

CIÊNCIAS AMBIENTAIS / BIOLÓGICAS /  
DA NATUREZA

---

# **(BIO)ESTATÍSTICA**

Prof<sup>a</sup>. Letícia Raposo  
profleticiaraposo@gmail.com

# ESTADÍSTICA DESCRIPTIVA BIVARIADA



## OBJETIVOS DA AULA

- Compreender os principais conceitos de estatística descritiva;
- Escolher o(s) método(s) adequado(s), incluindo tabelas, gráficos e/ou medidas-resumo, para descrever o comportamento de cada tipo de variável;
- Representar a frequência da ocorrência de um conjunto de observações por meio das tabelas de distribuição de frequências;



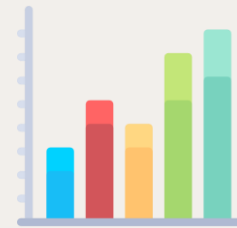
## OBJETIVOS DA AULA

- Representar a distribuição de uma variável com gráficos;
- Utilizar medidas de posição para representar os dados;
- Medir a variabilidade de um conjunto de dados por meio das medidas de dispersão;
- Utilizar medidas de assimetria e curtose para caracterizar a forma da distribuição dos elementos da população amostrados em torno da média.
- Gerar tabelas, gráficos e medidas-resumo por meio do R.



- INTERPRETAÇÃO (NÃO HÁ CONCLUSÕES NESTA ETAPA).
- ANÁLISE EXPLORATÓRIA DOS DADOS: OBSERVAR DETERMINADOS ASPECTOS RELEVANTES E COMEÇAR A DELINEAR HIPÓTESES A RESPEITO DA ESTRUTURA DO UNIVERSO EM ESTUDO.

# ESTATÍSTICA DESCRITIVA



---

**ORGANIZAR, RESUMIR E APRESENTAR OS DADOS (TABELAS, GRÁFICOS E MEDIDAS-RESUMO).**

# INTRODUÇÃO

**É POR MEIO DA EXPLORAÇÃO DOS DADOS QUE VOCÊ TERÁ UM CONHECIMENTO MAIS ELABORADO SOBRE ELES E ENTENDERÁ MELHOR O QUE PODE FAZER COM AS INFORMAÇÕES PARA ALCANÇAR OS OBJETIVOS DA PESQUISA.**

---

OS DADOS ESTÃO LHE DIZENDO ALGO IMPORTANTE?

VALE A PENA FAZER UMA ANÁLISE?

VOCÊ PRECISA COLETAR MAIS DADOS?

# EXEMPLO

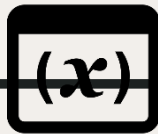
Escores de superação de pacientes vítimas de acidente vascular cerebral e ataque cardíaco.

**COMO VOCÊ IRIA  
DESCREVÊ-LOS A UM  
AMIGO QUE NÃO PODE VÊ-  
LOS?**



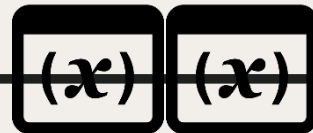
Acidente Vascular Cerebral		Ataque Cardíaco	
39	27	27	27
26	1	29	23
26	25	27	26
9	23	27	35
14	23	27	35
28	40	22	32
21	9	29	32
26	13	23	22
23	13	29	25
18	21	30	30

# ESTATÍSTICA DESCRITIVA



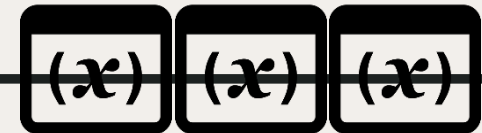
UNIVARIADA

Estuda uma única  
variável.



BIVARIADA

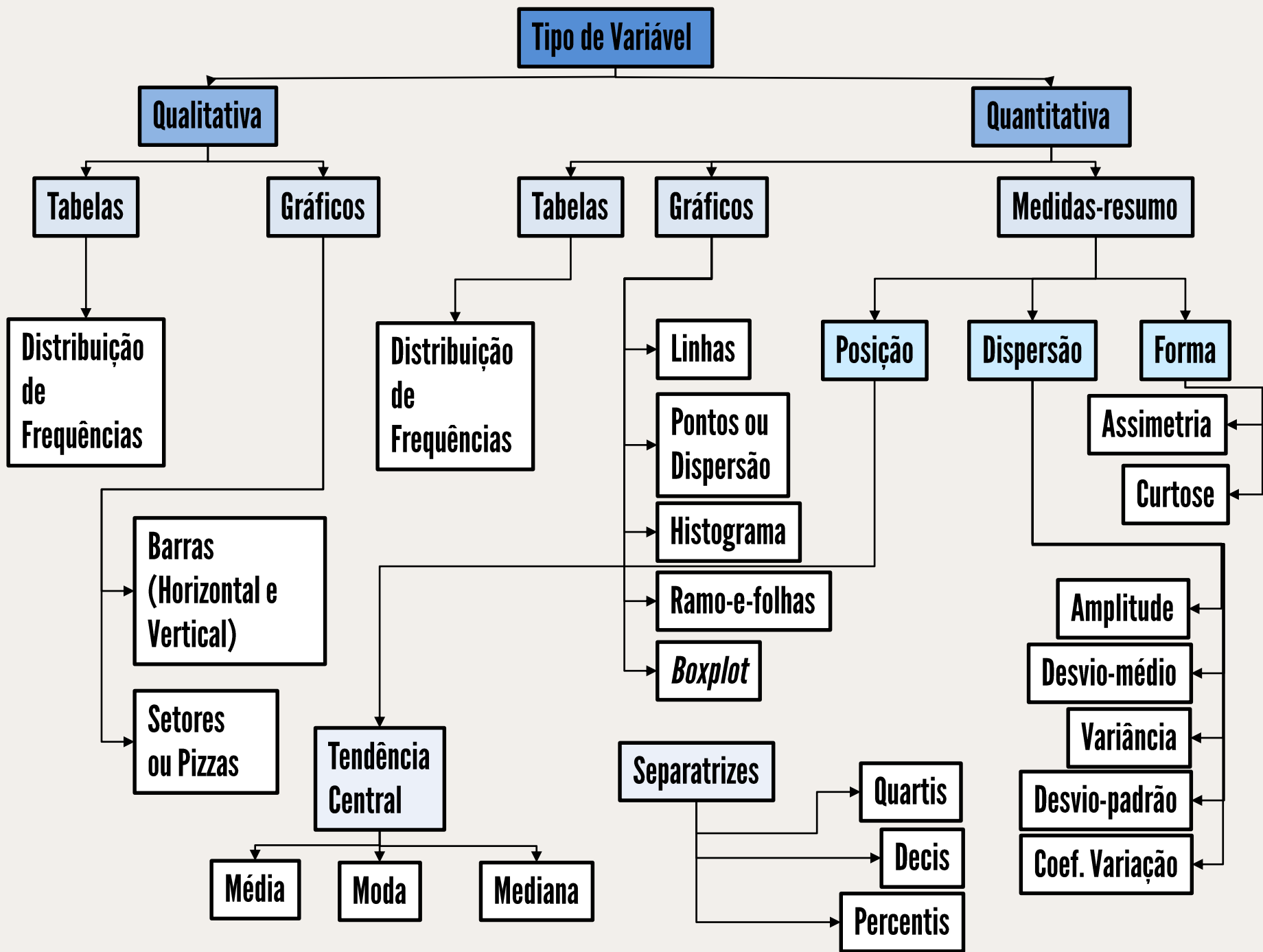
Estuda duas  
variáveis.



MULTIVARIADA

Estuda mais de duas  
variáveis.





# DESCRIÇÃO E EXPLORAÇÃO DE DADOS – VARIÁVEIS QUALITATIVAS



# TABELA DE DISTRIBUIÇÃO DE FREQUÊNCIAS

Organização dos dados de acordo com as ocorrências dos diferentes resultados observados.

Tipo Sanguíneo	$F_i$	$Fr_i$	$F_{ac}$	$Fr_{ac}$
A+	15	25	15	25
A-	2	3,33	17	28,33
B+	6	10	23	38,33
B-	1	1,67	24	40
AB+	1	1,67	25	41,67
AB-	1	1,67	26	43,33
O+	32	53,33	58	96,67
O-	2	3,33	60	100
Total	60	100		

*Frequência absoluta ( $F_i$ )*

*Frequência relativa ( $Fr_i$ )*

*Frequência acumulada ( $F_{ac}$ )*

*Frequência relativa acumulada ( $Fr_{ac}$ )*

# REPRESENTAÇÕES GRÁFICAS

- VISUALIZAÇÃO MAIS SUGESTIVA QUE AS TABELAS.
  - PERMITE INTERPRETAÇÃO RÁPIDA E OBJETIVA DOS DADOS.
  - FORMA ALTERNATIVA DE DISTRIBUIÇÃO DE FREQUÊNCIAS.
- 

GRÁFICO DE BARRAS



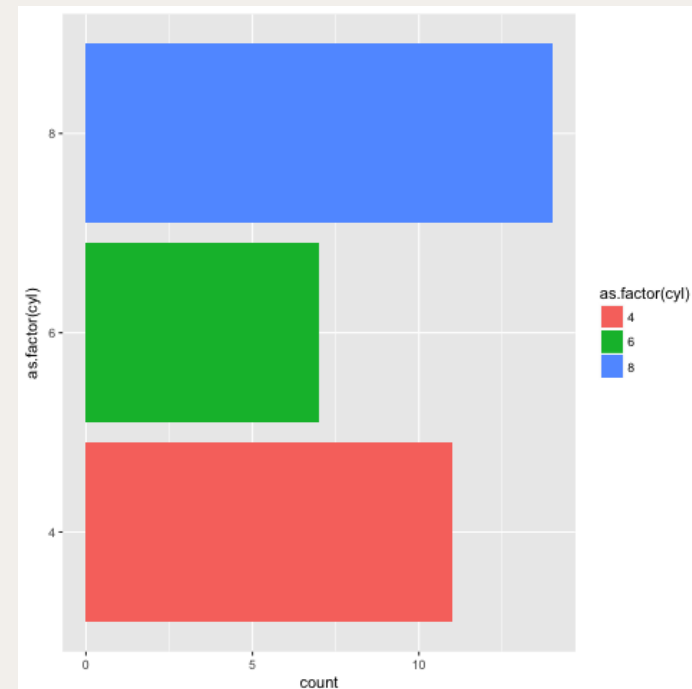
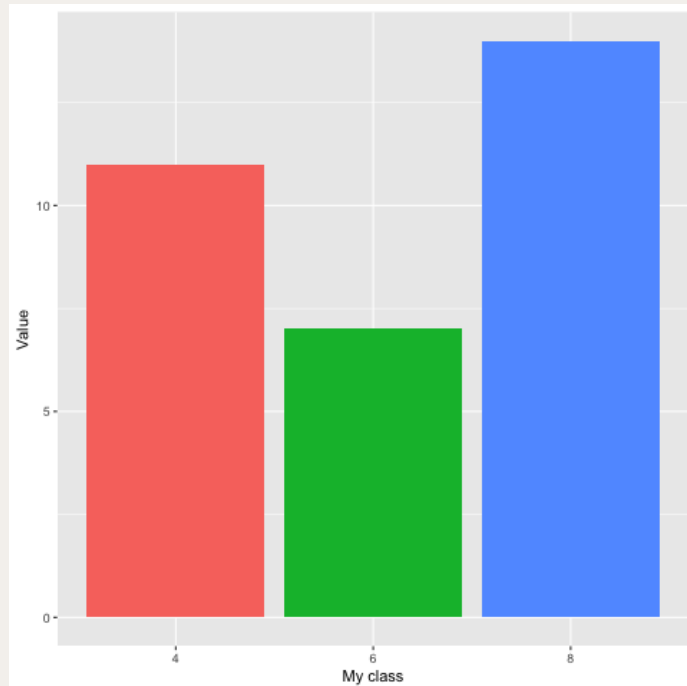
GRÁFICO DE SETORES



