Inferência Estatística

Estimadores

Distribuição Amostral

Teorema Limite Central

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

estimação

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 de possíveis problemas:

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. u

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

Que estimador usar nos exemplos acima?

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

= função dos dados da amostra.

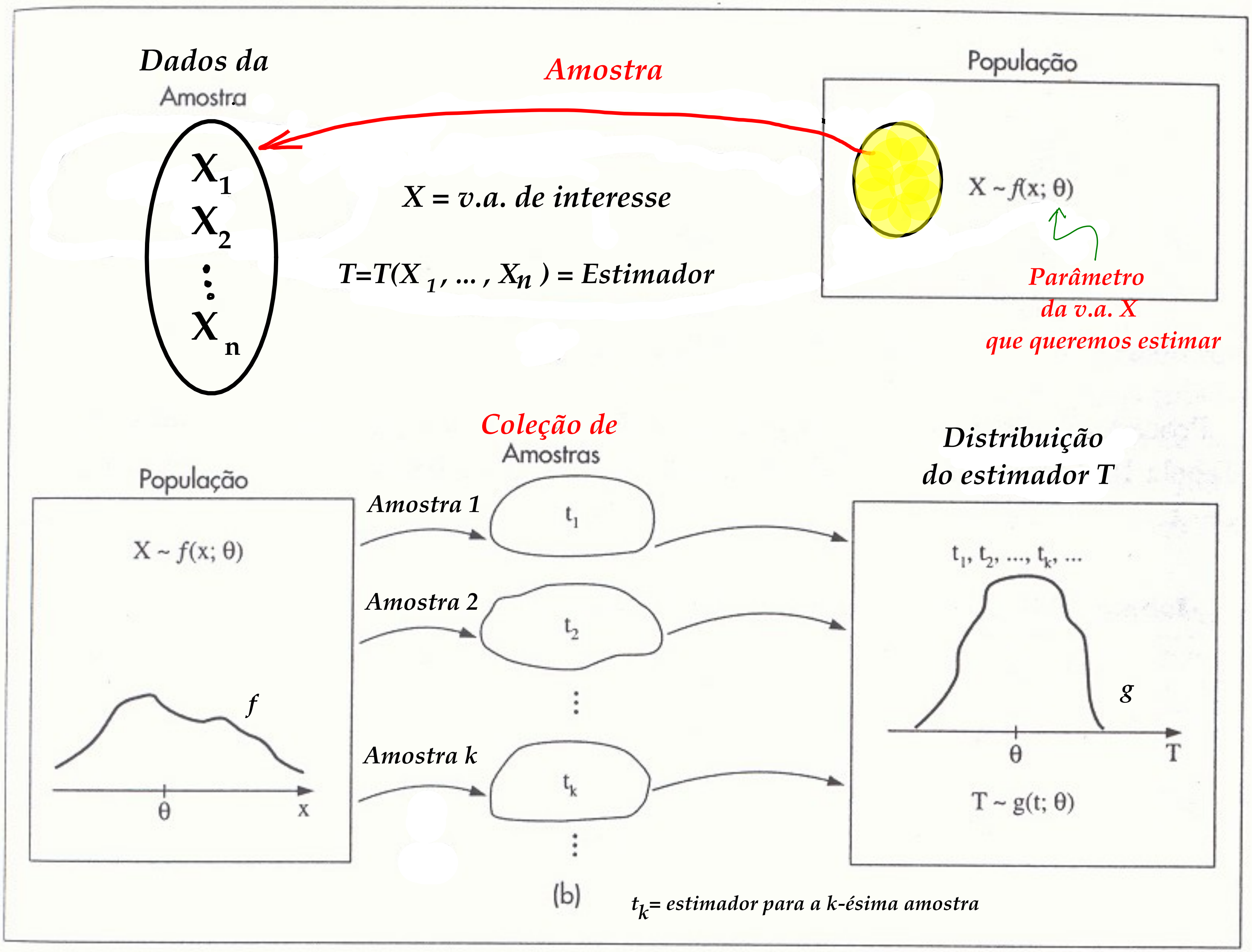
Exemplos:

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.

Distribuição Amostral do Estimador



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

(Verifique!)

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

$$\overline{X} = \frac{X_1 + X_2}{2}$$ = ***Média Amostral***

sendo

X_1: valor selecionado na primeira extração,

X_2: valor selecionado na segunda extração.

(X_1, X_2) = "amostra aleatória com n=2"

$\overline{X}$ = "estimador da média μ_x a partir dessa amostra"

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	

Portanto:

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

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

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

(Verifique!)

