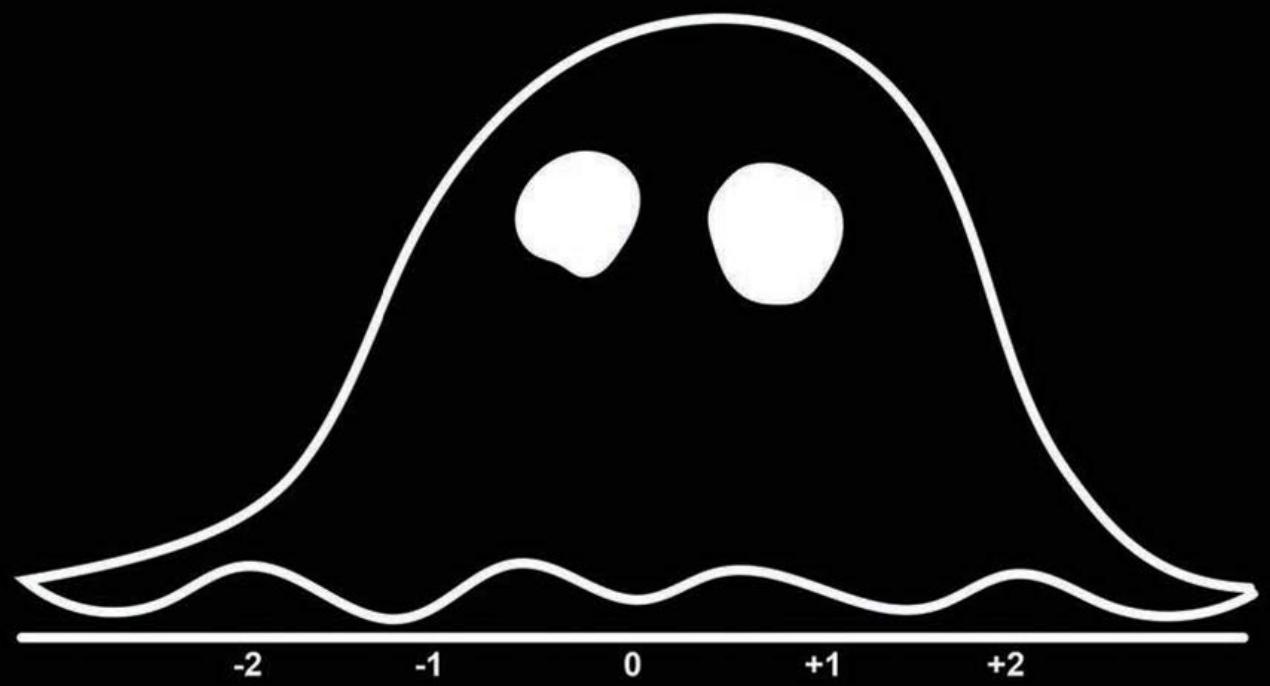


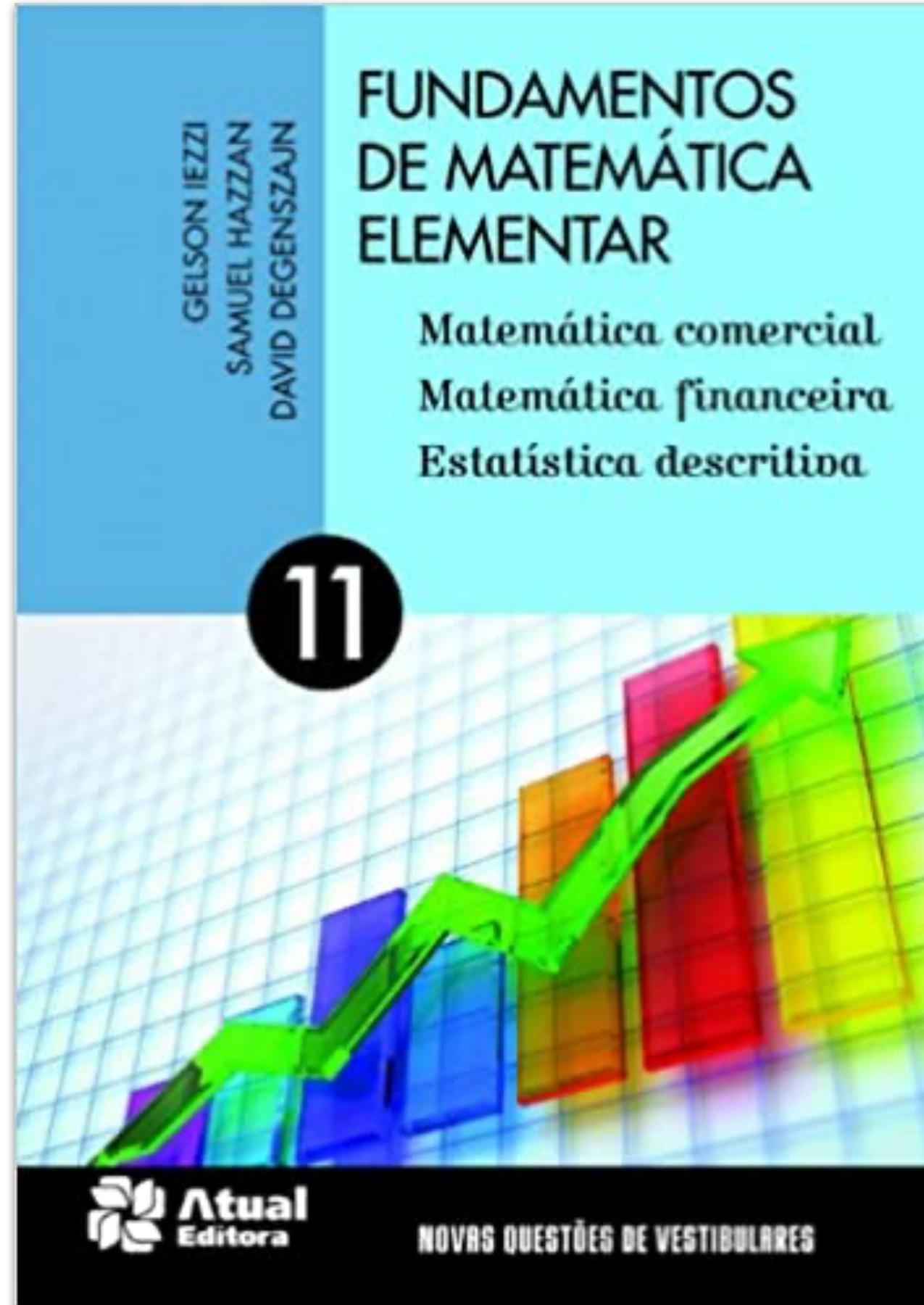
NORMAL DISTRIBUTION



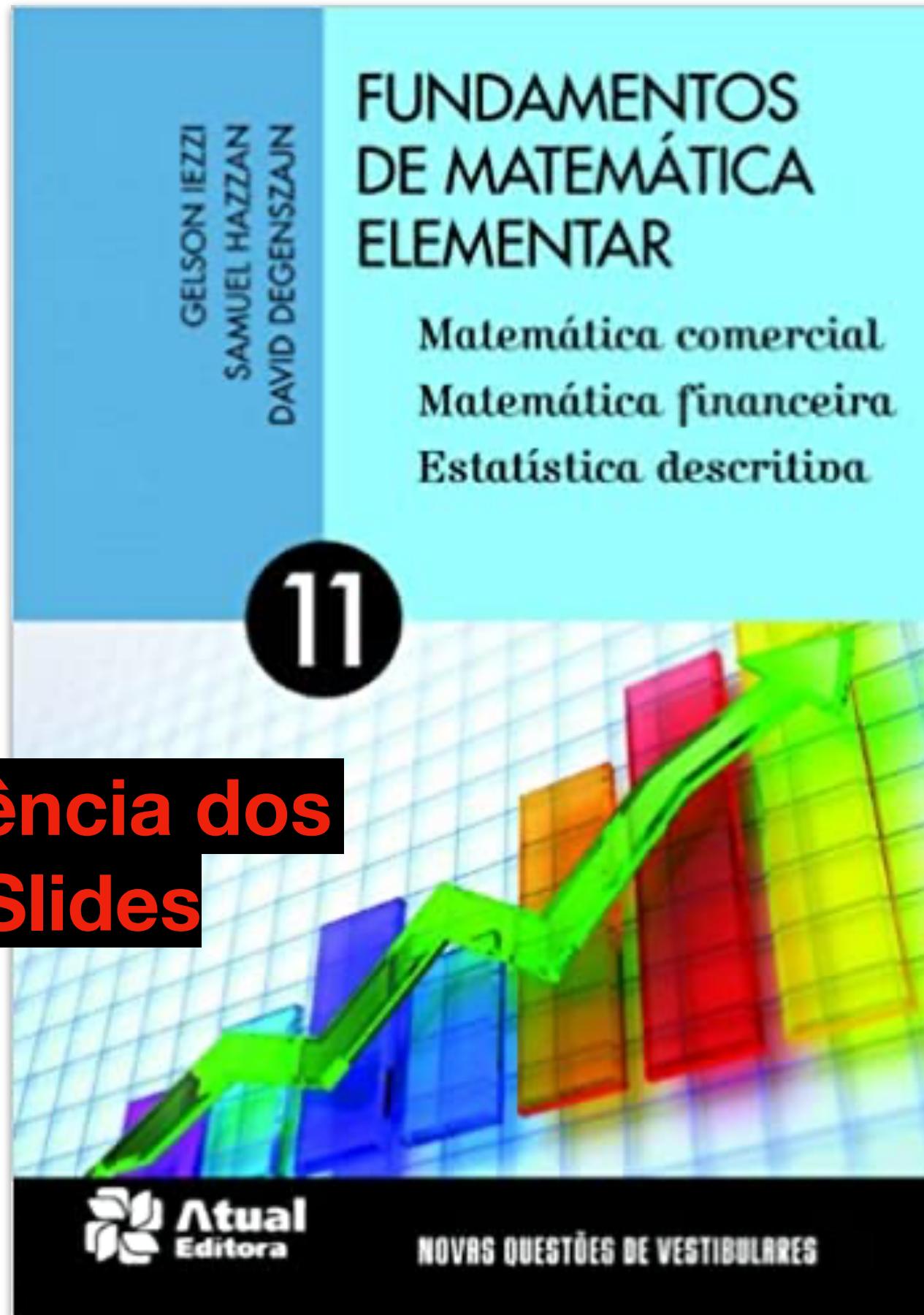
PARANORMAL DISTRIBUTION

Introdução à Estatística

Estatística Descritiva



Principal Referência dos Textos dos Slides



CAPÍTULO III Estatística Descritiva

I. Introdução

Imagine que, um mês antes de uma eleição presidencial, a federação das indústrias de determinado estado encomendou a um instituto especializado uma pesquisa cujo objetivo consistiu em detectar a intenção de voto do eleitor e levantar o perfil socioeconômico dos eleitores de cada um dos candidatos.

O que o instituto fez?

