Capítulo 2 | Estatística descritiva

Exemplo

Número de clientes aguardando atendimento em *call center* a cada 10 minutos.

4 4 1 2 9 9 9 9 10 7 8 8 11 14 20 23 22 21 24 25 29 29 28 27 13 12 20 16 17 20 20 18 16 15 115

Exemplo

Número de clientes aguardando atendimento em *call center* a cada 10 minutos.

4 4 **1** 2 9 9 9 9 10 7 8 8 11 14 20 23 22 21 24 25 **29** **29** 28 27 13 12 20 16 17 20 20 18 16 15 11 5

Máximos, mínimos,

Média, moda etc.

Exemplo

Número de clientes aguardando atendimento em *call center* a cada 10 minutos.

4 4 **1** 2 9 9 9 9 10 7 8 8 11 14 20 23 22 21 24 25 **29** **29** 28 27 13 12 20 16 17 20 20 18 16 15 11 5

Máximos, mínimos,

Média, moda etc.

Descrição do capítulo

- 2.1 Distribuição da frequência e seus gráficos
- 2.2 Mais gráficos e representações
- 2.3 Medidas de tendência central
- 2.4 Medidas de variação
- 2.5 Medidas de posição

Seção 2.1

Distribuição de frequência e seus gráficos

Distribuição de frequência

Distribuição de frequência

- Tabela que mostra **classes** ou **intervalos** de dados com uma contagem do número de entradas em cada classe
- A **frequência, *f*,** de uma classe é o número de entradas de dados na classe

Classe	Frequência, *f*
1 – 5	5
6 – 10	8
11 – 15	6
16 – 20	8
21 – 25	5
26 – 30	4

Tamanho da classe 6 – 1 = 5

Limite inferior da classe

Limite superior da classe

Construindo uma distribuição de frequência

1. Decida o número de classes.
   - Geralmente entre 5 e 20; do contrário, pode ser difícil detectar padrões
2. Encontre o tamanho da classe.
   - Determine a variação dos dados
   - Divida a variação pelo número de classes
   - *Arredonde para cima para o próximo número conveniente*

3. Encontre os limites da classe.
   - Você pode usar a entrada de menor valor como o limite inferior da primeira classe
   - Encontre os limites inferiores remanescentes (adicione o tamanho da classe ao limite inferior da classe precedente)
   - Encontre o limite superior da primeira classe. Lembre-se de que as classes não podem ter limites iguais
   - Encontre os limites superiores remanescentes

4. Faça um registro para cada entrada de dados na fileira da classe apropriada.
5. Conte os registros para encontrar a frequência total *f* para cada classe.

Exemplo: construindo uma distribuição de frequência

A amostra seguinte lista o número de minutos que 50 usuários da internet passaram conectados durante a sessão mais recente. Construa uma distribuição de frequência para as sete classes.

50	40	41	17	11	7	22	44	28	21	19	23	37	51	54	42	86
41	78	56	72	56	17	7	69	30	80	56	29	33	46	31	39	20
18	29	34	59	73	77	36	39	30	62	54	67	39	31	53	44	

Solução: construindo uma distribuição de frequência

50	40	41	17	11	7	22	44	28	21	19	23	37	51	54	42	86
41	78	56	72	56	17	7	69	30	80	56	29	33	46	31	39	20
18	29	34	59	73	77	36	39	30	62	54	67	39	31	53	44	

1. Número de classes = 7 (dados)
2. Encontre o tamanho da classe

$$\frac{\text{max} - \text{min}}{\#\text{classes}} = \frac{86 - 7}{7} \approx 11.29$$

Arredondando para cima: 12

3. Use 7 (valor mínimo) como o primeiro limite mínimo. Adicione o tamanho da classe, 12, para definir o limite mínimo da próxima classe.

   $7 + 12 = 19$

   Encontre os limites mínimos restantes.

Tamanho da classe = 12

Limite mínimo	Limite máximo
7	
19	
31	
43	
55	
67	
79	

O limite máximo da primeira classe é 18 (um a menos que o limite mínimo da segunda classe).

Some o tamanho da classe, 12, para definir o limite máximo da próxima classe.

$$18 + 12 = 30$$

Encontre os limites máximos restantes.

Limite mínimo	Limite máximo
7	18
19	30
31	42
43	54
55	66
67	78
79	90

Tamanho da classe = 12

4. Faça um registro para cada entrada de dado na fileira da classe apropriada.
5. Conte os registros para encontrar a frequência total *f* para cada classe.

Classe	Registro	Frequência, *f*
7 – 18	~~IIII~~ I	6
19 – 30	~~IIII~~ ~~IIII~~	10
31 – 42	~~IIII~~ ~~IIII~~ III	13
43 – 54	~~IIII~~ III	8
55 – 66	~~IIII~~	5
67 – 78	~~IIII~~ I	6
79 – 90	II	2

$\Sigma f = 50$

Determinando o ponto médio

Ponto médio de uma classe

$$\frac{(\text{Limite mínimo da classe}) + (\text{Limite máximo da classe})}{2}$$

Classe	Ponto médio	Frequência, f
7 – 18	$\frac{7+18}{2} = 12.5$	6
19 – 30	$\frac{19+30}{2} = 24.5$	10
31 – 42	$\frac{31+42}{2} = 36.5$	13

Tamanho da classe = 12

Larson/Farber 4th ed.

Determinando a frequência relativa

Frequência relativa de uma classe

- Porção da porcentagem dos dados que se encaixa em um classe em particular
- $$\text{Frequência relativa} = \frac{\text{Frequência da classe}}{\text{Tamanho da amostragem}} = \frac{f}{n}$$

Classe	Frequência, f	Frequência relativa
7 – 18	6	$\frac{6}{50} = 0.12$
19 – 30	10	$\frac{10}{50} = 0.20$
31 – 42	13	$\frac{13}{50} = 0.26$

Frequência acumulada de uma classe

A soma das frequências daquela classe e de todas as classes anteriores.

Classe	Frequência, *f*	Frequência acumulada
7 – 18	6	6
19 – 30	+ 10	16
31 – 42	+ 13	29

Distribuição de frequência expandida

Classe	Frequência, f	Ponto médio	Frequência relativa	Frequência acumulada
7 – 18	6	12.5	0.12	6
19 – 30	10	24.5	0.20	16
31 – 42	13	36.5	0.26	29
43 – 54	8	48.5	0.16	37
55 – 66	5	60.5	0.10	42
67 – 78	6	72.5	0.12	48
79 – 90	2	84.5	0.04	50

$\Sigma f = 50$ $\quad\quad$ $\sum \frac{f}{n} = 1$

Gráficos de distribuição de frequência

Histograma de frequência

- Um gráfico de barras que representa a distribuição da frequência
- O eixo horizontal é quantitativo e mede os valores dos dados
- O eixo vertical mede as frequências das classes
- Barras consecutivas precisam se tocar

Fronteiras de classes

- Os números que separam as classes sem formar espaços entre elas
- A distância do limite superior da primeira classe para o limite inferior da segunda é 19 – 18 = 1
- A metade dessa distância é 0,5

Classe	Fronteiras de classes	Frequência, f
7 – 18	6.5 – 18.5	6
19 – 30		10
31 – 42		13

- Fronteira inferior da primeira classe = 7 – 0.5 = 6.5
- Fronteira superior da primeira classe = 18 + 0.5 = 18.5

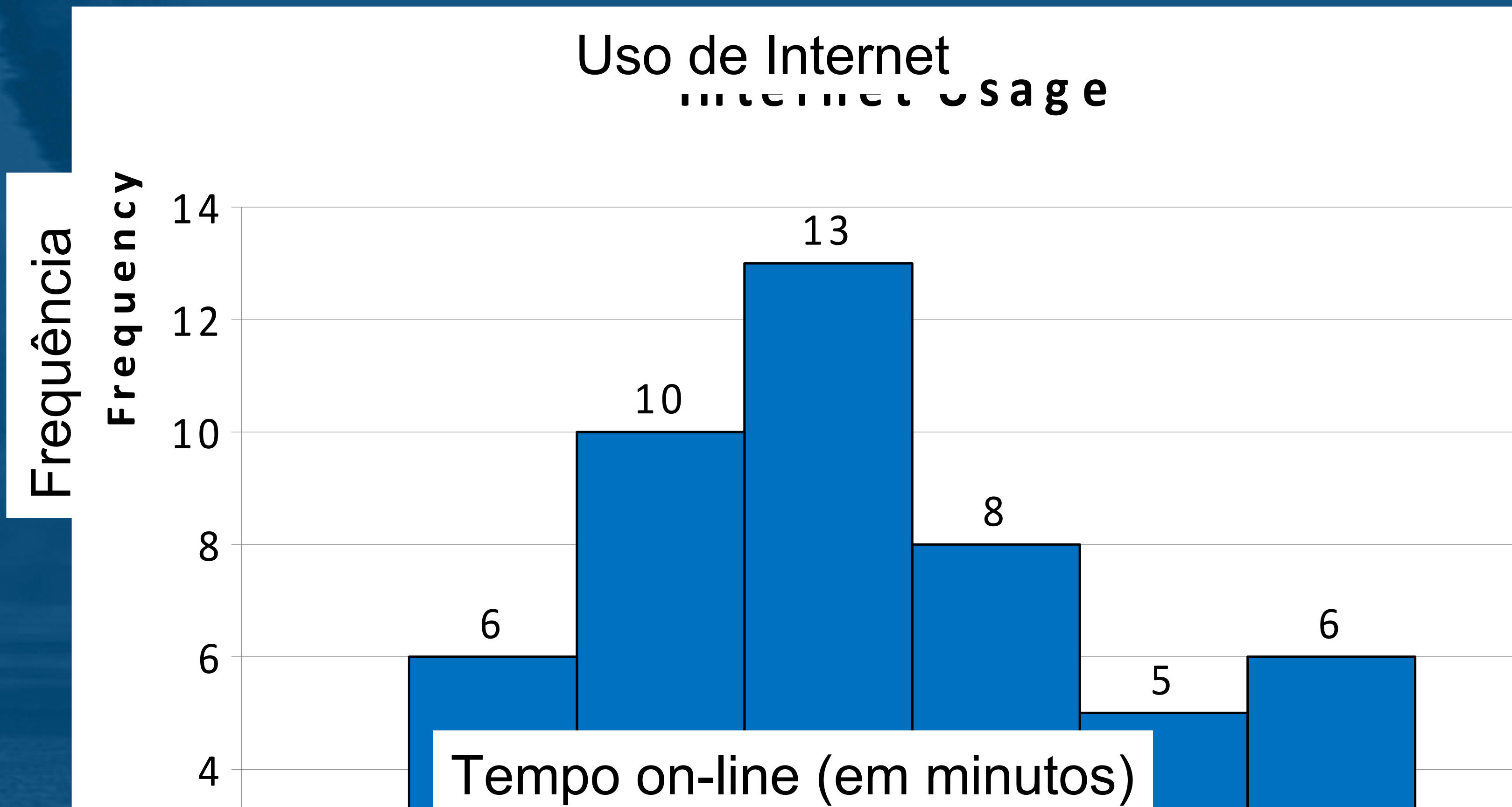
Classe	Fronteiras de classes	Frequência, f
7 – 18	6.5 – 18.5	6
19 – 30	18.5 – 30.5	10
31 – 42	30.5 – 42.5	13
43 – 54	42.5 – 54.5	8
55 – 66	54.5 – 66.5	5
67 – 78	66.5 – 78.5	6
79 – 90	78.5 – 90.5	2

Exemplo: histograma de frequência

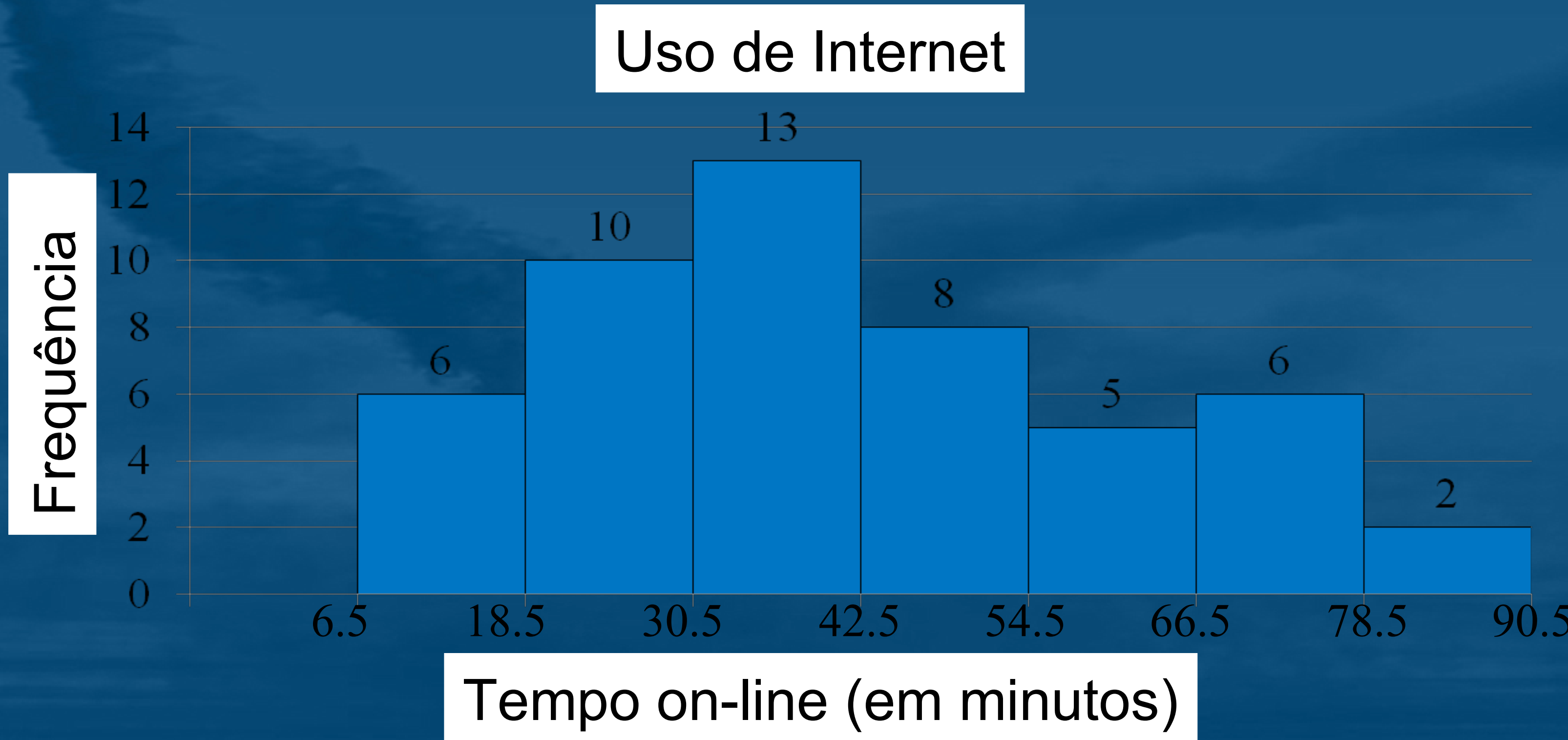
Construa um histograma de frequência para a distribuição da frequência do uso da internet.

