Chapter 63
The SURVEYSELECT Procedure

Chapter Table of Contents

OVERVIEW . . . 3275

GETTING STARTED . . . 3276
- Simple Random Sampling . . . 3277
- Stratified Sampling . . . 3279
- Stratified Sampling with Control Sorting . . . 3282

SYNTAX . . . 3283
- PROC SURVEYSELECT Statement . . . 3284
- CONTROL Statement . . . 3294
- ID Statement . . . 3295
- SIZE Statement . . . 3295
- STRATA Statement . . . 3295

DETAILS . . . 3296
- Missing Values . . . 3296
- Sorting by CONTROL Variables . . . 3296
- Sample Selection Methods . . . 3297
  - Simple Random Sampling . . . 3298
  - Unrestricted Random Sampling . . . 3298
  - Systematic Random Sampling . . . 3299
  - Sequential Random Sampling . . . 3299
  - PPS Sampling without Replacement . . . 3300
  - PPS Sampling with Replacement . . . 3302
  - PPS Systematic Sampling . . . 3302
  - PPS Sequential Sampling . . . 3303
  - Brewer's PPS Method . . . 3304
  - Murthy's PPS Method . . . 3305
  - Sampford's PPS Method . . . 3306
- Output Data Set . . . 3306
- Displayed Output . . . 3309
- ODS Table Names . . . 3310

EXAMPLES . . . 3310
- Example 63.1 Replicated Sampling . . . 3310
- Example 63.2 PPS Selection of Two Units Per Stratum . . . 3313

Example 63.3 PPS (Dollar-Unit) Sampling 3315
REFERENCES 3319

Chapter 63
The SURVEYSELECT Procedure

Overview

The SURVEYSELECT procedure provides a variety of methods for selecting probability-based random samples. The procedure can select a simple random sample or a 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, or list of units from which the sample is to be selected. You also specify the selection method, the desired sample size or sampling rate, and other selection parameters. The SURVEYSELECT procedure selects the sample, producing an output data set that contains the selected units, their selection probabilities, and sampling weights. When you are selecting 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.

The SURVEYSELECT procedure 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 on probability sampling methods, refer to Kish (1987, 1965), Kalton (1983), and Cochran (1977).

The SURVEYSELECT procedure provides the following equal probability sampling methods:

- simple random sampling
- unrestricted random sampling (with replacement)
- systematic random sampling
- sequential random sampling

This procedure also provides the following probability proportional to size (PPS) 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 very large input data sets or sampling frames, which may occur in practice for large-scale sample surveys.

The SURVEYSELECT procedure can perform stratified sampling, selecting samples independently within the specified strata, or nonoverlapping subgroups of the survey population. Stratification controls the distribution of the sample size in the strata. It is widely used in practice towards 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 towards improving the precision of the overall estimates. When you are using a systematic or sequential selection method, the SURVEYSELECT procedure also can sort by control variables within strata for the additional control of implicit stratification.

The SURVEYSELECT procedure provides replicated sampling, where the total sample is composed of a set of replicates, each 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 compute standard errors for the combined sample estimates.

Getting Started

In this example, an Internet service provider wants to conduct 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. Figure 63.1 displays the first ten observations of the data set Customers.

Internet Service Provider Customers
(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

Figure 63.1. Customers Data Set (First 10 Observations)

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 the following 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 using simple random sampling.

```
title1 'Customer Satisfaction Survey';
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 63.2 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. Since the SEED= option is not specified in the PROC SURVEYSELECT statement, the seed value is obtained using the time of day from the computer's clock. You can specify SEED=39647 to reproduce this sample.

```
                    Customer Satisfaction Survey

                     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
```

Figure 63.2. Sample Selection Summary

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 63.3 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.

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

Figure 63.3. Customer Sample (First 20 Observations)

Stratified Sampling

In this section, stratification is added to the sample design for the customer satisfaction survey. The sampling frame, or 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.

```
proc freq data=Customers;
   tables State*Type;
run;
```

Figure 63.4 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.

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

Figure 63.4. Stratification of Customers by State and Type

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

```
                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
```

Figure 63.5. Sample Selection Summary

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 63.6 displays the first 30 observations of the output data set SampleStrata, which contains the sample of 120 customers, 15 customers from each of the 8 strata. The variable SelectionProb contains the selection probability for each customer in the sample. Since 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 since the 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 is 0.021246 for customers in the second stratum. The variable SamplingWeight contains the sampling weights, which are computed as inverse selection probabilities.

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

Figure 63.6. Customer Sample (First 30 Observations)

Stratified Sampling with Control Sorting

The next sample design for the customer satisfaction survey uses stratification by State. The sampling frame is also sorted by Type and Usage before sample selection, to provide additional control over the distribution of the sample. Customers are then selected by systematic random sampling within strata. The following PROC SURVEYSELECT statements select a probability sample of customers from the Customers data set using this design.

