Agregação com Pandas e Conceitos de Estatística

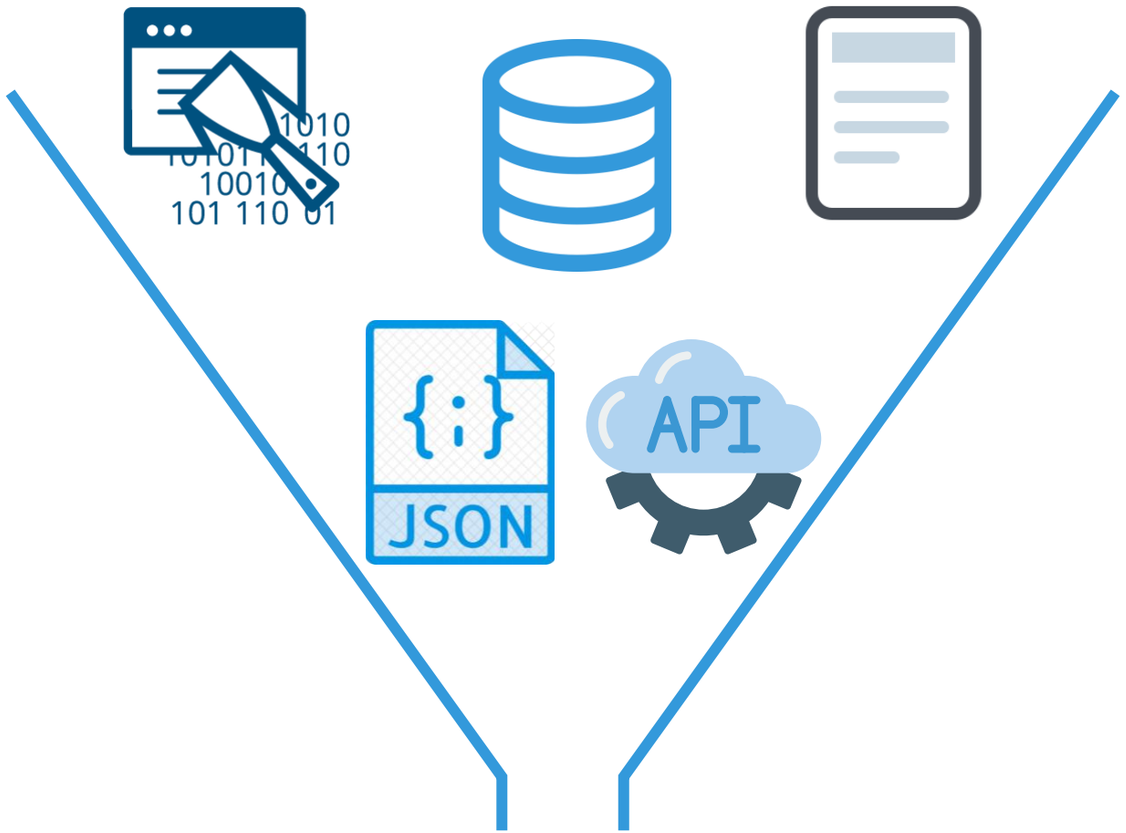


Imagem: https://clearbit.com/our-data

Professor: Alex Pereira

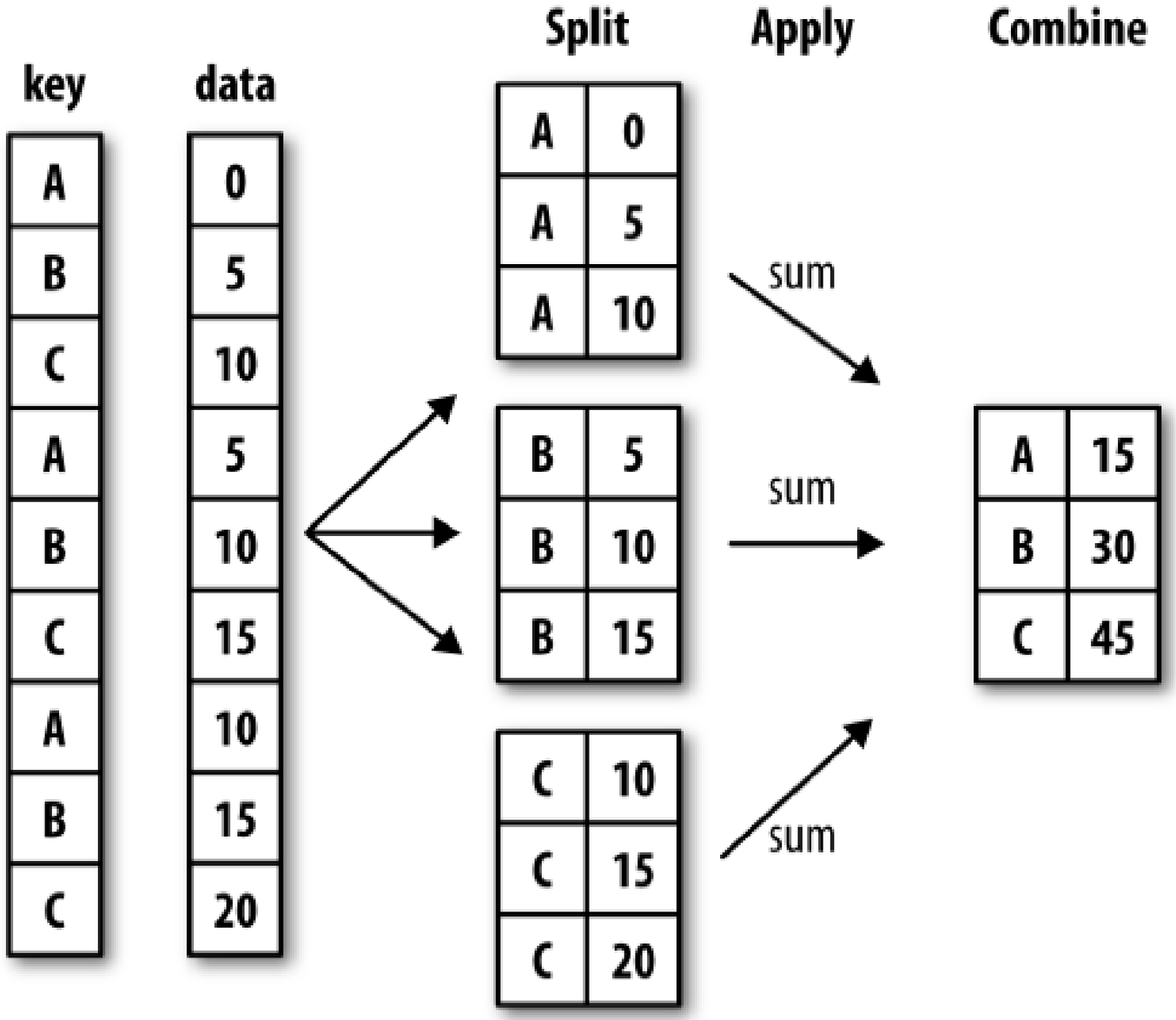
Agregação com Pandas

OrderDetailID	OrderID	ProductID	Quantity
1	10248	1	2
2	10248	2	10
3	10248	7	5
4	10249	4	5
5	10249	1	4
6	10250	2	5
7	10250	1	6
8	10250	4	15

GROUP BY ProductID
Somando a coluna
Quantity;

ProductID	Qtd
1	12
2	15
4	20
7	5

split-apply-combine (por Hadley Wickham)



Exemplo de Group By com Pandas

```
In [10]: df = pd.DataFrame({'key1' : ['a', 'a', 'b', 'b', 'a'],
   ....:                    'key2' : ['one', 'two', 'one', 'two', 'one'],
   ....:                    'data1' : np.random.randn(5),
   ....:                    'data2' : np.random.randn(5)})

In [11]: df
Out[11]:
      data1     data2 key1 key2
0 -0.204708  1.393406    a  one
1  0.478943  0.092908    a  two
2 -0.519439  0.281746    b  one
3 -0.555730  0.769023    b  two
4  1.965781  1.246435    a  one

In [12]: grouped = df['data1'].groupby(df['key1'])

In [13]: grouped
Out[13]: <pandas.core.groupby.SeriesGroupBy object at 0x7faa31537390>
```

```
In [14]: grouped.mean()
Out[14]:
key1
a    0.746672
b   -0.537585
Name: data1, dtype: float64
```

Analogia com SQL:

Agregação com duas colunas

```
In [15]: means = df['data1'].groupby([df['key1'], df['key2']]).mean()

In [16]: means
Out[16]:
key1  key2
a     one      0.880536
      two      0.478943
b     one     -0.519439
      two     -0.555730
Name: data1, dtype: float64
```

Curiosidade

```
In [17]: means.unstack()
Out[17]:
key2         one       two
key1
a       0.880536  0.478943
b      -0.519439 -0.555730
```

Analogia com SQL:

Agregação com vetor do tamanho do índice

- Equivalente a adicionar duas colunas ao dataframe
  - e depois realizar a agregação por estas colunas

```
In [18]: states = np.array(['Ohio', 'California', 'California', 'Ohio', 'Ohio'])

In [19]: years = np.array([2005, 2005, 2006, 2005, 2006])

In [20]: df['data1'].groupby([states, years]).mean()
Out[20]:
California  2005    0.478943
            2006   -0.519439
Ohio        2005   -0.380219
            2006    1.965781
Name: data1, dtype: float64
```

Aplicando a métrica em todas as colunas de dados

```
In [21]: df.groupby('key1').mean()
Out[21]:
          data1     data2
key1
a      0.746672  0.910916
b     -0.537585  0.525384
```

Analogia com SQL:

```
In [22]: df.groupby(['key1', 'key2']).mean()
Out[22]:
               data1     data2
key1 key2
a    one    0.880536  1.319920
     two    0.478943  0.092908
b    one   -0.519439  0.281746
     two   -0.555730  0.769023
```

Analogia com SQL:

Mostrando o tamanho dos grupos

```
In [23]: df.groupby(['key1', 'key2']).size()
Out[23]:
key1  key2
a     one     2
      two     1
b     one     1
      two     1
dtype: int64
```

Agrupando com uma função

- As funções são aplicadas sobre os valores do índice
  - e o valor retornado dá nome aos grupos

	a	b	c	d	e
Joe	1.007189	-1.296221	0.274992	0.228913	1.352917
Steve	0.886429	-2.001637	-0.371843	1.669025	-0.438570
Wes	-0.539741	NaN	NaN	-1.021228	-0.577087
Jim	0.124121	0.302614	0.523772	0.000940	1.343810
Travis	-0.713544	-0.831154	-2.370232	-1.860761	-0.860757

```
In [44]: people.groupby(len).sum()
Out[44]:
```

	a	b	c	d	e
3	0.591569	-0.993608	0.798764	-0.791374	2.119639
5	0.886429	-2.001637	-0.371843	1.669025	-0.438570
6	-0.713544	-0.831154	-2.370232	-1.860761	-0.860757

