Planos amostrais para seleção indireta de elementos de uma população.

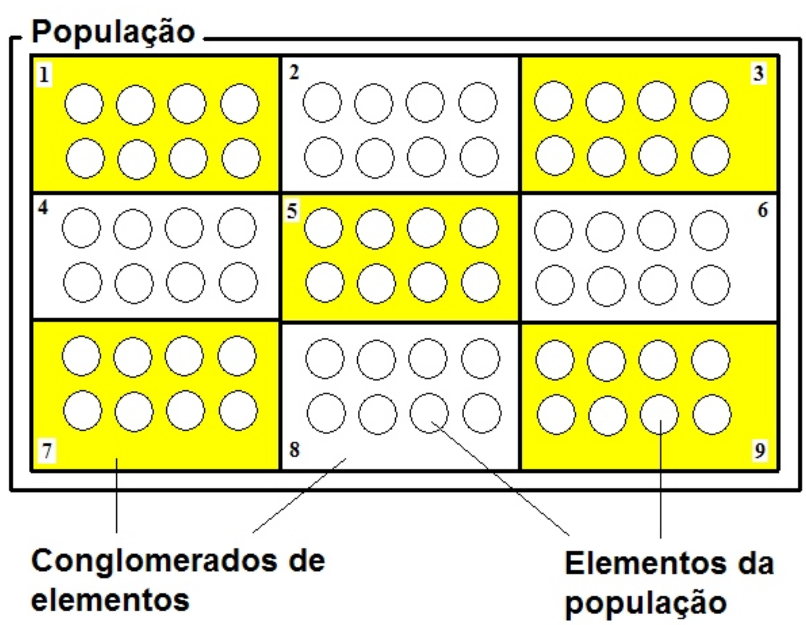
A implementação de planos amostrais para seleção indireta de elementos é motivada pela disponibilidade de um cadastro do tipo b).

Nesta parte, serão vistos os seguintes tópicos:

- Amostragem de Conglomerados - AC
- Amostragem em Dois Estágios -A2
- Amostragem em Múltiplos Estágios

Nomenclatura

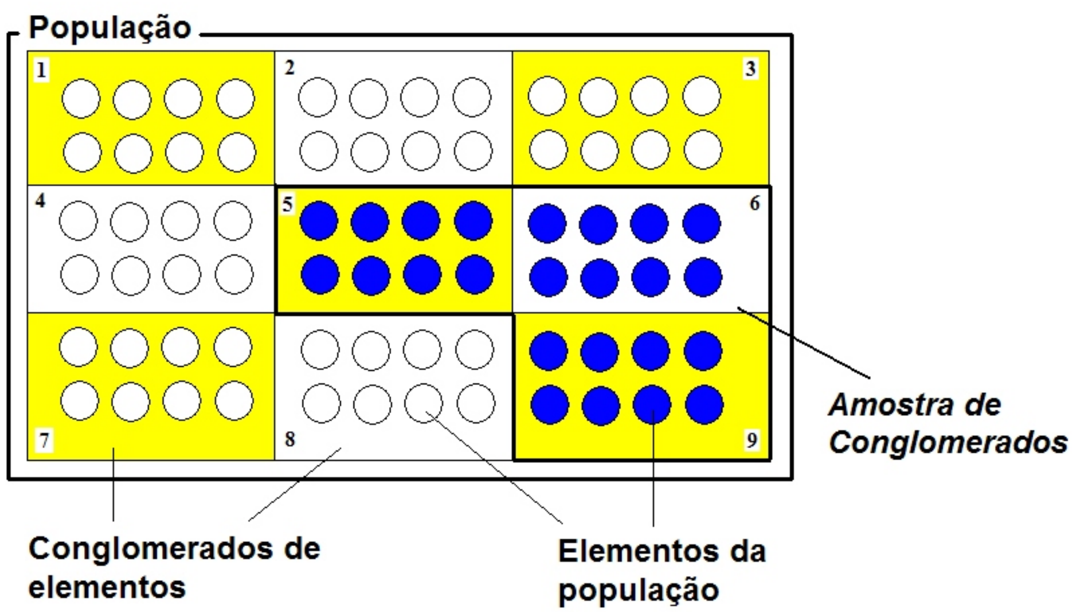
Num cadastro do tipo b), os elementos da população são listados em grupos, chamados na literatura de conglomerados.



Nessa situação, exitem vários cenários possíveis de como obter uma amostra de elementos da população sob estudo.

