Paper 007-2013

Line Sampling Macro for Multistage Sampling

Charley Jiang, University of Michigan; James M. Lepkowski, University of Michigan; Richard Valliant, University of Michigan; James Wagner, University of Michigan

ABSTRACT

The Sample Survey is today one of the most important methods for collecting information about the social and economic world. The success of the method can be attributed in part to probability selection of the sample. Probability samples can be complex in design, and require careful attention to detail to assure correct execution.

In the SAS® world, one tool, PROC SURVEYSELECT, is widely used for probability sample selection (see http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#surveyselect_toc.htm). However, the procedures implemented in SURVEYSELECT are basic selection tools that must be assembled into larger systems for complex probability samples, particularly multistage samples. This paper describes the development and operation of a set of sampling macros built around SURVEYSELECT for sampling the ultimate stage units in a multistage sample. Examples of several situations where the macro can be most beneficial are also given.

INTRODUCTION

There are many basic sampling techniques available for selecting a sample for a survey. For example, simple random sampling chooses elements directly with an equal probability of selection to each sample frame unit (see Cochran, 1977, for details on probability sampling techniques). On the other hand, cluster sampling chooses groups of elements first, and then selects elements from among those clusters chosen in the first stage.

Many surveys use sample designs that combine sampling techniques. They may first employ stratification and then selection of clusters within strata. For example, in area sampling in the U.S., counties are often used as first stage units, and are implicitly clusters of housing units located in them.

Subsequently, second stage clusters may be identified and selected within each first stage cluster in the sample. For example, in the U.S., Census blocks are often used in area probability samples as a second stage cluster within counties. The second stage clusters themselves are grouped into strata, and a second stage sample selected from each stratum formed within a first stage cluster. In such area probability samples, these second stage units are often called segments, and typically consist of several Census blocks joined together as a single second stage sampling unit.

This process can be repeated within second stage units, identifying third stage clusters within second stage units. For example, in selected Census block housing units and households (the collection of people who usually reside at a particular housing unit) are identified, listed, and selected.

The subsampling may extend to four or more stages in some designs, ending with the final selection of housing units, or persons within sample housing units. These designs may also employ unequal probabilities of selection and other complexities introduced into selection to make it practical to select a sample in multiple stages.

The examples given here are referred to as units within "area samples"(see Groves et al. 2009). Area sampling uses multiple stages of sampling in situations where a list of elements is completely lacking, or partly lacking. Identifying housing units within selected segments or Census blocks allows assembly of lists of housing units on a much smaller scale than trying to obtain or list all housing units across all first stage units in the entire country. The method is widely used, including, for example, in the National Crime Victimization Survey (http://www.bjs.gov/index.cfm?ty=dcdetail&iid=245). The target population is the civilian non-institutionalized household population ages 12 years old and older in the United States. There is no accurate list of names and addresses of such persons. So the sample design is based on multiple stages of sampling and area probability selection. These designs also reduce the costs of sending interviewers all over a large area, instead sending interviewers to relatively compact locations like a Census block and attempting to conduct interviews at multiple households.

Based on these two main cost advantages, multiple stage sampling is applied in many large scale area and national surveys.

NORMAL SAMPLING PROCEDURE IN SAS

In complex multistage sample designs, there are requirements that a simple application of SURVEYSELECT procedures cannot meet. For instance, in some strata or stages the probability proportionate to size (PPS) selection

method may be used, while in other strata or stages a systematic selection procedure (SYS) may be required. SURVEYSELECT can be applied each time separately to each stratum, or each stage, but of course is limited to one selection method in each instance. Also, SURVEYSELECT has limitations for sample size and sampling rates that can be applied. Applying different SURVEYSELECT procedures in sub-databases such as strata or stages yields subsamples that must then be combined to get the whole sample assembled. Further, a final sample selection report must be assembled describing results across repeated applications of different procedures, a costly and inconvenient process.

Faced with these constraints, we sought a powerful tool to combine all processes in one step that could be executed at one time. We developed a SAS Macro, the "The Line Sampling Macro," to meet the various requirements of such complex designs. The Macro is designed to select the sample at the last stage of a multistage sample, after accounting for all the prior steps of selection for each ultimate sampling unit. For example, in an area sample of several stages (e.g., PSU, block group or segments), at the last stage households have been listed and a sample is to be selected from among them across all PSU's and segments. The sampling rate for each household is determined by the sampling statistician, and specified in the database of listed households. For example, an equal probability sample of households is desired, the sampling rate of households may differ from one segment to another in order to account for unequal selection probabilities of units in the prior stages. We refer to these last stage units as 'lines' in the selection process. While in this illustration the lines are households, they could be any ultimate stage sampling unit in the application of the macro.

Effectively, then, listed lines have been combined across the entire set of sample PSU's and segments, and a sample of lines is to be selected in one pass through the consolidated file of all lines. The sampling statistician assigns a measure of size (MOS) to each line in such a way that when prior stage probabilities are multiplied by the probability of selecting a line, a desired overall sampling rate is obtained for each sample line in the data base. The final stage selection from this consolidated file of lines uses probability proportional to the MOS sample selection.

