0.5.

RANDOM

determines the rounding direction (up or down) randomly. Each rounding direction is equally likely (with probability of 0.5).

UP

CEILING

rounds up to the smallest integer that is not less than the target sample size. This option is the default.

SAMPRATE=*value* **|** **(***values***)|** *SAS-data-set*

RATE=*value* **|** **(***values***)|** *SAS-data-set*

specifies the sampling rate, which is the proportion of units to select for the sample. You can provide a single sampling rate *value* for the entire sample selection, or you can provide stratum sampling rates by specifying *values* or a *SAS-data-set*.

The sampling rate value must be a positive number. The stratum sampling rate values and the stratum sampling rates that you provide in the *SAS-data-set* must be nonnegative numbers. You can specify a sampling rate as a number between 0 and 1. Or you can specify a rate in percentage form as a number between 1 and 100, which PROC SURVEYSELECT converts to a proportion. The procedure treats the value 1 as 100% instead of 1%.

This option is available for the equal probability selection methods, as follows:

- For METHOD=SYS (systematic random sampling), PROC SURVEYSELECT computes the selection interval as the inverse of the sampling rate. For more information, see the section "Systematic Random Sampling" on page 10223.
- For METHOD=BERNOULLI (Bernoulli sampling), the procedure uses the sampling rate as the inclusion probability. For more information, see the section "Bernoulli Sampling" on page 10225.
- For the other equal probability selection methods (METHOD=SRS, METHOD=URS, METHOD=SYS, and METHOD=SEQ), the procedure converts the sampling rate to an integer-valued sample size before selecting the sample.

To convert the sampling rate to an integer-valued sample size, PROC SURVEYSELECT multiplies the total number of units (in the stratum or data set) by the sampling rate that you specify to obtain the target sample size. If this number is not an integer, the target sample size is rounded to a positive integer. By default, PROC SURVEYSELECT always rounds the target sample size up to the nearest integer. You can specify other types of rounding by using the ROUND= option.

You cannot specify the SAMPRATE= option together with the SAMPSIZE= option.

You can provide sampling rates by specifying one of the following forms:

SAMPRATE=*value*

RATE=*value*

specifies a single sampling rate *value*, which must be a positive number. If you request a stratified sample design by specifying the STRATA statement, PROC SURVEYSELECT uses the rate value for all strata.

SAMPRATE=(*values***)**

RATE=(*values***)**

specifies a list of stratum sampling rate *values*. You can separate the values with blanks or commas, and you must enclose the list of values in parentheses. The number of stratum sampling rate values should equal the number of strata in the input data set.

The order of the stratum sampling rate values must match the order of the stratum groups in the DATA= input data set. When you specify a list of values, the input data set must be sorted by the STRATA variables in ascending order; you cannot use the DESCENDING or NOTSORTED option in the STRATA statement.

The stratum sampling rate values must be nonnegative numbers. If you specify a stratum sampling rate of 0, PROC SURVEYSELECT does not select a sample from the stratum. This has the effect of subsetting the input data set before sample selection; the stratum that you omit is not included in the sampling frame or represented in the sample.

SAMPRATE=*SAS-data-set*

RATE=*SAS-data-set*

names a *SAS-data-set* that contains stratum sampling rates. You should provide the sampling rates in the data set variable named _RATE_. Each observation in the data set should correspond to a stratum group, which is determined by the values of the STRATA variables.

This data set, which is a secondary input data set, must contain all stratification variables that you specify in the STRATA statement. The data set must also contain all stratum groups that appear in the DATA= input data set. The order of the stratum groups in the SAMPRATE= data set must match the order of the groups in the DATA= data set. If formats are associated with the STRATA variables, the formats must be consistent in the two data sets. For more information, see the section "Secondary Input Data Set" on page 10236. You can name only one secondary input data set in each invocation of PROC SURVEYSELECT.

The stratum sampling rates must be nonnegative numbers. If you specify a stratum sampling rate of 0, PROC SURVEYSELECT does not select a sample from the stratum. This has the effect of subsetting the input data set before sample selection; the stratum that you omit is not included in the sampling frame or represented in the sample.

SAMPSIZE=*n* **|(***values***)|** *SAS-data-set*

N=*n* **| (***values***)|** *SAS-data-set*

specifies the sample size, which is the number of units to select for the sample. You can provide a single sample size *n* for the entire sample selection, or you can provide stratum sample sizes by specifying *values* or a *SAS-data-set*.

The value of *n* must be a positive integer. The stratum sample size values and the stratum sample sizes that you provide in the *SAS-data-set* must be nonnegative numbers. For selection methods that select without replacement, the sample size must not exceed the total number of units in the data set (or the number of units in the stratum, if you specify a STRATA statement).

This option specifies the number of sampling units to select. If you do not specify a SAMPLINGUNIT statement, PROC SURVEYSELECT defines sampling units as observations and selects the number of observations that you specify. If you specify a SAMPLINGUNIT statement, PROC SURVEYSELECT defines sampling units as groups of observations (clusters) and selects the number of clusters that you specify.

If you specify SAMPSIZE=*n* and the ALLOC= option in the STRATA statement, PROC SURVEYSELECT allocates the sample size *n* among the strata according to the allocation method that you request. For more information, see the section "Sample Size Allocation" on page 10232. You cannot specify SAMPSIZE=*values* or SAMPSIZE=*SAS-data-set* when you use the ALLOC= option. You cannot specify SAMPSIZE= with the MARGIN= option, which determines stratum sample sizes that provide the specified margin of error. For more information, see the section "Specifying the Margin of Error" on page 10235.

You cannot specify both the SAMPSIZE= option and the SAMPRATE= option.

You can provide sample size values by specifying one of the following forms:

SAMPSIZE=*n*

N=*n*

specifies a single sample size value *n*, which must be a positive integer. If you request a stratified sample design, PROC SURVEYSELECT selects *n* units from each stratum (unless you also specify the ALLOC= option in the STRATA statement, which allocates the total sample size among the strata).

For methods that select without replacement, the sample size *n* must not exceed the number of units in the stratum unless you also specify the SELECTALL option. If you specify the SELECTALL option, PROC SURVEYSELECT selects all stratum units when the stratum sample size exceeds the total number of units in the stratum.

SAMPSIZE=(*values*)

N=(*values*)

specifies a list of stratum sample size *values*. You can separate the values with blanks or commas, and you must enclose the list of values in parentheses. The number of sample size values must equal the number of strata in the input data set.

The order of the stratum sample size values must match the order of the stratum groups in the DATA= input data set. When you specify a list of values, the input data set must be sorted by the STRATA variables in ascending order; you cannot use the DESCENDING or NOTSORTED option in the STRATA statement.

The values of the stratum sample sizes must be nonnegative numbers. If you specify a stratum sample size of 0, PROC SURVEYSELECT does not select a sample from the stratum. This has the effect of subsetting the input data set before sample selection; the stratum that you omit is not included in the sampling frame or represented in the sample.

SAMPSIZE=*SAS-data-set*

N=*SAS-data-set*

names a *SAS-data-set* that contains stratum sample sizes. You should provide the sample sizes in the data set variable named _NSIZE_ or SampleSize. Each observation in the data set should correspond to a stratum group, which is determined by the values of the STRATA variables.

This data set, which is a secondary data set, must contain all stratification variables that you specify in the STRATA statement. The data set must also contain all stratum groups that appear in the DATA= input data set. The order of the stratum groups in the SAMPSIZE= data set must match the order of the groups in the DATA= data set. If formats are associated with the STRATA variables, the formats must be consistent in the two data sets. For more information, see the section "Secondary Input Data Set" on page 10236. You can name only one secondary input data set in each invocation of PROC SURVEYSELECT.

The stratum sample sizes must be nonnegative numbers. If you specify a stratum sample size of 0, PROC SURVEYSELECT does not select a sample from the stratum. This has the effect of subsetting the input data set before sample selection; the stratum that you omit is not included in the sampling frame or represented in the sample.

SEED < =*value* | *SAS-data-set* >

specifies the initial seed for random number generation. You can provide a single seed *value* for the entire sample selection, or you can provide stratum initial seeds by specifying a *SAS-data-set*. To initialize random number generation, a seed must be a positive integer. If you do not specify this option, or if you specify an initial seed that is negative or 0, PROC SURVEYSELECT uses the time of day from the computer's clock to obtain an initial seed. For more information, see the section "Random Number Generation" on page 10221.

PROC SURVEYSELECT displays the value of the initial seed in the "Sample Selection Summary" table. To reproduce the same sample in a subsequent execution of PROC SURVEYSELECT, you can specify the same initial seed in the SEED= option (for the same input data set and sample selection parameters).

If you specify a STRATA statement, you can provide stratum initial seeds in a *SAS-data-set*. If you do not provide stratum initial seeds and if you use the RANUNI random number generator, PROC SURVEYSELECT generates random numbers continuously across strata (from a single pseudorandom number stream). You can specify the OUTSEED option to include the stratum initial seeds in the output data set.

If you do not provide stratum initial seeds and if you use the (default) Mersenne twister random number generator, PROC SURVEYSELECT generates separate pseudorandom number streams for the strata by default. You can specify the STRATUMSEED=NONE option to use a single Mersenne twister stream across all strata. For more information, see the STRATUMSEED= option.

You can provide initial seeds by specifying one of the following forms:

SEED

indicates that stratum initial seeds are provided in a secondary input data set that you name in another option (for example, the SAMPSIZE=*SAS-data-set* option). You should provide the initial seeds in the data set variable named _SEED_ or InitialSeed. For more information, see the section "Secondary Input Data Set" on page 10236. You can name only one secondary input data set in each invocation of PROC SURVEYSELECT.

SEED=*value*

specifies a single initial seed *value* for random number generation. To initialize random number generation, the value must be a positive integer.

SEED=*SAS-data-set*

names a *SAS-data-set* that contains stratum initial seeds. You should provide the stratum initial seeds in the data set variable named _SEED_ or InitialSeed. Each observation in the data set should correspond to a stratum group, which is determined by the values of the STRATA variables.

This data set, which is a secondary input data set, must contain all stratification variables that you specify in the STRATA statement. The data set must also contain all stratum groups that appear in the DATA= input data set. The order of the stratum groups in the SEED= data set must match the order of the groups in the DATA= data set. If formats are associated with the STRATA variables, the formats must be consistent in the two data sets. For more information, see the section "Secondary Input Data Set" on page 10236. You can name only one secondary input data set in each invocation of PROC SURVEYSELECT.

The OUTSEED option includes the stratum initial seeds in the OUT= output data set. You can reproduce the same sample in a subsequent execution of PROC SURVEYSELECT by specifying the same stratum initial seeds (for the same input data set and sample selection parameters). If you need to reproduce the same sample for only a subset of the strata, you can use the same initial seeds for the strata in the subset.

SELECTALL

requests that PROC SURVEYSELECT select all stratum units when the stratum sample size exceeds the total number of units in the stratum. By default, PROC SURVEYSELECT does not allow you to specify a stratum sample size that is greater than the total number of units in the stratum unless you are using a with-replacement selection method.

The SELECTALL option is available for the following without-replacement selection methods: METHOD=SRS, METHOD=SYS, METHOD=SEQ, METHOD=PPS, and METHOD=PPS_SAMPFORD.

The SELECTALL option is not available for with-replacement selection methods, with-minimum-replacement methods, or those PPS methods that select two units per stratum.

SORT=NEST | SERP

specifies the type of sorting to perform when you specify a CONTROL statement for systematic or sequential sample selection.

SORT=NEST requests nested sorting, and SORT=SERP requests hierarchic serpentine sorting. For more information, see the section "Sorting by CONTROL Variables" on page 10220.

By default, SORT=SERP. Where there is only one CONTROL variable, the two types of sorting are equivalent.

The SORT= option is available when you specify a CONTROL statement for systematic or sequential selection methods (METHOD=SYS, METHOD=PPS_SYS, METHOD=SEQ, and METHOD=PPS_SEQ). When you specify a CONTROL statement, PROC SURVEYSELECT sorts the input data set by the CONTROL variables within strata before selecting the sample.

When you specify a CONTROL statement, you can use the OUTSORT= option to name an output data set that contains the sorted input data set. Otherwise, if you do not specify the OUTSORT= option, the sorted data set replaces the input data set.

The SORT= option and the CONTROL statement are not available when you specify a SAMPLINGUNIT statement.

STATS

includes the selection probability and sampling weight in the OUT= output data set for equal probability selection methods when you do not specify a STRATA statement. By default, the output data set does not include these values for equal probability selection methods unless you specify a STRATA statement. The STATS option applies to the following selection methods: METHOD=SRS, METHOD=URS, METHOD=SYS, METHOD=SEQ, METHOD=BALBOOTSTRAP, and METHOD=BERNOULLI.

In addition to the selection probability and sampling weight, the STATS option includes the following statistics in the output data set for METHOD=BERNOULLI: total number of sampling units, expected sample size, actual sample size, and adjusted sampling weight. For more information, see the section "Bernoulli Sampling" on page 10225.

For PPS selection methods, the output data set contains selection probabilities and sampling weights by default. The STATS option has no effect for PPS methods.

For more information about the contents of the output data set, see the section "Sample Output Data Set" on page 10237.

STRATUMSEED=NONE | RESTORE

controls stratum initial seeds for the Mersenne twister random number generator, which PROC SURVEYSELECT uses by default unless you specify the RANUNI option.

By default, PROC SURVEYSELECT uses separate, independent pseudorandom number streams for the strata. You can store the stratum initial seeds in the output data set by specifying the OUTSEED option. You can reproduce the entire sample (over all strata) by specifying the same initial seed in the SEED= option. You can also reproduce individual stratum samples by specifying the corresponding stratum initial seeds in the SEED= option.

The STRATUMSEED= option applies only to stratified selection, which you request by specifying a STRATA statement. The STRATUMSEED= option applies only to the Mersenne twister random number generator; you cannot specify this option together with the RANUNI option (which requests the RANUNI random number generator).

You cannot specify the STRATUMSEED= option when you provide stratum initial seeds by using the SEED=*SAS-data-set* option.

For more information, see the section "Random Number Generation" on page 10221.

You can specify one of the following keywords:

NONE

uses a single pseudorandom number stream across all strata instead of separate pseudorandom number streams for the strata.

When you specify this option, stratum-level initial seeds are not available; you can reproduce the entire sample (over all strata), but you cannot reproduce individual stratum samples separately (apart from the overall sample). To reproduce an entire sample that the procedure selects when you specify the STRATUMSEED=NONE option, you can specify the same SEED= value, input data set, and selection parameters (along with the STRATUMSEED=NONE option). The OUTSEED option, which stores stratum initial seeds in the output data set, is not available when you specify STRATUMSEED=NONE.

RESTORE

reproduces the stratum initial seeds that PROC SURVEYSELECT uses in releases before SAS/STAT 14.1 for the Mersenne twister random number generator. To reproduce a stratified sample that PROC SURVEYSELECT selects in releases before SAS/STAT 14.1, you can specify STRATUMSEED=RESTORE along with the same SEED= value, input data set, and selection parameters.

CONTROL Statement

CONTROL *variables* **;**

The CONTROL statement names one or more *variables* for sorting the input data set before sample selection. The CONTROL variables can be character or numeric. If you also specify a STRATA statement, PROC SURVEYSELECT sorts by CONTROL variables within strata.

Control sorting is available for systematic and sequential selection methods (METHOD=SYS, METHOD=PPS_SYS, METHOD=SEQ, and METHOD=PPS_SEQ). Ordering the sampling units before systematic or sequential selection can provide additional control over the distribution of the sample.

Control sorting is not available when you use a SAMPLINGUNIT statement, which defines groups of observations as units (clusters) for sample selection. See the description of the SAMPLINGUNIT statement for information about ordering clusters before systematic or sequential selection.

By default (or if you specify the SORT=SERP option in the PROC SURVEYSELECT statement), PROC SURVEYSELECT uses hierarchic serpentine sorting by the CONTROL variables. If you specify the SORT=NEST option, the procedure uses nested sorting. For more information about serpentine and nested sorting, see the section "Sorting by CONTROL Variables" on page 10220.

You can use the OUTSORT= option in the PROC SURVEYSELECT statement to name an output data set that contains the sorted input data set. If you do not specify the OUTSORT= option when you use the CONTROL statement, then the sorted data set replaces the input data set.

FREQ Statement

FREQ *variable* **;**

The FREQ statement names a numeric *variable* that contains the frequency of occurrence of each observation. If you use a FREQ statement, PROC SURVEYSELECT assumes that an observation represents n observations, where n is the value of the FREQ variable for the observation. The FREQ statement is not available when you specify a SAMPLINGUNIT statement.

The FREQ statement is available only for sample allocation when no sample is selected, which you can request by specifying the ALLOC= and NOSAMPLE options in the STRATA statement. The ALLOC= option requests allocation of the total sample size among the strata, and the NOSAMPLE option requests that no sample be selected after allocation. When you specify the NOSAMPLE option, the procedure computes stratum sample sizes according to the allocation method that you request, but does not select the sample. For more information, see the section "Sample Size Allocation" on page 10232.

The sum of the FREQ variable values (frequencies) represents the total number of sampling units. The sum of the frequencies in a stratum represents the total number of sampling units in the stratum. When you use a FREQ statement, the sample size allocation is based on the expanded total and stratum frequencies.

Values of the FREQ variable must be nonmissing and nonnegative. If a value of the FREQ variable is 0, PROC SURVEYSELECT ignores the observation. If a value of the FREQ variable is not an integer, PROC SURVEYSELECT uses only the integer portion as the frequency of the observation.

ID Statement

ID *variables* **;**

The ID statement names one or more *variables* from the DATA= input data set to include in the OUT= output data set of selected units. If there is no ID statement, PROC SURVEYSELECT includes all variables from the input data set in the output data set. The ID variables can be either character or numeric.

SAMPLINGUNIT | CLUSTER Statement

SAMPLINGUNIT | CLUSTER *variables < / options >* **;**

The SAMPLINGUNIT statement names one or more *variables* that identify the sampling units as groups of observations (clusters). The combinations of categories of SAMPLINGUNIT variables define the sampling units. If there is a STRATA statement, sampling units are nested within strata.

When you use a SAMPLINGUNIT statement to define units (clusters), PROC SURVEYSELECT selects a sample of these units by using the selection method and design parameters that you specify in the PROC SURVEYSELECT statement. If you do not use a SAMPLINGUNIT statement, then PROC SURVEYSELECT uses the input data set observations as sampling units by default.

The SAMPLINGUNIT 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 SAMPLINGUNIT variables determine the SAMPLINGUNIT variable levels. Thus, you can use formats to group values into levels. For more information, see the FORMAT procedure in the *Base SAS Procedures Guide* and the FORMAT statement and SAS formats in *SAS Formats and Informats: Reference*.

