EXPERIMENTAL DESIGN AND SURVEY SAMPLING METHODS USING THE SAS® MACRO UTILITY

Paul Johnson, University of Wales, Aberystwyth, Wales, Great Britain

ABSTRACT

This work involves using SAS® macros for the selection, estimation and transformation of sample survey data. The macros are for use in "Experimental Design and Survey Sampling Methods". The macros are set up for use in:

a) Selecting the sample size for estimating population parameters of interest (the mean μ, the total τ and the proportion P) in stratified random sampling.
b) Estimating population parameters of interest (the mean μ, the total τ and the proportion P) in stratified random sampling.
c) Estimating population parameters of interest (the mean μ ,the total τ and the proportion P) in cluster sampling.
d) Estimating population parameters of interest (the population ratio R, the ratio mean μ_y and the ratio total τ_y) using ratio estimation.
e) Estimating the population proportion in such a way as not to identify individual responses. This is the random response model.
f) Calculating the number of callbacks that should be made to make optimal precise population parameters [a solution to the issue of nonresponse].
g) Finding the power transformation needed to transform the data to follow more closely the normal distribution.
h) Calculating the cut-off points of the strata such that an optimal allocation of the sample between strata is made.[Cumulative Square Root of the Frequency Method]

The library contains SAS macro routines for the selection, estimation and transformation of sample survey data. The set of macros can be found in eight subdirectories namely: (A,B,C,D,E,F,G and H).The series of Statistical Analysis System macros have been placed in the Internet Computational Biology Gopher at Johns Hopkins University and are available from the Biodiversity and Biological Collections Gopher as of March 1994.[© Copyright 1994, Paul Johnson]

INTRODUCTION

The random response model macro is useful when nonresponse becomes an issue due to the sensitivity of the questions asked. The callback macro solves the problem of how many callbacks should be made when there is nonresponse. The Boxcox macro determines the power transformation needed to transform the data to follow more closely the normal distribution. The strata cut-off determination macro calculates the cut-off points such that an optimal allocation of the sample between strata is made. The cumulative square root of the frequency method is used for the optimal allocation procedure. The programs pertain to the analysis of many varied data types. For illustrative purposes an energy survey data example and a second example is provided for by Silverstein (1989).It involves the effect of parity on the later development of non-insulin -dependent diabetes mellitus or impaired glucose tolerance.

The subdirectories and macros are:

A Sample Size Determination Macros

SELECTING THE SAMPLE SIZE FOR ESTIMATING POPULATION PARAMETERS OF INTEREST (THE MEAN μ, THE TOTAL τ AND THE PROPORTION P) IN STRATIFIED RANDOM SAMPLING

These are a series of SAS macros to find the sample size required to estimate the population parameter with a Bound B on the error of estimation. The macros are:

%SIZEMEAN: sample size determination for estimating the population mean μ

%SIZETOTL: sample size determination for estimating the population total τ

%SIZEPROP: sample size determination for estimating the population proportion P

%macro sizemean;

This program contains the MACRO in order to find the sample size required to estimate the population mean with a Bound B on the error of estimation. The program reads the input SAS or ASCII data set. The name of the data set is input by the user. The macro calculates the required sample size based on a stratified estimator. The macro then does an optimal allocation of the sample between strata. (Neyman Allocation).[See Neyman 1934]. For the allocation 3 functions are used in obtaining the strata sample sizes. Often the ceiling function is used where a 0.3 calculated value is not possible and hence is taken as 1, or 14.4 for example is taken as 15. This provides for a level of precision less than α.The update sample size required taking into account the certainty strata can be large. If cost is important then the floor function provides for a lower bound with a 0.3 taken as a 0. However if we sample 0 then we have no reliable estimator for that particular industry/strata combination. The level of precision will be > α.A rounding function provides for a compromise between the lower and upper bounds. In this case 0.4 is taken as 0 and 0.8 is taken as 1. It provides for a 'rounding' of the two extremes. The sample size required to estimate the mean with a bound B on the error of the estimation is output.α is the level of precision required and is input by the user. $100(1 - \alpha)\%$ is a measure of the confidence built into the estimation of the MEAN. The bound on the error of estimation and the required sample size determination formulae can be found in Elementary Survey Sampling by Scheaffer, Mendenhall and Ott (1990).The examples given assume the data is from an energy study with an industry code. Each industry contain strata. One particular number in the strata is a certainty strata. This is one in which we wish to sample all elements so that this strata has 100% coverage. The variable of interest is Kilowatt Hours (KWH) of usage. The data does not have to be from an energy study. A second example is provided for by Silverstein (1989).It involves the effect of parity on the later development of non-insulin- dependent diabetes mellitus or impaired glucose tolerance.

