
PUC-SP
Mineração de Dados
Prof. Dr. Daniel Rodrigues da Silva
PUC-SP

PUC-SP

Introdução a Mineração de Dados Com Python

PUC-SP

Bibliografia básica:

Introdução à mineração de dados : conceitos básicos, algoritmos e aplicações. Leandro Nunes de Castro e Daniel Gomes Ferrari. Saraiva, 2016.

Inteligência Artificial - Uma Abordagem de Aprendizado de Máquina . André Carlos Ponce de Leon Ferreira et al. 2 Ed. LTC, 2024.

Estatística Aplicada - Larson / Farber – Editora Pearson, 2015

PUC-SP

Ementa

- Revisão de Estatística Descritiva
- Conceitos de Mineração de Dados
- Análise Exploratória
- Análise Preditiva,
- Agrupamentos
- Associações.

Critérios de Avaliação

Pelo menos 75% de presença

Média final maior ou igual a 5,0

$$MF = \frac{N_1 + N_2}{2} \quad ; \quad N_i = \frac{P_i + A_i}{2} \ , \quad i = 1\,,2$$

Onde:

$P_i = nota\ do\ projeto\ do\ bimestre$

$A_i = nota\ da\ atividade/prova\ do\ bimestre$

PUC-SP

Revisão de Estatística Descritiva

Texto baseado no livro:

***Estatística Aplicada* - Larson / Farber – Editora Pearson – 2015**

PUC-SP

Distribuição de frequência e seus gráficos

PUC-SP

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 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	

Distribuição de frequência

Distribuição de frequência

Uma 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*

PUC-SP

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

PUC-SP

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.

PUC-SP

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 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

$$\frac{max - min}{\#classes} = \frac{86 - 7}{7} \approx 11{,}29$$

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	|||| ||	6
19 – 30	|||| |||| ||	10
31 – 42	|||| |||| ||||		13
43 – 54	|||| ||||	8
55 – 66	||||		5
67 – 78	|||| ||	6
79 – 90	||	2

$\Sigma f = 50$

Determinando o ponto médio

Ponto médio de uma classe =

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

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

Determinando a frequência relativa

Frequência relativa de uma classe

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

Classe	Frequência, f	Frequência relativa
7 – 18	6	$\frac{6}{50} = 0{,}12$
19 – 30	10	$\frac{10}{50} = 0{,}20$
31 – 42	13	$\frac{13}{50} = 0{,}26$

PUC-SP

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

$\Sigma f = 50$

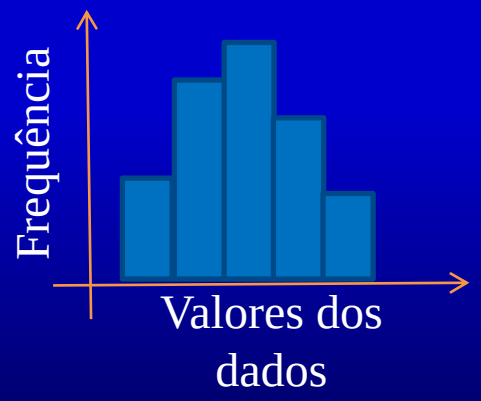
$\Sigma \frac{f}{n} = 1$

Gráficos de distribuição de frequência

PUC-SP

Histograma de frequência

- Um gráfico de barras que representa a distribuição de 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

- 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		10
31 – 42		13

PUC-SP

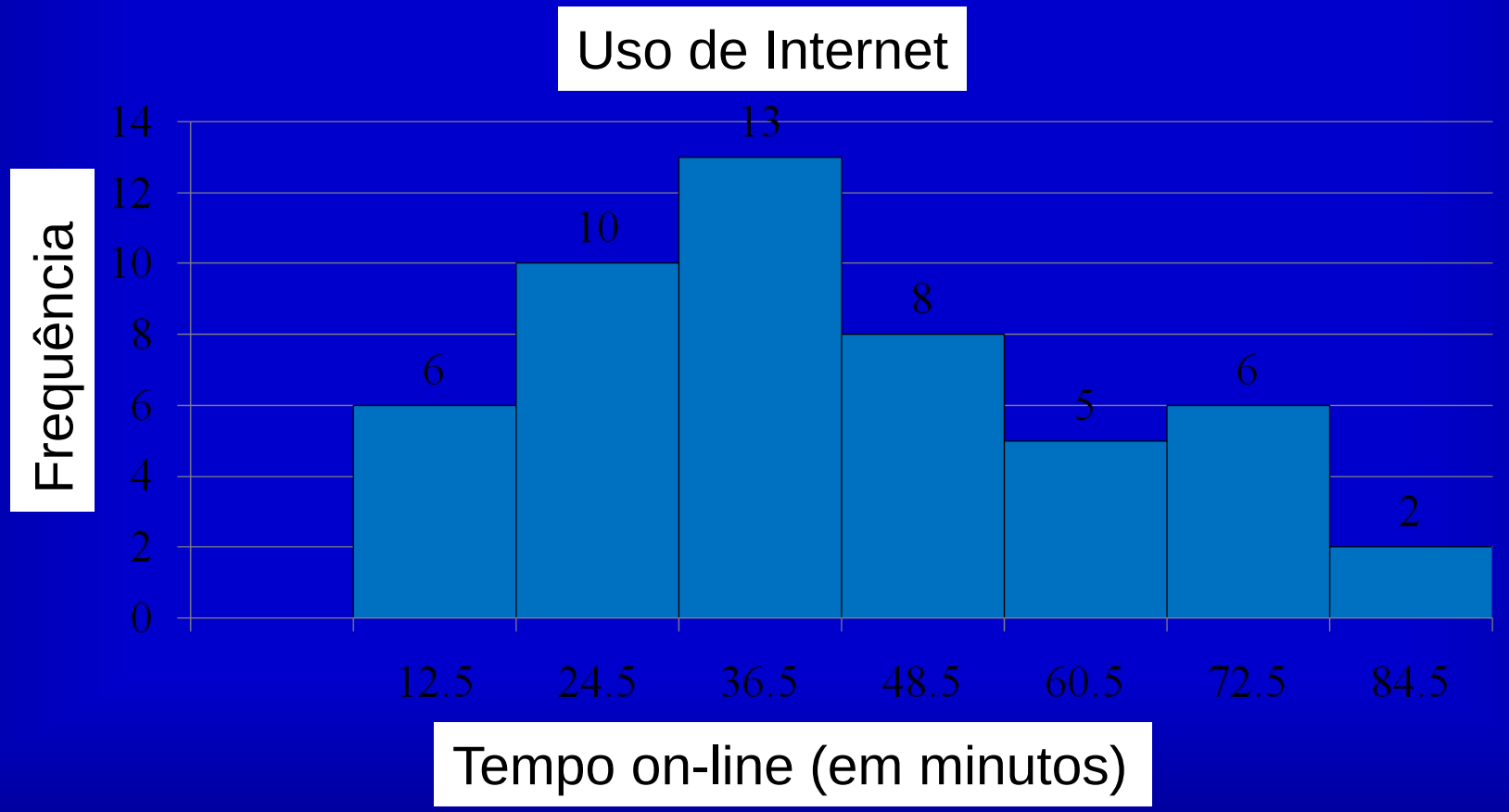
Classe	Fronteiras de classes	Frequência, *f*
7 – 18	6.5 – 18.5	6
19 – 30	18.5 – 30.5	10
31 – 42	30.5 – 42.5	13
43 – 54	42.5 – 54.5	8
55 – 66	54.5 – 66.5	5
67 – 78	66.5 – 78.5	6
79 – 90	78.5 – 90.5	2

