Sampling in SAS ® using PROC SURVEYSELECT

Drew Doyle and Rachael Becker

Abstract

This paper examines the various sampling options that are available in SAS ® through PROC SURVEYSELECT. We will not be covering all of the possible sampling methods or options that SURVEYSELECT features. Instead we will look at Simple Random Sampling, Stratified Random Sampling, Cluster Sampling, Systematic Sampling, and Sequential Random Sampling.

Introduction

Sampling is an essential part of statistics, there are many ways to take samples using SAS.

This paper will discuss the different sampling options that are available through the use of PROC SURVEY.

PROC SURVEYSELECT is an essential tool because it allows statisticians to sample finite populations and draw accurate conclusions when appropriate samples are taken.

Data Set

The data sets that we will be drawing samples from for all of our examples are included at the end of this paper.

Simple Random Sampling

What is simple random sampling without replacement?

Simple random sampling without replacement is the process of sampling that gives each observation the same probability of being selected. After a unit is selected it cannot be selected again. If you had a data set with 93 observations and wanted to take a sample of 6 observations, then the total number of possible samples is 762245484.

How do you do it in SAS using PROC SURVEY SELECT?

Suppose you have a data set that contains students' ID number, the year they are in college, their grade in a course, and what section of the course they attended. You want to know how course grade relates to the amount of financial aid the student is receiving, but you don't have the time to collect the information on all of the students so you want to take a random sample of the students and use the results to get further information.

To take this sample in SAS, your code would look like this:

```
Proc SurveySelect
      data = Example
      method = srs
      n = 15
      out = Example_SRS
      seed = 50460
;
Run;
```

Where the data option specifies the data set that you want the sample to be taken from, the METHOD option (SRS) is simple random sampling, the number of observations you want in your sample is

N, the OUT option indicates the name for the sample data set, and the optional SEED option is used for replication purposes. If you forget to use the SEED option on your first run SAS will tell you the seed that was used to create the sample.

Two things will happen when the above proc is run:

1.) A data set called Exampl_SRS will be created that has 15 observations and contains all of the variables for those observations that existed in the Example data set.

2.) The result viewer will display the following table:

Selection Method	Simple Random Sampling

Input Data Set	EXAMPLE
Random Number Seed	50460
Sample Size	15
Selection Probability	0.3
Sampling Weight	3.333333
Output Data Set	EXAMPLE_SRS

What is simple random sampling with replacement?

Simple random sampling can also mean that each random sample taken has the same probability of being created. If you have a data set with 93 observations in the population and only want a sample that contains 6 observations, and you take a sample with replacement, the number of possible samples is 646990183449 (note this is a much larger number than the number of possible samples for SRS without replacement), and all of these samples have the same probability of being chosen.

How do you do it in SAS using PROC SURVEY SELECT?

```
Proc SurveySelect
      data = Example
      method = urs
      n = 15
      out = Example_SRS_replacement
      seed = 50460
      outhits
;
Run;
```

The URS method specifies that it is unrestricted random sampling. Note that even though the same seed was used for both codes, a different sample was selected because the sampling method was changed. Also notice the extra option OUTHITS, without this option data sets created using URS may or may not have the number of observations specified in the N option. This is because some of the observations selected may be duplicates, if you want the duplicate observations to appear in the data set then the OUTHITS option needs to be used. The new data set Example_SRS_replacement has a new variable NumberHits which tells how many times an observation was duplicated.

Stratified Random Sampling

What is stratified random sampling?

Stratification is the process of sampling within smaller subgroups, or stratum, of a larger population. In stratified random sampling after the population is divided

into their perspective stratum a simple random sampling method without replacement is applied.

Why stratify?

Stratification may help to improve the accuracy of an estimate and its appropriateness depends on the data that is being analyzed. The biggest reason for the support of stratification is that it takes a skewed data set and can break it down into less skewed groups. In other words, it is a way of taking data that is dissimilar and creating smaller groups that are similar.

How do you stratify?

