

# Teoria da Amostragem

## I- Noções fundamentais sobre amostragem.

**Amostragem** é todo o processo de recolha de uma parte, geralmente pequena, dos elementos que constituem um dado conjunto. Da análise dessa parte pretende obter-se informações para todo o conjunto.

Vejamos agora algumas noções básicas da teoria da amostragem:

--- **População** -- é a colecção de todos os elementos com uma dada característica comum.

Num processo de amostragem é importante distinguir entre **população objectivo** -- é a totalidade dos elementos em estudo e relativamente aos quais se pretende obter certo tipo de informação e **população inquirida** -- aquela sobre a qual é efectivamente feita a amostragem.

A população objectivo e população inquirida devem coincidir. Se isso não acontecer deve ter-se em conta que as extrapolações apresentadas neste texto dizem respeito à população inquirida.

--- **Característica** ou **atributo da população** -- é a informação relativa à população que se pretende estudar.

As características podem ser de natureza quantitativa e neste caso consideram-se escalas numéricas nas quais as variáveis se podem classificar em

- contínuas (referem-se a medições, pesagens, etc.);
- discretas (referem-se a contagens),

ou de natureza qualitativa e neste caso classificam-se em

- nominais (ex: sexo, espécie de uma dada planta ou animal, etc...);
- ordinais (ex: items de valores de uma dada classificação).

--- **População de amostras** -- é o conjunto de todas as amostras possíveis.

--- **Estatística** -- é uma função da amostra aleatória que não contém parâmetros desconhecidos.

--- **Unidade de amostragem** ou **unidade estatística** -- é o elemento da população considerada e sobre o qual vai ser estudada a característica de interesse -

exemplos: um animal, uma planta, um objecto, uma família, uma exploração agrícola, um bairro, etc.

O objectivo principal da teoria da amostragem é obter uma amostra que seja uma representação honesta da população e que conduza à estimação das características da população com grande precisão.

Algumas das vantagens que podemos desde já apontar ao usarmos um processo de amostragem no estudo de um dado problema são:

- a) redução dos custos e maior rapidez no apuramento dos resultados;
- b) maior profundidade na recolha de elementos;
- c) resolve o problema de estudar características que são destrutivas;
- d) minimiza os erros associados à recolha de informação (na recolha, registo e tratamento de informação há sempre erros associados. A recolha de um número menor de elementos faz, obviamente, diminuir as possibilidades deste tipo de erro).

Qual o processo a adoptar na recolha de elementos a incluir na amostra?

Isto constitui o que se designa por **plano de amostragem**.

Vejamos quais são as fases principais de um plano de amostragem adequado:

- definição dos objectivos do estudo;
- escolha dos dados úteis a recolher, o que significa:
  - definição da unidade de amostragem;
  - definição da escala de valores para a característica em estudo;
- definição da população ou universo;
- escolha do método de amostragem;
- definição do nível de precisão ou erro de amostragem admitido.

Definida a população há que decidir sobre o processo a adoptar na recolha dos elementos a incluir na amostra, isto é, o método de amostragem. Tais processos podem ser globalmente classificados em:

**1--métodos não aleatórios** ou **dirigidos** - nestes métodos a construção da amostra é feita a partir de informação à priori sobre a população estudada, tentando que a amostra seja um espelho fiel dessa população. Por assentarem em bases empíricas, tais métodos não permitem calcular a precisão das estimativas obtidas a partir da amostra.

Os métodos não aleatórios mais conhecidos são a amostragem orientada, a amostragem por conveniência e a amostragem por quotas.

**2 --métodos aleatórios** ou **probabilísticos** quando cada elemento da população tem uma probabilidade conhecida de fazer parte da amostra. Estes métodos possibilitam a determinação da distribuição de probabilidade, pelo menos assintoticamente, do estimador de interesse, conseqüentemente a determinação da sua variância e permitem por isso quantificar o erro de amostragem decorrente da utilização de apenas uma parte da população.

Destes métodos iremos estudar a amostragem aleatória simples, a amostragem estratificada, a amostragem por conglomerados e a amostragem multietápica.

Outros desenvolvimentos além dos que irão ser aqui abordados podem ver-se em Cochran (1977).

Dada uma população, seja  $\theta$  a característica de interesse e seja  $\hat{\theta}$  um estimador construído a partir de uma amostra aleatória.

As propriedades de um estimador são de grande interesse para a sua caracterização. A **variância** do estimador é de importância fundamental em amostragem porque do seu valor depende:

- a precisão do estimador;
- o tamanho da amostra para obter a precisão desejada;
- a escolha do melhor método de selecção da amostra.

Ao falarmos na importância da variância de um estimador estamos a pensar em estimadores centrados. Acontece muitas vezes que, alguns estimadores usados são enviesados.

Sendo assim, se pretendermos comparar dois estimadores, um centrado e outro não ou dois enviesados, a medida adequada é o **erro quadrático médio(EQM)**, assim definido:

$$EQM(\hat{\theta}) = E\left[(\hat{\theta} - \theta)^2\right] = Var[\hat{\theta}] + (E[\hat{\theta}] - \theta)^2 .$$

O erro quadrático médio é a medida da exactidão, rigor (em inglês *accuracy*) do estimador considerado, relativamente ao que se está a estudar, enquanto a variância é a medida da precisão (do inglês *precision*) do afastamento ao valor esperado do estimador (medida obtida ´estimada` por sucessivas réplicas do procedimento de amostragem). **EQM** e **variância** coincidem, como se sabe, se o estimador é centrado.

## Amostragem de uma população finita

Consideremos uma população  $P$ , constituída por  $N$  indivíduos. Designemos por  $X$  a característica em estudo que supomos assumir os seguintes valores

$A_1, A_2, \dots, A_N$  para todos os elementos da população.

Em geral interessa-nos conhecer aspectos ou parâmetros caracterizadores da população, tais como:

$$\text{Valor Médio} \quad \mu = \mu_X = \sum_{i=1}^N \frac{A_i}{N} \quad (1)$$

$$\text{Variância} \quad \sigma_X^2 = E[(X - \mu)^2] = \sum_{i=1}^N \frac{(A_i - \mu)^2}{N} = \sum_{i=1}^N \frac{A_i}{N} - \mu^2 \quad (2)$$

$$\text{ou} \quad \sigma_X'^2 = \sum_{i=1}^N \frac{(A_i - \mu)^2}{N - 1} = \frac{N}{N - 1} \sigma_X^2 \quad (3)$$

$$\text{Total} \quad T = X_T = \sum A_i = n\mu \quad (4)$$

$$\text{Razão de dois totais} \quad \frac{X_T}{Y_T} \quad (5)$$

**Proporção**  $P$  dos elementos da população que possuem um certo atributo.

## Amostragem aleatória simples com reposição

Se considerarmos uma população com  $N$  elementos, num processo de amostragem com reposição, cada elemento tem a mesma probabilidade  $1/N$  de ser seleccionado. Sendo assim, qualquer amostra de dimensão  $n$  tem probabilidade  $1/N^n$  de ser seleccionada.

Seja então  $X_1, X_2, \dots, X_n$  uma amostra aleatória retirada com reposição de uma população com  $N$  elementos com valores  $A_i$  ( $i = 1, \dots, N$ ) e  $x_1, x_2, \dots, x_n$  a correspondente amostra observada.