Outras métricas para aplicar na agregação

Função	Descrição
count	Número de valores não NA no grupo
sum	Soma de valores não NA
mean	Média de valores não NA
median	Mediana de valores não NA
std, var	Desvio padrão e variância não enviesada (n-1 no denominador)
min, max	Mínimo e Máximo de valores não NA
prod	Produto de valores não NA
first, last	Primeiro e último valores não NA

Aplicando várias métricas ao mesmo tempo

```
   total_bill   tip smoker  day    time  size   tip_pct
0       16.99  1.01     No  Sun  Dinner     2  0.059447
1       10.34  1.66     No  Sun  Dinner     3  0.160542
2       21.01  3.50     No  Sun  Dinner     3  0.166587
3       23.68  3.31     No  Sun  Dinner     2  0.139780
4       24.59  3.61     No  Sun  Dinner     4  0.146808
5       25.29  4.71     No  Sun  Dinner     4  0.186240
```

```
In [60]: grouped = tips.groupby(['day', 'smoker'])
In [61]: grouped_pct = grouped['tip_pct']
In [63]: grouped_pct.agg(['mean', 'std', peak_to_peak])
Out[63]:
                mean       std  peak_to_peak
day  smoker
Fri  No      0.151650  0.028123      0.067349
     Yes     0.174783  0.051293      0.159925
Sat  No      0.158048  0.039767      0.235193
     Yes     0.147906  0.061375      0.290095
Sun  No      0.160113  0.042347      0.193226
     Yes     0.187250  0.154134      0.644685
```

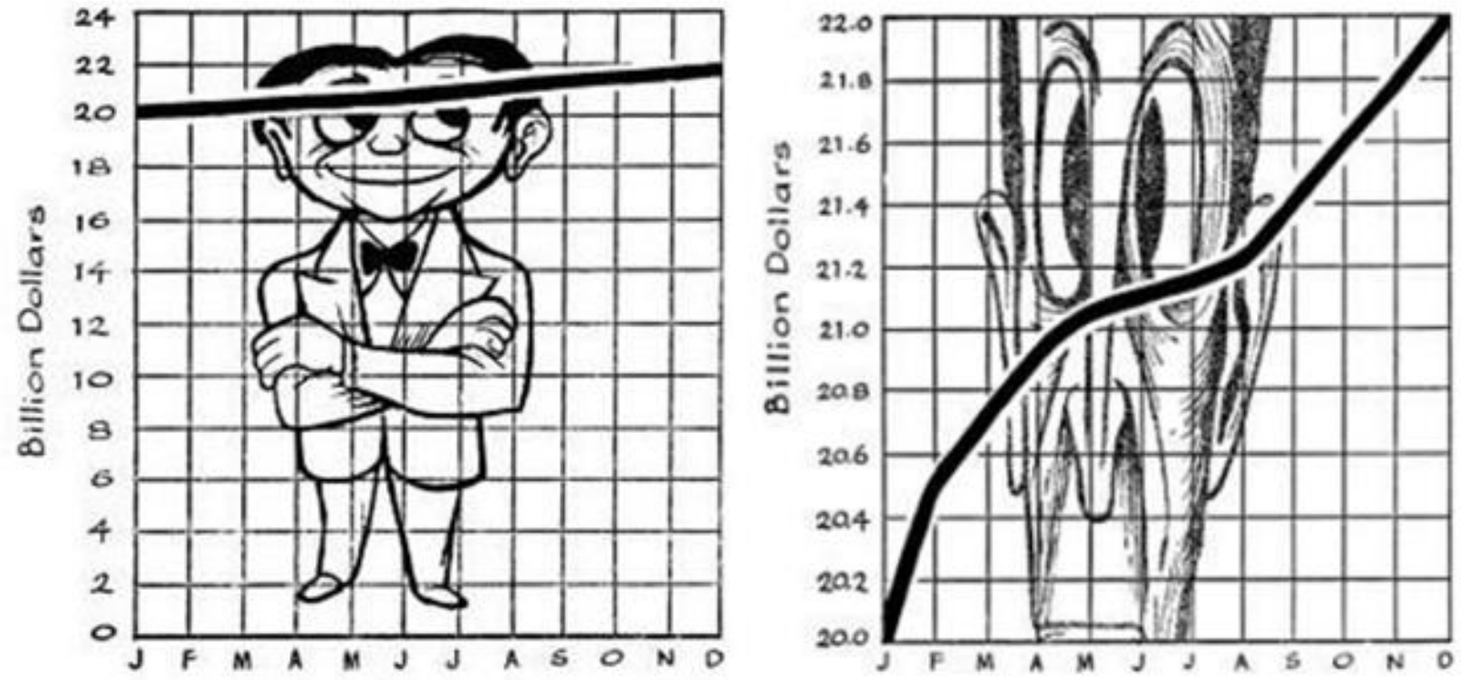
Várias métricas em diferentes colunas

```
In [72]: grouped.agg({'tip_pct' : ['min', 'max', 'mean', 'std'],
    ....:                 'size' : 'sum'})
Out[72]:
                tip_pct                                     size
                    min       max      mean       std      sum
day  smoker
Fri  No        0.120385  0.187735  0.151650  0.028123        9
     Yes       0.103555  0.263480  0.174783  0.051293       31
Sat  No        0.056797  0.291990  0.158048  0.039767      115
     Yes       0.035638  0.325733  0.147906  0.061375      104
Sun  No        0.059447  0.252672  0.160113  0.042347      167
     Yes       0.065660  0.710345  0.187250  0.154134       49
Thur No        0.072961  0.266312  0.160298  0.038774      112
     Yes       0.090014  0.241255  0.163863  0.039389       40
```

Intervalo
(20 min)

Conceitos básicos de Estatística Descritiva

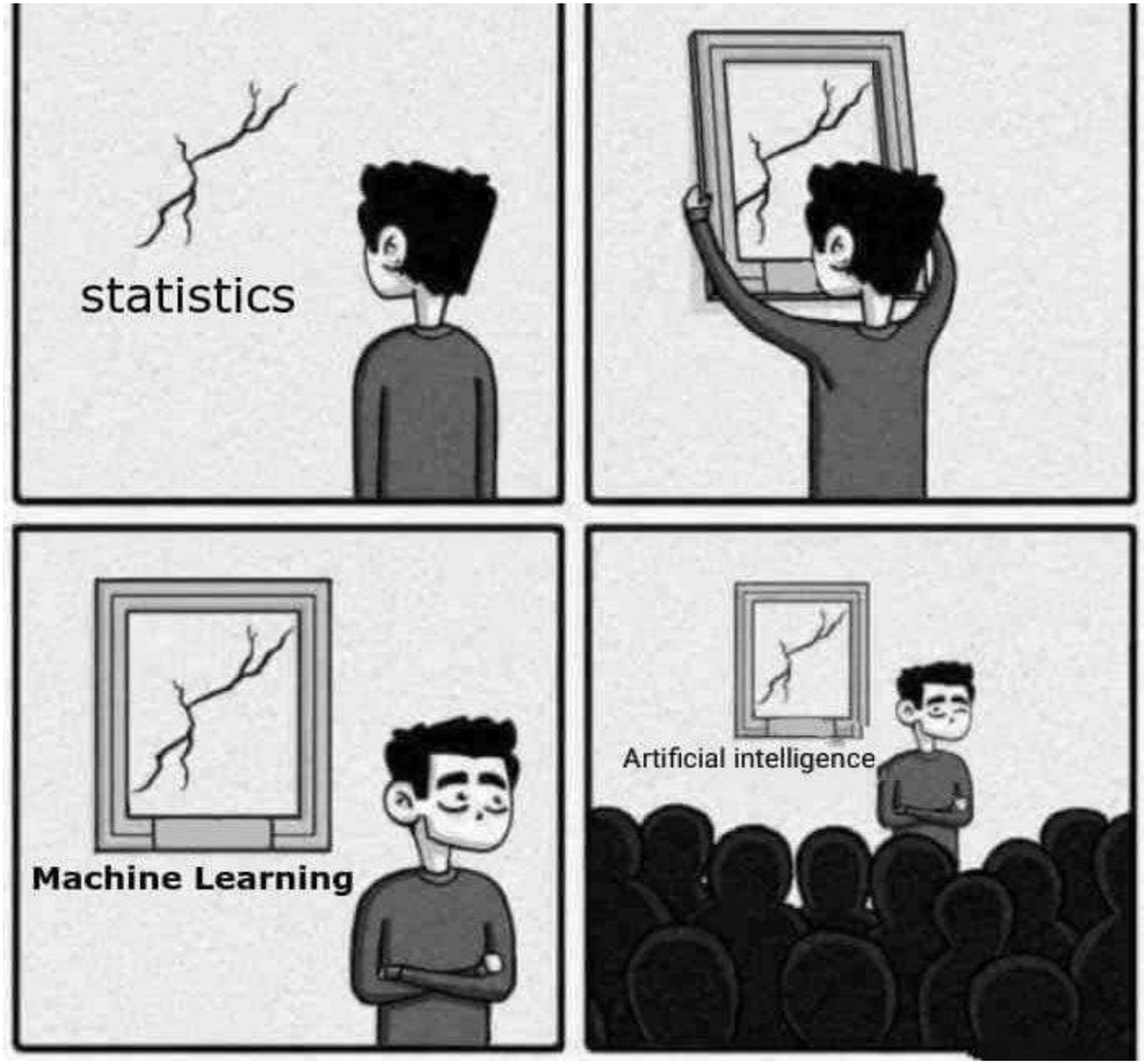
Livro: How To Lie With Statistics (Darrell Huff)



"É fácil mentir com estatísticas, mas é difícil dizer a verdade sem elas"

(Andrejs Dunkels / Matemático / 1939-1998)

