Unidade IV – Inferência estatística

4.1. Introdução e histórico
4.2. Conceitos fundamentais
4.3. Distribuições amostrais e Teorema central do limite
4.4. Estimação de parâmetros
4.5. Testes de hipóteses
4.6. Quebras das pressuposições no processo de inferência
4.7. Testes de qui-quadrado

Profª. Clause Piana

4.1. Introdução e histórico

Inferência Estatística

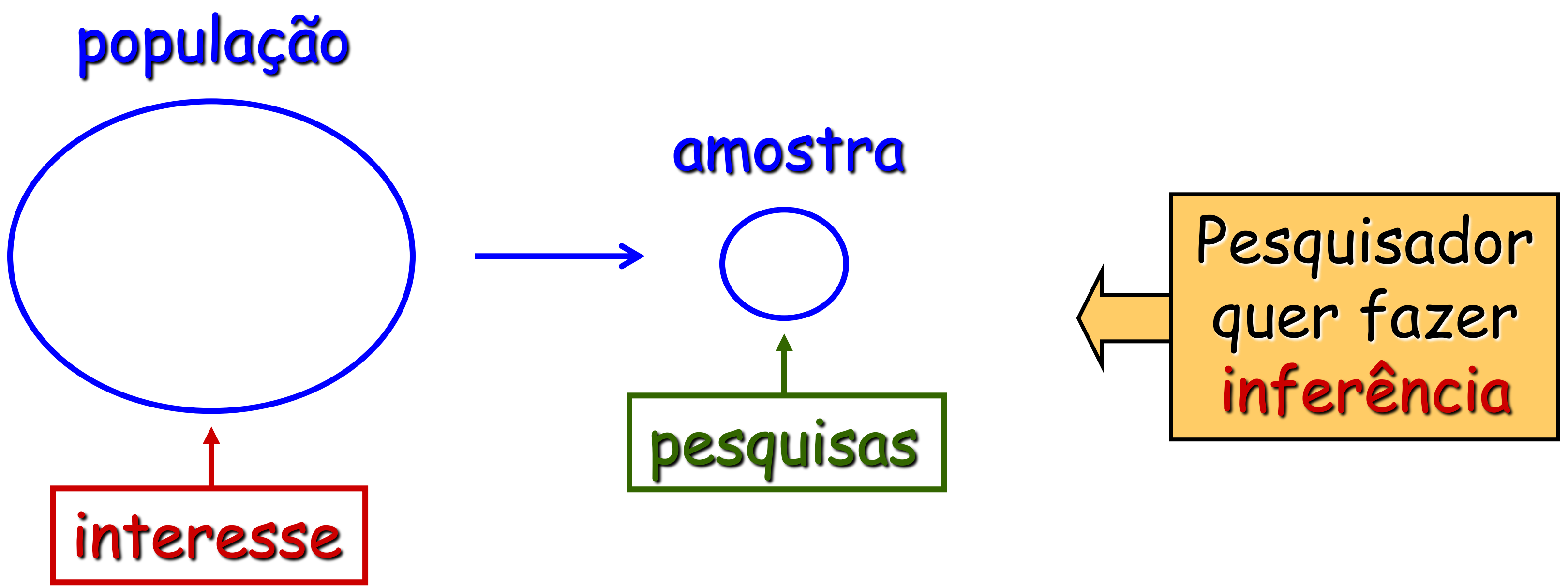
"Não é preciso beber toda a garrafa para saber se o vinho é bom."

Idéia fundamental

↓

Dar informação sobre o **todo** com base no conhecimento da **parte**

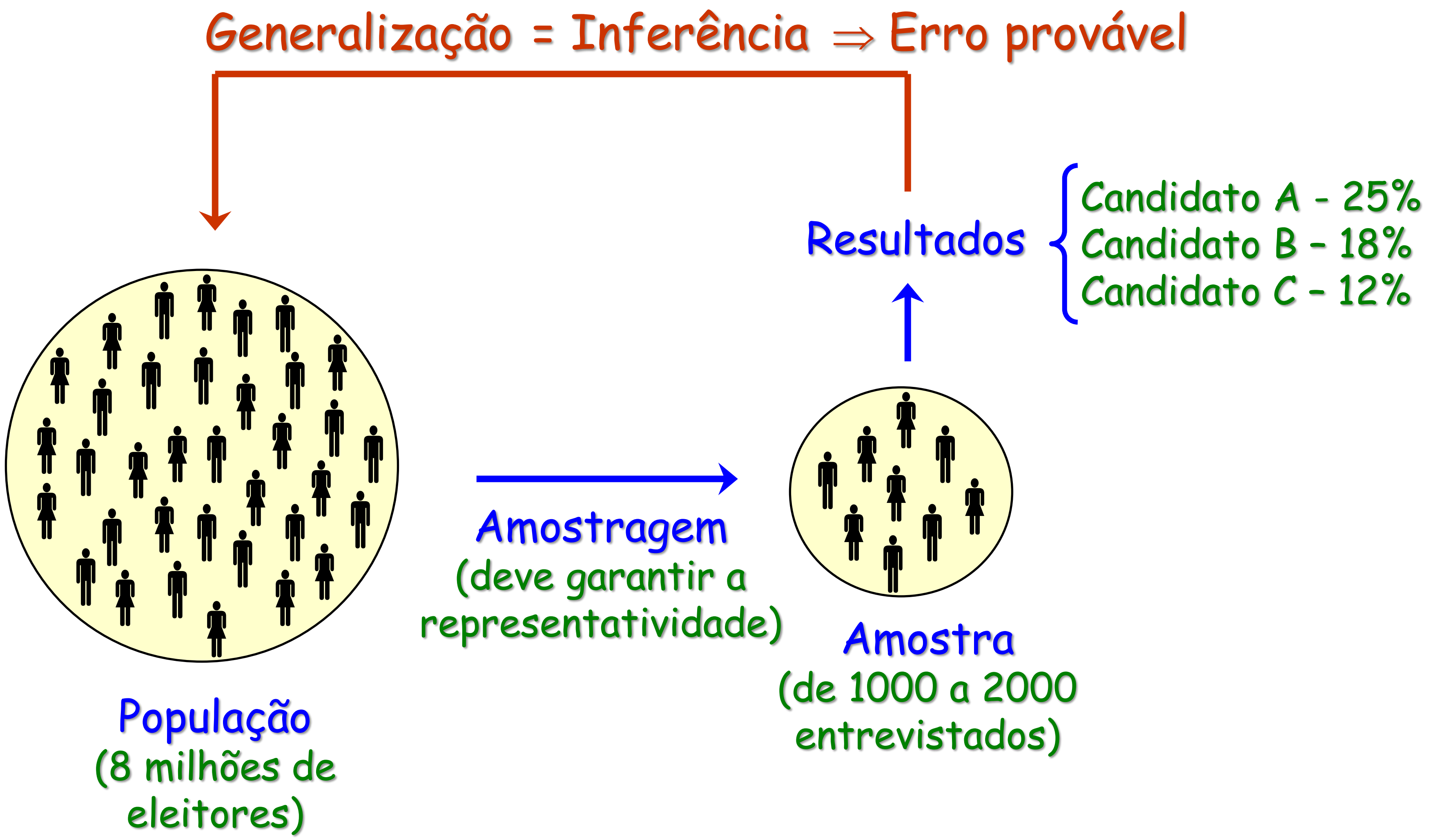
⇨ Pesquisas são feitas com **amostras**, mas o pesquisador quer **estender os resultados** que obteve para toda a **população**.



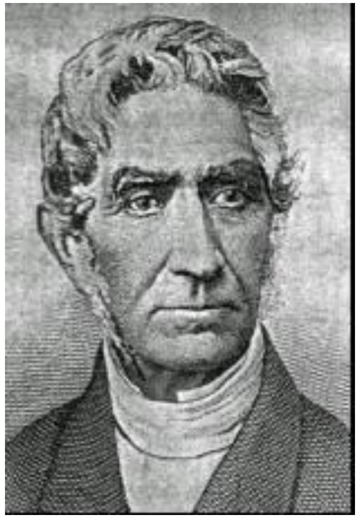
Inferência estatística

Conceito: É o conjunto de procedimentos estatísticos que têm por finalidade generalizar conclusões de uma amostra para uma população.

Exemplo: Pesquisas eleitorais no Rio Grande do Sul



Um pouco de história...



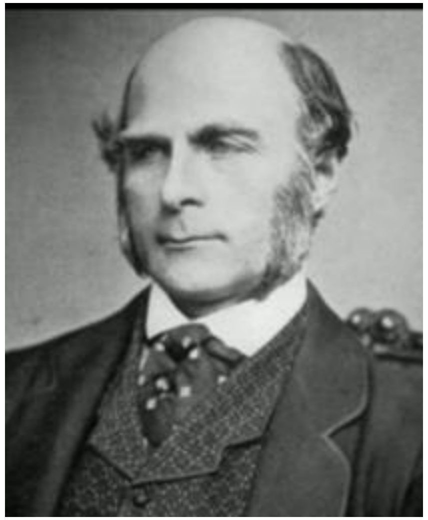
Lambert Quetelet
(1796 – 1874)

O casamento entre a estatística e o cálculo das probabilidades deve-se, em grande parte, ao astrônomo belga **Lambert Adolphe Jacques Quetelet**. Sua pesquisa abrangeu uma ampla gama de disciplinas científicas: meteorologia, astronomia, matemática, estatística, demografia, sociologia, criminologia e história da ciência. Na época, a nova ciência da probabilidade e estatística era utilizada principalmente em astronomia para estudar os erros de medição. Quetelet foi um dos primeiros a utiliza-la nas ciências sociais, planejando o que ele chamou de "física social".

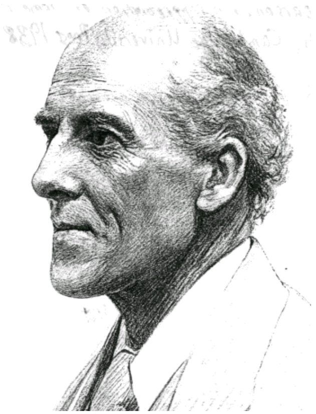
Ele estava consciente da enorme complexidade dos fenômenos sociais, mas seu objetivo era compreender as leis estatísticas subjacentes a fenômenos como taxas de criminalidade, taxas de casamento ou taxas de suicídio. Ele queria explicar os valores dessas variáveis por outros fatores sociais. Trabalhando como estatístico para as pesquisas censitárias de seu país, desenvolveu as ideias de "homem médio", que foi apresentado como um tipo ideal e abstrato que poderia ser visto como um padrão para análises sociológicas. Quetelet foi o precursor do estudo da demografia e o criador do índice de massa corporal.

Após Quételet, a Estatística teve um desenvolvimento sem precedentes, sendo o fenômeno da regularidade observado em muitos campos de pesquisa.

O inglês **Francis Galton**, considerado um dos maiores cientistas da história, foi antropólogo, eugenista, geógrafo, inventor, meteorologista, proto-geneticista, psicometrista e estatístico. Responsável pelo desenvolvimento de metodologias estatísticas essenciais como a **correlação** e a **regressão**, foi o primeiro a empregar o termo **regressão** para designar o fenômeno de retorno à média. Num estudo comparativo entre estaturas de pais e filhos, Galton observou que, quando os pais eram mais altos do que a média, os filhos tendiam a ser menores do que eles e, quando os pais eram menores do que a média, os filhos tendiam a ser maiores do que eles.