Cada elemento da amostra  $X_i$  pode tomar qualquer valor  $A_i$  com probabilidade  $1/N$ .

Um estimador centrado para  $\mu$  é, como sabemos,  $\bar{X} = \sum_{i=1}^n \frac{X_i}{n}$ .

Tem-se ainda  $Var(\bar{X}) = \frac{\sigma^2}{n}$ . (6)

A  $\sqrt{Var(\bar{X})} = \frac{\sigma}{\sqrt{n}}$  chama-se **erro padrão** da média.

Como regra geral não se conhece  $\sigma^2$  não é possível saber o valor do erro padrão. Há então que determinar um estimado de  $\sigma^2$ . Vamos lembrar que

$S'^2 = \sum_{i=1}^n \frac{(X_i - \bar{X})^2}{n-1}$  é um estimador centrado de  $\sigma^2$ .

Efectivamente

$$E[S'^2] = E\left[\sum_{i=1}^n \frac{(X_i - \bar{X})^2}{n-1}\right] = E\left[\frac{\sum_{i=1}^n X_i^2 - n\bar{X}^2}{n-1}\right] = \frac{\sum_{i=1}^n E[X_i^2] - nE[\bar{X}^2]}{n-1}.$$

Relembrando que  $Var[X] = E[X^2] - E^2[X]$  tem-se

$$E[S'^2] = \frac{\sum(\sigma^2 + \mu^2) - n(\sigma^2/n + \mu^2)}{n-1} = \frac{n\sigma^2 + n\mu^2 - \sigma^2 - n\mu^2}{n-1} = \sigma^2.$$

Num processo de amostragem, é necessário calcular a dimensão da amostra a recolher, de modo a obter a estimativa de interesse, com um erro inferior a  $\epsilon$ , fixado um nível de confiança.

Quando a dimensão da amostra aumenta, aumenta a precisão do estimador, mas também os custos de amostragem.

Idealmente deve estabelecer-se a precisão desejada e então escolher a dimensão da amostra.

Como se sabe, um intervalo de confiança para  $\mu$  a  $(1-\alpha)100\%$  de confiança, no caso de uma amostra aleatória obtida com reposição é

$$\left[ \bar{x} - t_{\alpha/2} \frac{s'}{\sqrt{n}}, \bar{x} + t_{\alpha/2} \frac{s'}{\sqrt{n}} \right] \quad (7)$$

determinado com base numa amostra de dimensão  $n$ .

Sendo assim, fixado o nível de precisão ou erro de amostragem ( $\epsilon$ ) e o nível de confiança ( $1-\alpha$ ) ou o risco ( $\alpha$ ) podemos determinar a dimensão da amostra a recolher por forma a termos um erro inferior a  $\epsilon$ . Para isso basta então exigir que

$$t_{\alpha/2} \frac{s'}{\sqrt{n}} \leq \epsilon \Rightarrow n \geq \left( \frac{t_{\alpha/2} s'}{\epsilon} \right)^2. \quad (8)$$

Porém, para calcular o valor  $t_{\alpha/2}$  é necessário saber o número de graus de liberdade ( $n-1$ ), e conseqüentemente a dimensão da amostra, que é afinal aquilo que pretendemos calcular. Por isso na prática costuma usar-se  $t_{\alpha/2}=2$  para um nível de significância de 5%. No que se refere ao valor  $s'$ , o desvio padrão da amostra, necessita de ser conhecido para se ter a dimensão da amostra.

O que se deverá fazer?

-- considerar uma amostragem de uma população semelhante e usar os valores de interesse desse estudo.

-- fazer um estudo piloto para, a partir dele obter estimativas dos parâmetros desconhecidos para podermos usar a fórmula (8).

-- considerar uma amostragem bi-etápica, isto é, obter uma primeira amostra de dimensão  $n_1$  e com desvio padrão  $s'_1$ . Para uma precisão  $\epsilon$ , a amostra final deverá ter um número de elementos  $n$ , dado por

$$n \geq \left( \frac{t_{\alpha/2} s'_1}{\epsilon} \right)^2 \left( 1 + \frac{2}{n_1} \right). \quad (9)$$

Se o valor resultante para  $n$  é tal que  $\frac{n}{N}$  é apreciável ( $>5\%$  ou  $>10\%$ ), deve considerar-se como dimensão de amostra a recolher o valor dado por

$$n^* \geq \frac{n}{1 + n/N}.$$

## Amostragem sem reposição

Neste caso a situação é diferente da anterior, porque os elementos vão ser incluídos na amostra sem reposição o que torna as variáveis aleatórias correspondentes aos valores da característica em estudo não independentes umas das outras. No entanto, no caso da população ser grande relativamente à dimensão da amostra extraída, pode considerar-se um esquema de amostragem em que aquelas variáveis são praticamente independentes.

Vejamos neste caso o estudo das propriedades dos estimadores da média e da variância da população.

Para facilitar consideremos as seguintes variáveis indicatrizes:

$$I_j \begin{cases} 1 & \text{se } A_j \text{ está na amostra} \\ 0 & \text{se } A_j \text{ não está na amostra} \end{cases}$$

Seja novamente  $(X_1, X_2, \dots, X_n)$  a amostra retirada desta vez sem reposição

então

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n} = \frac{\sum_{j=1}^N A_j I_j}{n}$$

(Note-se que se  $A_j$  está na amostra  $A_j I_j = X_j$ )

Vamos então calcular o valor médio e a variância de  $\bar{X}$ . Para isso vamos estudar a v. a.  $I_j$ .

$$P[I_j = 1] = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N},$$

$$E[I_j] = 0 \times P(I_j = 0) + 1 \times P(I_j = 1) = \frac{n}{N}; \quad \text{donde}$$

$$E[\bar{X}] = \frac{E\left[\sum_{j=1}^N A_j I_j\right]}{n} = \frac{1}{n} \sum A_j E[I_j] = \frac{1}{n} \sum_{j=1}^N A_j \frac{n}{N} = \mu$$

Portanto  $\bar{X}$  é estimador centrado de  $\mu$ .

Calculemos agora a variância de  $\bar{X}$ .

$$\text{Var}[\bar{X}] = \text{Var}\left[\frac{1}{n} \sum_{j=1}^N A_j I_j\right] = \frac{1}{n^2} \text{Var}\left[\sum A_j I_j\right]$$

Ora atendendo a que os  $I_j$  não são independentes tem-se

$$\text{Var}\left[\sum_{j=1}^N A_j I_j\right] = \sum A_j^2 \text{Var}[I_j] + \sum_{i \neq j} A_i A_j \text{Cov}(I_i, I_j)$$

$$\text{Ora } \text{Var}[I_j] = E[I_j^2] - E^2[I_j] = \frac{n}{N} - \frac{n^2}{N^2} = \frac{Nn - n^2}{N^2} = \frac{n}{N} \left(1 - \frac{n}{N}\right) \quad (10)$$

$$\text{Cov}(I_i, I_j) = E[I_i I_j] - E[I_i] \cdot E[I_j] = \frac{\binom{N-2}{n-2}}{\binom{N}{n}} - \frac{n}{N} \cdot \frac{n}{N} = \frac{n(n-1)}{N(N-1)} - \left(\frac{n}{N}\right)^2$$

o que, após pequenos cálculos dá

$$\text{Cov}(I_i, I_j) = -\frac{n}{N} \left(1 - \frac{n}{N}\right) \left(\frac{1}{N-1}\right). \quad (11)$$

Por curiosidade vejamos que a correlação é assim dada.

$$\rho(I_i, I_j) = \frac{\text{Cov}(I_i, I_j)}{\sqrt{\text{Var}(I_i)\text{Var}(I_j)}} = -\frac{1}{N-1}. \quad (12)$$

Observe-se que a covariância tende para zero quando  $N \rightarrow \infty$ , o que explica a quase independência para populações grandes.

