

Selecting a Stratified Sample with PROC SURVEYSELECT

Diana Suhr, University of Northern Colorado

Abstract

After seeing a presentation on PROC SURVEYSELECT at SAS® Colorado Day 2002, a routine was developed to select stratified samples determined by population parameters. SAS code and examples will be shown to select samples stratified on 1, 2, and 3 variables. The routine uses PROC FREQ and PROC SURVEYSELECT with STRATA.

Introduction

Selecting random samples representative of the population is essential for research studies. Definitions, a checklist for conducting a survey, and examples of selecting stratified random samples are provided in this paper. Annotated examples shown determine sample size for each strata and stratify on 1, 2, and 3 variables. Before PROC SURVEYSELECT was available, the *ranuni* function with several data steps was used to obtain stratified samples. Appendix A illustrates a *ranuni* method to select stratified samples.

Sampling

A sample is a group selected from a population. Inferences about a population can be made from information obtained in a sample when the sample is representative of the population. Samples based on planned randomness are called probability samples. Probability sampling has a certain amount of randomness built in so that bias or unbiasedness can be established and probability statements can be made about the accuracy of the methods (Scheaffer, Mendenhall, & Ott, 1996). Randomization inherent in probability sampling helps balance out variables that cannot be controlled or measured directly.

Simple random sampling consists of selecting a group of n units such that each sample of n units has the same chance of being selected.

Stratified random sampling occurs when the population is divided into groups, or strata, according to selected variables (e.g., gender, income) and a simple random sample is selected from each group.

Ratio estimators use responses from variables of interest incorporated with responses from an auxiliary variable (e.g., ratio of entertainment expense to total household expense when estimating the average yearly amount spent on entertainment).

Cluster sampling takes a simple random sample of groups and then samples items within the selected clusters.

Systematic sampling selects every n th observation in a list (e.g., every 10th or 15th name).

Unlike simple random sampling, *quota sampling* selects subjects one at a time until desired percentages are reached. Polls of the 1948 U.S. presidential election illustrate an example of quota sampling. Respondents were chosen according to gender, age, income, education, and factors related to political views. However, the polls underestimated the popularity of Harry Truman and overestimated the popularity of Thomas E. Dewey because Republicans were over represented in the poll. It is impossible to control for all variables in quota sampling.

Convenience sampling results when a group of people are selected because they are available. This type of sampling could limit inferences, result in bias and provide a sample unrepresentative of the population.

Planning a Survey

The following checklist could be followed when planning, administering, and analyzing a survey.

- 1) Statement of objectives. State objectives clearly and concisely. Refer to objectives regularly in the design, implementation, and analysis of the survey.
- 2) Measurement instrument. Select an appropriate measurement instrument(s) to answer research questions and meet objectives.
- 3) Data analysis. Outline the analyses to answer research questions/objectives.
- 4) Sample design. Define the target population and sampling variables. Choose a sample design so the sample provides sufficient information to meet objectives of the survey.
- 5) Method of measurement. Determine methods of measurement (e.g., interview, mailed questionnaire, direct observation, web survey).
- 6) Selection and training of survey administrators. Teach those collecting data/administering survey how to properly and accurately collect data.
- 7) Data organization. A plan is necessary for small or large surveys. The organizational plan includes data management and a codebook.
- 8) Pilot study. Provides an opportunity to field-test measurement instrument, survey administrators, management of survey and make modifications.

Sample selection can be accomplished easily with PROC SURVEYSELECT.

PROC SURVEYSELECT SYNTAX

```
PROC SURVEYSELECT <options>;  
    SIZE variable;  
    STRATA variables;  
    CONTRAL variables;  
    ID variables;
```

The sample selection method by default is SRS, simple random sampling. When SIZE is specified, METHOD=PPS. Sample size is indicated with SAMPSIZE= n or SAMPSIZE= SAS-data-set.

STRATA is similar to a BY variables statement and the input data set must be sorted by STRATA variables.

CONTROL lists variables to sort the input data set. If STRATA is specified, input data is sorted by control variables within STRATA.

SIZE names one and only one size variable and contains sizes to be used with probability proportional to size methods.