```
title1 'Customer Satisfaction Survey';
title2 'Stratified Sampling with Control Sorting';
proc surveyselect data=Customers method=sys seed=1234
      rate=.02 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 SEED=1234 option specifies the initial seed for random number generation. The RATE=.02 option specifies a sampling rate of 2% for each stratum.

Figure 63.7 displays the output from PROC SURVEYSELECT, which summarizes the sample selection. A sample of 271 customers is selected, 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 used by default since SORT=NEST is not specified. See the section "Sorting by CONTROL Variables" on page 3296 for a description of serpentine sorting. The output data set SampleControl contains the sample of customers.

```
          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
```

Figure 63.7. Sample Selection Summary

Syntax

The following statements are available in PROC SURVEYSELECT.

PROC SURVEYSELECT *options* **;**
STRATA *variables* **;**
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 SURVEYSELECT statement is required.

The SIZE statement identifies the variable that contains the size measures. It is required for any selection method that is probability proportional to size (PPS).

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 formed by the STRATA variables. 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 for the CONTROL, ID, 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*n* convention.

The PROC SURVEYSELECT statement also specifies the sample selection method, the sample size, and other sample design parameters. If you do not specify a selection method, PROC SURVEYSELECT uses simple random sampling (METHOD=SRS) if there is no SIZE statement. If you specify a SIZE statement but do not specify a selection method, PROC SURVEYSELECT uses probability proportional to size selection without replacement (METHOD=PPS). You must specify the sample size or sampling rate unless you request a method that selects two units from each stratum (METHOD=PPS_BREWER or METHOD=PPS_MURTHY).

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 specify 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=, and CERTSIZE= options. You can name only one secondary input data set in each invocation of the procedure.

The following table lists the options available with the PROC SURVEYSELECT statement. Descriptions follow in alphabetical order.

Table 63.1. PROC SURVEYSELECT Statement Options

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

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

CERTSIZE

requests automatic selection of those units with size measures greater than or equal to the stratum certainty size measures, which you provide in the secondary input data set variable _CERTSIZE_. Use the CERTSIZE option when you have already named the secondary input data set in another option, such as SAMPSIZE=*SAS-data-set*, SAMPRATE=*SAS-data-set*, MAXSIZE=*SAS-data-set*, or MINSIZE=*SAS-data-set*. You can name only one secondary input data set in each invocation of the procedure.

If any size measure is greater than or equal to the certainty size measure for its stratum, then PROC SURVEYSELECT selects this unit with certainty. Each certainty size measure must be a positive number. The CERTSIZE option is available for METHOD=PPS and METHOD=PPS_SAMPFORD.

If you want to specify a single certainty size measure in the PROC SURVEYSELECT statement, use the CERTSIZE=*certain* option.

CERTSIZE=*certain*

specifies the certainty size measure. PROC SURVEYSELECT selects with certainty any unit with size measure greater than or equal to the value *certain*. The certainty size measure must be a positive number. This option is available for METHOD=PPS and METHOD=PPS_SAMPFORD.

If you request a stratified sample design with a STRATA statement and specify the CERTSIZE= option, PROC SURVEYSELECT uses the certainty size *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 for each stratum.

CERTSIZE=*SAS-data-set*
names a SAS data set that contains the certainty size measures for the strata. PROC SURVEYSELECT selects with certainy any unit with size measure greater than or equal to the certainty size measure for its stratum. Each certainty size measure must be a positive number. This option is available for METHOD=PPS and METHOD=PPS_SAMPFORD.

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 should have a variable named _CERTSIZE_ that contains the certainty size measure for each stratum.

CERTSIZE=P=*p*
specifies the certainty proportion. PROC SURVEYSELECT selects with certainty any unit with size measure greater than or equal to the proportion *p* of the total size for all units in the stratum. The procedure repeats this process with the remaining units until no more certainty units are selected. This option is available for METHOD=PPS and METHOD=PPS_SAMPFORD.

The certainty proportion 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%, and not the percentage form 1%.

If you request a stratified sample design with a STRATA statement and specify the CERTSIZE=P= option, PROC SURVEYSELECT uses the same certainty proportion *p* for all strata.

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*, or list of units from which the sample is selected.

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 more information on the contents of the output data set, see the section “Output Data Set” on page 3306.

MAXSIZE
requests size measure adjustment by stratum maximum size measures, which you provide in the secondary input data set variable _MAXSIZE_. Use the MAXSIZE option when you have already named the secondary input data set

in another option, such as SAMPSIZE=*SAS-data-set*, SAMPRATE=*SAS-data-set*, MINSIZE=*SAS-data-set*, or CERTSIZE=*SAS-data-set*. You can name only one secondary input data set in each invocation of the procedure.

