Universidade de Évora
Curso de Mestrado em Matemática Aplicada 1995/97

TÉCNICAS DE AMOSTRAGEM E SUB-AMOSTRAGEM

Dissertação realizada por:
Carla Maria Lopes da Silva Afonso dos Santos

ÉVORA
1997

Universidade de Évora
Curso de Mestrado em Matemática Aplicada 1995/97

TÉCNICAS DE AMOSTRAGEM E SUB-AMOSTRAGEM

87262

Dissertação realizada por:
Carla Maria Lopes da Silva Afonso dos Santos

ÉVORA
1997

Autora: Carla Maria Lopes da Silva Afonso dos Santos

Orientador: Professor Doutor João Tiago Mexia

Co-orientador: Professor Doutor Carlos Braumann

ÍNDICE

1. INTRODUÇÃO

Apesar de o principal objectivo desta dissertação ser a apresentação de resultados novos no domínio da sub-amostragem, pretende-se igualmente apresentar, de uma forma consistente, as principais técnicas de amostragem.

Começa-se pelo estudo da amostragem simples, com e sem reposição, analisando, separadamente, as modalidades em que a probabilidade de escolha, dos elementos da população, é igual, e é variável.

Em seguida considera-se a amostragem estratificada e a amostragem para agregados, assim como um caso especial desta última, a amostragem sistemática.

Finalmente, dado que a teoria, referente às técnicas de sub-amostragem, estava apenas desenvolvida para, quando muito, três níveis de fraccionamento, generalizámos a mesma para qualquer número (k) de níveis, abordando os casos em que as probabilidades de escolha são iguais e proporcionais. Para tal, utilizou-se uma notação que permite referenciar, sem ambiguidades, o nível de estrutura a que pertence determinada sub-população e qual a posição que ocupa, assim como, considerando as decomposições sucessivas da população inicial, saber, exactamente, quais as sub-populações que deram origem a essa sub-população.

2. AMOSTRAGEM ALEATÓRIA SIMPLES

A amostragem aleatória simples, apesar de na prática ser pouco usada, é um método de amostragem de grande importância pois funciona como ponto de partida para o estudo de outros métodos.

Neste método a população é considerada como um todo e as unidades estatísticas que vão pertencer à amostra são seleccionadas ao acaso de entre todos os elementos da população. Cada unidade estatística tem uma probabilidade conhecida e não nula de figurar na amostra, podendo esta probabilidade ser igual para todos os elementos ou variar entre eles.

A amostragem aleatória simples, com probabilidades iguais, pressupõe um total desconhecimento dos caracteres a estudar na população e como consequência todas as unidades estatísticas têm a mesma probabilidade, de serem escolhidas para a amostra.

As unidades estatísticas podem ser retiradas, da população, com ou sem reposição.

A selecção é considerada com reposição se retiramos um elemento da população ,e após a sua observação o devolvemos à população. Assim, a população não é alterada, de extracção para extracção, e portanto, cada unidade estatística pode aparecer mais do que uma vez na amostra.

As extracções são independentes, visto que, a extracção de uma determinada unidade estatística não depende da extracções anteriores.

A selecção considera-se sem reposição se retiramos um elemento da população , observamo-lo e não o devolvemos à população. A população é alterada de extracção para extracção , ou seja , a extracção da segunda unidade estatística é feita duma população que difere da população inicial porque já não contem a unidade estatística retirada em primeiro lugar. Neste caso, cada elemento só pode aparecer uma vez na amostra. A extracção de uma determinada unidade estatística depende das extracções anteriores , pelo que, as extracções são acontecimentos dependentes.

2.1. AMOSTRAGEM COM PROBABILIDADES IGUAIS, SEM REPOSIÇÃO, PARA CARACTERÍSTICAS NUMÉRICAS

Este é um método de selecção de n elementos de entre N existentes na população, no qual em cada tiragem qualquer elemento da população tem igual probabilidade de ser escolhido. Para além disso cada elemento que é extraído é removido da população para as tiragens subsequentes.

Os n elementos escolhidos constituem a amostra.

Sejam y_i , i=1,2,...,N, os valores da característica numérica , para a qual se está a amostrar , os quais indicam o número de ordem dos diferentes elementos da população e Y_j , j=1,2,...,n , as variáveis aleatórias que dão os resultados obtidos nas várias tiragens. Estas variáveis tomam como valores , os valores da característica numérica nos elementos que vão sendo seleccionados.

Ao considerar que as probabilidades de selecção são iguais para todos os elementos da população temos

(2.1) $$P(Y_1 = y_i) = \frac{1}{N}$$

e para $j > 1$ temos

(2.2) $$\begin{cases} P\left[Y_j = y_i \mid \bigcap_{j'=1}^{j-1} (Y_{j'} \neq y_i)\right] = \dfrac{1}{N-j+1} \\ P\left[Y_j = y_i \mid \bigcup_{j'=1}^{j-1} (Y_{j'} = y_i)\right] = 0 \end{cases}.$$

Para provar que a probabilidade de selecção é igual em todas as extracções , ou seja, que o j-ésimo elemento (y_j) tem a mesma probabilidade de ser seleccionado em qualquer extracção, consideremos, em primeiro lugar, a probabilidade deste elemento ser seleccionado na primeira extracção, que é:

(2.3) $$P_j^1 = P\left(Y_1 = y_j\right) = \frac{1}{N} \quad j = 1,\ldots,N$$

A probabilidade de que y_j seja seleccionado na segunda extracção é:

(2.4) $$\begin{aligned} P_j^2 &= P\left(Y_2 = y_j\right) = \sum_{k \neq j} P\left(Y_1 = y_k \,,\, Y_2 = y_j\right) = \\ &= \sum_{k \neq j} P\left(Y_2 = y_j \mid Y_1 = y_k\right) \cdot P\left(Y_1 = y_k\right) = \\ &= \sum_{k \neq j} \frac{1}{N-1} \cdot \frac{1}{N} = (N-1)\frac{1}{N-1} \cdot \frac{1}{N} = \frac{1}{N} \quad j = 1,\ldots,N \end{aligned}$$

A probabilidade de que y_j seja seleccionado na terceira extracção é:

(2.5) $$\begin{aligned} P_j^3 &= P\left(Y_3 = y_j\right) = \sum_{j \neq k \neq m} P\left(Y_1 = y_k \,,\, Y_2 = y_m, Y_3 = y_j\right) = \\ &= \sum_{j \neq k \neq m} P\left[Y_3 = y_j \mid (Y_1 = y_k, Y_2 = y_m)\right] \cdot P\left(Y_2 = y_m \mid Y_1 = y_k\right) \cdot P\left(Y_1 = y_k\right) = \\ &= \sum_{j \neq k \neq m} P\left[Y_3 = y_j \mid (Y_1 = y_k, Y_2 = y_m)\right] \cdot \sum_{k \neq m} \frac{1}{N} \cdot \frac{1}{N-1} = \\ &= \sum_{j \neq k \neq m} \frac{1}{N-2} \cdot (N-1) \cdot \frac{1}{N-1} \cdot \frac{1}{N} = \\ &= (N-2) \cdot \frac{1}{N-2} \cdot (N-1) \cdot \frac{1}{N-1} \cdot \frac{1}{N} = \frac{1}{N} \quad j = 1,\ldots,N \end{aligned}$$

Por recorrência, conclui-se que,

(2.6) $$P_j^k = P\left(Y_k = y_j\right) = \frac{1}{N} \quad ; \quad j = 1,\ldots,N; \quad 1 \le k \le n; \quad n \le N$$

é a probabilidade de y_j ser seleccionado na k-ésima extracção.

Pode ainda provar-se que, no método de amostragem aleatória simples com probabilidades iguais e sem reposição, a probabilidade de que um qualquer elemento pertença a uma amostra, de dimensão n, é $\frac{n}{N}$.

Para que o elemento y_j pertença a uma amostra, é necessário que tenha sido escolhido uma e uma só vez, visto tratar-se de extracções sem reposição.

Seja

(2.7) $$P_j = P(y_j \in \text{amostra}) = 1 - P(y_j \notin \text{amostra})$$

Mas a probabilidade de y_j não pertencer à amostra é igual à probabilidade dos N-1 elementos y_k , com $k \neq j$,poderem pertencer à amostra.

Assim

(2.8) $$P(y_j \notin \text{amostra}) = \frac{{}^{N-1}C_n}{{}^{N}C_n}$$

e portanto

(2.9) $$P(y_j \in \text{amostra}) = 1 - \frac{{}^{N-1}C_n}{{}^{N}C_n} = \frac{{}^{N}C_n - {}^{N-1}C_n}{{}^{N}C_n} = \frac{{}^{N-1}C_{n-1}}{{}^{N}C_n} = \frac{n}{N},$$

pois ${}^{N-1}C_n + {}^{N-1}C_{n-1} = {}^{N}C_n$.

Para determinar quantas amostras diferentes, de n elementos , é possível obter com os N elementos da população há que ter em consideração se a ordem pela qual os elementos são seleccionados tem ou não interesse para a constituição da amostra, e também que, neste caso, a amostragem é feita sem reposição.

Se ignorarmos a ordem pela qual os elementos vão sendo obtidos, isto é, se considerarmos iguais as amostras que contêm os mesmos elementos, obtidos por diferentes ordens, o número de amostras possíveis é:

(2.10) $$ {}^{N}C_n = \frac{N!}{(N-n)!n!} .$$

Se tiver interesse a ordem pela qual se vão obtendo os elementos que irão compor a amostra, isto é, se considerarmos distintas as amostras compostas pelos mesmos elementos, mas obtidos por ordens diferentes, o número de amostras possíveis será :

(2.11) $$ {}^{N}A_n = \frac{N!}{(N-n)!} .$$

É fácil ver que todas as amostras distintas têm a mesma probabilidade de serem seleccionadas.

Consideremos um conjunto de n elementos específicos $\{y_1,\ldots,y_n\}$, que constituem uma determinada amostra. Os valores dessa amostra correspondem, tendo em consideração que a amostragem é sem reposição, a elementos distintos da população.

A probabilidade de esta amostra ser colhida é:

(2.12)

$$P\left[\bigcap_{j=1}^{n}(Y_j=y_j)\right]=P\left[Y_n=y_n|\bigcap_{j=1}^{n-1}(Y_j=y_j)\right]\cdot P\left[\bigcap_{j=1}^{n-1}(Y_j=y_j)\right]=\ \ldots\ =P\left[Y_n=y_n|\bigcap_{j=1}^{n-1}(Y_j=y_j)\right]\cdot$$

$$\cdot P\left[Y_{n-1}=y_{n-1}|\bigcap_{j=1}^{n-2}(Y_j=y_j)\right]\cdot P\left[Y_2=y_2|Y_1=y_1\right]\cdot P(Y_1=y_1).$$

Como todos os elementos têm igual probabilidade de serem escolhidos, a probabilidade de ser obtida uma amostra específica , de n elementos com uma determinada ordem, é:

(2.13) $$P\left[\bigcap_{j=1}^{n}(Y_j=y_j)\right]=\frac{1}{N-n+1}\cdot\frac{1}{N-n+2}\cdot\ \ldots\ \cdot\frac{1}{N-1}\cdot\frac{1}{N}=\frac{(N-n)!}{N!}$$

Se, pelo contrário, a ordem pela qual são obtidos os elementos que irão compor a amostra for ignorada, a probabilidade de se obter essa amostra passa a ser:

(2.14) $$P\left[\bigcap_{j=1}^{n}\left(Y_j=y_j\right)\right]=\frac{n!(N-n)!}{N!}=\frac{1}{{}^{N}C_n}$$

onde n! indica por quantas ordens diferentes podem ser obtidos os n elementos.

Consideremos agora os acontecimentos, que se verificam quando o i-ésimo elemento é escolhido na j-ésima extracção,

(2.15) $$A_{j,i}=(Y_j=y_i)\ ;\ j=1,\ldots,n\ ,\ i=1,\ldots,N\ ,$$

que como já vimos têm todos probabilidade $\frac{1}{N}$.

Assim as variáveis aleatórias Y_j , j = 1,..., n , são identicamente distribuídas , correspondendo-lhe o esquema

(2.16) $$Y_j\begin{cases}y_1,\ldots,y_N\\ \frac{1}{N},\ldots,\frac{1}{N}\end{cases};\ j=1,\ldots,n\ .$$

Considerando

(2.17) $$T=\sum_{i=1}^{N}y_i$$

e

(2.18) $$S=\sum_{i=1}^{N} y_i^2 \quad ,$$

obtém-se o valor médio da característica (o parâmetro que se pretende estimar)

(2.19) $$\mu=\mu\left(y_j\right)=\frac{T}{N} \quad , j=1,\ldots,n$$

e a variância da característica

(2.20) $$\sigma^2=\sigma^2\left(Y_j\right)=\mu\left(Y_j^{\,2}\right)-\mu^2\left(Y_j\right)=\frac{1}{N}\left(S-\frac{T^2}{N}\right) \quad ; j=1,\ldots,n,$$

com

(2.21) $$\mu\left(Y_j^{\,2}\right)=\frac{S}{N} \quad ; j=1,\ldots,n \ .$$

Para calcular $P\left[\left(Y_j=y_i\right)\cap\left(\left(Y_{j'}=y_{i'}\right)\right)\right]$, $i\neq i'$, consideremos as amostras que é possível obter fixados dois elementos. Para construir essas amostras, há que obter amostras de dimensão n-2 da população original, uma vez excluídos os dois elementos dados.

Se considerarmos, com importância, a ordem pela qual vão sendo obtidos os elementos da amostra, existem

(2.22) $$\frac{(N-2)!}{\left[(N-2)!-(n-2)!\right]}=\frac{(N-2)!}{(N-n)!}$$

amostras de dimensão n-2.

Se ignorarmos essa ordem, existirão

(2.23) $$\frac{(N-2)!}{(n-2)!\left[(N-2)!-(n-2)!\right]}=\frac{(N-2)!}{(n-2)!(N-n)!} \ .$$

Obtemos então, para o primeiro caso:

(2.24) $$P\left[\left(Y_j=y_i\right)\cap\left(Y_{j'}=y_{i'}\right)\right]=\frac{(N-2)!}{(N-n)!}\frac{(N-n)!}{N!}=\frac{1}{N(N-1)}$$

e para o segundo:

(2.25) $$P\left[\left(Y_j=y_i\right)\cap\left(Y_{j'}=y_{i'}\right)\right]=\frac{(N-2)!}{(n-2)!(N-n)!}\frac{n!(N-n)!}{N!}=\frac{n(n-1)}{N(N-1)}$$

De modo a não haver uma duplicação desnecessária de cálculos, devido à grande semelhança dos dois casos, de agora em diante apenas se efectuará o estudo do caso em que a ordem de selecção dos elementos é tida em consideração.

Obtemos então:

$$\mu(Y_j;Y_{j'}) = \frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{i'=1}^{N} Y_i Y_{i'} = \frac{1}{N(N-1)}\sum_{i=1}^{N} Y_i(T-Y_i) = \tag{2.26}$$
$$= \frac{1}{N(N-1)}\left(T\sum_{i=1}^{N} Y_i - S\right) = \frac{T^2-S}{N(N-1)}$$

já que, devido à amostragem ser sem reposição, $P\left[\left(Y_j = y_i\right)\cap\left(Y_j = y_{i'}\right)\right] = 0$, e

$$\sigma(Y_j;Y_{j'}) = \mu(Y_j;Y_{j'}) - \mu(Y_j)\mu(Y_{j'}) = \frac{T^2-S}{N(N-1)} - \frac{T^2}{N^2} = \tag{2.27}$$
$$= -\frac{1}{N(N-1)}\left(S - \frac{T^2}{N}\right) = -\frac{\sigma^2}{N-1}$$

Assim dado o vector aleatório

$$Y' = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} \tag{2.28}$$

tem-se

$$\mu'(Y') = \begin{bmatrix} \mu \\ \vdots \\ \mu \end{bmatrix} \tag{2.29}$$

bem como

$$\sigma^2(Y') = \sigma^2 \begin{bmatrix} 1 & \cdots & -\frac{1}{N-1} \\ \vdots & & \vdots \\ -\frac{1}{N-1} & \cdots & 1 \end{bmatrix} \tag{2.30}$$

Pretendemos agora obter o estimador linear , centrado com variância mínima, de μ.
Dado

$$\mu^* = k'^{T} Y' = \sum_{j=1}^{n} k_j Y_j \tag{2.31}$$

tem-se

(2.32) $$\mu(\mu^*) = k'^{T}\mu'(Y') = \sum_{j=1}^{n} k_j \mu = \mu \sum_{j=1}^{n} k_j .$$

Para que μ^* seja centrado é necessário que

(2.33) $$\sum_{j=1}^{n} k_j = 1 .$$

Por outro lado

(2.34) $$\sigma^2(\mu^*) = k'^{T}\sigma^2(Y')k' =$$

$$= \begin{bmatrix} k_1 & \cdots & k_n \end{bmatrix} \left(\sigma^2 \begin{bmatrix} 1 & \cdots & -\frac{1}{N-1} \\ \vdots & & \vdots \\ -\frac{1}{N-1} & \cdots & 1 \end{bmatrix} \right) \begin{bmatrix} k_1 \\ \vdots \\ k_n \end{bmatrix} =$$

$$= \sigma^2 \left(\sum_{j=1}^{n} k_j^2 - \frac{1}{N-1} \sum_{j=1}^{n} \sum_{\substack{j'=1 \\ j \neq j'}}^{n} k_j k_{j'} \right) =$$

$$= \sigma^2 \left(\frac{N}{N-1} \sum_{j=1}^{n} k_j^2 - \frac{1}{N-1} \sum_{j=1}^{n} \sum_{j'=1}^{n} k_j k_{j'} \right) =$$

$$= \sigma^2 \left(\frac{N}{N-1} \sum_{j=1}^{n} k_j^2 - \frac{1}{N-1} \left(\sum_{j=1}^{n} k_j \right) \left(\sum_{j'=1}^{n} k_{j'} \right) \right).$$

Para estimadores centrados que satisfazem a condição (2.33), a expressão (2.34) reduz-se a

(2.35) $$\sigma^2(\mu^*) = \frac{\sigma^2}{N-1} \left(N \sum_{j=1}^{n} k_j^2 - 1 \right)$$

O problema limita-se então a minimizar $\sum_{j=1}^{n} k_j^2$, sob a condição (2.33).

Utilizando multiplicadores de Lagrange, com uma função auxiliar do tipo

(2.36) $$g(k_1, \cdots, k_n, \lambda) = \sum_{j=1}^{n} k_j^2 - \lambda \left(\sum_{j=1}^{n} k_j - 1 \right)$$

procuramos a solução do sistema

(2.37) $$\begin{cases} \dfrac{\partial g}{\partial k_j}(k_1, \cdots, k_n, \lambda) = 2k_j - \lambda = 0 & ; j = 1, \cdots, n \\ \dfrac{\partial g}{\partial \lambda}(k_1, \cdots, k_n, \lambda) = \sum_{j=1}^{n} k_j - 1 = 0 \end{cases}$$

que é

(2.38) $$k_j = \frac{1}{n} \quad ; j = 1, \cdots, n .$$

Substituindo (2.38) na expressão (2.31), obtemos então

(2.39) $$\mu^* = \sum_{j=1}^{n} k_j Y_j = \frac{\sum_{j=1}^{n} Y_j}{n} = \overline{Y}_n$$

pelo que se conclui que, o estimador centrado ,com variância mínima, é a média amostral.

Substituindo os k_j , j=1,...,n , por $\frac{1}{n}$, na expressão (2.35), vem

(2.40) $$\sigma^2(\overline{Y}_n) = \frac{\sigma^2}{N-1}\left(N\sum_{j=1}^{n}\frac{1}{n^2} - 1\right) = \frac{\sigma^2}{N-1}\left(\frac{N}{n} - 1\right) = \frac{\sigma^2}{N-1}\frac{N-n}{n} = \frac{\sigma^2}{n}\frac{N-n}{N-1}.$$

2.2. AMOSTRAGEM COM PROBABILIDADES IGUAIS,SEM REPOSIÇÃO, PARA ATRIBUTOS

Admitamos agora que, a amostragem é feita com o objectivo de estimar a probabilidade, p, de um elemento, extraído ao acaso da população, possuir determinado atributo. A percentagem dos elementos da população com esse atributo será 100p .

De modo a podermos aplicar os resultados obtidos para características numéricas, consideramos uma característica numérica X que toma os valores 1 para os elementos da população que possuem o atributo e 0 para os elementos que não possuem. Isto é, a variável aleatória $X_i = 1$ se o elemento y_i possui o atributo e $X_i = 0$ se o elemento y_i não possui o atributo.

Seja M o número de elementos da população que possuem o atributo , então temos

(2.41) $$p = \frac{M}{N}$$

e

(2.42) $$T = S = M$$

logo

(2.43) $$\mu = \frac{M}{N} = p$$

e

(2.44) $$\sigma^2 = \frac{1}{N}\left(M - \frac{M^2}{N}\right) = p(1-p).$$

Numa amostra de dimensão n , onde m elementos possuem o atributo tem-se

(2.45) $$p^* = \overline{Y}_n = \frac{m}{n}.$$

Como já vimos anteriormente , no caso das características numéricas

(2.46) $$\mu(\overline{Y}_n) = \mu$$

e

(2.47) $$\sigma^2(\overline{Y}_n) = \frac{N-n}{N-1}\frac{\sigma^2}{n}$$

logo

(2.48) $$\mu(p^*) = p$$

e

(2.49) $$\sigma^2(p^*) = \frac{N-n}{N-1}\frac{p(1-p)}{n},$$

portanto p* é estimador centrado de p.

