Processo de Amostragem

➢ População versus Amostra

- População:
  - conjunto de unidades com uma ou mais característica em comum
  - N é a dimensão da população

- Amostra:
  - Subconjunto da população
  - N é a dimensão da amostra

- Ao processo de recolha de informação de toda a população dá-se o nome de censo ou recenseamento

- Por outro lado, ao processo de recolha de informação a um subconjunto da população dá-se o nome de amostragem ou sondagem

- Cada uma das características em estudo é representada por uma variável X_i

➢ Inferência Estatística

- É um processo de raciocínio indutivo, em que se procuram tirar conclusões indo do particular, para o geral. É um tipo de raciocínio contrário ao tipo de raciocínio matemático, essencialmente dedutivo.

- Utiliza-se quando se pretende estudar características da ***população***, estudando só alguns elementos dessa população, ou seja, uma **amostra**.

- Serve para, a partir das propriedades verificadas na **amostra**, inferir propriedades para a ***população***, com a indicação da precisão (o erro cometido) dessas inferências

- **Exemplo**

  População : alunos inscritos no ISCTE-IUL no ano de 2014/2015
  Seja p a percentagem de alunos que pratica regularmente desporto.
  Recolhida uma amostra de 10 alunos, com reposição:

  - se conhecermos o valor de p, por exemplo p=0,298, podemos calcular a probabilidade de haver x alunos a praticar desporto, nos 10 alunos seleccionados. A variável X, que representa o número de alunos em 10 que pratica desporto, é bem modelada por uma Binomial, com parâmetros 10 e 0,298.
  - se não conhecermos o valor de p, vamos utilizar o número x de alunos, que praticam desporto, nos 10 seleccionados, para "estimar" p, e temos um problema de Inferência Estatística.

- ***Parâmetro –*** é uma característica numérica da população e assume um valor fixo. Por exemplo, a média da população (μ), a variância da população (σ^2), o desvio populacional (σ)

- Quando se pretende estimar (obter um valor aproximado) dum ***parâmetro*** considera-se uma função conveniente, que só dependa dos valores da amostra – **estatística**, a que se dá o nome de **estimador** do parâmetro em estudo, ou seja, um estimador é uma função dos elementos da amostra. Por exemplo, a média amostral ($\overline{X}$), a variância da amostra (S^2), o desvio padrão amostral (S)

- Para cada amostra que se recolhe, obtém-se um valor dessa função, que se chama **estimativa.** Dado que duma população se podem retirar muitas amostras estamos perante valores que variam de amostra para amostra**.** Também se utiliza o termo estatística como significado de estimativa.

- Concluindo, as estatísticas são características numéricas da amostra, por oposição aos parâmetros que são características numéricas da população.

- Exemplo:

  Um bom **estimador** para o **parâmetro μ** duma população é a **estatística $\overline{X}$** cuja **estimativa é $\overline{x}$**

➢ Fases do processo de amostragem

1. Definição clara dos objectivos

2. Definição dos dados a recolher e escolha do instrumento de recolha (Ex.: inquérito)

3. Definição do processo adequado ao tipo de dados e instrumento de recolha:
   - Estabelecer plano de amostragem
   - Construção da amostra:
     - População alvo/inquirida
     - Método de secção da amostra
     - Dimensão da amostra

- Uma amostra que não seja representativa da População diz-se **enviesada** e a sua utilização pode dar origem a interpretações erradas.

- Métodos de Amostragem aleatória versus Métodos de Amostragem não aleatória

➢ Métodos de amostragem

- Métodos de Amostragem Aleatória:
  - Aleatória simples
  - Sistemática
  - Estratificada
  - Por clusters
  - Multi-etápica
  - Multi-fásica

- Métodos de Amostragem Não Aleatória
  - Por conveniência
  - Intencional
  - Snowball
  - Por quotas

➢ Amostragem aleatória simples

- Cada elemento da população *tem a mesma probabilidade* de ser seleccionado para a amostra. Numa amostra não aleatória, alguns elementos da população podem não poder ser seleccionados para a amostra.