If any size measure exceeds the maximum size measure for its stratum, then PROC SURVEYSELECT adjusts this size measure downward to equal the maximum size measure. Each maximum size measure must be a positive number. The MAXSIZE option is available whenever you specify a SIZE statement for probability proportional to size selection and a STRATA statement for stratification.

If you want to specify a single maximum size value in the PROC SURVEYSELECT statement, use the MAXSIZE=*max* option.

MAXSIZE=*max*

specifies the maximum allowable size measure. If any size measure exceeds the value *max*, then PROC SURVEYSELECT adjusts this size measure to equal *max*. The maximum size measure must be a positive number. This option is available whenever you specify a SIZE statement for selection with probability proportional to size.

If you request a stratified sample design with a STRATA statement and specify the MAXSIZE= 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 for each stratum.

MAXSIZE=*SAS-data-set*

names a SAS data set that contains the maximum allowable size measures for the strata. If any size measure exceeds the maximum size measure for its stratum, then PROC SURVEYSELECT adjusts this size measure downward to equal the maximum size measure. Each maximum size measure must be a positive number. This option is available whenever you specify a SIZE statement for probability proportional to size selection and a STRATA statement for stratified selection.

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 should have a variable named _MAXSIZE_ that contains the maximum size measure for each stratum.

METHOD=*name*
M=*name*

specifies the method for sample selection. If you do not specify the METHOD= option, by default PROC SURVEYSELECT uses simple random sampling (METHOD=SRS) if there is no SIZE statement. If you specify a SIZE statement, the default selection method is probability proportional to size without replacement (METHOD=PPS). Valid values for *name* are as follows:

PPS requests selection with probability proportional to size and without replacement. See the section "PPS Sampling without Replacement" on page 3300 for details. If you specify METHOD=PPS, you must name the size measure variable in the SIZE 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 3304 for details. If you specify METHOD=PPS_BREWER, you must name the size measure variable in the SIZE statement. You do not need to specify the sample size with the SAMPSIZE= option, since 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 3305 for details. If you specify METHOD=PPS_MURTHY, you must name the size measure variable in the SIZE statement. You do not need to specify the sample size with the SAMPSIZE= option, since 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 3306 for details. If you specify METHOD=PPS_SAMPFORD, you must name the size measure variable in the SIZE 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 3303 for details. If you specify METHOD=PPS_SEQ, you must name the size measure variable in the SIZE statement.

PPS_SYS requests systematic selection with probability proportional to size. See the section "PPS Systematic Sampling" on page 3302 for details on this method. If you specify METHOD=PPS_SYS, you must name the size measure variable in the SIZE statement.

PPS_WR requests selection with probability proportional to size and with replacement. See the section "PPS Sampling with Replacement" on page 3302 for details on this method. If you specify METHOD=PPS_WR, you must name the size measure variable in the SIZE statement.

SEQ requests sequential selection according to Chromy's method. If you specify METHOD=SEQ and do not specify a size measure with the SIZE statement, PROC SURVEYSELECT uses sequential zoned selection with equal probability and without replacement. See the section "Sequential Random Sampling" on page 3299 for details on this method. If you specify METHOD=SEQ and also name a size measure in the SIZE 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 3303 for details on this method.

SRS requests simple random sampling, which is selection with equal probability and without replacement. See the section "Simple Random Sampling" on page 3298 for details. This method is the default if you do not specify the METHOD= option and also do not specify a SIZE statement.

SYS requests systematic random sampling. If you specify METHOD=SYS and do not specify a size measure with the SIZE statement, PROC SURVEYSELECT uses systematic selection with equal probability. See the section "Systematic Random Sampling" on page 3299 for details on this method. If you specify METHOD=SYS and also name a size measure in the SIZE statement, PROC SURVEYSELECT uses METHOD=PPS_SYS, which is systematic selection with probability proportional to size. See the section "PPS Systematic Sampling" on page 3302 for details.

URS requests unrestricted random sampling, which is selection with equal probability and with replacement. See the section "Unrestricted Random Sampling" on page 3298 for details.

MINSIZE

requests size measure adjustment by the stratum minimum size measures, which you provide in the secondary input data set variable _MINSIZE_. Use the MINSIZE option when you have already named the secondary input data set in another option, such as SAMPSIZE=*SAS-data-set*, SAMPRATE=*SAS-data-set*, MAXSIZE=*SAS-data-set*, or CERTSIZE=*SAS-data-set*. You can name only one secondary input data set in each invocation of the procedure.

If any size measure is less than the minimum size measure for its stratum, then PROC SURVEYSELECT adjusts this size measure upward to equal the minimum size measure. Each minimum size measure must be a positive number. The MINSIZE option is available whenever you specify a SIZE statement for probability proportional to size selection and a STRATA statement for stratification.