Exemplo: histograma de frequência

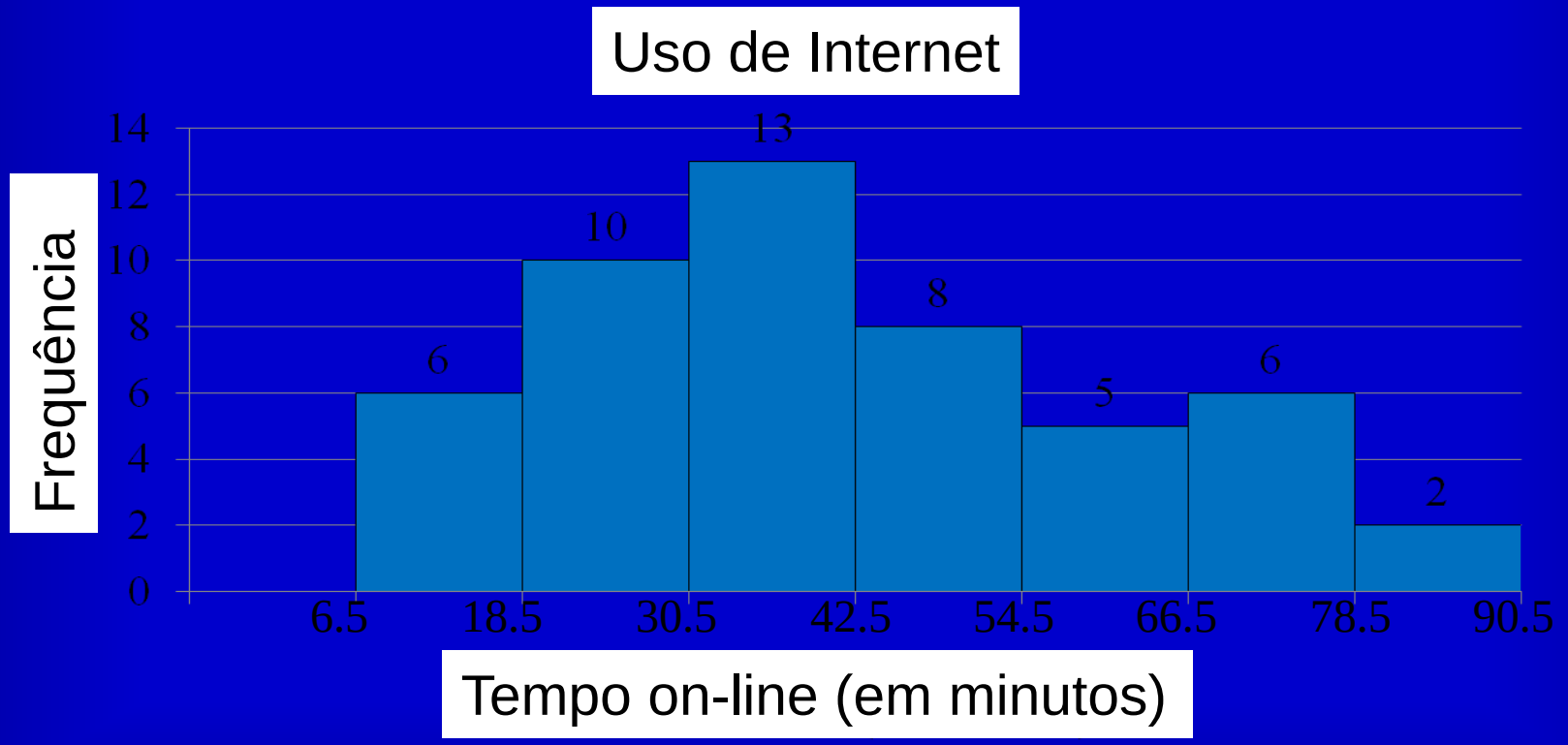
Construa um histograma de frequência para a distribuição da frequência do uso da internet.

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)



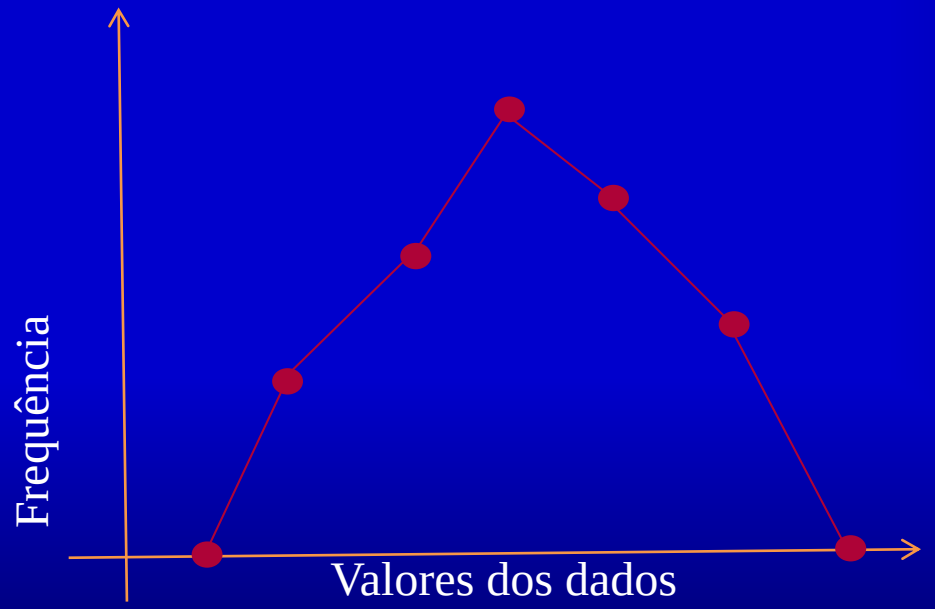
É visível que mais da metade dos assinantes passaram entre 19 e 54 minutos na Internet em sua sessão mais recente.

Gráficos de distribuições de frequência

PUC-SP

Polígono de frequência

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



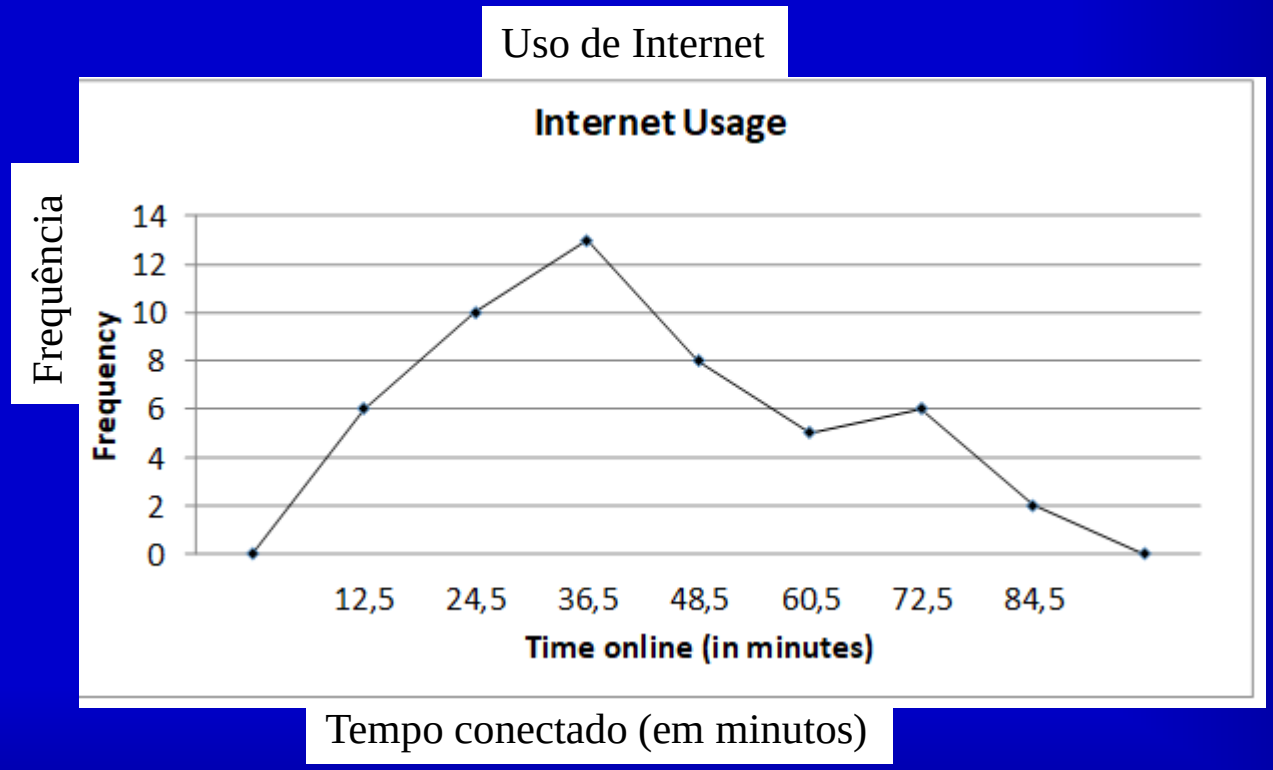
Exemplo: polígono de frequência

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.