O sinal negativo no coeficiente de correlação também se interpreta com facilidade, bastando pensar que o facto de na amostra se observar um elemento com a característica A. diminui a probabilidade de se observar outro com essa mesma característica.

Calculemos então

$$\begin{aligned} \text{Var}(\bar{X}) &= \frac{1}{n^2} \text{Var}\left(\sum_{j=1}^N A_j I_j\right) = \frac{1}{n^2} \left[ \sum A_j^2 \text{Var}(I_j) + \sum_{j \neq k} A_j A_k \text{Cov}(I_j, I_k) \right] = \\ &= \frac{1}{n^2} \left[ \sum_{j=1}^N A_j^2 \frac{n}{N} \left(1 - \frac{n}{N}\right) - \sum_{j \neq k} A_j A_k \frac{n}{N} \left(1 - \frac{n}{N}\right) \frac{1}{N-1} \right] \end{aligned}$$



Atendendo a que  $\sum_{k \neq j} A_j A_k$  se pode escrever como

$$\sum_{j \neq k} A_j A_k = \sum_j A_j \left( \sum_{k \neq j} A_k \right) \quad \text{com} \quad \sum_{k \neq j} A_k = (A_1 + \dots + A_N) - A_j = N\mu - A_j$$

vem

$$\sum_{j \neq k} A_j A_k = \sum_j A_j (N\mu - A_j) = N^2 \mu^2 - \sum_j A_j^2,$$

após o que, considerando a substituição, se tem

$$\begin{aligned} \text{Var}(\bar{X}) &= \frac{1}{n^2} \left[ \frac{n}{N} \left( 1 - \frac{n}{N} \right) \sum A_j^2 - \frac{n}{N} \left( 1 - \frac{n}{N} \right) \frac{1}{N-1} (N^2 \mu^2 - \sum A_j^2) \right] = \\ &= \frac{1}{n^2} \frac{n}{N} \left( 1 - \frac{n}{N} \right) \left[ \sum A_j^2 - \frac{1}{N-1} (N^2 \mu^2 - \sum A_j^2) \right] = \frac{N-n}{N^2 n} \frac{N \sum A_j^2 - N^2 \mu^2}{N-1} = \\ &= \frac{N-n}{N-1} \frac{1}{n} \frac{\sum A_j^2 - N \mu^2}{N} = \frac{N-n}{N-1} \frac{\sigma^2}{n}. \end{aligned} \quad (13)$$

Observe-se que

$$\frac{N-n}{N-1} \frac{\sigma^2}{n} < \frac{\sigma^2}{n} \quad \text{isto é}$$

$$\begin{array}{ccc} \text{Var}(\bar{X}) & < & \text{Var}(\bar{X}) \\ \text{s/ reposição} & & \text{c/ reposição} \end{array}$$

Sendo assim, quer dizer que a amostragem sem reposição é mais eficiente do que a amostragem com reposição para estimar o valor médio.

Se  $N$  é grande comparativamente a  $n$ , a fracção  $\frac{N-n}{N-1}$  não difere muito de 1 e a diferença na eficiência torna-se desprezável.

Ao factor  $\frac{N-n}{N-1}$  chama-se **correção de população finita** e a

$f = \frac{n}{N}$  chama-se **fracção de amostragem.**

A expressão da variância acima deduzida pode ser apresentada usando a variância corrigida  $\sigma^2$ , isto é,

$$\text{Var}(\bar{X}) = \frac{N-n}{N-1} \frac{N-1}{N} \frac{\sigma'^2}{n} = \frac{N-n}{N} \frac{\sigma'^2}{n} = (1-f) \frac{\sigma'^2}{n}. \quad (14)$$

Vimos que no caso da amostragem com reposição  $S'^2$  era um estimador centrado de  $\sigma^2$ , veremos agora que no caso da amostragem sem reposição  $S'^2$  é estimador centrado de  $\sigma^2$ .

Ora

$$\begin{aligned} E[S'^2] &= E\left[\sum_{i=1}^n \frac{(X_i - \bar{X})^2}{n-1}\right] = \frac{1}{n-1} E\left[\sum (X_i - \mu)^2 - n(\bar{X} - \mu)^2\right] = \\ &= \frac{1}{n-1} \left[\sum E(X_i - \mu)^2 - nE(\bar{X} - \mu)^2\right] = \frac{1}{n-1} \left[n\sigma^2 - n \frac{N-n}{N-1} \frac{\sigma^2}{n}\right] = \\ &= \frac{1}{n-1} \left[\frac{n\sigma^2(N-1) - (N-n)\sigma^2}{N-1}\right] = \frac{1}{n-1} \left[\frac{N(n-1)\sigma^2}{N-1}\right] = \frac{N}{N-1} \sigma^2 = \sigma'^2 \end{aligned}$$

logo  $S'^2$  é estimador centrado de  $\sigma^2$  na amostragem sem reposição.

Neste caso uma estimativa do erro padrão é:

$$s' \sqrt{\frac{1-f}{n}} \quad (15)$$

### Intervalos de confiança para $\mu$

Vejamos o seguinte exemplo, Barnett (1994).

Consideremos uma população com  $N=25$  elementos, todos conhecidos:

5 2 4 1 5 8 8 6 6 8 9 10 7 11 9 14 12 8 14 11 9  
8 11 10 15

Para esta população tem-se  $\mu=8.44$  e  $\sigma^2=12.42$  e dela é extraída aleatoriamente, sem reposição, uma amostra de 5 elementos. Seja por exemplo a amostra obtida

10 15 8 11 5

Para esta amostra tem-se  $\bar{x} = 9.8$  e  $Var(\bar{X}) = (1-f)\frac{\sigma^2}{5} = 1.9872$

Barnett (1994) apresenta o resultados obtidos quando, para aquela população se geram 500 amostras de dimensão 5. Verificou que

$$\bar{\bar{x}}_{500} = 8.46 \approx \mu \quad \text{e} \quad \bar{s}'^2 = 1.94 \approx Var(\bar{X}).$$

Tendo em conta o que foi acabado de observar, pode pensar-se numa extensão do Teorema Limite Central ao caso de populações finitas.

Assim pode considerar-se

$$\bar{X} \approx N(\mu, (1-f)\sigma'^2 / n) \quad (16)$$

Este resultado pode ser razoavelmente aceite mesmo em presença de assimetria na população. Como uma regra grosseira para uso daquela distribuição aproximada em populações enviesadas à direita require-se que

$$n > 25G_1^2 \quad \text{com} \quad G_1 = \sum_{i=1}^N \frac{(A_i - \mu)^3}{N\sigma'^3} \quad (\text{coeficiente de assimetria para populações finitas})$$

e que  $f$  não seja demasiado grande, ver Cochran(1977).

Sendo assim, nas condições anteriores pode usar-se a distribuição normal para fazer inferências sobre  $\mu$ .