If you want to specify a single minimum size value in the PROC SURVEYSELECT statement, use the MINSIZE=*min* option.

MINSIZE=*min*

specifies the minimum allowable size measure. If any size measure is less than the value *min*, then PROC SURVEYSELECT adjusts this size measure upward to equal *min*. The minimum size measure must be a positive number. This option is available whenever you specify a SIZE statement for selection with probability proportional to size.

If you request a stratified sample design with a STRATA statement and specify the MINSIZE= 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 for each stratum.

MINSIZE=*SAS-data-set*

names a SAS data set that contains the minimum allowable size measures for the strata. If any size measure is less than the minimum size measure for its stratum, then PROC SURVEYSELECT adjusts this size measure upward to equal the minimum size measure. Each minimum size measure must be a positive number. This option is available whenever you specify a SIZE statement for probability proportional to size selection and a STRATA statement for stratified selection.

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 should have a variable named _MINSIZE_ that contains the minimum size measure for each stratum.

NMAX=*n*

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

The maximum sample size *n* must be a positive integer. The NMAX= option is available only with the SAMPRATE= option, which may 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 desired stratum sample size from the specified sampling rate and the total number of units in the stratum. If this sample size is less than the value NMIN=*n*, then PROC SURVEYSELECT selects the minimum of *n* units.

The minimum sample size *n* must be a positive integer. The NMIN= option is available only with the SAMPRATE= option, which may 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 15, "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 selected for the sample, as well as design information and selection statistics, depending on the selection method and output options

you specify. See the descriptions for the options JTPROBS, OUTHITS, OUTSIZE, and STATS. For information on the contents of the output data set, see the section "Output Data Set" on page 3306.

OUTHITS

includes a separate observation in the output data set for each selection when the same unit is selected more than once. By default, the output data set contains only one observation for each selected unit, even if it is selected more than once, and the variable NumberHits contains the number of hits or selections for that unit. The OUTHITS option is available for selection methods that select with replacement or with minimum replacement (METHOD=URS, METHOD=PPS_WR, METHOD=PPS_SYS, and METHOD=PPS_SEQ).

OUTSIZE

includes additional design and sampling frame parameters in the output data set. If you specify the OUTSIZE option, PROC SURVEYSELECT includes the sample size or sampling rate in the output data set. When you request the OUTSIZE option and also specify the SIZE statement, the procedure outputs the size measure total for the sampling frame. If you do not specify the SIZE statement, the procedure outputs the total number of sampling units in the frame. Also, PROC SURVEYSELECT includes the minimum size measure if you specify the MINSIZE= option, the maximum size measure if you specify the MAXSIZE= option, and the certainty size measure if you specify the CERTSIZE= option.

If you have a stratified design, the output data set includes the stratum-level values of these parameters. Otherwise, the output data set includes the overall population-level values.

For information on the contents of the output data set, see the section "Output Data Set" on page 3306.

OUTSORT=*SAS-data-set*

names an output data set that contains the sorted input data set. This 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). PROC SURVEYSELECT sorts the input data set by the CONTROL variables within strata before selecting the sample.

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.

REP=*nrep*

specifies the number of sample replicates. If you specify the REP= option, PROC SURVEYSELECT selects *nrep* independent samples, each with the same specified sample size or sampling rate and the same sample design.

You can use replicated sampling to provide a simple method of variance estimation for any form of statistic, as well as to evaluate variable nonsampling errors such as interviewer differences. Refer to Kish (1965), Kish (1987), and Kalton (1983) for information on replicated sampling.

SAMPRATE=*r*
RATE=*r*

specifies the sampling rate, which is the proportion of units selected 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%, and not the percentage form 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 3299 for details. For other selection methods, PROC SURVEYSELECT converts the sampling rate *r* to the sample size before selection, multiplying the rate by the number of units in the stratum or frame and rounding up to the nearest integer.

If you request a stratified sample design with a 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 sampling rates for the strata. 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. If 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 options in the STRATA statement.

Each stratum sampling rate value 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%, and not the percentage form 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 stratum sampling rate as the interval for the stratum. See the section "Systematic Random Sampling" on page 3299 for details on systematic sampling. For other selection methods, PROC SURVEYSELECT converts the stratum sampling rate to a stratum sample size before selection, multiplying the rate by the number of units in the stratum and rounding up to the nearest integer.

SAMPRATE=*SAS-data-set*
RATE=*SAS-data-set*

names a SAS data set that contains sampling rates for the strata. This 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 SAMPRATE= data set should have a variable _RATE_ that contains the sampling rate for each stratum.

