Estimação de uma razão

Até aqui tem-se considerado a estimação de uma **única característica da população** --- usando uma **amostragem aleatória simples sem reposição**.

Mas, muitas vezes pretende-se obter informação sobre várias características --- recolhem-se dados multivariados.

Considere-se o **caso bivariado**, que permite nalgumas circunstâncias aproveitar a estrutura de correlação para melhorar os estimadores.

Consideremos a amostra aleatória constituída por
n pares de valores (X_i, Y_i) obtida por amostragem aleatória simples.

Suponhamos que, para a população de N indivíduos pretendemos estimar a razão

$$\boxed{R = X_T / Y_T = \mu_X / \mu_Y .}$$

Para isso dispomos então de uma amostra com os valores $(x_1, y_1) \ldots (x_n, y_n)$

sendo o **estimador de *R*** **mais usado**

$$\boxed{R^* = \overline{X} / \overline{Y}}$$

Prova-se que, no caso de grandes amostras, $\boxed{R^*}$ é assintoticamente normal com valor médio e variância assintóticos assim definidos:

$$\boxed{E\left[R^*\right] \cong R = \mu_X / \mu_Y ;}$$

$$\boxed{Var\left[R^*\right] \cong \frac{1-f}{n\overline{y}^2} \sum_{1}^{N} \frac{(X_i - RY_i)^2}{N-1} = \frac{1-f}{n\overline{y}^2}\left[\sigma'^2_X - 2R\sigma'_{XY} + R^2\sigma'^2_Y\right]}$$

Uma estimativa de $\boxed{Var\left[R^*\right]}$ é

$$\boxed{s'^2\left[R^*\right]=\frac{1-f}{n\bar{y}^2}\sum_1^n\frac{(x_i-r^*y_i)^2}{n-1}}$$

$$r^*=\bar{x}/\bar{y}$$

Para grandes amostras um intervalo a (1-α)100% de confiança para *R* é

$$\boxed{r^*-z_{\alpha/2}s'(R^*)<R<r^*+z_{\alpha/2}s'(R^*)}$$

Acontece por vezes que ao estudarmos duas características para cada unidade de amostragem, para uma delas é **conhecido o total dos valores** dessa característica.

Seja então $R = X_T / Y_T = \mu_X / \mu_Y$ e suponhamos que $\boxed{Y_T}$ é conhecido.

Neste caso é possível estimar o valor médio, μ_X

$\boxed{\mu_X = R\mu_Y}$, usando o **estimador de razão,** assim definido

$$\boxed{\overline{X}_R = \frac{\overline{X}}{\overline{Y}}\mu_Y = R^*\mu_Y}$$

O estimador $\boxed{\overline{X}_R}$ é assintoticamente centrado e para grandes amostras tem-se

$$\boxed{Var\left[\overline{X}_R\right] \cong \frac{1-f}{n}\sum_{1}^{N}\frac{(X_i - RY_i)^2}{N-1} = \frac{1-f}{n}\left[\sigma'^2_X - 2R\sigma'_{XY} + R^2\sigma'^2_Y\right]}$$

Uma pergunta natural:

--Em que circunstâncias será o estimador da razão preferível ao estimador habitual da média?

Isto é

Será $\overline{X}_R$ mais ou menos eficiente do que $\overline{X}$?

Em que condições $\boxed{Var\left[\overline{X}_R\right] < Var\left[\overline{X}\right]?}$

Ora tem-se

$$\frac{1-f}{n}\left[\sigma'^2_X - 2R\sigma'_{XY} + R^2\sigma'^2_Y\right] < \frac{1-f}{n}\sigma'^2_X$$

$$\Downarrow$$

$$2R\rho\sigma'_X\ \sigma'_Y > R^2\sigma'^2_Y$$

$$\Downarrow$$

$$\rho > \frac{R}{2}\frac{\sigma'_Y}{\sigma'_X} \Rightarrow \rho > \frac{1}{2}\frac{CV_Y}{CV_X},$$

onde *CV* designa coeficiente de variação .

