A SAS® MACRO FOR SIMPLE RANDOM SAMPLING

Chang Jiang
Highmark Blue Cross Blue Shield, Pittsburgh, PA

ABSTRACT
The methodology of simple random sampling is widely used for selecting a sample from a population to develop estimates of the values of population characteristics. This paper presents a SAS macro that makes the implementation of simple random sampling easy and fast.

INTRODUCTION
Information contained in a sample can be used to obtain the information about a population. Several methods are available for selecting a sample from a population. One of the most common is simple random sampling.

This paper presents a SAS macro for implementing simple random sampling. It can be a handy tool for SAS programmers and analysts who need to pull a sample from a population. The sampling scheme implemented by the macro is a simple version of the sampling methodology developed by the National Committee for Quality Assurance (NCQA), which has been widely used by managed care organizations such as Highmark for HEDIS 2000\(^1\) data collection to measure and report the quality of managed care.

THE SCHEME FOR SIMPLE RANDOM SAMPLING
The sampling method consists of the following steps:

1) Determine the eligible member population (EM)
Develop a list of eligible members (EM) from which the sample will be draw.

2) Determine the minimum required sample size (MRSS)
MRSS is the number of members in the sample.

3) Determine the final number of members to sample (FSS)
The final sample size includes the MRSS (from step 2) plus an adequate number of additional records in order to make substitutions for incomplete records. The over-sampling rates may range from 5 percent to 20 percent. The final sample size (FSS) is calculated by the following formula:

\[
FSS = MRSS + (MRSS \times \text{oversampling rate}) \\
\text{round up to the next whole number, where EM=>FSS=>MRSS}
\]

4) Calculate \(N\)
\(N\) is the number used in the formula to determine which member will start the sample. \(N\) is calculated using the equation:

\[
N = \frac{EM}{FSS} \text{ (round down to a whole number)}
\]

5) Calculate \(START\)
The sample should be selected from a single pass through the member list. \(START\) is the starting point from which to draw the final sample, it could have many values and still allow only one pass through the list. \(START\) is calculated as follows:

\[
START = (\text{RAND} \times N) \\
\text{(Round per the .5 rule to the nearest whole number greater than 0)}
\]

Where RAND is a random number, that has a value between 0-1 and could be selected from a Random Number table.

6) Select the sample
Starting with the member corresponding to the number \(START\), choose every \(i\)th member using the following formula until the minimum required sample size (MRSS) is met.

\[
\text{ith member} = START + [(i-1) \times (EM/FSS)] \\
\text{(Rounding down to a whole number)}
\]

For \(i=2,3,4,...,FSS\)

\(^1\) HEDIS 2000 is the latest edition of the Health Plan Employer Data and Information Set. It is the most widely used set of performance measures in the managed care industry.
This becomes the primary list of sampled members. Then continue choosing every ith member until the final sample size (FSS) is met. This set of members becomes the auxiliary list of sampled members (i.e., the over sample).

**THE MACRO FOR SAMPLING**

/********************************/
/*Positional Parameters:
* EMDS is the input data set of
* eligible member population (EM), from
* which the sample will be draw.
* SAMPLE is the output sample.
* RAND is the randomization number for
* determining START.
*
* Keyword Parameters:
* MRSS is the minimum required sample
* size.
* OVERSAM is the oversampling rate, and
* the default is 5%.
*/
/*****************************/

%Macro SAMPLE
(EMDS,SAMPLE,RAND,MRSS=,OVERSAM=0.05);

DATA _NULL_;   
FSS=CEIL(&MRSS*(1+&OVERSAM));   
CALL SYMPUT('FSS',LEFT(PUT(FSS,8.)));   
RUN;   /* get the number of FSS and store it in &FSS */

DATA _NULL_;   
IF 0 THEN SET &EMDS NOBS=EM;   
CALL SYMPUT('EM', LEFT(PUT(EM,8.)));   
STOP;   /* get the number of EM and store it in &EM at compile time */

DATA &EMDS;  
SET &EMDS;  
OBSNUM=_N_;  /*use OBSNUM to track chosen members */
RUN;

DATA _NULL_;   
N=FLOOR(&EM/&FSS);   
START=MAX(ROUND(&RAND*N),1);   /* round START using .5 rule */
CALL SYMPUT('N', LEFT(PUT(N,8.)));   
CALL SYMPUT('START',LEFT(PUT(START,8.)));   
RUN;

EXAMPLE
The example below will illustrate the use of the macro.

