Amostragem e Análise Ambiental
2021/2022

Módulo
Amostragem e Reamostragem

Manuela Neves

Técnicas de Amostragem

1. Recolher dados ambientais: amostragem e monitorização.

A recolha de dados de modo eficaz e adequado é de importância central no estudo de problemas ambientais.

O acto de conduzir um estudo começa com a obtenção de uma **amostra** de valores de uma variável de interesse – **variável aleatória – característica em estudo.**

Conceitos

- **Amostragem** é todo o processo de recolha de uma parte (representativa), geralmente pequena, dos elementos que constituem um dado **conjunto (população) finito** . Da análise, tratamento e interpretação dessa parte pretende obter-se informações para todo o conjunto.

População é o conjunto de todos os "indivíduos" ou unidades estatísticas sobre os quais se pretende estudar alguma característica, como por exemplo pessoas, alojamentos, escolas, empresas, cidades, regiões, países… A população é conceptualizada por um **modelo**.

- **população objectivo** -- é a totalidade dos elementos em estudo e relativamente aos quais se pretende obter certo tipo de informação.

- **população inquirida** -- aquela sobre a qual é efectivamente feita a amostragem.

- **Amostra** é um subconjunto de elementos extraídos da população

- **Unidade de amostragem** ou **unidade estatística** -- é o elemento sobre o qual vai ser estudada a **característica** de interesse.

As características são as **variáveis de interesse**, que podem ser de **natureza quantitativa** e neste caso consideram-se escalas numéricas nas quais as variáveis se podem **"medir".**

-- contínuas
-- discretas

ou de **natureza qualitativa** e neste caso **"observam-se"** atributos.

-- nominais

-- ordinais

Exemplo

Pretendemos registar o valor do **diâmetro de uma árvore** ou a **taxa de propagação de um fogo florestal**. O objecto sobre o qual se recolhe os dados chama-se ***unidade de amostragem*.**

Nos exemplos referidos a unidade de amostragem é – a **árvore** e o **fogo florestal**, respectivamente.

1º caso -- A população é o conjunto de todos os diâmetros de todas as árvores de uma espécie.

2º caso – A população é a taxa de propagação de todos os fogos florestais num país (região).

Rate of spread is the horizontal distance that the flame zone moves per unit of time (feet per minute) and usually refers to the **front** or **head** of the fire segment of the fire perimeter. However, rate of spread can be measured from any point on the fire perimeter in a irection that is perpendicular to the perimeter. Because rate of spread can vary significantly over the area of the fire, it is generally taken to be an average value over some given period of time. (Em http://www.forestencyclopedia.net/p/p478)

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

- **Parâmetro –** característica numérica da população

  Exemplo: **valor médio, variância, o mínimo, o máximo, a amplitude, o total…**

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

- **Estimador -** uma **função da amostra aleatória** que não contém parâmetros desconhecidos e que é usada para estimar um parâmetro

- **Estimativa –** é o **valor de um estimador** calculado usando uma amostra concreta.

Nota: Parâmetros e estimativas são **números** → a diferença é que um refere-se à população e outro à amostra observada.

Não sabemos, e regra geral não é possível conhecer o valor dos parâmetros – total de animais selvagens de uma dada espécie que existem numa região

A **teoria da amostragem** seguida da **inferência estatística** permitem-nos tomar decisões (tirar conclusões) sobre a população.

Como dissémos a **Amostragem** é o processo de recolha de uma parte representativa dos elementos que constituem **um população finita**

***Exemplo*:** Numa mata quer saber-se como estão a desenvolver-se as árvores:

- quantas variedades há;
- diâmetro médio ao fim de 10 anos
- proporção de árvores com uma dada doença
- etc.

Mas outros problemas surgem na área **do ambiente**

- A **poluição** é um dos maiores questões do nosso tempo -- surge de muitas formas:
  - descarga de substâncias radioactivas em termos de contaminação da água;
  - a exposição diária aos fumos do tráfego.
- **Poluição da água** ou do ar em certas áreas pode ser medida regularmente numa vasta rede de estações de monitorização. Isto pode ser feito retirando uma amostra ao acaso dessa rede de locais

- **Poluição sonora…**

Mas vamos agora falar de amostragem relembrando alguns conceitos que já demos na **Estatística** do semestre passado

…

apenas para ajudar a arrumar as ideias…

Amostragem de uma população finita

Notações:

População $\mathbb{P}$, constituída por **N** "indivíduos".