- Cada amostra de dimensão n tem a mesma probabilidade de ser seleccionada que qualquer outra da mesma dimensão

- Implica a existência de uma lista exaustiva da população

- Duas formas de proceder:
  - Lotaria
  - Nº aleatórios

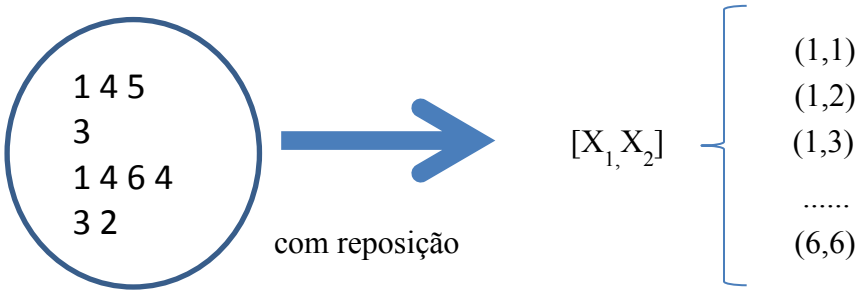
➢ Amostras aleatórias

Pretendemos estudar uma certa característica, que designamos por X, duma população.

Seja N a dimensão (tamanho) da população. Uma amostra de dimensão n dessa população X será $(X_1, X_2, \ldots, X_n)$. O valor observado de X para o i-ésimo elemento da amostra é x_i, ou seja, x_i é uma concretização de X_i

Cada vez que recolhemos uma amostra da mesma dimensão, obtemos valores diferentes para cada elemento, ou seja, uma amostra aleatória é uma variável aleatória multidimensional

- N=tamanho da população=10 n=tamanho da amostra=2



Nota: A amostra é uma variável aleatória bidimensional

➢ Exemplo

Num serviço de reparação de electrodomésticos são prestados três tipos de reparações com custos para o cliente de €40, €45 e €50, respectivamente. Seja a variável aleatória X - valor cobrado ao cliente.

Suponha que são retiradas amostras de tamanho n = 2, **com reposição**. Q**uantas** e **quais** são as possíveis amostras retiradas da população?

N=3 (40,45,50), n = 2: $[X_1, X_2]$, $k = N^n = 3^2 = 9$

em que :

k = número de amostras possíveis
N = tamanho da população
n = tamanho da amostra

Amostra	[X1,X2]
1	[40,40]
2	[40,45]
3	[40,50]
4	[45,40]
5	[45,45]
6	[45,50]
7	[50,40]
8	[50,45]
9	[50,50]

➢ Distribuição de uma amostra aleatória

- As variáveis aleatórias $X_1, X_2, \ldots, X_n$ assumem os mesmos valores de X, pois são elementos de uma amostra, todos retirados da mesma população, ou seja,

$$f(x_1) = f(x_2) = \ldots = f(x_n) = f(x)$$

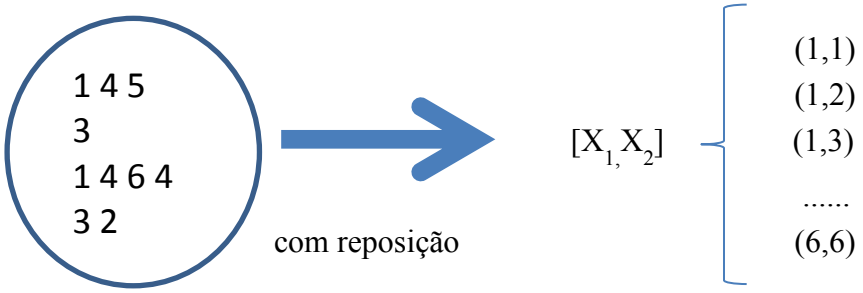
- Como $(X_1, X_2, \ldots, X_n)$ é uma amostra recolhida segundo um processo aleatório (amostragem aleatória simples) os seus elementos são variáveis aleatórias independentes entre si.

