



THE  
POWER  
TO KNOW.

# **SAS/STAT<sup>®</sup> 9.3 User's Guide**

## **The SURVEYSELECT Procedure**

### **(Chapter)**



This document is an individual chapter from *SAS/STAT® 9.3 User's Guide*.

The correct bibliographic citation for the complete manual is as follows: SAS Institute Inc. 2011. *SAS/STAT® 9.3 User's Guide*. Cary, NC: SAS Institute Inc.

Copyright © 2011, SAS Institute Inc., Cary, NC, USA

All rights reserved. Produced in the United States of America.

**For a Web download or e-book:** Your use of this publication shall be governed by the terms established by the vendor at the time you acquire this publication.

The scanning, uploading, and distribution of this book via the Internet or any other means without the permission of the publisher is illegal and punishable by law. Please purchase only authorized electronic editions and do not participate in or encourage electronic piracy of copyrighted materials. Your support of others' rights is appreciated.

**U.S. Government Restricted Rights Notice:** Use, duplication, or disclosure of this software and related documentation by the U.S. government is subject to the Agreement with SAS Institute and the restrictions set forth in FAR 52.227-19, Commercial Computer Software-Restricted Rights (June 1987).

SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513.

1st electronic book, July 2011

SAS® Publishing provides a complete selection of books and electronic products to help customers use SAS software to its fullest potential. For more information about our e-books, e-learning products, CDs, and hard-copy books, visit the SAS Publishing Web site at [support.sas.com/publishing](http://support.sas.com/publishing) or call 1-800-727-3228.

SAS® and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are registered trademarks or trademarks of their respective companies.

# Chapter 91

## The SURVEYSELECT Procedure

### Contents

---

Overview: SURVEYSELECT Procedure . . . . .	<b>7590</b>
Getting Started: SURVEYSELECT Procedure . . . . .	<b>7591</b>
Simple Random Sampling . . . . .	7592
Stratified Sampling . . . . .	7594
Stratified Sampling with Control Sorting . . . . .	7598
Syntax: SURVEYSELECT Procedure . . . . .	<b>7599</b>
PROC SURVEYSELECT Statement . . . . .	7599
CONTROL Statement . . . . .	7616
ID Statement . . . . .	7617
SAMPLINGUNIT   CLUSTER Statement . . . . .	7617
SIZE Statement . . . . .	7618
STRATA Statement . . . . .	7619
Details: SURVEYSELECT Procedure . . . . .	<b>7624</b>
Missing Values . . . . .	7624
Sorting by CONTROL Variables . . . . .	7625
Sample Selection Methods . . . . .	7626
Simple Random Sampling . . . . .	7627
Unrestricted Random Sampling . . . . .	7627
Systematic Random Sampling . . . . .	7627
Sequential Random Sampling . . . . .	7628
PPS Sampling without Replacement . . . . .	7629
PPS Sampling with Replacement . . . . .	7631
PPS Systematic Sampling . . . . .	7631
PPS Sequential Sampling . . . . .	7631
Brewer's PPS Method . . . . .	7633
Murthy's PPS Method . . . . .	7633
Sampford's PPS Method . . . . .	7634
Sample Size Allocation . . . . .	7634
Proportional Allocation . . . . .	7635
Optimal Allocation . . . . .	7635
Neyman Allocation . . . . .	7636
Specifying the Margin of Error . . . . .	7636
Secondary Input Data Set . . . . .	7638
Sample Output Data Set . . . . .	7639

Allocation Output Data Set . . . . .	7642
Displayed Output . . . . .	7643
ODS Table Names . . . . .	7646
Examples: SURVEYSELECT Procedure . . . . .	<b>7647</b>
Example 91.1: Replicated Sampling . . . . .	7647
Example 91.2: PPS Selection of Two Units per Stratum . . . . .	7650
Example 91.3: PPS (Dollar-Unit) Sampling . . . . .	7653
Example 91.4: Proportional Allocation . . . . .	7656
References . . . . .	<b>7659</b>

---

## 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 sample design that includes stratification, clustering, and unequal probabilities of selection. 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.

To select a sample with PROC SURVEYSELECT, you input a SAS data set that contains the sampling frame, which is 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 also specify the selection method, the desired 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. When you select a sample in multiple stages, you invoke the procedure separately for each stage of selection, inputting the frame and selection parameters for each current 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 in a stratum, has the same probability of being selected for the sample. In PPS sampling, a unit's selection probability is proportional to its size measure. For details about probability sampling methods, see Lohr (2010), Kish (1965, 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

This procedure also provides 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

The procedure uses fast, efficient algorithms for these sample selection methods. Thus, it performs well even for large input data sets or 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, with stratification you can ensure adequate sample sizes for subgroups of interest, including small subgroups, or you can use stratification toward improving the precision of the overall estimates. When you use a systematic or sequential selection method, PROC SURVEYSELECT can also 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, 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 replication to estimate standard errors for combined sample estimates and to perform a variety of other resampling and simulation tasks.

---

## Getting Started: SURVEYSELECT Procedure

In this example, an Internet service provider conducts a customer satisfaction survey. The survey population consists of the company's current subscribers. The company plans to select a sample of customers from this population, interview the selected customers, and then make inferences about the entire 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. It contains one observation for each customer, with a total of 13,471 observations.

The following PROC PRINT statements display the first 10 observations of the data set Customers and produce Figure 91.1:

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

**Figure 91.1** 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	

In the SAS data set Customers, the variable CustomerID uniquely identifies each customer. The variable State contains the state of the customer's address. The company has customers in four states: Georgia (GA), Alabama (AL), Florida (FL), and South Carolina (SC). The variable Type equals 'Old' if the customer has subscribed to the service for more than one year; otherwise, the variable Type equals 'New'. The variable Usage contains the customer's average monthly service usage, in minutes.

The following sections illustrate the use of PROC SURVEYSELECT for probability sampling with three different designs for the customer satisfaction survey. All three designs are one-stage, with customers as the sampling units. The first design is simple random sampling without stratification. In the second design, customers are stratified by state and type, and the sample is selected by simple random sampling within strata. In the third design, customers are sorted within strata by usage, and the sample is selected by systematic random sampling within strata.

---

## 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 unit has an equal probability of selection, and sampling is without replacement. Without-replacement sampling means that a unit cannot be selected more than once. The N=100 option specifies a sample size of 100 customers. The OUT= option stores the sample in the SAS data set named SampleSRS.

Figure 91.2 displays the output from PROC SURVEYSELECT, which summarizes the sample selection. A sample of 100 customers is selected from the data set Customers by simple random sampling. With simple random sampling and no stratification in the sample design, the selection probability is the same for all units in the sample. In this sample, the selection probability for each customer equals 0.007423, which is the sample size (100) divided by the population size (13,471). The sampling weight equals 134.71 for each customer in the sample, where the weight 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. (This information is always included in the output data set for more complex designs.)

The random number seed is 39647. PROC SURVEYSELECT uses this number as the initial seed for random number generation. Because the SEED= option is not specified in the PROC SURVEYSELECT statement, the seed value is obtained by using the time of day from the computer's clock. You can specify SEED=39647 to reproduce this sample.

**Figure 91.2** 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 91.3 displays the first 20 observations of the output data set SampleSRS, which contains the sample of customers. This data set includes all the variables from the DATA= 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 91.3** Customer Sample (First 20 Observations)

Customer Satisfaction Survey Sample of 100 Customers, Selected by SRS (First 20 Observations)					
Obs	CustomerID	State	Type	Usage	
1	036-89-0212	FL	New	74	
2	045-53-3676	AL	New	411	
3	050-99-2380	GA	Old	167	
4	066-93-5368	AL	Old	1232	
5	082-99-9234	FL	New	90	
6	097-17-4766	FL	Old	131	
7	110-73-1051	FL	Old	102	
8	111-91-6424	GA	New	247	
9	127-39-4594	GA	New	61	
10	162-50-3866	FL	New	100	
11	162-56-1370	FL	New	224	
12	167-21-6808	SC	New	60	
13	168-02-5189	AL	Old	7553	
14	174-07-8711	FL	New	284	
15	187-03-7510	SC	New	21	
16	190-78-5019	GA	New	185	
17	200-75-0054	GA	New	224	
18	201-14-1003	GA	Old	3437	
19	207-15-7701	GA	Old	24	
20	211-14-1373	AL	Old	88	

---

## Stratified Sampling

In this section, stratification is added to the sample design for the customer satisfaction survey. The sampling frame, which is the list of all customers, is stratified by State and Type. This divides the sampling frame into nonoverlapping subgroups formed from the values of the State and Type variables. 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 91.4 presents the 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.

**Figure 91.4** Stratification of Customers by State and Type

Customer Satisfaction Survey			
Strata of Customers			
The FREQ Procedure			
Table of State by Type			
State	Type		
Frequency			
Percent			
Row Pct			
Col Pct	New	Old	Total
AL	1238	706	1944
	9.19	5.24	14.43
	63.68	36.32	
	14.43	14.43	
FL	2170	1370	3540
	16.11	10.17	26.28
	61.30	38.70	
	25.29	28.01	
GA	3488	1940	5428
	25.89	14.40	40.29
	64.26	35.74	
	40.65	39.66	
SC	1684	875	2559
	12.50	6.50	19.00
	65.81	34.19	
	19.63	17.89	
Total	8580	4891	13471
	63.69	36.31	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. The N=15 option specifies a sample size of 15 customers for 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=1953 option specifies '1953' as the initial seed for random number generation.

Figure 91.5 displays the output from PROC SURVEYSELECT, which summarizes the sample selection. A total of 120 customers are selected.

**Figure 91.5** 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 91.6 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 in this design, the selection probability equals 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 is 0.021246. The variable SamplingWeight contains the sampling weights, which are computed as inverse selection probabilities.

**Figure 91.6** Customer Sample (First 30 Observations)

Customer Satisfaction Survey Sample Selected by Stratified Design (First 30 Observations)						
Obs	State	Type	CustomerID	Usage	Selection Prob	Sampling Weight
1	AL	New	002-26-1498	1189	0.012116	82.5333
2	AL	New	070-86-8494	106	0.012116	82.5333
3	AL	New	121-28-6895	76	0.012116	82.5333
4	AL	New	131-79-7630	265	0.012116	82.5333
5	AL	New	211-88-4991	108	0.012116	82.5333
6	AL	New	222-81-3742	83	0.012116	82.5333
7	AL	New	238-46-3776	278	0.012116	82.5333
8	AL	New	370-01-0671	123	0.012116	82.5333
9	AL	New	407-07-5479	1580	0.012116	82.5333
10	AL	New	550-90-3188	177	0.012116	82.5333
11	AL	New	582-40-9610	46	0.012116	82.5333
12	AL	New	672-59-9114	66	0.012116	82.5333
13	AL	New	848-60-3119	28	0.012116	82.5333
14	AL	New	886-83-4909	170	0.012116	82.5333
15	AL	New	993-31-7677	64	0.012116	82.5333
16	AL	Old	124-60-0495	80	0.021246	47.0667
17	AL	Old	128-54-9590	56	0.021246	47.0667
18	AL	Old	204-05-4017	17	0.021246	47.0667
19	AL	Old	210-68-8704	4363	0.021246	47.0667
20	AL	Old	239-75-4343	430	0.021246	47.0667
21	AL	Old	317-70-6496	452	0.021246	47.0667
22	AL	Old	365-37-1340	21	0.021246	47.0667
23	AL	Old	399-78-7900	108	0.021246	47.0667
24	AL	Old	404-90-6273	824	0.021246	47.0667
25	AL	Old	421-04-8548	1332	0.021246	47.0667
26	AL	Old	604-48-0587	16	0.021246	47.0667
27	AL	Old	774-04-0162	318	0.021246	47.0667
28	AL	Old	849-66-4156	79	0.021246	47.0667
29	AL	Old	937-69-9106	182	0.021246	47.0667
30	AL	Old	985-09-8691	24	0.021246	47.0667

## Stratified Sampling with Control Sorting

The next sample design for the customer satisfaction survey uses stratification by State and also control sorting by Type and Usage within State. After stratification and control sorting, customers are selected by systematic random sampling within strata. Selection by systematic sampling, together with control sorting before selection, spreads the sample uniformly over the range of type and usage values within each stratum (state). The following PROC SURVEYSELECT statements select a probability sample of customers from the Customers data set according to 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. The RATE=.02 option specifies a sampling rate of 2% for each stratum. The SEED=1234 option specifies the initial seed for random number generation.