Classe	Fronteiras de classes	Ponto médio	Frequência, f
7 – 18	6.5 – 18.5	12.5	6
19 – 30	18.5 – 30.5	24.5	10
31 – 42	30.5 – 42.5	36.5	13
43 – 54	42.5 – 54.5	48.5	8
55 – 66	54.5 – 66.5	60.5	5
67 – 78	66.5 – 78.5	72.5	6
79 – 90	78.5 – 90.5	84.5	2

Solução: histograma de frequência (usando pontos médios)



Solução: histograma de frequência (usando fronteiras de classes)



É visível que mais da metade dos assinantes passaram entre 19 e 54 minutos na Internet em sua sessão mais recente.

Gráficos de distribuições de frequência

Polígono de frequência

Um gráfico em linha que enfatiza a mudança contínua nas frequências.

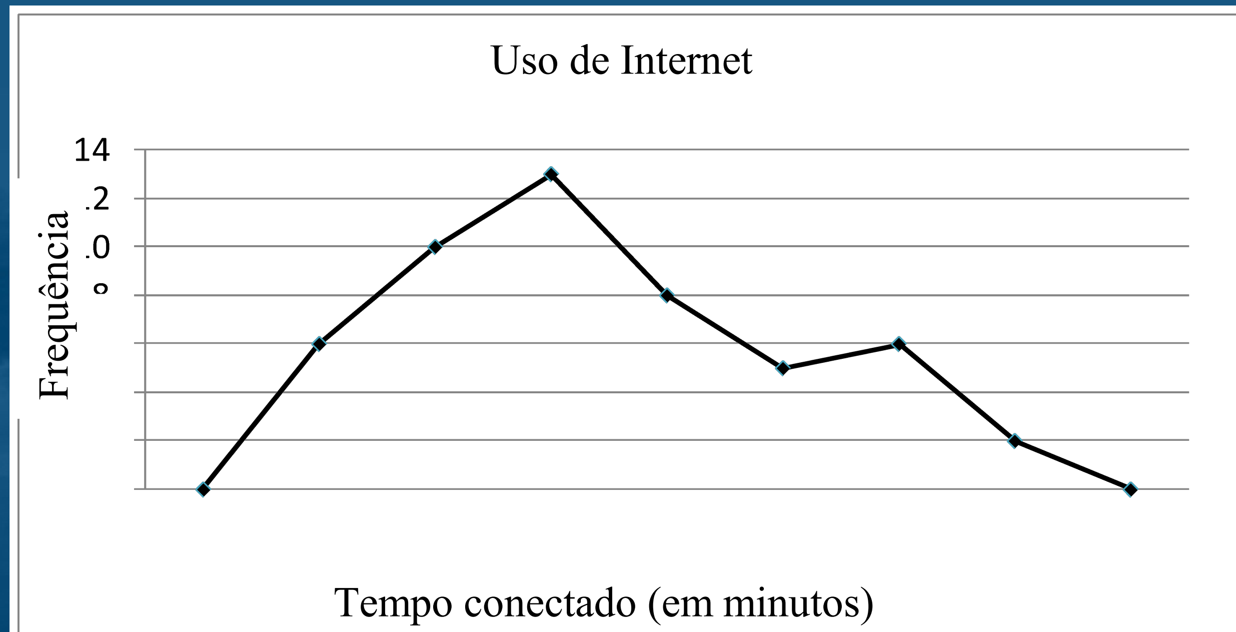
Exemplo: polígono de frequência

Construa um polígono de frequência para a distribuição da frequência do uso de Internet.

Classe	Ponto médio	Frequência, f
7 – 18	12.5	6
19 – 30	24.5	10
31 – 42	36.5	13
43 – 54	48.5	8
55 – 66	60.5	5
67 – 78	72.5	6
79 – 90	84.5	2

Solução: polígono de frequência

O gráfico deve começar e terminar no eixo horizontal, então estenda o lado esquerdo até o tamanho de uma classe antes do ponto médio da primeira classe e estenda o lado direito até o tamanho de uma classe depois do ponto médio da última classe.



Pode-se perceber que a frequência dos assinantes aumenta até 36,5 minutos e então diminui.

Gráficos de distribuição de frequência

Histograma de frequência relativa

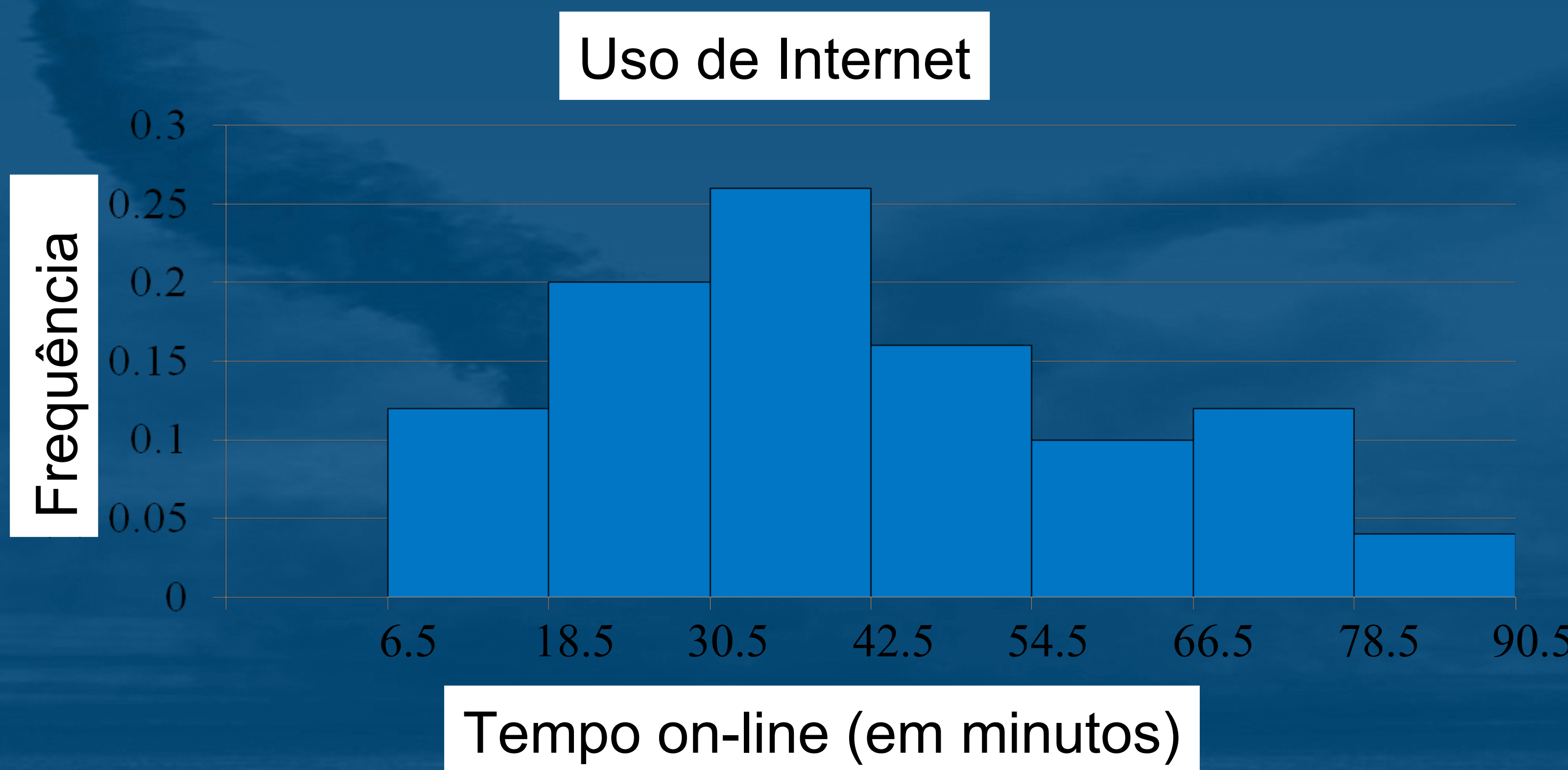
- Tem o mesmo formato e eixo horizontal que o histograma de frequência correspondente
- O eixo vertical mede as **frequências** relativas e não as frequências

Exemplo: histograma de frequência relativa

Construa um histograma de frequência relativa para a distribuição da frequência do uso de Internet.

Classe	Fronteiras de classes	Frequência, *f*	Frequência relativa
7 – 18	6.5 – 18.5	6	0.12
19 – 30	18.5 – 30.5	10	0.20
31 – 42	30.5 – 42.5	13	0.26
43 – 54	42.5 – 54.5	8	0.16
55 – 66	54.5 – 66.5	5	0.10
67 – 78	66.5 – 78.5	6	0.12
79 – 90	78.5 – 90.5	2	0.04

Solução: histograma de frequência relativa



A partir do gráfico, pode-se perceber que 20% dos assinantes de Internet passaram entre 18,5 minutos e 30,5 minutos conectados.

Gráficos de distribuição de frequências

Gráfico de frequências cumulativas, frequência acumulada ou ogiva

- Gráfico de linhas que demonstra a frequência cumulativa de cada classe em sua fronteira superior
- As fronteiras superiores são demarcadas no eixo horizontal
- As frequências cumulativas são demarcadas no eixo vertical

Construindo uma ogiva

1. Construa uma distribuição de frequência que inclua frequências cumulativas como uma das colunas.
2. Especifique os eixos horizontal e vertical.
   - O eixo horizontal consiste nas fronteiras superiores das classes
   - O eixo vertical mede as frequências cumulativas
3. Pontos que representam as fronteiras superiores das classes e suas frequências cumulativas correspondentes.

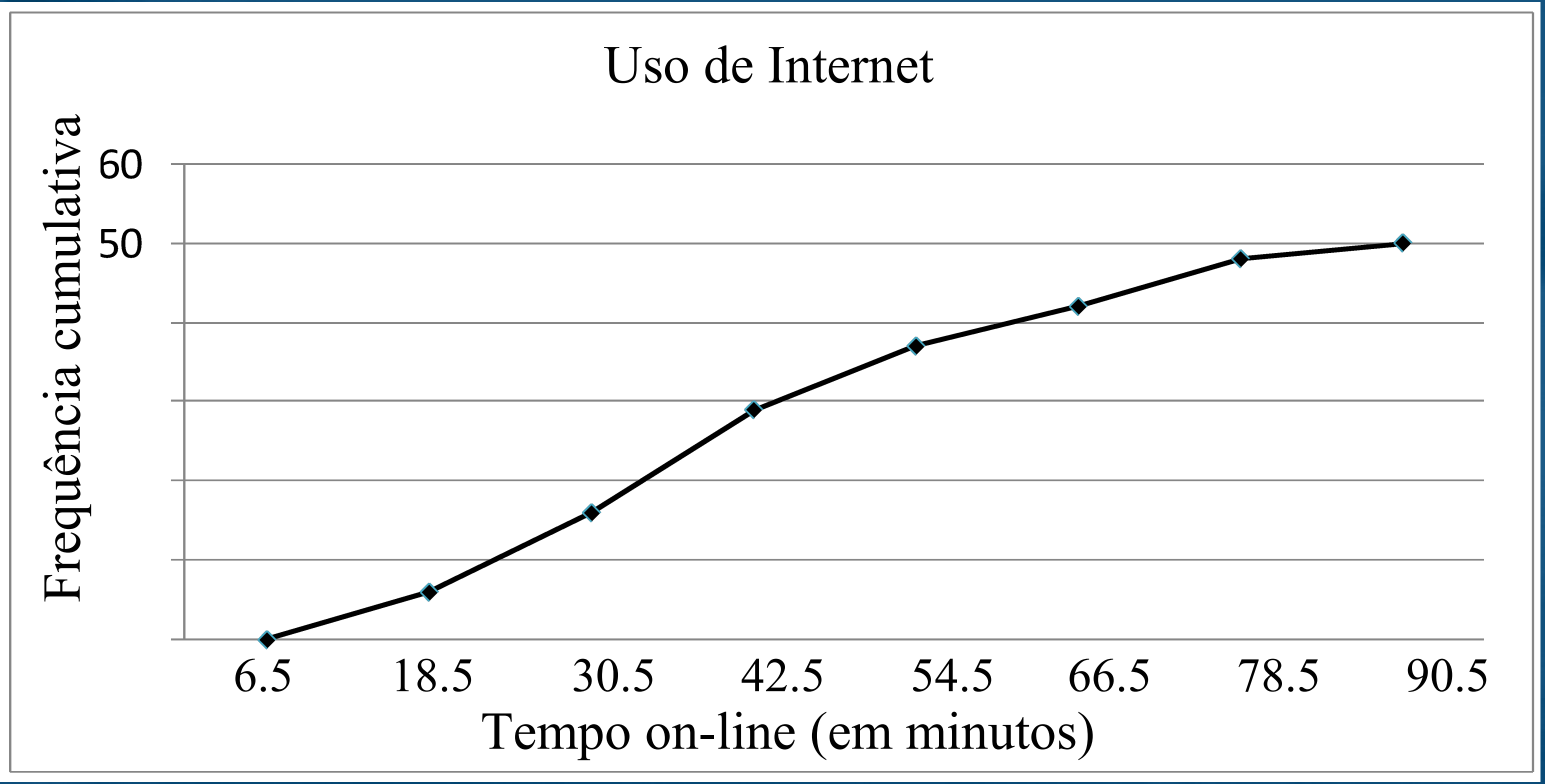
4. Conecte os pontos da esquerda com os da direita.
5. O gráfico deve começar na fronteira inferior da primeira classe (frequência cumulativa é zero) e deve terminar na fronteira superior da última classe (frequência cumulativa igual ao tamanho da amostra).