X a **característica em estudo** que supomos tomar os seguintes valores

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

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

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

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

ou <u>variância corrigida</u> definida como $$\sigma_X^{'2} = \sum_{i=1}^{N} \frac{(A_i - \mu)^2}{N-1} = \frac{N}{N-1}\sigma_X^2$$

Total $$T = X_T = \sum_{i=1}^{N} A_i = N\mu$$

Razão de dois totais $$X_T / Y_T$$

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

Uma **amostra aleatória simples** é um subconjunto de indivíduos (a amostra) seleccionado totalmente ao acaso a partir de um conjunto maior (a população) por um processo que garanta que:

1. Todos os indivíduos da população têm a mesma probabilidade de ser escolhidos para a amostra—diz-se **com reposição**
2. Cada subconjunto possível de indivíduos (amostra) tem a mesma probabilidade de ser escolhido que qualquer outro subconjunto de indivíduos – diz-se **sem reposição**

Amostragem aleatória simples com reposição (uma breve revisão)

População com N elementos, num processo de amostragem com reposição, cada elemento tem a mesma probabilidade $1/N$ de ser seleccionado.

Qualquer amostra de dimensão n tem probabilidade $1/N^n$ de ser seleccionada.

Seja então $X_1, X_2, ..., X_n$ **uma amostra aleatória** retirada com reposição de uma população com N elementos $A_i \ (i = 1, ..., N)$ e

$x_1, x_2, \ldots, x_n$ a correspondente amostra observada.

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

Um estimador centrado para μ $\overline{X} = \sum_{i=1}^{n} \frac{X_i}{n}$

Tem-se

$$Var(\overline{X}) = \frac{\sigma^2}{n}.$$

A uma **estimativa** de $\sqrt{Var(\overline{X})} = \frac{\sigma}{\sqrt{n}}$ chama-se **erro padrão** da média

Há então que considerar um estimador de σ^2

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

Efectivamente

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

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

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

É necessário calcular **a dimensão da amostra a recolher**, de modo a obter a estimativa de interesse, com **um erro inferior a ε,** 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.

O intervalo de confiança para μ a (1-α)100%

no caso de uma amostra aleatória obtida com reposição, determinado com base numa amostra de dimensão ***n*** , supondo a normalidade da população subjacente é

$$\left]\bar{x} - t_{\alpha/2(n-1)}\frac{s'}{\sqrt{n}}, \bar{x} + t_{\alpha/2(n-1)}\frac{s'}{\sqrt{n}}\right[$$

Sendo assim, fixado o nível de precisão ou erro de amostragem (ε) e o nível de confiança (1-α) ou o risco (α)

Para determinar a dimensão da amostra a recolher por forma a termos um erro inferior a ε basta exigir que

$$t_{\alpha/2}\frac{s'}{\sqrt{n}} \leq \varepsilon \Rightarrow n \geq \left(\frac{t_{\alpha/2}\, s'}{\varepsilon}\right)^2 \qquad (8)$$

Como calcular **o valor $t_{\alpha/2}$???**

é necessário saber o número de graus de liberdade (n-1), e consequentemente a dimensão da amostra, que é afinal aquilo que pretendemos calcular!!!!!!!

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 ε, a amostra final deverá ter um número de elementos n, dado por

$$n \geq \left(\frac{t_{\alpha/2}\, s'_1}{\varepsilon} \right)^2 \left(1 + \frac{2}{n_1} \right)$$

Se o valor resultante para n é tal que n/N é apreciável

(>5% ou >10%), deve considerar-se como dimensão de amostra a recolher o valor dado por

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

Note-se que há aqui dois conceitos cruciais:

Exactidão (*accuracy*) (**medida pelo erro quadrático médio**)– avalia a proximidade de um estimador ao verdadeiro valor do parâmetro.

Precisão (*precision*) (medida pela variância) -- é o "grau de concordância" de um conjunto de medições entre si.