Figure 91.7 displays the output from PROC SURVEYSELECT, which summarizes the sample selection. A sample of 271 customers is selected by using systematic random sampling within strata determined by State. The sampling frame Customers is sorted by control variables Type and Usage within strata. The type of sorting is serpentine, which is the default when SORT=NEST is not specified. See the section “[Sorting by CONTROL Variables](#)” on page 7625 for a description of serpentine sorting. The sorted data set replaces the input data set. (To leave the input data set unsorted and store the sorted input data in another data set, use the OUTSORT= option.) The output data set SampleControl contains the sample of customers.

**Figure 91.7** 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
Input Data Set	CUSTOMERS
Random Number Seed	1234
Stratum Sampling Rate	0.02
Number of Strata	4
Total Sample Size	271
Output Data Set	SAMPLECONTROL

---

## Syntax: SURVEYSELECT Procedure

The following statements are available in PROC SURVEYSELECT:

```

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

```

The **PROC SURVEYSELECT** statement invokes the procedure and optionally 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 rest of this section gives detailed syntax information about the **CONTROL**, **ID**, **SAMPLINGUNIT**, **SIZE**, and **STRATA** statements in alphabetical order after the description of the **PROC SURVEYSELECT** statement.

---

## PROC SURVEYSELECT Statement

```

PROC SURVEYSELECT options ;

```

The **PROC SURVEYSELECT** statement invokes the procedure and optionally 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<sub>n</sub>** 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 do specify a **SIZE** statement (or the **PPS** option), PROC SURVEYSELECT uses probability proportional to size selection without replacement (**METHOD=PPS**) by default. See the description of the **METHOD=** option for more information.

You must specify the sample size or sampling rate except when you request a method that selects two units from each stratum (**METHOD=PPS\_BREWER** or **METHOD=PPS\_MURTHY**) or when you specify the **MARGIN=** option in the **STRATA** statement for sample allocation. 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 can also provide stratum sampling rates, minimum size measures, maximum size measures, and certainty size measures in the secondary input data set. See the descriptions of the **SAMPSIZE=**, **SAMPRATE=**, **MINSIZE=**, **MAXSIZE=**, **CERTSIZE=**, and **CERTSIZE=P=** options for more information. You can name only one secondary input data set in each invocation of the procedure. See the section “Secondary Input Data Set” on page 7638 for details.

Table 91.1 lists the *options* available in the PROC SURVEYSELECT statement. Descriptions of the *options* follow in alphabetical order.

**Table 91.1** PROC SURVEYSELECT Statement Options

<b>Task</b>	<b>Options</b>
Specify the input data set	<b>DATA=</b>
Specify output data sets	<b>OUT=</b> <b>OUTSORT=</b>
Suppress displayed output	<b>NOPRINT</b>
Specify selection method	<b>METHOD=</b>
Specify sample size	<b>SAMPSIZE=</b> <b>SELECTALL</b>
Specify sampling rate	<b>SAMPRATE=</b> <b>NMIN=</b> <b>NMAX=</b>
Specify number of replicates	<b>REPS=</b>
Adjust size measures	<b>MINSIZE=</b> <b>MAXSIZE=</b>
Specify certainty size measures	<b>CERTSIZE=</b> <b>CERTSIZE=P=</b>
Specify type of sorting	<b>SORT=</b>
Specify random number seed	<b>SEED=</b>
Control OUT= contents	<b>JTPROBS</b> <b>OUTALL</b> <b>OUTHITS</b> <b>OUTSEED</b> <b>OUTSIZE</b> <b>STATS</b>

You can specify the following *options* in the PROC SURVEYSELECT statement:

### **CERTSIZE**

requests certainty selection, where the certainty size values are provided in the secondary input data set. Use the CERTSIZE option when you have already named the secondary data set in another option, such as the *SAMPsize=SAS-data-set* option. See the section “Secondary Input Data Set” on page 7638 for details.

The CERTSIZE option is available for *METHOD=PPS* and *METHOD=PPS\_SAMPFORD*. The CERTSIZE option is not available with the *SAMPLINGUNIT* statement.

In certainty selection, PROC SURVEYSELECT automatically selects all sampling units that have size measures greater than or equal to the stratum certainty size values. After identifying the certainty units, PROC SURVEYSELECT selects the remainder of the sample according to the method that is specified in the *METHOD=* option.

You provide the stratum certainty size values in the secondary input data set variable *\_CERTSIZE\_*. Each certainty size value must be a positive number. The variable *Certain* in the *OUT=* data set identifies the certainty selections, which have selection probabilities equal to 1.

If you want to specify a single certainty size value for all strata, you can use the *CERTSIZE=certain* option.

#### **CERTSIZE=*certain***

specifies the certainty size value, which must be a positive number. PROC SURVEYSELECT automatically selects all sampling units that have size measures greater than or equal to the value *certain*. After identifying the certainty units, PROC SURVEYSELECT selects the remainder of the sample according to the method that is specified in the *METHOD=* option.

The *CERTSIZE=* option is available for *METHOD=PPS* and *METHOD=PPS\_SAMPFORD*. The *CERTSIZE=* option is not available with the *SAMPLINGUNIT* statement.

The variable *Certain* in the *OUT=* data set identifies the certainty selections, which have selection probabilities equal to 1.

If you request a stratified sample design with the *STRATA* statement and specify the *CERTSIZE=certain* option, PROC SURVEYSELECT uses the value *certain* for all strata. If you do not want to use the same certainty size for all strata, use the *CERTSIZE=SAS-data-set* option to specify a certainty size value for each stratum.

#### **CERTSIZE=*SAS-data-set***

names a SAS data set that contains certainty size values for the strata. PROC SURVEYSELECT automatically selects all sampling units that have size measures greater than or equal to the stratum certainty size values. After identifying the certainty units, PROC SURVEYSELECT selects the remainder of the sample according to the method that is specified in the *METHOD=* option.

The *CERTSIZE=* option is available for *METHOD=PPS* and *METHOD=PPS\_SAMPFORD*. The *CERTSIZE=* option is not available with the *SAMPLINGUNIT* statement.

You provide the stratum certainty size values in the *CERTSIZE=* data set variable *\_CERTSIZE\_*. Each certainty size value must be a positive number. The variable *Certain* in the *OUT=* data set identifies the certainty selections, which have selection probabilities equal to 1.

The CERTSIZE= input data set should 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 CERTSIZE= data set as in the DATA= data set. The CERTSIZE= data set must include a variable named `_CERTSIZE_` that contains the certainty size value for each stratum. The CERTSIZE= data set is a secondary input data set. See the section “[Secondary Input Data Set](#)” on page 7638 for details. You can name only one secondary input data set in each invocation of the procedure.

If you want to specify a single certainty size value for all strata, you can use the `CERTSIZE=certain` option.

### **CERTSIZE=P**

requests certainty proportion selection, where the stratum certainty proportions are provided in the secondary input data set. Use the CERTSIZE=P option when you have already named the secondary data set in another option, such as the `SAMPsize=SAS-data-set` option. See the section “[Secondary Input Data Set](#)” on page 7638 for details.

The CERTSIZE=P option is available for `METHOD=PPS` and `METHOD=PPS_SAMPFORD`. The CERTSIZE=P option is not available with the `SAMPLINGUNIT` statement.

In certainty proportion selection, PROC SURVEYSELECT automatically selects all sampling units that have size measures greater than or equal to the stratum certainty proportion of the total stratum size. The procedure repeats this process with the remaining units until no more certainty units are selected. After identifying the certainty units, PROC SURVEYSELECT selects the remainder of the sample according to the method that is specified in the `METHOD=` option.

You provide the stratum certainty proportions in the secondary input data set variable `_CERTP_`. Each certainty proportion must be a positive number. You can specify a proportion value as a number between 0 and 1. Or you can specify a proportion value in percentage form as a number between 1 and 100, and PROC SURVEYSELECT converts that number to a proportion. The procedure treats the value 1 as 100% instead of 1%.

The variable `Certain` in the `OUT=` data set identifies the certainty selections, which have selection probabilities equal to 1.

If you want to specify a single certainty proportion for all strata, you can use the `CERTSIZE=P=p` option.

### **CERTSIZE=P=p**

specifies the certainty proportion. PROC SURVEYSELECT automatically selects all sampling units that have size measures greater than or equal to the proportion  $p$  of the total stratum size. The procedure repeats this process with the remaining units until no more certainty units are selected. After identifying the certainty units, PROC SURVEYSELECT selects the remainder of the sample according to the method that is specified in the `METHOD=` option.

The CERTSIZE=P= option is available for `METHOD=PPS` and `METHOD=PPS_SAMPFORD`. The CERTSIZE=P= option is not available with the `SAMPLINGUNIT` statement.

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

The variable `Certain` in the `OUT=` data set identifies the certainty selections, which have selection probabilities equal to 1.

If you request a stratified sample design with the `STRATA` statement and specify the `CERTSIZE=P= $\rho$`  option, PROC SURVEYSELECT uses the certainty proportion  $\rho$  for all strata. If you do not want to use the same certainty proportion for all strata, use the `CERTSIZE=P=SAS-data-set` option to specify a certainty proportion for each stratum.

#### **CERTSIZE=P=SAS-data-set**

names a SAS data set that contains certainty proportions for the strata. PROC SURVEYSELECT automatically selects all sampling units with size measures greater than or equal to the certainty proportion of the total stratum size. The procedure repeats this process with the remaining units until no more certainty units are selected. After identifying the certainty units, PROC SURVEYSELECT selects the remainder of the sample according to the method that is specified in the `METHOD=` option.

The `CERTSIZE=P=` option is available for `METHOD=PPS` and `METHOD=PPS_SAMPFORD`. The `CERTSIZE=P=` option is not available with the `SAMPLINGUNIT` statement.

You provide the stratum certainty proportions in the `CERTSIZE=P=` data set variable `_CERTP_`. Each certainty proportion must be a positive number. You can specify a proportion value as a number between 0 and 1. Or you can specify a proportion value in percentage form as a number between 1 and 100, and PROC SURVEYSELECT converts that number to a proportion. The procedure treats the value 1 as 100% instead of 1%.

The variable `Certain` in the `OUT=` data set identifies the certainty selections, which have selection probabilities equal to 1.

The `CERTSIZE=P=` input data set should 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 `CERTSIZE=P=` data set as in the `DATA=` data set. The `CERTSIZE=P=` data set must include a variable named `_CERTP_` that contains the certainty proportion for each stratum. The `CERTSIZE=P=` data set is a secondary input data set. See the section “[Secondary Input Data Set](#)” on page 7638 for details. You can name only one secondary input data set in each invocation of the procedure.

If you want to specify a single certainty proportion for all strata, you can use the `CERTSIZE=P= $\rho$`  option.

#### **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).

#### **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 details about computation of joint selection probabilities for a particular sampling method, see the method description in the section “[Sample Selection Methods](#)” on page 7626. For more information about the contents of the output data set, see the section “[Sample Output Data Set](#)” on page 7639.

## MAXSIZE

requests adjustment of size measures according to the stratum maximum size values provided in the secondary input data set. Use the MAXSIZE option when you have already named the secondary input data set in another option, such as the [SAMPSIZE=SAS-data-set](#) option. See the section “[Secondary Input Data Set](#)” on page 7638 for details.

The MAXSIZE option is available when you use size measures for any PPS selection method and also include a [STRATA](#) statement. You provide size measures by specifying the [SIZE](#) statement or the [PPS](#) option in the [SAMPLINGUNIT](#) statement.

You provide the stratum maximum size values in the secondary input data set variable `_MAXSIZE_`. Each maximum size value must be a positive number.