Nas condições atrás referidas um intervalo a  $(1-\alpha)100\%$  de confiança para  $\mu$  será então

$$\bar{x} - z_{\alpha/2} \sigma' \sqrt{\frac{1-f}{n}} < \mu < \bar{x} + z_{\alpha/2} \sigma' \sqrt{\frac{1-f}{n}} \quad (17)$$

sendo  $z_{\alpha/2}$  tal que  $P(|Z| > z_{\alpha/2}) = \alpha$ .

Porém na prática  $\sigma'$  não é conhecido e sendo assim considera-se  $s'$  como uma estimativa para  $\sigma'$ , o que é razoável desde que  $n$  grande, continuando a usar-se a aproximação à normal.

Se  $n$  não é suficientemente grande ( $n < 40$ ) e não se conhece  $\sigma'$ , o melhor é usar a distribuição  $t$ , donde um intervalo a  $(1-\alpha)100\%$  de confiança para  $\mu$  será então

$$\bar{x} - t_{\alpha/2(n-1)} s' \sqrt{\frac{1-f}{n}} < \mu < \bar{x} + t_{\alpha/2(n-1)} s' \sqrt{\frac{1-f}{n}} \quad (18)$$

sendo  $t_{\alpha/2(n-1)}$  tal que  $P\left(T > t_{\alpha/2(n-1)}\right) = \alpha$ , com  $T$  v.a. com distribuição  $t$  de Student.

Por exemplo em sondagens referem-se a populações grandes ( $N > 1000$ ) com amostras  $n > 100$  e por isso estamos em condições de usar a normal na construção de intervalos de confiança.

### Escolha da dimensão da amostra

Quando a dimensão da amostra aumenta, aumenta a precisão, mas há que ter em conta que também o custo de amostragem aumenta. Sendo assim há que criar-se uma situação de compromisso: a situação ideal seria escolher  $n$  de modo a ter precisão máxima com custo mínimo.

Neste caso pretendemos determinar o mínimo valor de  $n$  que permita estimar  $\mu$  de modo a ter uma precisão  $d$ .

Pretende-se então que

$$P\left\{|\bar{X} - \mu| \geq d\right\} < \alpha$$

Vimos já que o intervalo de confiança a  $(1-\alpha)100\%$  para  $\mu$  era

$$\bar{x} - z_{\alpha/2} \sigma' \sqrt{\frac{1-f}{n}} < \mu < \bar{x} + z_{\alpha/2} \sigma' \sqrt{\frac{1-f}{n}}$$

Basta então exigir que

$$z_{\alpha/2} \sigma' \sqrt{\frac{1-f}{n}} \leq d \Leftrightarrow z_{\alpha/2} \sigma' \sqrt{\frac{1-n/N}{n}} \leq d \Leftrightarrow z_{\alpha/2} \sigma' \sqrt{\frac{N-n}{Nn}} \leq d \Leftrightarrow \frac{N-n}{Nn} \leq \left(\frac{d}{z_{\alpha/2} \sigma'}\right)^2$$

$$\Leftrightarrow N (z_{\alpha/2} \sigma')^2 - n (z_{\alpha/2} \sigma')^2 - nNd^2 \leq 0 \Leftrightarrow n \left[ (z_{\alpha/2} \sigma')^2 + Nd^2 \right] \geq N (z_{\alpha/2} \sigma')^2$$

$$\Leftrightarrow n \geq \frac{N (z_{\alpha/2} \sigma')^2}{(z_{\alpha/2} \sigma')^2 + Nd^2} \Leftrightarrow n \geq \frac{\left(\frac{z_{\alpha/2} \sigma'}{d}\right)^2}{\left(\frac{z_{\alpha/2} \sigma'}{d}\right)^2 \frac{1}{N} + 1}$$

isto é, a dimensão da amostra é

$$n \geq \left(\frac{z_{\alpha/2} \sigma'}{d}\right)^2 \left[ \left(\frac{z_{\alpha/2} \sigma'}{d}\right)^2 \frac{1}{N} + 1 \right]^{-1} \quad (18)$$

Como primeira aproximação para  $n$  regra geral considera-se  $n \geq n_0 = \left( \frac{z_{\alpha/2} \sigma'}{d} \right)^2$ . No caso de  $\frac{n_0}{N}$  ter um valor muito elevado então deve usar-se como dimensão de amostra a recolher

$$n \geq n_0 \left[ \frac{n_0}{N} + 1 \right]^{-1}$$

Observe-se que, regra geral, mais uma vez se desconhece  $\sigma'$ , devendo então substituí-lo por  $s'$ .

Para isso seria necessário conhecer previamente a amostra que é aquilo que não se conhece. Há basicamente quatro atitudes a tomar:

**Recorrendo a estudos piloto**, que nos permitam uma primeira estimativa para  $\sigma'$ .

**Recorrendo a estudos prévios** da mesma população ou de populações semelhantes. É comum nas mais variadas áreas de interesse: medicina, educação, haver estudos de características semelhantes em populações semelhantes. Nesse caso uma medida da variabilidade obtida em situações semelhantes pode dar uma indicação de  $\sigma^2$ .

**Fazendo a selecção em duas fases**. É este o procedimento mais fiável, embora possa não ser praticável em termos administrativos ou de custos. Como se processa?

Tira-se uma amostra aleatória com  $n_1$  elementos e calcula-se  $s_1'^2$  como estimativa de  $\sigma^2$ . Necessitamos agora de verificar se a dimensão  $n_1$  é inadequada para obtermos a precisão requerida. Para isso aumenta-se a amostra com outra de dimensão  $(n - n_1)$  onde  $(n - n_1)$  é escolhida usando  $s_1'^2$  como uma estimativa inicial para  $\sigma^2$ . Cochran (1977) e Barnett (1994) propõem neste caso que se ignore a correcção de população finita (1-f) devendo a dimensão total da amostra ser pela mesma expressão definida em (9), isto é,

$$n \geq \left( \frac{t_{\alpha/2} s_1'}{d} \right)^2 \left( 1 + \frac{2}{n_1} \right).$$

**A partir de considerações práticas sobre a estrutura da população.** Pode acontecer ter-se alguma informação sobre a estrutura da população, por exemplo, pode haver razões que nos levem a suspeitar tratar-se de uma população de Poisson. Sendo assim  $\sigma^2 \cong \mu$ .

## Estimação do total T

Há muitas situações em que pretendemos estimar um total : a produção anual de trigo, etc.

$$\text{Dado que } T = X_T = N\mu \quad (19)$$

o estimador mais usado é

$$X_T^* = N\bar{X} \quad (20)$$

$$\text{sendo } E[X_T^*] = N\mu = X_T \quad \text{e} \quad \text{Var}[X_T^*] = N^2(1-f)\frac{\sigma'^2}{n} .$$

Nas mesmas condições referidas atrás, pode também aqui usar-se a aproximação à normal, tendo-se

$$X_T^* \approx N\left(X_T, N^2(1-f)\frac{\sigma'^2}{n}\right) \quad (21)$$

para construir intervalos de confiança para  $X_T$  e ainda determinar a dimensão da amostra necessária para obter certa precisão na estimação de  $X_T$ .