2.3. AMOSTRAGEM COM PROBABILIDADES IGUAIS, COM REPOSIÇÃO, PARA CARACTERÍSTICAS NUMÉRICAS

Neste caso, o elemento que é escolhido, em cada extracção, é devolvido à população ,pelo que esta mantém-se igual em todas as extracções. Todos os elementos têm igual probabilidade de escolha, em todas as extracções , quer tenha ou não sido escolhido anteriormente.

A probabilidade de um qualquer elemento y_i ser seleccionado na j-ésima extracção é então:

(2.50) $$P_i^j = P\left(Y_j = y_i\right) = \frac{1}{N} \quad ; j = 1,\cdots,n \quad ; i = 1,\cdots,N.$$

As variáveis aleatórias Y_j , j=1,...,n são independentes e identicamente distribuídas , correspondendo-lhes o esquema

(2.51) $$Y_j \begin{cases} y_1, \cdots, y_N \\ \frac{1}{N}, \cdots, \frac{1}{N} \end{cases} ; j = 1, \cdots, n$$

Podemos provar agora que, no método de amostragem aleatória simples com probabilidades iguais e com reposição, a probabilidade P_i de que o elemento y_i , i=1,...,N , pertença a uma qualquer amostra de dimensão n, é $1-\left(1-\frac{1}{N}\right)^n$.

A probabilidade de y_i pertencer à amostra é a probabilidade de y_i ter sido escolhido, pelo menos uma vez, para pertencer à amostra.

Atendendo a (2.50) a probabilidade de y_i não ser escolhido na j-ésima extracção é:

(2.52) $$P\left(Y_j \neq y_i\right) = 1 - \frac{1}{N}.$$

Tendo em conta que a dimensão da amostra é n , se y_i não foi escolhido para a amostra não pode ter sido escolhido em nenhuma das n extracções. Como as extracções são independentes, a probabilidade de y_i não ser escolhido em nenhuma extracção é :

(2.53) $$1 - P_i = \prod_{i=1}^{n}\left(1-\frac{1}{N}\right) = \left(1-\frac{1}{N}\right)^n$$

A probabilidade de y_i ser seleccionado pelo menos uma vez é então:

(2.54) $$P_i = 1-\left(1-P_i\right) = 1-\left(1-\frac{1}{N}\right)^n \qquad \text{c. q. d.}$$

Como se pode ver facilmente, também neste caso se verificam as expressões (2.19) e (2.20). Visto o estimador linear centrado , com variância mínima , do valor médio comum a n variáveis aleatórias independentes e identicamente distribuídas ser a respectiva média aritmética, tomamos $\mu^* = \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$, provamos que ele é estimador centrado de μ e que a sua variância é $\frac{\sigma^2}{n}$.

Para tal, calculemos μ (μ^*). Aplicando as propriedades da esperança matemática, obtemos

(2.55) $$\mu\left(\mu^*\right) = \mu\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mu\left(Y_i\right).$$

Para calcular $\mu(Y_i)$, usamos a definição que diz que, o valor esperado de uma variável, V, é a soma dos produtos obtidos, multiplicando cada possível valor de V pela sua probabilidade, somados sobre todos os possíveis valores de V, e também (2.50). Assim temos:

(2.56) $$\mu\left(Y_i\right) = \sum_{j=1}^{N} y_j \cdot P_j^i = \frac{1}{N}\sum_{j=1}^{N} y_j = \mu$$

então

(2.57) $$\mu\left(\mu^*\right) = \frac{1}{n}\cdot n \cdot \mu = \mu ,$$

o que prova que, μ^* é estimador centrado de μ .

Vamos então calcular a variância de μ^*. Por definição de μ^* e aplicando as propriedades da variância temos

(2.58) $$\sigma^2(\mu^*) = \sigma^2\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n^2}\sigma^2\left(\sum_{i=1}^{n} Y_i\right)$$

como as extracções são independentes, temos

(2.59) $$\sigma^2(\mu^*) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2(Y_i),$$

mas por definição

(2.60) $$\sigma^2(Y_i) = \mu\left[\left(Y_i - \mu(Y_i)\right)^2\right]$$

e como já vimos em (2.56), $\mu(Y_i) = \mu$, logo:

(2.61) $$\sigma^2(Y_i) = \sum_{j=1}^{N}\left(y_j - \mu\right)^2 P_j^i = \frac{1}{N}\sum_{j=1}^{N}\left(y_j - \mu\right)^2 = \sigma^2$$

Substituindo (2.61) em (2.59) temos:

(2.62) $$\sigma^2(\mu^*) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{\sigma^2}{n}.$$

Comparando as variâncias $\sigma^2(\mu^*)$ obtidas para a amostragem com e sem reposição , vimos que ,utilizando amostras da mesma dimensão , a amostragem aleatória simples com probabilidades iguais e sem reposição conduz a estimadores com menor variância, pois $\frac{N-n}{N-1} < 1$.

2.4. AMOSTRAGEM COM PROBABILIDADES IGUAIS, COM REPOSIÇÃO, PARA ATRIBUTOS

Para ser possível aplicar os resultados obtidos para as características numéricas consideramos novamente uma característica numérica que toma os valores 1 ou 0.

Apesar de a amostragem ser agora com reposição, as expressões (2.19) , (2.20) e (2.45), relativas à amostragem para atributos sem reposição, continuam a ser válidas, e portanto p^* é, também neste caso, um estimador centrado de p , ou seja , $\mu(p^*) = \mu = p$.

Como já vimos, em (2.62) e (2.44), $\sigma^2(\overline{Y}) = \frac{\sigma^2}{n}$ e $\sigma^2 = p(1-p)$, pelo que a variância de p^* é agora

(2.63) $$\sigma^2(p^*) = \frac{p(1-p)}{n}.$$

2.5. PRÉ - AMOSTRAGEM

Como já vimos , a amostragem com reposição , apesar de ser de mais fácil aplicação , conduz a estimadores com maior variância .

Quando se pretende obter resultados com uma dada precisão , realiza-se uma pré - amostragem com reposição para determinar a dimensão a dar à amostra definitiva.

Sejam $Y_1,\ldots,Y_n$ as observações da pré - amostra , pode admitir-se que

(2.64) $$D = \sum_{j=1}^{n} Y_i^2 - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n}$$

se distribui , pelo menos aproximadamente , como um chi-quadrado com n-1 graus de liberdade e que

(2.65) $$t = \frac{\overline{Y} - \mu}{\sqrt{\frac{D}{n(n-1)}}}$$

tem densidade *t* de Student, com n-1 graus de liberdade.

Assim os intervalos de confiança (aproximados) para μ são

(2.66) $$\overline{Y} - t_{n-1,q}\sqrt{\frac{D}{n(n-1)}} \le \mu \le \overline{Y} + t_{n-1,q}\sqrt{\frac{D}{n(n-1)}}$$

com a semi - amplitude

(2.67) $$A = t_{n-1,q}\sqrt{\frac{D}{n(n-1)}}.$$

Suponhamos que se pretende uma dada semi - amplitude $\overline{A}$. Terá então que se resolver a equação em ordem a n, de modo a determinar a dimensão de amostra conveniente para obter a precisão pré - definida.

Seja

(2.68) $$\overline{A} = t_{n-1,q}\sqrt{\frac{D}{n(n-1)}}$$

fazendo

(2.69) $$k = \left(\frac{\overline{A}}{t_{n-1},q}\right)^2$$

ou seja

(2.70) $$k = \frac{D}{n(n-1)}$$

obtém-se

(2.71) $$kn^2 - kn - D = 0$$

vindo

(2.72) $$n = \frac{k + \sqrt{k^2 + 4kD}}{2k}$$

(a outra solução não serve por ser negativa).

Na prática o valor de n obtido por esta fórmula é arredondado para cima. Quando n>120 a densidade t coincide com a normal reduzida, podendo-se utilizar as tabelas desta última. Quando n>30 pode-se, em geral, realizar testes t para hipóteses definidas a partir de μ.

No caso de a amostragem ser por atributos, quando $30 \leq n \leq 120$ pode admitir-se que

(2.73) $$t = \frac{p^* - p}{\sqrt{\frac{p^*(1-p^*)}{n}}}$$

tem a densidade t com n-1 graus de liberdade e que , para n>120, é normal reduzida.

Obtêm-se então os intervalos de confiança para p dados por

(2.74) $$p^* - t_{n-1,q}\sqrt{\frac{p^*(1-p^*)}{n}} \leq p \leq p^* + t_{n-1,q}\sqrt{\frac{p^*(1-p^*)}{n}}$$

podendo igualmente realizar-se testes t para hipóteses definidas a partir de p.

A semi - amplitude do intervalo anterior é

(2.75) $$A = t_{n-1,q}\sqrt{\frac{p^*(1-p^*)}{n}}\ .$$

Se pretendermos obter uma determinada semi - amplitude , digamos $\overline{A}$, terá de se resolver a equação seguinte em ordem a n

(2.76) $$\overline{A} = t_{n-1,q}\sqrt{\frac{p^*(1-p^*)}{n}}$$

Fazendo

(2.77) $$k = \frac{p^*(1-p^*)}{n}$$

vem

(2.78) $$n = \frac{p^*(1-p^*)}{k}$$

valor este que, tal como no caso referido anteriormente, se arredonda para cima.

Consideremos a função $g(p) = p(1-p)$ cujo gráfico está representado na figura 1 .

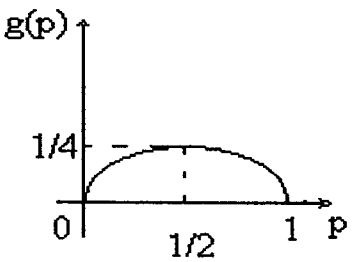


Figura 1

verificando-se

(2.79) $$\begin{cases} g\left(\frac{1}{2}-q\right) = g\left(\frac{1}{2}+q\right) \\ g(0) = g(1) = 0 \\ g\left(\frac{1}{2}\right) = \frac{1}{4} \end{cases}$$

Concluímos então que, quanto mais próximo p^* estiver de ½ , mais observações é preciso tomar, para se obter a precisão pretendida.

2.6. UTILIZAÇÃO DE TABELAS DE NÚMEROS ALEATÓRIOS

A utilização de tabelas de números aleatórios é um método simples e satisfatório de obter uma a mostra aleatória simples, desde que se consiga indexar os elementos da população por qualquer ordem conveniente.

As tabelas de números aleatórios tanto podem ser lidas na horizontal como na vertical, podendo ser tomados algarismos isolados como combinações de 2, 3 ou mais dígitos.

Partindo de determinado ponto (escolhido aleatoriamente), percorrem-se as linhas (colunas) da tabela num qualquer sentido, escolhendo-se os números que aparecem ao fim de um intervalo previamente estabelecido. Esse números indicarão quais os elementos da população que irão compor a amostra.

Apesar da facilidade de utilização destas tabelas há que ter em consideração que, em utilizações repetidas, se deve variar o ponto de partida , o sentido e o passo com que os números são seleccionados.

Há ainda que evitar utilizar a mesma tabela um número exagerado de vezes. Este procedimento assegura que todas as possíveis combinações têm igual probabilidade de serem seleccionadas.

3. AMOSTRAGEM SIMPLES COM PROBABILIDADES VARIÁVEIS

A amostragem simples, com probabilidades variáveis, pressupõe um conhecimento, à priori, da população e o uso desse conhecimento para atribuir ,às unidades estatísticas, diferentes probabilidades de pertencerem à amostra.

Nos esquemas de selecção amostral discutidos até agora, todas as unidades da população tinham a mesma hipótese de serem seleccionadas para a amostra. Não é essencial que assim seja, a única exigência, que se deve fazer, é que as probabilidades de inclusão sejam conhecidas e não nulas. Pode, de facto, ser demonstrado que é possível atingir uma precisão mais elevada com probabilidades variáveis.

3.1. SEM REPOSIÇÃO

Como já vimos anteriormente, na amostragem sem reposição, os elementos já escolhidos deixam de estar disponíveis para as tiragens seguintes.

Consideremos

(3.1) $$P_i^j = P(Y_j = y_i) \;;\; j = 1,\ldots,n \;,\; i = 1,\ldots,N$$

a probabilidade do i-ésimo elemento da população ser escolhido na j-ésima tiragem, e

(3.2) $$\Pi_i = \sum_{j=1}^{n} P_i^j \;;\; i = 1,\ldots,N \;.$$

Seja

(3.3) $$A_i = \bigcup_{j=1}^{n} \left(Y_j = y_i\right),$$

o acontecimento que se verifica quando o i-ésimo elemento da população é escolhido. Como os acontecimentos que figuram nesta reunião são incompatíveis dois a dois, tem-se:

(3.4) $$P\left(A_i\right) = \sum_{j=1}^{n} P_i^j = \Pi_i \;;\; i = 1,\ldots,N \;.$$

Seja X_i , i=1,...,N , uma variável aleatória que toma os valores 1 ou 0 consoante o i-ésimo elemento da população é ou não escolhido. Esta variável tem o esquema:

$$X_i = \begin{cases} 0 & 1 \\ 1-\Pi_i & \Pi_i \end{cases} , \quad i = 1,\ldots,N \tag{3.5}$$

logo

$$\mu(X_i) = \Pi_i \quad ; \quad i = 1,\ldots,N \; . \tag{3.6}$$

Por outro lado tem-se:

$$\sum_{i=1}^{N} X_i = n \tag{3.7}$$

visto haver sempre n elementos escolhidos. Assim:

$$\sum_{i=1}^{n} \Pi_i = \sum_{i=1}^{N} \mu(X_i) = \mu\left(\sum_{i=1}^{N} X_i\right) = n \; . \tag{3.8}$$

Como já vimos, existem

$$L = \frac{N!}{(N-n)!} \tag{3.9}$$

amostras possíveis, às quais podemos atribuir índices r , r=1,...,L .

Seja C_i , i=1,...,N o conjunto dos índices das amostras que contêm o i-ésimo elemento. Dada uma função h(Y), dos valores da característica numérica para a qual estamos a amostrar, sendo q_r a probabilidade de se colher a amostra com índice r, teremos:

$$\mu\left(\sum_{j=1}^{n} h(Y_j)\right) = \sum_{r=1}^{L} q_r \left(\sum_{j=1}^{n} h(Y_j)\right)_{(r)} \tag{3.10}$$

sendo $\left(\sum_{j=1}^{n} h(Y_j)\right)_{(r)}$ a soma dos valores da função relativamente à amostra com índice r . Logo, agrupando, no segundo membro, os termos em que intervém o i-ésimo elemento da população vem:

$$\mu\left(\sum_{j=1}^{n} h(Y_j)\right) = \sum_{i=1}^{N} \left(\sum_{r \in C_i} q_r\right) h(y_i) = \sum_{i=1}^{N} \Pi_i h(y_i) \tag{3.11}$$

visto que Π_i , a probabilidade de o i-ésimo elemento ser escolhido, é a soma das probabilidades das amostras que contêm esse elemento.

Utilizando estas igualdades, podemos agora obter um estimador linear centrado de μ (o valor médio da característica em estudo, $\mu = \frac{1}{N}\sum_{i=1}^{N} y_i$).

Seja

$$h(Y_j) = \frac{Y_j}{\Pi_j} \; ; \; j = 1,\ldots,n \; , \tag{3.12}$$

então a expressão do estimador pretendido será:

$$\mu^* = \frac{1}{N}\sum_{j=1}^{n} \frac{Y_j}{\Pi_j} \tag{3.13}$$

pois como se pode ver

$$\mu(\mu^*) = \frac{1}{N}\mu\left(\sum_{j=1}^{n} \frac{Y_j}{\Pi_j}\right) = \frac{1}{N}\sum_{i=1}^{N} \Pi_i \frac{y_i}{\Pi_i} = \frac{1}{N}\sum_{i=1}^{N} y_i = \mu \; . \tag{3.14}$$

Por outro lado, sendo $\Pi_{i,i'}$, com $i \neq i'$, a probabilidade de os elementos com índices i e i' serem ambos escolhidos para a amostra, tem-se:

$$\Pi_{i,i'} = \sum_{r \in C_i \cap C_{i'}} q_r \; ; \; i,i' = 1,\ldots,N \; ; \; i \neq i' \; , \tag{3.15}$$

ou seja, $\Pi_{i,i'}$ é a soma das probabilidades correspondentes às amostras que contêm ambos os elementos indicados e essas amostras têm os índices em $C_i \cap C_{i'}$.

Dada agora uma função g(Y,Y') de pares de valores da característica, temos

$$\mu\left(\sum_{\substack{j=1 \\ j \neq j'}}^{n}\sum_{j'=1}^{n} g(Y_j; Y_{j'})\right) = \sum_{r=1}^{L} q_r \left(\sum_{\substack{j=1 \\ j \neq j'}}^{n}\sum_{j'=1}^{n} g(Y_j; Y_{j'})\right)_{(r)} \; . \tag{3.16}$$

Agrupando os termos do segundo membro em que intervém o mesmo par de elementos da população vem:

$$\mu\left(\sum_{\substack{j=1 \\ j \neq j'}}^{n}\sum_{j'=1}^{n} g(Y_j; Y_{j'})\right) = \sum_{\substack{i=1 \\ i \neq i'}}^{N}\sum_{i'=1}^{N}\left(\sum_{r \in C_i \cap C_{i'}} q_r\right) g(y_i; y_{i'}) = \sum_{\substack{i=1 \\ i \neq i'}}^{N}\sum_{i'=1}^{N} \Pi_{i,i'} g(y_i; y_{i'}) \tag{3.17}$$

Consideremos uma variável aleatória

$$U = \sum_{j=1}^{n} h(Y_j) \tag{3.18}$$

para a qual se calcula a variância. Então:

$$U^2 = \sum_{j=1}^{n} h^2(Y_j) + \sum_{\substack{j=1 \\ j \neq j'}}^{n} \sum_{j'=1}^{n} h(Y_j)h(Y_{j'}). \tag{3.19}$$

De (3.11) vem

$$\mu(U) = \sum_{i=1}^{N} \Pi_i h(y_i) \tag{3.20}$$

e considerando também (3.17) vem

$$\mu(U^2) = \sum_{i=1}^{N} \Pi_i h^2(y_i) + \sum_{\substack{i=1 \\ i \neq i'}}^{N} \sum_{i'=1}^{N} \Pi_{i,i'} h(y_i)h(y_{i'}) \tag{3.21}$$

e

$$\mu^2(U) = \sum_{i=1}^{N} \Pi_i^2 h^2(y_i) + \sum_{\substack{i=1 \\ i \neq i'}}^{N} \sum_{i'=1}^{N} \Pi_i \Pi_{i'} h(y_i)h(y_{i'}) \tag{3.22}$$

.

Temos então

$$\begin{aligned}(3.23)\ \sigma^2(U) &= \mu(U^2) - \mu^2(U) = \\ &= \sum_{i=1}^{N} \Pi_i h^2(y_i) + \sum_{\substack{i=1 \\ i \neq i'}}^{N} \sum_{i'=1}^{N} \Pi_{i,i'} h(y_i)h(y_{i'}) - \sum_{i=1}^{N} \Pi_i^2 h^2(y_i) - \sum_{\substack{i=1 \\ i \neq i'}}^{N} \sum_{i'=1}^{N} \Pi_i \Pi_{i'} h(y_i)h(y_{i'}) = \\ &= \sum_{i=1}^{N} \Pi_i (1 - \Pi_i) h^2(y_i) + \sum_{\substack{i=1 \\ i \neq i'}}^{N} \sum_{i'=1}^{N} (\Pi_{i,i'} - \Pi_i \Pi_{i'}) h(y_i)h(y_{i'})\end{aligned}$$

Considerando (3.11) e (3.17), tem-se $\mu\left(\sum_{j=1}^{n} (1 - \Pi_j) h^2(Y_j)\right) = \sum_{i=1}^{N} \Pi_i (1 - \Pi_i) h^2(y_i)$, e $\mu\left(\sum_{\substack{j=1 \\ j \neq j'}}^{n} \sum_{j'=1}^{n} \frac{\Pi_{j,j'} - \Pi_j \Pi_{j'}}{\Pi_{j,j'}} h(Y_j)h(Y_{j'})\right) = \sum_{\substack{i=1 \\ i \neq i'}}^{N} \sum_{i'=1}^{N} (\Pi_{i,i'} - \Pi_i \Pi_{i'}) h(y_i)h(y_{i'})$. Então, dado o valor médio de uma soma ser a soma dos valores médios, vê-se que,

$$\sigma^{2*}(U) = \sum_{j=1}^{n} (1 - \Pi_j) h^2(y_j) + \sum_{j=1}^{n} \sum_{j'=1}^{n} \frac{\Pi_{j,j'} - \Pi_j \Pi_{j'}}{\Pi_{j,j'}} h(y_j)h(y_{j'}) \tag{3.24}$$

é estimador centrado de $\sigma^2(U)$.

Usando (3.23) e (3.24) é possível calcular $\sigma^2(\mu^*)$ e obter um estimador centrado para esta variância.