There are different ways to stratify within a data set. You can stratify by numeric variables if you create groups for the stratum, you can also stratify by character variables. For example if we are analyzing a data set that has home type as one variable and square footage as another variable, then it may be best to stratify the data by home type before trying to find any estimates about the population because we can assume that houses will often have more square footage than apartments. Not stratifying in this example could produce a larger variance because there could be a lot of variability across the strata.

How do you do it in SAS using PROC SURVEY SELECT?

The code for stratified random sampling is similar to the code for random sampling. The main difference is that now there is an extra option added called STRATA. After using the option statement you will need to list the variable that contains the strata that the data should be separated by. If you are stratifying by a numeric value, you first need to create a new variable and then use IF THEN logic to separate the data into the correct stratum. For our example we will be using character variables by which to stratify.

It is necessary to sort the data by the strata before the sample is taken.

```
Proc Sort data = Example;
      by Year Class;
Run;
```

This example is considered a bad example because there are ERRORS in the log, but we chose to show it because it still produces output, this highlights the necessity of checking the log.

```
/*bad example*/
Proc SurveySelect
      data = Example
      method = srs
      n = 2
      out =
Example_Stratification_bad
      seed = 52988
;
      strata Year Class;
Run;
```

```
89   /*bad example*/
90   Proc SurveySelect
91       data = Example
92       method = srs
93       n = 2
94       out = Example_Stratification_bad
95       seed = 52988
96   ;
97       strata Year Class;
98   Run;

NOTE: The sample size equals the number of
sampling units. All units are included in
the sample.
```

Input Data Set	EXAMPLE
Random Number Seed	52988
Stratum Sample Size	2
Number of Strata	11
Total Sample Size	22
Output Data Set	EXAMPLE_STRATIFICATION_BAD

```
NOTE: The above message was for the
following stratum:
      Year=Freshman Class=1.
NOTE: The sample size equals the number of
sampling units. All units are included in
the sample.
NOTE: The above message was for the
following stratum:
      Year=Freshman Class=2.
ERROR: The sample size, 2, is greater than
the number of sampling units, 1.
NOTE: The above message was for the
following stratum:
      Year=Freshman Class=3.
NOTE: The SAS System stopped processing this
step because of errors.
WARNING: The data set
WORK.EXAMPLE_STRATIFICATION_BAD may be
incomplete.  When this step was
         stopped there were 22 observations
and 6 variables.
NOTE: PROCEDURE SURVEYSELECT used (Total
process time):
      real time           0.17 seconds
      cpu time            0.10 seconds
```

The notes above inform the user that there were problems taking samples from the Freshman class. The code told SAS to stratify by the variables Year and Class, the order of the variables in the STRATA statement signifies how the stratification will be done. First there will be strata for the four different years and then within the years, there will be a strata for the three class groups. When N was specified at 2 the problem was caused because the number of Freshman in class three is only one, so the sample size exceeded the population size. Freshman for class three will not be represented in the output data set.

The result viewer will output the following information:

Selection Method	Simple Random Sampling
Strata Variables	Year
	Class

A better example of stratification is listed below. Notice how the sample size, N, does not exceed the number of observations within each strata for the population.

```
/*good example*/
Proc SurveySelect
      data = Example
      method = srs
      n = 3
      out =
Example_Stratification_good
      seed = 62493
;
      strata Year;
Run;
```

This is the output generated in the result viewer. This table gives the details about the data set that was created.

Selection Method	Simple Random Sampling
Strata Variable	Year

Input Data Set	EXAMPLE
Random Number Seed	62493
Stratum Sample Size	3
Number of Strata	4
Total Sample Size	12
Output Data Set	EXAMPLE_STRATIFICATION_GOOD

The data set that was created contains two new variables: SelectionProb and SamplingWeight.

There are also options available to specify specific sample sizes for each stratum, we will not explore that option in this paper.

Cluster Sampling

What is cluster sampling?

