Interactive Sampling Using SAS/AF® Frames

Keith Cranford, Marquee Associates, LLC, Austin, TX
Don Boudreaux, SAS Institute Inc., Austin, TX

Abstract

This paper presents a SAS/AF application which allows you to sample an existing SAS data set using three common sampling techniques. You may also produce estimates of a population mean and its associated standard error. These estimates use the appropriate statistical sampling theory to adjust for population and sample sizes.

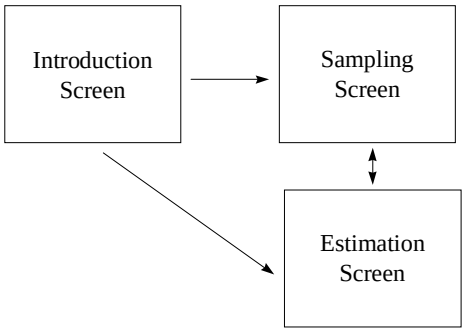
Introduction

In many applications SAS users are faced with very large data sets. Sampling can be used in these situations to provide smaller, more manageable data sets. Although the SAS System provides many tools that can be used for sampling these data sets, there is not a single procedure or Institute provided application for doing this. You are responsible for pulling these tools together. Also, SAS statistical estimation procedures such as PROC MEANS are only applicable for simple random samples from large data sets. For more complicated sampling techniques, you must write your own estimation programs.

The SAS/AF Frame application presented here addresses these two issues. First, the application provides an easy to use interface for sampling an existing SAS data set. Three common sampling techniques are provided: simple random sampling, stratified random sampling and cluster sampling. Secondly, the application allows the user to estimate the population mean of a variable of interest using the appropriate formulas based on the sampling technique and adjusted for the population and sample sizes. A flow diagram of the application is shown in Figure 1. Examples of the Sampling screen can be found in figures 2, 3, 5 and 6, and an example of the Estimation screen can be found in figure 4.

The basis of this application is outlined in two articles in *Observations* (see References) . The first article examined the efficiencies of different simple random sampling programs. The second contained macros for the above mentioned sampling techniques, utilizing the most efficient sampling programs. Also, for each technique macros were presented for estimating a population mean with its associated standard error. SAS/AF was then used to provide an interface to using the macros. The application was developed in SAS 6.11 using several of the new Frame objects available in this release such as the Data Table and the External File Viewer. This paper presents this application, including a brief description of the sampling techniques.

Figure 1: Application Flow Diagram



Introduction Screen

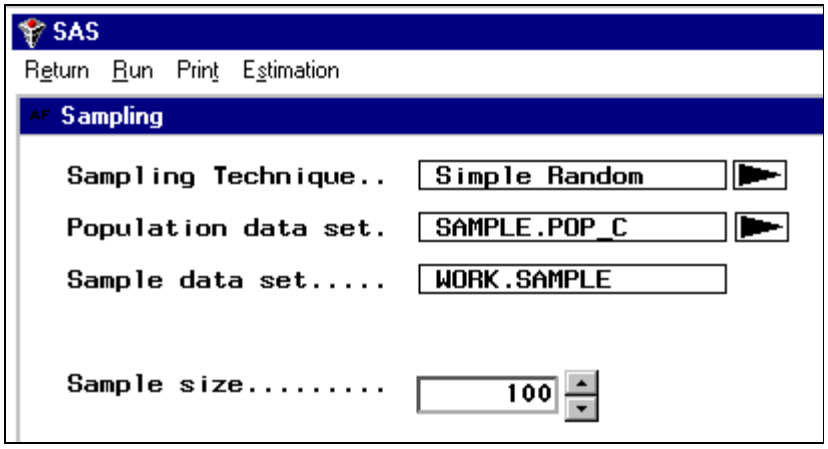
On invocation the application presents an 'Introduction' screen. This screen gives a summary of the function of the application. The application uses custom pull-down menus to navigate among screens. In this screen you have the choice to 'Exit' the application, proceed to the 'Sampling' screen, proceed to the 'Estimation' screen, or receive 'Help' on the current screen. We will examine the Sampling and Estimation screens in detail.

