UMA ROTINA PARA ESTIMATIVAS DE AMOSTRA EM DOIS ESTÁGIOS USANDO O SAS®[1]

Vera Lúcia Ferraz dos Santos Francisco[2]
Lilian Cristina Anefalos[3]
Maria de Lourdes Sumiko Sueyoshi[4]
Sérgio Augusto Galvão Cézar[5]

1 - INTRODUÇÃO

Define-se como levantamento censitário o que utiliza todos os elementos da população; por amostragem aquele que abrange parte desse total, podendo ser probabilística, ou seja, com seleção dos elementos baseada numa probabilidade conhecida (não nula) e conseqüente cálculo de variância e erros amostrais, ou não, caso contrário.

Assim, para se obter alguma informação ou se medir determinado parâmetro de uma dada população, como por exemplo, preço, salário, produção e outros, necessita-se fazer uma coleta de dados nas unidades de interesse.

Como a garantia da aleatoriedade do processo, além de outras características vantajosas como as viabilidades econômica e operacional, se dá através de amostragem probabilística, esta é a alternativa mais recomendada. A partir daí deve-se estabelecer o esquema amostral a ser adotado, que consiste na seleção dos elementos da população e estimação dos seus valores verdadeiros.

Há casos, porém, que nem sempre é possível se ter acesso à relação de toda população, denominada sistema de referência ou cadastro, para a seleção da amostra. Uma solução para esse problema seria trabalhar com outros elementos de uma lista correlata, para a formação de grupos, através de amostragem casual por conglomerados onde procura-se aumentar a heterogeneidade dentro de cada um deles e minimizar as diferenças entre as suas médias.

Nesse delineamento pode haver um, dois ou mais estágios, com uso de conglomerados naturais, como por exemplo: a) propriedades agrícolas no primeiro estágio e plantas no segundo estágio e b) municípios no primeiro estágio e propriedades agrícolas no segundo estágio. Sua principal vantagem, em relação a outros esquemas amostrais mais comuns, é a redução dos custos, aliada à maior facilidade de planejamento e organização do levantamento, apesar de geralmente produzir estimadores menos eficientes que as amostras estratificada e casual simples, fórmulas mais complexas e aleatoriedade do tamanho da amostra. Para maiores detalhes sobre o estudo da teoria de amostragem sugerem-se COCHRAN (1965) e KISH (1965), entre outros.

O presente trabalho tem como objetivo desenvolver uma rotina para o cálculo de estimativas (média e variância da média) em amostras de dois estágios, com conglomerados de tamanhos desiguais, sorteados com probabilidades iguais e subamostragem proporcional ao tamanho da unidade de primeiro estágio.

2 - ESTIMADORES

Esse esquema amostral aplica-se, principalmente, a levantamentos fitossanitários e pode ser ilustrado com o seguinte exemplo: pretende-se estimar o percentual de laranjeiras afetadas por CVC (Clorose Variegada de Citros) de uma determinada região, em diversos níveis de infestação. Nesse caso, devem ser

[1]Trabalho apresentado no 4º Congresso Brasileiro de Usuários SAS, Piracicaba, 4 a 7 de abril de 1995. Recebido em 08/03/95. Liberado para publicação em 28/04/95.

[2]Estatístico, Pesquisador Científico do Instituto de Economia Agrícola.

[3]Engenheiro Agrônomo, Pesquisador Científico do Instituto de Economia Agrícola.

[4]Matemático, Pesquisador Científico do Instituto de Economia Agrícola.

[5]Engenheiro Agrônomo, MS, Pesquisador Científico do Instituto de Economia Agrícola.

sorteadas aleatoriamente *n* propriedades agrícolas, com igual probabilidade de seleção, dentro do cadastro de propriedades citrícolas (unidade conglomerada no 1º estágio). O segundo estágio refere-se ao sorteio das plantas de acordo com uma proporção fixa (f_2) em relação ao número total de plantas da propriedade. Deve-se ressaltar que o conhecimento desse número só é necessário para as propriedades sorteadas no 1º estágio.