According to ISO/FDIS 3534-4, Cluster sampling is part of a population divided into mutually exclusive groups related in a certain manner. So for our example, we were looking for the average price of textbooks, and instead of surveying 30 people and only having thirty observations, we could use cluster sampling, using students ID number as the grouping variable, and sample only fifteen students, but have 60 data points (each student had the price for four textbooks).

Cluster sampling uses a simple random sample to select a group and all items within the group are selected. Cluster sampling is used because it can be a cheaper way to get more data. For example if a researcher wants to find out how much the average college textbook costs, they could take a simple random sample and get one response for each person that they sample or they could use cluster sampling and find out the cost of all of the textbooks purchased by that student. Using clustering sampling could mean that more data is obtained without the hassle of dealing with more observations.

How do you do it in SAS using PROC SURVEY SELECT?

The coding for cluster sampling in SAS is the same as the code for simple random sampling without replacement. However, it is necessary to add the samplingunit statement in order to specify which variable the clustering occurs on. For our example the data is clustered by observation or the students ID number (IDNo).

```
Proc SurveySelect
      data = Example2
      method = srs
      sampsize = 5
      out = Example_Clustering
      seed = 7162010
;
      samplingunit IDNo
;
Run;
```

Selection Method	Simple Random Sampling
Sampling Unit Variable	IDNo

Input Data Set	EXAMPLE2
Random Number Seed	7162010
Sample Size	5
Selection Probability	0.1

Sampling Weight	10
Output Data Set	EXAMPLE_CLUSTERING

Systematic Sampling

What is systematic sampling?

Systematic sampling selects items by taking every nth observation. SAS uses the following formula to decide on how to determine what iteration it uses.

$$K = \frac{N}{n}$$

$$Kth = \frac{Total\ \#\ in\ the\ Population}{\#\ of\ Observation\ in\ the\ Sample}$$

How do you do it in SAS using PROC SURVEY SELECT?

The METHOD option used for systematic random sampling is SYS.

```
Proc SurveySelect
      data = Example3
      method = sys
      n = 15
      out = Example_Systematic
      seed = 31636
;
Run;
```

Selection Method	Systematic Random Sampling

Input Data Set	EXAMPLE3
Random Number Seed	31636
Sample Size	15
Selection Probability	0.3
Sampling Weight	3.333333
Output Data Set	EXAMPLE_SYSTEMATIC

Sequential Random Sampling

What is sequential random sampling?

Sequential random sampling is the method of sampling that spreads the data out throughout the strata. This method of sampling takes the population size of each stratum into account. The basic difference between sequential random sampling and stratified random sampling is that sequential random sampling distributes the data to each strata appropriately without having to add extra options (it is possible to achieve this using stratified random sampling, but it would require the calculation of the appropriate proportions for n and then a special option statement that indicates the level of n desired for each stratum).

How do you do it in SAS using PROC SURVEY SELECT?

The METHOD option for sequential sampling is SEQ. Using the Option SORT = NEST the PROC will do nested sorting eliminating the PROC SORT statement. Also, CONTROL and STRATA statements are available with this method of sampling. If you utilize the SORT option then a CONTROL statement is required. The STRATA statement tells SAS to take a sample within the groups specified by the statement.

```
Proc SurveySelect
      data = Example3
      method = seq
      n = 1
      out = Example_Sequential
      seed = 31636
      sort = nest
;
      control Name;
      strata NoSib;
```

```
Run;
```

Selection Method	Sequential Random Sampling
	With Equal Probability
Strata Variable	NoSib
Control Variable	Name

Input Data Set	EXAMPLE3
Random Number Seed	31636
Stratum Sample Size	1
Number of Strata	8
Total Sample Size	8
Output Data Set	EXAMPLE_SEQUENTIAL

Conclusion

PROC SURVEYSELECT is an essential tool because it allows statisticians to obtain samples that are appropriate for statistical analysis.

PROC SURVEYSELECT has many more options and uses than what was described in this paper. There are many other papers and texts which should be read if you really want to understand all of the options and applications of PROC SURVEYSELECT.