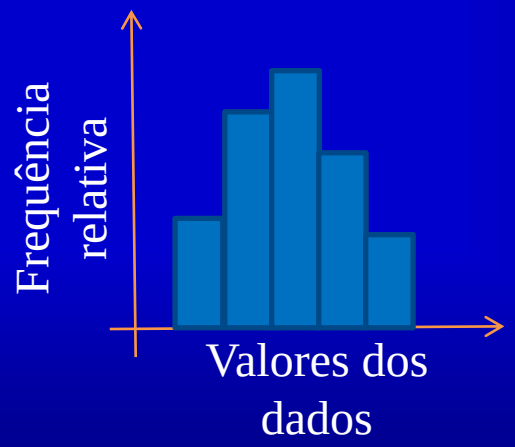


Pode-se perceber que a frequência dos assinantes aumenta até 36,5 minutos e então diminui.

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 absolutas.

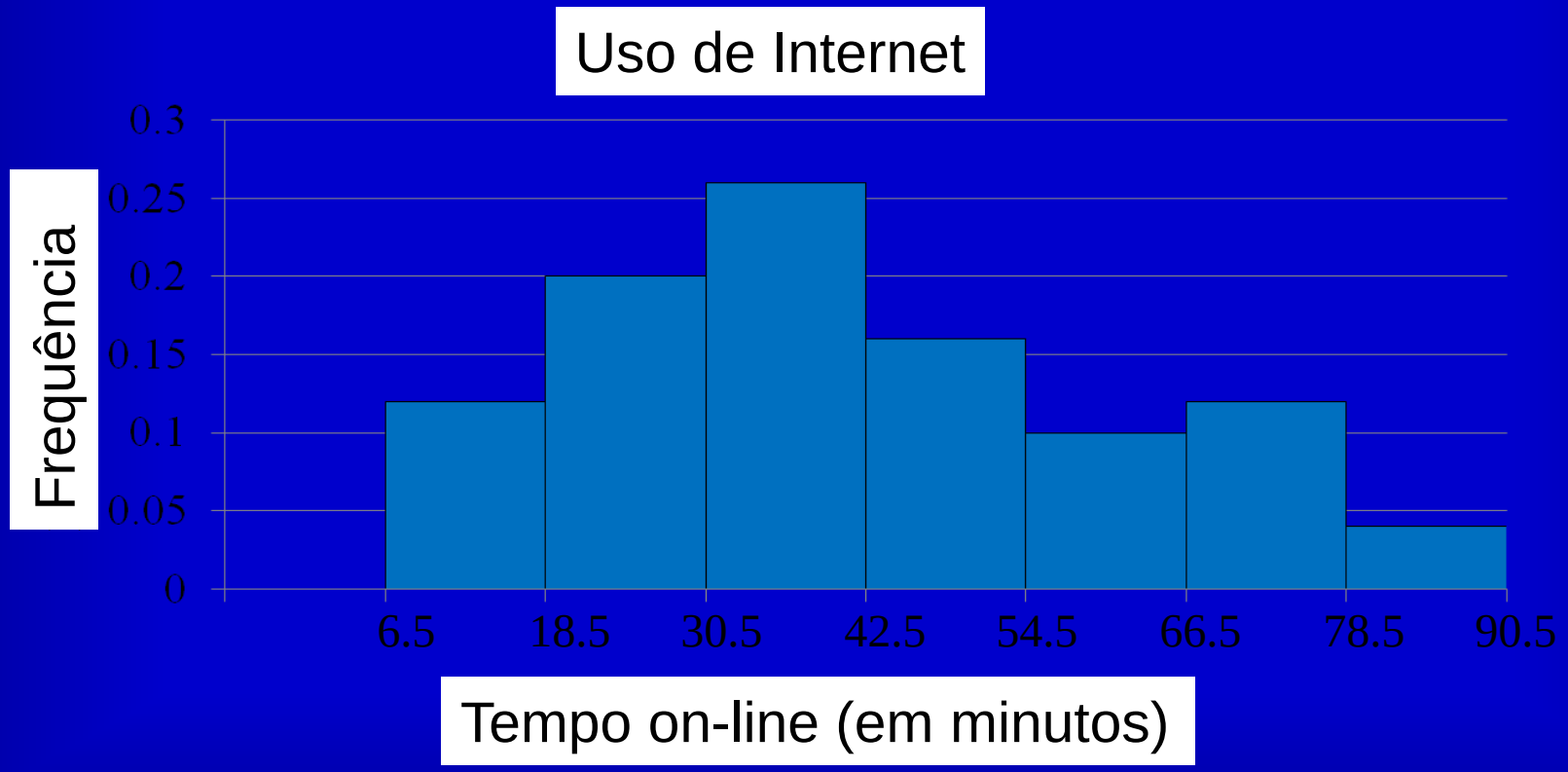


Exemplo: histograma de frequência relativa

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.

PUC-SP

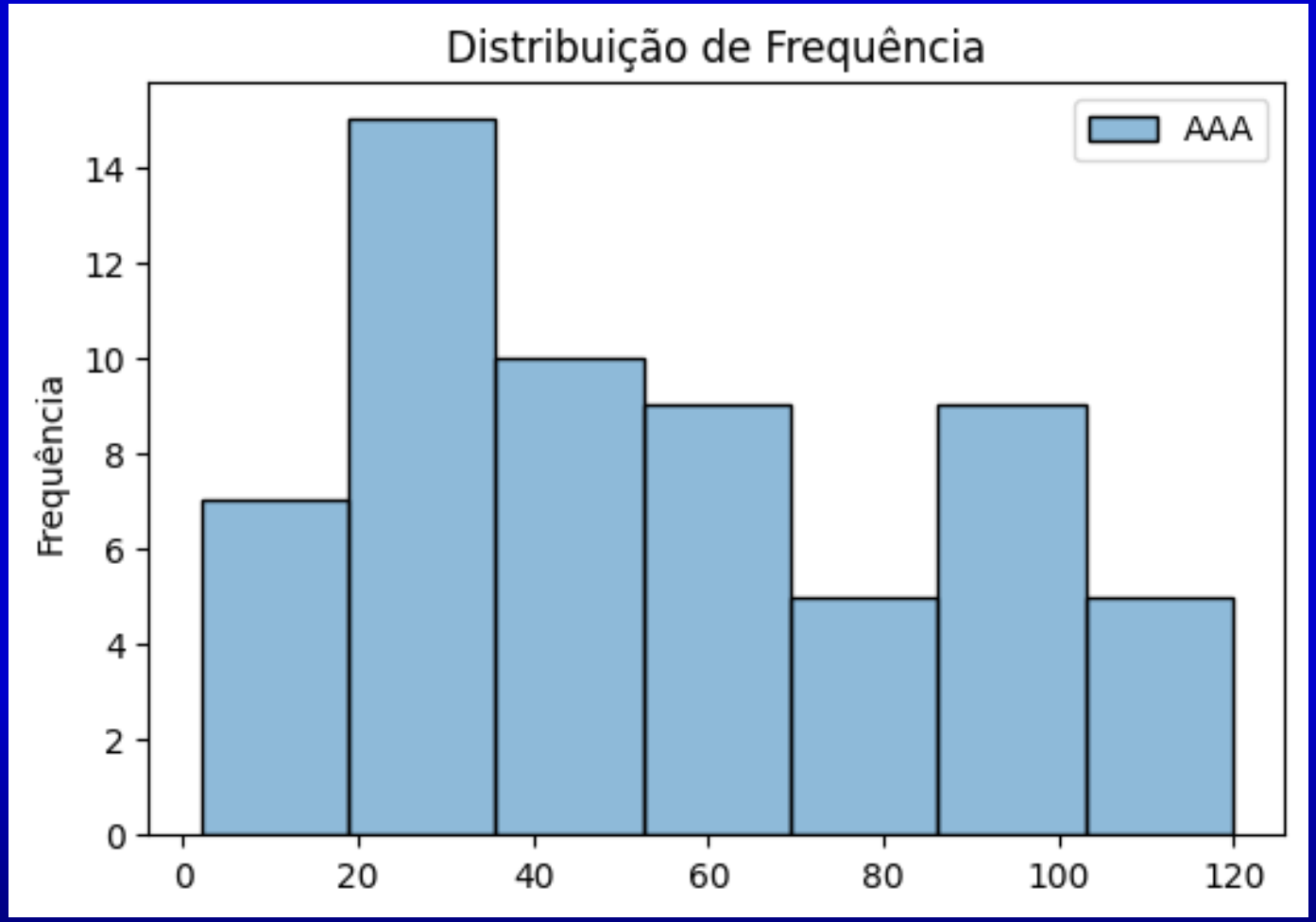
Exemplo 2: A amostra seguinte lista o número de minutos que 60 usuários de TV a cabo assistiram algum conteudo do seu pacote nas últimas duas horas. Construa uma distribuição de frequência para 8 classes e construa um histograma.