Francis Galton
(1822 – 1911)



Karl Pearson
(1857– 1936)

Estimulado pelos trabalhos de Galton, em 1890, o inglês **Karl Pearson** iniciou o estudo sobre relacionamento entre variáveis e, em 1900, deduziu a distribuição qui-quadrado, dentro de um contexto de ajustamento de distribuições. Em 1901, junto com Galton fundou a revista Biométrica cujo objetivo era o desenvolvimento da teoria estatística. O pensamento de Pearson fundamentou muitos dos métodos estatísticos "clássicos" que são de uso comum atualmente. Além das grandes contribuições, que o colocam como um dos fundadores da estatística, Pearson é lembrado também pela longa, ácida e muitas vezes rancorosa disputa que manteve com Fisher, em parte motivada por discordâncias filosóficas sobre a estatística.

Galton e os Métodos Estatísticos

Merece destaque o interesse de Galton pelas medidas e pela estatística. Ao longo de sua carreira, ele nunca parecia plenamente satisfeito com um problema até descobrir alguma maneira de quantificar os dados e analisá-los estatisticamente. Ele não se limitou a aplicar métodos estatísticos; também os desenvolveu.

O estatístico belga, **Adolph Quetelet**, tinha sido o primeiro a aplicar a dados biológicos e sociais métodos estatísticos e a curva normal de probabilidade. A curva normal fora usada em trabalhos sobre a distribuição de medidas e erros na observação científica, mas o princípio da distribuição normal só veio a ser aplicado à variabilidade humana quando Quetelet demonstrou que características antropométricas de amostras aleatórias de pessoas geravam tipicamente uma curva normal. Ele mostrou que medidas da estatura de dez mil sujeitos se aproximavam da curva normal de distribuição, e usou a expressão *l´homme moyen* (o homem médio) para exprimir a descoberta de que a maioria dos indivíduos se aglomera em torno da média ou centro de distribuição, e que um número cada vez menor vai sendo encontrado à medida que nos aproximamos dos extremos.

Galton ficou impressionado com os dados de Quetelet e supôs que resultados semelhantes poderiam ser encontrados para características mentais. Ele descobriu, por exemplo, que as notas dadas em exames universitários seguiam a mesma distribuição da curva normal dos dados de medida física de Quetelet. Devido à simplicidade da curva normal e à sua coerência em inúmeros traços, Galton propôs que um grande conjunto de medidas ou valores de características humanas poderia ser significativamente definido e resumido por dois números: o valor médio da distribuição (a média) e a dispersão ou gama de variação em torno desse valor médio (o desvio padrão).

A obra de Galton na estatística produziu uma das mais importantes medidas da ciência, a **correlação**. O primeiro relato sobre o que ele denominou "co-relações" apareceu em 1888. As técnicas modernas de determinação da validade e da confiabilidade de testes, bem como os métodos de análise fatorial, são resultados diretos da descoberta, por Galton, da correlação, produzida quando ele observou que as características herdadas tendem a regredir na direção da média. Por exemplo, **ele observou que os homens altos não são, em média, tão altos quanto os pais, enquanto os filhos de homens muito baixos são, em média, mais altos do que os pais**. Ele concebeu o meio gráfico de representar as propriedades básicas do coeficiente de correlação e desenvolveu uma fórmula de cálculo, hoje em desuso.

Galton aplicou o método da correlação a variações de medidas físicas, demonstrando, por exemplo, uma correlação entre a altura do corpo e o comprimento da cabeça. Com o estímulo de Galton, seu aluno **Karl Pearson** desenvolveu a fórmula matemática usada ainda hoje para o cálculo do coeficiente de correlação, chamada de coeficiente de correlação do produto-momento de Pearson. O símbolo do coeficiente de correlação, r, vem da primeira letra da palavra *regressão*, em reconhecimento à descoberta de Galton da tendência de as características humanas herdadas regredirem na direção da média ou mediana. A correlação é uma ferramenta fundamental das ciências sociais e do comportamento, bem como da engenharia e das ciências naturais. A partir da obra pioneira de Galton, foram desenvolvidas muitas outras técnicas estatísticas.

(Extraído do texto "As diferenças individuais: Francis Galton", de **Suely Vieira Lopes**, Pontifícia Universidade Católica de Goiás.)

William Gosset
(1857- 1936)

Em 1908, o inglês William Gosset, aluno de Pearson e também conhecido pelo pseudônimo Student, descobriu a **distribuição t** no intuito de resolver problemas relativos a pequenas amostras.

A partir de 1920, o matemático inglês **Ronald Fisher** trouxe contribuições valiosas à Estatística. Com os resultados de Gosset, Fisher descobriu rapidamente as distribuições amostrais dos coeficientes de correlação, de regressão e a distribuição da razão entre duas variâncias (distribuição F).

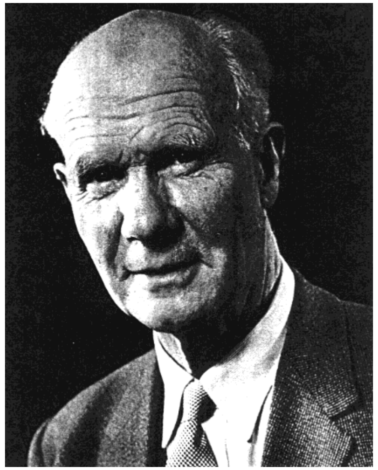
No período de 1920 a 1939, **Fisher** trabalhou na Estação Experimental de Rothamstead, na Inglaterra, onde preocupou-se com o fato de que, em situações experimentais, uma variável era explicada por várias outras, o que tornava impossível o estudo isolado de cada uma.

Para contornar este problema, desenvolveu esquemas experimentais de modo que os efeitos pudessem ser estudados isoladamente. No mesmo período, estendeu e deu ideia mais precisa a técnica chamada **análise da variância**, uma das mais poderosas utilizadas na Estatística.

Pelos trabalhos que desenvolveu em Rothamstead, Fisher é considerado o **pai da Estatística Experimental**.

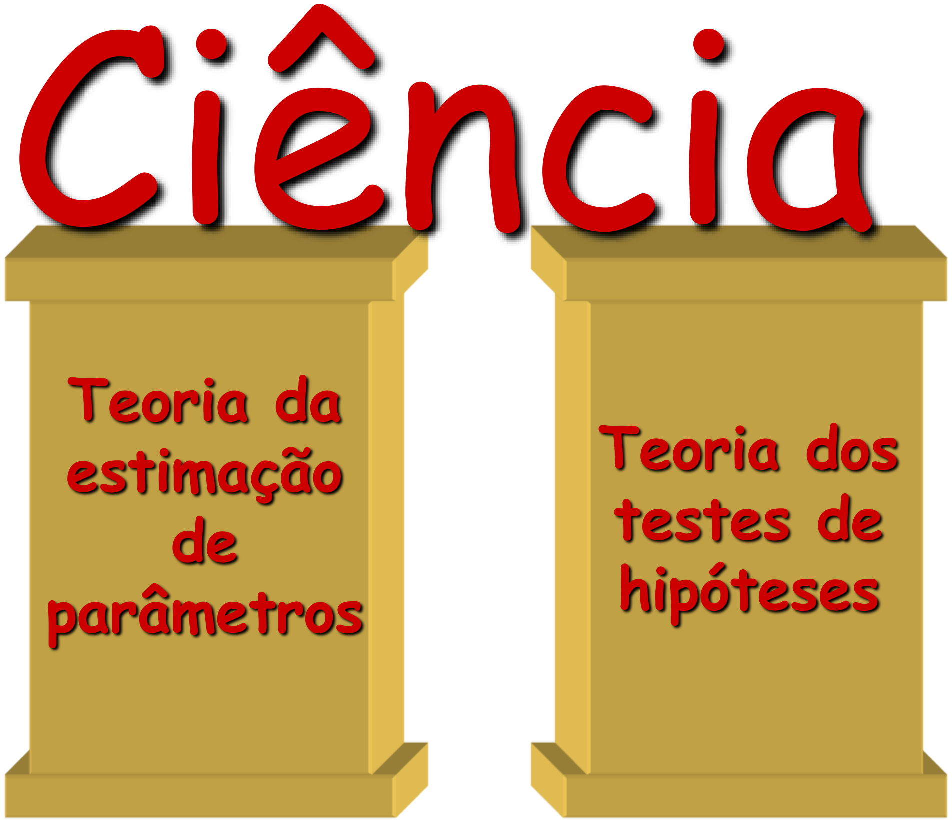
Ronald Fisher
(1890 – 1962)

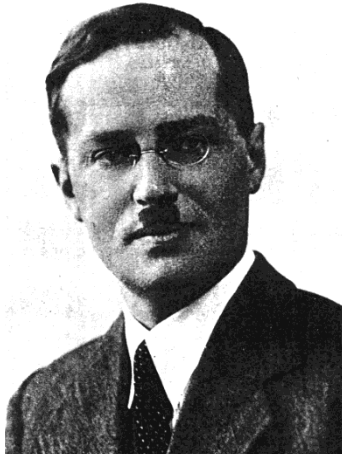
Após 1925, emergiram dois campos de extrema importância na Inferência Estatística, considerados os pilares da ciência: a **teoria da estimação de parâmetros**, desenvolvida por Fisher, e a **teoria dos testes de hipóteses**, sob a inspiração de **Egon Pearson**, único filho de Karl Pearson, e **Jerzy Neyman**.



Egon Pearson
(1895-1980)

Ronald Fisher
(1890-1962)





Jerzy Neyman
(1894-1981)

4.2 Conceitos fundamentais

- População estatística
- Amostra aleatória
- Estatística
- Distribuição amostral

4.2 Conceitos fundamentais

População é o conjunto de todos os indivíduos ou elementos que atendem a determinadas características definidoras. Estas características dependem do objetivo do estudo.

Exemplo:

Pesquisa eleitoral no Rio Grande do Sul

Objetivo: Conhecer a preferência eleitoral no estado

População: Todos eleitores votantes no RS

Amostra é um subconjunto da população retirado com o objetivo de representá-la.

Exemplo:

Pesquisa eleitoral no Rio Grande do Sul

Amostra: Conjunto de 1.000 a 2.000 eleitores votantes no RS que serão entrevistados pelos pesquisadores

Amostragem é o método de seleção que empregamos para obtenção de amostras.

Amostragem

- **Probabilística:** Todos os elementos da população têm probabilidade conhecida e diferente de zero de participar da amostra
- **Não probabilística:** Presença dos elementos na amostra deve-se a outros critérios

A amostragem **probabilística** é a mais recomendável porque garante a **imparcialidade** da amostra. Qualquer discrepância entre população e amostra é atribuída ao acaso.