FEATURES OF THE MACRO

The macro has a number of features that extend its use to a wide range of complex sampling applications. These features of the macro include the following:

- Different sample size and sampling rates for subgroups. Separate overall sampling rates can be specified for different subgroups of last stage unit defined by such characteristics as age, race-ethnicity, and sex.

- Sorting. The file can be sorted by (up to) 10 variables to provide implicit stratification when using systematic selection. Several methods of sorting are allowed.

- Minimum and maximum sample sizes per unit. The minimum and maximum total sample size can be specified for each unit at the stage of sampling prior to line sampling. For example, if the prior stage is a segment in an area sample, and housing units are to be selected, the minimum and maximum number of sample housing units in each segment can be specified. During the operation of a survey, Minimum, maximum sample sizes is a very convenient option.

- Selection methods. Four methods of probability proportional to size sampling are available.

The Macro system includes three main sections: 1) definition of macro variables, 2) specification of input databases, and 3) specification of output databases.

In the definition of macro variables, there is a clear definition of a set of macro variables which will be used in the system. All macro variables are either required (a value must be specified) or optional (the user decides whether to specify a value for a variable).

The following is a partial list of the 'definition' macro variables:

Macro variable	Description	Allowable values	Required
&LIB	library in which input database are stored	Valid SAS library name	Yes
&OUTLIB	library in which output database and files are stored	Valid SAS library name	No
&INF_DAT	file with one record for each line in the last-stage sampling frame	User specified file name	Yes
&INF_DPROB	file with desired, final selection probabilities; these can be different for different types of lines	User specified file name	Yes

&INF_PPROB	file with selection probabilities for stages of sampling that precede line sampling	User specified file name	Yes
&CONTROL	Sorting variables used in PROC SURVEYSELECT. Sorting variables only have an effect for the selection methods PPS_SEQ, and PPS_SYS.	Field name(s) in &INF_DAT	No
&NMIN	Minimum total sample size per segment (or within whatever the stage of sampling is prior to line sampling). Note that this sample size is across all values of &SUB1, …, &SUBk that occur within a segment	Positive integer "NULL" (or, possibly, . for missing) is allowed	No
&NMAX	Maximum sample size per segment (or within whatever the stage of sampling is prior to line sampling)	Positive integer	No
&OUTFRM	Name of the output data set that contains all records in the line frame. This data set contains all units in the entire frame from which the line sample is selected plus fields from &INF_DPROB and &INF_PPROB. An indicator variable for sample selection is included, 1= sampled, 0 = not.	User specified file name "NULL" (or, possibly, . for missing) is allowed	No
&OUTFORM	Indicate the type of output report file.	PDF, HTML, RTF,XLS,LST, TXT	No
&PRT_INREC	Number of records to be printed from file M2; default value = 100.	Positive integer "NULL" (or, possibly, . for missing) is allowed	No
&PRT_FRMREC	Number of records to be printed from file &OUTFRM; default value = 100.	Positive integer "NULL" (or, possibly, . for missing) is allowed	No
&PRT_SAMREC	Number of records to be printed from file &SAMFILE; default value = 100	Positive integer "NULL" (or, possibly, . for missing) is allowed	No
&SAMFILE	Names the output data set that contains the sample from proc surveyselect	User specified file name	Yes
&SEEDNO	Specifies the initial seed for random number generation. The value must be a positive integer. If &SEEDNO is missing,clock time is used for the random number seed. The seed that is used is printed as part of the SAS output	Positive integer "NULL" (or, possibly, . for missing) is allowed	No
&SEG	Name of field in &INF_DAT that identifies segments within first-stage units This identifier must be unique, i.e., segment numbers cannot be reused within PSU's. If necessary, a user can create a unique segment ID by concatenating PSU, segment, or other fields before inputting to linesamp	Numeric or Character	Yes
&AGG	Field that identifies the prior stage aggregate unit that contains each line; This could be the concatenation of fields that identify stratum, PSU, block group, etc.	Numeric or Character	Yes
&SELMETH	Method of sample selection to be used in proc surveyselect. If PPS_SEQ or PPS_SYS are specified, then at least one &CONTROL variable must be specified.	PPS, PPS_SAMPFORD, PPS_SEQ, PPS_SYS	Yes
&SORTMETH	Method of sorting used in systematic sample selection. &SORTMETH must be non-null if &CONTROL is non-null.	NEST, SERP, "NULL" (or, possibly, . for missing) is allowed	No

&STRATA	Name of field in &INF_DAT and &INF_DPROB that identifies stratum of first-stage units or another field that will be treated as defining strata for sample selection.	"NULL" (or, possibly, . for missing) is allowed	No
&SUB1, &SUB2, ..., &SUB10	Fields in &INF_DPROB that define subgroups or domains for which different overall selection rates are desired. "NULL" (or, possibly, . for missing) is allowed, in which case all lines have the same desired overall selection probability. Up to 10 subgroup fields are allowed.	Field name(s) in &INF_DPROB	Yes

Table 1: partial list of the 'definition' macro variables

There are three key databases which must also be input to the macro. The first is the final selection probability database, which has final selection probabilities which can differ for different types of lines. The name of this database is associated with the macro variable &INF_DPROB. This database will have as many lines as there are combinations of &STRATA and &SUB1-&SUB10. If there are no strata or subgroups that are sampled at different rates, the file &INF_DPROB will have 1 record only.