Exemplo: ogiva

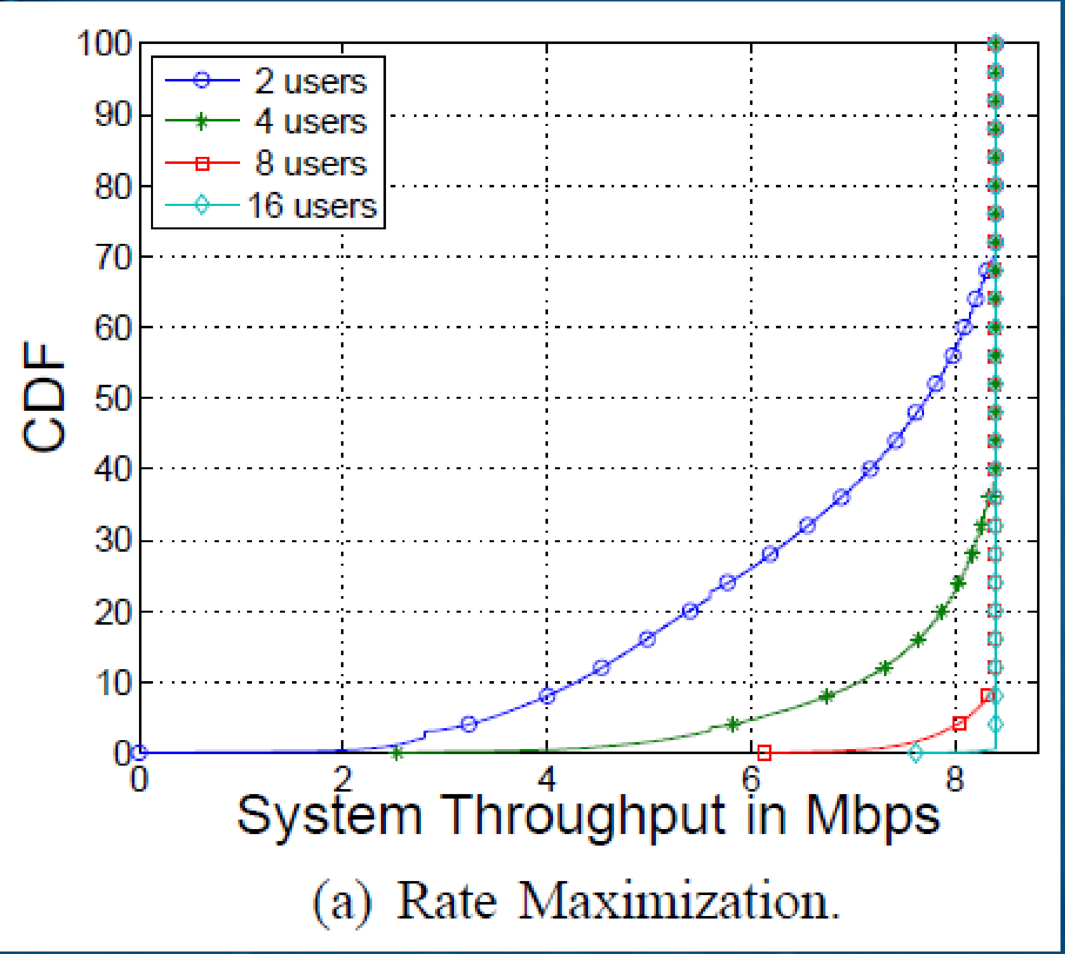
Construir uma ogiva para a distribuição de frequência do uso da Internet.

Classe	Fronteiras de classes	Frequência, *f*	Frequência cumulativa
7 – 18	6.5 – 18.5	6	6
19 – 30	18.5 – 30.5	10	16
31 – 42	30.5 – 42.5	13	29
43 – 54	42.5 – 54.5	8	37
55 – 66	54.5 – 66.5	5	42
67 – 78	66.5 – 78.5	6	48
79 – 90	78.5 – 90.5	2	50

Solução: ogiva



A partir da ogiva, pode-se perceber que aproximadamente 40 assinantes passaram 60 minutos ou menos conectados em sua última sessão. O maior aumento no uso ocorre entre 30,5 minutos e 42,5 minutos.



Fonte: Maciel et al. *"Scheduling-strategies for Coordinated Multi-Point Systems"*. SBRT`09.

Seção 2.2

Mais gráficos e representações

Fazendo gráficos de conjuntos de dados quantitativos (1)

Diagrama de ramos-e-folhas

- Cada número é separado em um **ramo** e uma **folha**
- Similar a um histograma
- Ainda contém os valores dos dados originais

26

Dados: 21, 25, 25, **26**, 27,
28, 30, 36, 36, 45

2	1	5	5	6	7	8
3	0	6	6			
4	5					

Exemplo: construindo um diagrama de ramos-e-folhas

Os números seguintes representam a quantidade de mensagens de texto enviadas no mês passado pelos usuários de celulares em um andar do dormitório da faculdade. Mostre os dados em um diagrama de ramos-e-folhas.

155	159	144	129	105	145	126	116	130	114	122	112	112	142	126
156	118	108	122	121	109	140	126	119	113	117	118	109	109	119
139	139	122	78	133	126	123	145	121	134	124	119	132	133	124
129	112	126	148	147										

Solução: construindo um diagrama de ramos-e-folhas

155	159	144	129	105	145	126	116	130	114	122	112	112	142	126
156	118	108	122	121	109	140	126	119	113	117	118	109	109	119
139	139	122	78	133	126	123	145	121	134	124	119	132	133	124
129	112	126	148	147										

- As entradas de dados variam do baixo 78 até um alto 159
- Use o dígito mais à direita como a folha
  - Por exemplo,
    78 = 7 | 8 e 159 = 15 | 9
- Liste os ramos, 7 até 15, à esquerda de uma linha vertical
- Para cada entrada de dados, liste uma folha à direita de seu ramo

Número de mensagens de texto enviadas

Chave: 5|5 = 155

7	8
8	
9	
10	5 8 9 9 9
11	6 4 2 2 8 8 9 3 7 8 9 9 2
12	9 6 2 6 2 1 6 2 6 3 1 4 4 9 6
13	0 9 9 3 4 2 3
14	4 5 2 0 5 8 7
15	5 9

Diagrama de ramos-e-folhas não ordenado

Número de mensagens de texto enviadas

Chave: 15|5 = 155

Inclui uma chave para identificar os valores dos dados.

7	8
8	
9	
10	5 8 9 9 9
11	2 2 2 3 4 6 7 8 8 8 9 9 9
12	1 1 2 2 2 3 4 4 6 6 6 6 6 9 9
13	0 2 3 3 4 9 9
14	0 2 4 5 5 7 8
15	5 9

Diagrama de ramos-e-folhas ordenado

Pode-se concluir pelos dados acima que mais de 50% dos usuários de celular enviaram entre 110 e 130 mensagens de texto.

Fazendo gráficos de conjuntos de dados quantitativos (2)

Diagrama de pontos

- Cada entrada de dados é posta usando um ponto acima de um eixo horizontal

Dados: 21, 25, 25, **26**, 27, 28, 30, 36, 36, 45

Exemplo: construindo um diagrama de pontos

Use um diagrama de pontos para organizar os dados das mensagens de texto.

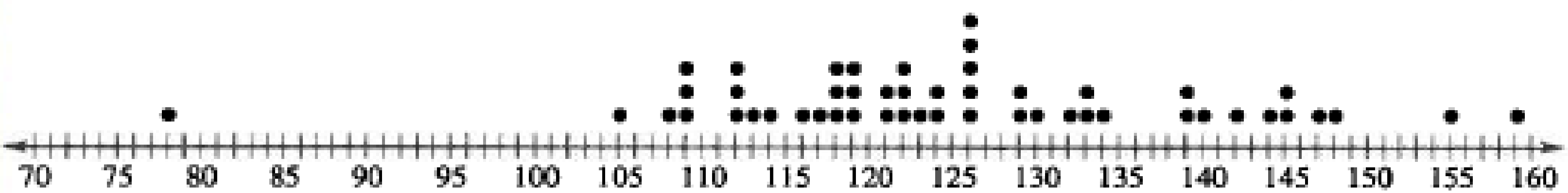
155	159	144	129	105	145	126	116	130	114	122	112	112	142	126
156	118	108	122	121	109	140	126	119	113	117	118	109	109	119
139	139	122	78	133	126	123	145	121	134	124	119	132	133	124
129	112	126	148	147										

- Para que cada entrada seja incluída no diagrama de pontos, o eixo horizontal deve incluir números entre 70 e 160
- Para reresentar uma entrada, insira um ponto acima da posição da entrada no eixo
- Se uma entrada for repetida, faça outro ponto acima do ponto anterior

Solução: construindo um diagrama de pontos

155	159	144	129	105	145	126	116	130	114	122	112	112	142	126
156	118	108	122	121	109	140	126	119	113	117	118	109	109	119
139	139	122	78	133	126	123	145	121	134	124	119	132	133	124
129	112	126	148	147										

Número de mensagens de texto enviadas

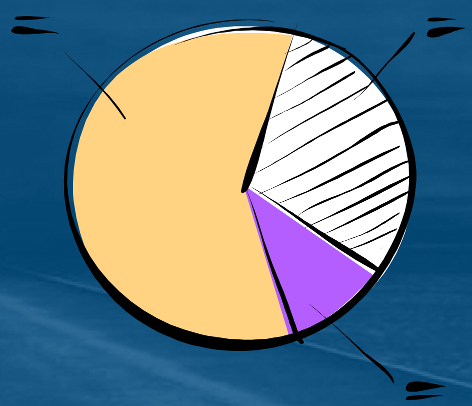


Maioria dos valores se agrupam entre 105 e 148;
Valor que mais ocorre é 126;
Pode-se notar também que 78 é um valor incomum.

Fazendo gráficos de conjunto de dados qualitativos (1)

Gráfico de pizza

- Um círculo é dividido em vários setores, que representam categorias
- A área de cada setor é proporcional à frequência de cada categoria



Exemplo: construindo um gráfico de pizza

O número de ocupantes de veículos motorizados mortos em acidentes em 2005 é exibido na tabela. Use um gráfico de pizza para organizar os dados. (*Fonte: U.S. Department of Transportation, National Highway Traffic Safety Administration.*)

Tipo de veículo	Mortes
Carros	18.440
Caminhões	13.778
Motocicletas	4.553
Outros	823

Solução: construindo um gráfico de pizza

- Encontre a frequência relativa de cada categoria.

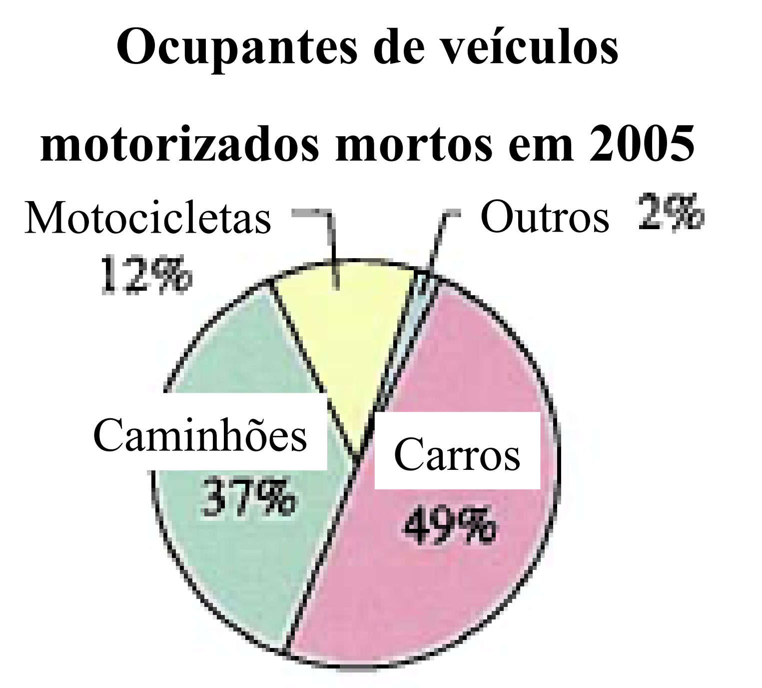
Tipo de Veículo	Frequência, f	Frequência relativa
Carros	18.440	$\frac{18440}{37594} \approx 0.49$
Caminhões	13.778	$\frac{13778}{37594} \approx 0.37$
Motocicletas	4.553	$\frac{4553}{37594} \approx 0.12$
Outros	823	$\frac{823}{37594} \approx 0.02$

- Construa o gráfico de pizza usando o ângulo central que corresponda à cada categoria.
  - Para encontrar o ângulo central, multiplique 360º pela frequência relativa da categoria.
  - Por exemplo, o ângulo central para carros é

$$360(0{,}49) \approx 176^{\circ}$$

Tipo de veículo	Frequência, f	Frequência relativa	Ângulo central
Carros	18.440	0,49	$360^\circ(0{,}49) \approx 176^\circ$
Caminhões	13.778	0,37	$360^\circ(0{,}37) \approx 133^\circ$
Motocicletas	4.553	0,12	$360^\circ(0{,}12) \approx 43^\circ$
Outros	823	0,02	$360^\circ(0{,}02) \approx 7^\circ$

Tipo de veículo	Frequência relativa	Ângulo central
Carros	0,49	176º
Caminhões	0,37	133º
Motocicletas	0,12	43º
Outros	0,02	7º

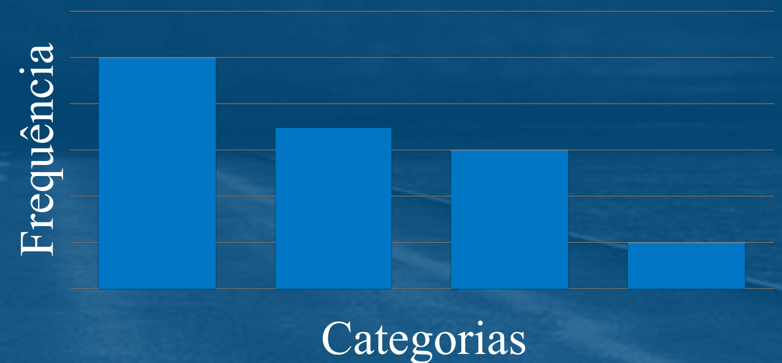


Conclui-se que a maioria dos acidentes fatais em veículos automotivos foram aqueles envolvendo carros.

Fazendo gráficos de conjunto de dados qualitativos (2)

Gráfico de Pareto

- Barras verticais com a altura representando uma frequência ou uma frequência relativa
- As barras são posicionadas por ordem decrescente de altura: barra mais alta posicionada à esquerda

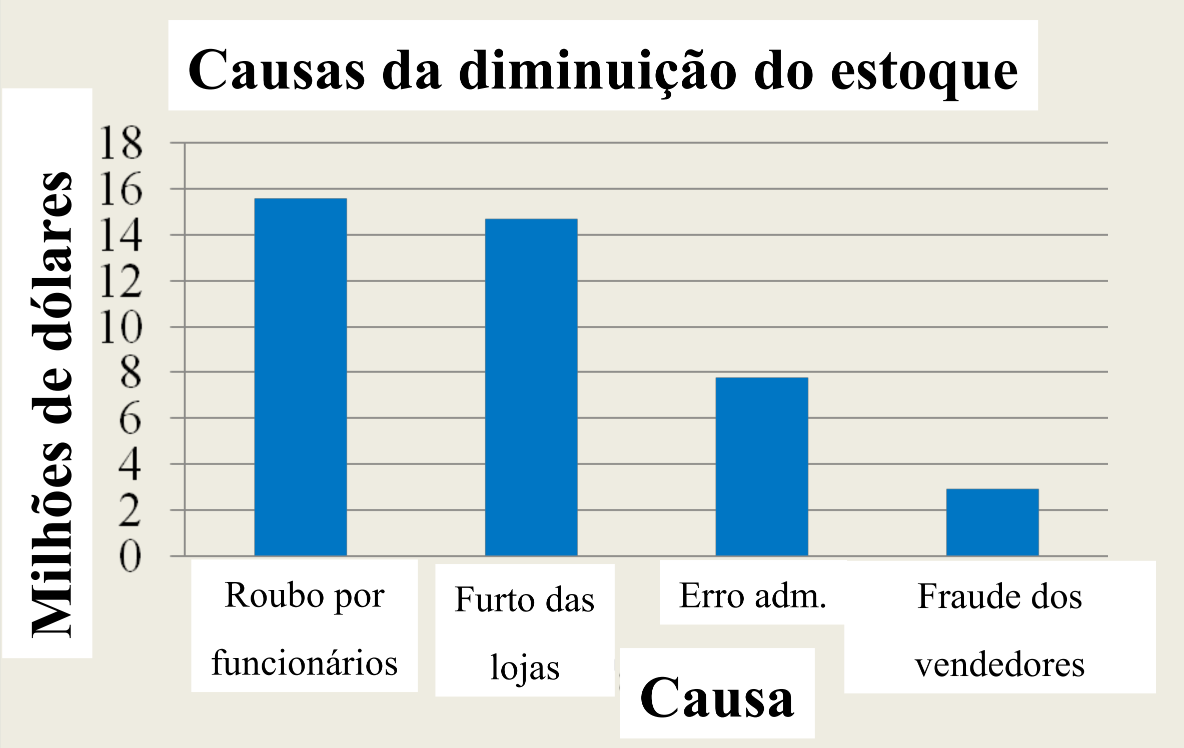


Exemplo: construindo um gráfico de Pareto

Recentemente, a indústria de varejo perdeu US$ 41 mi com redução nos estoques. A redução de estoque é uma perda de estoque por meio de quebra, roubo de carga, roubo em lojas e assim por diante. As causas da redução de estoque são erro administrativo (7,8 mi), roubo por funcionários (15,6 mi), furto das lojas (14,7 mi) e fraude dos vendedores (2,9 mi). Se você fosse um varejista, para qual causa de redução de estoque você olharia primeiro? (*Fonte: National Retail Federation and Center for Retailing Education, University of Florida*.)