Each sampling rate value 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%, and not the percentage form 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 stratum sampling rate as the interval for the stratum. See the section "Systematic Random Sampling" on page 3299 for details. For other selection methods, PROC SURVEYSELECT converts the stratum sampling rate to the stratum sample size before selection, multiplying the rate by the number of units in the stratum and rounding up to the nearest integer.

SAMPSIZE=*n*
N=*n*

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

If you request a stratified sample design with a STRATA statement and specify the SAMPSIZE=*n* option, PROC SURVEYSELECT selects *n* units from each stratum. 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 different sample sizes for the strata.

SAMPSIZE=(*values***)**
N=(*values***)**

specifies sample sizes for the strata. 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. If 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 options in the STRATA statement.

Each stratum sample size value must be a positive integer. For methods that select without replacement, the sample size for a stratum must not exceed the number of units in that stratum.

SAMPSIZE=*SAS-data-set*
N=*SAS-data-set*
names a SAS data set that contains the sample sizes for the strata. This 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 should have a variable _NSIZE_ that contains the sample size for each stratum. Each sample size value must be a positive integer. For methods that select without replacement, the stratum sample size must not exceed the number of units in the stratum.

SEED=*number*
specifies the initial seed for random number generation. The value of the SEED= option must be a positive integer. 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.

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 3296 for descriptions of serpentine and nested sorting. Where there is only one CONTROL variable, the two types of sorting are equivalent.

This 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). PROC SURVEYSELECT sorts the input data set by the CONTROL variables within strata before selecting the sample.

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 folowing 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 on the contents of the output data set, see the section "Output Data Set" on page 3306.

CONTROL Statement

CONTROL *variables* **;**

The CONTROL statement names variables for sorting the input data set. The CONTROL variables can be character or numeric.

PROC SURVEYSELECT sorts the input data set by the CONTROL variables before selecting the sample. 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).

By default, PROC SURVEYSELECT uses hierarchic serpentine sorting by the CONTROL variables. If you specify the SORT=NEST option, the procedure uses nested sorting. See the description for the SORT= option. For more information on serpentine and nested sorting, see the section "Sorting by CONTROL Variables" on page 3296.

You can use the OUTSORT= option 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 variables from the DATA= input data set to be included in the OUT= data set of selected units. If there is no ID statement, PROC SURVEYSELECT includes all variables from the DATA= data set in the OUT= data set. The ID variables can be character or numeric.

SIZE Statement

SIZE *variable* **;**

The SIZE statement names one and only one size measure variable, which contains the size measures to be used when sampling with probability proportional to size. The SIZE variable must be numeric. When the value of an observation's SIZE variable is missing or nonpositive, that observation has no chance of being selected for the sample.

The SIZE statement is required for all PPS selection methods, which include METHOD=PPS, METHOD=PPS_BREWER, METHOD=PPS_MURTHY METHOD=PPS_SAMPFORD, METHOD=PPS_SEQ, METHOD=PPS_SYS, and METHOD=PPS_WR.

STRATA Statement

STRATA *variables* **;**

You can specify a STRATA statement with PROC SURVEYSELECT to partition the input data set into nonoverlapping groups defined by the STRATA variables. PROC SURVEYSELECT then selects independent samples from these strata, according to the selection method and design parameters specified in the PROC SURVEYSELECT statement.

For information on the use of stratification in sample design, refer to Kalton (1983), Kish (1987), and Cochran (1977).

The *variables* are one or more variables in the input data set. 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 of the STRATA variables. So 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 using the SORT procedure with the STRATA variables in a BY statement.
- Specify the option NOTSORTED or DESCENDING in the STRATA statement for the SURVEYSELECT procedure (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 using the DATASETS procedure.

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

Details

Missing Values

If an observation has a missing or nonpositive value for the SIZE variable, PROC SURVEYSELECT excludes that observation from the sample selection. The procedure writes a note to the log giving the number of observations omitted due to missing or nonpositive size measures.

PROC SURVEYSELECT treats missing STRATA variable values like any other STRATA variable value. The missing values form a separate stratum.

If a value of _NSIZE_ is missing in the SAMPSIZE= input data set, then PROC SURVEYSELECT writes an error message to the log and does not select a sample from that stratum. The procedure treats missing values of _NRATE_, _MINSIZE_, _MAXSIZE_, and _CERTSIZE_ similarly.

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, giving some benefits of proportionate stratification.