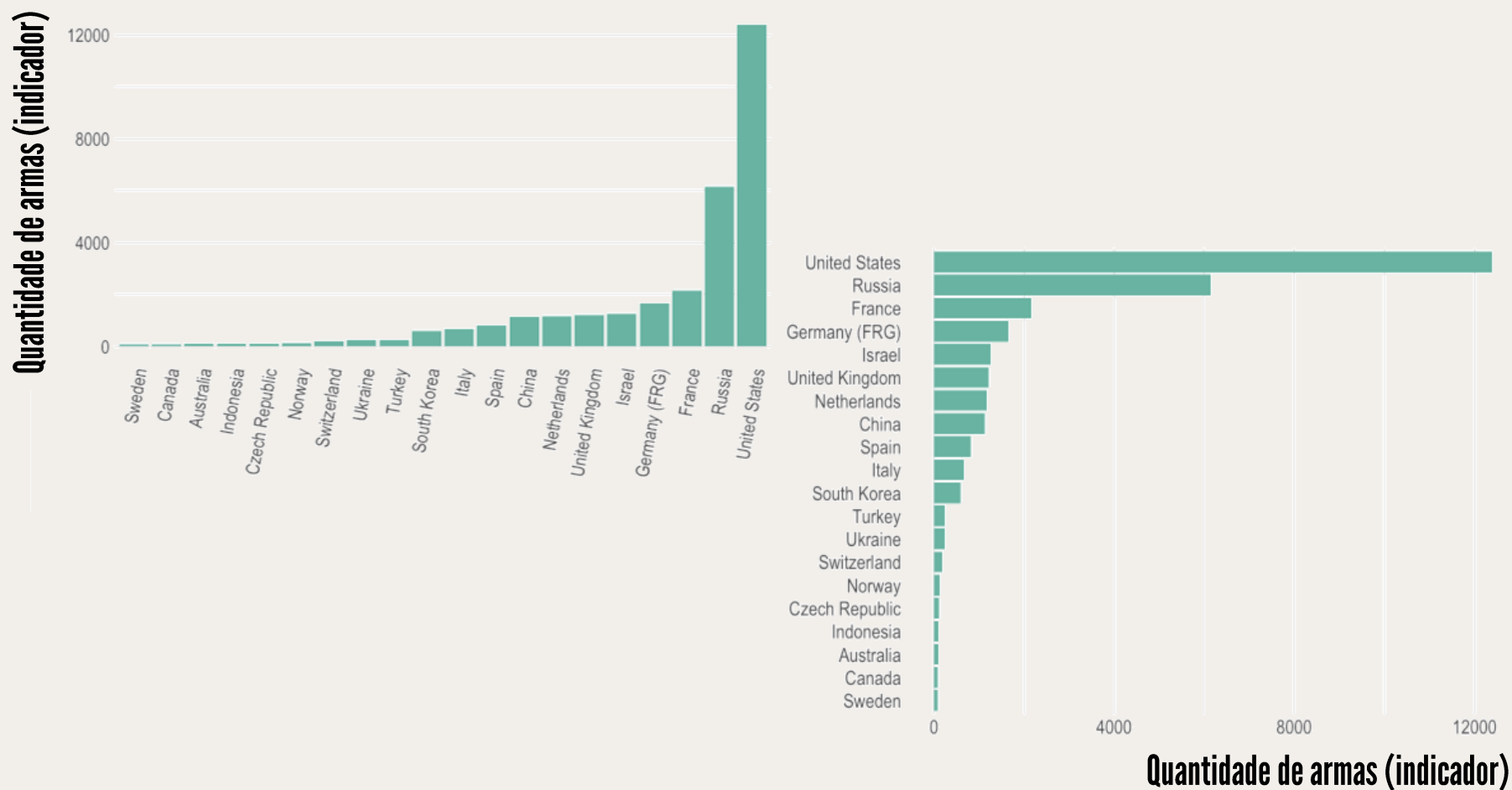
# GRÁFICO DE BARRAS



- Representa, por meio de barras, as frequências absolutas ou relativas de cada possível categoria.
- Cada entidade da variável categórica é representada como uma barra.
- O tamanho da barra representa seu valor numérico.



# BARRAS HORIZONTAIS: QUANDO OS NOMES SÃO EXTENSOS

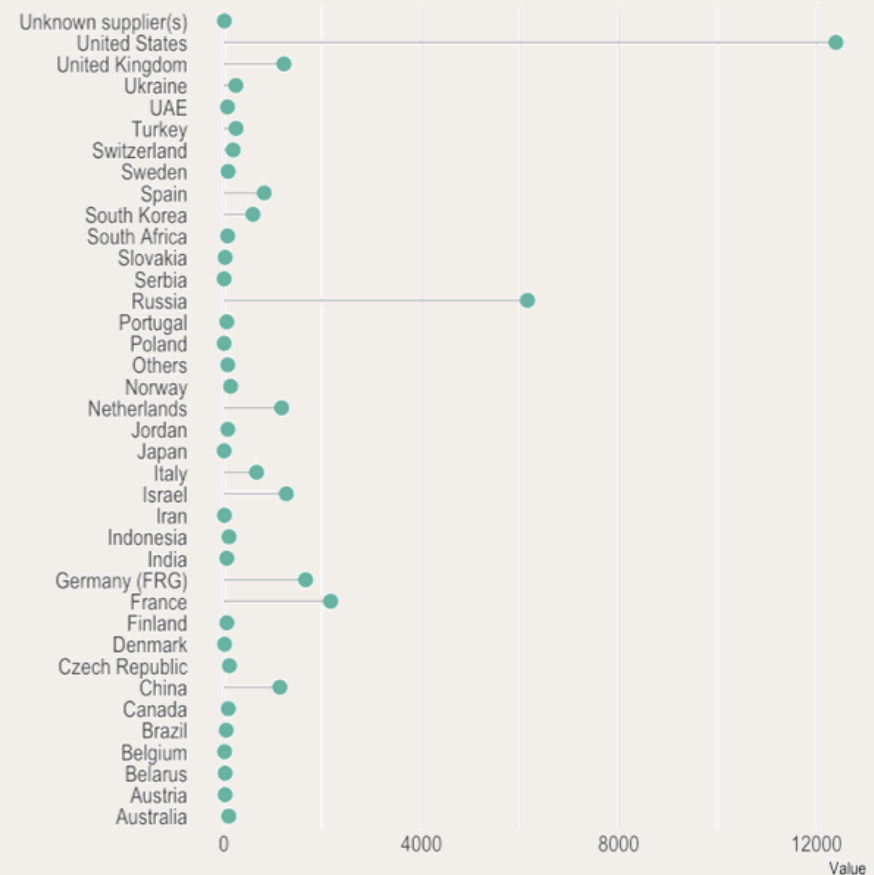


[https://www.data-to-viz.com/caveat/hard\\_label.html](https://www.data-to-viz.com/caveat/hard_label.html)

# POR QUE VOCÊ DEVE ORDENAR OS SEUS DADOS?

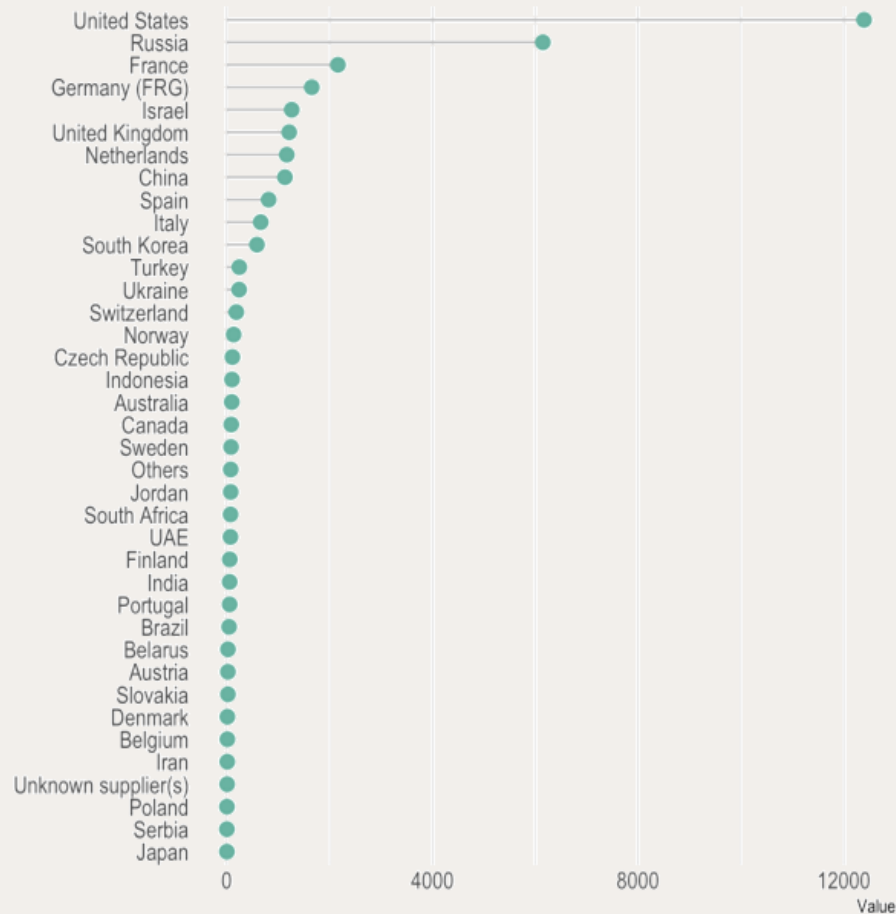


Por padrão, a maioria das ferramentas de visualização de dados ordenará os grupos de variáveis categóricas usando ordem alfabética ou usando a ordem de aparição em sua tabela de entrada.



[https://www.data-to-viz.com/caveat/order\\_data.html](https://www.data-to-viz.com/caveat/order_data.html)

# POR QUE VOCÊ DEVE ORDENAR OS SEUS DADOS?



É claro que, às vezes, a ordem dos grupos deve ser definida por suas características e não por seus valores, como os meses do ano.