Solução: construindo um gráfico de Pareto

Causa	US$ (em milhões)
Erro adm.	7,8
Roubo por funcionários	15,6
Furto das lojas	14,7
Fraude dos vendedores	2,9

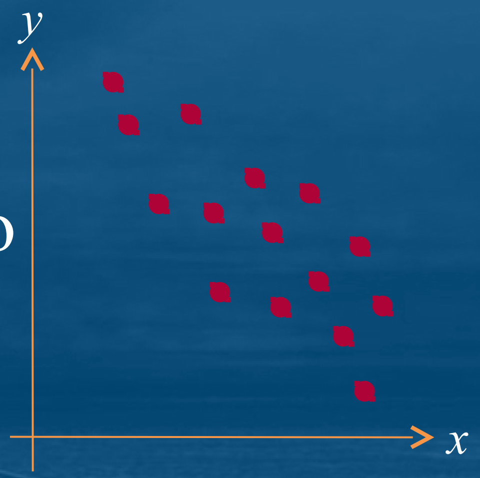


Pelo gráfico, é fácil ver que, das causas da diminuição do estoque, o roubo por funcionários deveria ser o primeiro a receber atenção.

Fazendo gráficos de conjunto de dados emparelhados (1)

Conjunto de dados emparelhados

- Cada entrada em um conjunto de dados corresponde à outra entrada em um segundo conjunto de dados
- Gráficos usam **um gráfico de dispersão**
  - Os pares ordenados são representados como pontos em um plano coordenado
  - Usado para representar a relação entre duas variáveis quantitativas



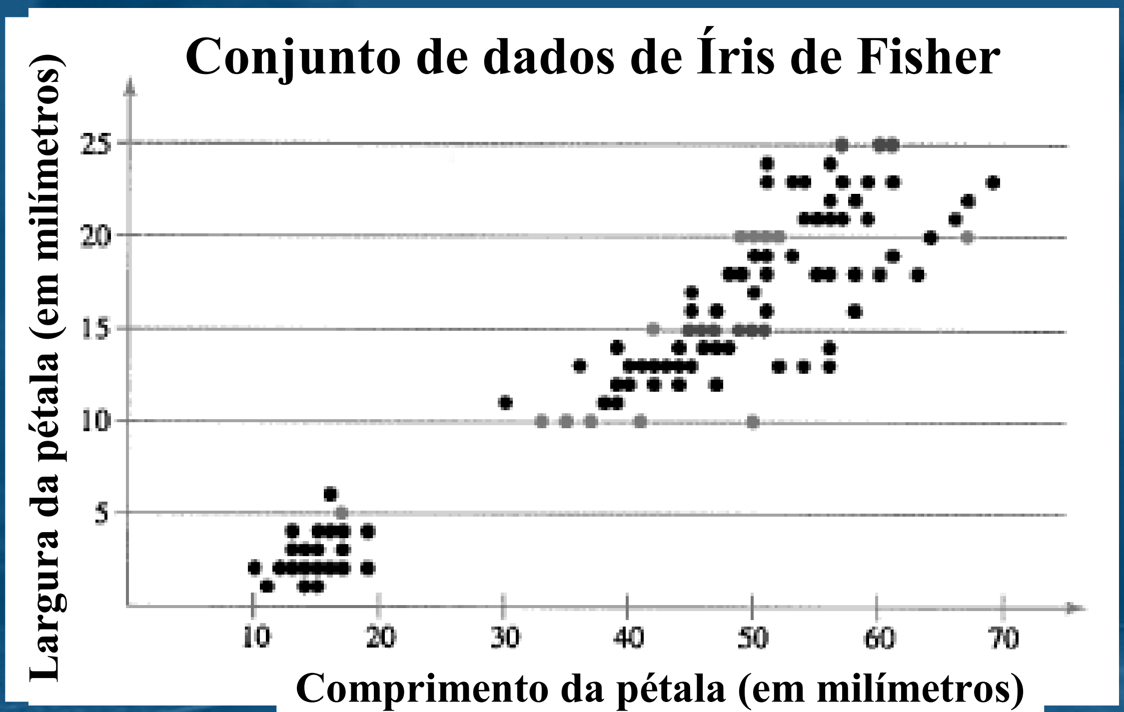
Exemplo: interpretando um gráfico de dispersão

O conjunto de dados de *Íris de Fisher* descreve várias características físicas para três espécies de íris, como o comprimento de pétalas e a sua largura (em mm).

No gráfico de dispersão mostrado, os comprimentos de pétalas formam o primeiro conjunto de dados e as larguras formam o segundo conjunto de dados.

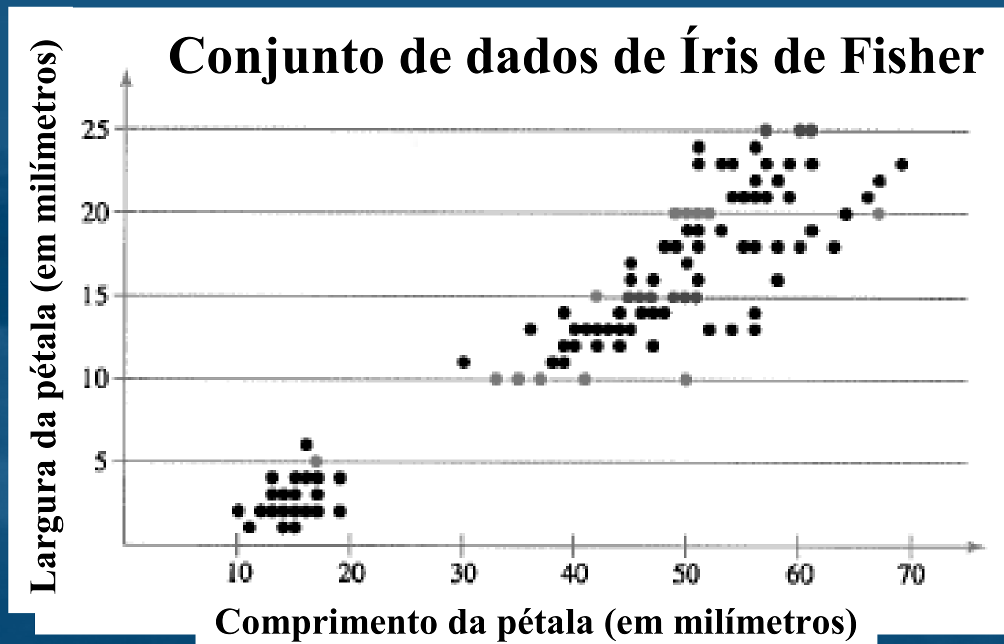
(*Fonte: Fisher, R. A., 1936*.)

Conforme o comprimento da pétala aumenta, o que tende a acontecer com a largura?



Cada ponto no esquema disperso representa o comprimento e a largura da pétala de uma flor.

Solução: interpretando um gráfico de dispersão



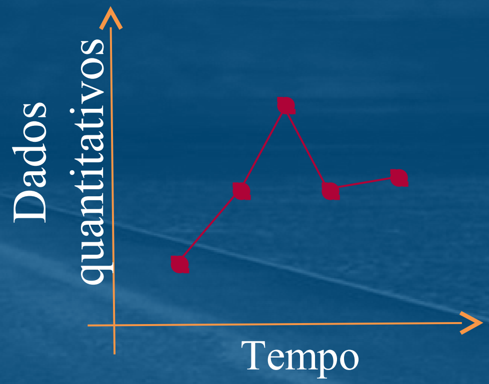
Interpretação

Nota-se que conforme o comprimento da pétala aumenta, a largura da pétala também tende a aumentar.

Fazendo gráficos de conjunto de dados

Série temporal

- Conjuntos de dados são compostos de entradas quantitativas tomadas em intervalos regulares em um período de tempo
    - Por exemplo, a quantidade de precipitações medidas a cada dia por um mês
- Usa um gráfico de **períodos de tempo**



Exemplo: construindo um gráfico de série temporal

A tabela lista o número de telefones celulares (em milhões) para os anos de 1995 até 2005.

Construa um gráfico de série temporal do número de telefones celulares.

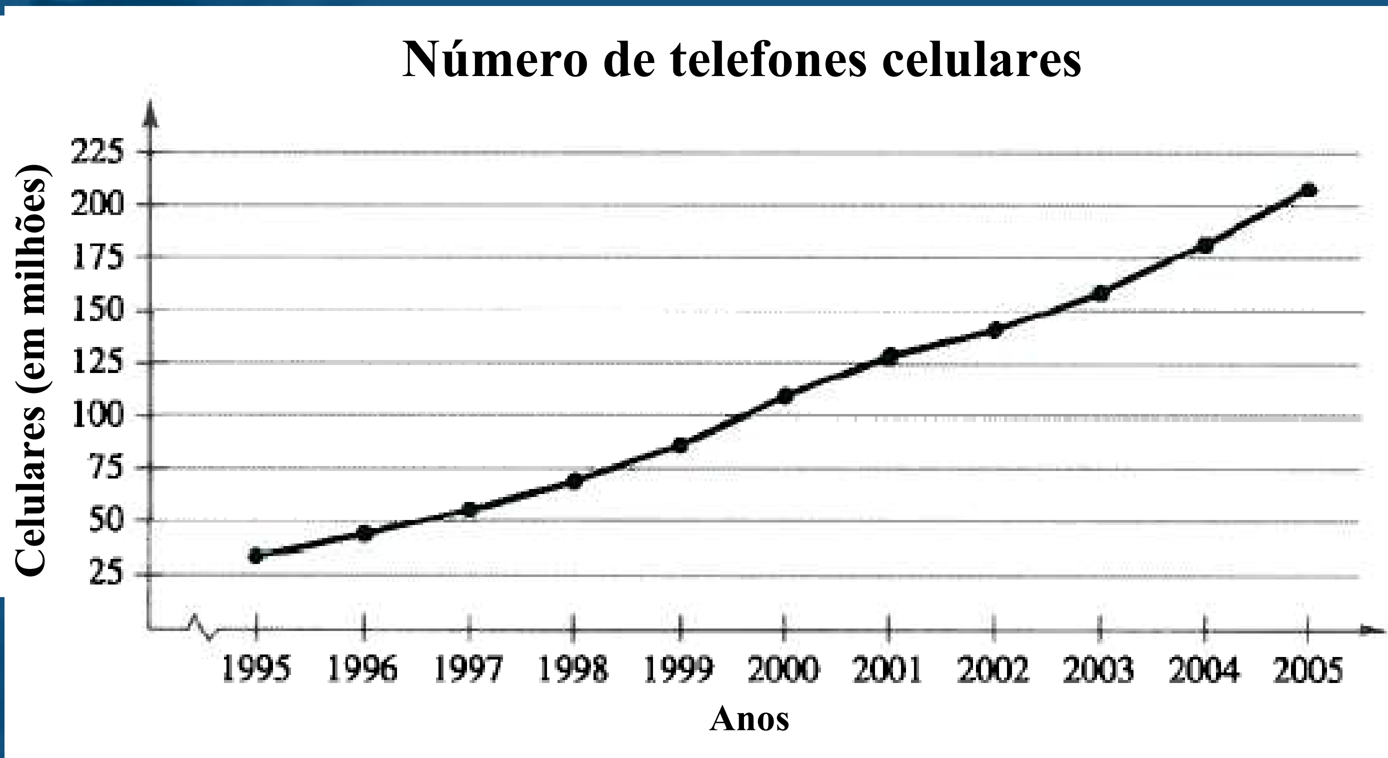
(*Fonte: Cellular Telecommunication & Internet Association*.)

Ano	Usuários (em milhões)
1995	33.8
1996	44.0
1997	55.3
1998	69.2
1999	86.0
2000	109.5
2001	128.4
2002	140.8
2003	158.7
2004	182.1
2005	207.9

Solução: construindo um gráfico de série temporal

- Faça com que o eixo horizontal represente os anos
- Deixe o eixo vertical representar o número de celulares (milhões)
- Marque os dados emparelhados e conecte-os com os segmentos de linha

Ano	Usuários (em milhões)
1995	33.8
1996	44.0
1997	55.3
1998	69.2
1999	86.0
2000	109.5
2001	128.4
2002	140.8
2003	158.7
2004	182.1
2005	207.9



Ano	Usuários (em milhões)
1995	33.8
1996	44.0
1997	55.3
1998	69.2
1999	86.0
2000	109.5
2001	128.4
2002	140.8
2003	158.7
2004	182.1
2005	207.9

O gráfico mostra que o número de celulares tem aumentado desde 1995, com aumentos ainda mais significativos recentemente.

Seção 2.3

Medidas de tendência central

Medidas de tendência central

- Um valor que revela o que é típico ou central para um conjunto de dados
- Medidas de tendência central mais comuns:
  - Média
  - Mediana
  - Moda

Medidas de tendência central: média

Média

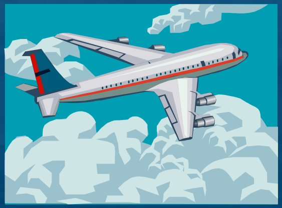
- A soma de todas as entradas de dados divididas pelo número de entradas
- **Notação sigma**: Σx = adicione todas as entradas (x) no conjunto de dados
- **Média populacional:** $\mu = \dfrac{\Sigma x}{N}$
- **Média amostral:** $\bar{x} = \dfrac{\Sigma x}{n}$

Exemplo: encontrando a média da amostra

Os preços (em dólares) para uma amostra de viagens feitas de Chicago, Illinois, para Cancun, México, são listados.