By default, the sorted data set replaces the input data set. Or 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, nested sorting and hierarchic serpentine sorting. If you specify the SORT=NEST option, then the procedure sorts by the CONTROL variables according to nested sorting. If you do not specify the SORT=NEST option, the procedure uses serpentine sorting by default. 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. Refer to the chapter on the SORT procedure in the *SAS Procedures Guide*. PROC SURVEYSELECT sorts within strata if you also specify a STRATA statement. The procedure first arranges the input observations in ascending order of the first CONTROL variable. Then within each level of the first control variable, the procedure arranges the observations in ascending order of the second CONTROL variable. This continues for all CONTROL variables 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 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 on serpentine sorting, refer to 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 probablity 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. Refer to Kish (1987, 1965), Kalton (1983), and Cochran (1977) for more information on 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 may be schools, hospitals, or geographical areas, and the final sampling units may be students, patients, or citizens. Cluster sampling can provide efficiencies in frame construction and other survey operations. Refer to Kalton (1983), Kish (1965), and the other references cited in the following sections for more information.

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

PROC SURVEYSELECT uses Floyd's ordered hash table algorithm for simple random sampling. This algorithm is fast, efficient, and appropriate for large data sets. Refer to Bentley and Floyd (1987) and Bentley and Knuth (1986). For additional information on simple random sampling algorithms, refer to McLeod and Bellhouse (1983) and Fann, Muller, and Rezucha (1962).

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 selections or hits 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.

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. PROC SURVEYSELECT chooses the first unit randomly from the entire stratum and then treats the stratum observations as a closed loop. This is done to obtain an unbiased variance estimator, as suggested by Lahiri (Murthy 1967). 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. Refer to Chromy (1979) and Williams and Chromy (1980). See the section "PPS Sequential Sampling" on page 3303 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, 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 3296 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). Using 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 or hits, where the expected number of selections ES_{hi} equals n_h/N_h for all units i in stratum h. The procedure computes

$$I_{hi} \;=\; Int\,(\sum_{j=1}^{i} ES_{hj}) \;=\; Int\,(i\,n_h/N_h)$$

$$F_{hi} \;=\; Frac\,(\sum_{j=1}^{i} ES_{hj}) \;=\; Frac\,(i\,n_h/N_h)$$

where Int denotes the integer part of the number, and $Frac$ denotes the fractional part.

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

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

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

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

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

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

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

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 $n_h Z_{hi}$. The procedure 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 nonnegativiy and stability of the Sen-Yates-Grundy variance estimator. Refer to 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 $M_1 \leq M_2 \leq \ldots \leq M_N$. Then the procedure selects the PPS sample of n observations as follows:

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

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

   $Z_j = M_j/M$, $T = \sum_{j=1}^{N-n} Z_j$, and, 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 below.

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,

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

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

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

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

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

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

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

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

If you request 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

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

and

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

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

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 $Z_{hi} = M_{hi}/M_{h\cdot}$. Because units are selected with replacement, a unit can be selected for the sample more than once. The expected number of selections or hits for unit i in stratum h equals $n_h Z_{hi}$. If you request 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

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

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 chooses the first unit randomly from the entire stratum with probability proportional to size and then treats the stratum observations as a closed loop. This is done to obtain an unbiased variance estimator, as suggested by Lahiri (Murthy 1967). 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 may be possible for a unit to be selected more than once. The expected number of selections or hits for unit i in stratum h equals $n_h M_{hi}/M_{h.} = n_h Z_{hi}$. Refer to Cochran (1977, pp. 265–266) and Madow (1949).

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. Refer to Chromy (1979) and Williams and Chromy (1980). Chromy's method selects units sequentially with probability proportional to size and with minimum replacement. Selection *with minimum replacement* means that the actual number of hits for a unit can equal the integer part of the expected number of hits for that unit, or the next largest integer. This can be compared to selection *without replacement*, where each unit can be selected only once, so the number of hits can equal 0 or one. 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 $0, 1, \cdots, n_h$, 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, 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 3296 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 selections or hits, although fewer than n_h distinct units may be selected. Beginning with the random start, the procedure accumulates the expected number of hits and computes

$$ES_{hi} = n_h Z_{hi}$$

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

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

where ES_{hi} represents the expected number of hits for unit i in stratum h; Int denotes the integer part of the number; and $Frac$ 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

$$\left(F_{hi} - F_{h(i-1)}\right) \,/\, \left(1 - F_{h(i-1)}\right)$$

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 $2M_{hi}/M_{h\cdot} = 2Z_{hi}$.

Brewer's algorithm first selects a unit with probability

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

where

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

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

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

Brewer's method requires that the relative size Z_{hi} be less than 0.5 for all units. Refer to Cochran (1977, pp. 261–263) and Brewer (1963). Brewer's method yields the same selection probabilities and joint selection probabilities as Durbin's method. Refer to Cochran (1977) and Durbin (1967).

