

Abstract

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. This paper presents a SAS/AF Frame application for producing samples and estimates of the population mean for three common sampling techniques: simple random sampling, stratified random sampling, and cluster sampling. For each of these sampling techniques, a sample data set is created by using the appropriate technique on an existing, population data set. Also, the application can provide an estimate of the population mean of a variable of interest with its associated standard error. These standard errors differ from those provided by SAS procedures such as PROC MEANS in that they assume a finite population and use the appropriate correction factor. This application was developed in SAS 6.11 using several of the new objects available in this release such as the Data Table and the External File Viewer.

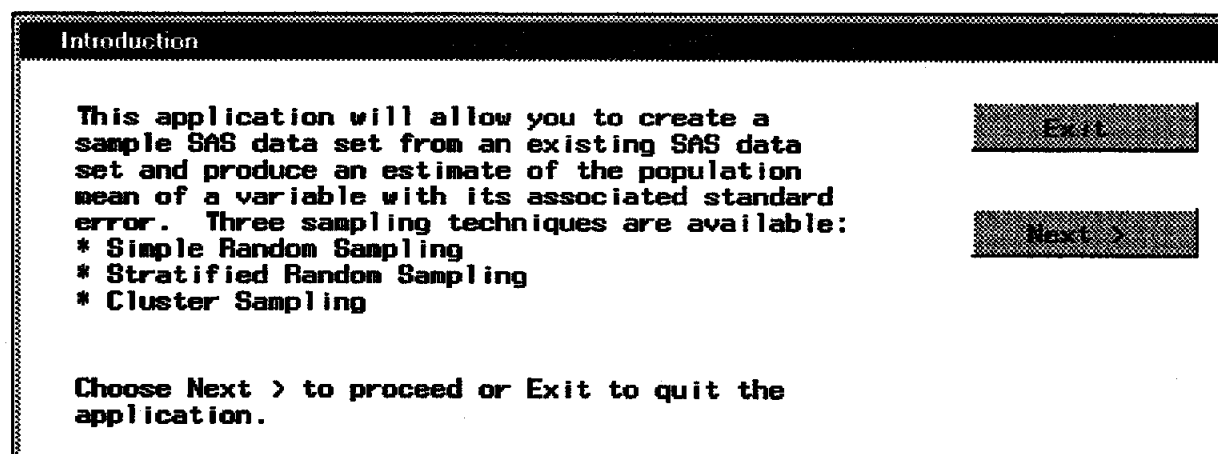
Introduction

In this article we present a SAS/AF Frame application to perform three of the most common sampling techniques. These techniques are simple random sampling, stratified random sampling and cluster sampling. This application is based on SAS macros we developed for an Observations article. The application allows you to produce a sample data set from an existing SAS data set and estimate the population mean of a variable of interest with its associated standard error.

The SAS/AF Frame application presented here steps you through the process of sampling from an existing SAS data set and estimating the population mean of a variable of interest. The result will be a sample SAS data set, which you could then use in other SAS procedures, and a statistical summary report saved to a text file. This can be done for any of the three sampling techniques mentioned above.