References

24555 - Using PROC SURVEYSELECT for single-stage cluster sampling. (n.d.). Retrieved July 17, 2015, from http://support.sas.com/kb/24/555.html

An, A. and Watts, D. (2000). New SAS Procedures for Analysis of Sample Survey Data. SUGI 23, Retrieved from http://www2.sas.com/proceedings/sugi23/Stats/p247.pdf

Diseker, R. and Permanente, K. (2004). Simplified Matched Case-Control Sampling using PROC SURVEYSELECT. SUGI 209-29, Retrieved from http://www2.sas.com/proceedings/sugi29/209-29.pdf

Frerichs, R.R. Rapid Surveys (unpublished), 2008, Retrieved from http://www.ph.ucla.edu/epi/rapidsurveys/RScourse/RSbook_ch3.pdf

Hadden, L. (2005). PROC SURVEYSELECT: A Simply Serpentine Solution for Complex Sample Designs. NESUGI 18, Retrieved from http://www.nesug.org/proceedings/nesug05/an/an5.pdf

Putnam, D. (2011). PROC SURVEY…Says!: Selecting and Analyzing Stratified Samples. SESUGI ST-05, Retrieved from http://analytics.ncsu.edu/sesug/2011/ST05.Putnam.pdf

Suhr, D. (2009). Selecting a Stratified Sample with PROC SURVEYSELECT. SUGI 058-2009, Retrieved from http://support.sas.com/resources/papers/proceedings09/058-2009.pdf

Thompson, Steven K. *Sampling*. 3rd ed. Hoboken, N.J.: John Wiley & Sons, 2012. Print.

Unknown Author, Retrieved from, http://www.math.wpi.edu/saspdf/stat/chap63.pdf

Contact Information

Drew Doyle:

Email: Drewdoyle@knights.ucf.edu

Rachael Becker:

Email: Leahcarbecker@knights.ucf.edu

Acknowledgements

We would like to thank Dr. Mark Johnson for sparking our interest in sampling methods. We would also like to say a special thank you to Kelcey Ellis, without her this paper would not exist.

Data Sets:

Example:

Obs	IDNo	Year	FinalGrade	Class
1	646	Freshman	50	1
2	986	Freshman	20	1
3	498	Freshman	35	2
4	516	Freshman	56	2
5	094	Freshman	61	3
6	132	Junior	85	1
7	103	Junior	86	1
8	464	Junior	52	1
9	984	Junior	78	1
10	140	Junior	48	2
11	180	Junior	57	2
12	949	Junior	54	2
13	098	Junior	68	3
14	815	Junior	87	3
15	849	Junior	88	3
16	401	Junior	64	3
17	979	Senior	46	1
18	239	Senior	48	1
19	059	Senior	64	1
20	169	Senior	80	1
21	868	Senior	60	1
22	462	Senior	36	1
23	724	Senior	70	1
24	252	Senior	91	1
25	907	Senior	35	2
26	432	Senior	64	2
27	279	Senior	95	2
28	754	Senior	92	2
29	041	Senior	84	2
30	630	Senior	64	2
31	674	Senior	45	2
32	970	Senior	73	2

Obs	IDNo	Year	FinalGrade	Class
33	075	Senior	86	2
34	937	Senior	45	3
35	049	Senior	42	3
36	327	Senior	85	3
37	920	Senior	91	3
38	016	Senior	81	3
39	540	Senior	61	3
40	270	Senior	59	3
41	497	Senior	72	3
42	818	Sophomore	74	1
43	069	Sophomore	68	1
44	209	Sophomore	87	1
45	190	Sophomore	75	2
46	841	Sophomore	98	2
47	747	Sophomore	85	2
48	013	Sophomore	72	3
49	641	Sophomore	67	3
50	068	Sophomore	78	3

Example_SRS:

Obs	IDNo	Year	FinalGrade	Class
1	986	Freshman	20	1
2	180	Junior	57	2
3	949	Junior	54	2
4	401	Junior	64	3
5	907	Senior	35	2
6	327	Senior	85	3
7	868	Senior	60	1
8	540	Senior	61	3
9	674	Senior	45	2
10	724	Senior	70	1
11	497	Senior	72	3