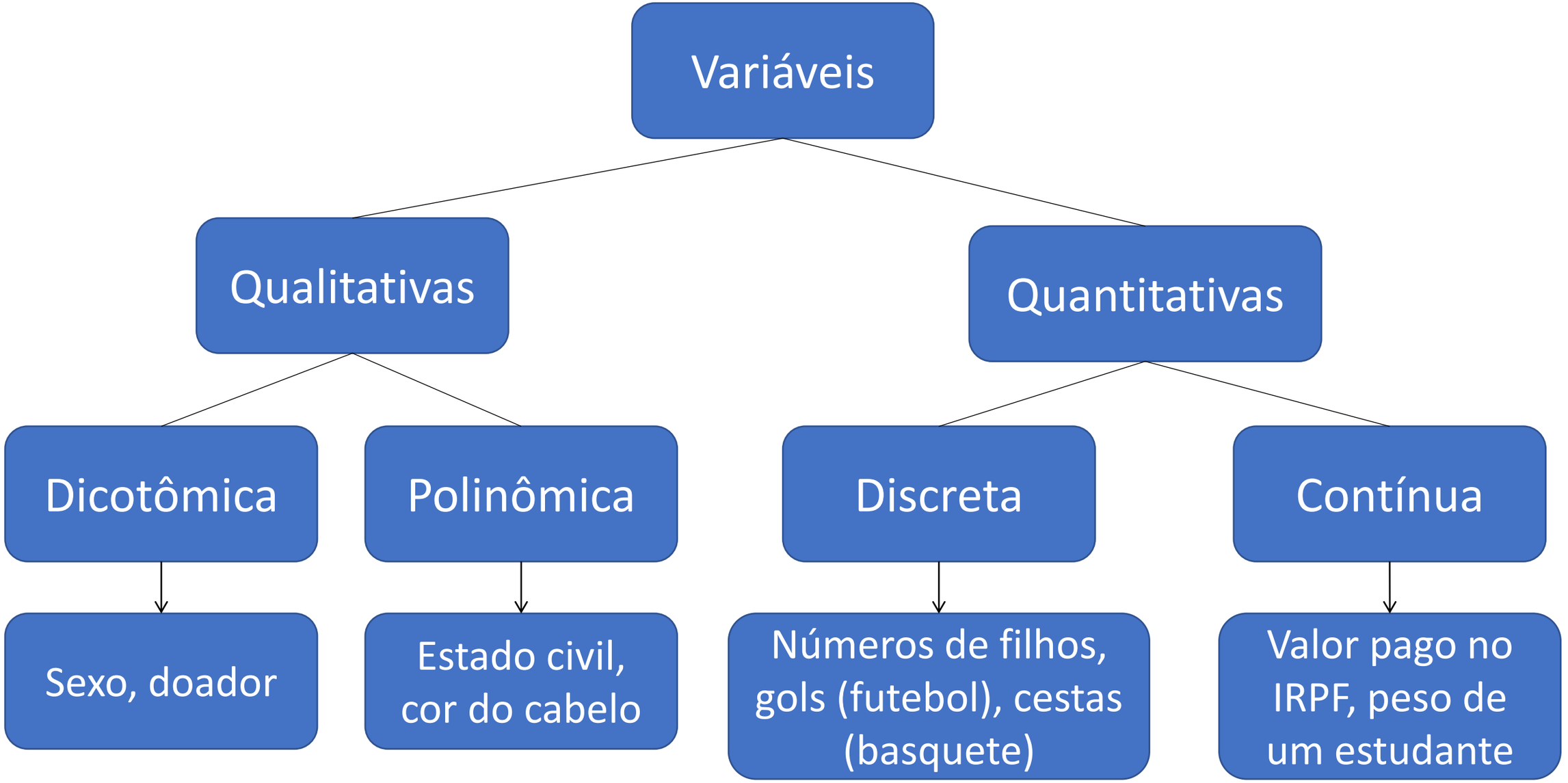
A Estatística é a base de outras ferramentas de análise de dados



Variáveis

- São características que podem ser observadas
- Quando coletar variáveis por meio de perguntas
  - Há quanto o Sr.(a) trabalha nessa empresa?
  - Qual seu estado civil?
- Elaborar perguntas que aceitam respostas precisas
  - Há quanto o Sr.(a) trabalha nessa empresa? ____ anos completos
  - Qual seu estado civil? ( ) solteiro ( ) casado ( ) viúvo ...
- Podem ser quantitativas ou qualitativas (categóricas)

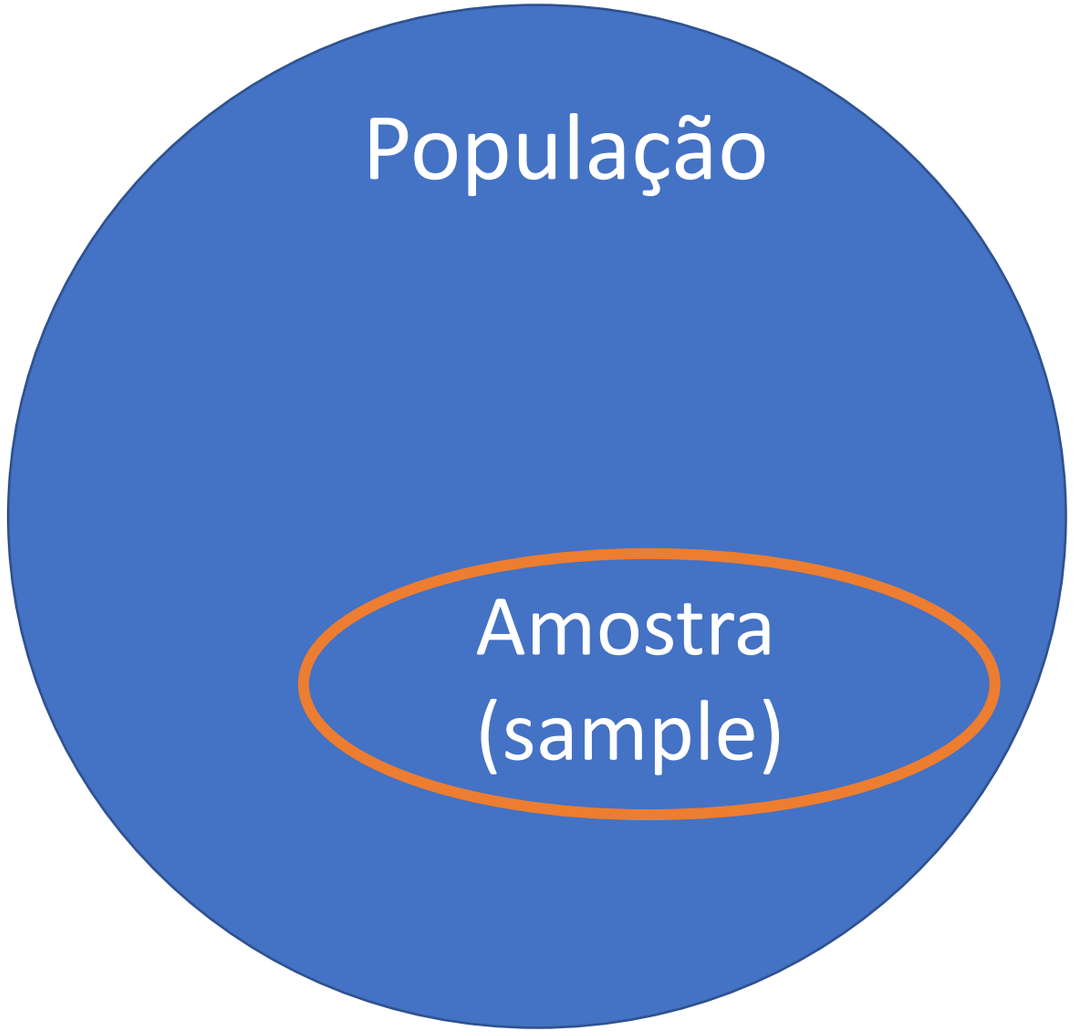
Variáveis

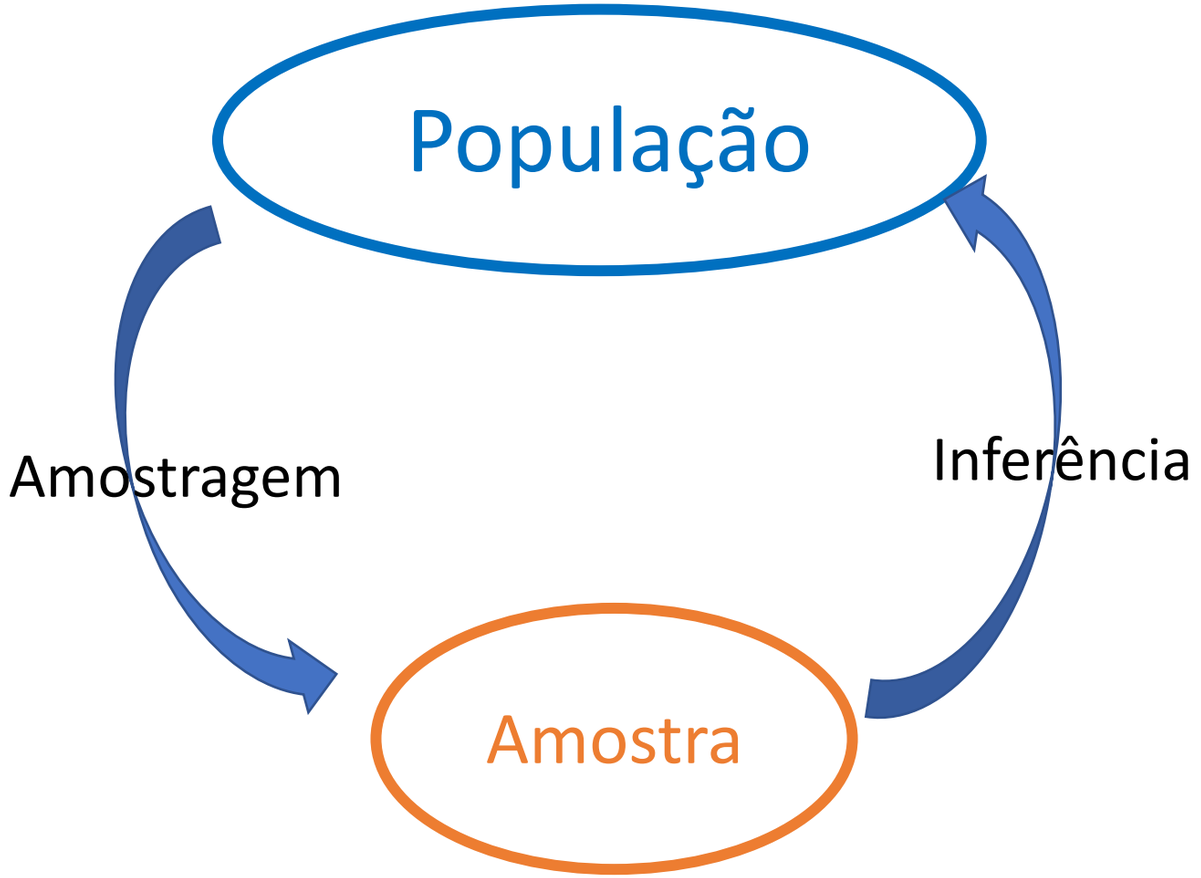

Variáveis
Qualitativas
Quantitativas
Dicotômica
Polinômica
Discreta
Contínua
Sexo, doador
Estado civil, cor do cabelo
Números de filhos, gols (futebol), cestas (basquete)
Valor pago no IRPF, peso de um estudante

População

- População Alvo
  - Conjunto de elementos que se quer abranger no estudo.
    - ✓ Exemplo: O conjunto de todos os indivíduos de uma Empresa, num determinado tempo.
- População Acessível (ou simplesmente População)
  - Conjunto de elementos (indivíduos) observáveis
    - ✓ Exemplo: funcionários que não estão de férias nem licença
      - Veja que a variável tempo é relevante.