The SAMPLINGUNIT statement is available with all selection methods except Poisson sampling (METHOD=POISSON) and sequential Poisson sampling (METHOD=SEQ_POISSON).

If you specify the PPS option in the SAMPLINGUNIT statement and do not specify a SIZE statement, the procedure computes sampling unit size as the number of observations in the sampling unit. If you specify a SIZE statement and a SAMPLINGUNIT statement, the procedure computes sampling unit size by summing the size measures of all observations in the sampling unit.

By default, PROC SURVEYSELECT sorts the input data set by the SAMPLINGUNIT variables within strata before sample selection. This groups the observations into sampling units and orders the sampling units by the SAMPLINGUNIT variables. If you do not want the procedure to sort the input data set by the SAMPLINGUNIT variables, then specify the PRESORTED option in the SAMPLINGUNIT statement. By using the PRESORTED option, you can provide the order of the sampling units for systematic and sequential selection methods. The CONTROL statement is not available with the SAMPLINGUNIT statement.

The SAMPLINGUNIT statement defines groups of observations (clusters) to use as sampling units, and PROC SURVEYSELECT selects a sample of these units. When you use a SAMPLINGUNIT statement, PROC SURVEYSELECT does not select samples of observations from within the sampling units (clusters). To select independent samples within groups, use the STRATA statement.

You can specify the following *options*:

PPS

computes a sampling unit's size measure as the number of observations in the sampling unit. The procedure then uses these size measures to select a sample according to the PPS selection method that you specify in the METHOD= option in the PROC SURVEYSELECT statement.

This option has no effect when you specify a SIZE statement. When you specify a SIZE statement, the procedure computes sampling unit size by summing the size measures of all observations that belong to the sampling unit.

PRESORTED

requests that PROC SURVEYSELECT not sort the input data set by the SAMPLINGUNIT variables within strata. By default, the procedure sorts the input data set by the SAMPLINGUNIT variables, which groups the observations into sampling units and orders the units by the SAMPLINGUNIT variables.

The PRESORTED option enables you to provide the order of the sampling units. For systematic and sequential selection methods, ordering provides additional control over the distribution of the sample and gives some benefits of proportionate stratification. Systematic and sequential methods include METHOD=SYS, METHOD=PPS_SYS, METHOD=SEQ, and METHOD=PPS_SEQ. For more information, see the descriptions of these methods in the section "Sample Selection Methods" on page 10222.

When you specify the PRESORTED option, the procedure treats the sampling unit groups as NOTSORTED. Like the BY statement option NOTSORTED, this does not mean that the data are unsorted by the SAMPLINGUNIT variables, but rather that the data are arranged in groups (according to values of the SAMPLINGUNIT variables) and that these groups are not necessarily in alphabetical or increasing numeric order. For more information about the BY statement NOTSORTED option, see *SAS Programmers Guide: Essentials*.

SIZE Statement

SIZE *variable* **;**

The SIZE statement names one and only one *variable* that contains size measures that are used for PPS selection. The SIZE variable must be numeric.

If you specify a SAMPLINGUNIT statement together with a SIZE statement, the procedure computes a sampling unit's size by summing the size measures of all observations that belong to the sampling unit. Alternatively, if you specify the PPS option in the SAMPLINGUNIT statement and do not specify a SIZE statement, the procedure computes sampling unit size as the number of observations in the sampling unit.

When the value of a sampling unit's size measure is missing or nonpositive, that sampling unit is excluded from the sample selection. For more information, see the section "Missing Values" on page 10219.

You can adjust the size measure values by using the MAXSIZE= option, the MINSIZE= option, or both of these options in the PROC SURVEYSELECT statement.

All PPS selection methods require size measures, which you can provide by specifying a SIZE statement (or by specifying the PPS option in the SAMPLINGUNIT statement). PPS selection methods include the following: METHOD=PPS, METHOD=PPS_BREWER, METHOD=PPS_MURTHY, METHOD=PPS_SAMPFORD, METHOD=PPS_SEQ, METHOD=PPS_SYS, METHOD=PPS_WR, and METHOD=SEQ_POISSON. For information about how size measures are used in sample selection, see the descriptions of PPS selection methods in the section "Sample Selection Methods" on page 10222.

A sampling unit's size measure, which you provide for PPS selection by specifying a SIZE statement, is not the same as the *sample size*. The sample size is the number of units to select for the sample; you specify the sample size in the SAMPSIZE= option in the PROC SURVEYSELECT statement.

For METHOD=POISSON, the variable that you specify in the SIZE statement provides inclusion probabilities for Poisson sampling. For more information, see the section "Poisson Sampling" on page 10226. When the value of the SIZE variable is missing, nonpositive, or greater than 1, the sampling unit is not included in the sample selection.

STRATA Statement

STRATA *variables < / options >* **;**

You can specify a STRATA statement to obtain stratified sampling. The STRATA statement names one or more *variables* that partition the input data set into nonoverlapping groups (strata). The combinations of levels of the STRATA variables define the strata. PROC SURVEYSELECT independently selects samples from the strata according to the selection method and design parameters that you specify in the PROC SURVEYSELECT statement. For information about stratification in sample design, see Lohr (2010), Kalton (1983), Kish (1965), Kish (1987), and Cochran (1977).

The STRATA variables are one or more variables in the DATA= input data set. These variables can be either character or numeric, but PROC SURVEYSELECT treats them as categorical variables. The formatted values of the STRATA variables determine the STRATA variable levels. Thus, you can use formats to group values into levels. For more information, see the FORMAT procedure in the *Base SAS Procedures Guide* and the FORMAT statement and SAS formats in *SAS Formats and Informats: Reference*.

The STRATA variables function much like BY variables, and PROC SURVEYSELECT expects the input data set to be sorted by the STRATA variables. The BY statement options DESCENDING and NOTSORTED are available in the STRATA statement. For more information about these BY statement options, see *SAS Programmers Guide: Essentials*.

If you specify a CONTROL statement or METHOD=PPS in the PROC SURVEYSELECT statement, the input data set must be sorted by the STRATA variables in ascending order. In this case, you cannot specify the NOTSORTED or DESCENDING option in the STRATA statement.

If your input data set is not sorted by the STRATA variables, use one of the following alternatives:

- Sort the data by using the SORT procedure with the STRATA variables in a BY statement.
- Specify the NOTSORTED or DESCENDING option in the STRATA statement (if you do not specify a CONTROL statement or METHOD=PPS in the PROC SURVEYSELECT 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 STRATA variables) and that these groups are not necessarily in alphabetical or increasing numeric order.
- Create an index on the STRATA variables by using the DATASETS procedure (in Base SAS software).

For more information about BY-group processing, see the discussion in *SAS Programmers Guide: Essentials*. For more information about the DATASETS procedure, see the discussion in the *Base SAS Procedures Guide*.

Table 123.2 summarizes the options available in the STRATA statement. Descriptions of the options follow in alphabetical order.

Table 123.2 STRATA Statement Options for Sample Allocation

Option	Description
ALLOC=*name*	Specifies the allocation method
ALLOC=(*values*)	Provides allocation proportions
ALLOCMIN=	Specifies the minimum sample size per stratum
ALPHA=	Specifies the confidence level for the MARGIN= option
COST=	Provides stratum costs
MARGIN=	Specifies the margin of error
NOSAMPLE	Allocates but does not select the sample
STATS	Displays additional allocation statistics
VAR=	Provides stratum variances

You can specify the following *options*:

ALLOC=*name* **|** **(***values***)|** *SAS-data-set*

specifies the allocation method *name* or specifies the stratum allocation proportions as a list of *values* or a *SAS-data-set*. You can use the ALLOC= option with any selection method (which you specify in the PROC SURVEYSELECT statement) except METHOD=PPS_BREWER and METHOD=PPS_MURTHY, either of which selects two units from each stratum.

You can specify the sample size allocation by using one of the following forms:

ALLOC=*name*

specifies the method for allocating the total sample size among the strata. You can specify one of the following values for *name*:

NEYMAN

requests Neyman allocation, which allocates the total sample size among the strata in proportion to the stratum sizes and variances. For more information, see the section "Neyman Allocation" on page 10234. If you specify ALLOC=NEYMAN, you must provide the stratum variances by also specifying the VAR= option.

OPTIMAL

OPT

requests optimal allocation, which allocates the total sample size among the strata in proportion to the stratum sizes, stratum variances, and stratum costs. For more information, see the section "Optimal Allocation" on page 10233. If you specify ALLOC=OPTIMAL, you must provide the stratum variances by also specifying the VAR= option, and you must provide the stratum costs by also specifying the COST= option.

PROPORTIONAL

PROP

requests proportional allocation, which allocates the total sample size in proportion to the stratum sizes, where stratum size is the number of sampling units in the stratum. For more information, see the section "Proportional Allocation" on page 10233.

ALLOC=(*values*)

specifies a list of stratum allocation proportion *values*. You can separate the values with blanks or commas, and you must enclose the list of values in parentheses. Each value should correspond to a stratum group, and the number of values must equal the number of strata in the input data set.

A stratum allocation proportion specifies the proportion of the total sample size to allocate to the stratum. The sum of the allocation proportions must be 1 or 100%.

The allocation proportions must be positive numbers. You can specify the proportion values as numbers between 0 and 1. Or you can specify the values in percentage form (as numbers between 1 and 100), and PROC SURVEYSELECT converts the numbers to proportions. PROC SURVEYSELECT treats the value 1 as 100% instead of 1%.

The order of the stratum allocation proportions must match the order of the stratum groups in the DATA= input data set. When you specify a list of proportion values, the input data set must be sorted by the STRATA variables in ascending order; you cannot use the DESCENDING or NOTSORTED option in the STRATA statement.

ALLOC=*SAS-data-set*

names a *SAS-data-set* that contains stratum allocation proportions. You should provide the stratum allocation proportions in the data set variable named _ALLOC_. Each observation in the data set should correspond to a stratum group, which is determined by the values of the STRATA variables.

A stratum allocation proportion specifies the proportion of the total sample size to allocate to the corresponding stratum. The sum of the allocation proportions must be 1 or 100%.

The allocation proportions must be positive numbers. You can specify the proportion values as numbers between 0 and 1. Or you can specify the values in percentage form (as numbers between 1 and 100), and PROC SURVEYSELECT converts the numbers to proportions. PROC SURVEYSELECT treats the value 1 as 100% instead of 1%.

The ALLOC= data set, which is a secondary input data set, must contain all stratification variables that you specify in the STRATA statement. The data set must also contain all stratum groups that appear in the DATA= input data set. The order of the stratum groups in the ALLOC= data set must match the order of the groups in the DATA= data set. If formats are associated with the STRATA variables, the formats must be consistent between the two data sets. For more information, see the section "Secondary Input Data Set" on page 10236. You can name only one secondary data set in each invocation of PROC SURVEYSELECT.

ALLOCMIN=*n*

specifies the minimum sample size to allocate to a stratum. If you specify ALLOCMIN=*n*, PROC SURVEYSELECT allocates at least *n* sampling units to each stratum.

The minimum stratum sample size *n* must be a positive integer. The value of *n* times the number of strata must not exceed the total sample size to be allocated. For without-replacement selection methods, the value of *n* must not exceed the number of sampling units in any stratum.

By default, PROC SURVEYSELECT allocates at least one sampling unit to each stratum.

ALPHA=α

specifies the confidence level that PROC SURVEYSELECT uses in the MARGIN= computations. For more information, see the section "Specifying the Margin of Error" on page 10235.

The value of α must be between 0 and 1; a confidence level of α produces a $100(1-\alpha)\%$ confidence interval. By default, ALPHA=0.05, which produces a 95% confidence interval.

COST < =*values* | *SAS-data-set* >

specifies the stratum-level costs that PROC SURVEYSELECT uses to compute optimal allocation when you specify ALLOC=OPTIMAL. For more information, see the section "Optimal Allocation" on page 10233. The stratum costs must be positive numbers. A stratum cost represents the per-unit cost, which is the survey cost of a single unit in the stratum.

You can provide stratum costs by specifying one of the following forms:

COST

indicates that stratum costs are provided in a secondary input data set that you name in another option (for example, the VAR=*SAS-data-set* option). You should provide the stratum costs in the data set variable named _COST_. For more information, see the section "Secondary Input Data Set" on page 10236. You can name only one secondary input data set in each invocation of PROC SURVEYSELECT.

COST=(*values*)

specifies a list of stratum cost *values*. You can separate the values with blanks or commas, and you must enclose the list of values in parentheses. Each value should correspond to a stratum group, and the number of values must equal the number of strata in the input data set.

The order of the stratum cost values must match the order of the stratum groups in the DATA= input data set. When you specify a list of values, the input data set must be sorted by the STRATA variables in ascending order; you cannot use the DESCENDING or NOTSORTED option in the STRATA statement.

COST=*SAS-data-set*

names a *SAS-data-set* that contains the stratum costs. You should provide the stratum costs in the data set variable named _COST_. Each observation in the data set should correspond to a stratum group, which is determined by the values of the STRATA variables.

This data set, which is a secondary data set, must contain all stratification variables that you specify in the STRATA statement. The data set must also contain all stratum groups that appear in the DATA= input data set. The order of the stratum groups in the COST= data set must match the order of the groups in the DATA= data set. If formats are associated with the STRATA variables, the formats must be consistent in the two data sets. For more information, see the section “Secondary Input Data Set” on page 10236. You can name only one secondary input data set in each invocation of PROC SURVEYSELECT.

MARGIN=*value*

specifies the margin of error for the estimate of the overall mean from the stratified sample. When you specify this option, PROC SURVEYSELECT determines the stratum sample sizes that achieve the margin *value* by using the allocation method or proportions that you specify in the ALLOC= option. For more information, see the section “Specifying the Margin of Error” on page 10235.

The *value* must be a positive number. When you specify this option, you must also provide the stratum variances in the VAR= option.

You can use the ALPHA= option to specify the confidence level for the MARGIN= computations. By default, ALPHA=0.05, which produces a 95% confidence interval.

You can specify the MARGIN= option with any allocation method (proportional, optimal, or Neyman) or with allocation proportions that you provide (ALLOC=(*values*) or ALLOC=*SAS-data-set*).

Allocation to achieve a specified margin is an alternative approach to the allocation of a specified total sample size. Therefore, when you specify the MARGIN= option, you cannot also specify a total sample size in the SAMPSIZE= option in the PROC SURVEYSELECT statement.

NOSAMPLE

requests that PROC SURVEYSELECT not select a sample after computing the allocation. When you specify this option, the OUT= output data set contains the stratum sample sizes that PROC SURVEYSELECT computes. For more information, see the section “Allocation Output Data Set” on page 10241. (By default, PROC SURVEYSELECT selects a sample after computing the allocation.)

STATS

displays sample allocation statistics. When you specify the MARGIN= option, the STATS option displays the expected margin of error for the allocation. For more information, see the section “Specifying the Margin of Error” on page 10235. When you specify ALLOC=OPTIMAL or ALLOC=NEYMAN but do not specify the MARGIN= option, the STATS option displays the expected variance, which is computed from the stratum variances that you provide and the allocated stratum sample sizes. When you specify ALLOC=OPTIMAL, the STATS option also displays the total stratum-level cost, which is computed from the stratum costs that you provide and the allocated stratum sample sizes.

VAR <=*values* **|** *SAS-data-set* **>**

specifies the stratum variances that PROC SURVEYSELECT uses to compute optimal allocation (ALLOC=OPTIMAL), Neyman allocation (ALLOC=NEYMAN), or allocation for a specified margin (MARGIN=). The stratum variances must be positive numbers.

You can provide stratum variances by specifying one of the following forms:

VAR

indicates that stratum variances are provided in a secondary input data set that you name in another option (for example, the COST=*SAS-data-set* option). You should provide the stratum variances in the data set variable named _VAR_. For more information, see the section “Secondary Input Data Set” on page 10236. You can name only one secondary input data set in each invocation of PROC SURVEYSELECT.

VAR=(*values*)

specifies a list of stratum variance *values*. You can separate the values with blanks or commas, and you must enclose the list of values in parentheses. Each value should correspond to a stratum group, and the number of values must equal the number of strata in the input data set.

The order of the stratum variance values must match the order of the stratum groups in the DATA= input data set. When you specify a list of values, the input data set must be sorted by the STRATA variables in ascending order; you cannot use the DESCENDING or NOTSORTED option in the STRATA statement.

VAR=*SAS-data-set*

names a *SAS-data-set* that contains the stratum variances. You should provide the stratum variances in the data set variable named _VAR_. Each observation in the data set should correspond to a stratum group, which is determined by the values of the STRATA variables.

This data set, which is a secondary data set, must contain all stratification variables that you specify in the STRATA statement. The data set must also contain all stratum groups that appear in the DATA= input data set. The order of the stratum groups in the VAR= data set must match the order of the groups in the DATA= data set. If formats are associated with the STRATA variables, the formats must be consistent in the two data sets. For more information, see the section "Secondary Input Data Set" on page 10236. You can name only one secondary input data set in each invocation of PROC SURVEYSELECT.

Details: SURVEYSELECT Procedure

Missing Values

PROC SURVEYSELECT treats missing values of STRATA and SAMPLINGUNIT variables like any other STRATA or SAMPLINGUNIT variable value. The missing values form a separate, valid variable level.

When you use a FREQ statement for sample size allocation, all values of the frequency variable must be nonmissing. If there is a missing or nonpositive frequency, PROC SURVEYSELECT does not perform the allocation.

When you specify a SIZE variable, any sampling units that have missing or nonpositive size measures are excluded from the sample selection. The procedure provides a log note that reports the number of observations omitted because of missing or nonpositive size measures.

If you do not use a SAMPLINGUNIT statement with the SIZE statement, your sampling units are input data set observations, and observations that have missing or nonpositive size measures are excluded from the sample selection. If you do use a SAMPLINGUNIT statement with the SIZE statement, the procedure computes sampling unit size by summing the size measures of all observations in the unit. When summing the observation size measures, the procedure omits any observations that have missing or nonpositive size measures. If the size of an entire sampling unit is missing or nonpositive, the procedure excludes that unit from the sample selection. When a sampling unit is selected, the output data set includes all observations that belong to the selected unit, regardless of whether an observation's size measure is missing.

If you provide stratum-level design or allocation information in a secondary input data set, the variable values should be nonmissing. For example, if a stratum value of _NSIZE_ (or SampleSize) in the SAMPSIZE= secondary input data set is missing or negative, PROC SURVEYSELECT cannot select a sample from the stratum. The procedure gives an error message and skips the stratum. Similarly, if other secondary data set variables have missing values for a stratum, a sample cannot be selected from the stratum. These variables include _NRATE_, _MINSIZE_, _MAXSIZE_, _CERTSIZE_, and _CERTP_. Additionally, if any of the sample allocation variables in the secondary input data set have missing or nonpositive values, PROC SURVEYSELECT cannot compute the sample allocation. Variables that provide information for allocation include _ALLOC_, _VAR_, and _COST_. For more information, see the section "Secondary Input Data Set" on page 10236.

