Amostragem

Objetivo:

- identifica e delineia os principais planos de amostragem;
- determina a dimensão da amostra mais adequada;
- utiliza software como meio de cálculo e como instrumento para a análise e tratamento estatístico de dados.

Tópicos a tratar:

- Amostragem aleatória e intervalos de confiança
- amostragem não-aleatória
- Determinação da dimensão da amostra

Amostragem – motivação (problema)

Qual a idade média dos estudantes da Universidade de Aveiro?

Duas hipóteses de abordagem:

- avaliar a idade de todos os estudantes (e dividir a soma pelo número total de estudantes)
- avaliar a idade de alguns estudantes tomados ao *acaso* (e dividir pelo número de estudantes inquiridos)

Discussão:

- a primeira solução parece ser a mais segura, mas é pouco cómoda e a sua precisão é ilusória; seria apenas uma "fotografia" e não isenta de erro porque a população altera-se (há novos estudantes e cancelamentos de matrícula)
- a segunda opção, apesar de considerar uma amostra, é *rigorosa* na medida em que as idades avaliadas são tomadas ao *acaso* e o número de idades avaliadas é *grande*

Conceitos elementares - revisão

População

Conjunto de indivíduos ou objetos que apresentam em comum determinadas características definidas para um estudo.

Amostra

Um subconjunto da população.
A dificuldade, e nalguns casos, a impossibilidade de estudar a população na totalidade justifica a relevância do estudo de uma população com recurso a amostras.

Conceitos elementares - revisão

Sondagem

Metodologia de pesquisa que permite o conhecimento momentâneo de uma população, numa perspetiva e quantificada. A recolha e a análise dos dados é realizada com base numa **amostra** de elementos que *deverá* permitir a extrapolação das interpretações à totalidade da população.

Censo/Recenseamento

Metodologia de pesquisa que capta a realidade de uma população num determinado momento.
É uma fotografia de todos os elementos de uma população.
Exemplos: Recenseamento Eleitoral, Recenseamento Militar e Recenseamento da População e Habitação

Sondagem *versus* Censo/Recenseamento

Vantagens de uma **sondagem** relativamente ao recenseamento:

- **custo** – mais económica porque os recursos e os meios são menores
- **tempo** – processo mais rápido na recolha e no tratamento
- **tipo de informação** – uma amostragem é a melhor opção quando se pretende características da população que não se reduzem a factos (p.e. opiniões, expetativas) porque permite um questionário mais detalhado
- **credibilidade dos dados** – um recenseamento não está isento de erros (respostas incompletas ou falsas, trabalho deficiente dos entrevistadores, etc.); as teorias estatística e de amostragem têm técnicas para lidar com os erros de amostragem e para se reconhecer a validade da metodologia de investigação

- por vezes uma sondagem é a única forma de estudar uma população, por exemplo, nem sempre todos os elementos estão acessíveis
- o recenseamento pode ser a opção correta quando a população é de pequena dimensão e é facilmente acessível

Qualidade numa sondagem e representatividade da amostra

Segundo Vicente et al. (2001), a *qualidade* de uma sondagem pode ser olhada de duas óticas:

ótica do cliente – aquele que solicita a sondagem - a qualidade traduz-se na credibilidade, precisão e validade da informação que lhe é fornecida;

ótica do produtor – a entidade que leva a cabo a sondagem - a qualidade associa-se à condução do processo de conceção e realização da sondagem por forma a obter resultados credíveis, precisos e válidos.

Uma **amostra representativa** do universo

- é uma espécie de maquete (*esboço em escala de redução*) que reflete, para o estudo concreto, as características mais relevantes da população, e consequentemente,
- as conclusões estatísticas obtidas a partir dela são similares às que se obteriam se a análise recaísse sobre todos os elementos da população.

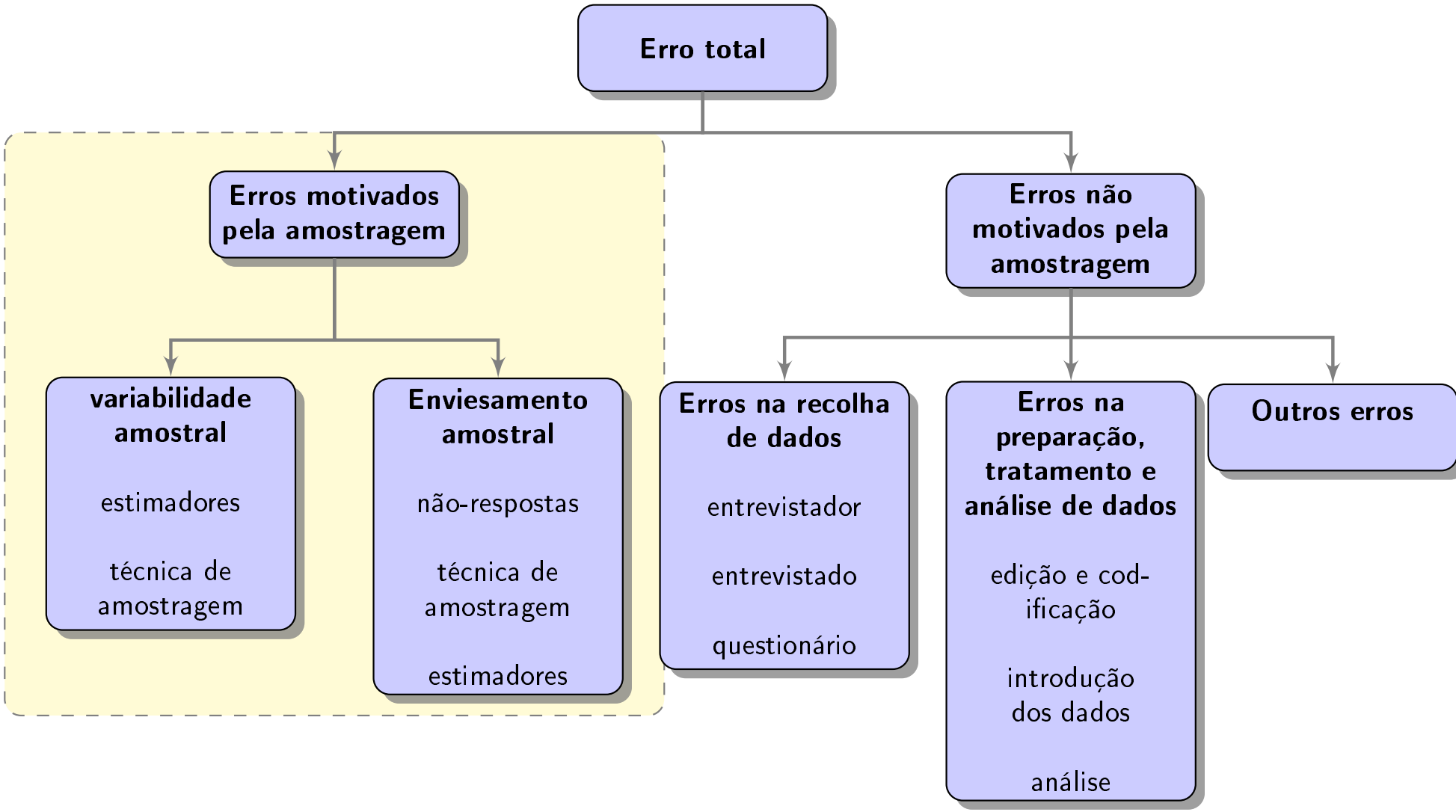
Exemplo – Eleições Autárquicas 2021

Resultados oficiais e algumas sondagens relativos às Eleições Autárquicas de 26 de setembro de 2021.

Data da divulgaç	Instituto	Fernando Medina (PS/Livre)	Carlos Moedas (PSD/CDS/MPT/PPM)	João Ferreira (CDU)	Beatriz Dias (BE)	Bruno Horta Soares (IL)
26/09/2021	**Resultados Oficiais**	**33,31%**	**34,26%**	**10,51%**	**6,20%**	**4,22%**
26/09/2021	Universidade Católica	31%-35%	32%-36%	10%-13%	5%-7%	3%-5%
26/09/2021	ICS-ISCTE	31,3%-36,3%	30,2%-35,2%	10,4%-13,4%	5,7%-8,7%	3,2%-6,2%
26/09/2021	Pitagórica	32,6%-38,6%	29,3%-35,3%	6,6%-12,6%	4,2%-8,2%	3,3%-7,3%
26/09/2021	Intercampus	33,2%-37,6%	29,9%-34,3%	10,1%-13,7%	5,4%-8,4%	2,8%-5,8%
23/09/2021	Pitagórica	40,6%	33,1%	7,6%	4,7%	3,2%
22/09/2021	Universidade Católica	37,0%	28,0%	11,0%	7,0%	3,0%
08/09/2021	Pitagórica	39,8%	32,6%	8,5%	6,8%	3,0%

Origens:
https://www.cmjornal.pt/politica/detalhe/medina-e-moreira-reeleitos-em-lisboa-e-porto-santana-regressa-a-figueira-ps-perde-coimbra
https://www.rtp.pt/noticias/autarquicas-2021/autarquicas-2021-projecoes-da-universidade-catolica_n1345928