Notas finais:

- O estimador da razão $\overline{X}_R$ resulta num estimador mais eficiente se ρ_{XY} for suficientemente elevado, mas…
- … se $CV_Y > 2CV_X$, então o ganho na eficiência só se verificava se $\rho_{XY} > 1$(isto não pode ser pois não ???!!!!).

  Portanto, nestas condições mesmo existindo uma relação linear perfeita entre X e Y, $\overline{X}_R$ não pode ser mais eficiente do que $\overline{X}$

Dois factores importantes para o aumento da eficiência dos estimadores da razão

--- a variabilidade da variável Y não pode ser muito maior que a de X

--- o coeficiente de correlação ρ_{XY} tem que ser positivo e elevado

Amostragem Estratificada

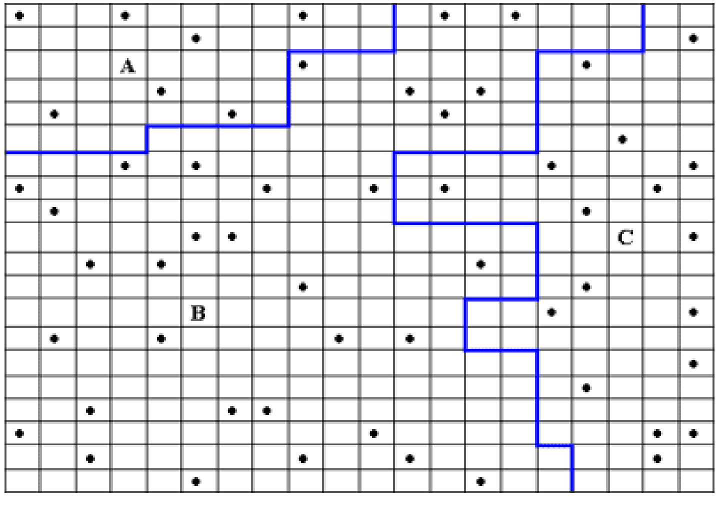
Exemplo

Suponhamos que se divide uma região em tipos (ou classes) de arbustos para se estudar uma dada característica desses arbustos.

Qual a vantagem?

Veremos que desta forma se pode ter estimativas mais precisas do valor médio da população do que ao usar a amostragem aleatória simples. Há também vantagem em dirigir o estudo para cada subpopulação ou estrato – modos diferentes de lidar com cada tipo de arbusto….

-- também pode haver questões de custos envolvidos…



Amostragem Estratificada – conceitos

População dividida em subpopulações ou **estratos.**

São várias as **razões** que levam a estratificar a população:

-- oferece maior garantia de representatividade;
-- permite obter estimativas com uma dada precisão para a variável de interesse em cada estrato;
-- permite resolver os problemas inerentes a cada estrato e que podem diferir de estrato para estrato;
-- a estratificação permite um aumento de precisão nas estimativas; essa precisão é tanto maior quanto mais homogéneos forem os estratos;
-- conveniências administrativas de organização do trabalho de recolha da informação.

População finita com N indivíduos

$a_1, \ldots, a_N$ os valores de uma dada característica para aqueles indivíduos.

População é dividida em k grupos ou **estratos** de dimensões conhecidas: $N_1, \ldots, N_k \quad (\sum N_i = N)$

Estrato	dimensão	elementos	valor médio	variância
S_1	N_1	$a_{11}a_{12}\cdots a_{1N_1}$	μ_1	$\sigma_1'^2$
S_2	N_2	$a_{21}a_{22}\cdots a_{2N_2}$	μ_2	$\sigma_2'^2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
S_k	N_k	$a_{k1}a_{k2}\cdots a_{kN_k}$	μ_k	$\sigma_k'^2$

$$N = \sum_{i=1}^{k} N_i$$

Valor médio

$$\boxed{\mu = \frac{1}{N}\sum_{i=1}^{k} N_i \mu_i = \sum_{i=1}^{k} W_i \mu_i}$$

Variância da população

$$\boxed{\sigma'^2 = \frac{1}{N-1}\left\{\sum_{1}^{k}(N_i - 1)\sigma_i'^2 + \sum_{1}^{k} N_i(\mu_i - \mu)^2\right\}}$$

onde $W_i = {}^{N_i}\!/_{N}$ é o "peso" em cada estrato.