Sorting by CONTROL Variables

If you specify a CONTROL statement, PROC SURVEYSELECT sorts the input data set by the CONTROL variables before selecting the sample. If you also specify a STRATA statement, the procedure sorts by CONTROL variables within strata. Sorting by CONTROL variables is available for systematic and sequential selection methods, which include METHOD=SYS, METHOD=PPS_SYS, METHOD=SEQ, and METHOD=PPS_SEQ. Sorting provides additional control over the distribution of the sample and gives some benefits of proportionate stratification.

Control sorting is not available when you use a SAMPLINGUNIT statement, which defines groups of observations as units (clusters) for sample selection. See the description of the SAMPLINGUNIT statement for information about ordering clusters before systematic or sequential selection.

When you specify a CONTROL statement, the sorted data set replaces the input data set by default. Alternatively, you can use the OUTSORT= option to name an output data set that contains the sorted input data set.

PROC SURVEYSELECT provides two types of sorting: hierarchic serpentine sorting and nested sorting. By default (or if you specify the SORT=SERP option), the procedure uses serpentine sorting. If you specify the SORT=NEST option, then the procedure sorts by the CONTROL variables according to nested sorting. These two types of sorting are equivalent when there is only one CONTROL variable.

If you request nested sorting, PROC SURVEYSELECT sorts observations in the same order as PROC SORT does for an ascending sort by the CONTROL variables. For more information, see the chapter "The SORT Procedure" in the *Base SAS Procedures Guide*. PROC SURVEYSELECT sorts within strata if you also specify a STRATA statement. The procedure first arranges the input observations in ascending order of the first CONTROL variable. Then within each level of the first control variable, the procedure arranges the observations in ascending order of the second CONTROL variable. This continues for all CONTROL variables that are specified.

In hierarchic serpentine sorting, PROC SURVEYSELECT sorts by the first CONTROL variable in ascending order. Then within the first level of the first CONTROL variable, the procedure sorts by the second CONTROL variable in ascending order. Within the second level of the first CONTROL variable, the procedure sorts by the second CONTROL variable in descending order. Sorting by the second CONTROL variable continues to alternate between ascending and descending sorting throughout all levels of the first CONTROL variable. If there is a third CONTROL variable, the procedure sorts by that variable within levels formed from the first two CONTROL variables, again alternating between ascending and descending sorting. This continues for all CONTROL variables that are specified. This sorting algorithm minimizes the change from one observation to the next with respect to the CONTROL variable values, thus making nearby observations more similar. For more information about serpentine sorting, see Chromy (1979) and Williams and Chromy (1980).

Random Number Generation

The probability sampling methods that PROC SURVEYSELECT provides use random numbers in their selection algorithms, as described in the following sections and in the references cited. PROC SURVEYSELECT uses a uniform random number function to generate streams of pseudorandom numbers from an initial starting point, or *seed*. You can use the SEED= option to specify the initial seed. If you do not specify the SEED= option, PROC SURVEYSELECT uses the time of day from the computer's clock to obtain the initial seed. For information about specifying initial seeds for strata, storing stratum seeds in the output data set, and reproducing samples, see the description of the SEED= option.

By default, PROC SURVEYSELECT uses the Mersenne twister uniform random number generator (Matsumoto and Nishimura 1998) The Mersenne twister generator has a very long period ($2^{19937} - 1$) and good statistical properties. The algorithm is a twisted generalized feedback shift register. This is the same random number generator that the RAND function provides for the uniform distribution. For more information, see *SAS Functions and CALL Routines: Reference*.

If you specify the RANUNI option, PROC SURVEYSELECT uses the RANUNI random number generator. This uniform random number generator is based on the method of Fishman and Moore (1982), which uses a prime modulus multiplicative generator with modulus 2^{31} and multiplier 397,204,094. This is the same uniform random number generator that the RANUNI function provides. For more information, see *SAS Functions and CALL Routines: Reference*.

When you use the RANUNI random number generator for stratified sampling, PROC SURVEYSELECT generates a single pseudorandom number stream across all strata. You can store the stratum initial seeds in the output data set by specifying the OUTSEED option, and you can use the stratum seeds to reproduce stratum samples (separately, apart from the entire sample).

When you use the Mersenne twister random number generator for stratified sampling, PROC SURVEYSELECT generates separate, independent pseudorandom number streams for the strata by default. To use a single Mersenne twister pseudorandom number stream across all strata, you can specify the STRATUMSEED=NONE option. When you specify this option, stratum initial seeds are not available in the output data set.

In releases before SAS/STAT 14.1, PROC SURVEYSELECT uses a different method to initialize the stratum (Mersenne twister) pseudorandom number streams. To reproduce stratified samples that PROC SURVEYSELECT selects by using the Mersenne twister random number generator in releases before SAS/STAT 14.1, you can specify the STRATUMSEED=RESTORE option (along with the same SEED= value, input data set, and selection parameters).

In releases before SAS/STAT 12.1, PROC SURVEYSELECT uses the RANUNI random number generator by default. To reproduce samples that PROC SURVEYSELECT selects in releases before SAS/STAT 12.1, you can specify the RANUNI option along with the same SEED= value (for the same input data set and selection parameters).

Sample Selection Methods

PROC SURVEYSELECT provides a variety of methods for selecting probability-based random samples. With probability sampling, each unit in the survey population has a known, positive probability of selection. This property of probability sampling avoids selection bias and enables you to use statistical theory to make valid inferences from the sample to the survey population. For more information about probability sampling, see Lohr (2010), Kish (1965), Kish (1987), Kalton (1983), and Cochran (1977).

In equal probability sampling, each unit in the sampling frame, or in a stratum, has the same probability of being selected for the sample. PROC SURVEYSELECT provides the following methods that select units with equal probability: simple random sampling, unrestricted random sampling, systematic random sampling, sequential random sampling, and Bernoulli sampling. In simple random sampling, units are selected *without replacement*, which means that a unit cannot be selected more than once. Both systematic and sequential equal probability sampling are also without replacement. In unrestricted random sampling, units are selected *with replacement*, which means that a unit can be selected more than once. In with-replacement sampling, the *number of hits* refers to the number of times a unit is selected.

In probability proportional to size (PPS) sampling, a unit's selection probability is proportional to its size measure. PROC SURVEYSELECT provides the following methods that select units with probability proportional to size (PPS): PPS sampling without replacement, PPS sampling with replacement, PPS systematic sampling, PPS sequential sampling, Brewer's method, Murthy's method, and Sampford's method. PPS sampling is often used in cluster sampling, where you select clusters (or groups of sampling units) of varying size in the first stage of selection. For example, clusters might be schools, hospitals, or geographical areas, and the final sampling units might be students, patients, or citizens. Cluster sampling can provide efficiencies in frame construction and other survey operations. For more information, see Lohr (2010), Kalton (1983), and Kish (1965), in addition to the other references cited in the following sections.

The following sections give detailed descriptions of the sample selection methods available in PROC SURVEYSELECT. In these sections, n_h denotes the sample size (the number of units in the sample) for stratum h, and N_h denotes the population size (number of units in the population) for stratum h, for $h = 1, 2, \ldots, H$. When the sample design is not stratified, n denotes the sample size, and N denotes the population size. For PPS sampling, M_{hi} represents the size measure for unit i in stratum h, $M_{h \cdot}$ is the total of all size measures for the population of stratum h, and $Z_{hi} = M_{hi} / M_{h \cdot}$ is the relative size of unit i in stratum h.

Simple Random Sampling

The method of simple random sampling (METHOD=SRS) selects units with equal probability and without replacement. Each possible sample of n different units out of N has the same probability of being selected. The selection probability for each individual unit is n/N. When you request stratified sampling by using a STRATA statement, PROC SURVEYSELECT selects samples independently within strata. The selection probability for a unit in stratum h is n_h/N_h for stratified simple random sampling.

By default, PROC SURVEYSELECT uses Floyd's ordered hash table algorithm for simple random sampling. This algorithm is fast, efficient, and appropriate for large data sets. For more information, see Bentley and Floyd (1987) and Bentley and Knuth (1986).

If there is not enough memory available for Floyd's algorithm, PROC SURVEYSELECT switches to the sequential algorithm of Fan, Muller, and Rezucha (1962), which requires less memory but might require more time to select the sample. When PROC SURVEYSELECT uses the alternative sequential algorithm, it writes a note to the log. To request the sequential algorithm, even if enough memory is available for Floyd's algorithm, you can specify METHOD=SRS2 in the PROC SURVEYSELECT statement.

Unrestricted Random Sampling

The method of unrestricted random sampling (METHOD=URS) selects units with equal probability and with replacement. Because units are selected with replacement, a unit can be selected for the sample more than once. The expected number of hits (selections) for each unit is n/N when sampling without stratification. For stratified sampling, the expected number of hits for a unit in stratum h is n_h/N_h. The expected number of hits exceeds one when the sample size n is greater than the population size N.

For unrestricted random sampling, by default, the output data set contains a single copy of each unit selected, even when a unit is selected more than once, and the variable NumberHits records the number of hits (selections) for each unit. If you specify the OUTHITS option, the output data set contains m copies of a sampling unit for which NumberHits is m; for example, the output data set contains three copies of a sampling unit that is selected three times (NumberHits is three). For information about the contents of the output data set, see the section "Sample Output Data Set" on page 10237.

Balanced Bootstrap Sampling

Balanced bootstrap sampling (METHOD=BALBOOTSTRAP) selects r bootstrap samples of N units, where N is the total number of sampling units (in the stratum or in the data set) and r is the number of samples (replicates) that you specify. Units are selected with equal probability and with replacement. The bootstrap selection is balanced so that the total number of selections (over all replicate samples) is r selections for each sampling unit. For more information, see Davison, Hinkley, and Schechtman (1986). PROC SURVEYSELECT performs balanced bootstrap selection by using the algorithm of Gleason (1988).

Systematic Random Sampling

Systematic random sampling (METHOD=SYS) selects units at a fixed interval throughout the sampling frame (or stratum) after a random start. If you request stratified sampling by specifying a STRATA statement, PROC SURVEYSELECT independently selects systematic samples from the strata. PROC SURVEYSELECT applies systematic selection to sampling units in the order of their appearance in the input data set, or in their sorted order if you specify a CONTROL statement.

This section describes equal-probability systematic sampling, where each sampling unit in the sampling frame (or stratum) has the same probability of selection. For information about PPS systematic sampling, see the section "PPS Systematic Sampling" on page 10229.

When you specify the sample size in the SAMPSIZE= option, PROC SURVEYSELECT computes the systematic selection interval as the ratio of the total number of sampling units to the sample size (N/n, or N_h/n_h for stratified sampling). The procedure uses a fractional systematic interval to provide the specified sample size exactly. The selection probability for each unit is computed as n/N (or n_h/N_h for stratified sampling).

When you specify the sampling rate in the SAMPRATE= option, PROC SURVEYSELECT computes the systematic selection interval as the inverse of the sampling rate. The selection probability for each unit is the sampling rate.

Instead of specifying the sample size or sampling rate, you can directly specify the systematic interval in the INTERVAL= option. When you specify the interval, PROC SURVEYSELECT computes the selection probability as the inverse of the interval value.

By default, PROC SURVEYSELECT randomly determines a starting value in the selection interval. Optionally, you can specify the starting value in the START= option. The random component of systematic sampling is the random selection of a starting value in the systematic interval. If you use the START= option to provide a purposely chosen (nonrandom) starting value, the resulting systematic selection does not provide a random, probability-based sample.

Systematic sampling controls the distribution of the sample by spreading the selections throughout the sampling frame (or stratum) at equal intervals and thus provides implicit stratification. You can specify a CONTROL statement to order the input data set by CONTROL variables before sample selection. If you also specify a STRATA statement, PROC SURVEYSELECT sorts by the CONTROL variables within strata. If you do not specify a CONTROL statement, PROC SURVEYSELECT applies systematic selection to the observations in the order in which they appear in the input data set.

Sequential Random Sampling

If you specify the METHOD=SEQ option and do not include a SIZE statement, PROC SURVEYSELECT uses the equal probability version of Chromy's method for sequential random sampling. This method selects units sequentially with equal probability and without replacement. For more information, see Chromy (1979) and Williams and Chromy (1980). For information about Chromy's PPS selection method, see the section "PPS Sequential Sampling" on page 10229.

Sequential random sampling controls the distribution of the sample by spreading it throughout the sampling frame or stratum, thus providing implicit stratification according to the order of units in the frame or stratum. You can use the CONTROL statement to sort the input data set by the CONTROL variables before sample selection. If you also use a STRATA statement, PROC SURVEYSELECT sorts by the CONTROL variables within strata. By default (or if you specify the SORT=SERP option), the procedure uses hierarchic serpentine ordering for sorting. If you specify the SORT=NEST option, the procedure uses nested sorting. See the section "Sorting by CONTROL Variables" on page 10220 for descriptions of serpentine and nested sorting. If you do not specify a CONTROL statement, PROC SURVEYSELECT applies sequential selection to the observations in the order in which they appear in the input data set.

Following Chromy's method of sequential selection, PROC SURVEYSELECT randomly chooses a starting unit from the entire stratum (or frame, if the design is not stratified). With this unit as the first one, the procedure treats the stratum units as a closed loop. This is done so that all pairwise (joint) selection probabilities are positive and an unbiased variance estimator can be obtained. The procedure numbers units sequentially from the random start to the end of the stratum and then continues from the beginning of the stratum until all units are numbered.

Beginning with the randomly chosen starting unit, PROC SURVEYSELECT accumulates the expected number of selections (hits), where the expected number of selections $\mathrm{E}(S_{hi})$ is n_h/N_h for all units i in stratum h. The procedure computes

$$I_{hi} = \mathrm{Int}\Big(\sum_{j=1}^{i} \mathrm{E}(S_{hj})\Big) = \mathrm{Int}(i\ n_h/N_h)$$

$$F_{hi} = \mathrm{Frac}\Big(\sum_{j=1}^{i} \mathrm{E}(S_{hj})\Big) = \mathrm{Frac}(i\ n_h/N_h)$$

where $\mathrm{Int}(\cdot)$ denotes the integer part of the number, and $\mathrm{Frac}(\cdot)$ denotes the fractional part.

Considering each unit sequentially, Chromy's method determines whether unit i is selected by comparing the total number of selections for the first (i – 1) units,

$$T_{h(i-1)} = \sum_{j=1}^{i-1} S_{hj}$$

with the value of $I_{h(i-1)}$.

If $T_{h(i-1)} = I_{h(i-1)}$, Chromy's method determines whether or not unit i is selected as follows. If $F_{hi} = 0$ or $F_{h(i-1)} > F_{hi}$, then unit i is selected with certainty. Otherwise, unit i is selected with probability

$$(F_{hi} - F_{h(i-1)})/(1 - F_{h(i-1)})$$

If $T_{h(i-1)} = (I_{h(i-1)} + 1)$, Chromy's method determines whether or not unit i is selected as follows. If $F_{hi} = 0$ or $F_{hi} > F_{h(i-1)}$, then the unit is not selected. Otherwise, unit i is selected with probability

$$F_{hi}/F_{h(i-1)}$$

Bernoulli Sampling

Bernoulli sampling, which you request by specifying the METHOD=BERNOULLI option, is an equal probability selection method for which the total sample size is not fixed. PROC SURVEYSELECT performs an independent random selection trial for each of the N sampling units in the input data set by using the constant inclusion probability (sampling rate) that you specify. You can specify a single value of the inclusion probability π to use for all N sampling units, or you can specify separate stratum-level values of π_h to use for the N_h units in each stratum.

You provide the inclusion probability (or probabilities) by specifying the SAMPRATE= option. For stratified sampling (which you request by using the STRATA statement), you can specify the same sampling rate for each stratum in the SAMPRATE=*value* option. Or you can specify different sampling rates for different strata by using the SAMPRATE=(*values*) or SAMPRATE=*SAS-data-set* option.

In Bernoulli sampling, the sample size n (number of units selected) is not fixed; it is a random variable that has a binomial distribution with parameters N and π. The possible values of n range from 0 to N. The expected value of the sample size is πN (or $\pi_h N_h$ for stratified sampling), and the variance of the sample size is $\pi(1 - \pi)N$.

For Bernoulli sampling, the selection probability is the inclusion probability that you specify in the SAMPRATE= option. PROC SURVEYSELECT computes the sampling weight as the inverse of the selection probability, which is $1/\pi$. For Bernoulli sampling, the procedure also computes an adjusted sampling weight as the ratio of the total number of sampling units to the actual sample size, N/n (or N_h/n_h for stratified sampling). The joint selection probability for any two distinct units is π^2. For more information, see Särndal, Swensson, and Wretman (1992).

You can specify the STATS option to include the following information in the OUT= output data set for METHOD=BERNOULLI: total number of sampling units, selection probability, expected sample size, actual sample size, sampling weight, and adjusted sampling weight.

Poisson Sampling

Poisson sampling, which you request by specifying the METHOD=POISSON option, is an unequal probability sampling method for which the total sample size is not fixed. A generalization of Bernoulli sampling, Poisson sampling also consists of independent random selection trials for the N sampling units in the input data set, but the sampling units can have different inclusion probabilities. You provide inclusion probabilities for Poisson sampling in the variable that you specify in the SIZE statement.

The expected value of the sample size for Poisson sampling is $\sum_i \pi_i$, where π_i is the inclusion probability for sampling unit i. The variance of the sample size is $\sum_i \pi_i(1 - \pi_i)$.

For Poisson sampling, the selection probability for unit i is the inclusion probability π_i that you specify by using the SIZE statement. PROC SURVEYSELECT computes the sampling weight for unit i as the inverse of the selection probability, which is $1/\pi_i$. The joint selection probability for any two distinct units i and j is $\pi_i \pi_j$. For more information, see Särndal, Swensson, and Wretman (1992).

Sequential Poisson Sampling

Sequential Poisson sampling, which you request by specifying the METHOD=SEQ_POISSON option, is a fixed-sample-size modification of Poisson sampling. For information about Poisson sampling, see the section "Poisson Sampling" on page 10226.

PROC SURVEYSELECT performs sequential Poisson sampling by using the method of Ohlsson (1998). A *transformed random number* is computed for each sampling unit as X_{hi}/M_{hi}, where M_{hi} is the size measure of unit i in stratum h and X_{hi} is a uniform random number (from the procedure's pseudorandom number stream). For more information about random number generation, see the SEED= option and the section "Random Number Generation" on page 10221.

The N_h transformed random numbers are ordered, and the stratum h sample consists of the n_h sampling units that correspond to the n_h smallest transformed random numbers.