Qual o preço médio dos voos?

872   432   397   427   388   782   397



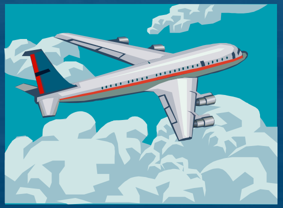
Solução: encontrando a média da amostra

872 432 397 427 388 782 397

- A soma dos preços dos voos é

$\Sigma x = 872 + 432 + 397 + 427 + 388 + 782 + 397 = 3.695$

- Para encontrar o preço médio, divida a soma dos preços pelo número de preços na amostra



$$\bar{x} = \frac{\Sigma x}{n} = \frac{3695}{7} \approx 527.9$$

O preço médio dos voos é cerca de $ 527,90.

Medidas de tendência central: mediana

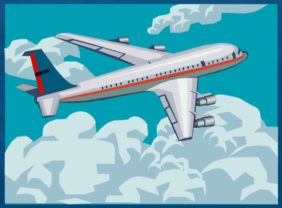
Mediana

- O valor que está no meio dos dados quando o conjunto dos dados é **ordenado**
- Mede o centro de um conjunto de dados ordenado dividindo-o em duas partes iguais
- Se conjunto de dados possui um número de entradas:
  - **ímpar**: mediana é o elemento do meio
  - **par**: mediana é a média dos 2 elementos centrais

Exemplo: encontrando a mediana

Os preços (em dólares) para uma amostra de viagens feitas de Chicago, Illinois, para Cancun, México, são listados. Encontre a mediana.

872 432 397 427 388 782 397



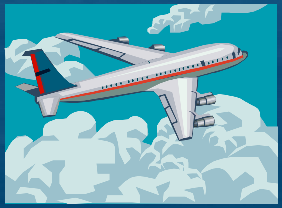
Solução: encontrando a mediana

872 432 397 427 388 782 397

- Primeiramente, ordene os dados

  388 397 397 427 432 782 872

- Existem sete entradas (um número ímpar), e a mediana é o elemento central, ou o quarto, do conjunto de dados

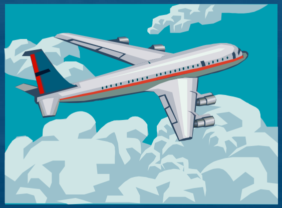


O preço mediano dos voos é $ 427.

Exemplo: encontrando a mediana

O preço dos voos em $ 432 não está mais disponível. Qual o preço mediano dos voos restantes?

872 397 427 388 782 397



Solução: encontrando a mediana

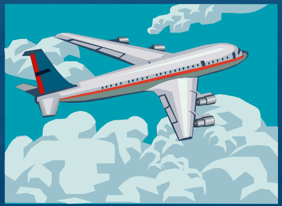
872 397 427 388 782 397

- Primeiramente, ordene os dados

388 397 $\underbrace{397 \quad 427}$ 782 872

- Há seis elementos (um número par), a mediana é a média das duas entradas centrais.

$$\text{Mediano} = \frac{397 + 427}{2} = 412$$



O preço mediano dos voos é $ 412.

Medidas de tendência central: moda

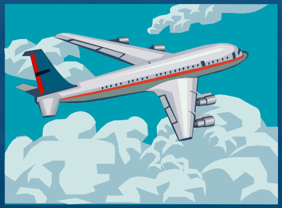
Moda

- A entrada que ocorre com a maior frequência
- Se não houver entradas repetidas, o conjunto de dados não tem moda
- Se duas entradas ocorrem com a mesma e mais alta frequência, cada entrada é um moda (**bimodal**)

Exemplo: encontrando a moda

Os preços (em dólares) para uma amostra de viagens feitas de Chicago, Illinois, para Cancun, México, são listados. Encontre a moda dos preços dos voos.

872 432 397 427 388 782 397



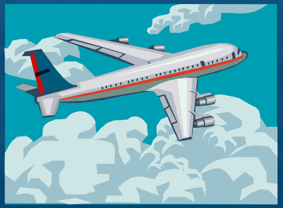
Solução: encontrando a moda

872 432 397 427 388 782 397

- Ordenar os dados ajuda a encontrar a moda

388 397 397 427 432 782 872

- A entrada de 397 ocorre duas vezes, enquanto as outras ocorrem somente uma vez.

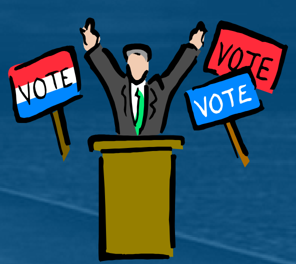


A moda dos preços dos voos é $ 397.

Exemplo: encontrando a moda

Em um debate político norte-americano uma amostra de membros da audiência foi questionada à respeito de seus partidos políticos. Suas respostas estão na tabela. Qual a moda de suas respostas?

Partido político	**Frequência, *f***
Democrata	34
Republicano	56
Outros	21
Não responderam	9



Solução: encontrando a moda

Partido político	Frequência, f
Democrata	34
Republicano	56
Outros	21
Não responderam	9

A moda é Republicano (a resposta com maior ocorrência). Nessa amostra havia mais republicanos do que pessoas de qualquer outro partido político.

Comparando a média, a mediana e a moda

- Todas as três medidas descrevem uma entrada típica de um conjunto de dados
- Vantagens de usar a média:
  - A média é uma medida confiável porque leva em conta cada entrada do conjunto de dados
- Desvantagens de usar a média:
  - Muito afetada por valores discrepantes (uma entrada que é muito distante das outras entradas no conjunto de dados)

Exemplo: comparando a média, a mediana e a moda

Encontre a média, a mediana e a moda da amostra de idades de uma classe. Qual medida de tendência central descreve melhor uma entrada típica desse conjunto de dados? Existe algum valor discrepante?

Idades em uma classe						
20	20	20	20	20	20	21
21	21	21	22	22	22	23
23	23	23	24	24	65	

Solução: comparando a média, a mediana e a moda

Idades em uma classe						
20	20	20	20	20	20	21
21	21	21	22	22	22	23
23	23	23	24	24	65	

Média: $\bar{x} = \dfrac{\Sigma x}{n} = \dfrac{20 + 20 + \ldots + 24 + 65}{20} \approx 23.8$ anos

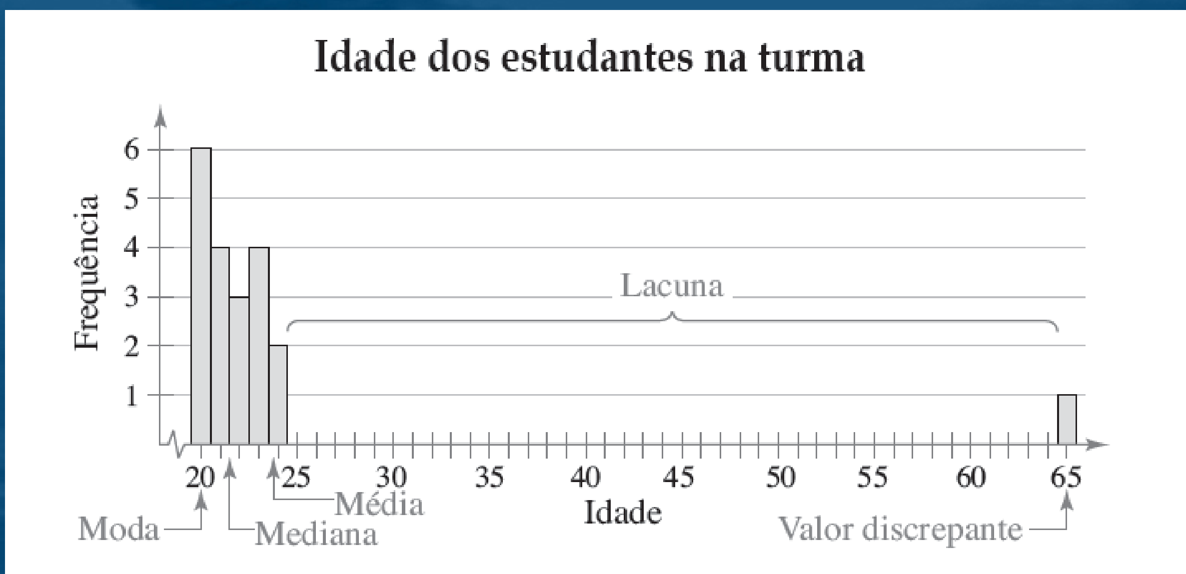
Mediana: $\dfrac{21 + 22}{2} = 21.5$ anos

Moda: 20 anos (a entrada que ocorre com a maior frequência)

Média ≈ 23,8 anos Mediana = 21,5 anos Moda = 20 anos

- A média leva todas as entradas em consideração, mas é influenciada pelo valor discrepante 65
- A mediana também leva todas as entradas em consideração, e não é afetada pelo valor discrepante
- Nesse caso a moda existe, mas não parece representar uma entrada típica

Algumas vezes uma comparação gráfica pode ajudar a decidir qual medida de tendência central melhor representa um conjunto de dados.



Nesse caso, parece que a **mediana** é o que melhor descreve o conjunto de dados.

Média ponderada

- A média de um conjunto de dados cujas entradas possuem pesos variantes
- $\bar{x} = \dfrac{\Sigma\,(x \times w)}{\Sigma\, w}$ em que w é o peso de cada entrada x

Exemplo: encontrando a média ponderada

Você está frequentando uma aula na qual sua nota é determinada com base em 5 fontes: 50% da média de seu exame, 15% do seu exame bimestral, 20% de seu exame final, 10% de seu trabalho no laboratório de informática e 5% de seus deveres de casa.

Suas notas são: 86 (média do exame), 96 (exame bimestral), 82 (exame final), 98 (laboratório) e 100 (dever de casa).

Qual é a média ponderada de suas notas?

Se a média mínima para um A é 90, você obteve uma nota A?

Solução: encontrando a média ponderada

Fonte	Notas, x	Peso, w	$x \cdot w$
Média do exame	86	0,50	86(0,50)= 43,0
Exame bimestral	96	0,15	96(0,15) = 14,4
Exame final	82	0,20	82(0,20) = 16,4
Laboratório	98	0,10	98(0,10) = 9,8
Dever de casa	100	0,05	100(0,05) = 5,0
		$\Sigma w = 1$	$\Sigma(x \cdot w) = 88{,}6$

$$\bar{x} = \frac{\Sigma\,(x \times w)}{\Sigma\, w} = \frac{88.6}{1} = 88.6$$

Sua média ponderada para essa aula foi 88,6. Você não tirou um A.

Média de dados agrupados

Média de uma distribuição de frequência

- Aproximada por

$$\bar{x} = \frac{\Sigma\,(x \times f)}{n} \qquad n = \Sigma f$$

em que x e f são, respectivamente, os pontos médios e as frequências de uma classe

Encontrando a média da distribuição de uma frequência

Em palavras	*Em símbolos*
1. Encontre o ponto médio de cada classe.	$x = \dfrac{(\text{Limite inferior}) + (\text{Limite superior})}{2}$
2. Encontre a soma dos produtos dos pontos médios e das frequências.	$\Sigma(x \times f)$
3. Encontre a soma das frequências.	$n = \Sigma f$
4. Encontre a média da distribuição das frequências.	$\bar{x} = \dfrac{\Sigma(x \times f)}{n}$

Exemplo: encontrando a média da distribuição de uma frequência

Use a distribuição de frequência para aproximar a média do número de minutos que uma amostra de internautas passou conectada em sua última sessão.

Classe	Ponto médio	Frequência, *f*
7 – 18	12,5	6
19 – 30	24,5	10
31 – 42	36,5	13
43 – 54	48,5	8
55 – 66	60,5	5
67 – 78	72,5	6
79 – 90	84,5	2

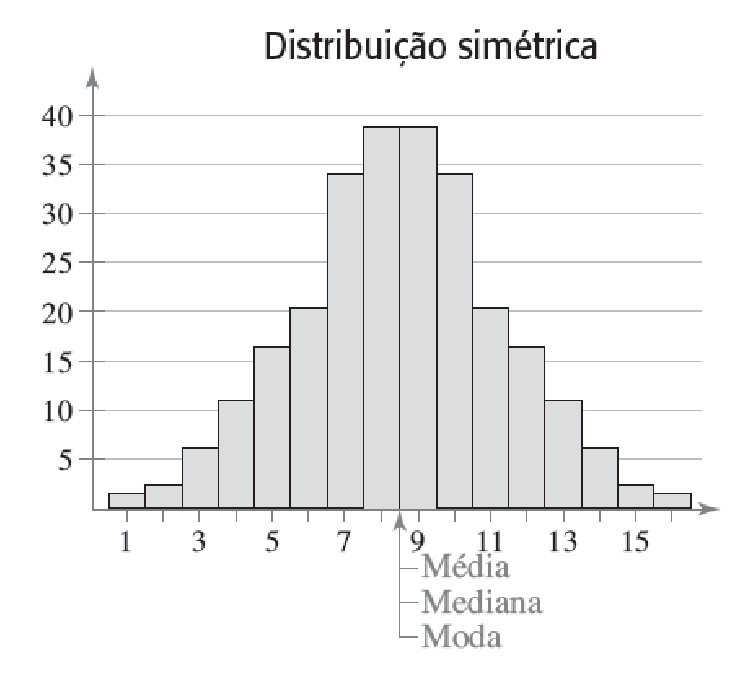
Classe	Ponto médio, x	Frequência, f	$(x \cdot f)$
7 – 18	12,5	6	12,5·6 = 75,0
19 – 30	24,5	10	24,5·10 = 245,0
31 – 42	36,5	13	36,5·13 = 474,5
43 – 54	48,5	8	48,5·8 = 388,0
55 – 66	60,5	5	60,5·5 = 302,5
67 – 78	72,5	6	72,5·6 = 435,0
79 – 90	84,5	2	84,5·2 = 169,0
		n = 50	$\Sigma(x \cdot f) = 2.089,0$

$$\bar{x} = \frac{\Sigma(x \times f)}{n} = \frac{2089}{50} \approx 41.8 \text{ minutos}$$

A forma das distribuições

Distribuição simétrica

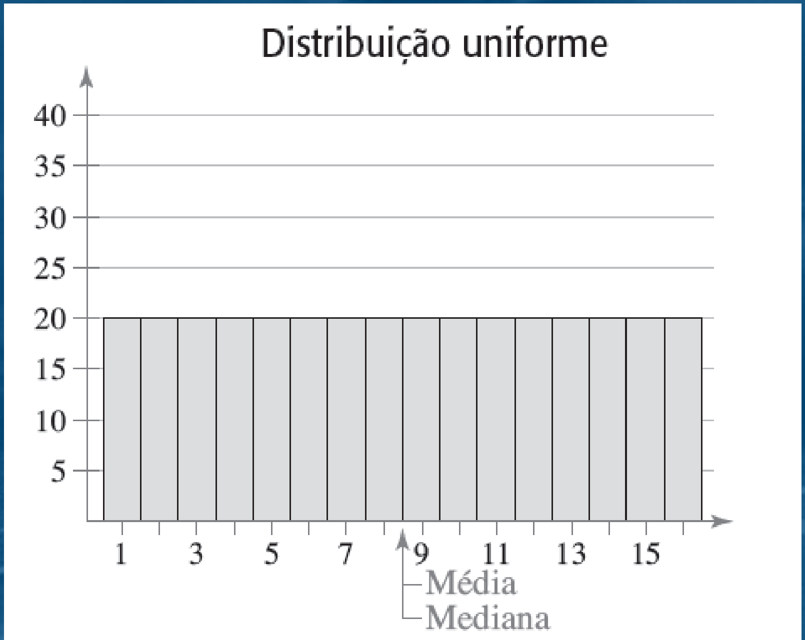
Uma linha vertical pode ser traçada do meio do gráfico de distribuição e as metades resultantes são quase idênticas.