An example input data set may read:

IND_CODE	STRATA	FKWH
1	1	16675
1	1	63840
6	4	98438
6	4	133267
6	5	244140
6	3	57179
20	4	139520
9	3	516960
19	7	2391522
9	2	255811
14	1	14563
19	1	18060
4	1	27855
20	7	9372838
6	2	20193
21	1	32828
6	2	27180
.	.	.
.	.	.
.	.	.

etc.

{ IND_CODE is the Industry Code, STRATA is the Strata ID #
FKWH is the Frame Kilowatt-Hours usage for the specific
IND_CODE/STRATA combination }.

Or an example data set may be:

ID_CODE	STRATA	CIAA
1	1	63
1	1	68
1	2	49
1	2	56
1	3	21
2	4	20
2	1	19
2	4	89
4	3	103
.	.	.
.	.	.
.	.	.

etc.

{ID_CODE corresponds to a particular degree of obesity
STRATA corresponds to a family history of diabetes indicator
CIAA represents the Competitive Insulin Autoantibody level}.

[See Silverstein et al. (1989)]

%macro sizetotl;

This program contains the MACRO in order to find the sample size required to estimate the population total with a Bound B on the error of estimation. The program reads the input SAS or ASCII data set. The name of the data set is input by the user. The macro calculates the required sample size based on a stratified estimator. The macro then does an optimal allocation of the sample between strata. (Neyman Allocation).[See Neyman 1934]. For the allocation the 3 functions as described above are used. These are the ceiling function, the floor function and the rounding function. The sample size required to estimate the total with a bound B on the error of the estimation is output.α is the level of precision required and is input by the user. $100(1 - \alpha)\%$ is a measure of the confidence built into the estimation of the TOTAL. The bound on the error of estimation and the required sample size determination formulae can be found in Elementary Survey Sampling by Scheaffer, Mendenhall and Ott (1990).The examples given are as above for the estimation of the mean. The first is from an energy study and the second involves the effect of parity on the later development of non-insulin- dependent diabetes mellitus or impaired glucose tolerance.

%macro sizeprop;

This program contains the MACRO in order to find the sample size required to estimate the population proportion with a Bound B on the error of estimation. The program reads the input SAS or ASCII data set. The name of the data set is input by the user. The macro calculates the required sample size based on a stratified estimator. The macro then does an optimal allocation of the sample between strata. (Neyman Allocation).[See Neyman 1934]. For the allocation the 3 functions as described above are used. These are the ceiling function, the floor function and the rounding function. The sample size required to estimate the proportion with a bound B on the error of the estimation is output.α is the level of precision required and is input by the user. $100(1 - \alpha)\%$ is a measure of the confidence built into the estimation of the PROPORTION. The bound on the error of estimation and the required sample size determination formulae can be found in Elementary Survey Sampling by Scheaffer, Mendenhall and Ott (1990).The first example given assume the data is from an energy study with an industry code. The second example involves the effect of parity on the later development of non-insulin-dependent diabetes mellitus or impaired glucose tolerance.