The second key database is the selection probability for stages of sampling. It contains selection probabilities for stages of sampling that precede line sampling. The name of this database is associated with macro variable &INF_PPROB. The number of records in this file will be the number of aggregate sample units (e.g., PSU and segment combinations) that precede line sampling.

Finally, there is the sample frame database with one record for each line in the last-stage sampling frame. The name of this database is associated with macro variable &INF_DAT.

There are two key output databases. The first is the sample database that contains the sample selected from the line sampling macro applied to the input database. The name of this database is associated with macro variable &SAMFILE.

Users can request the following variables which are all based on fields produced by PROC SURVEYSELECT in &SAMFILE: 1) STRATA; 2) CONTROL, 3) ZONE (the selection required for METHOD = PPS_SEQ), 4) SIZE, 5) NumberHits (the number of hits or selections, included for selection methods that are with replacement or with minimum replacement such as PPS_SYS and PPS_SEQ), 6) ExpectedHits (the expected number of hits or selections, included for selection methods that are with replacement or with minimum replacement that may select the same unit more than once such as PPS_SYS and PPS_SEQ), 7) SelectionProb (the conditional probability of selection given all prior stages of selection, included for selection methods that are without replacement), and 8) SamplingWeight (the sampling weight, the inverse of ExpectedHits or SelectionProb, a conditional weight that will weight sample lines up to the line sample frame only).

Part of the output includes fields computed for each record in &SAMFILE, including 1) LineProb, the selection probability of a line within the final stage of selection; 2) LineWt, the inverse of LineProb; and 3) FPROB, the overall probability of selection (the product of PPROB from file &INF_PPROB and LineProb).

The output also includes the sample frame database that contains not only one record for each line from the input sample frame database, but also a sample flag showing whether or not the line is selected and a sample weight. This database has all the fields in &INF_DAT, and adds new fields. The name of this database is associated with macro variable &OUTFRM.

Furthermore, the output includes brief report with PDF, HTML, RTF,XLS,LST or TXT options. The report shows description of frame, sample, weight and summary of sampling process.

The process of running the macro is as follows. First, set up a library for the input database, and associate it with macro variable &LIB. The value of &LIB should be a valid SAS library name. The three required input databases must be stored in &LIB. Second, give a valid value to all the required macro variables with the individual user's specific requirements. Third, assign valid values to other optional Macro variables. Lastly, set up the options for the specific sampling task, run the Line Sampling Macro program, and carefully check the output.

ILLUSTRATION

Consider the following illustration of how to apply the macro to a national survey database. Figure 1 gives twenty-nine observations from the input database, and Table 2 the complete dictionary of the example data. SID is the unique ID for each line. There are 3023 lines across PSU's, SSU's, and segments. The frequencies of lines across different stages are given in Figure 2.

