

Amostragem e Análise Ambiental 2019/2020

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 direction 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 dissemos 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 lembrando 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 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:

Valor Médio

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

Variância

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

ou variância corrigida 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

$$\frac{X_T}{Y_T}$$

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, \dots, X_n **uma amostra aleatória** retirada com reposição de uma

população com N elementos 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 μ

$$\bar{X} = \sum_{i=1}^n \frac{X_i}{n}$$

Tem-se

$$\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$$

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

Há então que considerar um estimador de σ^2

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

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

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 $(1-\alpha)100\%$

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 $(1-\alpha)$ ou o risco (α)

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

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

Como 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!!!!!!

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.



Accuracy & Precision



Accurate
but , not precise



Precise
but , not accurate



Accurate
and **Precise**

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.

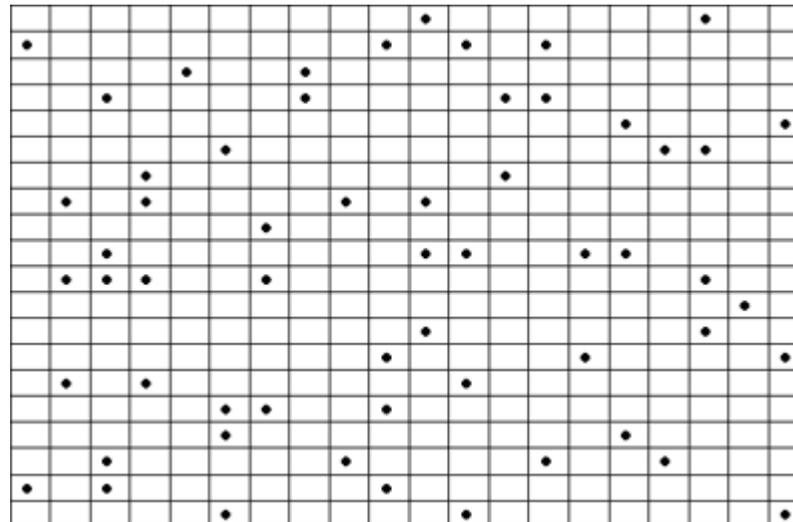
Só uma nota sobre delineamento experimental

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 simples** 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, \dots, 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**.

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

Seja novamente (X_1, X_2, \dots, 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

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

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

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

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

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

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

Calculemos agora a variância de \bar{X} .

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

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

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

Ora

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

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

o que, após pequenos cálculos dá

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

Por curiosidade vejamos que a correlação é assim dada.

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

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

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

Observe-se

$$\begin{aligned}
\frac{N-n}{N-1} \frac{\sigma^2}{n} &< \frac{\sigma^2}{n} \quad \text{isto é} \\
\text{Var}(\bar{X}) &< \text{Var}(\bar{X}) \\
\text{s/ reposi\c{c}\~{a}o} & \quad \quad \quad \text{c/ reposi\c{c}\~{a}o}
\end{aligned}$$

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 $\frac{N - n}{N - 1}$ 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 σ'^2 , isto é,

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

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

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

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

Neste caso o **erro padrão** do estimador \bar{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 $\mu=8.44$ 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

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

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

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

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

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

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 μ .**

Nas condições atrás referidas **um intervalo a $(1-\alpha)100\%$ 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 $P(|Z| > z_{\alpha/2}) = \alpha$.

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

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

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

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

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

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

isto é, a dimensão da amostra é

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

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 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

$$X_T^* = N \bar{X}$$

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

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

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

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

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

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

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

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

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 r de indivíduos** que satisfazem a propriedade.

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

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

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:

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

$$Y_T = \sum_1^N Y_i = R,$$

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

$$\mu_Y = \frac{\sum_1^N Y_i}{N} = \frac{R}{N} = P$$

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

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

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, \dots, Y_n , cuja média é

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

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

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

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

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

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

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

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

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

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

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 R de elementos da amostra que possuem aquele atributo, tem distribuição hipergeométrica, i.e.,

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

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 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,

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

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**

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

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 $\hat{P} = \frac{\hat{R}}{n}$ 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** , 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

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

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 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

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

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

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

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

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

Podemos tomar como primeira aproximação

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

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

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

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

Ora sabe-se que

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

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

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

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