- Primeiramente, dimensionou uma amostra da população e fez a coleta de dados por meio de uma pesquisa de campo. A escolha da amostra é, em geral, complexa, pois deve-se levar em conta, entre outros fatores, o tempo e o custo da pesquisa, o número de eleitores de cada cidade do estado, a camada social à qual o entrevistado pertence, o local onde será feita a entrevista. É imprescindível que a amostra seja representativa, a fim de não haver comprometimento na análise dos resultados.
- Num segundo momento, organizou em tabelas os dados brutos coletados, construiu gráficos para apresentar os resultados obtidos e divulgou-os nos meios de comunicação. É preciso também associar ao conjunto de informações medidas de tendência central e medidas de variabilidade (ou dispersão dos dados em relação a valores centrais).
- Por fim, fez a análise confirmatória dos dados, isto é, verificou a margem de erro com que os resultados da amostra refletiram, de fato, a intenção de votos de toda a população de eleitores.

A ciência que se dedica a esse trabalho é a **Estatística**.

11 | Fundamentos de Matemática Elementar



Imagine que, um mês antes de uma eleição presidencial, a federação das indústrias de determinado estado encomendou a um instituto especializado uma pesquisa cujo objetivo consistiu em detectar a intenção de voto do eleitor e levantar o perfil socioeconômico dos eleitores de cada um dos candidatos.

O que o instituto fez?

Primeiramente, dimensionou uma amostra da população e fez a coleta de dados por meio de uma pesquisa de campo. A escolha da amostra é, em geral, complexa, pois deve-se levar em conta, entre outros fatores, o tempo e o custo da pesquisa, o número de eleitores de cada cidade do estado, a camada social à qual o entrevistado pertence, o local onde será feita a entrevista. É imprescindível que a **amostra seja representativa**, a fim de não haver comprometimento na análise dos resultados.

Num segundo momento, organizou em tabelas os dados brutos coletados, construiu gráficos para apresentar os resultados obtidos e divulgou-os nos meios de comunicação. É preciso também associar ao conjunto de informações medidas de tendência central e medidas de variabilidade (ou dispersão dos dados em relação a valores centrais).

Por fim, fez a análise confirmatória dos dados, isto é, verificou a margem de erro com que os resultados da amostra refletiram, de fato, a intenção de votos de toda a população de eleitores.

A ciência que se dedica a esse trabalho é a Estatística.

Primeiramente, dimensionou uma amostra da população e fez a coleta de dados por meio de uma pesquisa de campo. A escolha da amostra é, em geral, complexa, pois deve-se levar em conta, entre outros fatores, o tempo e o custo da pesquisa, o número de eleitores de cada cidade do estado, a camada social à qual o entrevistado pertence, o local onde será feita a entrevista. É imprescindível que a amostra seja representativa, a fim de não haver comprometimento na análise dos resultados.

Num segundo momento, organizou em tabelas os dados brutos coletados, construiu gráficos para apresentar os resultados obtidos e divulgou-os nos meios de comunicação. É preciso também associar ao conjunto de informações medidas de tendência central e medidas de variabilidade (ou dispersão dos dados em relação a valores centrais).

Por fim, fez a análise confirmatória dos dados, isto é, verificou a margem de erro com que os resultados da amostra refletiram, de fato, a intenção de votos de toda a população de eleitores.

Primeiramente, dimensionou uma amostra da população e fez a coleta de dados por meio de uma pesquisa de campo. A escolha da amostra é, em geral, complexa, pois deve-se levar em conta, entre outros fatores, o tempo e o custo da pesquisa, o número de eleitores de cada cidade do estado, a camada social à qual o entrevistado pertence, o local onde será feita a entrevista. É imprescindível que a amostra seja representativa, a fim de não haver comprometimento na análise dos resultados.

Num segundo momento, organizou em tabelas os dados brutos coletados, construiu gráficos para apresentar os resultados obtidos e divulgou-os nos meios de comunicação. É preciso também associar ao conjunto de informações medidas de tendência central e medidas de variabilidade (ou dispersão dos dados em relação a valores centrais).

Por fim, fez a análise confirmatória dos dados, isto é, verificou a margem de erro com que os resultados da amostra refletiram, de fato, a intenção de votos de toda a população de eleitores.

Primeiramente, dimensionou uma amostra da população e fez a coleta de dados por meio de uma pesquisa de campo. A escolha da amostra é, em geral, complexa, pois deve-se levar em conta, entre outros fatores, o tempo e o custo da pesquisa, o número de eleitores de cada cidade do estado, a camada social à qual o entrevistado pertence, o local onde será feita a entrevista. É imprescindível que a amostra seja representativa, a fim de não haver comprometimento na análise dos resultados.

Num segundo momento, organizou em tabelas os dados brutos coletados, construiu gráficos para apresentar os resultados obtidos e divulgou-os nos meios de comunicação. É preciso também associar ao conjunto de informações medidas de tendência central e medidas de variabilidade (ou dispersão dos dados em relação a valores centrais).

Por fim, fez a análise confirmatória dos dados, isto é, verificou a margem de erro com que os resultados da amostra refletiram, de fato, a intenção de votos de toda a população de eleitores.

A **Estatística Descritiva** é utilizada também para se **organizar e resumir informações relativas a uma população inteira**, como ocorre, por exemplo, nos censos demográficos efetuados pelo Instituto Brasileiro de Geografia e Estatística (IBGE).

Variável

Sexo	Idade	Frequência semanal	Estado civil	Meio de transporte	Tempo de permanência	Renda familiar mensal (em salários mínimos)
Masculino	26	2	casado	carro	30 min	13,3
Masculino	23	1	solteiro	ônibus	35 min	11,8
Feminino	41	5	viúva	a pé	2h50min	8,9
Masculino	49	3	separado	a pé	45 min	13,9
Feminino	19	5	solteira	carro	1 h	11,6
Feminino	20	4	solteira	a pé	1h20min	16,0
Masculino	27	3	solteiro	carro	45 min	19,5
Masculino	38	3	casado	a pé	2h15min	9,3
Feminino	50	7	casada	a pé	45 min	12,4
Masculino	52	2	solteiro	a pé	1h40min	10,7
Feminino	48	4	casada	a pé	1h15min	14,7
Masculino	28	4	casado	a pé	1 h	16,6
Masculino	36	1	casado	carro	1h30min	12,5
Feminino	31	3	solteira	ônibus	2 h	8,2
Masculino	56	3	viúvo	a pé	30 min	15,4
Feminino	41	6	solteira	carro	2h30min	18,8
Masculino	44	1	casado	ônibus	50 min	12,1
Feminino	29	2	separada	a pé	40 min	5,0
Masculino	31	3	casado	ônibus	2h45min	7,6

Variáveis

Quantitativas

Qualitativas

Variáveis Qualitativas

- Algumas variáveis, como sexo, estado civil e meio de transporte utilizado para chegar ao parque, apresentam como resposta um atributo, qualidade ou preferência do entrevistado. Variáveis dessa natureza recebem o nome de variáveis qualitativas.
- Se considerarmos, por exemplo, a variável meio de transporte utilizado, dizemos que carro, ônibus e a pé correspondem às realizações ou valores assumidos por essa variável.

Variáveis Quantitativas

- **Discretas**
 - São aquelas cujos valores são obtidos por contagem e representados por elementos de um conjunto finito ou enumerável. No exemplo, a variável frequência semanal é discreta, e seus valores são 1, 2, 3, 4, 5, 6 ou 7.
- **Contínuas**
 - São aquelas cujos valores são obtidos por mensuração e representados por valores pertencentes a um intervalo real. As variáveis idade, tempo de permanência e renda familiar mensal são contínuas e seus valores se distribuem em determinado intervalo real. A variável tempo de permanência, por exemplo, tem seus valores (em horas) pertencentes ao intervalo [0,5; 3[.

- 255.** Ao se cadastrar em um site de comércio eletrônico, o usuário deve preencher um questionário com estas oito perguntas:
1. Você tem computador em casa?
 2. Quantas vezes por semana você acessa a Internet?
 3. Numa escala de zero a 10, qual seu índice de confiança na segurança do comércio eletrônico?
 4. Quantos cartões de crédito você possui?
 5. A residência em que vive é própria ou alugada?
 6. Qual é o provedor que você utiliza para acessar a rede?
 7. Qual é o tempo médio de acesso à Internet?
 8. Já comprou algum produto via Internet?

Cada uma das questões anteriores define uma variável. Classifique-as como qualitativas ou quantitativas.

Tabelas de Frequênciā

Tabelas de Frequência

Sexo	Idade	Frequência semanal	Estado civil	Meio de transporte	Tempo de permanência	Renda familiar mensal (em salários mínimos)
Masculino	26	2	casado	carro	30 min	13,3
Masculino	23	1	solteiro	ônibus	35 min	11,8
Feminino	41	5	viúva	a pé	2h50min	8,9
Masculino	49	3	separado	a pé	45 min	13,9
Feminino	19	5	solteira	carro	1 h	11,6
Feminino	20	4	solteira	a pé	1h20min	16,0
Masculino	27	3	solteiro	carro	45 min	19,5
Masculino	38	3	casado	a pé	2h15min	9,3
Feminino	50	7	casada	a pé	45 min	12,4
Masculino	52	2	solteiro	a pé	1h40min	10,7
Feminino	48	4	casada	a pé	1h15min	14,7
Masculino	28	4	casado	a pé	1 h	16,6
Masculino	36	1	casado	carro	1h30min	12,5
Feminino	31	3	solteira	ônibus	2 h	8,2
Masculino	56	3	viúvo	a pé	30 min	15,4
Feminino	41	6	solteira	carro	2h30min	18,8
Masculino	44	1	casado	ônibus	50 min	12,1
Feminino	29	2	separada	a pé	40 min	5,0
Masculino	31	3	casado	ônibus	2h45min	7,6

Tabelas de Frequência

- A simples leitura dos dados brutos da tabela anteriormente apresentada não nos fornece as condições necessárias à determinação do perfil do frequentador do parque, uma vez que as informações não estão devidamente organizadas.
- O primeiro procedimento que possibilita uma leitura mais resumida dos dados é a construção de tabelas de frequência.

Tabelas de Frequência

- Para cada variável estudada, conte o número de vezes que um dos seus valores ocorre.
 - Frequência Absoluta (n_i)
- Dos 20 entrevistados, você coletou os seguintes valores:
 - separado ($n_1 = 3$);
 - solteiro ($n_2 = 7$);
 - casado ($n_3 = 8$);
 - viúvo ($n_4 = 2$).

Tabelas de Frequência

- Para cada variável estudada, conte o número de vezes que um dos seus valores ocorre.
 - Frequência Absoluta (n_i)
- Dos 20 entrevistados, você coletou os seguintes valores:
 - separado ($n_1 = 3$);
 - solteiro ($n_2 = 7$);
 - casado ($n_3 = 8$);
 - viúvo ($n_4 = 2$).

$$n_1 + n_2 + n_3 + n_4 = \sum_{i=1}^4 n_i = 20$$

Tabelas de Frequência

- Por exemplo, em situações onde o tamanho de duas amostras são diferentes, seria interessante usar outra frequência além da absoluta:
 - Frequência Relativa

$$f_i = \frac{n_i}{n}$$

Tabelas de Frequência

- Por exemplo, em situações onde o tamanho de duas amostras são diferentes, seria interessante usar outra frequência além da absoluta:
 - Frequência Relativa

$$f_i = \frac{n_i}{n}$$

Observações:

- 1^a) Como $n_i \leq n$, segue que, para cada i , $0 \leq f_i \leq 1$. Por esse motivo, é comum a frequência relativa ser expressa em porcentagem.
- 2^a) A soma das frequências relativas dos valores assumidos por determinada variável é sempre igual a 1.