Os estimadores, cujas fórmulas foram extraídas de COCHRAN (1965), são os seguintes:

$$\bar{Y} = \frac{1}{n\bar{M}} \sum_{i=1}^{n} M_i \bar{y}_i$$

fornece estimativa não viesada da média (proporção); e

$$v(\bar{Y}) = \frac{1 - f_1}{n\bar{M}^2} \frac{\sum_{i=1}^{n} (M_i \bar{y}_i - \bar{y})^2}{n - 1} + \frac{f_1(1 - f_2)}{n\bar{m}\bar{M}} \sum_{i=1}^{n} M_i$$

fornece a variância da estimativa da média, onde:

$$\bar{y} = \frac{\sum_{i=1}^{n} M_i \bar{y}_i}{n}$$

As equações acima contêm os seguintes símbolos ou expressões:

$f_1 = \frac{n}{N}$ 4: representa a relação entre o número de conglomerados incluídos na amostra e o número total de conglomerados;

$f_2 = \frac{m_i}{M_i}$ 5: representa a proporção constante de elementos incluídos na amostra em relação ao número de elementos de cada conglomerado;

n 6 :número de conglomerados da amostra;

N 7 :número de conglomerados da população;

$\bar{M} = \frac{1}{n} \sum_{i=1}^{n} M_i$ 8 , onde:

M_i 9 :número de elementos do conglomerado i.

$\bar{m} = \frac{1}{n} \sum_{i=1}^{n} m_i$ 10 , onde:

m_i 11 :número de elementos incluídos na amostra para o conglomerado i;

$\bar{y}_i$ 12 :média da caraterística procurada no conglomerado i;

S_{2i}^2 13 :variância dentro do conglomerado i, estimada por:

onde:

p 15 :proporção da característica em questão; e, $q = 1 - p$ 16.

3 - ROTINA SAS®

A rotina, baseada em comandos MACRO (SAS INSTITUTE, 1990), tendo por finalidade calcular as variáveis citadas anteriormente, compõe-se de oito partes.

3.1 - MACRO CAL1

3.1.1 - Descrição

a) criação do SASdataset CVC02x, onde x é o nível de infestação e CVC01 é o SASdataset com as informações do levantamento;

b) Cálculo de:

i) média da característica no conglomerado i:

$$YBARi = \frac{O_i}{m_i}$$

ii) variância dentro do conglomerado i:

$$S2i = \frac{m_i}{m_i - 1} pq = \frac{O_i m_i - O_i^2}{(m_i - 1) m_i}$$

iii) estimativa do total de plantas com característica no conglomerado i:

$$MiYBARi = M_i \frac{O_i}{m_i}$$

iv) número de plantas do conglomerado i multiplicado pela variância dentro do conglomerado i

(variável utilizada para cálculo da variância dentro dos conglomerados):

$$MiS2i = M_i \frac{m_i}{m_i - 1} pq$$

3.1.2 - Sintaxe

```
%MACRO Cal1;
    DATA Cvc.Cvc02&cvc;
        SET Cvc.Cvc01;
        KEEP QUEST Mi MOBi
    OBCVC&cvc YBARi S2i
MiYBARi MiS2i;
        YBARi = OBCVC&cvc/MOBi;
        IF (OBCVC&cvc-1) > 0 then
            S2i = (OBCVC&cvc *
      MOBi - (OBCVC&cvc *
  OBCVC&cvc)) / (MOBi*
  (MOBi - 1) );
        ELSE
          S2i=0;
        MiYBARi = Mi * (OBCVC&cvc
        / MOBi);
        MiS2i = Mi * S2i;
    RUN;
%MEND Cal1;
```

3.2 - MACRO TOTAL

3.2.1 - Descrição

Cálculo de valores totais, a partir dos seguintes componentes:

a) arq: nome do SASdataset de entrada;

b) var: nome da variável que se deseja calcular o valor total; e

c) total: nome da variável que contém o total.