Cenário 1

Uma amostra de conglomerados é retirada de acordo com um plano amostral p e todos os elementos pertencentes aos conglomerados selecionados compõem a amostra.



☞ *Amostragem de Conglomerados*

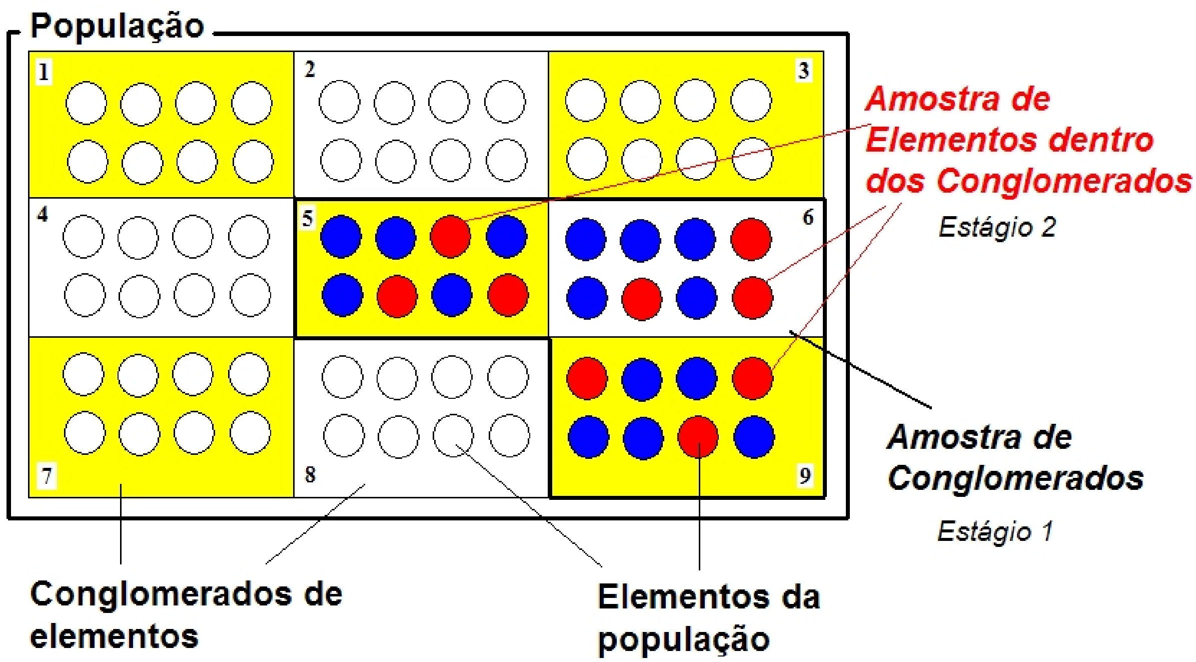
Cenário 2

Estágio 1: Uma amostra s_I de conglomerados é retirada de acordo com o plano $p_I(.)$

Unidade amostral primária: ______________

Estágio 2: Uma amostra s_i de elementos é retirada de cada conglomerado i selecionado, de acordo com o plano $p_{II}(. \mid s_I)$

Unidade amostral secundária: _______________



☞ *Amostragem em Dois Estágios*

Cenário 3

Estágio 1: Uma amostra s_I de conglomerados maiores é retirada de acordo com o plano $p_I(.)$

Unidade amostral primária: CONGLOMERADO MAIOR

Estágio 2: Uma amostra s_i de conglomerados menores é retirada de cada conglomerado maior i selecionado, de acordo com o plano $p_{II}(.\,|s_I)$

Unidade amostral secundária: CONGLOMERADO MENOR

Faça um esquema gráfico representando o Cenário 3.

☞ *Amostragem de Conglomerados em Dois Estágios*

Um exemplo de cenário mais complexo

13. Amostragem de Conglomerados - AC

(ou amostragem de conglomerados em um estágio)

Conjunto dos elementos da população:

$$U = \{1, 2, \ldots, N\}$$

Conjunto de N_I conglomerados listados no cadastro:

$$U_I = \{1, 2, \ldots, N_I\}$$

População representada em função dos conglomerados:

$$U = \bigcup_{i \in U_I} U_i \quad ,$$

onde cada U_i representa um conglomerado de N_i elementos da população.

O total de elementos da população é dado por

$$N = \sum_{i \in U_I} N_i .$$

Estratégia amostral

Uma AC é definida como segue:

i) Selecione uma amostra S_I, de n_I (ou n_{S_I}, se o tamanho da amostra for uma variável aleatória) conglomerados de U_I, de acordo com o plano amostral p_I ;

ii) Observe a variável de interesse y em todos os elementos que compõem os conglomerados selecionados para compor a amostra.