Exemplos de métodos de amostragem

Amostragem probabilística:

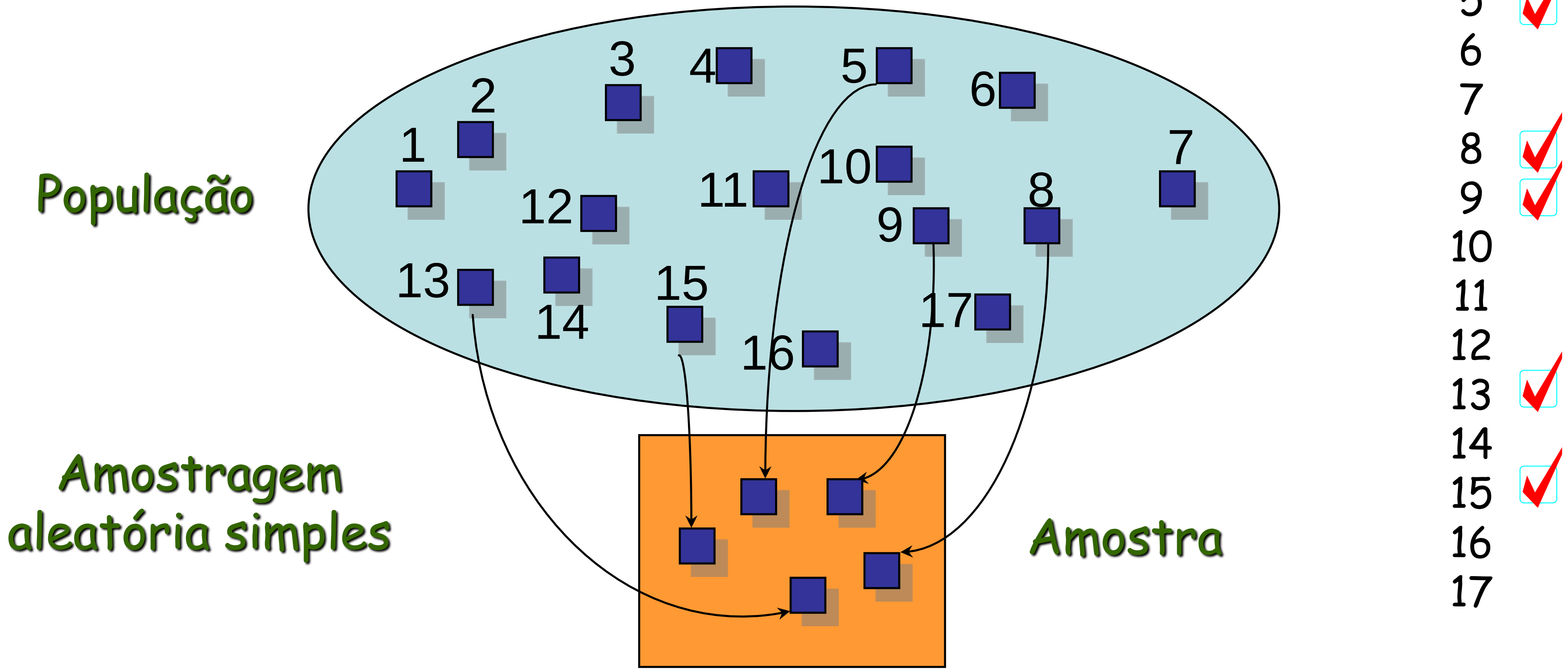
- Amostragem aleatória simples
- Amostragem aleatória estratificada
- Amostragem aleatória por conglomerados
- Amostragem aleatória sistemática

Amostragem não probabilística:

- Amostragem de conveniência
- Amostragem de julgamento
- Amostragem por quota
- Amostragem a esmo ou sem norma
- Amostragem acidental

Obtenção de uma amostra aleatória simples

1. Identificação das unidades da população
2. Definição do tamanho da amostra
3. Escolha das unidades da amostra por processo aleatório

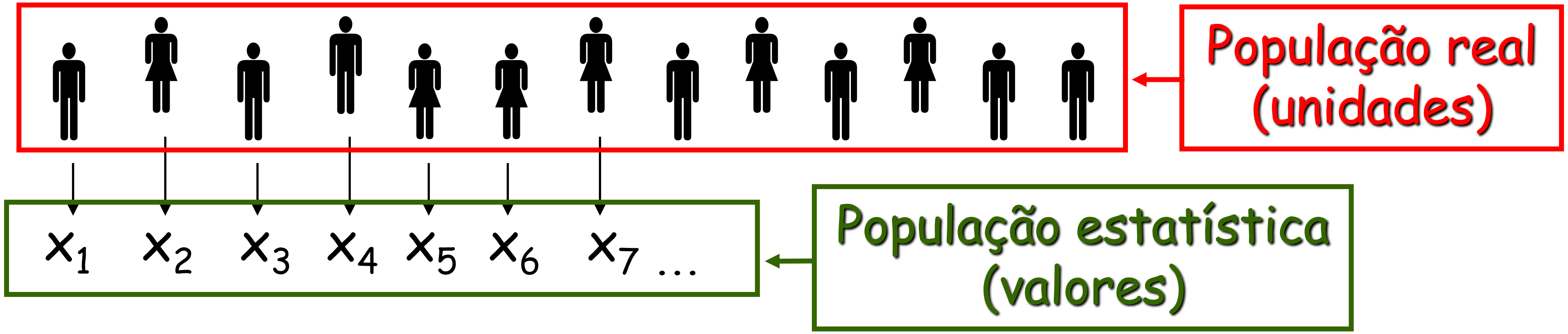


População

- **Real:** é o conjunto de todos os indivíduos ou elementos que atendem a determinadas características definidoras
- **Estatística:** é o conjunto de todos os valores de uma variável aleatória

Exemplo:

Variável: medida da estatura da população (cm)



População estatística é o conjunto de todos os valores de uma variável aleatória, cuja distribuição de probabilidade é conhecida ou passível de ser obtida.

⇨ Para utilizar os conceitos de probabilidade em estatística é essencial saber qual é a distribuição de probabilidade da variável em estudo.

Variável aleatória **X = número de peças defeituosas** → distribuição binomial? / distribuição de Poisson? / distribuição normal? / outra distribuição?

Variável aleatória **Y= quantidade de chuva (mm)** → distribuição binomial? / distribuição de Poisson? / distribuição normal? / outra distribuição?

População estatística e amostra aleatória

População estatística é o conjunto de todos os valores de uma variável aleatória, cuja distribuição de probabilidade é conhecida ou passível de ser obtida.

⇨ Para utilizar os conceitos de probabilidade em estatística é essencial saber qual é a distribuição de probabilidade da variável em estudo.

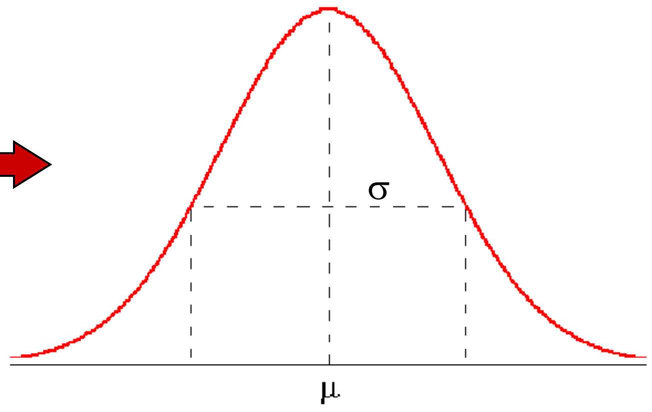
Amostra aleatória é aquela cujos elementos $[X_1, X_2, ..., X_n]$ são todos **independentes** entre si e têm a mesma distribuição de probabilidade da população (X).

⇨ A escolha **com reposição** serve muitas vezes para garantir a independência entre os elementos da amostra.

População estatística e amostra aleatória

População estatística (X) ⟷ Distribuição de probabilidade

$\mu \quad \sigma_X^2 \quad \sigma_X$ ⟷



↓

Amostra aleatória da distribuição

$X_1, X_2, \ldots, X_n$

N = tamanho da população

n = tamanho da amostra

$\longrightarrow$ $[X_1, X_2, ..., X_n]$ $\leftarrow$ amostra aleatória

com reposição

Como os valores que compõem a amostra são aleatórios, a amostra é uma variável aleatória.

4 1 4 6 3 2 1 5 4 3

$n = 2$

$[X_1, X_2] \rightarrow \begin{cases} (1, 1) \\ (1, 2) \\ (1, 3) \\ \dots \\ (6, 6) \end{cases}$

$X = x$	1	2	3	4	5	6	Σ
$P(X = x)$	$\frac{2}{10}$	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{3}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	1

distribuição da população (X)

$X_1 = x_1$	1	2	3	4	5	6	Σ
$P(X_1 = x_1)$	$\frac{2}{10}$	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{3}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	1

$X_2 = x_2$	1	2	3	4	5	6	Σ
$P(X_2 = x_2)$	$\frac{2}{10}$	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{3}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	1

X_1 e X_2 têm a mesma distribuição da população

4 1 4 6 3 2 1 5 4 3 → $n = 2$ $[X_1, X_2]$

$$\begin{cases} (1,1) \\ (1,2) \\ (1,3) \\ \ldots \\ (6,6) \end{cases}$$

$X = x$	1	2	3	4	5	6	Σ
$P(X = x)$	$\frac{2}{10}$	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{3}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	1

$X_1 = x_1$	1	2	3	4	5	6	Σ
$P(X_1 = x_1)$	$\frac{2}{10}$	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{3}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	1

$X_2 = x_2$	1	2	3	4	5	6	Σ
$P(X_2 = x_2)$	$\frac{2}{10}$	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{3}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	1

E as medidas descritivas?

$$E(X) = \sum x\,p(x) = 3{,}3$$

$$V(X) = E(X^2) - \mu^2 = 2{,}41$$

$$E(X_1) = \sum x_1\,p(x_1) = 3{,}3$$

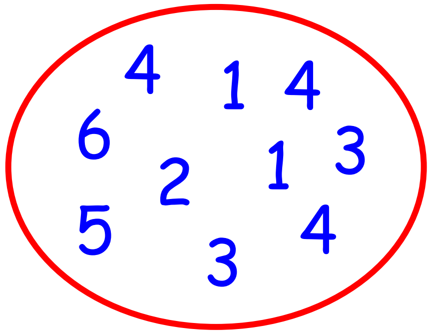
$$V(X_1) = E(X_1^2) - \mu^2 = 2{,}41$$

$$E(X_2) = \sum x_2\,p(x_2) = 3{,}3$$

$$V(X_2) = E(X_2^2) - \mu^2 = 2{,}41$$

Amostra aleatória

- Todos os seus elementos ($X_{i's}$) são independentes entre si
- Todos os $X_{i's}$ têm a mesma distribuição de probabilidade da população
- Todos os $X_{i's}$ têm a mesma média e mesma variância que a população: $E(X_i) = \mu$ e $V(X_i) = \sigma^2$



Amostra aleatória

$[X_1, X_2]$

Variável aleatória

Quantas?

$k = N^n = 6^2 = 36$

onde:

k = número de amostras possíveis

N = tamanho da população

n = tamanho da amostra

Quais?

Amostra aleatória

$[X_1, X_2]$

↓

Variável aleatória

↓

distribuição de probabilidade

(1, 1)	(4, 1)
(1, 2)	(4, 2)
(1, 3)	(4, 3)
(1, 4)	(4, 4)
(1, 5)	(4, 5)
(1, 6)	(4, 6)
(2, 1)	(5, 1)
(2, 2)	(5, 2)
(2, 3)	(5, 3)
(2, 4)	(5, 4)
(2, 5)	(5, 5)
(2, 6)	(5, 6)
(3, 1)	(6, 1)
(3, 2)	(6, 2)
(3, 3)	(6, 3)
(3, 4)	(6, 4)
(3, 5)	(6, 5)
(3, 6)	(6, 6)

Como a amostra é uma variável aleatória qualquer função da amostra (soma, média, variância) também será uma variável aleatória.

Amostra: $[X_1, X_2, \ldots, X_n]$ ← Variável aleatória

$\downarrow \qquad \downarrow$

$\bar{X} \qquad S^2$ ← Variáveis aleatórias

$\bar{x}_1$	s_1^2
$\bar{x}_2$	s_2^2
...	...
$\bar{x}_k$	s_k^2

Se temos k amostras de mesmo tamanho n, temos k médias e variâncias

Como a amostra é uma variável aleatória qualquer função da amostra (soma, média, variância) também será uma variável aleatória.