De fato:

$$\sum_i f_i = \sum_i \frac{n_i}{n} = \frac{1}{n} \sum_i n_i = \frac{1}{n} \cdot n = 1$$

Estado civil	Frequência absoluta (n_i)	Frequência relativa (f_i)	Porcentagem (%)
Separado	3	$\frac{3}{20} = 0,15$	15
Solteiro	7	$\frac{7}{20} = 0,35$	35
Casado	8	$\frac{8}{20} = 0,40$	40
Viúvo	2	$\frac{2}{20} = 0,10$	10
Total	20	1,0	100

Estado civil	Frequência absoluta (n_i)	Frequência relativa (f_i)	Porcentagem (%)
Separado	3	$\frac{3}{20} = 0,15$	15
Solteiro	7	$\frac{7}{20} = 0,35$	35
Casado	8	$\frac{8}{20} = 0,40$	40
Viúvo	2	$\frac{2}{20} = 0,10$	10
Total	20	1,0	100

Em alguns casos, porém, pode ocorrer que os valores assumidos por uma variável pertençam a determinado intervalo real, não havendo praticamente repetição (coincidência) de valores.

Isso ocorre com as variáveis idade, tempo de permanência no parque e renda familiar mensal. Esta última tem seus valores variando no intervalo [5, 20].

Nesse caso, construímos uma tabela de frequência em que os dados estarão agrupados em classes (ou intervalos) de valores.

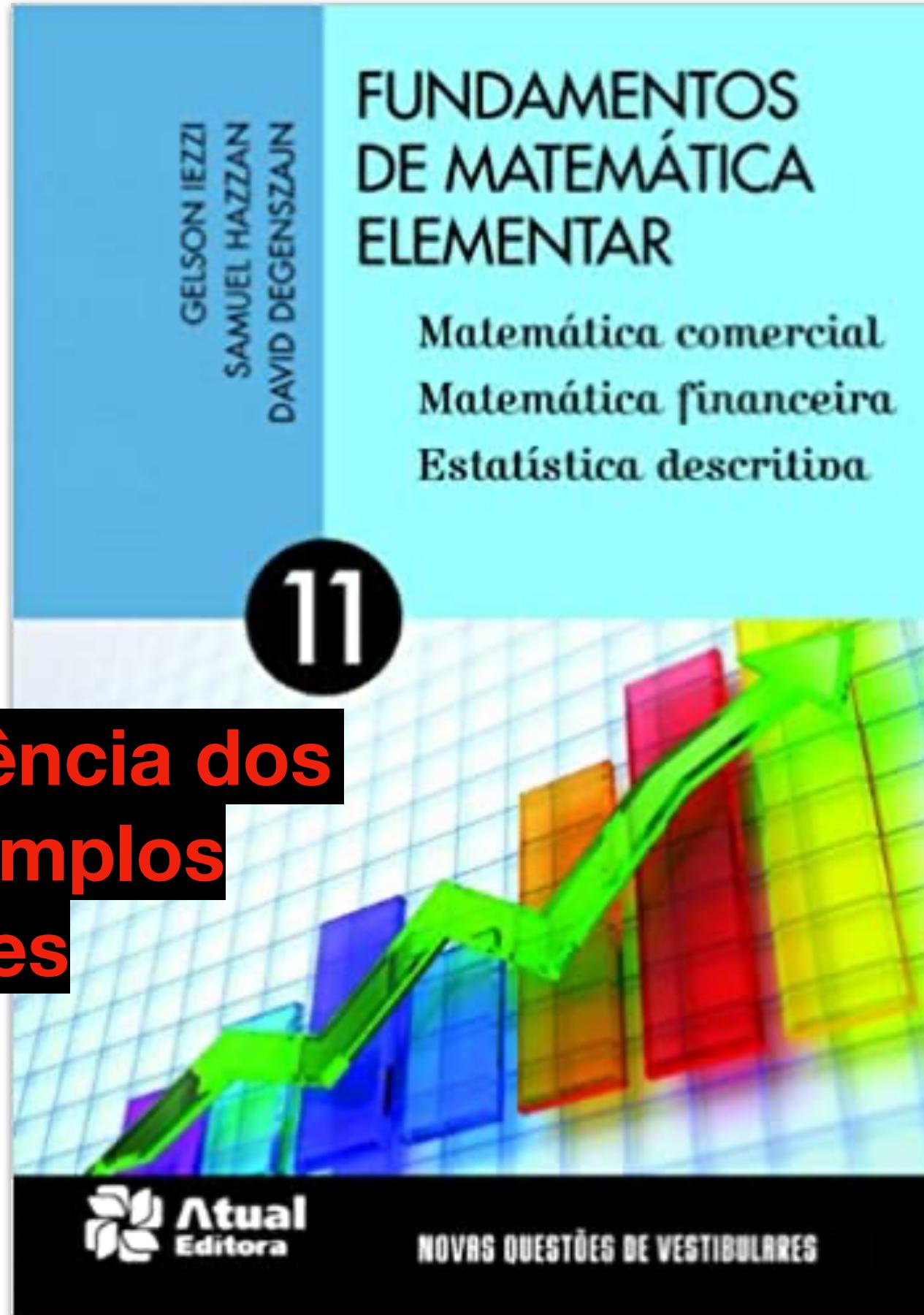
$$[a, b[= \{x \in \mathbb{R} \mid a \leq x < b\}$$

$$[0, 18[= \{x \text{ pertencente aos números Reais} \mid 0 \leq x < 18\}$$

Renda familiar mensal (em salários mínimos)	Frequência absoluta (n_i)	Frequência relativa (f_i)	Porcentagem (%)
5 \vdash 8	2	$\frac{2}{20} = 0,10$	10
8 \vdash 11	5	$\frac{5}{20} = 0,25$	25
11 \vdash 14	7	$\frac{7}{20} = 0,35$	35
14 \vdash 17	4	$\frac{4}{20} = 0,20$	20
17 \vdash 20	2	$\frac{2}{20} = 0,10$	10
Total	20	1,0	100

-) A amplitude do intervalo $a \vdash b$ é dada pela diferença $b - a$. (No exemplo que será fornecido a seguir, a amplitude de cada uma das classes da renda familiar é igual a 3.)

Principal Referência dos Textos e Exemplos dos Slides



CAPÍTULO III Estatística Descritiva

I. Introdução

Imagine que, um mês antes de uma eleição presidencial, a federação das indústrias de determinado estado encomendou a um instituto especializado uma pesquisa cujo objetivo consistiu em detectar a intenção de voto do eleitor e levantar o perfil socioeconômico dos eleitores de cada um dos candidatos.

O que o instituto fez?

- Primeiramente, dimensionou uma amostra da população e fez a coleta de dados por meio de uma pesquisa de campo. A escolha da amostra é, em geral, complexa, pois deve-se levar em conta, entre outros fatores, o tempo e o custo da pesquisa, o número de eleitores de cada cidade do estado, a camada social à qual o entrevistado pertence, o local onde será feita a entrevista. É imprescindível que a amostra seja representativa, a fim de não haver comprometimento na análise dos resultados.
- Num segundo momento, organizou em tabelas os dados brutos coletados, construiu gráficos para apresentar os resultados obtidos e divulgou-os nos meios de comunicação. É preciso também associar ao conjunto de informações medidas de tendência central e medidas de variabilidade (ou dispersão dos dados em relação a valores centrais).
- Por fim, fez a análise confirmatória dos dados, isto é, verificou a margem de erro com que os resultados da amostra refletiram, de fato, a intenção de votos de toda a população de eleitores.

A ciência que se dedica a esse trabalho é a **Estatística**.

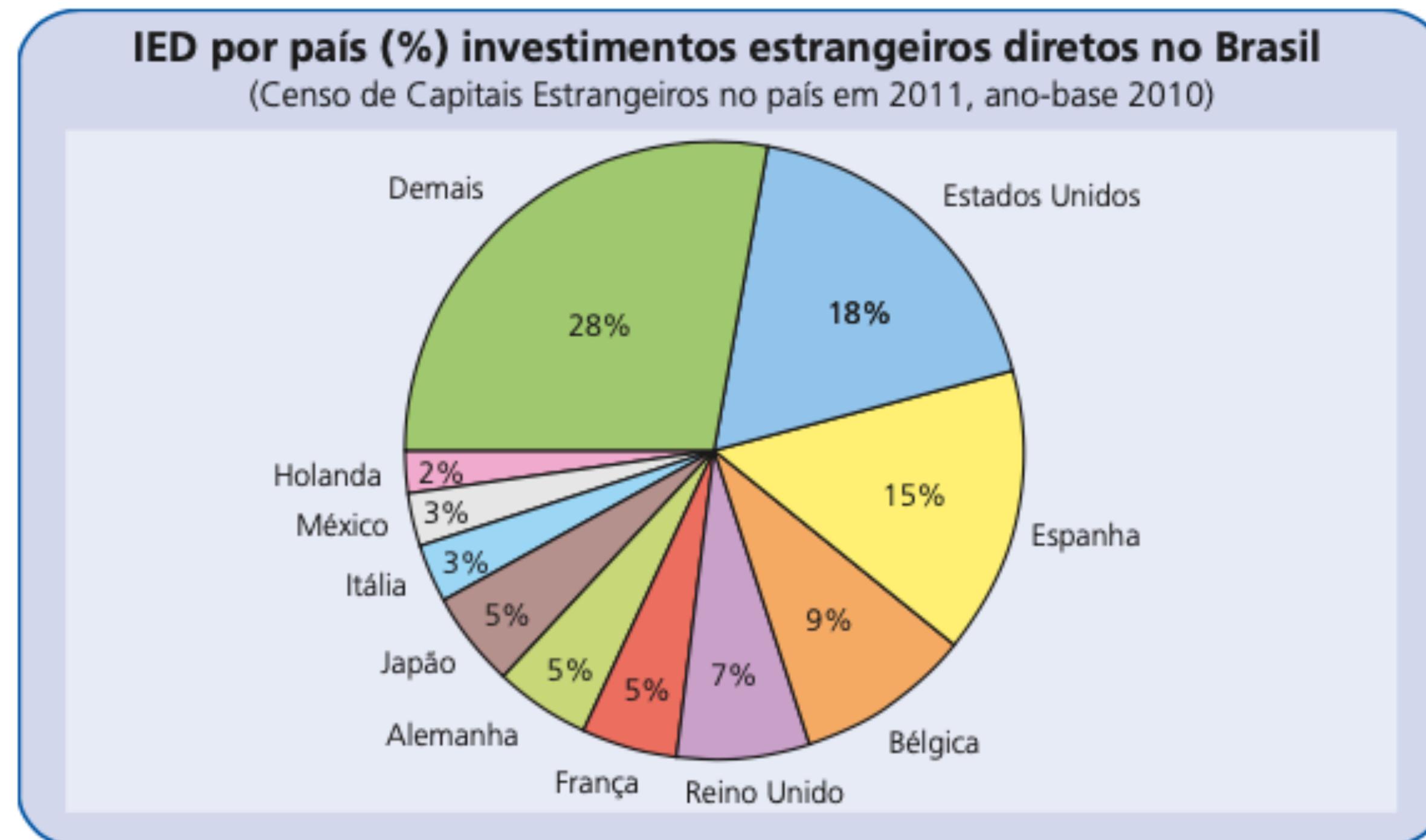
11 | Fundamentos de Matemática Elementar

Representação Gráfica

Representação Gráfica

- No **mundo ideal**, não use gráficos que têm nome de comida.
- No **mundo real**, use o que seu chefe insistir e garanta seu emprego (enquanto precisar).