Sampling Screen

Upon choosing 'Sampling' from the Introduction Screen, the Sampling screen appears as shown in Figure 2. This screen allows you to sample an existing SAS data set using any of the three sampling techniques mentioned above. This is done by providing some basic information and then choosing 'Run' from the menu. The sampled data set is then displayed at the bottom of the screen for your perusal. The other menu selections allow you to 'Return' to the previous screen, 'Print' the sample data set, or proceed to the 'Estimation' screen.

The first information needed is the sampling technique you wish to use. This can be selected by clicking on the arrow to the right of the input field for the Sampling Technique. You have the choice of Simple Random

Figure 2: Sampling Screen for Simple Random Sample



(default), Stratified, and Cluster. The input field is then filled in automatically. In the next two fields you indicate which population data set to sample and the name of the resulting sample data set. The arrow to the right of the Population data set pops up a selection list of available data sets, or you may enter the name directly. In Figure 2, SAMPLE.POP_C is the data set from which we wish to sample. The default sample data set is WORK.SAMPLE, but you may change this by typing directly in the input field.

The remaining information on this screen varies depending on the sampling technique chosen. In the following sections each technique will be addressed. Simple random sampling will be discussed in detail to illustrate both the sampling and estimation aspects of the application. Stratified random sampling and cluster sampling will then be discussed by referring to their differences to simple random sampling.

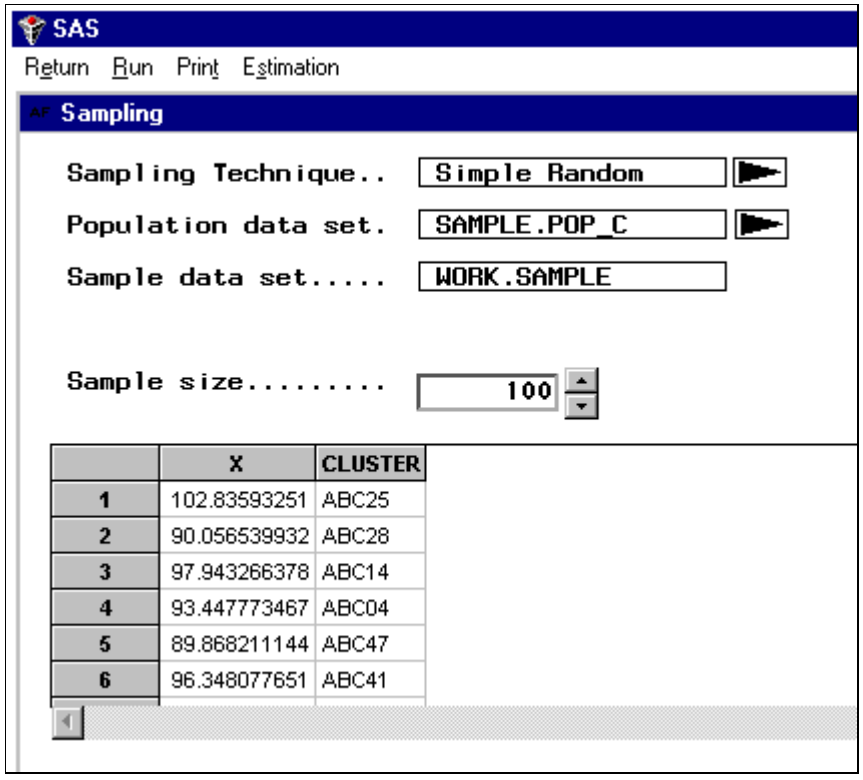
Simple Random Sample

The simplest and most common sampling technique is simple random sampling. A **simple random sample** is one in which all samples of a certain size, denoted by n, have the same chance of being selected from a population of a given size, denoted by N. From a practical standpoint only one sample is selected and each element of that sample is selected one at a time. As each element is selected, if all remaining elements have the same chance of being selected, then the statistical properties of a simple random sample are maintained. Conceptually, this is done by first assigning each element of the population a number from 1 to N. Then a random number table or a computer program is used to determine which element is selected from those not already drawn.