De facto tem-se

$$\sigma'^2 = \frac{1}{N-1}\sum_{i,j}\left(a_{ij}-\mu\right)^2 = \frac{1}{N-1}\sum_{i=1}^{k}\sum_{j=1}^{N_i}\left(a_{ij}-\mu\right)^2 = \frac{1}{N-1}\sum_{i}\left[\sum_{j=1}^{N_i}\left(a_{ij}-\mu_i+\mu_i-\mu\right)^2\right]$$

$$= \frac{1}{N-1}\sum_{i}\left[\sum_{j=1}^{N_i}\left(a_{ij}-\mu_i\right)^2 + N_i\left(\mu_i-\mu\right)^2 + 2\left(\mu_i-\mu\right)\underbrace{\sum_{j=1}^{N_i}\left(a_{ij}-\mu_i\right)}_{=0}\right]$$

$$\boxed{\sigma'^2 = \frac{1}{N-1}\left[\sum_{i=1}^{k}(N_i-1)\sigma_i'^2 + \sum_{i=1}^{k} N_i(\mu_i-\mu)^2\right]}.$$

Para cada estrato *i* tem-se

$$\boxed{\mu_i = \frac{1}{N_i}\sum_{j=1}^{N_i} a_{ij} \qquad \text{e} \qquad \sigma_i'^2 = \frac{1}{N_i - 1}\sum_{j=1}^{N_i}\left(a_{ij} - \mu_i\right)^2}$$

A amostragem aleatória estratificada consiste em tirar de cada estrato uma amostra aleatória de tamanho pré-fixado:

$$\boxed{n_1, n_2, \ldots, n_k} \qquad \left(\sum_{i}^{k} n_i = n\right)$$

tendo como <u>elementos em cada estrato i</u>

$$\boxed{x_{i1}, x_{i2}, \ldots, x_{in_i}}$$

A média e a variância do i-ésimo estrato são:

$$\boxed{\bar{x}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij} \quad \text{e} \quad s_i'^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}\left(x_{ij} - \bar{x}_i\right)^2}$$

A $\boxed{f_i = \frac{n_i}{N_i}}$ chama-se **fracção de amostragem** em cada estrato.

Há dois problemas que se colocam neste tipo de amostragem:

1- Como se divide a população em estratos.

2- Qual o número de elementos a escolher em cada estrato? **Afectação**.

Destes dois problemas o mais simples é o segundo e é esse que começaremos a tratar.

Fixada a dimensão da amostra a recolher, seja $\underline{n}$, um dos modos que à primeira vista parece mais razoável consiste em seleccionar em cada estrato um número de elementos proporcional à dimensão do estrato, i.e.,

$$\frac{n_i}{n} = \frac{N_i}{N} \quad \text{donde} \quad n_i = n\frac{N_i}{N}$$

Verifica-se portanto que

$$\boxed{f_i = \frac{n_i}{N_i} = \frac{n}{N}}$$

É habitual designar esta afectação por **afectação proporcional.**

Estimação do valor médio

O estimador do valor médio é a <u>média empírica estratificada</u> assim definida

$$\boxed{\overline{X}_{st} = \sum_{i=1}^{k} W_i \overline{X}_i = \sum_{i=1}^{k} \frac{N_i \overline{X}_i}{N}}$$

A média empírica estratificada $\neq$ média aritmética,

$$\overline{X}_{st} = \sum_{i=1}^{k} W_i \overline{X}_i = \sum_{i=1}^{k} \frac{N_i \overline{X}_i}{N} \quad \neq \quad \overline{X}' = \sum_{i=1}^{k} \frac{n_i \overline{X}_i}{n}$$

O primeiro é um estimador centrado, enquanto o segundo não é.

$$E\left[\overline{X}_{st}\right] = \sum_{1}^{k} W_i \mu_i = \mu \quad \text{enquanto} \quad E\left[\overline{X}'\right] = \frac{1}{n} \sum_{1}^{k} n_i \mu_i \neq \mu$$

$\overline{X}'$ só será estimador centrado se $\frac{n_i}{n} = \frac{N_i}{N}$, ou seja, no caso da afectação ser proporcional.