3.2.2 - Sintaxe

```
%MACRO Total(arq,total,var);
    DATA _NULL_;
        SET &arq;
        total + &var;
        CALL SYMPUT ("&total",total);
    RUN;
%MEND Total;
```

3.3 - MACRO CAL2

3.3.1 - Descrição

Criação e cálculo da variável intermediária (*DIFi*), para obtenção da variância entre os conglomerados, onde:

3.3.2 - Sintaxe

```
%MACRO Cal2;
    DATA Cvc.Cvc02&cvc(replace=yes);
        SET Cvc.Cvc02&cvc;
    DIFi = ( ((Mi * OBCVC&cvc) / MOBi ) -
        (&TMiYi / &enep)) ** 2;
    RUN;
%MEND Cal2;
```

3.4 - MACRO MEDIA

3.4.1 - Descrição

Cálculo da média (proporção) através da criação da variável *media*.

3.4.2 - Sintaxe

```
%MACRO Media (media,num,den);
    DATA _NULL_;
        media = &num / &den;
        CALL SYMPUT("&media",
    media);
    RUN;
%MEND Media;
```

3.5 - MACRO VARIANI

3.5.1 - Descrição

Cálculo da variância entre os conglomerados, ou variância intraclasse (variável *varian*):

$$varian = \frac{1 - \frac{n}{N}}{\frac{\left(\sum_{i=1}^{n} M_i\right)^2}{n}} \frac{\sum_{i=1}^{n} \left(M_i \bar{y}_i - \frac{\sum_{i=1}^{n} M_i \bar{y}_i}{n}\right)^2}{n - 1}$$

onde:
eneg: número de conglomerados da população (*N*);
enep: número de conglomerados da amostra no primeiro estágio (*n*);
f2: proporção de unidades incluídas na amostra para cada conglomerado $\frac{m_i}{M_i}$ 23;

$$t1 = \sum_{i=1}^{n} M_i \; 24$$

$$t2 = \sum_{1=1}^{n} m_i \; 25$$

t3: total de *MiS2i*; e
t4: total de *DIFiQ*;

3.5.2 - Sintaxe

```
%MACRO Variani (varian,enep,eneg,f2,t1,
t2,t3,t4);
        DATA _NULL_;
                varian=
(((&eneg - &enep) * &enep)/ (&eneg * (&t1**2))) *
                (&t4 / (&enep-1));
                CALL SYMPUT ("&varian",
        varian);
        RUN;
%MEND Variani;
```

3.6 - MACRO VARIANE

3.6.1 - Descrição

Cálculo da variância dentro dos conglomerados, ou variância interna (variável *varian*):

onde:
eneg: número de conglomerados da população (*N*);
enep: número de conglomerados da amostra no primeiro estágio (*n*);
f2: proporção de unidades incluídas na amostra para cada conglomerado $\frac{m_i}{M_i}$ 27;

$$t1 = \sum_{i=1}^{n} M_i \; 28$$

$$t2 = \sum_{1=1}^{n} m_i \; 29$$

t3: total de *MiS2i*; e
t4: total de *DIFiQ*;

3.6.2 - Sintaxe

```
%MACRO Variane (varian,enep,eneg,f2,t1,
t2,t3,t4);
DATA _NULL_;
  varian=  ((((&enep **2 )*(1 - &f2)) * &t3)/
(&eneg * &t2 * &t1));
  CALL SYMPUT ("&varian",varian);
RUN;
%MEND Variane;
```

3.7 - MACRO CALGERAL

3.7.1 - Descrição

Cálculo de: variância da estimativa da média, desvio padrão e coeficiente de variação (percentual), onde:

vari: variância intraclasse;
vare: variância interna;
media: média;
des: desvio padrão; e
coef: coeficiente de variação percentual.

3.7.2 - Sintaxe

```
%MACRO Calgeral (vari,vare,media,var,des,
coef);
        DATA _NULL_;
                var  = &vari + &vare;
```

```
            des  = sqrt(var);
            coef = (des / &media)*100;
            CALL SYMPUT("&var",var);
            CALL SYMPUT("&des",des);
            CALL SYMPUT("&coef",coef);
        RUN;
    %MEND Calgeral;
```

3.8 - PROCEDIMENTO CONGLO

3.8.1 - Descrição

Contém as chamadas das macros e atribuições dos valores: f_2 , N e nível de infestação.

