Chapter 6
The SURVEYSELECT Procedure

Chapter Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>OVERVIEW</td>
<td>241</td>
</tr>
<tr>
<td>GETTING STARTED</td>
<td>242</td>
</tr>
<tr>
<td>Simple Random Sampling</td>
<td>243</td>
</tr>
<tr>
<td>Stratified Sampling</td>
<td>245</td>
</tr>
<tr>
<td>Stratified Sampling with Control Sorting</td>
<td>248</td>
</tr>
<tr>
<td>SYNTAX</td>
<td>249</td>
</tr>
<tr>
<td>PROC SURVEYSELECT Statement</td>
<td>250</td>
</tr>
<tr>
<td>CONTROL Statement</td>
<td>258</td>
</tr>
<tr>
<td>ID Statement</td>
<td>259</td>
</tr>
<tr>
<td>SIZE Statement</td>
<td>259</td>
</tr>
<tr>
<td>STRATA Statement</td>
<td>259</td>
</tr>
<tr>
<td>DETAILS</td>
<td>260</td>
</tr>
<tr>
<td>Missing Values</td>
<td>260</td>
</tr>
<tr>
<td>Sorting by CONTROL Variables</td>
<td>260</td>
</tr>
<tr>
<td>Sample Selection Methods</td>
<td>261</td>
</tr>
<tr>
<td>Simple Random Sampling</td>
<td>262</td>
</tr>
<tr>
<td>Unrestricted Random Sampling</td>
<td>262</td>
</tr>
<tr>
<td>Systematic Random Sampling</td>
<td>262</td>
</tr>
<tr>
<td>Sequential Random Sampling</td>
<td>263</td>
</tr>
<tr>
<td>PPS Sampling without Replacement</td>
<td>264</td>
</tr>
<tr>
<td>PPS Sampling with Replacement</td>
<td>266</td>
</tr>
<tr>
<td>PPS Systematic Sampling</td>
<td>266</td>
</tr>
<tr>
<td>PPS Sequential Sampling</td>
<td>267</td>
</tr>
<tr>
<td>Brewer’s PPS Method</td>
<td>268</td>
</tr>
<tr>
<td>Murthy’s PPS Method</td>
<td>269</td>
</tr>
<tr>
<td>Sampford’s PPS Method</td>
<td>269</td>
</tr>
<tr>
<td>Output Data Set</td>
<td>270</td>
</tr>
<tr>
<td>Displayed Output</td>
<td>272</td>
</tr>
<tr>
<td>ODS Table Names</td>
<td>273</td>
</tr>
<tr>
<td>EXAMPLES</td>
<td>273</td>
</tr>
<tr>
<td>Example 6.1 Replicated Sampling</td>
<td>274</td>
</tr>
<tr>
<td>Example 6.2 PPS Selection of Two Units Per Stratum</td>
<td>276</td>
</tr>
</tbody>
</table>
Example 6.3 PPS (Dollar-Unit) Sampling ........................................ 280

REFERENCES ................................................................. 283
Chapter 6
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
Chapter 6. The SURVEYSELECT Procedure

- 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 6.1 displays the first ten observations of the data set Customers.
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 6.2 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.

<table>
<thead>
<tr>
<th>Customer Satisfaction Survey</th>
</tr>
</thead>
<tbody>
<tr>
<td>The SURVEYSELECT Procedure</td>
</tr>
<tr>
<td>Selection Method</td>
</tr>
<tr>
<td>Input Data Set</td>
</tr>
<tr>
<td>Random Number Seed</td>
</tr>
<tr>
<td>Sample Size</td>
</tr>
<tr>
<td>Selection Probability</td>
</tr>
<tr>
<td>Sampling Weight</td>
</tr>
<tr>
<td>Output Data Set</td>
</tr>
</tbody>
</table>

**Figure 6.2. 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.

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

Figure 6.3 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.
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 6.4 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