[https://www.data-to-viz.com/caveat/order\\_data.html](https://www.data-to-viz.com/caveat/order_data.html)

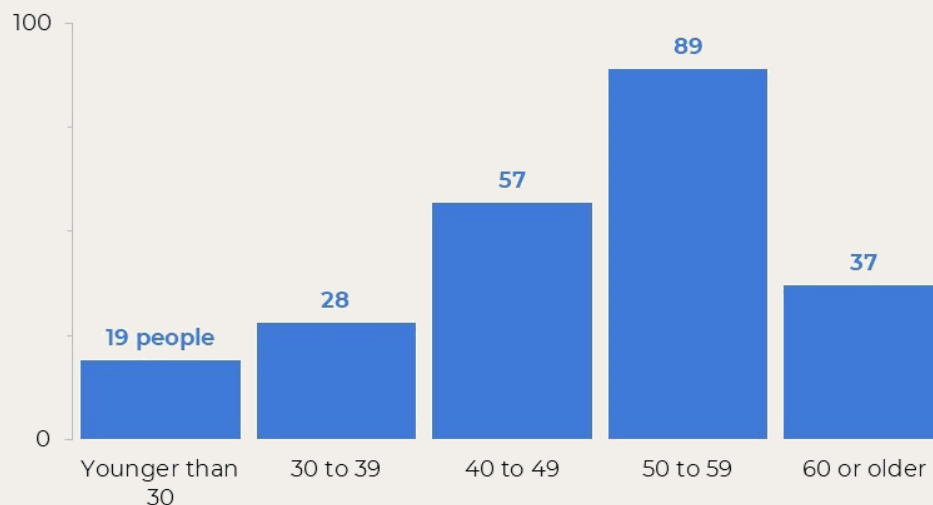


# QUANDO USAR BARRAS HORIZONTAIS OU VERTICAIS?



Barras verticais: quando estiver criando gráficos de variáveis ordinais.

- Dica: organizar as categorias ordinais da esquerda para a direita para ser visualizada uma sequência.

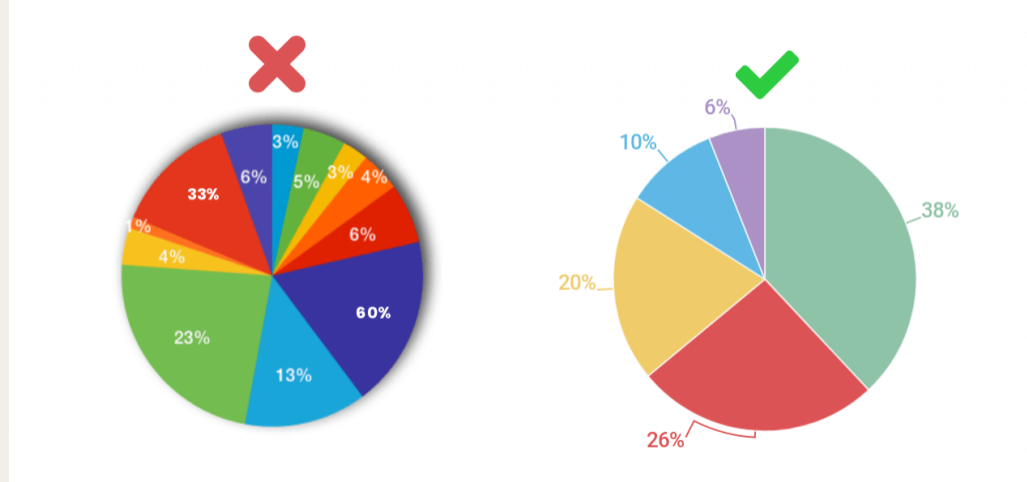
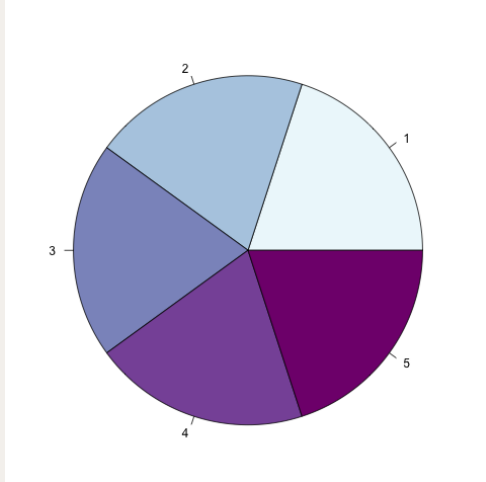


<https://depictdatastudio.com/when-to-use-horizontal-bar-charts-vs-vertical-column-charts/>

# GRÁFICO DE SETORES (PIZZA)



- Representa as frequências relativas de cada possível categoria.
- É frequentemente usado para mostrar porcentagem, onde a soma dos setores é igual a 100%.

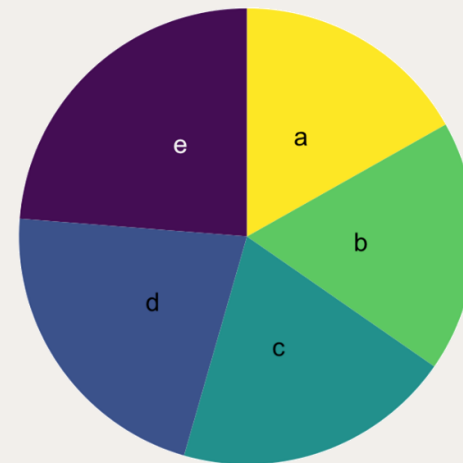


# GRÁFICO DE SETORES (PIZZA)



No gráfico de pizza adjacente, tente descobrir qual grupo é o maior e tente ordená-los por valor.

Você provavelmente terá dificuldades para fazê-lo e é por isso que os gráficos de pizza devem ser evitados.

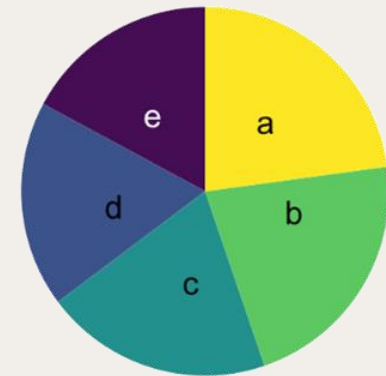
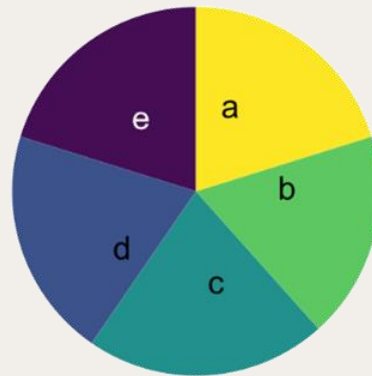
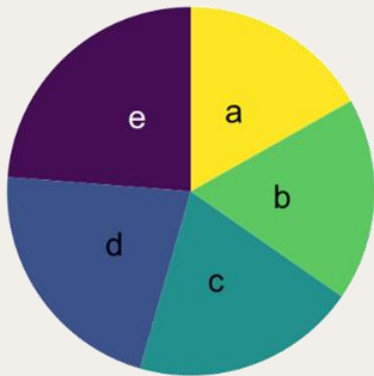


<https://www.data-to-viz.com/caveat/pie.html>

# GRÁFICO DE SETORES (PIZZA)

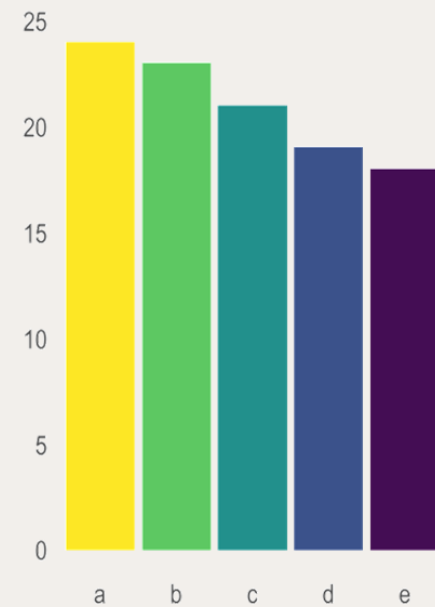
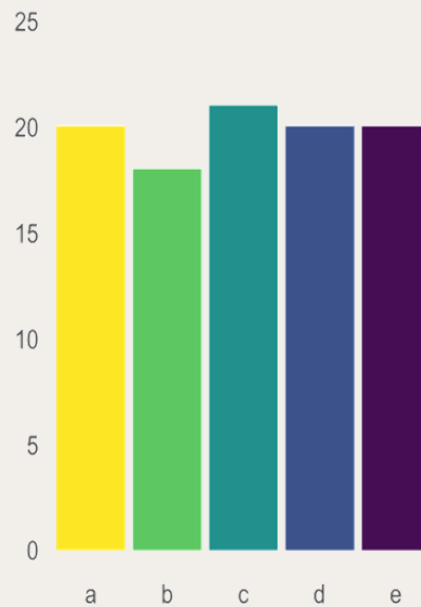
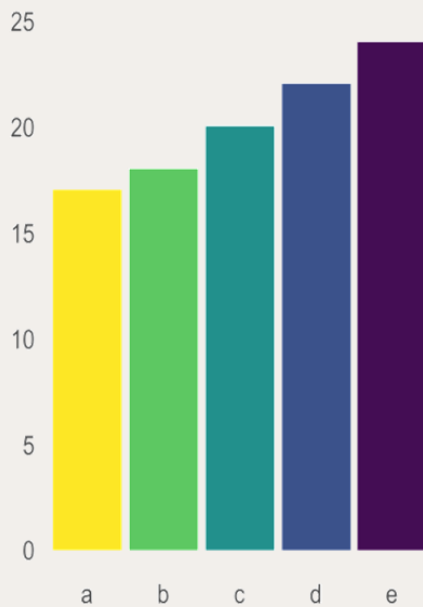


Se você ainda não está convencido, vamos tentar comparar vários gráficos de pizza. Mais uma vez, tente descobrir qual grupo tem o maior valor nesses três gráficos. Além disso, tente descobrir qual é a evolução do valor entre os grupos.