20	55	5	64	78	49	91	87	18	83	33	39	30	31	59	85	102	24	27	28
92	108	98	67	85	120	48	19	32	69	24	59	6	49	116	37	92	43	101	60
55	107	25	33	57	25	17	49	24	101	14	45	73	109	91	2	11	47	21	38

Distribuição de frequência expandida

Classe		Ponto médio	Freq Abs	Freq rel	Acum Abs	Acum rel
2	16	9	5	0,0833	5	0,0833
17	31	24	14	0,2333	19	0,3167
32	46	39	8	0,1333	27	0,4500
47	61	54	11	0,1833	38	0,6333
62	76	69	4	0,0667	42	0,7000
77	91	84	7	0,1167	49	0,8167
92	106	99	6	0,1000	55	0,9167
107	121	114	5	0,0833	60	1,0000
			60	1		

Distribuição de frequência expandida



PUC-SP

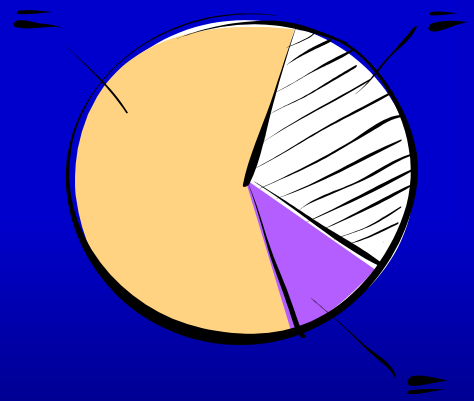
Mais gráficos e representações

Fazendo gráficos de conjunto de dados qualitativos

PUC-SP

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 (porcentagem) 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	$\frac{823}{37594} \approx 0{,}02$

PUC-SP

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.

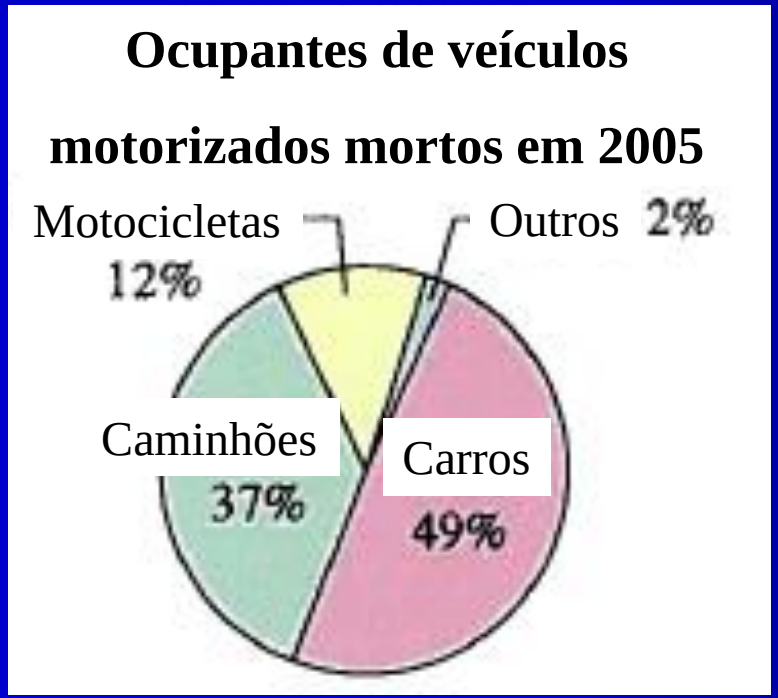
Por exemplo, o ângulo central para carros é

$$360(0{,}49) \approx 176º$$

PUC-SP

Tipo de veículo	Frequência, *f*	Frequência relativa	Ângulo central
Carros	18.440	0,49	360º(0,49)≈176º
Caminhões	13.778	0,37	360º(0,37)≈133º
Motocicletas	4.553	0,12	360º(0,12)≈43º
Outros	823	0,02	360º(0,02)≈7º

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

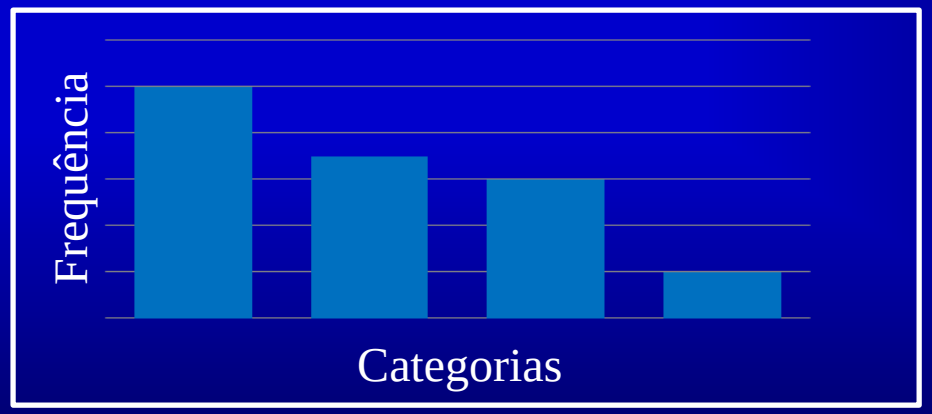


A partir do gráfico, pode-se concluir que a maioria dos acidentes fatais em veículos automotivos foram aqueles envolvendo carros.

Fazendo gráficos de conjunto de dados qualitativos

Gráfico de Pareto

- Um gráfico de barras verticais no qual a altura de cada barra representa uma frequência ou uma frequência relativa
- As barras são posicionadas por ordem decrescente de altura; a barra mais alta é posicionada à esquerda