3.8.2 - Sintaxe

```
* INCLUSAO DO ARQUIVO CONTENDO AS MACROS;
%INCLUDE 'macros.pgm';

options MPRINT;
LIBNAME cvc 'cvc';
* INDICAÇAO DO NIVEL DE CVC (NOSSO EXEMPLO =1) ;
%LET cvc=1;
/* * * * * * *
* CRIACAO DO ARQUIVO CVC02X ONDE X=NIVEL DE CVC
* E VARIAVEIS BASICAS
* * */
%CAL1;
/* * * * * * *
* CALCULO DE VALORES TOTAIS
* * */
%LET TMiYi   =0;
%LET TMiS2i  =0;
%LET TMi     =0;
%LET TMOBi   =0;
%LET ENEP    =0;
%TOTAL(cvc.cvc02&cvc,TMiYi,MiYBARi);
%PUT TMIYI = &TMIYI;
%TOTAL(cvc.cvc02&cvc,TMiS2i,MiS2i);
%PUT TMIS2I = &TMIS2I;
%TOTAL(cvc.cvc02&cvc,TMi,Mi);
%PUT TMI = &TMI;
%TOTAL(cvc.cvc02&cvc,TMOBi,MOBi);
%PUT TMOBI = &TMOBI;
%TOTAL(cvc.cvc02&cvc,ENEP,1);
%PUT ENEP = &ENEP;
/* * * * * * *
* CALCULO DA VARIAVEL INTERMEDIARIA DIFi;
* * */
%CAL2;
%LET TDIFiQ=0;
%TOTAL(cvc.cvc02&cvc,TDIFiQ,DIFi);
%PUT TDIFiQ = &TDIFiQ;
/* * * * * * *
* CALCULO DA MEDIA;
* * */
%LET media=0;
%MEDIA (media,&TMiYi,&TMi);
%PUT media = &media;
/* * * * * * *
* CALCULO DA VARIANCIA INTERNA E INTRACLASSE;
* * */
%LET variani=0;
%LET variane=0;
%LET eneg=20000;
%LET f2=1/500;
```

```
%VARIANI(variani,&enep,&eneg,&f2,
&TMi,&TMOBi,&TMiS2i,&TDIFiQ);
%VARIANE(variane,&enep,&eneg,&f2,
&TMi,&TMOBi,&TMiS2i,&TDIFiQ);
/* * * * * * *
* CALCULOS GERAIS;
* * *;
%LET varian   = 0;
%LET desvio   = 0;
%LET coef     = 0;
%CALGERAL(&variani,&variane,&media,varian,
desvio,coef);
%PUT varian = &varian;
%PUT desvio = &desvio;
%PUT coef % = &coef;
```

LITERATURA CITADA

COCHRAN, Willian G. **Técnicas de amostragem** Rio de Janeiro, Fundo de Cultura, 1965. 555p.

KISH, Leslie. **Survey sampling.** New York, J.

SAS INSTITUTE, SAS® guide to macro processing: version 6. 2.ed. Cary, NC; SAS Institute Inc., 1990. 319p.

UMA ROTINA PARA ESTIMATIVAS DE AMOSTRA EM DOIS ESTÁGIOS USANDO O SAS®

SINOPSE: O objetivo deste trabalho é apresentar uma rotina, a partir de comandos MACRO do Statistical Analysis Software (SAS®), para amostras por conglomerados. Utiliza-se como exemplo prático o esquema de cálculo da estimativa do percentual de laranjeiras afetadas por CVC (Clorose Variegada de Citros) em amostra de dois estágios, com grupos de tamanhos desiguais, sorteados com probabilidades iguais e subamostragem proporcional ao número total de plantas na propriedade citrícola.

Palavras-chave: amostragem casual por conglomerados, SAS.

A ROUTINE FOR TWO STAGE SAMPLE ESTIMATES USING SAS®

ABSTRACT: This paper deals with using SAS® (Statistical Analysis Software) MACRO commands in order to analyse cluster samples. An application is presented: the case study of a two stage sample, with unequal clusters, used to estimative the occurrence of CVC (Citrus Variegated Chlorosis).

Key-words: cluster sampling, SAS.