B Estimation Macros [Stratified Random Sampling]

ESTIMATING POPULATION PARAMETERS OF INTEREST
(THE MEAN μ ,THE TOTAL τ AND THE PROPORTION P)
IN STRATIFIED RANDOM SAMPLING

These are a series of SAS macros to estimate the population parameter with a Bound B on the error of estimation. The estimated variances of the estimators are calculated. The macros are:

%CALCMEAN	estimating the population mean	μ
%CALCTOTL	estimating the population total	τ
%CALCPROP	estimating the population proportion	P

%macro calcmean;

This program contains the MACRO in order to estimate the population mean with a Bound B on the error of estimation. The program reads the input SAS or ASCII data set. The name of the data set is input by the user. The macro calculates the estimate of the mean based on a stratified estimator. The variance estimate of the estimator is calculated.α is the level of precision required and is input by the user. $100(1 - \alpha)\%$ is a measure of the confidence built into the estimation of the MEAN. The bound on the error of estimation, the variance estimate and the mean estimate formulae can all be found in Elementary Survey Sampling by Scheaffer, Mendenhall and Ott (1990).The examples are described above Cochran (1977) provides for a further discussion of this estimation problem.

%macro calctotl;

This program contains the MACRO in order to estimate the population total with a Bound B on the error of estimation. The program reads the input SAS or ASCII data set. The name of the data set is input by the user. The macro calculates the estimate of the total based on a stratified estimator. The variance estimate of the estimator is calculated.α is the level of precision required and is input by the user. $100(1 - \alpha)\%$ is a measure of the confidence built into the estimation of the TOTAL.

The bound on the error of estimation, the variance estimate and the total estimate formulae can all be found in Elementary Survey Sampling by Scheaffer, Mendenhall and Ott (1990).The examples are described above. Cochran (1977) provides for a further discussion of this estimation problem.

%macro calcprop;

This program contains the MACRO in order to estimate the population proportion with a Bound B on the error of estimation. The program reads the input SAS or ASCII data set. The name of the data set is input by the user. The macro calculates the estimate of the proportion based on a stratified estimator. The variance estimate of the estimator is calculated.α is the level of precision required and is input by the user. $100(1 - \alpha)\%$ is a measure of the confidence built into the estimation of the PROPORTION. The bound on the error of estimation, the variance estimate and the proportion estimate formulae can all be found in Elementary Survey Sampling by Scheaffer, Mendenhall and Ott (1990).The examples are described above. Cochran (1977) provides for a further discussion of this estimation problem.

C Estimation Macros [Cluster Sampling]

ESTIMATING POPULATION PARAMETERS OF INTEREST
(THE MEAN μ ,THE TOTAL τ AND THE PROPORTION P)
IN CLUSTER SAMPLING

These are a series of SAS macros to estimate the population mean, the population total and the population proportion with a Bound Bon the error of estimation. The macros are:

%CLUSMEAN	estimating the population mean	μ
%CLUSTOTL	estimating the population total	τ
%CLUSPROP	estimating the population proportion	P

%macro clusmean;

This program contains the MACRO in order to estimate the population mean with a Bound B on the error of estimation. The program reads the input SAS or ASCII data set. The name of the data set is input by the user. The macro estimates the population mean based on a cluster sampled estimator. The variance of the estimator is estimated and the Bound on the error of estimation is also calculated. α is the level of precision required and is input by the user. $100(1 - \alpha)\%$ is a measure of the confidence built into the estimation of the MEAN. The bound on the error of estimation, the estimator for the mean and its variance formulae can be found in Elementary Survey Sampling (1990).

%macro clustotl;

This program contains the MACRO in order to estimate the population total with a Bound B on the error of estimation. The program reads the input SAS or ASCII data set. The name of the data set is input by the user. The macro estimates the population total based on a cluster sampled estimator. The variance of the estimator is estimated and the Bound on the error of estimation is also calculated. α is the level of precision required and is input by the user. $100(1 - \alpha)\%$ is a measure of the confidence built into the estimation of the TOTAL. The bound on the error of estimation, the estimator for the total and its variance formulae can be found in Elementary Survey Sampling (1990).