Erros das sondagens



Etapas de um plano amostral

- **Definição da população alvo** – especificação de limites geográficos e temporais do estudo, definição da unidade amostral (pessoas, famílias, empresas, peças, ...)
- **Organização da base de sondagem** – listagem dos elementos da qual se vai selecionar a amostra; por vezes é difícil ou impossível constituir estas listagens (o universo que se pretende estudar nem sempre coincide com o efetivamente estudado ...)
- **Definição de um processo de amostragem** – seleção de um procedimento amostral / método de amostragem
- **Determinação da dimensão da amostra** – n.º de elementos a incluir na amostra depende, entre outros fatores, da homogeneidade da população, orçamento, tempo e recursos disponíveis e a precisão esperada dos resultados
- **Seleção dos elementos e recolha da informação** – após a identificação dos elementos a incluir na amostra há que *contactá-los*, conforme a sua natureza, no sentido da recolha dos dados

Notações e definições básicas

- $\mathcal{P}$ - População de dimensão N de **indivíduos** u_i, $i = 1, ..., N \rightarrow u_1, u_2, ..., u_N$
- X **característica** em estudo associada à população
- $X(u_i) = x_i$ valor da característica X no indivíduo u_i

  $x_1, x_2, ..., x_N \rightarrow$ medidas da característica X na população
- Características usualmente estudadas:
  - média da população $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$
  - variância da população $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$
  - desvio padrão populacional $\sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$
  - proporção π de membros da população que pertencem a determinada categoria de classificação de X

Notações e definições básicas

N é elevado (geralmente) $\Longrightarrow$ inquérito por amostragem

Objetivo: estimar 1 ou mais características a partir de uma amostra **concreta** de n $(n < N)$ elementos da população $\rightarrow x_1,\ x_2,\ \ldots,\ x_n$

$$\frac{n}{N} \rightarrow \textbf{fração de amostragem}$$

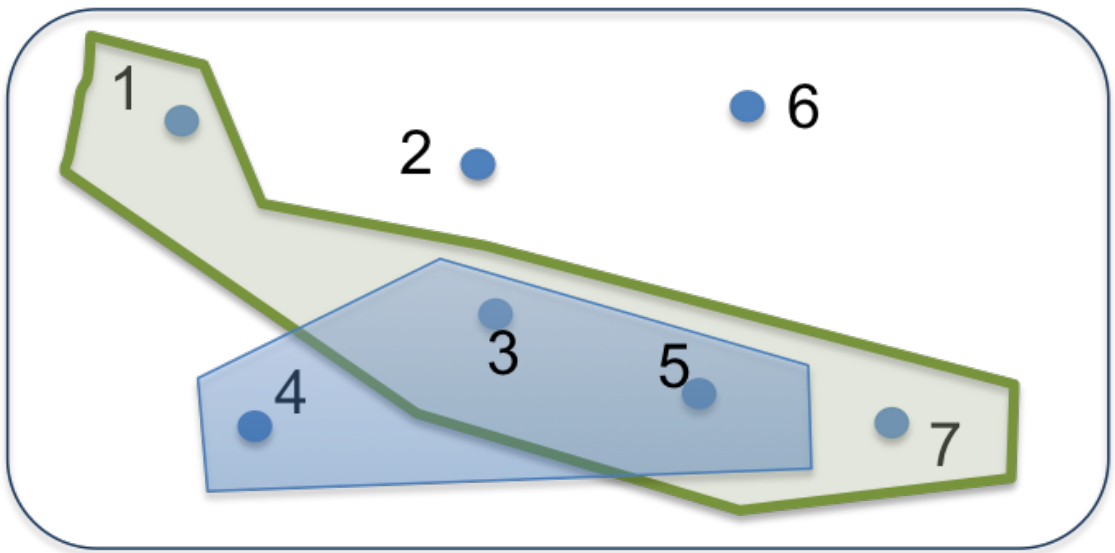
- média da amostra $\widehat{\mu} = \overline{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$
- variância da amostra $\widehat{\sigma}^2 = s^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$
- desvio padrão amostral $\widehat{\sigma} = s = \sqrt{s^2}$
- frequência relativa $\widehat{\pi} = f$ de membros da amostra que pertencem a determinada categoria de classificação de X

Exemplo

Seja $\mathcal{P} = \{1, 2, 3, 4, 5, 6, 7\}$ a população a estudar e duas amostras

$A = \{1, 3, 5, 7\}$ e $B = \{3, 4, 5\}$

$N = 7,\ n_A = 4,\ n_B = 3$



Exemplo

parâmetro (população)
μ – média populacional
σ – desvio padrão populacional
π_{par} – %. de elem. pares da pop.

estatística (amostra concreta)
$\widehat{\mu} = \overline{x}$ média amostral
$\widehat{\sigma} = s$ desvio padrão amostral
$\widehat{\pi}_{par} = f_{par}$ % dos elem. pares da amostra

parâmetros	estimativas e erros amostrais			
$\mu = 4$	$\overline{x}_A = 4$	erro $= 0$	$\overline{x}_B = 4$	erro $= 0$
$\sigma = 2$	$s_A = 2,57$	erro $= -0,57$	$s_B = 1$	erro $= 1$
$\pi_{par} = 42,86\%$	$f_{par,A} = 0\%$	erro $= 43\%$	$f_{par,B} = 33\%$	erro $= 10\%$

- as amostras A e B são representativas de $\mathcal{P}$ em relação à média μ
- a amostra A é mais representativa do que B em relação ao desvio padrão σ
- a amostra B é mais representativa do que A em relação à proporção π_{par}

Amostragem aleatória *versus* não aleatória

Como escolher uma amostra de modo a que esta seja **representativa** da população?

- Uma amostra é **aleatória** ou probabilística se for recolhida por um processo tal que assegura que todo e qualquer elemento (ou grupo de elementos) da população tem probabilidade calculável e diferente de zero de ser selecionado.
- O termo aleatório toma aqui o seu significado estatístico – cada elemento da população tem oportunidade de ser escolhido, ou seja, nenhum elemento é *a priori* excluído.
- Com um processo **não aleatório** há elementos da população que não têm possibilidade de serem escolhidos. A amostra não aleatória surge quando a inclusão dos elementos é determinada por um critério subjetivo (por exemplo, uma opinião pessoal) e não pela rigorosa aplicação da **teoria das probabilidades**.

Num processo de amostragem o que é verdadeiramente importante é a obtenção de uma amostra que seja **representativa**, que pode não ser necessariamente aleatória.

Amostra aleatória – notação

Dada uma população $\mathcal{P}$ com N elementos, várias amostras aleatórias de dimensão $n < N$ podem ser obtidas.

Ao quociente $FA = \frac{n}{N}$ dá-se o nome de fração de amostragem

Quando nos referimos a uma amostra, de entre as possíveis, e não a uma amostra em concreto, falamos numa **amostra aleatória** $\rightarrow (X_1, X_2, ..., X_n)$.

Relativamente a uma amostra aleatória falamos em **estimadores** (variáveis aleatórias):

- média amostral, $\overline{X}$
- variância amostral, S^2
- desvio padrão amostral, S
- frequência relativa de uma característica na amostra, f

Em regra, usamos **letras gregas** para os parâmetros (valores numéricos da população), **letras maiúsculas** para os estimadores associados a amostras aleatórias, e **letras minúsculas** para os valores observados dos estimadores (estatísticas ou **estimativas**) para uma amostra em concreto.

Métodos de amostragem aleatórios

- Amostragem aleatória simples (com e sem reposição)
- Amostragem sistemática
- Amostragem estratificada
- Amostragem por *clusters*, aglomerados ou grupos
- Amostragem multi-etápica ou multi-etapas
- Amostragem multi-fásica ou multi-fases

Tarefa

Realização da tarefa sobre amostragem

Amostragem aleatória simples (com reposição)

Dada uma população de dimensão N, referir-nos-emos a uma amostra aleatória, de dimensão n, com reposição, como um conjunto dos N^n conjuntos diferentes de n elementos, todos com igual probabilidade de serem selecionados.

- Na amostragem com reposição, sempre que um elemento é selecionado, é reposto na população.
- O tratamento estatístico das propriedades dos estimadores é mais simples na amostragem com reposição do que amostragem sem reposição, uma vez que existe **independência** entre os elementos da amostra.