Notação adicional:

Conjunto dos elementos que compõem a amostra:

$$S = \bigcup_{i \in S_I} U_i$$

Tamanho da amostra:

$$n_S = \sum_{i \in S_I} N_i$$

Note que o tamanho de amostra observado depende dos tamanhos dos conglomerados. Em geral, n_S é uma variável aleatória. Em que situações n_S será um valor fixo ?

Probabilidades de inclusão de primeira e segunda ordem dos conglomerados

$$\pi_{Ii} = \sum_{s_I \ni i} p_I(s_I)$$

$$\pi_{Iij} = \sum_{s_I \ni i,j} p_I(s_I)$$

Probabilidades de inclusão de primeira e segunda ordem dos elementos

$$\pi_k = Prob(k \in S) = Prob(i \in S_I) = \pi_{Ii},$$

para todo $k \in U_i$

$$\pi_{kl} = Prob(k, l \in S) = \begin{cases} \pi_{Ii}, \ se \ k, l \in U_i \\ \pi_{Iij}, \ se \ k \in U_i, l \in U_j \end{cases}$$

π_{kl} Pode assumir então dois valores: π_{Ii} se ambos, k e l pertencem ao mesmo conglomerado (U_i); e π_{Iij}, se k e l pertencem conglomerados distintos (U_i e U_j).

Estimação

Considere a estimação do total populacional de uma população finita agrupada em conglomerados. Seguindo a notação desenvolvida até o momento, tem-se:

$$t = \sum_{k \in U} y_k = \sum_{i \in U_I} \left(\sum_{k \in U_i} y_k \right) = \sum_{i \in U_i} t_i$$

<u>Resultado 13.1:</u> Em amostragem de conglomerados (AC), o estimador de H-T assume a forma

$$\hat{t}_{\pi} = \sum_{i \in S_I} \check{t}_i = \sum_{i \in S_I} \frac{t_i}{\pi_{Ii}}$$

e tem como variância

$$Var_{AC}(\hat{t}_{\pi}) = \sum_{i \in U_I} \sum_{j \in U_I} \Delta_{Iij} \check{t}_i \check{t}_j .$$

Além disso,

$$\hat{Var}_{AC}(\hat{t}_{\pi}) = \sum_{i \in S_I} \sum_{j \in S_I} \check{\Delta}_{Iij} \check{t}_i \check{t}_j$$

é um estimador centrado para $Var_{AC}(\hat{t}_{\pi})$.

Prova:

13.1 Amostragem aleatória simples de conglomerados

Considere que n_I conglomerados são selecionados sob um plano AAS. Denote este estratégia amostral por AAC (amostragem aleatória simples de conglomerados). Assim,

$$p_I(S_I) = \binom{N_I}{n_I}^{-1},$$

$$\pi_{Ii} = \frac{n_I}{N_I} \qquad , \qquad \pi_{Iij} = \frac{n_I(n_I - 1)}{N_I(N_I - 1)},$$

e a prova do seguinte resultado segue dos resultados obtidos para um plano AAS.

<u>Resultado 13.2:</u> Em amostragem aleatória simples de conglomerados (AAC), o estimador de H-T assume a forma

$$\hat{t}_\pi = N_I \bar{t}_{S_I},$$

onde $\bar{t}_{S_I} = \frac{1}{n_I} \sum_{i \in S_I} t_i$.

A sua variância é dada por

$$Var_{AAC}(\hat{t}_\pi) = N_I^2 \left(1 - \frac{n_I}{N_I}\right) \frac{S_{tU_I}^2}{n_I},$$

onde $S_{tU_I}^2 = \frac{1}{N_I - 1} \sum_{i \in U_I} (t_i - \bar{t}_{U_I})^2$ e

$$\bar{t}_{U_I} = \frac{1}{N_I} \sum_{i \in U_I} t_i .$$

Além disso,

$$\hat{Var}_{AAC}(\hat{t}_\pi) = N_I^2 (1 - \frac{n_I}{N_I}) \frac{S_{tS_I}^2}{n_I},$$

(onde $S_{tS_I}^2 = \frac{1}{N_I - 1} \sum_{i \in S_I} (t_i - \bar{t}_{S_I})^2$)

é um estimador centrado para $Var_{AAC}(\hat{t}_\pi)$.