Amostra: $[X_1, X_2, ..., X_n]$ ← **Variável aleatória**

$\bar{X}$ S^2 ← **Variáveis aleatórias**

Exemplo:

$k = 6^2 = 36$

Amostra	$[X_1, X_2]$	$\bar{X}$	S^2
1	(1, 1)	$\bar{x}_1 = 1$	$s_1^2 = 0$
2	(1, 2)	$\bar{x}_2 = 1{,}5$	$s_2^2 = 0{,}5$
3	(1, 3)	$\bar{x}_3 = 2$	$s_3^2 = 2$
...	...	...	...
36	(6, 6)	$\bar{x}_k = 6$	$s_k^2 = 0$

Estatística e distribuição amostral

Amostra: $[X_1, X_2, ..., X_n]$

$\bar{X}$ X_+, S^2 ← Estatísticas

Estatística é qualquer valor obtido em função da amostra.

⇨ Como as estatísticas são variáveis aleatórias, também terão alguma distribuição de probabilidade com média, variância, etc.

Distribuição amostral é a distribuição de probabilidade de uma estatística.

4.3 Distribuições amostrais e Teorema central do limite

Amostra: $[X_1, X_2, \ldots, X_n]$

$\bar{X}$ X_+ ← Estatísticas

Qual é a distribuição amostral da Média e da Soma?

$$\bar{X} = \frac{\sum X_i}{n}$$

$$X_+ = \sum X_i$$

Exemplo:

O mecânico de uma oficina de regulagem para carros com 4, 6 e 8 cilindros, cobra pelo serviço 40, 45 e 50 reais, respectivamente. Seja a variável X = valor cobrado pelo mecânico, com a seguinte distribuição de probabilidade:

$X = x$	40	45	50	Σ
$P(X = x)$	0,2	0,3	0,5	1

a) Determine a média e a variância da população.

b) Supondo a retirada de uma amostra de tamanho n=2, com reposição, determine quantas e quais são as possíveis amostras e qual a probabilidade associada a cada uma; construa a distribuição amostral da média; e calcule o valor esperado e a variância da média.

c) Supondo a retirada de uma amostra de tamanho n=3, com reposição, determine quantas e quais são as possíveis amostras e qual a probabilidade associada a cada uma; construa a distribuição amostral da média; e calcule o valor esperado e a variância da média.

Distribuição de probabilidade da população

X = x	40	45	50	Σ
P(X = x)	0,2	0,3	0,5	1

$$E(X) = \mu = \sum_{x \in S_X} x\, p(x)$$
$$= 40 \times 0{,}2 + 45 \times 0{,}3 + 50 \times 0{,}5 = 46{,}5$$

$$V(X) = \sigma^2 = E(X^2) - \mu^2$$
$$= \left(40^2 \times 0{,}2 + 45^2 \times 0{,}3 + 50^2 \times 0{,}5\right) - 46{,}5^2 = 15{,}25$$

$X = x$	40	45	50	Σ
$P(X = x)$	0,2	0,3	0,5	1

$n = 2 \rightarrow [X_1, X_2] \rightarrow k = N^n = 3^2 = 9$ ← quantas

onde:

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

Variável aleatória

Estatística

quais

Amostra	$[X_1, X_2]$	$P[X_1, X_2]$	$\bar{X}$
1	(40, 40)	0,04	40
2	(40, 45)	0,06	42,5
3	(40, 50)	0,10	45
4	(45, 40)	0,06	42,5
5	(45, 45)	0,09	45
6	(45, 50)	0,15	47,5
7	(50, 40)	0,10	45
8	(50, 45)	0,15	47,5
9	(50, 50)	0,25	50

Distribuição amostral da média das amostras de tamanho 2

$\bar{X}=\bar{x}$	40	42,5	45	47,5	50	Σ
$P(\bar{X}=\bar{x})$	0,04	0,12	0,29	0,3	0,25	1

$$E(\bar{X})=\mu_{\bar{X}}=\sum_{\bar{x}\in S_{\bar{X}}}\bar{x}\,p(\bar{x})$$

$$=40\times 0,04+42,5\times 0,12+45\times 0,29+47,5\times 0,3+50\times 0,25$$

$$=46,5$$

$\boxed{= \mu \ \text{(média da população)}}$

$$V(\bar{X})=\sigma^2_{\bar{X}}=E(\bar{X}^2)-\mu^2_{\bar{X}}$$

$$=\left(40^2\times 0,04+42,5^2\times 0,12+\ldots+50^2\times 0,25\right)-46,5^2$$

$$=7,625$$

$\boxed{= \text{metade de } \sigma^2 \ \text{(variância da população)}}$

Quantas?

$n = 3 \rightarrow [X_1, X_2, X_3] \rightarrow k = N^n = 3^3 = 27$

Quais?

Amostra	$[X_1, X_2, X_3]$	$P[X_1, X_2, X_3]$	$\overline{X}$	Amostra	$[X_1, X_2, X_3]$	$P[X_1, X_2, X_3]$	$\overline{X}$
1	(40, 40, 40)	0,008	40	15	(45, 45, 50)	0,045	46,7
2	(40, 40, 45)	0,012	41,7	16	(45, 50, 40)	0,030	45
3	(40, 40, 50)	0,020	43,3	17	(45, 50, 45)	0,045	46,7
4	(40, 45, 40)	0,012	41,7	18	(45, 50, 50)	0,075	48,3
5	(40, 45, 45)	0,018	46,7	19	(50, 40, 40)	0,020	43,3
6	(40, 45, 50)	0,030	45	20	(50, 40, 45)	0,030	45
7	(40, 50, 40)	0,020	43,3	21	(50, 40, 50)	0,050	46,7
8	(40, 50, 45)	0,030	45	22	(50, 45, 40)	0,030	45
9	(40, 50, 50)	0,050	48,3	23	(50, 45, 45)	0,045	46,7
10	(45, 40, 40)	0,012	41,7	24	(50, 45, 50)	0,075	48,3
11	(45, 40, 45)	0,018	43,3	25	(50, 50, 40)	0,020	46,7
12	(45, 40, 50)	0,030	45	26	(50, 50, 45)	0,075	48,3
13	(45, 45, 40)	0,018	43,3	27	(50, 50, 50)	0,125	50
14	(45, 45, 45)	0,027	45				

Distribuição amostral da média das amostras de tamanho 3

$\bar{X}=\bar{x}$	40	41,7	43,3	45	46,7	48,3	50	Σ
$P(\bar{X}=\bar{x})$	0,008	0,036	0,114	0,207	0,285	0,225	0,125	1

$$E(\bar{X}) = \mu_{\bar{X}} = \sum_{\bar{x}\in S_{\bar{X}}} \bar{x}\, p(\bar{x})$$

$$= 40\times 0,008 + 41,7\times 0,036 + \ldots + 50\times 0,125$$

$$= 46,5$$

= μ (média da população)

$$V(\bar{X}) = \sigma^2_{\bar{X}} = E(\bar{X}^2) - \mu^2_{\bar{X}}$$

$$= (40\times 0,008 + 41,7\times 0,036 + \ldots + 50\times 0,125) - 46,5^2$$

$$= 5,083$$

= um terço de σ^2 (variância da população)

População

$$E(X) = \mu = 40 \times 0,2 + 45 \times 0,3 + 50 \times 0,5 = 46,5$$

$$V(X) = \sigma^2 = \left(40^2 \times 0,2 + 45^2 \times 0,3 + 50^2 \times 0,5\right) - 46,5^2 = 15,25$$

Amostras de tamanho n = 2

$$E(\bar{X}) = \mu_{\bar{x}} = 46,5 \longleftarrow \boxed{= \mu}$$

$$V(\bar{X}) = \sigma^2_{\bar{X}} = 7,625 \longleftarrow \boxed{= \sigma^2/2}$$

Amostras de tamanho n = 3

$$E(\bar{X}) = \mu_{\bar{x}} = 46,5 \longleftarrow \boxed{= \mu}$$

$$V(\bar{X}) = \sigma^2_{\bar{X}} = 5,083 \longleftarrow \boxed{= \sigma^2/3}$$

Teoremas

$$E(\bar{X}) = \mu$$

$$V(\bar{X}) = \frac{\sigma^2}{n}$$

$$\sigma_{\bar{X}} = \sqrt{V(\bar{X})} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$$

Resultados importantes

Relacionando as medidas da distribuição amostral de $\bar{X}$ com as medidas da distribuição populacional, verificamos que:

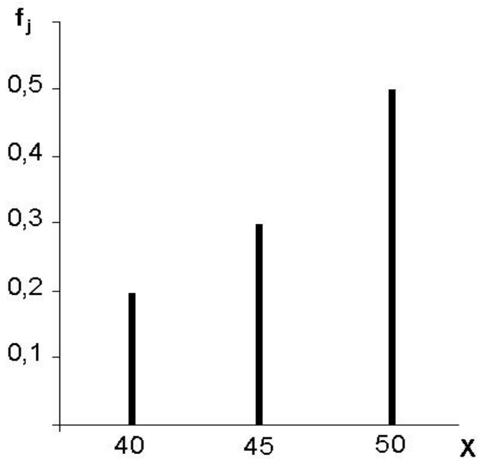
- A média da **soma** de todas as k amostras aleatórias possíveis, de um mesmo tamanho n, extraídas de uma população, é igual à média da população multiplicada pelo tamanho da amostra.

$$E(X_+) = n\mu$$

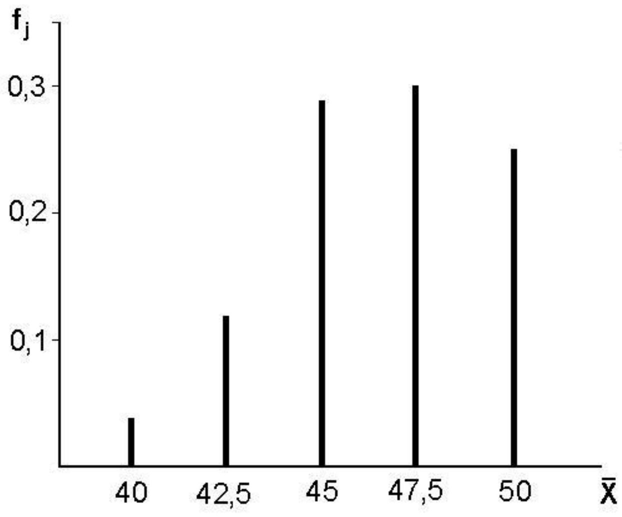
- A variância da **soma** de todas as k amostras aleatórias possíveis, de um mesmo tamanho n, extraídas de uma população, é igual à variância da população multiplicada pelo tamanho da amostra.