<https://www.data-to-viz.com/caveat/pie.html>

# GRÁFICO DE SETORES (PIZZA)



<https://www.data-to-viz.com/caveat/pie.html>

# GRÁFICO DE SETORES (PIZZA)



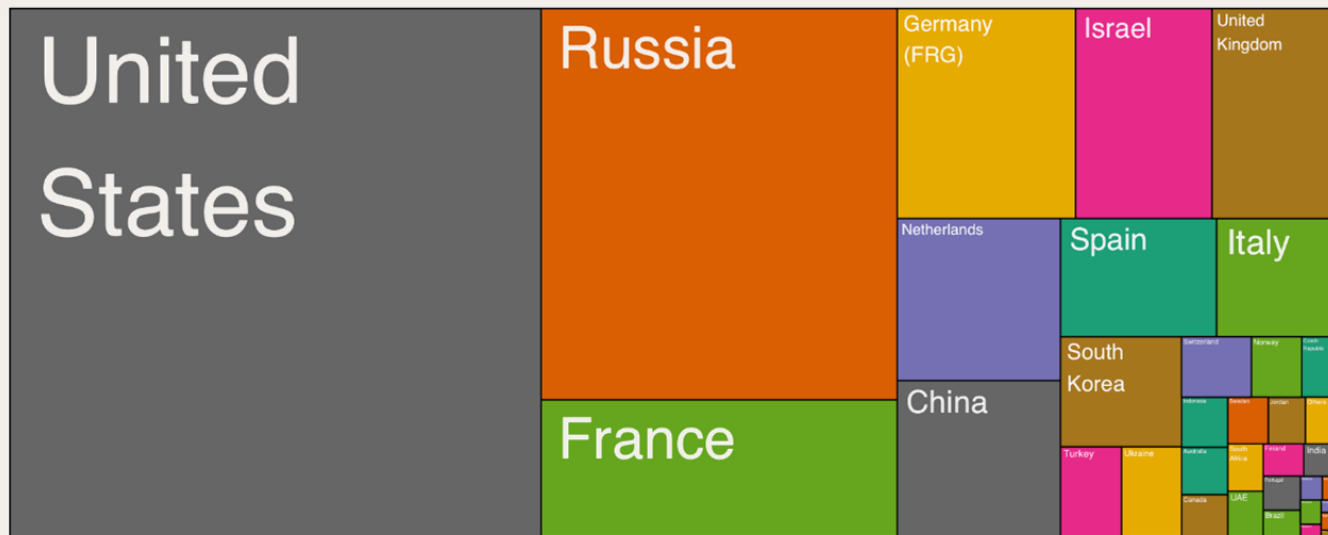
## Erros comuns:

- Usar em 3D;
- Legenda ao lado e não referenciada diretamente a cada setor;
- Porcentagens que não somam 100%;
- Muitos itens;
- Muitos gráficos de setores lado a lado.

# GRÁFICO DE SETORES (PIZZA)



Outra possibilidade seria criar um *treemap* se seu objetivo é descrever do que o todo é composto.



<https://www.data-to-viz.com/caveat/pie.html>

# DESCRIÇÃO E EXPLORAÇÃO DE DADOS – VARIÁVEIS QUANTITATIVAS





# VARIÁVEIS

---

## DISCRETA

- Seus possíveis valores podem ser listados.
- Ex: número de filhos de um casal, número de bactérias em uma placa de Petri.
- Normalmente resultam em alguma contagem.

## CONTÍNUA

- Pode assumir qualquer valor num intervalo.
- Ex: peso de um indivíduo.
- Costumam ser geradas por um instrumento de mensuração.

# VARIÁVEIS DISCRETAS

Tabela de distribuição de frequências:

- Similares às dos dados categorizados, desde que não haja grande quantidade de diferentes valores observados.
- No lugar das possíveis categorias devem constar os possíveis valores numéricos.

Representação gráfica: similar à das variáveis qualitativas.

# VARIÁVEIS CONTÍNUAS

Tabela de distribuição de frequências:

Para as variáveis contínuas, não faz muito sentido contar as repetições de cada valor, pois, considerando que dificilmente os valores se repetem, não chegaríamos a um resumo apropriado.

Podemos construir distribuições de frequências agrupando resultados em classes pré-estabelecidas.

- As classes são mutuamente exclusivas;
- Todo valor observado deve pertencer a uma e apenas uma classe;
- Na apresentação de uma tabela de frequências, é comum colocar o ponto médio das classes (ex: 40 |-- 50, o ponto médio é 45) → ponto médio representa o valor típico da classe.

# TABELA DE DISTRIBUIÇÃO DE FREQUÊNCIAS


Notas dos 30 alunos na disciplina de Estatística.

4,2	3,9	5,7	6,5	4,6	6,3	8,0	4,4	5,0	5,5
6,0	4,5	5,0	7,2	6,4	7,2	5,0	6,8	4,7	3,5
6,0	7,4	8,8	3,8	5,5	5,0	6,6	7,1	5,3	4,7

Dados anteriores ordenados de forma crescente.

3,5	3,8	3,9	4,2	4,4	4,5	4,6	4,7	4,7	5
5	5	5	5,3	5,5	5,5	5,7	6	6	6,3
6,4	6,5	6,6	6,8	7,1	7,2	7,2	7,4	8	8,8

# TABELA DE DISTRIBUIÇÃO DE FREQUÊNCIAS



E quantas classes  
eu uso?

- O número de classes a ser usada é uma escolha arbitrária.
  - Maior o conjunto de dados → mais classes podem ser usadas.
  - Em geral, são usadas de 5 a 20 classes.
- Usar, aproximadamente,  $\sqrt{n}$  classes, em que  $n$  é a quantidade de valores.

$$\sqrt{30} \approx 5,48 \rightarrow 5 \text{ classes.}$$

$$8,8 - 3,5 = 5,3; \frac{5,3}{5} \approx 1,06 \approx 1$$

- Expressão de Sturges: usar  $1 + 3,3 \cdot \log(n)$

$$1 + 3,3 \cdot \log(30) \approx 5,87 \rightarrow 6 \text{ classes}$$

$$\frac{5,3}{6} \approx 0,88 \approx 1$$

# TABELA DE DISTRIBUIÇÃO DE FREQUÊNCIAS

Distribuição de frequências para o exemplo da tabela anterior.

<b>Classe</b>	<b><math>F_i</math></b>	<b><math>F_i(\%)</math></b>	<b><math>F_{ac}</math></b>	<b><math>F_{ac}(\%)</math></b>
<b>[3,5;4,5)</b>	<b>5</b>	<b>16,67</b>	<b>5</b>	<b>16,67</b>
<b>[4,5;5,5)</b>	<b>9</b>	<b>30</b>	<b>14</b>	<b>46,67</b>
<b>[5,5;6,5)</b>	<b>7</b>	<b>23,33</b>	<b>21</b>	<b>70</b>
<b>[6,5;7,5)</b>	<b>7</b>	<b>23,33</b>	<b>28</b>	<b>93,33</b>
<b>[7,5;8,5)</b>	<b>1</b>	<b>3,33</b>	<b>29</b>	<b>96,67</b>
<b>[8,5;9,5)</b>	<b>1</b>	<b>3,33</b>	<b>30</b>	<b>100</b>
<b>Soma</b>	<b>30</b>	<b>100</b>		

# REPRESENTAÇÕES GRÁFICAS

- VISUALIZAÇÃO MAIS SUGESTIVA QUE AS TABELAS.
  - PERMITE INTERPRETAÇÃO RÁPIDA E OBJETIVA DOS DADOS.
  - FORMA ALTERNATIVA DE DISTRIBUIÇÃO DE FREQUÊNCIAS.
- 

HISTOGRAMA



GRÁFICO DE DENSIDADES



GRÁFICO DE RAMO-E-FOLHAS

Stem	Leaf
0	5
1	6, 7
2	8, 3, 6
3	4, 5, 9, 5, 5, 8, 5
4	7, 7, 7, 8
5	5, 4
6	0

BOXPLOT



# HISTOGRAMA



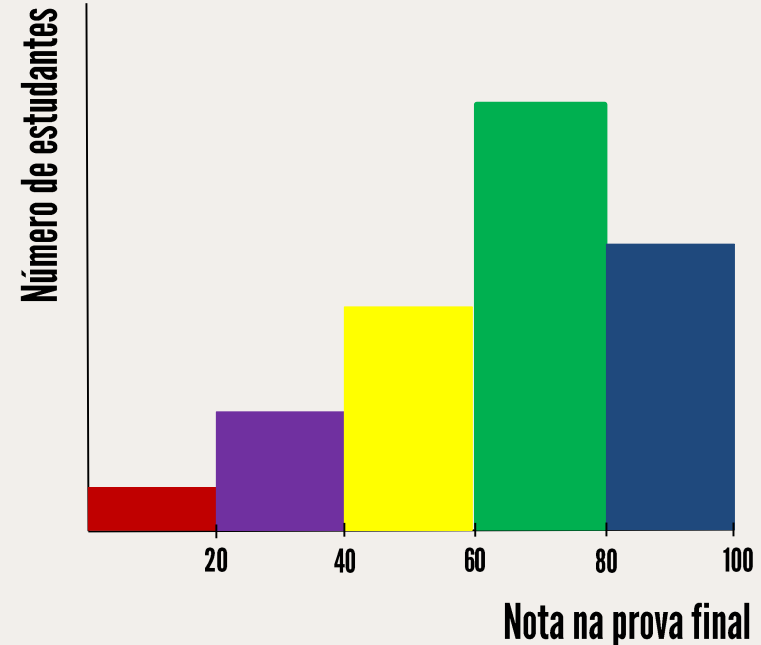
- São retângulos justapostos, feitos sobre as classes da variável em estudo.
- A altura de cada retângulo é proporcional à frequência (absoluta, relativa ou acumulada) observada da correspondente classe.
- Permite identificar a distribuição e a frequência dos dados. O histograma divide a variável contínua em grupos (eixo x) e fornece a frequência (eixo y) em cada grupo.



**Atenção:** percebam que o gráfico de barras possui as barras separadas, enquanto o histograma apresenta as barras justapostas.

As barras dos histogramas são normalmente chamadas de “bins”.

- Tente vários tamanhos de *bins*, isso pode levar a conclusões diferentes.
- Não use larguras de *bins* diferentes.