<table>
<thead>
<tr>
<th>State</th>
<th>Type</th>
<th>Frequency</th>
<th>Percent</th>
<th>Row Pct</th>
<th>Col Pct</th>
<th>New</th>
<th>Old</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AL</td>
<td></td>
<td>1238</td>
<td>706</td>
<td>1944</td>
<td>9.19</td>
<td>5.24</td>
<td>14.43</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>63.68</td>
<td>36.32</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>14.43</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FL</td>
<td></td>
<td>2170</td>
<td>1370</td>
<td>3540</td>
<td>16.11</td>
<td>10.17</td>
<td>26.28</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>61.30</td>
<td>38.70</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>25.29</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GA</td>
<td></td>
<td>3488</td>
<td>1940</td>
<td>5428</td>
<td>25.89</td>
<td>14.40</td>
<td>40.29</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>64.26</td>
<td>35.74</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>40.65</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SC</td>
<td></td>
<td>1684</td>
<td>875</td>
<td>2559</td>
<td>12.50</td>
<td>6.50</td>
<td>19.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>65.81</td>
<td>34.19</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>19.63</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>8580</td>
<td>4891</td>
<td>13471</td>
<td>63.69</td>
<td>36.31</td>
<td>100.00</td>
<td></td>
</tr>
</tbody>
</table>

Figure 6.4. 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.

```sas
title1 'Customer Satisfaction Survey';
title2 'Stratified Sampling';
proc surveyselect data=Customers method=srs n=15
   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.

Figure 6.5 displays the output from PROC SURVEYSELECT, which summarizes the sample selection. A total of 120 customers are selected.
Stratified Sampling

The SURVEYSELECT Procedure

Selection Method: Simple Random Sampling
Strata Variables: State, Type

<table>
<thead>
<tr>
<th>Input Data Set</th>
<th>CUSTOMERS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Number Seed</td>
<td>1953</td>
</tr>
<tr>
<td>Stratum Sample Size</td>
<td>15</td>
</tr>
<tr>
<td>Number of Strata</td>
<td>8</td>
</tr>
<tr>
<td>Total Sample Size</td>
<td>120</td>
</tr>
<tr>
<td>Output Data Set</td>
<td>SAMPLESTRATA</td>
</tr>
</tbody>
</table>

**Figure 6.5. Sample Selection Summary**

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

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

Figure 6.6 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.
Chapter 6. The SURVEYSELECT Procedure

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

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

Figure 6.6. 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 6.7 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 260 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

<table>
<thead>
<tr>
<th>Selection Method</th>
<th>Systematic Random Sampling</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strata Variable</td>
<td>State</td>
</tr>
<tr>
<td>Control Variables</td>
<td>Type, Usage</td>
</tr>
<tr>
<td>Control Sorting</td>
<td>Serpentine</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Input Data Set</th>
<th>CUSTOMERS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Number Seed</td>
<td>1234</td>
</tr>
<tr>
<td>Stratum Sampling Rate</td>
<td>0.02</td>
</tr>
<tr>
<td>Number of Strata</td>
<td>4</td>
</tr>
<tr>
<td>Total Sample Size</td>
<td>271</td>
</tr>
<tr>
<td>Output Data Set</td>
<td>SAMPLECONTROL</td>
</tr>
</tbody>
</table>
```

Figure 6.7. 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, and maximum size measures in the secondary input data set. See the descriptions of the SAMPSIZE=, SAMPRATE=, MINSIZE=, and MAXSIZE= 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 6.1. PROC SURVEYSELECT Statement Options

<table>
<thead>
<tr>
<th>Task</th>
<th>Options</th>
</tr>
</thead>
<tbody>
<tr>
<td>Specify the input data set</td>
<td>DATA=</td>
</tr>
<tr>
<td>Specify output data sets</td>
<td>OUT=</td>
</tr>
<tr>
<td></td>
<td>OUTSORT=</td>
</tr>
<tr>
<td>Suppress displayed output</td>
<td>NOPRINT</td>
</tr>
<tr>
<td>Specify selection method</td>
<td>METHOD=</td>
</tr>
<tr>
<td>Specify sample size</td>
<td>SAMPsize=</td>
</tr>
<tr>
<td>Specify sampling rate</td>
<td>SAMPrate=</td>
</tr>
<tr>
<td>Specify number of replicates</td>
<td>REP=</td>
</tr>
<tr>
<td>Adjust size measures</td>
<td>MINsize=</td>
</tr>
<tr>
<td></td>
<td>MAXsize=</td>
</tr>
<tr>
<td>Specify sorting type</td>
<td>SORT=</td>
</tr>
<tr>
<td>Specify random number seed</td>
<td>SEED=</td>
</tr>
<tr>
<td>Control OUT= contents</td>
<td>JTPROBS</td>
</tr>
<tr>
<td></td>
<td>OUTSIZE</td>
</tr>
<tr>
<td></td>
<td>STATS</td>
</tr>
</tbody>
</table>

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

**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 270.

**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*, or **MINsize=** *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 264 for details. If you specify METHOD=PPS, you must name the size measure variable in the SIZE statement.