O formato das distribuições

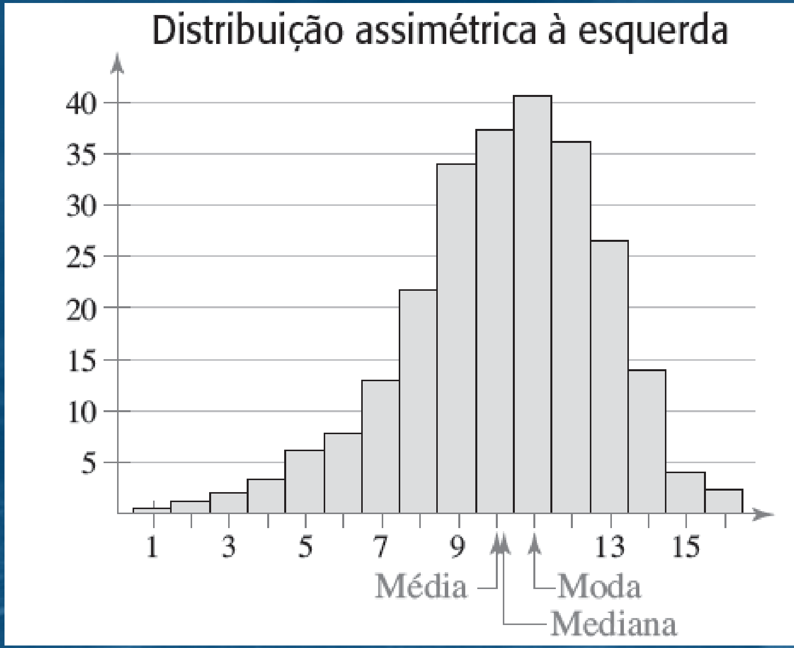
Distribuição uniforme (retangular)

- Todas as entradas têm frequências iguais ou quase iguais
- Simétrica



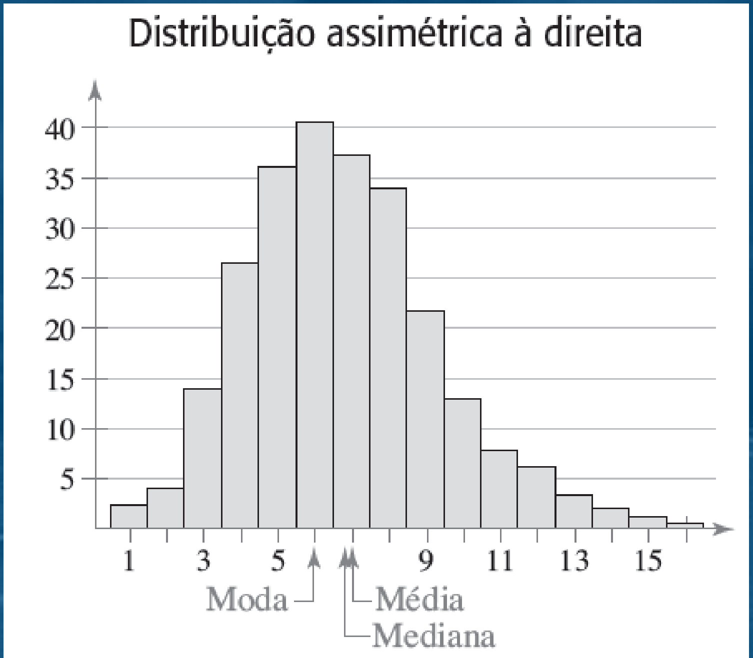
Distribuição assimétrica à esquerda (assimétrica negativamente)

- A “cauda” do gráfico se alonga mais à esquerda
- A média fica à esquerda da mediana



Distribuição assimétrica à direita (positivamente assimétrica)

- A "cauda" do gráfico se alonga mais à direita
- A média fica à direita da mediana



Seção 2.4

Medidas de variação

Variância

Variação

- A diferença entre as entradas máxima e mínima em um conjunto de dados
- Os dados precisam ser quantitativos
- Variação = (Entrada máx.) – (Entrada mín.)

Exemplo: encontrando a variação

Uma corporação contratou 10 graduados. Os salários (anuais) iniciais de cada um são demonstrados abaixo. Encontre a variação dos salários iniciais.

Salários iniciais (milhares de dólares)									
41	38	39	45	47	41	44	41	37	42

- Ordenar os dados ajuda a encontrar o menor e o maior salário

37   38   39   41   41   41   42   44   45   47

Mínimo   Máximo

- Variação = (Entrada máx.) – (Entrada mín.)

  = 47 – 37 = 10

A variação dos salários iniciais é igual a 10.

Desvio, variância e desvio padrão

Desvio

- A diferença entre a entrada de dados, *x*, e a média do conjunto de dados
- Conjunto de dados da população:
  - Desvio de $x = x - \mu$
- Conjunto de dados da amostra:
  - Desvio de $x = x - \bar{x}$

Exemplo: encontrando o desvio

Uma corporação contratou 10 graduados. Os salários iniciais de cada um são demonstrados abaixo. Encontre a variação dos salários iniciais.

Salários iniciais (milhares de dólares)									
41	38	39	45	47	41	44	41	37	42

Solução:

- Primeiro, determine a média dos salários iniciais.

$$\mu = \frac{\Sigma x}{N} = \frac{415}{10} = 41.5$$

Solução: encontrando o desvio

Determine o desvio para cada entrada.

Salário ($ 1.000s), x	Desvio: $x - \mu$
41	41 – 41,5 = –0,5
38	38 – 41,5 = –3,5
39	39 – 41,5 = –2,5
45	45 – 41,5 = 3,5
47	47 – 41,5 = 5,5
41	41 – 41,5 = –0,5
44	44 – 41,5 = 2,5
41	41 – 41,5 = –0,5
37	37 – 41,5 = –4,5
42	42 – 41,5 = 0,5
$\Sigma x = 415$	$\Sigma(x - \mu) = 0$

Analisando...

$$\Sigma(x - \mu) = \Sigma x - \Sigma\mu = N\mu - N\mu = 0$$

Ou seja, o desvio médio não captura qualquer informação de variação dos dados.

Desvio, variância e desvio padrão

Variância da população

- $$\sigma^2 = \frac{\Sigma(x-\mu)^2}{N}$$

Soma dos quadrados, SQ_x

Desvio padrão da população

- $$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\Sigma(x-\mu)^2}{N}}$$

Encontrando a variância populacional e o desvio padrão

Em palavras	*Em símbolos*
1. Encontre a média do conjunto de dados da população.	$\mu = \dfrac{\Sigma x}{N}$
2. Encontre o desvio de cada entrada.	$x - \mu$
3. Eleve os desvios ao quadrado.	$(x - \mu)^2$
4. Some para obter a soma dos quadrados.	$SS_x = \Sigma(x - \mu)^2$

Em palavras	*Em símbolos*
5. Divida por N para obter **a variância populacional**.	$\sigma^2 = \frac{\Sigma(x-\mu)^2}{N}$
6. Encontre a raiz quadrada para obter **o desvio padrão populacional**.	$\sigma = \sqrt{\frac{\Sigma(x-\mu)^2}{N}}$

Exemplo: encontrando o desvio padrão da população

Uma corporação contratou 10 graduados. Os salários iniciais de cada um são demonstrados abaixo. Encontre a variação dos salários iniciais.

Salários iniciais (milhares de dólares)

41 38 39 45 47 41 44 41 37 42

Lembrar $\mu = 41{,}5$.

Solução: encontrando o desvio padrão da população

- Determine SQ_x
- $N = 10$

Salário, x	Desvio: $x - \mu$	Quadrados: $(x - \mu)^2$
41	$41 - 41{,}5 = -0{,}5$	$(-0{,}5)^2 = 0{,}25$
38	$38 - 41{,}5 = -3{,}5$	$(-3{,}5)^2 = 12{,}25$
39	$39 - 41{,}5 = -2{,}5$	$(-2{,}5)^2 = 6{,}25$
45	$45 - 41{,}5 = 3{,}5$	$(3{,}5)^2 = 12{,}25$
47	$47 - 41{,}5 = 5{,}5$	$(5{,}5)^2 = 30{,}25$
41	$41 - 41{,}5 = -0{,}5$	$(-0{,}5)^2 = 0{,}25$
44	$44 - 41{,}5 = 2{,}5$	$(2{,}5)^2 = 6{,}25$
41	$41 - 41{,}5 = -0{,}5$	$(-0{,}5)^2 = 0{,}25$
37	$37 - 41{,}5 = -4{,}5$	$(-4{,}5)^2 = 20{,}25$
42	$42 - 41{,}5 = 0{,}5$	$(0{,}5)^2 = 0{,}25$
	$\Sigma(x - \mu) = 0$	$SS_x = 88{,}5$

Variância da população

- $\sigma^2 = \dfrac{\Sigma(x-\mu)^2}{N} = \dfrac{88.5}{10} \approx 8.9$

Desvio padrão da população

- $\sigma = \sqrt{\sigma^2} = \sqrt{8.85} \approx 3.0$

O desvio padrão da população é cerca de 3,0 ou $ 3.000.

Desvio, variância e desvio padrão

Variância da amostra

- $$s^2 = \frac{\Sigma (x - \bar{x})^2}{n - 1}$$

Desvio padrão da amostra

- $$s = \sqrt{s^2} = \sqrt{\frac{\Sigma (x - \bar{x})^2}{n - 1}}$$

Encontrando a variância e o desvio padrão da amostra

Em palavras	*Em símbolos*
1. Encontre a média do conjunto de dados da amostra.	$\bar{x} = \frac{\Sigma x}{n}$
2. Encontre o desvio de cada entrada.	$x - \bar{x}$
3. Eleve cada desvio ao quadrado.	$(x - \bar{x})^2$
4. Some-os para obter a soma dos quadrados.	$SS_x = \Sigma(x - \bar{x})^2$

Em palavras	*Em símbolos*
5. Divida por $n - 1$ para obter **a variância da amostra**.	$s^2 = \dfrac{\Sigma (x - \bar{x})^2}{n - 1}$
6. Encontre a raiz quadrada para obter **o desvio padrão da amostra**.	$s = \sqrt{\dfrac{\Sigma (x - \bar{x})^2}{n - 1}}$

É comum adotar uma representação alternativa para o cálculo da variância.

Observemos o numerador da variância populacional:

$$\Sigma(x - \mu)^2 = \Sigma(x^2 - 2x\mu + \mu^2) =$$
$$= \Sigma x^2 - \Sigma 2x\mu + \Sigma \mu^2 =$$
$$= \Sigma x^2 - 2\mu \Sigma x + \mu^2 \Sigma 1 =$$
$$= \Sigma x^2 - 2\mu N\mu + \mu^2 N =$$
$$= \Sigma x^2 - \mu^2 N$$

Exemplo: encontre o desvio padrão da amostra

Os salários inicias são para uma filial da empresa em Chicago. A empresa tem várias outras filiais e você planeja usar os salários iniciais de Chicago para estimar os salários iniciais da população maior. Encontre o desvio padrão dos salários iniciais da amostra.

Salários iniciais (milhares de dólares)

41 38 39 45 47 41 44 41 37 42

Solução: encontrando o desvio padrão da população

- Determine SQ_x
- $n = 10$

Salário, x	Desvio: $x - \mu$	Quadrados: $(x - \mu)^2$
41	$41 - 41{,}5 = -0{,}5$	$(-0{,}5)^2 = 0{,}25$
38	$38 - 41{,}5 = -3{,}5$	$(-3{,}5)^2 = 12{,}25$
39	$39 - 41{,}5 = -2{,}5$	$(-2{,}5)^2 = 6{,}25$
45	$45 - 41{,}5 = 3{,}5$	$(3{,}5)^2 = 12{,}25$
47	$47 - 41{,}5 = 5{,}5$	$(5{,}5)^2 = 30{,}25$
41	$41 - 41{,}5 = -0{,}5$	$(-0{,}5)^2 = 0{,}25$
44	$44 - 41{,}5 = 2{,}5$	$(2{,}5)^2 = 6{,}25$
41	$41 - 41{,}5 = -0{,}5$	$(-0{,}5)^2 = 0{,}25$
37	$37 - 41{,}5 = -4{,}5$	$(-4{,}5)^2 = 20{,}25$
42	$42 - 41{,}5 = 0{,}5$	$(0{,}5)^2 = 0{,}25$
	$\Sigma(x - \mu) = 0$	$SS_x = 88{,}5$

Variância da amostra

- $s^2 = \dfrac{\Sigma(x - \bar{x})^2}{n - 1} = \dfrac{88.5}{10 - 1} \approx 9.8$

Desvio padrão da amostra

- $s = \sqrt{s^2} = \sqrt{\dfrac{88.5}{9}} \approx 3.1$

O desvio padrão da amostra é de aproximadamente 3,1 ou $ 3.100.

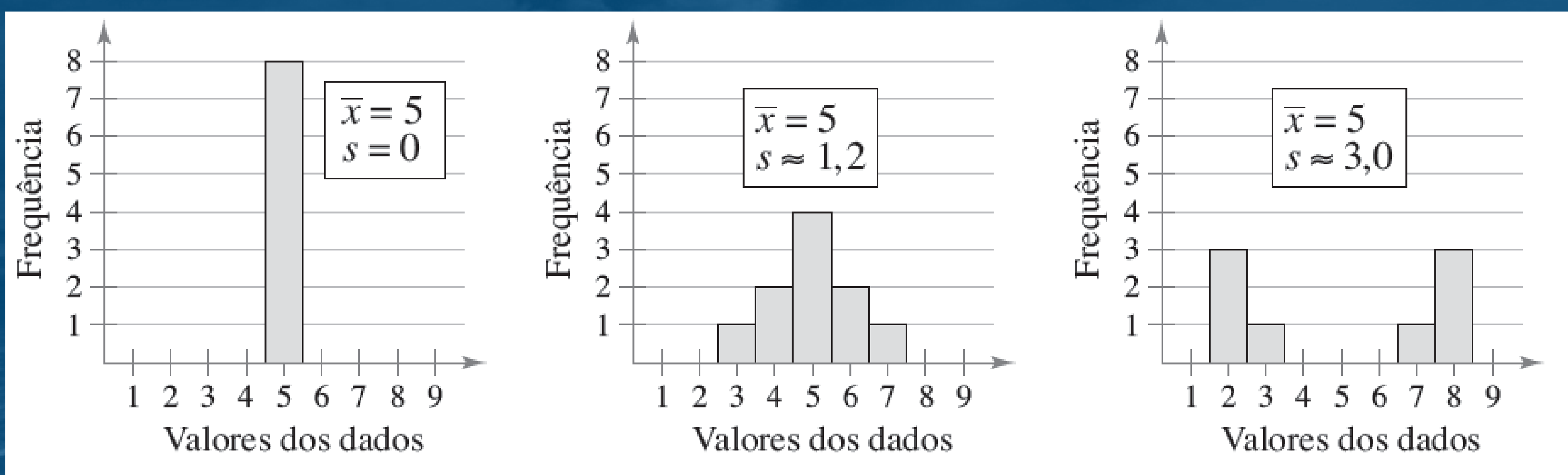
Exemplo: usando tecnologia para encontrar o desvio padrão

A amostra dos aluguéis de escritórios (em dólares por pé quadrado ao ano) no distrito comercial central de Miami é exibida na tabela. Use uma calculadora ou um computador para encontrar a média dos aluguéis e o desvio padrão da amostra
(Adaptado de: Cushman & Wakefield Inc.)