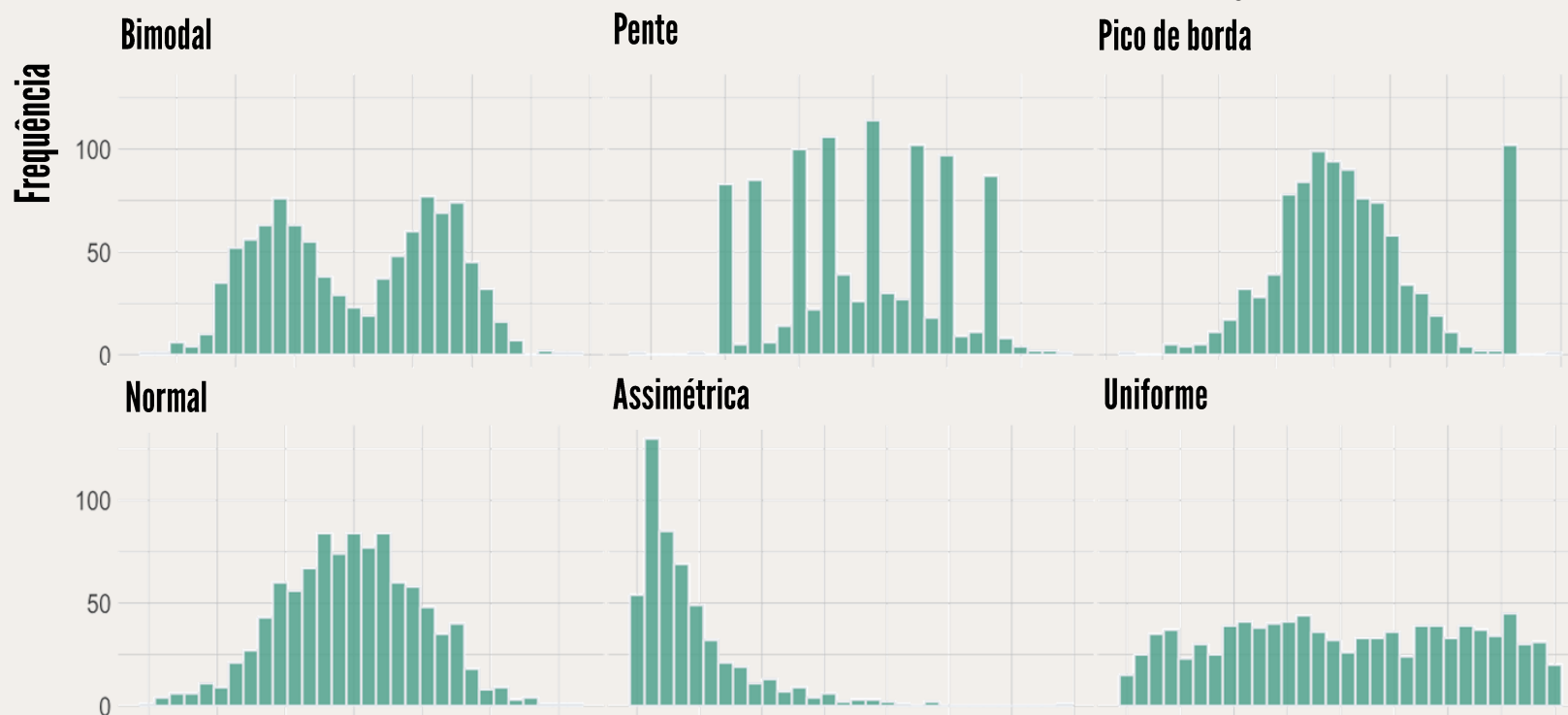




# HISTOGRAMA



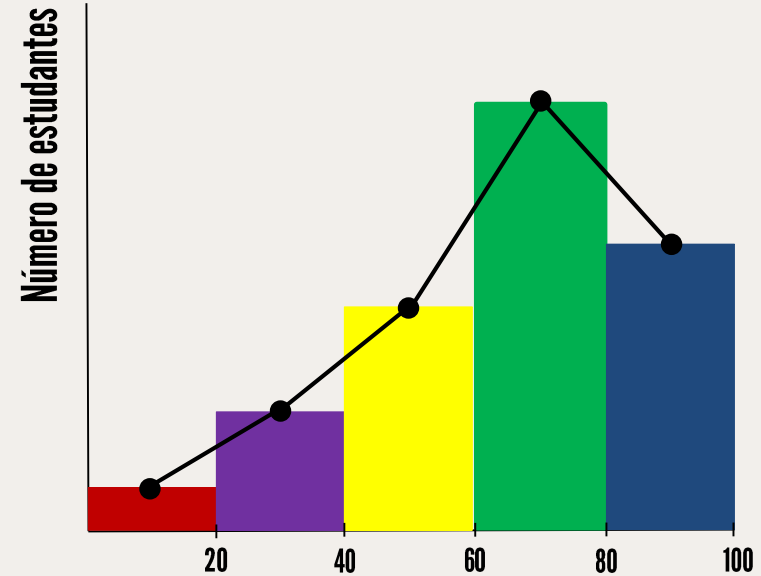
- Usado para estudar a distribuição de uma ou algumas variáveis.
- Verificar a distribuição de suas variáveis é provavelmente a primeira tarefa que você deve fazer quando receber um novo conjunto de dados. Ele fornece uma boa quantidade de informações.



<https://www.data-to-viz.com/graph/histogram.html>

# POLÍGONO DE FREQUÊNCIAS

- Representação gráfica alternativa.
- Toma-se o ponto médio ( $x$ ) e a frequência de cada classe ( $f$ ).
- Os pares  $(x, f)$  são plotados em um gráfico.
- Observe que os pontos são os valores médios de cada *bin* do histograma.

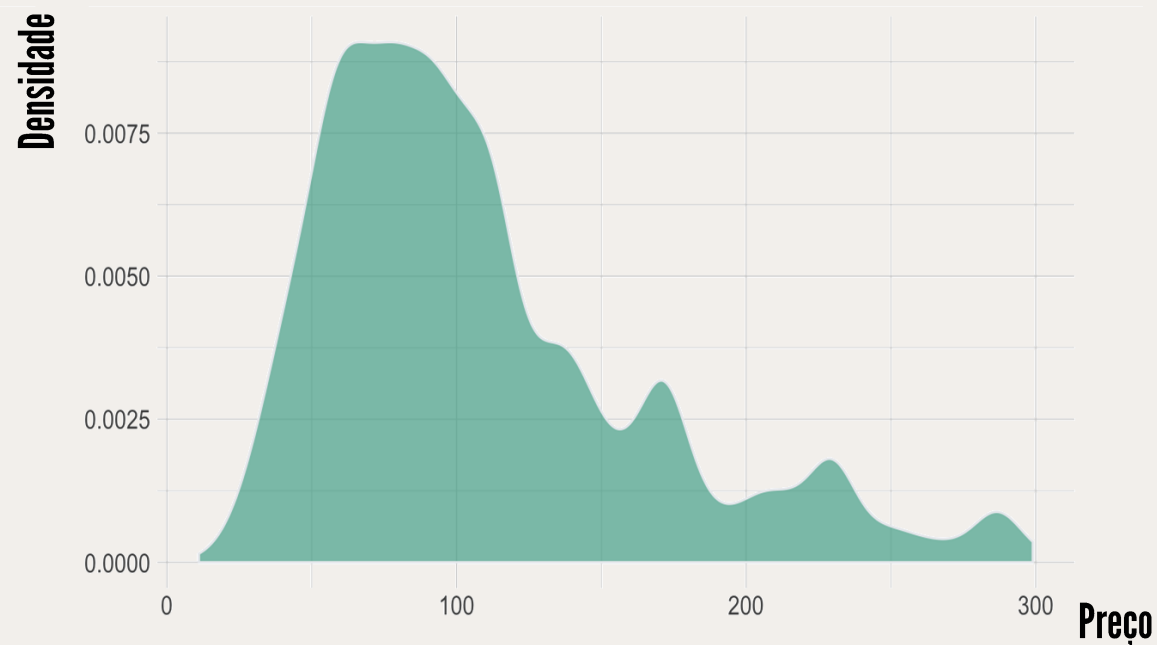


# GRÁFICO DE DENSIDADES



- Representação da distribuição de uma variável numérica.
- É uma versão suavizada do histograma e é usada no mesmo conceito.

**Distribuição dos preço por noite dos apartamentos do Airbnb**



<https://www.data-to-viz.com/graph/density.html>

# RAMO-E-FOLHAS

Stem	Leaf
0	5
1	6, 7
2	8, 3, 6
3	4, 5, 9, 5, 5, 8, 5
4	7, 7, 7, 8
5	5, 4
6	0

- Útil quando a quantidade de dados não é muito grande (centena de observações).
- Fornece a forma da distribuição de frequências e ainda preserva a magnitude aproximada dos valores.
- Os dados ficam ordenados crescentemente, o que facilita a obtenção de algumas medidas descritivas.
- Geralmente, os ramos compõem os dígitos mais significativos e as folhas os menos significativos.

# RAMO-E-FOLHAS

Notas dos 30 alunos na disciplina de Estatística.

4,2	3,9	5,7	6,5	4,6	6,3	8,0	4,4	5,0	5,5
6,0	4,5	5,0	7,2	6,4	7,2	5,0	6,8	4,7	3,5
6,0	7,4	8,8	3,8	5,5	5,0	6,6	7,1	5,3	4,7

Dados anteriores ordenados de forma crescente.

3,5	3,8	3,9	4,2	4,4	4,5	4,6	4,7	4,7	5
5	5	5	5,3	5,5	5,5	5,7	6	6	6,3
6,4	6,5	6,6	6,8	7,1	7,2	7,2	7,4	8	8,8

Stem	Leaf
0	5
1	6, 7
2	8, 3, 6
3	4, 5, 9, 5, 5, 8, 5
4	7, 7, 7, 8
5	5, 4
6	0

1. ORDENAR OS DADOS DE FORMA CRESCENTE

2. DEFINIR O NÚMERO DE DÍGITOS INICIAIS (RAMO E FOLHAS)
3. CONSTRUIR OS RAMOS (ESQUERDA DA LINHA VERTICAL)
4. COLOCAR AS FOLHAS CORRESPONDENTES AOS RESPECTIVOS RAMOS (LADO DIREITO DA LINHA VERTICAL)

3	5 8 9
4	2 4 5 6 7 7
5	0 0 0 0 3 5 5 7
6	0 0 3 4 5 6 8
7	1 2 2 4
8	0 8

HISTOGRAMA DEITADO

# BOXPLOT



- Representação gráfica de cinco medidas de posição ou localização de determinada variável:
  - Limite inferior;
  - Primeiro quartil ( $Q_1$ );
  - Segundo quartil ( $Q_2$ ) ou mediana ( $Md$ );
  - Terceiro quartil ( $Q_3$ );
  - Limite superior.
- Permite avaliar a simetria e distribuição dos dados, e também propicia a perspectiva visual da presença ou não de dados discrepantes (outliers univariados).

# BOXPLOT

Valores

Intervalo interquartilico (IQR) =  $Q3 - Q1$



Outlier

Limite superior =  $Q3 + 1,5 \times IQR$

3º quartil (Q1)

Mediana = 2º quartil (Q2)

1º quartil (Q1)

Limite inferior =  $Q1 - 1,5 \times IQR$

Outliers



# MEDIDAS-RESUMO

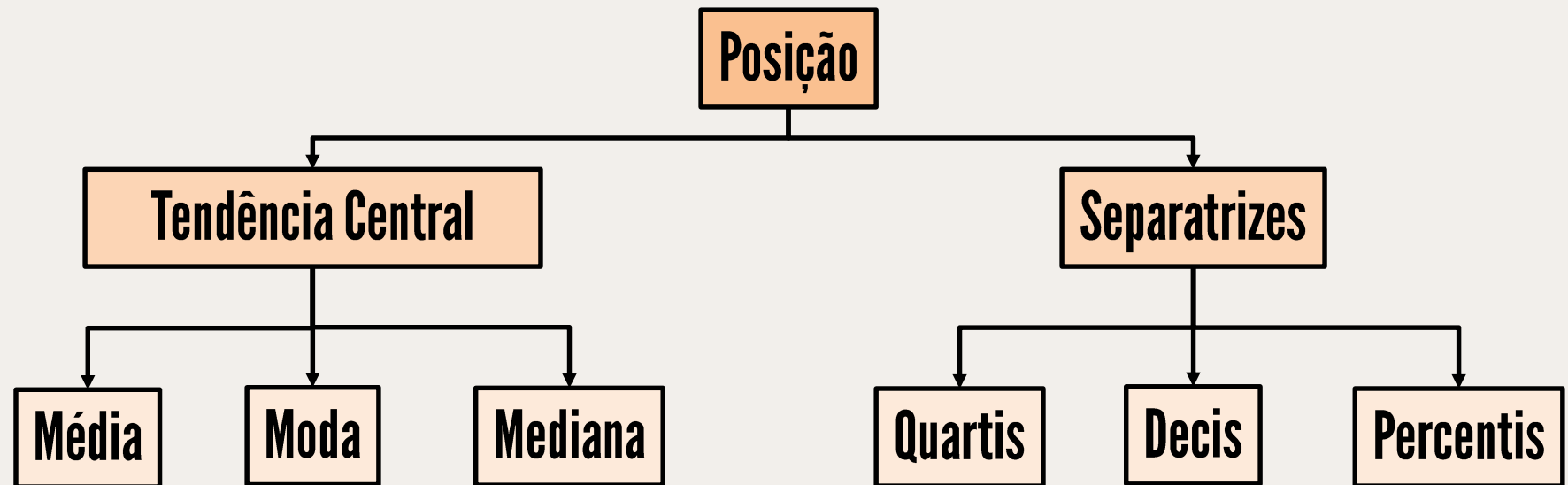
---



- DESCREVER E EXPLORAR DADOS QUANTITATIVOS POR MEIO DE FORMAS ALTERNATIVAS ÀS DISTRIBUIÇÕES DE FREQUÊNCIAS.
- CALCULAR E INTERPRETAR CERTAS MEDIDAS QUE DESCREVEM INFORMAÇÕES ESPECÍFICAS DE UM CONJUNTO DE VALORES.