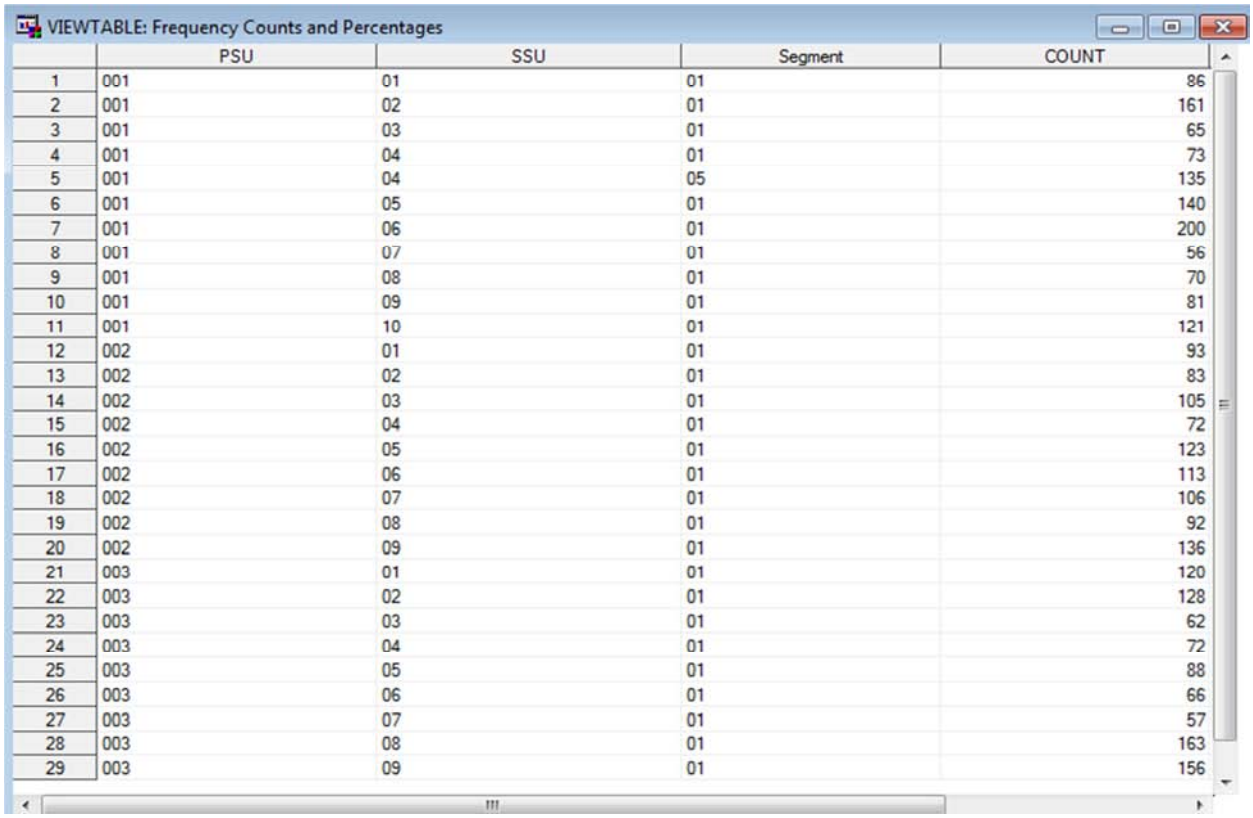
VIEWTABLE: Work.Input

	SID	PSU	SSU	Segment	state	tract	block	seg	domai
1	001010100111	001	01	01	MA	300100	2023	0010101	1
2	001020100111	001	02	01	MA	311600	1007	0010201	1
3	001030100111	001	03	01	MA	322100	4012	0010301	1
4	001040100111	001	04	01	MA	387201	1005	0010401	1
5	001040500111	001	04	05	MA	387201	1002	0010405	1
6	001050100111	001	05	01	MA	322100	1001	0010501	1
7	001060100111	001	06	01	MA	321600	4000	0010601	1
8	001070100111	001	07	01	MA	385100	9002	0010701	1
9	001080100111	001	08	01	MA	351200	8003	0010801	1
10	001090100111	001	09	01	MA	352900	1004	0010901	1
11	001100100111	001	10	01	MA	373900	2004	0011001	1
12	002010100111	002	01	01	MA	51100	6003	0020101	1
13	002020100111	002	02	01	MA	170100	6006	0020201	1
14	002030100111	002	03	01	MA	100500	1009	0020301	1
15	002040100111	002	04	01	MA	92100	4005	0020401	1
16	002050100111	002	05	01	MA	100200	3007	0020501	1
17	002060100111	002	06	01	MA	92200	4006	0020601	1
18	002070100111	002	07	01	MA	110502	1012	0020701	1
19	002080100111	002	08	01	MA	90900	1003	0020801	1
20	002090100111	002	09	01	MA	402	1005	0020901	1
21	003010100111	003	01	01	MA	801101	1015	0030101	1
22	003020100111	003	02	01	MA	801503	2036	0030201	1
23	003030100111	003	03	01	MA	801602	1027	0030301	1
24	003040100111	003	04	01	MA	801605	2015	0030401	1
25	003050100111	003	05	01	MA	812101	5000	0030501	1
26	003060100111	003	06	01	MA	800500	3015	0030601	1
27	003070100111	003	07	01	MA	813401	5016	0030701	1
28	003080100111	003	08	01	MA	813700	6005	0030801	1
29	003090100111	003	09	01	MA	813404	1000	0030901	1

Figure1. Initial lines in the example database showing ID number, PSU, second stage sampling unit, and segment numbers, and ancillary information about the line

Alphabetic List of Variables and Attributes					
Variable	Type	Len	Format	Informat	Label
PSU	Char	3	$3	$3	Primary sample Unit
SID	Char	12			
SSU	Char	2	$3	14	Second sample Unit
Segment	Char	2	14	14	
Domain	Num	1	1	1	
Seg	Char	10			Segment
block	Char	4			Block
city	Char	40			City
state	Char	2	$255	$255	State
tract	Char	6			Tract number

Table2. Dictionary of variables and attributes for fields in the input database in Figure 1