- Concluindo, $X_1, X_2, \ldots, X_n$ são independentes e identicamente distribuídas sendo a função distribuição conjunta dada pelo produto das marginais; ou seja,

$$f(x_1, x_2, \ldots, x_n) = f(x_1) \times f(x_2) \times \ldots \times f(x_n) = \prod_{i=1}^{n} f(x_i)$$

➢ Amostras aleatórias

- N=tamanho da população=10 n=tamanho da amostra=2



[X1, X2]	$\bar{X}$	S^2
(1, 1)	1	0
(1, 2)	1,5	0,5
(1, 3)	2	2
. . .	. . .	. . .
(6, 6)	6	0

[X1,X2] é uma amostra

$\bar{X}$ e S^2 são estatísticas e também variáveis aleatórias

Se temos k amostras do mesmo tamanho n, temos k médias e k variâncias - são estimativas

➢ Distribuição de uma amostra aleatória

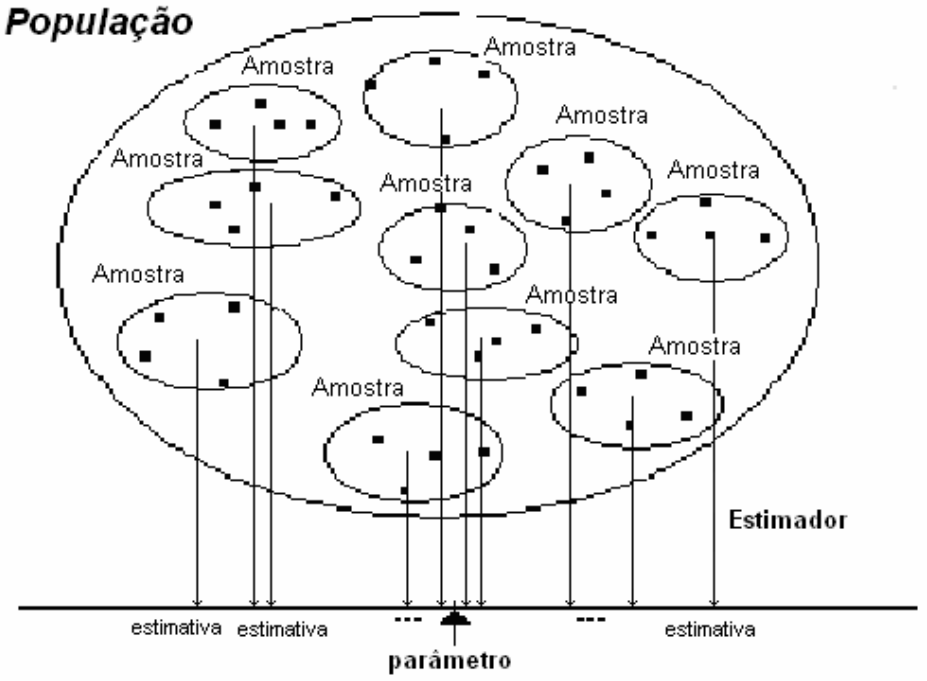
- As estatísticas são variáveis aleatórias, logo também terão alguma distribuição de probabilidade com média, variância, etc.

- **Distribuição amostral** é a **distribuição de probabilidade de uma estatística**.

- O estudo de um estimador é feito a partir da sua **distribuição de amostragem**, ou seja, da distribuição dos valores obtidos pelo estimador, quando se consideram todas as amostras possíveis, utilizando um determinado esquema de amostragem.

Como se comportam todas estas **estimativas**, relativamente ao ***parâmetro***, em estudo?

É necessário estudar a **distribuição de amostragem** do estimador



➢ Exemplo

Num serviço de reparação de electrodomésticos são prestados três tipos de reparações com custos para o cliente de €40, €45 e €50, respectivamente. Seja a variável aleatória X = valor cobrado ao cliente, com a seguinte distribuição de probabilidade:

X=x	40	45	50
P[X=x]	0,2	0,3	0,5

a) Determine a **média** e a **variância da população**.