Se  $n > 50$  um intervalo de confiança para  $X_T$  a  $(1-\alpha)100\%$  é

$$x_T^* - z_{\alpha/2} N \sigma' \sqrt{\frac{1-f}{n}} < X_T < x_T^* + z_{\alpha/2} N \sigma' \sqrt{\frac{1-f}{n}} \quad (22)$$

Se  $n$  pequeno, digamos inferior a 50, substitui-se  $z_{\alpha/2}$  por  $t_{\alpha/2(n-1)}$  .

### Escolha de $n$

Fixada uma precisão  $d$ , para um nível de significância  $\alpha$ , pretende-se que

$$P\left[|X_T^* - X_T| < d\right] \geq 1 - \alpha$$

donde, e tendo em conta o intervalo de confiança escrito acima, terá que exigir-se

$$\begin{aligned} z_{\alpha/2} N \sigma' \sqrt{\frac{1-f}{n}} \leq d &\Leftrightarrow (z_{\alpha/2} N \sigma')^2 \frac{1-n/N}{n} \leq d^2 \Leftrightarrow \frac{1-n/N}{n} \leq \left(\frac{d}{z_{\alpha/2} N \sigma'}\right)^2 \\ &\Leftrightarrow \frac{N-n}{n} \leq N \left(\frac{d}{z_{\alpha/2} N \sigma'}\right)^2 \Leftrightarrow N \leq n \left[1 + \frac{1}{N} \left(\frac{d}{z_{\alpha/2} \sigma'}\right)^2\right] \end{aligned}$$

donde se tem

$$n \geq N \left[ 1 + \frac{1}{N} \left( \frac{d}{z_{\alpha/2} \sigma'} \right)^2 \right]^{-1}. \quad (23)$$

Mais uma vez estaremos em presença das mesmas dificuldades que surgiram anteriormente aquando da determinação da dimensão da amostra. As considerações sobre os procedimentos a usar deverão ser aqui tidas em conta.

Como primeira aproximação podemos considerar

$$n_0 \geq N^2 \left( \frac{z_{\alpha/2} \sigma'}{d} \right)^2.$$

Se  $\frac{n_0}{N}$  grande deve considerar-se  $n_0 \geq n_0 \left( 1 + \frac{n_0}{N} \right)^{-1}$ .

### Estimação de uma proporção P

No estudo de uma dada característica X, pretende-se estimar P, a proporção de elementos com uma dada propriedade.

Exemplo: Na população de estudantes de uma dada Universidade, qual a proporção dos que vivem em quartos alugados?

Retirando uma amostra aleatória de dimensão n, conta-se o número r de indivíduos que satisfazem a propriedade.

Sendo assim uma estimativa de P, pode ser dada por

$$\hat{p} = r / n$$

Ora o modo mais simples de obter propriedades para o estimador  $\hat{P}$  é usar as propriedades já estudadas anteriormente para o estimador do valor médio, bastando para isso considerar o seguinte:

Suponhamos que P representa a proporção de elementos de uma população finita de dimensão N, que verificam uma dada característica A. Pode construir-se a seguinte variável aleatória auxiliar associada a cada elemento da população:

$$Y_i = \begin{cases} 1 & \text{se o elemento da população verifica a propriedade A} \\ 0 & \text{se o elemento da população não verifica a propriedade A} \end{cases}$$

$Y_T = \sum_1^N Y_i = R$ , onde  $R$  é o número de elementos da população que verificam  $A$ .

$$\mu_Y = \frac{\sum_1^N Y_i}{N} = \frac{R}{N} = P \quad (24)$$

$P$  é então a média da variável  $Y$  na população;  $\hat{p}$  será então a média da amostra observada .

Para estudar a eficiência do estimador  $\hat{P}$ , estamos de novo na situação de considerar as propriedades da média de uma amostra para estimar a média da população.

Consideremos então a amostra aleatória  $Y_1, Y_2, \dots, Y_n$ , cuja média é

$$\bar{Y} = \frac{\sum_1^n Y_i}{n} = \frac{\hat{R}}{n} = \hat{P}, \quad (25)$$

sendo a verdadeira proporção,  $P$ , correspondente ao valor médio da variável  $Y$

$$P = \mu_Y = \frac{\sum_{i=1}^N Y_i}{N} = \frac{R}{N} \quad (26)$$

com variância

$$\sigma_Y'^2 = \frac{\sum_1^N (Y_i - \mu_Y)^2}{N-1} = \frac{\sum_1^N Y_i^2 - N\mu_Y^2}{N-1} = \frac{NP - NP^2}{N-1} = \frac{NP(1-P)}{N-1}. \quad (27)$$

Portanto 
$$E[\hat{P}] = \frac{\sum_{i=1}^n E[Y_i]}{n} = \frac{nP}{n} = P$$

logo  $\hat{P}$  é um estimador centrado.

$$Var[\hat{P}] = (1-f) \frac{\sigma_Y'^2}{n} = (1-f) \frac{NP(1-P)}{n(N-1)} = \left( \frac{N-n}{N-1} \right) \frac{P(1-P)}{n}. \quad (28)$$

Porém, mais uma vez estamos na situação de ter nas definições anteriores parâmetros desconhecidos, isto é,  $P$  é desconhecido, e por isso não é possível calcular



$\sigma'^2$ . Então terá que ser estimado, usando o estimador centrado de  $\sigma'^2$ ,  $S'^2$ , cuja estimativa é

$$s'^2_Y = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 = n\hat{p}\hat{q}/(n-1) \quad (29)$$

Donde, um estimador centrado de  $Var[\hat{P}]$  é

$$S'^2(\hat{P}) = (1-f)\hat{P}\hat{Q}/(n-1) \quad (30)$$

É de referir que este estimador não resulta da substituição dos valores da amostra, na expressão da variância da população, que vimos ser

$$Var[\hat{P}] = \frac{N-n}{N-1} \frac{P(1-P)}{n},$$

como se poderia pensar, embora a diferença seja muito pequena.

Se  $f$  é desprezável, tem-se

$$S'^2(\hat{P}) = \hat{P}\hat{Q}/(n-1). \quad (31)$$

que acontece em particular quando estamos a amostrar uma população infinita.

### Intervalos de confiança para P

Ao recolher atributos ou características para estimar P, sabemos mais acerca da distribuição de amostragem de  $\hat{P}$  do que nas situações correspondentes para estimar  $\mu$  ou  $X_T$ . De facto a distribuição exacta de  $\hat{P}$  é conhecida. O número  $R$  de elementos da amostra que possuem aquele atributo, tem distribuição hipergeométrica, i.e.,

$$P[\text{haver } r \text{ elementos}] = \frac{\binom{R}{r} \binom{N-R}{n-r}}{\binom{N}{n}}; \quad \max(0, n-R+N) \leq r \leq \min(R, n)$$

Porém, na prática, o conhecimento da distribuição exacta do número de elementos da amostra possuindo aquela característica não é muito importante, em face dos cálculos pesados que esta distribuição envolve.

É portanto útil procurar aproximações para a distribuição do estimador, agora num espírito mais pragmático do que teórico. Uma possibilidade consiste em usar a distribuição binomial como uma aproximação da hipergeométrica -- se  $n$  é pequeno relativamente a  $R$  e a  $(N-R)$ , a "falta de reposição" pode ser "ignorada", donde

$$\hat{R} \approx B(n, P)$$

Embora possamos usar esta distribuição binomial para construir intervalos de confiança para  $P$ , também esta envolve cálculos pesados (excepto se  $n$  é pequeno).