Gráfico de setores



Fonte: Banco Central do Brasil e Ministério da Fazenda. Disponível em: <www.fazenda.gov.br>. Acesso em: 2 abr. 2013.

**Pizza engorda, não é saudável e normalmente tem massa grossa.
Logo, um gráfico de pizza também deve ser uma merda.**

Assista à Masterclass de Data Storytelling

Histograma

Distrito Federal	605,4
Santa Catarina	348,7
São Paulo	442,7
Rio Grande do Sul	357,7
Rio de Janeiro	413,9
Paraná	321,4
Mato Grosso do Sul	287,5
Goiás	286,0
Mato Grosso	288,1

Minas Gerais	276,6
Espírito Santo	289,6
Amapá	211,4
Roraima	232,5
Rondônia	233,8
Pará	168,6
Amazonas	173,9
Tocantins	172,6
Pernambuco	183,8

Rio Grande do Norte	176,2
Ceará	156,2
Acre	180,7
Bahia	160,2
Sergipe	163,5
Paraíba	150,2
Piauí	129,0
Alagoas	139,9
Maranhão	110,4

Fonte: *Folha de S.Paulo*, 3 out. 2003.

Histograma

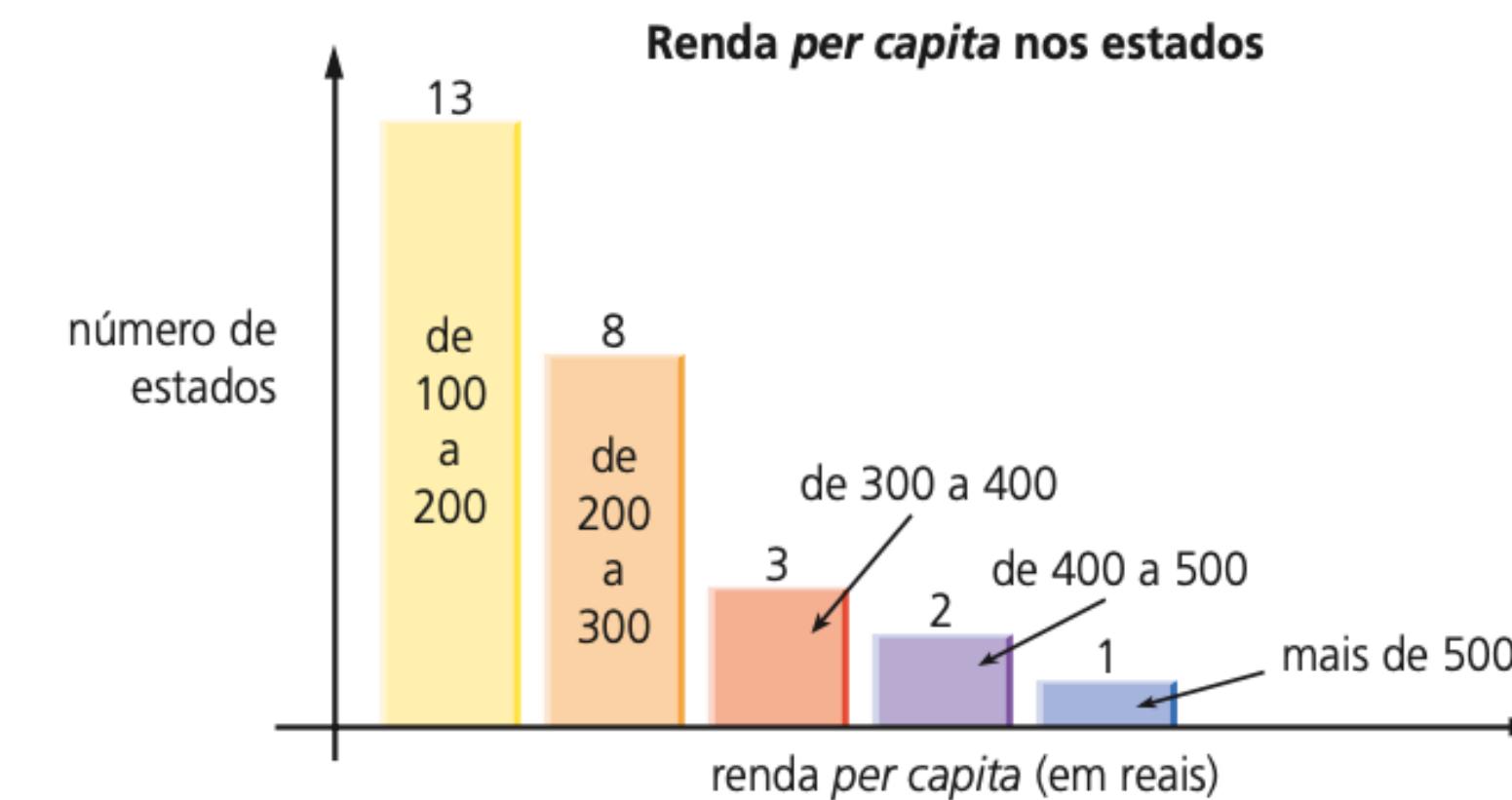
- Agrupando os valores em cinco classes de intervalos, é possível construir uma tabela de frequência.

Distrito Federal	605,4
Santa Catarina	348,7
São Paulo	442,7
Rio Grande do Sul	357,7
Rio de Janeiro	413,9
Paraná	321,4
Mato Grosso do Sul	287,5
Goiás	286,0
Mato Grosso	288,1

Minas Gerais	276,6
Espírito Santo	289,6
Amapá	211,4
Roraima	232,5
Rondônia	233,8
Pará	168,6
Amazonas	173,9
Tocantins	172,6
Pernambuco	183,8

Rio Grande do Norte	176,2
Ceará	156,2
Acre	180,7
Bahia	160,2
Sergipe	163,5
Paraíba	150,2
Piauí	129,0
Alagoas	139,9
Maranhão	110,4

Fonte: Folha de S.Paulo, 3 out. 2003.



Medidas de Centralidade e Variabilidade

Média Aritmética

- Seja x uma variável quantitativa e x_1, x_2, \dots, x_n , os valores assumidos por x .
- Define-se a média aritmética de x , como a divisão da soma de todos esses valores pelo número de valores.

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

Média Aritmética

Exemplos:

- 1º) Um aluno, preparando-se para o exame vestibular, fez 12 simulados no cursinho ao longo do ano. Em cada simulado, o número de questões era oitenta. Os valores seguintes correspondem às pontuações obtidas nesses exames:

$$56 - 52 - 61 - 53 - 48 - 68$$

$$49 - 59 - 61 - 62 - 60 - 55$$

Qual é a média aritmética desses valores?

Temos:

$$\bar{x} = \frac{\sum_{i=1}^{12} x_i}{12} = \frac{56 + 52 + \dots + 60 + 55}{12} = \frac{684}{12} = 57$$

A nota média obtida por esse aluno é 57 pontos. Qual é o significado desse valor?

Caso o aluno apresentasse a mesma pontuação (desempenho) em todos os simulados, essa pontuação deveria ser 57 pontos a fim de que fosse obtida a pontuação total de 684 pontos, equivalente à soma dos pontos obtidos efetivamente nas 12 provas.

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

Propriedades:

Vamos estudar agora duas propriedades da média aritmética.

Sejam x_1, x_2, \dots, x_n os valores assumidos por uma variável x e \bar{x} a média aritmética correspondente.

Se a cada x_i ($i = 1, 2, \dots, n$) adicionarmos uma constante real c , a média aritmética fica adicionada de c unidades.

Essa propriedade pode ser facilmente demonstrada.

Consideremos que os novos valores assumidos por essa variável sejam:
 $x_1 + c, x_2 + c, \dots, x_n + c$.

A nova média (\bar{x}') é dada por:

$$\begin{aligned}\bar{x}' &= \frac{\sum_{i=1}^n (x_i + c)}{n} = \frac{(x_1 + c) + (x_2 + c) + \dots + (x_n + c)}{n} = \\ &= \frac{(x_1 + x_2 + \dots + x_n)}{n} + \underbrace{\frac{(c + c + \dots + c)}{n}}_{n \text{ vezes}} + \frac{\sum_{i=1}^n x_i}{n} = \frac{n \cdot c}{n}\end{aligned}$$

isto é:

$$\bar{x}' = \bar{x} + c$$

Se multiplicarmos cada x_i ($i = 1, 2, \dots, n$) por uma constante real c , a média aritmética fica multiplicada por c .

Para demonstrar essa segunda propriedade, consideremos que os novos valores assumidos por essa variável sejam: cx_1, cx_2, \dots, cx_n .

A nova média (\bar{x}') é dada por:

$$\bar{x}' = \frac{\sum_{i=1}^n (cx_i)}{n} = \frac{cx_1 + cx_2 + \dots + cx_n}{n} = \frac{c \cdot (x_1 + x_2 + \dots + x_n)}{n} = c \cdot \frac{\sum_{i=1}^n x_i}{n}$$

isto é:

$$\bar{x}' = c \cdot \bar{x}$$

Média Aritmética Ponderada

- Seja x uma variável quantitativa e x_1, x_2, \dots, x_k , com frequências absolutas iguais a n_1, n_2, \dots, n_k . A média aritmética ponderada de x , é definida como a divisão de todos os produtos $x_i n_i (i = 1, 2, \dots, k)$ pela soma das frequências:

$$\bar{x} = \frac{\sum_{i=1}^k x_i \cdot n_i}{\sum_{i=1}^k n_i} = \frac{x_1 \cdot n_1 + x_2 \cdot n_2 + \dots + x_k \cdot n_k}{n_1 + n_2 + \dots + n_k}$$

Média Aritmética Ponderada

- Lembrando que a frequência relativa f_i , é definida por $f_i = \frac{n_i}{\sum_{i=1}^k n_i}$, é possível também expressar a média por:

$$\bar{x} = \sum_{i=1}^k x_i \cdot f_i = x_1 \cdot f_1 + x_2 \cdot f_2 + \dots + x_k \cdot f_k$$

Exemplos:

- 1º) Um feirante possuía 50 kg de maçã para vender em uma manhã. Começou a vender as frutas por R\$ 2,50 o quilo e, com o passar das horas, reduziu o preço em duas ocasiões para não haver sobras. A tabela seguinte informa a quantidade de maçãs vendidas em cada período, bem como os diferentes preços cobrados pelo feirante.

Período	Preço por quilo (em reais)	Número de quilos de maçã vendidos
Até às 10 h	2,50	32
Das 10 h às 11 h	2,00	13
Das 11 h às 12 h	1,40	5

Naquela manhã, por quanto foi vendido, em média, o quilo da maçã?
Sendo \bar{p} o preço médio do quilo da maçã, temos:

$$\bar{p} = \frac{\overbrace{2,50 + 2,50 + \dots + 2,50}^{32 \text{ vezes}} + \overbrace{2,0 + \dots + 2,0}^{13 \text{ vezes}} + \overbrace{1,40 + 1,40 + \dots + 1,40}^{5 \text{ vezes}}}{32 + 13 + 5}$$

isto é:

$$\bar{p} = \frac{2,50 \cdot 32 + 2,00 \cdot 13 + 1,40 \cdot 5}{50} = \frac{113}{50} \cong 2,26 \text{ reais}$$

Ou seja, 2,26 reais é o preço médio do quilo de maçãs vendido.