Although this algorithm produces a sample of the fixed size that you specify, the sample selection is considered to be only approximately probability proportional to size (PPS); it is not strictly PPS. For more information, see Ohlsson (1998). The (approximate) selection probability for unit i in stratum h is computed as $n_h Z_{hi}$, where n_h is the sample size for stratum h and Z_{hi} is the relative size of unit i in stratum h. The relative size is computed as $M_{hi}/M_{h\cdot}$, which is the ratio of the size measure for unit i in stratum h (M_{hi}) to the total of all size measures for stratum h ($M_{h\cdot}$)

The relative size of each sampling unit cannot exceed $1/n_h$ because the selection probability (n_h times the relative size) cannot exceed 1. This requirement can be expressed as $Z_{hi} \le 1/n_h$, or equivalently as $M_{hi} \le M_{h\cdot}/n_h$. If your size measures do not meet this requirement, you can adjust the size measures by using the MAXSIZE= or MINSIZE= option. Or you can select the larger units with certainty by using the CERTSIZE= or CERTSIZE=P= option. Alternatively, you can use a selection method that does not have a relative size restriction, such as PPS with minimum replacement (METHOD=PPS_SEQ).

PPS Sampling without Replacement

When you specify the METHOD=PPS option, PROC SURVEYSELECT selects units with probability proportional to size and without replacement. The selection probability for unit i in stratum h is $n_h Z_{hi}$, where n_h is the sample size for stratum h and Z_{hi} is the relative size of unit i in stratum h. The relative size Z_{hi} is computed as $M_{hi} / M_{h\cdot}$, which is the ratio of the size measure of unit i in stratum h to the total of all size measures in stratum h.

Because selection probabilities cannot exceed 1, the relative size for each unit must not exceed $1/n_h$ for METHOD=PPS. This requirement can be expressed as $Z_{hi} \le 1/n_h$, or equivalently as $M_{hi} \le M_{h\cdot}/n_h$. If your size measures do not meet this requirement, you can adjust the size measures by using the MAXSIZE= or MINSIZE= option. Or you can request certainty selection for the larger units by using the CERTSIZE= or CERTSIZE=P= option. Alternatively, you can use a selection method that does not have this relative size restriction, such as PPS with minimum replacement (METHOD=PPS_SEQ).

PROC SURVEYSELECT performs PPS selection by using the Hanurav-Vijayan algorithm. Hanurav (1967) introduced this algorithm for the selection of two units per stratum, and Vijayan (1968) generalized it for the selection of more than two units. This algorithm enables computation of joint selection probabilities and provides joint selection probability values that usually ensure nonnegativity and stability of the Sen-Yates-Grundy variance estimator. For more information, see Fox (1989), Golmant (1990), and Watts (1991).

The notation in the remainder of this section drops the stratum subscript h for simplicity. If you specify a stratified design, n now denotes the sample size for the current stratum, N denotes the stratum population size, M_i denotes the size measure for unit i in the stratum, and M denotes the total of size measures in the stratum. For a stratified design, PROC SURVEYSELECT selects samples independently within strata by using the same selection method in each stratum.

PROC SURVEYSELECT performs the Hanurav-Vijayan selection algorithm as described by Fox (1989, p. 169). For the definition of $P_k^{(i)}$, see Golmant (1990). The sampling units are first sorted in ascending order by size measure so that $M_1 \le M_2 \le \cdots \le M_N$. The procedure then selects a PPS sample of n units as follows:

1. The procedure randomly chooses one of the integers $1, 2, \ldots, n$ with probability $\theta_1, \theta_2, \ldots, \theta_n$, where

   $$\theta_i = n \, (Z_{N-n+i+1} - Z_{N-n+i}) \, (T + i \, Z_{N-n+1}) \, / \, T$$

   where $Z_j = M_j / M$ and

   $$T = \sum_{j=1}^{N-n} Z_j$$

   By definition, $Z_{N+1} = 1/n$ to ensure that $\sum_{i=1}^{n} \theta_i = 1$.

2. If the integer i is selected in step 1, the procedure includes the last $(n - i)$ units in the sample (where the units are ordered by their size measures). The procedure then selects the remaining i units by following steps 3 through 6.

3. The procedure defines new normed size measures for the remaining $(N - n + i)$ units that were not selected in steps 1 and 2:

   $$Z_j^*(i) = \begin{cases} Z_j \, / \, (T + i \, Z_{N-n+1}) & \text{for } j = 1, \ldots, N-n+1 \\ Z_{N-n+1} \, / \, (T + i \, Z_{N-n+1}) & \text{for } j = N-n+2, \ldots, N-n+i \end{cases}$$

4. The procedure selects the next unit from the first $(N - n + 1)$ units with probability proportional to $a_j(1)$, where

$$\begin{aligned} a_1(1) &= i\, Z_1^*(i) \\ a_j(1) &= i\, Z_j^*(i) \textstyle\prod_{k=1}^{j-1} \left(1 - (i-1)\, P_k^{(i)}\right) \quad \text{for } j = 2, \ldots, N - n + 1 \end{aligned}$$

   and

$$P_k^{(i)} = M_k / (M_{k+1} + M_{k+2} + \cdots + M_{N-n+i})$$

5. Where j_1 denotes the unit that is selected in step 4, the procedure selects the next unit from units $(j_1 + 1)$ through $(N - n + 2)$ with probability proportional to $a_j(2, j_1)$, where

$$a_{j_1+1}(2, j_1) = (i-1)\, Z_{j_1+1}^*(i)$$

$$a_j(2, j_1) = (i-1)\, Z_j^*(i) \prod_{k=j_1+1}^{j-1} \left(1 - (i-2)\, P_k^{(i)}\right) \quad \text{for } j = j_1 + 2, \ldots, N - n + 2$$

6. The procedure repeats step 5 until all n sample units are selected.

If you specify the JTPROBS option, PROC SURVEYSELECT computes the joint selection probabilities for all pairs of selected units in each stratum. The joint selection probability for units i and j is

$$P_{(ij)} = \sum_{r=1}^{n} \theta_r K_{ij}^{(r)}$$

where

$$K_{ij}^{(r)} = \begin{cases} 1 & N - n + r < i \le N - 1 \\ r\, Z_{N-n+1} \,/\, (T + r\, Z_{N-n+1}) & N - n < i \le N - n + r, \quad j > N - n + r \\ r\, Z_i \,/\, (T + r\, Z_{N-n+1}) & 1 \le i \le N - n, \quad j > N - n + r \\ \pi_{ij}^{(r)} & j \le N - n + r \end{cases}$$

$$\pi_{ij}^{(r)} = r(r-1)\, P_i^{(r)}\, Z_j^*(r) \prod_{k=1}^{i-1} (1 - P_k^{(r)})$$

$$P_k^{(r)} = M_k / (M_{k+1} + M_{k+2} + \cdots + M_{N-n+r})$$

PPS Sampling with Replacement

If you specify the METHOD=PPS_WR option, PROC SURVEYSELECT selects units with probability proportional to size and with replacement. The procedure makes n_h independent random selections from the stratum of N_h units, selecting with probability $Z_{hi} = M_{hi} / M_{h\cdot}$. Because units are selected with replacement, a unit can be selected for the sample more than once. The expected number of hits (selections) for unit i in stratum h is $n_h Z_{hi}$. If you specify the JTPROBS option, PROC SURVEYSELECT computes the joint expected number of hits for all pairs of selected units in each stratum. The joint expected number of hits for units i and j in stratum h is

$$P_{h(ij)} = \begin{cases} n_h(n_h - 1) Z_{hi} Z_{hj} & \text{for } j \ne i \\ n_h(n_h - 1) Z_{hi} Z_{hi} / 2 & \text{for } j = i \end{cases}$$

PPS Systematic Sampling

If you specify the METHOD=PPS_SYS option, PROC SURVEYSELECT selects the sample by using systematic random sampling with probability proportional to size. Systematic sampling selects units at a fixed interval throughout the sampling frame (or stratum) after a random start. If you request stratified sampling by specifying a STRATA statement, PROC SURVEYSELECT independently selects systematic samples from the strata. PROC SURVEYSELECT applies systematic selection to sampling units in the order of their appearance in the input data set, or in their sorted order if you specify a CONTROL statement.

When you specify the sample size in the SAMPSIZE= option, PROC SURVEYSELECT computes the systematic selection interval as the ratio of the total size to the sample size (M/n, or $M_{h\cdot}/n_h$ for stratified sampling). The procedure uses a fractional systematic interval to provide the specified sample size exactly. Depending on the sample size and the values of the size measures, it might be possible for a sampling unit to be selected more than once. The expected number of hits (selections) for unit i in stratum h is computed as $n_h M_{hi} / M_{h\cdot} = n_h Z_{hi}$. For more information, see Cochran (1977, pp. 265–266) and Madow (1949).

Instead of specifying the sample size for systematic sampling, you can directly specify the systematic interval in the INTERVAL= option. When you specify the interval, PROC SURVEYSELECT computes the expected number of hits as the inverse of the interval value.

By default, PROC SURVEYSELECT randomly determines a starting value in the selection interval. Optionally, you can specify the starting value in the START= option. The random component of systematic sampling is the random selection of a starting value in the systematic interval. If you use the START= option to provide a purposely chosen (nonrandom) starting value, the resulting systematic selection does not provide a random, probability-based sample.

Systematic sampling controls the distribution of the sample by spreading the selections throughout the sampling frame (or stratum) at equal intervals and thus provides implicit stratification. You can specify a CONTROL statement to order the input data set by the CONTROL variables before sample selection. If you also specify a STRATA statement, PROC SURVEYSELECT sorts by the CONTROL variables within strata. If you do not specify a CONTROL statement, PROC SURVEYSELECT applies systematic selection to the observations in the order in which they appear in the input data set.

PPS Sequential Sampling

If you specify the METHOD=PPS_SEQ option, PROC SURVEYSELECT uses Chromy's method of sequential random sampling. For more information, see Chromy (1979) and Williams and Chromy (1980). Chromy's method selects units sequentially with probability proportional to size and with minimum replacement. Selection *with minimum replacement* means that the actual number of hits for a unit can equal the integer part of the expected number of hits for that unit, or the next largest integer. This can be compared to selection *without replacement*, where each unit can be selected only once, so the number of hits can equal 0 or 1. The other alternative is selection *with replacement*, where there is no restriction on the number of hits for each unit, so the number of hits can equal $0, 1, \ldots, n_h$, where n_h is the stratum sample size.

Sequential random sampling controls the distribution of the sample by spreading it throughout the sampling frame or stratum, thus providing implicit stratification according to the order of units in the frame or stratum. You can use the CONTROL statement to sort the input data set by the CONTROL variables before sample selection. If you also use a STRATA statement, PROC SURVEYSELECT sorts by the CONTROL variables within strata. By default (or if you specify the SORT=SERP option), the procedure uses hierarchic serpentine ordering to sort the sampling frame by the CONTROL variables within strata. If you specify the SORT=NEST option, the procedure uses nested sorting. See the section "Sorting by CONTROL Variables" on page 10220 for descriptions of serpentine and nested sorting. If you do not specify a CONTROL statement, PROC SURVEYSELECT applies sequential selection to the observations in the order in which they appear in the input data set.

According to Chromy's method of sequential selection, PROC SURVEYSELECT first chooses a starting unit randomly from the entire stratum, with probability proportional to size. The procedure uses this unit as the first one and treats the stratum observations as a closed loop. This is done so that all pairwise (joint) expected number of hits are positive and an unbiased variance estimator can be obtained. The procedure numbers observations sequentially from the random start to the end of the stratum and then continues from the beginning of the stratum until all units are numbered.

Beginning with the randomly chosen starting unit, Chromy's method partitions the ordered stratum sampling frame into n_h zones of equal size. There is one selection from each zone and a total of n_h hits (selections), although fewer than n_h distinct units might be selected. Beginning with the random start, the procedure accumulates the expected number of hits and computes

$$\mathrm{E}(S_{hi}) = n_h Z_{hi}$$

$$I_{hi} = \mathrm{Int}\Big(\sum_{j=1}^{i} \mathrm{E}(S_{hj}) \Big)$$

$$F_{hi} = \mathrm{Frac}\Big(\sum_{j=1}^{i} \mathrm{E}(S_{hj}) \Big)$$

where $\mathrm{E}(S_{hi})$ represents the expected number of hits for unit *i* in stratum *h*, $\mathrm{Int}(\cdot)$ denotes the integer part of the number, and $\mathrm{Frac}(\cdot)$ denotes the fractional part.

Considering each unit sequentially, Chromy's method determines the actual number of hits for unit *i* by comparing the total number of hits for the first (*i* – 1) units,

$$T_{h(i-1)} = \sum_{j=1}^{i-1} S_{hj}$$

with the value of $I_{h(i-1)}$.

If $T_{h(i-1)} = I_{h(i-1)}$, Chromy's method determines the total number of hits for the first *i* units as follows. If $F_{hi} = 0$ or $F_{h(i-1)} > F_{hi}$, then $T_{hi} = I_{hi}$. Otherwise, $T_{hi} = I_{hi} + 1$ with probability

$$(F_{hi} - F_{h(i-1)})/(1 - F_{h(i-1)})$$

And the number of hits for unit *i* is $T_{hi} - T_{h(i-1)}$.

If $T_{h(i-1)} = (I_{h(i-1)} + 1)$, Chromy's method determines the total number of hits for the first *i* units as follows. If $F_{hi} = 0$, then $T_{hi} = I_{hi}$. If $F_{hi} > F_{h(i-1)}$, then $T_{hi} = I_{hi} + 1$. Otherwise, $T_{hi} = I_{hi} + 1$ with probability

$$F_{hi} / F_{h(i-1)}$$

Brewer's PPS Method

Brewer's method (METHOD=PPS_BREWER) selects two units from each stratum, with probability proportional to size and without replacement. The selection probability for unit i in stratum h is $2M_{hi}/M_{h\cdot} = 2Z_{hi}$. (Because selection probabilities cannot exceed 1, the relative size for each unit, Z_{hi}, must not exceed $1/2$.)

Brewer's algorithm first selects a unit with probability

$$\frac{Z_{hi}(1 - Z_{hi})}{D_h(1 - 2Z_{hi})}$$

where

$$D_h = \sum_{i=1}^{N_h} \frac{Z_{hi}(1 - Z_{hi})}{1 - 2Z_{hi}}$$

Then a second unit is selected from the remaining units with probability

$$\frac{Z_{hj}}{1 - Z_{hi}}$$

where unit i is the first unit selected. The joint selection probability for units i and j in stratum h is

$$P_{h(ij)} = \frac{2Z_{hi}Z_{hj}}{D_h}\left(\frac{1 - Z_{hi} - Z_{hj}}{(1 - 2Z_{hi})(1 - 2Z_{hj})}\right)$$

For more information, see Cochran (1977, pp. 261–263) and Brewer (1963). Brewer's method yields the same selection probabilities and joint selection probabilities as Durbin's method (Cochran 1977; Durbin 1967).

Murthy's PPS Method

Murthy's method (METHOD=PPS_MURTHY) selects two units from each stratum, with probability proportional to size and without replacement. The selection probability for unit i in stratum h is

$$P_{hi} = Z_{hi}\left(1 + K_h - (Z_{hi}/(1 - Z_{hi})\right)$$

where $Z_{hi} = M_{hi}/M_{h\cdot}$ and

$$K_h = \sum_{j=1}^{N_h}\left(Z_{hj}/(1 - Z_{hj})\right)$$

Murthy's algorithm first selects a unit with probability Z_{hi}. Then a second unit is selected from the remaining units with probability $Z_{hj}/(1 - Z_{hi})$, where unit i is the first unit selected. The joint selection probability for units i and j in stratum h is

$$P_{h(ij)} = Z_{hi}Z_{hj}\left(\frac{2 - Z_{hi} - Z_{hj}}{(1 - Z_{hi})(1 - Z_{hj})}\right)$$

For more information, see Cochran (1977, pp. 263–265) and Murthy (1957).

Sampford's PPS Method

Sampford's method (METHOD=PPS_SAMPFORD) is an extension of Brewer's method that selects more than two units from each stratum, with probability proportional to size and without replacement. The selection probability for unit i in stratum h is $n_h M_{hi} / M_{h\cdot} = n_h Z_{hi}$. (Because selection probabilities cannot exceed 1, the relative size for each unit, Z_{hi}, must not exceed $1/n_h$.)

Sampford's method first selects a unit from stratum h with probability Z_{hi}. Then subsequent units are selected with probability proportional to

$$\lambda_{hi} = Z_{hi} \;/\; (1 - n_h Z_{hi})$$

and with replacement. If the same unit appears more than once in the sample of size n_h, then Sampford's algorithm rejects that sample and selects a new sample. The sample is accepted if it contains n_h distinct units.

If you specify the JTPROBS option, PROC SURVEYSELECT computes the joint selection probabilities for all pairs of selected units in each stratum. The joint selection probability for units i and j in stratum h is

$$P_{h(ij)} = K_h \, \lambda_{hi} \, \lambda_{hj} \sum_{t=2}^{n_h} \Big(\left[t - n_h (Z_{hi} + Z_{hj}) \right] \, L_{h,(n_h - t)}(\overline{ij}) \Big) \;/\; n_h^{t-2}$$

where

$$K_h = 1 \;/\; \sum_{t=1}^{n_h} \left(t \; L_{h,(n_h - t)} \;/\; n_h^t \right)$$

$$L_{h,m} = \sum_{S_h(m)} \lambda_{hi_1} \, \lambda_{hi_2} \cdots \lambda_{hi_m}$$

and $S_h(m)$ denotes all possible samples of size m, for $m = 1, 2, \ldots, N_h$. The sum $L_{h,m}(\overline{ij})$ is defined similarly to $L_{h,m}$ but sums over all possible samples of size m that do not include units i and j. For more information, see Cochran (1977, pp. 262–263) and Sampford (1967).

Sample Size Allocation

If you specify the ALLOC= option in the STRATA statement, PROC SURVEYSELECT allocates the total sample size among the strata according to the method that you request. PROC SURVEYSELECT provides proportional allocation (ALLOC=PROPORTIONAL), optimal allocation (ALLOC=OPTIMAL), and Neyman allocation (ALLOC=NEYMAN). For more information about these allocation methods, see Lohr (2010), Kish (1965), and Cochran (1977). You can also directly provide the allocation proportions by using the ALLOC=(*values*) option or the ALLOC=*SAS-data-set* option. Then PROC SURVEYSELECT allocates the sample size among the strata according to the proportions that you provide. Allocation proportions are the relative stratum sample sizes, n_h/n, where n_h is the sample size for stratum h and n is the total sample size.

You can use the SAMPSIZE=*n* option in the PROC SURVEYSELECT statement to specify the total sample size to allocate among the strata. Or you can specify the required margin of error in the MARGIN= option in the STRATA statement, and PROC SURVEYSELECT computes the stratum sample sizes necessary to achieve that margin of error for the allocation method that you request. For more information, see the section "Specifying the Margin of Error" on page 10235.