Seja

(3.25) $$h(Y_j)=\frac{Y_j}{\Pi_j} \quad ; \quad j=1,\ldots,n$$

atendendo a (3.23) obtém-se

(3.26) $$\sigma^2(\mu^*)=\frac{1}{N^2}\sigma^2\left(\sum_{j=1}^{n}\frac{Y_j}{\Pi_j}\right)=\frac{1}{N^2}\left[\sum_{i=1}^{N}(1-\Pi_i)\frac{y_i^2}{\Pi_i}+\sum_{i=1}^{N}\sum_{i'=1}^{N}(\Pi_{i,i'}-\Pi_i\Pi_{i'})\frac{y_i}{\Pi_i}\frac{y_{i'}}{\Pi_{i'}}\right]$$

e atendendo a (3.24) obtém-se

(3.27) $$\sigma^{2*}(\mu^*)=\frac{1}{N^2}\sigma^{2*}\left(\sum_{j=1}^{n}\frac{Y_j}{\Pi_j}\right)=\frac{1}{N^2}\left[\sum_{j=1}^{n}(1-\Pi_j)\frac{Y_j^2}{\Pi_j^2}+\sum_{\substack{j=1\\ j\neq j'}}^{n}\sum_{j'=1}^{n}\frac{\Pi_{j,j'}-\Pi_j\Pi_{j'}}{\Pi_{j,j'}}\frac{Y_j}{\Pi_j}\frac{Y_{j'}}{\Pi_{j'}}\right].$$

Para escolhermos as probabilidades que levam a minimizar $\sigma^2(\mu^*)$, temos que obter uma expressão alternativa para (3.23). Para tal comecemos por observar que a variável aleatória $X_{i,i'}=X_iX_{i'}$; $i,i'=1,\ldots,N$ e $i\neq i'$,toma o valor 1 quando ambos os elementos da população com índices i e i' são escolhidos e o valor 0 quando pelo menos um deles não é escolhido. Assim esta variável terá o esquema

(3.28) $$X_{i,i'}\begin{cases}0 & 1\\ 1-\Pi_{i,i'} & \Pi_{i,i'}\end{cases} \quad ; \quad i,i'=1,\ldots,N \ , \ i\neq i'$$

vindo

(3.29) $$\mu(X_{i,i'})=\Pi_{i,i'} \quad ; \quad i,i'=1,\ldots,N \ , \ i\neq i' \ .$$

Observe-se agora que

(3.30) $$\sum_{\substack{i'=1\\ i'\neq i}}^{N}X_{i,i'}=\sum_{\substack{i'=1\\ i'\neq i}}^{N}X_iX_{i'}=X_i\sum_{\substack{i'=1\\ i'\neq i}}^{N}X_{i'}=X_i(n-X_i)=nX_i-X_i^2=$$
$$=nX_i-X_i=(n-1)X_i \quad ; \quad i=1,\ldots,N$$

visto que $\sum_{i=1}^{N}X_i=n$ e que, como X_i toma apenas os valores 0 e 1, $X_i^2=X_i$. Logo, atendendo à expressão anterior:

(3.31) $$\sum_{\substack{i'=1\\ i'\neq i}}^{N}\Pi_{i,i'}=\sum_{\substack{i'=1\\ i'\neq i}}^{N}\mu(X_{i,i'})=\mu\left(\sum_{\substack{i'=1\\ i'\neq i}}^{N}X_{i,i'}\right)=\mu[(n-1)X_i]=$$
$$=(n-1)\mu(X_i)=(n-1)\Pi_i \quad ; \quad i=1,\ldots,N$$

visto que $\mu(X_i) = \Pi_i$.

Como $\sum_{i=1}^{N} \Pi_i = n$ vem

$$(3.32) \qquad \sum_{\substack{i'=1 \\ i' \neq i}}^{N} \left(\Pi_i \Pi_{i'} - \Pi_{i,i'}\right) = \Pi_i \sum_{\substack{i'=1 \\ i' \neq i}}^{N} \Pi_{i'} - \sum_{\substack{i'=1 \\ i' \neq i}}^{N} \Pi_{i,i'} = \Pi_i (n - \Pi_i) - (n-1)\Pi_i = \Pi_i (1 - \Pi_i).$$

Atendendo à expressão anterior, (3.23) passa a ter a forma:

$$(3.33) \qquad \sigma^2(U) = \sum_{\substack{i=1 \\ i \neq i'}}^{N} \left[\sum_{i'=1}^{N} \left(\Pi_i \Pi_{i'} - \Pi_{i,i'}\right) \right] h^2(y_i) - \sum_{i=1}^{N} \sum_{i'=1}^{N} \left(\Pi_i \Pi_{i'} - \Pi_{i,i'}\right) h(y_i) h(y_{i'})$$

e como

$$(3.34) \qquad \sum_{\substack{i=1 \\ i \neq i'}}^{N} \left[\sum_{i'=1}^{N} \left(\Pi_i \Pi_{i'} - \Pi_{i,i'}\right) \right] h^2(y_i) = \sum_{\substack{i'=1 \\ i \neq i'}}^{N} \left[\sum_{i=1}^{N} \left(\Pi_i \Pi_{i'} - \Pi_{i,i'}\right) \right] h^2(y_i)$$

vem

(3.35)

$$\sigma^2(U) = \frac{1}{2} \sum_{\substack{i=1 \\ i \neq i'}}^{N} \left[\sum_{i'=1}^{N} \left(\Pi_i \Pi_{i'} - \Pi_{i,i'}\right) \right] h^2(y_i) - \sum_{i=1}^{N} \sum_{i'=1}^{N} \left(\Pi_i \Pi_{i'} - \Pi_{i,i'}\right) h(y_i) h(y_{i'}) +$$

$$+ \frac{1}{2} \sum_{\substack{i'=1 \\ i \neq i'}}^{N} \left[\sum_{i=1}^{N} \left(\Pi_i \Pi_{i'} - \Pi_{i,i'}\right) \right] h^2(y_{i'}) =$$

$$= \frac{1}{2} \sum_{\substack{i=1 \\ i \neq i'}}^{N} \left[\sum_{i'=1}^{N} \left(\Pi_i \Pi_{i'} - \Pi_{i,i'}\right) \right] \left[h^2(y_i) - 2h(y_i) h(y_{i'}) + h^2(y_{i'}) \right] =$$

$$= \frac{1}{2} \sum_{\substack{i=1 \\ i \neq i'}}^{N} \left[\sum_{i'=1}^{N} \left(\Pi_i \Pi_{i'} - \Pi_{i,i'}\right) \right] \left[h(y_i) - h(y_{i'}) \right]^2$$

Assim, como para μ^* se tem $h(Y_j) = \frac{Y_j}{\Pi_j}$, vem

$$(3.36) \qquad \sigma^2(\mu^*) = \frac{1}{2N^2} \sum_{i=1}^{N} \sum_{i'=1}^{N} \left(\Pi_i \Pi_{i'} - \Pi_{i,i'}\right) \left[\frac{y_i}{\Pi_i} - \frac{y_{i'}}{\Pi_{i'}} \right]^2,$$

que mostra que, se as probabilidades Π_i forem proporcionais aos y_i, $\sigma^2(\mu^*)$ é nula.

Atendendo a (3.24) e (3.32), obtemos

(3.37) $$\sigma^{2^*}(U)=\frac{1}{2}\sum_{j=1}^{n}\sum_{j'=1}^{n}\frac{\Pi_j\Pi_{j'}-\Pi_{j,j'}}{\Pi_{j,j'}}\left[h\left(Y_j\right)-h\left(Y_{j'}\right)\right]^2$$

então

(3.38) $$\sigma^{2^*}\left(\mu^*\right)=\frac{1}{2N^2}\sum_{j=1}^{n}\sum_{j'=1}^{n}\frac{\Pi_j\Pi_{j'}-\Pi_{j,j'}}{\Pi_{j,j'}}\left[\frac{Y_j}{\Pi_j}-\frac{Y_{j'}}{\Pi_{j'}}\right]^2 .$$

3.2. COM REPOSIÇÃO

Como já vimos anteriormente, na amostragem com reposição, os elementos que já foram escolhidos continuam disponíveis para as tiragens seguintes. O número de amostras possíveis é portanto $L=N^n$, pois, para cada uma das tiragens, existem N hipóteses de escolha.

Consideremos novamente as expressões (3.1) e (3.2).

Como, neste caso, há reposição, os acontecimentos $\left(Y_j=y_i\right)$, $j=1,\ldots,n$, não são, para um dado i, incompatíveis dois a dois, logo não se podem interpretar os Π_i como probabilidades.

Veremos a seguir que, os Π_i são agora os valores médios das variáveis que dão o número de vezes que cada elemento da população é escolhido.

Seja X_i^j a variável aleatória que toma o valor 1 ou 0 conforme o i-ésimo elemento da população é ou não escolhido, na j-ésima tiragem.

A variável aleatória X_i, que dá o número de vezes que o i-ésimo elemento é escolhido, é então da forma :

(3.39) $$X_i=\sum_{j=1}^{n}X_i^j .$$

Visto que, em cada tiragem, apenas um elemento é escolhido, temos:

(3.40) $$\sum_{i=1}^{N}X_i^j=1 \;;\; j=1,\ldots,n .$$

À variável aleatória X_i^j corresponde então o esquema:

(3.41) $$X_i^j\begin{cases}0 & 1\\ 1-P_i^j & P_i^j\end{cases};\; j=1,\ldots,n \;,\; i=1,\ldots,N .$$

Logo

(3.42) $$\mu\left(X_i^j\right)=P_i^j \;;\; j=1,\ldots,n \;,\; i=1,\ldots,N ,$$

pelo que

(3.43) $$\Pi_i = \sum_{j=1}^{n} P_i^j = \mu\left(\sum_{j=1}^{n} X_i^j\right) = \mu(X_i) \ ; \ i = 1,\ldots,N \ .$$

Embora, neste caso, a variável aleatória não tome só os valores 0 ou 1 , mas possa tomar valores de 1 até n, continuamos a ter

(3.44) $$\sum_{i=1}^{N} X_i = n$$

visto que, em cada tiragem, é escolhido apenas um único elemento. Então:

(3.45) $$\sum_{i=1}^{N} \Pi_i = \sum_{i=1}^{N} \mu(X_i) = \mu\left(\sum_{i=1}^{N} X_i\right) = n \ .$$

Seja $C_{i,s}$, s = 0, . . .,n , o conjunto dos índices das amostras em que o i-ésimo elemento da população figura s vezes. A soma das probabilidades de se obterem essas amostras dá-nos a probabilidade do i-ésimo elemento da população ser escolhido s vezes, logo

(3.46) $$P(X_i = s) = \sum_{r \in C_{i,s}} q_r \ ; \ s = 0,1,\ldots,n \ , \ i = 1,\ldots,N \ .$$

Consideremos ainda uma função $h(Y)$, dos valores da característica numérica para a qual se está a amostrar. Continua a ter-se

(3.47) $$\mu\left(\sum_{j=1}^{n} h(Y_j)\right) = \sum_{r=1}^{L} q_r \left(\sum_{j=1}^{n} h(Y_j)\right)_{(r)} ,$$

mas, se quisermos reagrupar os termos do segundo membro, para cada elemento da população, teremos de atender ao número de vezes que o elemento é escolhido. Como no segundo membro $h(Y_j)$ figura s vezes, nas amostras pertencentes a $C_{i,s}$, vem então:

(3.48) $$\mu\left(\sum_{j=1}^{n} h(Y_j)\right) = \sum_{i=1}^{N}\left(\sum_{s=1}^{n} s \cdot \left(\sum_{r \in C_{i,s}} q_r\right)\right) h(y_i) = \sum_{i=1}^{N}\left[\sum_{s=1}^{n} P(X_i = s) \cdot s\right] h(y_i) = \sum_{i=1}^{N} \mu(X_i) h(y_i) = \sum_{i=1}^{N} \Pi_i h(y_i) ,$$

o que mostra que a expressão (3.11) se mantém, só que agora os Π_i são valores médios e não probabilidades.

Raciocinando como para a amostragem sem reposição, verifica-se que (3.13) é, também neste caso, a expressão do estimador centrado para o valor médio μ.

Vamos agora mostrar que a expressão (3.17) também se aplica neste tipo de amostragem. Como estamos a considerar a reposição dos elementos, com $j \neq j'$, os acontecimentos $(Y_j = y_i)$ e $(Y_{j'} = y_{i'})$ são independentes, pelo que as variáveis aleatórias X_i^j e $X_{i'}^{j'}$ também são independentes.

Recordemos que X_i^j toma os valores 1 e 0 consoante o i-ésimo elemento da população é ou não escolhido na j-ésima tiragem, continua pois a verificar-se a expressão (3.42).

Note-se que, como X_i^j e $X_{i'}^{j'}$ são, com $j \neq j'$, independentes se tem

$$\mu\left(X_i^j X_{i'}^{j'}\right) = \mu\left(X_i^j\right)\mu\left(X_{i'}^{j'}\right) \; ; \; j,j' = 1,\ldots,N \; , \; i,i' = 1,\ldots,N \; ; \; j \neq j' \tag{3.49}$$

enquanto que, como, em cada tiragem, só é escolhido um elemento,

$$X_i^j \cdot X_{i'}^j = 0 \; ; \; j = 1,\ldots,n \; , \; i \neq i' \; . \tag{3.50}$$

Finalmente, como X_i^j apenas toma os valores 0 e 1, temos

$$X_i^{j^2} = X_i^j \; ; \; j = 1,\ldots,n \; , \; i = 1,\ldots,N \, . \tag{3.51}$$

Consideremos agora

$$\Pi_{i,i'} = \sum_{j=1}^{n}\sum_{j'=1}^{n} P_i^j P_{i'}^{j'} \; ; \; i,i' = 1,\ldots,N \; , \tag{3.52}$$

quando $i \neq i'$ tem-se, com $j \neq j'$

(3.53)

$$\Pi_{i,i'} = \sum_{j=1}^{n}\sum_{j'=1}^{n} \mu\left(X_i^j\right)\mu\left(X_{i'}^{j'}\right) = \sum_{j=1}^{n}\sum_{j'=1}^{n} \mu\left(X_i^j X_{i'}^{j'}\right) =$$

$$= \mu\left(\sum_{j=1}^{n}\sum_{j'=1}^{n} X_i^j X_{i'}^{j'}\right) == \mu\left(\sum_{j=1}^{n}\sum_{j'=1}^{n} X_i^j X_{i'}^{j'}\right) =$$

$$= \mu\left(\sum_{j=1}^{n}\sum_{j'=1}^{n} X_i^j X_{i'}^{j'}\right) = \mu\left[\left(\sum_{j=1}^{n} X_i^j\right)\left(\sum_{j'=1}^{n} X_{i'}^{j'}\right)\right] = \mu\left(X_i X_{i'}\right)$$

enquanto que, se $i = i'$, vem, com $j \neq j'$

(3.54)

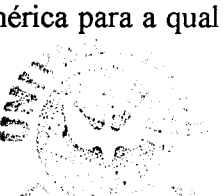
$$\Pi_{i,i'} = \sum_{j=1}^{n}\sum_{j'=1}^{n} \mu\left(X_i^j\right)\mu\left(X_{i'}^{j'}\right) = \sum_{j=1}^{n}\sum_{j'=1}^{n} \mu\left(X_i^j X_{i'}^{j'}\right) =$$

$$= \mu\left(\sum_{j=1}^{n}\sum_{j'=1}^{n} X_i^j X_{i'}^{j'}\right) = \mu\left(\sum_{j=1}^{n}\sum_{j'=1}^{n} X_i^j X_i^{j'}\right) =$$

$$= \mu\left(\sum_{j=1}^{n}\sum_{j'=1}^{n} X_i^j X_i^{j'} - \sum_{j=1}^{n} X_i^{j^2}\right) = \mu\left[\left(\sum_{j=1}^{n} X_i^j\right)\left(\sum_{j'=1}^{n} X_i^j\right) - \sum_{j=1}^{n} X_i^j\right] =$$

$$= \mu\left(X_i^2 - X_{i'}\right) = \mu\left[X_i\left(X_i - 1\right)\right]$$

Dada agora uma função $g(Y,Y')$ de pares de valores da característica numérica para a qual se está a amostrar, temos



(3.55) $$\mu\left(\sum_{j=1}^{n}\sum_{j'=1}^{n} g\left(Y_j, Y_{j'}\right)\right) = \sum_{r=1}^{L'} q_r \left(\sum_{j=1}^{n}\sum_{j'=1}^{n} g\left(y_j, y_{j'}\right)\right)_{(r)} .$$

Ao agruparmos, os termos do segundo membro, segundo os pares de elementos da população, temos de considerar, os pares de elementos distintos em separado dos pares de elementos repetidos. Para uma amostra com $r \in C_{i,s} \cap C_{i',s'}$ em que o i-ésimo elemento aparece s vezes e o i'-ésimo elemento aparece s' vezes, o par $\left(y_i, y_{i'}\right)$ ocorre $s \cdot s'$ vezes $(s + s' \leq n)$. Para uma amostra com $r \in C_{i,s}$ o par $\left(y_i, y_{i'}\right)$ aparece s(s-1) vezes, logo

(3.56)

$$\mu\left(\sum_{j=1}^{n}\sum_{j'=1}^{n} g\left(Y_j, Y_{j'}\right)\right) = \sum_{i=1}^{N}\sum_{i'=1}^{N}\left[\sum_{s=1}^{n-1}\sum_{s'=1}^{n-s}\left(\sum_{r \in C_{i,s} \cap C_{i',s'}} q_r\right) s \cdot s'\right] g\left(y_i, y_{i'}\right) + \sum_{i=1}^{N}\left[\sum_{s=1}^{N}\left(\sum_{r \in C_{i,s}} q_r\right) s(s-1)\right] g\left(y_i, y_{i'}\right) =$$

$j \neq j'$ $\qquad$ $i \neq i'$

$$= \sum_{\substack{i=1 \\ i \neq i'}}^{N}\sum_{i'=1}^{N} \mu\left(X_i X_{i'}\right) g\left(y_i, y_{i'}\right) + \sum_{i=1}^{N} \mu\left[X_i\left(X_i - 1\right)\right] g\left(y_i, y_i\right) =$$

$$= \sum_{\substack{i=1 \\ i \neq i'}}^{N}\sum_{i'=1}^{N} \Pi_{i,i'} g\left(y_i, y_{i'}\right) + \sum_{i=1}^{N} \Pi_{i,i'} g\left(y_i, y_i\right) =$$

$$= \sum_{i=1}^{N}\sum_{i'=1}^{N} \Pi_{i,i'} g\left(y_i, y_{i'}\right)$$

pelo que a expressão (3.17) continua a aplicar-se, só que agora, por ser com reposição, se pode ter i = i'.

Consideremos novamente as expressões (3.18) , (3.19) , (3.20) , (3.21) e (3.22),vê-se que:

(3.57)
$$\begin{cases} \mu\left(\sum_{j=1}^{n} h^2\left(Y_j\right)\right) = \sum_{i=1}^{N} \Pi_i h^2\left(y_i\right) \\ \mu\left(\sum_{j=1}^{n}\sum_{j'=1}^{n} \dfrac{\Pi_{j,j'} - \Pi_j \Pi_{j'}}{\Pi_{j,j'}} h\left(Y_j\right) h\left(Y_{j'}\right)\right) = \sum_{i=1}^{N}\sum_{i'=1}^{N} \left(\Pi_{i,i'} - \Pi_i \Pi_{i'}\right) h\left(y_i\right) h\left(y_{i'}\right) \end{cases}$$

o que mostra que

(3.58) $$\sigma^{2*}(U) = \sum_{j=1}^{n} h^2\left(Y_j\right) + \sum_{j=1}^{n}\sum_{j'=1}^{n} \frac{\Pi_{j,j'} - \Pi_j \Pi_{j'}}{\Pi_{j,j'}} h\left(Y_j\right) h\left(Y_{j'}\right)$$

é um estimador centrado de $\sigma^2(U)$.

Em particular, para μ^*, com $h\left(Y_j\right) = \dfrac{Y_j}{\Pi_j}$; $j = 1, \ldots, n$, as expressões (3.57) e (3.58) dão

(3.59)
$$\begin{cases} \sigma^2(\mu^*) = \dfrac{1}{N^2}\left[\sum_{i=1}^{N}\dfrac{y_i^2}{\Pi_i} + \sum_{i=1}^{N}\sum_{i'=1}^{N}\left(\Pi_{i,i'} - \Pi_i\Pi_{i'}\right)h(y_i)h(y_{i'})\right] \\ \sigma^{2*}(\mu^*) = \dfrac{1}{N^2}\left[\sum_{j=1}^{n}\dfrac{Y_j^2}{\Pi_j^2} + \sum_{j=1}^{n}\sum_{j'=1}^{n}\dfrac{\Pi_{j,j'} - \Pi_j\Pi_{j'}}{\Pi_{j,j'}}\dfrac{Y_j}{\Pi_j}\dfrac{Y_{j'}}{\Pi_{j'}}\right] \end{cases}$$

Tal como foi feito para a amostragem com reposição, procuremos uma expressão alternativa para $\sigma^2(U)$.