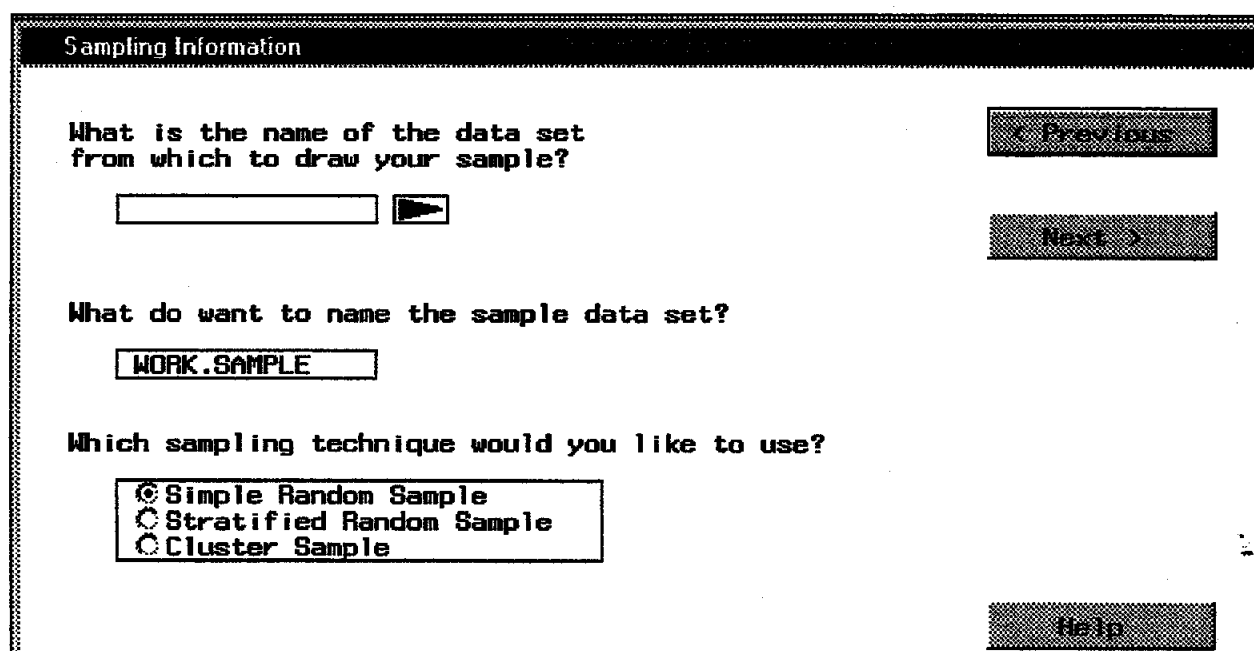
On invocation the application presents an 'Introduction' screen shown in Figure 1 that gives a summary of the function of the application. You may click on the 'Next >' button to proceed to the next screen or the 'Exit' button to quit the application. This will be consistent throughout the application; you either choose to proceed or return to the previous screen.

Figure 1: Introduction Screen



Upon choosing 'Next' the 'Sampling Information' screen appears as in Figure 2. On this screen you will provide basic information needed for all of the sampling techniques. Information particular to a technique will be requested on subsequent screens. First, you must select the data set from which to draw the sample. The name of the population data set can either be entered directly, or clicking on the arrow provides a selection list of all data sets available in the SAS session. In this example, the SAMPLE.POP_C data set is selected. See Figure 3. You

Figure 2: Initial Sampling Information Screen



The screenshot shows a window titled "Sampling Information". It contains three questions with corresponding input fields and buttons. The first question asks for the data set name, with an empty text box and a right-pointing arrow button. The second question asks for the sample data set name, with a text box containing "WORK.SAMPLE". The third question asks for the sampling technique, with three radio buttons: "Simple Random Sample" (selected), "Stratified Random Sample", and "Cluster Sample". Navigation buttons include "< Previous" at the top right, "Next >" at the middle right, and "Help" at the bottom right.

Sampling Information

What is the name of the data set from which to draw your sample?