When a size measure exceeds the specified maximum value for its stratum, PROC SURVEYSELECT adjusts the size measure downward to equal the maximum size value. If your sampling units are individual observations, the variable `AdjustedSize` in the `OUT=` data set contains the adjusted size measures.

If you use a [SAMPLINGUNIT](#) statement to define sampling units (clusters), then the procedure applies the MAXSIZE adjustment to the sampling unit size. The sampling unit size equals the number of observations in the sampling unit if you specify the [PPS](#) option, or the sum of the observation size measures if you specify a [SIZE](#) statement. The output data set variable `UnitSize` contains the adjusted sampling unit size measures.

If you want to specify a single maximum size value for all strata, you can use the [MAXSIZE=max](#) option.

## MAXSIZE=max

specifies the maximum size value. The value of *max* must be a positive number.

When a size measure exceeds the value *max*, PROC SURVEYSELECT adjusts the size measure downward to equal *max*. If your sampling units are individual observations, the variable `AdjustedSize` in the `OUT=` data set contains the adjusted size measures.

If you use a [SAMPLINGUNIT](#) statement to define sampling units (clusters), then the procedure applies the MAXSIZE adjustment to the sampling unit size. The sampling unit size equals the number of observations in the sampling unit if you specify the [PPS](#) option, or the sum of the observation size measures if you specify a [SIZE](#) statement. The output data set variable `UnitSize` contains the adjusted sampling unit size measures.

The [MAXSIZE=max](#) option is available when you use size measures for any PPS selection method. You provide size measures by specifying the [SIZE](#) statement or the [PPS](#) option in the [SAMPLINGUNIT](#) statement.

If you request a stratified sample design with the [STRATA](#) statement and specify the [MAXSIZE=max](#) option, PROC SURVEYSELECT uses the maximum size *max* for all strata. If you do not want to use the same maximum size for all strata, use the [MAXSIZE=SAS-data-set](#) option to specify a maximum size value for each stratum.

**MAXSIZE=SAS-data-set**

names a SAS data set that contains maximum size values for the strata. You provide the stratum maximum size values in the MAXSIZE= data set variable `_MAXSIZE_`. Each maximum size value must be a positive number.

The MAXSIZE=SAS-data-set option is available when you use size measures for any PPS selection method and also include a STRATA statement. You provide size measures by specifying the SIZE statement or the PPS option in the SAMPLINGUNIT statement.

When a size measure exceeds the maximum size value for its stratum, PROC SURVEYSELECT adjusts the size measure downward to equal the maximum size value. If your sampling units are individual observations, the variable AdjustedSize in the OUT= data set contains the adjusted size measures.

If you use a SAMPLINGUNIT statement to define sampling units (clusters), then the procedure applies the MAXSIZE adjustment to the sampling unit size. The sampling unit size equals the number of observations in the sampling unit if you specify the PPS option, or the sum of the observation size measures if you specify a SIZE statement. The output data set variable UnitSize contains the adjusted sampling unit size measures.

The MAXSIZE= input data set should 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 MAXSIZE= data set as in the DATA= data set. The MAXSIZE= data set must include a variable named `_MAXSIZE_` that contains the maximum size value for each stratum. The MAXSIZE= data set is a secondary input data set. See the section “[Secondary Input Data Set](#)” on page 7638 for details. You can name only one secondary input data set in each invocation of the procedure.

If you want to specify a single maximum size value for all strata, you can use the MAXSIZE=max option.

**METHOD=name****M=name**

specifies the method for sample selection.

If you do not specify the METHOD= option, 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 do specify a SIZE statement (or the PPS option), PROC SURVEYSELECT uses probability proportional to size selection without replacement (METHOD=PPS) by default.

The following values are available for the METHOD= option:

**PPS**

requests selection with probability proportional to size and without replacement. See the section “[PPS Sampling without Replacement](#)” on page 7629 for details. If you specify METHOD=PPS, 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. See the section “[Brewer’s PPS Method](#)” on page 7633 for details. If you specify METHOD=PPS\_BREWER,

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 with 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. See the section "[Murthy's PPS Method](#)" on page 7633 for details. If you specify METHOD=PPS\_MURTHY, 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 with 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. See the section "[Sampford's PPS Method](#)" on page 7634 for details. If you specify METHOD=PPS\_SAMPFORD, 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. See the section "[PPS Sequential Sampling](#)" on page 7631 for details. If you specify METHOD=PPS\_SEQ, you must name a size measure variable in the SIZE statement or specify the PPS option in the SAMPLINGUNIT statement.

#### **PPS\_SYS**

requests systematic selection with probability proportional to size. See the section "[PPS Systematic Sampling](#)" on page 7631 for details. If you specify METHOD=PPS\_SYS, you must name a size measure variable in the SIZE statement or specify the PPS option in the SAMPLINGUNIT statement.

#### **PPS\_WR**

requests selection with probability proportional to size and with replacement. See the section "[PPS Sampling with Replacement](#)" on page 7631 for details. If you specify METHOD=PPS\_WR, you must name a size measure variable in the SIZE statement or specify the PPS option in the SAMPLINGUNIT statement.

#### **SEQ**

requests sequential selection according to Chromy's method. If you specify METHOD=SEQ 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. See the section "[Sequential Random Sampling](#)" on page 7628 for details.

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. See the section "[PPS Sequential Sampling](#)" on page 7631 for more information.

**SRS**

requests simple random sampling, which is selection with equal probability and without replacement. See the section “[Simple Random Sampling](#)” on page 7627 for details. METHOD=SRS is the default 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**

requests systematic random sampling. If you specify METHOD=SYS and do not specify a SIZE statement (or the PPS option in the SAMPLINGUNIT statement), PROC SURVEYSELECT uses systematic selection with equal probability. See the section “[Systematic Random Sampling](#)” on page 7627 for more information.

If you specify METHOD=SYS and also specify a SIZE statement (or the PPS option in the SAMPLINGUNIT statement), PROC SURVEYSELECT uses METHOD=PPS\_SYS, which is systematic selection with probability proportional to size. See the section “[PPS Systematic Sampling](#)” on page 7631 for details.

**URS**

requests unrestricted random sampling, which is selection with equal probability and with replacement. See the section “[Unrestricted Random Sampling](#)” on page 7627 for details.

**MINSIZE**

requests adjustment of size measures according to the stratum minimum size values provided in the secondary input data set. Use the MINSIZE option when you have already named the secondary input data set in another option, such as the *SAMPsize=SAS-data-set* option. See the section “[Secondary Input Data Set](#)” on page 7638 for details.

The MINSIZE option is available when you use size measures for any PPS selection method and also include a STRATA statement. You provide size measures by specifying the SIZE statement or the PPS option in the SAMPLINGUNIT statement.

You provide the stratum minimum size values in the secondary input data set variable `_MINSIZE_`. Each minimum size value must be a positive number.

When a size measure is less than the specified minimum value for its stratum, PROC SURVEYSELECT adjusts the size measure upward to equal the minimum size value. If your sampling units are individual observations, the variable AdjustedSize in the OUT= data set contains the adjusted size measures.

If you use a SAMPLINGUNIT statement to define sampling units (clusters), then the procedure applies the MINSIZE adjustment to the sampling unit size. The sampling unit size equals the number of observations in the sampling unit if you specify the PPS option, or the sum of the observation size measures if you specify a SIZE statement. The output data set variable UnitSize contains the adjusted sampling unit size measures.

If you want to specify a single minimum size value for all strata, you can use the *MINSIZE=min* option.

**MINSIZE=*min***

specifies the minimum size value. The value of *min* must be a positive number.

When a size measure is less than the value *min*, PROC SURVEYSELECT adjusts the size measure upward to equal *min*. If your sampling units are individual observations, the variable AdjustedSize in the OUT= data set contains the adjusted size measures.

If you use a **SAMPLINGUNIT** statement to define sampling units (clusters), then the procedure applies the MINSIZE adjustment to the sampling unit size. The sampling unit size equals the number of observations in the sampling unit if you specify the **PPS** option, or the sum of the observation size measures if you specify a **SIZE** statement. The output data set variable UnitSize contains the adjusted sampling unit size measures.

The MINSIZE=*min* option is available when you use size measures for any PPS selection method. You provide size measures by specifying the **SIZE** statement or the **PPS** option in the **SAMPLINGUNIT** statement.

If you request a stratified sample design with the **STRATA** statement and specify the MINSIZE=*min* option, PROC SURVEYSELECT uses the minimum size *min* for all strata. If you do not want to use the same minimum size for all strata, use the **MINSIZE=SAS-data-set** option to specify a minimum size value for each stratum.

**MINSIZE=SAS-data-set**

names a SAS data set that contains minimum size values for the strata. You provide the stratum minimum size values in the MINSIZE= data set variable `_MINSIZE_`. Each minimum size value must be a positive number.

The MINSIZE=SAS-data-set option is available when you use size measures for any PPS selection method and also include a **STRATA** statement. You provide size measures by specifying the **SIZE** statement or the **PPS** option in the **SAMPLINGUNIT** statement.

When a size measure is less than the minimum size value for its stratum, PROC SURVEYSELECT adjusts the size measure upward to equal the minimum size measure. If your sampling units are individual observations, the variable AdjustedSize in the OUT= data set contains the adjusted size measures.

If you use a **SAMPLINGUNIT** statement to define sampling units (clusters), then the procedure applies the MINSIZE adjustment to the sampling unit size. The sampling unit size equals the number of observations in the sampling unit if you specify the **PPS** option, or the sum of the observation size measures if you specify a **SIZE** statement. The output data set variable UnitSize contains the adjusted sampling unit size measures.

The MINSIZE= input data set should 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 MINSIZE= data set as in the DATA= data set. The MINSIZE= data set must include a variable named `_MINSIZE_` that contains the minimum size measure for each stratum. The MINSIZE= data set is a secondary input data set. See the section “**Secondary Input Data Set**” on page 7638 for details. You can name only one secondary input data set in each invocation of the procedure.

If you want to specify a single minimum size value for all strata, you can use the **MINSIZE=*min*** option.

**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***, then 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 can be used with equal probability selection methods (**METHOD=SRS**, **METHOD=URS**, **METHOD=SYS**, and **METHOD=SEQ**).

**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***, then 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 can be used with equal probability selection methods (**METHOD=SRS**, **METHOD=URS**, **METHOD=SYS**, and **METHOD=SEQ**).

**NOPRINT**

suppresses the display of all output. You can use the **NOPRINT** option when you want only to create an output data set. Note that this option temporarily disables the Output Delivery System (ODS). For more information, see Chapter 20, “Using the Output Delivery System.”

**OUT=SAS-*data-set***

names the output data set that contains the sample. 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, in addition to design information and selection statistics, depending on the selection method and output options that you request. See descriptions of the options **JTPROBS**, **OUTALL**, **OUTHITS**, **OUTSEED**, **OUTSIZE**, and **STATS**, which specify information to include in the output data set. See the section “Sample Output Data Set” on page 7639 for details about the contents of the output data set.

By default, the 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 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.

If you specify the **NOSAMPLE** option in the **STRATA** statement, PROC SURVEYFREQ 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. See the section “Allocation Output Data Set” on page 7642 for details.

**OUTALL**

includes all observations from the **DATA=** input data set 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 from the input data set and also contains a variable that indicates each observation's selection status. The variable **Selected** equals 1 for an observation that is selected for the sample, and equals 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 7639.

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

**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** equals  $m$ . For example, with the **OUTHITS** option a unit that is selected three times is represented by three copies in the output data set.

A sampling unit can be selected more than once by with-replacement and with-minimum-replacement selection methods, which include **METHOD=URS**, **METHOD=PPS\_WR**, **METHOD=PPS\_SYS**, and **METHOD=PPS\_SEQ**. The **OUTHITS** option is available for these selection methods.

See the section “[Sample Output Data Set](#)” on page 7639 for details about the contents of the output data set.

**OUTSEED**

includes the initial seed for each stratum in the **OUT=** output data set. The variable **InitialSeed** contains the stratum initial seeds. See the section “[Sample Output Data Set](#)” on page 7639 for details about the contents of the output data set.