Obtenção de uma amostra aleatória simples com reposição

1. Numerar consecutivamente os elementos do universo de 1 a N;

2. Selecionar n elementos, com reposição, através de um procedimento aleatório (ou pseudo-aleatório):
   - método da lotaria;
   - consulta de tabelas de números aleatórios;
     Nota: atender ao número de algarismos de N
   - geração de números pseudo-aleatórios (informaticamente)

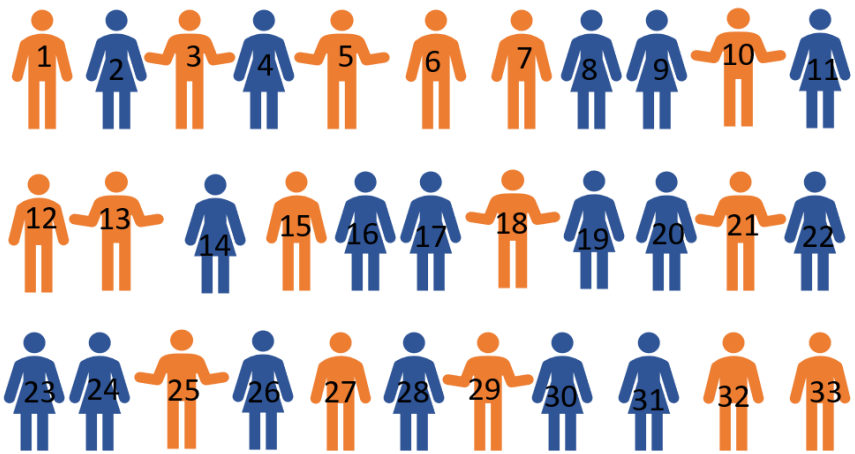
3. Estabelecer a correspondência entre os números selecionados e os elementos do universo

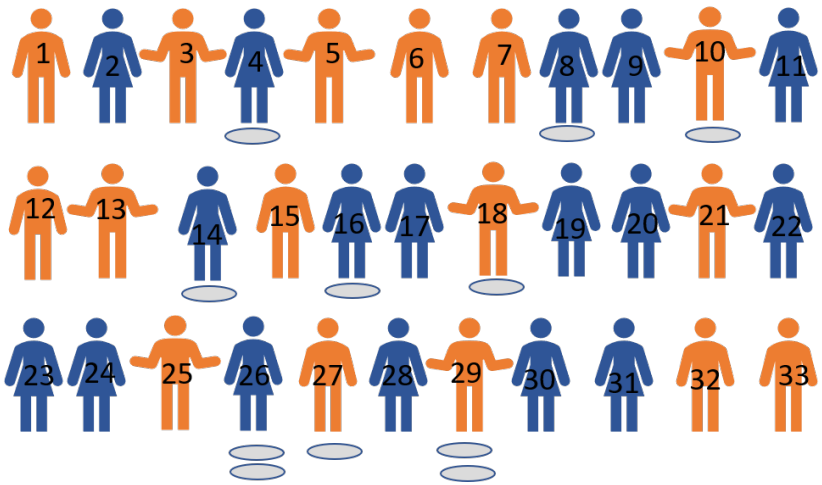
Obtenção de uma amostra aleatória simples com reposição

Selecionar uma amostra aleatória simples com reposição de 11 elementos a partir da tabela de n.ºs pseudo-aleatórios, a partir da posição 00–25 na horizontal (esq.-dir)

Tabela A. Números aleatórios

	00-04	00-09	10-14	15-19	20-24	25-29	30-34	35-39	40-44	45-49
00	25498	89023	51126	64820	23789	04134	14083	18386	61241	68406
01	80893	67443	75532	59516	79297	42914	49916	27068	60130	66369
02	66902	26158	36809	89941	61070	66880	97932	08193	54808	41963
03	51986	16493	46039	00750	26587	29645	86926	29426	59362	10374
04	61482	55273	25189	40426	71495	70014	85536	07432	54599	11280





Amostra aleatória simples com reposição - estimadores

Neste caso, utilizam-se os estimadores usuais associados a populações infinitas:

- média da amostra $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
- variância da amostra $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$
- desvio padrão amostral $S = \sqrt{S^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}$
- frequência relativa f de membros da amostra que pertencem a determinada categoria de classificação de X

Neste caso, havendo independência na seleção dos elementos da amostra não há diferença entre população finita ou infinita.

Distribuições de Amostragem e Intervalos de Confiança

média amostral – amostragem com reposição ou população infinita

- Distribuição de amostragem da média amostral $\overline{X}$ de uma **população normal** $X \sim N(\mu, \sigma^2)$
  - σ^2 conhecida (**raro**) $\longrightarrow \sqrt{n}\dfrac{\overline{X}-\mu}{\sigma} \sim N(0,1) \longrightarrow \mu = \overline{X} \pm z_{1-\frac{\alpha}{2}} \dfrac{\sigma}{\sqrt{n}}$
  - σ^2 desconhecida, é substituída por S^2
    - $\sqrt{n}\dfrac{\overline{X}-\mu}{S} \sim t_{(n-1)} \longrightarrow \mu = \overline{X} \pm t_{1-\frac{\alpha}{2};n-1} \dfrac{S}{\sqrt{n}}$
    - Se $n \geq 30$ pode-se utilizar a aproximação dada pelo T.L.C. $\overline{X} \overset{a}{\sim} N\left(\mu, \dfrac{S^2}{n}\right)$

      ou $\sqrt{n}\dfrac{\overline{X}-\mu}{S} \overset{a}{\sim} N(0,1) \longrightarrow \mu = \overline{X} \pm z_{1-\frac{\alpha}{2}} \dfrac{S}{\sqrt{n}}$
- **população <u>não</u> normal**

  Não se pode admitir a normalidade de X mas pelo Teorema Limite Central (TLC), considera-se a distribuição assintótica

$$\sqrt{n}\frac{\overline{X}-\mu}{S} \overset{a}{\sim} N(0,1) \longrightarrow \mu = \overline{X} \pm z_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}$$

Distribuição de Amostragem e Intervalo de Confiança

frequência relativa – amostragem com reposição ou população infinita

Seja $\widehat{\pi} = f = \frac{x}{n}$ a frequência relativa dos elementos da amostra que têm certa característica de interesse, sendo

- p a verdadeira proporção na população (desconhecida)
- $\widehat{\pi} = f$ estimador de π
- n a dimensão da amostra

então para $(\ n \geq 30)$

$$\widehat{\pi} \overset{a}{\sim} N\left(\pi, \frac{\pi(1-\pi)}{n}\right) \text{ ou } \frac{\widehat{\pi} - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \overset{a}{\sim} N(0,1)$$

$$\pi = \widehat{\pi} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\pi(1-\pi)}{n}}$$

Sendo a verdadeira proporção π desconhecida usa-se a estimativa $\widehat{\pi} = f$.

Exercício

Pretende-se fazer um estudo sobre o custo médio de um quarto por noite baseado numa amostra aleatória simples, com reposição, com 30 dos 42 hotéis existentes.

n. hotel	1	2	3	4	5	6	7	8	9	10	11	12	13	14
€/noite	35	40	38	41	50	30	35	40	32	39	51	39	41	53
n. hotel	15	16	17	18	19	20	21	22	23	24	25	26	27	28
€/noite	56	70	37	46	30	54	51	71	65	32	62	47	38	56
n. hotel	29	30	31	32	33	34	35	36	37	38	39	40	41	42
€/noite	56	47	48	56	38	63	68	49	39	43	54	45	73	61

1. Obtenha o parâmetro populacional pretendido.
2. Obtenha duas amostras, com reposição, de 30 observações, utilizando uma tabela de n.ºs pseudo-aleatórios e com um *software*.
3. Para cada uma das amostras obtidas:
   a) calcule um intervalo de confiança a 95% para o custo médio de um quarto, por noite, num hotel de 3 estrelas na cidade.
   Nota: considere população normal e a variância conhecida e desconhecida.
   b) calcule um intervalo de confiança a 95% para a % de hotéis na cidade com preço, por noite, igual ou superior a 50€.
   Verifique se o IC contém o verdadeiro parâmetro.

Amostragem aleatória simples (sem reposição)

A amostragem sem reposição é a mais aplicada

- Uma amostra aleatória simples (sem reposição) de n elementos retirada do universo com N elementos é tal que qualquer das C_n^N amostras possíveis tem a mesma probabilidade de ser selecionada, sendo essa probabilidade igual a $1/C_n^N$, e,
- cada um dos N elementos do universo tem a mesma probabilidade de ser selecionado, sendo igual a n/N.
  Atendendo a que a amostragem nas sondagens é **sem reposição**, existem C_n^{N-1} de amostras que não incluem o elemento.
  Assim, a probabilidade de um elemento não ser incluído na amostra é

$$\frac{C_n^{N-1}}{C_n^N} = \ldots = \frac{N-n}{N},$$

  logo, a probabilidade do elemento ser selecionado é $1 - \dfrac{N-n}{N} = \dfrac{n}{N}$.

Algumas considerações

A amostragem sem reposição é mais eficiente do que a com reposição.
Isto parece intuitivo uma vez que se recolhermos informação de elementos que já tinham sido selecionados, não estamos a acrescentar nova informação.