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

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

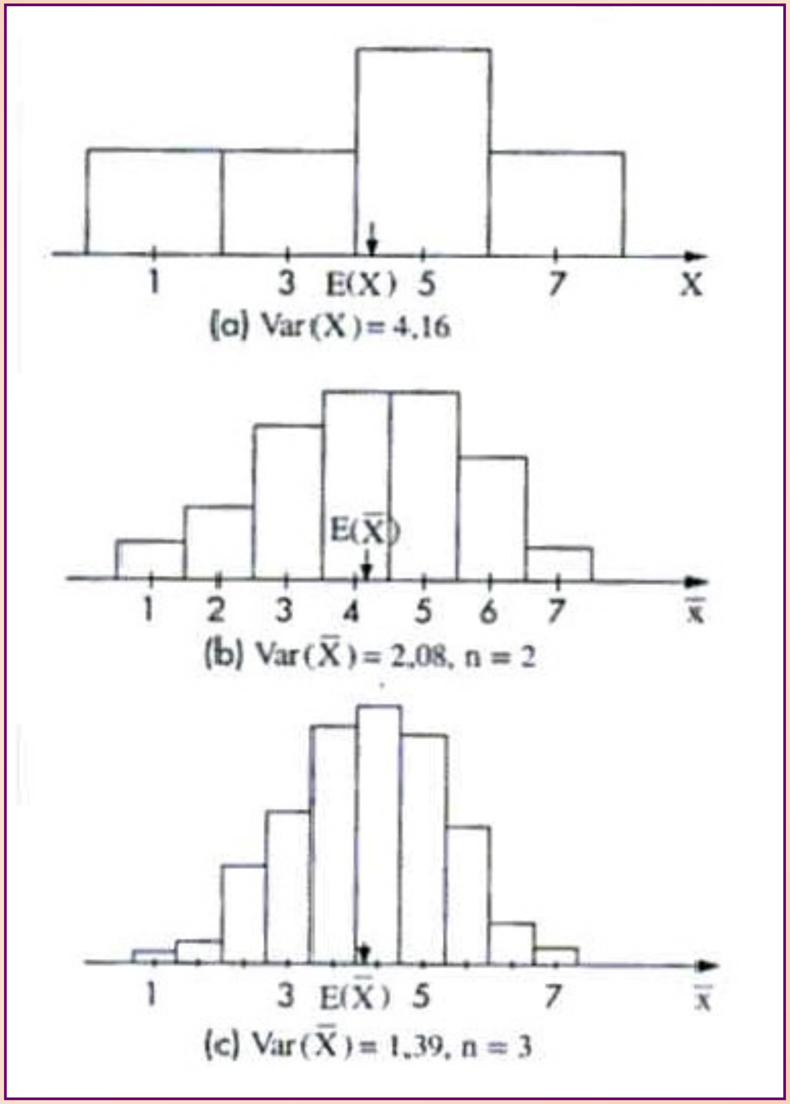
$\bar{x}$	$P(\overline{X} = \bar{x})$
1	1/125
5/3	3/125
7/3	9/125
3	16/125
11/3	24/125
13/3	27/125
5	23/125
17/3	15/125
19/3	6/125
7	1/125