Vejamos agora

$$\boxed{Var\left[\overline{X}_{st}\right] = \sum_{1}^{k} W_i^2 (1 - f_i) \sigma_i^2 / n_i}$$

Visto $\boxed{Var\left[\overline{X}_{st}\right] = \sum_{1}^{k} W_i^2 \; Var(\overline{X}_i)}$, pois $Cov(\overline{X}_i, \overline{X}_j) = 0$.

Observação:

Vimos que no caso da afectação proporcional $\overline{X}_{st}$ e $\overline{X}'$ coincidiam, no entanto estes dois estimadores **não apresentam a mesma variância**. Efectivamente

$$\boxed{Var\left[\overline{X}'\right]=\frac{1}{n^2}\sum_{1}^{k} n_i\,(1-f_i)\sigma_i^{'2}}.$$

Vejamos várias expressões para a variância, em certos casos particulares:

1. Se $f_i=\frac{n_i}{N_i}$ for desprezável $Var\left[\overline{X}_{st}\right]=\sum_{1}^{k} W_i^2\ \sigma_i^{'2}/n_i$;

2. Se $w_i=\frac{n_i}{n}=\frac{N_i}{N}$ -- afectação proporcional $Var\left[\overline{X}_{st}\right]=\frac{1-f}{n}\sum_{1}^{k} W_i\ \ \sigma_i^{'2}$;

3. Se a amostragem é proporcional e a variância é constante , i.e., $\sigma_i^{'2}=\sigma^{'2}$, então

$$Var\left[\overline{X}_{st}\right]=\frac{1-f}{n}\sigma^{'2}.$$

Estimação do Total da População

Um estimador centrado para o total X_T da população é

$$\boxed{X_T^* = N\,\overline{X}_{st} = \sum_1^k N_i \overline{X}_i}$$

Facilmente se verifica que se trata de um estimador centrado, sendo a sua variância dada por

$$\boxed{Var\left[X_T^*\right] = \sum_1^k N_i^2 (1 - f_i)\sigma_i'^2 / n_i}$$

Intervalos de Confiança

Um intervalo de confiança para μ a $(1-\alpha)100\%$ é

$$\boxed{\bar{x}_{st} - z_{\alpha/2} s'(\bar{x}_{st}) < \mu < \bar{x}_{st} + z_{\alpha/2} s'(\bar{x}_{st})}$$

e um intervalo de confiança para X_T a (1-α)100% é

$$\boxed{N\bar{x}_{st} - z_{\alpha/2} N s'(\bar{x}_{st}) < X_T < N\bar{x}_{st} + z_{\alpha/2} N s'(\bar{x}_{st})}$$

Se em cada estrato são recolhidas poucas observações o procedimento usual consiste em considerar $t_{\alpha/2}$ em vez de $z_{\alpha/2}$, sendo o número de graus de liberdade dado por

$$n = \frac{\left(\sum_{i=1}^{k} g_i s'^2_i\right)^2}{\sum_{i=1}^{k} g_i^2 s'^4_i \Big/ (n_i - 1)} \qquad \text{com} \qquad g_i = \frac{N_i (N_i - n_i)}{n_i}$$

Observação: Vejamos em que condições a amostragem estratificada é preferível à amostragem aleatória simples, i.e, em que condições

$$Var\left[\overline{X}_{st}\right] < Var\left[\overline{X}\right]$$

Temos:

$$Var\left[\overline{X}\right] = (1-f)\frac{\sigma'^2}{n}$$ e

$$Var\left[\overline{X}_{st}\right] = \sum_{i=1}^{k} W_i^2 (1-f_i)\frac{\sigma_i'^2}{n_i}.$$

Numa primeira fase consideremos que estamos no caso de **afectação proporcional,** $\boxed{f_i = f}$

$$Var\left[\overline{X}_{st}\right]=\frac{(1-f)}{n}\sum_{i=1}^{k}\frac{N_i}{N}\sigma_i'^2 \qquad Var\left[\overline{X}\right]-Var\left[\overline{X}_{st}\right]=\frac{1-f}{n}\left(\sigma'^2-\frac{1}{N}\sum_{i=1}^{k}N_i\sigma_i'^2\right)$$

vimos porém que

$$\sigma'^2=\frac{1}{N-1}\left(\sum_{i=1}^{k}(N_i-1)\sigma_i'^2+\sum_{i=1}^{k}N_i(\mu_i-\mu)^2\right)$$