Obs	IDNo	Year	FinalGrade	Class
12	252	Senior	91	1
13	818	Sophomore	74	1
14	841	Sophomore	98	2
15	068	Sophomore	78	3

Example_SRS_Replacement:

Obs	IDNo	Year	FinalGrade	Class	NumberHits
1	986	Freshman	20	1	1
2	464	Junior	52	1	1
3	907	Senior	35	2	3
4	907	Senior	35	2	3
5	907	Senior	35	2	3
6	041	Senior	84	2	1
7	462	Senior	36	1	1
8	724	Senior	70	1	1
9	970	Senior	73	2	1
10	818	Sophomore	74	1	1
11	190	Sophomore	75	2	1
12	641	Sophomore	67	3	1
13	069	Sophomore	68	1	2
14	069	Sophomore	68	1	2
15	747	Sophomore	85	2	1

Example_Stratification_bad:

Obs	Year	Class	IDNo	FinalGrade	SelectionProb	SamplingWeight
1	Freshman	1	646	50	1.00000	1.0
2	Freshman	1	986	20	1.00000	1.0
3	Freshman	2	498	35	1.00000	1.0

Obs	Year	Class	IDNo	FinalGrade	SelectionProb	SamplingWeight
4	Freshman	2	516	56	1.00000	1.0
5	Junior	1	132	85	0.50000	2.0
6	Junior	1	984	78	0.50000	2.0
7	Junior	2	180	57	0.66667	1.5
8	Junior	2	949	54	0.66667	1.5
9	Junior	3	815	87	0.50000	2.0
10	Junior	3	401	64	0.50000	2.0
11	Senior	1	239	48	0.25000	4.0
12	Senior	1	252	91	0.25000	4.0
13	Senior	2	754	92	0.22222	4.5
14	Senior	2	041	84	0.22222	4.5
15	Senior	3	270	59	0.25000	4.0
16	Senior	3	497	72	0.25000	4.0
17	Sophomore	1	069	68	0.66667	1.5
18	Sophomore	1	209	87	0.66667	1.5
19	Sophomore	2	190	75	0.66667	1.5
20	Sophomore	2	747	85	0.66667	1.5
21	Sophomore	3	641	67	0.66667	1.5
22	Sophomore	3	068	78	0.66667	1.5

Example_Stratification_good:

Obs	Year	IDNo	FinalGrade	Class	SelectionProb	SamplingWeight
1	Freshman	646	50	1	0.60000	1.66667
2	Freshman	516	56	2	0.60000	1.66667
3	Freshman	094	61	3	0.60000	1.66667
4	Junior	949	54	2	0.27273	3.66667
5	Junior	815	87	3	0.27273	3.66667
6	Junior	849	88	3	0.27273	3.66667
7	Senior	868	60	1	0.12000	8.33333

Obs	Year	IDNo	FinalGrade	Class	SelectionProb	SamplingWeight
8	Senior	674	45	2	0.12000	8.33333
9	Senior	075	86	2	0.12000	8.33333
10	Sophomore	841	98	2	0.33333	3.00000
11	Sophomore	013	72	3	0.33333	3.00000
12	Sophomore	641	67	3	0.33333	3.00000

Example2:

Obs	IDNo	Year	PriceBook1	PriceBook2	PriceBook3	PriceBook4
1	646	Freshman	150	234	342	203
2	498	Freshman	134	205	50	73
3	094	Freshman	98	56	52	68
4	986	Freshman	125	95	98	75
5	516	Freshman	200	230	193	50
6	098	Junior	68	83	103	112
7	132	Junior	85	14	41	5
8	140	Junior	48	286	183	45
9	815	Junior	87	75	57	174
10	103	Junior	86	168	30	45
11	180	Junior	57	178	45	20
12	849	Junior	88	22	154	160
13	464	Junior	52	45	50	67
14	949	Junior	54	174	287	40
15	401	Junior	64	45	150	40
16	984	Junior	78	293	268	98
17	907	Senior	35	293	14	63
18	937	Senior	45	50	154	98
19	979	Senior	46	114	180	160
20	432	Senior	64	50	172	200
21	049	Senior	42	13	40	80
22	239	Senior	48	154	114	64
23	279	Senior	95	172	50	190