To reproduce the same sample for any stratum in a subsequent execution of PROC SURVEYSELECT, you can specify the same stratum initial seed with the **SEED=SAS-data-set** option, along with the same sample selection parameters. See the section “[Sample Selection Methods](#)” on page 7626 for information about initial seeds and random number generation in PROC SURVEYSELECT.

The “[Sample Selection Summary](#)” table displays the initial random number seed for the entire sample selection, which is the same as 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. For PPS selection methods, the **OUTSIZE** option includes the total size measure in the output data set. If you do not specify size measures, or if you use a **SAMPLINGUNIT** statement, the **OUTSIZE** option includes the total number of sampling units.

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, the certainty proportion if you specify the **CERTSIZE=P=** option.

For more information about the contents of the output data set, see the section “[Sample Output Data Set](#)” on page 7639.

#### **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 with the **OUTSORT=** option, then the sorted data set replaces the input data set.

#### **REPS=nreps**

specifies the number of sample replicates. The value of *nreps* must be a positive integer.

When you specify the **REPS=** option, PROC SURVEYSELECT selects *nreps* independent samples, each with the same sample size or sampling rate and the same sample design that you request. The variable **Replicate** in the **OUT=** data set contains the sample replicate number.

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

#### **SAMPRATE=r**

##### **RATE=r**

specifies the sampling rate, which is the proportion of units to select for the sample. The sampling rate *r* must be a positive number. You can specify *r* as a number between 0 and 1. Or you can specify *r* in percentage form as a number between 1 and 100, and PROC SURVEYSELECT converts that number to a proportion. The procedure treats the value 1 as 100% instead of 1%.

The **SAMPRATE=** option is available only for equal probability selection methods (**METHOD=SRS**, **METHOD=URS**, **METHOD=SYS**, and **METHOD=SEQ**). For systematic random sampling (**METHOD=SYS**), PROC SURVEYSELECT uses the inverse of the sampling rate *r* as the interval. See the section “[Systematic Random Sampling](#)” on page 7627 for details. For other selection methods, PROC SURVEYSELECT converts the sampling rate *r* to the sample size before selection by multiplying the total number of units in the stratum or frame by the sampling rate and rounding up to the nearest integer.

If you request a stratified sample design with the **STRATA** statement and specify the **SAMPRATE=r** option, PROC SURVEYSELECT uses the sampling rate *r* for each stratum. If you do not want to use the same sampling rate for each stratum, use the **SAMPRATE=(values)** option or the **SAMPRATE=SAS-data-set** option to specify a sampling rate for each stratum.

**SAMPRATE=**(values)

**RATE=**(values)

specifies stratum sampling rates, where the stratum sampling rate is the proportion of units to select from the stratum. You can separate *values* with blanks or commas. The number of SAMPRATE= values must equal the number of strata in the input data set.

List the stratum sampling rate values in the order in which the strata appear in the input data set. When you use the SAMPRATE=(*values*) option, 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.

Each stratum sampling rate value must be a nonnegative. You can specify a rate value as a number between 0 and 1. Or you can specify a rate value in percentage form as a number between 1 and 100, and PROC SURVEYSELECT converts that number to a proportion. The procedure treats the value 1 as 100% instead of 1%.

To select a sample from a stratum, the value of the stratum sampling rate must be positive. If you specify a stratum sampling rate of 0, then 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.

The SAMPRATE= option is available only for equal probability selection methods (**METHOD=SRS**, **METHOD=URS**, **METHOD=SYS**, and **METHOD=SEQ**). For systematic random sampling (**METHOD=SYS**), PROC SURVEYSELECT uses the inverse of the stratum sampling rate as the interval for the stratum. See the section “[Systematic Random Sampling](#)” on page 7627 for details about systematic sampling. For other selection methods, PROC SURVEYSELECT converts the stratum sampling rate to a stratum sample size before selection by multiplying the total number of units in the stratum by the sampling rate and rounding up to the nearest integer.

**SAMPRATE=**SAS-data-set

**RATE=**SAS-data-set

names a SAS data set that contains stratum sampling rates, where the stratum sampling rate is the proportion of units to select from the stratum. The SAMPRATE= data set should include a variable `_RATE_` that contains the stratum sampling rates.

Each sampling rate value must be a nonnegative number. You can specify a rate value as a number between 0 and 1. Or you can specify a rate value in percentage form as a number between 1 and 100, and PROC SURVEYSELECT converts that number to a proportion. The procedure treats the value 1 as 100% instead of 1%.

To select a sample from a stratum, the value of the stratum sampling rate must be positive. If you specify a stratum sampling rate of 0, then 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.

The SAMPRATE= input data set should 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 SAMPRATE= data set as in the DATA= data set.

The SAMPRATE= option is available only for equal probability selection methods (**METHOD=SRS**, **METHOD=URS**, **METHOD=SYS**, and **METHOD=SEQ**). For systematic random sampling

(METHOD=SYS), PROC SURVEYSELECT uses the inverse of the stratum sampling rate as the interval for the stratum. See the section “[Systematic Random Sampling](#)” on page 7627 for details. For other selection methods, PROC SURVEYSELECT converts the stratum sampling rate to the stratum sample size before selection by multiplying the total number of units in the stratum by the sampling rate and rounding up to the nearest integer.

### **SAMPSIZE=*n***

#### **N=*n***

specifies the sample size, which is the number of units to select for the sample. The sample size *n* must be a positive integer. For selection methods that select without replacement, the sample size *n* must not exceed the number of units in the input data set.

If you do not specify a [SAMPLINGUNIT](#) statement, then your sampling units are observations, and PROC SURVEYSELECT selects *n* observations. If you use a [SAMPLINGUNIT](#) statement to define sampling units as groups of observations (clusters), then the procedure selects *n* clusters.

If you specify the [SAMPSIZE=\*n\*](#) option and request stratified selection with the [STRATA](#) statement, PROC SURVEYSELECT selects *n* units from each stratum unless you also specify the [ALLOC=](#) option in the [STRATA](#) statement to allocate the total sample size among the strata.

If you specify the [ALLOC=](#) option in the [STRATA](#) statement and the [SAMPSIZE=\*n\*](#) option, PROC SURVEYSELECT allocates the total sample size *n* among the strata according to the allocation method that you request. See the section “[Sample Size Allocation](#)” on page 7634 for details. If you specify the [MARGIN=](#) option with the [ALLOC=](#) option in the [STRATA](#) statement, PROC SURVEYSELECT determines the stratum sample sizes that provide the requested margin of error for the allocation. Therefore, you cannot use the [SAMPSIZE=](#) option with the [MARGIN=](#) option.

For methods that select without replacement, the sample size *n* must not exceed the number of units in any stratum. If you do not want to select the same number of units from each stratum, use the [SAMPSIZE=\(\*values\*\)](#) option or the [SAMPSIZE=SAS-data-set](#) option to specify a sample size for each stratum.

For without-replacement selection methods, by default, PROC SURVEYSELECT does not allow you to specify a stratum sample size that is greater than the total number of units available in the stratum. If you specify the [SELECTALL](#) option, PROC SURVEYSELECT selects all stratum units when the stratum sample size exceeds the number of units in the stratum.

### **SAMPSIZE=(*values*)**

#### **N=(*values*)**

specifies stratum sample sizes, where the stratum sample size is the number of units to select from the stratum. You can separate *values* with blanks or commas. The number of [SAMPSIZE=](#) values must equal the number of strata in the input data set.

List the stratum sample size values in the order in which the strata appear in the input data set. When you use the [SAMPSIZE=\(\*values\*\)](#) option, 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.

Each stratum sample size value must be a nonnegative integer. To select a sample from a stratum, the value of the stratum sample size must be positive. If you specify a stratum sample size of 0, then 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.

For without-replacement selection methods, by default, PROC SURVEYSELECT does not allow you to specify a stratum sample size that is greater than the total number of units available in the stratum. If you specify the [SELECTALL](#) option, PROC SURVEYSELECT selects all stratum units when the stratum sample size exceeds the number of units in the stratum.

### **SAMPSIZE=SAS-data-set**

#### **N=SAS-data-set**

names a SAS data set that contains stratum sample sizes, where the stratum sample size is the number of units to select from the stratum. The SAMPSIZE= input data set should include a variable named `_NSIZE_` or `SampleSize` that contains the stratum sample sizes.

Each stratum sample size value must be a nonnegative integer. To select a sample from a stratum, the value of the stratum sample size must be positive. If you specify a stratum sample size of 0, then 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.

The SAMPSIZE= input data set should 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 SAMPSIZE= data set as in the DATA= data set. The SAMPSIZE= data set is a secondary input data set. See the section “[Secondary Input Data Set](#)” on page 7638 for details. You can name only one secondary input data set in each invocation of the procedure.

For without-replacement selection methods, by default, PROC SURVEYSELECT does not allow you to specify a stratum sample size that is greater than the total number of units available in the stratum. If you specify the [SELECTALL](#) option, PROC SURVEYSELECT selects all stratum units when the stratum sample size exceeds the number of units in the stratum.

### **SEED**

indicates that stratum-level initial seeds are included in the secondary input data set. Use the SEED option when you have already named the secondary input data set in another option, such as the [SAMPSIZE=SAS-data-set](#) option. See the section “[Secondary Input Data Set](#)” on page 7638 for details. You can name only one secondary input data set in each invocation of the procedure.

You provide the stratum initial seeds in the secondary input data set variable named `_SEED_` or `InitialSeed`. The initial seeds must be positive integers.

See the description of the [SEED=SAS-data-set](#) option for more information about initial seeds for random number generation.

### **SEED=number**

specifies the initial seed for random number generation. The SEED= value must be a positive integer. If you do not specify the SEED= option, or if the SEED= value is negative or zero, PROC SURVEYSELECT uses the time of day from the computer’s clock to obtain the initial seed. See the section “[Sample Selection Methods](#)” on page 7626 for more information.

Whether or not you specify the SEED= option, PROC SURVEYSELECT displays the value of the initial seed in the “Sample Selection Summary” table. If you need to reproduce the same sample

in a subsequent execution of PROC SURVEYSELECT, you can specify this same seed value in the SEED= option, along with the same sample selection parameters, and PROC SURVEYSELECT will reproduce the sample.

If you request a stratified sample design with the STRATA statement, you can use the SEED=SAS-data-set option to specify an initial seed for each stratum. Otherwise, PROC SURVEYSELECT generates random numbers continuously across strata from the random number stream initialized by the SEED= value, as described in the section “Sample Selection Methods” on page 7626.

You can use the OUTSEED option to include the stratum initial seeds in the output data set.

#### **SEED=SAS-data-set**

names a SAS data set that contains initial seeds for the strata. You provide the stratum seeds in the SEED= input data set variable `_SEED_` or `InitialSeed`.

The initial seed values must be positive integers. If the initial seed value for the first stratum is not a positive integer, PROC SURVEYSELECT uses the time of day from the computer’s clock to obtain the initial seed. If the initial seed value for a subsequent stratum is not a positive integer, PROC SURVEYSELECT continues to use the random number stream already initialized by the seed for the previous stratum. See the section “Sample Selection Methods” on page 7626 for more information.

The SEED= input data set should 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 SEED= data set as in the DATA= data set. The SEED= data set is a secondary input data set. See the section “Secondary Input Data Set” on page 7638 for details. You can name only one secondary input data set in each invocation of the procedure.

You can use the OUTSEED option to include the stratum initial seeds in the output data set.

Whether or not you specify the SEED= option, PROC SURVEYSELECT displays the value of the initial seed in the “Sample Selection Summary” table. If you need to reproduce the same sample in a subsequent execution of PROC SURVEYSELECT, you can specify this same seed value in the SEED= option, along with the same sample selection parameters, and PROC SURVEYSELECT will reproduce the sample.

If you specify initial seeds by strata with the SEED=SAS-data-set option, you can reproduce the same sample in a subsequent execution of PROC SURVEYSELECT by specifying these same stratum initial seeds, along with the same 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 those 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 by CONTROL variables. The option SORT=NEST requests nested sorting, and SORT=SERP requests hierarchic serpentine sorting. The default is SORT=SERP. See the section “[Sorting by CONTROL Variables](#)” on page 7625 for descriptions of serpentine and nested sorting. 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.