Neste caso,

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

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

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

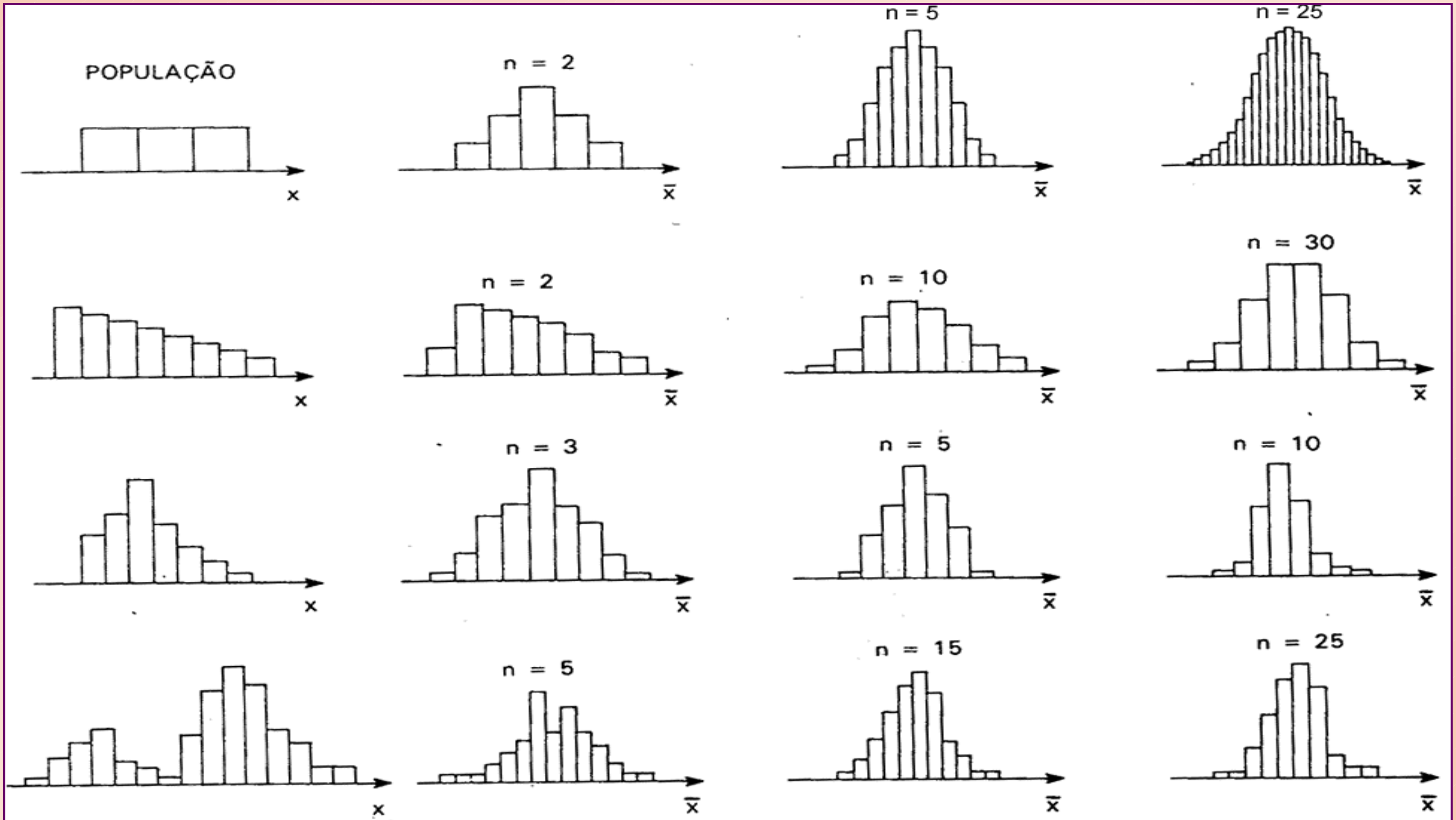
$$E(\bar{X}) = 4{,}2 = \mu_x \, ,$$

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* $\bar{X}$ para amostras de algumas populações



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

Esses gráficos sugerem que, ***se***

$E(X) = \mu_x$ *e* $Var(X) = \sigma_x^2$

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

$$\overline{X} \underset{\text{(aprox.)}}{\sim} N\left(\mu_x ; \frac{\sigma_x^2}{n}\right)$$

ou

$$\frac{\overline{X} - \mu_x}{\sigma_x / \sqrt{n}} \underset{\text{(aprox.)}}{\sim} N(0 ; 1)$$

Valem aproximadamente, para "n grande"

Teorema Limite Central

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

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

Comentários:

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

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

erro padrão da média.