ID lists identifier variables to be included in the OUT= data set.

Formatting Data

```
PROC FORMAT;
  VALUE LVLFMT
    1='FRESHMAN'
    2='SOPHOMORE'
    3='JUNIOR'
    4='SENIOR';
  VALUE COLGFMT
    1 = 'ARTS & SCI'
    2 = 'EDUCATION'
    3 = 'HHS'
    4 = 'BUSINESS'
    5 = 'PVA'
    6 = 'GRAD SCH'
    7 = 'UNDECLARED';
```

Reading Data

```
DATA RAWSUB;
  INFILE RAWSUB;          - - - - ->
  INPUT ID 1-4
    LEVEL 6
    GEND $8
    MAJCOLG 27;
  FORMAT LEVEL LVLFMT.
    MAJCOLG COLGFMT.;
```

Example #1

```
PROC FREQ DATA = RAWSUB;  - - - - ->
  TABLES GEND/OUT=NEWFREQ NOPRINT;
DATA NEWFREQ2 ERROR;
  SET NEWFREQ;
  SAMPNUM=(PERCENT * 500)/100; - - - ->
  _NSIZE_= ROUND(SAMP,1);
  SAMPNUM=ROUND(SAMPNUM,.01);
  IF _NSIZE_=0 THEN OUTPUT ERROR;
  IF _NSIZE_=0 THEN DELETE;
OUTPUT NEWFREQ2;
DATA NEWFREQ3;
  SET NEWFREQ2;
  KEEP GEND _NSIZE_;      - - - - ->
PROC SORT DATA = NEWFREQ3;
  BY GEND;
PROC SORT DATA = RAWSUB;
  BY GEND;
PROC SURVEYSELECT DATA=RAWSUB
  OUT=SAMPFL
  SAMPSIZE=NEWFREQ3;
  STRATA GEND;
  ID ID GEND;
PROC FREQ DATA = SAMPFL;
  TABLES GEND/OUT=SAMPFREQ NOPRINT;
PROC PRINT DATA=SAMPFREQ;
TITLE 'SAMPLE FREQUENCIES';
PROC PRINT DATA = ERROR;
TITLE 'STRATA DELETED';
PROC DELETE DATA = NEWFREQ NEWFREQ2
  NEWFREQ3 SAMPFL SAMPFREQ ERROR;
```

Annotations

Formatting Data

The PROC FORMAT statement creates “lvlfmt” to describe level (classification) as freshman, sophomore, junior, or senior and “colgfmt” to describe major college as arts & sciences, education, health & human sciences, business, performing & visual arts, graduate school or undeclared.

Reading Data

Data is read from an external file. Formats are “attached” in the data step. A format statement could be included in a procedure rather than in the data step.

Example #1

PROC FREQ calculates gender frequencies and percentages for the total population (data=rawsub) and are not printed (noprnt).

Strata sizes are determined in a DATA step. Sample size is 500 in this example. PROC SURVEYSELECT options SAMPSIZE= specifies the name of the data set containing sample sizes. _NSIZE_, specifies sample size, must be a positive integer, and is rounded off to an integer in the data step. If _NSIZE_ is not a positive integer, it is deleted from the sample size data set and an “error” data set is created.

Gender and sample/strata sizes are kept to read into the PROC SURVEYSELECT procedure.

The sample/strata size data set and the population data set are sorted by gender.

PROC SURVEYSELECT stratifies on gender, creates an output data set named “SAMPFL”, and keeps identification variables “ID” and “GENDER”.

Frequencies are output and not printed with PROC FREQ. Values, counts, and percentages are printed with PROC PRINT.

If sample frequencies are equal to zero, an error message is printed.

Data sets are deleted with PROC DELETE.