$$V(X_+) = n\sigma^2$$

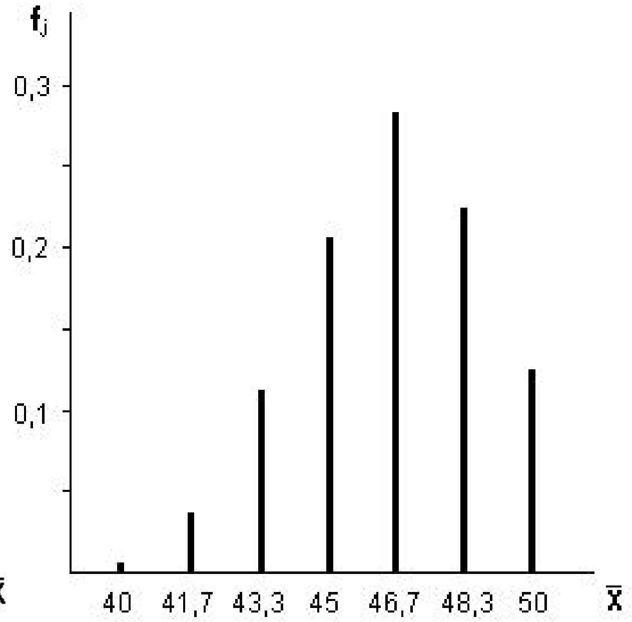
Formato da distribuição da população



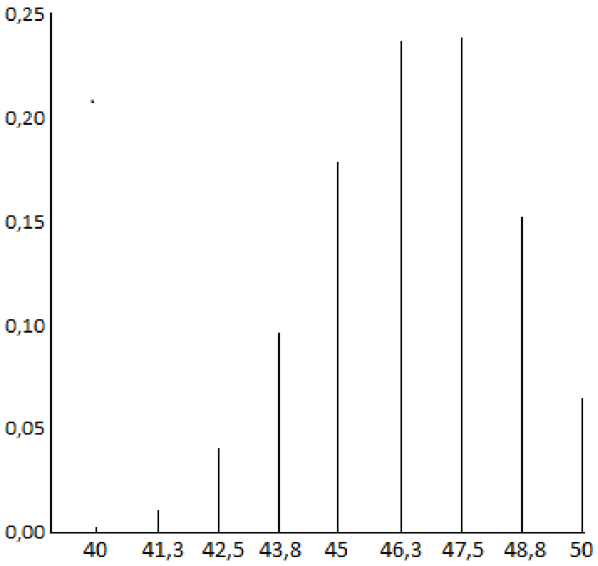
Formato da distribuição da média



Distribuição da média das amostras de tamanho 2



Distribuição da média das amostras de tamanho 3



Distribuição da média das amostras de tamanho 4

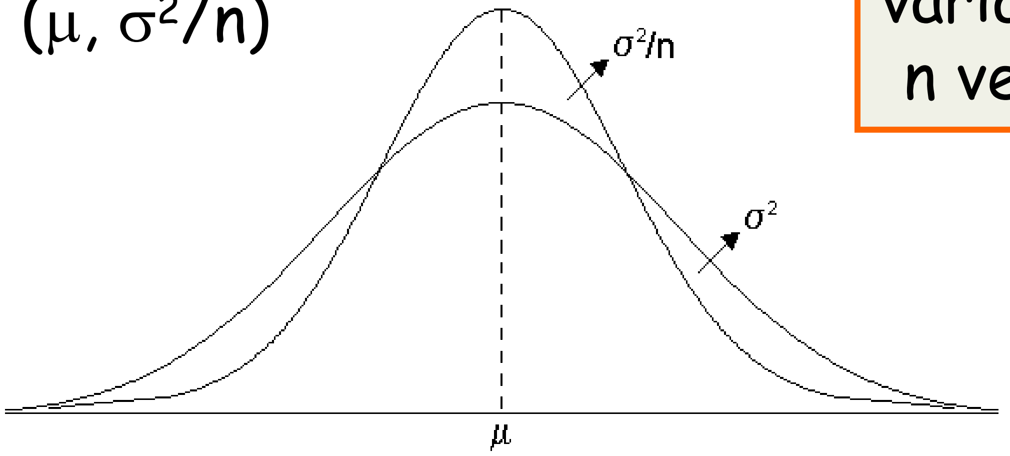
Profª. Clause Piana

Qual é a distribuição da média?

1. Se a população (X) de onde foi extraída a amostra aleatória tiver distribuição normal, então a distribuição amostral da média será normal.

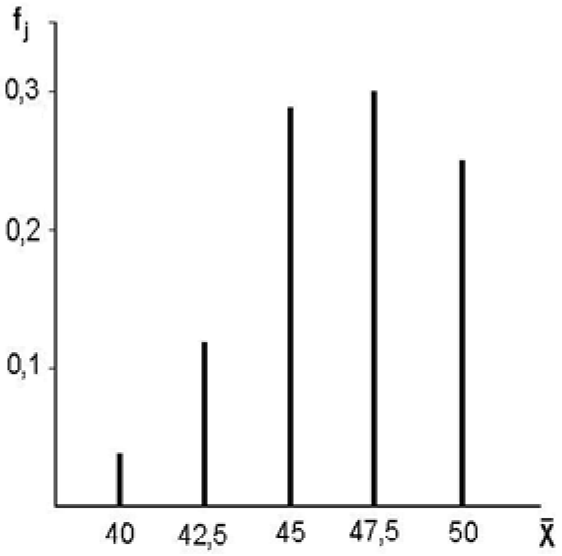
se $X \sim N(\mu, \sigma^2)$,

então, $\bar{X} \sim N(\mu, \sigma^2/n)$

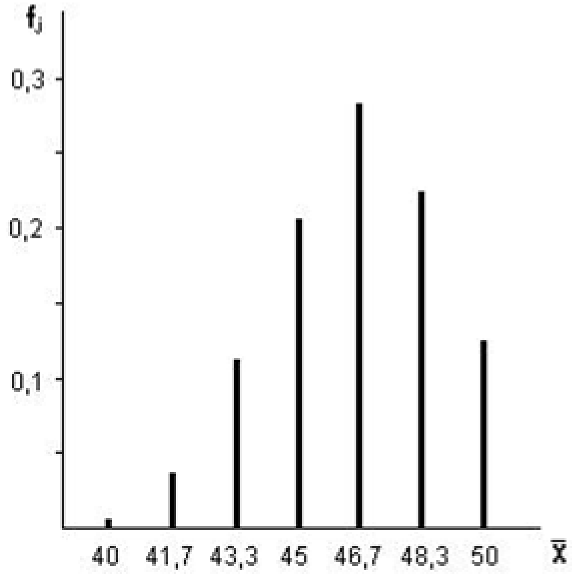


As médias são iguais, mas a variância de $\bar{X}$ é n vezes menor.

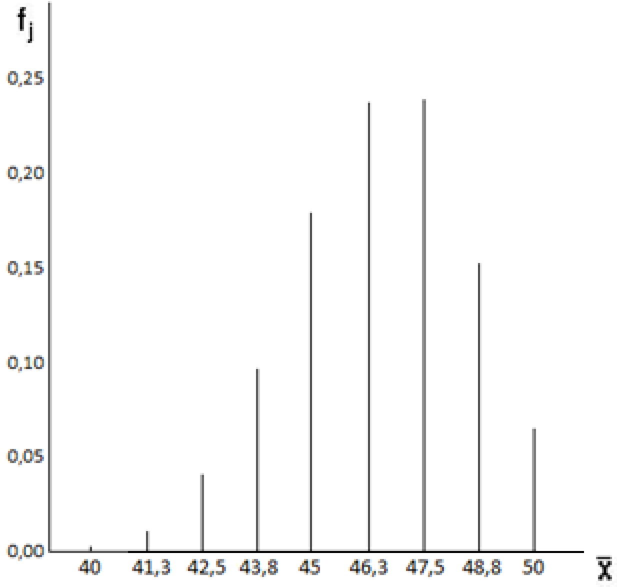
2. Se a população (X) de onde foi extraída a amostra aleatória não tiver distribuição normal, então a distribuição amostral da média se aproximará da normal à medida que o tamanho da amostra (n) cresce.



Distribuição da média das amostras de tamanho 2

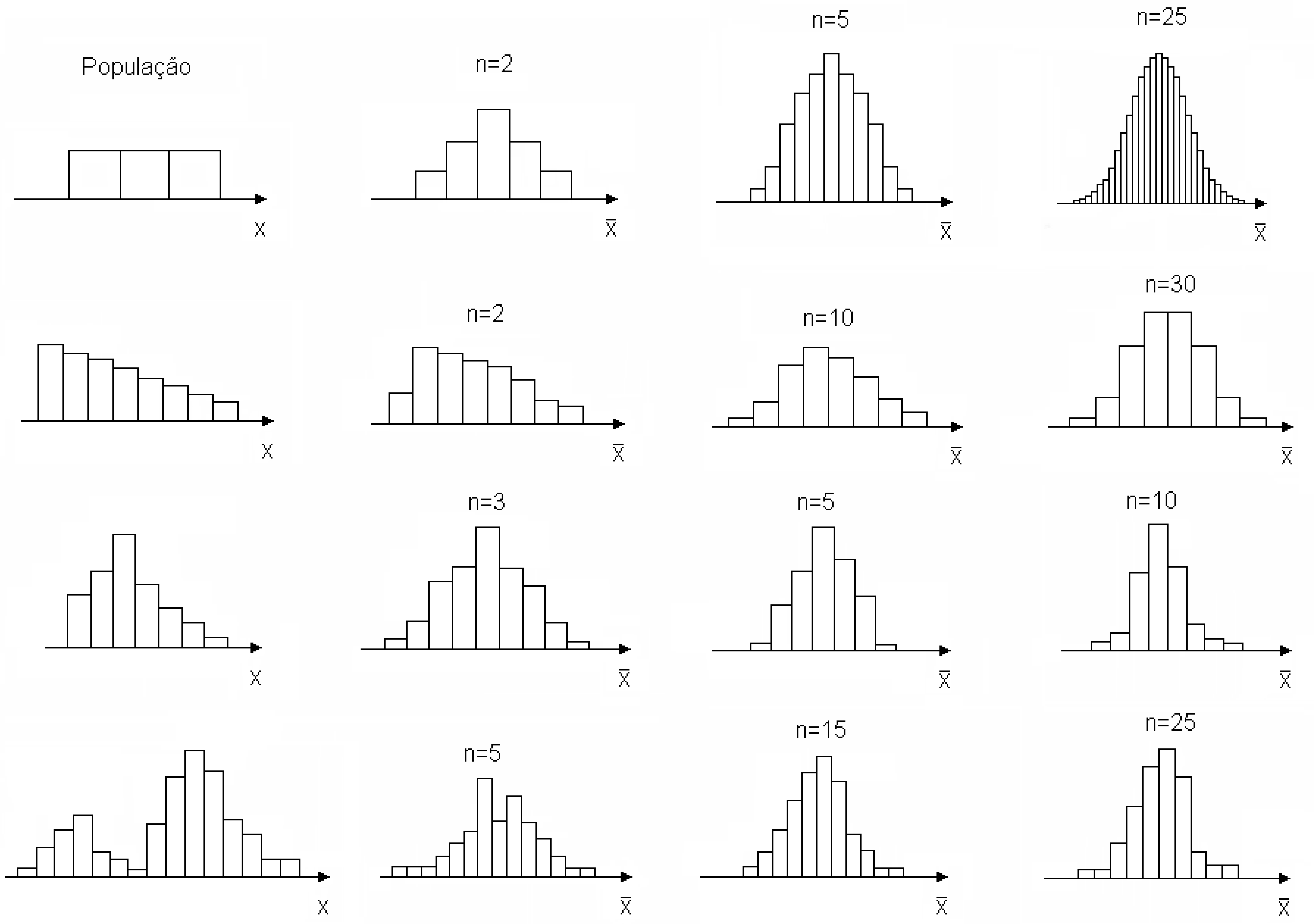


Distribuição da média das amostras de tamanho 3



Distribuição da média das amostras de tamanho 4

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



Histogramas correspondentes às distribuições de $\bar{X}$ para amostras extraídas de algumas populações.

Qual é a distribuição da média?