Amostragem (Sampling)





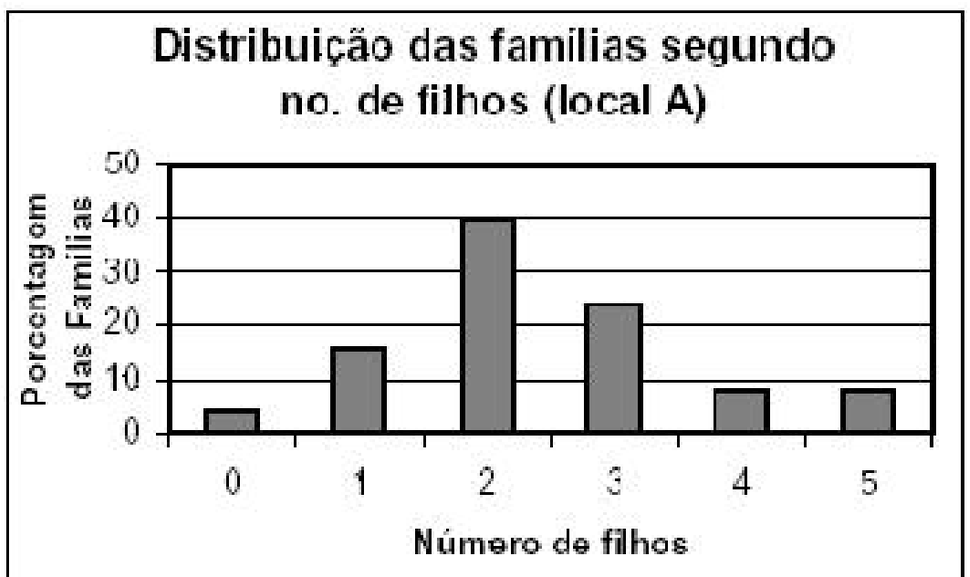
Amostragem

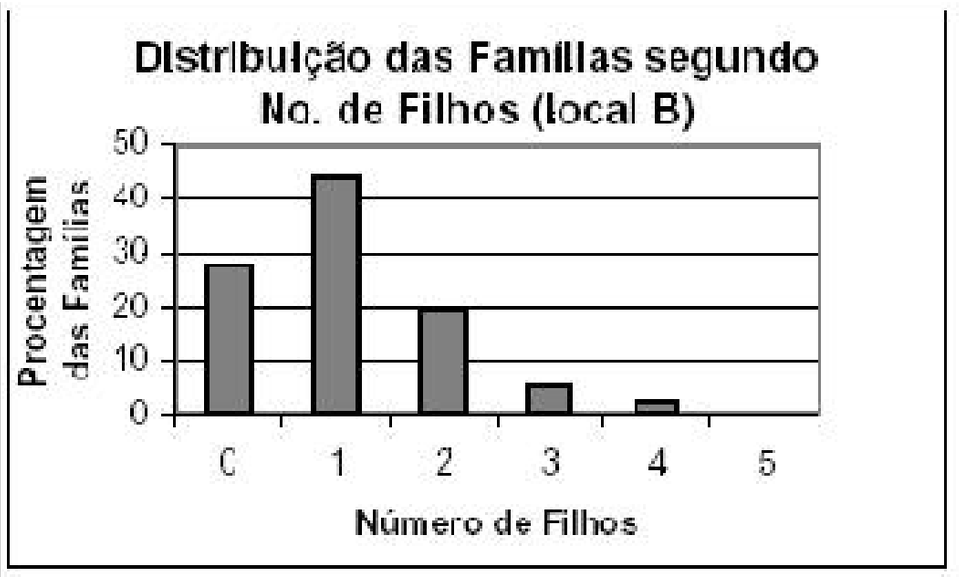
- Por que amostrar ?
  - Viabilizar o custo.
    - ✓ Entrevistar 1000 pessoas para fazer uma pesquisa eleitoral quinzenal com margem de erro de 5%.
  - Não consumir todo o estoque (experimentar uma sopa)
- Uma amostra deve ter as mesmas características da população subjacente (que está representando)
- Amostragem pode ser:
  - Com reposição: Um membro poderá ser escolhido mais de uma vez
    - ✓ Retirar bolas de uma urna (devolvendo-as)
  - Sem reposição: Um membro poderá ser escolhido apenas uma vez
    - ✓ Loteria, sorteio, bingo
- Útil para elaborar estimativas

Distribuição de Frequências

- Compreende a organização dos dados de acordo com as ocorrências dos diferentes resultados observados

Número da classe	Salário do mês – R$	Número de empregados
1	1 000 → 2 000	20
2	2 000 → 3 000	18
3	3 000 → 4 000	9
4	4 000 → 5 000	3

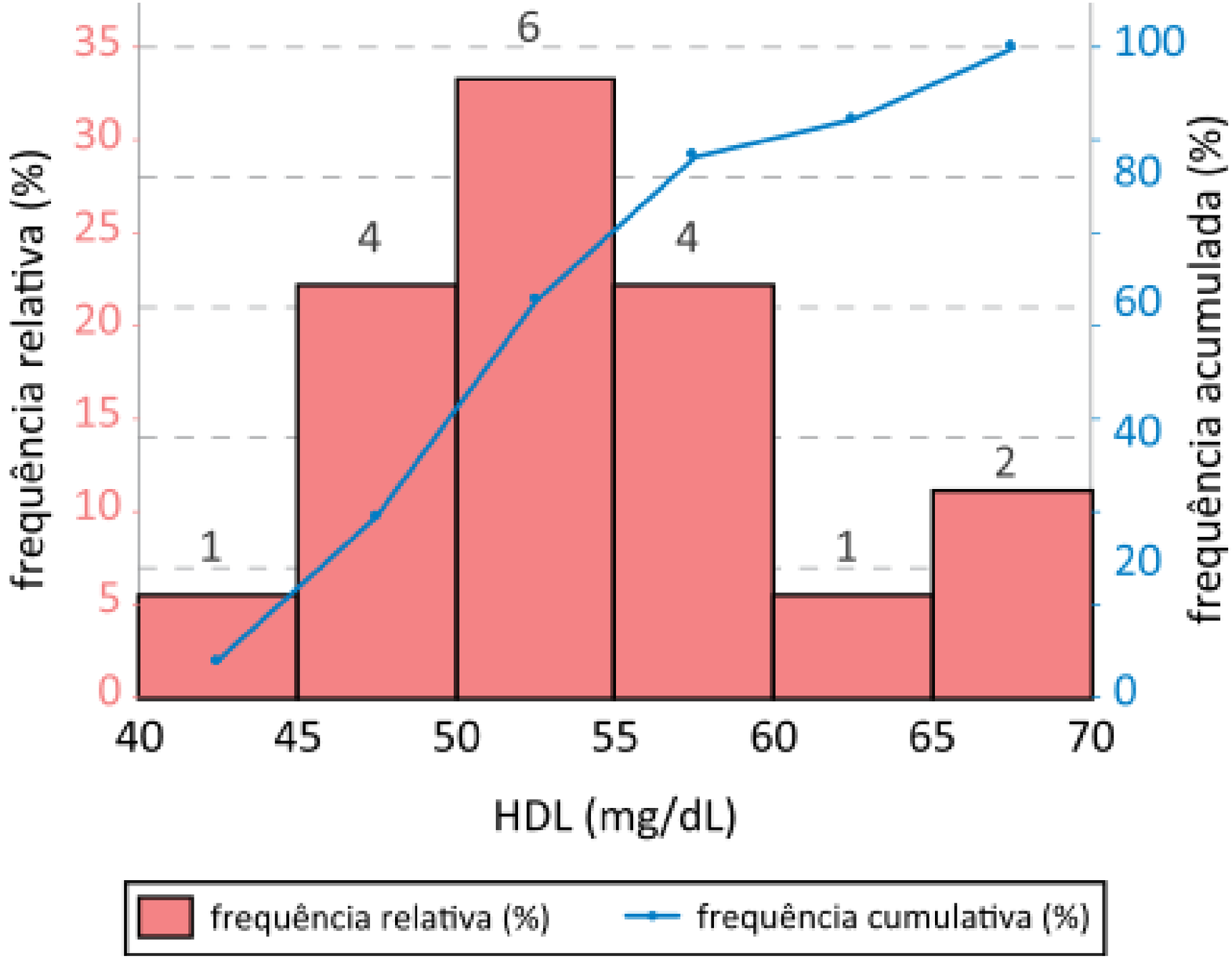




Distribuição de frequências (Variável contínua)

coleta aleatória		dados ordenados		intervalo	frequência	classe
pacientes	HDL (mg/dL)	pacientes	HDL (mg/dL)			
1	55	7	44	HDL<45	1	1
2	57	8	45	45≤HDL <50	4	2
3	53	16	46			
4	49	14	47			
5	54	4	49			
6	52	9	50	50≤HDL <55	6	3
7	44	10	52			
8	45	6	52			
9	50	13	53			
10	52	3	53			
11	55	5	54			
12	67	1	55	55≤HDL <60	4	4
13	53	11	55			
14	47	2	57			
15	65	18	59			
16	46	17	64	60≤HDL <65	1	5
17	64	15	65	65≤HDL	2	6
18	59	12	67			

Histograma


35
30
25
20
15
10
5
0
frequência relativa (%)
100
80
60
40
20
0
frequência acumulada (%)
6
4
4
2
1
1
40
45
50
55
60
65
70
HDL (mg/dL)
frequência relativa (%)
frequência cumulativa (%)

Medidas de tendência central

(Introdução ao R)

Média Aritmética

- Média aritmética

$$\mu = \frac{\sum_1^n X_i}{n} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

Mediana

- É o valor que divide uma **distribuição** ao meio.
  - Metade dos valores (os menores) fica de um lado e a metade (os maiores) fica de outro.
- Procedimento de cálculo
  - Ordena-se os valores, e escolhe-se o valor do centro
    - ✓ Qual é a mediana de: 1 1 1 4 20 680 2300
  - Com uma quantidade par de números, calcula-se a média dos dois números centrais
    - ✓ Qual é a mediana de: 1 1 1 4 20 680
- Consegue filtrar valores extremos (outliers)

Média vs Mediana: Exemplo fictício

- Imagine você num bar com mais 8 clientes presentes
  - Considere que a renda anual de cada um dos clientes seja esta:
  - 15 15 16 18 20 20 21 21 84
    - ✓ Média = 25.5
    - ✓ Mediana = 20
- O Bill Gates entra no bar (renda anual de 10 milhões)
  - 15 15 16 18 20 20 21 21 24 $10x10^6$
    - ✓ Média ~ 1.1 milhão
    - ✓ Mediana = 20
- Transmitiria a mensagem correta sobre o ambiente,
  - dizer que no bar onde você toma cerveja a renda média anual dos frequentadores é
    - ✓ um pouco mais de 1 milhão ?

Moda

- É o valor mais frequente de uma distribuição de frequência
- Útil como tendência central para variáveis qualitativas
  - **Sim, Sim, Sim, Sim**, Não, Não, Não sei, Não sei

Média vs Mediana

- A mediana consegue filtrar valores extremos (outliers)
- Uma boa análise estatística
  - Apresenta as duas métricas
- Qual é a mais apropriada depende se os valores extremos são outliers
  - Ou são parte da mensagem que você quer transmitir

Desvio padrão e Variância

- São medidas de dispersão (espalhamento)
  - Em relação ao valor médio
- São medidas quantitativas para expressar
  - o quanto os elementos distam da média
- Exemplo:
  - Peso médio dos passageiros de um avião que carrega competidores de uma maratona;
  - Peso médio de passageiros de um voo comercial comum.
    - ✓ Crianças, jovens, adultos
  - O peso pode ser parecido, mas a dispersão dos pesos em relação a média será parecida ?