Assim:

(3.60)
$$\sum_{i'=1}^{N}\Pi_{i,i'} = \Pi_{i,i} + \sum_{\substack{i=1\\ i'\neq i}}^{N}\Pi_{i,i'} = \Pi_{i,i} + \sum_{\substack{i=1\\ i'\neq i}}^{N}\mu(X_iX_{i'}) =$$

$$= \Pi_{i,i} + \mu\left(\sum_{\substack{i=1\\ i'\neq i}}^{N}X_iX_{i'}\right) = \Pi_{i,i} + \mu\left(X_i\sum_{\substack{i=1\\ i'\neq i}}^{N}X_{i'}\right) = \Pi_{i,i} + \mu\left[X_i(n - X_i)\right] =$$

$$= \Pi_{i,i} + n\mu(X_i) - \mu(X_i^2) = \Pi_{i,i} + n\Pi_i - (\Pi_{i,i} + \Pi_i) = (n-1)\Pi_i$$

$$i = 1,\ldots,N$$

logo

(3.61)
$$\sum_{i'=1}^{N}\left(\Pi_i\Pi_{i'} - \Pi_{i,i'}\right) = \Pi_i\sum_{i'=1}^{N}\Pi_{i'} - \sum_{i'=1}^{N}\Pi_{i,i'} = n\Pi_i - (n-1)\Pi_i = \Pi_i \;\; ; \;\; i = 1,\ldots,N$$

pelo que

(3.62)
$$\sigma^2(U) = \frac{1}{2}\sum_{i=1}^{N}\left[\sum_{i'=1}^{N}\left(\Pi_i\Pi_{i'} - \Pi_{i,i'}\right)\right]h^2(y_i) - \sum_{i=1}^{N}\sum_{i'=1}^{N}\left(\Pi_i\Pi_{i'} - \Pi_{i,i'}\right)h(y_i)h(y_{i'}) +$$

$$+\frac{1}{2}\sum_{i'=1}^{N}\left[\sum_{i=1}^{N}\left(\Pi_i\Pi_{i'} - \Pi_{i,i'}\right)\right]h^2(y_{i'}) =$$

$$= \frac{1}{2}\sum_{i=1}^{N}\sum_{i'=1}^{N}\left(\Pi_i\Pi_{i'} - \Pi_{i,i'}\right)\left[h^2(y_i) - 2h(y_i)h(y_{i'}) + h^2(y_{i'})\right] =$$

$$= \frac{1}{2}\sum_{i=1}^{N}\sum_{i'=1}^{N}\left(\Pi_i\Pi_{i'} - \Pi_{i,i'}\right)\left[h(y_i) - h(y_{i'})\right]^2$$

assim, para μ^* ter-se-à

(3.63)
$$\sigma^2(\mu^*) = \frac{1}{2N^2}\sum_{i=1}^{N}\sum_{i'=1}^{N}\left(\Pi_i\Pi_{i'} - \Pi_{i,i'}\right)\left[\frac{y_i}{\Pi_i} - \frac{y_{i'}}{\Pi_{i'}}\right]^2$$

pelo que, se os valores médios Π_i forem proporcionais aos y_i, $\sigma^2(\mu^*)$ seria nula.

Atendendo a (3.58) e a (3.62),obtemos

$$\sigma^{2*}(U)=\frac{1}{2}\sum_{\substack{j=1\\ j\neq j'}}^{n}\sum_{j'=1}^{n}\frac{\Pi_j\Pi_{ji'}-\Pi_{j,j'}}{\Pi_{j,j'}}\left[h(Y_j)-h(Y_{j'i'})\right]^2 \tag{3.64}$$

vindo,

$$\sigma^{2*}(\mu^*)=\frac{1}{2N^2}\sum_{j=1}^{n}\sum_{ji'=1}^{n}\frac{\Pi_j\Pi_{j'}-\Pi_{j,j'}}{\Pi_{j,j'}}\left[\frac{Y_j}{\Pi_j}-\frac{Y_{j'}}{\Pi_{ji'}}\right]^2 . \tag{3.65}$$

Considerando agora um caso particular, em que as probabilidades de escolha dos vários elementos da população não variam de tiragem para tiragem (embora possam variar de elemento para elemento), ou seja, $P_i^j=p_i$; $j=1,\ldots,n$, $i=1,\ldots,N$ .Temos então:

$$\begin{cases}\Pi_i=np_i \ ; \ i=1,\ldots,N\\ \Pi_{i,i'}=n(n-1)p_ip_{i'} \ ; \ i,i'=1,\ldots,N\end{cases} \tag{3.66}$$

logo

$$\Pi_i\Pi_{i'}-\Pi_{i,i'}=np_ip_{i'} \ ; \ i,i'=1,\ldots,N \ , \tag{3.67}$$

tendo em conta (3.62) e (3.64) obtemos

$$\sigma^2(U)=\frac{1}{2}\sum_{i=1}^{N}\sum_{i'=1}^{N}np_ip_{i'}\left[h(y_i)-h(y_{i'})\right]^2 \tag{3.68}$$

e

$$\sigma^{2*}(U)=\frac{1}{2}\sum_{\substack{j=1\\ j\neq j'}}^{n}\sum_{j'=1}^{n}\frac{\left[h(Y_i)-h(Y_{j'})\right]^2}{n-1} . \tag{3.69}$$

Neste caso tem-se

$$\mu^*=\frac{1}{N}\sum_{j=1}^{n}\frac{Y_j}{np_j} \ , \tag{3.70}$$

$$\sigma^2(\mu^*)=\frac{1}{2N^2n}\sum_{i=1}^{N}\sum_{i'=1}^{N}p_ip_{i'}\left(\frac{y_i}{p_i}-\frac{y_{i'}}{p_{i'}}\right)^2 \tag{3.71}$$

e

(3.72) $$\sigma^{2^*}(\mu^*) = \frac{1}{2N^2 n^2 (n-1)} \sum_{\substack{j=1 \\ j \neq j'}}^{n} \sum_{j'=1}^{n} \left(\frac{Y_j}{p_j} - \frac{Y_{j'}}{p_{j'}} \right)^2$$

Se as probabilidades p_i forem proporcionais nos y_i, $\sigma^2(\mu^*)$ é nula.

4. AMOSTRAGEM ESTRATIFICADA

Já foi visto que, na amostragem aleatória simples, a variância do estimador, da média da população, $\overline{Y}$, depende da variabilidade da característica Y na população, independentemente da dimensão da amostra.

Se a população é bastante heterogénea, e a característica em estudo difere, consideravelmente, entre certos subconjuntos da população inicial, é aconselhável dividir a população. Essa divisão, chamada estratificação, deve ser feita de forma a que cada um desses subconjuntos (os estratos) seja constituído por indivíduos de alguma forma semelhantes. Os estratos podem ser definidos considerando diversas características. Em populações humanas podemos referir, por exemplo, a idade, o sexo, a classe social ou o grau de instrução.

A estratificação pode produzir um ganho de precisão nos estimadores das características da população inteira. Ao dividir uma população heterogénea em estratos, internamente homogéneos, nos quais existem apenas pequenas variações de indivíduo para indivíduo, é possível obter estimadores precisos, em cada estrato, usando uma pequena amostra de elementos de cada um desses estratos.

Obtidos os estimadores para cada estrato, estes podem então ser combinados num estimador preciso para a população inteira.

O principal objectivo, da teoria da amostragem estratificada, é portanto saber qual a melhor escolha para as dimensões das amostras, retiradas de cada estrato, por forma a obter a melhor precisão.

Consideremos uma divisão de uma população, de dimensão N, em k estratos, com dimensões $N_1,\ldots,N_k$, então:

$$N = \sum_{t=1}^{k} N_t \quad . \tag{4.1}$$

Se forem colhidas, dos vários estratos, sub-amostras independentes com dimensões $n_1,\ldots,n_k$, a amostra total terá dimensão

$$n = \sum_{t=1}^{k} n_t \quad . \tag{4.2}$$

Admitindo que as sub-amostras são colhidas sem reposição e com probabilidades iguais, a probabilidade de um elemento, do estrato t, ser escolhido será dada por:

$$\Pi_t = \frac{n_t}{N_t} \quad ; \quad t = 1,\ldots,N_t \tag{4.3}$$

visto que, em cada uma das n_t tiragens, realizadas nesse estrato, cada um dos seus elementos tem probabilidade $\frac{1}{N_t}$ de ser escolhido.

Sejam $Y_{t,j}$; j=1,...,n_t , as variáveis aleatórias que dão os resultados obtidos para o estrato t, atendendo à expressão (3.13), obtemos

$$\mu^* = \frac{1}{N}\sum_{t=1}^{k}\sum_{j=1}^{n_t}\frac{Y_{t,j}}{\Pi_t} = \frac{1}{N}\sum_{t=1}^{k}\frac{N_t}{n_t}\sum_{j=1}^{n_t}Y_{t,j} = \frac{1}{N}\sum_{t=1}^{k}N_t\overline{Y}_t \tag{4.4}$$

onde

$$\overline{Y}_t = \frac{1}{n_t}\sum_{j=1}^{n_t}Y_{t,j} \;;\; t=1,\ldots,k \tag{4.5}$$

é a média da sub-amostra colhida do estrato t .

Devido à grande homogeneidade de cada estrato, podemos considerá-lo como uma população. Assim, tendo em conta as devidas adaptações, podemos escrever a variância do estrato t utilizando (2.40). Temos então:

$$\sigma^2\left(\overline{Y}_t\right) = \frac{\sigma_t^2}{n_t}\frac{N_t - n_t}{N_t - 1} \;;\; t=1,\ldots,k \tag{4.6}$$

Visto que as sub-amostras são colhidas independentemente, em cada estrato, as médias, $\overline{Y}_1,\ldots,\overline{Y}_k$,são independentes, pelo que

$$\sigma^2\left(\mu^*\right) = \sigma^2\left(\frac{1}{N}\sum_{t=1}^{k}N_t\overline{Y}_t\right) = \frac{1}{N^2}\sum_{t=1}^{k}N_t^2\sigma^2\left(\overline{Y}_t\right) = \frac{1}{N^2}\sum_{t=1}^{k}N_t^2\frac{\sigma_t^2}{n_t}\frac{N_t - n_t}{N_t - 1} \tag{4.7}$$

Para N_t grande, $\frac{1}{N_t - 1} \cong \frac{1}{N_t}$, pelo que podemos considerar a aproximação, a $\sigma^2\left(\mu^*\right)$, dada por:

$$\begin{aligned}\sigma^2\left(\mu^*\right) &= \frac{1}{N^2}\sum_{t=1}^{k}N_t^2\frac{\sigma_t^2}{n_t}\left(1-\frac{n_t}{N_t}\right) = \frac{1}{N^2}\sum_{t=1}^{k}\left(N_t^2\frac{\sigma_t^2}{n_t} - N_t\sigma_t^2\right) = \\ &= -\frac{1}{N^2}\sum_{t=1}^{k}N_t\sigma_t^2 + \sum_{t=1}^{k}\frac{N_t^2\sigma_t^2}{N^2}\frac{1}{n_t}\end{aligned} \tag{4.8}$$

Fazendo

$$\begin{cases} V_0 = -\frac{1}{N^2}\sum_{t=1}^{k}N_t\sigma_t^2 \\ V_t = \frac{N_t^2}{N^2}\sigma_t^2 \;;\; t=1,\ldots,k \\ W_t = n_t \;;\; l=1,\ldots,k \end{cases} \tag{4.9}$$

vem

$$\sigma^2(\mu^*) = V_0 + \sum_{t=1}^{k} \frac{V_t}{W_t} \quad . \tag{4.10}$$

Note-se que $V_0, V_1, \ldots, V_k$ apenas dependem dos parâmetros dos estratos $\left(N_t \text{ e } \sigma_t^2\right)$, enquanto que os $W_1, \ldots, W_k$ coincidem com as dimensões das sub-amostras.

Por outro lado, quando se realiza uma amostragem estratificada, para além de um custo fixo C_0, podemos ter custos $C_1, \ldots, C_k$ para cada estrato. Assim o custo total será

$$C = C_0 + \sum_{t=1}^{k} C_t n_t = \sum_{t=0}^{k} C_t n_t \quad , \quad \text{com } n_0 = 1 \tag{4.11}$$

Pela desigualdade de Cauchy, podemos escrever

$$\left(\sigma^2(\mu^*) - V_0\right)(C - C_0) = \left(\sum_{t=1}^{k} \frac{V_t}{W_t}\right)\left(\sum_{t=1}^{k} C_t W_t\right) \geq \left(\sqrt{\sum_{t=1}^{k} C_t V_t}\right)^2 \geq \sum_{t=1}^{k} C_t V_t \tag{4.12}$$

Visto que, a igualdade em (4.12) é obtida apenas quando se verifica a condição

$$\frac{\left(\frac{V_t}{W_t}\right)}{W_t C_t} = k^2 \quad \text{(constante para todos os valores de t)}, \tag{4.13}$$

e que segundo membro de (4.12),é independente de W_t, então a escolha de W_t para satisfazer esta condição, que rescrevemos

$$W_t^2 = \frac{V_t}{k^2 C_t} \quad , \quad \forall t \ , \tag{4.14}$$

minimizará $\sigma^2(\mu^*) \cdot C$.

Considerando (4.9) e (4.14), vem:

$$n_t = W_t = \frac{1}{k}\sqrt{\frac{V_t}{C_t}} = \frac{N_t \sigma_t}{kN\sqrt{C_t}} \quad ; \quad t = 1, \ldots, k \tag{4.15}$$

e portanto

$$n = \sum_{t=1}^{k} n_t = \frac{1}{kN} \sum_{t=1}^{k} \frac{N_t \sigma_t}{\sqrt{C_t}} \tag{4.16}$$

o que dá

(4.17)
$$\frac{n_t}{n} = \frac{\frac{N_t\sigma_t}{\sqrt{C_t}}}{\sum_{t'=1}^{k}\frac{N_{t'}\sigma_{t'}}{\sqrt{C_{t'}}}} \quad ; \quad t = 1,\ldots,k$$

ou seja

(4.18)
$$n_t = n\frac{\frac{N_t\sigma_t}{\sqrt{C_t}}}{\sum_{t'=1}^{k}\frac{N_{t'}\sigma_{t'}}{\sqrt{C_{t'}}}} \quad ; \quad t = 1,\ldots,k \quad .$$

Note-se que $\sigma^2(\mu^*)$ e C foram expressas como somas de duas componentes, uma fixa V_0 e C_0 e outra dada pelas funções $\sum_{t=1}^{k}\frac{V_t}{n_t}$ e $\sum_{t=1}^{k}C_t n_t$.

Fixando o custo total C, para minimizar $(\sigma^2(\mu^*) - V_0)(C - C_0)$ basta minimizar $\sigma^2(\mu^*)$. Se fixarmos $\sigma^2(\mu^*)$ basta minimizar C.

Se os custos $C_1,\ldots,C_k$, correspondentes aos diferentes estratos, forem desprezados ou considerandos iguais ,C_t=C' obtêm-se expressões simplificadas para as dimensões dos estratos, n_t, da forma:

(4.19)
$$n_t = n\frac{N_t\sigma_t}{\sum_{t'=1}^{k}N_{t'}\sigma_{t'}} \quad ; \quad t = 1,\ldots,k \quad .$$

As expressões obtidas aqui podem também aplicar-se à amostragem estratificada por atributos.

Consideremos p_t a probabilidade de os elementos de um qualquer estrato t possuírem o atributo que se está a estudar, tem-se então:

(4.20)
$$\sigma_t = \sqrt{p_t(1-p_t)}$$

vindo, quando são considerados os custos C_t correspondentes aos diferentes estratos,

(4.21)
$$n_t = n\frac{\frac{N_t\sqrt{p_t(1-p_t)}}{\sqrt{C_t}}}{\sum_{t'=1}^{k}\frac{N_{t'}\sqrt{p_{t'}(1-p_{t'})}}{\sqrt{C_{t'}}}} \quad ; \quad t = 1,\ldots,k$$

ou então, ignorando esses custos,

(4.22) $$n_t = n\frac{N_t\sqrt{p_t(1-p_t)}}{\sum_{t'=1}^{k} N_{t'}\sqrt{p_{t'}(1-p_{t'})}} \quad ; \quad t=1,\ldots,k \quad .$$

Comparemos agora as variâncias, dos estimadores de μ, obtidas usando a amostragem estratificada $\sigma_E^2(\mu^*)$ e a amostragem, sem reposição, com probabilidades iguais $\sigma_s^2(\mu^*)$, utilizando em ambos os casos amostras da mesma dimensão, n.

Sejam $y_{t,i}$ os valores da característica numérica, para a qual se está a amostrar, nos elementos do estrato t, os valores médios dos estratos serão:

(4.23) $$\mu_t = \frac{1}{N_t}\sum_{i=1}^{N_t} y_{t,i} \quad ; \quad t=1,\ldots,k$$

e o valor médio da população será dado por:

(4.24) $$\mu = \frac{1}{N}\sum_{t=1}^{k}\sum_{i=1}^{N_t} y_{t,i} \quad .$$

As variâncias dos estratos e da população podem agora ser redefinidas como

(4.25) $$\begin{cases} \sigma_t^2 = \dfrac{1}{N_t-1}\displaystyle\sum_{i=1}^{N_t}(y_{t,i}-\mu_t)^2 \quad ; \quad t=1,\ldots,k \\ \sigma^2 = \dfrac{1}{N-1}\displaystyle\sum_{t=1}^{k}\sum_{i=1}^{N_t}(y_{t,i}-\mu)^2 \end{cases}$$

Obtém-se assim, considerando (2.40) e

(4.26) $$\begin{aligned} (N-1)\sigma^2 &= \sum_{t=1}^{k}\sum_{i=1}^{N_t}(y_{t,i}-\mu)^2 = \sum_{t=1}^{k}\sum_{i=1}^{N_t}\left[(y_{t,i}-\mu_t)+(\mu_t-\mu)\right]^2 = \\ &= \sum_{t=1}^{k}\sum_{i=1}^{N_t}\left[(y_{t,i}-\mu_t)^2+2(\mu_t-\mu)(y_{t,i}-\mu_t)+(\mu_t-\mu)^2\right] = \\ &= \sum_{t=1}^{k}\sum_{i=1}^{N_t}(y_{t,i}-\mu_t)^2+2\sum_{t=1}^{k}\left[(\mu_t-\mu)\sum_{i=1}^{N_t}(y_{t,i}-\mu_t)\right]+\sum_{t=1}^{k}N_t(\mu_t-\mu)^2 = \\ &= \sum_{t=1}^{k}(N_t-1)\sigma_t^2+\sum_{t=1}^{k}N_t(\mu_t-\mu)^2 \end{aligned}$$

com $\sum_{i=1}^{N_t}(y_{t,i}-\mu_t)=0$; $t=1,\ldots,k$ (visto que a soma dos desvios para a média é nula),

(4.27) $$\sigma_S^2(\mu^*) = \frac{\sigma^2}{n}\frac{N-n}{N} = \frac{N-n}{nN(N-1)}\left[\sum_{t=1}^{k}(N_t-1)\sigma_t^2 + \sum_{t=1}^{k}N_t(\mu_t-\mu)^2\right] .$$

Consideremos agora que a amostragem estratificada é realizada com critério proporcional, isto é, que

(4.28) $$\frac{n_t}{N_t} = \frac{n}{N} \;;\; t = 1,\ldots,k \;.$$

Atendendo a (4.8) vem então:

(4.29) $$\sigma_E^2(\mu^*) = \frac{1}{N^2}\sum_{t=1}^{k}N_t^2\frac{\sigma_t^2}{n_t}\left(1-\frac{n_t}{N_t}\right) = \frac{1}{N^2}\frac{N}{n}\left(1-\frac{n}{N}\right)\sum_{t=1}^{k}N_t\sigma_t^2 = \frac{N-n}{nN^2}\sum_{t=1}^{k}N_t\sigma_t^2 ,$$

logo

(4.30)
$$\sigma_S^2(\mu^*) - \sigma_E^2(\mu^*) = \frac{N-n}{nN}\left\{\frac{1}{N-1}\left[\sum_{t=1}^{k}(N_t-1)\sigma_t^2 + \sum_{t=1}^{k}N_t(\mu_t-\mu)^2\right] - \frac{1}{N}\sum_{t=1}^{k}N_t\sigma_t^2\right\} =$$
$$= \frac{N-n}{nN}\left\{\left[\frac{1}{N-1}\sum_{t=1}^{k}(N_t-1)\sigma_t^2 - \frac{1}{N}\sum_{t=1}^{k}N_t\sigma_t^2\right] + \frac{1}{N-1}\sum_{t=1}^{k}N_t(\mu_t-\mu)^2\right\}$$

onde

(4.31) $$\frac{1}{N-1}\sum_{t=1}^{k}(N_t-1)\sigma_t^2 - \frac{1}{N}\sum_{t=1}^{k}N_t\sigma_t^2 = \frac{1}{N(N-1)}\sum_{t=1}^{k}(N_t-N)\sigma_t^2 < 0 \;.$$

Considerando, o caso particular em que $\mu_t = \cdots = \mu_k$, isto é, $\mu_t = \mu$, tem-se

(4.32) $$\frac{1}{N-1}\sum_{t=1}^{k}N_t(\mu_t-\mu)^2 = 0,$$

pelo que,

(4.33) $$\sigma_s^2(\mu^*) < \sigma_E^2(\mu^*) \;.$$

Vê-se assim que, para que a amostragem estratificada conduza a bons resultados, os valores médios, dos estratos, devem diferir bastante entre si, isto é, a variância entre os estratos, $\frac{1}{N-1}\sum_{t=1}^{k}N_t(\mu_t-\mu)^2$, deve ser grande, enquanto que cada estrato deve ser homogéneo, isto é, a variância, dentro de cada estrato, deve ser pequena.