%macro clusprop;

This program contains the MACRO in order to estimate the population proportion with a Bound B on the error of estimation. The program reads the input SAS or ASCII data set. The name of the data set is input by the user. The macro estimates the population proportion based on a cluster sampled estimator. The variance of the estimator is estimated and the Bound on the error of estimation is also calculated. α is the level of precision required and is input by the user. $100(1 - \alpha)\%$ is a measure of the confidence built into the estimation of the PROPORTION. The bound on the error of estimation, the estimator for the proportion and its variance formulae can be found in Elementary Survey Sampling (1990).

D Estimation Macros [Ratio Estimation]

ESTIMATING POPULATION PARAMETERS OF INTEREST (THE POPULATION RATIO R, THE RATIO MEAN μ_y AND THE RATIO TOTAL τ_y) RATIO ESTIMATION

These are a series of SAS macros to estimate the population ratio, the population ratio mean and the population ratio total with a Bound on the error of estimation. The macros are:

%RATIO	estimating the population ratio	R
%RATIOMN	estimating the population ratio mean	μ_y
%RATIOTL	estimating the population ratio total	τ_y

%macro ratio;

This program contains the MACRO in order to estimate the population ratio with a Bound B on the error of estimation. The program reads the input SAS or ASCII data set. The name of the data set is input by the user. The macro estimates the population ratio based on ratio estimation techniques. The variance of the estimator is estimated and the Bound on the error of estimation is also calculated. α is the level of precision required and is input by the user. $100(1 - \alpha)\%$ is a measure of the confidence built into the estimation of the RATIO. The bound on the error of estimation, the estimator for the ratio and its variance formulae can be found in Elementary Survey Sampling (1990).

%macro ratiomn;

This program contains the MACRO in order to estimate the population ratio mean with a Bound B on the error of estimation. The program reads the input SAS or ASCII data set. The name of the data set is input by the user. The macro estimates the population ratio mean based on ratio estimation techniques. The variance of the estimator is estimated and the Bound on the error of estimation is also calculated. α is the level of precision required and is input by the user. $100(1 - \alpha)\%$ is a measure of the confidence built into the estimation of the RATIO MEAN. The bound on the error of estimation, the estimator for the ratio mean and its variance formulae can be found in Elementary Survey Sampling (1990).

%macro ratiotl;

This program contains the MACRO in order to estimate the population ratio total with a Bound B on the error of estimation. The program reads the input SAS or ASCII data set. The name of the data set is input by the user. The macro estimates the population ratio total based on ratio estimation techniques. The variance of the estimator is estimated and the Bound on the error of estimation is also calculated. α is the level of precision required and is input by the user. $100(1 - \alpha)\%$ is a measure of the confidence built into the estimation of the RATIO TOTAL. The bound on the error of estimation, the estimator for the ratio total and its variance formulae can be found in Elementary Survey Sampling (1990).

E Random Response Model Macro

ESTIMATING THE POPULATION PROPORTION IN SUCH A WAY AS NOT TO IDENTIFY INDIVIDUAL RESPONSES THE RANDOM RESPONSE MODEL

This is a SAS macro to estimate the population proportion P with a bound B on the error of estimation when it is necessary to not identify the individual responses. The macro is:

%RRESP	estimating the population proportion P using the random response model.

%macro rresp;

This program contains the MACRO in order to estimate the population proportion with a Bound B on the error of estimation. The program reads the input SAS or ASCII data set. The name of the data set is input by the user. The macro estimates the population proportion based on the random response model. The variance of the estimator is estimated and the Bound on the error of estimation is also calculated. α is the level of precision required and is input by the user. $100(1 - \alpha)\%$ is a measure of the confidence built into the estimation of the PROPORTION. The bound on the error of estimation, the estimator of the proportion and its variance formulae can be found in Elementary Survey Sampling (1990). Persons being interviewed about personal medical problems or habits often refuse to answer or give correct answers. For example some persons may not respond truthfully to "Have you taken drugs intravenously in the past year?". This question would perhaps be one in an AIDS health study. This macro estimates the proportion of people who have some characteristic of interest without obtaining direct answers from the people interviewed. The method is due to S.L.Warner (1965). Warner showed that it is possible to estimate this proportion without revealing his/her personal status with respect to the question. Cochran (1977) presents an alternative to the Warner method.