Variância: Média do Desvio quadrático

Desvio quadrático

Grupo 1	Altura (μ = 175 cm)	Média = Valor absoluto de $(x_n - \mu)^*$	$(x_n - \mu)^2$
Nick	185	10	100
Elana	165	10	100
Dinah	170	5	25
Rebecca	173	2	4
Ben	183	8	64
Charu	175	0	0
		Total = 35	Total = 293
			Variância = $^{293}/_{6}$ = 48,83
			Desvio padrão = $\sqrt{48,83}$ = 6,988 = 7

Fonte: WHEELAN, C. Estatística, o que é, para que serve, como funciona. Zahar, 2016.

Desvio padrão

- Variância $$v = \frac{\sum(X_i - \mu)^2}{n-1}$$
- Desvio padrão $$\sigma = \sqrt{\frac{\sum(X_i - \mu)^2}{n-1}}$$
- Ou $$\sigma = \sqrt{v} \quad ou \quad \sigma^2 = v$$

- Obs.: quando se trata de toda a população, alguns autores
  - Usam N no lugar de n-1 da fórmula.

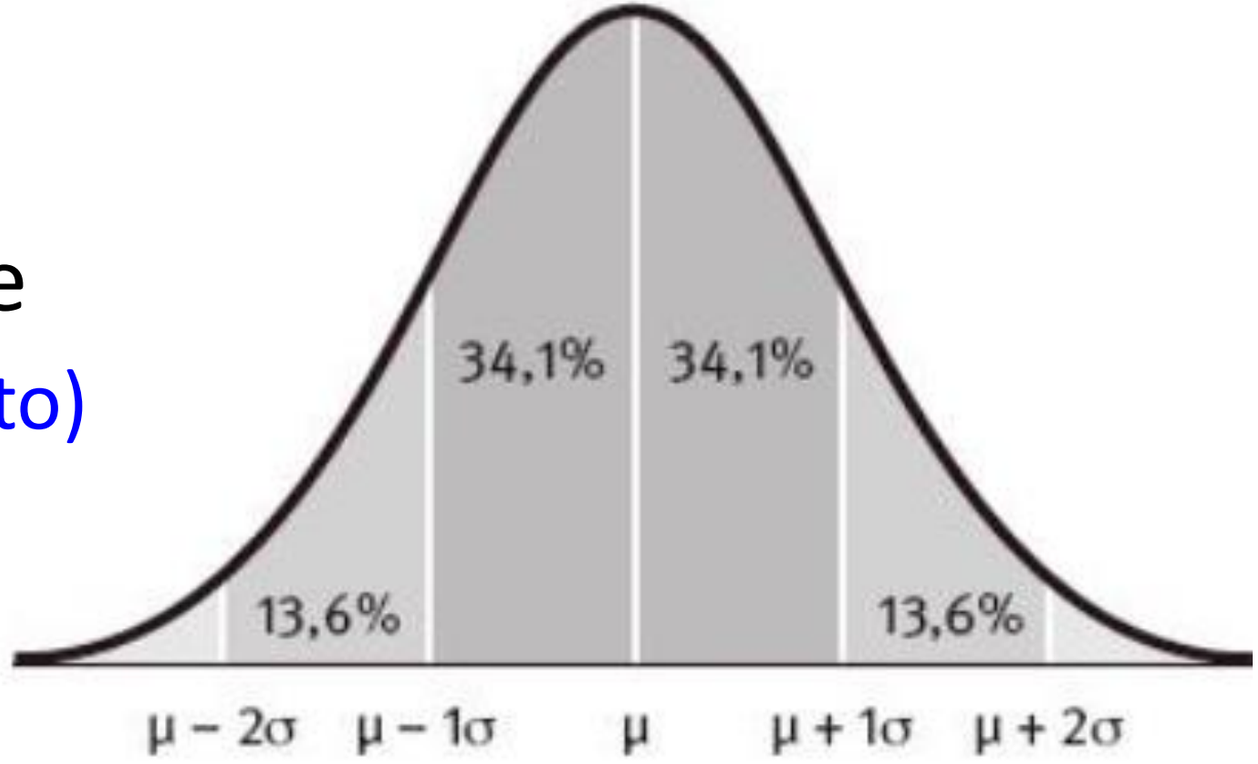
Os dois grupos têm a mesma média

Grupo 2	Altura ($\mu = 175$ cm)	Média = Valor absoluto de $(x_n - \mu)^*$	$(x_n - \mu)^2$
Sahar	163	12	144
Maggie	170,5	4,5	20,25
Faisal	174	1	1
Ted	175	0	0
Jeff	180,5	5,5	30,25
Narciso	187	12	144
		Total = 35	Total = 339,5
			Variância = $339{,}5/6 = 56{,}583$
			Desvio padrão = $\sqrt{56{,}583} = 7{,}522 = 7{,}5$

Fonte: WHEELAN, C. Estatística, o que é, para que serve, como funciona. Zahar, 2016.

Intepretação do desvio padrão

- Numa distribuição Normal (curva caracterizada por μ e σ)
  - 68,2% das medições estão dentro de 1 σ da média
  - 95,4% estão dentro de 2 σ
  - 99,7% estão dentro de 3 σ
- Útil quando você desconhece
  - Os valores envolvidos (contexto)

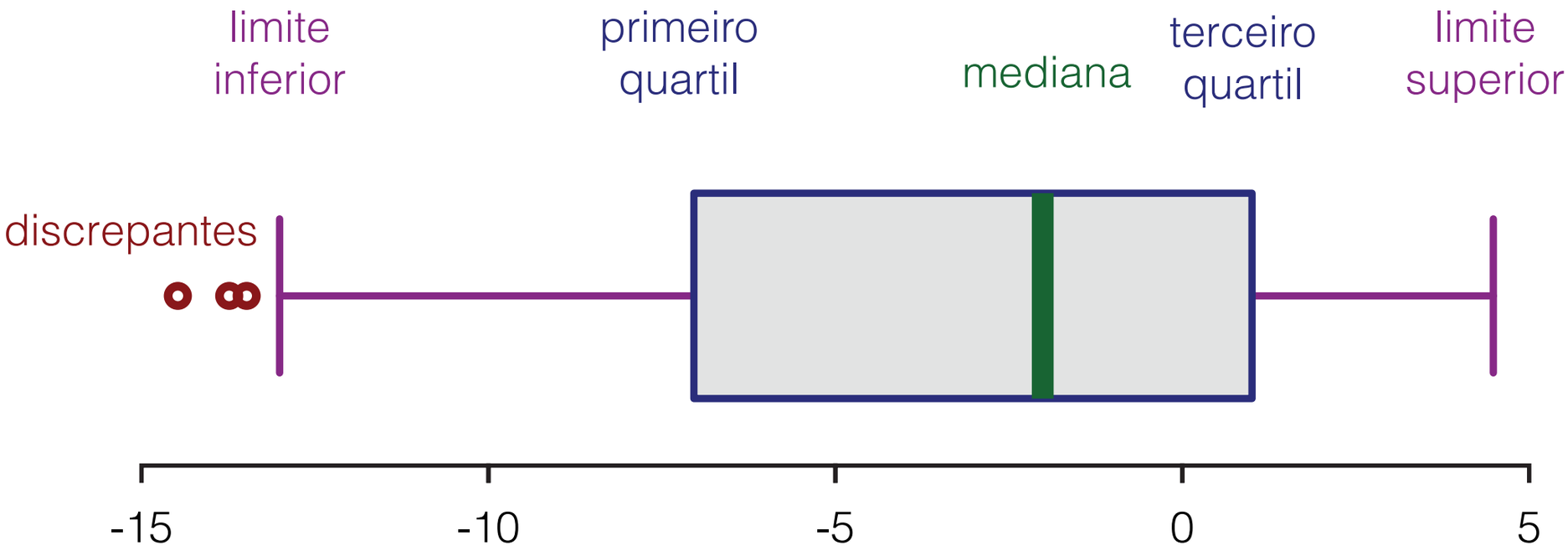


Fonte: WHEELAN, C. Estatística, o que é, para que serve, como funciona. Zahar, 2016.

Quartis

- Divide uma distribuição em 4 partes iguais
  - Cada parte tem ¼ da amostra (ou da população)
- Como calcular os quartis
- $Q_{1/4}$ = arredondar 0.25*(N+1)
- $Q_{2/4}$
  - Se N for par:
    - $Q_{2/4}$ = média dos itens na posição (N/2) e (N/2)+1
  - Se N for ímpar:
    - $Q_{2/4}$ = o item na posição (N+1)/2
- $Q_{3/4}$ = arredondar 0.75*(N+1)

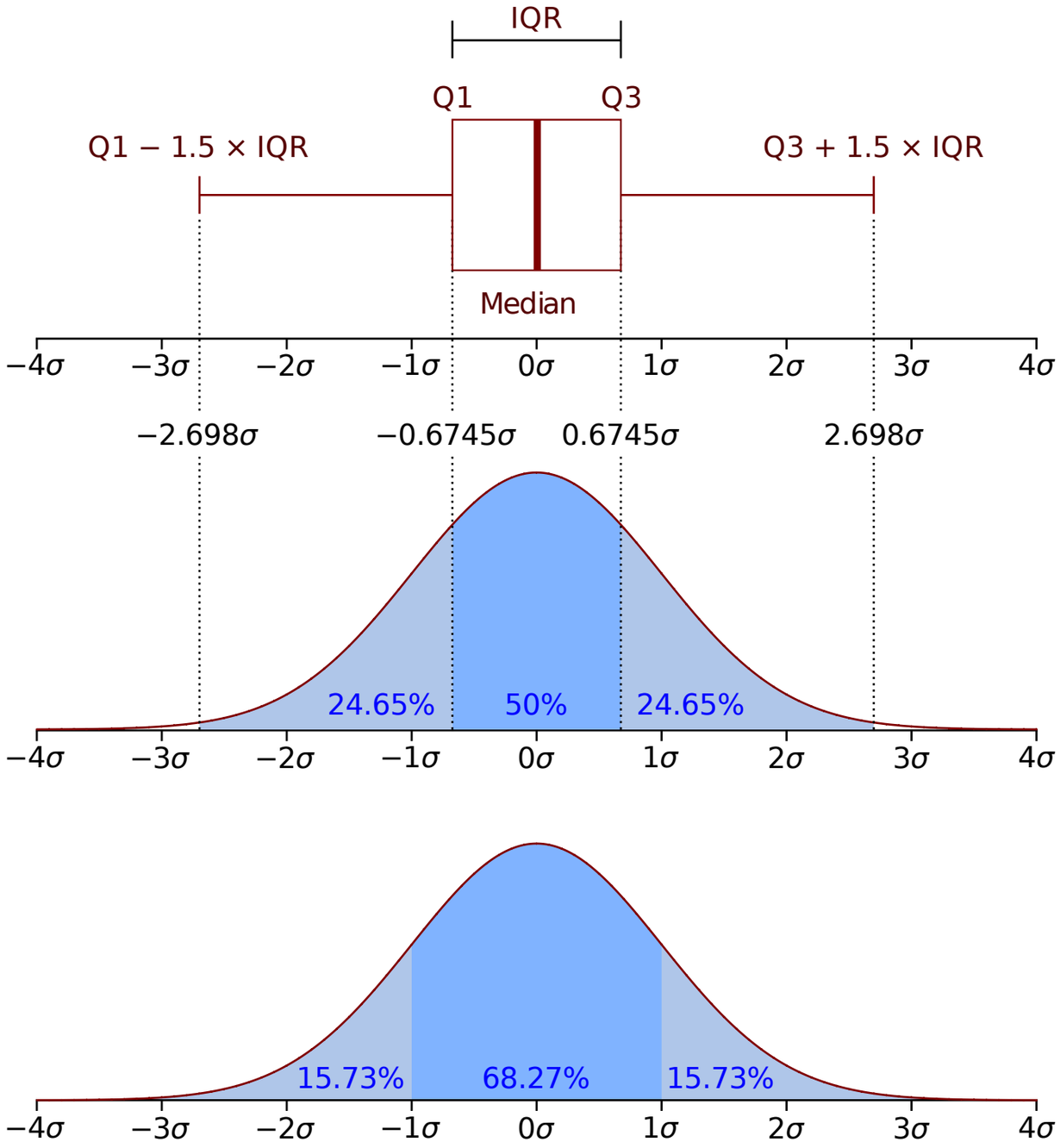
Diagrama de Caixa (boxplot)