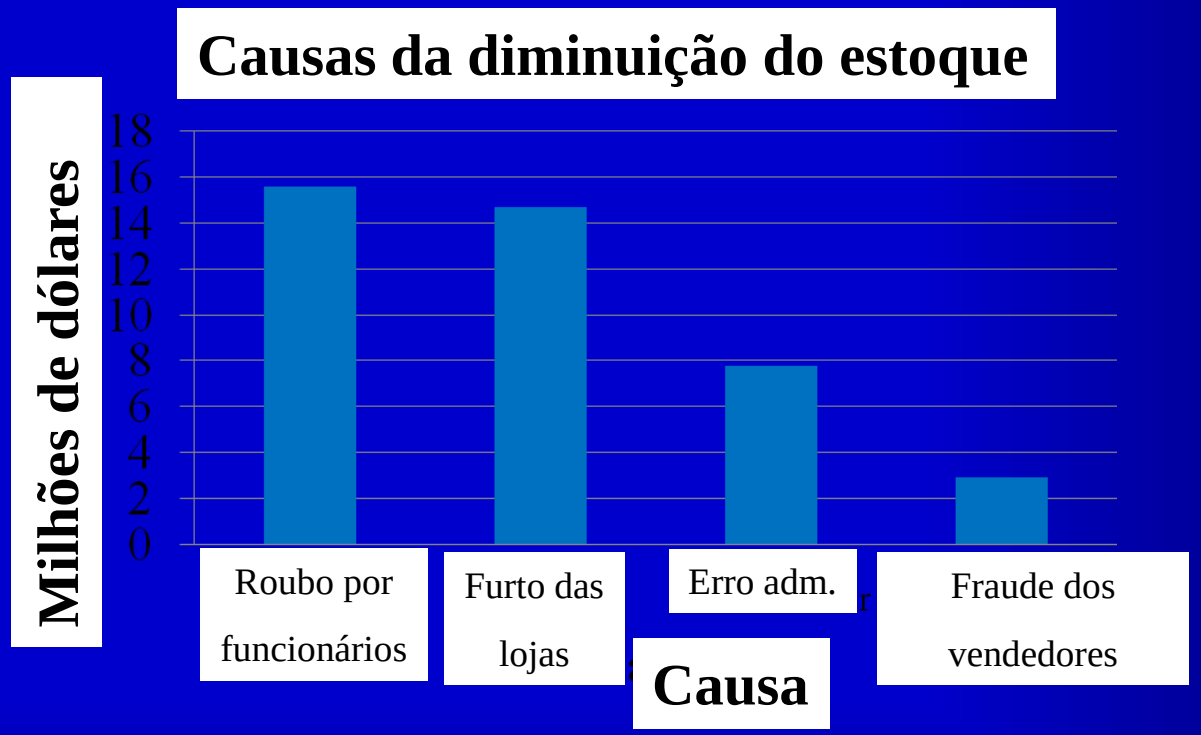
Exemplo: construindo um gráfico de Pareto

Recentemente, a indústria de varejo perdeu 41 milhões 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 milhões), roubo por funcionários (15,6 milhões), furto das lojas (14,7 milhões) e fraude dos vendedores (2,9 milhões). Se você fosse um varejista, para qual causa de redução de estoque você olharia primeiro?

(*Fonte: University of Florida*.)

Solução: 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.

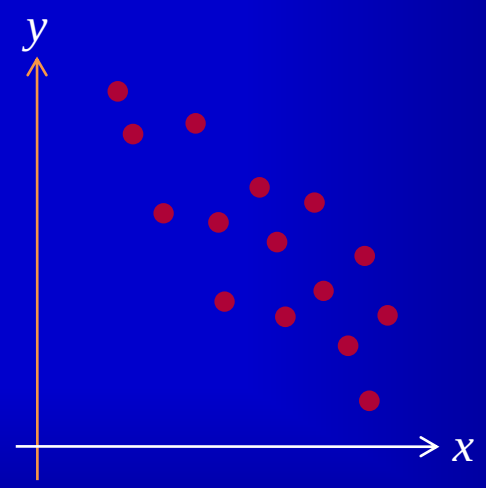
Fazendo gráficos de conjunto de dados emparelhados

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

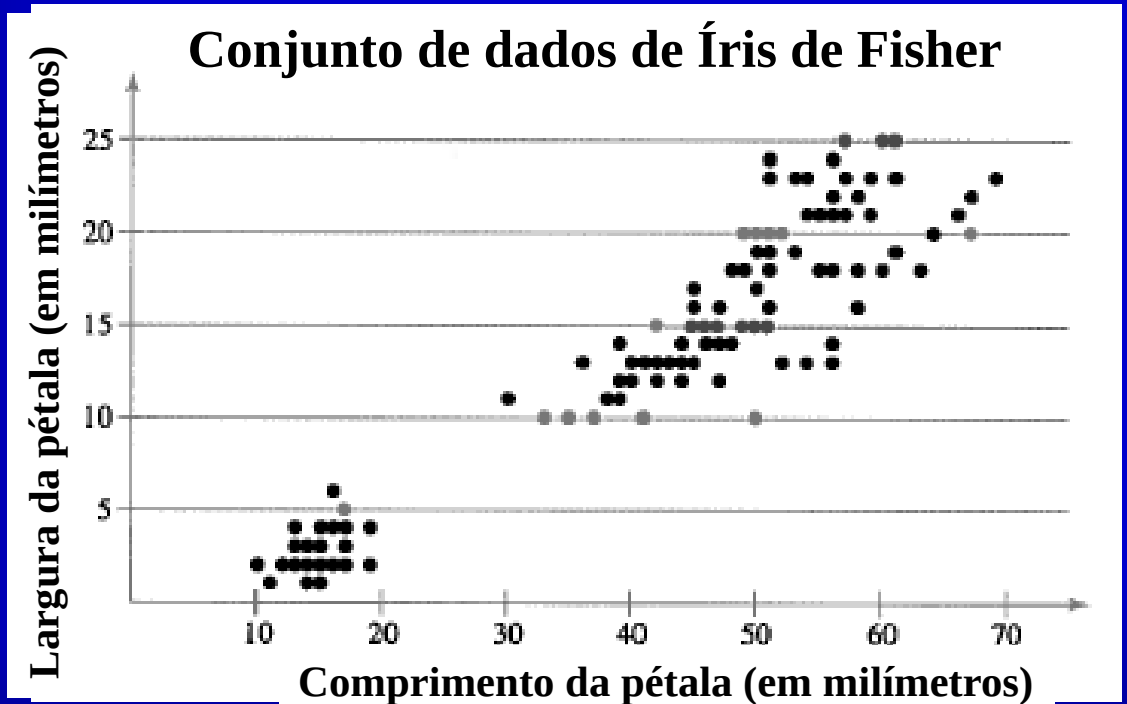


Exemplo: interpretando um gráfico de dispersão

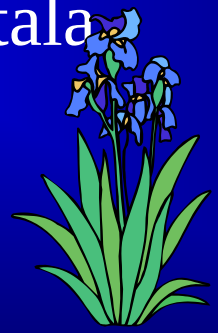
O estatístico britânico Ronald Fisher apresentou um famoso conjunto de dados chamado de conjunto de dados de Íris de Fisher. Esse conjunto de dados descreve várias características físicas, como o comprimento de pétalas e a sua largura (em milímetros) para três espécies de íris (flor). No gráfico de dispersão mostrado, os comprimentos de pétalas formam o primeiro conjunto de dados e as larguras formam o segundo conjunto de dados. Conforme o comprimento da pétala aumenta, o que tende a acontecer com a largura? (*Fonte: Fisher, R. A., 1936*.)

PUC-SP

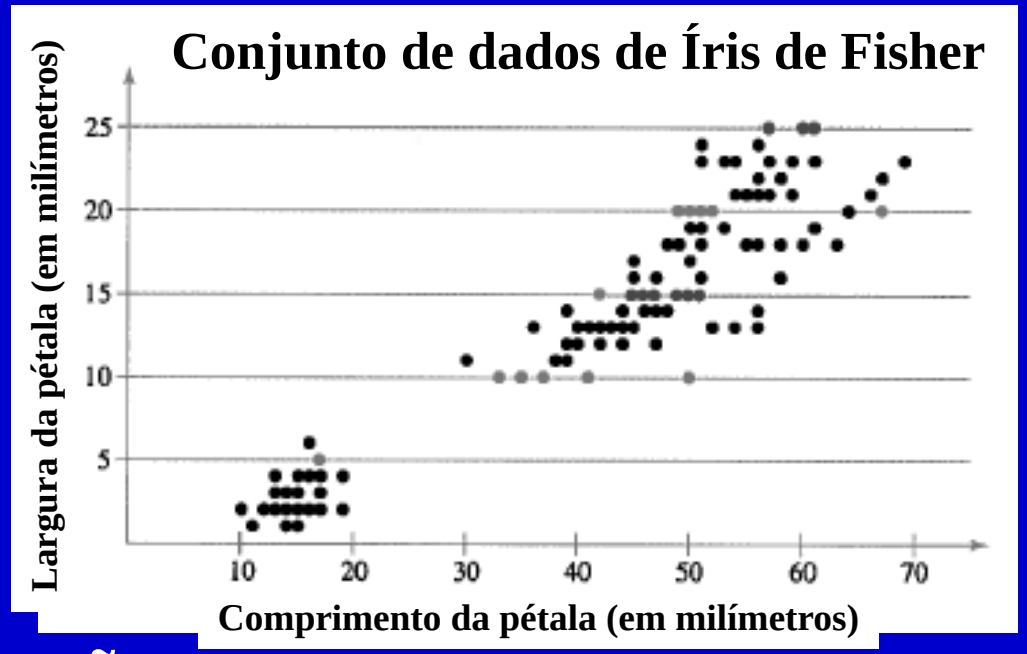
Conforme o comprimento da pétala aumenta, o que tende a acontecer com a largura?