5. AMOSTRAGEM PARA AGREGADOS

De modo a não alongar, desnecessariamente, a discussão deste tipo de amostragem, consideramos apenas o seu caso mais elementar, que, apesar da sua simplicidade, ilustra, perfeitamente, a teoria da amostragem para agregados.

Consideremos agora uma divisão da população em agregados disjuntos dois a dois e escrevamos iθi' quando " i " e " i' " são índices de elementos que pertencem ao mesmo agregado.

Sejam Π_i a probabilidade de o elemento com índice i ser escolhido e $\Pi_{i,i'}$ a probabilidade de o par de elementos i e i' ser escolhido.

Se $\Pi_i = \Pi_{i'} = \Pi_{i,i'}$, com iθi', vê-se que a probabilidade de o elemento i' ser escolhido, quando o elemento i é escolhido, é dada por

$$\frac{\Pi_{i,i'}}{\Pi_i} = 1 \tag{5.1}$$

logo quando um elemento de um agregado é escolhido os restantes elementos do agregado são escolhidos (com probabilidade 1). Assim, para se ter $\Pi_i = \Pi_{i'} = \Pi_{i,i'}$, com iθi', a amostra tem de ser constituída por agregados completos, isto é, que abranjam toda a população.

Inversamente, se a amostra for constituída por agregados completos, dado iθi', as probabilidades de os elementos i e i' serem escolhidos, isoladamente ou em conjunto, são iguais à probabilidade de ser escolhido o agregado a que pertencem.

Assim, com iθi', tem-se $\Pi_i = \Pi_{i'} = \Pi_{i,i'}$ se e só se a amostra for constituída por agregados completos.

Admitamos que se constituem N_1 agregados com N_2 elementos cada e que a amostra é formada por n_1 agregados, tem-se então

$$\begin{cases} N = N_1 N_2 \\ n = n_1 N_2 \end{cases} \tag{5.2}$$

logo

$$\frac{n_1}{N_1} = \frac{n}{N} \tag{5.3}$$

A escolha da amostra reduz-se à escolha dos agregados que a vão compor .Trata-se pois de escolher uma amostra de n_1 agregados dentro da população dos agregados.

Suponhamos que esta amostragem é realizada com probabilidades iguais e sem reposição. A probabilidade de se escolher qualquer agregado será $\frac{n_1}{N_1}$. Como a probabilidade de se escolher um elemento é igual à probabilidade de se escolher o agregado a que pertence, vem

(5.4) $$\Pi_i = \frac{n_1}{N_1} = \frac{n}{N} \quad ; \quad i = 1,\ldots,N$$

vindo igualmente

(5.5) $$i\theta i' \quad ; \quad \Pi_{i,i'} = \frac{n_1}{N_1} = \frac{n}{N} \quad .$$

Se $i\theta i'$, isto é, se os elementos i e i' pertencem a agregados diferentes, a probabilidade de serem ambos escolhidos é igual à probabilidade de os dois agregados a que pertencem serem escolhidos. Admitindo que os agregados são escolhidos pela técnica da amostragem aleatória simples, com probabilidades iguais e sem reposição, havendo N_1 agregados e destes sendo escolhidos n_1, haverá, adaptando (3.9),

(5.6) $$L = \frac{N_1!}{(N_1 - n_1)!}$$

amostras possíveis formadas por n1 agregados, todos com probabilidades iguais.

Dados dois agregados, podemos agrupar as amostras, que os contêm, segundo as posições dos dois agregados nessa amostra. Haverá $n(n-1)$ destas classes de amostras cada uma com

(5.7) $$L' = \frac{(N_1 - 2)!}{(N_1 - n_1)!}$$

amostras, visto que as amostras de cada classe correspondem às amostras de dimensão n_1-2 que se podem extrair da população dos agregados, uma vez excluídos os dois agregados dados. Assim haverá $n_1(n_1 - 1)L'$ amostras contendo os dois agregados. Como todas as amostras têm probabilidades iguais teremos

(5.8) $$i\theta i' \quad ; \quad \Pi_{i,i'} = \frac{n_1(n_1 - 1)\dfrac{(N_1 - 2)!}{(N_1 - n_1)!}}{\dfrac{N_1!}{(N_1 - n_1)!}} = \frac{n_1(n_1 - 1)}{N_1(N_1 - 1)} \quad .$$

O estimador de μ é então dado por

(5.9) $$\mu^* = \frac{1}{N}\sum_{j=1}^{n}\frac{Y_j}{\Pi_j} = \frac{1}{N}\sum_{j=1}^{n}\frac{N}{n}Y_j = \overline{Y}$$

onde $\overline{Y}$ é a média da amostra.

Por outro lado, atendendo a (3.36), com $\Pi_i = \Pi_{i'} = \frac{n}{N}$, temos

(5.10) $$\sigma^2(\mu^*) = \frac{1}{2N^2}\sum_{i=1}^{N}\sum_{i'=1}^{N}\left(\Pi_i\Pi_{i'} - \Pi_{i,i'}\right)\frac{N^2}{n^2}\left(y_i - y_{i'}\right)^2 .$$

Como

(5.11) $$\begin{cases} i\theta i' \;; \Pi_i\Pi_{i'} - \Pi_{i,i'} = \dfrac{n^2}{N^2} - \dfrac{n}{N} = \dfrac{n}{N}\left(\dfrac{n}{N} - 1\right) = \dfrac{n_1}{N_1}\left(\dfrac{n_1}{N_1} - 1\right) \\ i\not\theta i \;; \Pi_i\Pi_{i'} - \Pi_{i,i'} = \dfrac{n^2}{N^2} - \dfrac{n_1}{N_1}\left(\dfrac{n_1 - 1}{N_1 - 1}\right) = \dfrac{n_1^2}{N_1^2} - \dfrac{n_1}{N_1} = \dfrac{n_1}{N_1}\left(\dfrac{N_1 - n_1}{N_1(N_1 - 1)}\right) \end{cases},$$

se pusermos

(5.12) $$\begin{cases} s_1 = \sum\sum_{i\theta i'}\left(y_i - y_{i'}\right)^2 \\ s_2 = \sum\sum_{i\not\theta i'}\left(y_i - y_{i'}\right)^2 \end{cases}$$

onde os dois somatórios são feitos para os pares (i,i') que satisfazem ou não a relação θ , vemos que

(5.13) $$s = \sum_{\substack{i=1 \\ i \neq i'}}^{N}\sum_{i'=1}^{N}\left(y_i - y_{i'}\right)^2 = s_1 + s_2$$

obtendo-se ainda, devido a (5.11):

(5.14)

$$\begin{aligned} \sigma^2(\mu^*) &= \frac{1}{2N^2}\frac{N^2}{n^2}\left[\frac{n_1}{N_1}\left(\frac{n_1}{N_1} - 1\right)s_1 + \frac{n_1}{N_1}\frac{N_1 - n_1}{N_1(N_1 - 1)}s_2\right] = \\ &= \frac{1}{2n^2}\left[\frac{n_1}{N_1}\left(\frac{n_1}{N_1} - 1\right)s_1 - \frac{n_1}{N_1}\frac{N_1 - n_1}{N_1(N_1 - 1)}s_2\right] = \\ &= \frac{1}{2n^2}\frac{n}{N}\left[\left(\frac{n_1}{N_1} - 1\right)s_1 - \frac{N_1 - n_1}{N_1(N_1 - 1)}\left(s - s_1\right)\right] = \\ &= \frac{1}{2nN}\left[\frac{N_1 - n_1}{N_1(N_1 - 1)}s - \frac{1}{N_1}\left(\frac{N_1 - n_1}{N_1 - 1} + N_1 - n_1\right)s_1\right] = \\ &= \frac{1}{2nN}\left[\frac{N_1 - n_1}{N_1(N_1 - 1)}s - \frac{(N_1 - n_1)N_1^2}{N_1^2(N_1 - 1)}s_1\right] = \\ &= \frac{1}{2nN}\frac{N_1 - n_1}{N_1(N_1 - 1)}\left(s - N_1 s_1\right) \end{aligned}$$

o que mostra que $\sigma^2(\mu^*)$ é tanto menor quanto maior for s_1. Observe-se que s depende apenas da população, enquanto que o valor de s_1 resulta da maneira como os agregados são constituídos. Vê-se assim que os agregados devem ser tão heterogéneos quanto possível, por forma a maximizar s_1 e, consequentemente, a minimizar $\sigma^2(\mu^*)$.

Voltando à expressão (5.10),como estamos a realizar uma amostragem sem reposição, vemos que quando $n_1>1$

$$\sigma^{2*}(\mu^*) = \frac{1}{2n^2}\sum_{j=1}^{n}\sum_{j'=1}^{n}\frac{\Pi_j\Pi_{j'} - \Pi_{j,j'}}{\Pi_{j,j'}}\left(Y_j - Y_{j'}\right)^2 \tag{5.15}$$

é o estimador centrado de $\sigma^2(\mu^*)$. Quando n=1 vê-se que, devido a (5.11), este estimador toma valores negativos pelo que não se deve utilizar. Em certos casos dispõe-se de um modelo matemático para a população, o que nos permite estimar $\sigma^2(\mu^*)$ mesmo tomando $n_1=1$.

Comparando a amostragem estratificada com a amostragem para agregados, vemos que a principal diferença entre elas é que, na primeira todos os estratos são amostrados, enquanto que na segunda os próprios agregados são sujeitos a um procedimento de selecção. É este facto que conduz os princípios, para os dois métodos, em direcções opostas:

- Na amostragem estratificada, a variabilidade amostral está confinada ao interior dos estratos.

- Na amostragem para agregados., existe apenas variabilidade entre os agregados, uma vez que, todos eles, são inteiramente amostrados.

Um caso particular da amostragem para agregados é a amostragem sistemática. Este tipo de amostragem é recomendado e adoptado, na prática, devido à sua atractiva simplicidade e utilidade, particularmente no estudo de populações instáveis, onde a amostragem tradicional não é aplicável.

Consideremos uma população, a qual é disposta (fisicamente ou por meio de uma lista) numa sequência simples de $N=N_1.N_2$ indivíduos. O agregado com índice h, h=1,..., N_1 , é constituído pelos elementos da população com índices $i = h+(m-1)N_1$; $m = 1,\ldots, N_2$ (Quadro 1).

AGREGADO	COMPOSIÇÃO DO AGREGADO
1	1 , N_1+1 , $2N_1$+1 , ... , $(N_2-1)N_1$+1
. . .	
h	h , h +N_1 , h +$2N_1$, ... , h+$(N_2-1)N_1$
. . .	
N_1	N_1 , $2N_1$, $3N_1$, ... , N_2N_1

Quadro 1

Neste tipo de amostragem, toma-se, em geral, um único agregado para constituir a amostra. A probabilidade de seleccionar um agregado é $\frac{1}{N_1}$,ou seja, é a probabilidade com que cada membro do agregado é seleccionado para a amostra.

Para se escolher o agregado que irá compor a amostra, de entre os primeiros N_1 elementos da população é seleccionado um , aleatoriamente, com probabilidades de selecção iguais.(Nesta escolha podem-se utilizar tabelas de números aleatórios).

Consideremos que foi o p-ésimo elemento o seleccionado, tomando-o como ponto de partida, os restantes elementos, que vão compor amostra, são obtidos por progressão aritmética de razão N_1. Assim, a amostra consiste no agregado de índice p, ou seja, nos elementos nas posições p , $p + N_1$, $p + 2\,N_1$, ... , $p + (N_2\text{-}1)\,N_1$.

A conveniência deste tipo de amostragem recai no facto de a selecção do primeiro membro da amostra determinar automaticamente a amostra inteira.

Prova-se que se a população for autocorrelacionada, isto é, se os elementos próximos tiverem tendência a ser semelhantes, a amostragem sistemática conduz a bons resultados.

6. SUB-AMOSTRAGEM

Nas diversas técnicas de amostragem estudadas até aqui, considerámos sempre que, os elementos, que vão compor a amostra, devem ser retirados directamente da população. Factores económicos e outros, baseados na dificuldade ou impossibilidade de o estudo abranger, de igual forma, toda a população, desencorajam a aplicação destas técnicas. Estes factores conduzem à necessidade de seleccionar grupos de indivíduos em vez de indivíduos directamente da população. mas isso apenas é válido em alguns tipos de pesquisa.

Suponhamos que cada grupo da população, aos quais chamaremos unidades primárias , pode ser dividido num determinado número de sub-unidades. Se ao seleccionar uma amostra de n unidades primárias, as sub-unidades, dentro de cada unidades seleccionada, dão resultados semelhantes, parece ser pouco económico medi-las todas.

Uma prática comum é seleccionar e medir uma amostra de sub-unidades de uma qualquer unidades escolhida, ou seja, o conjunto de elementos a ser estudado é obtido por sub-amostragem.

A sub-amostragem tem uma grande variedade de aplicações, que vão bem para além do objectivo imediato dos estudos por amostragem. Sempre que qualquer processo envolva testes químicos, físicos ou biológicos, ele pode ser executado numa pequena quantidade de material, é portanto indiferente que os resultados obtidos sejam provenientes de uma sub-amostra de uma grande quantidade ou de uma amostra.

Quanto à escolha das unidades que vão ser observadas, tanto pode ser feita com probabilidades iguais como com probabilidades proporcionais à sua dimensão.

Várias regras se podem considerar para determinar as fracções de amostragem e sub-amostragem, e vários métodos de estimação estão disponíveis. As vantagens dos diferentes métodos dependem da natureza da população, dos custos e dos dados suplementares que estão à nossa disposição.

Apesar de o tipo de sub-amostragem mais usual ser para três níveis, este processo pode se estendido sucessivamente a mais um nível tanto quanto for necessário, se as sub-unidades anteriormente obtidas em vez de serem enumeradas completamente, forem elas também amostradas.

Tendo em conta que a teoria de sub-amostragem com probabilidades proporcionais existente, praticamente só se encontra desenvolvida para um número reduzido de níveis de fraccionamento, pretende-se obter uma teoria geral deste tipo de sub-amostragem, válida para qualquer número de níveis, partindo da generalização da sub-amostragem com probabilidades iguais.

6.1. RESULTADOS PRÉVIOS

Seja X uma variável aleatória cuja distribuição depende de uma condição C. Esta variável aleatória terá o valor médio condicional $\mu(X|C)$ e a variância condicional

$$\sigma^2(X|C) = \mu(X^2|C) - \mu^2(X|C). \tag{6.1}$$

O valor médio, não condicional de X, obtém-se descondicionalizando $\mu(X|C)$, vindo

$$\mu(X) = \mu_C(\mu(X|C)), \tag{6.2}$$

obtendo-se , analogamente, $\mu(X^2) = \mu_C(\mu(X^2|C))$, pelo que

$$\begin{aligned}
\sigma^2(X) &= \mu(X^2) - \mu^2(X) = \\
&= \mu_C(\mu(X^2|C)) - \mu_C^2(\mu(X|C)) = \\
&= \mu_C(\mu(X^2|C)) - \mu_C(\mu^2(X|C)) + \mu_C(\mu^2(X|C)) - \mu_C^2(\mu(X|C)) = \\
&= \mu_C(\mu(X^2|C) - \mu^2(X|C)) + \mu_C(\mu^2(X|C)) - \mu_C^2(\mu(X|C)) = \\
&= \mu_C(\sigma^2(X|C)) + \mu_C(\mu^2(X|C)) - \mu_C^2(\mu(X|C))
\end{aligned} \tag{6.3}$$

fazendo $Z = \mu(X|C)$ temos

$$\begin{aligned}
\mu_C(\sigma^2(X|C)) + \mu_C(\mu^2(X|C)) - \mu_C^2(\mu(X|C)) &= \mu_C(\sigma^2(X|C)) + \mu_C(Z^2) - \mu_C^2(Z) = \\
&= \mu_C(\sigma^2(X|C)) + \sigma_C^2(Z)
\end{aligned}$$

e portanto

$$\sigma^2(X) = \mu_C(\sigma^2(X|C)) + \sigma_C^2(\mu(X|C)) \tag{6.4}$$

6.2. NOTAÇÕES

De modo a ser possível a generalização da teoria da sub-amostragem para qualquer número de níveis, é necessário criar um terminologia de índices que permita referenciar, sem ambiguidades, a posição e génese de uma dada sub-população, qualquer que ela seja, isto é, qual o nível de estrutura a que essa sub-população pertence, dentro desse nível qual a posição que ocupa e quais as sub-populações, que por decomposições sucessivas, lhe deram origem.

Vamos considerar neste estudo que a sub-amostragem é efectuada em k níveis de fraccionamento.

Por forma a clarificar, à partida, como a população inicial é decomposta, sucessivamente, até ao nível k, recorreu-se à elaboração de um esquema que está representado na figura 2.

A identificação das sub-populações, dos vários níveis, a partir do primeiro é feita através da atribuição de vectores. Consideremos as sub-populações de nível h, com h=1,...,k . Podemos identificá-las por vectores $\vec{i}^{\,h}$ com h componentes, ou seja:

$$\vec{i}^{\,h} = (i_1, \cdots, i_h) \; ; \; h = 1, \ldots, k \; , \tag{6.5}$$

indicando as componentes de $\vec{i}^{\,h}$ as sucessivas opções que levam à escolha da população correspondente.

Consideremos ainda o índice p , $1 \leq p \leq h$, e representemos por $\vec{i}_p^{\,h}$ o vector obtido tomando as p primeiras componentes do vector $\vec{i}^{\,h}$, temos então:

$$\vec{i}^{\,h} = (\vec{i}_p^{\,h} | i_{p+1}, \cdots, i_h) \; ; \; 1 \leq p \leq h \; . \tag{6.6}$$

No caso em que temos p = h-1 vem

$$\vec{i}^{\,h} = (\vec{i}^{\,h-1} | i_h) \; ; \; 1 \leq p \leq h \; . \tag{6.7}$$

(Na situação particular de p=0 toma-se $\vec{i}_0^{\,h} = \vec{i}^{\,0}$).

Atendendo às expressões (6.5) e (6.7) as sub-populações de nível 1 serão identificadas pelo vector $\vec{i}^{\,1} = (i_1)$, as sub-populações de nível 2 pelo vector $\vec{i}^{\,2} = (i_1, i_2) = (\vec{i}^{\,1} | i_2)$, as de nível 3 pelo vector $\vec{i}^{\,3} = (i_1, i_2, i_3) = (\vec{i}^{\,2} | i_3)$ e assim sucessivamente até às populações de nível k que serão identificadas pelo vector $\vec{i}^{\,k} = (\vec{i}^{\,k-1} | i_k)$.

Deste modo, $P(\vec{i}^{\,1})$ representará qualquer sub-população de nível 1, $P(\vec{i}^{\,2})$ representará qualquer sub-população de nível 2,e assim sucessivamente até $P(\vec{i}^{\,k-1})$, que representará por sua vez qualquer sub-população de nível k-1, que finalmente, por decomposição, origina, directamente, os elementos. Por extensão de linguagem ,estes elementos serão considerados como uma sub-população, de nível k, $P(\vec{i}^{\,k})$, assim como a própria população inicial é considerada como uma sub-população de nível zero, isto é, $P(\vec{i}^{\,0})$.

Facilmente se verifica que, dada uma sub-população de nível h, ela está contida noutras sub-populações existentes em níveis anteriores a h, ou seja , com p<h, $P(\vec{i}^{\,h})$ é sub-população de $P(\vec{i}_p^{\,h})$.

Seja $X\,(X_h \; , \; h = 0,1,\ldots,k)$ o conjunto dos vectores representativos das sub-populações (de nível h , h=0,1,...,k) . Tem-se $X_0 = \{\vec{i}^{\,0}\}$ e o seu único vector corresponde à população.

Dado $\vec{i}^{\,h}$ $(h = 0,1,\ldots,k-1)$, $\Omega(\vec{i}^{\,h})$ representa o conjunto dos vectores $\vec{i}^{\,h+1}$ cujas primeiras h componentes são as componentes de $\vec{i}^{\,h}$. Tais vectores correspondem às sub-populações em que se decompõe uma determinada população genérica $P(\vec{i}^{\,h})$. Relembrando que

$\vec{i}_p^{\,h}$ representa os vectores obtidos tomando as p primeiras componentes de $\vec{i}^{\,h}$, no caso de p=h-1 tem-se $\vec{i}^{\,h} \in \Omega\left(\vec{i}_{h-1}^{\,h}\right)$.