Murthy's PPS Method

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

$$P_{hi} \quad = \quad Z_{hi}\,[1+K-(Z_{hi}/(1-Z_{hi})]$$

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

$$K \quad = \quad \sum_{j=1}^{N}[Z_{hj}/(1-Z_{hj})]$$

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

$$P_{h(ij)} \quad = \quad Z_{hi}\;Z_{hj}\;\frac{2-Z_{hi}-Z_{hj}}{(1-Z_{hi})\,(1-Z_{hj})}$$

Murthy's method requires that the relative size Z_{hi} be less than 0.5 for all units. Refer to Cochran (1977, pp. 263–265) and Murthy (1957).

Sampford's PPS Method

Sampford's method (METHOD=PPS_SAMPFORD) is an extension of Brewer's method that selects more than two units from each stratum, with probability proportional to size and without replacement. The selection probability for unit i in stratum h equals

$$P_{hi} = n_h \frac{M_{hi}}{M_{h\cdot}} = n_h \, Z_{hi}$$

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

$$\frac{Z_{hi}}{1 - n_h \, Z_{hi}}$$

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.

The joint selection probability for units i and j in stratum h equals

$$P_{h(ij)} = K_h \, \lambda_i \, \lambda_j \sum_{t=2}^{n_h} \left[t - n_h \, (P_{hi} + P_{hj}) \, L_{n_h - t}(ij) \right] \, / \, n_h^{t-2}$$

where

$$\lambda_i = \frac{Z_{hi}}{1 - n_h \, Z_{hi}}$$

$$L_m = \sum_{S(m)} \lambda_{i_1} \lambda_{i_2} \cdots \lambda_{i_m}$$

where $S(m)$ denotes all possible samples of size m, for $m = 1, 2, \ldots, N_h$. The sum $L_m(ij)$ is defined similarly to L_m but sums over all possible samples of size m that do not include units i and j, and

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

Sampford's method requires that the relative size Z_{hi} be less than $1/n_h$ for all units. Refer to Cochran (1977, pp. 262–263) and Sampford (1967).

Output Data Set

PROC SURVEYSELECT creates a SAS data set that contains the sample of selected units. You can specify the name of this 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 output data set contains an observation for each unit selected for the sample. If you specify the OUTHITS option for methods that may select the same unit more than once (that is, methods that select with replacement or with minimum replacement), the output data set contains a separate observation for each selection. If you do not specify the OUTHITS option, the output data set contains only one observation for each selected unit, even if the unit is selected more than once, and the variable NumberHits contains the number of hits or selections for that unit.

The output data set 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:

- STRATA variables
- Replicate, which is the sample replicate number. This variable is included when you request replicated sampling with the REP= option.
- ID variables
- CONTROL variables
- Zone, which is the selection zone. This variable is included for METHOD=PPS_SEQ.
- SIZE variable
- AdjustedSize, which is the adjusted size measure. This variable is included if you request adjusted sizes with the MINSIZE= option or the MAXSIZE= option.
- Certain, which indicates certainty selection. This variable is included if you specify the CERTSIZE= option. It equals 1 for units included with certainty because their size measures exceed the certainty size measure. Otherwise, it equals 0.
- NumberHits, which is the number of hits or 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 for other methods:

- ExpectedHits, which is the expected number of hits or 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).
- 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 request 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 request 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 request 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 population-level values of these variables.

- MinimumSize, which is the minimum size measure specified with the MINSIZE= option. This variable is included if you request the MINSIZE= option.
- MaximumSize, which is the maximum size measure specified with the MAXSIZE= option. This variable is included if you request the MAXSIZE= option.
- CertaintySize, which is the certainty size measure specified with the CERTSIZE= option. This variable is included if you request the CERTSIZE= option.
- Total, which is the total number of sampling units in the stratum. This variable is included if there is no SIZE statement.
- TotalSize, which is the total of size measures in the stratum. This variable is included if there is a SIZE statement.

- TotalAdjSize, which is the total of adjusted size measures in the stratum. This variable is included if there is a SIZE statement and if you request adjusted sizes with the MAXSIZE= option or the 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=BREWER or METHOD=MURTHY, which select two units from each stratum.

Displayed Output

By default, PROC SURVEYSELECT displays two tables that summarize the sample selection. You can suppress display of these tables by using the NOPRINT option.

PROC SURVEYSELECT creates an output data set that contains the units selected for the sample. The procedure does not display this output data set. 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
- Size Measure variable, if you specify a SIZE statement
- Minimum Size Measure, if you specify the MINSIZE=*min* option
- Maximum Size Measure, if you specify the MAXSIZE=*max* option
- Certainty Size Measure, if you specify the CERTSIZE=*certain* option
- Strata Variables, if you specify a STRATA statement
- Control Variables, if you specify a CONTROL statement
- type of Control Sorting, Serpentine or Nested, if you specify a CONTROL 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
- Selection Probability, if you specify METHOD=SRS, METHOD=SYS, or METHOD=SEQ and do not specify a STRATA statement
- Expected Number of Hits, if you specify METHOD=URS and do not specify a STRATA statement
- Sampling Weight for equal probability selection methods (METHOD=SRS, METHOD=URS, METHOD=SYS, METHOD=SEQ) if you do not specify a STRATA statement
- Number of Strata, if you specify a STRATA statement
- Number of Replicates, if you specify the REP= option
- Total Sample Size, if you specify a STRATA statement or the REP= option
- Output Data Set name

ODS Table Names

PROC SURVEYSELECT assigns a name to each table it creates. You can use these names to reference the table when using the Output Delivery System (ODS) to select tables and create output data sets. These names are listed in the following table. For more information on ODS, see Chapter 15, "Using the Output Delivery System."