Dizemos que se trata de uma média aritmética ponderada dos preços, em que o “fator de ponderação” (que também pode ser chamado de “peso”) corresponde à quantidade de maçãs vendidas (frequência absoluta) em cada período.

Mediana

- Em 2002, a população brasileira era constituída por aproximadamente 175 milhões de habitantes.
- A área da superfície do território brasileiro é 8514204,8 km².

Estado	Densidade demográfica
Acre	3,7
Alagoas	101,3
Amapá	3,3
Amazonas	1,8
Bahia	23,2
Ceará	50,9
Distrito Federal	352,2
Espírito Santo	67,2
Goiás	14,7

Estado	Densidade demográfica
Maranhão	17,0
Mato Grosso	2,8
Mato Grosso do Sul	5,8
Minas Gerais	30,5
Pará	5,0
Paraíba	61,1
Paraná	48,0
Pernambuco	80,3
Piauí	11,3

Estado	Densidade demográfica
Rio de Janeiro	328,0
Rio Grande do Norte	52,2
Rio Grande do Sul	36,1
Rondônia	5,8
Roraima	1,5
Santa Catarina	56,1
São Paulo	149,0
Sergipe	81,1
Tocantins	4,2

$$\frac{175 \text{ milhões de habitantes}}{8,514 \text{ milhões de km}^2} \approx 20,6 \text{ habitantes/km}^2$$

Mediana

- Em 2002, a população brasileira era constituída por aproximadamente 175 milhões de habitantes.
- A área da superfície do território brasileiro é 8514204,8 km².

Estado	Densidade demográfica
Acre	3,7
Alagoas	101,3
Amapá	3,3
Amazonas	1,8
Bahia	23,2
Ceará	50,9
Distrito Federal	352,2
Espírito Santo	67,2
Goiás	14,7

Estado	Densidade demográfica
Maranhão	17,0
Mato Grosso	2,8
Mato Grosso do Sul	5,8
Minas Gerais	30,5
Pará	5,0
Paraíba	61,1
Paraná	48,0
Pernambuco	80,3
Piauí	11,3

Estado	Densidade demográfica
Rio de Janeiro	328,0
Rio Grande do Norte	52,2
Rio Grande do Sul	36,1
Rondônia	5,8
Roraima	1,5
Santa Catarina	56,1
São Paulo	149,0
Sergipe	81,1
Tocantins	4,2

$$\frac{175 \text{ milhões de habitantes}}{8,514 \text{ milhões de km}^2} \cong 20,6 \text{ habitantes/km}^2$$

$$\bar{x} = \frac{\sum_{i=1}^{27} x_i}{27} = \frac{1594,1}{27} \cong 59,04 \text{ habitantes/km}^2$$

Mediana

- Em 2002, a população brasileira era constituída por aproximadamente 175 milhões de habitantes.
- A área da superfície do território brasileiro é 8514204,8 km².

Estado	Densidade demográfica
Acre	3,7
Alagoas	101,3
Amapá	3,3
Amazonas	1,8
Bahia	23,2
Ceará	50,9
Rio Grande do Ceará	352,2
Espírito Santo	67,2
Goiás	14,7

Estado	Densidade demográfica
Maranhão	17,0
Mato Grosso	2,8
Mato Grosso do Sul	5,8
Minas Gerais	30,5
Pará	5,0
Paraíba	61,1
Paraná	48,0
Pernambuco	80,3
Piauí	11,3

Estado	Densidade demográfica
Rio de Janeiro	328,0
Rio Grande do Norte	52,2
Rio Grande do Sul	36,1
Rondônia	5,8
Roraima	1,5
Santa Catarina	56,1
São Paulo	149,0
Sergipe	81,1
Tocantins	4,2

$$\frac{175 \text{ milhões de habitantes}}{8,514 \text{ milhões de km}^2} \approx 20,6 \text{ habitantes/km}^2$$

$$\bar{x}' = \frac{1594,1 - 352,2 - 328,0}{25} = \frac{913,9}{25} \approx 36,6 \text{ habitantes/km}^2$$

Mediana

- Em 2002, a população brasileira era constituída por aproximadamente 175 milhões de habitantes.
- A área da superfície do território brasileiro é 8514204,8 km².

Estado	Densidade demográfica
Acre	3,7
Alagoas	101,3
Amapá	3,3
Amazonas	1,8
Bahia	23,2
Ceará	50,9
Rio Grande do Ceará	352,2
Espírito Santo	67,2
Goiás	14,7

Estado	Densidade demográfica
Maranhão	17,0
Mato Grosso	2,8
Mato Grosso do Sul	5,8
Minas Gerais	30,5
Pará	5,0
Paraíba	61,1
Paraná	48,0
Pernambuco	80,3
Piauí	11,3

Estado	Densidade demográfica
Rio de Janeiro	328,0
Rio Grande do Norte	52,2
Rio Grande do Sul	36,1
Rondônia	5,8
Roraima	1,5
Santa Catarina	56,1
Se	149,0
Sergipe	81,1
Tocantins	4,2

$$\frac{175 \text{ milhões de habitantes}}{8,514 \text{ milhões de km}^2} \cong 20,6 \text{ habitantes/km}^2$$

$$\bar{x}'' = \frac{913,9 - 149}{24} = \frac{764,9}{24} \cong 31,9 \text{ habitantes/km}^2$$

Mediana

1, 3, 1000

- Como vimos, a média aritmética pode ser muito afetada quando encontramos valores discrepantes em um conjunto de dados, podendo se tornar uma medida de centralidade pouco representativa do resumo dos dados.
- Para contornar questões dessa natureza, definiremos, a seguir, uma medida de centralidade mais resistente aos valores discrepantes (em inglês, chamados *outliers*) denominada mediana.

Sejam $x_1 \leq x_2 \leq \dots \leq x_n$ os n valores ordenados de uma variável x .

A **mediana** desse conjunto de valores – indicada por Me – é definida por:

$$Me = \begin{cases} x_{\left(\frac{n+1}{2}\right)}, & \text{se } n \text{ é ímpar} \\ \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}, & \text{se } n \text{ é par} \end{cases}$$

OUTLIER: Atenção

- Não necessariamente, um outlier significa um valor errado ou lançado incorretamente.
- Por exemplo, *outliers* podem ser valores corretos que têm o potencial de afetar a média e troná-la pouco representativa como insumo do resumo dos dados.

Mediana

- Essa definição garante que a mediana seja um valor que divide o conjunto de dados em duas partes nas quais o número de elementos é o mesmo e de modo que o número de valores menores ou iguais à mediana seja igual ao número de valores maiores ou iguais a ela.

Sejam $x_1 \leq x_2 \leq \dots \leq x_n$ os n valores ordenados de uma variável x .

A **mediana** desse conjunto de valores – indicada por M_e – é definida por:

$$M_e = \begin{cases} x_{\left(\frac{n+1}{2}\right)}, & \text{se } n \text{ é ímpar} \\ \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}, & \text{se } n \text{ é par} \end{cases}$$

Moda

Seja x uma variável quantitativa que assume os valores x_1, x_2, \dots, x_k , com frequências absolutas iguais a n_1, n_2, \dots, n_k , respectivamente. Se o máximo entre n_1, n_2, \dots, n_k é igual a n_j , $j \in \{1, 2, \dots, k\}$, dizemos que a moda – indicada por M_o – é igual ao valor x_j .

Ou seja:

A moda de um conjunto de valores corresponde ao valor que ocorre mais vezes.

Moda

Seja x uma variável quantitativa que assume os valores x_1, x_2, \dots, x_k , com frequências absolutas iguais a n_1, n_2, \dots, n_k , respectivamente. Se o máximo entre n_1, n_2, \dots, n_k é igual a n_j , $j \in \{1, 2, \dots, k\}$, dizemos que a moda – indicada por M_o – é igual ao valor x_j .

Ou seja:

A moda de um conjunto de valores corresponde ao valor que ocorre mais vezes.

Exemplos:

Vamos determinar a moda dos seguintes conjuntos de valores.

1º) 6 – 9 – 12 – 9 – 4 – 5 – 9

A moda é $M_o = 9$, pois há três valores iguais a 9.

2º) 12 – 13 – 19 – 13 – 14 – 12 – 16

Há duas modas, 12 e 13, pois cada um desses valores ocorre com maior frequência (duas vezes). Dizemos que se trata de uma distribuição **bimodal**.

3º) 4 – 29 – 15 – 13 – 18 – 20 – 21 – 26 – 9

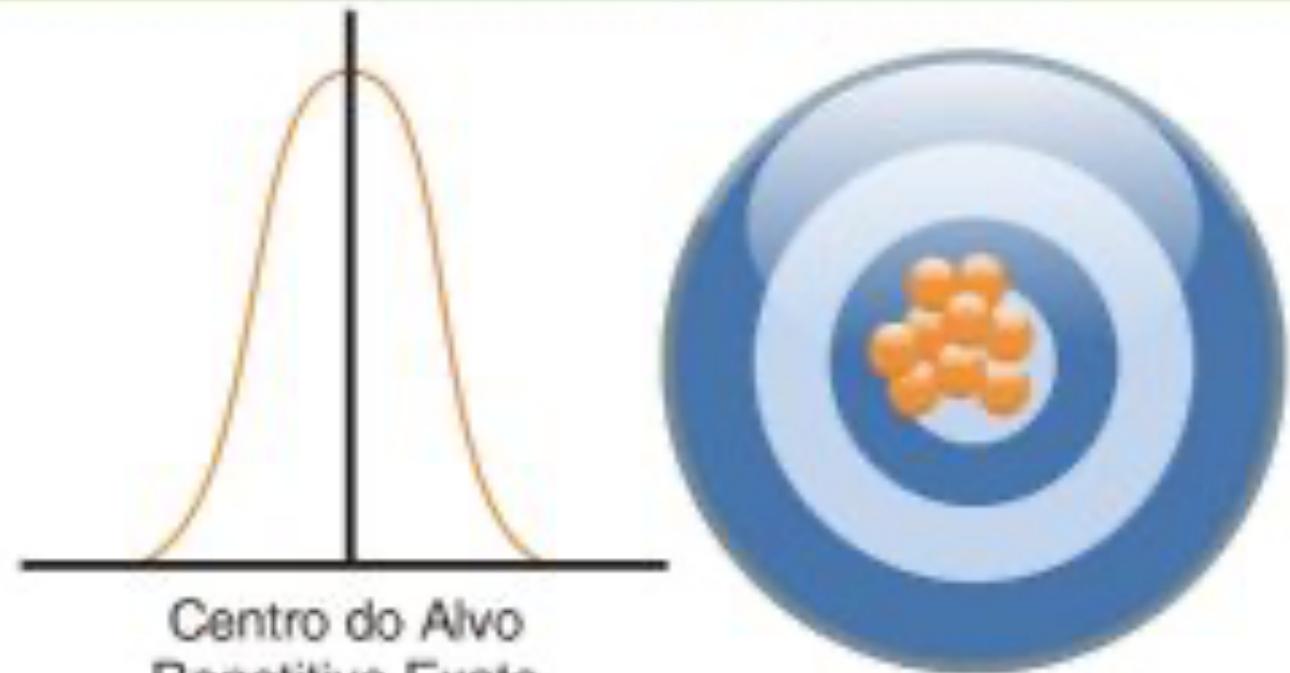
Nesse caso, todos os valores “aparecem” com a mesma frequência unitária. Assim, não há moda nessa distribuição.