**PPS_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 268 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.
PROC SURVEYSELECT Statement • 253

PSM_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 269 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.

PSM_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 269 for details. If you specify METHOD=PPS_SAMPFORD, you must name the size measure variable in the SIZE statement.

PSM_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 267 for details. If you specify METHOD=PPS_SEQ, you must name the size measure variable in the SIZE statement.

PSM_SYS requests systematic selection with probability proportional to size. See the section “PPS Systematic Sampling” on page 266 for details on this method. If you specify METHOD=PPS_SYS, you must name the size measure variable in the SIZE statement.

PSM_WR requests selection with probability proportional to size and with replacement. See the section “PPS Sampling with Replacement” on page 266 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 263 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 267 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 262 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 262 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 266 for details.

URS requests unrestricted random sampling, which is selection with equal probability and with replacement. See the section “Unrestricted Random Sampling” on page 262 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, or MAXSIZE=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.

**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 the chapter titled “Using the Output Delivery System” in SAS/STAT User’s Guide.

**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\textit{n}, where \textit{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, OUTSIZE, and STATS. For information on the contents of the output data set, see the section “Output Data Set” on page 270.

**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 and the maximum size measure if you specify the MAXSIZE= 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 270.

**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 \textit{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 262 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 262 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 262 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 260 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 following equal probability selection methods: METHOD=SRS, METHOD=URS, METHOD=SYS, and METHOD=SEQ. For PPS selection methods and stratified designs, the output data set contains selection probabilities and sampling weights by default. For more information on the contents of the output data set, see the section “Output Data Set” on page 270.