VIEWTABLE: Frequency Counts and Percentages

	PSU	SSU	Segment	COUNT
1	001	01	01	86
2	001	02	01	161
3	001	03	01	65
4	001	04	01	73
5	001	04	05	135
6	001	05	01	140
7	001	06	01	200
8	001	07	01	56
9	001	08	01	70
10	001	09	01	81
11	001	10	01	121
12	002	01	01	93
13	002	02	01	83
14	002	03	01	105
15	002	04	01	72
16	002	05	01	123
17	002	06	01	113
18	002	07	01	106
19	002	08	01	92
20	002	09	01	136
21	003	01	01	120
22	003	02	01	128
23	003	03	01	62
24	003	04	01	72
25	003	05	01	88
26	003	06	01	66
27	003	07	01	57
28	003	08	01	163
29	003	09	01	156

Figure2. Frequency of the number of lines by combinations of PSU, SSU, and segment for the illustration input database

In this illustration, the PSUs, SSUs, and segments have already been selected. A sample of lines is to be selected so that each line has an overall selection probability of 0.15.

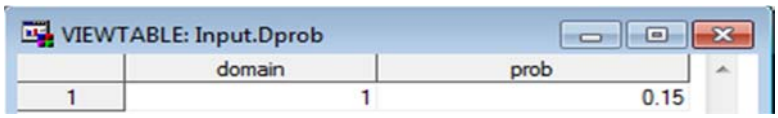
The PSU, SSU, and segment are given in Table 3. The SSU and segment selection probabilities are conditional on the previous steps. In particular, the SSU selection probability is the selection probability within the PSU, and not the product of the PSU and conditional SSU probabilities.

PSU	PSU Selection Probability	SSU	SSU Selection Probability within PSU	Segment	Segment Selection Probability within SSU
001	0.65	01	0.81	01	0.82
002	0.60	02	0.82	01	0.82
003	0.70	03	0.75	01	0.82
001	0.65	04	0.66	01	0.82
002	0.60	05	0.77	01	0.82
003	0.70	06	0.64	01	0.82
001	0.65	07	0.57	01	0.82
002	0.60	08	0.87	01	0.82
003	0.70	09	0.77	01	0.82
001	0.65	04	0.66	05	0.80
002	0.60	01	0.81	01	0.82
003	0.70	02	0.82	01	0.82
001	0.65	10	0.76	01	0.82

Table3. PSU, SSU, and segment probabilities in the illustration

We seek to use selection probabilities in Table 3 to select a sample of lines, generate output sample files, and obtain accompanying reports on the sampling operation.

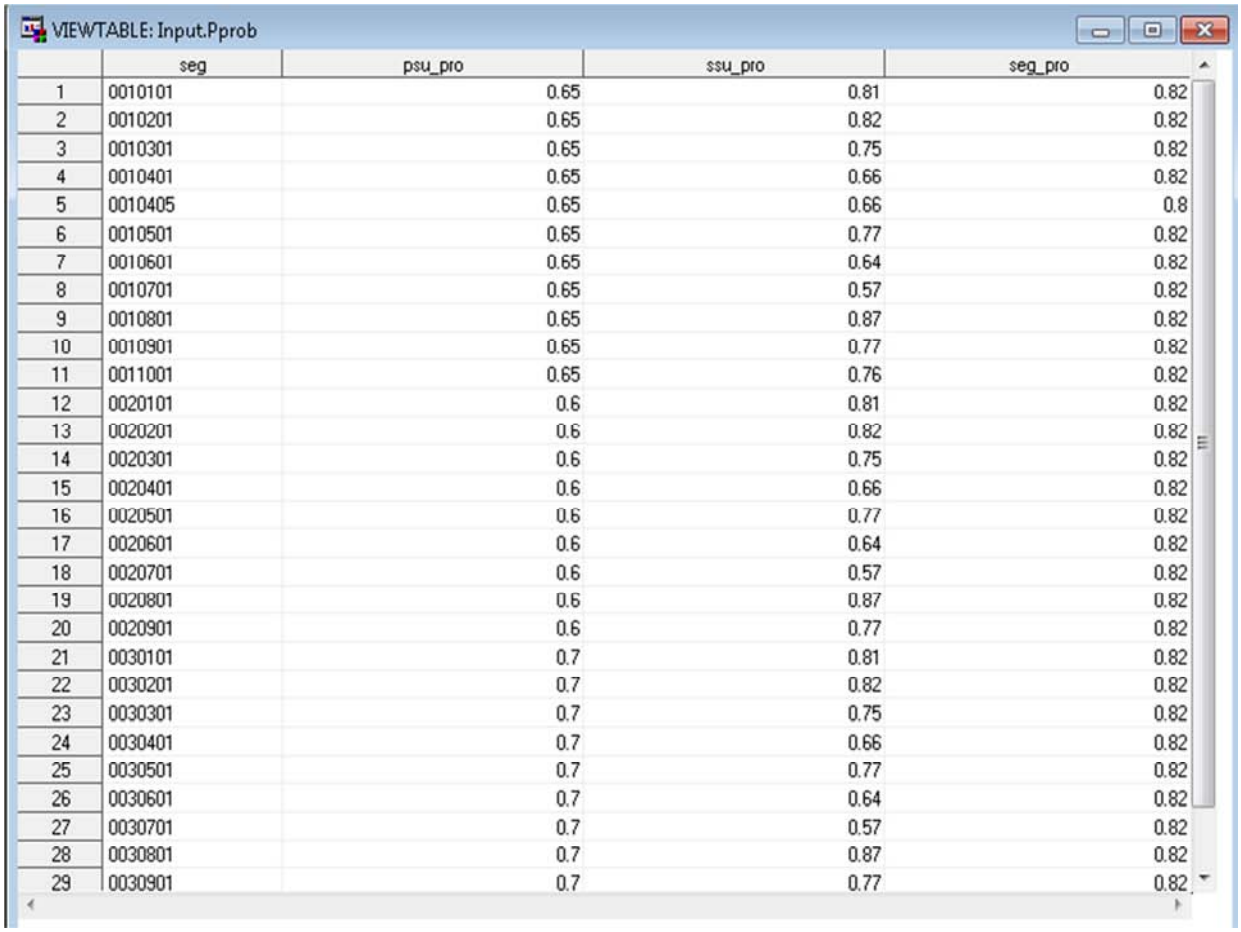
The process of implementation of the macro to these data is as follows. Set up a library for the input database: libname input 'L:\user\example';. Then the user sets up the three required input databases, one by one. For the 1) final selection probability database (&INF_DPROB,the final design probability equals 0.15) (see Figure 3).