Obs	IDNo	Year	PriceBook1	PriceBook2	PriceBook3	PriceBook4
24	327	Senior	85	172	60	77
25	059	Senior	64	114	40	80
26	754	Senior	92	57	154	190
27	920	Senior	91	204	12	200
28	169	Senior	80	293	40	50
29	041	Senior	84	46	39	190
30	016	Senior	81	103	180	74
31	868	Senior	60	268	277	80
32	630	Senior	64	112	212	70
33	540	Senior	61	97	147	55
34	462	Senior	36	297	48	150
35	674	Senior	45	225	125	40
36	270	Senior	59	161	102	51
37	724	Senior	70	90	84	100
38	970	Senior	73	81	154	25
39	497	Senior	72	100	150	60
40	252	Senior	91	65	98	70
41	075	Senior	86	15	10	15
42	013	Sophomore	100	25	162	72
43	818	Sophomore	74	150	68	70
44	190	Sophomore	75	65	90	230
45	641	Sophomore	67	50	80	95
46	069	Sophomore	68	50	95	120
47	841	Sophomore	98	80	25	30
48	068	Sophomore	78	250	60	75
49	209	Sophomore	87	89	67	61
50	747	Sophomore	125	40	50	85

Example_Clustering:

Obs	IDNo	Year	PriceBook1	PriceBook2	PriceBook3	PriceBook4
1	401	Junior	64	45	150	40
2	462	Senior	36	297	48	150

Obs	IDNo	Year	PriceBook1	PriceBook2	PriceBook3	PriceBook4
3	630	Senior	64	112	212	70
4	641	Sophomore	67	50	80	95
5	815	Junior	87	75	57	174

Example3:

Obs	Name	NoSib
1	Sarah	3
2	Samantha	2
3	Michael	4
4	Chad	3
5	Carol	2
6	Justin	1
7	Eric	6
8	Alissa	3
9	Joel	2
10	Kathy	4
11	John	7
12	Mark	4
13	Tracy	3
14	Michelle	1
15	David	2
16	Caleb	0
17	Daniel	4
18	Jonah	2
19	Kristen	0
20	Meaghan	5
21	Taylor	1
22	Courtney	0
23	Nicole	2
24	Marie	1
25	Dennis	5

Obs	Name	NoSib
26	Denise	2
27	Lyn	3
28	Carlton	6
29	Robert	0
30	Franklin	3
31	Chester	2
32	Myron	5
33	Marilyn	1
34	Chance	2
35	Ryan	5
36	Jenny	2
37	Samuel	3
38	Richard	4
39	Grant	2
40	Spencer	2
41	Walter	0
42	Rob	1
43	Reed	0
44	Jacob	1
45	Alex	3
46	Allen	1
47	Martin	6
48	Anthony	3
49	Jeff	3
50	James	2

Example_Systematic:

Obs	Name	NoSib
1	Michael	4
2	Eric	6
3	Kathy	4

Obs	Name	NoSib
4	Tracy	3
5	Daniel	4
6	Meaghan	5
7	Nicole	2
8	Lyn	3
9	Franklin	3
10	Marilyn	1
11	Samuel	3
12	Spencer	2
13	Reed	0
14	Martin	6
15	James	2

Example_Sequential:

Obs	NoSib	Name	SelectionProb	SamplingWeight
1	0	Kristen	0.16667	6
2	1	Marilyn	0.12500	8
3	2	Grant	0.07692	13
4	3	Tracy	0.10000	10
5	4	Richard	0.20000	5
6	5	Dennis	0.25000	4
7	6	Carlton	0.33333	3
8	7	John	1.00000	1