**CONTROL Statement**

```sas
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 260.

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 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. 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 SAMPSCALE= 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 _NRA TE_ , _MINSIZE_ , and _MAXSIZE_ 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
Sample Selection Methods

PROC SURVEYSELECT provides a variety of methods for selecting probability-based random samples. With probability sampling, each unit in the survey population has a known, positive probability of selection. This property of probability sampling avoids selection bias and enables you to use statistical theory to make valid inferences from the sample to the survey population. 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 ran-
dom 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 \( \frac{N}{n} \) or \( \frac{N_h}{n_h} \) for stratified sampling. The selection probability for each unit equals \( \frac{n}{N} \) or \( \frac{n_h}{N_h} \) 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 of 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 267 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 260 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

$$I_{hi} = \text{Int} \left( \sum_{j=1}^{i} ES_{hi} \right) = \text{Int} \left( i \frac{n_h}{N_h} \right)$$

$$F_{hi} = \text{Frac} \left( \sum_{j=1}^{i} ES_{hi} \right) = \text{Frac} \left( i \frac{n_h}{N_h} \right)$$

where $\text{Int}$ denotes the integer part of the number, and $\text{Frac}$ 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

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

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 nonnegativity 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 \( \theta_i \), where
\[
\theta_i = n \left( Z_{N-n+i+1} - Z_{N-n+i} \right) (T + i Z_{N-n+1}) / T
\]
\[
Z_j = M_j / M, \quad T = \sum_{j=1}^{N-n} Z_j, \quad \text{and, by definition, } Z_{N+1} = 1/n \text{ to ensure that } \sum_{i=1}^{n} \theta_i = 1.
\]

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,
\[
Z_j^* = Z_j / (T + i Z_{N-n+1}) \quad \text{for } j = 1, \ldots, N-n+1
\]
\[
Z_j^* = Z_{N-n+1} / (T + i Z_{N-n+1}) \quad \text{for } j = N-n+2, \ldots, N-n+i
\]

4. The procedure selects the next unit from the first \( (N-n+1) \) stratum units with probability proportional to \( a_j(1) \), where
\[
a_1(1) = i Z_1^* \\
a_j(1) = i Z_j^* \prod_{k=1}^{j-1} [1 - (i - 1) P_k] \quad \text{for } j = 2, \ldots, N-n+1
\]
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

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

where

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

and

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

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 $Z_{hi} = M_{hi} / M_h$. 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 260 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

\[ ES_{hi} = n_h Z_{hi} \]

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

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

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

\[ \frac{(F_{hi} - F_{h(i-1)})}{(1 - F_{h(i-1)})} \]

And the number of hits for unit \( i \) equals \( T_{hi} - T_{h(i-1)} \).

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

\[ \frac{F_{hi}}{F_{h(i-1)}} \]

**Brewer’s PPS Method**

Brewer’s method (METHOD=PPS_BREWER) selects two units from each stratum, with probability proportional to size and without replacement. The selection probability for unit \( i \) in stratum \( h \) equals \( 2M_{hi}/M_h = 2Z_{hi} \).

Brewer’s algorithm first selects a unit with probability

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

where

\[ D_h = \sum_{i=1}^{N_h} \frac{Z_{hi} (1 - Z_{hi})}{1 - 2Z_{hi}} \]
Then a second unit is selected from the remaining units with probability

\[
\frac{Z_{hj}}{1 - Z_{hi}}
\]

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

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

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 \( 2M_{hi} / M_h = 2Z_{hi} \).

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

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

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

\[
P_{hi} = n_h \frac{M_{hi}}{M_h} = n_h Z_{hi}
\]

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

\[
K_h = \left( \sum_{t=1}^{n_h} \frac{t L_{n_h - t}}{n_h^t} \right)^{-1}
\]

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 one observation for each unit selected for the sample. 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.
- **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 to compute joint probabilities of selection 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. This variable is included if you request the MINSIZE= option.
- **MaximumSize**, which is the maximum size measure. This variable is included if you request the MAXSIZE= 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
- 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= option or the SAMPRATE= 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= option
- Sampling Rate Data Set, if you specify the SAMPRATE=SAS-data-set 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 the chapter titled “Using the Output Delivery System” in *SAS/STAT User’s Guide*.

<table>
<thead>
<tr>
<th>ODS Table Name</th>
<th>Description</th>
<th>Statement</th>
<th>Option</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method</td>
<td>Sample Selection Method</td>
<td>PROC</td>
<td>default</td>
</tr>
<tr>
<td>Summary</td>
<td>Sample Selection Summary</td>
<td>PROC</td>
<td>default</td>
</tr>
</tbody>
</table>
Chapter 6. The SURVEYSELECT Procedure

Examples

Example 6.1. Replicated Sampling

This example uses the Customers data set from the section “Getting Started” on page 242. 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.

```sas
title1 'Customer Satisfaction Survey';
title2 'Replicated Sampling';
proc surveyselect data=Customers method=seq
   rep=4 n=(8 12 20 10) 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.

Figure 6.1.1 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 6.1.1. Sample Selection Summary

<table>
<thead>
<tr>
<th>Customer Satisfaction Survey</th>
<th>Replicated Sampling</th>
</tr>
</thead>
<tbody>
<tr>
<td>The SURVEYSELECT Procedure</td>
<td></td>
</tr>
<tr>
<td>Selection Method</td>
<td>Sequential Random Sampling</td>
</tr>
<tr>
<td>With Equal Probability</td>
<td></td>
</tr>
<tr>
<td>Strata Variable</td>
<td>State</td>
</tr>
<tr>
<td>Control Variables</td>
<td>Type, Usage</td>
</tr>
<tr>
<td>Control Sorting</td>
<td>Serpentine</td>
</tr>
<tr>
<td>Input Data Set</td>
<td>CUSTOMERS</td>
</tr>
<tr>
<td>Random Number Seed</td>
<td>40070</td>
</tr>
<tr>
<td>Number of Strata</td>
<td>4</td>
</tr>
<tr>
<td>Number of Replicates</td>
<td>4</td>
</tr>
<tr>
<td>Total Sample Size</td>
<td>200</td>
</tr>
<tr>
<td>Output Data Set</td>
<td>SAMPLEREP</td>
</tr>
</tbody>
</table>

The following PROC PRINT statements display the selected customers for the first stratum, State = ‘AL’, from the output data set SampleRep.