Fonte: https://commons.wikimedia.org/wiki/File:Elements_of_a_boxplot_pt.svg

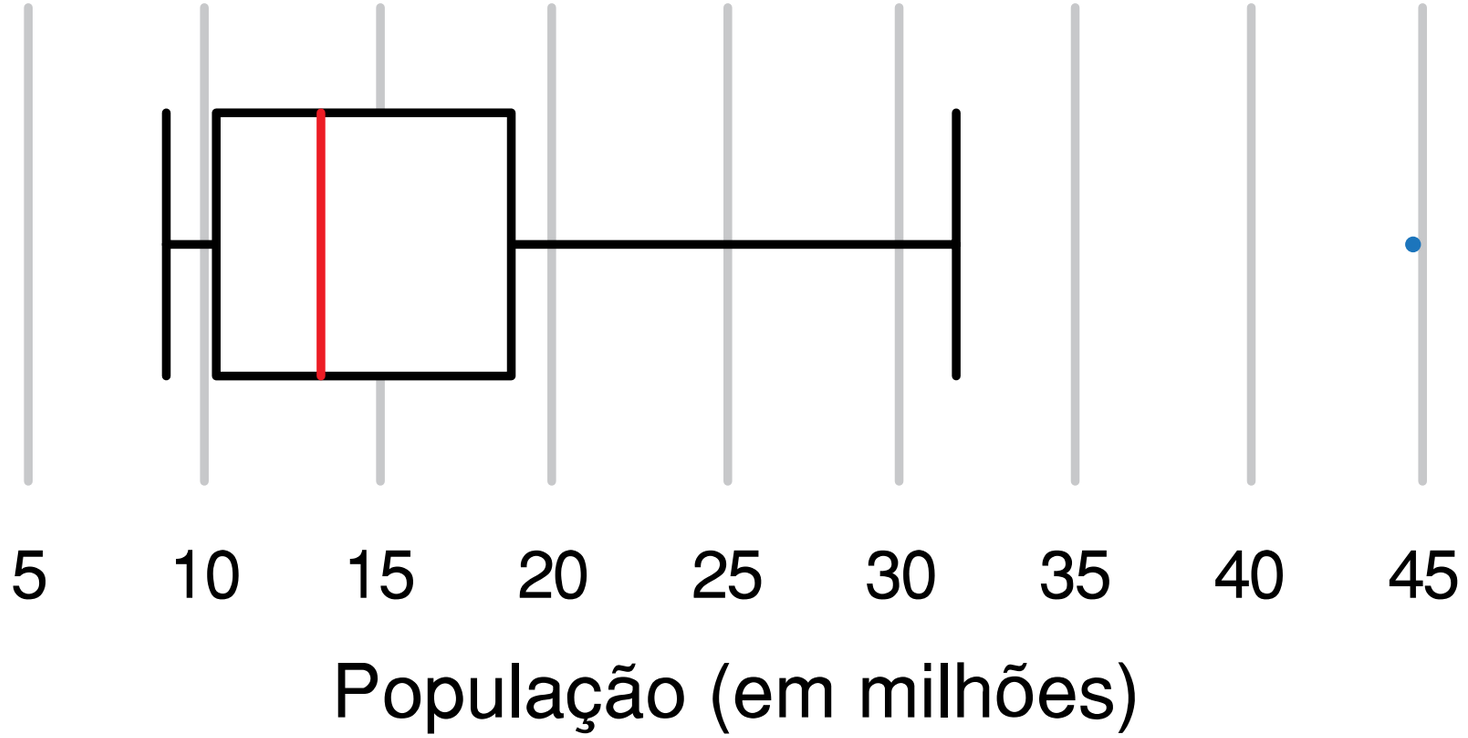
Diagrama de Caixa (boxplot)

- Distribuição Normal
  - $N(0, 1\sigma^2)$



Fonte: https://commons.wikimedia.org/wiki/File:Diagrama_de_caixa_-_Popula%C3%A7%C3%A3o.svg

População dos estados brasileiros segundo o IBGE



Fonte 1: ftp://ftp.ibge.gov.br/Estimativas_de_Populacao/Estimativas_2016/estimativa_dou_2016_20160913.pdf

Fonte 2: https://commons.wikimedia.org/wiki/File:Diagrama_de_caixa_-_Popula%C3%A7%C3%A3o.svg

Distribuição de renda de duas localidades

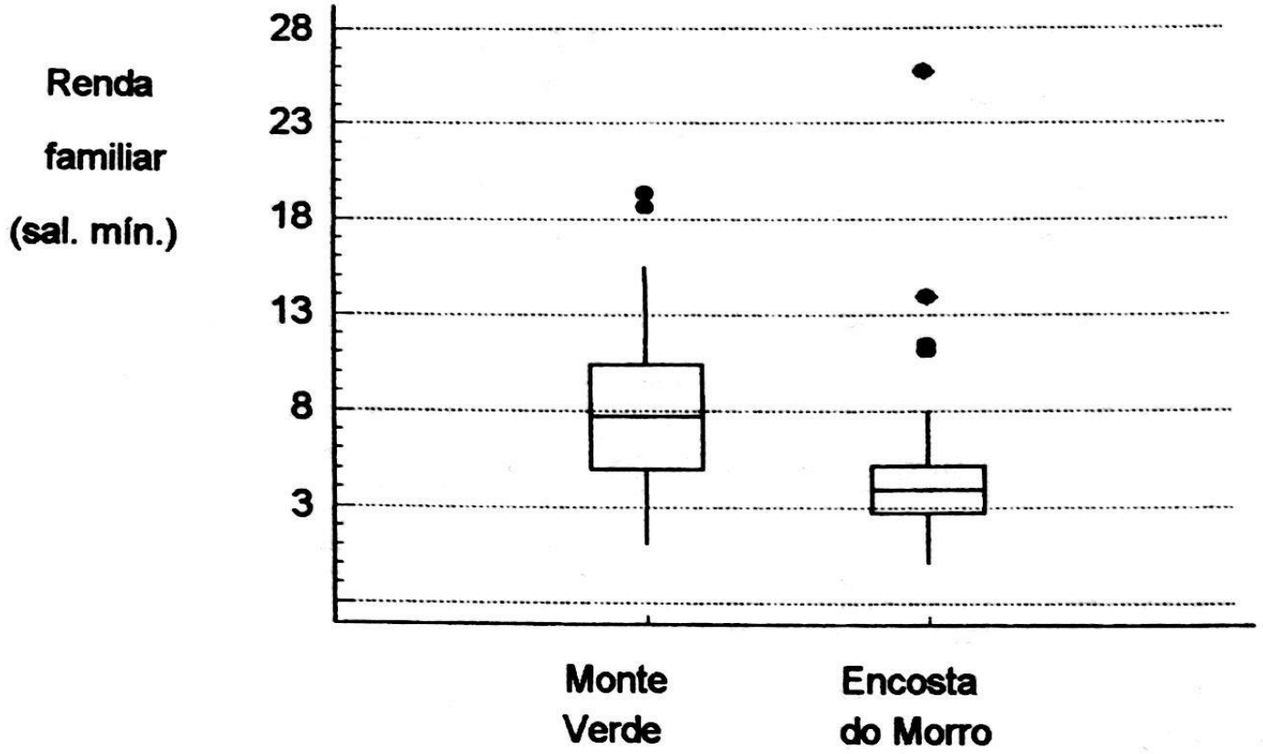


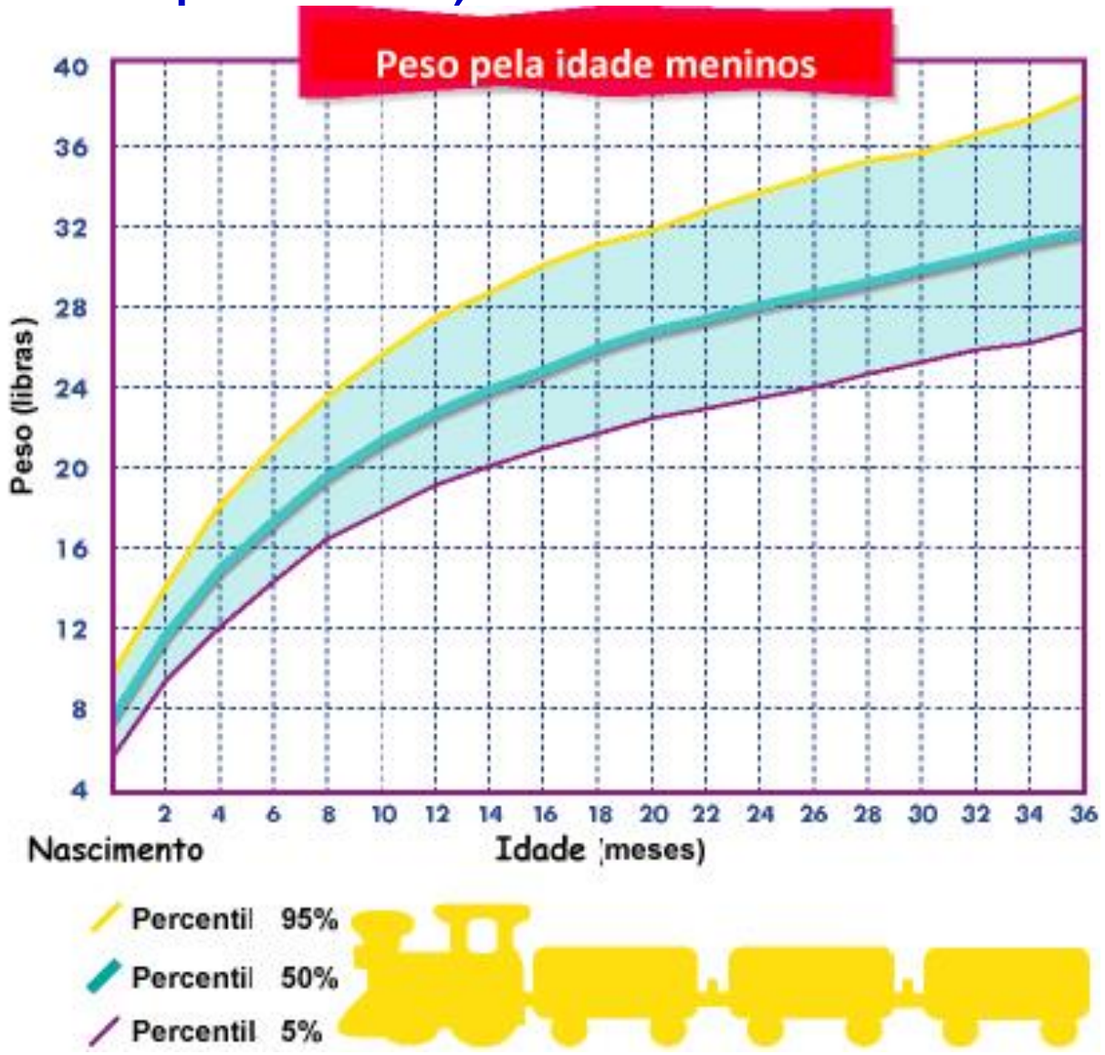
FIG. 6.10 Representação das distribuições de renda do Exemplo 6.4 em *diagramas em caixas*.