Variância



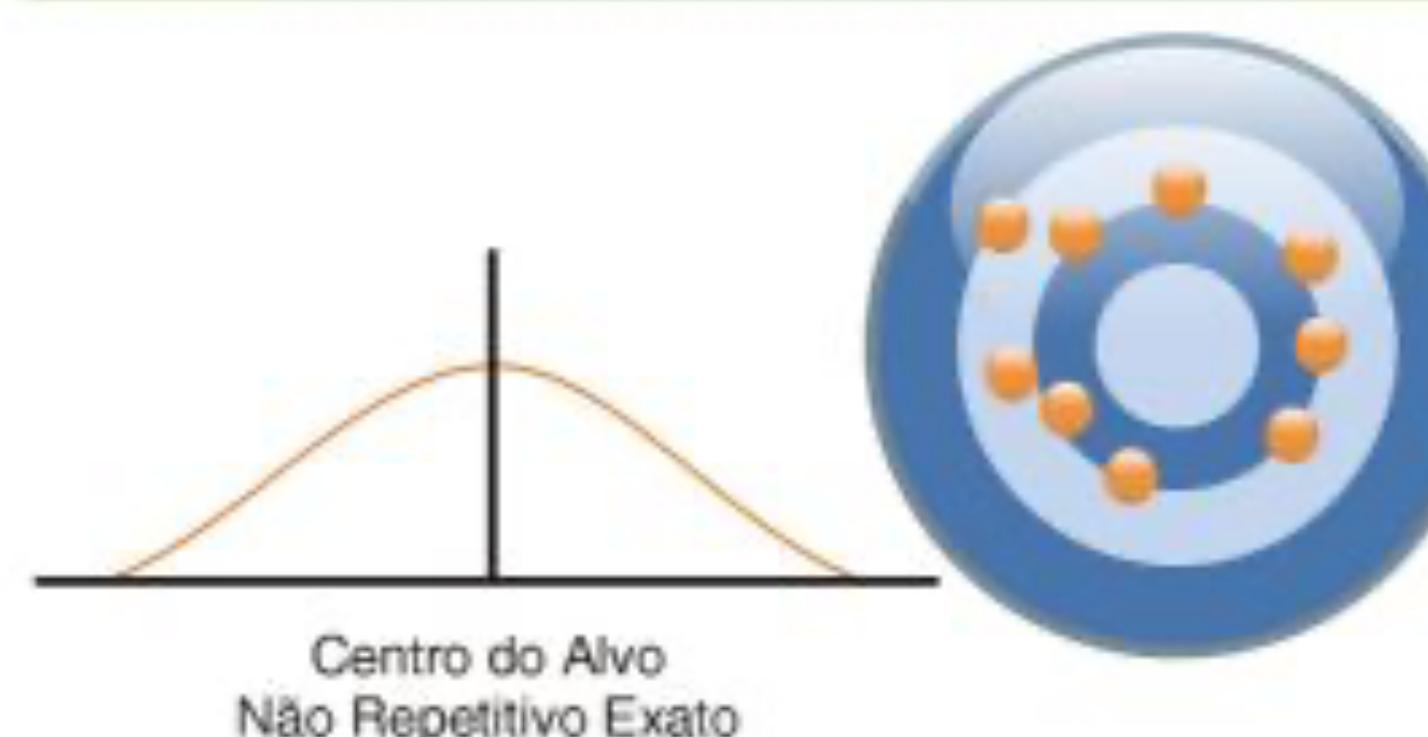


Exato e preciso



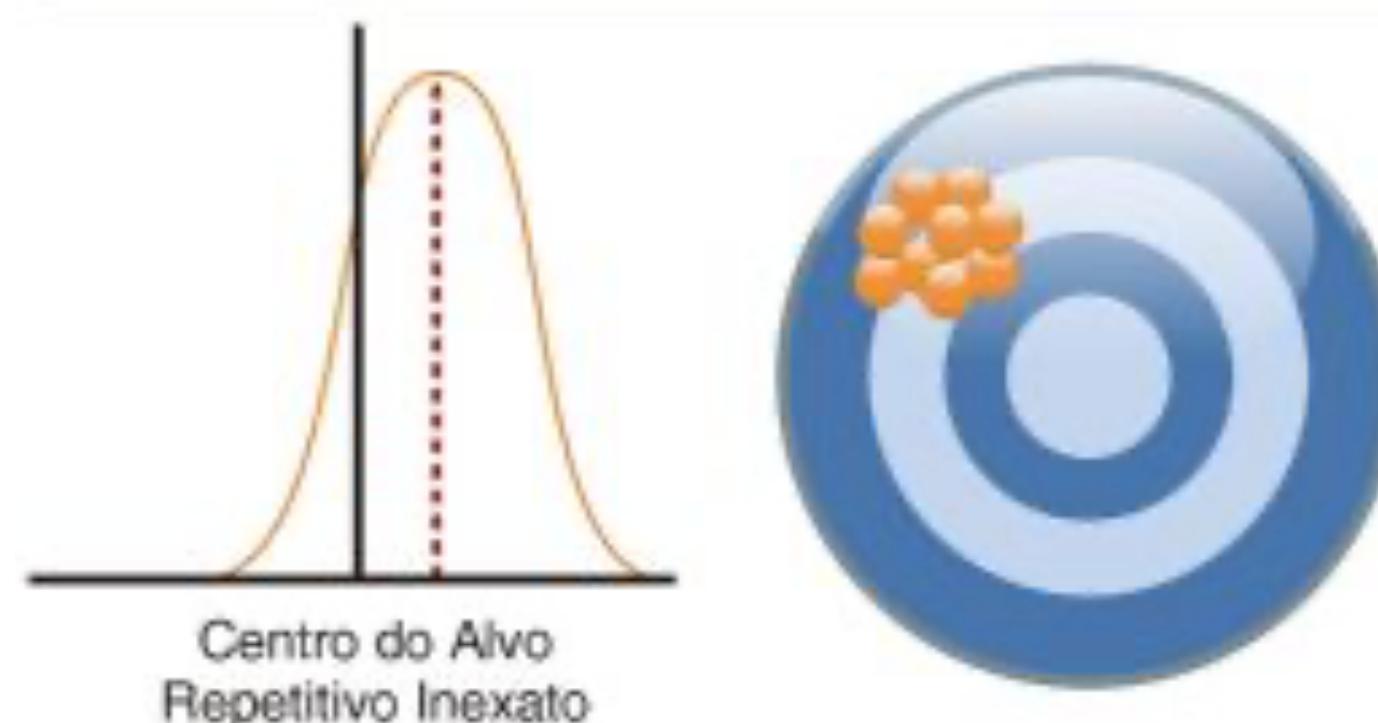
Centro do Alvo
Repetitivo Exato

Exato mas não preciso



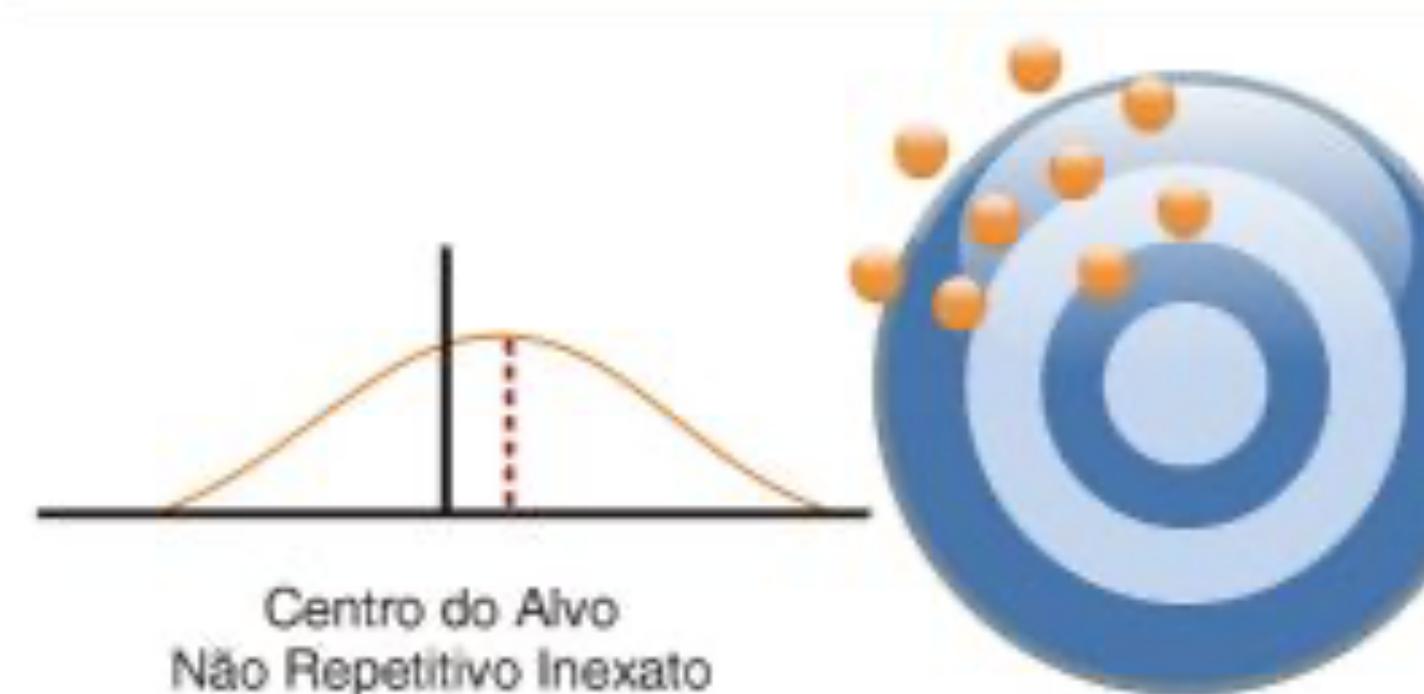
Centro do Alvo
Não Repetitivo Exato

Preciso mas não exato

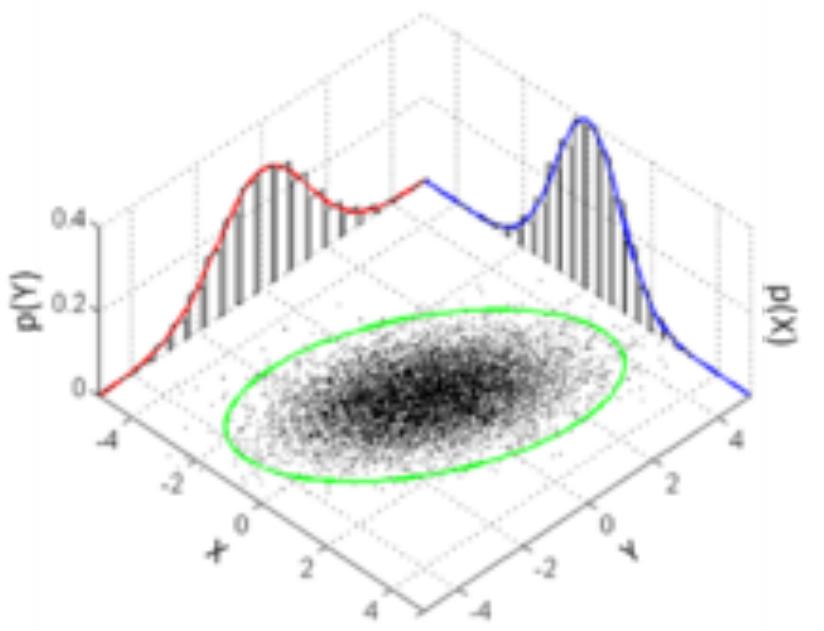
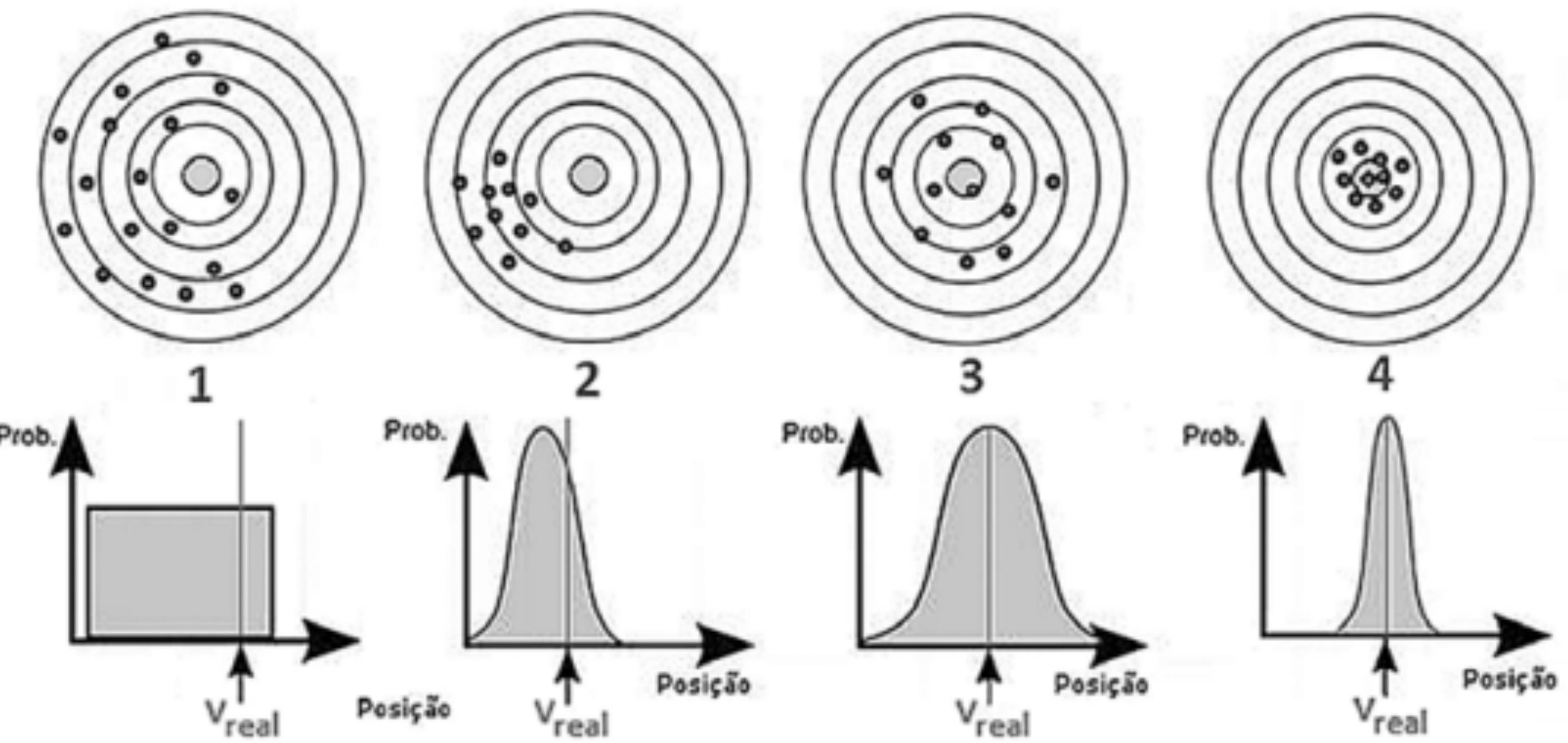


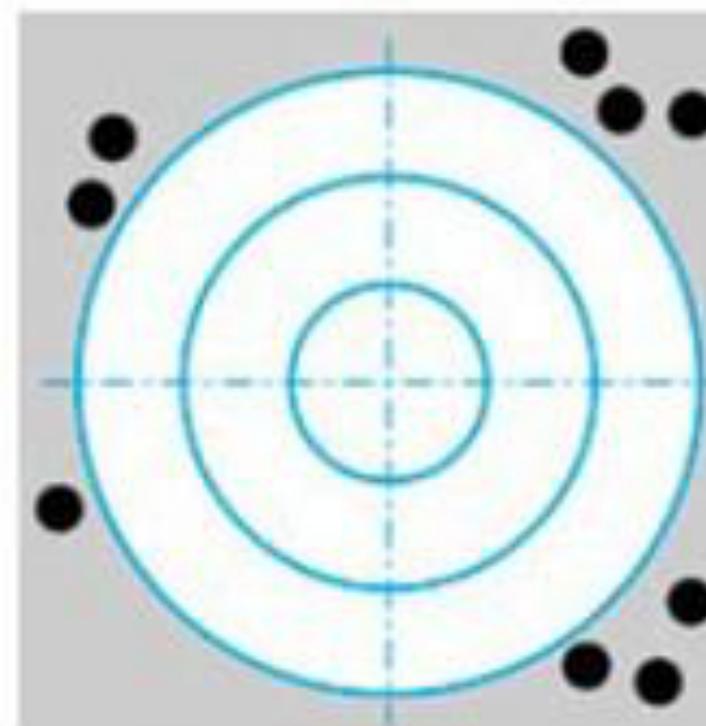
Centro do Alvo
Repetitivo Inexato

Não preciso e não exato



Centro do Alvo
Não Repetitivo Inexato



**Baixa exatidão:**

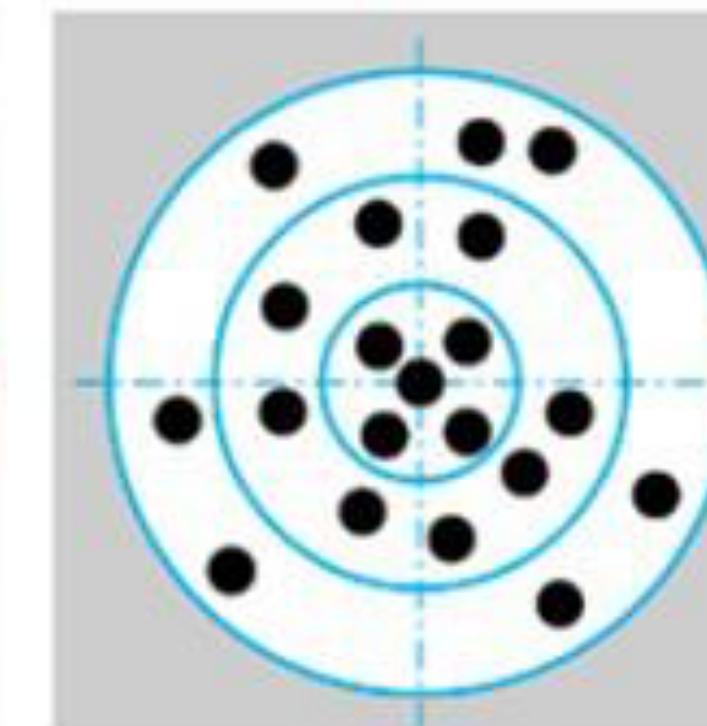
Os resultados estão longe do centro.

Baixa reproduzibilidade:

Os resultados estão muito dispersos.

Resultado:

Estes instrumentos volumétricos são de baixa qualidade.

**Boa exatidão:**

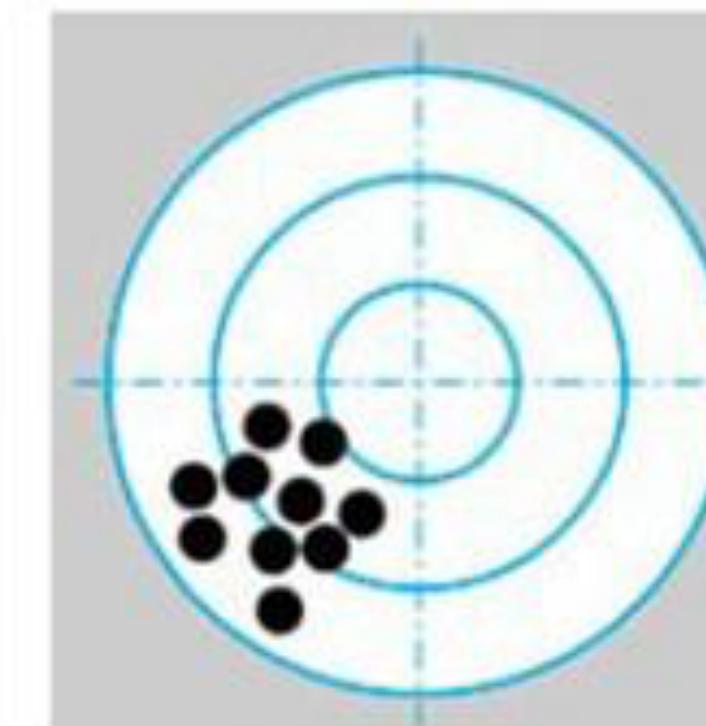
Os resultados estão distribuídos regularmente ao redor do centro.

Baixa reproduzibilidade:

Não há grandes erros, mas os resultados estão muito dispersos.

Resultado:

Todas as desvios têm a "mesma" probabilidade. Os instrumentos volumétricos cujos valores ultrapassam os limites de erro devem ser retirados.

**Baixa exatidão:**

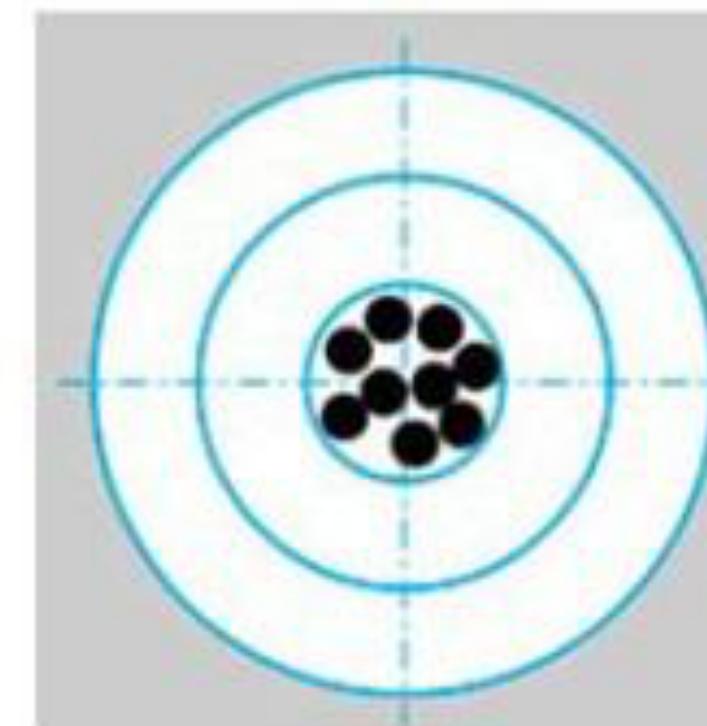
Apesar de todos os resultados estarem muito próximos entre si, a meta (valor nominal) não foi alcançada.