VIEWTABLE: Input.Dprob

	domain	prob
1	1	0.15

Figure3. Example input.Dprob with only one record, and overall probability 0.15

For the selection probability for stages of sampling database &INF_PPROB, there is one record for each unique combination of &SUB1 to &SUB10, where these combinations represent domains. This file should contain the probabilities associated with prior stages of selection as shown in Table 3 above. Figure 4 shows some of the lines in the PPROB database below.


VIEWTABLE: Input.Pprob

	seg	psu_pro	ssu_pro	seg_pro
1	0010101	0.65	0.81	0.82
2	0010201	0.65	0.82	0.82
3	0010301	0.65	0.75	0.82
4	0010401	0.65	0.66	0.82
5	0010405	0.65	0.66	0.8
6	0010501	0.65	0.77	0.82
7	0010601	0.65	0.64	0.82
8	0010701	0.65	0.57	0.82
9	0010801	0.65	0.87	0.82
10	0010901	0.65	0.77	0.82
11	0011001	0.65	0.76	0.82
12	0020101	0.6	0.81	0.82
13	0020201	0.6	0.82	0.82
14	0020301	0.6	0.75	0.82
15	0020401	0.6	0.66	0.82
16	0020501	0.6	0.77	0.82
17	0020601	0.6	0.64	0.82
18	0020701	0.6	0.57	0.82
19	0020801	0.6	0.87	0.82
20	0020901	0.6	0.77	0.82
21	0030101	0.7	0.81	0.82
22	0030201	0.7	0.82	0.82
23	0030301	0.7	0.75	0.82
24	0030401	0.7	0.66	0.82
25	0030501	0.7	0.77	0.82
26	0030601	0.7	0.64	0.82
27	0030701	0.7	0.57	0.82
28	0030801	0.7	0.87	0.82
29	0030901	0.7	0.77	0.82

Figure4. Values in the &INF_PPROB database, where &AGG is seg, &PPROB1 is the PSU probability, PPROB2 is the SSU probability, and PPROB3 is the segment probability

The program operates on these databases, and internally multiplies the PSU, SSU, and segment probabilities to get a combined probability of selection for a segment. For example, the selection probability of PSU 001, SSU 01, and Segment 01 will be computed as PPROB = 0.65*0.81*0.82 = 0.43173. Using the example database as the sample frame database, the program carries out a brief check on the first two databases to be sure that the ratio &DPROB/&PPROB is less than or equal to 1 for all 10 segments.

Thus, we have now given valid values to all required Macro variables. Figure 5 summarizes the macro variables that have been illustrated in Figures 3, 4, and Table 3.

&LIB	&INF_DAT	&INF_DPROB
&INF_PPROB	&SAMFILE	&AGG
&SUB1-&SUB10 (at least one)	&DPROB	&SELMETH
&PPROB1-&PPROB10 (at least one)		

Figure 5. Summary of macro variables set in the illustration in Figures 4, 5, and 6

With all of these parameters set, and the database available, the program code is set up as follows:

```
options mprint mlogic;
libname input "L:\example";
libname output "L:\example\outnew";
%LINESAMPLING(LIB =input,PROJ_NAME =, VERSION=,
CONTROL=, INF_DPROB=dprob, INF_PPROB=pprob,
INF_DAT=example,   NMIN=, NMAX=,
OUTLIB=output,   OUTFRM=fullframe,
OUTDIR=L:\example\outnew,
PRT_INREC=10, PRT_FRMREC=10, PRT_SAMREC=10,
SAMFILE=selectedsample, SEEDNO=0, AGG=seg,
LINEID=sid,SELMETH=PPS_SYS,SORTMETH=,
STRATA=,  SUB1=domain,
PPROB1=psu_pro,PPROB2=ssu_pro,PPROB3=seg_pro );
```

The output is in the sample database and the sample frame database. Figure 6 shows the output from the sample database with computed probabilities by line. In this database, there are three different probability selection rates for the different stages. Also, the final selection rate, last stage line weight, and final weight are included.