If you choose a simple random sample, the application needs only the required sample size which you provide in the input field next to 'Sample size.' You may either enter the sample size directly or use the "spinner arrow," to select the appropriate sample size. Clicking on 'Run' menu item prompts the application to produce the sample data set. At this point, a set of macros are used to sample the population data set and produce the sample data set with the name you specified. At the same time a secondary file is created to hold needed information about the sample, such as the sample data set name, the sample size, and the population size. This file is used later in producing the statistical estimates.

Once the sample data set is created you have the opportunity to browse the data set as seen in Figure 3. This is done through a SAS/AF Data Table object. The data set is opened in browse mode, which allows you to quickly check the sample data set for proper sample size and appropriate variables. You may also print the sample data set by clicking on the 'Print' menu item.

Figure 3: Sampling Screen with Sample Data Set View



	X	CLUSTER
1	102.83593251	ABC25
2	90.056539932	ABC28
3	97.943266378	ABC14
4	93.447773467	ABC04
5	89.868211144	ABC47
6	96.348077651	ABC41

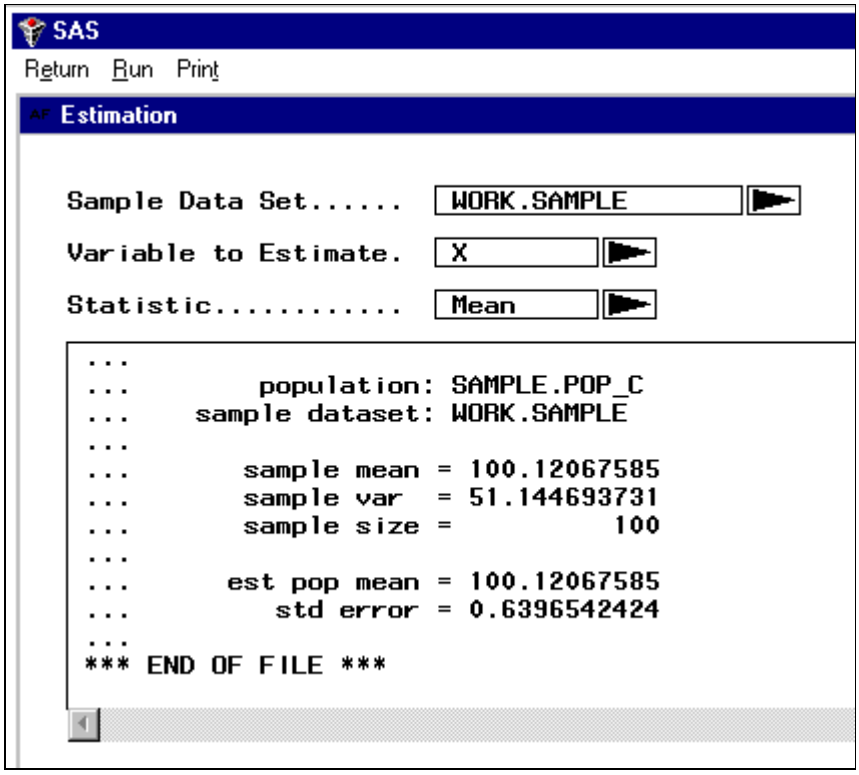
If you then wish to use this sample data set as input for SAS procedures, you may choose 'Return' to return to the Introduction screen without producing any statistical estimates, and exit the application. However, clicking on the 'Estimation' menu item allows you to estimate the population mean and its associated standard error within the application.

When the Estimation screen appears, the Sample Data Set is automatically filled in. This allows you, however, to enter a previous sample data set for analysis as well. As seen earlier in Figure 1, the Estimation screen can be accessed directly from the Introduction screen, without passing through the Sampling screen. The next input is the name of the variable to be estimated. In Figure 4, X has been chosen as the variable to estimate. Again, you may input this directly or click on the arrow to produce a variable list. This list gives only the numeric variables in the data set since a mean is being estimated. Lastly, the statistic you want to estimate is entered. Currently, the only choice is the mean, but this allows for future enhancements to include other statistics such as proportions.

To produce the estimate choose 'Run' from the menu. The application then uses the appropriate macro to produce a report such as the one shown in Figure 4. This is an External File Viewer object, viewing the file TEMP.TXT in the current working directory. You may also print this report by clicking on the 'Print' menu item. In this example, the report includes the sample mean, variance, and sample size along with the estimated population mean and the standard error of the mean. Note that the standard error given here differs from that which would be obtained from PROC MEANS or PROC UNIVARIATE since the finite population correction factor, derived from statistical sampling theory, is used in these calculations. This report will differ slightly depending on the sampling technique that is used, but the way the application works is the same.