Proportional Allocation

When you specify the ALLOC=PROPORTIONAL option in the STRATA statement, PROC SURVEYSELECT allocates the total sample size among the strata in proportion to the stratum sizes, where the stratum size is the number of sampling units in the stratum. The allocation proportion of the total sample size for stratum h is

$$f_h^* = N_h / N$$

where N_h is the number of sampling units in stratum h and N is the total number of sampling units for all strata. PROC SURVEYSELECT computes the target sample size for stratum h as

$$n_h^* = f_h^* \times n$$

where n is the total sample size that you specify in the SAMPSIZE= option in the PROC SURVEYSELECT statement.

If you specify a minimum stratum sample size n_{min} in the ALLOCMIN= option in the STRATA statement, then all stratum sample sizes are required to be at least n_{min}. By default, all stratum sample sizes are required to be at least 1 (to ensure that at least one sampling unit is selected from each stratum). If a target sample size is less than the required minimum value, PROC SURVEYSELECT sets the target sample size equal to the minimum value.

PROC SURVEYSELECT computes the allocated stratum sample sizes n_h (which must be integers) by rounding the target sample size values in order of the fractional parts until the total sample size n is achieved.

PROC SURVEYSELECT provides the target allocation proportions f_h^* in the output data set variable AllocProportion. The variable ActualProportion contains the actual proportions for the allocated sample sizes n_h. For stratum h, the actual proportion is computed as

$$f_h = n_h / n$$

where n_h is the allocated sample size for stratum h and n is the total sample size. The actual proportions f_h can differ from the target allocation proportions f_h^* because of rounding and the requirement that $n_h \geq 1$ (or $n_h \geq n_{min}$).

Optimal Allocation

When you specify the ALLOC=OPTIMAL option in the STRATA statement, PROC SURVEYSELECT allocates the total sample size among the strata in proportion to stratum sizes, stratum costs, and stratum variances. You provide the stratum costs and variances in the COST= and VAR= options, respectively.

Optimal allocation minimizes the overall variance for a specified cost, or equivalently minimizes the overall cost for a specified variance. For more information, see Lohr (2010), Cochran (1977), and Kish (1965). For optimal allocation, PROC SURVEYSELECT computes the proportion of the total sample size for stratum h as

$$f_h^* = \frac{N_h S_h}{\sqrt{C_h}} \Big/ \sum_{i=1}^{H} \frac{N_i S_i}{\sqrt{C_i}}$$

where N_h is the number of sampling units in stratum h, S_h is the standard deviation within stratum h, C_h is the unit cost within stratum h, and H is the total number of strata.

PROC SURVEYSELECT computes the target sample size for stratum h as

$$n_h^* = f_h^* \times n$$

where n is the total sample size that you specify in the SAMPSIZE= option in the PROC SURVEYSELECT statement.

If you specify a minimum stratum sample size n_{min} in the ALLOCMIN= option in the STRATA statement, then all stratum sample sizes are required to be at least n_{min}. By default, all stratum sample sizes are required to be at least 1 (to ensure that at least one sampling unit is selected from each stratum). If a target sample size is less than the required minimum value, PROC SURVEYSELECT sets the target sample size equal to the minimum value.

For without-replacement selection methods, a stratum sample size cannot exceed the number of sampling units in the stratum (N_h). If a target stratum sample size exceeds the number of units in the stratum, PROC SURVEYSELECT allocates the number of available units (N_h) to the stratum and allocates the remaining sample size proportionally among the remaining strata.

PROC SURVEYSELECT computes the allocated stratum sample sizes n_h (which must be integers) by rounding the target sample size values in order of the fractional parts until the total sample size n is achieved.

PROC SURVEYSELECT provides the target allocation proportions f_h^* in the output data set variable AllocProportion. The variable ActualProportion contains the actual proportions for the allocated sample sizes n_h. For stratum h, the actual proportion is computed as

$$f_h = n_h / n$$

where n_h is the allocated sample size for stratum h and n is the total sample size. The actual proportions f_h can differ from the target allocation proportions f_h^* because of rounding and the requirement that $n_h \geq 1$ (or $n_h \geq n_{min}$).

Neyman Allocation

When you specify the ALLOC=NEYMAN option in the STRATA statement, PROC SURVEYSELECT allocates the total sample size among the strata in proportion to stratum sizes and stratum variances. Neyman allocation is a special case of optimal allocation (which is described in the section "Optimal Allocation" on page 10233), where the costs per unit are the same for all strata. For Neyman allocation, PROC SURVEYSELECT computes the proportion of the total sample size for stratum h as

$$f_h^* = N_h S_h \; / \; \sum_{i=1}^{H} N_i S_i$$

where N_h is the number of sampling units in stratum h, S_h is the standard deviation within stratum h, and H is the total number of strata.

PROC SURVEYSELECT computes the target sample size for stratum h as

$$n_h^* = f_h^* \times n$$

where n is the total sample size that you specify in the SAMPSIZE= option in the PROC SURVEYSELECT statement.

The allocated sample sizes n_h are computed from the target sample sizes as described in the section "Optimal Allocation" on page 10233.

Specifying the Margin of Error

Instead of specifying the total sample size to allocate among the strata, you can specify the margin of error for the estimate of the overall mean from the stratified sample. Based on the requested allocation method and the stratum variances that you provide, PROC SURVEYSELECT computes the stratum sample sizes that are required to achieve this margin of error. You specify the margin of error in the MARGIN= option in the STRATA statement, and you provide stratum variances in the VAR= option. You can use the MARGIN= option with any allocation method (proportional, optimal, or Neyman) or with allocation proportions that you provide (ALLOC=(*values*) or ALLOC=*SAS-data-set*).

The margin of error e is the half-width of the $100(1-\alpha)\%$ confidence interval for the overall mean based on the stratified sample,

$$e = z_{\alpha/2} \sqrt{\mathrm{Var}(\bar{y}_{str})}$$

where $\mathrm{Var}(\bar{y}_{str})$ is the variance of the estimate of the mean from the stratified sample and $z_{\alpha/2}$ is the $100(1-\alpha/2)$th percentile of the standard normal distribution. You can specify the value of α in the ALPHA= option in the STRATA statement. By default, PROC SURVEYSELECT uses a 95% confidence interval (ALPHA=0.05).

For the specified margin of error e, PROC SURVEYSELECT computes the target stratum sample sizes n_h^* for without-replacement selection methods as

$$n_h^* = f_h^* \left(\sum_{i=1}^{H} N_i^2 S_i^2 / f_i^* \right) \Big/ \left((e\, N / z_{\alpha/2})^2 + \sum_{i=1}^{H} N_i S_i^2 \right)$$

where N_i is the number of sampling units in stratum i, S_i^2 is the variance within stratum i, N is the total number of sampling units for all strata, and H is the total number of strata.

The values of f_h^* are the stratum allocation proportions, which PROC SURVEYSELECT computes according to the allocation method that you request. For more information, see the sections “Proportional Allocation” on page 10233, “Optimal Allocation” on page 10233, and “Neyman Allocation” on page 10234.

For with-replacement selection methods, PROC SURVEYSELECT computes the target stratum sample sizes as

$$n_h^* = f_h^* \left(\sum_{i=1}^{H} N_i^2 S_i^2 / f_i^* \right) \Big/ \left(e\, N / z_{\alpha/2} \right)^2$$

For more information, see Lohr (2010, p. 91), Cochran (1977, Chapter 5), and Arkin (1984, Chapter 10).

The target sample size values n_h^* might not be integers, but the stratum sample sizes are required to be integers. PROC SURVEYSELECT rounds all fractional target sample sizes up to integer sample sizes. If you specify a minimum stratum sample size n_{min} in the ALLOCMIN= option in the STRATA statement, then all stratum sample sizes n_h are required to be at least n_{min}.

For without-replacement selection methods, a stratum sample size cannot exceed the number of units in the stratum. If a target stratum sample size does exceed the number of units in the stratum, the procedure sets $n_h = N_h$ for that stratum, removes the stratum from the variance computation (because it contributes nothing to the sampling error), revises the allocation proportions f_h^* for the remaining strata, and computes the stratum sample sizes again. If a stratum sample size equals the number of units in its stratum, the procedure also removes that stratum from the variance computation and revises the sample sizes for the remaining strata. For more information, see Cochran (1977, p. 104) and Arkin (1984, p. 176).

When you specify the STATS option with the MARGIN= option in the STRATA statement, PROC SURVEYSELECT displays the expected margin of error for the sample allocation. The expected margin of error (for the overall mean based on the stratified sample) is computed from the stratum sizes (N_i), the stratum variances that you provide (S_i^2), and the allocated stratum sample sizes that the procedure computes (n_i). For without-replacement selection methods, the expected margin of error is

$$e = (z_{\alpha/2}/N) \sqrt{\sum_{i=1}^{H} (N_i^2 S_i^2 / n_i)\,(1 - n_i/N)}$$

For with-replacement selection methods, the expected margin of error is

$$e = (z_{\alpha/2}/N) \sqrt{\sum_{i=1}^{H} (N_i^2 S_i^2 / n_i)}$$

The expected margin of error should be less than or equal to the value specified in the MARGIN= option. Any difference between the expected margin and the specified value is due to rounding the target stratum sample sizes up to integer values and increasing stratum sample sizes to equal the required minimum value (ALLOCMIN=).

Secondary Input Data Set

The primary input data set for PROC SURVEYSELECT is the DATA= data set, which contains the list of units from which the sample is selected. You can use a secondary input data set to provide stratum-level design and selection information, such as sample sizes or rates, certainty size values, or stratum costs. This secondary input data set is sometimes called the SAMPSIZE= input data set. You can provide stratum sample sizes in the _NSIZE_ (or SampleSize) variable in the SAMPSIZE= data set.

The secondary input data set must contain all the STRATA variables, with the same type and length as in the DATA= data set. The STRATA groups should appear in the same order in the secondary data set as in the DATA= data set. You can name only one secondary data set in each invocation of PROC SURVEYSELECT.

You must name the secondary input data set in the appropriate PROC SURVEYSELECT or STRATA option, and use the designated variable name to provide the stratum-level values. For example, if you want to provide stratum-level costs for sample allocation, you name the secondary data set in the COST=*SAS-data-set* option in the STRATA statement. The data set must include the stratum costs in a variable named _COST_. You can use the secondary input data set for more than one option if it is appropriate for your design. For example, the secondary data set can include both stratum costs and stratum variances, which are required for optimal allocation (ALLOC=OPTIMAL).

Instead of using a separate secondary input data set, you can include secondary information in the DATA= data set along with the sampling frame. When you include secondary information in the DATA= data set, name the DATA= data set in the appropriate options, and include the required variables in the DATA= data set.

Table 123.3 lists the available secondary data set variables, together with their descriptions and the corresponding options.

Table 123.3 PROC SURVEYSELECT Secondary Data Set Variables

Variable	Description	Statement	Option
ALLOC	Allocation proportion	STRATA	ALLOC=
CERTP	Certainty proportion	PROC	CERTSIZE=P=
CERTSIZE	Certainty size	PROC	CERTSIZE=
COST	Cost	STRATA	COST=
MAXSIZE	Maximum size	PROC	MAXSIZE=
MINSIZE	Minimum size	PROC	MINSIZE=
NSIZE	Sample size	PROC	SAMPSIZE=
RATE	Sampling rate	PROC	SAMPRATE=
SEED	Random number seed	PROC	SEED=
VAR	Variance	STRATA	VAR=

Sample Output Data Set

PROC SURVEYSELECT selects a sample and creates a SAS data set that contains the sample of selected units unless you specify the NOSAMPLE option in the STRATA statement or the GROUPS= option in the PROC SURVEYSELECT statement. When you specify the NOSAMPLE option, PROC SURVEYSELECT allocates the total sample size among strata but does not select a sample; the output data set contains the allocated sample sizes. For more information, see the section “Allocation Output Data Set” on page 10241. When you specify the GROUPS= option, PROC SURVEYSELECT randomly assigns observations to groups and does not select a sample. For more information, see the section “Random Assignment Output Data Set” on page 10242.

You can specify the name of the sample output data set in the OUT= option in the PROC SURVEYSELECT statement. If you omit the OUT= option, the data set is named DATA*n*, where *n* is the smallest integer that makes the name unique.

The output data set contains the units that are selected for the sample. These units are either observations or groups of observations (clusters) that you define by specifying the SAMPLINGUNIT statement. If you do not specify the SAMPLINGUNIT statement to define units (clusters), then PROC SURVEYSELECT uses observations as sampling units by default.

By default, the output data set contains only those units that are selected for the sample. If you specify the OUTALL option, the output data set includes all observations from the input data set and also contains a variable that indicates each observation’s selection status. For an observation that is selected, the value of the variable Selected is 1; for an observation that is not selected, the value of Selected is 0. The OUTALL option is available for equal probability selection methods.

By default, the output data set contains a single copy of each selected unit, even if the unit is selected more than once, and the variable NumberHits records the number of hits (selections) for each unit. A unit can be selected more than once if you use a with-replacement or with-minimum-replacement selection method (METHOD=URS, METHOD=PPS_WR, METHOD=PPS_SYS, or METHOD=PPS_SEQ). If you specify the OUTHITS option, the output data set includes a distinct copy of each selection in the output data set; for example, the output data set includes three copies of a unit that is selected three times (NumberHits is three).

The output data set also contains design information and selection statistics, depending on the selection method and output options that you specify. The output data set can include the following variables:

- Selected, which indicates whether or not the observation is selected for the sample. This variable is included if you specify the OUTALL option. For an observation that is selected, the value of the variable Selected is 1; for an observation that is not selected, the value of Selected is 0.
- STRATA variables, which you specify in the STRATA statement.
- Replicate, which is the sample replicate number. This variable is included when you specify the REPS= option. You can specify a different name for this variable in the REPNAME= option.
- SAMPLINGUNIT (CLUSTER) variables, which you specify in the SAMPLINGUNIT statement.
- ID variables, which you name in the ID statement.
- CONTROL variables, which you specify in the CONTROL statement.
- Zone, which is the selection zone. This variable is included for METHOD=PPS_SEQ.
- SIZE variable, which you specify in the SIZE statement.
- AdjustedSize, which is the adjusted size measure. This variable is included if you request adjusted sizes with the MINSIZE= or MAXSIZE= option when your sampling units are observations.
- UnitSize, which is the sampling unit (or cluster) size measure. This variable is included if you specify the SAMPLINGUNIT statement.
- Certain, which indicates certainty selection. This variable is included if you specify the CERTSIZE= or CERTSIZE=P= option. For units that are selected with certainty (because their size measures exceed the certainty size value or the certainty proportion), the value of Certain is 1; for other units, the value of Certain is 0.
- NumberHits, which is the number of hits (selections). This variable is included for selection methods that are with replacement or with minimum replacement (METHOD=URS, METHOD=PPS_WR, METHOD=PPS_SYS, and METHOD=PPS_SEQ).

The output data set includes the following variables if you request a PPS selection method or if you specify the STATS option in the PROC SURVEYSELECT statement for other methods:

- ExpectedHits, which is the expected number of hits (selections). This variable is included for selection methods that are with replacement or with minimum replacement, where the same unit can be selected more than once (METHOD=URS, METHOD=BALBOOTSTRAP, METHOD=PPS_WR, METHOD=PPS_SYS, and METHOD=PPS_SEQ).
- SelectionProb, which is the probability of selection. This variable is included for selection methods that are without replacement.
- SamplingWeight, which is the sampling weight. The value of this variable is the inverse of ExpectedHits or SelectionProb.

If you specify the STATS or OUTSIZE option for METHOD=BERNOULLI, the output data set contains the following variables. If you specify a STRATA statement, the output data set includes stratum-level values of these variables; otherwise, the output data set includes overall values.

- Total, which is the total number of sampling units
- SelectionProb, which is the selection probability that you specify in the SAMPRATE= option
- ExpectedN, which is the expected value of the sample size
- SampleSize, which is the actual sample size

If you specify the STATS option for METHOD=BERNOULLI, the output data set also contains the following variable:

- AdjSamplingWeight, which is the adjusted sampling weight

For METHOD=PPS_BREWER and METHOD=PPS_MURTHY, either of which selects two units from each stratum with probability proportional to size, the output data set contains the following variable:

- JtSelectionProb, which is the joint probability of selection for the two units selected from the stratum.

If you specify the JTPROBS option to compute joint probabilities of selection for METHOD=PPS or METHOD=PPS_SAMPFORD, then the output data set contains the following variables:

- Unit, which is an identification variable that numbers the selected units sequentially within each stratum.
- JtProb_1, JtProb_2, JtProb_3, . . . , where the variable JtProb_1 contains the joint probability of selection for the current unit and unit 1. Similarly, JtProb_2 contains the joint probability of selection for the current unit and unit 2, and so on.

If you specify the JTPROBS option for METHOD=PPS_WR, then the output data set contains the following variables:

- Unit, which is an identification variable that numbers the selected units sequentially within each stratum.
- JtHits_1, JtHits_2, JtHits_3, . . . , where the variable JtHits_1 contains the joint expected number of hits for the current unit and unit 1. Similarly, JtHits_2 contains the joint expected number of hits for the current unit and unit 2, and so on.

If you specify the OUTSIZE option, the output data set contains the following variables. If you specify a STRATA statement, the output data set includes stratum-level values of these variables; otherwise, the output data set includes overall values.

- MinimumSize, which is the minimum size measure that you specify in the MINSIZE= option. This variable is included if you specify the MINSIZE= option.
- MaximumSize, which is the maximum size measure that you specify in the MAXSIZE= option. This variable is included if you specify the MAXSIZE= option.
- CertaintySize, which is the certainty size measure that you specify in the CERTSIZE= option. This variable is included if you specify the CERTSIZE= option.
- CertaintyProp, which is the certainty proportion that you specify in the CERTSIZE=P= option. This variable is included if you specify the CERTSIZE=P= option.
- Total, which is the total number of sampling units in the stratum. This variable is included if you do not specify a SIZE statement or a SAMPLINGUNIT statement.
- TotalSize, which is the total of size measures in the stratum. This variable is included if you specify a SIZE statement or the PPS option in the SAMPLINGUNIT statement.
- TotalAdjSize, which is the total of adjusted size measures in the stratum. This variable is included if you request adjusted sizes in the MAXSIZE= or MINSIZE= option.
- SamplingRate, which is the sampling rate. This variable is included if you specify the SAMPRATE= option.
- SampleSize, which is the sample size. This variable is included if you specify the SAMPSIZE= option, or if you specify METHOD=PPS_BREWER or METHOD=PPS_MURTHY, either of which selects two units from each stratum.
- Interval, which is the specified systematic interval. This variable is included if you specify the INTERVAL= option for METHOD=SYS or METHOD=PPS_SYS.
- NCertain, which is the number of certainty units. This variable is included if you specify the CERTSIZE= or CERTSIZE=P= option and CERTUNITS=OUTPUT.