Na maioria das aplicações acha-se conveniente usar a aproximação pela normal, isto é,

$$\hat{P} \sim N\left(P, (1-f) \frac{PQ}{n}\right) \quad (32)$$

A aproximação à normal é razoável desde que:

- $n$  não seja muito grande relativamente a  $R$  e a  $N-R$ .
- o menor dos valores  $nP$  e  $nQ$  não seja muito pequeno,  $\min(nP, nQ) > 30$  é uma regra empírica habitualmente considerada.
- se  $P$  está próximo de  $1/2$ , então os valores pequenos de  $nP$  e  $nQ$  são assegurados pelos seus estimadores centrados  $n\hat{P}$  e  $n\hat{Q}$ .

Sendo assim um intervalo de confiança para  $P$  será

$$\hat{P} - z_{\alpha/2} \sqrt{\frac{(1-f)\hat{P}\hat{Q}}{n-1}} < P < \hat{P} + z_{\alpha/2} \sqrt{\frac{(1-f)\hat{P}\hat{Q}}{n-1}} \quad (33)$$

resultante da substituição de  $\text{var}(\hat{P})$  pelo seu estimador centrado

$$S'^2(\hat{P}) = (1-f) \frac{\hat{P}\hat{Q}}{n-1}. \quad (34)$$

### **Escolha do tamanho da amostra para estimar uma proporção**

Como vimos um estimador para  $P$  é  $\hat{P} = \frac{\hat{R}}{n}$  com

$$E[\hat{P}] = P \quad \text{e} \quad \text{Var}[\hat{P}] = \frac{N-n}{N-1} \frac{PQ}{n}.$$

A  $\text{Var}[\hat{P}]$  atinge o seu máximo para  $P=Q=1/2$ .

Quando se pretende determinar o tamanho da amostra para obter uma dada precisão na estimação de  $P$ , o que é que se pretende?

- a) o valor absoluto do erro ser inferior a um dado valor, ou  
 b) o valor relativo do erro?

a) Se pretendemos fixar um valor máximo para o erro absoluto, então

$$s.e.[\hat{P}] = \sqrt{\frac{PQ}{n}} \leq d \quad (\text{supondo } N \text{ grande, portanto } (1-f) \cong 1)$$

b) Se pretendemos fixar um valor máximo para o erro relativo, então

$$s.e.[\hat{P}]/P = \sqrt{\frac{Q}{nP}} \leq \varepsilon \quad (\text{supondo } N \text{ grande, portanto } (1-f) \cong 1)$$

Observe-se que o erro relativo não é mais do que o coeficiente de variação, por isso a condição expressa atrás é equivalente a dizer que pretendemos o coeficiente de variação não superior a  $\varepsilon$ .

Sendo assim, escolher o tamanho da amostra de modo a assegurar certos limites ao erro padrão ou ao coeficiente de variação é o mesmo que assegurar que

$$P\{|\hat{P} - P| > d\} \leq \alpha \quad \text{ou} \quad P\{|\hat{P} - P| > \xi P\} \leq \alpha$$

ou seja, considerando a aproximação pela normal, viria

$$s.e.[\hat{P}] = \sqrt{\frac{PQ}{n}} \leq \frac{d}{z_{\alpha/2}} \quad \text{ou} \quad s.e.[\hat{P}]/P = \sqrt{\frac{Q}{nP}} \leq \frac{\xi}{z_{\alpha/2}} .$$

Aqui, na determinação de  $n$  (dimensão da amostra), temos uma facilidade que não tínhamos no caso da estimação de  $\mu$  ou  $T$ , porque independentemente do valor que  $P$  possa assumir, podemos ter sempre um limite superior.

Para a primeira desigualdade tem-se

$$n \geq \frac{PQ z_{\alpha/2}^2}{d^2} ,$$

mas  $PQ$  tem como valor máximo 1/4, quando  $P=1/2$ , então

$$n \geq \frac{z_{\alpha/2}^2}{4d^2}$$

satisfaz a desigualdade pretendida.

No que respeita à segunda desigualdade já não é possível majorá-la.

Os resultados apresentados até aqui consideravam *f* desprezável. Mas **se *f* não é desprezável**, terá que considerar-se a fórmula exacta para

$$\text{Var}[\hat{P}] = \frac{N-n}{N-1} \frac{PQ}{n}, \text{ donde}$$

$$z_{\alpha/2} \sqrt{\frac{N-n}{N-1} \frac{PQ}{n}} \leq d \Leftrightarrow \frac{N-n}{N-1} \frac{PQ}{n} \leq \frac{d^2}{z_{\alpha/2}^2} \Leftrightarrow \frac{N-n}{n} \leq \frac{N-1}{PQ} \left( \frac{d}{z_{\alpha/2}} \right)^2 \Leftrightarrow$$

$$n \geq N \left[ 1 + \frac{N-1}{PQ} \left( \frac{d}{z_{\alpha/2}} \right)^2 \right]^{-1} \Leftrightarrow n \geq \frac{PQ z_{\alpha/2}^2}{d^2} \left[ 1 + \frac{1}{N} \left\{ PQ \left( \frac{z_{\alpha/2}}{d} \right)^2 - 1 \right\} \right]^{-1}$$

Podemos tomar como primeira aproximação

$$n_0 \geq \frac{PQ z_{\alpha/2}^2}{d^2} \quad (35)$$

porém se  $\frac{n_0}{N}$  é grande, deve considerar-se  $n_0 \geq n_0 \left( 1 + \frac{n_0 - 1}{N} \right)^{-1}$ .

Vejamos o caso de se pretender uma precisão proporcional a P:

Ora sabe-se que  $\text{Var}(\hat{P}) = \frac{N-n}{N-1} \frac{PQ}{n}$  e pretende-se que

$$z_{\alpha/2} \sqrt{\frac{N-n}{N-1} \frac{PQ}{n}} \leq \xi P \Leftrightarrow \frac{N-n}{N-1} \frac{PQ}{n} \leq \frac{\xi^2 P^2}{z_{\alpha/2}^2} \Leftrightarrow \frac{N-n}{n} \leq (N-1) \frac{\xi^2 P^2}{PQ z_{\alpha/2}^2} \Leftrightarrow$$

$$\Leftrightarrow \frac{N}{n} \leq 1 + (N-1) \frac{\xi^2 P}{Q z_{\alpha/2}^2} \Leftrightarrow n \geq N \left[ 1 + (N-1) \frac{\xi^2 P}{Q z_{\alpha/2}^2} \right]^{-1} \Leftrightarrow n \geq \frac{Q}{P} \left( \frac{z_{\alpha/2}}{\xi} \right)^2 \left[ 1 + \frac{1}{N} \left( \frac{Q z_{\alpha/2}^2}{P \xi^2} - 1 \right) \right]^{-1}.$$

Como primeira aproximação pode considerar-se  $n_0 \geq \frac{Q z_{\alpha/2}^2}{P \xi^2}$ . (37)

De novo se  $\frac{n_0}{N}$  é grande, deve considerar-se  $n_0 \geq n_0 \left( 1 + \frac{n_0 - 1}{N} \right)^{-1}$ .