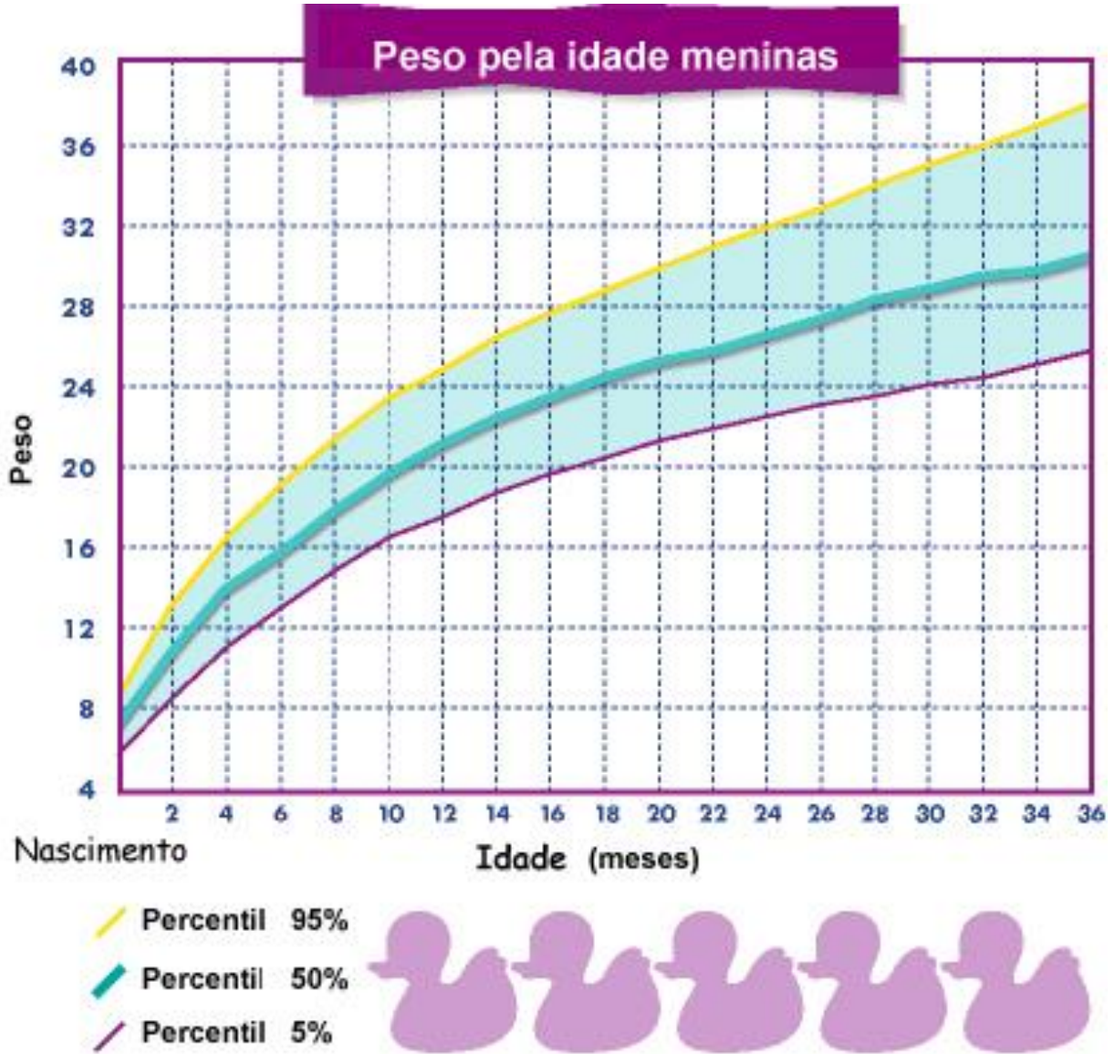
Fonte: BARBETTA, P. A. Estatística Aplicada às Ciências Sociais. UFSC, 6a. ed. 2006.

Fonte: https://www.agravidez.com/percentis.html

Percentil

- Divide-se a distribuição em 100 partes
  - 1º percentil, os 1% menores valores

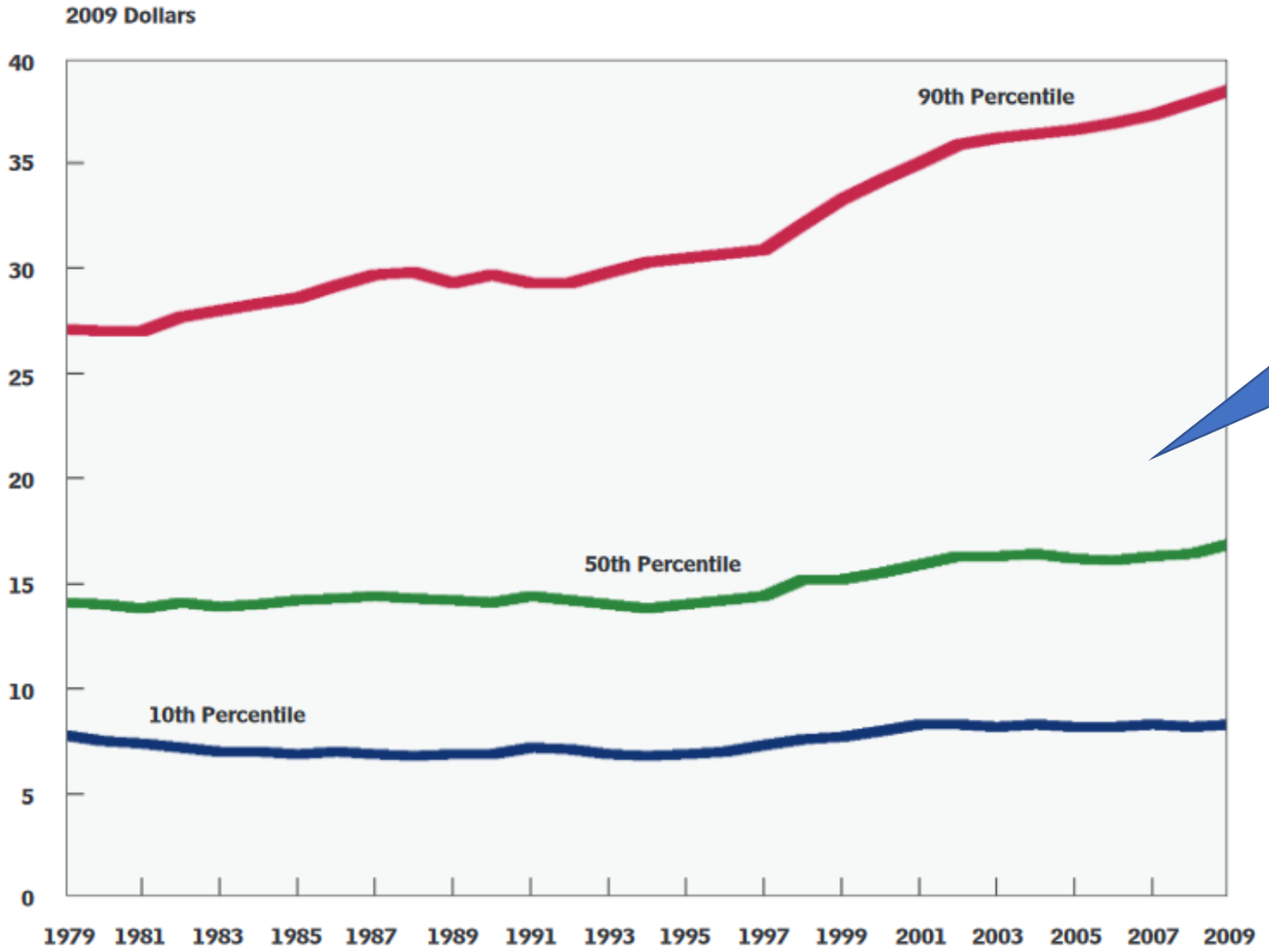




Examinar a saúde econômica da classe média americana

- Segundo Jeff Grogger
  - PhD em Economia e professor de Política Pública na Univ. de Chicago
- E Alan Krugger
  - Chefe do conselho de assessores econômicos do presidente Obama
- Duas boas medidas para avaliar a saúde econômica da classe média:
  - As mudanças no salário mediano (corrigido pela inflação) durante as últimas décadas; e
  - As mudanças nos salários no 25º e 75º percentis
    - ✓ Esses valores podem ser interpretados como os limites inferior e superior da classe média
- Renda é diferente de salário. Qual delas é mais apropriada ?