If you specify the OUTSEED option, the output data set contains the following variable:

- InitialSeed, which is the initial seed for the stratum.

If you specify the ALLOC= option in the STRATA statement, the output data set contains the following variables:

- Total, which is the total number of sampling units in the stratum.
- Variance, which is the stratum variance. This variable is included if you specify the VAR, VAR=(*values*), or VAR=*SAS-data-set* option for the ALLOC=OPTIMAL, ALLOC=NEYMAN, or MARGIN= allocation option.
- Cost, which is the stratum cost. This variable is included if you specify the COST, COST=(*values*), or COST=*SAS-data-set* option for ALLOC=OPTIMAL.

- AllocProportion, which is the target allocation proportion (the proportion of the total sample size to allocate to the stratum). PROC SURVEYSELECT computes this proportion by using the allocation method that you specify.
- SampleSize, which is the sample size allocated to the stratum.
- ActualProportion, which is the actual proportion allocated to the stratum. The value of ActualProportion equals the allocated stratum sample size divided by the total sample size. This value can differ from the target AllocProportion because of rounding and other restrictions. For more information, see the section "Sample Size Allocation" on page 10232.

Allocation Output Data Set

When you specify the NOSAMPLE option in the STRATA statement, PROC SURVEYSELECT allocates the total sample size among the strata but does not select the sample. In this case, the OUT= data set contains the allocated sample sizes.

You can specify the name of the allocation output data set with the OUT= option in the PROC SURVEYSELECT statement. If you omit the OUT= option, the data set is named DATA*n*, where *n* is the smallest integer that makes the name unique.

The allocation output data set contains one observation for each stratum. The data set can include the following variables:

- STRATA variables, which you specify in the STRATA statement.
- Total, which is the total number of sampling units in the stratum.
- Variance, which is the stratum variance. This variable is included if you specify the VAR, VAR=(*values*), or VAR=*SAS-data-set* option for the ALLOC=OPTIMAL, ALLOC=NEYMAN, or MARGIN= allocation option.
- Cost, which is the stratum cost. This variable is included if you specify the COST, COST=(*values*), or COST=*SAS-data-set* option for ALLOC=OPTIMAL.
- AllocProportion, which is the target allocation proportion (the proportion of the total sample size to allocate to the stratum). PROC SURVEYSELECT computes this proportion by using the allocation method that you specify.
- SampleSize, which is the sample size allocated to the stratum.
- ActualProportion, which is the actual proportion allocated to the stratum. The value of ActualProportion equals the allocated stratum sample size divided by the total sample size. This value can differ from the target AllocProportion because of rounding and other restrictions. For more information, see the section "Sample Size Allocation" on page 10232.

Random Assignment Output Data Set

When you specify the GROUPS= option, PROC SURVEYSELECT provides random assignment of the observations in the DATA= input data set. The OUT= output data set contains all observations in the input data set and identifies the assigned groups. If you do not specify an ID statement, the output data set contains all variables in the input data set. If you specify an ID statement, PROC SURVEYSELECT copies those variable that you specify from the input data set to the output data set.

You can specify the name of the output data set in the OUT= option in the PROC SURVEYSELECT statement. If you omit the OUT= option, the data set is named DATA*n*, where *n* is the smallest integer that makes the name unique.

The random assignment output data set can include the following variables:

- STRATA variables, if you specify a STRATA statement
- Replicate, which is the replicate identification number. This variable is included when you specify the REPS= option.
- ID variables, if you specify an ID statement
- GroupID, which is the group identification number. If you specify a STRATA statement, PROC SURVEYSELECT performs random assignment independently within strata, and the groups are nested within strata.
- InitialSeed, which is the initial seed for random number generation

If you specify the OUTSIZE option, the random assignment output data set also includes the following variables:

- Total, which is the total number of units in the data set, or the total in the stratum if you specify a STRATA statement
- NGroups, which is the number of groups in the data set, or the number in the stratum if you specify a STRATA statement
- GroupSize, which is the number of units in the observation's group

Displayed Output

By default, PROC SURVEYSELECT displays two tables that summarize the sample selection: the “Sample Selection Method” table and the “Sample Selection Summary” table.

If you request sample allocation but no sample selection, PROC SURVEYSELECT displays two tables that summarize the allocation: the “Sample Allocation Method” table and the “Sample Allocation Summary” table.

If you request random assignment, the procedure displays the “Random Assignment” table.

You can suppress display of these tables by specifying the NOPRINT option.

PROC SURVEYSELECT creates an output data set that contains the units that are selected for the sample. Or if you request sample allocation but no sample selection, PROC SURVEYSELECT creates an output data set that contains the sample size allocation results. If you request random assignment, the procedure creates an output data set that contains the assignments. For more information, see the sections “Sample Output Data Set” on page 10237, “Allocation Output Data Set” on page 10241, and “Random Assignment Output Data Set” on page 10242. The procedure does not display the output data set that it creates. Use PROC PRINT, PROC REPORT, or any other SAS reporting tool to display the output data set.

Sample Selection Method Table

PROC SURVEYSELECT displays the following information in the “Sample Selection Method” table:

- Selection Method
- Sampling Unit Variables, if you specify a SAMPLINGUNIT statement
- Size Measure variable, if you specify a SIZE statement
- Size Measure: Number of Observations, if you specify the PPS option in the SAMPLINGUNIT statement and do not specify a SIZE statement
- Minimum Size Measure, if you specify the MINSIZE= option
- Maximum Size Measure, if you specify the MAXSIZE= option
- Certainty Size Measure, if you specify the CERTSIZE= option
- Certainty Proportion, if you specify the CERTSIZE=P= option
- Strata Variables, if you specify a STRATA statement
- Control Variables, if you specify a CONTROL statement
- Control Sorting (Serpentine or Nested), if you specify a CONTROL statement
- Allocation (Proportional, Neyman, Optimal, or Input), if you specify the ALLOC= option in the STRATA statement
- Margin of Error, if you specify the MARGIN= option in the STRATA statement
- Confidence Level, if you specify the ALPHA= option in the STRATA statement

Sample Selection Summary Table

PROC SURVEYSELECT displays the following information in the “Sample Selection Summary” table:

- Input Data Set name
- Sorted Data Set name, if you specify the OUTSORT= option
- Random Number Seed
- Sample Size or Stratum Sample Size, if you specify the SAMPSIZE=*n* option
- Sample Size Data Set, if you specify the SAMPSIZE=*SAS-data-set* option
- Sampling Rate or Stratum Sampling Rate, if you specify the SAMPRATE=*value* option for METHOD=SRS, METHOD=URS, METHOD=SYS, or METHOD=SEQ.
- Selection Probability or Stratum Selection Probability, if you specify the SAMPRATE=*value* option for METHOD=BERNOULLI
- Sampling Rate Data Set, if you specify the SAMPRATE=*SAS-data-set* option
- Minimum Sample Size or Stratum Minimum Sample Size, if you specify the NMIN= option in the SAMPRATE= option
- Maximum Sample Size or Stratum Maximum Sample Size, if you specify the NMAX= option in the SAMPRATE= option
- Number of Certainty Units, if you specify the CERTSIZE= or CERTSIZE=P= option and do not specify a STRATA statement
- Specified Start, if you specify the START= option for METHOD=SYS or METHOD=PPS_SYS
- Random Start, if you specify the DETAILS option for METHOD=SYS or METHOD=PPS_SYS and do not specify a STRATA statement or the REPS= option
- Specified Interval, if you specify the INTERVAL= option for METHOD=SYS or METHOD=PPS_SYS
- Systematic Interval, if you specify the DETAILS option for METHOD=SYS or METHOD=PPS_SYS and do not specify a STRATA statement or the REPS= option
- Sample Size, if you specify the INTERVAL= option for METHOD=SYS or METHOD=PPS_SYS and do not specify a STRATA statement or the REPS= option
- Allocation Input Data Set name, if you specify the ALLOC=*SAS-data-set* option in the STRATA statement
- Variance Input Data Set name, if you specify the VAR=*SAS-data-set* option in the STRATA statement
- Cost Input Data Set name, if you specify the COST=*SAS-data-set* option in the STRATA statement
- Selection Probability, if you specify METHOD=SRS, METHOD=SYS, or METHOD=SEQ and do not specify a SIZE statement or a STRATA statement
- Expected Number of Hits, if you specify METHOD=URS and do not specify a STRATA statement

- Total Number of Units, if you specify METHOD=BERNOULLI or METHOD=POISSON and do not specify a STRATA statement
- Expected Sample Size, if you specify METHOD=BERNOULLI or METHOD=POISSON and do not specify a STRATA statement
- Sample Size, if you specify METHOD=BERNOULLI or METHOD=POISSON and do not specify a STRATA statement
- Sampling Weight, if you specify an equal probability selection method (METHOD=SRS, METHOD=URS, METHOD=SYS, METHOD=SEQ, or METHOD=BERNOULLI) and do not specify a STRATA statement
- Adjusted Sampling Weight, if you specify METHOD=BERNOULLI and do not specify a STRATA statement
- Number of Strata, if you specify a STRATA statement
- Stratum Minimum Sample Size, if you specify the ALLOCMIN= option in the STRATA statement
- Number of Replicates, if you specify the REPS= option
- Total Sample Size, if you specify a STRATA statement or the REPS= option
- Expected Margin of Error, if you specify the STATS option with the MARGIN= option in the STRATA statement
- Expected Variance, if you specify the STATS option without the MARGIN= option in the STRATA statement for ALLOC=OPTIMAL or ALLOC=NEYMAN
- Total Stratum Costs, if you specify the STATS option with ALLOC=OPTIMAL in the STRATA statement
- Output Data Set name

Sample Allocation Method Table

If you specify the NOSAMPLE option in the STRATA statement, PROC SURVEYSELECT allocates the total sample among the strata but does not select the sample. When you specify the NOSAMPLE option, PROC SURVEYSELECT displays the “Sample Allocation Method” table and the “Sample Allocation Summary” table. The “Sample Allocation Method” table includes the following information:

- Allocation (Proportional, Neyman, Optimal, or Input)
- Margin of Error, if you specify the MARGIN= option in the STRATA statement
- Confidence Level, if you specify the ALPHA= option in the STRATA statement
- Sampling Unit Variables, if you specify a SAMPLINGUNIT statement
- Strata Variables
- Frequency Variable
- Selection Method, if you specify the METHOD= option

Sample Allocation Summary Table

PROC SURVEYSELECT displays the following information in the “Sample Allocation Summary” table.

- Input Data Set name
- Allocation Input Data Set name, if you specify the ALLOC=*SAS-data-set* option in the STRATA statement
- Variance Input Data Set name, if you specify the VAR=*SAS-data-set* option in the STRATA statement
- Cost Input Data Set name, if you specify the COST=*SAS-data-set* option in the STRATA statement
- Number of Strata
- Stratum Minimum Sample Size, if you specify the ALLOCMIN= option in the STRATA statement
- Total Sample Size
- Expected Margin of Error, if you specify the STATS option with the MARGIN= option in the STRATA statement
- Expected Variance, if you specify the STATS option without the MARGIN= option in the STRATA statement for ALLOC=OPTIMAL or ALLOC=NEYMAN
- Total Stratum Costs, if you specify the STATS option with ALLOC=OPTIMAL in the STRATA statement
- Allocation Output Data Set name

Random Assignment Table

If you specify the GROUPS= option, PROC SURVEYSELECT displays the following information in the “Random Assignment” table:

- Input Data Set name
- Strata Variables, if you specify a STRATA statement
- Random Number Seed
- Number of Groups
- Total Number of Units, if you specify the GROUPS= option and do not specify a STRATA statement
- Number of Units per Group, if you specify the GROUPS= option and do not specify a STRATA statement
- Number of Replicates, if you specify the REPS= option
- Number of Strata, if you specify a STRATA statement
- Total Number of Groups, if you specify a STRATA statement or the REPS= option
- Output Data Set name

ODS Table Names

PROC SURVEYSELECT assigns a name to each table that it creates. You can use these names to refer to tables when you use the Output Delivery System (ODS) to select tables and create output data sets. For more information about ODS, see Chapter 22, "Using the Output Delivery System." Table 123.4 lists the table names.

Table 123.4 ODS Tables Produced by PROC SURVEYSELECT

ODS Table Name	Description	Statement	Option
Groups	Random assignment summary	PROC	GROUPS=
Method	Sample selection method	PROC	Default
Method	Sample allocation method	STRATA	NOSAMPLE
Summary	Sample selection summary	PROC	Default
Summary	Sample allocation summary	STRATA	NOSAMPLE

Examples: SURVEYSELECT Procedure

Example 123.1: Replicated Sampling

This example uses the Customers data set from the section "Getting Started: SURVEYSELECT Procedure" on page 10180. The data set Customers contains an Internet service provider's current subscribers, and the service provider wants to select a sample from this population for a customer satisfaction survey.

This example illustrates replicated sampling, which selects multiple samples from the survey population according to the same design. You can use replicated sampling to provide a simple method of variance estimation, or to evaluate variable nonsampling errors such as interviewer differences. For information about replicated sampling, see Lohr (2010), Wolter (2007), Kish (1965), Kish (1987), and Kalton (1983).

This design includes four replicates, which each have a sample size of 50 customers. The sampling frame is stratified by State and sorted by Type and Usage within strata. Customers are selected by sequential random sampling with equal probability within strata. The following PROC SURVEYSELECT statements select a probability sample of customers from the Customers data set by using this design:

```
title1 'Customer Satisfaction Survey';
title2 'Replicated Sampling';
proc surveyselect data=Customers method=seq n=(8 12 20 10)
                  reps=4 seed=40070 ranuni out=SampleRep;
   strata State;
   control Type Usage;
run;
```

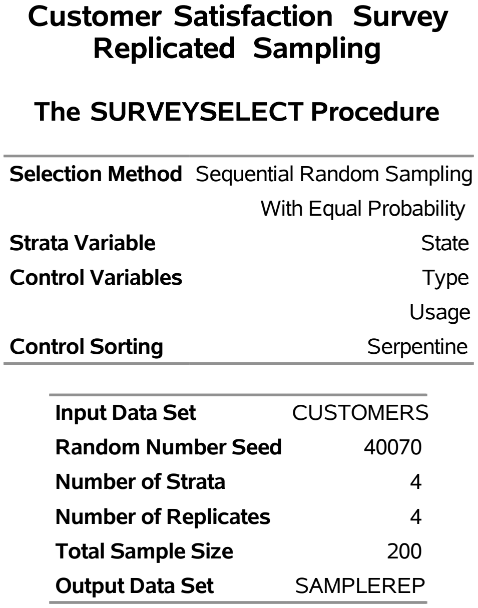
The STRATA statement names the stratification variable State. The CONTROL statement names the control variables Type and Usage.

In the PROC SURVEYSELECT statement, the METHOD=SEQ option requests sequential random sampling. The REPS= option specifies 4 replicates of this sample. The N=(8 12 20 10) option specifies the stratum sample sizes in each replicate. The N= option lists the stratum sample sizes in the same order as the strata appear in the Customers data set, which is sorted by State. The sample size of 8 customers corresponds to the first stratum, (State = ‘AL’); the sample size of 12 customers corresponds to the second stratum (State = ‘FL’), and so on.

The SEED= option specifies 40070 as the initial seed for random number generation. The RANUNI option requests random number generation by the RANUNI random number generator, which PROC SURVEYSELECT uses in releases before SAS/STAT 12.1. (Beginning in SAS/STAT 12.1, PROC SURVEYSELECT uses the Mersenne twister random number generator by default.) You can specify the RANUNI option along with the same SEED= value to reproduce a sample that PROC SURVEYSELECT selects in releases before SAS/STAT 12.1. To reproduce a sample by using the RANUNI and SEED= options, you must also specify the same input data set and sample selection parameters.

Output 123.1.1 displays the output from PROC SURVEYSELECT, which summarizes the sample selection. A total of 200 customers is selected in 4 replicates. PROC SURVEYSELECT selects each replicate by using sequential random sampling within strata that are determined by State. The sampling frame Customers is sorted by the control variables Type and Usage within strata, according to hierarchic serpentine sorting. The output data set SampleRep contains the sample.

Output 123.1.1 Sample Selection Summary



```
                Customer Satisfaction Survey
                     Replicated Sampling

                The SURVEYSELECT Procedure

        Selection Method  Sequential Random Sampling
                                 With Equal Probability
        Strata Variable                          State
        Control Variables                         Type
                                                 Usage
        Control Sorting                     Serpentine

            Input Data Set          CUSTOMERS
            Random Number Seed          40070
            Number of Strata                4
            Number of Replicates            4
            Total Sample Size             200
            Output Data Set         SAMPLEREP
```

The following PROC PRINT statements display the selected customers for the first stratum (State = ‘AL’) in the output data set SampleRep:

```
title1 'Customer Satisfaction Survey';
title2 'Sample Selected by Replicated Design';
title3 '(First Stratum)';
proc print data=SampleRep;
   where State = 'AL';
run;
```

Output 123.1.2 displays the 32 sample customers in the first stratum (State = ‘AL’) from the output data set SampleRep, which includes the entire sample of 200 customers. The variable SelectionProb contains the selection probability, and SamplingWeight contains the sampling weight. Because customers are selected with equal probability within strata in this design, all customers in the same stratum have the same selection probability. These selection probabilities and sampling weights apply to a single replicate, and the variable Replicate contains the sample replicate number.

Output 123.1.2 Customer Sample (First Stratum)

Customer Satisfaction Survey
Sample Selected by Replicated Design
(First Stratum)

Obs	State	Replicate	CustomerID	Type	Usage	SelectionProb	SamplingWeight
1	AL	1	882-37-7496	New	572	.004115226	243
2	AL	1	581-32-5534	New	863	.004115226	243
3	AL	1	980-29-2898	Old	571	.004115226	243
4	AL	1	172-56-4743	Old	128	.004115226	243
5	AL	1	998-55-5227	Old	35	.004115226	243
6	AL	1	625-44-3396	New	60	.004115226	243
7	AL	1	627-48-2509	New	114	.004115226	243
8	AL	1	257-66-6558	New	172	.004115226	243
9	AL	2	622-83-1680	New	22	.004115226	243
10	AL	2	343-57-1186	New	53	.004115226	243
11	AL	2	976-05-3796	New	110	.004115226	243
12	AL	2	859-74-0652	New	303	.004115226	243
13	AL	2	476-48-1066	New	839	.004115226	243
14	AL	2	109-27-8914	Old	2102	.004115226	243
15	AL	2	743-25-0298	Old	376	.004115226	243
16	AL	2	722-08-2215	Old	105	.004115226	243
17	AL	3	668-57-7696	New	200	.004115226	243
18	AL	3	300-72-0129	New	471	.004115226	243
19	AL	3	073-60-0765	New	656	.004115226	243
20	AL	3	526-87-0258	Old	672	.004115226	243
21	AL	3	726-61-0387	Old	150	.004115226	243
22	AL	3	632-29-9020	Old	51	.004115226	243
23	AL	3	417-17-8378	New	56	.004115226	243
24	AL	3	091-26-2366	New	93	.004115226	243
25	AL	4	336-04-1288	New	419	.004115226	243
26	AL	4	827-04-7407	New	650	.004115226	243
27	AL	4	317-70-6496	Old	452	.004115226	243
28	AL	4	002-38-4582	Old	206	.004115226	243
29	AL	4	181-83-3990	Old	33	.004115226	243
30	AL	4	675-34-7393	New	47	.004115226	243
31	AL	4	228-07-6671	New	65	.004115226	243
32	AL	4	298-46-2434	New	161	.004115226	243

Example 123.2: PPS Selection of Two Units per Stratum

This example describes hospital selection for a survey by using PROC SURVEYSELECT. A state health agency plans to conduct a statewide survey of a variety of different hospital services. The agency plans to select a probability sample of individual discharge records within hospitals by using a two-stage sample design. First-stage units are hospitals, and second-stage units are patient discharges during the study period. Hospitals are stratified first according to geographic region and then by rural/urban type and size of hospital. Two hospitals are selected from each stratum with probability proportional to size.