## Estimação de uma razão

Consideremos a amostra aleatória constituída por  $n$  pares de valores  $(X_i, Y_i)$  obtida por amostragem aleatória simples. Suponhamos que pretendemos estimar a razão

$$R = X_T / Y_T = \mu_X / \mu_Y. \quad (38)$$

Para isso dispomos então de uma amostra com os valores  $(x_1, y_1) \dots (x_n, y_n)$  e seja então o estimador de  $R$ ,

$$R^* = \bar{X} / \bar{Y}. \quad (39)$$

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

$$\begin{aligned} E[R^*] &\cong R = \mu_X / \mu_Y; \\ \text{Var}[R^*] &\cong \frac{1-f}{n\bar{y}^2} \sum_1^N \frac{(X_i - RY_i)^2}{N-1} = \frac{1-f}{n\bar{y}^2} [\sigma_X^2 - 2R\sigma_{XY} + R^2\sigma_Y^2] \end{aligned} \quad (40)$$

Uma estimativa de  $\text{Var}[R^*]$  é

$$s'^2[R^*] = \frac{1-f}{n\bar{y}^2} \sum_1^n \frac{(x_i - r^* y_i)^2}{n-1}, \quad (41)$$

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

Para grandes amostras um intervalo a  $(1-\alpha)100\%$  de confiança para  $R$  é

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

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  $Y_T$  é conhecido. Neste caso é possível estimar o valor médio  $\mu_X$ ,  $\mu_X = R\mu_Y$ , usando o **estimador da razão**, assim definido

$$\bar{X}_R = \frac{\bar{X}}{\bar{Y}} \mu_Y = R^* \mu_Y \quad (42)$$

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

$$\text{Var}[\bar{X}_R] \cong \frac{1-f}{n} \sum_1^N \frac{(X_i - RY_i)^2}{N-1} = \frac{1-f}{n} [\sigma_X^2 - 2R\sigma_{XY} + R^2\sigma_Y^2].$$

Uma vez construído o estimador da razão, coloca-se uma pergunta natural:  
 --Em que circunstâncias será o estimador da razão preferível ao estimador habitual da média? Será  $\bar{X}_R$  mais ou menos eficiente do que  $\bar{X}$ ?

Isto é, em que condições  $\text{Var}[\bar{X}_R] < \text{Var}[\bar{X}]$ ?

Ora tem-se

$$\begin{aligned} \frac{1-f}{n} [\sigma_X^2 - 2R\sigma_{XY} + R^2\sigma_Y^2] &< \frac{1-f}{n} \sigma_X^2 \\ \Downarrow \\ 2R\rho\sigma_X\sigma_Y &> R^2\sigma_Y^2 \\ \Downarrow \\ \rho &> \frac{R\sigma_Y}{2\sigma_X} \Rightarrow \rho > \frac{1}{2} \frac{CV_Y}{CV_X}, \end{aligned}$$

onde  $CV$  designa coeficiente de variação .

## Amostragem Estratificada

Suponhamos que temos a população dividida em subpopulações ou **estratos**. (Esta divisão regra geral é feita com base numa variável dita de estratificação).

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 homogêneos forem os estratos;
- conveniências administrativas de organização do trabalho de recolha da informação.

Suponhamos então que dispomos de uma população finita com  $N$  indivíduos (note que são as nossas unidades de amostragem) e sejam  $a_1, \dots, a_N$  os valores de uma dada característica para aqueles indivíduos. Suponhamos que a população é dividida em  $k$  grupos ou **estratos** de dimensões conhecidas:  $N_1, \dots, N_k$  ( $\sum N_i = N$ ), assim caracterizados:

Estrato	dimensão	elementos	valor médio	variância
$S_1$	$N_1$	$a_{11}a_{12} \cdots a_{1N_1}$	$\mu_1$	$\sigma_1^2$
$S_2$	$N_2$	$a_{21}a_{22} \cdots a_{2N_2}$	$\mu_2$	$\sigma_2^2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$S_k$	$N_k$	$a_{k1}a_{k2} \cdots a_{kN_k}$	$\mu_k$	$\sigma_k^2$

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

**Valor médio** (43)

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

**Variância da população** (44)

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

onde  $W_i = \frac{N_i}{N}$  é o “peso “ em cada estrato.

De facto tem-se

$$\begin{aligned}\sigma'^2 &= \frac{1}{N-1} \sum_{i,j} (a_{ij} - \mu)^2 = \frac{1}{N-1} \sum_{i=1}^k \sum_{j=1}^{N_i} (a_{ij} - \mu)^2 = \frac{1}{N-1} \sum_i \left[ \sum_{j=1}^{N_i} (a_{ij} - \mu_i + \mu_i - \mu)^2 \right] \\ &= \frac{1}{N-1} \sum_i \left[ \sum_{j=1}^{N_i} (a_{ij} - \mu_i)^2 + N_i (\mu_i - \mu)^2 + 2(\mu_i - \mu) \underbrace{\sum_{j=1}^{N_i} (a_{ij} - \mu_i)}_{=0} \right] = \\ &= \frac{1}{N-1} \left[ \sum_{i=1}^k (N_i - 1) \sigma_i'^2 + \sum_{i=1}^k N_i (\mu_i - \mu)^2 \right].\end{aligned}$$

Para cada estrato  $i$  tem-se

$$\mu_i = \frac{1}{N_i} \sum_{j=1}^{N_i} a_{ij} \quad \text{e} \quad \sigma_i'^2 = \frac{1}{N_i - 1} \sum_{j=1}^{N_i} (a_{ij} - \mu_i)^2 \quad (45)$$

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

$$n_1, n_2, \dots, n_k \quad \left( \sum_i n_i = n \right)$$

tendo como elementos em cada estrato  $i$

$$x_{i1}, x_{i2}, \dots, x_{in_i}$$

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

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

A  $f_i = \frac{n_i}{N_i}$  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? É isto que nós designaremos por afecção.



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

Fixada a dimensão da amostra a recolher, seja  $n$ , 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

$$f_i = \frac{n_i}{N_i} = \frac{n}{N} \quad (46)$$

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

### Estimação do valor médio

O estimador do valor médio é a média empírica estratificada assim definida

$$\bar{X}_{st} = \sum_{i=1}^k W_i \bar{X}_i = \sum_{i=1}^k \frac{N_i}{N} \bar{X}_i. \quad (47)$$

Observe-se que, a média empírica estratificada não é o mesmo que a média aritmética, assim definida

$$\bar{X}' = \sum_{i=1}^k \frac{n_i}{n} \bar{X}_i \quad (48)$$

pois o primeiro é um estimador centrado, enquanto o segundo não é. Vejamos

$$E[\bar{X}_{st}] = \sum_1^k W_i \mu_i = \mu \quad \text{enquanto} \quad E[\bar{X}'] = \frac{1}{n} \sum_1^k n_i \mu_i \neq \mu$$

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

Vejamos agora

$$Var[\bar{X}_{st}] = \sum_1^k W_i^2 (1 - f_i) \sigma_i^2 / n_i$$

pois

$Var[\bar{X}_{st}] = \sum_1^k W_i^2 Var(\bar{X}_i)$ , visto que na amostragem estratificada os diferentes estratos as médias não estão correlacionadas, logo  $Cov(\bar{X}_i, \bar{X}_j) = 0$ .