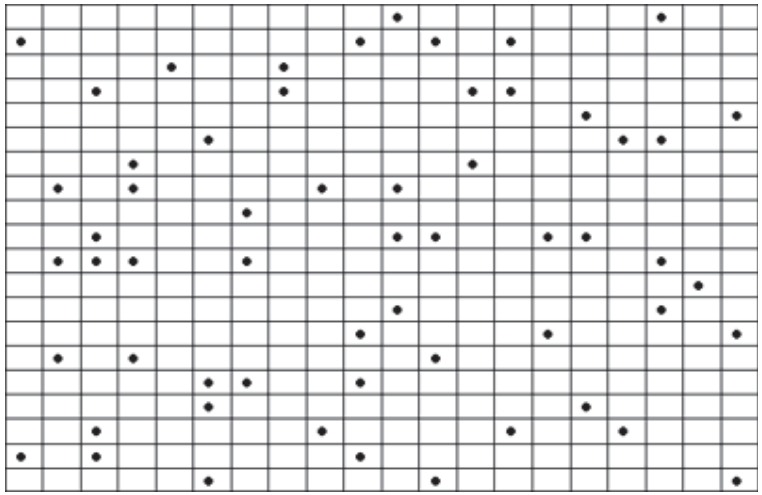
Viés (*bias*) – reflecte a tendência das medições se desviarem do verdadeiro valor sistematicamente numa direcção --- **erros sistemáticos.** Estes erros são causados pela falta de calibração dos instrumentos.
Em contrapartida **erros aleatórios** são devidos a flutuações estatísticas que ocorrem quando se mede uma quantidade. Estes produzem uma dispersão dos valores em torno da valor do centro.

Só uma nota sobre delineamento experimental

We could measure the effects of a fertilizer on tree growth on a national forest, but we would have little control over temperature, humidity, insects etc… where as if we conducted the same experiment in a green house we could control all of these variables.

Mas na **amostragem aleatória simple**s o procedimento mais comum e adequado (veremos mais adiante porquê) é o que

- considera que qualquer combinação possível de unidades experimentais tem a mesma probabilidade de ser recolhida
- e não cada unidade experimental é a que tem a mesma probabilidade de ser recolhida.



Amostragem aleatória simples sem reposição

Relembre-se que se designou por

$A_1, A_2, \ldots, A_N$ os valores da característica em estudo na população

Agora 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 **se podem considerar** ***praticamente*** **independentes**.

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

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

Para facilitar consideremos as seguintes variáveis indicatrizes:

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

portanto a média amostral é

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

Vamos então calcular o **valor médio** e a **variância** de $\overline{X}$.
Só umas notas….

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

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

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

Portanto $\overline{X}$ **é estimador centrado de** μ**.**

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

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

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

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

Ora

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

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

o que, após pequenos cálculos dá

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

Por curiosidade vejamos que a correlação é assim dada.

$$\rho\left(I_i, I_j\right) = \frac{Cov(I_i, I_j)}{\sqrt{Var(I_i)Var(I_j)}} = -\frac{1}{N-1}.$$

Observe-se que a covariância tende para zero quando $N \to \infty$, 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

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

$$= \frac{1}{n^2}\left[\sum_{j=1}^{N} A_j^2 \frac{n}{N}\left(1-\frac{n}{N}\right) - \sum_{j \neq k} A_j A_k \frac{n}{N}\left(1-\frac{n}{N}\right)\frac{1}{N-1}\right]$$

Atendendo a que

$$\sum_{k \neq j} A_j A_k$$

se pode escrever como

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

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

$$Var(\bar{X}) = \frac{1}{n^2}\left[\frac{n}{N}\left(1-\frac{n}{N}\right)\sum A_j^2 - \frac{n}{N}\left(1-\frac{n}{N}\right)\frac{1}{N-1}\left(N^2\mu^2 - \sum A_j^2\right)\right] =$$

$$= \frac{1}{n^2}\frac{n}{N}\left(1-\frac{n}{N}\right)\left[\sum A_j^2 - \frac{1}{N-1}\left(N^2\mu^2 - \sum A_j^2\right)\right] = \frac{N-n}{N^2 n}\frac{N\sum A_j^2 - N^2\mu^2}{N-1} =$$

$$= \frac{N-n}{N-1}\frac{1}{n}\frac{\sum A_j^2 - N\ \mu^2}{N} = \frac{N-n}{N-1}\frac{\sigma^2}{n}.$$

Observe-se

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

$$Var(\bar{X}) \quad < \quad Var(\bar{X})$$

s/ reposição **c/ reposição**

que

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 $\boxed{\frac{N-n}{N-1}}$ chama-se **correcção de população finita** e a

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

A expressão da variância acima deduzida pode ser apresentada usando a variância corrigida σ'^2, isto é,

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

Vimos que no caso da **amostragem com reposição** S'^2 era um estimador centrado de σ^2,

Pode ver-se que no caso da **amostragem sem reposição** S'^2 é estimador centrado de σ'^2.

Ora

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

$$=\frac{1}{n-1}\left[\sum E(X_i-\mu)^2-nE(\overline{X}-\mu)^2\right]=\frac{1}{n-1}\left[n\sigma^2-n\frac{N-n}{N-1}\frac{\sigma^2}{n}\right]=$$

$$=\frac{1}{n-1}\left[\frac{n\sigma^2(N-1)-(N-n)\sigma^2}{N-1}\right]=\frac{1}{n-1}\left[\frac{N(n-1)\sigma^2}{N-1}\right]=\frac{N}{N-1}\sigma^2=\sigma'^2$$

logo S'^2 é estimador centrado de σ'^2 na amostragem sem reposição.