The data set HospitalFrame contains all hospitals in the first geographical region of the state:

```
data HospitalFrame;
   input Hospital$ Type$ SizeMeasure @@;
   if (SizeMeasure < 20) then Size='Small ';
      else if (SizeMeasure < 50) then Size='Medium';
      else Size='Large ';
   datalines;
034 Rural  0.870   107 Rural  1.316
079 Rural  2.127   223 Rural  3.960
236 Rural  5.279   165 Rural  5.893
086 Rural  0.501   141 Rural 11.528
042 Urban  3.104   124 Urban  4.033
006 Urban  4.249   261 Urban  4.376
195 Urban  5.024   190 Urban 10.373
038 Urban 17.125   083 Urban 40.382
259 Urban 44.942   129 Urban 46.702
133 Urban 46.992   218 Urban 48.231
026 Urban 61.460   058 Urban 65.931
119 Urban 66.352
;
```

In the SAS data set HospitalFrame, the variable Hospital identifies the hospital. The variable Type equals ‘Urban’ if the hospital is located in an urban area, and ‘Rural’ otherwise. The variable SizeMeasure contains the hospital’s size measure, which is constructed from past data on service utilization for the hospital together with the target sampling rates for each service. This size measure reflects the amount of relevant survey information expected from the hospital. For information about this type of size measure, see Drummond et al. (1982). The value of the variable Size is ‘Small’, ‘Medium’, or ‘Large’, depending on the value of the hospital’s size measure.

The following PROC PRINT statements display the data set Hospital Frame and produce Output 123.2.1:

```
title1 'Hospital Utilization Survey';
title2 'Sampling Frame, Region 1';
proc print data=HospitalFrame;
run;
```

Output 123.2.1 Sampling Frame

Hospital Utilization Survey
Sampling Frame, Region 1

Obs	Hospital	Type	SizeMeasure	Size
1	034	Rural	0.870	Small
2	107	Rural	1.316	Small
3	079	Rural	2.127	Small
4	223	Rural	3.960	Small
5	236	Rural	5.279	Small
6	165	Rural	5.893	Small
7	086	Rural	0.501	Small
8	141	Rural	11.528	Small
9	042	Urban	3.104	Small
10	124	Urban	4.033	Small
11	006	Urban	4.249	Small
12	261	Urban	4.376	Small
13	195	Urban	5.024	Small
14	190	Urban	10.373	Small
15	038	Urban	17.125	Small
16	083	Urban	40.382	Medium
17	259	Urban	44.942	Medium
18	129	Urban	46.702	Medium
19	133	Urban	46.992	Medium
20	218	Urban	48.231	Medium
21	026	Urban	61.460	Large
22	058	Urban	65.931	Large
23	119	Urban	66.352	Large

The following PROC SURVEYSELECT statements select a probability sample of hospitals from the HospitalFrame data set by using a stratified design with PPS selection of two units from each stratum:

```
title1 'Hospital Utilization Survey';
title2 'Stratified PPS Sampling';
proc surveyselect data=HospitalFrame method=pps_brewer
                  seed=48702 out=SampleHospitals;
   size SizeMeasure;
   strata Type Size notsorted;
run;
```

The STRATA statement names the stratification variables Type and Size. The NOTSORTED option specifies that observations with the same STRATA variable values are grouped together but are not necessarily sorted in alphabetical or increasing numerical order. In the HospitalFrame data set, Size = ‘Small’ precedes Size = ‘Medium’.

In the PROC SURVEYSELECT statement, the METHOD=PPS_BREWER option requests sample selection by Brewer’s method, which selects two units per stratum with probability proportional to size. The SEED= option specifies 48702 as the initial seed for random number generation. The SIZE statement names SizeMeasure as the size measure variable. It is not necessary to specify the sample size in the N= option because Brewer’s method selects two units from each stratum.

Output 123.2.2 displays the output from PROC SURVEYSELECT. A total of 8 hospitals are selected from the 4 strata. The data set SampleHospitals contains the selected hospitals.

Output 123.2.2 Sample Selection Summary

Hospital Utilization Survey
Stratified PPS Sampling

The SURVEYSELECT Procedure

Selection Method	Brewer's PPS Method
Size Measure	SizeMeasure
Strata Variables	Type Size

Input Data Set	HOSPITALFRAME
Random Number Seed	48702
Stratum Sample Size	2
Number of Strata	4
Total Sample Size	8
Output Data Set	SAMPLEHOSPITALS

The following PROC PRINT statements display the sample hospitals and produce Output 123.2.3:

```
title1 'Hospital Utilization Survey';
title2 'Sample Selected by Stratified PPS Design';
proc print data=SampleHospitals;
run;
```

Output 123.2.3 Sample Hospitals

Hospital Utilization Survey
Sample Selected by Stratified PPS Design

Obs	Type	Size	Hospital	SizeMeasure	SelectionProb	SamplingWeight	JtSelectionProb
1	Rural	Small	165	5.893	0.37447	2.67046	0.22465
2	Rural	Small	141	11.528	0.73254	1.36511	0.22465
3	Urban	Small	190	10.373	0.42967	2.32739	0.25370
4	Urban	Small	038	17.125	0.70934	1.40975	0.25370
5	Urban	Medium	083	40.382	0.35540	2.81374	0.08953
6	Urban	Medium	133	46.992	0.41357	2.41795	0.08953
7	Urban	Large	026	61.460	0.63445	1.57617	0.31940
8	Urban	Large	119	66.352	0.68495	1.45996	0.31940

The variable SelectionProb contains the selection probability for each hospital in the sample. The variable JtSelectionProb contains the joint probability of selection for the two sample hospitals in the same stratum. The variable SamplingWeight contains the sampling weight component for this first stage of the design. The final-stage weight components, which correspond to patient record selection within hospitals, can be multiplied by the hospital weight components to obtain the overall sampling weights.

Example 123.3: PPS (Dollar-Unit) Sampling

A small company wants to audit employee travel expenses in an effort to improve the expense reporting procedure and possibly reduce expenses. The company does not have resources to examine all expense reports and wants to use statistical sampling to objectively select expense reports for audit.

The data set TravelExpense contains the dollar amount of all employee travel expense transactions during the past month:

```
data TravelExpense;
   input ID$ Amount @@;
   if (Amount < 500) then Level='1_Low ';
      else if (Amount > 1500) then Level='3_High';
      else Level='2_Avg ';
   datalines;
110  237.18   002  567.89   234  118.50
743   74.38   411 1287.23   782  258.10
216  325.36   174  218.38   568 1670.80
302  134.71   285 2020.70   314   47.80
139 1183.45   775  330.54   425  780.10
506  895.80   239  620.10   011  420.18
672  979.66   142  810.25   738  670.85
192  314.58   243   87.50   263 1893.40
496  753.30   332  540.65   486 2580.35
614  230.56   654  185.60   308  688.43
784  505.14   017  205.48   162  650.42
289 1348.34   691   30.50   545 2214.80
517  940.35   382  217.85   024  142.90
478  806.90   107  560.72
;
```

In the SAS data set TravelExpense, the variable ID identifies the travel expense report. The variable Amount contains the dollar amount of the reported expense. The variable Level equals '1_Low', '2_Avg', or '3_High', depending on the value of Amount.

In the sample design for this audit, expense reports are stratified by Level. This ensures that each of these expense levels is included in the sample and also permits a disproportionate allocation of the sample, selecting proportionately more of the expense reports from the higher levels. Within strata, the sample of expense reports is selected with probability proportional to the amount of the expense, thus giving a greater chance of selection to larger expenses. In auditing terms, this is known as monetary-unit sampling. For more information, see Wilburn (1984).

PROC SURVEYSELECT requires that the input data set be sorted by the STRATA variables. The following PROC SORT statements sort the TravelExpense data set by the stratification variable Level.

```
proc sort data=TravelExpense;
   by Level;
run;
```

Output 123.3.1 displays the sampling frame data set TravelExpense, which contains 41 observations.

Output 123.3.1 Sampling Frame

Travel Expense Audit

Obs	ID	Amount	Level
1	110	237.18	1_Low
2	234	118.50	1_Low
3	743	74.38	1_Low
4	782	258.10	1_Low
5	216	325.36	1_Low
6	174	218.38	1_Low
7	302	134.71	1_Low
8	314	47.80	1_Low
9	775	330.54	1_Low
10	011	420.18	1_Low
11	192	314.58	1_Low
12	243	87.50	1_Low
13	614	230.56	1_Low
14	654	185.60	1_Low
15	017	205.48	1_Low
16	691	30.50	1_Low
17	382	217.85	1_Low
18	024	142.90	1_Low
19	002	567.89	2_Avg
20	411	1287.23	2_Avg
21	139	1183.45	2_Avg
22	425	780.10	2_Avg
23	506	895.80	2_Avg
24	239	620.10	2_Avg
25	672	979.66	2_Avg
26	142	810.25	2_Avg
27	738	670.85	2_Avg
28	496	753.30	2_Avg
29	332	540.65	2_Avg
30	308	688.43	2_Avg
31	784	505.14	2_Avg
32	162	650.42	2_Avg
33	289	1348.34	2_Avg
34	517	940.35	2_Avg
35	478	806.90	2_Avg
36	107	560.72	2_Avg
37	568	1670.80	3_High
38	285	2020.70	3_High
39	263	1893.40	3_High
40	486	2580.35	3_High
41	545	2214.80	3_High

The following PROC SURVEYSELECT statements select a probability sample of expense reports from the TravelExpense data set by using the stratified design with PPS selection within strata:

```
title1 'Travel Expense Audit';
title2 'Stratified PPS (Dollar-Unit) Sampling';
proc surveyselect data=TravelExpense method=pps n=(6 10 4)
                  seed=47279 out=AuditSample;
   size Amount;
   strata Level;
run;
```

The STRATA statement names the stratification variable Level. The SIZE statement specifies the size measure variable Amount. In the PROC SURVEYSELECT statement, the METHOD=PPS option requests sample selection with probability proportional to size and without replacement. The N=(6 10 4) option specifies the stratum sample sizes by listing the sample sizes in the same order as the strata appear in the TravelExpense data set. The sample size of 6 corresponds to the first stratum (Level = '1_Low'); the sample size of 10 corresponds to the second stratum (Level = '2_Avg'); and the sample size of 4 corresponds to the last stratum (Level = '3_High'). The SEED= option specifies 47279 as the initial seed for random number generation.

Output 123.3.2 displays the output from PROC SURVEYSELECT. A total of 20 expense reports are selected for audit. The data set AuditSample contains the sample of travel expense reports.

Output 123.3.2 Sample Selection Summary

```
                 Travel Expense Audit
        Stratified PPS (Dollar-Unit) Sampling

             The SURVEYSELECT Procedure

     Selection Method  PPS, Without Replacement
     Size Measure                        Amount
     Strata Variable                      Level

      Input Data Set          TRAVELEXPENSE
      Random Number Seed              47279
      Number of Strata                    3
      Total Sample Size                  20
      Output Data Set           AUDITSAMPLE
```

The following PROC PRINT statements display the audit sample, which is shown in Output 123.3.3:

```
title1 'Travel Expense Audit';
title2 'Sample Selected by Stratified PPS Design';
proc print data=AuditSample;
run;
```

Output 123.3.3 Audit Sample

Travel Expense Audit
Sample Selected by Stratified PPS Design

Obs	Level	ID	Amount	SelectionProb	SamplingWeight
1	1_Low	024	142.90	0.23949	4.17553
2	1_Low	614	230.56	0.38640	2.58797
3	1_Low	110	237.18	0.39750	2.51574
4	1_Low	782	258.10	0.43256	2.31183
5	1_Low	192	314.58	0.52721	1.89676
6	1_Low	216	325.36	0.54528	1.83392
7	2_Avg	239	620.10	0.42503	2.35278
8	2_Avg	308	688.43	0.47186	2.11925
9	2_Avg	496	753.30	0.51633	1.93676
10	2_Avg	478	806.90	0.55307	1.80810
11	2_Avg	142	810.25	0.55536	1.80063
12	2_Avg	517	940.35	0.64454	1.55151
13	2_Avg	672	979.66	0.67148	1.48925
14	2_Avg	139	1183.45	0.81116	1.23280
15	2_Avg	411	1287.23	0.88229	1.13341
16	2_Avg	289	1348.34	0.92418	1.08204
17	3_High	568	1670.80	0.64385	1.55316
18	3_High	263	1893.40	0.72963	1.37056
19	3_High	545	2214.80	0.85348	1.17167
20	3_High	486	2580.35	0.99435	1.00568

Example 123.4: Proportional Allocation

This example uses the Customers data set from the section "Getting Started: SURVEYSELECT Procedure" on page 10180. The data set Customers contains an Internet service provider's current subscribers, and the service provider wants to select a sample from this population for a customer satisfaction survey. This example illustrates proportional allocation, which allocates the total sample size among the strata in proportion to the strata sizes.

The section "Getting Started: SURVEYSELECT Procedure" on page 10180 gives an example of stratified sampling, where the list of customers is stratified by State and Type. Figure 123.4 displays the strata in a table of State by Type for the 13,471 customers. There are four states and two levels of Type, forming a total of eight strata. A sample of 15 customers was selected from each stratum by using the following PROC SURVEYSELECT statements:

```
title1 'Customer Satisfaction Survey';
title2 'Stratified Sampling';
proc surveyselect data=Customers method=srs n=15
                  seed=1953 out=SampleStrata;
   strata State Type;
run;
```

The STRATA statement names the stratification variables State and Type. In the PROC SURVEYSELECT statement, the N= option specifies a sample size of 15 customers in each stratum.

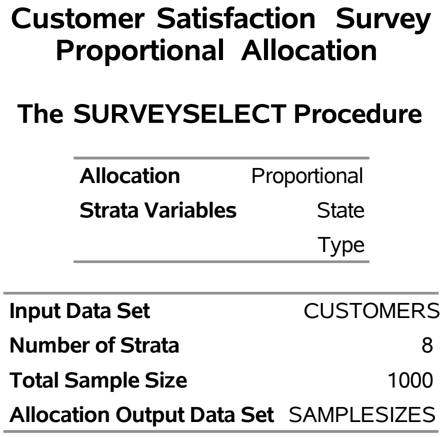
Instead of specifying the number of customers to select from each stratum, you can specify the total sample size and request allocation of the total sample size among the strata. The following PROC SURVEYSELECT statements request proportional allocation, which allocates the total sample size in proportion to the stratum sizes:

```
title1 'Customer Satisfaction Survey';
title2 'Proportional Allocation';
proc surveyselect data=Customers n=1000
                  out=SampleSizes;
   strata State Type / alloc=prop nosample;
run;
```

The STRATA statement names the stratification variables State and Type. In the STRATA statement, the ALLOC=PROP option requests proportional allocation. The NOSAMPLE option requests that no sample be selected after the procedure computes the sample size allocation. In the PROC SURVEYSELECT statement, the N= option specifies a total sample size of 1000 customers to be allocated among the strata.

Output 123.4.1 displays the output from PROC SURVEYSELECT, which summarizes the sample allocation. The total sample size of 1000 is allocated among the eight strata by using proportional allocation. The allocated sample sizes are stored in the SAS data set SampleSizes.

Output 123.4.1 Proportional Allocation Summary



```
        Customer Satisfaction  Survey
           Proportional  Allocation

         The SURVEYSELECT Procedure

        Allocation            Proportional
        Strata Variables             State
                                      Type

     Input Data Set               CUSTOMERS
     Number of Strata                     8
     Total Sample Size                 1000
     Allocation Output Data Set SAMPLESIZES
```

The following PROC PRINT statements display the allocation output data set SampleSizes, which is shown in Output 123.4.2:

```
title1 'Customer Satisfaction Survey';
title2 'Proportional Allocation';
proc print data=SampleSizes;
run;
```

Output 123.4.2 Stratum Sample Sizes

```
               Customer Satisfaction  Survey
                  Proportional  Allocation

Obs  State  Type  Total  AllocProportion  SampleSize  ActualProportion
  1  AL     New    1238          0.09190          92             0.092
  2  AL     Old     706          0.05241          52             0.052
  3  FL     New    2170          0.16109         161             0.161
  4  FL     Old    1370          0.10170         102             0.102
  5  GA     New    3488          0.25893         259             0.259
  6  GA     Old    1940          0.14401         144             0.144
  7  SC     New    1684          0.12501         125             0.125
  8  SC     Old     875          0.06495          65             0.065
```

The output data set SampleSizes includes one observation for each of the eight strata, which are identified by the stratification variables State and Type. The variable Total contains the number of sampling units in the stratum, and the variable AllocProportion contains the proportion of the total sample size to allocate to the stratum. The variable SampleSize contains the allocated stratum sample size. For the first stratum (State=‘AL’ and Type=‘New’), the total number of sampling units is 1238 customers, the allocation proportion is 0.09190, and the allocated sample size is 92 customers. The sum of the allocated sample sizes equals the requested total sample size of 1000 customers.

The output data set also includes the variable ActualProportion, which contains actual stratum proportions of the total sample size. The actual proportion for a stratum is the stratum sample size divided by the total sample size. For the first stratum (State=‘AL’ and Type=‘New’), the actual proportion is 0.092, whereas the allocation proportion is 0.09190. The target sample sizes computed from the allocation proportions are often not integers, and PROC SURVEYSELECT uses a rounding algorithm to obtain integer sample sizes and maintain the requested total sample size. Because of rounding and other restrictions, the actual proportions can differ from the target allocation proportions. For more information, see the section “Sample Size Allocation” on page 10232.