Example #2

```
PROC FREQ DATA = RAWSUB;
  TABLES LEVEL*GEND
  /OUT=NEWFREQ NOPRINT;
DATA NEWFREQ2 ERROR;
  SET NEWFREQ;
  SAMPNUM=(PERCENT * 500)/100;
  _NSIZE_= ROUND(SAMPNUM,1);
  SAMPNUM=ROUND(SAMPNUM, .01);
  IF _NSIZE_=0 THEN OUTPUT ERROR;
  IF _NSIZE_=0 THEN DELETE;
  OUTPUT NEWFREQ2;
DATA NEWFREQ3;
  SET NEWFREQ2;
  KEEP LEVEL GEND _NSIZE_;
PROC SORT DATA = NEWFREQ3;
  BY LEVEL GEND;
PROC SORT DATA = RAWSUB;
  BY LEVEL GEND;
PROC SURVEYSELECT DATA=RAWSUB
  OUT=SAMPFL
  SAMPsize=NEWFREQ3;
  STRATA LEVEL GEND;
  ID ID LEVEL GEND;
PROC FREQ DATA = SAMPFL;
  TABLES LEVEL * GEND
  /OUT=SAMPFREQ NOPRINT;
PROC PRINT DATA=SAMPFREQ;
TITLE2 'SAMPLE FREQUENCIES';
PROC PRINT DATA = ERROR;
TITLE2 'STRATA DELETED';
PROC DELETE DATA = NEWFREQ NEWFREQ2
  NEWFREQ3 SAMPFL SAMPFREQ ERROR;
```

Example #3

```
PROC FREQ DATA = RAWSUB;
  TABLES LEVEL * GEND * MAJCOLG
  /OUT=NEWFREQ NOPRINT;
DATA NEWFREQ2 ERROR;
  SET NEWFREQ;
  SAMPNUM=(PERCENT * 500)/100;
  _NSIZE_= ROUND(SAMPNUM,1);
  SAMPNUM=ROUND(SAMPNUM, .01);
  IF _NSIZE_=0 THEN OUTPUT ERROR;
  IF _NSIZE_=0 THEN DELETE;
  OUTPUT NEWFREQ2;
DATA NEWFREQ3;
  SET NEWFREQ2;
  KEEP LEVEL GEND MAJCOLG _NSIZE_;
PROC SORT DATA = NEWFREQ3;
  BY LEVEL GEND MAJCOLG;
PROC SORT DATA = RAWSUB;
  BY LEVEL GEND MAJCOLG;
PROC SURVEYSELECT DATA=RAWSUB
  OUT=SAMPFL
  SAMPsize=NEWFREQ3;
  STRATA LEVEL GEND MAJCOLG;
  ID ID LEVEL GEND MAJCOLG;
PROC FREQ DATA = SAMPFL;
  TABLES LEVEL * GEND * MAJCOLG
  /OUT=SAMPFREQ NOPRINT;
PROC PRINT DATA = SAMPFREQ;
PROC PRINT DATA = ERROR;
PROC DELETE DATA = NEWFREQ NEWFREQ2
  NEWFREQ3 SAMPFL SAMPFREQ ERROR;
```

Example #2 - Annotations

The same procedures are followed to determine strata sizes when stratifying on two variables.

Determine population frequencies and percentages.

Determine sample size (positive integers) for a sample of 500.

Create error data set.

Delete strata size if equal to 0.

Keep level, gender, and strata sizes.

Sort population data set and strata data set by level and gender.

PROC SURVEYSELECT selects a random sample stratified on level and gender, creates an output data set, and keeps id level and gender as identifiers.

Check sample frequencies and percentages.

Print an error report.

Delete data sets with PROC DELETE.

Example #3

Use the same procedures to stratify on three variables, level, gender, and major college.

Determine population frequencies and percentages.

Determine strata sizes (positive integers) for a sample of 500.

Create error data set.

Delete strata size if equal to 0.

Keep level, gender, major college, and strata sizes.

Sort population data set and strata size data set by level, gender and major college.

PROC SURVEYSELECT selects a random sample stratified on level, gender, and major college, creates an output data set, and keeps id level gender and major college as identifiers.

Check sample frequencies and percentages.

Print an error report.

Delete data sets with PROC DELETE.

Conclusion

PROC SURVEYSELECT makes sampling quick, efficient, and easy. It is flexible and allows for random or proportional sampling with or without replacement.