Se o tamanho dos estratos é grande

$$\boxed{\frac{N_i-1}{N-1}\doteq\frac{N_i}{N}\doteq\frac{N_i}{N-1}} \qquad (*)$$

donde $\sigma'^2=\frac{1}{N}\left(\sum_{i=1}^{k}N_i\sigma_i'^2+\sum_{i=1}^{k}N_i(\mu_i-\mu)^2\right)$

$$\Rightarrow Var\left[\overline{X}\right]-Var\left[\overline{X}_{st}\right]=\frac{1-f}{Nn}\sum_{i=1}^{k}N_i(\mu_i-\mu)^2=\frac{1-f}{n}\sum_{i=1}^{k}W_i(\mu_i-\mu)^2>0$$

excepto se μ_i todos iguais.

Conclusão: o estimador da média na amostragem estratificada será sempre mais eficiente do que o estimador da média na amostragem aleatória simples, ou melhor, é tanto mais eficiente quanto maior for a variação nas médias dos estratos.

Porém, se acontece que os estratos não são suficientemente grandes que permitam que se verifique (*), deve considerar-se

$$\sigma'^2 = \frac{1}{N-1}\left(\sum_{i=1}^{k}(N_i - 1)\sigma_i'^2 + \sum_{i=1}^{k} N_i(\mu_i - \mu)^2\right)$$

$$\Rightarrow Var\left[\overline{X}\right] - Var\left[\overline{X}_{st}\right] = \frac{1-f}{n(N-1)}\left[\sum_{i=1}^{k} N_i(\mu_i - \mu)^2 - \frac{1}{N}\sum_{i=1}^{k}(N - N_i)\sigma_i'^2\right]$$

Sendo assim, podemos dizer que

$\overline{X}_{st}$ é mais eficiente do que $\overline{X}$ se

$$\boxed{\sum_{i=1}^{k} N_i (\mu_i - \mu)^2 > \frac{1}{N} \sum_{i=1}^{k} (N - N_i) \sigma_i'^2}$$

Informalmente pode dizer-se que quanto maior for a variabilidade entre os estratos e menor for a variabilidade dentro de cada estrato, maior será o ganho potencial ao considerar a amostra estratificada para estimar a média populacional.

Escolha óptima do tamanho da amostra a recolher em cada estrato

Objectivo:

Saber como escolher a dimensão da amostra de modo a satisfazer uma certa precisão ou um certo custo .

Suponhamos que no processo de amostragem há:

C_0 --- custo base da amostragem;

c_i --- custo de cada observação individual no estrato *i*.

Custo total C_T é dado por $\boxed{C_T = C_0 + \sum_1^k n_i c_i}$.

Que valores escolher para $n_1, n_2, \ldots, n_k$ de modo a:

a) minimizar $Var(\overline{X}_{st})$, para um custo total C_T;

b) minimizar o custo total, para um dado valor de $Var(\overline{X}_{st})$.

a) Variância mínima para custo fixo.

Pretendemos determinar $n_1, n_2, \ldots, n_k$ que minimize

$$\boxed{Var\left[\overline{X}_{st}\right] = \sum_{i=1}^{k} W_i^2 \frac{\sigma_i'^2}{n_i} - \frac{1}{N}\sum_{i=1}^{k} W_i \sigma_i'^2 \quad \text{sujeito a} \quad \sum_{i=1}^{k} c_i n_i = C_T - c_0}$$

Usando o método dos multiplicadores de Lagrange, temos a *Lagrangeana* assim definida

$$L = \sum_{i=1}^{k} W_i^2 \frac{\sigma_i'^2}{n_i} - \frac{1}{N}\sum_{i=1}^{k} W_i \sigma_i'^2 + \lambda\left(\sum_{i=1}^{k} c_i n_i - C_T + c_0\right)$$