The SORT= option and the CONTROL statement are not available with a SAMPLINGUNIT statement. See the descriptions of the CONTROL and SAMPLINGUNIT statements for more information.

When you specify a CONTROL statement, you can also 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.

**STATS**

includes selection probabilities and sampling weights in the OUT= output data set for equal probability selection methods when you do not specify a STRATA statement. This option is available for the following equal probability selection methods: METHOD=SRS, METHOD=URS, METHOD=SYS, and METHOD=SEQ. For PPS selection methods and stratified designs, the output data set contains selection probabilities and sampling weights by default. For more information about the contents of the output data set, see the section “[Sample Output Data Set](#)” on page 7639.

---

## CONTROL Statement

**CONTROL variables ;**

The CONTROL statement names 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 7625.

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.

---

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

You can use a **SAMPLINGUNIT** statement with any equal probability or PPS selection method. If you specify the **PPS** option in the **SAMPLINGUNIT** statement and do not specify a **SIZE** statement, then the procedure computes sampling unit size as the number of observations in the sampling unit. If you specify a **SIZE** statement with a **SAMPLINGUNIT** statement, then 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.

Note that 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* in the SAMPLINGUNIT statement after a slash (/):

### 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 with 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**. See the descriptions of these methods in the section “Sample Selection Methods” on page 7626 for more information.

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 Language Reference: Concepts*.

---

## 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 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 use 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. See the section “Missing Values” on page 7624 for more information.

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`, and `METHOD=PPS_WR`. For details about how size measures are used in sample selection, see the descriptions of PPS selection methods in the section “Sample Selection Methods” on page 7626.

Note that 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 with the `SAMPSIZE=` option in the `PROC SURVEYSELECT` statement.

---

## STRATA Statement

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

The `STRATA` statement names variables that partition the input data set into nonoverlapping subgroups (strata). The combinations of levels of `STRATA` variables define the strata. `PROC SURVEYSELECT` then selects independent samples from these strata, according to the selection method and design parameters that you specify in the `PROC SURVEYSELECT` statement. For information about the use of stratification in sample design, see Lohr (2010), Kalton (1983), Kish (1965, 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 the procedure 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. See the discussion of the `FORMAT` procedure in the *Base SAS Procedures Guide* and the discussions of the `FORMAT` statement and SAS formats in *SAS Language Reference: Dictionary*.

The `STRATA` variables function much like `BY` variables, and `PROC SURVEYSELECT` expects the input data set to be sorted in order of the `STRATA` variables.

If you specify a `CONTROL` statement, or if you specify `METHOD=PPS`, the input data set must be sorted in ascending order by the `STRATA` variables. This means you cannot use the `STRATA` option `NOTSORTED` or `DESCENDING` when you specify a `CONTROL` statement or `METHOD=PPS`.

If your input data set is not sorted by the `STRATA` variables in ascending order, 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 (when you do not specify a `CONTROL` statement or `METHOD=PPS`). 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 Language Reference: Concepts*. For more information about the DATASETS procedure, see the discussion in the *Base SAS Procedures Guide*.

## Allocation Options

The STRATA options request allocation of the total sample size among the strata. You can use the ALLOC= option to specify the allocation method. Available allocation methods include proportional allocation (ALLOC=PROP), optimal allocation (ALLOC=OPTIMAL), and Neyman allocation (ALLOC=NEYMAN). See the section “Sample Size Allocation” on page 7634 for details about these methods.

Instead of requesting that PROC SURVEYSELECT compute the sample allocation, you can provide the allocation proportions by using the ALLOC=(values) option or the ALLOC=SAS-data-set option. Then PROC SURVEYSELECT allocates the total sample size among the strata according to the proportions that you provide. Allocation proportions are relative stratum sample sizes,  $n_h/n$ , where  $n_h$  is the stratum  $h$  sample size and  $n$  is the total sample size.

You can use the SAMPSIZE= option in the PROC SURVEYSELECT statement to specify the total sample size to be allocated among the strata. Alternatively, you can specify the desired margin of error in the MARGIN= option, and the procedure determines the stratum sample sizes that are required to achieve that margin. See the section “Specifying the Margin of Error” on page 7636 for details.

When you request sample allocation, by default PROC SURVEYSELECT computes the allocation of the total sample size among the strata and then selects the sample. If you specify the NOSAMPLE option, the procedure computes the allocation but does not select the sample. In this case the OUT= output data set contains the stratum sample sizes that are computed according to the specified allocation method. See the section “Allocation Output Data Set” on page 7642 for details.

You can use the ALLOC= option with any selection method except METHOD=PPS\_BREWER and METHOD=PPS\_MURTHY, which select two units from each stratum.

Table 91.2 summarizes the options available in the STRATA statement. Descriptions of the options follow in alphabetical order.

**Table 91.2** STRATA Statement Options for Sample Allocation

Task	Options
Specify the allocation method	ALLOC=name
Provide allocation proportions	ALLOC=(values) ALLOC=SAS-data-set
Specify the margin of error	MARGIN= ALPHA=
Provide stratum costs and variances	COST= VAR=
Specify the minimum sample size per stratum	ALLOCMIN=
Allocate but do not select the sample	NOSAMPLE
Display additional allocation statistics	STATS

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

**ALLOC=*name***

specifies the method for allocating the total sample size among the strata. The following values for *name* are available:

**PROPORTIONAL | PROP**

requests proportional allocation, which allocates the total sample size in proportion to the stratum sizes, where the stratum size is the number of sampling units in the stratum. See the section “[Proportional Allocation](#)” on page 7635 for details.

**OPTIMAL | OPT**

requests optimal allocation, which allocates the total sample size among the strata in proportion to stratum sizes, stratum variances, and stratum costs. See the section “[Optimal Allocation](#)” on page 7635 for more information. If you specify ALLOC=OPTIMAL, you must provide the stratum variances with the [VAR=\(values\)](#), [VAR=SAS-data-set](#), or [VAR](#) option. You must provide the stratum costs with the [COST=\(values\)](#), [COST=SAS-data-set](#), or [COST](#) option.

**NEYMAN**

requests Neyman allocation, which allocates the total sample size among the strata in proportion to the stratum sizes and variances. See the section “[Neyman Allocation](#)” on page 7636 for more information. If you specify ALLOC=NEYMAN, you must provide the stratum variances with the [VAR=\(values\)](#), [VAR=SAS-data-set](#), or [VAR](#) option.

**ALLOC=(*values*)**

lists stratum allocation proportions. You can separate *values* with blanks or commas.

Each allocation proportion specifies the percent of the total sample size to allocate to the corresponding stratum. The number of ALLOC= values must equal the number of strata in the input data set. The sum of the allocation proportions must equal 1.

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

List the allocation proportions in the order in which the strata appear in the input data set. If you use the ALLOC=(*values*) option, 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 provide the stratum allocation proportions in the ALLOC= data set variable `_ALLOC_`.

Each allocation proportion specifies the percent of the total sample size to allocate to the corresponding stratum. The sum of the allocation proportions must equal 1.

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

The ALLOC= data set should contain all the STRATA variables, with the same type and length as in the DATA= input data set. The STRATA groups should appear in the same order in the ALLOC= data set as in the DATA= data set. The ALLOC= data set is a secondary input data set. See the section “[Secondary Input Data Set](#)” on page 7638 for details. You can name only one secondary data set in each invocation of the procedure.

**ALLOCMIN=*n***

specifies the minimum sample size to allocate to a stratum. When you specify ALLOCMIN=*n*, PROC SURVEYSELECT allocates at least *n* sampling units to each stratum. If you do not specify the ALLOCMIN= option, PROC SURVEYSELECT allocates at least one sampling unit to each stratum by default.

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

**ALPHA= $\alpha$** 

specifies the level of the confidence interval for the [MARGIN=](#) determination of stratum sample sizes. See the section “[Specifying the Margin of Error](#)” on page 7636 for details.

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

**COST**

indicates that stratum costs are included in the secondary input data set. Use the COST option when you have already named the secondary input data set in another option, such as the [VAR=SAS-data-set](#) option. You provide the stratum costs in the secondary input data set variable `_COST_`.

A stratum cost represents the per-unit cost (the survey cost of a single unit in the stratum). Each stratum cost must be a positive number. Cost values are required if you specify the [ALLOC=OPTIMAL](#) option.

**COST=(*values*)**

specifies stratum costs, which are required if you specify the [ALLOC=OPTIMAL](#) option. You can separate *values* with blanks or commas.

A stratum cost represents the per-unit cost (the survey cost of a single unit in the stratum). Each stratum cost must be a positive number.

The number of COST= values must equal the number of strata in the input data set. List the stratum costs in the order in which the strata appear in the input data set. If you use the COST=*values* option, 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 provide the stratum costs in the COST= data set variable `_COST_`.

A stratum cost represents the per-unit cost (the survey cost of a single unit in the stratum). Each stratum cost must be a positive number. Stratum costs are required if you specify the [ALLOC=OPTIMAL](#) option.

The `COST=` data set should contain all the STRATA variables, with the same type and length as in the `DATA=` input data set. The STRATA groups should appear in the same order in the `COST=` data set as in the `DATA=` data set. The `COST=` data set is a secondary input data set. See the section “[Secondary Input Data Set](#)” on page 7638 for details. You can name only one secondary input data set in each invocation of the procedure.

### **MARGIN=***value*

specifies the desired margin of error for estimating the overall mean from the stratified sample. When you specify the `MARGIN=` option, PROC SURVEYSELECT determines the stratum sample sizes required to achieve this margin for the allocation method or proportions that you specify in the `ALLOC=` option. See the section “[Specifying the Margin of Error](#)” on page 7636 for details.

The `MARGIN=` *value* must be a positive number. When you specify the `MARGIN=` option, you must also provide the stratum variances in the `VAR=(values)`, `VAR=SAS-data-set`, or `VAR` option.

You can use the `ALPHA=` option to set the level of the confidence interval that the `MARGIN=` computation uses. The default of `ALPHA=0.05` specifies a 95% confidence interval.

You can request the `MARGIN=` option for any allocation method (proportional, optimal, or Neyman) or for allocation proportions that you provide (`ALLOC=(values)` or `ALLOC=SAS-data-set`). When you use the `MARGIN=` option, you cannot specify a total sample size in the `SAMPSIZE=` option in the PROC SURVEYSELECT statement.

### **NOSAMPLE**

requests that PROC SURVEYSELECT allocate the total sample size among the strata but not select the sample. When you specify the `NOSAMPLE` option, the `OUT=` output data set contains the stratum sample sizes that PROC SURVEYSELECT computes. See the section “[Allocation Output Data Set](#)” on page 7642 for details.

### **STATS**

displays statistics for the sample allocation. If you specify the `MARGIN=` option, the `STATS` option displays the expected margin of error for the allocation. See the section “[Specifying the Margin of Error](#)” on page 7636 for details. If you request `ALLOC=OPTIMAL` or `ALLOC=NEYMAN` without 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. If you request `ALLOC=OPTIMAL`, the `STATS` option also displays the total stratum cost, which is computed from the stratum costs that you provide and the allocated stratum sample sizes.

### **VAR**

indicates that stratum variances are included in the secondary input data set. Use the `VAR` option when you have already named the secondary input data set in another option, such as the `COST=SAS-data-set` option. You provide the stratum variances in the secondary input data set variable `_VAR_`.

Each stratum variance must be a positive number. Stratum variances are required if you specify the `ALLOC=OPTIMAL`, `ALLOC=NEYMAN`, or `MARGIN=` option.

### **VAR=(values)**

lists stratum variances, which are required if you specify the `ALLOC=OPTIMAL`, `ALLOC=NEYMAN`, or `MARGIN=` option. You can separate *values* with blanks or commas.

Each stratum variance must be a positive number. The number of VAR= values must equal the number of strata in the input data set. List the stratum variances in the order in which the strata appear in the input data set. If you use the VAR=(values) option, 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 provide the stratum variances in the VAR= data set variable `_VAR_`.

Each stratum variance must be a positive number. Stratum variances are required if you specify the ALLOC=OPTIMAL, ALLOC=NEYMAN, or MARGIN= option.