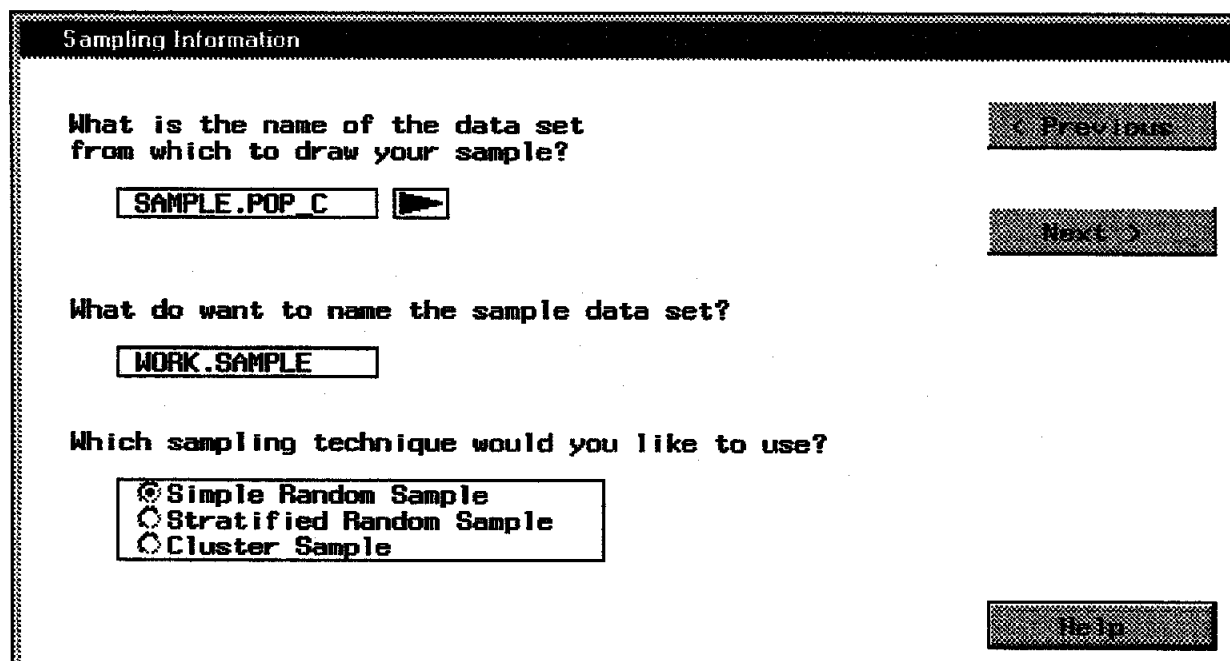
What do want to name the sample data set?

Which sampling technique would you like to use?

☒ Simple Random Sample
☐ Stratified Random Sample
☐ Cluster Sample

may also select the name of the resulting sample data set. By default, this is set to WORK.SAMPLE; however, any valid data set name may be entered. Both of these fields must be completed before the application allows you to go to the next screen. If you fail to do so, a message window will appear giving instructions. Finally, you select the sampling technique. Simple random sampling is selected by default, but you can choose stratified random or cluster sampling via the radio buttons. After completing the 'Sampling Information' screen, you then click on the 'Next >' button to proceed to the appropriate sampling screen or the '< Previous' to return to the 'Introduction' screen. On this and subsequent screens you also have the choice of a 'Help' button. Clicking on 'Help' will give you additional information for that particular screen. For example, on the 'Sampling Information' screen, this includes requirements for the data set names you provide and definitions for the sampling techniques.

Figure 3: Completed Sampling Information Screen



This screenshot is identical in layout to Figure 2, but the input fields are now populated. The first text box contains "SAMPLE.POP_C" and the second contains "WORK.SAMPLE". The "Simple Random Sample" radio button remains selected.

Sampling Information

What is the name of the data set from which to draw your sample?



What do want to name the sample data set?

Which sampling technique would you like to use?

☒ Simple Random Sample
☐ Stratified Random Sample
☐ Cluster Sample

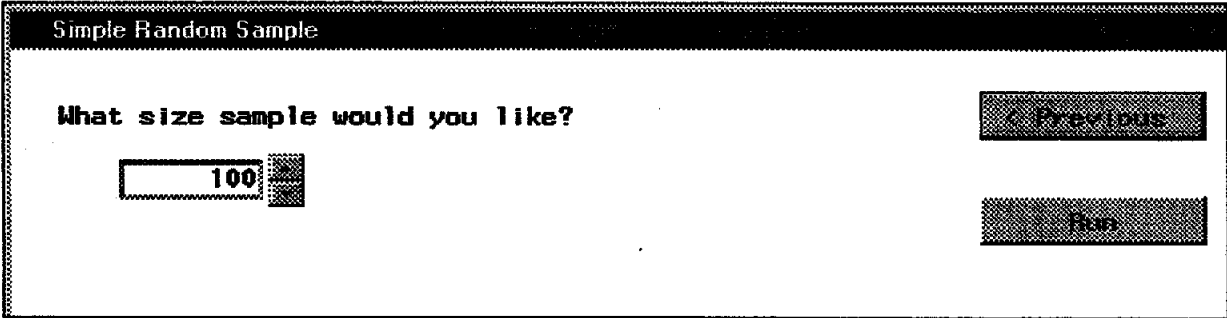
Each of the sampling techniques require different information, so the subsequent sampling screens are different for each technique. Each of these are presented in the following sections.