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



Solução: interpretando um gráfico de dispersão



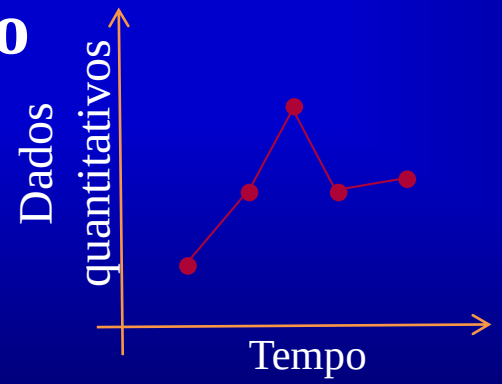
Interpretação

Partindo do gráfico de dispersão, pode-se notar 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*.)

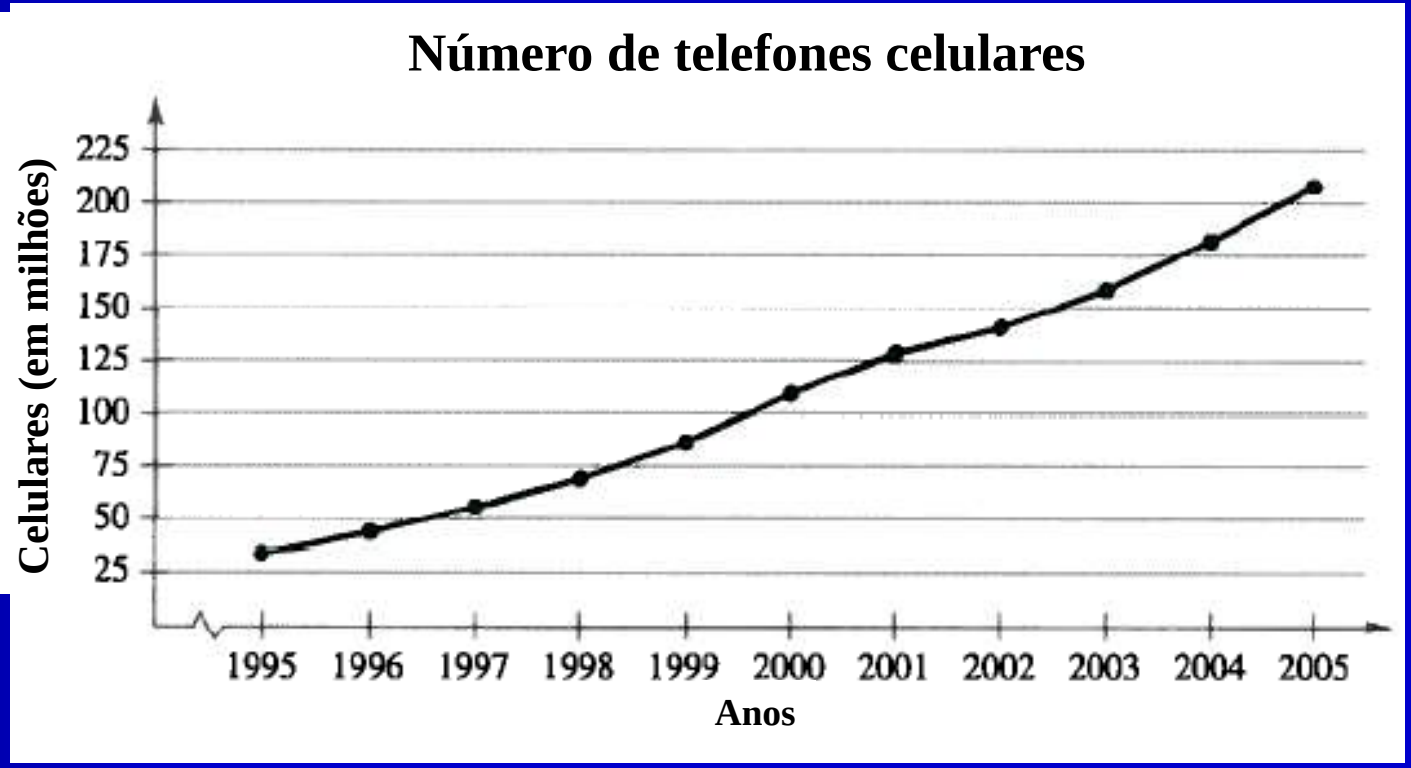
Ano	Usuários (em milhões)
1995	33.8
1996	44.0
1997	55.3
1998	69.2
1999	86.0
2000	109.5
2001	128.4
2002	140.8
2003	158.7
2004	182.1
2005	207.9

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 (em 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

PUC-SP



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.

PUC-SP

Medidas de tendência central

Medidas de tendência central

- Um valor que representa uma entrada de um conjunto de dados como típico ou central
- Medidas de tendência central mais comuns:
  - Média
  - Mediana
  - Moda

Medidas de tendência central: média

Média

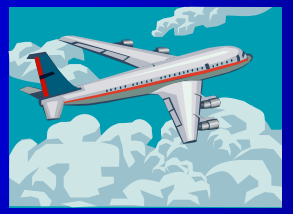
- A soma de todas as entradas de dados divididas pelo número de entradas
- **Notação sigma**: Σx = adicione todas as entradas (x) no conjunto de dados
- **Média populacional**: $\mu = \frac{\sum x}{n}$
- **Média amostral:** $\bar{x} = \frac{\sum x}{n}$

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

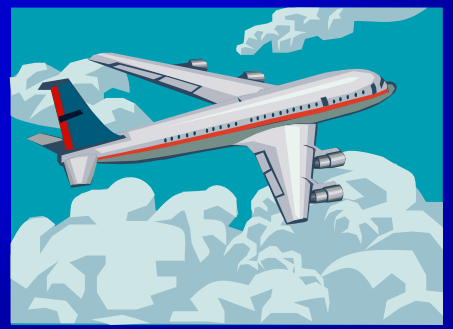
872 432 397 427 388 782 397

- 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

- $\bar{x} = \frac{\sum x}{n} = \frac{3695}{7} \approx 527,9$



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

Medidas de tendência central

PUC-SP

Mediana

- O valor que está no meio dos dados quando o conjunto dos dados é **ordenado**

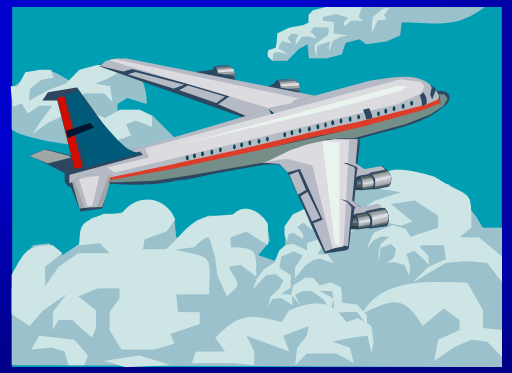
- Mede o centro de um conjunto de dados ordenado dividindo-o em duas partes iguais

- Se o conjunto de dados possui um número de entradas:
  - **ímpar**: o mediano é o elemento do meio
  - **par**: o mediano será a média dos dois 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

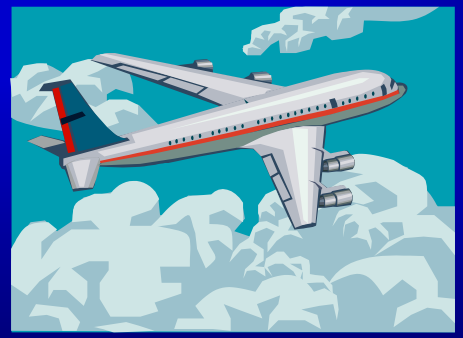
872 432 397 427 388 782 397

- 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

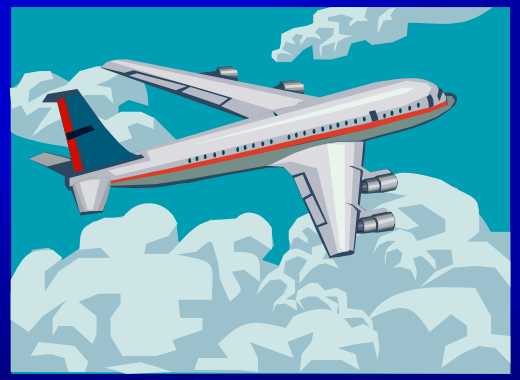
O preço mediano dos voos é $ 427.