- A amostragem aleatória simples (a.a.s.) é raramente considerada por si só numa operação de amostragem, apesar de ser conceptualmente muito fácil
- Quando o universo é de grande dimensão, este tipo de amostragem implica muito tempo para a escolha da amostra e torna-se mais complicada porque implica que todos os elementos da população sejam enumerados
- Se os elementos do universo estão geograficamente dispersos e a sondagem implica entrevista pessoais, este tipo de amostragem torna-se dispendiosa e morosa, uma vez que corremos o risco de obter uma amostra muito dispersa geograficamente
- Uma amostra aleatória simples sem reposição pode ser obtida de modo análogo à com reposição, sendo que, neste caso, não há a reposição dos elementos selecionados

Obtenção de uma amostra aleatória simples sem reposição

Selecionar uma amostra aleatória simples de 11 elementos a partir da tabela de n.ºs pseudo-aleatórios, a partir da posição 00–25 na horizontal (esq.-dir)

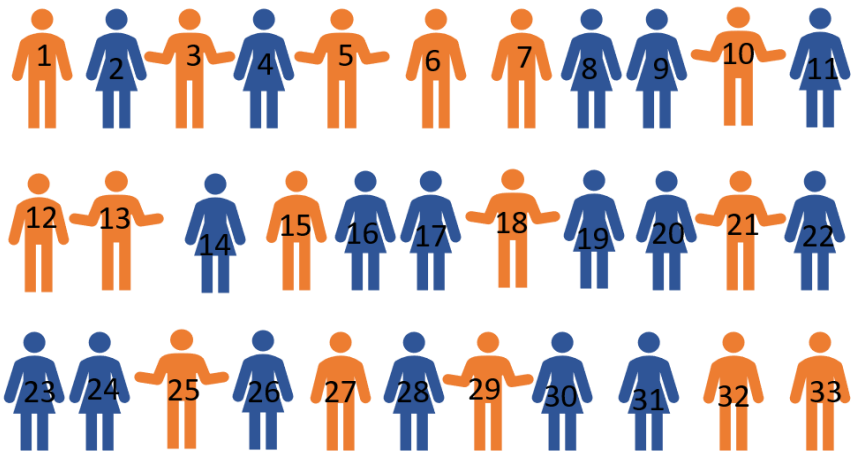
Vamos convencionar que:

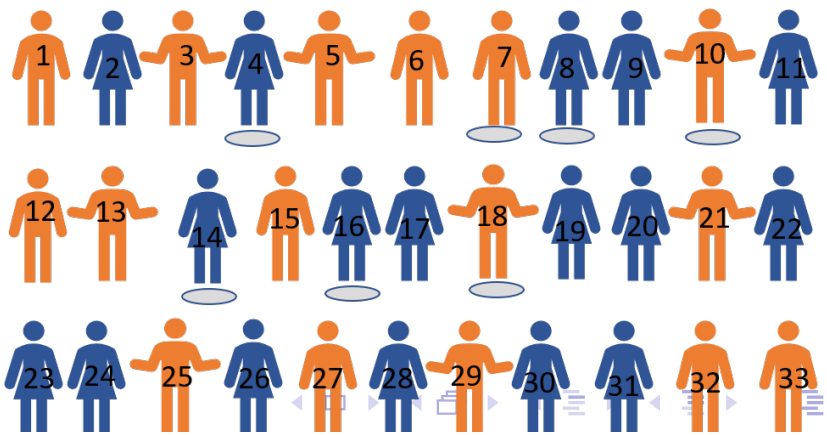
na horizontal iremos ler um número (algarismos consecutivos) por cada coluna de valores (pode ser efetuado de outra forma)

na vertical iremos ler algarismos consecutivos, um por linha

Tabela A. Números aleatórios

	00-04	00-09	10-14	15-19	20-24	25-29	30-34	35-39	40-44	45-49
00	25498	89023	51126	64820	23789	04134	14083	18386	61241	68406
01	80893	67443	75532	59516	79297	42914	49916	27068	60130	66369
02	66902	26158	36809	89941	61070	66880	97932	08193	51808	41963
03	51986	16493	46039	00750	26587	29645	86926	29426	59362	10374
04	61482	55273	25189	40426	71495	70014	85536	07432	54599	11280





Distribuições de Amostragem e Intervalos de Confiança

amostragem sem reposição e população finita

Há que introduzir um **fator de correção** para populações finitas dado por

$$FC = \sqrt{\frac{N-n}{N-1}}$$

- Distribuição de amostragem da média amostral $\overline{X}$ de uma **população normal** $X \sim N(\mu, \sigma^2)$
  - σ^2 conhecida (**raro**) $\longrightarrow \sqrt{n}\dfrac{\overline{X}-\mu}{\sigma} \sim N(0,1) \longrightarrow \mu = \overline{X} \pm z_{1-\frac{\alpha}{2}} \dfrac{\sigma}{\sqrt{n}} \sqrt{\dfrac{N-n}{N-1}}$
  - σ^2 desconhecida, é substituída por S^2
    - $\sqrt{n}\dfrac{\overline{X}-\mu}{S} \sim t_{(n-1)} \longrightarrow \mu = \overline{X} \pm t_{1-\frac{\alpha}{2};n-1} \dfrac{S}{\sqrt{n}} \sqrt{\dfrac{N-n}{N-1}}$
    - Se $n \geq 30$ pode-se usar o T.L.C. $\overline{X} \overset{a}{\sim} N\left(\mu, \dfrac{S^2}{n}\right)$

      $\longrightarrow \mu = \overline{X} \pm z_{1-\frac{\alpha}{2}} \dfrac{S}{\sqrt{n}} \sqrt{\dfrac{N-n}{N-1}}$

Distribuições de Amostragem e Intervalos de Confiança

amostragem sem reposição e população finita

- **população não normal**
  Não se pode admitir a normalidade de X mas pelo Teorema Limite Central (TLC), considera-se a distribuição assimtótica da média amostral é

$$\sqrt{n}\frac{\overline{X}-\mu}{S} \overset{a}{\sim} N(0,1) \longrightarrow \mu = \overline{X} \pm z_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

- Intervalo de confiança aproximado para uma **proporção** π ($nf > 10$ ou $n \geq 30$)

$$\widehat{\pi} \overset{a}{\sim} N\left(\pi, \frac{\pi(1-\pi)}{n}\frac{N-n}{N-1}\right) \longrightarrow \pi = \widehat{\pi} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\widehat{\pi}(1-\widehat{\pi})}{n}} \sqrt{\frac{N-n}{N-1}}$$

Sendo a verdadeira proporção π desconhecida usa-se a estimativa $\widehat{\pi}$.

Exercício

Cada um dos 45 aviões foram avaliados relativamente ao seu desempenho ecológico.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0,43	0,75	0,52	0,65	0,57	0,66	0,75	0,68	0,65	0,57	0,64	0,76	0,66	0,57	0,57
16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
0,7	0,54	0,49	0,65	0,7	0,57	0,72	0,59	0,59	0,54	0,74	0,64	0,6	0,61	0,54
31	32	33	34	35	36	37	38	39	40	41	42	43	44	45
0,62	0,43	0,72	0,47	0,65	0,58	0,68	0,51	0,67	0,57	0,62	0,61	0,56	0,67	0,58

1. Uma transportadora aérea pretende comprar 10 aviões.
   a) Selecione duas amostras aleatória simples, sem reposição, de 10 aviões cada, através de uma tabela de números aleatórios ((10,25); vertical) e através de *software*.
   b) Calcule intervalos de confiança a 90% para o indicador ecológico médio de um avião para cada amostra.
   Nota: considere população normal e a variância desconhecida.
   c) Calcule o verdadeiro parâmetro e verifique se os IC anteriores o contém.
2. Dos 45 aviões produzidos, 35 foram vendidos a companhias aéreas.
   a) Suponha que os aviões vendidos correspondem a uma a.a.s., sem reposição, e obtenha duas amostras através de uma tabela de n.ºs aleatórios e de *software*.
   b) Calcule um IC aproximado para a % de aviões produzidos pelo fabricante nesse ano com um índice ecológico inferior a $0,60$ e verifique se o verdadeiro parâmetro se encontra no IC obtido.

Amostragem sistemática

Uma amostra sistemática é obtida selecionando aleatoriamente um elemento de entre os primeiros k elementos da população ordenados segundo uma listagem, e adicionando sucessivamente o elemento que está na k-ésima posição seguinte até prefazer o número de elementos n da amostra.

Admitindo que existe um registo enumerado da população, a recolha de uma amostra sistemática consiste em:

- Calcular o **intervalo de amostragem** k, sendo a parte inteira do quociente $\frac{N}{n}$
- Se o quociente $\frac{N}{n}$ for inteiro, mostra-se que a probabilidade de qualquer elemento ser selecionado é igual a $\frac{n}{N}$
- selecionar aleatoriamente um número r entre 1 e k
- partindo desse número, adicionar sucessivamente o valor k, ficando assim selecionando os elementos r, $r+k$, $r+2k$, ..., $r+(n-1)k$

Algumas considerações

- A amostra sistemática não é uma amostra aleatória simples, uma vez que nem todas as amostras possíveis, de dimensão n, têm a mesma probabilidade de serem selecionadas
- A amostragem sistemática pode colocar em causa a representatividade da amostra se a listagem da população tiver sido realizada segundo uma periodicidade ou critério de regularidade (por exemplo, faturações diárias de uma loja, etc)
- Este tipo de metodologia é aplicada, por exemplo, na realização de entrevistas em locais públicos (centros comerciais, ruas, entradas de edifícios, ...), faz-se o questionário à r-ésima pessoa que passa e, depois, as pessoas são abordadas em intervalos sistemáticos de r pessoas $\longrightarrow$ **mas não é uma amostragem aleatória no verdadeiro sentido do termo**