Neste caso **o erro padrão** do estimador $\overline{X}$ é:

$$s'\sqrt{\frac{1-f}{n}}$$

Intervalos de confiança para μ

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 μ=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

$$\bar{x} = 9.8 \qquad \text{e} \qquad Var(\overline{X}) = (1-f)\frac{\sigma^2}{5} = 1.9872$$

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

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

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

Assim pode considerar-se

$$\overline{X} \approx N\left(\mu, (1-f)\sigma'^2 / n\right)$$

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

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

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 μ.

Nas condições atrás referidas **um intervalo a (1-α)100% de confiança para μ** será então

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

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

Porém na prática σ **não é conhecido** e sendo assim considera-se s' como uma estimativa para σ, 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 σ, o melhor é usar a distribuição t, donde um

intervalo a $(1-\alpha)100\%$ de confiança para μ será então

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

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 μ de modo a ter uma precisão *d*.

Pretende-se então que

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

Vimos já que o intervalo de confiança a (1-α)100% para μ era

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

Basta então exigir que

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

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

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

isto é, a dimensão da amostra é

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

Como primeira aproximação para n, regra geral, considera-se

$$n \ge n_0 = \left(\frac{z_{\alpha/2}\,\sigma'}{d}\right)^2.$$

No caso de 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 σ, 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 σ.

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 σ'^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 σ'^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,

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

Estimação do total T

$$T = X_T = N\mu$$

o estimador mais usado é

$$X_T^* = N\,\overline{X}$$

$$E\left[X_T^*\right] = N\,\mu = X_T \qquad \text{e} \qquad Var\left[X_T^*\right] = N^2\,(1-f)\,\frac{\sigma'^2}{n}$$

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

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

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

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

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 α, pretende-se que

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

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

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

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

donde se tem

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

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 $n \geq n_0\left(1+\frac{n_0}{N}\right)^{-1}$.

Estimação de uma proporção P

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

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

Sendo assim uma **estimativa de P**, pode ser dada por

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

Propriedades para o estimador $\hat{P}$???

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:

$$Y_i = \begin{cases} 1 & \text{se o elemento da pop. verifica a propriedad e A} \\ 0 & \text{se o elemento da pop. não verifica a propriedad e A} \end{cases}$$

$$, \quad Y_Y = \sum_{1}^{N} Y_i = K$$

K é o número de elementos da população que verificam A.

$$\mu_Y = \frac{\sum_{1}^{N} Y_i}{N} = \frac{K}{N} = P$$

P é então a média da variável Y na população, para a qual

$$\sigma_Y^{'2} = \frac{\sum_{1}^{N} (Y_i - \mu_Y)^2}{N-1} = \frac{\sum Y_i^2 - N\mu_Y^2}{N-1} = \frac{NP - NP^2}{N-1} = \frac{NP(1-P)}{N-1}$$

Para **estudar o 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 $Y_1, Y_2, \ldots, Y_n$, cuja média é

$$\bar{Y} = \frac{\sum_{1}^{n} Y_i}{n} = \frac{\hat{K}}{n} = \hat{P},$$

$\hat{P}$ será então a média da amostra aleatória .

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

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

Vejamos a variância de $\bar{Y} = \hat{P}$

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

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 σ'^2.

Então terá que ser estimado, usando o **estimador centrado de** σ'^2**,** S'^2

Uma estimativa é

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

Sendo um estimador centrado de $Var\left[\hat{P}\right]$

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

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

Se f é desprezável, tem-se

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

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 μ ou X_T.

A **distribuição exacta de $\hat{P}$ é conhecida**.

O número K de elementos da amostra que possuem aquele atributo, tem distribuição hipergeométrica, i.e.,

$$P[\text{haver } k \text{ elementos }] = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}; \qquad \max(0, n-K+N) \le r \le \min(K, n)$$

Mas… os cálculos que esta distribuição envolve são pesados

É portanto útil procurar aproximações para a distribuição do estimador, agora num espírito mais pragmático do que teórico.

Possibilidades:

--- usar a distribuição binomial como uma aproximação da hipergeométrica

se n é pequeno relativamente a K e a $(N\text{-}K)$, a "falta de reposição" pode ser "ignorada", donde

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

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,

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

A **aproximação à normal é razoável desde que**:

-- n não seja muito grande relativamente a K e a $N\text{-}K$.