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

Figure 6.1.2 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.
Example 6.2. 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 ';

SAS OnlineDoc™: Version 7-1
Example 6.2. PPS Selection of Two Units Per Stratum

```sas
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`.

```sas
title1 'Hospital Utilization Survey';
title2 'Sampling Frame, Region 1';
proc print data=HospitalFrame;
run;
```
Output 6.2.1.  Sampling Frame

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

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.

```sas
%title 'Hospital Utilization Survey';
proc surveyselect data=HospitalFrame method=pps_brewer
   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 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 6.2.2 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 6.2.2.  Sample Selection Summary

Hospital Utilization Survey

The SURVEYSELECT Procedure

<table>
<thead>
<tr>
<th>Selection Method</th>
<th>Brewer’s PPS Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size Measure</td>
<td>SizeMeasure</td>
</tr>
<tr>
<td>Strata Variables</td>
<td>Type</td>
</tr>
<tr>
<td>Size</td>
<td></td>
</tr>
</tbody>
</table>

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.

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

Output 6.2.3.  Sample Hospitals

<table>
<thead>
<tr>
<th>Obs</th>
<th>Type</th>
<th>Size</th>
<th>Hospital</th>
<th>Size Measure</th>
<th>Selection Prob</th>
<th>Sampling Weight</th>
<th>Jt Selection Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Rural</td>
<td>Small</td>
<td>079</td>
<td>2.127</td>
<td>0.13516</td>
<td>7.39868</td>
<td>0.01851</td>
</tr>
<tr>
<td>2</td>
<td>Rural</td>
<td>Small</td>
<td>236</td>
<td>5.279</td>
<td>0.33545</td>
<td>2.98106</td>
<td>0.01851</td>
</tr>
<tr>
<td>3</td>
<td>Urban</td>
<td>Small</td>
<td>006</td>
<td>4.249</td>
<td>0.17600</td>
<td>5.68181</td>
<td>0.01454</td>
</tr>
<tr>
<td>4</td>
<td>Urban</td>
<td>Small</td>
<td>195</td>
<td>5.024</td>
<td>0.20810</td>
<td>4.80533</td>
<td>0.01454</td>
</tr>
<tr>
<td>5</td>
<td>Urban</td>
<td>Medium</td>
<td>133</td>
<td>46.992</td>
<td>0.41357</td>
<td>2.41795</td>
<td>0.11305</td>
</tr>
<tr>
<td>6</td>
<td>Urban</td>
<td>Medium</td>
<td>218</td>
<td>48.231</td>
<td>0.42448</td>
<td>2.35584</td>
<td>0.11305</td>
</tr>
<tr>
<td>7</td>
<td>Urban</td>
<td>Large</td>
<td>026</td>
<td>61.460</td>
<td>0.63445</td>
<td>1.57617</td>
<td>0.31505</td>
</tr>
<tr>
<td>8</td>
<td>Urban</td>
<td>Large</td>
<td>058</td>
<td>65.931</td>
<td>0.68060</td>
<td>1.46929</td>
<td>0.31505</td>
</tr>
</tbody>
</table>

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 6.3. 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.

```sas
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.

```sas
proc sort data=TravelExpense;
  by Level;
run;
```
Example 6.3.  PPS (Dollar-Unit) Sampling

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

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

**Output 6.3.1.  Sampling Frame**

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

```sas
title1 'Travel Expense Audit';
proc surveyselect data=TravelExpense method=pps
    n=(6 10 4) 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 appear in the TravelExpense data set. The sample size of 6 corresponds to the first stratum, Level = ‘1_Low’, the sample size of 10 corresponds to the second stratum, Level = ‘2_Avg’, and 4 corresponds to the last stratum, Level = ‘3_High’.