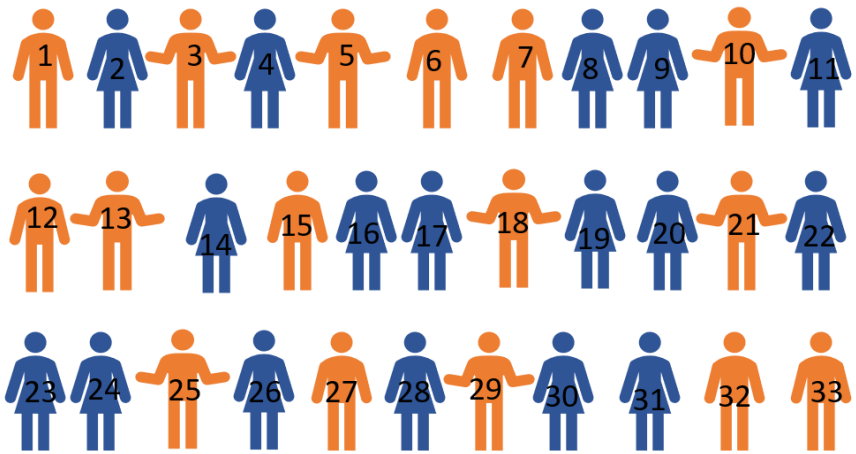
Exemplo 1

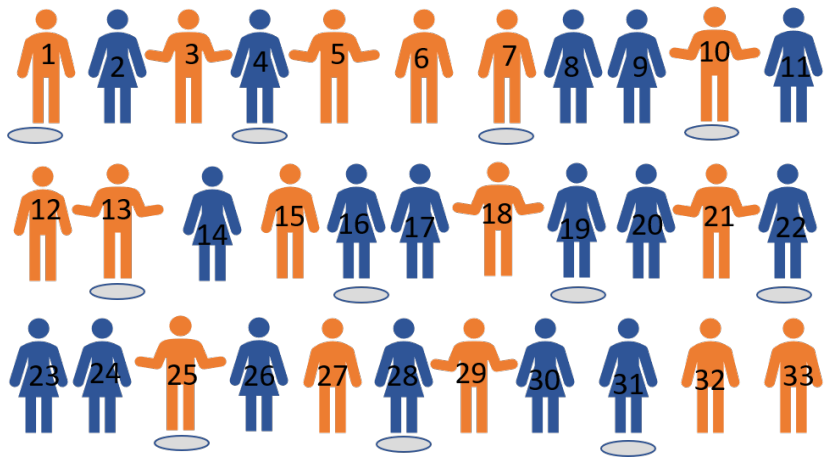
Selecionar uma amostra sistemática de 9 elementos a partir da tabela de n.ºs pseudo-aleatórios, a partir da posição 10–10 na vertical (cima–baixo)

	00-04	00-09	10-14
00	25498	89023	51126
01	80893	67443	75532
02	66902	26158	36809
03	61986	16493	46039
04	61482	55273	25189
05	76825	55025	08681
06	84742	50574	10346
07	25099	87744	96203
08	27557	92191	53609
09	62317	14085	54050
10	94965	63794	75264
11	53284	61416	15307

$$N = 33;\ k = \left[\frac{33}{9}\right] = [3, (6)] = 3$$

A seleção do 1.º elemento é feita entre os números 1 e 3





Exemplo 2

Considere a seguinte população com 560 elementos ($N = 560$), dos quais se apresentam os primeiros 12 na tabela seguinte:

obs.	32	45	56	76	34	44	61	59	50	48	63	61	...
i	1	2	3	4	5	6	7	8	9	10	11	12	...

Pretende-se determinar uma amostra sistemática de dimensão 120.

- $N = 560$, $n = 120$, $\frac{560}{120} \approx 4,67$
  logo o intervalo de amostragem é $k = 4$ (parte inteira)
- por exemplo, vamos selecionar o 1.º elemento, de entre os 4 primeiros elementos, aleatoriamente a partir da tabela de n.ºs aleatórios (vertical, posição da linha 20, coluna 19) $\longrightarrow 8, 8,$ **3** $, ...$
- partindo da posição 3, adicionamos sucessivamente o valor $k = 4$, ficando assim selecionamos os elementos nas posições
  3 $, 3 + 4 =$ **7** $, 3 + 2 \times 4 =$ **11** $, ..., 3 + (120 - 1) \times 4 =$ **479**
- A amostra é constituída pelos elementos 56, 61, 63, ...

obs.	32	45	**56**	76	34	44	**61**	59	50	48	**63**	61	...	**?**
i	1	2	**3**	4	5	6	**7**	8	9	10	**11**	12	...	**479**

Exercícios

1. Obtenha uma amostra aleatória sistemática de dimensão 3 das primeiras 12 letras do alfabeto.
2. Uma empresa produz por mês, 500 lotes de determinadas peças, os quais são codificados de 1 a 500. Para efeitos de controlo de qualidade, pretende-se selecionar uma amostra sistemática de 2% dos lotes produzidos. Determine uma amostra nestas condições indicando os códigos identificadores dos lotes a incluir na amostra.

Amostragem estratificada

Na amostragem aleatória estratificada assume-se que a população é constituída por k estratos, de dimensões N_1, N_2, ..., N_k, e:

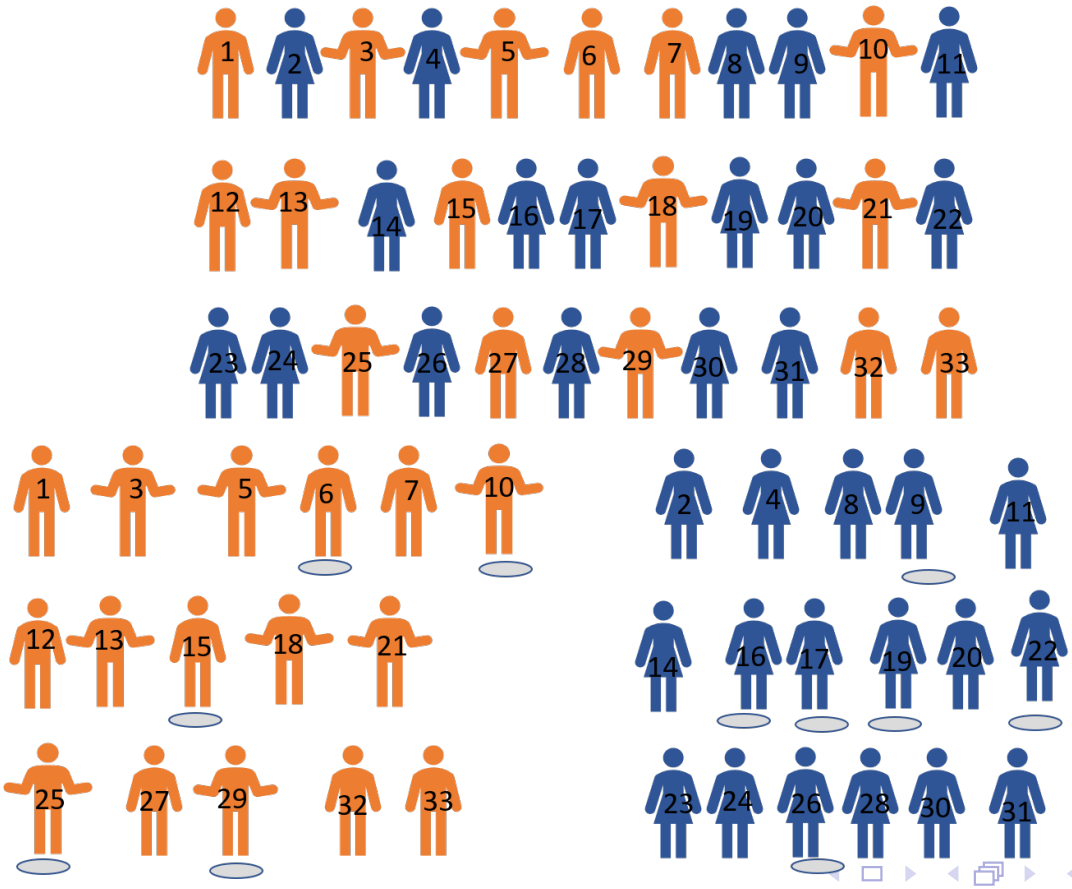
- a estratificação da população define uma partição, isto é, todo o elemento da população pertence a um único estrato (estratos mutuamente exclusivos)
- a população pode ser estratificada por uma ou mais variáveis qualitativas, nominais ou ordinais
- é retirada uma **amostra aleatória simples** de n_i elementos de cada estrato i
- a amostra total de n elementos é a reunião das sub-amostras, retiradas de cada estrato
- A identificação da variável ou variáveis a considerar para a estratificação é importante para garantir, tanto quanto possível, a representatividade da amostra para o estudo concreto

Exemplo 3

Selecionar uma amostra aleatória estratificada, pelo sexo, de 11 elementos. Usar a tabela de n.ºs pseudo-aleatórios, Masc - 00–25 (vert) Fem-00–35 (vert)

$N = 33;\ f_F = 51,5\%\ f_M = 48,5\%$

$n_F = 6\ n_M = 5$



Exemplos de Estratificação

- a população dos estudantes da ESTGA pode ser considerada estratificada por **curso** (GQ, GP, etc.)
- o universo dos eleitores em determinada eleição pode ser estratificado por **região geográfica** (norte, centro, ...) e por **sexo** (masculino, feminino), neste caso são consideradas **duas** variáveis na estratificação
- as peças produzidas por uma fábrica podem ser estratificadas pela sua **dimensão** (diferentes calibres, pesos, etc)
- se pretendermos estudar a *igualdade de género* numa organização, naturalmente que pelo menos a variável **sexo** deve ser tida em consideração na estratificação dos seus colaboradores;
- se pretendermos investigar o desempenho escolar de um conjunto de estudantes, certamente que uma estratificação pela **cor dos olhos** não será relevante para o estudo

Amostragem estratificada com afetação proporcional

Como definir as frações de amostragem de cada estrato?
Por outras palavras, quantos elementos, n_i, com $i = 1, ..., k$, de cada estrato devem ser selecionados?