Preço dos aluguéis		
35,00	33,50	37,00
23,75	26,50	31,25
36,50	40,00	32,00
39,25	37,50	34,75
37,75	37,25	36,75
27,00	35,75	26,00
37,00	29,00	40,50
24,50	33,00	38,00

Interpretando o desvio padrão

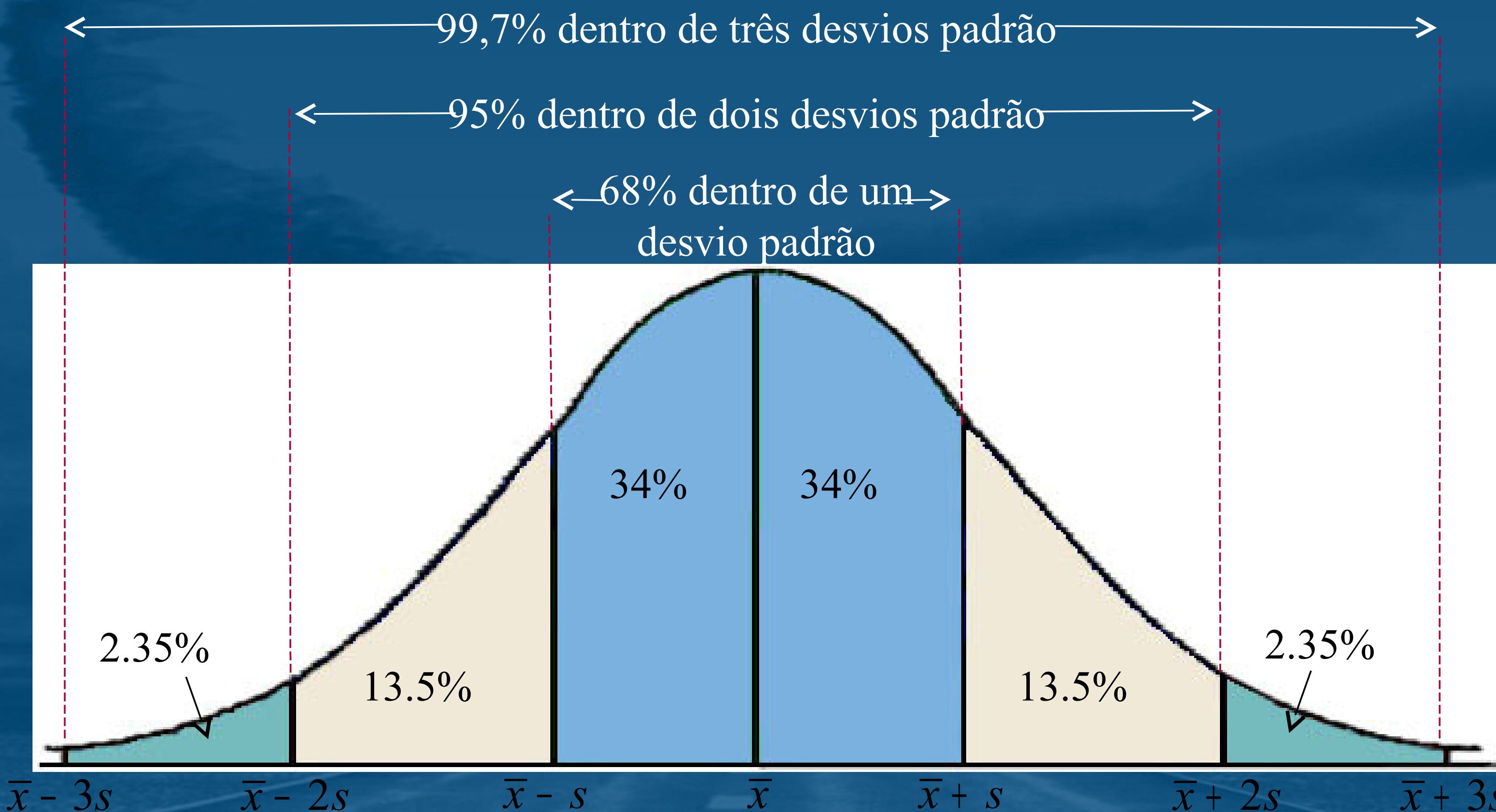
- Desvio padrão é a medida do valor típico que uma entrada desvia da média
- Quanto mais as entradas estão espalhadas, maior o desvio padrão



Interpretando desvio padrão: Regra Empírica (Regra 68 – 95 – 99.7)

Para dados com uma distribuição em formato de sino (simétrico), o desvio padrão tem as seguintes características:

- Cerca de **68%** dos dados estão dentro de um desvio padrão da média
- Cerca de **95%** dos dados estão dentro de dois desvios padrão da média
- Cerca de **99,7%** dos dados estão dentro de três desvios padrão da média

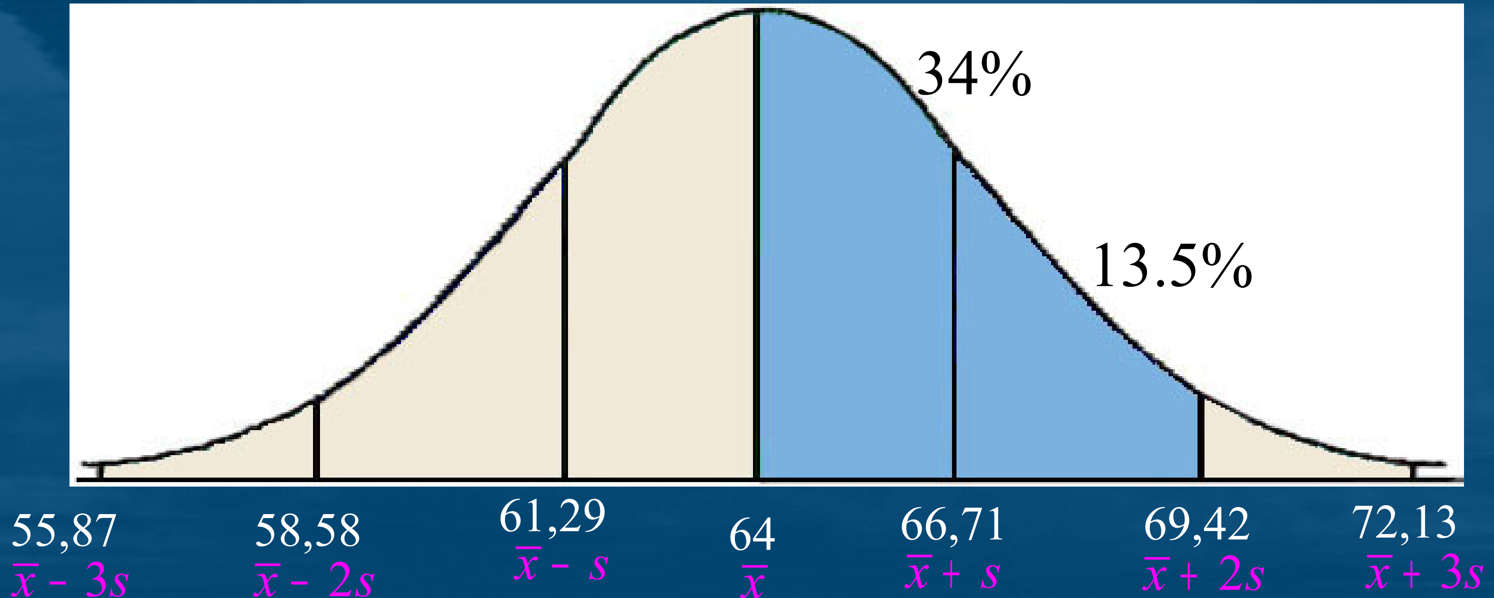

99,7% dentro de três desvios padrão
95% dentro de dois desvios padrão
68% dentro de um desvio padrão
34%
34%
13.5%
13.5%
2.35%
2.35%
$\bar{x} - 3s$
$\bar{x} - 2s$
$\bar{x} - s$
$\bar{x}$
$\bar{x} + s$
$\bar{x} + 2s$
$\bar{x} + 3s$

Exemplo: usando a Regra Empírica

Em uma pesquisa conduzida pelo National Center for Health Statistics, a amostragem média da estatura feminina nos Estados Unidos (20-29 anos) era de 64 polegadas, com um desvio padrão da amostragem de 2,71 polegadas. Estime a porcentagem de mulheres que estão entre 64 e 69,42 polegadas.

Solução: usando a Regra Empírica

- Porque a distribuição tem formato de sino, você pode usar a Regra Empírica.



34% + 13,5% = 47,5% das mulheres estão entre 64 e 69,42 polegadas.

Teorema de Chebychev

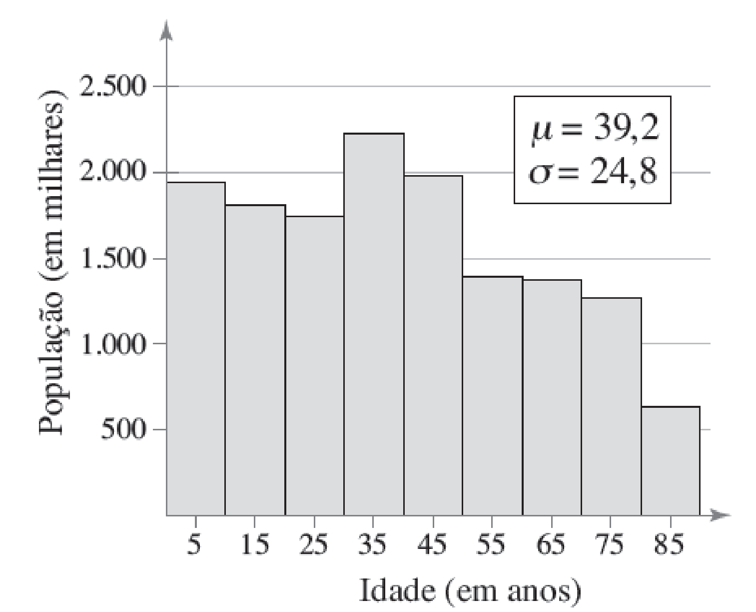
- A porção de qualquer conjunto de dados postos dentro de k desvios padrão ($k > 1$) da média é no mínimo:

$$1 - \frac{1}{k^2}$$

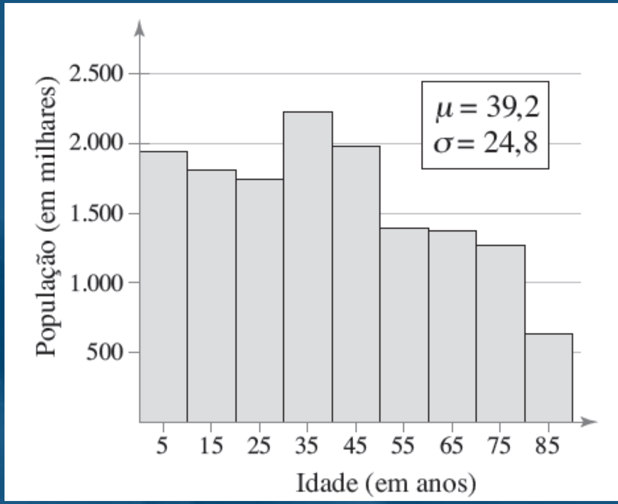
- $k = 2$: Em qualquer conjunto de dados, pelo menos 75% dos dados estão dentro de dois desvios padrão da média.
- $k = 3$: Em qualquer conjunto de dados, pelo menos 88.9% dos dados estão dentro de três desvios padrão da média.

Exemplo: usando o Teorema de Chebychev

A distribuição de idades na Flórida é mostrada no histograma. Aplique o Teorema de Chebychev aos dados usando $k = 2$. O que se pode concluir?



Solução: usando o Teorema de Chebychev



$k = 2$: $\mu - 2\sigma = 39{,}2 - 2(24{,}8) = -10{,}4$ (use 0 já que idade não pode ser negativa) $\mu + 2\sigma = 39{,}2 + 2(24{,}8) = 88{,}8$

Pelo menos 75% da população da flórida está entre 0 e 88,8 anos de idade.

Desvio padrão para dados agrupados

Desvio padrão de uma amostra para uma distribuição de frequência

$$s = \sqrt{\frac{\Sigma (x - \bar{x})^2 f}{n - 1}}$$

em que $n = \Sigma f$ (o número de entradas no conjunto de dados)

Quando uma distribuição de frequência tem classes, estime a média da amostra e o desvio padrão usando o ponto médio de cada classe.

Exemplo: encontrando o desvio padrão para dados agrupados

Você coleta uma amostragem aleatória do número de crianças por por casa em uma região. Encontre a média da amostra e o desvio padrão da amostra do conjunto de dados.

Número de crianças em 50 casas				
1	3	1	1	1
1	2	2	1	0
1	1	0	0	0
1	5	0	3	6
3	0	3	1	1
1	1	6	0	1
3	6	6	1	2
2	3	0	1	1
4	1	1	2	2
0	3	0	2	4

Solução: encontrando o desvio padrão para dados agrupados

- Primeiro, construa a distribuição da frequência
- Encontre a média da distribuição da frequência

x	f	xf
0	10	0(10) = 0
1	19	1(19) = 19
2	7	2(7) = 14
3	7	3(7) =21
4	2	4(2) = 8
5	1	5(1) = 5
6	4	6(4) = 24
	$\Sigma f = 50$	$\Sigma(xf) = 91$

$$\bar{x} = \frac{\Sigma xf}{n} = \frac{91}{50} \approx 1.8$$

A média da amostra é de cerca de 1,8 criança.

