AMOSTRAGEM ALEATÓRIA
DISTRIBUIÇÕES POR AMOSTRAGEM

Quando se pretende estudar uma determinada população, analisam-se certas características ou variáveis dessa população. Essas variáveis poderão ser discretas ou contínuas e ficam caracterizadas se conhecermos a sua função de probabilidade ou função densidade de probabilidade. Assim, identificada a distribuição e respectivos parâmetros, conheceremos o comportamento da v.a. X .

Porém, a determinação dos parâmetros de uma população impõe que se estudem todos os elementos que a constituem, o que só é possível para populações finitas não muito numerosas. Efectivamente, o custo associado ao estudo de toda uma população é por vezes tão elevado, que a melhor alternativa consiste em seleccionar uma amostra dessa população e estimar os parâmetros necessários a partir dos valores amostrais. Isto é, a partir do estudo da amostra tiram-se conclusões que se pretendem válidas para a população como um todo.

Contudo, nem todas as amostras permitem que, a partir dos seus resultados, se faça uma generalização a toda a população com uma certa credibilidade. No fundo, pretende-se que a amostra seleccionada seja um micro-cosmos da respectiva população e daí que nos debrucemos a partir de agora, apenas sobre o método de **amostragem aleatória**.

Este método de selecção de amostras, a que já fizemos referência no início da disciplina, garante que todos os elementos da população têm as mesmas hipóteses de serem integrados na amostra, evitando-se assim qualquer

enviesamento da selecção, isto é, qualquer tendência sistemática para subrepresentar ou sobrerepresentar na amostra alguns elementos da população.

Consideremos então que se pretende estudar a característica X de uma população e que X tem uma fdp $f_X(x)$ (se estivessemos a trabalhar com uma função de probabilidade o processo era análogo).

Se for retirada dessa população uma amostra A_1 de dimensão n obteremos $\left(x_1^1, x_2^1, x_3^1, ..., x_n^1\right)$ em que o k-ésimo elemento da amostra X_k^1 $(k = 1, 2, ..., n)$ é um valor do conjunto de todos os valores que X pode assumir.

Se retirarmos agora sucessivamente amostras de dimensão n da nossa população obteremos:

Amostra 1 (A_1) : $\left(x_1^1, x_2^1, x_3^1, ..., x_n^1\right)$

Amostra 2 (A_2) : $\left(x_1^2, x_2^2, x_3^2, ..., x_n^2\right)$

...

Amostra s (A_s) : $\left(x_1^s, x_2^s, x_3^s, ..., x_n^s\right)$

...

Então podemos considerar que temos uma amostra tipo:

$$\left(X_1, X_2, X_3, ..., X_n\right)$$

que por gerar as diferentes amostras $(A_1, A_2, ..., A_s, ...)$ pode ser considerada como uma variável aleatória n-dimensional com função densidade de probabilidade conjunta $f_{X_1 X_2 ... X_n}(x_1, x_2, ..., x_n)$.

Admitamos agora que, no caso de uma população finita, cada amostra é obtida executando o seguinte procedimento:

1) Numerar de 1 até N todos os elementos da população.

2) Colocar bolas numeradas de 1 até N numa máquina de extracção de números do totoloto.

3) Seleccionar uma bola (a probabilidade de seleccionar uma dada bola é igual à probabilidade de seleccionar outra qualquer) e incluir na amostra o elemento da população com o número correspondente.

4) Repor a bola na máquina.

5) Repetir os passos 3) e 4) tantas vezes quantas as necessárias para seleccionar a amostra da dimensão pretendida (n). Admite-se que as extracções sucessivas são independentes.

Usando este processo de amostragem, as funções de probabilidade das variáveis $X_1, X_2, X_3, ..., X_n$ são idênticas às da variável original X, isto é:

$$p_{X_1}(x) = p_{X_2}(x) = ... = p_{X_n}(x) = p_X(x) \qquad \forall_x$$

Isto traduz-se dizendo que o **processo de amostragem é aleatório** e as **amostras** assim obtidas dizem-se **aleatórias**.

No procedimento descrito admitiu-se que a **amostragem era efectuada com reposição** o que resulta no facto das variáveis $X_1, X_2, X_3, ..., X_n$ serem independentes, isto é:

$$p_{X_1 X_2 ... X_n}(x_1, x_2, ..., x_n) = p_X(x_1) \cdot p_X(x_2) \cdot ... \cdot p_X(x_n) \quad \forall_{x_1, x_2, ..., x_n}$$

Sempre que esta situação se verificar, diz-se que o processo de amostragem e as amostras são **aleatórios simples**.

NOTA: Como resultado da reposição o que está em causa é retirar n amostras de dimensão unitária a partir de n populações finitas idênticas, de dimensão N. Isto é equivalente a retirar, sem reposição, uma amostra de dimensão n a partir de uma população infinita constituída por elementos idênticos aos da população original finita, figurando nas mesmas proporções.

Exemplo: Considerar uma população constituída pelas alturas de 5 indivíduos, expressas em metros:

{ 1,70 ; 1,74 ; 1,75 ; 1,75 ; 1,82 }

Definir uma v.a. X como a altura de um dos indivíduos seleccionados ao acaso. Então a função de probabilidade de X é dada por:

x	1,70	1,74	1,75	1,82
$p_X(x)$	1/5	1/5	2/5	1/5

Considerem-se agora amostras de dimensão n = 2, obtidas a partir da população designando-se por X_1 e X_2 as réplicas da v.a. X relativas ao primeiro e segundo elementos da amostra.

Seleccionando as amostras pelo procedimento anteriormente apresentado teremos a seguinte função de probabilidade conjunta (e funções de probabilidade marginais):

		X_2				
		1,70	1,74	1,75	1,82	$p_{X_1}(x_1)$
X_1	1,70	1/25	1/25	2/25	1/25	1/5
	1,74	1/25	1/25	2/25	1/25	1/5
	1,75	2/25	2/25	4/25	2/25	2/5
	1,82	1/25	1/25	2/25	1/25	1/5
	$p_{X_2}(x_2)$	1/5	1/5	2/5	1/5	

Da análise da tabela conclui-se que, efectivamente:

$$p_{X_1}(x) = p_{X_2}(x) = p_X(x) \qquad \forall_x$$

e

$$p_{X_1X_2}(x_1, x_2) = p_X(x_1) \cdot p_X(x_2)$$

Consideremos agora que as amostras eram recolhidas por um processo idêntico ao anterior com excepção da reposição da bola extraída, isto é, a **amostra é recolhida sem reposição.** Neste caso e dado que se admitiu que a população é finita, as variáveis $X_1, X_2, X_3, ..., X_n$ deixam de ser independentes.

Isto significa que, existem valores de $X_1, X_2, X_3, ..., X_n$ para os quais:

$$p_{X_1 X_2 ... X_n}(x_1, x_2, ..., x_n) \neq p_X(x_1) \cdot p_X(x_2) \cdot ... \cdot p_X(x_n)$$

ou de outra maneira, valores de $X_1, X_2, X_3, ..., X_n$ para os quais:

$$p_{X_2|X_1}(x_2|x_1) \neq p_{X_2}(x_2)$$

$$p_{X_3|X_1 X_2}(x_3|x_1, x_2) \neq p_{X_3}(x_3)$$

....

$$p_{X_n|X_1 X_2 ... X_{n-1}}(x_n|x_1, x_2, ..., x_{n-1}) \neq p_{X_n}(x_n)$$

Neste caso a função de probabilidade conjunta e as funções de probabilidade marginais são dadas por:

		X_2				
		1,70	1,74	1,75	1,82	$p_{X_1}(x_1)$
X_1	1,70	0	1/20	2/20	1/20	1/5
	1,74	1/20	0	2/20	1/20	1/5
	1,75	2/20	2/20	2/20	2/20	2/5
	1,82	1/20	1/20	2/20	0	1/5
	$p_{X_2}(x_2)$	1/5	1/5	2/5	1/5	

E podemos verificar que, apesar de:

$$p_{X_1}(x) = p_{X_2}(x) = p_X(x) \qquad \forall_x$$

temos agora:

$$p_{X_1X_2}(x_1,x_2) \neq p_X(x_1)\cdot p_X(x_2)$$

isto é, as variáveis X_1 e X_2 não são independentes mas, dado que no procedimento de selecção da amostra, nenhum elemento da população é tratado de modo diferente dos restantes, as probabilidades dos diferentes elementos da população virem a pertencer à amostra são iguais.

NOTA: Quando a dimensão da população tende para infinito e a dimensão da amostra se mantém finita, a dependência entre as variáveis $X_1, X_2, X_3, ..., X_n$ tende a desaparecer. Assim, quando a população for infinita, é indiferente realizar a amostragem com ou sem reposição, isto é, estaremos sempre numa situação de amostragem aleatória simples.

A noção de amostragem aleatória pode ser facilmente generalizado ao caso de variáveis contínuas, que pressupõem populações infinitas ou, na prática, populações finitas de dimensão muito elevada.

Seja então uma população deste tipo e consideremos uma característica a estudar desta população que representamos pela v.a. contínua X, com fdp $f_X(x)$. Seleccionemos amostras de dimensão n a partir da nossa população:

$$(X_1, X_2, X_3, ..., X_n)$$

em que cada um dos X_k $(k = 1, 2, ..., n)$ representa uma réplica da v.a. X relativa ao k-ésimo elemento de cada amostra.

O processo de amostragem e as amostras obtidas dizem-se aleatórios (simples) se forem satisfeitas as seguintes condições:

- $f_{X_1}(x) = f_{X_2}(x) = ... = f_{X_n}(x) = f_X(x) \qquad \forall_x$
- $f_{X_1 X_2 ... X_n}(x_1, x_2, ..., x_n) = f_X(x_1) \cdot f_X(x_2) \cdot ... \cdot f_X(x_n) \qquad \forall_{x_1, x_2, ..., x_n}$

PARÂMETROS E ESTATÍSTICAS

Um **parâmetro** é uma característica duma população, isto é, um valor que embora possa ser desconhecido é fixo.

Uma **estatística** é uma característica da amostra, isto é, um valor que caracteriza uma dada amostra e que é variável de amostra para amostra, ou seja, uma variável aleatória.

Exemplo: Se para cada uma das amostras $A_1, A_2, ..., A_s, ...$ referidas anteriormente, calcularmos a respectiva média, iremos obter:

$$\overline{x}^1, \overline{x}^2, ..., \overline{x}^s, ...$$

Podemos então considerar que a média amostral $\overline{X}$ é uma variável aleatória (amostral), que assume um dado valor concreto $\overline{x}^i$ para cada amostra A_i.

Designa-se por **estimativa** o valor que uma estatística assume para uma dada amostra concreta.

Seja $(X_1, X_2, ..., X_n)$ uma amostra aleatória de uma v.a. X e sejam $(x_1, x_2, ..., x_n)$ os valores concretos assumidos por essa amostra. Seja H uma função definida sobre a n-upla $(x_1, x_2, ..., x_n)$. Define-se $Y = H(X_1, X_2, ..., X_n)$ como uma estatística, que assume o valor $y = H(x_1, x_2, ..., x_n)$. Isto é, uma estatística é uma função real da amostra. Então, uma estatística é uma variável aleatória e fará sentido falar da sua distribuição de probabilidade, do seu valor esperado e da variância.

Sempre que uma v.a. for de facto uma estatística, isto é, uma função da amostra, designa-se a sua distribuição de probabilidade por **distribuição amostral**.

Seja $(X_1, X_2, ..., X_n)$ uma amostra aleatória de uma v.a. X. Algumas estatísticas importantes são:

- Média amostral : $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$

- Variância amostral : $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$

- Proporção amostral de uma população de Bernoulli:

  $P = \frac{Y}{n}$ em que Y representa o nº de elementos de um dado tipo incluídos na amostra

- Mínimo da amostra : $K = \text{mín}(X_1, X_2, ..., X_n)$

- Máximo da amostra : $M = \text{máx}\left(X_1, X_2, ..., X_n \right)$

- Amplitude da amostra : R = M – K

- j – ésima maior observação na amostra : $X_n^{(j)}$

$$j = 1, 2, ..., n$$

É óbvio que: $j = 1 \Rightarrow X_n^{(1)} = M$

$j = n \Rightarrow X_n^{(n)} = K$

DISTRIBUIÇÕES POR AMOSTRAGEM

Como foi já referido existe uma diferença fundamental entre parâmetros e estatísticas. Efectivamente, para uma determinada população e uma dada variável aleatória sobre ela definida, os parâmetros da distribuição correspondente (valor esperado, variância, ...) são valores fixos. Por outro lado, as estatísticas (média amostral, variância amostral, ...) variam de amostra para amostra.

Dada esta variabilidade das estatísticas, que são afinal novas variáveis aleatórias, interessa definir a sua distribuição probabilística através das respectivas funções de probabilidade ou densidade de probabilidade. **As distribuições das estatísticas designam-se por distribuições por amostragem ou distribuições amostrais.**

Exemplo: Consideremos uma população com 4 elementos, que correspondem aos seguintes valores da variável aleatória $Y:\{2,4,6,6\}$.

A função de probabilidade da variável aleatória Y, bem como o seu valor esperado e variância são:

Y	2	4	6
P (Y = y)	1/4	1/4	1/2

$E(Y) = \mu_Y = 2 \,.\, (1/4) + 4 \,.\, (1/4) + 6 \,.\, (1/2) = 4{,}5$

$V(Y) = \sigma^2_Y = (2 - 4{,}5)^2 \,.\, (1/4) + (4 - 4{,}5)^2 \,.\, (1/4) + (6 - 4{,}5)^2 \,.\, (1/2) = 2{,}75$

Pretende-se agora definir a distribuição por amostragem de $\overline{Y}$, média amostral, calculada com base em amostras de dimensão 2 obtidas por um processo aleatório sem reposição. Considerando todas as amostras de dimensão 2 que podem obter-se recorrendo ao processo indicado, as respectivas médias amostrais e as correspondentes probabilidades de ocorrência, temos:

Amostra	$\overline{y}$	**Prob. de ocorrência**
2,4	3	1/4 . 1/3 = 1/12
2,6	4	1/4 . 2/3 = 2/12
4,2	3	1/4 . 1/3 = 1/12
4,6	5	1/4 . 2/3 = 2/12
6,2	4	2/4 . 1/3 = 2/12
6,4	5	2/4 . 1/3 = 2/12
6,6	6	2/4 . 1/3 = 2/12

A partir desta tabela podemos obter a função de probabilidade da média amostral, isto é, a distribuição de $\overline{Y}$.

$\bar{\mathbf{Y}}$	3	4	5	6
$P(\bar{\mathbf{Y}}=\bar{\mathbf{y}})$	1/6	1/3	1/3	1/6

Concluindo-se que:

$E(\bar{\mathbf{Y}}) = \mu_{\bar{\mathbf{Y}}} = 3 . (1/6) + 4 . (1/3) + 5 . (1/3) + 6 . (1/6) = 4,5$

$V(\bar{\mathbf{Y}}) = \sigma^2_{\bar{\mathbf{Y}}} = (3 – 4,5)^2 . (1/6) + (4 – 4,5)^2 . (1/3) +$
$+ (5 – 4,5)^2 . (1/3) + (6 – 4,5)^2 . (1/6) = 0,917$

Neste caso foi possível especificar de um modo simples, a distribuição por amostragem de $\bar{Y}$ a partir da distribuição da variável original porque a população é bastante pequena.

Se este método não for aplicável, devido por exemplo à elevada dimensão da população, poderemos ainda recorrer a métodos que nos permitem obter de uma forma precisa ou aproximada a distribuição por amostragem de uma determinada estatística.

Exemplo: ver aplicação de métodos analíticos e teorema do limite central.

Quando através de métodos analíticos não se consegue deduzir a distribuição por amostragem de uma estatística, pode obter-se uma ideia aproximada (que pode até ser muito próxima) daquela distribuição recorrendo ao **método de Monte Carlo,** que consiste no seguinte procedimento**:**

i) Geram-se amostras aleatórias, constituídas por n observações, de uma população com uma dada distribuição:

$$\left(\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_n\right),\ \left(\mathbf{x}_1^{'},\mathbf{x}_2^{'},...,\mathbf{x}_n^{'}\right),\ \left(\mathbf{x}_1^{''},\mathbf{x}_2^{''},...,\mathbf{x}_n^{''}\right),\ ...$$

ii) Calcula-se para cada amostra o valor particular da estatística:

$$\mathbf{t}=\mathbf{T}\left(\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_n\right),\ \mathbf{t}^{'}=\mathbf{T}\left(\mathbf{x}_1^{'},\mathbf{x}_2^{'},...,\mathbf{x}_n^{'}\right),\mathbf{t}^{''}=\mathbf{T}\left(\mathbf{x}_1^{''},\mathbf{x}_2^{''},...,\mathbf{x}_n^{''}\right),...$$

iii) Com o conjunto dos valores particulares da estatística, $\mathbf{t},\mathbf{t}^{'},\mathbf{t}^{''},...$, constrói-se um histograma.

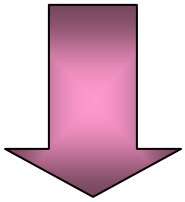
iv) Toma-se para representação (aproximada) da distribuição por amostragem o histograma a que pode ajustar-se uma função densidade (função probabilidade) ou a partir do qual se pode obter a média, a variância, a moda, os percentis, etc.

O último problema a resolver consiste em saber como gerar amostras aleatórias provenientes de uma população com uma dada distribuição.

O problema é resolvido recorrendo a números aleatórios: gerados em computador (números pseudo-aleatórios obtidos de um modo geral recorrendo a geradores lineares

congruenciais) ou a tabelas de números aleatórios já disponíveis.
A geração de números aleatórios é um assunto complexo e uma discussão detalhada deste tema está fora do âmbito da disciplina, porém, é de salientar a importância dos geradores de números aleatórios que constituem o núcleo central da simulação computacional.

A geração de números aleatórios com uma distribuição qualquer é conseguida a partir da geração e da transformação de números aleatórios de uma variável aleatória uniforme no intervalo (0,1).



Método da Transformação Inversa

MÉTODO DA TRANSFORMAÇÃO INVERSA
V.A. Contínuas

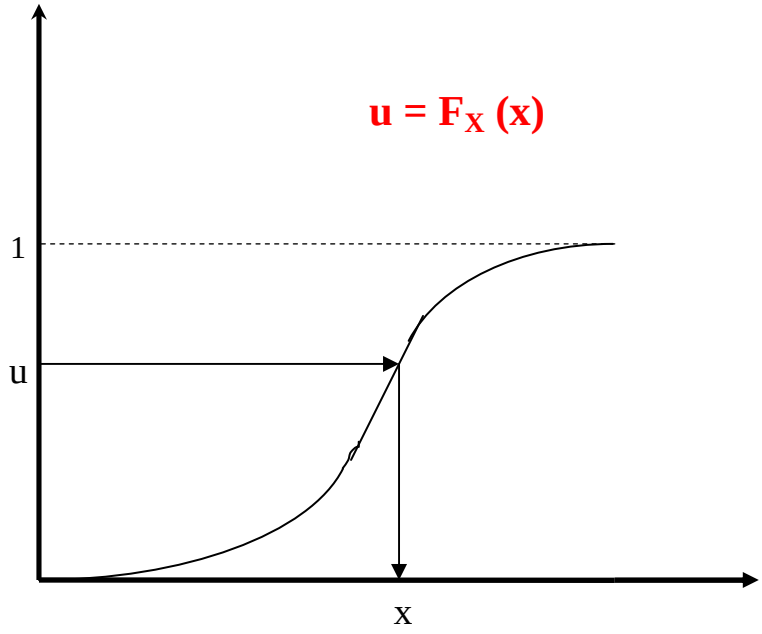
Admita-se que a variável aleatória X é contínua e tem uma função de distribuição acumulada F_X com inversa F_X^{-1}.
Então um algoritmo para gerar uma realização da variável aleatória X é o seguinte:

Dada a função de distribuição acumulada $F_X(x) = P(X \leq x)$

** Gerar uma realização u da v.a. $U \sim U(0,1)$

** Obter a realização $x = F_X^{-1}(x)$ da v.a. X.

A figura abaixo ilustra o método da transformação inversa para variáveis aleatórias contínuas.



EXEMPLOS:

i) Geração de realizações da distribuição uniforme U(a,b).

Pretende-se assim determinar $x = F_X^{-1}(u)$ para o caso de $X \sim U(a,b)$.

Atendendo a que a função de distribuição acumulada de $X \sim U(a,b)$ é:

$$F_X(x) = \begin{cases} 0 & x < a \\ \dfrac{x-a}{b-a} & a \le x \le b \\ 1 & x > b \end{cases}$$

Fazendo $u = F_X(x)$ e resolvendo em ordem a x, obtém-se:

$$x = a + u(b-a)$$

ii) Geração de realizações da distribuição exponencial E(λ).

Pretende-se determinar $x = F_X^{-1}(u)$ para o caso de $X \sim E(\lambda)$.

Atendendo a que a função de distribuição acumulada de $X \sim E(\lambda)$ é:

$$F_X(x) = \begin{cases} 1 - e^{-\lambda x} & x \ge 0 \\ 0 & x < 0 \end{cases}$$

Fazendo $u = F_X(x) = 1 - e^{-\lambda x}$ e resolvendo em ordem a x, obtém-se:

$$x = -\frac{\ln(1-u)}{\lambda}$$

iii) Geração de realizações da distribuição normal $N(0,1)$

Uma vez que não é possível obter analiticamente $x = F_X^{-1}(u)$ recorre-se por exemplo ao método de Box e Muller que, a partir de duas realizações u e v de uma variável aleatória U~(0,1) as transforma em duas observações normais estandardizadas $\left(N(0,1)\right)$ usando as transformações:

$$x_1 = \sqrt{-2\ln u}\cos(2\pi v)$$

$$x_2 = \sqrt{-2\ln u}\,\mathrm{sen}(2\pi v)$$

Se se pretender que as observações geradas sigam uma distribuição normal qualquer $Y \sim N(\mu,\sigma^2)$, apenas haverá que modificar convenientemente a expressão para x obtida anteriormente do seguinte modo:

$$y = \mu + \sigma x_1$$

MÉTODO DA TRANSFORMAÇÃO INVERSA
V.A. Discretas

Admita-se que a variável aleatória X é discreta, que assume os valores $x_1, x_2, \ldots$ e que $x_1 < x_2 < \ldots$. Neste caso, a função de distribuição acumulada F_X é dada por:

$$F_X(x) = P(X \leq x) = \sum_{x_i < x} f(x_i)$$

em que $f_X(x_i) = P(X = x_i)$ é a função de probabilidade.

Então um algoritmo para gerar uma realização da variável aleatória X é o seguinte:

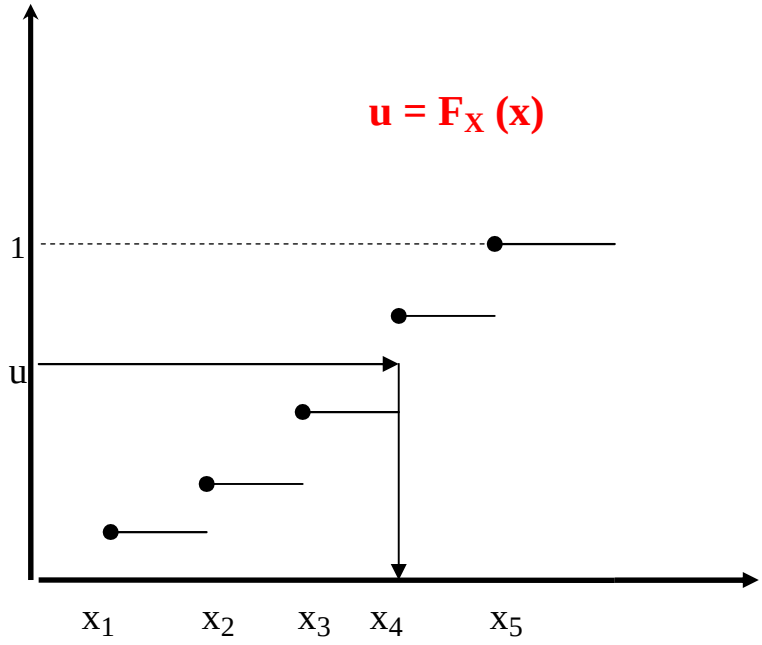
Dada a função de distribuição acumulada $F_X(x) = P(X \leq x)$

** Gerar uma realização u da v.a. $U \sim U(0,1)$

** Determinar o menor inteiro positivo α tal que $u \leq F_X(x_\alpha)$

** Obter a realização $x = x_\alpha$ da v.a. X.

A figura abaixo ilustra o método da transformação inversa para variáveis aleatórias discretas.



EXEMPLO:

i) Geração de realizações da distribuição de Bernoulli de parâmetro p.

A função de probabilidade da v.a. de Bernoulli é como já vimos:

$$p_X(x) = \begin{cases} 1-p & x = 0 \\ p & x = 1 \\ 0 & \text{outros valores} \end{cases}$$

e a função de distribuição acumulada é:

$$F_X(x) = \begin{cases} 0 & x < 0 \\ 1-p & 0 \le x < 1 \\ 1 & x \ge 1 \end{cases}$$

Gerada uma realização u da v.a. $U \sim U(0,1)$ então se $0 \le u \le 1-p$ temos x = 0 , se pelo contrário $1-p < u \le 1$ obtemos x = 1.

DISTRIBUIÇÃO DA MÉDIA AMOSTRAL
(DISTRIBUIÇÃO AMOSTRAL DE $\overline{X}$)

Para uma v.a. X , representativa de uma dada população, e para amostras de dimensão n, a média amostral é definida como vimos, por:

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

Atendendo às propriedades do valor esperado temos :

$$E(\overline{X}) = E\left(\frac{1}{n} \sum_{i=1}^{n} X_i \right) = \frac{1}{n} E\left(\sum_{i=1}^{n} X_i \right) = \frac{1}{n} \sum_{i=1}^{n} E(X_i)$$

$$= \frac{1}{n} \sum_{i=1}^{n} E(X) = \frac{1}{n} (n \cdot E(X)) = E(X)$$

isto é:

$$\mu_{\overline{X}} = \mu_X$$

Para definir a variância, é necessário estabelecer a distinção entre **amostras aleatórias simples** (população finita com reposição ou população infinita) e **amostras aleatórias que não sejam simples** (população finita sem reposição).

No caso das amostras aleatórias simples, as variáveis X_i são independentes. Então, atendendo às propriedades da variância, temos que:

$$\mathrm{Var}\left(\overline{X}\right)=\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)=\frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n}X_i\right)$$

$$=\frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}\left(X_i\right)=\frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}\left(X\right)=\frac{1}{n^2}\left(n\cdot\mathrm{Var}\left(X\right)\right)$$

$$=\frac{\mathrm{Var}\left(X\right)}{n}$$

ou seja:

$$\sigma^2_{\overline{X}}=\frac{1}{n}\cdot\sigma^2_X$$

Se as amostras aleatórias não forem simples é possível demonstrar que a variância da média amostral é dada por:

$$\sigma^2_{\overline{X}}=\frac{N-n}{N-1}\cdot\frac{1}{n}\cdot\sigma^2_X$$ (população finita, amostragem sem reposição).

Em que N e n representam respectivamente a dimensão da população e da amostra.

O termo $\frac{N-n}{N-1}$ ($n \leq N$), designa-se por factor de correcção (ou de redução) para populações finitas. De um modo geral este factor é inferior à unidade visto que $1 < n < N$.

Nota:

- $\mathbf{N \to \infty}$: o factor de correcção tende para a unidade e a amostra, que se admite ter dimensão finita, é aleatória simples, haja ou não reposição.

- n = 1 : o factor de redução é igual à unidade. Neste caso, tendo a amostra apenas um elemento, não há qualquer diferença entre a amostragem com ou sem reposição. A média amostral coincide com a variável original ($\overline{X} = X$).

- N = n : a amostra coincide com a população e portanto a média amostral corresponde ao valor esperado da variável original, sendo portanto uma constante. O factor de redução e a variância da média amostral são nulos.

Calculado o valor esperado e a variância da média amostral, falta **especificar a forma da distribuição de** $\overline{X}$**.**

Se $X \sim N\left(\mu_X, \sigma_X^2\right)$ então a média amostral é uma combinação linear das variáveis X_i todas elas com distribuição $N\left(\mu_X, \sigma_X^2\right)$ e independentes entre si (o facto de X ser Normal presume que a população seja infinita, e que portanto, as amostras aleatórias sejam simples). Neste caso, como já vimos, a média amostral $\overline{X}$ segue uma distribuição Normal.

Se a distribuição de X não for Normal, podemos recorrer ao teorema do limite central, o qual estabelece que " a soma de n v.a. independentes tende para a distribuição Normal quando $n \to \infty$ " , isto é:

$$\overline{X} \underset{n \to \infty}{\rightarrow} N\left(\mu_X, \frac{\sigma_X^2}{n}\right) \quad \text{ou} \quad \frac{\overline{X} - \mu_X}{\frac{\sigma_X}{\sqrt{n}}} \underset{n \to \infty}{\rightarrow} N(0,1)$$

DISTRIBUIÇÃO DA VARIÂNCIA AMOSTRAL (DISTRIBUIÇÃO AMOSTRAL DE S^2)

Se a população em análise for caracterizada por uma v.a. contínua $X \sim N(\mu, \sigma^2)$ e se desta população forem obtidas amostras aleatórias de dimensão n com variância amostral S^2, então podemos afirmar que:

i) $\overline{X}$ e S^2 são independentes

ii) A v.a.: $\frac{n-1}{\sigma^2} S^2 \sim \chi^2_{n-1}$

As afirmações anteriores correspondem ao enunciado de um teorema cuja demonstração integral está fora do âmbito desta disciplina. Podemos porém salientar alguns aspectos dessa demonstração.

Demonstração: Comecemos por verificar que:

$$\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2=\sum_{i=1}^{n}\left(X_i-\mu\right)^2-n.\left(\overline{X}-\mu\right)^2 \qquad \text{(ver)}$$

Se agora dividirmos ambos os membros da equação por σ^2 obtemos:

$$\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2=\frac{n-1}{\sigma^2}\cdot S^2+\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right)^2$$

atendendo a que:

$$S^2=\frac{1}{n-1}\cdot\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2$$

Então, com base nos teoremas anteriormente apresentados temos no primeiro membro uma v.a. com distribuição qui-quadrado com n graus de liberdade e no segundo termo do segundo membro uma v.a. com distribuição qui-quadrado com 1 grau de liberdade.

Nota: Quando uma amostra aleatória i.i.d. é obtida de uma distribuição $N\left(\mu,\sigma^2\right)$, as variáveis aleatórias:

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}\left(X_i-\mu\right)^2\sim\chi_n \quad \text{e} \quad \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2=\frac{(n-1)S^2}{\sigma^2}\sim\chi_{n-1}$$

Isto é, a substituição da média populacional (μ) pela média amostral $\overline{X}$, resulta na perda de um grau de liberdade.

DISTRIBUIÇÃO DA PROPORÇÃO AMOSTRAL

(DISTRIBUIÇÃO AMOSTRAL DE $P = \frac{Y}{n}$)

Consideremos uma população constituída apenas por elementos de dois tipos. O valor p, que corresponde à proporção de um dos dois tipos na população, designa-se por proporção binomial (a proporção do outro tipo é q = 1 – p) . Se extrairmos uma amostra de dimensão n desta população ela irá incluir y elementos de um dos tipos. Então $\frac{y}{n}$ é uma proporção para uma dada amostra. Esta estimativa corresponde a um valor particular da **estatística** (**ou estimador**) proporção amostral , $P = \frac{Y}{n}$.

Se o processo de amostragem for aleatório simples (o que pressupõe que a população é infinita ou caso contrário, que a amostragem é efectuada com reposição) então a v.a. Y, que representa o número de elementos de um dado tipo numa qualquer amostra, pode ser interpretada como o número de "sucessos" em n experiências de Bernoulli. Como vimos já, nesta situação Y apresenta uma distribuição b ~ (n,p). O teorema do limite central permite-nos afirmar que para valores de n suficientemente elevados podemos aproximar a distribuição de Y por uma distribuição Normal com $\mu = n \cdot p$ e $\sigma^2 = n \cdot p \cdot (1 - p)$. Então podemos concluir que:

$$P = \frac{Y}{n} \sim N\left(p, \frac{p \cdot (1 - p)}{n} \right)$$

DISTRIBUIÇÃO DO MÁXIMO E DO MÍNIMO DA AMOSTRA

Seja uma variável aleatória X com função densidade de probabilidade $f_X(x)$ e função de distribuição $F_X(x)$.
Dada uma amostra aleatória de tamanho n de X e sendo:

$$\mathbf{K = Mín}(\mathbf{X_1, X_2, ..., X_n})$$

$$\mathbf{M = Máx}(\mathbf{X_1, X_2, ..., X_n})$$

Então:

$$\mathbf{f_M(m) = n\left[F_X(m)\right]^{n-1} f_X(m)}$$

$$\mathbf{f_K(k) = n\left[1 - F_X(k)\right]^{n-1} f_X(k)}$$

Nota: Sugere-se a demonstração destes resultados.