An example input data set may read:

RESPONSE
no
yes
yes
yes
no
yes
no
.
.
.
etc.

where a yes indicates the person has seen a card randomly which is in agreement to his group identity. The cards contain a fixed proportion theta of cards with a '+' on them. Hence if a person belongs to the '+' group and sees a card with a '+' on it he/she responds yes. Otherwise he/she responds no. The card is replaced before the next person draws. The interviewer does not see the card and simply records whether the response is yes or no.

F The Callback Macro

CALCULATING THE NUMBER OF CALLBACKS THAT SHOULD BE MADE TO MAKE OPTIMAL PRECISE POPULATION PARAMETERS A SOLUTION TO THE NONRESPONSE PROBLEM

This is a SAS macro to estimate the number of callbacks to be made. The initial cost estimate of sampling, the cost estimate for a standard response and the cost estimate for a callback response are input variables. The number of callbacks made is a value that minimizes the expected cost of sampling for a fixed value of the variance estimate of an unbiased estimator of the population mean. Nonresponse is an important problem to consider in any survey. Suppose out of a size n only n1 responses are obtained. If n1 is much smaller than n (i.e., n1 << n) then callbacks need to be made.

%CALLBACK estimating the number of callbacks to be made when the problem of nonresponse is to be addressed.

%macro callback;

This program contains the MACRO in order to estimate the number of callbacks that should be made to make optimal precise population parameters. The program reads the input SAS or ASCII data set. The name of the data set is input by the user. The macro estimates the number of callbacks to be made by minimizing the expected cost of sampling for a fixed value of the variance estimate of an unbiased estimator of the population mean. The formulae and methodology for estimating the number of callbacks can be found in Elementary Survey Sampling (1990). The method provides for an optimal rule for allocating resources to callbacks. Cochran (1977) provides reference material for further discussion. The example given assume the data is from an energy study with an industry code. Each industry contain strata. One particular number in the strata is a certainty strata. This is one in which we wish to sample all elements so that this strata has 100% coverage. The variable of interest is Kilowatt Hours (KWH) of usage. The data does not have to be from an energy study. A second example data set is the effect of parity on the later development of non-insulin-dependent diabetes mellitus or impaired glucose tolerance. The example is provided for by Silverstein et al. (1989).

G The Boxcox Macro

BOX-COX TRANSFORMATION

This is a SAS macro to find the power transformation needed to transform data to follow more closely the normal distribution. It is understood that the transformation obtained by maximizing the log-likelihood usually improves the approximation to normality.

%macro boxcox;

This program contains the MACRO in order to find the BOX-COX Power Transformation (Box and Cox 1964).The program reads an input SAS created (SSD) data set from the user. The data needs to consist of a vector of observations. The program then performs the BOX-COX Procedure to find a transformation that will aid in transforming data to follow more closely the normal distribution. The user inputs the lower bound where the search is to begin and the number of searches to be made. The increment for the search is also input by the user. Under each of the search transformations the log-likelihood function is evaluated. The transformation that produces the largest value for the log-likelihood function is chosen over the search grid. In the case when the largest value of the function is found is at either the lower or upper end of the grid then a new search should be made increasing the grid size (i.e., start at a different lower bound and/or increase the number of searches to be made).Choosing the transformation that produces the largest value for the log-likelihood function is equivalent to choosing the transformation that produces the largest value for the likelihood function. This is due to the fact that the logarithm is a strictly increasing function. Hence the maximum of the likelihood and the maximum of the log-likelihood occur at the same point. This transformation chosen (called LAMBDA) is the required power transformation. Johnson and Wichern (1982) discuss using the Box-Cox method to choose from a broad class of transformations.

H Strata Cut-off Point Determination Macro

CALCULATING THE CUT-OFF POINTS OF THE STRATA SUCH THAT AN OPTIMAL ALLOCATION OF THE SAMPLE BETWEEN STRATA IS MADE
CUMULATIVE SQUARE ROOT OF THE FREQUENCY METHOD

This is a SAS macro to calculate the cut-off points using the cumulative square root of the frequency method. This method works well for delineating strata. The macro is:

%OPTIMAL calculating the cut-off points.