Consideremos agora

$$\begin{cases} N\left(\vec{i}^{\,h}\right) = \#\left[P\left(\vec{i}^{\,h}\right)\right] \; ; \; h = 0,1,\ldots,k \\ N'\left(\vec{i}^{\,h}\right) = \#\left[\Omega\left(\vec{i}^{\,h}\right)\right] \; ; \; h = 0,1,\ldots,k-1 \end{cases}, \tag{6.8}$$

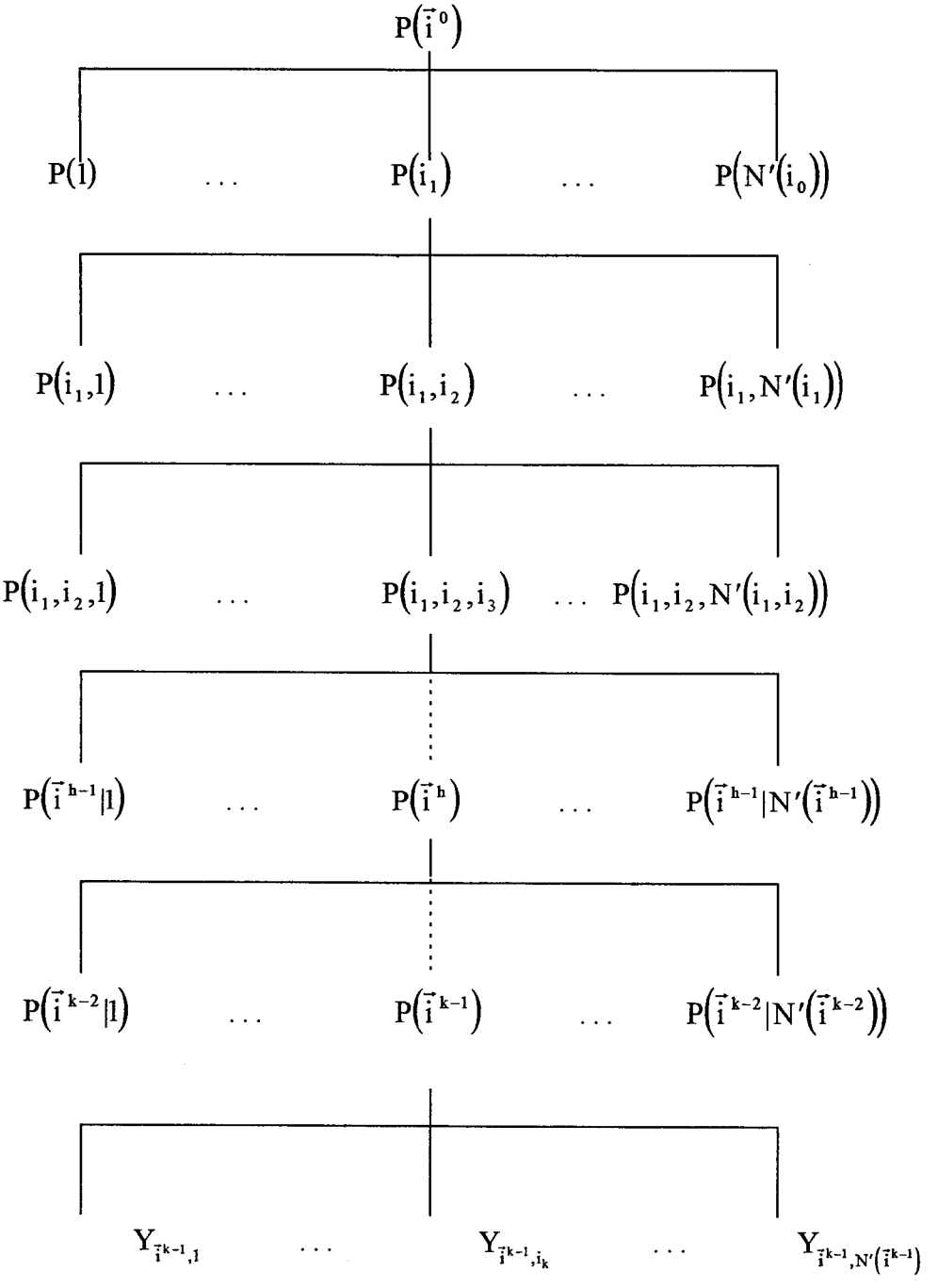


Figura 2 - Caso geral da decomposição de uma população em k níveis de fraccionamento.

ou seja, $N(\vec{i}^{\,h})$ representa a dimensão de $P(\vec{i}^{\,h})$, estando definida para h=0,1,...,k e $N'(\vec{i}^{\,h})$ dá o número de sub-populações em que $P(\vec{i}^{\,h})$ se decompõe. $N(\vec{i}^{\,h})$ está definida apenas para h=0,1,...,k-1 , visto a passagem do penúltimo para o último nível de estrutura, k-1 e k respectivamente, resultar numa decomposição directa de qualquer das sub-populações $P(\vec{i}^{\,k-1})$ nos respectivos elementos, $P(\vec{i}^{\,k})$. Então:

$$\begin{cases} N(\vec{i}^{\,k}) = 1 \\ N(\vec{i}^{\,k-1}) = N'(\vec{i}^{\,k-1}) \end{cases} \tag{6.9}$$

e

$$N(\vec{i}^{\,h}) = \sum_{\vec{i}^{\,h+1} \in \Omega(\vec{i}^{\,h})} N(\vec{i}^{\,h+1}) \ , \tag{6.10}$$

isto é, a dimensão de uma sub-população é igual à soma das dimensões das várias sub-populações resultantes da sua decomposição.

Atendendo a (6.10), obtemos a expressão

$$N(\vec{i}^{\,0}) = \sum_{\vec{i}^{\,1} \in \Omega(\vec{i}^{\,0})} N(\vec{i}^{\,1}), \tag{6.11}$$

que mostra que a dimensão da população, $N(\vec{i}^{\,0})$, pode ser dada em função das dimensões, $N(\vec{i}^{\,1})$, das sub-populações de nível 1.

Se quisermos reportar a dimensão da população ao nível k (último nível, sendo este constituído por elementos), é válida a seguinte expressão:

$$N(\vec{i}^{\,0}) = \sum_{\vec{i}^{\,1} \in X_1} \sum_{\vec{i}^{\,2} \in X_2} \cdots \sum_{\vec{i}^{\,k-2} \in X_{k-2}} \sum_{\vec{i}^{\,k-1} \in \Omega(\vec{i}^{\,k-2})} N(\vec{i}^{\,k-1}) \ . \tag{6.12}$$

As expressões seguintes são apresentadas com o objectivo de melhor clarificar o significado de $\mathbf{X}_h$ e de $\Omega(\vec{i}^{\,h})$.

$$
(6.13)\quad
\begin{cases}
\mathbf{X}_0 = \left\{\vec{i}^{\,0}\right\} \\
\mathbf{X}_1 = \left\{\vec{i}^{\,1}\right\} = \left\{ \left(i_1\right) \right\} = \Omega\left(\vec{i}^{\,0}\right) = \left\{1,\cdots,N'\left(\vec{i}^{\,0}\right)\right\} \\
\mathbf{X}_2 = \left\{\vec{i}^{\,2}\right\} = \left\{ \left(i_1,i_2\right) \right\} = \bigcup_{\vec{i}^{\,1}\in X_1} \Omega\left(\vec{i}^{\,1}\right) = \bigcup_{\vec{i}^{\,1}\in X_1}^{N'\left(\vec{i}^{\,0}\right)} \left\{\left(\vec{i}^{\,1}|1\right),\ldots,\left(\vec{i}^{\,1}|N'\left(\vec{i}^{\,1}\right)\right)\right\} \\
\mathbf{X}_3 = \left\{\vec{i}^{\,3}\right\} = \left\{ \left(i_1,i_2,i_3\right) \right\} = \bigcup_{\vec{i}^{\,2}\in X_2} \Omega\left(\vec{i}^{\,2}\right) = \bigcup_{\vec{i}^{\,2}\in X_2}^{N'\left(\vec{i}^{\,1}\right)} \left\{\left(\vec{i}^{\,2}|1\right),\ldots,\left(\vec{i}^{\,2}|N'\left(\vec{i}^{\,2}\right)\right)\right\} \\
\vdots \\
\mathbf{X}_{h+1} = \left\{\vec{i}^{\,h+1}\right\} = \left\{ \left(i_1,\ldots,i_{h+1}\right) \right\} = \bigcup_{\vec{i}^{\,h}\in X_h} \Omega\left(\vec{i}^{\,h}\right) = \bigcup_{\vec{i}^{\,h}\in X_h}^{N'\left(\vec{i}^{\,h-1}\right)} \left\{\left(\vec{i}^{\,h}|1\right),\ldots,\left(\vec{i}^{\,h}|N'\left(\vec{i}^{\,h}\right)\right)\right\} \\
\vdots \\
\mathbf{X}_k = \left\{\vec{i}^{\,k}\right\} = \left\{ \left(i_1,\ldots,i_k\right) \right\} = \bigcup_{\vec{i}^{\,k-1}\in X_{k-1}} \Omega\left(\vec{i}^{\,k-1}\right) = \bigcup_{\vec{i}^{\,k-1}\in X_{k-1}}^{N'\left(\vec{i}^{\,k-2}\right)} \left\{\left(\vec{i}^{\,k-1}|1\right),\ldots,\left(\vec{i}^{\,k-1}|N'\left(\vec{i}^{\,k-1}\right)\right)\right\}
\end{cases}
$$

À semelhança do que foi feito anteriormente para as sub-populações, iremos agora criar uma notação para as amostras.

Dado existirem k níveis de fraccionamento, a colheita da amostra é feita em k etapas. Em cada uma destas etapas colhem-se sub-populações, do nível correspondente, de entre as contidas nas sub-populações escolhidas na etapa precedente.

Representaremos pois, as sub-populações escolhidas por $P\left(\vec{j}^{\,h}\right)$ sendo $n\left(\vec{j}^{\,h}\right), T\left(\vec{j}^{\,h}\right), \overline{Y}\left(\vec{j}^{\,h}\right)$ respectivamente o número, o total e a média das observações colhidas em $P\left(\vec{j}^{\,h}\right)$. Observe-se que $T\left(\vec{j}^{\,k}\right)$ e $\overline{Y}\left(\vec{j}^{\,k}\right)$ coincidem com a única observação $Y\left(\vec{j}^{\,k}\right)$ colhida em $P\left(\vec{j}^{\,k}\right)$. Os vectores $\vec{j}^{\,h}$ são aleatórios, podendo ocorrer mais de uma vez, se se estiver a amostrar com reposição.

Representaremos ainda por $\mathbf{X}_h^*$ o conjunto dos vectores $\vec{j}^{\,h}$ e por $A\left(\vec{i}^{\,h}\right)$ o acontecimento que se verifica quando $P\left(\vec{i}^{\,h}\right)$ é escolhida. Observe-se que, com p<h , $A\left(\vec{i}^{\,h}\right)$ implica $A\left(\vec{i}_p^{\,h}\right)$.

Importa ainda referir que, as várias amostragens feitas na mesma etapa são independentes.

6.3. SUB-AMOSTRAGEM COM PROBABILIDADES IGUAIS SEM REPOSIÇÃO

Admitamos que a população se pode decompor em $N'\left(\vec{i}^{\,0}\right)$ sub-populações, de nível 1, das quais a i_1-ésima sub-população contem $N'\left(\vec{i}^{\,1}\right)$ sub-populações, de nível 2, e assim sucessivamente até se chegar à sub-população de nível k-1 que é composta por $N'\left(\vec{i}^{\,k-2}\right)$ elementos.

Consideremos o acontecimento $A\left(\vec{i}^{\,h}\right)$, com h<k, que se verifica quando a população $P\left(\vec{i}^{\,h}\right)$ é escolhida.

Na fase, da sub-amostragem, em que se verifica $A\left(\vec{i}^{\,h}\right)$, são escolhidas, com probabilidades iguais e sem reposição, $n'\left(\vec{i}^{\,h}\right)$ sub-populações, de nível h+1, sub-populações estas que estão contidas em $P\left(\vec{i}^{\,h}\right)$.

Como é obvio, a obtenção das sub-populações de nível h+1 depende, directamente, da sub-população, de nível h, que lhe deu origem, assim, temos as probabilidades condicionais:

$$\Pi\left(\vec{i}^{\,h+1}\middle|\vec{i}_h^{\,h+1}\right)=P\left[A\left(\vec{i}^{\,h+1}\right)\middle|A\left(\vec{i}_h^{\,h+1}\right)\right]=\frac{n'\left(\vec{i}_h^{\,h+1}\right)}{N'\left(\vec{i}_h^{\,h+1}\right)} \quad . \tag{6.14}$$

Tendo em atenção o que se referiu para o nível h+1, facilmente se vê que o mesmo se verifica para qualquer nível. Logo tem-se:

$$A\left(\vec{i}^{\,h+1}\right)\Rightarrow A\left(\vec{i}_h^{\,h+1}\right)\Rightarrow\cdots\Rightarrow A\left(\vec{i}_p^{\,h+1}\right) \text{ , com } p=0,\ldots,h \text{ ,}$$

isto é, a ocorrência de um acontecimento, num dado nível, resulta da ocorrência de um acontecimento do nível anterior, sucessivamente até ao nível zero.

Obtemos então

$$\Pi\left(\vec{i}^{\,h+1}\right)=\Pi\left(\vec{i}^{\,h+1}\middle|\vec{i}_h^{\,h+1}\right)\cdots\Pi\left(\vec{i}_{p+1}^{\,h+1}\middle|\vec{i}_p^{\,h+1}\right)\Pi\left(\vec{i}_p^{\,h+1}\right)=\frac{n'\left(\vec{i}_h^{\,h+1}\right)}{N'\left(\vec{i}_h^{\,h+1}\right)}\cdots\frac{n'\left(\vec{i}_p^{\,h+1}\right)}{N'\left(\vec{i}_p^{\,h+1}\right)}\Pi\left(\vec{i}_p^{\,h+1}\right), \tag{6.15}$$

representando $\Pi\left(\vec{i}^{\,h}\right)$ a probabilidade de $A\left(\vec{i}^{\,h}\right)$. Como $\Pi\left(\vec{i}^{\,0}\right)=1$, podemos escrever

$$\Pi\left(\vec{i}^{\,h+1}\right)=\prod_{p=0}^{h}\frac{n'\left(\vec{i}_p^{\,h+1}\right)}{N'\left(\vec{i}_p^{\,h+1}\right)} \tag{6.16}$$

tendo-se ainda

$$\Pi\left(\vec{i}^{\,h+1}\right)=\Pi\left(\vec{i}_h^{\,h+1}\right)\frac{n'\left(\vec{i}_h^{\,h+1}\right)}{N'\left(\vec{i}_h^{\,h+1}\right)} \tag{6.17}$$

6.3.1. CONSTRUÇÃO DO ESTIMADOR

Tendo em conta que, cada sub-população, de qualquer nível, apenas pode ser escolhida uma vez, dado que estamos a amostrar sem reposição, obtém-se o seguinte estimador para valor médio:

$$\mu^{*}=\frac{1}{N\left(\vec{j}^{\,0}\right)}\sum_{j_1=1}^{n'\left(\vec{j}^{\,0}\right)}\sum_{j_2=1}^{n'\left(\vec{j}^{\,1}\right)}\cdots\sum_{j_k=1}^{n'\left(\vec{j}^{\,k-1}\right)}\frac{Y\left(j_1,j_2,\ldots,j_k\right)}{\Pi\left(j_1,j_2,\ldots,j_k\right)}=\frac{1}{N\left(\vec{j}^{\,0}\right)}\sum_{\vec{j}^{\,k}\in X_k}\frac{Y\left(\vec{j}^{\,k}\right)}{\Pi\left(\vec{j}^{\,k}\right)} \tag{6.18}$$

onde $N(\vec{j}^{\,0})$ representa a dimensão da população.

De modo a simplificar a notação, representaremos, a partir daqui, a dimensão da população, $N(\vec{j}^{\,0})$, apenas por N.

Os vectores que representam as sub-populações de nível h+1, escolhidas entre as contidas em $P(\vec{j}^{\,h})$, pertencem a $\Omega(\vec{j}^{\,h}) \cap X^*_{h+1}$ Usando (6.17), tem-se pois:

$$(6.19) \qquad \mu^* = \frac{1}{N} \sum_{\vec{j}^{\,k-1} \in X^*_{k-1}} \frac{1}{\Pi(\vec{j}^{\,k-1})} \frac{N'(\vec{j}^{\,k-1})}{n'(\vec{j}^{\,k-1})} \sum_{j^k \in \Omega(\vec{j}^{\,k-1}) \cap X^*_k} Y(\vec{j}^{\,k}) = \frac{1}{N} \sum_{\vec{j}^{\,k-1} \in X^*_{k-1}} \frac{N'(\vec{j}^{\,k-1})}{\Pi(\vec{j}^{\,k-1})} \overline{Y}(\vec{j}^{\,k-1}) .$$

Considerando que as sub-populações, pertencentes a qualquer nível h (h=0,1,...,k-1), se subdividem todas em igual número de sub-populações do nível seguinte, isto é,

$$(6.20) \qquad \begin{cases} N'(\vec{i}^{\,h}) = N'_h \;\; ; \;\; h = 0,1,\ldots,k-1 \\ n'(\vec{i}^{\,h}) = n'_h \;\; ; \;\; h = 0,1,\ldots,k-1 \end{cases} ,$$

é então possível obter uma expressão simplificada do estimador μ^*.

Assim, sendo n a dimensão da amostra,

$$(6.21) \qquad \begin{cases} N = \prod_{h=0}^{k-1} N'_h \\ n = \prod_{h=0}^{k-1} n'_h \end{cases} ,$$

logo, de (6.16), (6.20) e (6.21) vem $\Pi(\vec{i}^{\,k}) = \frac{n}{N}$, e portanto o estimador para o valor médio vem:

$$(6.22) \qquad \mu^* = \frac{1}{N} \sum_{\vec{j}^{\,k} \in X^*_k} \frac{N}{n} Y(\vec{j}^{\,k}) = \overline{Y} ,$$

onde $\overline{Y}$ é a média global da amostra.

6.3.2. VARIÂNCIA DO ESTIMADOR

Como foi visto, em (6.4), a variância de X é a soma do valor médio da variância condicional com a variância do valor médio condicional.

Quando se chega ao nível k (k-ésima etapa da amostragem) a condição C corresponde aos resultados dos k-1 níveis anteriores pelo que a representamos por C_{k-1}, então:

(6.23) $$\sigma^2(\mu^*) = \mu_{C_{k-1}}\left[\sigma^2(\mu^*|C_{k-1})\right] + \sigma^2_{C_{k-1}}\left[\mu(\mu^*|C_{k-1})\right].$$

Utilizando os símbolos μ_h e σ^2_h para indicar o valor médio e a variância, relativos à etapa h, sendo

(6.24) $$\begin{cases} \sigma^2(\mu^*|C_{k-1}) = \sigma^2_k(\mu^*) \\ \mu(\mu^*|C_{k-1}) = \mu_k(\mu^*) \end{cases},$$

como podemos considerar $\mu_{(k)}(\mu^*)$ uma variável aleatória, cuja distribuição depende de C_{k-2}, vem

(6.25) $$\sigma^2_{C_{k-1}}\left[\mu_k(\mu^*)\right] = \mu_{C_{k-2}}\left[\sigma^2(\mu_k(\mu^*)|C_{k-2})\right] + \sigma^2_{C_{k-2}}\left[\mu(\mu_k(\mu^*)|C_{k-2})\right].$$

Analogamente ao que fizemos anteriormente vamos pôr

(6.26) $$\begin{cases} \sigma^2\left[\mu_k(\mu^*)|C_{k-2}\right] = \sigma^2_{k-1}\left[\mu_k(\mu^*)\right] \\ \mu\left[\mu_k(\mu^*)|C_{k-2}\right] = \mu_{k-1}\left[\mu_k(\mu^*)\right] \end{cases}$$

obtendo-se assim

(6.27) $$\sigma^2(\mu^*) = \mu_{C_{k-1}}\left[\sigma^2_k(\mu^*) + \mu_{C_{k-2}}\left(\sigma^2_{k-1}(\mu_k)\right)\right] + \sigma^2_{C_{k-2}}\left[\mu_{k-1}\mu_k(\mu^*)\right]$$

Pode ainda concluir-se que:

(6.28) $$\mu_{C_h}(w) = \mu_1(\ldots\mu_h(w)) \;\; ; \;\; h = 1,\ldots,k.$$

Representando por $\nabla_h \;\; ; \;\; h = 1,\ldots,k$, o cálculo dos valores médios, para as etapas com índices $h' \neq h$ e da variância para a etapa com índice h, obtém-se:

(6.29) $$\sigma^2(\mu^*) = \sum_{h=1}^{k} \nabla_h(\mu^*),$$

ou seja, a variância do estimador $\sigma^2(\mu^*)$ é composta por tantas componentes quantas as etapas da sub-amostragem.

Note-se que, a validade deste resultado se verifica para qualquer técnica de sub-amostragem.