Usualmente, considera-se a **amostragem estratificada por afetação proporcional**, isto é,

$$\frac{n_i}{n} = \frac{N_i}{N}, \text{ com } i = 1, ..., k$$

Neste caso, temos cada estrato tem o mesmo "*peso*" na amostra com que tem na população.

Amostragem estratificada com afetação proporcional

- População constituída pelos elementos X_{ij}, $j = 1, ..., N_i$ e $i = 1, ..., k$
- Cada estrato tem uma média e uma variância

$$\mu_i = \frac{1}{N_i} \sum_{j=1}^{N_i} X_{ij} \text{ e } \sigma_i^2 = \frac{1}{N_i} \sum_{j=1}^{N_i} (X_{ij} - \overline{X}_i)^2$$

- Estimador da média populacional μ - média amostral estratificada

$$\overline{X}_{est} = \sum_{i=1}^{k} \frac{n_i}{n} \overline{X}_i = \sum_{i=1}^{k} \omega_i \overline{X}_i, \text{ média ponderada com pesos } \omega_i = \frac{n_i}{n}$$

que, fazendo amostragem sem reposição em cada estrato, tem variância

$$\sigma^2_{\overline{X}_{est}} = \sum_{i=1}^{k} \left(\frac{n_i}{n}\right)^2 \frac{\sigma_i^2}{n_i} \frac{N_i - n_i}{N_i - 1}$$

Como as variâncias populacionais de cada estrato são, em regra, desconhecidas, são substituídas pelas variâncias amostrais de cada estrato, $\widehat{\sigma}_i^2 = S_i^2$.

Amostragem estratificada com afetação proporcional

Na amostragem estratificada com afetação proporcional:

- a média amostral é igual à média amostral estratificada, $\overline{X} = \overline{X}_{est}$
- a amostragem estratificada é tanto mais eficiente quanto maior for a variabilidade entre os estratos e menor for a variabilidade dentro de cada um dos estratos.

Uma amostragem estratificada é pelo menos tão eficiente como a amostragem aleatória simples

(no sentido em que permite obter estimativas mais precisas dos parâmetros populacionais)

Exemplo

Considere a base de dados (*contacts.sav*) com arquivo de 70 contactos realizados por um grupo de vendedores de computadores para organizações a diferentes empresas. Entre outras variáveis são conhecidas: o *valor da última venda feita*, o *tempo desde a última venda*, e a *dimensão da empresa*.
Pretende-se contactar apenas 25 das empresas para possíveis novas vendas através de uma amostragem estratificada (com afetação proporcional) pela variável *dimensão da empresa*.

- $N = 70 \longrightarrow n = 25$
- obtenção da proporção de cada estrato na população para aplicar à amostra

dimensão	N_i	N_i/N	n_i
muito pequena	22	31,4%	$31,4\% \times 25 \approx 8$
pequena	29	41,4%	$41,4\% \times 25 \approx 10$
média	17	24,3%	$24,3\% \times 25 \approx 6$
grande	2	2,9%	$2,9\% \times 25 \approx 1$
	$N = 70$		$n = 25$

Devido aos arredondamentos poderá haver ajustamentos ou a dimensão final da amostra pode ser superior ou inferior em uma unidade ao valor previsto.

Exemplo-alternativa

dimensão	N_i	N_i/N	n_i
muito pequena	22	31,4%	$31,4\% \times 25 \approx 8$
pequena	29	41,4%	$41,4\% \times 25 \approx 10$
média	17	24,3%	$24,3\% \times 25 \approx 6$
grande	2	2,9%	$2,9\% \times 25 \approx 1$
	$N = 70$		$n = 25$

- $N = 70 \longrightarrow n = 25$
- obtenção da fração de amostragem $\frac{n}{N} = \frac{25}{70} = 35,71\%$

dimensão	N_i	n_i
muito pequena	22	$35,71\% \times 22 \approx 8$
pequena	29	$35,71\% \times 29 \approx 10$
média	17	$35,71\% \times 17 \approx 6$
grande	2	$35,71\% \times 2 \approx 1$
	$N = 70$	$n = 25$

Exemplo

Um universo de 1296 eleitores foram inquiridos antes e depois de um debate político. Os resultados estão disponíveis na base de dados *debate.sav*. Pretende-se fazer um estudo mais restrito a uma amostra destes eleitores sobre as razões que os levaram a manter ou alterar a sua preferência.
Pretende-se determinar uma amostra com cerca de 5% dos elementos estratificada pela *idade* e pelo *sexo*.

Exemplo

Um universo de 1296 eleitores foram inquiridos antes e depois de um debate político. Os resultados estão disponíveis na base de dados *debate.sav*. Pretende-se fazer um estudo mais restrito a uma amostra destes eleitores sobre as razões que os levaram a manter ou alterar a sua preferência.
Pretende-se determinar uma amostra com cerca de 5% dos elementos estratificada pela *idade* e pelo *sexo*.

$N = 1296$, fração de amostragem de **5%**

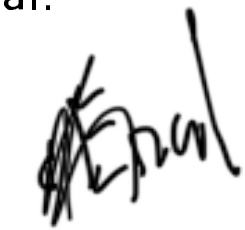
idade	<31	31-45	46-60	>60	
M	92	166	142	211	
F	110	166	182	227	Total 1296

Dimensão das subamostras a recolher aleatoriamente

idade	<31	31-45	46-60	>60	
M	5	8	7	11	
F	6	8	9	11	Total 65

Exercícios

1. As famílias são classificadas mediante o rendimento mensal (R) global:
   - classe A: $R > 3600€$
   - classe B: $1775€ < R \leq 3600€$
   - classe C: $650€ < R \leq 1775€$
   - classe D: $R \leq 650€$

   Em determinada freguesia existem 60 famílias da classe A, 90 da Classe B, 120 da Classe C e 480 da Classe D.
   a) Determine quantas famílias de cada classe devem participar num estudo em dois cenários: $n = 50$ e $n = 30$.
   b) Admitindo que as famílias estão numeradas sequencialmente dentro de cada classe, obtenha uma amostra estratificada de afetação proporcional através da tabela de números aleatórios, a partir da posição 00-10, para o cenário $n = 30$.
2. Certos investigadores indicam que o n.º de horas por semana no instagram é diferenciado consoante o sexo e a idade.
   a) Obtenha uma amostra aleatória adequada com cerca de 60% dos estudantes em aula.
   b) Calcule um intervalo de confiança a 90% para o n.º médio de horas semanal no instagram dos estudantes em aula no último mês. (Terá de inquirir os elementos selecionados.)

Amostragem por *Clusters*

Na **amostragem por *clusters*** não é necessário identificar individualmente todos os elementos da população mas apenas que se disponha de uma listagem completa de **grupos** ou ***clusters*** de elementos individuais.

- este tipo de amostragem está orientada para a seleção de grupos de elementos e não de elementos individuais
- os grupos são mutuamente exclusivos e exaustivos que, geralmente correspondem a um agrupamento "natural" existente na população

Metodologia

- aplica-se a **amostragem aleatória** aos **grupos**
- ficam incluídos na amostra **todos** os elementos pertencentes aos *clusters* selecionados na amostragem aleatória

Amostragem por clusters

Exemplos: ***cluster***	**unidade elementar**	**exemplo de aplicação**
turma	aluno	estudar a opinião dos estudantes de uma escola sobre os seus gostos musicais
centro de saúde	utentes	estimar o tempo médio de espera para atendimento numa consulta
departamento	colaborador	estudar a opinião dos colaboradores sobre as suas chefias
zona geográfica de vendas	vendedor	estimar o montante de vendas para o ano seguinte

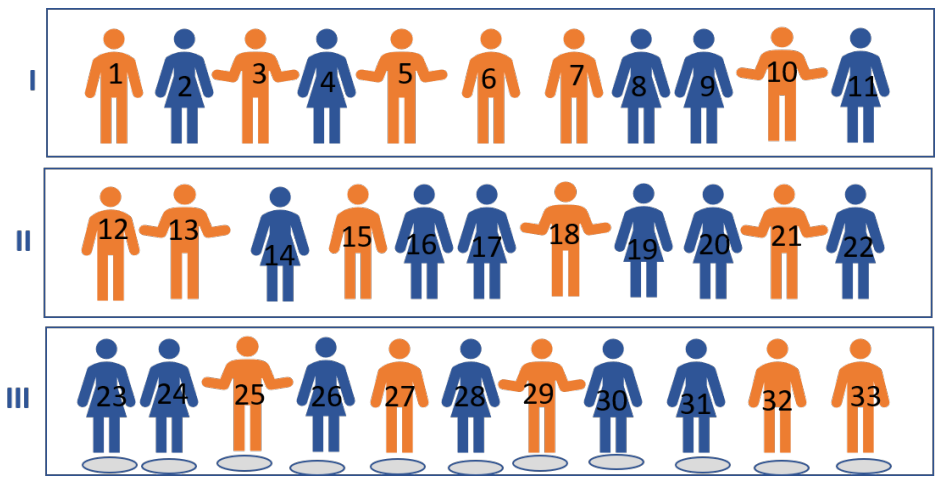
Algumas considerações:

- a amostragem por *clusters* tem duas vantagens relevantes em estudos de populações humanas que cobrem uma larga área geográfica: a **facilidade com se põe em prática** e o **menor custo**
- se os elementos de um *cluster* forem homogéneos relativamente a algumas características, a representatividade da amostra pode nem sempre ficar assegurada uma vez que existe alguma redundância na amostra

Exemplo 1

Selecionar uma amostra aleatória por clusters Usar a tabela de n.ºs pseudo-aleatórios, 00–10 (vert)

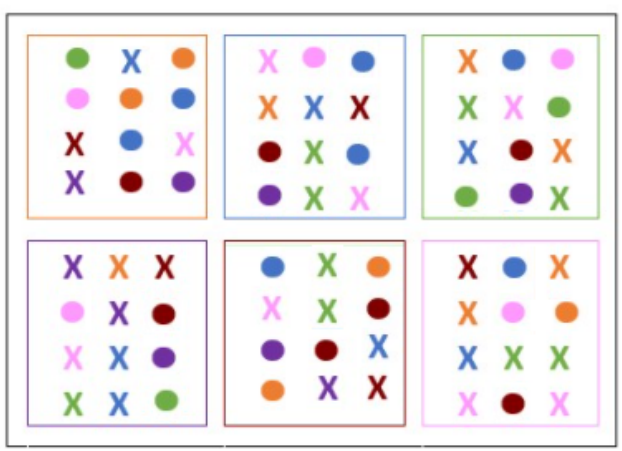
	00-04	05-09	10-14
00	2 5 4 9 8	8 9 0 2 3	~~X~~ 1 1 2 6
01	8 0 8 9 3	6 7 4 4 3	~~X~~ 5 5 3 2
02	6 6 9 0 2	2 6 1 5 8	3 6 8 0 9
03	6 1 9 8 6	1 6 4 9 3	4 6 0 3 9
04	6 1 4 8 2	5 5 2 7 3	2 5 1 8 9

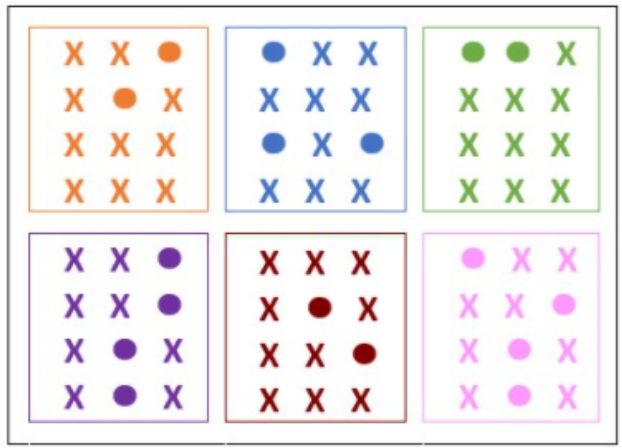


Amostragem estratificada versus amostragem por *clusters*

Em regra, a amostragem por *clusters* é tão mais eficiente quanto maior for a variabilidade **dentro** dos grupos.
Os grupos devem ser heterogéneos internamente e semelhantes entre si.

Quanto mais semelhantes forem os elementos dentro de um *cluster*, melhores serão os resultados se esse *cluster* for usado como um estrato numa amostra estratificada e piores se forem usados como unidades amostrais na amostragem por *clusters*.





Exemplo

Considere uma organização com departamentos identificados com as letras de A a G com o n.º de colaboradores em cada departamento indicado na tabela

A	B	C	D	E	F	G
23	15	31	16	26	28	20

Pretende-se fazer um estudo sobre os agregados familiares dos colaboradores da empresa. Para isso, pretende-se obter uma amostra por *clusters*, que não exceda os 70 elementos, de modo a não ser necessária a deslocação a todos os departamentos.

Iremos selecionando aleatoriamente, sem reposição
departamento numerados A-1, B-2, ..., G-7; tabela de n.º pseudo-aleatórios, por exemplo, em coluna, cima-baixo, a partir da posição 25-34:

n.º pseudo aleatório	7	3	9	6
dep.	G	C	–	F
n.º elementos	20	31	–	**28**, considerando F excede-se os 70

Todos os colaboradores dos departamentos G e C irão participar no estudo, obtendo-se uma amostra com $n = 51$.

Exercícios

1. Considere o universo dos estudantes que pertencem às comissões de curso das licenciaturas da ESTGA (GQ, GP, GC, SCE, EMI, TI).
   Note que cada comissão de curso tem 3 membros discentes.
   a) Selecione uma amostra por grupos com 6 estudantes, e cujos grupos são definidos pelos cursos a que cada estudante está inscrito.
   b) Caso o estudo pretenda estudar as competências linguísticas dos estudantes das comissões de curso, o tipo de amostragem adotado dá garantias sobre a representatividade anterior? Justifique.

Exercícios

2. Considere a produção de 10 lotes, cada lote com 2 peças de cada uma das 3 máquinas a operar.

Lote	1	2	3	4	5	6	7	8	9	10
máq. A	0,51	0,54	0,52	0,50	0,50	0,44	0,50	0,57	0,49	0,51
	0,50	0,51	0,46	0,48	0,51	0,51	0,52	0,52	0,53	0,48
máq. B	1,20	1,44	1,20	1,32	1,22	1,10	1,25	1,10	1,20	1,26
	1,09	1,26	1,28	1,50	1,33	1,12	1,22	1,20	1,05	1,20
máq. C	1,53	1,49	1,61	1,73	1,62	1,76	1,64	1,61	1,64	1,56
	1,72	1,59	1,45	1,61	1,56	1,59	1,64	1,60	1,44	1,51

   a) Pretende-se obter uma amostra por grupos com 30% das peças produzidas. Caso se pretenda estudar o calibre médio de todas as peças produzidas, qual das variáveis *lote* ou *máquina* deve ser considerada para a definição dos *clusters*? Justifique.
   b) Obtenha uma amostra por grupos nas condições da alínea anterior.
   c) Calcule um IC a 95% para o calibre médio de uma peça produzida na empresa e verifique se este contém o parâmetro populacional.
   d) Calcule um IC a 90% para a proporção de peças com calibre superior a $1,1$ e verifique se este contém o parâmetro populacional.

Nota: para efeitos dos IC considere que a amostra é uma a.a.s..

Amostragem multi-etapas

A amostragem multi-etapas é uma extensão da **amostragem por *clusters***.

A maior flexibilidade vem do facto de se selecionar uma amostra aleatória de cada grupo em vez de todos os elementos integrarem a amostra.

- As vantagens deste tipo de amostragem são as enunciadas para a amostragem por grupos.
- Este método pode ter várias fases sucessivas de amostragem até alcançar uma amostra de unidades elementares.

Unidade Amostral Primária	Unidade Amostral Secundária	Unidade Amostral Terciária	Unidade Amostral Quaternária
bairro	quarteirão	prédio	habitação
página	linha de texto		
país	concelho	freguesia	

Amostragem multi-fases

A amostragem multi-fases não deve ser confundida com a amostragem multi-etapas

- Na amostragem multi-fases, em cada fase de amostragem está sempre em causa o mesmo tipo de unidade amostral.
- Neste tipo de amostragem existe diversas fases de amostragem, realizadas mediante um processo aleatório, mas sempre sobre a mesma unidade amostral.

O procedimento pode ser descrito nos seguintes pontos:

- selecionar uma amostra aleatória de elementos para participarem numa 1ª fase do estudo;
- selecionar uma subamostra aleatória da amostra inicial, na qual os elementos podem ser alvo de um novo inquérito com maior nível de profundidade e de detalhe.

Métodos de amostragem não aleatórios

Estes métodos não são aconselháveis quando se pretende extrapolar os resultados e conclusões obtidos com a amostra para o universo.
Podem ser úteis no início de um estudo, por exemplo, para testar as primeiras versões de um questionário.

- amostragem intencional
- amostragem "bola de neve"
- amostragem por conveniência
- amostragem por quotas

Amostragem intencional

Uma **amostra intencional** é uma amostra composta de elementos selecionados deliberadamente por quem conduz o estudo.

- A escolha dos elementos é determinada através de um critério subjetivo – a opinião do investigador.
- É utilizada em estudos exploratórios em que importa recolher opiniões e ideias relevantes, por exemplo, na escolha de peritos para se pronunciarem sobre determinada matéria.
- Pode ser útil:
  - na obtenção de uma amostra reduzida
  - quando é impossível obter-se uma amostra aleatória
  - quando se pretende obter deliberadamente uma amostra enviesada, por exemplo, na avaliação de modificações num produto pode fazer sentido primeiro estudar os consumidores que, pelas suas características estariam recetivos à mudança

Amostragem "bola de neve"

A **amostragem "bola de neve"** é uma forma de amostragem intencional em que o investigador pede a cada um dos elementos inicialmente selecionados por si nomes de outros elementos que podem ser igualmente inquiridos.