The VAR= data set should contain all the STRATA variables, with the same type and length as in the DATA= input data set. The STRATA groups should appear in the same order in the VAR= data set as in the DATA= data set. The VAR= data set is a secondary input data set. See the section “[Secondary Input Data Set](#)” on page 7638 for details. You can name only one secondary input data set in each invocation of the procedure.

---

## 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 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 due to 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 `SAMP-SIZE=` 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_`. See the section “Secondary Input Data Set” on page 7638 for details.

---

## 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. See the chapter “The SORT Procedure” in the *Base SAS Procedures Guide* for more information. 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).

---

## 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. See Lohr (2010), Kish (1965, 1987), Kalton (1983), and Cochran (1977) for more information about probability sampling.

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, and sequential random 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. See Lohr (2010), Kalton (1983), Kish (1965), and the other references cited in the following sections for more information.

All the probability sampling methods provided by PROC SURVEYSELECT 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 pseudo-random 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. PROC SURVEYSELECT generates uniform random numbers according to the method of Fishman and Moore (1982), which uses a prime modulus multiplicative generator with modulus  $2^{31}$  and multiplier 397204094. PROC SURVEYSELECT uses the same uniform random number generator as the RANUNI function. For more information about the RANUNI function, see *SAS Language Reference: Dictionary*.

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  $h = 1, 2, \dots, H$ . 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  $h$ ,  $M_h$  is the total of all size measures for the population of stratum  $h$ , and  $Z_{hi} = M_{hi}/M_h$  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 equals  $n/N$ . When you request stratified sampling with a **STRATA** statement, PROC SURVEYSELECT selects samples independently within strata. The selection probability for a unit in stratum  $h$  equals  $n_h/N_h$  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. See Bentley and Floyd (1987) and Bentley and Knuth (1986) for details.

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 equals  $n/N$  when sampling without stratification. For stratified sampling, the expected number of hits for a unit in stratum  $h$  equals  $n_h/N_h$ . Note that 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** equals  $m$ . For example, with the **OUTHITS** option a unit that is selected three times is represented by three copies in the output data set. For information about the contents of the output data set, see the section "Sample Output Data Set" on page 7639.

## Systematic Random Sampling

The method of systematic random sampling (**METHOD=SYS**) selects units at a fixed interval throughout the sampling frame or stratum after a random start. If you specify the sample size (or the stratum sample sizes) with the **SAMPSIZE=** option, PROC SURVEYSELECT uses a fractional interval to provide exactly the specified sample size. The interval equals  $N/n$ , or  $N_h/n_h$  for stratified sampling. The selection probability for each unit equals  $n/N$ , or  $n_h/N_h$  for stratified sampling. If you specify the sampling rate (or the stratum sampling rates) with the **SAMPRATE=** option, PROC SURVEYSELECT uses the inverse of the rate as the interval for systematic selection. The selection probability for each unit equals the specified rate.

Systematic random sampling controls the distribution of the sample by spreading it throughout the sampling frame or stratum at equal intervals, thus providing implicit stratification. You can use the **CONTROL** statement to order 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. 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 option **METHOD=SEQ** 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. See Chromy (1979) and Williams and Chromy (1980) for details. See the section "PPS Sequential Sampling" on page 7631 for a description of Chromy's PPS selection method.

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 7625 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  $E(S_{hi})$  equals  $n_h/N_h$  for all units  $i$  in stratum  $h$ . The procedure computes

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

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

where  $\text{Int}(\cdot)$  denotes the integer part of the number, and  $\text{Frac}(\cdot)$  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  $T_{h(i-1)} = I_{h(i-1)}$ , Chomy'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  $T_{h(i-1)} = (I_{h(i-1)} + 1)$ , Chomy'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

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

## PPS Sampling without Replacement

If you specify the option `METHOD=PPS`, PROC SURVEYSELECT selects units with probability proportional to size and without replacement. The selection probability for unit  $i$  in stratum  $h$  equals  $n_h Z_{hi}$ , 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 equals  $M_{hi} / M_h$ , 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  $h$  ( $M_h$ ).

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  $Z_{hi} \leq 1/n_h$ , or equivalently,  $M_{hi} \leq M_h/n_h$ . 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 uses the Hanurav-Vijayan algorithm for PPS selection without replacement. 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. The 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. See Fox (1989), Golmant (1990), and Watts (1991) for details.

Notation in the remainder of this section drops the stratum subscript  $h$  for simplicity, but selection is still done independently within strata if you specify a stratified design. For a stratified design,  $n$  now denotes the sample size for the current stratum,  $N$  denotes the stratum population size, and  $M_i$  denotes the size measure for unit  $i$  in the stratum. If the design is not stratified, this notation applies to the entire sampling frame.

According to the Hanurav-Vijayan algorithm, PROC SURVEYSELECT first orders units within the stratum in ascending order by size measure, so that  $M_1 \leq M_2 \leq \dots \leq M_N$ . Then the procedure selects the PPS sample of  $n$  observations as follows:

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

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

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  $i$  is the integer selected in step 1, the procedure includes the last  $(n - i)$  units of the stratum in the sample, where the units are ordered by size measure as described previously. The procedure then selects the remaining  $i$  units according to steps 3 through 6.
3. The procedure defines new normed size measures for the remaining  $(N - n + i)$  stratum units that were not selected in steps 1 and 2:

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

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

$$\begin{aligned} a_1(1) &= iZ_1^* \\ a_j(1) &= iZ_j^* \prod_{k=1}^{j-1} (1 - (i-1)P_k) \quad \text{for } j = 2, \dots, N - n + 1 \end{aligned}$$

and

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

5. If stratum unit  $j_1$  is the unit selected in step 4, then 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

$$\begin{aligned} a_{j_1+1}(2, j_1) &= (i-1)Z_{j_1+1}^* \\ a_j(2, j_1) &= (i-1)Z_j^* \prod_{k=j_1+1}^{j-1} (1 - (i-2)P_k) \quad \text{for } j = j_1 + 2, \dots, N - n + 2 \end{aligned}$$

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  $j$  in the stratum equals

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

where

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

$$\pi_{ij}^{(r)} = \frac{r(r-1)}{2} P_i Z_j \prod_{k=1}^{i-1} (1 - P_k)$$

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

## PPS Sampling with Replacement

If you specify the option `METHOD=PPS_WR`, 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  $Z_{hi} = M_{hi}/M_h$ . 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  $h$  equals  $n_h Z_{hi}$ . 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  $h$  equals

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

## PPS Systematic Sampling

If you specify the option `METHOD=PPS_SYS`, PROC SURVEYSELECT selects units by systematic random sampling with probability proportional to size. Systematic sampling selects units at a fixed interval throughout the stratum or sampling frame after a random start. PROC SURVEYSELECT uses a fractional interval to provide exactly the specified sample size. The interval equals  $M_h/n_h$  for stratified sampling and  $M/n$  for sampling without stratification. Depending on the sample size and the values of the size measures, it might be possible for a unit to be selected more than once. The expected number of hits (selections) for unit  $i$  in stratum  $h$  equals  $n_h M_{hi}/M_h = n_h Z_{hi}$ . See Cochran (1977, pp. 265–266) and Madow (1949) for details.

Systematic random sampling controls the distribution of the sample by spreading it throughout the sampling frame or stratum at equal intervals, thus providing implicit stratification. You can use the `CONTROL` statement to order 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. 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 option `METHOD=PPS_SEQ`, PROC SURVEYSELECT uses Chromy's method of sequential random sampling. See Chromy (1979) and Williams and Chromy (1980) for details. 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  $0, 1, \dots, n_h$ , 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 7625 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

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

$$I_{hi} = \text{Int}\left(\sum_{j=1}^i E(S_{hj})\right)$$

$$F_{hi} = \text{Frac}\left(\sum_{j=1}^i E(S_{hj})\right)$$

where  $E(S_{hi})$  represents the expected number of hits for unit  $i$  in stratum  $h$ ,  $\text{Int}(\cdot)$  denotes the integer part of the number, and  $\text{Frac}(\cdot)$  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  $i$  equals  $T_{hi} - T_{h(i-1)}$ .

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  $h$  equals  $2M_{hi}/M_h = 2Z_{hi}$ . (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  $h$  equals

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

See Cochran (1977, pp. 261–263) and Brewer (1963) for details. Brewer's method yields the same selection probabilities and joint selection probabilities as Durbin's method. See Cochran (1977) and Durbin (1967) for details.

## 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  $h$  equals

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

where  $Z_{hi} = M_{hi}/M_h$  and

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

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  $h$  equals

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

See Cochran (1977, pp. 263–265) and Murthy (1957) for details.

## 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  $h$  equals  $n_h M_{hi} / M_h = n_h Z_{hi}$ . (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  $h$  equals

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

where

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

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

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

---

## 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=PROP`), optimal allocation (`ALLOC=OPTIMAL`), and Neyman allocation (`ALLOC=NEYMAN`). See Lohr (2010), Kish (1965), and Cochran (1977) for more information about these allocation methods. 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 desired 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. See the section “Specifying the Margin of Error” on page 7636 for details.

## Proportional Allocation

When you specify the **ALLOC=PROP** 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  $h$  equals

$$f_h^* = N_h/N$$

where  $N_h$  is the number of sampling units in stratum  $h$  and  $N$  is the total number of sampling units for all strata. If you specify the total sample size  $n$  in the **SAMPsize=** option in the PROC SURVEYSELECT statement, the procedure computes the target sample size for stratum  $h$  as

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

The target sample size values,  $n_h^*$ , might not be integers, but the stratum sample sizes are required to be integers. PROC SURVEYSELECT uses a rounding algorithm to convert the  $n_h^*$  to integer values  $n_h$  and maintain the requested total sample size  $n$ . The rounding algorithm includes the restriction that all values of  $n_h$  must be at least 1, so that at least one unit is selected from each stratum. If you specify a minimum stratum sample size  $n_{min}$  in the **ALLOCmin=** option in the **STRATA** statement, then all values of  $n_h$  are required to be at least  $n_{min}$ . For without-replacement selection methods, PROC SURVEYSELECT also requires that each stratum sample size must not exceed the total number of sampling units in the stratum,  $n_h \leq N_h$ . If a target stratum sample size exceeds the number of units in the stratum, PROC SURVEYSELECT allocates the maximum number of units,  $N_h$ , to the stratum, and then allocates the remaining total sample size proportionally among the remaining strata.

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  $f_h^*$  due to rounding, the requirement that  $n_h \geq 1$  (or  $n_h \geq n_{min}$ ), and the requirement that  $n_h \leq N_h$  for without-replacement selection methods.

## 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 by using 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. See Lohr (2010), Cochran (1977), and Kish (1965) for details. For optimal

allocation, PROC SURVEYSELECT computes the proportion of the total sample size for stratum  $h$  as

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

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.

If you specify the total sample size  $n$  in the **SAMPSIZE=** option in the **PROC SURVEYSELECT** statement, the procedure computes the target sample size for stratum  $h$  as

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

As described in the section “**Proportional Allocation**” on page 7635, the values of  $n_h^*$  are converted to integer sample sizes  $n_h$  by using a rounding algorithm that requires the sum of the stratum sample sizes to equal  $n$ . The final stratum sample sizes  $n_h$  are also required to be at least 1, or at least  $n_{min}$  if you specify a minimum stratum sample size in the **ALLOCMIN=** option in the **STRATA** statement. For without-replacement selection methods, the final sample sizes cannot exceed the stratum sizes.

## 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 (described in the section “**Optimal Allocation**” on page 7635), where the costs per unit are the same for all strata. For Neyman allocation, the proportion of the total sample size for stratum  $h$  is computed as

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

If you specify the total sample size  $n$  in the **SAMPSIZE=** option in the **PROC SURVEYSELECT** statement, the procedure computes the target sample size for stratum  $h$  as  $n_h^* = f_h^* \times n$ . The  $n_h^*$  are converted to integer sample sizes  $n_h$  by using a rounding algorithm that requires the sum of the stratum sizes to equal  $n$ . The final sample sizes  $n_h$  are required to be at least 1, or at least  $n_{min}$  if you specify a minimum sample size in the **ALLOCMIN=** option in the **STRATA** statement. For without-replacement selection methods, the final sample sizes must not exceed the stratum sizes.