To use the allocated sample sizes in a later invocation of PROC SURVEYSELECT, you can name the allocation data set in the N=*SAS-data-set* option, as shown in the following PROC SURVEYSELECT statements:

```
title1 'Customer Satisfaction Survey';
title2 'Stratified Sampling';
proc surveyselect data=Customers method=srs n=SampleSizes
                  seed=1953 out=SampleStrata;
   strata State Type;
run;
```

References

Arkin, H. (1984). *Handbook of Sampling for Auditing and Accounting*. New York: McGraw-Hill.

Bentley, J. L., and Floyd, R. W. (1987). “Programming Pearls: A Sample of Brilliance.” *Communications of the Association for Computing Machinery* 30:754–757.

Bentley, J. L., and Knuth, D. E. (1986). “Literate Programming.” *Communications of the Association for Computing Machinery* 29:364–369.

Brewer, K. W. R. (1963). “A Model of Systematic Sampling with Unequal Probabilities.” *Australian Journal of Statistics* 5:93–105.

Cassell, D. L. (2007). “Don’t Be Loopy: Re-sampling and Simulation the SAS Way.” In *Proceedings of the SAS Global Forum 2007 Conference*. Cary, NC: SAS Institute Inc. http://www2.sas.com/proceedings/forum2007/183-2007.pdf.

Chromy, J. R. (1979). “Sequential Sample Selection Methods.” In *Proceedings of the Survey Research Methods Section*, 401–406. Washington, DC: American Statistical Association.

Cochran, W. G. (1977). *Sampling Techniques*. 3rd ed. New York: John Wiley & Sons.

Davison, A. C., Hinkley, D. V., and Schechtman, E. (1986). “Efficient Bootstrap Computation.” *Biometrika* 73:555–566.

Drummond, D., Lessler, J., Watts, D., and Williams, S. (1982). “A Design for Achieving Prespecified Levels of Representation for Multiple Domains in Health Record Samples.” In *Proceedings of the Fourth Conference on Health Survey Research Methods.* DHHS Publication No. (PHS) 84-3346, 233–248. Washington, DC: National Center for Health Services Research.

Durbin, J. (1967). “Design of Multi-stage Surveys for the Estimation of Sampling Errors.” *Journal of the Royal Statistical Society, Series C* 16:152–164.

Fan, C. T., Muller, M. E., and Rezucha, I. (1962). “Development of Sampling Plans by Using Sequential (Item by Item) Selection Techniques and Digital Computers.” *Journal of the American Statistical Association* 57:387–402.

Fishman, G. S., and Moore, L. R. (1982). “A Statistical Evaluation of Multiplicative Congruential Generators with Modulus $(2^{31} - 1)$.” *Journal of the American Statistical Association* 77:129–136.

Fox, D. R. (1989). “Computer Selection of Size-Biased Samples.” *American Statistician* 43:168–171.

Gleason, J. R. (1988). “Algorithms for Balanced Bootstrap Simulations.” *American Statistician* 42:263–266.

Golmant, J. (1990). “Correction: Computer Selection of Size-Biased Samples.” *American Statistician* 44:194.

Hanurav, T. V. (1967). “Optimum Utilization of Auxiliary Information: π_{ps} Sampling of Two Units from a Stratum.” *Journal of the Royal Statistical Society, Series B* 29:374–391.

Kalton, G. (1983). *Introduction to Survey Sampling.* Vol. 07-035 of University Paper Series on Quantitative Applications in the Social Sciences. Beverly Hills, CA: Sage Publications.

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

Kish, L. (1987). *Statistical Design for Research.* New York: John Wiley & Sons.

Lohr, S. L. (2010). *Sampling: Design and Analysis*. 2nd ed. Boston: Brooks/Cole.

Madow, W. G. (1949). “On the Theory of Systematic Sampling, II.” *Annals of Mathematical Statistics* 20:333–354.

Matsumoto, M., and Nishimura, T. (1998). "Mersenne Twister: A 623-Dimensionally Equidistributed Uniform Pseudo-random Number Generator." *ACM Transactions on Modeling and Computer Simulation* 8:3–30.

McLeod, A. I., and Bellhouse, D. R. (1983). "A Convenient Algorithm for Drawing a Simple Random Sample." *Journal of the Royal Statistical Society, Series C* 32:182–183.

Murthy, M. N. (1957). "Ordered and Unordered Estimators in Sampling without Replacement." *Sankhyā* 18:379–390.

Murthy, M. N. (1967). *Sampling Theory and Methods*. Calcutta: Statistical Publishing Society.

Ohlsson, E. (1998). "Sequential Poisson Sampling." *Journal of Official Statistics* 14:149–162.

Sampford, M. R. (1967). "On Sampling without Replacement with Unequal Probabilities of Selection." *Biometrika* 54:499–513.

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

Vijayan, K. (1968). "An Exact π_{ps} Sampling Scheme: Generalization of a Method of Hanurav." *Journal of the Royal Statistical Society, Series B* 30:556–566.

Watts, D. L. (1991). "Correction: Computer Selection of Size-Biased Samples." *American Statistician* 45:172.

Wilburn, A. J. (1984). *Practical Statistical Sampling for Auditors*. New York: Marcel Dekker.

Williams, R. L., and Chromy, J. R. (1980). "SAS Sample Selection Macros." In *Proceedings of the Fifth Annual SAS Users Group International Conference*. Cary, NC: SAS Institute Inc. https://support.sas.com/resources/papers/proceedings-archive/SUGI80/Sugi-80-71%20Williams%20Chromy.pdf.

Wolter, K. M. (2007). *Introduction to Variance Estimation*. 2nd ed. New York: Springer.

Subject Index

allocation
    of sample size (SURVEYSELECT), 10179, 10215, 10232

balanced bootstrap sampling
    SURVEYSELECT procedure, 10223
Bernoulli sampling
    SURVEYSELECT procedure, 10225
Brewer's selection method
    SURVEYSELECT procedure, 10231, 10250

Chromy's selection method
    SURVEYSELECT procedure, 10224, 10229
cluster sampling
    SURVEYSELECT procedure, 10212, 10222
clusters
    SURVEYSELECT procedure, 10178, 10212
control sorting
    SURVEYSELECT procedure, 10186, 10211, 10220, 10247

dollar-unit sampling
    SURVEYSELECT procedure, 10253

Hanurav-Vijayan selection method
    SURVEYSELECT procedure, 10227

initial seed
    SURVEYSELECT procedure, 10208

joint selection probabilities
    SURVEYSELECT procedure, 10192

margin of error
    SURVEYSELECT procedure, 10235
missing values
    SURVEYSELECT procedure, 10219
multistage sampling
    SURVEYSELECT procedure, 10178
Murthy's selection method
    SURVEYSELECT procedure, 10231

Neyman allocation
    SURVEYSELECT procedure, 10215, 10234

optimal allocation
    SURVEYSELECT procedure, 10215, 10233

Poisson sampling
    SURVEYSELECT procedure, 10226
population
    SURVEYSELECT procedure, 10178
PPS sampling
    SURVEYSELECT procedure, 10178, 10213, 10222
PPS sampling, with replacement
    SURVEYSELECT procedure, 10228
PPS sampling, without replacement
    SURVEYSELECT procedure, 10227
PPS sequential sampling
    SURVEYSELECT procedure, 10229
PPS systematic sampling
    SURVEYSELECT procedure, 10229
probability sampling
    SURVEYSELECT procedure, 10178
proportional allocation
    SURVEYSELECT procedure, 10216, 10233, 10256

random sampling
    SURVEYSELECT procedure, 10178
replicated sampling
    SURVEYSELECT procedure, 10179, 10203, 10247
replication, *see* replicated sampling

Sampford's selection method
    SURVEYSELECT procedure, 10232
sample
    SURVEYSELECT procedure, 10178
sample allocation
    SURVEYSELECT procedure, 10179, 10215, 10232
sample design
    SURVEYSELECT procedure, 10178
sample selection
    SURVEYSELECT procedure, 10178
sample selection methods
    SURVEYSELECT procedure, 10194, 10222
sample size
    SURVEYSELECT procedure, 10206
sample size allocation
    SURVEYSELECT procedure, 10179, 10215, 10232
sampling
    SURVEYSELECT procedure, 10178
sampling frame
    SURVEYSELECT procedure, 10178, 10192
sampling rate

  - SURVEYSELECT procedure, 10205
- sampling units
  - SURVEYSELECT procedure, 10178, 10212, 10222
- sampling weights
  - SURVEYSELECT procedure, 10181
- sampling with replacement
  - SURVEYSELECT procedure, 10222
- sampling without replacement
  - SURVEYSELECT procedure, 10222
- seed
  - initial (SURVEYSELECT), 10208
- sequential Poisson sampling
  - SURVEYSELECT procedure, 10226
- sequential random sampling
  - SURVEYSELECT procedure, 10224, 10247
- serpentine sorting
  - SURVEYSELECT procedure, 10220
- simple random sampling
  - SURVEYSELECT procedure, 10181, 10222
- size measures
  - PPS sampling (SURVEYSELECT), 10213, 10250
- sorting
  - SURVEYSELECT procedure, 10220
- strata
  - SURVEYSELECT procedure, 10179, 10183, 10214
- stratification, *see* stratified sampling
- stratified sampling
  - SURVEYSELECT procedure, 10179, 10183, 10214
- survey sampling
  - sample selection (SURVEYSELECT), 10178
  - SURVEYSELECT procedure, 10178
- survey weights, *see* sampling weights
- SURVEYSELECT procedure, 10178
  - allocation, 10215, 10232
  - allocation output data set, 10241
  - balanced bootstrap sampling, 10223
  - Bernoulli sampling, 10225
  - Brewer's selection method, 10231, 10250
  - certainty size measure, 10189
  - certainty size proportion, 10190
  - Chromy's selection method, 10224, 10229
  - cluster sampling, 10212
  - control sorting, 10186, 10211, 10220, 10247
  - displayed output, 10243
  - dollar-unit sampling, 10253
  - Hanurav-Vijayan selection method, 10227
  - initial seed, 10208
  - introductory example, 10180
  - joint selection probabilities, 10192
  - margin of error, 10235
  - maximum size measure, 10193
  - minimum size measure, 10198
  - missing values, 10219
  - Murthy's selection method, 10231
  - nested sorting, 10220
  - Neyman allocation, 10215, 10234
  - ODS table names, 10247
  - optimal allocation, 10215, 10233
  - output data sets, 10237, 10241
  - Poisson sampling, 10226
  - PPS sampling, with replacement, 10228
  - PPS sampling, without replacement, 10227
  - PPS sequential sampling, 10229
  - PPS systematic sampling, 10229
  - proportional allocation, 10216, 10233, 10256
  - replicated sampling, 10179, 10203, 10247
  - Sampford's selection method, 10232
  - sample output data set, 10237
  - sample selection methods, 10194, 10222
  - sample size, 10206
  - sample size allocation, 10179, 10215, 10232
  - sampling rate, 10205
  - sampling units, 10212
  - secondary input data set, 10236
  - sequential Poisson sampling, 10226
  - sequential random sampling, 10224, 10247
  - serpentine sorting, 10220
  - simple random sampling, 10181, 10222
  - size measures, 10213, 10250
  - sorting, 10220
  - strata, 10183, 10214
  - stratified sampling, 10179, 10183, 10214
  - systematic random sampling, 10186, 10223
  - unrestricted random sampling, 10223
  - with-replacement sampling, 10222
  - without-replacement sampling, 10222
- systematic random sampling
  - SURVEYSELECT procedure, 10186, 10223

- unrestricted random sampling
  - SURVEYSELECT procedure, 10223

- weighting, *see also* sampling weights
- weights, *see* sampling weights
- with-replacement sampling
  - SURVEYSELECT procedure, 10222
- without-replacement sampling
  - SURVEYSELECT procedure, 10222

Syntax Index

ALLOC= option
  STRATA statement (SURVEYSELECT), 10215
ALLOC=NEYMAN option
  STRATA statement (SURVEYSELECT), 10215
ALLOC=OPTIMAL option
  STRATA statement (SURVEYSELECT), 10215
ALLOC=PROPORTIONAL option
  STRATA statement (SURVEYSELECT), 10216
ALLOCMIN= option
  STRATA statement (SURVEYSELECT), 10217
ALPHA= option
  STRATA statement (SURVEYSELECT), 10217

CERTSIZE= option
  PROC SURVEYSELECT statement, 10189
CERTSIZE=P= option
  PROC SURVEYSELECT statement, 10190
CLUSTER statement
  SURVEYSELECT procedure, 10212
CONTROL statement
  SURVEYSELECT procedure, 10211
COST= option
  STRATA statement (SURVEYSELECT), 10217

DATA= option
  PROC SURVEYSELECT statement, 10192

FREQ statement
  SURVEYSELECT procedure, 10211

ID statement
  SURVEYSELECT procedure, 10212

JTPROBS option
  PROC SURVEYSELECT statement, 10192

MARGIN= option
  STRATA statement (SURVEYSELECT), 10218
MAXSIZE= option
  PROC SURVEYSELECT statement, 10193
METHOD= option
  PROC SURVEYSELECT statement, 10194
METHOD=BALBOOTSTRAP option
  PROC SURVEYSELECT statement, 10194
METHOD=BERNOULLI option
  PROC SURVEYSELECT statement, 10194
METHOD=CHROMY option
  PROC SURVEYSELECT statement, 10196, 10197
METHOD=POISSON option
  PROC SURVEYSELECT statement, 10195
METHOD=PPS option
  PROC SURVEYSELECT statement, 10195
METHOD=PPS_BREWER option
  PROC SURVEYSELECT statement, 10195
METHOD=PPS_MURTHY option
  PROC SURVEYSELECT statement, 10195
METHOD=PPS_SAMPFORD option
  PROC SURVEYSELECT statement, 10195
METHOD=PPS_SEQ option
  PROC SURVEYSELECT statement, 10196
METHOD=PPS_SYS option
  PROC SURVEYSELECT statement, 10196
METHOD=PPS_WR option
  PROC SURVEYSELECT statement, 10197
METHOD=SEQ option
  PROC SURVEYSELECT statement, 10197
METHOD=SEQ_POISSON option
  PROC SURVEYSELECT statement, 10197
METHOD=SRS option
  PROC SURVEYSELECT statement, 10197
METHOD=SYS option
  PROC SURVEYSELECT statement, 10197
METHOD=URS option
  PROC SURVEYSELECT statement, 10198
MINSIZE= option
  PROC SURVEYSELECT statement, 10198

NMAX= option
  PROC SURVEYSELECT statement, 10199
NMIN= option
  PROC SURVEYSELECT statement, 10200
NOPRINT option
  PROC SURVEYSELECT statement, 10200
NOSAMPLE option
  STRATA statement (SURVEYSELECT), 10218

OUT= option
  PROC SURVEYSELECT statement, 10200
OUTALL option
  PROC SURVEYSELECT statement, 10201
OUTHITS option
  PROC SURVEYSELECT statement, 10201
OUTSEED option
  PROC SURVEYSELECT statement, 10202
OUTSIZE option
  PROC SURVEYSELECT statement, 10202
OUTSORT= option

  PROC SURVEYSELECT statement, 10203

PPS option
  SAMPLINGUNIT statement (SURVEYSELECT), 10213
PRESORTED option
  SAMPLINGUNIT statement (SURVEYSELECT), 10213
PROC SURVEYSELECT statement, 10188, *see* SURVEYSELECT procedure

RANUNI option
  PROC SURVEYSELECT statement, 10203
REPS= option
  PROC SURVEYSELECT statement, 10203
ROUND= option
  PROC SURVEYSELECT statement, 10204

SAMPLINGUNIT statement
  SURVEYSELECT procedure, 10212
SAMPRATE= option
  PROC SURVEYSELECT statement, 10205
SAMPSIZE= option
  PROC SURVEYSELECT statement, 10206
SEED= option
  PROC SURVEYSELECT statement, 10208
SELECTALL option
  PROC SURVEYSELECT statement, 10209
SIZE statement
  SURVEYSELECT procedure, 10213
SORT= option
  PROC SURVEYSELECT statement, 10209
STATS option
  PROC SURVEYSELECT statement, 10209
  STRATA statement (SURVEYSELECT), 10218
STRATA statement
  SURVEYSELECT procedure, 10214
STRATUMSEED= option
  PROC SURVEYSELECT statement, 10210
SURVEYSELECT procedure
  syntax, 10187
SURVEYSELECT procedure, CLUSTER statement, 10212
SURVEYSELECT procedure, CONTROL statement, 10211
SURVEYSELECT procedure, FREQ statement, 10211
SURVEYSELECT procedure, ID statement, 10212
SURVEYSELECT procedure, PROC SURVEYSELECT statement, 10188
  CERTSIZE= option, 10189
  CERTSIZE=P= option, 10190
  DATA= option, 10192
  JTPROBS option, 10192
  MAXSIZE= option, 10193
  METHOD= option, 10194
  METHOD=BALBOOTSTRAP option, 10194
  METHOD=BERNOULLI option, 10194
  METHOD=CHROMY option, 10196, 10197
  METHOD=POISSON option, 10195
  METHOD=PPS option, 10195
  METHOD=PPS_BREWER option, 10195
  METHOD=PPS_MURTHY option, 10195
  METHOD=PPS_SAMPFORD option, 10195
  METHOD=PPS_SEQ option, 10196
  METHOD=PPS_SYS option, 10196
  METHOD=PPS_WR option, 10197
  METHOD=SEQ option, 10197
  METHOD=SEQ_POISSON option, 10197
  METHOD=SRS option, 10197
  METHOD=SYS option, 10197
  METHOD=URS option, 10198
  MINSIZE= option, 10198
  NMAX= option, 10199
  NMIN= option, 10200
  NOPRINT option, 10200
  OUT= option, 10200
  OUTALL option, 10201
  OUTHITS option, 10201
  OUTSEED option, 10202
  OUTSIZE option, 10202
  OUTSORT= option, 10203
  RANUNI option, 10203
  REPS= option, 10203
  ROUND= option, 10204
  SAMPRATE= option, 10205
  SAMPSIZE= option, 10206
  SEED= option, 10208
  SELECTALL option, 10209
  SORT= option, 10209
  STATS option, 10209
  STRATUMSEED= option, 10210
SURVEYSELECT procedure, SAMPLINGUNIT statement, 10212
  PPS option, 10213
  PRESORTED option, 10213
SURVEYSELECT procedure, SIZE statement, 10213
SURVEYSELECT procedure, STRATA statement, 10214
  ALLOC= option, 10215
  ALLOC=NEYMAN option, 10215
  ALLOC=OPTIMAL option, 10215
  ALLOC=PROPORTIONAL option, 10216
  ALLOCMIN= option, 10217
  ALPHA= option, 10217
  COST= option, 10217
  MARGIN= option, 10218
  NOSAMPLE option, 10218
  STATS option, 10218
  VAR= option, 10218, 10219

VAR= option
STRATA statement (SURVEYSELECT), 10218, 10219