Table 63.2. ODS Tables Produced in PROC SURVEYSELECT

ODS Table Name	Description	Statement	Option
Method	Sample selection method	PROC	default
Summary	Sample selection summary	PROC	default

Examples

Example 63.1. Replicated Sampling

This example uses the Customers data set from the section "Getting Started" on page 3276. 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. Refer to Kish (1965), Kish (1987), and Kalton (1983) for information on replicated sampling.

This design includes 4 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 using this design.

```
title1 'Customer Satisfaction Survey';
title2 'Replicated Sampling';
proc surveyselect data=Customers method=seq
      rep=4 n=(8 12 20 10)
      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 REP=4 option specifies 4 replicates of this sample. The N=(8 12 20 10) option specifies 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 8 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.

Figure 63.1.1 displays the output from PROC SURVEYSELECT, which summarizes the sample selection. A total of 200 customers is selected in 4 replicates. PROC SURVEYSELECT selects each replicate using sequential random sampling within strata determined by State. The sampling frame Customers is sorted by control variables Type and Usage within strata, according to hierarchic serpentine sorting. The output data set SampleRep contains the sample.

Output 63.1.1. 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;
```

Figure 63.1.2 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. Since 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 63.1.2. 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 63.2. PPS Selection of Two Units Per Stratum

A state health agency plans to conduct a state-wide survey of a variety of different hospital services. The agency plans to select a probability sample of individual discharge records within hospitals using a two-stage sample design. First stage units are hospitals, and second stage units are patient discharges during the study time 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. This example describes hospital selection for this survey using PROC SURVEYSELECT.

The data set HospitalFrame contains all hospitals in the first geographical region of this 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 urbanized 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. Refer to Drummond et al. (1982) for details on 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.

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

Output 63.2.1. 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, using a stratified design with PPS selection of two units from each stratum.

```
title1 'Hospital Utilization Survey';
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 specifies the size measure variable. It is not necessary to specify the sample size with the N= option, since Brewer's method always selects two units from each stratum.

Figure 63.2.2 displays the output from PROC SURVEYSELECT. A total of 8 hospitals were selected from the 4 strata. The data set SampleHospitals contains the selected hospitals.

Output 63.2.2. Sample Selection Summary

```
                    Hospital Utilization Survey

                     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.

```
title1 'Hospital Utilization Survey';
title2 'Sample Hospitals, Region 1';
proc print data=SampleHospitals;
run;
```

Output 63.2.3. 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 63.3. 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. Refer to Wilburn (1984).

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

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

The following PROC PRINT statements display the sampling frame data set TravelExpense, which contains 41 observations.

```
title1 'Travel Expense Audit';
proc print data=TravelExpense;
run;
```

Output 63.3.1. 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 using the stratified design with PPS selection within strata.

```
title1 'Travel Expense Audit';
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 that the strata apear 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.

Figure 63.3.2 displays the output from PROC SURVEYSELECT. A total of 20 expense reports is selected for audit. The data set AuditSample contains the sample of travel expense reports.

Output 63.3.2. Sample Selection Summary

```
                    Travel Expense Audit

                 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.

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

Output 63.3.3. 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

References

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.

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, Inc.

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, D.C.: 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.

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: π_{ps} Sampling of Two Units from a Stratum," *Journal of the Royal Statistical Society, Series B*, 29, 374–391.

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

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

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

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," *Sankhya*, 18, 379–390.

Murthy, M.N. (1967), *Sampling Theory and Methods*, Calcutta, India: 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 π_{ps} Sampling Scheme: Generalization of a Method of Hanurav," *Journal of the Royal Statistical Society, Series B*, 30, 556–566.

Watts, D.L. (1991), "Correction: Computer Selection of Size-Biased Samples," *The American Statistician*, 45(2), 172.

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

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.

The correct bibliographic citation for this manual is as follows: SAS Institute Inc., *SAS/STAT® User's Guide, Version 8*, Cary, NC: SAS Institute Inc., 1999.

SAS/STAT® User's Guide, Version 8

ISBN 1–58025–494–2

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

1st printing, October 1999