1. Se a população (X) de onde foi extraída a amostra aleatória tiver distribuição normal, então a distribuição amostral da média será normal.

$$\text{se } X \sim N(\mu, \sigma^2),$$

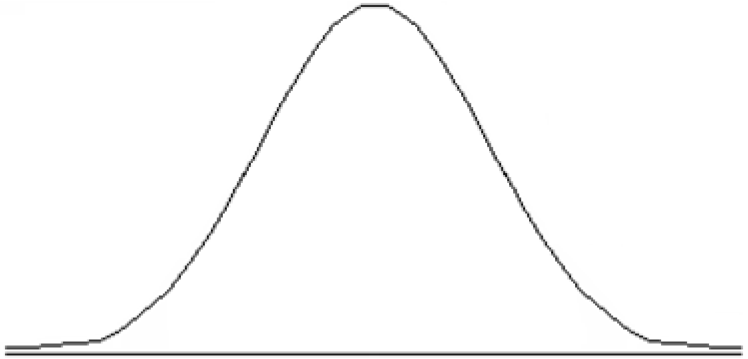
$$\text{então, } \bar{X} \sim N(\mu, \sigma^2/n)$$

2. Se a população (X) de onde foi extraída a amostra aleatória não tiver distribuição normal, então a distribuição amostral da média se aproximará da normal à medida que o tamanho da amostra (n) cresce.

Quando n cresce, a distribuição de $\bar{X}$ tende para $N(\mu, \sigma^2/n)$

Qual é a distribuição amostral da média?

$X \sim N(\mu, \sigma^2)$



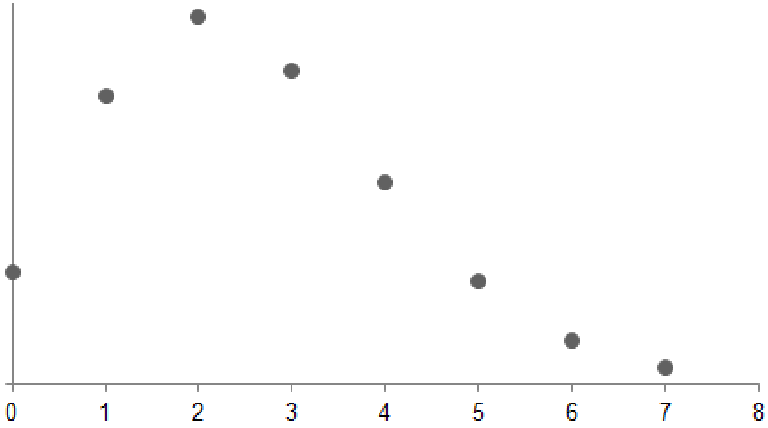
μ

Amostra aleatória

$[X_1, X_2, ..., X_n]$

$\bar{X}$ $\longrightarrow$ $N(\mu, \sigma^2/n)$

$X \sim Bin(n, \pi)$



Amostra aleatória

$[X_1, X_2, ..., X_n]$

$\bar{X}$ $\longrightarrow$ $N(\mu, \sigma^2/n)$

Quando n é grande

Qual é a distribuição amostral da soma?

Amostra: $[X_1, X_2, ..., X_n]$

$$X_+ = \sum X_i \longleftarrow \boxed{\text{Estatística}}$$

Exemplo:

O mecânico de uma oficina de regulagem para carros com 4, 6 e 8 cilindros, cobra pelo serviço 40, 45 e 50 reais, respectivamente. Seja a variável X = valor cobrado pelo mecânico, com a seguinte distribuição de probabilidade:

$X = x$	40	45	50	Σ
$P(X = x)$	0,2	0,3	0,5	1

a) Determine a média e a variância da população.

b) Supondo a retirada de uma amostra de tamanho n=2, com reposição, determine quantas e quais são as possíveis amostras e qual a probabilidade associada a cada uma; construa a distribuição amostral da média; e calcule o valor esperado e a variância da média.

c) Supondo a retirada de uma amostra de tamanho n=3, com reposição, determine quantas e quais são as possíveis amostras e qual a probabilidade associada a cada uma; construa a distribuição amostral da média; e calcule o valor esperado e a variância da média.

d) Supondo a retirada de uma amostra de tamanhos 2 e 3, com reposição, construa a distribuição amostral da soma; e calcule o valor esperado e a variância da soma.

d) Supondo a retirada de uma amostra de tamanhos 2 e 3, com reposição, construa a distribuição amostral da soma; e calcule o valor esperado e a variância da soma.

População

$E(X) = \mu = 46,5$

$V(X) = \sigma^2 = 15,25$

n=2

$[X_1, X_2]$	$P[X_1, X_2]$	$X_+ = \sum X_i$
(40, 40)	0,04	80
(40, 45)	0,06	85
(40, 50)	0,10	90
(45, 40)	0,06	85
(45, 45)	0,09	90
(45, 50)	0,15	95
(50, 40)	0,10	90
(50, 45)	0,15	95
(50, 50)	0,25	100

Distribuição amostral da soma em amostras de tamanho 2

$X_+ = x_+$	80	85	90	95	100	$\sum$
$P(X_+ = x_+)$	0,04	0,12	0,29	0,3	0,25	1

$$E(X_+) = \mu_{X_+} = \sum_{\bar{x} \in S_{\bar{X}}} x_+ \, p(x_+) = 80 \times 0,04 + 85 \times 0,12 + 90 \times 0,29 + 95 \times 0,3 + 100 \times 0,25 = 93 = 2\mu$$

$$V(X_+) = \sigma^2_{X_+} = E(X_+^2) - \mu^2_{X_+} = (80^2 \times 0,04 + 85^2 \times 0,12 + \ldots + 100^2 \times 0,25) - 93^2 = 30,5 = 2\sigma^2$$

Distribuição de probabilidade da população

$X = x$	40	45	50	Σ
$P(X = x)$	0,2	0,3	0,5	1

Distribuição amostral da soma em amostras de tamanho 2

$X_+ = x_+$	80	85	90	95	100	Σ
$P(X_+ = x_+)$	0,04	0,12	0,29	0,3	0,25	1

Distribuição amostral da soma em amostras de tamanho 3

$X_+ = x_+$	120	125	130	135	140	145	150	Σ
$P(X_+ = x_+)$	0,008	0,036	0,114	0,207	0,285	0,225	0,125	1

$E(X_+) = 139{,}5$ $V(X_+) = 45{,}75$

População

$$E(X) = \mu = 40 \times 0,2 + 45 \times 0,3 + 50 \times 0,5 = 46,5$$

$$V(X) = \sigma^2 = (40^2 \times 0,2 + 45^2 \times 0,3 + 50^2 \times 0,5) - 46,5^2 = 15,25$$

Amostras de tamanho n = 2

$$E(X_+) = 93 \leftarrow \boxed{= 2\mu}$$

$$V(X_+) = 30,5 \leftarrow \boxed{= 2\sigma^2}$$

Amostras de tamanho n = 3

$$E(X_+) = 139,5 \leftarrow \boxed{= 3\mu}$$

$$V(X_+) = 45,75 \leftarrow \boxed{= 3\sigma^2}$$

Teoremas

$$E(X_+) = n\mu$$

$$V(X_+) = n\sigma^2$$

Resultados importantes

Relacionando as medidas da distribuição amostral de X_+ com as medidas da distribuição populacional, verificamos que:

- A média da soma de todas as k amostras aleatórias possíveis, de um mesmo tamanho n, extraídas de uma população, é igual à média da população multiplicada pelo tamanho da amostra.

$$E(X_+) = n\mu$$

- A variância da soma de todas as k amostras aleatórias possíveis, de um mesmo tamanho n, extraídas de uma população, é igual à variância da população multiplicada pelo tamanho da amostra.

$$V(X_+) = n\sigma^2$$

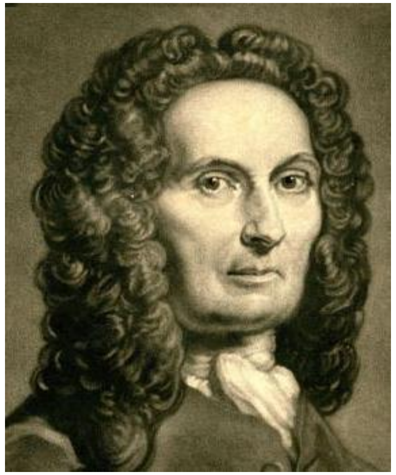
Estes resultados verificados no exemplo também podem ser derivados do teorema fundamental da estatística paramétrica:
o **Teorema Central do Limite**.

Em 1733, De Moivre publicou um panfleto em latim com a obtenção da aproximação da distribuição binomial por uma função que mais tarde se tornaria a função densidade de probabilidade da distribuição normal.

Em 1809 e 1816, Gauss obteve técnicas baseadas na distribuição normal que se tornariam a metodologia padrão durante o século XIX.

A derivação da distribuição normal feita por Gauss considerava uma variável aleatória resultante da **soma** de um grande número de erros independentes.

Tanto o resultado de De Moivre quanto a derivação de Gauss são casos particulares do Teorema Central do Limite.



Abraham De Moivre
(1667 - 1754)

Carl Gauss
(1777 –1855)

Teorema central do limite (TCL): se $(X_1, X_2, ..., X_n)$ é uma amostra aleatória de X, para **n grande** a distribuição da **soma** X_+ da amostra se aproxima da distribuição normal com média **$n\mu$** e variância **$n\sigma^2$**.
Se o teorema é verdadeiro, temos:

$$X_+ \sim N(n\mu, n\sigma^2) \longrightarrow \frac{X_+ - n\mu}{\sqrt{n\sigma^2}} = Z \sim N(0,1) \quad \textbf{(soma padronizada)}$$

Pode-se verificar que o numerador e o denominador de Z podem ser divididos por n sem que a variável seja alterada

$$\textbf{soma padronizada} \quad \boxed{\frac{X_+ - n\mu}{\sqrt{n\sigma^2}}} = \frac{\dfrac{X_+ - n\mu}{n}}{\dfrac{\sqrt{n\sigma^2}}{n}} = \frac{\dfrac{X_+}{n} - \dfrac{n\mu}{n}}{\sigma n^{1/2}\dfrac{1}{n}} = \frac{\bar{X} - \mu}{\sigma n^{1/2} n^{-1}} = \frac{\bar{X} - \mu}{\sigma n^{-1/2}} = \boxed{\frac{\bar{X} - \mu}{\dfrac{\sigma}{\sqrt{n}}}} \quad \textbf{média padronizada}$$

Assim, pode-se dizer que, à medida que n cresce, a distribuição da **média** $\bar{X}$ da amostra também se aproxima da distribuição normal.