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. This is typically done by first assigning each element of the population a number from 1 to N . Then a random number table or a computer program (in this case the SAS System) 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 on the 'Simple Random Sample' screen shown in Figure 4. You may either enter the sample size directly or use the "spinner arrow," to select the appropriate sample size. Clicking on Run prompts the application to produce the sample data set. At this point, a set of SAS macros are used to sample the population data set and produce the sample data set with the name provided on the 'Sampling Information' screen. At the same time a second data set is created to hold needed information about the sample, such as the sample data set name, the sample size, and the population size. This temporary data set is used later in producing the statistical estimates.

Figure 4: Simple Random Sampling Screen



The image shows a graphical user interface window titled "Simple Random Sample". Inside the window, the text "What size sample would you like?" is displayed. Below this text is a text input field containing the number "100". To the right of the input field is a small square button with a spinner arrow. Further to the right are two larger rectangular buttons: "Previous" at the top and "Run" at the bottom.

Once the sample data set is created you have the opportunity to browse the data set as seen in Figure 5. 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. The current application includes all variables from the population data set, but could be modified to include only selected variables. Other enhancements will be discussed in a later section.

Figure 5: Browse the Sample Data Set Screen

	X	CLUSTER
1	98.493206528	ABC38
2	102.97794507	ABC20
3	109.3158859	ABC23
4	102.93748974	ABC15
5	103.26272507	ABC27
6	110.35834401	ABC17
7	105.92512011	ABC06
8	92.813312189	ABC24
9	107.29561293	ABC24
10	92.189934051	ABC33
11	113.60537176	ABC47
12	103.3000533	ABC13
13	101.28481106	ABC38
14	106.35272974	ABC48

Buttons: Previous, Estimates, Help

If you wish to then use this sample data set as input for SAS procedures, you may choose 'Previous' to return to the previous screen without producing any statistical estimates. However, clicking on the 'Estimates' button allows you to estimate the population mean and its associated standard error within the application. The initial 'Estimate of the Mean' screen is shown in Figure 6. The only required input here is the name of the