Boa reproduzibilidade:

Todos os resultados estão muito próximos entre si.

Resultado:

Produção mal controlada, com variação sistemática. Os instrumentos volumétricos cujos valores ultrapassam os limites de erro devem ser retirados.

**Boa exatidão:**

Todos os resultados estão muito próximos do centro, ou seja, do valor nominal.

Boa reproduzibilidade:

Todos os resultados estão muito próximos entre si.

Resultado:

A fabricação está perfeitamente orientada, através de um controle de qualidade ao longo do processo de fabricação. Mínimo desvio sistemático e estreita dispersão. O limite de erro permitido não é alcançado. Estes instrumentos devem ser mantidos.

Variância

Em certo país, o governo financia um programa de assistência às famílias de baixa renda. Cada família recebe, de cinco em cinco semanas, a quantia de 100 UM (unidades monetárias) para comprar produtos de alimentação em estabelecimentos conveniados. O coordenador desse projeto selecionou em uma pequena cidade quatro famílias e acompanhou a distribuição dos gastos semana a semana.

Observe a tabela:

	1^a semana	2^a semana	3^a semana	4^a semana	5^a semana	Total (valor do benefício)
Família I	20 UM	100 UM				
Família II	20 UM	24 UM	20 UM	16 UM	20 UM	100 UM
Família III	12 UM	28 UM	24 UM	20 UM	16 UM	100 UM
Família IV	36 UM	32 UM	20 UM	8 UM	4 UM	100 UM

Como cada família gasta 100 UM no período de cinco semanas, a média semanal de gastos é $\frac{100}{5} = 20$ UM.

Variância

- Se essa análise for limitada à média semanal de gastos, estarão sendo omitidas informações importantes em relação à homogeneidade ou heterogeneidade dos gastos semanais de cada família. Para revelar o grau de variabilidade de um conjunto de dados há necessidade de uma medida específica, a **variância**, definida a seguir.

Seja x uma variável quantitativa que assume os valores x_1, x_2, \dots, x_n e \bar{x} a média aritmética correspondente a esses valores.

A variância desses valores – indicada por $\text{Var}(x)$ ou σ^2 – é definida por:

$$\sigma^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n} = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n}$$

Variância

- Notemos que cada termo do numerador corresponde ao quadrado da diferença entre um valor observado e o valor médio. Essa diferença traduz o quanto um valor observado se distancia do valor médio, sendo, portanto, uma medida do grau de variabilidade dos dados em estudo.

Seja x uma variável quantitativa que assume os valores x_1, x_2, \dots, x_n e \bar{x} a média aritmética correspondente a esses valores.

A variância desses valores – indicada por $\text{Var}(x)$ ou σ^2 – é definida por:

$$\sigma^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n} = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n}$$

Variância

- Família I:

$$\sigma^2 = \frac{(20 - 20)^2 + (20 - 20)^2 + (20 - 20)^2 + (20 - 20)^2 + (20 - 20)^2}{5} = 0$$

- Família II:

$$\begin{aligned}\sigma^2 &= \frac{(20 - 20)^2 + (24 - 20)^2 + (20 - 20)^2 + (16 - 20)^2 + (20 - 20)^2}{5} = \frac{16 + 16}{5} = \\ &= 6,4 \text{ UM}^2\end{aligned}$$

- Família III:

$$\begin{aligned}\sigma^2 &= \frac{(12 - 20)^2 + (28 - 20)^2 + (24 - 20)^2 + (20 - 20)^2 + (16 - 20)^2}{5} = \\ &= \frac{64 + 64 + 16 + 16}{5} = 32 \text{ UM}^2\end{aligned}$$

- Família IV:

$$\begin{aligned}\sigma^2 &= \frac{(36 - 20)^2 + (32 - 20)^2 + (20 - 20)^2 + (8 - 20)^2 + (4 - 20)^2}{5} = \\ &= \frac{256 + 144 + 144 + 256}{5} = 160 \text{ UM}^2\end{aligned}$$

	1ª semana	2ª semana	3ª semana	4ª semana	5ª semana	Total (valor do benefício)
Família I	20 UM	100 UM				
Família II	20 UM	24 UM	20 UM	16 UM	20 UM	100 UM
Família III	12 UM	28 UM	24 UM	20 UM	16 UM	100 UM
Família IV	36 UM	32 UM	20 UM	8 UM	4 UM	100 UM

O aumento no valor da variância nesses cálculos revela uma variabilidade crescente de gastos semanais em relação à média (20 UM).

Nessa situação, a **unidade de variância** é UM^2 e o **gasto semanal médio** é expresso em UM, o que gera uma incompatibilidade. Para uniformizar as unidades, definiremos mais adiante o desvio padrão σ .

Desvio Padrão

Sejam x_1, x_2, \dots, x_n os valores assumidos por uma variável x . Chamamos **desvio padrão** de x – indicado por $DP(x)$ ou σ – a raiz quadrada da variância de x .

$$\sigma = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}} = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n}}$$

Desvio Padrão

Sejam x_1, x_2, \dots, x_n os valores assumidos por uma variável x . Chamamos **desvio padrão** de x – indicado por $DP(x)$ ou σ – a raiz quadrada da variância de x .

$$\sigma = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}} = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n}}$$

- Família I:

$$\sigma^2 = \frac{(20 - 20)^2 + (20 - 20)^2 + (20 - 20)^2 + (20 - 20)^2 + (20 - 20)^2}{5} = 0$$

- Família II:

$$\begin{aligned}\sigma^2 &= \frac{(20 - 20)^2 + (24 - 20)^2 + (20 - 20)^2 + (16 - 20)^2 + (20 - 20)^2}{5} = \\ &= 6,4 \text{ UM}^2\end{aligned}$$

- Família III:

$$\begin{aligned}\sigma^2 &= \frac{(12 - 20)^2 + (28 - 20)^2 + (24 - 20)^2 + (20 - 20)^2 + (16 - 20)^2}{5} = \\ &= \frac{64 + 64 + 16 + 16}{5} = 32 \text{ UM}^2\end{aligned}$$

- Família IV:

$$\begin{aligned}\sigma^2 &= \frac{(36 - 20)^2 + (32 - 20)^2 + (20 - 20)^2 + (8 - 20)^2 + (4 - 20)^2}{5} = \\ &= \frac{256 + 144 + 144 + 256}{5} = 160 \text{ UM}^2\end{aligned}$$

Desvio Padrão

Sejam x_1, x_2, \dots, x_n os valores assumidos por uma variável x . Chamamos **desvio padrão** de x – indicado por $DP(x)$ ou σ – a raiz quadrada da variância de x .

$$\sigma = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}} = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n}}$$

- Família I:

$$\sigma^2 = \frac{(20 - 20)^2 + (20 - 20)^2 + (20 - 20)^2 + (20 - 20)^2 + (20 - 20)^2}{5} = 0$$

- Família II:

$$\begin{aligned}\sigma^2 &= \frac{(20 - 20)^2 + (24 - 20)^2 + (20 - 20)^2 + (16 - 20)^2 + (20 - 20)^2}{5} = \frac{16 + 16}{5} = \\ &= 6,4 \text{ UM}^2\end{aligned}$$

- Família III:

$$\begin{aligned}\sigma^2 &= \frac{(12 - 20)^2 + (28 - 20)^2 + (24 - 20)^2 + (20 - 20)^2 + (16 - 20)^2}{5} = \\ &= \frac{64 + 64 + 16 + 16}{5} = 32 \text{ UM}^2\end{aligned}$$

- Família IV:

$$\begin{aligned}\sigma^2 &= \frac{(36 - 20)^2 + (32 - 20)^2 + (20 - 20)^2 + (8 - 20)^2 + (4 - 20)^2}{5} = \\ &= \frac{256 + 144 + 144 + 256}{5} = 160 \text{ UM}^2\end{aligned}$$

Exemplo:

Na situação considerada na introdução do estudo da variância (ver p. 126), o desvio padrão dos gastos de cada família é dado por:

- família I: $\sigma^2 = 0 \Rightarrow \sigma = 0 \text{ UM}$
- família II: $\sigma^2 = 6,4 \text{ UM}^2 \Rightarrow \sigma = \sqrt{6,4 \text{ UM}^2} \cong 2,53 \text{ UM}$
- família III: $\sigma^2 = 32 \text{ UM}^2 \Rightarrow \sigma = \sqrt{32 \text{ UM}^2} \cong 5,66 \text{ UM}$
- família IV: $\sigma^2 = 160 \text{ UM}^2 \Rightarrow \sigma = \sqrt{160 \text{ UM}^2} \cong 12,65 \text{ UM}$

Outra expressão para variância e desvio padrão

$$\sigma^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$$

$$\sigma^2 = \frac{\sum_{i=1}^n (x_i^2 - 2 \cdot x_i \cdot \bar{x} + \bar{x}^2)}{n} = \frac{\sum_{i=1}^n x_i^2 - 2 \cdot \bar{x} \cdot \sum_{i=1}^n x_i + n \cdot \bar{x}^2}{n}$$

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

$$\sigma^2 = \frac{1}{n} \cdot \left[\sum_{i=1}^n x_i^2 - 2 \cdot \frac{\sum_{i=1}^n x_i}{n} \cdot \sum_{i=1}^n x_i + n \cdot \frac{\left(\sum_{i=1}^n x_i \right)^2}{n^2} \right] \Rightarrow$$

$$\Rightarrow \sigma^2 = \frac{1}{n} \cdot \left[\sum_{i=1}^n x_i^2 - \frac{2}{n} \left(\sum_{i=1}^n x_i \right)^2 + \frac{1}{n} \cdot \left(\sum_{i=1}^n x_i \right)^2 \right]$$

$$\sigma^2 = \frac{1}{n} \cdot \left[\sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i \right)^2}{n} \right]$$

$$\sigma = \sqrt{\frac{1}{n} \cdot \left[\sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i \right)^2}{n} \right]},$$

Exemplo:

Os dados seguintes referem-se aos gastos mensais com ônibus e metrô (expressos em reais) que um estudante universitário tem durante um semestre:

$$42 - 50 - 54 - 48 - 56 - 59$$

Aplicando a expressão anterior para o cálculo do desvio padrão (σ) das despesas, temos:

$$\sum_{i=1}^6 x_i = 42 + 50 + 54 + 48 + 56 + 59 = 309$$

e

$$\sum_{i=1}^6 x_i^2 = 42^2 + 50^2 + 54^2 + 48^2 + 56^2 + 59^2 = 16101$$

$$\sigma^2 = \frac{1}{n} \cdot \left[\sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i \right)^2}{n} \right]$$

$$\sigma = \sqrt{\frac{1}{n} \cdot \left[\sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i \right)^2}{n} \right]}.$$

Daí:

$$\sigma^2 = \frac{1}{6} \cdot \left[16101 - \frac{309^2}{6} \right] = 31,25 \text{ (reais)}^2 \Rightarrow \sigma \approx 5,59 \text{ reais}$$

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$