The other sampling techniques are presented in the following sections. The Estimation screen is identical to the simple random sample (except for differences in the reports), so this will not be discussed. However, different information is needed to produce the samples, so the discussion will center around the Sampling screen.

Figure 4: Estimation Screen for Simple Random Sample



Stratified Random Sample

Many times the elements of the population under consideration can be put into distinct groups. In this case it may make more sense to consider each group separately and sample within each group, than to sample from the entire population as a single entity. If a simple random sample is selected from each group, called a stratum, the result is a **stratified random sample**.

In the application you can produce a stratified random sample by selecting 'Stratified' for the Sampling Technique on the Sampling screen. This is shown in Figure 5 using the data set SAMPLE.POP_S from which to sample. Next, you must enter a character variable that determines your strata, or groups. This can be either entered directly, or you can click on the arrow and a variable list will be displayed. In this example, GROUP is the variable that determines the strata.

At this point, a data set containing information about the strata is displayed in a Data Table object. The first two variables, which are protected columns, are the strata variable you specified and the population size for each strata. The last variable, for which you need to supply values, is the required sample size for each stratum. The data set is in edit mode so you can enter these values. The Data Table works similar to a spreadsheet, making this easy. Figure 5 illustrates entering 5, 7, and 13 for the strata groupA, groupB, and groupC, respectively. After entering the sample sizes, you choose 'Run' to produce the sample data set, which is displayed in the Data

Figure 5: Sampling Screen for Stratified Random Sample

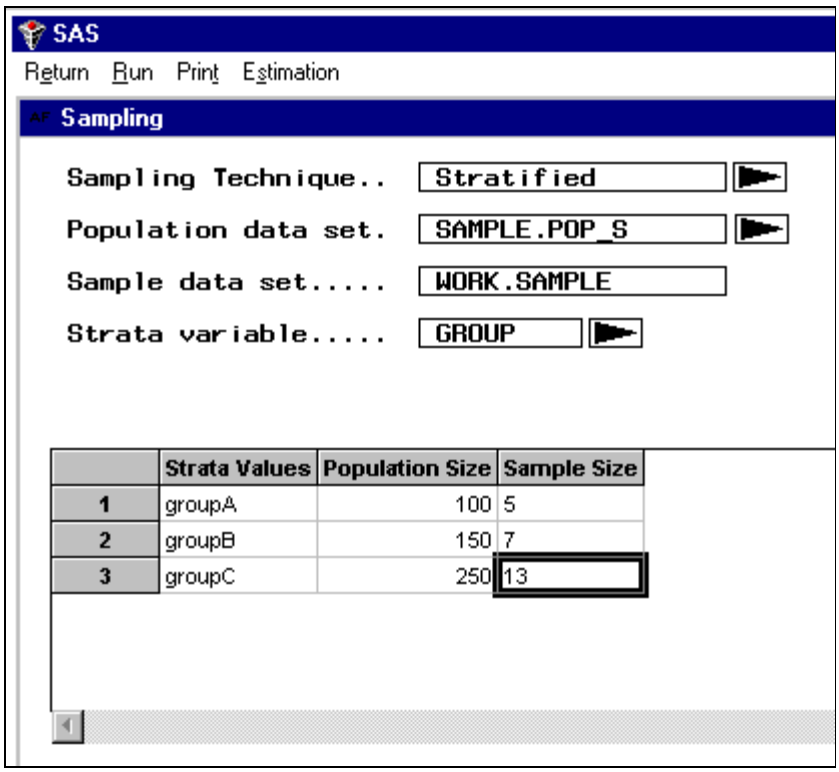


Table object shown earlier. You may then browse the sample data set and estimate a population mean in the same manner as the simple random sample.

Cluster Sampling

Sometimes it is difficult or costly to obtain a list of all elements in a population. This situation makes the previously discussed techniques either very costly or impossible to obtain. To handle this a third sampling technique, called cluster sampling, is sometimes used. A **cluster sample** is one in which each sampling unit is a collection, or cluster, of elements, instead of individual elements.