Figure 6: Initial Estimate of the Mean Screen

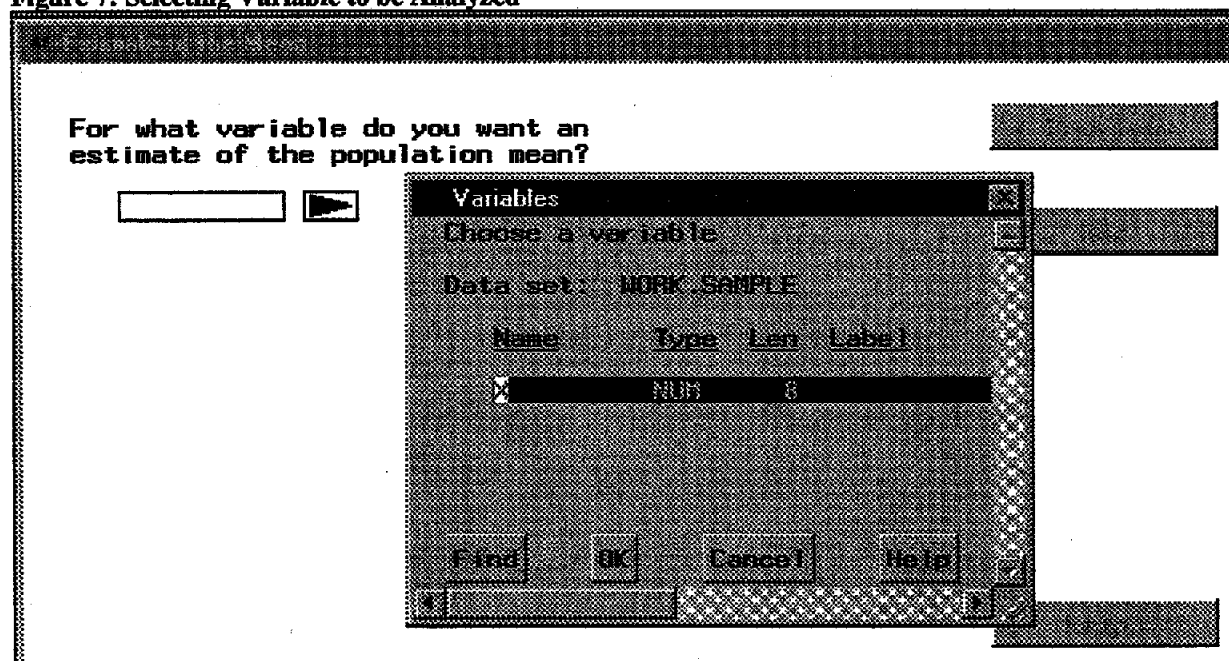
Estimate of the Mean

For what variable do you want an estimate of the population mean?

Buttons: Previous, Run

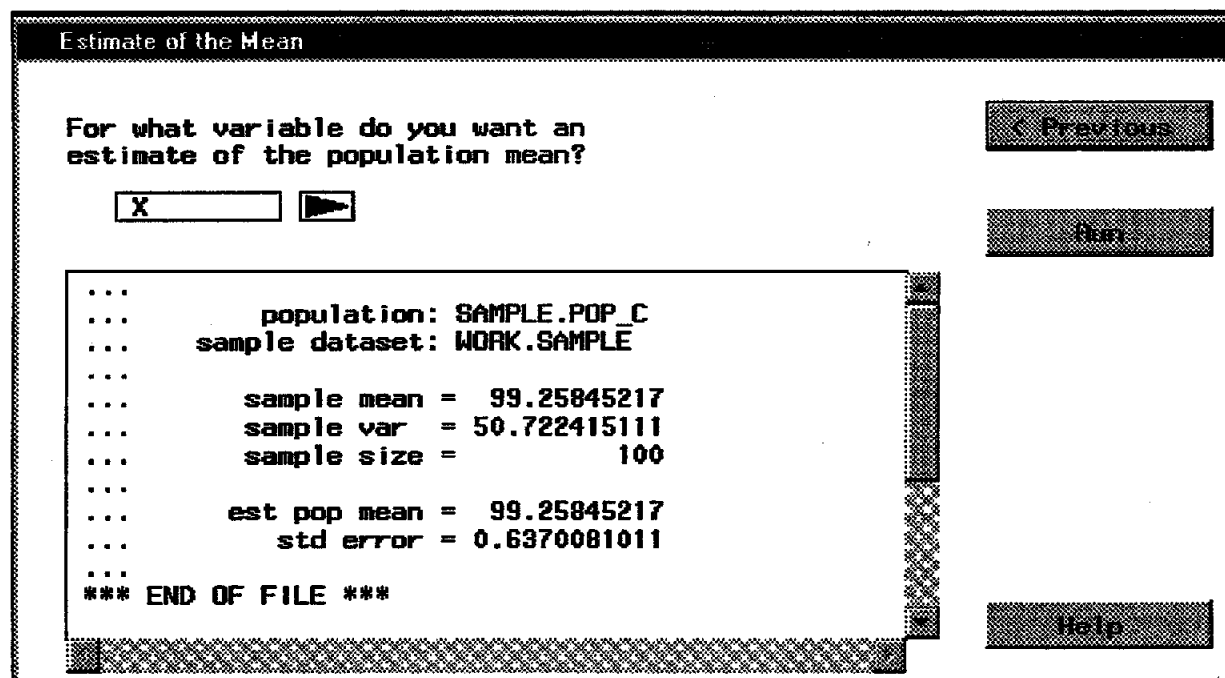
variable to be analyzed. Again, you may input this directly or click on the arrow to produce a variable list. As can be seen in Figure 7, this list only gives the numeric variables in the data set since a mean is being estimated. In this example there is only one numeric variable, X.