## Specifying the Margin of Error

Instead of specifying the total sample size to allocate among the strata, you can specify the desired margin of error for estimating 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,

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

where  $\text{Var}(\bar{y}_{str})$  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  $\alpha$  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

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

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. See the sections “Proportional Allocation” on page 7635, “Optimal Allocation” on page 7635, and “Neyman Allocation” on page 7636 for details.

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

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

See Lohr (2010, page 91), Cochran (1977, Chapter 5), and Arkin (1984, Chapter 10) for more information.

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. See Cochran (1977, page 104) and Arkin (1984, page 176) for details.

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

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

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

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

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 the procedure.

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

**Table 91.3** PROC SURVEYSELECT Secondary Data Set Variables

Variable	Description	Statement	Option
<code>_ALLOC_</code>	Allocation proportion	STRATA	ALLOC=
<code>_CERTP_</code>	Certainty proportion	PROC	CERTSIZE=P=
<code>_CERTSIZE_</code>	Certainty size	PROC	CERTSIZE=
<code>_COST_</code>	Cost	STRATA	COST=
<code>_MAXSIZE_</code>	Maximum size	PROC	MAXSIZE=
<code>_MINSIZE_</code>	Minimum size	PROC	MINSIZE=
<code>_NSIZE_</code>	Sample size	PROC	SAMPSIZE=
<code>_RATE_</code>	Sampling rate	PROC	SAMPRATE=
<code>_SEED_</code>	Random number seed	PROC	SEED=
<code>_VAR_</code>	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. If you specify the **NOSAMPLE** option, PROC SURVEYSELECT allocates the total sample size among the strata but does not select the sample. When you specify the **NOSAMPLE** option, the output data set contains the allocated sample sizes. See the section “Allocation Output Data Set” on page 7642 for details.

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. But 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. The variable `Selected` equals 1 for an observation selected for the sample, and equals 0 for an observation not selected. 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 selected unit in the output data set. For example, with the **OUTHITS** option a unit that is selected three times is represented by three copies in the output data set.

The output data set also contains design information and selection statistics, depending on the selection method and output options 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. Selected equals 1 for an observation that is selected for the sample, or 0 for an observation that is not selected.
- STRATA variables, which you specify in the **STRATA** statement.
- Replicate, which is the sample replicate number. This variable is included when you request replicated sampling with the **REPS=** 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. Certain equals 1 for units that are included with certainty because their size measures exceed the certainty size value or the certainty proportion; otherwise, Certain equals 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=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. This variable equals the inverse of ExpectedHits or SelectionProb.

For `METHOD=PPS_BREWER` and `METHOD=PPS_MURTHY`, which select 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 contains the overall values.

- `MinimumSize`, which is the minimum size measure specified with the `MINSIZE=` option. This variable is included if you specify the `MINSIZE=` option.
- `MaximumSize`, which is the maximum size measure specified with the `MAXSIZE=` option. This variable is included if you specify the `MAXSIZE=` option.
- `CertaintySize`, which is the certainty size measure specified with the `CERTSIZE=` option. This variable is included if you specify the `CERTSIZE=` option.
- `CertaintyProp`, which is the certainty proportion specified with 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 there is no `SIZE` statement, or if you specify a `SAMPLINGUNIT` statement.
- `TotalSize`, which is the total of size measures in the stratum. This variable is included if there is a `SIZE` statement, or if you specify 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 with 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`, which selects two units from each stratum.

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 specified allocation method.
- 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 due to rounding and other restrictions. See the section “Sample Size Allocation” on page 7634 for details.

---

## 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 specified allocation method.
- 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 due to rounding and other restrictions. See the section “[Sample Size Allocation](#)” on page 7634 for details.

---

## 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.

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. (See the sections “[Sample Output Data Set](#)” on page 7639 and “[Allocation Output Data Set](#)” on page 7642 for information about these output data sets.) 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.

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

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=r` option
- 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 with the `SAMPRATE=` option
- Maximum Sample Size or Stratum Maximum Sample Size, if you specify the `NMAX=` option with the `SAMPRATE=` 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
- Sampling Weight, if you specify an equal probability selection method (`METHOD=SRS`, `METHOD=URS`, `METHOD=SYS`, or `METHOD=SEQ`) 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

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
- Selection Method, if you specify the `METHOD=` option

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

---

## 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 20, “Using the Output Delivery System.” Table 91.4 lists the table names.

**Table 91.4** ODS Tables Produced by PROC SURVEYSELECT

ODS Table Name	Description	Statement	Option
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 91.1: Replicated Sampling

This example uses the Customers data set from the section “Getting Started: SURVEYSELECT Procedure” on page 7591. 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. See Lohr (2010), Wolter (2007), Kish (1965, 1987), and Kalton (1983) for information about replicated sampling.

This design includes four replicates, each with 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 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=4 option specifies four replicates of this sample. The N=(8 12 20 10) option lists the stratum sample sizes for each replicate. The N= option lists the stratum sample sizes in the same order as the strata appear in the Customers data set, which has been sorted by State. The sample size of eight customers corresponds to the first stratum, State = ‘AL’. The sample size 12 corresponds to the next stratum, State = ‘FL’, and so on. The SEED=40070 option specifies ‘40070’ as the initial seed for random number generation.

Output 91.1.1 displays the output from PROC SURVEYSELECT, which summarizes the sample selection. A total of 200 customers is selected in four replicates. PROC SURVEYSELECT selects each replicate by using sequential random sampling within strata 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 91.1.1** 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', from 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 91.1.2** displays the 32 sample customers of 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 91.1.2** Customer Sample (First Stratum)

Customer Satisfaction Survey Sample Selected by Replicated Design (First Stratum)							
Obs	State	Replicate	CustomerID	Type	Usage	Selection Prob	Sampling Weight
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 91.2: 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 desired sampling rates for each service. This size measure reflects the amount of relevant survey information expected from the hospital. See Drummond et al. (1982) for details about this type of size measure. The variable Size equals '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 91.2.1](#):

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

**Output 91.2.1** Sampling Frame

Hospital Utilization Survey Sampling Frame, Region 1				
Obs	Hospital	Type	Size Measure	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=48702 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 with the N= option, because Brewer's method always selects two units from each stratum.

Output 91.2.2 displays the output from PROC SURVEYSELECT. A total of eight hospitals were selected from the four strata. The data set SampleHospitals contains the selected hospitals.

### Output 91.2.2 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 91.2.3:

```

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

```

### Output 91.2.3 Sample Hospitals

Hospital Utilization Survey Sample Selected by Stratified PPS Design							
Obs	Type	Size	Hospital	Size Measure	Selection Prob	Sampling Weight	Jt Selection Prob
1	Rural	Small	079	2.127	0.13516	7.39868	0.01851
2	Rural	Small	236	5.279	0.33545	2.98106	0.01851
3	Urban	Small	006	4.249	0.17600	5.68181	0.01454
4	Urban	Small	195	5.024	0.20810	4.80533	0.01454
5	Urban	Medium	133	46.992	0.41357	2.41795	0.11305
6	Urban	Medium	218	48.231	0.42448	2.35584	0.11305
7	Urban	Large	026	61.460	0.63445	1.57617	0.31505
8	Urban	Large	058	65.931	0.68060	1.46929	0.31505

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 91.3: 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. See Wilburn (1984) for details.

`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 91.3.1** displays the sampling frame data set `TravelExpense`, which contains 41 observations.

**Output 91.3.1** 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, 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 4 corresponds to the last stratum, Level = '3\_High'. The SEED=47279 option specifies '47279' as the initial seed for random number generation.

Output 91.3.2 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 91.3.2 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 91.3.3:

```

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

```

**Output 91.3.3** Audit Sample

Travel Expense Audit					
Sample Selected by Stratified PPS Design					
Obs	Level	ID	Amount	Selection Prob	Sampling Weight
1	1_Low	654	185.60	0.31105	3.21489
2	1_Low	017	205.48	0.34437	2.90385
3	1_Low	382	217.85	0.36510	2.73896
4	1_Low	614	230.56	0.38640	2.58797
5	1_Low	782	258.10	0.43256	2.31183
6	1_Low	775	330.54	0.55396	1.80518
7	2_Avg	784	505.14	0.34623	2.88823
8	2_Avg	332	540.65	0.37057	2.69853
9	2_Avg	002	567.89	0.38924	2.56909
10	2_Avg	239	620.10	0.42503	2.35278
11	2_Avg	738	670.85	0.45981	2.17479
12	2_Avg	496	753.30	0.51633	1.93676
13	2_Avg	425	780.10	0.53470	1.87022
14	2_Avg	478	806.90	0.55307	1.80810
15	2_Avg	672	979.66	0.67148	1.48925
16	2_Avg	139	1183.45	0.81116	1.23280
17	3_High	568	1670.80	0.64385	1.55316
18	3_High	263	1893.40	0.72963	1.37056
19	3_High	285	2020.70	0.77869	1.28421
20	3_High	486	2580.35	0.99435	1.00568

**Example 91.4: Proportional Allocation**

This example uses the Customers data set from the section “Getting Started: SURVEYSELECT Procedure” on page 7591. 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 7591 gives an example of stratified sampling, where the list of customers is stratified by State and Type. Figure 91.4 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=15 option specifies a sample size of 15 customers for 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=1000 option specifies a total sample size of 1000 customers to be allocated among the strata.

Output 91.4.1 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 91.4.1 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 91.4.2:

```

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

```

**Output 91.4.2** Stratum Sample Sizes

Customer Satisfaction Survey Proportional Allocation						
Obs	State	Type	Total	Alloc Proportion	Sample Size	Actual Proportion
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 equals 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, while 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. Due to rounding and other restrictions, the actual proportions can differ from the target allocation proportions. See the section “[Sample Size Allocation](#)” on page 7634 for details.

If you want 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*, Third Edition, New York: McGraw-Hill.
- Bentley, J. L. and Floyd, R. (1987), "A Sample of Brilliance," *Communications of the Association for Computing Machinery*, 30, 754–757.
- Bentley, J. L. and Knuth, D. (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," *Proceedings of the SAS Global Forum 2007 Conference*, Cary, NC: SAS Institute Inc.
- Chromy, J. R. (1979), "Sequential Sample Selection Methods," *Proceedings of the American Statistical Association, Survey Research Methods Section*, 401–406.
- Cochran, W. G. (1977), *Sampling Techniques*, Third Edition, New York: John Wiley & Sons.
- 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," *Proceedings of the Fourth Conference on Health Survey Research Methods*, DHHS Publication No. (PHS) 84-3346, Washington, DC: National Center for Health Services Research, 233–248.
- Durbin, J. (1967), "Design of Multi-stage Surveys for the Estimation of Sampling Errors," *Applied Statistics*, 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," *The American Statistician*, 43(3), 168–171.
- Golmant, J. (1990), "Correction: Computer Selection of Size-Biased Samples," *The American Statistician*, 44(2), 194.
- Hanurav, T. V. (1967), "Optimum Utilization of Auxiliary Information:  $\pi_{pS}$  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*, Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07-035, Beverly Hills, CA and London: 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*, Second Edition, Pacific Grove, CA: Duxbury Press.
- Madow, W. G. (1949), "On the Theory of Systematic Sampling, II," *Annals of Mathematical Statistics*, 20, 333–354.
- McLeod, A. I. and Bellhouse, D. R. (1983), "A Convenient Algorithm for Drawing a Simple Random Sample," *Applied Statistics*, 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.
- Sampford, M. R. (1967), "On Sampling without Replacement with Unequal Probabilities of Selection," *Biometrika*, 54, 499–513.
- Vijayan, K. (1968), "An Exact  $\pi_{ps}$  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," *The American Statistician*, 45(2), 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," *Proceedings of the Fifth Annual SAS Users Group International Conference*, 5, 392–396.
- Wolter, K. M. (2007), *Introduction to Variance Estimation*, Second Edition, New York: Springer-Verlag.