- $E(X) = \mu = \sum x\, p(x) = 40x0,2 + 45\,x0,3 + 50+0,5 = 46,5$
- $Var[X] = \sigma^2 = E(X^2) - \mu^2 = \sum x^2\, p(x) - 46,5^2$
  $= (40^2\, x\, 0,2 + 45^2\, x\, 0,3 + 50^2\, x\, 0,5) - 46,5^2 = 15,25$

b) Suponha que são retiradas amostras de tamanho n = 2, **com reposição**.
Q**uantas** e **quais** são as possíveis amostras retiradas da população e qual a **probabilidade** associada a cada uma?
Determine a **média e a variância da distribuição amostral da média**

N=3 (40,45,50), n = 2: [X**1**, X**2**], $k = N^n = 3^2 = 9$

em que : k = número de amostras possíveis
N = tamanho da população
n = tamanho da amostra

Amostra	[X1,X2]	P[X1,X2]	$\bar{X}$
1	[40,40]	0,2x0,2	40
2	[40,45]	0,2x0,3	42,5
3	[40,50]	0,2x0,5	45
4	[45,40]	0,3x0,2	42,5
5	[45,45]	0,3x0,3	45
6	[45,50]	0,3x0,5	47,5
7	[50,40]	0,5x0,2	45
8	[50,45]	0,5x0,3	47,5
9	[50,50]	0,5x0,5	50

$\bar{X}=x$	40	42,5	45	47,5	50
$P[\bar{X}=x]$	0,04	2x0,06	2x0,1 + 0,09	2x0,15	0,25

$$E(\bar{X}) = \mu_X = \sum \bar{x}\, p(\bar{x}) = 40x0,04 + 42,5\ x0,12 + 45x0,29 + 47,5x0,3 + 50*0,25 = 46,5$$

Média (μ) da população

$$Var[\bar{X}] = E(\bar{X}^2) - \mu_X^{\,2} = \sum \bar{x}^2\, p(\bar{x}) - 46,5^{\,2} = (40^{\,2} \times 0,04 + 42,5^{\,2} \times 0,12 + 45^{\,2} \times 0,29 + 50^{\,2} \times 0,25) - 46,5^{\,2} = 7,625 = 15,25/2$$

Metade da variância (σ^2) da população

c) Suponha que são retiradas amostras de tamanho n = 3, **com reposição**.

N=3 (40,45,50), n = 3: [X**1**, X**2**,**X3**], $k = N^{n} = 3^{3} = 27$

Amostra	**[X1,X2,X3]**	**P[X1,X2,X3]**	$\overline{X}$
1	[40,40,40]	0,2x0,2x0,2	40
2	[40,40,45]	0,2x0,2x0,3	41,67
3	[40,40,50]	0,2x0,2x0,5	43,33
4	[40,45,40]	0,2x0,3x0,2	41,67
5	[40,45,45]	0,2x0,3x0,3	43,33
6	[40,45,50]	0,2x0,3x0,5	45
7	[40,50,40]	0,2x0,5x0,2	43,33
8	[40,50,45]	0,2x0,5x0,3	45
9	[40,50,50]	0,2x0,5x0,5	46,67
10	[45,40,40]	0,3x0,2x0,2	41,67
11	[45,40,45]	0,3x0,2x0,3	43,33
12	[45,40,50]	0,3x0,2x0,5	45
13	[45,45,40]	0,3x0,3x0,2	43,33
14	[45,45,45]	0,3x0,3x0,3	45
15	[45,45,50]	0,3x0,3x0,5	46,67
16	[45,50,40]	0,3x0,5x0,2	45
17	[45,50,45]	0,3x0,5x0,3	46,67
18	[45,50,50]	0,3x0,5x0,5	48,33

Amostra	[X1,X2,X3]	P[X1,X2,X3]	$\overline{X}$
19	[50,40,40]	0,5x0,2x0,2	43,33
20	[50,40,45]	0,5x0,2x0,3	45
21	[50,40,50]	0,5x0,2x0,5	46,67
22	[50,45,40]	0,5x0,3x0,2	45
23	[50,45,45]	0,5x0,3x0,3	46,67
24	[50,45,50]	0,5x0,3x0,5	48,33
25	[50,50,40]	0,5x0,5x0,2	46,67
26	[50,50,45]	0,5x0,5x0,3	48,33
27	[50,50,50]	0,5x0,5x0,5	50