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

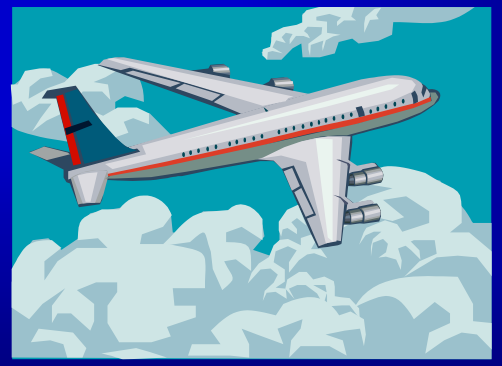
- Primeiramente, ordene os dados

  388 397 397 427 782 872

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

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

O preço mediano dos voos é $ 412.



Medidas de tendência central: moda

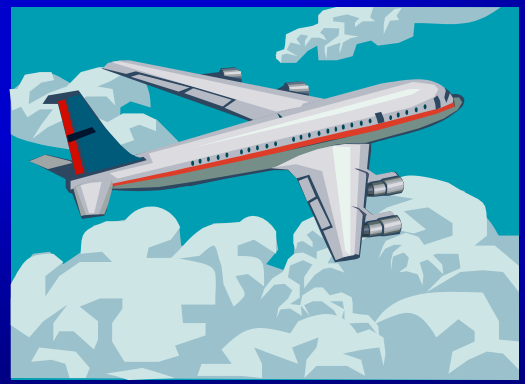
Moda

- A entrada de dados que ocorre com 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



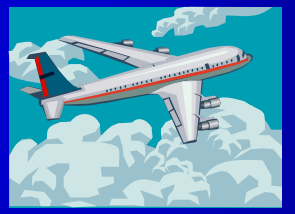
Solução: encontrando a moda

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.

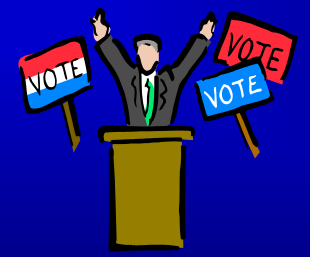


A moda dos preços dos voos é $ 397.

Exemplo: encontrando a moda

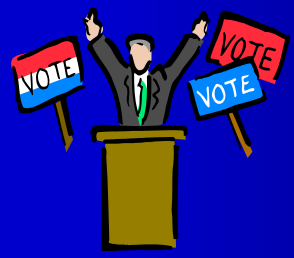
Em um debate político uma amostra de membros da audiência foi questionada à respeito de seus partidos políticos. Suas respostas estão na tabela. Qual a moda de suas respostas?

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

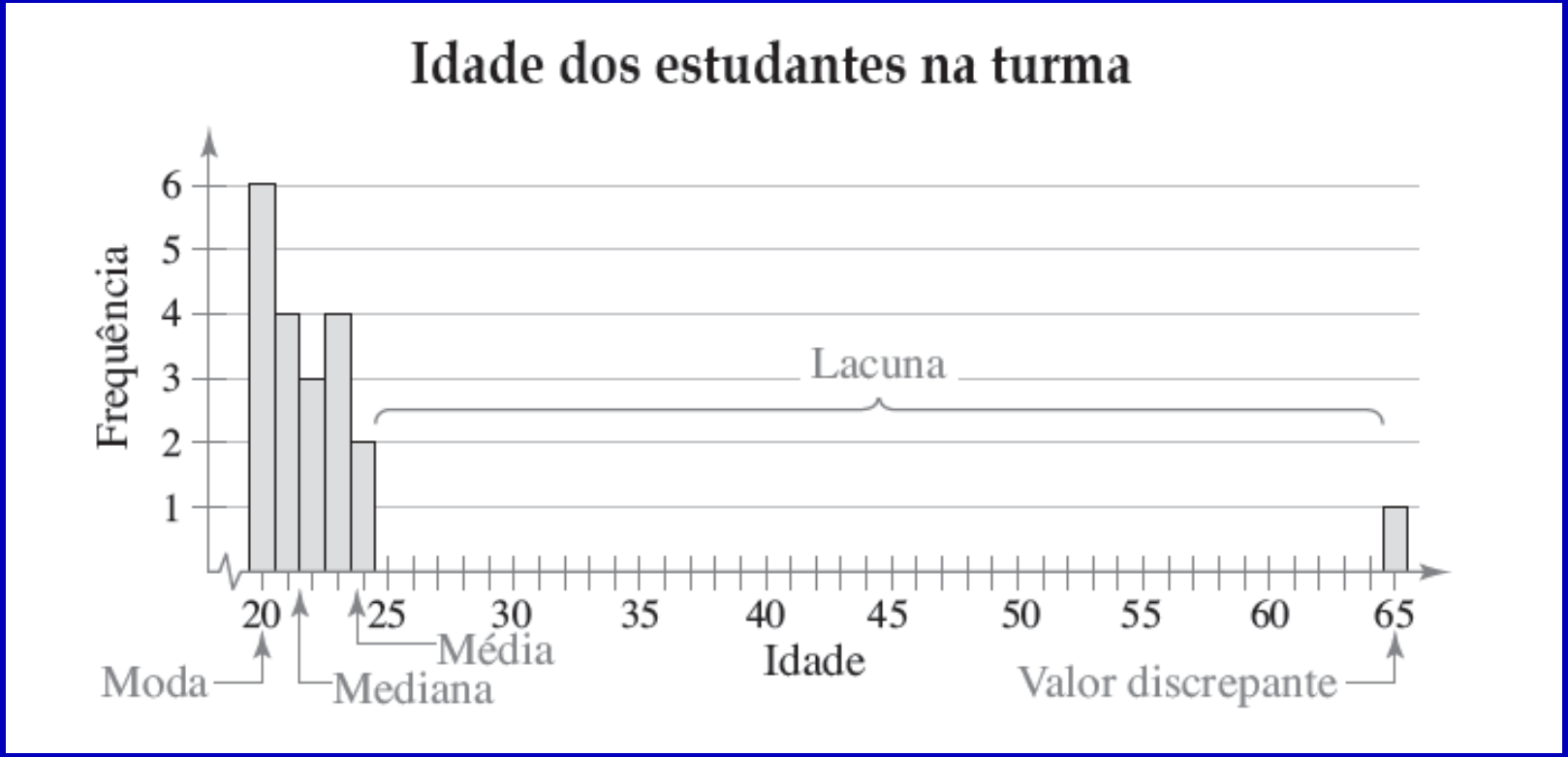
Mediana: $\frac{21 + 22}{2} = 21.5$

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, por isso 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

$$\overline{x} = \frac{\Sigma(x \cdot w)}{\Sigma w}$$

Onde 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
		$\Sigma w = 1$	$\Sigma(x \cdot w) = 88,6$

$$\overline{x} = \frac{\Sigma(x \cdot w)}{\Sigma w} = \frac{88.6}{1} = 88.6$$

A média ponderada para essa disciplina foi 88,6. Não tirou um A.