%macro optimal;

This program contains the MACRO in order to calculate the cut-off points of the strata such that an optimal allocation of the sample between strata is made. The optimal allocation is made to minimize the total variability based upon the variances within each possible level makeup as a result of the choosing of the number of strata and of the strata cut-off points chosen. The program reads an input SAS or ASCII data set. The name of the data set is input by the user. The macro calculates the cut-off points based on the cumulative square root of the frequency method. Formulae for the calculations can be found in Elementary Survey Sampling (1990).Further examples and results can be found in Cochran (1977). The cumulative square root of the frequency method ,due to Dalenius and Hodges (1959), provides for minimum variance stratification.

SAMPLE PROGRAM

BOX-COX TRANSFORMATION

This is a SAS macro to find the power transformation needed to transform data to follow more closely the normal distribution. It is understood that the transformation obtained by maximizing the log-likelihood usually improves the approximation to normality.

```
%macro boxcox;

/* This program contains the MACRO in order to find the BOX-COX
Power Transformation (Box and Cox 1964).The program reads an input
SAS created (SSD) data set from the user. The data needs to consist of a
vector of observations. The program then performs the BOX-COX
Procedure to find a transformation that will aid in transforming data to
follow more closely the normal distribution. The user inputs the lower
bound where the search is to begin and the number of searches to be
made. The increment for the search is also input by the user. Under each
of the search transformations the log-likelihood function is evaluated. The
transformation that produces the largest value for the log-likelihood
function is chosen over the search grid. In the case when the largest
value of the function is found is at either the lower or upper end of the
grid then a new search should be made increasing the grid size (i.e., start
at a different lower bound and/or increase the number of searches to be
made).Choosing the transformation that produces the largest value for
the log-likelihood function is equivalent to choosing the transformation
that produces the largest value for the likelihood function. This is due to
the fact that the logarithm is a strictly increasing function. Hence the
maximum of the likelihood and the maximum of the log-likelihood occur
at the same point. This transformation chosen (called LAMBDA) is the
required power transformation. Johnson and Wichern (1982) discuss
using the Box-Cox method to choose from a broad class of
transformations.

USAGE:    %BOXCOX

INPUT:
   1. Input the number 1 if a SAS (*.SSD) data set is to be used in
      finding the Box-Cox transformation for a vector of
      observations. Input the number 2 if an ASCII text file is to be
      used.

   2. Input the direotory location for the data set with quotes
      (e.g., 'C:\MYSAS\' if the data set is to be found on the hard drive
      C: and in the sub-directory named SAS).

For a SAS data set (*.SSD):

   3. Input the file directory and file name (e.g., OLD TESTBOX if the
      sas data set is called TESTBOX.ssd and was created in the sas
      library [i.e., libname] called OLD). Input the data variable name
      [the variable for which the transformation is required] (e.g., X).
      Go To 4.

For an ASCII text data file (e.g., *.txt,*.dat, etc.):

   3. Input the name for the ASCII text file (e.g., 'TESTBOX.DAT')

   4. The lower bound for which the search to maximize the log-
      likelihood is to be made (e.g., -1.2).
   5. The number of searches to be made (e.g., 10).
   6. The incremental unit to be added at each stage of the search
      beginning at the lower bound (e.g., 0.2).

The program asks for this information as the macro runs.

An example input data set may be the Competitive Insulin Autoantibody
(CIAA) of subjects at high risk for development of Type 1 Diabetes
Mellitus [See Vardi et al. (1987)].
Let X represent the CIAA level.

        X

       63
       71
       72
       44
        .
      etc.

OUTPUT:  LAMBDA  (the power transformation to be used to make
                  the original data more normally
                  distributed).
         MAXL_LIK (the value of the log-likelihood for the
                  LAMDA transformation; this value is
                  the maximum log-likelihood over the
                  grid search).

  NOTE:  For each run of the Boxcox macro to clear parameter
settings it may be necessary to exit the SAS environment.
                                                              */

/*      Information on retrieval of data is obtained          */

%put FIRST PRESS THE ENTER (RETURN) KEY;
%put ENTER DIRECTORY LOCATION FOR DATA SET, THEN
PRESS F10;
%put PLEASE REMEMBER TO USE OPENING AND CLOSING
QUOTES e.g., 'c:\mysas\';
%put THEN PRESS F10 --- THE SUBMIT BUTTON;