Começando por calcular a componente $\nabla_k(\mu^*)$, usando (6.19) e (2.40) temos então:

(6.30) $$\begin{aligned} \sigma^2_k(\mu^*) &= \frac{1}{N^2} \sum_{\vec{j}^{k-1} \in X^*_{k-1}} \frac{N'^2\left(\vec{j}^{k-1}\right)}{\Pi^2\left(\vec{j}^{k-1}\right)} \sigma^2\left(\overline{Y}\left(\vec{j}^{k-1}\right)\right) = \\ &= \frac{1}{N^2} \sum_{\vec{j}^{k-1} \in X^*_{k-1}} \frac{N'^2\left(\vec{j}^{k-1}\right)}{\Pi^2\left(\vec{j}^{k-1}\right)} \frac{\sigma^2\left(\vec{j}^{k-1}\right)}{n'\left(\vec{j}^{k-1}\right)} \frac{N\left(\vec{j}^{k-1}\right) - n'\left(\vec{j}^{k-1}\right)}{N\left(\vec{j}^{k-1}\right) - 1} \end{aligned}$$

onde

(6.31) $$\sigma^2\left(\vec{j}^{k-1}\right)=\frac{1}{N\left(\vec{j}^{k-1}\right)}\left[\sum_{\vec{i}^{k}\in\Omega\left(\vec{j}^{k-1}\right)} Y^2\left(\vec{i}^{k}\right)-\frac{T^2\left(\vec{j}^{k-1}\right)}{N\left(\vec{j}^{k-1}\right)}\right]$$

é a variância da sub-população $P\left(\vec{j}^{k-1}\right)$, visto que os vectores, que correspondem aos seus elementos, são vectores de $\Omega\left(\vec{j}^{k-1}\right)$, onde a característica em estudo toma os valores $Y\left(\vec{i}^{k}\right)$ com $\vec{i}^{k}\in\Omega\left(\vec{j}^{k-1}\right)$.

Como os vectores representativos das sub-populações de nível k-1, escolhidos entre as contidas em $P\left(\vec{j}^{k-2}\right)$, são os vectores que pertencem a $\Omega\left(\vec{j}^{k-2}\right)\cap X^{*}_{k-1}$ e como, devido a (6.17), se tem:

(6.32) $$\Pi\left(\vec{i}^{h+1}\right)=\Pi\left(\vec{i}^{h}\right)\frac{n'\left(\vec{i}^{h}\right)}{N'\left(\vec{i}^{h}\right)}\ ;\ \text{com}\ \vec{i}^{h+1}\in\Omega\left(\vec{i}^{h}\right),$$

podemos escrever (6.30) na forma

(6.33)

$$\sigma_k^2\left(\mu^*\right)=\frac{1}{N^2}\sum_{\vec{j}^{k-2}\in X^{*}_{k-2}}\frac{1}{\Pi^2\left(\vec{j}^{k-2}\right)}\frac{N'\left(\vec{j}^{k-2}\right)}{n'\left(\vec{j}^{k-2}\right)}\cdot\sum_{\vec{j}^{k-1}\in\Omega\left(\vec{j}^{k-2}\right)\cap X_{k-1}}\frac{\dfrac{N'^2\left(\vec{j}^{k-1}\right)\sigma^2\left(\vec{j}^{k-1}\right)\left[N\left(\vec{j}^{k-1}\right)-n'\left(\vec{j}^{k-1}\right)\right]}{n'\left(\vec{j}^{k-1}\right)\left[N\left(\vec{j}^{k-1}\right)-1\right]}}{\dfrac{n'\left(\vec{j}^{k-2}\right)}{N'\left(\vec{j}^{k-2}\right)}}.$$

Para calcularmos os sucessivos valores médios, que acabarão por dar $\nabla_k\left(\mu^*\right)$, procederemos sempre da mesma maneira, utilizando (3.11) e (6.32). Assim temos:

(6.34)

$$\mu_{k-1}\left(\sigma_k^2\left(\mu^*\right)\right)=$$

$$=\frac{1}{N^2}\sum_{\vec{j}^{k-2}\in X^{*}_{k-2}}\frac{1}{P^2\left(\vec{j}^{k-2}\right)}\frac{N'\left(\vec{j}^{k-2}\right)}{n'\left(\vec{j}^{k-2}\right)}\cdot\sum_{\vec{i}^{k-1}\in\Omega\left(\vec{j}^{k-2}\right)}N^2\left(\vec{i}^{k-1}\right)\frac{s^2\left(\vec{i}^{k-1}\right)}{n'\left(\vec{i}^{k-1}\right)}\frac{N\left(\vec{i}^{k-1}\right)-n'\left(\vec{i}^{k-1}\right)}{N\left(\vec{i}^{k-1}\right)-1}=$$

$$= \frac{1}{N^2} \sum_{\vec{j}^{k-3} \in X^*_{k-3}} \frac{1}{P^2\left(\vec{j}^{k-3}\right)} \frac{N'\left(\vec{j}^{k-3}\right)}{n'\left(\vec{j}^{k-3}\right)} \sum_{\vec{j}^{k-2} \in X_{k-2} \cap \Omega\left(\vec{j}^{k-3}\right)} \frac{N\left(\vec{j}^{k-2}\right)}{n'\left(\vec{j}^{k-2}\right)} \cdot$$

$$\cdot \sum_{\vec{i}^{k-1} \in \Omega\left(\vec{j}^{k-2}\right)} \frac{N^2\left(\vec{i}^{k-1}\right) \frac{s^2\left(\vec{i}^{k-1}\right)}{n'\left(\vec{i}^{k-1}\right)} \frac{N\left(\vec{i}^{k-1}\right) - n'\left(\vec{i}^{k-1}\right)}{N\left(\vec{i}^{k-1}\right) - 1}}{\frac{n'\left(\vec{j}^{k-3}\right)}{N'\left(\vec{j}^{k-3}\right)}} \cdot$$

Analogamente tem-se

$$\mu_{k-2}\left[\mu_{k-1}\left(\sigma_k^2\left(\mu^*\right)\right)\right] = \frac{1}{N^2} \sum_{\vec{j}^{k-3} \in X^*_{k-3}} \frac{1}{\Pi^2\left(\vec{j}^{k-3}\right)} \frac{N'\left(\vec{j}^{k-3}\right)}{n'\left(\vec{j}^{k-3}\right)} \sum_{\vec{i}^{k-2} \in \Omega\left(\vec{j}^{k-3}\right)} \frac{N'\left(\vec{i}^{k-2}\right)}{n'\left(\vec{i}^{k-2}\right)} \cdot$$

(6.35)

$$\cdot \sum_{\vec{i}^{k-1} \in \Omega\left(\vec{i}^{k-2}\right)} N^2\left(\vec{i}^{k-1}\right) \frac{\sigma^2\left(\vec{i}^{k-1}\right)}{n'\left(\vec{i}^{k-1}\right)} \frac{N\left(\vec{i}^{k-1}\right) - n'\left(\vec{i}^{k-1}\right)}{N\left(\vec{i}^{k-1}\right) - 1}$$

e portanto

$$\nabla_k\left(\mu^*\right) = \mu_1\left[\ldots \mu_{k-1}\left(\sigma_k^2\left(\mu^*\right)\right)\right] = \frac{N'\left(\vec{i}^{0}\right)}{N^2 n'\left(\vec{i}^{0}\right)} \sum_{\vec{i}^{1} \in \mathbf{X}_1} \frac{N'\left(\vec{i}^{1}\right)}{n'\left(\vec{i}^{1}\right)} \sum_{\vec{i}^{2} \in \Omega\left(\vec{i}^{1}\right)} \frac{N'\left(\vec{i}^{2}\right)}{n'\left(\vec{i}^{2}\right)} \cdot \ldots$$

(6.36)

$$\ldots \cdot \sum_{\vec{i}^{k-2} \in \Omega\left(\vec{i}^{k-3}\right)} \frac{N'\left(\vec{i}^{k-2}\right)}{n'\left(\vec{i}^{k-2}\right)} \cdot \sum_{\vec{i}^{k-1} \in \Omega\left(\vec{i}^{k-2}\right)} N^2\left(\vec{i}^{k-1}\right) \frac{\sigma^2\left(\vec{i}^{k-1}\right)}{n'\left(\vec{i}^{k-1}\right)} \frac{N\left(\vec{i}^{k-1}\right) - n'\left(\vec{i}^{k-1}\right)}{N\left(\vec{i}^{k-1}\right) - 1}$$

Calculemos agora uma das $\nabla_h\left(\mu^*\right)$ com $1 < h < k$.

De (6.19) vem

$$\mu_k\left(\mu^*\right) = \frac{1}{N} \sum_{\vec{j}^{k-1} \in X^*_{k-1}} \frac{N'\left(\vec{j}^{k-1}\right)}{\Pi\left(\vec{j}^{k-1}\right)} \mu\left[\overline{Y}\left(\vec{j}^{k-1}\right)\right] = \frac{1}{N} \sum_{\vec{j}^{k-1} \in X^*_{k-1}} \frac{T\left(\vec{j}^{k-1}\right)}{\Pi\left(\vec{j}^{k-1}\right)} \tag{6.37}$$

visto que

$$N'\left(\vec{j}^{k-1}\right)\mu\left(\overline{Y}\left(\vec{j}^{k-1}\right)\right) = N\left(\vec{j}^{k-1}\right)\mu\left(\overline{Y}\left(\vec{j}^{k-1}\right)\right) = T\left(\vec{j}^{k-1}\right). \tag{6.38}$$

Portanto

$$\mu_{k-1}\left(\mu_k\left(\mu^*\right)\right) = \mu_{k-1}\left(\frac{1}{N} \sum_{\vec{j}^{k-1} \in \mathbf{X}^*_{k-1}} \frac{T\left(\vec{j}^{k-1}\right)}{\Pi\left(\vec{j}^{k-1}\right)}\right) = \tag{6.39}$$

$$= \frac{1}{N}\mu_{k-1}\left(\sum_{\vec{j}^{k-2}\in X^*_{k-2}} \frac{1}{\Pi\left(\vec{j}^{k-2}\right)} \sum_{\vec{j}^{k-1}\in X^*_{k-1}\cap\Omega\left(\vec{j}^{k-2}\right)} \frac{T\left(\vec{j}^{k-1}\right)}{\frac{n'\left(\vec{j}^{k-2}\right)}{N'\left(\vec{j}^{k-2}\right)}}\right) =$$

$$= \frac{1}{N}\sum_{\vec{j}^{k-2}\in X^*_{k-2}} \frac{1}{\Pi\left(\vec{j}^{k-2}\right)}\mu_{k-1}\left(\sum_{\vec{j}^{k-1}\in X^*_{k-1}\cap\Omega\left(\vec{j}^{k-2}\right)} \frac{T\left(\vec{j}^{k-1}\right)}{\Pi\left(\vec{j}^{k-2}\right)}\right) =$$

$$= \frac{1}{N}\sum_{\vec{j}^{k-2}\in X^*_{k-2}} \frac{1}{\Pi\left(\vec{j}^{k-2}\right)} \sum_{\vec{j}^{k-1}\in\Omega\left(\vec{j}^{k-2}\right)} T\left(\vec{j}^{k-1}\right) =$$

$$= \frac{1}{N}\sum_{\vec{j}^{k-2}\in X^*_{k-2}} \frac{T\left(\vec{j}^{k-2}\right)}{\Pi\left(\vec{j}^{k-2}\right)}$$

e

$$(6.40)\qquad \mu_{h+1}\left(\cdots\mu_k\left(\mu^*\right)\right) = \frac{1}{N}\sum_{\vec{j}^{h}\in X^*_{h}} \frac{T\left(\vec{j}^{h}\right)}{\Pi\left(\vec{j}^{h}\right)} = \frac{1}{N}\sum_{\vec{j}^{h-1}\in X^*_{h-1}} \frac{1}{\Pi\left(\vec{j}^{h-1}\right)} \frac{N'\left(\vec{j}^{h-1}\right)}{n'\left(\vec{j}^{h-1}\right)} \sum_{\vec{j}^{h}\in X^*_{h}\cap\Omega\left(\vec{j}^{h-1}\right)} T\left(\vec{j}^{h}\right)$$

Como

$$(6.41)\qquad \sigma^2\left[\frac{1}{n'\left(\vec{j}^{h-1}\right)} \sum_{\vec{j}^{h}\in X^*_{h}\cap\Omega\left(\vec{j}^{h-1}\right)} T\left(\vec{j}^{h}\right)\right] = \frac{\sigma^2\left(\vec{j}^{h-1}\right)}{n'\left(\vec{j}^{h-1}\right)} \frac{N'\left(\vec{j}^{h-1}\right) - n'\left(\vec{j}^{h-1}\right)}{N'\left(\vec{j}^{h-1}\right) - 1},$$

visto tratar-se da média de uma amostra de totais de sub-populações, obtida sem reposição e com probabilidades iguais, tem-se ainda:

$$(6.42)\qquad \sigma^2\left(\vec{j}^{h-1}\right) = \frac{1}{N'\left(\vec{j}^{h-1}\right)}\left[\sum_{\vec{i}^{h}\in\Omega\left(\vec{j}^{h-1}\right)} T^2\left(\vec{i}^{h}\right) - \frac{T^2\left(\vec{j}^{h-1}\right)}{N'\left(\vec{j}^{h-1}\right)}\right]$$

logo

$$(6.43)\qquad \sigma^2_h\left[\mu_{h+1}\left(\cdots\mu_k\left(\mu^*\right)\right)\right] = \frac{1}{N^2}\sum_{\vec{j}^{h-1}\in X^*_{h-1}} \frac{1}{\Pi^2\left(\vec{j}^{h-1}\right)} \frac{N'^2\left(\vec{j}^{h-1}\right)\sigma^2\left(\vec{j}^{h-1}\right)\left[N'\left(\vec{j}^{h-1}\right) - n'\left(\vec{j}^{h-1}\right)\right]}{n'\left(\vec{j}^{h-1}\right)\left[N'\left(\vec{j}^{h-1}\right) - 1\right]},$$

Raciocinando como atrás, obtemos:

(6.44)

$$\nabla_h(\mu^*) = \frac{N'(\vec{i}^0)}{N^2 n'(\vec{i}^0)} \sum_{\vec{i}^1 \in X_1} \frac{N'(\vec{i}^1)}{n'(\vec{i}^1)} \sum_{\vec{i}^2 \in \Omega(\vec{i}^1)} \frac{N'(\vec{i}^2)}{n'(\vec{i}^2)} \cdot \ldots \cdot$$

$$\cdot \sum_{\vec{i}^{h-1} \in \Omega(\vec{i}^{h-2})} N^2(\vec{i}^{h-1}) \frac{\sigma^2(\vec{i}^{h-1})}{n'(\vec{i}^{h-1})} \frac{N'(\vec{i}^{h-1}) - n'(\vec{i}^{h-1})}{N'(\vec{i}^{h-1}) - 1}$$

com $h = 2, \ldots, k$.

Finalmente para obtermos $\nabla_1(\mu^*)$ vemos que se tem, atendendo à expressão (6.40),

(6.45)
$$\mu_2(\cdots \mu_k(\mu^*)) = \frac{1}{N} \sum_{\vec{j}^1 \in X^*} \frac{1}{\Pi(\vec{j}^0)} \frac{N'(\vec{i}^0)}{n'(\vec{i}^0)} \sum_{\vec{j}^1 \in X_i^*} T(\vec{j}^1),$$

pelo que, raciocinando como para obter (6.44) e atendendo a que $\vec{j}^0 = \vec{i}^0$ e que $\Pi(\vec{i}^0) = 1$, vem

(6.46)
$$\nabla_1(\mu^*) = \frac{1}{N^2} N'^2(\vec{i}^0) \frac{\sigma^2(\vec{i}^0)}{n'(\vec{i}^0)} \frac{N'(\vec{i}^0) - n'(\vec{i}^0)}{N'(\vec{i}^0) - 1}$$

com $\sigma^2(\vec{i}^0) = \frac{1}{N'(\vec{i}^0)} \left[\sum_{\vec{i}^1 \in X_i} T^2(\vec{i}^0) - \frac{T^2(\vec{i}^0)}{N'(\vec{i}^0)} \right]$, onde $\Omega(\vec{i}^0) = \mathbf{X}_1$.

6.4. SUB-AMOSTRAGEM COM PROBABILIDADES PROPORCIONAIS

Nesta técnica as sub-amostras da etapa h, h=1,...,k , têm dimensões n'_h , $h = 1, \ldots, k$, sendo obtidas com reposição e com probabilidades proporcionais, salvo na última etapa em que a selecção é feita sem reposição e com probabilidades iguais.

Tendo-se escolhido $P(\vec{j}^h)$, com $h = 0, \ldots, k-2$, as probabilidades das sub-populações $P(\vec{i}^{h+1})$, com $\vec{i}^{h+1} \in \Omega(\vec{i}^h)$,serem escolhidas, em cada tiragem, são $\frac{N(\vec{i}^{h+1})}{N(\vec{i}^h)}$. Facto que determina o nome adoptado para esta técnica.

Seja agora $\Pi(\vec{i}^h)$ o valor médio da variável aleatória $X(\vec{i}^h)$, que nos dá o número de vezes que $P(\vec{i}^h)$é escolhida. Observe-se que dado estarmos a trabalhar com reposição, salvo na última etapa, se $X(\vec{i}^h) > 1$ e se $h < k-1$ então colhem-se $X(\vec{i}^h)$ sub-amostras em $P(\vec{i}^h)$.

Se $\vec{i}^{h+1} \in \Omega(\vec{i}^h)$, o número de vezes que $P(\vec{i}^{h+1})$ é escolhida, é a soma dos números de vezes que correspondem às várias sub-amostras colhidas em $P(\vec{i}^h)$. Sendo $X(\vec{i}^{h+1}, m)$,

com $m = 1,\ldots, X\left(\vec{i}^{\,h}\right)$, as variáveis aleatórias que dão o número de vezes que $P\left(\vec{i}^{\,h+1}\right)$ é escolhida, teremos

$$X\left(\vec{i}^{\,h+1}\right) = \sum_{m=1}^{X\left(\vec{i}^{\,h}\right)} X\left(\vec{i}^{\,h+1}, m\right) \; ; \; h = 1,\ldots, k-1. \tag{6.47}$$

Sabendo que

$$\mu\left[X\left(\vec{i}^{\,h+1}, m\right)\right] = n'_{h+1} \frac{N\left(\vec{i}^{\,h+1}\right)}{N\left(\vec{i}^{\,h}\right)} \; ; \quad h = 1,\ldots, k-1 \, , \tag{6.48}$$

e que, dado o valor tomado por $X\left(\vec{i}^{\,h}\right)$ poder ser interpretado como uma condição de que depende a distribuição de $X\left(\vec{i}^{\,h+1}\right)$, com $\vec{i}^{\,h+1} \in \Omega\left(\vec{i}^{\,h}\right)$, teremos portanto

(6.49)

$$\begin{aligned}
\Pi\left(\vec{i}^{\,h+1}\right) &= \mu\left[X\left(\vec{i}^{\,h+1}\right)\right] = \sum_t P\left[X\left(\vec{i}^{\,h}\right) = t\right] \cdot \mu\left[X\left(\vec{i}^{\,h+1}\right) | X_m = t\right] = \\
&= \sum_t P\left[X\left(\vec{i}^{\,h}\right) = t\right] \cdot \mu\left[\sum_{m=1}^{t} X\left(\vec{i}^{\,h+1}, m\right)\right] = \sum_t P\left[X\left(\vec{i}^{\,h}\right) = t\right] \cdot t \cdot n'_{h+1} \frac{N\left(\vec{i}^{\,h+1}\right)}{N\left(\vec{i}^{\,h}\right)} \\
&= n'_{h+1} \frac{N\left(\vec{i}^{\,h+1}\right)}{N\left(\vec{i}^{\,h}\right)} \mu\left[X\left(\vec{i}^{\,h}\right)\right] = n'_{h+1} \frac{N\left(\vec{i}^{\,h+1}\right)}{N\left(\vec{i}^{\,h}\right)} \Pi\left(\vec{i}^{\,h}\right), \quad \text{com} \quad h = 1,\ldots, k-1 \, .
\end{aligned}$$

Por outro lado, utilizando os resultados apresentados anteriormente, obtém-se

$$\Pi\left(\vec{i}^{\,1}\right) = n'_1 \frac{N\left(\vec{i}^{\,1}\right)}{N\left(\vec{i}^{\,0}\right)} . \tag{6.50}$$

De (6.49) e (6.50), vem

$$\Pi\left(\vec{i}^{\,h+1}\right) = n'_1 \cdot \ldots \cdot n'_{h+1} \frac{N\left(\vec{i}^{\,h+1}\right)}{N\left(\vec{i}^{\,0}\right)} \; ; \; h = 1,\ldots, k-1 . \tag{6.51}$$

Observemos que a dimensão da amostra é dada por

$$n = n'_1 \cdot \ldots \cdot n'_k \tag{6.52}$$

logo

$$\Pi\left(\vec{i}^{\,k}\right) = \frac{n}{N\left(\vec{i}^{\,0}\right)} . \tag{6.53}$$

6.4.1. CONSTRUÇÃO DO ESTIMADOR

Os resultados anteriores permitem construir a seguinte expressão para o estimador do valor médio:

$$\mu^* = \frac{1}{N(\vec{i}^0)} \sum_{\vec{j}^k \in X_k^*} \frac{Y(\vec{j}^k)}{\Pi(\vec{j}^k)} = \frac{1}{n} \sum_{\vec{j}^k \in X_k^*} Y(\vec{j}^k) = \overline{Y}, \tag{6.54}$$

que indica que, o cálculo do estimador se reduz à determinação da média aritmética da amostra.

Relembremos, contudo, que esta generalização, para qualquer número (k) de níveis, só é possível desde que se satisfaçam as seguintes condições:

- Em todas as etapas, excepto a última, a amostragem é com probabilidades proporcionais às dimensões das sub-populações;

- Na última etapa, a amostragem é com probabilidades iguais.