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n}) \longrightarrow \frac{\bar{X} - \mu}{\dfrac{\sigma}{\sqrt{n}}} = Z \sim N(0,1) \quad \textbf{(média padronizada)}$$

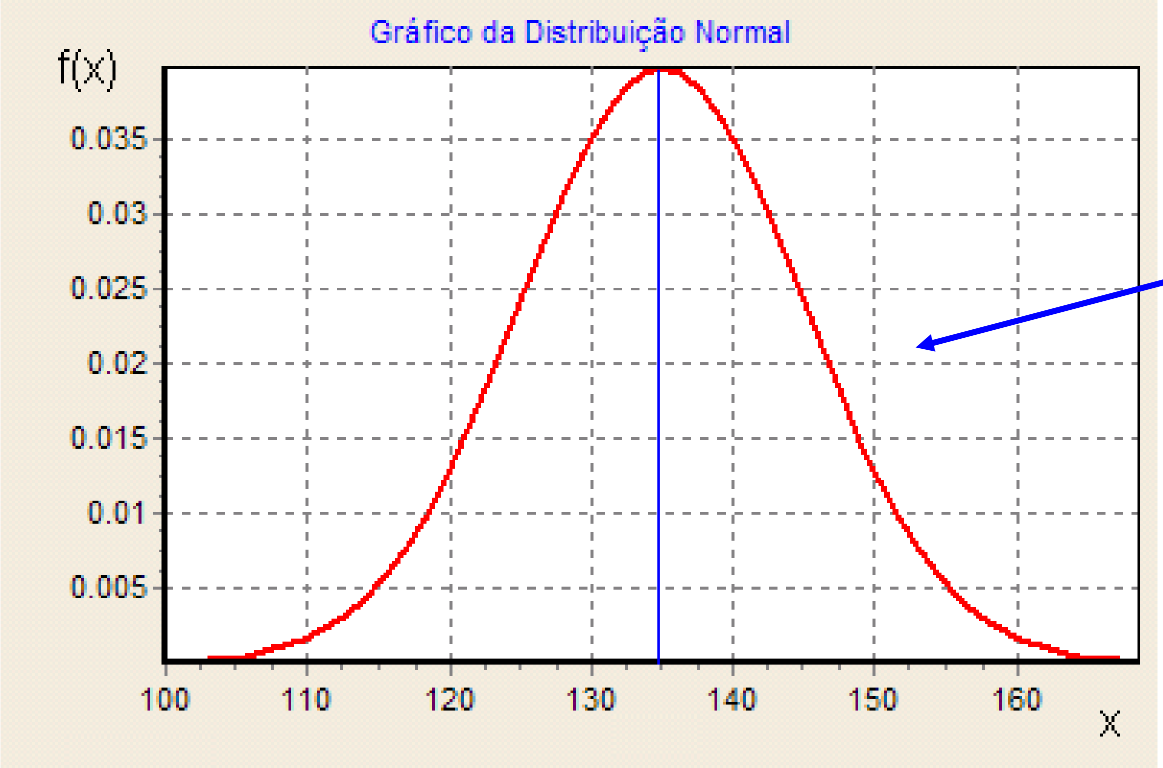
Teorema central do limite (TCL)

- A importância da distribuição normal na estatística se deve em grande parte a este teorema
- Distribuições importantes como a binomial e a Poisson (soma de variáveis Bernoulli) se aproximam naturalmente da normal
- Se a distribuição binomial é simétrica, a aproximação (ou convergência) para a normal é mais rápida

Uma regra às vezes utilizada é que a aproximação da binomial para a normal é boa se n é tal que $n\pi \geq 5$ e $n(1-\pi) \geq 5$.

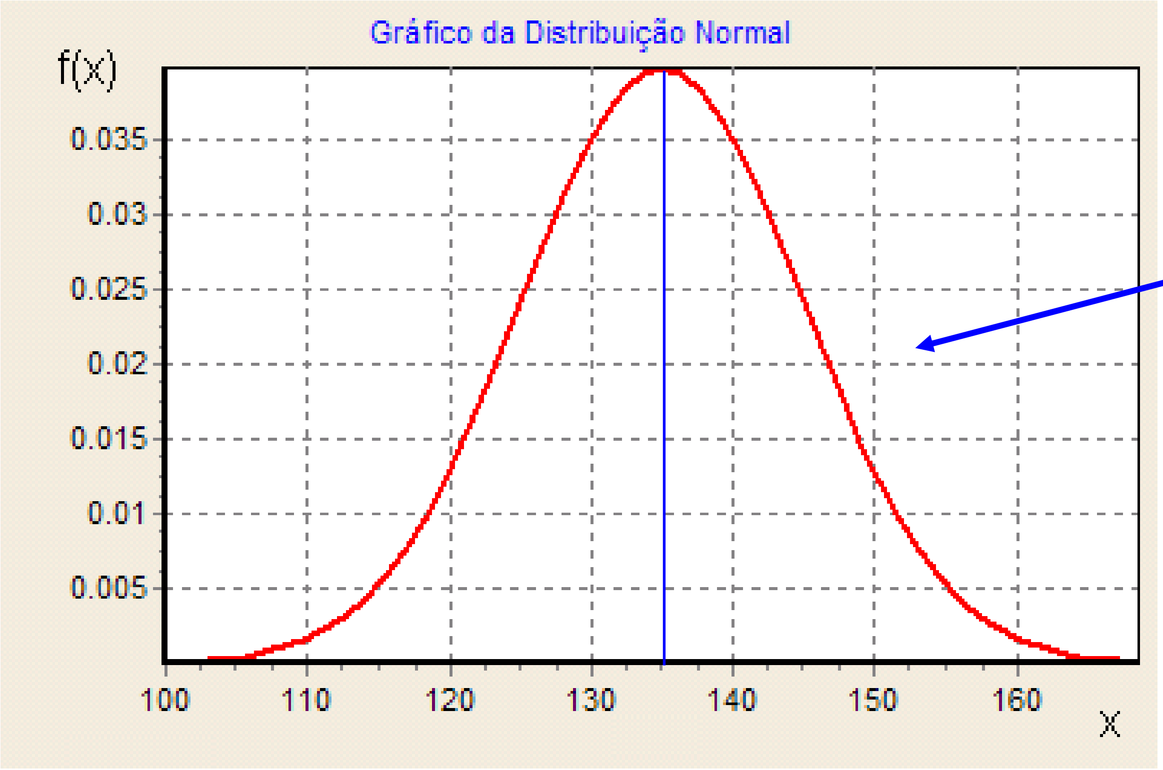
Essencialmente, esse é o resultado obtido por De Moivre em 1733 para o caso particular π=0,5 e generalizado mais tarde por Laplace para qualquer valor de π. Por esse motivo, é denominado teorema de De Moivre-Laplace.