Figure 6.3.2 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 6.3.2. Sample Selection Summary**

<table>
<thead>
<tr>
<th>Travel Expense Audit</th>
</tr>
</thead>
<tbody>
<tr>
<td>The SURVEYSELECT Procedure</td>
</tr>
<tr>
<td>Selection Method</td>
</tr>
<tr>
<td>Size Measure</td>
</tr>
<tr>
<td>Strata Variable</td>
</tr>
<tr>
<td>Input Data Set</td>
</tr>
<tr>
<td>Random Number Seed</td>
</tr>
<tr>
<td>Number of Strata</td>
</tr>
<tr>
<td>Total Sample Size</td>
</tr>
<tr>
<td>Output Data Set</td>
</tr>
</tbody>
</table>

The following PROC PRINT statements display the audit sample.

```sas
title1 'Travel Expense Audit Sample';
title2 'Sample Selected by Stratified PPS Design';
proc print data=AuditSample;
run;
```
Output 6.3.3.  Audit Sample

<table>
<thead>
<tr>
<th>Obs</th>
<th>Level</th>
<th>ID</th>
<th>Amount</th>
<th>Selection Prob</th>
<th>Sampling Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1_Low</td>
<td>654</td>
<td>185.60</td>
<td>0.31105</td>
<td>3.21489</td>
</tr>
<tr>
<td>2</td>
<td>1_Low</td>
<td>017</td>
<td>205.48</td>
<td>0.34437</td>
<td>2.90385</td>
</tr>
<tr>
<td>3</td>
<td>1_Low</td>
<td>382</td>
<td>217.85</td>
<td>0.36510</td>
<td>2.73896</td>
</tr>
<tr>
<td>4</td>
<td>1_Low</td>
<td>614</td>
<td>230.56</td>
<td>0.38640</td>
<td>2.58797</td>
</tr>
<tr>
<td>5</td>
<td>1_Low</td>
<td>782</td>
<td>258.10</td>
<td>0.43256</td>
<td>2.31183</td>
</tr>
<tr>
<td>6</td>
<td>1_Low</td>
<td>775</td>
<td>330.54</td>
<td>0.55396</td>
<td>1.80518</td>
</tr>
<tr>
<td>7</td>
<td>2_Avg</td>
<td>784</td>
<td>505.14</td>
<td>0.34623</td>
<td>2.88823</td>
</tr>
<tr>
<td>8</td>
<td>2_Avg</td>
<td>332</td>
<td>540.65</td>
<td>0.37057</td>
<td>2.69853</td>
</tr>
<tr>
<td>9</td>
<td>2_Avg</td>
<td>002</td>
<td>567.89</td>
<td>0.38924</td>
<td>2.56909</td>
</tr>
<tr>
<td>10</td>
<td>2_Avg</td>
<td>239</td>
<td>620.10</td>
<td>0.42503</td>
<td>2.35278</td>
</tr>
<tr>
<td>11</td>
<td>2_Avg</td>
<td>738</td>
<td>670.85</td>
<td>0.45981</td>
<td>2.17479</td>
</tr>
<tr>
<td>12</td>
<td>2_Avg</td>
<td>496</td>
<td>753.30</td>
<td>0.51633</td>
<td>1.93676</td>
</tr>
<tr>
<td>13</td>
<td>2_Avg</td>
<td>425</td>
<td>780.10</td>
<td>0.53470</td>
<td>1.87022</td>
</tr>
<tr>
<td>14</td>
<td>2_Avg</td>
<td>478</td>
<td>806.90</td>
<td>0.55307</td>
<td>1.80810</td>
</tr>
<tr>
<td>15</td>
<td>2_Avg</td>
<td>672</td>
<td>979.66</td>
<td>0.67148</td>
<td>1.49295</td>
</tr>
<tr>
<td>16</td>
<td>2_Avg</td>
<td>139</td>
<td>1183.45</td>
<td>0.81116</td>
<td>1.23280</td>
</tr>
<tr>
<td>17</td>
<td>3_High</td>
<td>568</td>
<td>1670.80</td>
<td>0.64385</td>
<td>1.55316</td>
</tr>
<tr>
<td>18</td>
<td>3_High</td>
<td>263</td>
<td>1893.40</td>
<td>0.72963</td>
<td>1.37056</td>
</tr>
<tr>
<td>19</td>
<td>3_High</td>
<td>285</td>
<td>2020.70</td>
<td>0.77869</td>
<td>1.28421</td>
</tr>
<tr>
<td>20</td>
<td>3_High</td>
<td>486</td>
<td>2580.35</td>
<td>0.99435</td>
<td>1.00568</td>
</tr>
</tbody>
</table>

References


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