VIEWTABLE: Work.Output_sample

	domain	prob	psu_pro	ssu_pro	seg_pro	PPROB	LINEWT	FINWT
1	1	0.15	0.6	0.57	0.82	0.28044	1.8696	6.666667
2	1	0.15	0.65	0.57	0.82	0.30381	2.0254	6.666667
3	1	0.15	0.6	0.64	0.82	0.31488	2.0992	6.666667
4	1	0.15	0.6	0.66	0.82	0.32472	2.1648	6.666667
5	1	0.15	0.7	0.57	0.82	0.32718	2.1812	6.666667
6	1	0.15	0.65	0.64	0.82	0.34112	2.27413	6.666667
7	1	0.15	0.65	0.66	0.8	0.3432	2.288	6.666667
8	1	0.15	0.65	0.66	0.82	0.35178	2.3452	6.666667
9	1	0.15	0.7	0.64	0.82	0.36736	2.44907	6.666667
10	1	0.15	0.6	0.75	0.82	0.369	2.46	6.666667
11	1	0.15	0.6	0.77	0.82	0.37884	2.5256	6.666667
12	1	0.15	0.6	0.81	0.82	0.39852	2.6568	6.666667
13	1	0.15	0.65	0.75	0.82	0.39975	2.665	6.666667
14	1	0.15	0.6	0.82	0.82	0.40344	2.6896	6.666667
15	1	0.15	0.65	0.76	0.82	0.40508	2.70053	6.666667
16	1	0.15	0.65	0.77	0.82	0.41041	2.73607	6.666667
17	1	0.15	0.6	0.87	0.82	0.42804	2.8536	6.666667
18	1	0.15	0.7	0.75	0.82	0.4305	2.87	6.666667
19	1	0.15	0.65	0.81	0.82	0.43173	2.8782	6.666667
20	1	0.15	0.65	0.82	0.82	0.43706	2.91373	6.666667
21	1	0.15	0.7	0.77	0.82	0.44198	2.94653	6.666667
22	1	0.15	0.65	0.87	0.82	0.46371	3.0914	6.666667
23	1	0.15	0.7	0.81	0.82	0.46494	3.0996	6.666667
24	1	0.15	0.7	0.82	0.82	0.47068	3.13787	6.666667
25	1	0.15	0.7	0.87	0.82	0.49938	3.3292	6.666667

Figure6. Output sample database showing computed probabilities

The selected sample frame database is shown for the illustration in Figure 7. In this database, the nsize_variable shows the total number of sample lines, while sam_indic shows which lines in the frame are selected.