-- 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**

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

Note-se que $\text{var}(\hat{P})$ é substituída pelo seu estimador centrado

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

Escolha do tamanho da amostra para estimar uma proporção

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

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

A $Var\left[\hat{P}\right]$ 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, pode pretender-se:

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

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

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

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

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

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 ε.

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

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

ou seja, considerando a aproximação pela normal, viria

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

Aqui, na determinação de *n* (dimensão da amostra), temos uma facilidade que não tinhamos no caso da estimação de μ 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

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

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

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

Podemos tomar como primeira aproximação

$$n_0 \geq \frac{PQ\, z_{\alpha/2}^2}{d^2}$$

porém se $\frac{n_0}{N}$ é grande, deve considerar-se

$$n \geq n_0 \left(1 + \frac{n_0 - 1}{N}\right)^{-1}$$

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

Ora sabe-se que

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

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

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

Como primeira aproximação pode considerar-se

$$n_0 \geq \frac{Q\, z_{\alpha/2}^2}{P\, \xi^2}.$$

De novo se $\frac{n_0}{N}$ é grande, deve considerar-se

$$n \geq n_0 \left(1 + \frac{n_0 - 1}{N}\right)^{-1}$$

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

$$R = X_T / Y_T = \mu_X / \mu_Y .$$

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

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

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

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

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

$$Var\left[R^*\right] \cong \frac{1-f}{n\bar{y}^2} \sum_1^N \frac{(X_i - RY_i)^2}{N-1} = \frac{1-f}{n\bar{y}^2}\left[\sigma'^2_X - 2R\sigma'_{XY} + R^2\sigma'^2_Y\right]$$

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

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

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

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

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

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, μ_X

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

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

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

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

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 $Var\left[\overline{X}_R\right] < Var\left[\overline{X}\right]$?

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 aleatória Estratificada

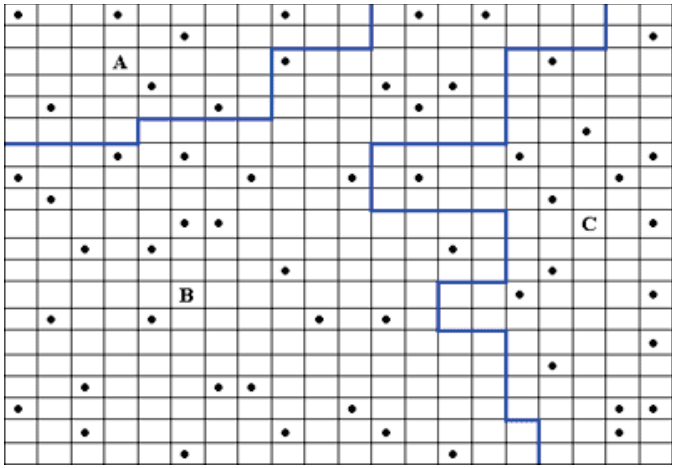
Exemplo

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

Variância da população

$$\sigma'^2 = \frac{1}{N-1}\left\{\sum_1^k (N_i - 1)\sigma_i'^2 + \sum_1^k N_i(\mu_i - \mu)^2\right\}$$

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

De facto tem-se

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

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

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

Para cada estrato *i* tem-se

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

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

$$n_1, n_2, \ldots, n_k \qquad \left(\sum_i^k n_i = n\right)$$

tendo como elementos em cada estrato i

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

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

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

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? **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 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

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

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

Estimação do valor médio

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

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

A média empírica estratificada ≠ 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.

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

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

Vejamos agora

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

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

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

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

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

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

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

Intervalos de Confiança

Um intervalo de confiança para μ a (1-α)100% é

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

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

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

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

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

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

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

Temos:

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

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

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

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

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

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

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

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

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

Sendo assim, podemos dizer que

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

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

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 do tamanho da amostra a recolher

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

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

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

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

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

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

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 $\boxed{n_i\sqrt{\lambda}=\dfrac{W_i\ \sigma_i'}{\sqrt{c_i}}}$,

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

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

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

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

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

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

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

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

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

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

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

Sendo assim deve tomar-se

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

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

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

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

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

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

$$E\left[\hat{P}_{st}\right] = P$$

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

Um **estimador desta variância** é :

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

Se N_i grande tem-se

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

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

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

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

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

Dimensão total da amostra

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

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

$$n_i = \frac{nW_i\sqrt{P_iQ_i}}{\sum W_i\sqrt{P_iQ_i}}.$$