Média de dados agrupados

Média de uma distribuição de frequência é calculada por:

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

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.	$x = \frac{\textbf{limite inferior} + \textbf{limite superior}}{2}$
2. Encontre a soma dos produtos dos pontos médios e das frequências.	$\Sigma(x \cdot f)$
3. Encontre a soma das frequências.	$n = \Sigma f$
4. Encontre a média da distribuição das frequências.	$\overline{x} = \frac{\Sigma(x \cdot f)}{n}$

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

PUC-SP

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

$$\overline{x} = \frac{\Sigma(x \cdot f)}{n} = \frac{2089}{50} \approx 41.8 \text{ minutos ;}$$

PUC-SP

Considerando o dataset “banco.csv”, determine:

- **A média**
- **A Mediana**
- **A moda**

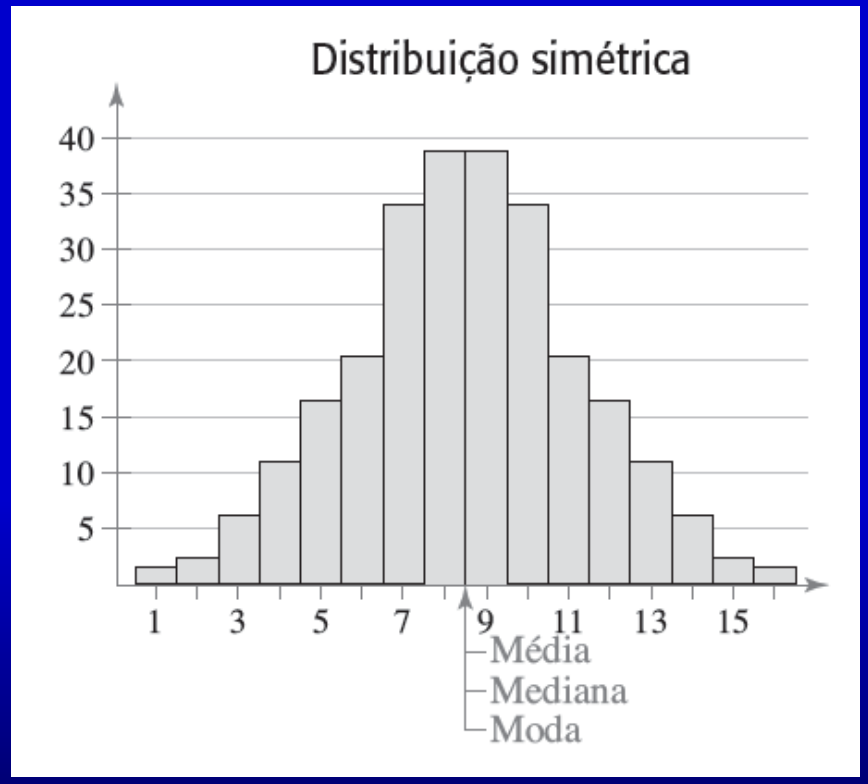
Para todas as colunas numéricas do arquivo


```
df = pd.read_csv('D:/A - PUC/Mineração/Dados/banco.csv' , sep = ';')

media = df['balance'].mean().round(4)

mediana = df['balance'].median()

moda = df['balance'].mode()

print('Média = ',media,'\nMediana = ',mediana,'\nModa = ',moda)
```

A forma das distribuições

Distribuição simétrica

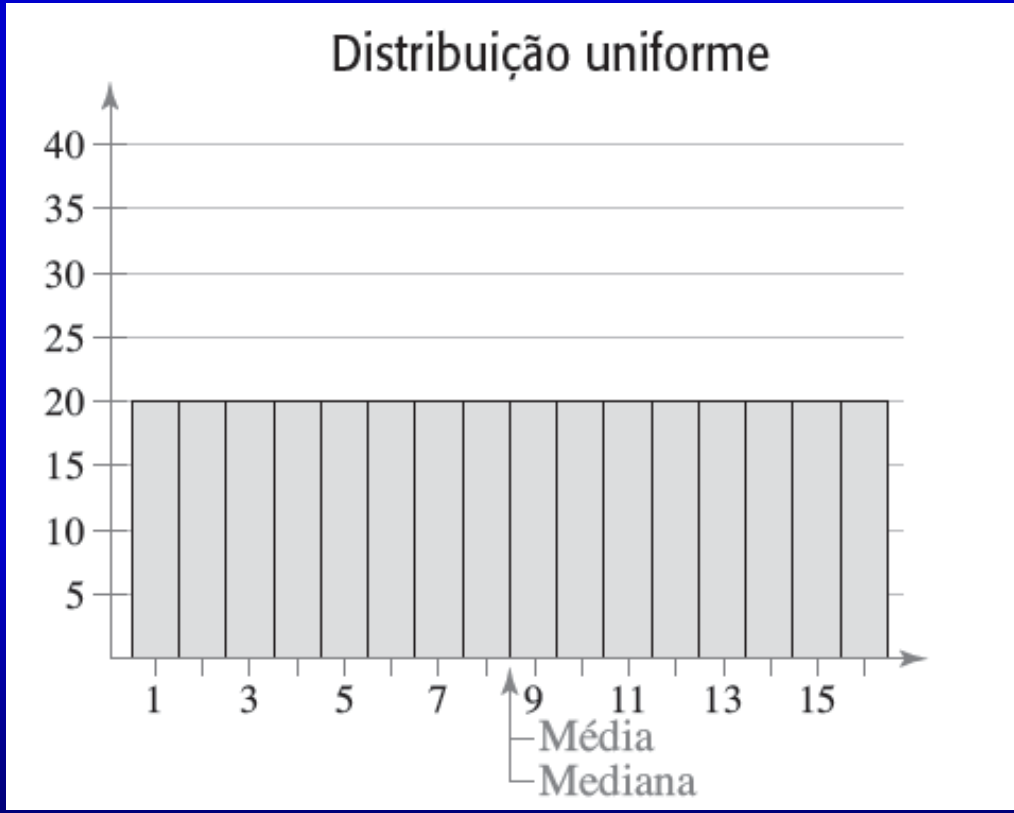
Uma linha vertical pode ser traçada do meio do gráfico de distribuição e as metades resultantes são quase idênticas.



O formato das distribuições

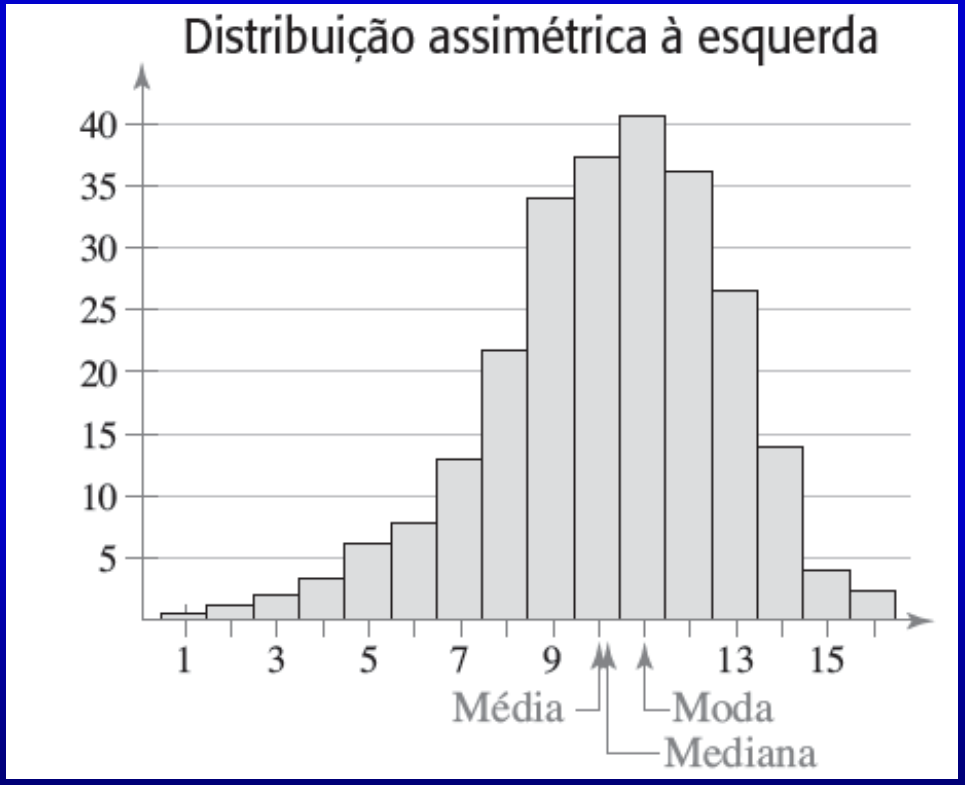
Distribuição uniforme (retangular)

- 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