# MEDIDAS DE POSIÇÃO OU LOCALIZAÇÃO



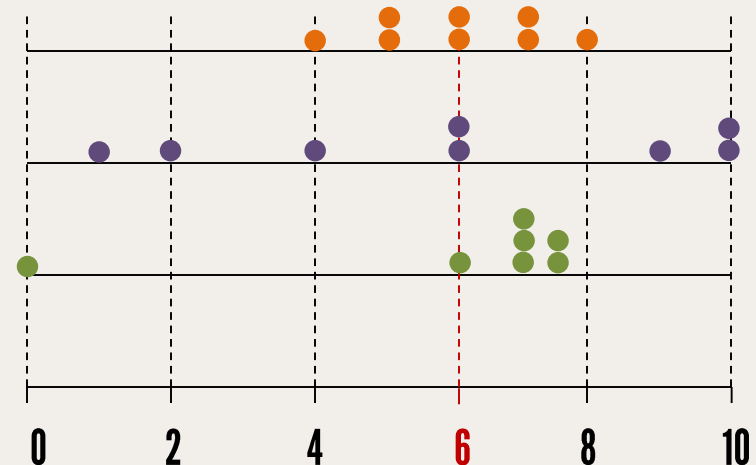
# MÉDIA ARITMÉTICA SIMPLES

- Média aritmética, ou simplesmente média, é a soma dos valores ( $\sum_{i=1}^n X_i$ ) dividida pelo número de valores observados ( $n$ ).

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

- A média é um resumo dos dados e, por isso, pode esconder informações relevantes.

Turma	Notas dos alunos	Média da turma
<b>A</b>	4 5 5 6 6 7 7 8	6,0
<b>B</b>	1 2 4 6 6 9 10 10	6,0
<b>C</b>	0 6 7 7 7 7,5 7,5	6,0



# MÉDIA ARITMÉTICA PONDERADA

- Em geral, a ponderação é feita sempre que precisamos dar mais importância a um caso do que a outro (atribuir pesos diferentes).

$$\bar{X} = \frac{\sum_{i=1}^n X_i \cdot p_i}{\sum_{i=1}^n p_i}$$

- Se os pesos estiverem expressos em termos percentuais (peso relativo -  $pr$ ), a fórmula passa a ser:

$$\bar{X} = \sum_{i=1}^n X_i \cdot pr_i$$

# MÉDIA ARITMÉTICA PARA DADOS DISCRETOS AGRUPADOS

- Quando existem valores repetidos, podemos calcular a média ponderando pelas respectivas frequências.
- Essa estratégia facilita bastante quando existem muitas repetições de valores.

$$\bar{X} = \frac{\sum_{i=1}^m X_i \cdot F_i}{n}$$

em que  $F_i$  são as frequências absolutas e  $m$  são os diferentes valores (dados agrupados).

$$\bar{X} = \frac{1 \times 9 + 2 \times 12 + \dots + 9 \times 3 + 10 \times 1}{120} = 4,62$$

Frequência absoluta dos entrevistados.

Notas	No. de entrevistados
1	9
2	12
3	15
4	18
5	24
6	26
7	5
8	7
9	3
10	1

# MÉDIA ARITMÉTICA PARA DADOS CONTÍNUOS AGRUPADOS

Quando os dados estão agrupados em classes, o cálculo de  $\bar{X}$  somente poderá ser feito de forma aproximada, usando o ponto médio de cada classe para representar os valores que ocorreram nessa classe.

$$\bar{X} \approx \frac{\sum_{i=1}^k X_i \cdot F_i}{n}$$

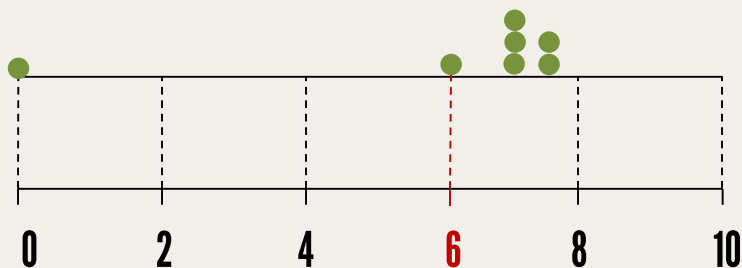
em que  $X_i$  é o ponto médio da classe  $i$  ( $i = 1, \dots, k$ ).

Classes	Frequência absoluta - $F_i$	Valor central da classe - $X_i$	$X_i \cdot F_i$
[0, 5[	9	2,5	22,5
[5, 10[	7	7,5	52,5
[10, 15[	4	12,5	50
[15, 20[	6	17,5	105
[20, 25[	5	22,5	112,5
Total	$\sum F_i = 31$		$\sum X_i \cdot F_i = 342,5$

$$\bar{X} \approx \frac{\sum_{i=1}^k X_i \cdot F_i}{n} \approx \frac{342,5}{31} \approx 11,05$$

# MÉDIA ARITMÉTICA

- A média resume o conjunto de dados em termos de uma posição central ou valor típico, mas, em geral, não fornece informação sobre outros aspectos da distribuição.
- Para melhorar o resumo dos dados, podemos apresentar, ao lado da média aritmética, uma medida de dispersão, como a variância ou o desvio padrão.
- A média aritmética é fortemente influenciada por valores discrepantes.



**O valor discrepante 0 puxa a média para baixo. Apesar de a média aritmética ser 6, o diagrama de pontos sugere que o valor 7 seja um valor mais típico para representar as notas da turma, pois, além de ser o valor mais frequente, ele é o valor do meio, deixando metade das notas abaixo dele e metade acima.**

# MEDIANA

- É uma medida de localização do centro da distribuição de um conjunto de dados ordenados de forma crescente.
- Seu valor separa a série em duas partes iguais, de modo que 50% dos elementos são menores ou iguais à mediana e os outros 50% são maiores ou iguais à mediana.

Se  $n$  for ímpar,  $Md(X)$  é o elemento central, se  $n$  for par,  $Md(x)$  é a soma dos dois elementos centrais divididos por dois.

5, 13, 9, 7, 1, 9, 2, 9, 11

1, 2, 5, 7, 9, 9, 9, 11, 13

Mediana

Ordem crescente  
↓

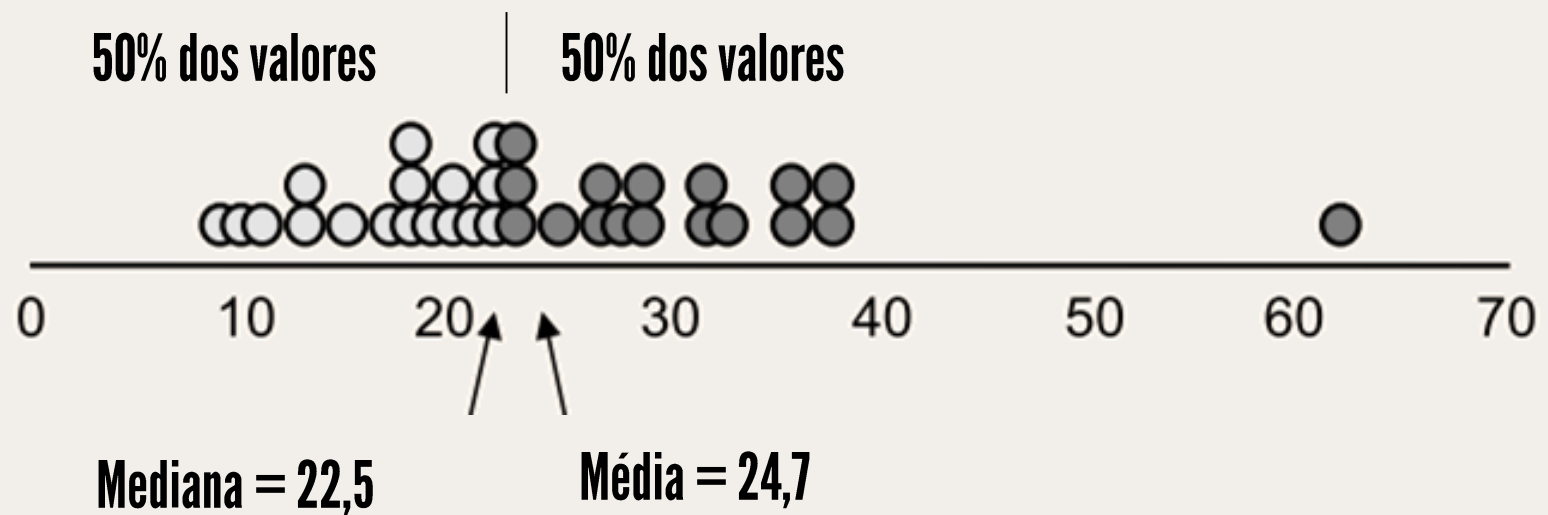
4, 3, 7, 8, 4, 5, 12, 4, 5, 3, 2, 3

2, 3, 3, 3, 4, 4, 4, 5, 5, 7, 8, 12

Mediana é a média dos  
dois números do meio.