Figure 7: Selecting Variable to be Analyzed



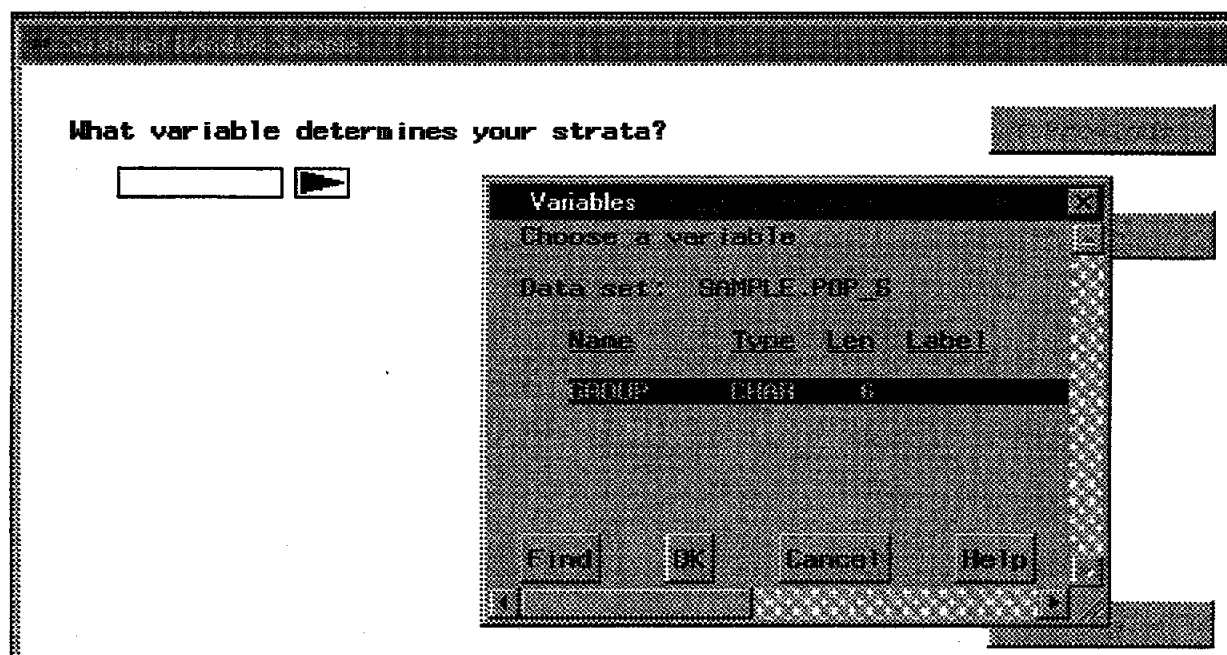
To produce the estimate click on Run. The application then uses the appropriate SAS macro to produce a report such as the one shown in Figure 8. This is an External File Viewer object, viewing the file TEMP.TXT in