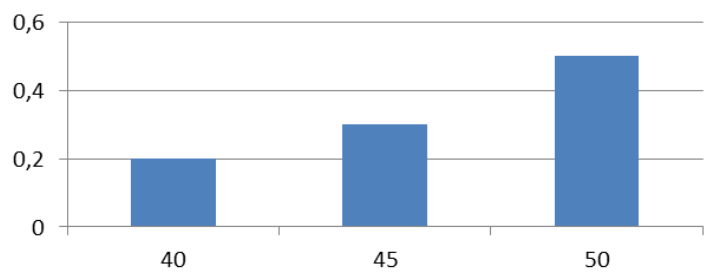
$E(\overline{X}) = 46,5$ → **Média (μ) da população**

$Var[\overline{X}] = 5,083 = 15,25/3$ → **Um terço da variância (σ^2) da população**

Assim,

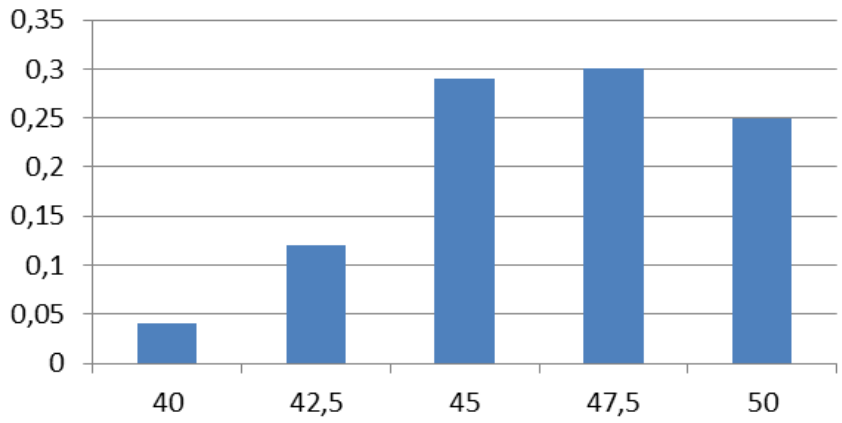
Distribuição de probabilidade da população

X=x	40	45	50
P[X=x]	0,2	0,3	0,5



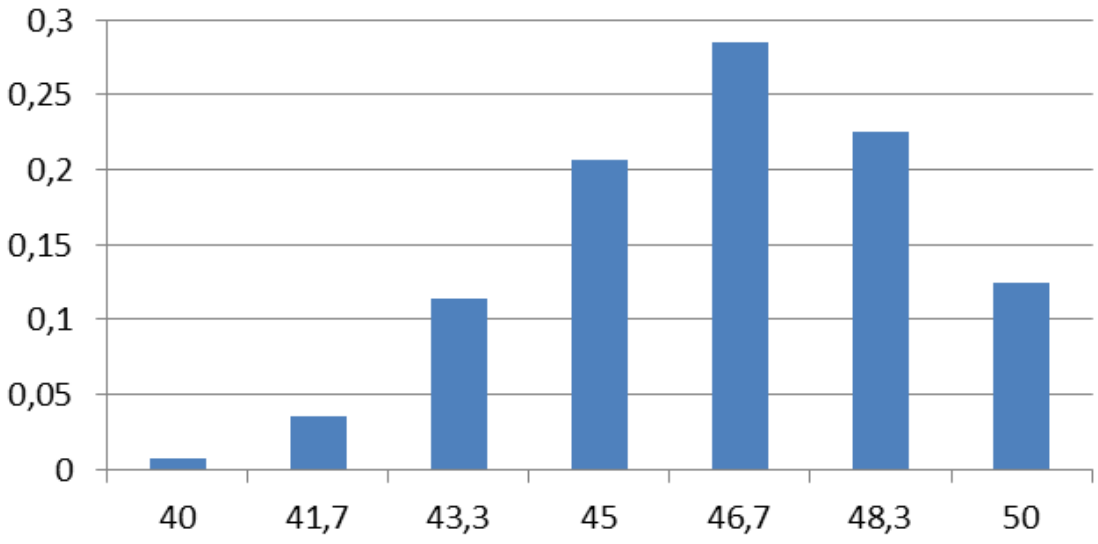
Distribuição amostral da média para amostras de dimensão n=2

$\overline{X}$=x	40	42,5	45	47,5	50
P[$\overline{X}$=x]	0,04	0,12	0,29	0,3	0,25



Distribuição amostral da média para amostras de dimensão n=3

$\overline{X}$=x	40	41,7	43,3	45	46,7	48,3	50
P[$\overline{X}$=x]	0,008	0,036	0,114	0,207	0,285	0,225	0,125



Comparando o diagrama de barras da população X com os da média para as amostras de tamanhos 2 e 3, observamos que, **mesmo que a distribuição da população não seja simétrica**, a distribuição amostral da média **tende para a simetria** à medida que o **tamanho da amostra aumenta**.

➢ Teorema do Limite Central

Suponhamos que se recolhe uma amostra de dimensão n de uma população X, com valor médio μ e desvio-padrão σ. A recolha da amostra deve ter em consideração o seguinte:

1. Se a população for finita a recolha é feita com reposição;
2. No caso de a população ter uma dimensão "suficientemente grande", a selecção da amostra pode ser feita sem reposição.

Então, se a dimensão da amostra for suficientemente grande (n > 30):

- A variável aleatória **soma de todos os X_i** , ($S_n = X_1 + X_2 + \ldots + X_n$), tem um comportamento aproximadamente normal:

$$Y_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} \overset{\circ}{\cap} N(0;1)$$

- A variável aleatória **média de todos os X_i ,** $\overline{(X)}$ **,** tem um comportamento aproximadamente normal.

$$\frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \overset{\circ}{\cap} N(0;1)$$

Nota que:

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

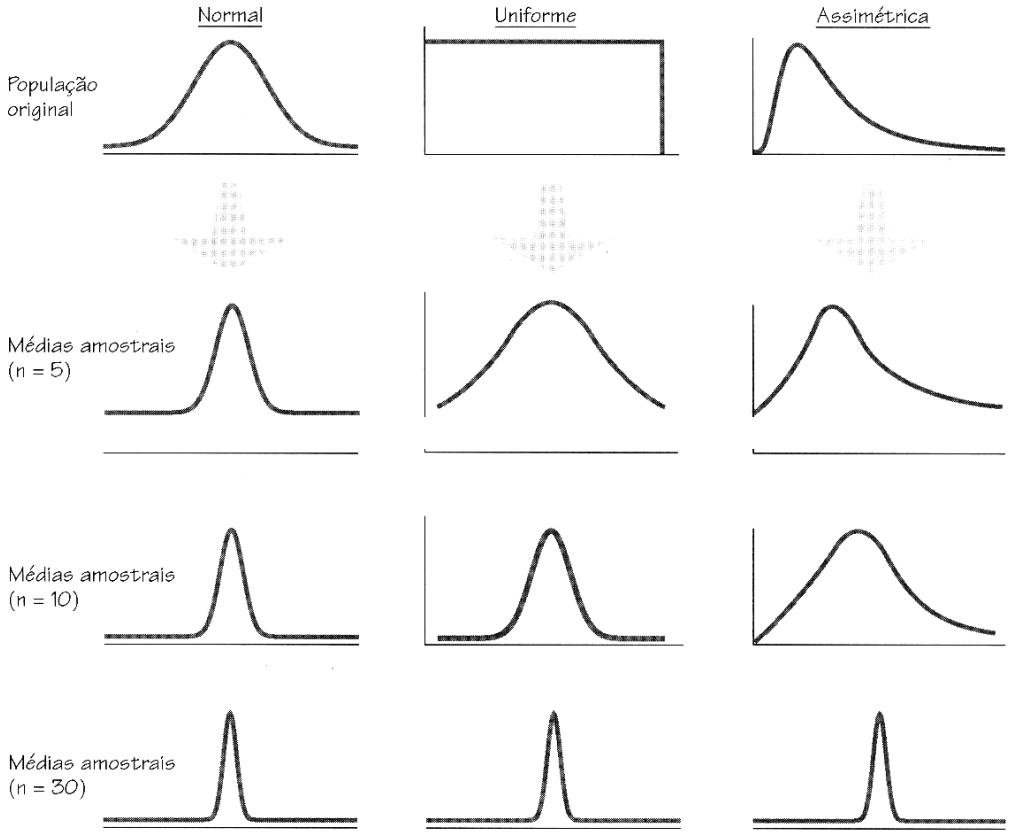


Fig. 5-21 Distribuição normal, uniforme e assimétrica.