%input name3;
%put &name3;
%put ENTER THE NUMBER 1 IF A SAS SSD DATA SET IS TO BE
USED;
%put ENTER THE NUMBER 2 IF AN ASCII TEXT DATA FILE IS
TO BE USED;
%put THEN PRESS F10 --- THE SUBMIT BUTTON;
%input ind;
%put &ind;
%if "&ind" = "1" %then %do;
%put ENTER NAMES FOR FILE DIRECTORY AND FILE NAME;
%put THEN PRESS F10 --- THE SUBMIT BUTTON;
%put e.g., old testbox ;
%input name1 name2 ;
%put &name1 &name2 ;

%put ENTER DATA VARIABLE NAME.;
%put (THE VARIABLE FOR WHICH A TRANSFORMATION IS
REQUIRED).;
%put THEN PRESS F10 --- THE SUBMIT BUTTON;
%put e.g., x;
%input x2;
%put &x2;
%end;
%if "&ind" = "2" %then %do;
%put ENTER THE NAME FOR THE ASCII TEXT DATA FILE;
%put PLEASE REMEMBER TO USE OPENING AND CLOSING
QUOTES e.g., 'testbox.dat';
%put THEN PRESS F10 --- THE SUBMIT BUTTON;
%put e.g., 'testbox.dat' ;
%input name1 ;
%put &name1 ;
%end;
```

```
/*      The search grid is obtained from information about the lower
bound, the number of  searches to be made and the incremental unit to be
added at each stage of the search    */

%put ENTER THE LOWER BOUND;
%put FOR WHICH THE SEARCH TO MAXIMIZE THE
LOGLIKELIHOOD IS TO BE MADE.;
%put THEN PRESS F10 --- THE SUBMIT BUTTON;
%put e.g., -1.2;
%input lower2;
%put &lower2;
%put ENTER THE NUMBER OF SEARCHES TO BE MADE.;
%put THEN PRESS F10 --- THE SUBMIT BUTTON;
%put e.g., 10;
%input number;
%put &number;
%put ENTER THE INCREMENTAL UNIT TO BE ADDED TO THE
LOWER BOUND.;
%put THIS UNIT WILL BE ADDED FOR THE REQUIRED NUMBER
OF SEARCHES.;
%put THEN PRESS F10 --- THE SUBMIT BUTTON;
%put e.g., 0.2;
%input increas;
%put &increas;

  /* Options set for pagesize, linesize, no page number, no date, etc. Can
be changed by the user  e.g., options pagesize = 60 linesize = 132 nodate
nonumber                                  */

options nodate nonumber;

  /*  Data Retrieval for a SAS (*.SSD) data set                  */

%if "&ind" = "1" %then %do;
 %let nn1= ;
    libname &name1.&nn1.&name3;
    data a;set &name1..&name2;
    data a;set a;
        xx =&x2;
%end;

  /*  Data Retrieval for an ASCII text data file                 */

%if "&ind" = "2" %then %do;
    data a;infile &name1;
        input xx;
%end;
    proc sort data=a;
        by xx;
    data a;set a;
        title1 'THE BOX-COX TRANSFORMATION ';
        title2 'The Lower and Upper Bounds for the Search Grid';
        ll=&lower2;
        incr=&increas;
        numb=&number;
        uu = ll+(numb-1)*incr;
    data aa;set a;
        nn = _N_;
        if nn = 1;
        drop nn;
    proc print data=aa split ='*' noobs ;
        var ll uu;
        label ll = 'Lower*Bound'
        uu = 'Upper*Bound';
%do i = 1 %to &number;
    data a;set a;
        lower=&lower2;
        increase =&increas;
%let j=%eval(&i-1);
        j=&j;

 /*     Transformation is increased by the increment at each stage of the
        search                  */

        lambda=lower+increase*j;
        if lambda=0 then lambda=0.001;
   data a;set a;
        y=log(xx);
        z=(xx**lambda-1)/lambda;
   proc means data=a noprint;
        var y z;
        id lambda;
        output out=b n=n1 n2 var=var1 var2 sum = sum1 sum2;

 /*  The log-likelihood is calculated at each stage of the search   */

    data b;set b;
        c91=var2*(n2-1)/n2;
        c71=sum1;
        c51=(-n2*log(c91)/2)+(lambda-1)*c71;
        if abs(lambda)<0.000001 then lambda=0;
        lambda&i=lambda;
        loglik&i=c51;
        loglik=c51;

    data dat&i;set b;
        keep lambda&i loglik&i;
%end;
    data cc;set dat1;
        lambda=lambda1;
        loglik=loglik1;
        keep loglik lambda;
%do j=2 %to &number;
    data cc&j;set dat&j;
        lambda=lambda&j;
        loglik=loglik&j;
        keep loglik lambda;
    proc append base=cc data=cc&j force;
%end;
    data cc;set cc;
        indicate='a';
    proc sort data=cc;
        by lambda;
    proc print data=cc noobs split ='*';
        var lambda loglik ;
        label loglik = 'Log*Likelihood';

  /*  The LAMDA power Transformation corresponding to the value
that  maximizes the log-likelihood is obtained                 */

title1 'The Log-Likelihood is Maximized giving the required Power
Transformation';
    proc means data=cc noprint;
        var loglik;
        output out=dd max=maxlik;
    data dd;set dd;
        indicate='a';
    data ee;merge cc dd;
        by indicate;
        if loglik=maxlik;
        drop indicate;
        maxl_lik=maxlik;
    proc print data =ee split = '*' noobs;
        var lambda maxl_lik;
        label maxl_lik = 'Maximum*Log-Likelihood';
```

```
title1 'THE BOX-COX TRANSFORMATION ';
title2 'If LAMBDA = 0 or close to 0 Then Use the LOGe Transformation';
title3 'If LAMBDA = 1 or close to 1 Then Use NO Transformation';
title4 'If LAMBDA = 0.5 or close to 0.5 Then Use the Square Root
Transformation';
title5 'If LAMBDA = -1 or close to -1 Then Use the Reciprocal
Transformation';
%put ;
%put For the next run of the Boxcox macro you will need to exit SAS.;
%put ;
        run;
%mend boxcox;
%boxcox
```