A cluster sample can be produced in the application by selecting 'Cluster' as the Sampling Technique on the Sampling screen as shown in Figure 6 (again using SAMPLE.POP_C as the population data set). You must then provide a character variable that determines the clusters. This can be input directly, or you can use the arrow to display a selection list of character variables in the data set. In this example, the values of CLUSTER determine the clusters.

Once the cluster variable is provided, you must give the number of clusters you wish to sample. This can be done directly, or you can use the "spinner arrow" to increase or decrease the number. In this example, 3 clusters will be sampled.

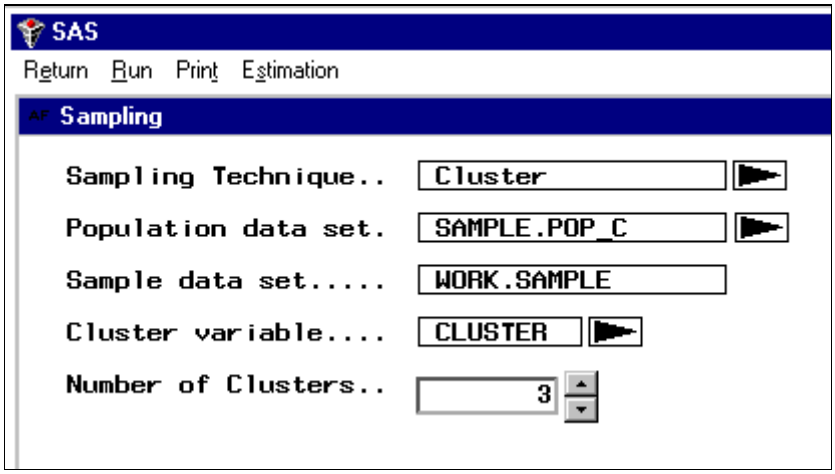
Finally, click on 'Run' to produce the sample data set. Again, the sample data set will be displayed in the Data Table object, and you can then calculate estimates of the mean of a variable. This is done as discussed earlier.

Summary and Future Enhancements

The application presented here provides you with a step-by-step method of producing samples based on three common sampling techniques. Additionally, you can estimate the population mean of a variable of interest with its associated standard error, correcting for a finite population. This is accomplished using SAS/AF Frame technology as an interface to a set of macros.

There are additional features which could be added. These include proportional sampling, in addition to fixed sample sizes; additional sampling techniques such as two-stage cluster sampling; and additional statistical estimates such as proportions. Also, a module for computing required sample sizes would be useful. If you have additional suggestions, please feel free to contact Keith at the address listed below.

Figure 6: Sampling Screen for Cluster Sample



References

Mendenhall, W., Ott L., Scheaffer R. (1971*), Elementary Survey Sampling*, Belmont, CA: Duxbury Press.

Cochran, William (1977), *Sampling Techniques, Third Edition*, New York: John Wiley & Sons, Inc.

Boudreaux, Don and Cranford Keith (1995), "Simple Random Sampling and Subsetting Strategies Using SAS Software," *Observations: The Technical Journal of SAS Software Users*, 4(4), 34-40.

Boudreaux, Don and Cranford Keith (1996), "SAS Macros for Simple Random, Stratified, and Cluster Sampling," *Observations: The Technical Journal of SAS Software Users*, 6(1), 31-43.

SAS Institute Inc. (1990), *SAS Language: Reference, Version 6, First Edition*, Cary, NC: SAS Institute Inc.

SAS Institute Inc. (1990), *SAS Guide to Macro Processing, Version 6, Second Edition*, Cary, NC: SAS Institute Inc.

SAS Institute Inc. (1989), *SAS Guide to the SQL Procedure: Usage and Reference, Version 6, First Edition*, Cary, NC: SAS Institute Inc.

SAS Institute Inc. (1990), *SAS Language and Procedures: Usage, Volume 2, Version 6, First Edition*, Cary, NC: SAS Institute Inc.

Keith Cranford can be contacted at:
Marquee Associates, LLC
3810 Medical Parkway, Suite 153
Austin, TX 78756
(512) 453-6140
kcranford@aol.com