Ordem crescente  
↓

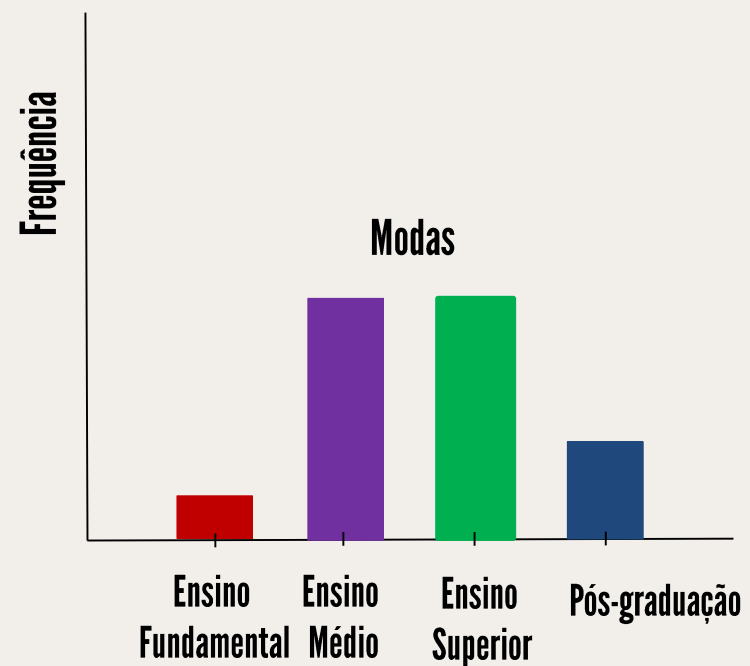
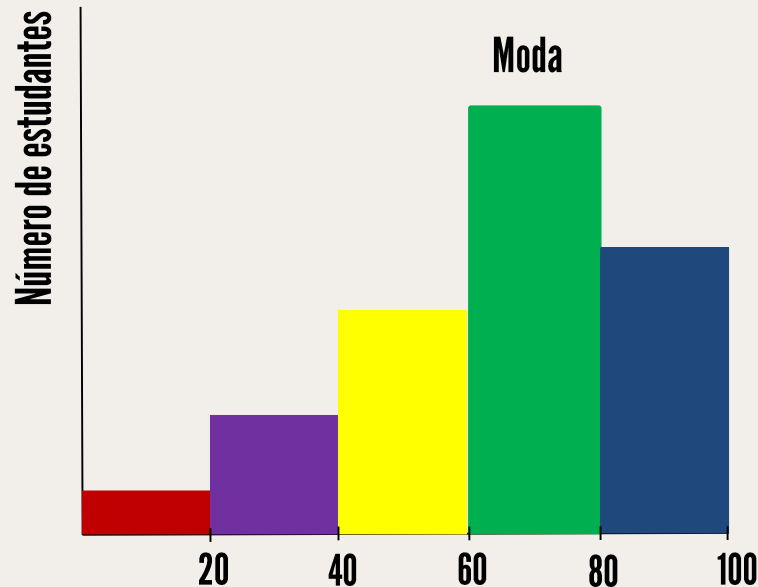
# COMPARAÇÃO ENTRE MÉDIA E MEDIANA





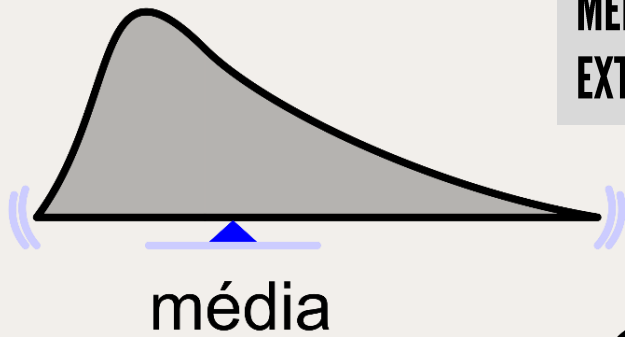
# MODA

- Corresponde à observação que ocorre com maior frequência.
- A moda é a única medida de posição que também pode ser utilizada para variáveis qualitativas, já que essas variáveis permitem apenas o cálculo de frequências.

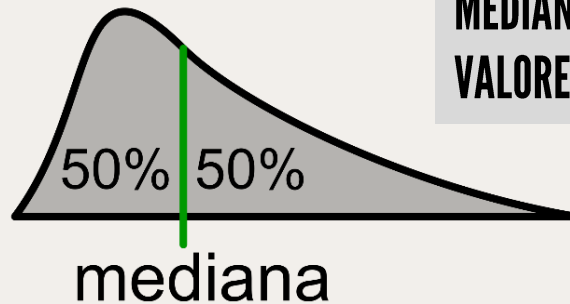


# RESUMINDO

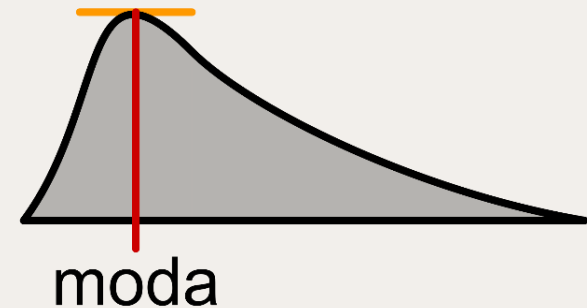
**MÉDIA:** INDICADA QUANDO NÃO HÁ VALORES EXTREMOS NOS DADOS.



**MEDIANA:** ÓTIMA QUANDO HÁ VALORES EXTREMOS.



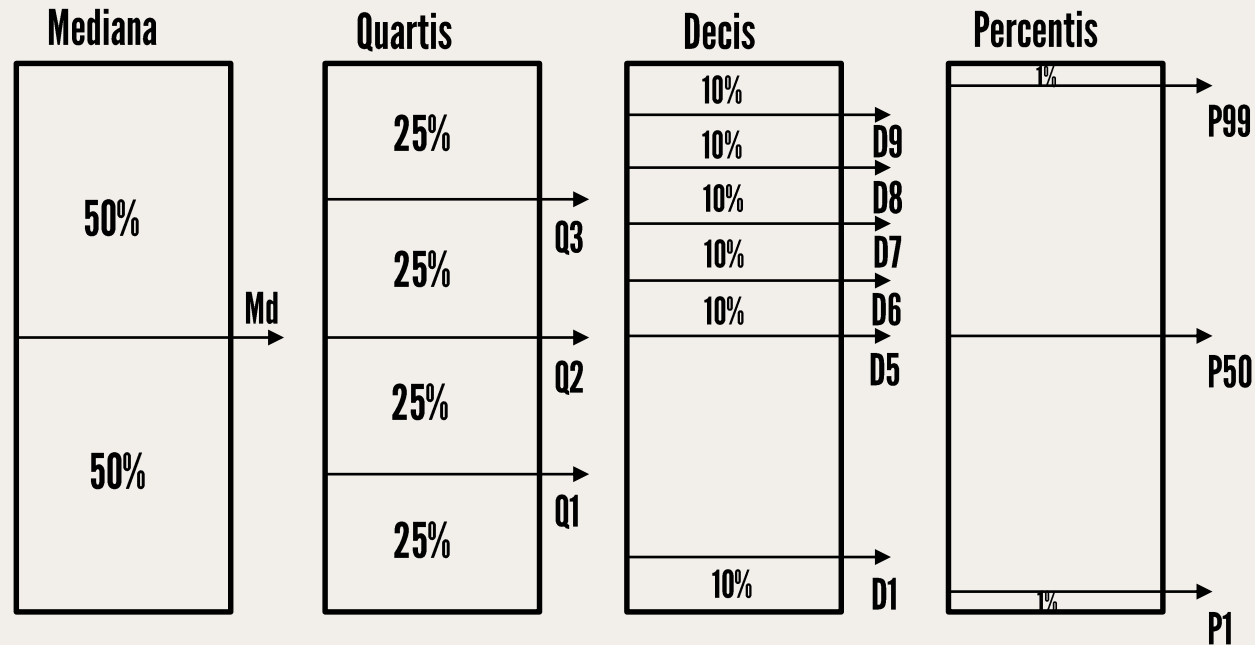
**MODA:** MAIS ÚTIL QUANDO HÁ DADOS CATEGÓRICOS.



# MEDIDAS SEPARATRIZES

- *Medidas de tendência central:* afetadas por valores extremos e, apenas com o uso destas medidas, não é possível que o pesquisador tenha uma ideia clara de como a dispersão e simetria dos dados se comportam.
- *Alternativa:* medidas separatrizes, como quartis, decis e percentis.

# MEDIDAS SEPARATRIZES

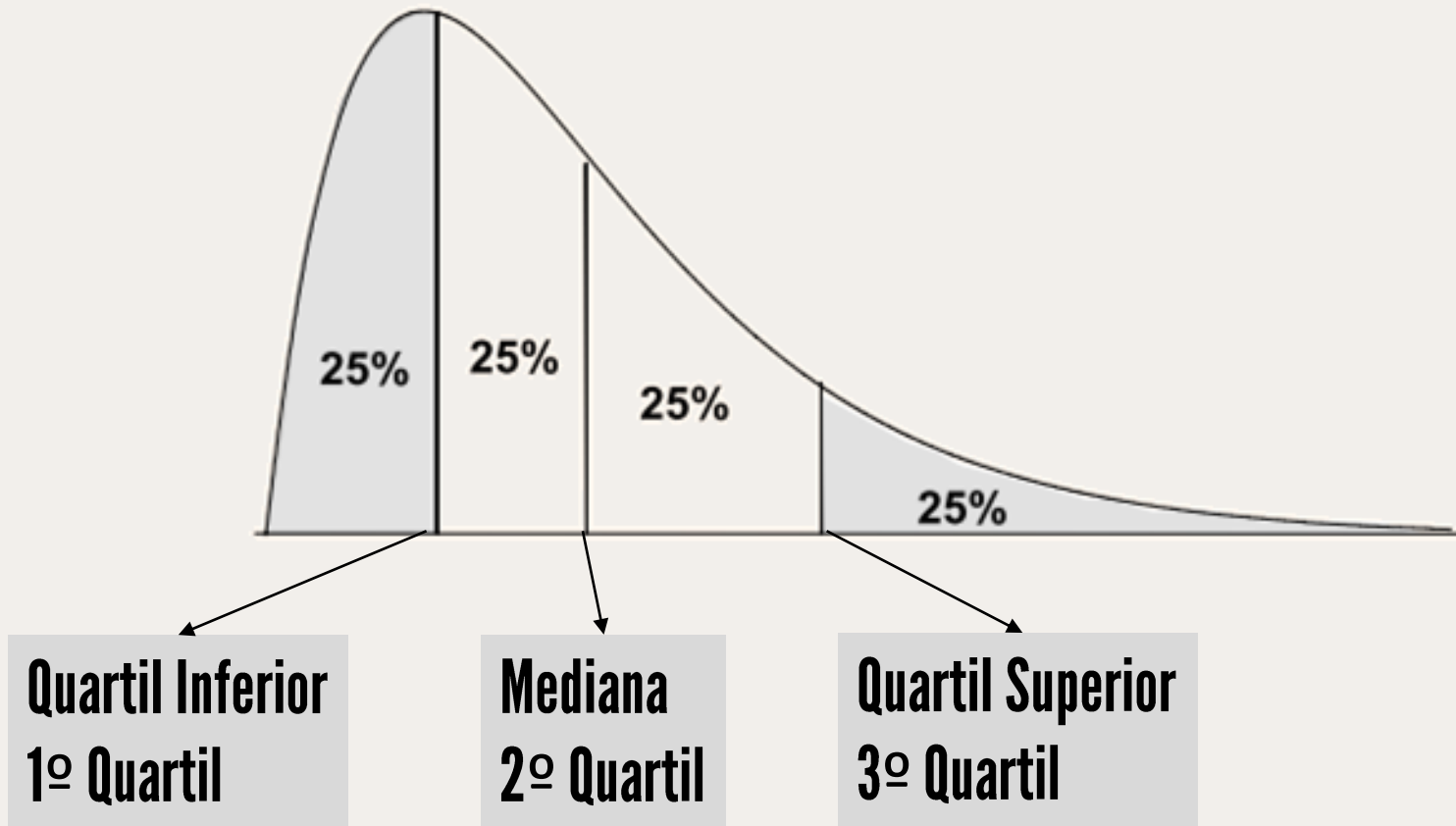


$$Pos(Q_i) = \left[ \frac{n}{4} \times i \right] + \frac{1}{2}, i = 1, 2, 3$$

$$Pos(D_i) = \left[ \frac{n}{10} \times i \right] + \frac{1}{2}, i = 1, 2, \dots, 9$$

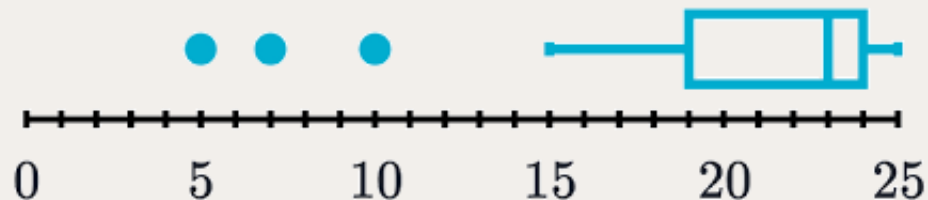
$$Pos(P_i) = \left[ \frac{n}{100} \times i \right] + \frac{1}{2}, i = 1, 2, \dots, 99$$

# QUARTIS



# IDENTIFICAÇÃO DE EXISTÊNCIA DE OUTLIERS UNIVARIADOS

- *Outliers*: observações que apresentam um grande afastamento das restantes ou são inconsistentes.
- *Possíveis causas*: erros de medição, de execução e variabilidade inerente aos elementos da população.

