How can I use PROC SURVEYSELECT in SAS? Can you provide examples?

Authored by

stats writer

June 24, 2024

PROC SURVEYSELECT is a SAS procedure that allows users to select a sample from a larger dataset based on specific criteria. This procedure is commonly used in survey sampling and can be useful for obtaining representative subsets of data for analysis. To use PROC SURVEYSELECT, users need to specify the sample size, sampling method, and other relevant parameters. Examples of how to use this procedure include selecting a random sample of individuals from a population or selecting a stratified sample based on certain characteristics. Overall, PROC SURVEYSELECT is a powerful tool for obtaining samples from large datasets and can be easily customized to fit different sampling needs.

Use PROC SURVEYSELECT in SAS (With Examples)

You can use PROC SURVEYSELECT to select a random sample from a dataset in SAS.

Here are three common ways to use this procedure in practice:

Example 1: Use PROC SURVEYSELECT to Select Simple Random Sample

```
proc surveyselectdata=my_data
out=my_sample
method=srs /*use simple random sampling*/n=5 /*select a total of 5 observations*/seed=1; /*set seed to make this example reproducible*/run;
```

This particular example selects 5 random observations from the entire dataset.

Example 2: Use PROC SURVEYSELECT to Select Stratified Random Sample

```
proc surveyselectdata=my_data
out=my_sample
method=srs /*use simple random sampling*/n=2 /*select 2 observations from each strata*/seed=1; /*set seed to make this example reproducible*/
strata grouping_var; /*specify variable to use for stratification*/run;
```

This particular example selects 2 random observations from each unique stratum in the dataset.

The strata statement specifies the variable to use for stratification.

Example 3: Use PROC SURVEYSELECT to Select Clustered Random Sample

```
proc surveyselectdata=my_data
out=my_sample
n=2 /*select 2 clusters*/seed=1; /*set seed to make this example reproducible*/
cluster grouping_var; /*specify variable to use for
```

stratification*/run;

This particular example selects 2 random clusters from the dataset and includes every observation from each cluster in the sample.

The cluster statement specifies the variable to use for clustering.

The following examples show how to use each method in practice with the following dataset in SAS that contains information about basketball players on various teams:

```
/*create dataset*/
data my_data;
input team $ points;
datalines;
A 12
A 14
A 22
A 35
A 40
B 12
B 10
```

```
B 29
B 33
C 40
C 25
C 11
C 10
C 15
;
run;

/*view dataset*/
proc printdata = my_data;
```

Obs	team	points
1	A	12
2	A	14
3	A	22
4	A	35
5	A	40
6	B	12
7	B	10
8	B	29
9	B	33
10	C	40
11	C	25
12	C	11
13	C	10
14	C	15

Example 1: Use PROC SURVEYSELECT to Select Simple Random Sample

```
proc surveyselectdata=my_data
out=my_sample
method=srs /*use simple random sampling*/n=5 /*select a total of 5 observations*/seed=1; /*set seed to make this example reproducible*/run;

/*view sample*/proc printdata=my_sample;
```

Obs	team	points
1	B	10
2	B	33
3	C	40
4	C	25
5	C	11

The resulting sample contains 5 randomly chosen from the entire dataset.

Example 2: Use PROC SURVEYSELECT to Select Stratified Random Sample

We can use the following syntax to perform stratified random sampling in which 2 observations are randomly chosen from each team to be included in the sample:

```
proc surveyselectdata=my_data
out=my_sample
method=srs /*use simple random sampling within strata*/n=2 /*select 2 observations from each strata*/seed=1; /*set seed to make this example reproducible*/
strata grouping_var; /*specify variable to use for stratification*/run;

/*view sample*/proc printdata=my_sample;
```

Obs	team	points	SelectionProb	SamplingWeight
1	A	35	0.4	2.5
2	A	40	0.4	2.5
3	B	29	0.5	2.0
4	B	33	0.5	2.0
5	C	25	0.4	2.5
6	C	10	0.4	2.5

The resulting sample contains 2 randomly chosen from each team.

Example 3: Use PROC SURVEYSELECT to Select Clustered Random Sample

We can use the following syntax to perform clustered

random sampling in which we use the teams as clusters and randomly select 2 clusters and include each observation from those clusters in the sample:

```
proc surveyselectdata=my_data
out=my_sample
n=2 /*select a total of 2 clusters*/seed=1; /*set seed to make this example reproducible*/
cluster grouping_var; /*specify variable to use for clustering*/run;

/*view sample*/proc printdata=my_sample;
```

Obs	team	points
1	A	12
2	A	14
3	A	22
4	A	35
5	A	40
6	B	12
7	B	10
8	B	29
9	B	33

This particular sample contains every observation from teams A and B, which were the two "clusters" randomly

chosen.

Note: You can find the complete documentation for PROC SURVEYSELECT .

The following tutorials explain how to perform other common tasks in SAS: