- Inferência Estatística - Distribuição Amostral

Objetivo: tirar conclusões sobre uma população com base na informação de uma amostra.

$\rightarrow$ estimação

$\rightarrow$ testes de hipóteses

***Parâmetro*:** quantidades desconhecidas da população e sobre as quais temos interesse.

Ex: μ - média da população

Estimador: combinação dos elementos da amostra, construída com a finalidade de representar, ou estimar, um parâmetro de interesse na população.

Ex: $\overline{X}$ - média da amostra (estimador de μ)

Estimativa: valor numérico assumido pelo estimador.

Ex: $\overline{x}$ é o valor de $\overline{X}$ *para a amostra observada.*

Estudamos algumas distribuições teóricas de probabilidade: distribuição *binomial* e *normal*.

Probabilidade $\Rightarrow$ os parâmetros da distribuição eram conhecidos $\Rightarrow$ calculamos probabilidades

Inferência $\Rightarrow$ os valores desses parâmetros não são conhecidos.

A amostra deve ser “representativa” da população da qual ela é selecionada.

Se não for, as conclusões extraídas sobre a população podem estar distorcidas ou viesadas.

Exemplos:

1. Fazer uma afirmação sobre o nível sérico médio de colesterol para todos os homens de 20 a 74 anos de idade $\Rightarrow$ amostramos somente homens acima de 60 anos $\Rightarrow$ é provável que nossa estimativa da média da população seja muito alta.

2. Estimar a proporção de eleitores que pretendem votar no candidato A $\Rightarrow$ amostra é selecionada dentro da USP.

$\rightarrow$ Que estimador usar nos exemplos acima?

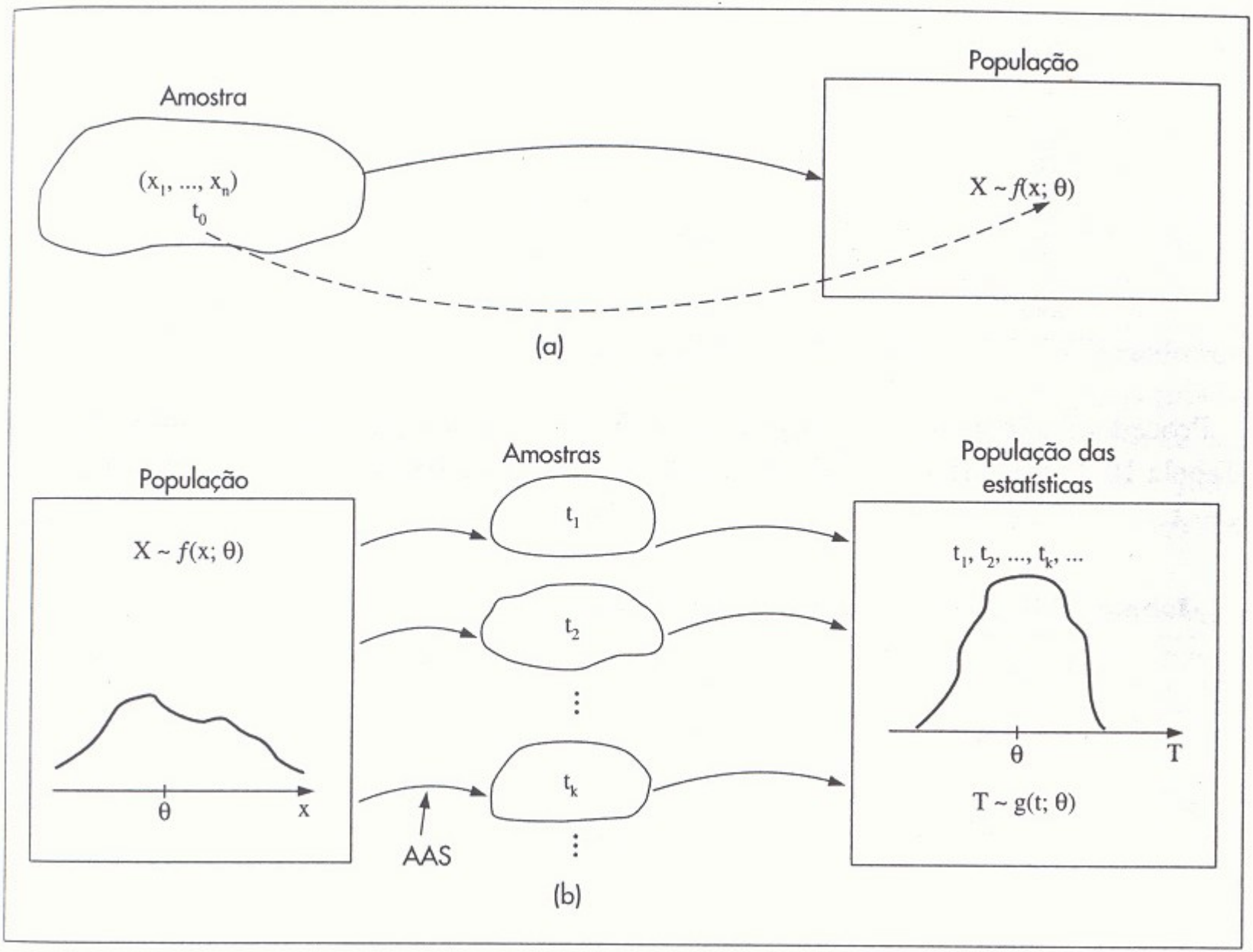
$(X_1, X_2,...,X_n) \Rightarrow$ representa uma amostra de tamanho n.
Estimador $\Rightarrow f(X_1, X_2,...,X_n)$.

Os estimadores $\bar{X}$ (média amostral) e $\hat{p}$ (proporção amostral) são intuitivos e têm boas propriedades.

Estimadores são funções de variáveis aleatórias e, portanto, eles também são variáveis aleatórias.

Conseqüentemente, têm uma distribuição de probabilidades, denominada ***distribuição amostral*** do estimador.

Figura 10.1: (a) Esquema de inferência sobre θ.
(b) Distribuição amostral da estatística T.



Distribuição amostral da média

Exemplo 1: Considere uma população em que uma variável X assume um dos valores do conjunto {1, 3, 5, 5, 7}. A distribuição de probabilidade de X é dada por

x	1	3	5	7
$P(X = x)$	1/5	1/5	2/5	1/5

É fácil ver que $\mu_x = E(X) = 4{,}2$

e $\sigma_x^2 = Var(X) = 4{,}16$.

Vamos relacionar todas as amostras possíveis de tamanho $n = 2$, selecionadas ao acaso e com reposição dessa população, e encontrar a distribuição da média amostral de

$$\overline{X} = \frac{X_1 + X_2}{2},$$

sendo

X_1: valor selecionado na primeira extração,

X_2: valor selecionado na segunda extração.

Amostra (X_1,X_2)	Probabilidade	Média Amostral
(1,1)	1/25	1
(1,3)	1/25	2
(1,5)	2/25	3
(1,7)	1/25	4
(3,1)	1/25	2
(3,3)	1/25	3
(3,5)	2/25	4
(3,7)	1/25	5
(5,1)	2/25	3
(5,3)	2/25	4
(5,5)	4/25	5
(5,7)	2/25	6
(7,1)	1/25	4
(7,3)	1/25	5
(7,5)	2/25	6
(7,7)	1/25	7
	1	

A distribuição de probabilidade de $\overline{X}$ para $n = 2$ é

$\overline{x}$	1	2	3	4	5	6	7
$P(\overline{X} = \overline{x})$	1/25	2/25	5/25	6/25	6/25	4/25	1/25

Neste caso, $E(\overline{X}) = 4{,}2 = \mu_X$

e $\mathrm{Var}(\overline{X}) = 2{,}08 = \dfrac{\sigma_X^2}{2}$.

Repetindo o mesmo procedimento, para amostras de tamanho $n = 3$, temos a seguinte distribuição de probabilidade de $\overline{X}$,

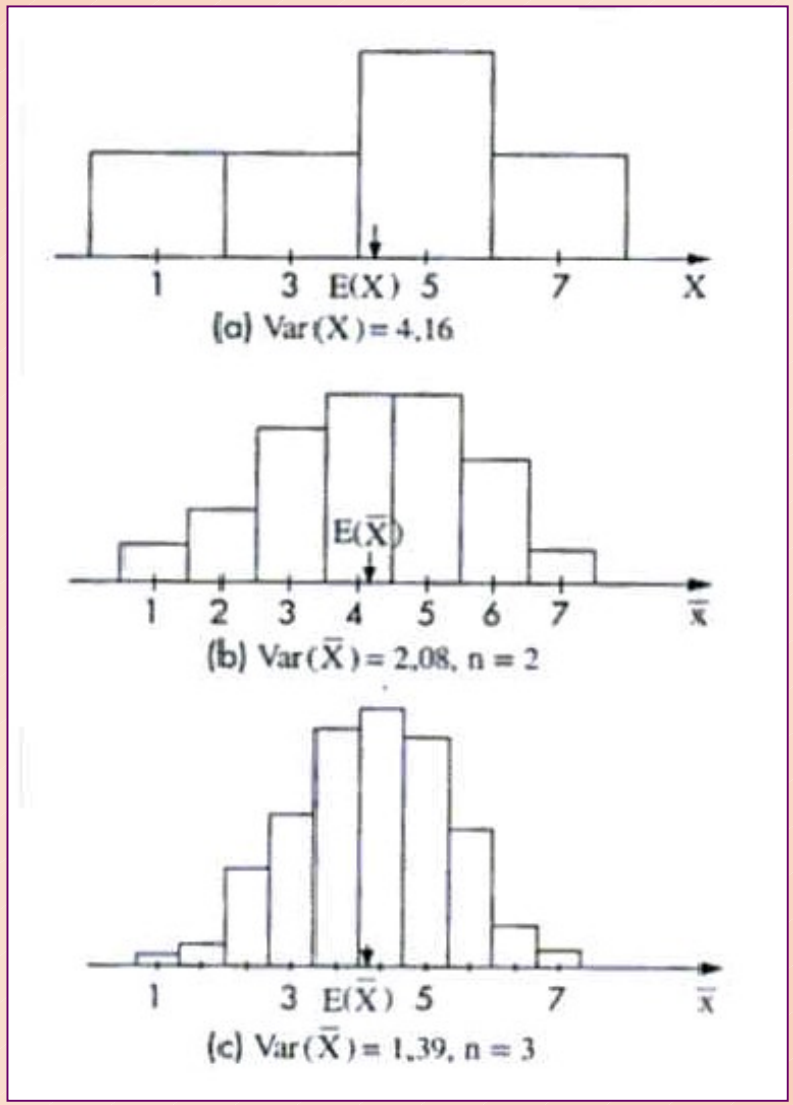
$\bar{x}$	$P(\overline{X} = \bar{x})$
1	1/125
5/3	3/125
7/3	9/125
3	16/125
11/3	24/125
13/3	27/125
5	23/125
17/3	15/125
19/3	6/125
7	1/125

Neste caso,

$$E(\overline{X}) = 4{,}2 = \mu_X$$

e $$\mathrm{Var}\left(\overline{X}\right) = 1{,}39 = \frac{\sigma_X^2}{3} \ .$$

Figura 1: Histogramas correspondentes às distribuições de X e de $\overline{X}$, para amostras de {1,3,5,5,7}.



Dos histogramas, observamos que

• conforme *n* aumenta, os valores de $\bar{x}$ tendem a se concentrar cada vez mais em torno de

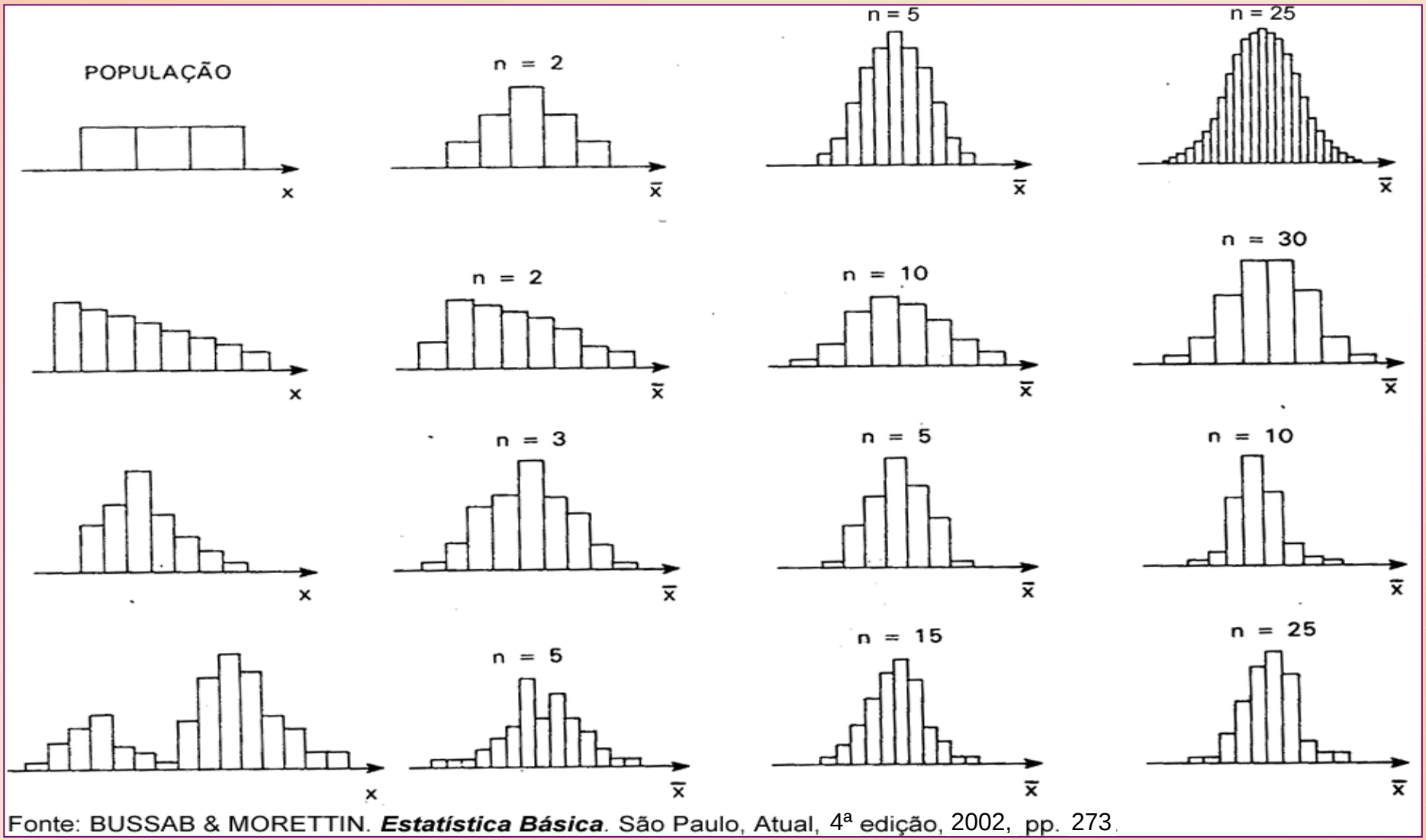
$$E(\bar{x}) = 4{,}2 = \mu_x \ ,$$

uma vez que a variância vai diminuindo;

• os casos extremos passam a ter pequena probabilidade de ocorrência;

• para *n* suficientemente grande, a forma do histograma *aproxima-se de uma distribuição normal.*

Figura 2: Histogramas correspondentes às *distribuições de* $\overline{x}$ para amostras de algumas populações



Fonte: BUSSAB & MORETTIN. ***Estatística Básica***. São Paulo, Atual, 4ª edição, 2002, pp. 273.

Esses gráficos sugerem que,

quando *n* aumenta, *independentemente da forma da distribuição* de *X* , a distribuição de probabilidade da média amostral $\bar{X}$ *aproxima-se* de uma distribuição normal.

Teorema do Limite Central

Seja X uma v. a. que tem média μ e variância σ^2. Para amostras $X_1, X_2, ..., X_n$, retiradas ao acaso e com reposição de X, a distribuição de probabilidade da média amostral $\overline{X}$ *aproxima-se*, *para n grande*, de uma distribuição normal, com média μ e variância σ^2 / n, ou seja,

$$\overline{X} \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right), \text{ para } n \text{ grande, aproximadamente.}$$

Comentários:

• Se a distribuição de X é normal, então $\overline{X}$ tem distribuição normal *exata*, ***para todo n***.

• O desvio padrão $\sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$ é denominado

erro padrão da média.

Considere uma amostra aleatória de tamanho *n* de uma variável N(10, 16).

→ Como se comporta $\overline{X}$ em função de *n* ?

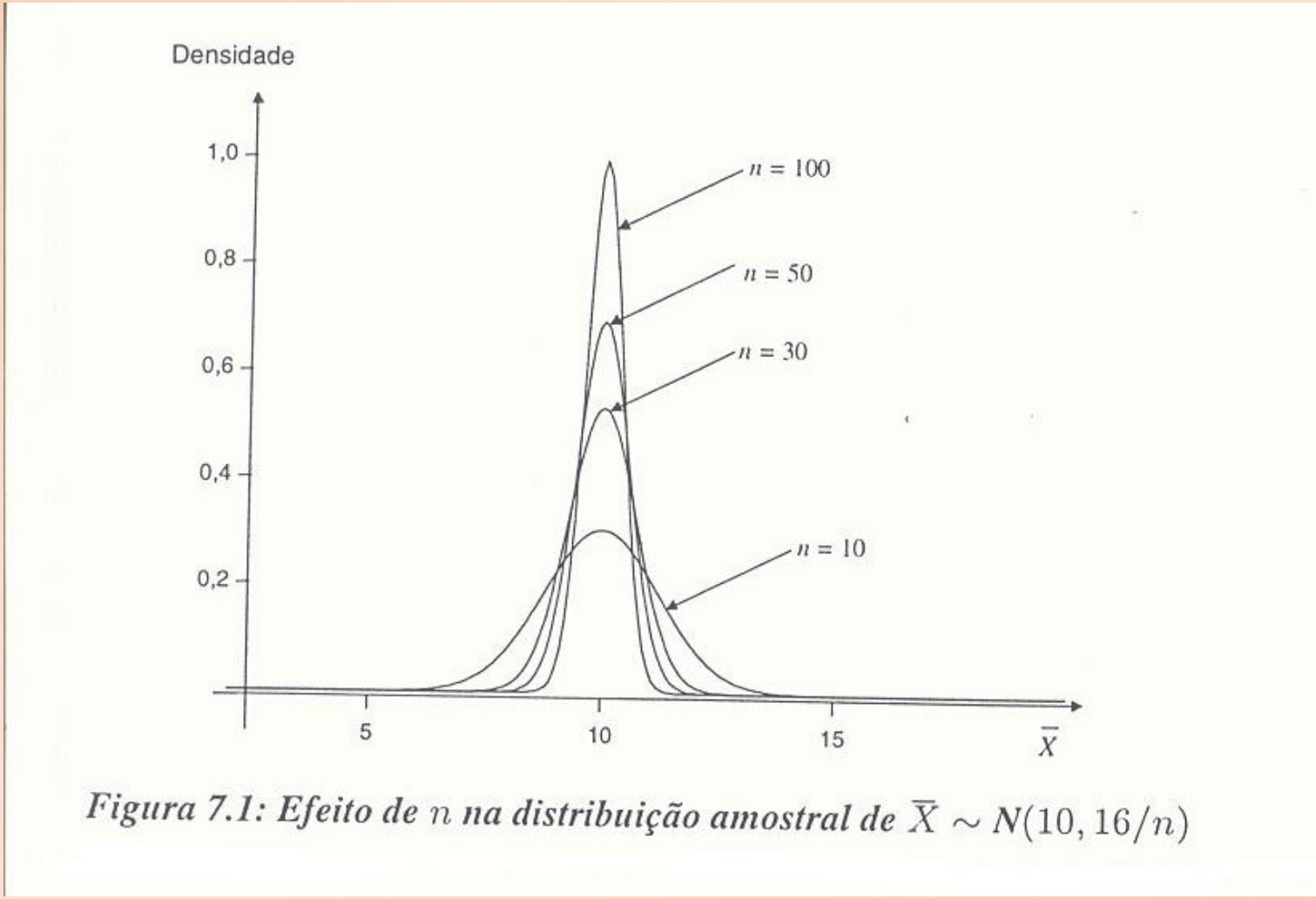


Figura 7.1: Efeito de n na distribuição amostral de $\overline{X} \sim N(10, 16/n)$

Exemplo 2:

Uma v.a. X tem média $\mu = 5{,}4$ e variância $\sigma^2 = 4{,}44$. Uma amostra com 40 observações é sorteada. Qual a probabilidade da média amostral ser maior do que 5?

$$X \rightarrow \begin{cases} \mu = 5{,}4 \\ \sigma^2 = 4{,}44 \end{cases}$$

Consideramos que $n = 40$ observações é uma amostra grande o suficiente para usar o Teorema do Limite Central. Assim,

$$\bar{X} \sim N\left(\mu;\ \frac{\sigma^2}{n}\right), \text{ isto é, } \bar{X} \sim N\left(5{,}4;\ \frac{4{,}44}{40}\right) \text{ e}$$

$$P(\bar{X} > 5) \cong P\left(Z > \frac{5 - 5{,}4}{\sqrt{\frac{4{,}44}{40}}}\right) = P(Z > -1{,}20) = A(1{,}20) = 0{,}8849,$$

lembrando que $Z \sim N(0, 1)$.

Exemplo 3:

Sabe-se que o faturamento diário de um posto de gasolina segue uma certa distribuição de média R$ 20 mil e desvio padrão de R$ 2 mil. Qual a probabilidade, em um período de 60 dias, do faturamento total ultrapassar R$ 1230 mil?

Seja X o faturamento diário de um posto de gasolina, em mil reais. Sabemos que

$$\mu = E(X) = 20$$
$$\sigma^2 = Var(X) = 4$$

Obtemos uma amostra aleatória de 60 valores de X, denotada por X_1, X_2, …, X_{60}, sendo X_i o faturamento do posto no dia i, i = 1, 2, …, 60.

Então,

$$P(X_1 + X_2 + \ldots + X_{60} > 1230) = P\left(\frac{X_1 + X_2 + \ldots + X_{60}}{60} > \frac{1230}{60}\right)$$

$$= P(\overline{X} > 20{,}5) \cong P\left(Z > \frac{20{,}5 - 20}{\sqrt{\frac{4}{60}}}\right) = P(Z > 1{,}94) = 0{,}026.$$

Exemplo 4: Considere que a distribuição dos níveis séricos de colesterol para todos os homens de 20 a 74 anos é normal com média μ = 211 mg/100ml e o desvio padrão σ = 46 mg/100ml.

Selecionamos amostras de tamanho 25 da população.

$\rightarrow$ Que proporção de amostras terá um valor médio maior do que 230 mg/100ml?

$$P(\overline{X} > 230) = ?$$

A distribuição da média amostral (n =25) é normal com média μ = 211 mg/100ml e erro padrão $\sigma/\sqrt{n}$ = 46/5 = 9,2mg/100ml.

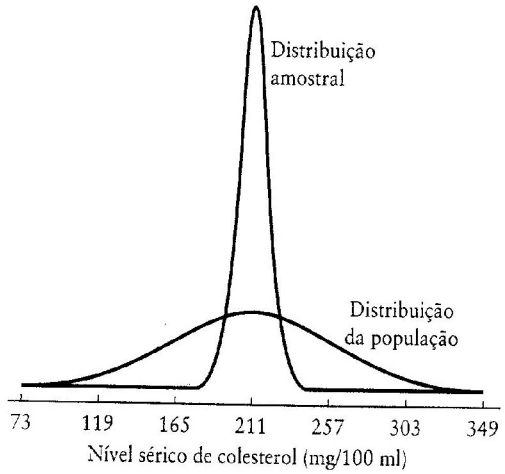


FIGURA 8.1

$$P(\overline{X} > 230) = P(Z > \frac{230-211}{9,2}) = P(Z > 2,07) = 0,019.$$

Somente 1,9% das amostras terão uma média maior do que 230 mg/100ml.

Equivalentemente, se selecionamos uma amostra de tamanho 25 da população de homens de 20 a 74 anos, a probabilidade de que o nível sérico médio de colesterol para essa amostra seja maior do que 230 mg/100ml é de 0,019.

→ Que valor médio de nível sérico de colesterol limita os 10% valores mais baixos da distrib. amostral?

$$P(\overline{X} > x) = 0,1 \Rightarrow P(Z > \frac{x-211}{9,2}) = 0,1.$$

$$\frac{x-211}{9,2} = -1,28 \Rightarrow x = 211 - 1,28 \times 9,2 = 199,2.$$

≅ 10% das amostras de tamanho 25 têm médias que são menores ou iguais a 199,2 mg/100ml.

→ Calcular os limites superior e inferior que incluem 95% das médias das amostras de tamanho 25.

$$P(\mu - x < \overline{X} < \mu + x) = 0{,}95 \Rightarrow$$

$$P\left(\frac{211 - x - 211}{9{,}2} < Z < \frac{211 + x - 211}{9{,}2}\right) = 0{,}95 \Rightarrow$$

$$P\left(\frac{-x}{9{,}2} < Z < \frac{x}{9{,}2}\right) = 0{,}95 \Rightarrow$$

$$\frac{x}{9{,}2} = 1{,}96 \Rightarrow x = 1{,}96 \times 9{,}2 = 18{,}03.$$

Limites: 211-18 = 193,0 e 211+18 = 229,0

$\cong$ 95% das médias das a.a. de tamanho 25 estão entre 193,0 mg/100ml e 229,0 mg/100ml.

$\Rightarrow$ se selecionamos uma a.a. de tamanho 25 e a amostra tem uma média maior que 229,0 ou menor que 193,0 mg/100ml então, ou a a.a. foi extraída de uma população diferente ou um evento raro se realizou.

→ Suponha que selecionamos amostras de tamanho 10 da população.

Nesse caso, o erro padrão de $\overline{X}$ é

$\sigma/\sqrt{n} = 46/\sqrt{10} = 14{,}5$ mg/100ml.

$$P(\mu - x < \overline{X} < \mu + x) = 0{,}95 \Rightarrow$$

$$\frac{x}{14{,}5} = 1{,}96 \Rightarrow x = 1{,}96 \times 14{,}5 = 28{,}5.$$

Limites: 211-28,5=182,5 e 211+28,5=239,5

$\Rightarrow \cong$ 95% das médias das a.a. de tamanho 10 estão entre 182,5 mg/100ml e 239,5 mg/100ml.

n	$\sigma/\sqrt{n}$	Intervalo contendo 95% das médias	Comprimento do intervalo
1	46,0	$120{,}8 \leq \overline{X} \leq 301{,}2$	180,4
10	14,5	$182{,}5 \leq \overline{X} \leq 239{,}5$	57,0
25	9,2	$193{,}0 \leq \overline{X} \leq 229{,}0$	36,0
50	6,5	$198{,}2 \leq \overline{X} \leq 223{,}8$	25,6
100	4,6	$102{,}0 \leq \overline{X} \leq 220{,}0$	18,0

Conforme o tamanho das amostras aumenta, a variabilidade entre as médias da amostra (erro padrão) diminui $\Rightarrow$ os limites englobando 95% dessas médias se aproximam.

comprimento do intervalo = limite superior - limite inferior.

Os intervalos que construímos foram simétricos ao redor da média da população de 211 mg/100ml.

Existem outros intervalos que contém a proporção apropriada de médias da amostra.

Suponha que desejamos construir um intervalo que contenha 95% das médias das amostras de tamanho 25.

$P(x_1 < \overline{X} < x_2) = 0{,}95$, mas com 1% da área acima de x_2 e 4% abaixo de x_1.

$$\frac{x_1 - 211}{9{,}2} = -1{,}75 \Rightarrow x_1 = 211 - 1{,}75 \times 9{,}2 = 194{,}9.$$

$$\frac{x_2 - 211}{9{,}2} = 2{,}32 \Rightarrow x_2 = 211 + 1{,}75 \times 9{,}2 = 232{,}9.$$

Podemos dizer que aproximadamente 95% das médias das amostras de tamanho 25 se encontram entre 194,9 mg/100ml e 232,3 mg/100ml.

Em geral, é preferível construir um intervalo simétrico.

intervalo assimétrico $\Rightarrow$ comprimento =

$232,3 - 194,9 = 37,4$ mg/100ml;

intervalo simétrico $\Rightarrow$ comprimento =

$229,0 - 193,0 = 36,0$ mg/100ml.

→ Qual deve ser o tamanho das amostras para que 95% de suas médias se encontrem a $\pm$ 5 mg/100ml da média μ da população?

Para responder isso, não é necessário conhecer o valor do parâmetro μ.

Precisamos encontrar o tamanho da amostra ***n*** para o qual $P(\mu - 5 < \overline{X} < \mu + 5) = 0{,}95$

$$\Rightarrow P\left(\frac{\mu - 5 - \mu}{46/\sqrt{n}} < Z < \frac{\mu + 5 - \mu}{46/\sqrt{n}}\right) = 0{,}95$$

$$\Rightarrow \frac{5}{46/\sqrt{n}} = 1{,}96 \Rightarrow \sqrt{n} = \frac{1{,}96 \times 46}{5} \Rightarrow n = 325{,}2.$$

Amostras de tamanho 326 seriam exigidas para que 95% das médias das amostra se encontrem a ±5 mg/100ml da média μ da população.

Ou, se selecionamos uma amostra de tamanho 326 da população e calculamos sua média, a probabilidade de que a média da amostra esteja a ± 5 mg/100ml da verdadeira média μ da população é 0,95.