Examinar a saúde econômica da classe média americana



Hourly Wages at Selected Percentiles for Workers Ages 16 to 64

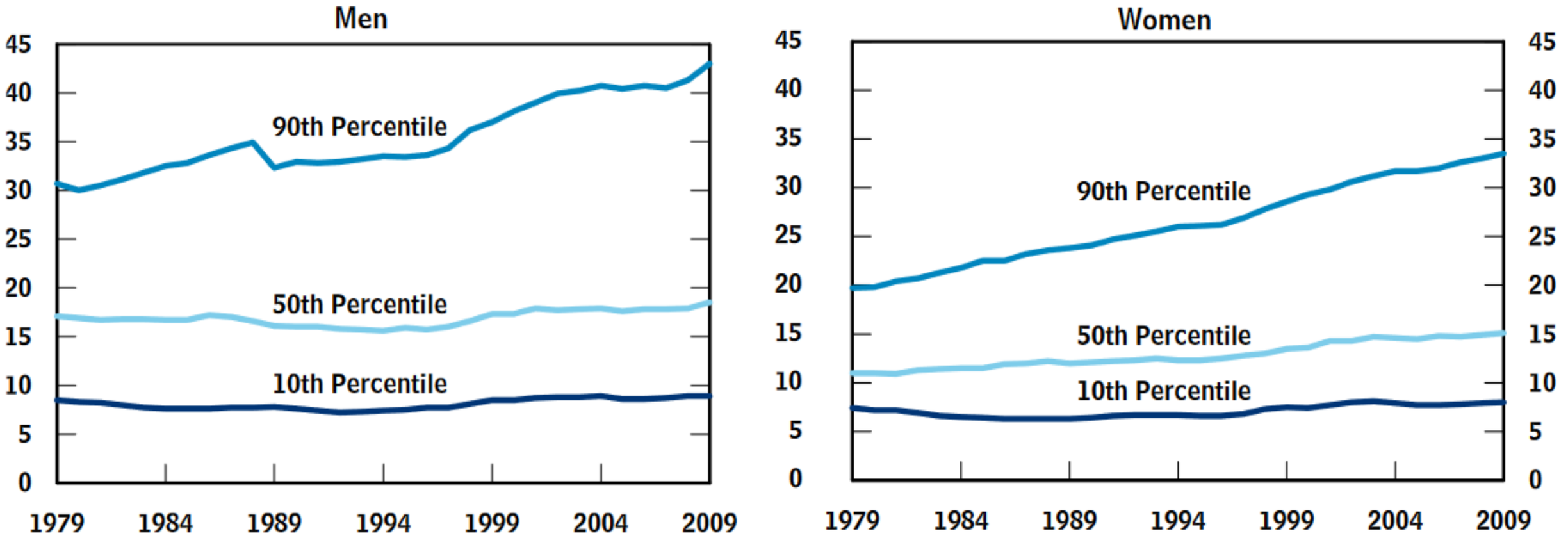
Compare o desempenho do 50º percentil com o desempenho do

Fonte: http://www.cbo.gov/sites/default/files/cbofiles/ftpdocs/120xx/doc12051/02-16-wagedispersion.pdf

Examinar a saúde econômica da classe média

Hourly Wages at Selected Percentiles for Men and Women Ages 16 to 64

(2009 dollars)



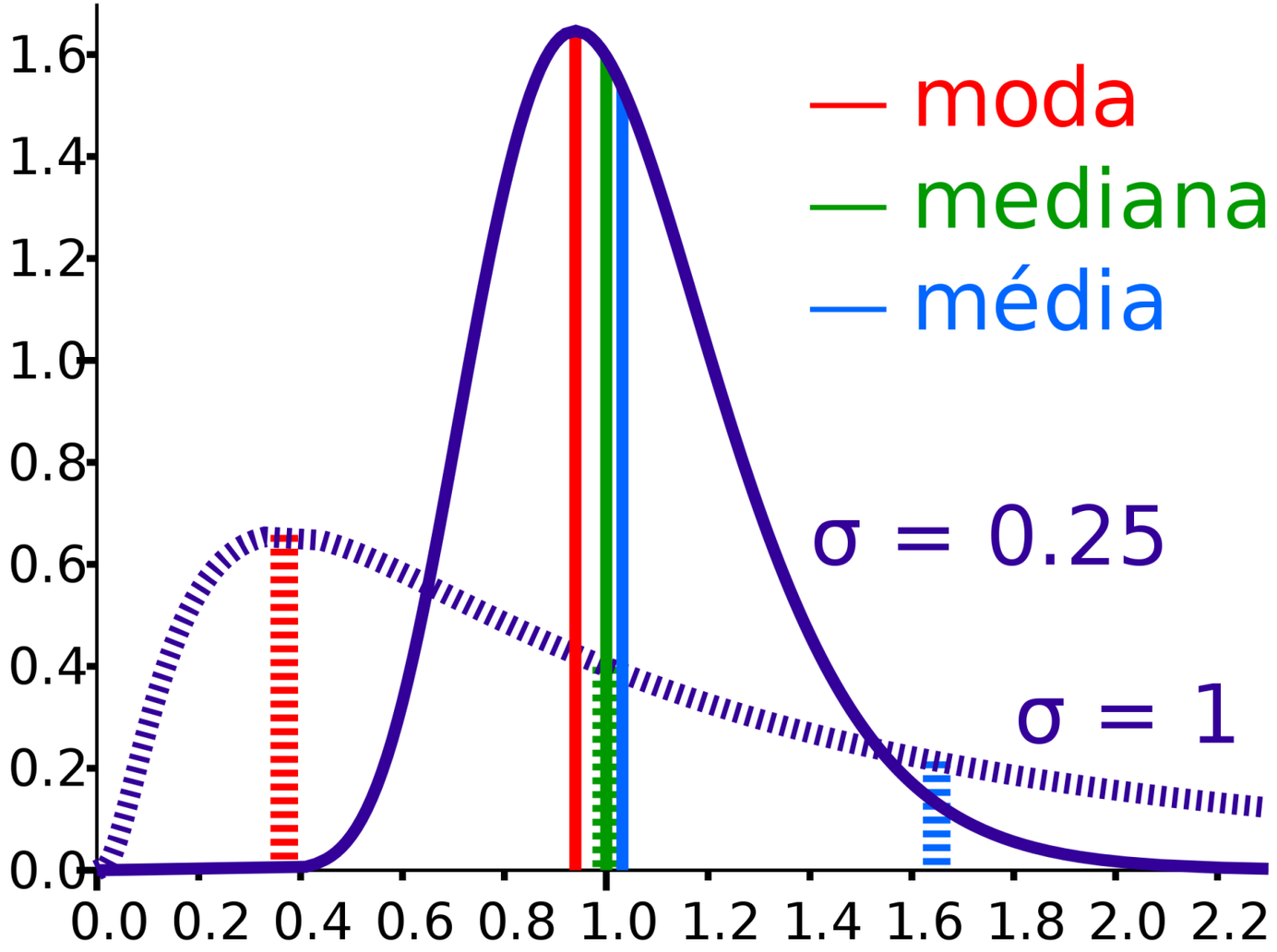
Source: Congressional Budget Office based on monthly data from Census Bureau, Current Population Survey, Outgoing Rotation Groups, 1979 to 2009.

Fonte: http://www.cbo.gov/sites/default/files/cbofiles/ftpdocs/120xx/doc12051/02-16-wagedispersion.pdf

Distribuições Assimétricas

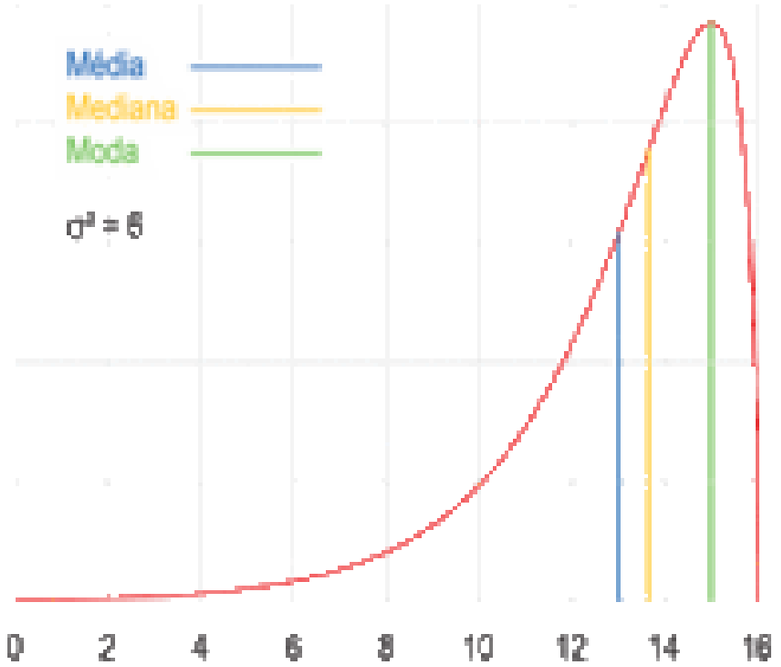
- Uma distribuição simétrica
  - Tem uma curva de frequência unimodal; e
  - Duas caudas simétricas em relação a uma linha vertical central
    - ✓ Nesta linha central estão a moda, média e mediana
- Numa distribuição assimétrica
  - Esses parâmetros não são coincidentes
    - ✓ A média sempre estará do lado da cauda mais longa
  - As caudas não são simétricas

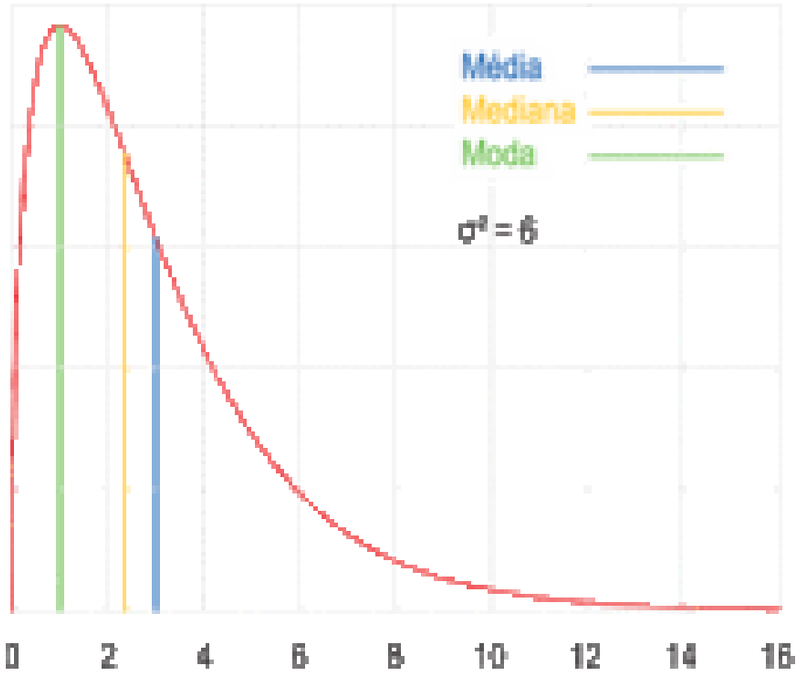
Média, Mediana e Moda de distribuições assimétricas



Fonte: https://pt.wikipedia.org/wiki/Mediana_(estat%C3%ADstica)

Média, Mediana e Moda de distribuições assimétricas





Fonte: https://pt.wikipedia.org/wiki/Mediana_(estat%C3%ADstica)

Apresentação dos Códigos Panda sobre Estatística

Prática no Jupyter Notebook

- Faça o restante dos exercícios da aula;
- Há exercícios extra.