Para se minimizar esta função teremos

$$\frac{\partial L}{\partial n_i} = -\sum_{i=1}^{k} W_i^2 \frac{\sigma_i'^2}{n_i^2} - \lambda \sum_{i=1}^{k} c_i = 0$$

$$\frac{\partial L}{\partial \lambda} = \sum_{i=1}^{k} c_i n_i - C_T + c_0 = 0$$

Da primeira equação tem-se

$$-\sum_{i=1}^{k}\left(W_i^2 \frac{\sigma_i'^2}{n_i^2} - \lambda c_i\right) = 0$$

onde para cada parcela se tem $\boxed{n_i \sqrt{\lambda} = \frac{W_i \ \sigma_i'}{\sqrt{c_i}}}$,

que multiplicando por c_i, dá $c_i n_i \sqrt{\lambda} = \sqrt{c_i} W_i \; \sigma_i'$
e efectuando a soma ao longo de todas os estratos:

$$(C_T - c_0)\sqrt{\lambda} = \sum_{i=1}^{k} \sqrt{c_i} W_i \; \sigma_i' \;\; \Rightarrow \sqrt{\lambda} = \frac{\sum_{i=1}^{k} \sqrt{c_i} W_i \; \sigma_i'}{C_T - c_0}$$

e dado que $n_i = \dfrac{W_i \; \sigma_i'}{\sqrt{c_i \lambda}}$ tem-se

$$\boxed{n_i = \frac{(C_T - c_0) W_i \sigma_i' \; / \sqrt{c_i}}{\sum_{i=1}^{k} W_i \sigma_i' \sqrt{c_i}}},$$

Esta é a dimensão óptima da amostra a recolher em cada estrato para um custo total fixo.

Observações:

-- As dimensões das amostras em cada estrato devem ser:

--- proporcionais ao tamanho do estrato e ao desvio padrão do estrato e

--- inversamente proporcionais à raíz quadrado do preço unitário de amostragem em cada estrato.

A **dimensão total** da amostra a recolher é

$$\boxed{n = \frac{(C_T - c_0)\sum_{i=1}^{k} W_i \sigma_i' \ / \sqrt{c_i}}{\sum_{i=1}^{k} W_i \sigma_i' \sqrt{c_i}}}$$

Caso particular

Se os custos c_i são os mesmos para todos os estratos tem-se

$C_T = c_0 + nc$ onde c é o custo unitário de amostragem (constante), donde

$$\boxed{n_i = n\frac{W_i\sigma_i'}{\sum_{i=1}^{k} W_i\sigma_i'} \qquad \text{com} \qquad n = \frac{C_T - c_0}{c}}$$

esta é a **dimensão óptima**, para n fixo.

Chama-se a esta afectação, **afectação de Neymann** ou **afectação óptima**, tendo então como variância mínima

$$\boxed{Var_{\min}\left[\overline{X}_{st}\right] = \frac{1}{n}\left(\sum_{i}^{k} W_i\sigma_i'\right)^2 - \frac{1}{N}\sum_{i=1}^{k} W_i\sigma_i'^2}.$$

b) Custo mínimo para variância fixa

Consideremos $Var\left[\overline{X}_{st}\right] = V$

--- qual a dimensão da amostra a recolher em cada estrato de modo a termos um custo mínimo?

Do que vimos atrás sabemos que $Var\left[\overline{X}_{st}\right]$ é minimizada quando os n_i são escolhidos proporcionalmente a $W_i\sigma_i' \ / \sqrt{c_i}$.

Para um dado V deverá haver um custo mínimo para o qual a afectação permitirá obter V como a variância mínima. Sendo assim a escolha dos n_i será aquela que satisfazendo a proporcionalidade acima referida, minimize o custo total, para um dado valor de $Var\left[\overline{X}_{st}\right]$, isto é,

$$n_i = k\frac{W_i\sigma_i'}{\sqrt{c_i}}$$

onde *k* deve ser escolhido de modo a assegurar que

$$\boxed{Var\left[\overline{X}_{st}\right] = \sum_{i=1}^{k} W_i^2\frac{\sigma_i'^2}{n_i} - \frac{1}{N}\sum_{i=1}^{k} W_i\sigma_i'^2 = V}.$$

Sendo assim deve tomar-se

$$\boxed{n_i = \left\{\frac{\sum_{i=1}^{k} W_i\sigma_i'\sqrt{c_i}}{V + \frac{1}{N}\sum_{i=1}^{k} W_i\sigma_i'^2}\right\} W_i\sigma_i' \,/\sqrt{c_i}}.$$

Mais uma Nota

No caso de pretendermos uma afectação proporcional, isto é, se $n_i = n\frac{N_i}{N}$, que **valor de *n*** se deve considerar?