6.4.2. VARIÂNCIA DO ESTIMADOR

Tal como sucede no caso da sub-amostragem com probabilidades iguais, a variância do estimador é dada por uma decomposição, em tantas componentes quantas as etapas da sub-amostragem, isto é,

$$\sigma^2(\overline{Y}) = \sigma^2(\mu^*) = \sum_{h=1}^{k} \nabla_{(h)}(\mu^*) \tag{6.55}$$

onde $\nabla_{(h)}(\mu^*)$; $h = 1,\ldots,k$ é a componente da etapa h (h = 1,...,k) , para $\sigma^2(\mu^*)$.

Observe-se agora que, com $n_{k-1} = n'_1 \cdot \ldots \cdot n'_{k-1}$, se tem

$$\mu^* = \frac{1}{n} \sum_{\vec{j}^k \in X_k^*} Y(\vec{j}^k) = \frac{1}{n_{k-1}} \sum_{\vec{j}^{k-1} \in X_{k-1}^*} \frac{1}{n'_k} \sum_{\vec{j}^k \in \Omega(\vec{j}^{k-1}) \cap X_k^*} Y(\vec{j}^k) = \frac{1}{n_{k-1}} \sum_{\vec{j}^{k-1} \in X_{k-1}^*} \overline{Y}(\vec{j}^{k-1}) \tag{6.56}$$

onde $\overline{Y}(\vec{j}^{k-1})$ representa a média das observações, tomadas em $P(\vec{j}^{h+1})$. Como as sub-amostras da última etapa são colhidas independentemente uma das outras sem reposição, as $\overline{Y}(\vec{j}^{k-1})$ são independentes.

Adaptando a este caso a expressão (2.40), a variância da média das observações é dada por

$$\sigma^2\left[\overline{Y}(\vec{j}^{k-1})\right] = \frac{\sigma^2(\vec{j}^{k-1})}{n'_k} \frac{N(\vec{j}^{k-1}) - n'_k}{N(\vec{j}^{k-1}) - 1}, \tag{6.57}$$

onde $\sigma^2\left(\vec{j}^{k-1}\right)$ continua a representar a variância da sub-população $P\left(\vec{j}^{k-1}\right)$, logo

$$(6.58) \qquad \sigma_k^2\left(\mu^*\right) = \frac{1}{n_{k-1}^2} \sum_{\vec{j}^{k-1} \in X_{k-1}^*} \sigma^2\left[\overline{Y}\left(\vec{j}^{k-1}\right)\right] = \frac{1}{n_{k-1}^2} \sum_{\vec{j}^{k-1} \in X_{k-1}^*} \frac{\sigma^2\left(\vec{j}^{k-1}\right)}{n'_k} \frac{N\left(\vec{j}^{k-1}\right) - n'_k}{N\left(\vec{j}^{k-1}\right) - 1} ,$$

então, recorrendo à expressão (3.11) utilizada na amostragem , com probabilidades variáveis, com reposição, vem:

(6.59)

$$\mu_{k-1}\left[\sigma_k^2\left(\mu^*\right)\right] = \mu_{k-1}\left(\frac{1}{n_{k-1}^2} \sum_{\vec{j}^{k-2} \in X_{k-2}^*} \sum_{\vec{j}^{k-1} \in X_{k-1} \cap \Omega\left(\vec{j}^{k-2}\right)} \frac{\sigma^2\left(\vec{j}^{k-1}\right)}{n'_k} \frac{N\left(\vec{j}^{k-1}\right) - n'_k}{N\left(\vec{j}^{k-1}\right) - 1}\right) =$$

$$= \frac{1}{n_{k-1}^2} \sum_{\vec{j}^{k-2} \in X_{k-2}^*} \mu_{k-1}\left[\sum_{\vec{j}^{k-1} \in X_{k-1} \cap \Omega\left(\vec{j}^{k-2}\right)} \frac{\sigma^2\left(\vec{j}^{k-1}\right)}{n'_k} \frac{N\left(\vec{j}^{k-1}\right) - n'_k}{N\left(\vec{j}^{k-1}\right) - 1}\right] =$$

$$= \frac{1}{n_{k-1}^2} \sum_{\vec{j}^{k-2} \in X_{k-2}^*} \sum_{\vec{i}^{k-1} \in \Omega\left(\vec{j}^{k-2}\right)} \frac{n'_{k-1} N\left(\vec{i}^{k-1}\right)}{N\left(\vec{j}^{k-2}\right)} \frac{\sigma^2\left(\vec{i}^{k-1}\right)}{n'_k} \frac{N\left(\vec{i}^{k-1}\right) - n'_k}{N\left(\vec{i}^{k-1}\right) - 1} .$$

Tendo em atenção a expressão(6.48), o valor médio da variável aleatória que dá o número de vezes que $P\left(\vec{j}^{h+1}\right)$ é escolhida, cada vez que se amostra para $P\left(\vec{j}^{k-2}\right)$ com $\vec{j}^{k-2} \in \Omega\left(\vec{j}^{k-1}\right)$, é

$$(6.60) \qquad \mu\left[X\left(\vec{j}^{k-1}, m\right)\right] = n'_{k-1} \frac{N\left(\vec{j}^{k-1}\right)}{N\left(\vec{j}^{k-2}\right)} \quad ; \quad m = 1, \ldots, X\left(\vec{j}^{k-2}\right) .$$

Continuando a raciocinar desta forma obtemos

(6.61)

$$\nabla_k\left(\mu^*\right) = \mu_1\left[\ldots \sigma_k^2\left(\mu^*\right)\right] = \frac{1}{n_{k-1}^2} \sum_{\vec{i}^1 \in X_1} \frac{n'_1 N\left(\vec{i}^1\right)}{N\left(\vec{i}^0\right)} \sum_{\vec{i}^2 \in \Omega\left(\vec{i}^1\right)} \frac{n'_2 N\left(\vec{i}^2\right)}{N\left(\vec{i}^1\right)} \ldots$$

$$\ldots \sum_{\vec{i}^{k-2} \in \Omega\left(\vec{i}^{k-3}\right)} \frac{n'_{k-2} N\left(\vec{i}^{k-2}\right)}{N\left(\vec{i}^{k-3}\right)} \sum_{\vec{i}^{k-1} \in \Omega\left(\vec{i}^{k-2}\right)} \frac{n'_{k-1} N\left(\vec{i}^{k-1}\right)}{N\left(\vec{i}^{k-2}\right)} \frac{\sigma^2\left(\vec{i}^{k-1}\right)}{n'_k} \frac{N\left(\vec{i}^{k-1}\right) - n'_k}{N\left(\vec{i}^{k-1}\right) - 1} =$$

$$= \frac{1}{nN\left(\vec{i}^0\right)} \sum_{\vec{i}^{k-1} \in X_{k-1}} N\left(\vec{i}^{k-1}\right) \sigma^2\left(\vec{i}^{k-1}\right) \frac{N\left(\vec{i}^{k-1}\right) - n'_k}{N\left(\vec{i}^{k-1}\right) - 1} .$$

Procuremos agora obter a expressão das $\nabla_h\left(\mu^*\right)$, $h = 2, \ldots, k-1$.

De (6.38) vem que $\mu\left[\overline{Y}\left(\vec{j}^{k-1}\right)\right]=\dfrac{T\left(\vec{j}^{k-1}\right)}{N\left(\vec{j}^{k-1}\right)}$ é o valor médio de $P\left(\vec{j}^{h+1}\right)$, obtém-se então:

(6.62)
$$\mu_k\left(\mu^*\right)=\frac{1}{n_{k-1}}\sum_{\vec{j}^{k-1}\in X^*_{k-1}}\frac{T\left(\vec{j}^{k-1}\right)}{N\left(\vec{j}^{k-1}\right)}.$$

Recorrendo novamente à expressão (3.11) vem

(6.63)
$$\mu_{k-1}\left[\mu_k\left(\mu^*\right)\right]=\mu_{k-1}\left[\frac{1}{n'_1\ldots n'_{k-1}}\sum_{\vec{j}^{k-2}\in X^*_{k-2}}\sum_{\vec{j}^{k-1}\in X^*_{k-1}\cap\Omega\left(\vec{j}^{k-2}\right)}\frac{T\left(\vec{j}^{k-1}\right)}{N\left(\vec{j}^{k-1}\right)}\right]=$$
$$=\frac{1}{n'_1\ldots n'_{k-1}}\sum_{\vec{j}^{k-2}\in X^*_{k-2}}\mu_{k-1}\left[\sum_{\vec{j}^{k-1}\in X^*_{k-1}\cap\Omega\left(\vec{j}^{k-2}\right)}\frac{T\left(\vec{j}^{k-1}\right)}{N\left(\vec{j}^{k-1}\right)}\right]=$$
$$=\frac{1}{n'_1\ldots n'_{k-1}}\sum_{\vec{j}^{k-2}\in X^*_{k-2}}\sum_{\vec{i}^{k-1}\in\Omega\left(\vec{j}^{k-2}\right)}n'_{k-1}\frac{N\left(\vec{i}^{k-1}\right)}{N\left(\vec{j}^{k-2}\right)}\frac{T\left(\vec{i}^{k-1}\right)}{N\left(\vec{i}^{k-1}\right)}=$$
$$=\frac{1}{n'_1\ldots n'_{k-2}}\sum_{\vec{j}^{k-2}\in X^*_{k-2}}\frac{T\left(\vec{j}^{k-2}\right)}{N\left(\vec{j}^{k-2}\right)}$$

com $\sum_{\vec{i}^{k-1}\in\Omega\left(\vec{j}^{k-2}\right)}T(\vec{i}^{k-1})=T\left(\vec{i}^{k-2}\right)$, e também

(6.64)
$$\mu_{h+1}\left[\ldots\mu_k\left(\mu^*\right)\right]=\frac{1}{n'_1\ldots n'_h}\sum_{\vec{j}^h\in X^*_h}\frac{T\left(\vec{j}^h\right)}{N\left(\vec{j}^h\right)}=\frac{1}{n'_1\ldots n'_h}\sum_{\vec{j}^h\in X^*_h}\mu\left(\vec{j}^h\right)$$

onde $\mu\left(\vec{j}^h\right)$ é o valor médio de $P\left(\vec{j}^h\right)$. Atendendo à independência das sub-amostragens realizadas na mesma etapa, vem

(6.65)
$$\sigma^2_h\left[\mu_{h+1}\left[\ldots\mu_k\left(\mu^*\right)\right]\right]=\frac{1}{n'^2_1\ldots n'^2_{h-1}}\sum_{\vec{j}^{h-1}\in X^*_{h-1}}\frac{1}{n'_h}\sum_{\vec{j}^h\in X^*_h\cap\Omega\left(\vec{j}^{h-1}\right)}\mu\left(\vec{j}^h\right)=$$
$$=\frac{1}{n'^2_1\ldots n'^2_{h-1}}\sum_{\vec{j}^{h-1}\in X^*_{h-1}}\sigma^2\left[W\left(\vec{j}^{h-1}\right)\right]$$

onde

(6.66)
$$W\left(\vec{j}^{h-1}\right)=\frac{1}{n'_h}\sum_{\vec{j}^h\in X^*_h\cap\Omega\left(\vec{j}^{h-1}\right)}\mu\left(\vec{j}^h\right)$$

é a média de uma amostra com dimensão n'_h, de valores médios de sub-populações $P\left(\vec{i}^{\,h}\right)$, com $\vec{i}^{\,h} \in \Omega\left(\vec{i}^{\,h-1}\right)$. Como esta amostra é colhida com reposição e probabilidades iguais tem-se

$$\sigma^2\left[W\left(\vec{j}^{\,h-1}\right)\right] = \frac{1}{n'_h} \sum_{\vec{i}^{\,h} \in \Omega\left(\vec{j}^{\,h-1}\right)} \frac{N\left(\vec{i}^{\,h}\right)}{N\left(\vec{j}^{\,h-1}\right)} \left[\mu\left(\vec{i}^{\,h}\right) - \mu\left(\vec{j}^{\,h-1}\right)\right]^2 \tag{6.67}$$

vindo

$$\sigma_h^2\left[\mu_{h+1}\left[\ldots\mu_k\left(\mu^*\right)\right]\right] = \frac{1}{n'^2_1 \ldots n'^2_{h-1} \ldots n'_h} \sum_{\vec{j}^{\,h-1} \in X^*_{h-1}} \sum_{\vec{i}^{\,h} \in \Omega\left(\vec{j}^{\,h-1}\right)} \frac{N\left(\vec{i}^{\,h}\right)}{N\left(\vec{j}^{\,h-1}\right)} \left[\mu\left(\vec{i}^{\,h}\right) - \mu\left(\vec{j}^{\,h-1}\right)\right]^2 \tag{6.68}$$

Utilizando mais uma vez a expressão (3.11), temos

(6.69)

$$\mu_{h-1}\left[\sigma_h^2\left[\mu_{h+1}\left(\ldots\mu_k\left(\mu^*\right)\right)\right]\right] =$$

$$= \frac{1}{n'^2_1 \ldots n'^2_{h-1}} \sum_{\vec{j}^{\,h-2} \in X^*_{h-2}} \mu_{h-1}\left[\sum_{\vec{j}^{\,h-1} \in X^*_{h-1} \cap \Omega\left(\vec{j}^{\,h-2}\right)} \frac{1}{n'_h} \sum_{\vec{i}^{\,h} \in \Omega\left(\vec{j}^{\,h-1}\right)} \frac{N\left(\vec{i}^{\,h}\right)}{N\left(\vec{j}^{\,h-1}\right)} \left[\mu\left(\vec{i}^{\,h}\right) - \mu\left(\vec{j}^{\,h-1}\right)\right]^2\right] =$$

$$= \frac{1}{n'^2_1 \ldots n'^2_{h-1}} \sum_{\vec{j}^{\,h-2} \in X^*_{h-2}} \sum_{\vec{i}^{\,h-1} \in \Omega\left(\vec{j}^{\,h-2}\right)} \frac{n'_{h-1} N\left(\vec{i}^{\,h-1}\right)}{N\left(\vec{j}^{\,h-2}\right)} \frac{1}{n'_h} \sum_{\vec{i}^{\,k} \in \Omega\left(\vec{i}^{\,h-1}\right)} \frac{N\left(\vec{i}^{\,h}\right)}{N\left(\vec{i}^{\,h-1}\right)} \left[\mu\left(\vec{i}^{\,h}\right) - \mu\left(\vec{j}^{\,h-1}\right)\right]^2 =$$

$$= \frac{1}{n'^2_1 \ldots n'^2_{h-2}\, n'_{h-1}\, n'_h} \sum_{\vec{j}^{\,h-2} \in X^*_{h-2}} \frac{1}{N\left(\vec{j}^{\,h-2}\right)} \sum_{\vec{i}^{\,h-1} \in \Omega\left(\vec{j}^{\,h-2}\right)} \sum_{\vec{i}^{\,h} \in \Omega\left(\vec{i}^{\,h-1}\right)} N\left(\vec{i}^{\,h}\right) \left[\mu\left(\vec{i}^{\,h}\right) - \mu\left(\vec{i}^{\,h-1}\right)\right]^2$$

e, em seguida, com $n_h = n'_1 \ldots n'_h$, $h = 1, \ldots, k$ tem-se

$$\nabla_h\left(\mu^*\right) = \mu_1\left[\ldots\mu_{h-1}\left[\sigma_h^2\left(\mu_{h+1}\ldots\mu_k\left(\mu^*\right)\right)\right]\right] =$$

$$= \frac{1}{n'_h N\left(\vec{i}^{\,0}\right)} \sum_{\vec{i}^{\,h} \in X_h} N\left(\vec{i}^{\,h}\right) \left[\mu\left(\vec{i}^{\,h}\right) - \mu\left(\vec{i}^{\,h-1}\right)\right]^2 \quad , \; h = 2, \ldots, k \tag{6.70}$$

Para obtermos $\nabla_1\left(\mu^*\right)$, observamos que, devido a (6.59) e (6.61), se tem

$$\mu_2\left[\ldots\mu_k\left(\mu^*\right)\right] = \frac{1}{n'_1} \sum_{\vec{j}^{\,1} \in X_1} \mu\left(\vec{j}^{\,1}\right) = W\left(\vec{j}^{\,0}\right) \tag{6.71}$$

deduzindo-se então

(6.72) $$\nabla_1(\mu^*) = \sigma_1^2\left[\mu_2 \ldots \mu_k(\mu^*)\right] = \frac{1}{n'_1}\sum_{\vec{i}^1 \in X_1}\frac{N\left(\vec{i}^1\right)}{N\left(\vec{i}^0\right)}\left[\mu\left(\vec{i}^1\right) - \mu\left(\vec{i}^0\right)\right]^2 .$$

Se considerarmos

(6.73) $$\begin{cases} A_h = \dfrac{1}{N\left(\vec{i}^0\right)}\displaystyle\sum_{\vec{i}^h \in X_h} N\left(\vec{i}^h\right)\left[\mu\left(\vec{i}^h\right) - \mu\left(\vec{i}^{h-1}\right)\right]^2 \;\; ; \; h = 1, \ldots, k-1 \\ A_k = \dfrac{1}{N\left(\vec{i}^0\right)}\displaystyle\sum_{\vec{i}^{h-1} \in X_k} N\left(\vec{i}^{k-1}\right)\sigma^2\left(\vec{i}^{k-1}\right)\frac{N\left(\vec{i}^{k-1}\right) - n'_k}{N\left(\vec{i}^{k-1}\right) - 1} \end{cases}$$

obtemos, atendendo a (6.50), (6.56), (6.65) e (6.67),

(6.74) $$\sigma^2(\mu^*) = \sum_{h=1}^{k}\frac{A_h}{n_h} = \sum_{h=1}^{k}\frac{A_h}{\prod_{m=1}^{h} n'_m}$$

6.5. CUSTO TOTAL

Nesta secção procura-se resolver o problema da escolha de $n'_1, n'_2, \ldots, n'_k$, para um custo total dado, adoptando para o mesmo uma expressão da forma

(6.75) $$C = C_0 + \sum_{h=1}^{k} C_h n_h = C_0 + \sum_{h=1}^{k} C_h \prod_{m=1}^{h} n'_m .$$

Dado C_0 ser um custo fixo, a obtenção do valor óptimo para n'_m, m=1,...,k, não depende do seu valor. Pode pois considerar-se C_0=0.

Assim, de (6.74) e (6.75), usando a desigualdade de Cauchy, temos

(6.76) $$\left(\sum_{h=1}^{k}\frac{A_h}{\prod_{m=1}^{h} n'_m}\right)\left(\sum_{h=1}^{k} C_h \prod_{m=1}^{h} n'_m\right) \geq \left(\sqrt{\sum_{h=1}^{k} A_h C_h}\right)^2 \geq \sum_{h=1}^{k} A_h C_h ,$$

atingindo-se o mínimo quando

(6.77) $$\frac{\dfrac{A_h}{\prod_{m=1}^{h} n'_m}}{C_h \prod_{m=1}^{h} n'_m} = k \text{ (constante)},$$

ou seja,

(6.78) $$\left(\prod_{m=1}^{h} n'_m\right)^2 = \frac{A_h}{C_h k}.$$

Como

(6.79) $$\prod_{m=1}^{h} n'_m = n_h$$

temos

(6.80) $$\frac{n_{h+1}}{n_h} = n'_{h+1}.$$

Substituindo (6.79) em (6.78) obtém-se

(6.81) $$n_h^2 = \frac{A_h}{C_h k}$$

e também

(6.82) $$n_{h+1}^2 = \frac{A_{h+1}}{C_{h+1} k}.$$

Atendendo a (6.80), usando (6.81) e (6.82), temos

(6.83) $$n'^2_{h+1} = \frac{\frac{A_{h+1}}{kC_{h+1}}}{\frac{A_h}{kC_h}} = \frac{A_{h+1}}{C_{h+1}} \frac{C_h}{A_h}.$$

Então o custo mínimo é atingido com

(6.84) $$n'_{h+1} = \sqrt{\frac{A_{h+1}}{C_{h+1}} \frac{C_h}{A_h}} \quad , \quad h = 1, \ldots, k-1.$$

BIBLIOGRAFIA

(1) Hansen, M. H. & Hurwitz, W. N. - The problem of non-response in sample survey, J.A.S.A., n.91 p.517 , 1948

(2) Hardy, G.H. , Littlewood , J.E. & Polya - Inequalities - Cambridge University Press , Cambridge , 1934

(3) Kendall, M. ; Stuart, A. & Ord, J.K. - The Advanced Theory of Statistics - Vol III , Charles Griffins & Co , London , 1983

(4) Kish , L. & Hess , I. - On variance of ratios and their differences in multistage samples - J.A.S.A., n.54 p.416 , 1954

(5) Kish , L. & Anderson,D.W. - Multivariate and multipropose stratification , J.A.S.A. , n.73 p.24 - 1978

(6) Kokan, A.R. & Khan,S. - Optimum allocation in multivariate surveys: na analytical solution, J.R. Stat. Soc. , B29 - 1967

Das referências bibliográficas apresentadas, destaca-se a indicada em (3) pois serviu de base a este estudo.

Da bibliografia existente sobre amostragem, terão especial interesse os artigos indicados em (1) , (4) , (5) e (6) .

É também referido, em (2) o livro onde vem desenvolvido o estudo da desigualdade de Cauchy