Figure 8: Estimate of the Mean Report



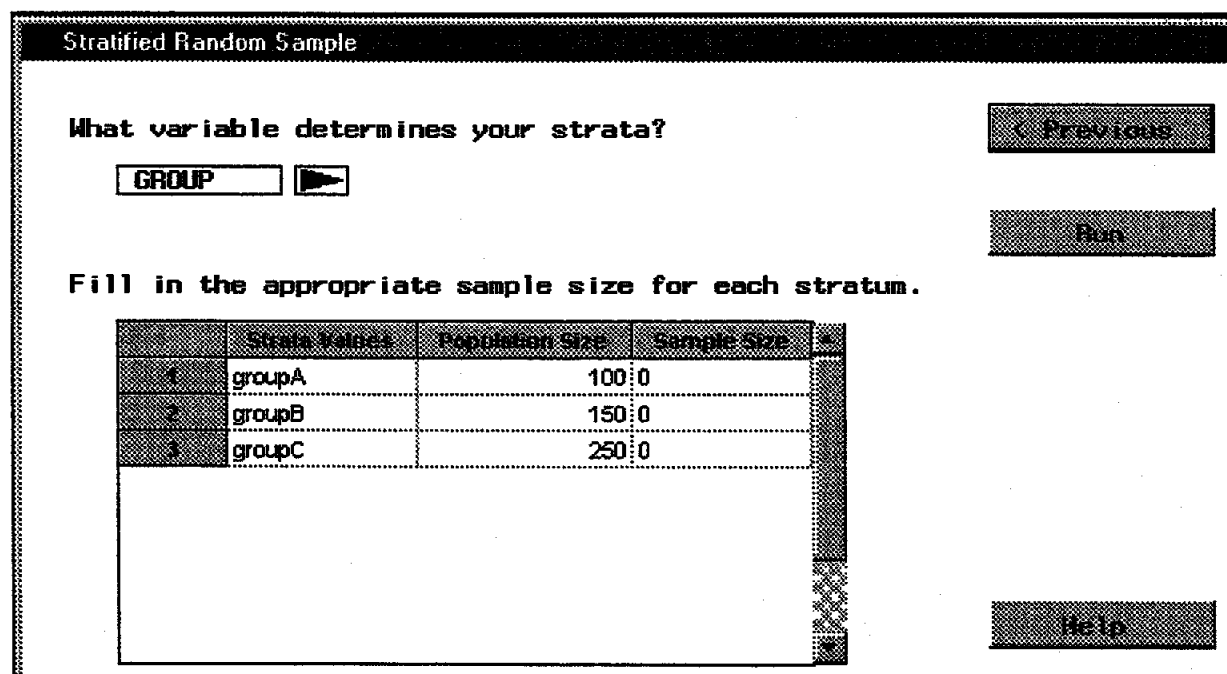
the current working directory. If you wish to save or print this report, that may easily be done using the SAS Program Editor or your favorite word processor. In this example, the report includes the sample mean, variance, and sample size along with the estimated population mean and the standard error 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 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.

Figure 11: Enter the Strata Variable



At this point, a data set containing information about the strata is displayed in a Data Table object as shown in Figure 12. The first two variables, which are protected columns, are the strata variable you specified and

Figure 12: Enter the Sample Sizes - Initially Set to Zero



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 13 illustrates entering 5, 7, and 13 for the strata groupA, groupB, and groupC, respectively. After entering the sample sizes, you click on Run to produce the sample data set, which is displayed in the 'Browse the Sample Data Set' screen as we had shown earlier. You may then browse the sample data set and estimate a population mean in the same manner as the simple random sample.

Figure 13: Enter the Sample Sizes

Stratified Random Sample

What variable determines your strata?

GROUP

▶

< Previous

Run

Fill in the appropriate sample size for each stratum.

	Strata Values	Population Size	Sample Size
1	groupA	100	5
2	groupB	150	7
3	groupC	250	13

Help

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 Sample' in the radio box on the 'Sampling Information' screen as shown in Figure 14 (again using SAMPLE.POP_C as the population data set) and clicking on 'Next.' The initial 'Cluster Sample' screen will appear as in Figure 15.

Figure 14: Completed Sample Information Screen

Sampling Information

What is the name of the data set from which to draw your sample?

SAMPLE.POP.C

▶

< Previous

Next >

What do want to name the sample data set?

WORK.SAMPLE

Which sampling technique would you like to use?