Relação entre amostragem sistemática (AS) e AAC

Relembre o plano de amostragem sistemática. Naquele contexto, a população era estruturada da seguinte forma (para o caso em que $c=0$):

S_1 (U_1)	$\cdots$	S_r (U_i)	$\cdots$	S_a (U_{N_I})
y_1	$\cdots$	y_r	$\cdots$	y_a
y_{1+a}	$\cdots$	y_{r+a}	$\cdots$	y_{2a}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$y_{1+(n-1)a}$	$\cdots$	$y_{r+(n-1)a}$	$\cdots$	y_N
t_{S_1} (t_1)	$\cdots$	t_{S_r} (t_i)	$\cdots$	t_{S_a} (t_{N_I})

Note que essa mesma estrutura pode ser aproveitada para o plano AAC, como sugere a notação entre parêntesis.

Deste modo, valem as seguintes observações:

i) Uma AS equivale a uma AAC com tamanho de amostra $n_I = 1$

ii) Se uma AS é empregada com mais de um r, as fórmulas de estimação de variância sob uma AAC fornecem um meio de estimar, sem tendência, a variância do estimador de H-T para o total populacional, sob uma AS.

Suponha que m valores de r sejam utilizados. Assim, $n_I = m$ e $N_I = m\,a$.

iii) O mesmo coeficiente de homogeneidade utilizado para analisar a eficiência de um plano AS será útil para analisar a eficiência de uma amostragem de conglomerados

Note que, no contexto de AS, foi usada a expressão

$$\delta = 1 - \frac{N-1}{N-a}\frac{SQD}{SQT},$$

que equivale a $\delta = 1 - \dfrac{QMD}{S_{yU}^2}$.

13.2 Análise da eficiência de um plano AAC

Tabela de ANOVA

F.V.	G.L.	S.Q.	Q.M.
Entre conglomerados	N_I-1	SQE	QME
Dentro de conglomerados	N-N_I	SQD	QMD
Total	N-1	$SQT = (N-1)S_{yU}^2$	

$$SQE = \sum_{i=1}^{N_I} N_i(\bar{y}_{U_i} - \bar{y}_U)^2 \ , \qquad QME = \frac{SQE}{N_I - 1}$$

$$SQD = \sum_{i=1}^{N_i} (N_i - 1) S_{yU_i}^2 \ , \qquad QME = \frac{SQD}{N - N_I}$$

Defina, então, o coeficiente de homogeneidade neste contexto, como

$$\delta = 1 - \frac{QMD}{S_{yU}^2}.$$

Note que δ satisfaz os seguintes limites:

$$\delta_{\min} = -\frac{N_I - 1}{N - N_I}, \text{ quando } SQE = 0$$

$$\delta_{\max} = 1, \text{ quando } SQD = 0.$$

Desta forma, valores pequenos para δ (SQD grande) implicam heterogeneidade entre elementos dentro dos conglomerados.

Por outro lado, valores grandes para δ (SQD pequeno) implicam homogeneidade entre elementos dentro dos conlgomerados.

Comparando uma AAC com uma AAS de tamanho

$$n = E_{AAC}(n_S) = \sum_{i \in U_I} N_i \pi_{Ii}$$

$$= \sum_{i \in U_I} N_i \frac{n_I}{N_I} = N \frac{n_I}{N_I}$$

é possível mostrar que

$$deff(AAC,\hat{t}_\pi) = \frac{Var_{AAC}(\hat{t}_\pi)}{Var_{AAS}(\hat{t}_\pi)}$$

$$= 1 + \frac{N - N_I}{N_I - 1}\delta + \frac{N_I Cov}{N S_{yU}^2}$$

onde $Cov = \frac{1}{N_I - 1} \sum_{i \in U_i} (N_i - \frac{N}{N_I}) N_i \bar{y}_{U_i}^2.$

Em particular, quando os conglomerados forem de mesmo tamanho $N_i = N/N_I$, tem-se

$$deff(AAC,\hat{t}_\pi) = 1 + \frac{N - N_I}{N_I - 1}\delta$$

Neste caso,

$$Var_{AAC} < Var_{AAS}$$

se e somente se $\delta < 0$

(Alta heterogeneidade dentro dos conglomerados)

Compare esse resultado com o de um plano AAE sob alocação proporcional ao tamanho do estrato.

Comentário:

Quando os conlgomerados são de mesmo tamanho, M, é possível fazer a análise de eficiência de uma AAC via coeficiente de correlação intra-classe:

$$\rho_{IC} = 1 - \frac{M}{M-1}\frac{SQD}{SQT}$$