References

- Scheaffer, R. L., Mendenhall III, W., & Ott, R. L. (1996). Elementary survey sampling, Fifth Edition. Belmont: Duxbury Press.
- SAS® Applications Guide, 1980 Edition, Cary, N.C.: SAS Institute.
- SAS® Language, Version 6. Cary, N.C.: SAS Institute, 1990.
- SAS® OnlineDoc, Version 8, SAS/STAT® User's Guide, Chapter 63. Cary, N.C.: SAS Institute, 1999.
- SAS® Language and Procedures, Version 6, First Edition. Cary, N.C.: SAS Institute, 1989.
-

Appendix A

Simple Random Sample

```
DATA SMPL;
  RETAIN K 100 N;
  DROP N K;
  IF _N_ EQ 1 THEN N=NUMOBS;
  SET FINAL POINT=_N_ NOBS-NUMOBS;
IF RANUNI(06) < K/N THEN DO;
  OUTPUT;
  K = K-1;
END;
  N = N-1;
IF N EQ 0 OR K EQ 0 OR _N_ =NUMOBS
  THEN STOP;
```

Stratified Random Sample

```
DATA FRFL;
  SET UGFL;
  IF CLSF EQ '1';
DATA FRSMPL;
  RETAIN K 72 N;
  DROP N K;
  IF _N_ EQ 1 THEN N=NUMOBS;
  SET FRFL POINT=_N_ NOBS-NUMOBS;
IF RANUNI(13) < K/N THEN DO;
  OUTPUT;
  K = K-1;
END;
  N = N-1;
IF N EQ 0 OR K EQ 0 OR _N_ =NUMOBS
  THEN STOP;
DATA SOFL;
  SET UGFL;
  IF CLSF EQ '2';

DATA SOSMPL;
  RETAIN K 98 N;
  DROP N K;
  IF _N_ EQ 1 THEN N=NUMOBS;
  SET SOFL POINT=_N_ NOBS-NUMOBS;
IF RANUNI(405) < K/N THEN DO;
  OUTPUT;
  K = K-1;
END;
  N = N-1;
IF N EQ 0 OR K EQ 0 OR _N_ =NUMOBS
  THEN STOP;
```

About the author

Diana Suhr is a Statistical Analyst in the Office of Institutional Research at the University of Northern Colorado. She earned a Ph.D. in Educational Psychology at UNC in 1999. The first programming language she learned was Fortran in 1970. She has been a SAS programmer since 1984.

Contact

Diana Suhr, Statistical Analyst
Institutional Research
University of Northern Colorado
Greeley, CO 80639
970-351-2193, diana.suhr@unco.edu

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

```
DATA JRFL;
  SET UGFL;
  IF CLSF EQ '3';
DATA JRSMPL;
  RETAIN K 62 N;
  DROP N K;
  IF _N_ EQ 1 THEN N=NUMOBS;
  SET JRFL POINT=_N_ NOBS-NUMOBS;
IF RANUNI(17) < K/N THEN DO;
  OUTPUT;
  K = K-1;
END;
  N = N-1;
IF N EQ 0 OR K EQ 0 OR _N_ =NUMOBS
  THEN STOP;

DATA SRFL;
  SET UGFL;
  IF CLSF IN('4', 'C');
DATA SRSMPL;
  RETAIN K 54 N;
  DROP N K;
  IF _N_ EQ 1 THEN N=NUMOBS;
  SET FINAL POINT=_N_ NOBS-NUMOBS;
IF RANUNI(613) < K/N THEN DO;
  OUTPUT;
  K = K-1;
END;
  N = N-1;
IF N EQ 0 OR K EQ 0 OR _N_ =NUMOBS
  THEN STOP;
```

```
DATA UGSMPL;
  SET FRSMPL SOSMPL JRSMPL SRSMPL;

PROC FREQ DATA = UGSMPL;
  TABLES CLSF;
  TITLE 'SAMPLE FREQUENCIES';
PROC FREQ DATA = UGFL;
  TABLES CLSF;
  TITLE 'POPULATION FREQUENCIES';
```

Note: The seed in ranuni is an integer $< 2^{31} - 1$.
If the seed is negative, the time of day is used.