Verificação das propriedades da média de uma amostra aleatória



população normal com μ=135 e σ=10

Processo de obtenção de uma amostra



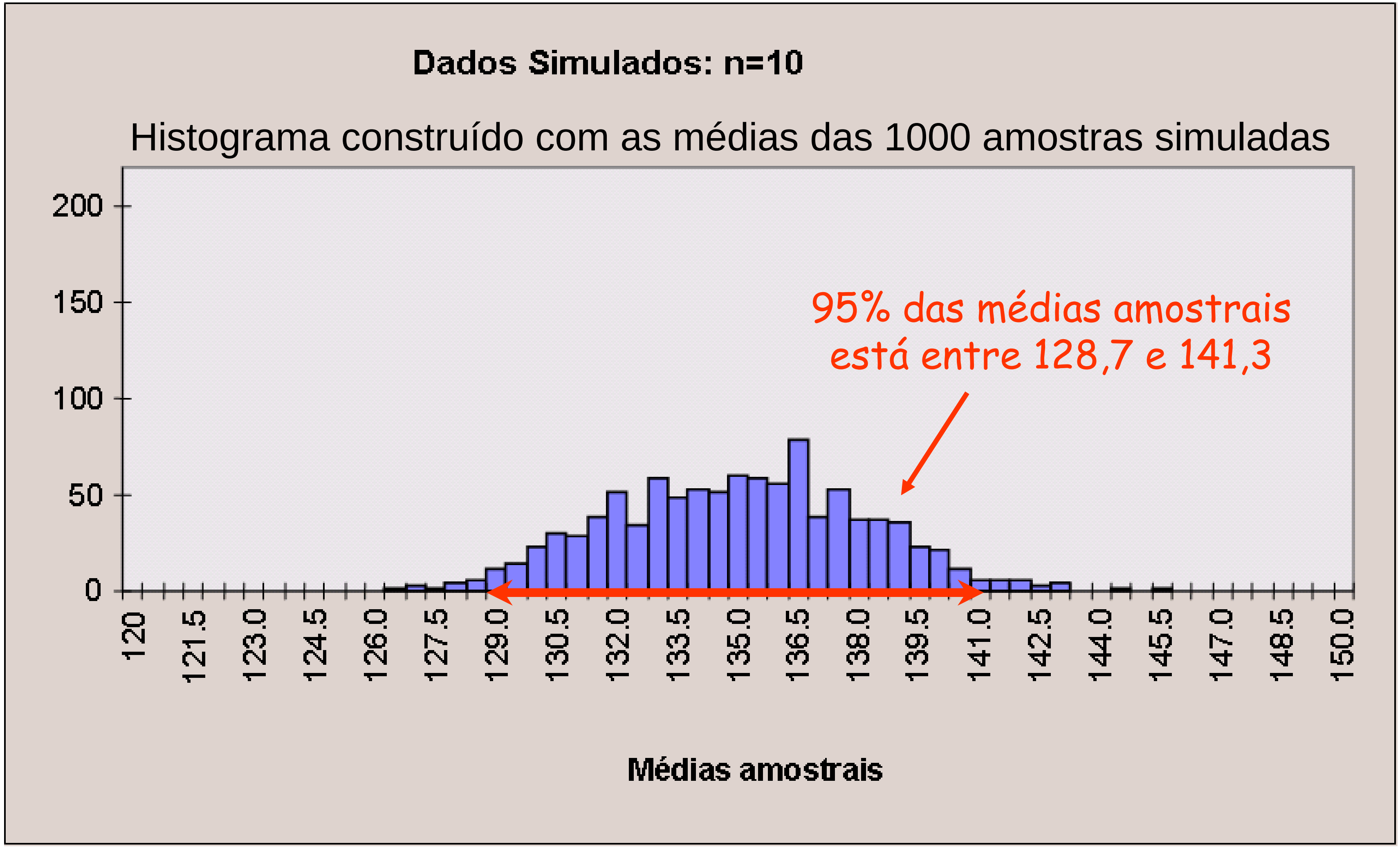
população normal com μ=135 e σ=10

X_1 X_2 X_3 X_4 … X_n

$\bar{X}$

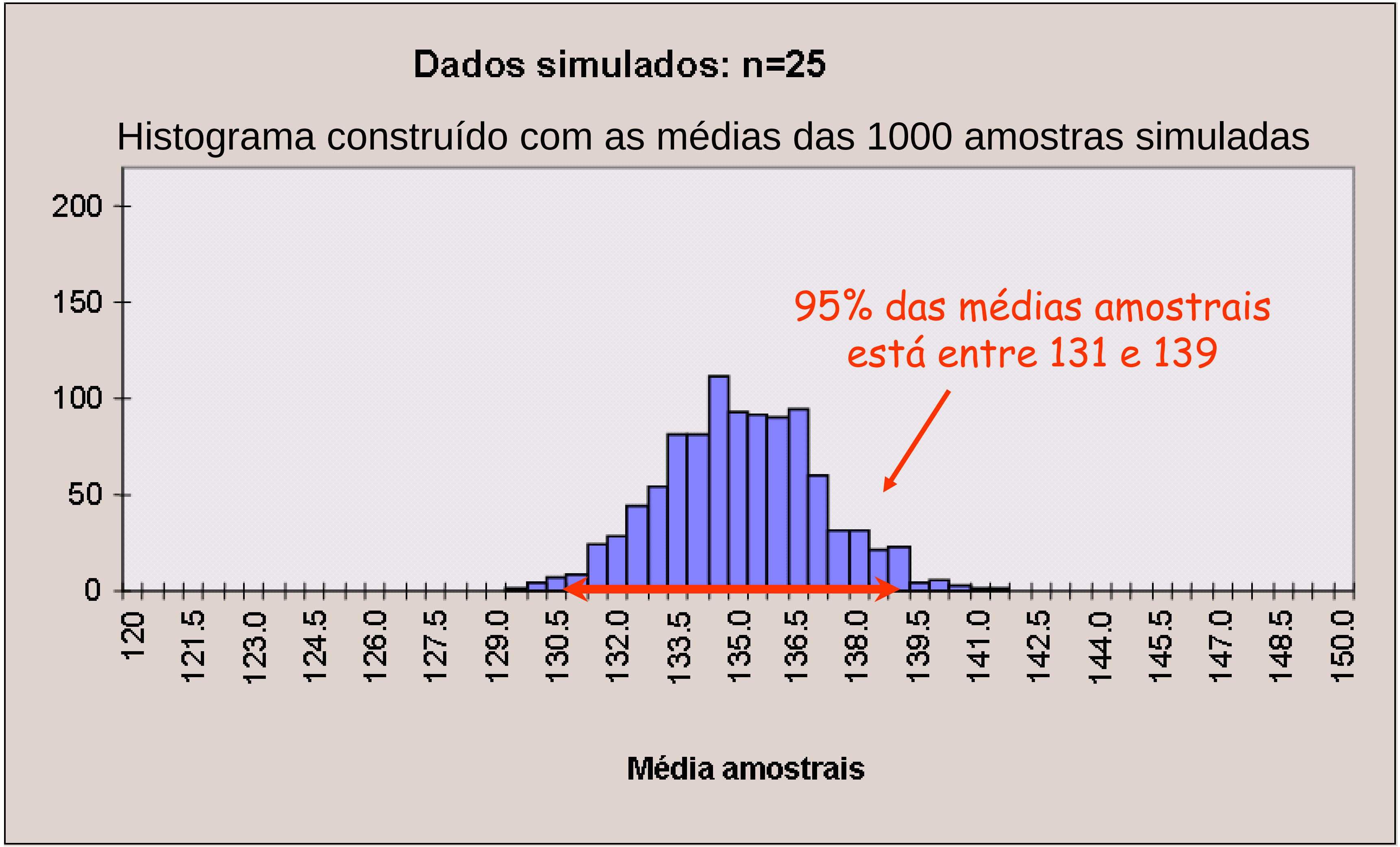
1000 amostras de mesmo tamanho n são coletadas (ou simuladas) desta população e a média de cada amostra é obtida

Distribuição amostral da média (n=10)



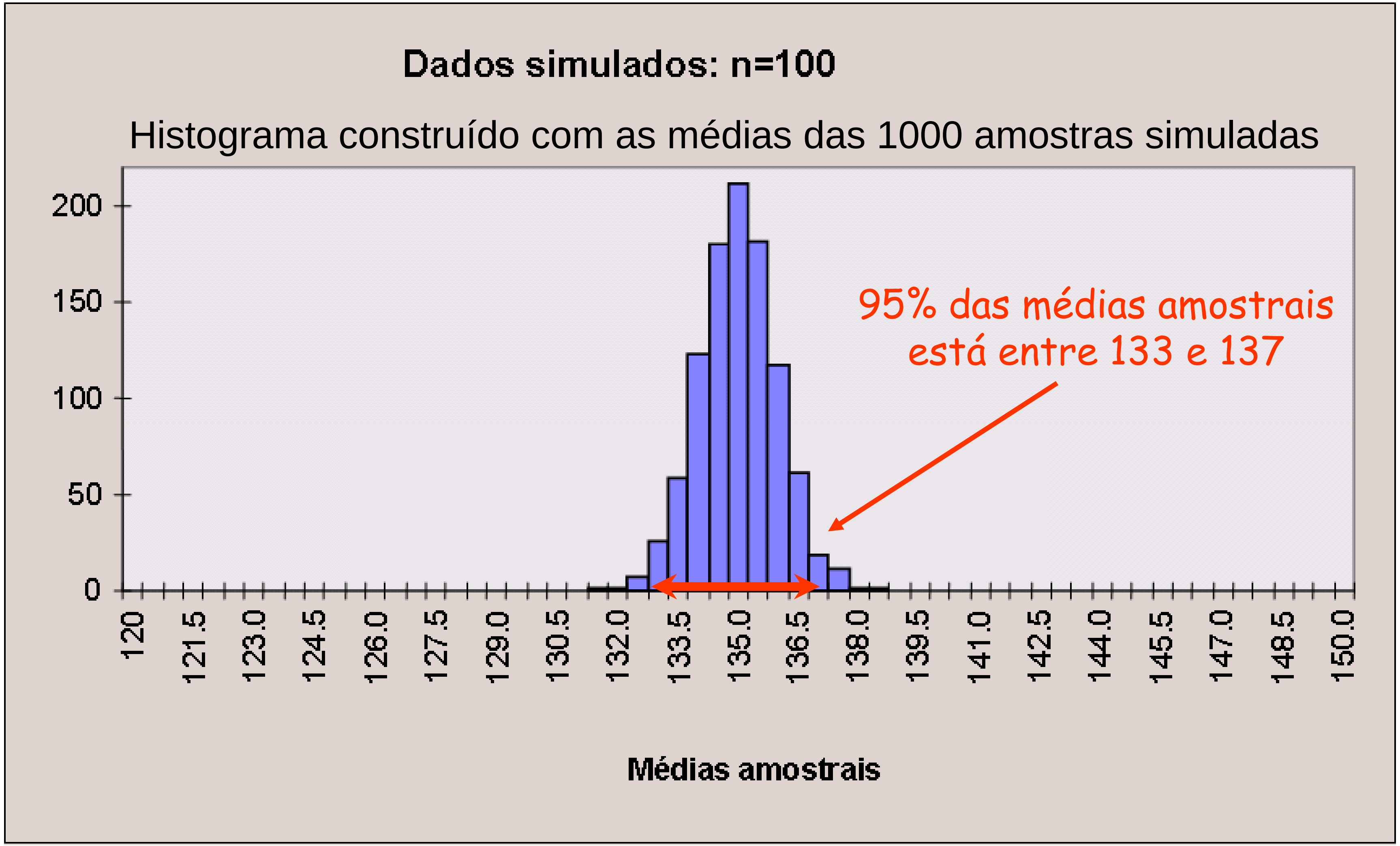
$$\mu_{\bar{X}} = \mu = 135 \qquad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{10}} = 3{,}16$$

Distribuição amostral da média (n=25)



$$\mu_{\bar{X}} = \mu = 135 \qquad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{25}} = 2$$

Distribuição amostral da média (n=100)



$$\mu_{\overline{X}} = \mu = 135 \qquad \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{10}} = 1$$

Resumo

Medidas das estatísticas Média e Soma

População

$$X \begin{cases} E(X) = \mu \\ V(X) = \sigma^2 \end{cases}$$

Amostra aleatória

$[X_1, X_2, ..., X_n]$

$$X_i \begin{cases} E(X_i) = \mu \\ V(X_i) = \sigma^2 \end{cases}$$

Média e variância da média e da soma de k amostras de mesmo tamanho n

$$\bar{X} = \frac{\sum X_i}{n} \begin{cases} E(\bar{X}) = \mu \\ V(\bar{X}) = \dfrac{\sigma^2}{n} \end{cases}$$

$$X_+ = \sum X_i \begin{cases} E(X_+) = n\mu \\ V(X_+) = n\sigma^2 \end{cases}$$

Teoremas importantes

$$X_i \quad \begin{cases} E(X_i) = \mu \\ V(X_i) = \sigma^2 \end{cases}$$

$$\overline{X} = \frac{\sum X_i}{n} \quad \begin{cases} E(\overline{X}) = \mu \\ V(\overline{X}) = \dfrac{\sigma^2}{n} \end{cases}$$

$$X_+ = \sum X_i \quad \begin{cases} E(X_+) = n\mu \\ V(X_+) = n\sigma^2 \end{cases}$$

Distribuição amostral das estatísticas Média e Soma

Teorema Central do Limite

Se n é grande a média e a soma tem distribuição normal

Amostra aleatória

$[X_1, X_2, ..., X_n] \Rightarrow \begin{cases} X_+ \longrightarrow X_+ \sim N(n\mu, n\sigma^2) \\ \bar{X} \longrightarrow \bar{X} \sim N(\mu, \frac{\sigma^2}{n}) \end{cases}$

Conseqüência $\Rightarrow$ As variáveis aleatórias Média e Soma podem ser padronizadas

$$\begin{cases} \dfrac{\bar{X}-\mu}{\dfrac{\sigma}{\sqrt{n}}} = Z \sim N(0,1) \\ \dfrac{X_+ - n\mu}{\sqrt{n\sigma^2}} = Z \sim N(0,1) \end{cases}$$

Resultados importantes

População

$$X \begin{cases} E(X) = \mu \\ V(X) = \sigma^2 \end{cases}$$

Amostra aleatória

$$[X_1, X_2, \ldots, X_n]$$

$$X_i \begin{cases} E(X_i) = \mu \\ V(X_i) = \sigma^2 \end{cases}$$

Distribuição amostral da média

$$\bar{X} = \frac{\sum X_i}{n} \begin{cases} E(\bar{X}) = \mu \\ V(\bar{X}) = \dfrac{\sigma^2}{n} \end{cases}$$

Resultados importantes

População

$$X \begin{cases} E(X) = \mu \\ V(X) = \sigma^2 \end{cases}$$

Amostra aleatória

$[X_1, X_2, ..., X_n]$

$$X_i \begin{cases} E(X_i) = \mu \\ V(X_i) = \sigma^2 \end{cases}$$

Estatísticas

$$\begin{cases} X_+ = \sum X_i \\ S^2 = \dfrac{\sum (X_i - \bar{X})}{n-1} \\ \bar{X} = \dfrac{\sum X_i}{n} \end{cases}$$

Resultados importantes

Distribuições amostrais da média e da soma das k amostras de mesmo tamanho n

População

$$X \begin{cases} E(X) = \mu \\ V(X) = \sigma^2 \end{cases}$$

Amostra aleatória

$[X_1, X_2, \ldots, X_n]$

$$X_i \begin{cases} E(X_i) = \mu \\ V(X_i) = \sigma^2 \end{cases}$$

$$\bar{X} = \frac{\sum X_i}{n} \begin{cases} E(\bar{X}) = \mu \\ V(\bar{X}) = \dfrac{\sigma^2}{n} \end{cases}$$

$$X_+ = \sum X_i \begin{cases} E(X_+) = n\mu \\ V(X_+) = n\sigma^2 \end{cases}$$

Teoremas importantes

$$X_i \qquad \begin{cases} E(X_i) = \mu \\ V(X_i) = \sigma^2 \end{cases}$$

$$\bar{X} = \frac{\sum X_i}{n} \qquad \begin{cases} E(\bar{X}) = \mu \\ V(\bar{X}) = \dfrac{\sigma^2}{n} \end{cases}$$

$$X_+ = \sum X_i \qquad \begin{cases} E(X_+) = n\mu \\ V(X_+) = n\sigma^2 \end{cases}$$

Bibliografia

BUSSAB, W.O.; MORETTIN, P.A. Estatística Básica. São Paulo: Saraiva. 2006. 526p.

FERREIRA, D.F. Estatística Básica. Lavras: Editora UFLA, 2005, 664p.

MEMÓRIA, J. M. P. Breve história da estatística. Brasília, DF: Embrapa Informação Tecnológica, 2004. 111p.

MLODINOW, L. O andar do bêbado. Como o acaso determina nossas vidas. Rio de Janeiro: Editora Zahar, 2009, 264p.

SILVEIRA JÚNIOR, P., MACHADO, A.A., ZONTA, E.P., SILVA, J.B. da Curso de Estatística v.1, Pelotas: Universidade Federal de Pelotas, 1989. 135p.

Sistema Galileu de Educação Estatística. Disponível em: http://www.galileu.esalq.usp.br/topico.html