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

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

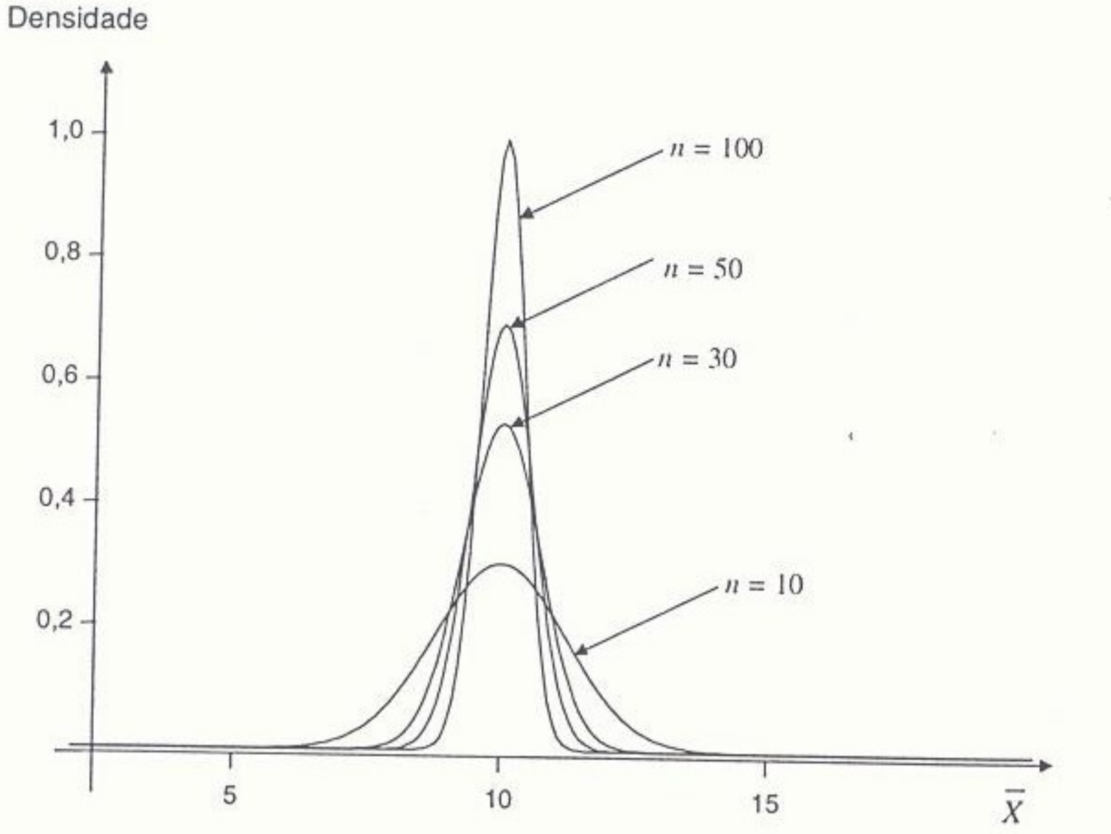


Figura 7.1: Efeito de n na distribuição amostral de $\overline{X} \sim N(10, 16/n)$

Exemplo 2:

Uma v.a. X tem média μ = 5,4 e variância σ^2 = 4,44. Uma amostra com 40 observações é sorteada. Qual a probabilidade da média amostral ser maior do que 5?

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

Consideramos que n = 40 observações é uma amostra grande o suficiente para usar o Teorema 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,

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

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

Aproximação Normal para a Binomial

Caso particular do TLC com amostra casual simples $(X_1, \ldots, X_n)$ de uma v.a. X com

$$X \sim \text{Bernoulli}(p), \quad \text{ou seja:} \begin{cases} P(X=1)=p \\ P(X=0)=1-p \end{cases}$$

temos $E(X) = \mu_x = p$

e $Var(X) = \sigma_x^2 = p(1-p)$

TLC:

$$\bar{X} \underset{\text{(aprox.)}}{\sim} N\left(p; \frac{p(1-p)}{n}\right)$$

ou

$$\frac{\bar{X} - p}{\sqrt{\frac{p(1-p)}{n}}} \underset{\text{(aprox.)}}{\sim} N(0;1)$$

Valem aproximadamente, para "n grande"

Exemplo da lista de exercícios para classe:

MAE116 - Noções de Estatística

Lista de exercícios 7 - Aproximação da Binomial pela Normal - Classe

Exercício 1
Uma vacina contra gripe é eficiente em 80% dos casos. Sorteamos ao acaso 200 pessoas vacinadas e pergunta-se a probabilidade *(estimada) de termos:*

(a) pelo menos 150 imunizados;
(b) menos de 140 imunizados;
(c) no máximo 20 não imunizados.

Suponha que X é a v.a. de interesse: "X=1" indica que a vacina foi eficiente e "X=0" indica que a vacina não foi eficiente.

Temos X~Bernoulli(p)

Para as 200 pessoas vacinadas, sejam $X_1, \ldots, X_{200}$ os correspondentes resultados.

$$X_i = \begin{cases} 1 & \text{se a vacina foi eficiente para a i-ésima pessoa vacinada e} \\ 0 & \text{caso contrário.} \end{cases}$$

O total de indivíduos imunizados, dentre os 200, é dado por

$$Y = X_1 + X_2 + \ldots + X_{200} \sim \text{Binomial}(n,p)$$

Pergunta (a) $P(Y > 149) = ?$

Note que $\dfrac{Y}{200}$ *é a média amostral de X para esta amostra com n=200 e*

use o TLC para estimar $P(Y > 149)$.