VIEWTABLE: Work.Output_frame

	SID	domain	prob	psu_pro	ssu_pro	seg_pro	PPROB	_NSIZE_	SAM_INDIC	LINEWT	FINWT
1	002070100111	1	0.15	0.6	0.57	0.82	0.28044	.	0	.	.
2	002070100211	1	0.15	0.6	0.57	0.82	0.28044	1165	1	1.8696	6.666667
3	001070100111	1	0.15	0.65	0.57	0.82	0.30381	.	0	.	.
4	001070100211	1	0.15	0.65	0.57	0.82	0.30381	1165	1	2.0254	6.666667
5	002060100111	1	0.15	0.6	0.64	0.82	0.31488	.	0	.	.
6	002060100211	1	0.15	0.6	0.64	0.82	0.31488	1165	1	2.0992	6.666667
7	002040100111	1	0.15	0.6	0.66	0.82	0.32472	.	0	.	.
8	002040100211	1	0.15	0.6	0.66	0.82	0.32472	1165	1	2.1648	6.666667
9	003070100111	1	0.15	0.7	0.57	0.82	0.32718	.	0	.	.
10	003070100211	1	0.15	0.7	0.57	0.82	0.32718	1165	1	2.1812	6.666667
11	001060100111	1	0.15	0.65	0.64	0.82	0.34112	.	0	.	.
12	001060100211	1	0.15	0.65	0.64	0.82	0.34112	1165	1	2.2741333333	6.666667
13	001040500111	1	0.15	0.65	0.66	0.8	0.3432	.	0	.	.
14	001040500311	1	0.15	0.65	0.66	0.8	0.3432	1165	1	2.288	6.666667
15	001040100211	1	0.15	0.65	0.66	0.82	0.35178	.	0	.	.
16	001040100111	1	0.15	0.65	0.66	0.82	0.35178	1165	1	2.3452	6.666667
17	003060100111	1	0.15	0.7	0.64	0.82	0.36736	.	0	.	.
18	003060100211	1	0.15	0.7	0.64	0.82	0.36736	1165	1	2.4490666667	6.666667
19	002030100211	1	0.15	0.6	0.75	0.82	0.369	.	0	.	.
20	002030100111	1	0.15	0.6	0.75	0.82	0.369	1165	1	2.46	6.666667
21	002050100211	1	0.15	0.6	0.77	0.82	0.37884	.	0	.	.
22	002050100111	1	0.15	0.6	0.77	0.82	0.37884	1165	1	2.5256	6.666667
23	002010100211	1	0.15	0.6	0.81	0.82	0.39852	.	0	.	.
24	002010100111	1	0.15	0.6	0.81	0.82	0.39852	1165	1	2.6568	6.666667
25	001030100111	1	0.15	0.65	0.75	0.82	0.39975	.	0	.	.
26	001030100211	1	0.15	0.65	0.75	0.82	0.39975	1165	1	2.665	6.666667
27	002020100211	1	0.15	0.6	0.82	0.82	0.40344	.	0	.	.
28	002020100111	1	0.15	0.6	0.82	0.82	0.40344	1165	1	2.6896	6.666667
29	001100100111	1	0.15	0.65	0.76	0.82	0.40508	.	0	.	.
30	001100100311	1	0.15	0.65	0.76	0.82	0.40508	1165	1	2.7005333333	6.666667
31	001050100111	1	0.15	0.65	0.77	0.82	0.41041	.	0	.	.
32	001050100311	1	0.15	0.65	0.77	0.82	0.41041	1165	1	2.7360666667	6.666667
33	002080100211	1	0.15	0.6	0.87	0.82	0.42804	.	0	.	.
34	002080100111	1	0.15	0.6	0.87	0.82	0.42804	1165	1	2.8536	6.666667
35	003030100111	1	0.15	0.7	0.75	0.82	0.4305	.	0	.	.
36	003030100211	1	0.15	0.7	0.75	0.82	0.4305	1165	1	2.87	6.666667
37	001010100111	1	0.15	0.65	0.81	0.82	0.43173	.	0	.	.
38	001010100211	1	0.15	0.65	0.81	0.82	0.43173	1165	1	2.8782	6.666667
39	001020100111	1	0.15	0.65	0.82	0.82	0.43706	.	0	.	.
40	001020100211	1	0.15	0.65	0.82	0.82	0.43706	1165	1	2.9137333333	6.666667
41	003050100111	1	0.15	0.7	0.77	0.82	0.44198	.	0	.	.
42	003050100211	1	0.15	0.7	0.77	0.82	0.44198	1165	1	2.9465333333	6.666667
43	001080100211	1	0.15	0.65	0.87	0.82	0.46371	.	0	.	.
44	001080100111	1	0.15	0.65	0.87	0.82	0.46371	1165	1	3.0914	6.666667
45	003010100211	1	0.15	0.7	0.81	0.82	0.46494	.	0	.	.
46	003010100111	1	0.15	0.7	0.81	0.82	0.46494	1165	1	3.0996	6.666667
47	003020100211	1	0.15	0.7	0.82	0.82	0.47068	.	0	.	.
48	003020100111	1	0.15	0.7	0.82	0.82	0.47068	1165	1	3.1378666667	6.666667

Figure7. Sample frame database after program operation showing the size and sample selection indicator.

CONCLUSION

In order to efficiently collect data from geographically dispread population, surveys employ multistage sample designs that can have complex sample selection methods. This paper has shown how a line sampling macro in SAS can efficiently select the final stage units in a multistage sample. It provides flexible input, and the output includes a sample, a frame with sample selection indicators and a description report. The illustration showed how the line sampling macro can be used to select the final stage units. While PROC SURVEYSELECT deals with basic sampling techniques, the line sampling macro is a good choice for more complicated sampling designs when multiple stages are needed, such as area sample selection of housing units.

The whole SAS Code and User Guide of line sampling macro for multistage sampling are available on University of Michigan Google Document.

REFERENCES

Cochran, W. (1997), Sampling Techniques, New York: Wiley

Robert M. Groves, Floyd J. Fowler, Jr., Mick P. Couper, James M. Lepkowski, Eleanor Singer, Roger Tourangeau (2009), Survey Methodology, Second Edition, New York: Wiley

SAS, PROC SURVEYSELECT
Available at http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/surveyselect_toc.htm

U.S Bureau of Justice Statistics, National Crime Victimization Survey (NCVS)
Available at http://www.bjs.gov/index.cfm?ty=dcdetail&iid=245

ACKNOWLEDGMENTS

Thanks to University of Michigan colleagues Dan Zahs, Frost Hubbard, both for their perceptive review of this paper and valuable suggestion on programming technique.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Charley Jiang
Institute for Social Research
University of Michigan
Phone: (734) 647-4610
Email: lijiang@umich.edu