- É muito utilizada quando se pretende chegar a populações pequenas e muito específicas, como por exemplo adeptos de um *hobbie* invulgar, indivíduos com determinada deficiência ou característica física.
- Tem o inconveniente de que os inquiridos tendem a indicar o nome de amigos ou pessoas próximas, o que pode levar a uma amostra com elementos muito similares no âmbito do interesse do estudo

Amostragem por conveniência

Na **amostragem por conveniência** os elementos são escolhidos porque se encontram onde os dados para o estudo estão a ser recolhidos.
A sua participação no estudo é como que "acidental".

- os inquéritos de rua são exemplos de uma amostragem por conveniência porque são favorecidos os indivíduos que passaram no local e momento da recolha dos dados
- os inquéritos a amigos e conhecidos são outro exemplo deste tipo de amostragem

Problemas

- forte possibilidade de a amostra ser enviesada (não representativa)
- amostragem sem bases científicas o que não permite fazer projeções para a população, mas pode ser servir para captar ideias gerais, identificar aspetos críticos ou na realização do pré-teste do questionário

Amostragem por quotas

A **amostragem por quotas** pode ser definida como uma **amostragem estratificada não aleatória**.
Isto é, em vez de se escolher uma amostra aleatória dentro de cada estrato escolhe-se uma amostra **não aleatória** de dimensão determinado pela fração de amostragem, a **quota**.

- Este tipo de amostragem justifica-se quando é difícil obter-se as listagens dos estratos da população.
- Apesar do n.º de elementos de cada estrato ser proporcional ao verificado na população as subamostras não são necessariamente representativas porque foram obtidos por métodos não aleatórios.
- Muitas vezes as amostras não-aleatórias são obtidas por **conveniência**, contudo, a consideração das **variáveis de controlo** que definem as quotas tornam este método melhor, em geral, que a amostragem simples por conveniência.

Exercícios

1. Identifique o método de amostragem em cada um dos seguintes casos:
   a) Polícia rodoviária escolhe os veículos numa estrada para parar, ver a documentação e fazer inspeção. Conveniencia
   b) Num estudo realizado por uma empresa de aluguer de veículos num aeroporto cada entrevistador recebeu a indicação de quantas entrevistas devem ser realizadas de acordo com os diversos perfis de clientes e cuja escolha ficou para o julgamento do entrevistador. Intencional / Por Quotas
   c) Um estudo sobre os "sem-abrigo" iniciou-se com 5 indivíduos que frequentavam uma cantina social. Cada um destes indivíduos indicou mais 3 "sem-abrigos" que foram posteriormente entrevistados. Bola de neve
   d) Para explicar os fenómenos meteorológicos extremos ocorridos em determinado dia, foram convidados os 3 especialistas mais reputados do IPMA. Intencional
2. Indique e descreva um processo de amostragem não aleatório adequado a cada um dos seguintes casos:
   a) Um estudo na ESTGA sobre o meio de transporte principal dos estudantes.
   b) Uma sondagem eleitoral no concelho de Águeda.

A questão mais comum ...

Num estudo estatístico, a questão mais frequente é

"Qual a dimensão da amostra a recolher?"

Infelizmente não há uma resposta simples a esta questão.

- O cálculo da dimensão da amostra pode ser feito matematicamente desde que os elementos sejam escolhidos por um procedimento aleatório.
- Nesta situação existem expressões que permitem calcular a dimensão da amostra com determinadas **precisão** e **confiança** requeridas.

Fatores determinantes na dimensão da amostra

- **Dimensão da população**
  - em regra, quanto maior for a população maior será a amostra mas **não existe uma proporcionalidade** entre a dimensão da população e a dimensão da amostra
  - a dimensão da população tem especial relevância nas amostragem sem reposição (população finita)
- **Variabilidade da população** no que diz respeito à característica em estudo
  - quanto maior for a dispersão da característica em estudo maior terá de ser a dimensão da amostra
- **Precisão** - neste contexto, a precisão é entendida como a **amplitude máxima** de um intervalo de confiança a calcular, em que o **erro máximo** é metade dessa amplitude
- **Nível de confiança** - neste contexto, é o nível de confiança adotado no cálculo do intervalo de confiança do parâmetro de interesse no estudo

De seguida, serão apresentadas fórmulas para o cálculo de n, visando à construção de intervalos de confiança para a média μ e proporção π a partir dos seus respetivos estimadores $\overline{X}$ e f.

Dimensão da amostra para estimar uma média μ

- População infinita ou amostragem com reposição $n \geq \dfrac{z^2_{1-\frac{\alpha}{2}} \cdot \sigma^2}{d^2}$
- População finita ou amostragem sem reposição $n \geq \dfrac{z^2_{1-\frac{\alpha}{2}} \cdot \sigma^2 \cdot N}{d^2(N-1) + z^2_{1-\frac{\alpha}{2}} \cdot \sigma^2}$

onde

- $z_{1-\frac{\alpha}{2}}$ é o quantil de probabilidade $1 - \frac{\alpha}{2}$ da normal padrão, associado a um IC com um nível de confiança $(1-\alpha) \times 100\%$
- d é a margem de erro tolerada (metade da amplitude do IC pretendido)
- σ^2 é a variância populacional da variável em estudo, em regra, é desconhecida pelo que deve ser substituída por um valor
  - indicado em especificações técnicas
  - obtido em estudos anteriores sobre a mesma variável
  - obtido numa amostra piloto

Exemplo

Uma indústria de discos metálicos pretende construir um intervalo de confiança a 95%, com uma amplitude máxima de 2cm, para o diâmetro médio dos discos. Numa amostra piloto obteve-se um desvio padrão amostral para o diâmetro de um disco de 3,4 cm.

$z_{1-\frac{0,05}{2}} = z_{0,975} = 1,96$, $d = \frac{2}{2} = 1$ e $\widehat{\sigma}^2 = s^2 = 3,4^2$ Vem de onde?

1. Não sendo conhecido o n.º total de discos produzidos ($N = \infty$)

$$n \geq \frac{1,96^2 \cdot 3,4^2}{1^2} = 44,41 \Longrightarrow n \geq 45$$

2. sabendo que foram produzidos $N = 100$ discos

$$n \geq \frac{1,96^2 \cdot 3,4^2 \cdot 100}{2^2(100-1) + 1,96^2 \cdot 3,4^2} = 30,97 \Longrightarrow n \geq 31$$

Dimensão da amostra para estimar uma proporção π

- População infinita ou amostragem com reposição $n \geq \dfrac{z^2_{1-\frac{\alpha}{2}} \cdot \widehat{\pi}(1-\widehat{\pi})}{d^2}$
- Pop. finita ou amostragem sem reposição $n \geq \dfrac{z^2_{1-\frac{\alpha}{2}} \cdot \widehat{\pi}(1-\widehat{\pi}) \cdot N}{d^2(N-1) + z^2_{1-\frac{\alpha}{2}} \cdot \widehat{\pi}(1-\widehat{\pi})}$

onde

- $z_{1-\frac{\alpha}{2}}$ é o quantil de probabilidade $1-\frac{\alpha}{2}$ da normal padrão, associado a um IC com um nível de confiança $(1-\alpha)\times 100\%$
- d é a margem de erro tolerada (metade da amplitude do IC pretendido)
- $\widehat{\pi}$ é uma estimativa da verdadeira proporção.
  Caso não se tenha uma estimativa prévia considera-se $\widehat{\pi} = f = 50\%$

Exemplo

Pretende-se obter um intervalo de confiança para proporção dos eleitores que indicam votar num candidato. Na última sondagem, a percentagem era de 34%. Pretende-se um nível de confiança de 90% e um erro máximo de 2%.

$z_{1-\frac{0,10}{2}} = z_{0,95} = 1,645$, $d = 2\% = 0,02$ e $\widehat{\pi} = f = 34\% = 0,34$

Considere:

1. população infinita, $N = \infty$

$$n \geq \frac{1,645^2 \cdot 0,34(1-0,34)}{0,02^2} = 1518,08 \Longrightarrow n \geq 1519$$

2. população finita e $N = 1500$

$$n \geq \frac{1,645^2 \cdot 0,34(1-0,34) \cdot 1500}{0,02^2(1500-1) + 1,645^2 \cdot 0,34(1-0,34)} = 754,74 \Longrightarrow n \geq 755$$

3. população finita com $N = 1500$ e que não existia uma sondagem anterior

$$n \geq \frac{1,645^2 \cdot 0,5(1-0,5) \cdot 1500}{0,02^2(1500-1) + 1,645^2 \cdot 0,5(1-0,5)} = 795,2 \Longrightarrow n \geq 796$$

Exercícios

1. Qual a dimensão de uma amostra aleatória simples que se deve obter para estimar o tempo médio de produção de uma peça por uma máquina atendendo a que $\sigma = 5$ segundos e pretende-se não errar em mais de 1,5 segundos, usando um nível de confiança de 99%.
   Refaça os cálculos considerando que a máquina produziu um total de 200 peças.
2. Num município existem cerca de 10 000 árvores. Quantas árvores devem ser analisadas de modo a estimar a proporção de árvores do concelho que precisam de ser podadas, se o objetivo é ter 90% de confiança e de não errar em mais do que 3%?
3. Um técnico da qualidade pretende estimar a % de artigos defeituosos de um grande lote de lâmpadas. Com base na sua experiência ele sabe que esta % deve rondar os 20%. Qual a dimensão da amostra se se pretender estimar essa % com tolerância de 1%, usando um nível de confiança de 95%.
   Refaça as contas considerando que o lote tem 1000 lâmpadas.