☐ Simple Random Sample
☐ Stratified Random Sample
☒ Cluster Sample

Help

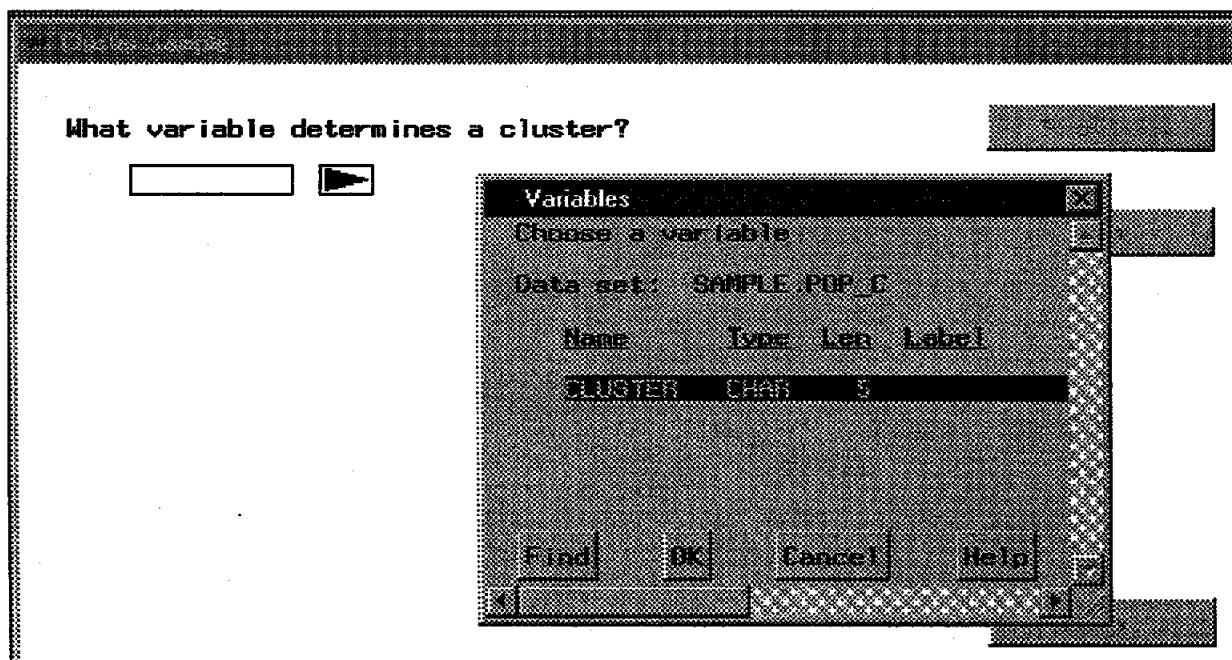
Figure 15: Initial Cluster Sample Screen



The screenshot shows a window titled "Cluster Sample". Inside, the text "What variable determines a cluster?" is followed by a text input field and a right-pointing arrow button. In the top right corner is a "Previous" button, and in the bottom right corner is a "Run" button.

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. See Figure 16. In this example, the values of CLUSTER determine the clusters.

Figure 16: Selection List for Cluster Variable



The screenshot shows the "Cluster Sample" window with a "Variables" dialog box open. The dialog box has a title bar "Variables" and a close button. It contains the text "Choose a variable" and "Data set: SAMPLE.POP.C". Below this is a table with columns "Name", "Type", "Len", and "Label". The table has one row with the value "CLUSTER" under "Name", "CHAR" under "Type", and "3" under "Len". At the bottom of the dialog box are buttons for "Find", "OK", "Cancel", and "Help".

Name	Type	Len	Label
CLUSTER	CHAR	3	

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. This is shown in Figure 17. In this example, 3 clusters will be sampled.

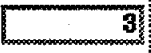
Figure 17: Number of Clusters to Sample

Cluster Sample

What variable determines a cluster?

CLUSTER 

How many clusters do you want to sample?

< Previous

Run

Finally, click on Run to produce the sample data set. Again, the 'Browse the Sample Data Set' screen will appear and you can view the sample and calculate estimates of the mean of a variable, if you desire. 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 SAS macros.

This was meant to be only a starting point as 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. If you have additional suggestions, please feel free to contact Keith at kcranford@aol.com.

References

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

Cochran, W, (1977). *Sampling Techniques*, 3rd edition, New York: John Wiley & Sons, Inc.

Kish, L. (1965). *Survey Sampling*, New York: John Wiley & Sons, Inc.

Boudreaux, D., Cranford K. (1995), *Simple Random Sampling and Subsetting Strategies Using SAS Software*, Observations, Volume 4, Number 4, Cary, NC: SAS Institute Inc.

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.

SAS Institute Inc. (1993), *Basis Statistics Using SAS/STAT Software Course Notes*, Cary, NC: SAS Institute Inc.

SAS Institute Inc. (1996), *Applying Statistics to Business Decisions: A Point and Click Approach Course Notes*, Cary, NC: SAS Institute Inc.