Determine a soma dos quadrados.

x	f	$x - \overline{x}$	$(x - \overline{x})^2$	$(x - \overline{x})^2 f$
0	10	0 – 1,8 = –1,8	$(–1,8)^2$ = 3,24	3,24(10) = 32,40
1	19	1 – 1,8 = –0,8	$(–0,8)^2$ = 0,64	0,64(19) = 12,16
2	7	2 – 1,8 = 0,2	$(0,2)^2$ = 0,04	0,04(7) = 0,28
3	7	3 – 1,8 = 1,2	$(1,2)^2$ = 1,44	1,44(7) = 10,08
4	2	4 – 1,8 = 2,2	$(2,2)^2$ = 4,84	4,84(2) = 9,68
5	1	5 – 1,8 = 3,2	$(3,2)^2$ = 10,24	10,24(1) = 10,24
6	4	6 – 1,8 = 4,2	$(4,2)^2$ = 17,64	17,64(4) = 70,56

$$\Sigma(x - \overline{x})^2 f = 145.40$$

Encontre o desvio padrão da amostra.

$$x-\bar{x} \qquad (x-\bar{x})^2 \qquad (x-\bar{x})^2 f$$

$$s=\sqrt{\frac{\Sigma(x-\bar{x})^2 f}{n-1}}=\sqrt{\frac{145.40}{50-1}}\approx 1.7$$

O desvio padrão é de cerca de 1,7 criança.

Seção 2.5

Medidas de posição

Quartilhos

- **Fractis** são números que particionam (dividem) um conjunto de dados ordenados em partes iguais
- **Quartis** dividem dados ordenados em quatro partes aproximadamente iguais
  - **Primeiro quartil, Q_1**: Cerca de um quarto dos dados cai em ou abaixo de Q_1
  - **Segundo quartil, Q_2**: Cerca de metade dos dados caem em ou abaixo de Q_2 (mediana)
  - **Terceiro quartil, Q_3**: Cerca de três quartos dos dados caem em ou abaixo de Q_3

Exemplo: encontrando quartis

As pontuações dos testes de 15 empregados matriculados em um curso de primeiros socorros são listadas. Encontre o primeiro, o segundo e o terceiro quartil das pontuações dos testes.

13 9 18 15 14 21 7 10 11 20 5 18 37 16 17

Solução:

- Q_2 divide o conjunto de dados em duas metades

$$\underbrace{5\ \ 7\ \ 9\ \ 10\ \ 11\ \ 13\ \ 14}_{\text{Metade inferior}}\ \ \underset{\uparrow\ Q_2}{15}\ \ \underbrace{16\ \ 17\ \ 18\ \ 18\ \ 20\ \ 21\ \ 37}_{\text{Metade superior}}$$

Solução: encontrando quartis

- O primeiro e o terceiro quartis são as medianas das metades inferior e superior do conjunto de dados

Metade inferior

Metade superior

5 7 9 10 11 13 14 15 16 17 18 18 20 21 37

↑ Q_1

↑ Q_2

↑ Q_3

Cerca de um quarto dos funcionários obteve nota 10 ou menor; cerca de metade deles obteve 15 ou menor; e cerca de três quartos obteve 18 ou menor.

Amplitude interquartil

Encontrando a amplitude interquartil (VIQ)

- A diferença entre o terceiro e o primeiro quartis
- VIQ = $Q_3 - Q_1$

Exemplo: encontrando a amplitude interquartil

Encontre a amplitude interquartil das notas dos testes.

Lembre-se: $Q_1 = 10$, $Q_2 = 15$ e $Q_3 = 18$

Solução:

- VIQ = $Q_3 - Q_1 = 18 - 10 = 8$

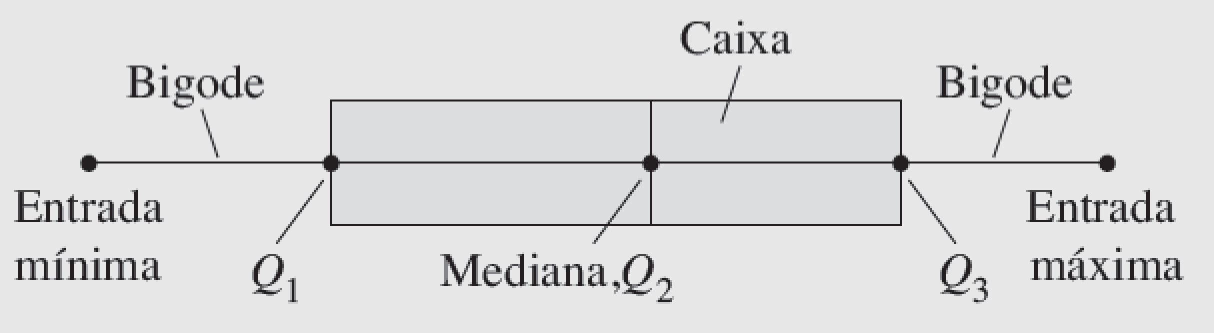
As notas dos testes na porção do meio do conjunto de dados variam no máximo em 8 pontos.

Gráfico de caixa-e-bigodes

- Ferramenta exploratória de análise de dados
- Destaca qualidades importantes do conjunto de dados
- Requer (**sumário de cinco números**):
  - Entrada mínima
  - Primeiro quartil Q_1
  - Mediana Q_2
  - Terceiro quartil Q_3
  - Entrada máxima

Desenhando um gráfico de caixa-e-bigodes

1. Encontre o sumário de cinco números do conjunto de dados.
2. Construa uma escala horizontal que cubra a variância dos dados.
3. Ponha os cinco números acima da escala horizontal.
4. Desenhe uma caixa acima da escala horizontal de Q_1 até Q_3 e desenhe uma linha vertical na caixa em Q_2.
5. Desenhe bigodes saindo da caixa para as entradas mínima e máxima.



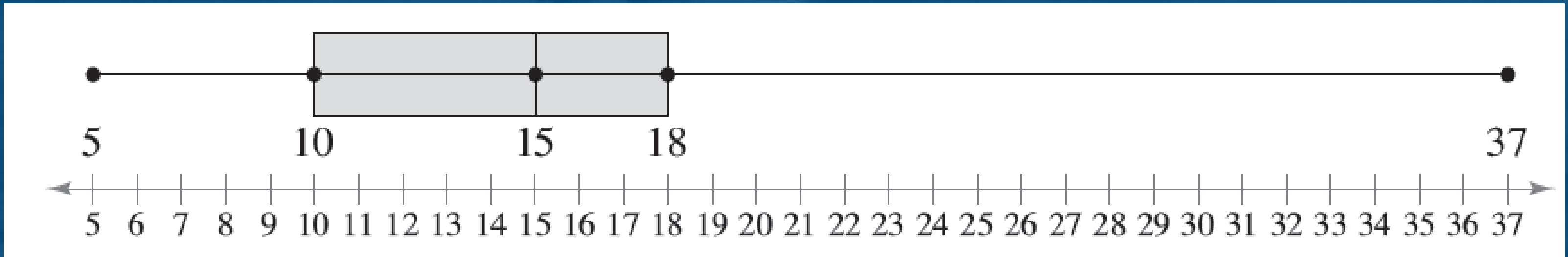
Exemplo: desenhando um gráfico de caixa-e-bigodes

Desenhe um gráfico de caixa-e-bigodes que represente as 15 pontuações dos testes.

Lembre-se: Mín. = 5 $Q_1 = 10$ $Q_2 = 15$ $Q_3 = 18$

Máx. = 37

Solução:



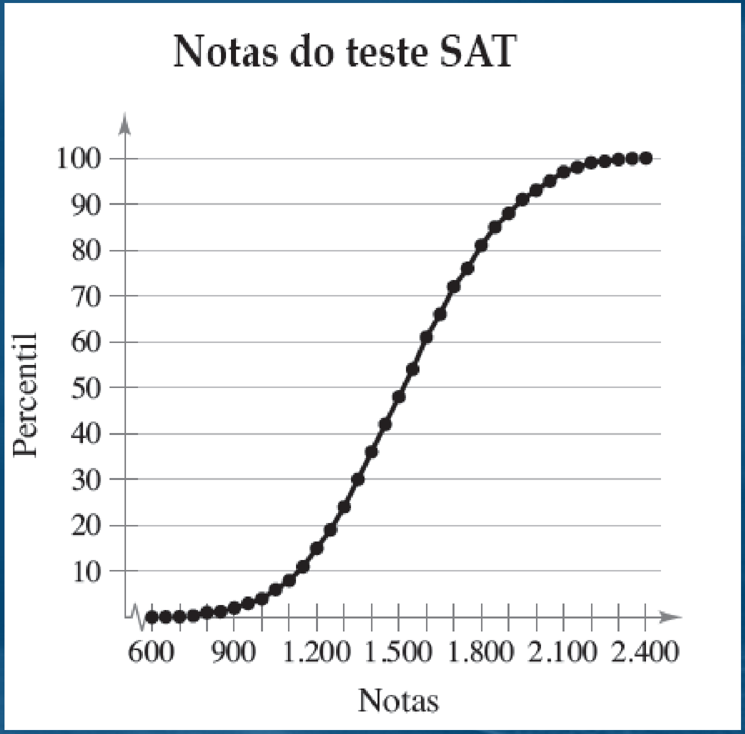
Cerca de metade das notas estão entre 10 e 18. Olhando para o comprimento do bigode direito, pode-se concluir que 37 é um possível valor discrepante.

Percentis e outros fractis

Fractis	**Sumário**	**Símbolos**
Quartis	Divide os dados em 4 partes iguais	Q_1, Q_2, Q_3
Decis	Divide os dados em 10 partes iguais	$D_1, D_2, D_3, \ldots, D_9$
Percentis	Divide os dados em 100 partes iguais	$P_1, P_2, P_3, \ldots, P_{99}$

Exemplo: interpretando percentis

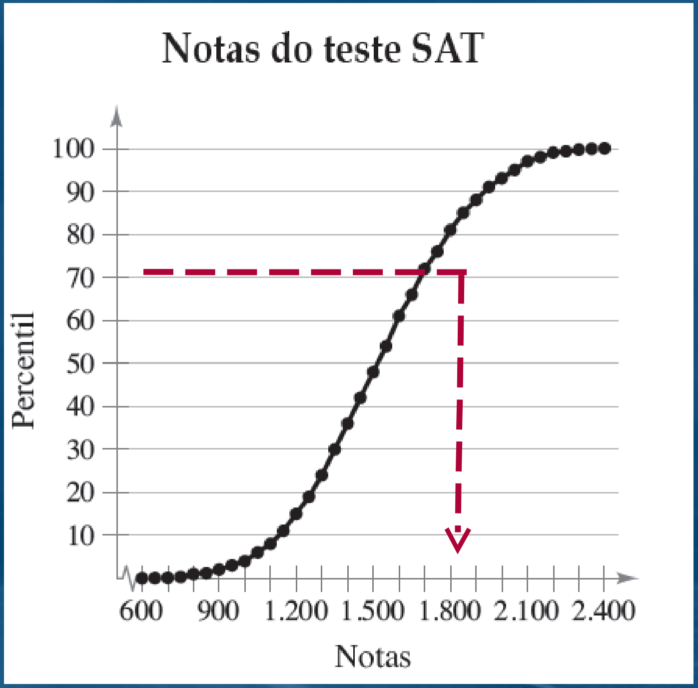
A ogiva representa a distribuição de frequência cumulativa para as provas do SAT (vestibular dos EUA) de estudantes em uma ano recente. Qual nota representa o 72º percentil? Como você deve interpretar isso? (*Fonte: College Board Online*.)



Solução: interpretando percentis

O 72º percentil corresponde à nota 1.700.

Isso significa que 72% dos alunos obtiveram resultados de 1.700 ou menos.



O escore padrão

Escore padrão (escore *z*)

- Representa o número de desvios padrão que um dado valor x cai da média μ.

- $$z = \frac{\text{valor} - \text{média}}{\text{desvio padrão}} = \frac{x - \mu}{\sigma}$$

Exemplo: comparando escores z de diferentes conjuntos de dados

Em 2007, o ator Forest Whitaker ganhou o Oscar de melhor ator, aos 45 anos de idade, por sua atuação no filme *O Último Rei da Escócia*. A atriz Helen Mirren ganhou o prêmio de melhor atriz aos 61 anos por seu papel em *A Rainha*. A idade média para todos os vencedores do prêmio de melhor ator é 43,7, com desvio padrão de 8,8. A idade média para as vencedoras do prêmio de melhor atriz é 36, com desvio padrão de 11,5. Encontre o escore *z* que corresponda à idade de cada ator ou atriz. Depois, compare os resultados.

Solução: comparando escores *z* de diferentes conjuntos de dados

- Forest Whitaker

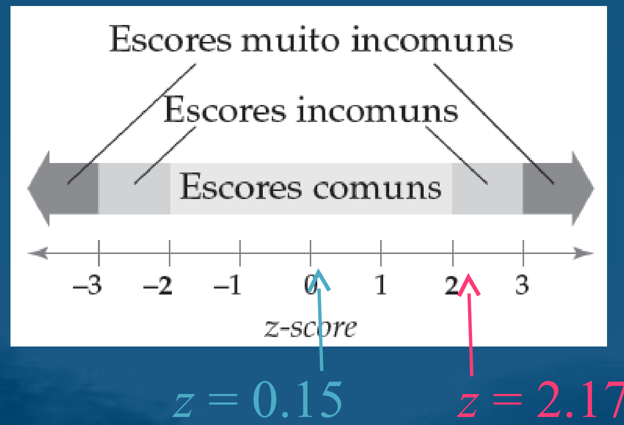
$$z = \frac{x - \mu}{\sigma} = \frac{45 - 43.7}{8.8} \approx 0.15$$

Desvio padrão 0,15 acima da média

- Helen Mirren

$$z = \frac{x - \mu}{\sigma} = \frac{61 - 36}{11.5} \approx 2.17$$

Desvio padrão 2,17 acima da média



Escores muito incomuns

Escores não comuns

Escores comuns

Escore *z*

O escore *z* correspondente à idade de Helen Mirren é mais de dois desvios padrão da média, então é considerado incomum. Comparado a outras vencedoras do prêmio de melhor atriz, ela é relativamente mais velha, enquanto a idade de Forest Whitaker é pouco acima da média dos ganhadores do prêmio de melhor ator.

Sumário

- Construímos distribuições de frequência e suas representações gráficas (histogramas, polígonos e ogivas)
- Fizemos gráficos de dados quantitativos usando um diagrama de ramos-e-folhas e diagrama de pontos
- Fizemos gráficos de dados qualitativos usando gráficos de pizza e de Pareto
- Fizemos gráficos de dados emparelhados usando gráficos de dispersão e gráficos de série temporal

Sumário

- Determinamos a média, a mediana e a moda de uma população e de uma amostra
- Determinamos a média ponderada de um conjunto de dados e a média de uma distribuição de frequência
- Descrevemos a forma da distribuição como simétrica, uniforme, ou assimétrica e comparamos a média e a mediana de cada um
- Determinamos a variância e o desvio padrão da população e da amostra
- Usamos a Regra Empírica e o Teorema de Chebychev para interpretar o desvio padrão

Sumário

- Determinamos os quartis de um conjunto de dados
- Determinamos a amplitude interquartil de um conjunto de dados
- Criamos gráfico de caixa-e-bigodes
- Interpretamos outros fractis, como percentis
- Determinamos e interpretamos o escore padrão (escore z)