### Observação:

Vimos que no caso proporcional  $\bar{X}_{st}$  e  $\bar{X}'$  coincidem, no entanto estes dois estimadores não apresentam a mesma variância. Efectivamente

$$Var[\bar{X}'] = \frac{1}{n^2} \sum_1^k n_i (1 - f_i) \sigma_i'^2.$$

Como exercício sugere-se a obtenção de expressões para a variância, em certos casos particulares:

1. Se  $f_i = \frac{n_i}{N_i}$  for desprezável  $Var[\bar{X}_{st}] = \sum_1^k W_i^2 \sigma_i'^2 / n_i$  ;
2. Se  $w_i = \frac{n_i}{n} = \frac{N_i}{N}$  -- caso proporcional  $Var[\bar{X}_{st}] = \frac{1-f}{n} \sum_1^k W_i \sigma_i'^2$  ;
3. Se a amostragem é proporcional e a variância é constante, i.e.,  $\sigma_i'^2 = \sigma^2$ , então  $Var[\bar{X}_{st}] = \frac{1-f}{n} \sigma^2$ .

### Estimação do Total da População

Um estimador centrado para o total  $X_T$  da população é

$$X_T^* = N \bar{X}_{st} = \sum_1^k N_i \bar{X}_i. \quad (49)$$

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

$$Var[X_T^*] = \sum_1^k N_i^2 (1 - f_i) \sigma_i'^2 / n_i. \quad (50)$$

### Intervalos de Confiança

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

$$\bar{x}_{st} - z_{\alpha/2} s'(\bar{x}_{st}) < \mu < \bar{x}_{st} + z_{\alpha/2} s'(\bar{x}_{st}) \quad (51)$$

e um intervalo de confiança para  $X_T$  a  $(1-\alpha)100\%$  é

$$N\bar{x}_{st} - z_{\alpha/2}Ns'(\bar{x}_{st}) < X_T < N\bar{x}_{st} + z_{\alpha/2}Ns'(\bar{x}_{st}) \quad (52)$$

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

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

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

$$Var[\bar{X}_{st}] < Var[\bar{X}]$$

Ora vejamos:

$$\text{Como sabemos} \quad Var[\bar{X}] = (1-f) \frac{\sigma^2}{n} \quad \text{e}$$

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

Numa primeira fase consideremos que estamos no caso de afecção proporcional,  $f_i = f$

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

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

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

$$\frac{N_i - 1}{N - 1} = \frac{N_i}{N} = \frac{N_i}{N - 1} \quad (53)$$

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

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

excepto se  $\mu_i$  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, 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 (53), deve considerar-se

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

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

Sendo assim, podemos dizer que

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

$$\sum_{i=1}^k N_i (\mu_i - \mu)^2 > \frac{1}{N} \sum_{i=1}^k (N - N_i) \sigma_i'^2 . \quad (54)$$

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

Nesta questão há dois pontos a ter em conta. Pretende-se saber como escolher a dimensão da amostra de modo a satisfazer uma certa precisão ou questões de custo .

Consideremos a situação de no processo de amostragem haver:

$C_0$  --- custo base da amostragem;

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

Sendo assim, o custo total  $C_T$  é dado por  $C_T = C_0 + \sum_1^k n_i c_i$  .

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

- minimizar  $Var(\bar{X}_{st})$ , para um custo total  $C_T$ ;
- minimizar o custo total, para um dado valor de  $Var(\bar{X}_{st})$ .

a) Variância mínima para custo fixo.

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

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

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

$$- \sum_{i=1}^k \left( W_i^2 \frac{\sigma_i'^2}{n_i^2} + \lambda c_i \right) = 0, \quad \text{onde para cada parcela se tem } n_i \sqrt{\lambda} = \frac{W_i \sigma_i'}{\sqrt{c_i}},$$

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  $n_i = \frac{W_i \sigma_i'}{\sqrt{c_i \lambda}}$  tem-se

$$n_i = \frac{(C_T - c_0) W_i \sigma_i' / \sqrt{c_i}}{\sum_{i=1}^k W_i \sigma_i' \sqrt{c_i}}, \quad (55)$$

sendo a dimensão total da amostra a recolher

$$n = \frac{(C_T - c_0) \sum_{i=1}^k W_i \sigma'_i / \sqrt{c_i}}{\sum_{i=1}^k W_i \sigma'_i \sqrt{c_i}} \quad (56)$$

Esta é a dimensão óptima da amostra a recolher em cada estrato para um custo total fixo. Observe-se que podemos resumir as seguintes observações:

-- As dimensões das amostras em cada estrato devem ser proporcionais ao tamanho do estrato; ao desvio padrão do estrato e inversamente proporcionais à raíz quadrado do preço unitário de amostragem em cada estrato.

### 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

$$n_i = n \frac{W_i \sigma'_i}{\sum_{i=1}^k W_i \sigma'_i} \quad \text{com} \quad n = \frac{C_T - c_0}{c} \quad (57)$$

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

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

$$Var_{\min} [\bar{X}_{st}] = \frac{1}{n} \left( \sum_{i=1}^k W_i \sigma'_i \right)^2 - \frac{1}{N} \sum_{i=1}^k W_i \sigma_i'^2. \quad (58)$$

### Custo mínimo para variância fixa

Consideremos  $Var[\bar{X}_{st}] = V$  e para este valor pretendemos saber qual a dimensão da amostra a recolher em cada estrato de modo a termos um custo mínimo.

Do que vimos atrás sabemos que  $Var[\bar{X}_{st}]$  é minimizada quando os  $n_i$  são escolhidos proporcionalmente a  $W_i \sigma'_i / \sqrt{c_i}$ . Sendo assim, 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  $Var[\bar{X}_{st}]$ , isto é,

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

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

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

Sendo assim deve tomar-se

$$n_i = \left\{ \frac{\sum_{i=1}^k W_i \sigma_i' \sqrt{c_i}}{V + \frac{1}{N} \sum_{i=1}^k W_i \sigma_i'^2} \right\} W_i \sigma_i' / \sqrt{c_i}. \quad (59)$$

Na expressão (56) encontramos a dimensão total de amostra a recolher no caso de afectação óptima. E no caso de pretendemos 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

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

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

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

## 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

$$Y_T = \sum_1^N Y_i \quad \text{donde} \quad P = \frac{\sum_{i=1}^N Y_i}{N} \quad (60)$$

Como estimador de  $P$  tem sentido considerar  $\bar{Y}_{st} = \sum_{i=1}^k \frac{N_i}{N} \bar{Y}_i = \sum_{i=1}^k W_i \hat{P}_i = \hat{P}_{st}$ ,

onde  $\bar{Y}_i = \hat{P}_i$  designa a proporção de indivíduos no estrato  $i$ , incluídos na amostra e verificando  $A$ . O estimador de  $P$  é tal que

$$\begin{aligned} E[\hat{P}_{st}] &= P \\ \text{Var}[\hat{P}_{st}] &= \sum_{i=1}^k \frac{W_i^2}{n_i} \left( \frac{N_i - n_i}{N_i - 1} \right) P_i (1 - P_i) \end{aligned} \quad (61)$$

Um estimador desta variância é :

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

Se  $N_i$  grande tem-se

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

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

se

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

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



$$n_i = \frac{nW_i\sqrt{P_iQ_i}}{\sum W_i\sqrt{P_iQ_i}}. \quad (62)$$

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

$$n_i = \frac{(C_T - c_0)W_i\sqrt{P_iQ_i/c_i}}{\sum W_i\sqrt{P_iQ_i c_i}}. \quad (63)$$