Nalguns casos é pre-fixado;

caso contrário, sendo *d*, o erro absoluto, considera-se

$$\boxed{n_0 \cong \frac{4\sum W_i \sigma_i'^2}{d^2} \qquad \text{se} \quad \alpha = 0.05}$$

caso a população seja finita, deve considerar-se a correcção

$$n = \frac{n_0}{1 + n_0 / N} \quad .$$

Estimação de Proporções

Seja P a proporção dos indivíduos na população, verificando uma dada característica, A

Definindo, como fizemos na amostragem aleatória simples, as variáveis aleatórias Y_i como

$$Y_i = \begin{cases} 1 & \text{se o elemento } i \text{ verifica A} \\ 0 & \text{se o elemento } i \text{ não verifica A} \end{cases}$$

Seja

$$\boxed{Y_T = \sum_{1}^{N} Y_i \quad \text{donde} \quad P = \frac{\sum_{i=1}^{N} Y_i}{N}}$$

Como **estimador de *P*** tem sentido considerar

$$\boxed{\overline{Y_{st}} = \sum_{i=1}^{k} \frac{N_i}{N} \overline{Y_i} = \sum_{i=1}^{k} W_i \hat{P}_i = \hat{P}_{st}}$$

onde $\overline{Y_i} = \hat{P}_i$ designa a proporção de individuos no estrato *i*, incluidos na amostra e verificando A. O estimador de *P* é tal que

$$\boxed{\begin{array}{l} E\left[\hat{P}_{st}\right] = P \\ Var\left[\hat{P}_{st}\right] = \sum_{i=1}^{k} \frac{W_i^2}{n_i} \left(\frac{N_i - n_i}{N_i - 1}\right) P_i (1 - P_i) \end{array}}$$

(comparar com slide 42)

Um **estimador desta variância** é :

$$\boxed{Va\hat{r}\left[\hat{P}_{st}\right] = S'^2\left[\hat{P}_{st}\right] = \sum_{i=1}^{k} \frac{W_i^2}{n_i - 1} \left(\frac{N_i - n_i}{N_i - 1}\right) \hat{P}_i (1 - \hat{P}_i)}$$

Se N_i grande tem-se

$$\boxed{Var\left[\hat{P}_{st}\right] = \sum \frac{W_i^2}{n_i}\left(1 - f_i\right) P_i \left(1 - P_i\right)}$$

Se estamos numa situação de afectação proporcional, isto é, se $\frac{n_i}{N_i} = \frac{n}{N}$ tem-se

$$\boxed{Var\left[\hat{P}_{st}\right] = \frac{N-n}{n} \sum \frac{W_i^2}{N_i - 1} P_i (1-P_i) \cong \frac{1-f}{n} \sum W_i \, P_i (1-P_i)}.$$

Dimensão da amostra em cada estrato de modo a minimizar a variância com custo fixo

No caso de $C_T = c_0 + \sum c_i n_i$, tem-se a dimensão da amostra a recolher em cada estrato

$$\boxed{n_i = \frac{(C_T - c_0) W_i \sqrt{P_i Q_i / c_i}}{\sum W_i \sqrt{P_i Q_i c_i}}}$$

Dimensão total da amostra

$$\boxed{n = \sum n_i = \frac{(C_T - c_0) \sum_{i=1}^{k} W_i \sqrt{P_i Q_i / c_i}}{\sum W_i \sqrt{P_i Q_i c_i}}}$$

Se considerarmos a **afectação de Neyman**, com *n* fixo ignorando custos tem-se

$$\boxed{n_i = \frac{n W_i \sqrt{P_i Q_i}}{\sum W_i \sqrt{P_i Q_i}}}.$$