DATA A;
  INPUT GROUP $ FREQ;
  CARDS;
  A 11
  B 12
  C 13
  D 14
  E 15
  F 16
  G 17
  H 18
  J 19
  K 20
  L 21
  M 22
  N 23
  P 24
  A 11
  B 12
  C 13
  D 14
  E 15
  F 16
  G 17
  H 18
  J 19

DATA &SAMPLE(DROP=I);  
LENGTH LIST $7;  
DO I=1 TO &FSS;  
  OBSIN=&START+FLOOR((I-1)*(&EM/&FSS));  
  SET &EMDS POINT=OBSIN;  
  /*draw members by their observation #*/  
  IF I <= &MRSS THEN LIST='PRIMARY';  
  ELSE LIST='AUXILIA';  
  OUTPUT;  
END;  
STOP;  
RUN;

%MEND SAMPLE;
The output sample of this macro is shown as follows:

<table>
<thead>
<tr>
<th>OBS</th>
<th>GROUP</th>
<th>FREQ</th>
<th>OBSNUM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>PRIMARY</td>
<td>A</td>
<td>11</td>
</tr>
<tr>
<td>2</td>
<td>PRIMARY</td>
<td>C</td>
<td>13</td>
</tr>
<tr>
<td>3</td>
<td>PRIMARY</td>
<td>F</td>
<td>16</td>
</tr>
<tr>
<td>4</td>
<td>PRIMARY</td>
<td>H</td>
<td>18</td>
</tr>
<tr>
<td>5</td>
<td>PRIMARY</td>
<td>L</td>
<td>21</td>
</tr>
<tr>
<td>6</td>
<td>PRIMARY</td>
<td>N</td>
<td>23</td>
</tr>
<tr>
<td>7</td>
<td>PRIMARY</td>
<td>B</td>
<td>12</td>
</tr>
<tr>
<td>8</td>
<td>PRIMARY</td>
<td>D</td>
<td>14</td>
</tr>
<tr>
<td>9</td>
<td>PRIMARY</td>
<td>G</td>
<td>17</td>
</tr>
<tr>
<td>10</td>
<td>PRIMARY</td>
<td>J</td>
<td>19</td>
</tr>
<tr>
<td>11</td>
<td>PRIMARY</td>
<td>M</td>
<td>22</td>
</tr>
<tr>
<td>12</td>
<td>PRIMARY</td>
<td>P</td>
<td>24</td>
</tr>
<tr>
<td>13</td>
<td>PRIMARY</td>
<td>D</td>
<td>14</td>
</tr>
<tr>
<td>14</td>
<td>PRIMARY</td>
<td>F</td>
<td>16</td>
</tr>
<tr>
<td>15</td>
<td>PRIMARY</td>
<td>J</td>
<td>19</td>
</tr>
<tr>
<td>16</td>
<td>PRIMARY</td>
<td>L</td>
<td>21</td>
</tr>
<tr>
<td>17</td>
<td>PRIMARY</td>
<td>P</td>
<td>24</td>
</tr>
<tr>
<td>18</td>
<td>PRIMARY</td>
<td>B</td>
<td>12</td>
</tr>
<tr>
<td>19</td>
<td>PRIMARY</td>
<td>E</td>
<td>15</td>
</tr>
<tr>
<td>20</td>
<td>PRIMARY</td>
<td>G</td>
<td>17</td>
</tr>
<tr>
<td>21</td>
<td>AUXILIA</td>
<td>K</td>
<td>20</td>
</tr>
<tr>
<td>22</td>
<td>AUXILIA</td>
<td>M</td>
<td>22</td>
</tr>
</tbody>
</table>

REFERENCE


ACKNOWLEDGMENT

SAS is a registered trademark of SAS Institute Inc. in the USA and other countries. © indicates USA registration.

CONTACT INFORMATION

Chang Jiang
Consumer & Product Analyses
Highmark Blue Cross Blue Shield
Fifth Avenue Place
120 Fifth Avenue, Ste P7205
Pittsburgh PA 15222
(412) 544-2432
chang.jiang@highmark.com