# Subject Index

## A

### allocation

of sample size (SURVEYSELECT), 7591, 7620, 7634

## B

### Brewer's selection method

SURVEYSELECT procedure, 7633, 7650

## C

### Chromy's selection method

SURVEYSELECT procedure, 7628, 7631

### cluster sampling

SURVEYSELECT procedure, 7617, 7626

### clusters

SURVEYSELECT procedure, 7590, 7617

### control sorting

SURVEYSELECT procedure, 7598, 7616, 7625, 7647

## D

### dollar-unit sampling

SURVEYSELECT procedure, 7653

## H

### Hanurav-Vijayan selection method

SURVEYSELECT procedure, 7629

## I

### initial seed

SURVEYSELECT procedure, 7614

## J

### joint selection probabilities

SURVEYSELECT procedure, 7603

## M

### margin of error

SURVEYSELECT procedure, 7636

### missing values

SURVEYSELECT procedure, 7624

### multistage sampling

SURVEYSELECT procedure, 7590

### Murthy's selection method

SURVEYSELECT procedure, 7633

## N

### Neyman allocation

SURVEYSELECT procedure, 7621, 7636

## O

### optimal allocation

SURVEYSELECT procedure, 7621, 7635

## P

### population

SURVEYSELECT procedure, 7590

### PPS sampling

SURVEYSELECT procedure, 7590, 7618, 7626

### PPS sampling, with replacement

SURVEYSELECT procedure, 7631

### PPS sampling, without replacement

SURVEYSELECT procedure, 7629

### PPS sequential sampling

SURVEYSELECT procedure, 7631

### PPS systematic sampling

SURVEYSELECT procedure, 7631

### probability sampling

SURVEYSELECT procedure, 7590

### proportional allocation

SURVEYSELECT procedure, 7621, 7635, 7656

## R

### random sampling

SURVEYSELECT procedure, 7590

### replicated sampling

SURVEYSELECT procedure, 7591, 7611, 7647

replication, *see* replicated sampling

## S

### Sampford's selection method

SURVEYSELECT procedure, 7634

### sample

- SURVEYSELECT procedure, 7590
- sample allocation
  - SURVEYSELECT procedure, 7591, 7620, 7634
- sample design
  - SURVEYSELECT procedure, 7590
- sample selection
  - SURVEYSELECT procedure, 7590
- sample selection methods
  - SURVEYSELECT procedure, 7605, 7626
- sample size
  - SURVEYSELECT procedure, 7613
- sample size allocation
  - SURVEYSELECT procedure, 7591, 7620, 7634
- sampling
  - SURVEYSELECT procedure, 7590
- sampling frame
  - SURVEYSELECT procedure, 7590, 7603
- sampling rate
  - SURVEYSELECT procedure, 7611
- sampling units
  - SURVEYSELECT procedure, 7590, 7617, 7626
- sampling weights
  - SURVEYSELECT procedure, 7593
- sampling with replacement
  - SURVEYSELECT procedure, 7626
- sampling without replacement
  - SURVEYSELECT procedure, 7626
- seed
  - initial (SURVEYSELECT), 7614
- sequential random sampling
  - SURVEYSELECT procedure, 7628, 7647
- serpentine sorting
  - SURVEYSELECT procedure, 7625
- simple random sampling
  - SURVEYSELECT procedure, 7592, 7627
- size measures
  - PPS sampling (SURVEYSELECT), 7618, 7650
- strata
  - SURVEYSELECT procedure, 7591, 7594, 7619
- stratification, *see* stratified sampling
- stratified sampling
  - SURVEYSELECT procedure, 7591, 7594, 7619
- survey sampling
  - sample selection (SURVEYSELECT), 7590
  - SURVEYSELECT procedure, 7590
- survey weights, *see* sampling weights
- SURVEYSELECT procedure, 7590
  - allocation, 7620, 7634
  - allocation output data set, 7642
  - Brewer's selection method, 7633, 7650
  - certainty size measure, 7601
  - certainty size proportion, 7602
  - Chromy's selection method, 7628, 7631
  - cluster sampling, 7617
  - control sorting, 7598, 7616, 7625, 7647
  - displayed output, 7643
  - dollar-unit sampling, 7653
  - Hanurav-Vijayan selection method, 7629
  - initial seed, 7614
  - introductory example, 7591
  - joint selection probabilities, 7603
  - margin of error, 7636
  - maximum size measure, 7604
  - minimum size measure, 7608
  - missing values, 7624
  - Murthy's selection method, 7633
  - nested sorting, 7625
  - Neyman allocation, 7621, 7636
  - ODS table names, 7646
  - optimal allocation, 7621, 7635
  - output data sets, 7639, 7642
  - PPS sampling, with replacement, 7631
  - PPS sampling, without replacement, 7629
  - PPS sequential sampling, 7631
  - PPS systematic sampling, 7631
  - proportional allocation, 7621, 7635, 7656
  - replicated sampling, 7591, 7611, 7647
  - Sampford's selection method, 7634
  - sample output data set, 7639
  - sample selection methods, 7605, 7626
  - sample size, 7613
  - sample size allocation, 7591, 7620, 7634
  - sampling rate, 7611
  - sampling units, 7617
  - secondary input data set, 7638
  - sequential random sampling, 7628, 7647
  - serpentine sorting, 7625
  - simple random sampling, 7592, 7627
  - size measures, 7618, 7650
  - strata, 7594, 7619
  - stratified sampling, 7591, 7594, 7619
  - systematic random sampling, 7598, 7627
  - unrestricted random sampling, 7627
  - with-replacement sampling, 7626
  - without-replacement sampling, 7626
- systematic random sampling
  - SURVEYSELECT procedure, 7598, 7627

**U**

- unrestricted random sampling
  - SURVEYSELECT procedure, 7627

**W**

- weighting, *see also* sampling weights
- weights, *see* sampling weights
- with-replacement sampling

SURVEYSELECT procedure, 7626  
without-replacement sampling  
SURVEYSELECT procedure, 7626



# Syntax Index

## A

- ALLOC= option
  - STRATA statement (SURVEYSELECT), 7621
- ALLOCMIN= option
  - STRATA statement (SURVEYSELECT), 7622
- ALPHA= option
  - STRATA statement (SURVEYSELECT), 7622

## C

- CERTSIZE= option
  - PROC SURVEYSELECT statement, 7601
- CERTSIZE=P= option
  - PROC SURVEYSELECT statement, 7602
- CLUSTER statement
  - SURVEYSELECT procedure, 7617
- CONTROL statement
  - SURVEYSELECT procedure, 7616
- COST= option
  - STRATA statement (SURVEYSELECT), 7622

## D

- DATA= option
  - PROC SURVEYSELECT statement, 7603

## I

- ID statement
  - SURVEYSELECT procedure, 7617

## J

- JTPROBS option
  - PROC SURVEYSELECT statement, 7603

## M

- MARGIN= option
  - STRATA statement (SURVEYSELECT), 7623
- MAXSIZE= option
  - PROC SURVEYSELECT statement, 7604
- METHOD= option
  - PROC SURVEYSELECT statement, 7605
- MINSIZE= option
  - PROC SURVEYSELECT statement, 7608

## N

- NMAX= option
  - PROC SURVEYSELECT statement, 7609
- NMIN= option
  - PROC SURVEYSELECT statement, 7609
- NOPRINT option
  - PROC SURVEYSELECT statement, 7609
- NOSAMPLE option
  - STRATA statement (SURVEYSELECT), 7623

## O

- OUT= option
  - PROC SURVEYSELECT statement, 7609
- OUTALL option
  - PROC SURVEYSELECT statement, 7610
- OUTHITS option
  - PROC SURVEYSELECT statement, 7610
- OUTSEED option
  - PROC SURVEYSELECT statement, 7610
- OUTSIZE option
  - PROC SURVEYSELECT statement, 7610
- OUTSORT= option
  - PROC SURVEYSELECT statement, 7611

## P

- PPS option
  - SAMPLINGUNIT statement (SURVEYSELECT), 7618
- PRESORTED option
  - SAMPLINGUNIT statement (SURVEYSELECT), 7618
- PROC SURVEYSELECT statement, 7599, *see* SURVEYSELECT procedure

## R

- REPS= option
  - PROC SURVEYSELECT statement, 7611

## S

- SAMPLINGUNIT statement
  - SURVEYSELECT procedure, 7617
- SAMPRATE= option
  - PROC SURVEYSELECT statement, 7611
- SAMPSIZE= option
  - PROC SURVEYSELECT statement, 7613

SEED= option  
    PROC SURVEYSELECT statement, 7614  
SELECTALL option  
    PROC SURVEYSELECT statement, 7615  
SIZE statement  
    SURVEYSELECT procedure, 7618  
SORT= option  
    PROC SURVEYSELECT statement, 7616  
STATS option  
    PROC SURVEYSELECT statement, 7616  
    STRATA statement (SURVEYSELECT), 7623  
STRATA statement  
    SURVEYSELECT procedure, 7619  
SURVEYSELECT procedure  
    syntax, 7599  
SURVEYSELECT procedure, CLUSTER statement,  
    7617  
SURVEYSELECT procedure, CONTROL statement,  
    7616  
SURVEYSELECT procedure, ID statement, 7617  
SURVEYSELECT procedure, PROC  
    SURVEYSELECT statement, 7599  
    CERTSIZE= option, 7601  
    CERTSIZE=P= option, 7602  
    DATA= option, 7603  
    JTPROBS option, 7603  
    MAXSIZE= option, 7604  
    METHOD= option, 7605  
    MINSIZE= option, 7608  
    NMAX= option, 7609  
    NMIN= option, 7609  
    NOPRINT option, 7609  
    OUT= option, 7609  
    OUTALL option, 7610  
    OUTHITS option, 7610  
    OUTSEED option, 7610  
    OUTSIZE option, 7610  
    OUTSORT= option, 7611  
    REPS= option, 7611  
    SAMPRATE= option, 7611  
    SAMPSIZE= option, 7613  
    SEED= option, 7614  
    SELECTALL option, 7615  
    SORT= option, 7616  
    STATS option, 7616  
SURVEYSELECT procedure, SAMPLINGUNIT  
    statement, 7617  
    PPS option, 7618  
    PRESORTED option, 7618  
SURVEYSELECT procedure, SIZE statement, 7618  
SURVEYSELECT procedure, STRATA statement,  
    7619  
    ALLOC= option, 7621  
    ALLOCMIN= option, 7622

ALPHA= option, 7622  
COST= option, 7622  
MARGIN= option, 7623  
NOSAMPLE option, 7623  
STATS option, 7623  
VAR= option, 7623

## V

VAR= option  
    STRATA statement (SURVEYSELECT), 7623

## Your Turn

---

We welcome your feedback.

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



# SAS® Publishing Delivers!

Whether you are new to the work force or an experienced professional, you need to distinguish yourself in this rapidly changing and competitive job market. SAS® Publishing provides you with a wide range of resources to help you set yourself apart. Visit us online at [support.sas.com/bookstore](http://support.sas.com/bookstore).

## SAS® Press

Need to learn the basics? Struggling with a programming problem? You'll find the expert answers that you need in example-rich books from SAS Press. Written by experienced SAS professionals from around the world, SAS Press books deliver real-world insights on a broad range of topics for all skill levels.

[support.sas.com/saspress](http://support.sas.com/saspress)

## SAS® Documentation

To successfully implement applications using SAS software, companies in every industry and on every continent all turn to the one source for accurate, timely, and reliable information: SAS documentation. We currently produce the following types of reference documentation to improve your work experience:

- Online help that is built into the software.
- Tutorials that are integrated into the product.
- Reference documentation delivered in HTML and PDF – free on the Web.
- Hard-copy books.

[support.sas.com/publishing](http://support.sas.com/publishing)

## SAS® Publishing News

Subscribe to SAS Publishing News to receive up-to-date information about all new SAS titles, author podcasts, and new Web site features via e-mail. Complete instructions on how to subscribe, as well as access to past issues, are available at our Web site.

[support.sas.com/spn](http://support.sas.com/spn)



**sas**

**THE  
POWER  
TO KNOW®**