REFERENCES

Box, G.E.P., and Cox, D.R. (1964). "An Analysis of Transformations," J. Roy. Statist. Soc. Ser. B 39:211-252.

Cochran, W.G. (1977). Sampling Techniques. 3rd ed. New York, Wiley.

Johnson, P. (1994). "Survey Sampling Methods for the Analysis of Clinical Trial Survey and Energy Survey Data," Internet Computational Biology Gopher, Johns Hopkins University.

Johnson, R.A., and D.Wichern. (1982) Applied Multivariate Statistical Analysis. Englewood Cliffs, N.J.:Prentice-Hall.

Neyman, J. (1934). "On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection," J. Roy. Statist. Soc. 97:558-606.

SAS Institute. Inc. (1993). SAS Guide to Macro Processing, Version 6, 2nd ed. SAS Institute. Inc., Cary, NC.

Scheaffer, R.L., Mendenhall, W., and Ott, L. (1990). Elementary Survey Sampling. Boston, Duxbury Press.

Silverstein D., Barrett-Conner, E., and Winguard D. (1989). "The effect of parity on the later development of non-insulin-dependent diabetes mellitus or impaired glucose tolerance," N. Engl. J. Med. 321:1214-1219.

SAS is a registered trademark of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.
Other brand and product names are registered trademarks or trademarks of their respective companies.