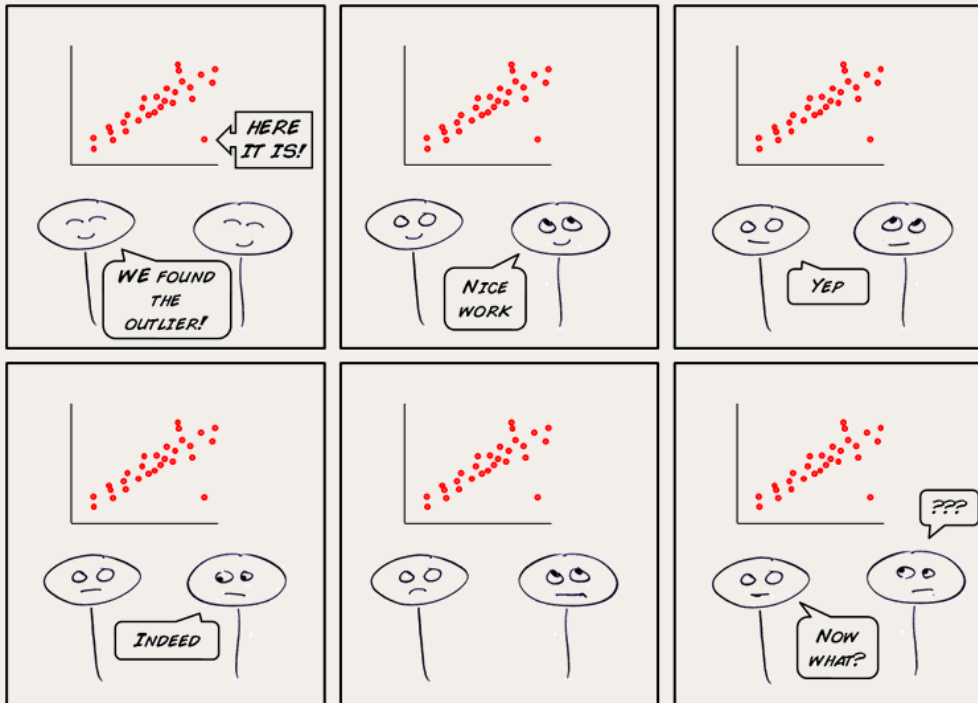


$$Q_1 - 1.5 \cdot \text{IQR}$$

$$Q_3 + 1.5 \cdot \text{IQR}$$

$$\text{IQR} = Q_3 - Q_1$$

# IDENTIFICAÇÃO DE EXISTÊNCIA DE OUTLIERS UNIVARIADOS

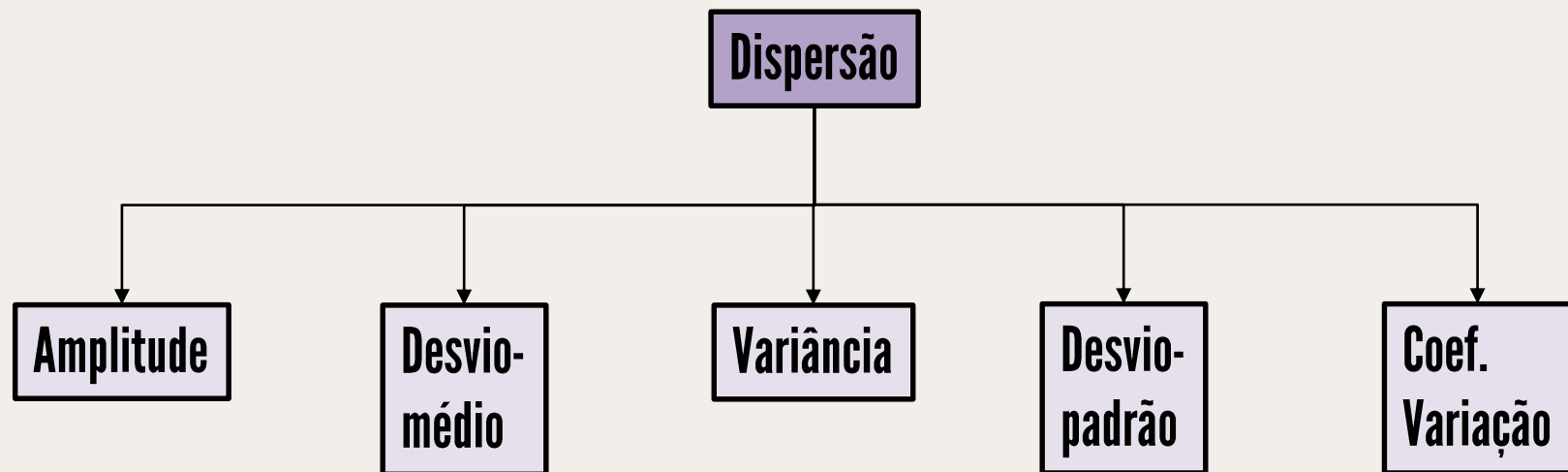


- Se forem erros (e você tem certeza disso!): remover.
  - Ex: peso do bebê ao nascer: 40kg.
- Se não forem erros, será necessário verificar o impacto que eles causam – avaliar com e sem *outliers*.
  - Inspecione o grupo de *outliers* - eles têm alguma característica especial que o grupo não-discrepante não tem?

Fonte: <https://optimalbi.com/blog/2014/02/24/how-do-you-solve-a-problem-like-an-outlier/>

**É preciso se certificar sobre as diferenças e dar explicações possíveis.**  
**É sempre uma boa ideia incluir um resumo do que se sabe sobre os *outliers*.**

# MEDIDAS DE DISPERSÃO OU VARIABILIDADE

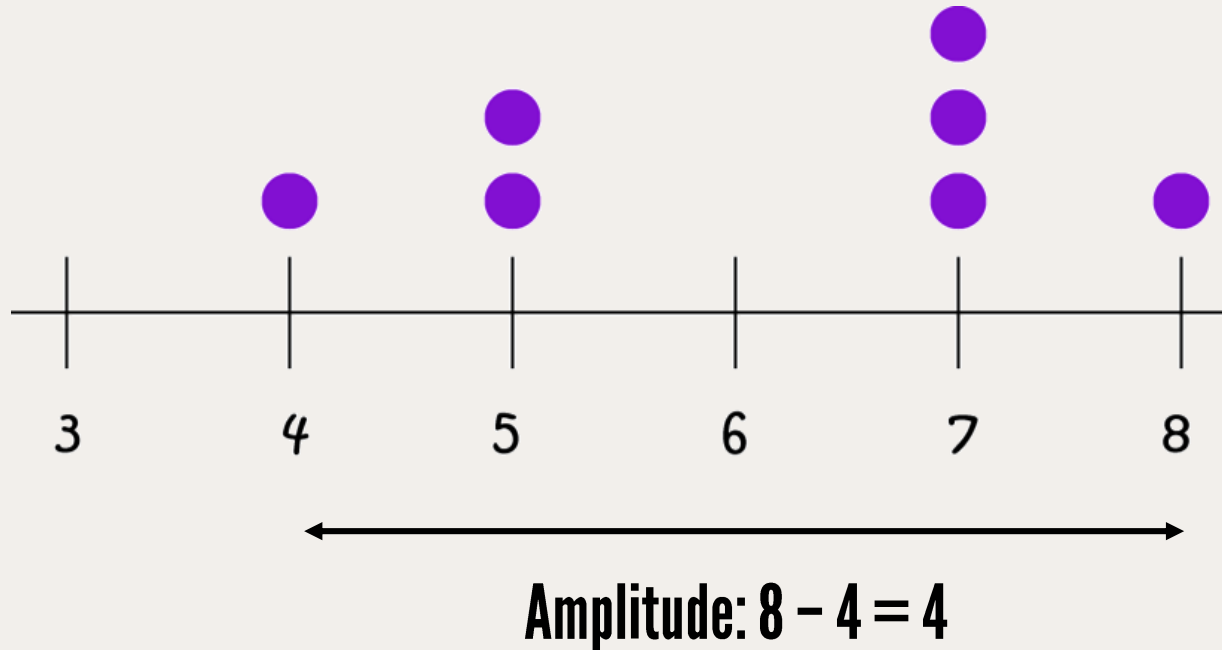




# AMPLITUDE

- Medida mais simples, representa a diferença entre o maior e o menor valor do conjunto de observações.
- Não informa como os valores variam entre as extremidades.

$$A = X_{m\acute{a}x} - X_{m\acute{i}n}$$



# DESVIO-MÉDIO

- *Desvio*: diferença entre cada valor observado e a média da variável.
  - Para dados populacionais:  $(X_i - \mu)$
  - Para dados amostrais:  $(X_i - \bar{X})$
- *Desvio-médio (desvio-médio absoluto)*: média aritmética dos desvios absolutos (em módulo).

$$D_m = \frac{\sum_{i=1}^N |X_i - \mu|}{N} \quad (\text{para a população})$$

$$D_m = \frac{\sum_{i=1}^n |X_i - \bar{X}|}{n} \quad (\text{para a amostra})$$

# VARIÂNCIA

- Medida de dispersão ou variabilidade que avalia o quanto os dados estão dispersos em relação à média aritmética.
- Quanto maior a variância, maior a dispersão dos dados.
- O valor tende a ser muito grande e de difícil interpretação.

$$\sigma^2 = \frac{\sum_{i=1}^N (X_i - \mu)^2}{N} = \frac{\sum_{i=1}^N X_i^2 - \frac{(\sum_{i=1}^N X_i)^2}{N}}{N} \quad (\text{para a população})$$

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1} = \frac{\sum_{i=1}^n X_i^2 - \frac{(\sum_{i=1}^n X_i)^2}{n}}{n - 1} \quad (\text{para a amostra})$$

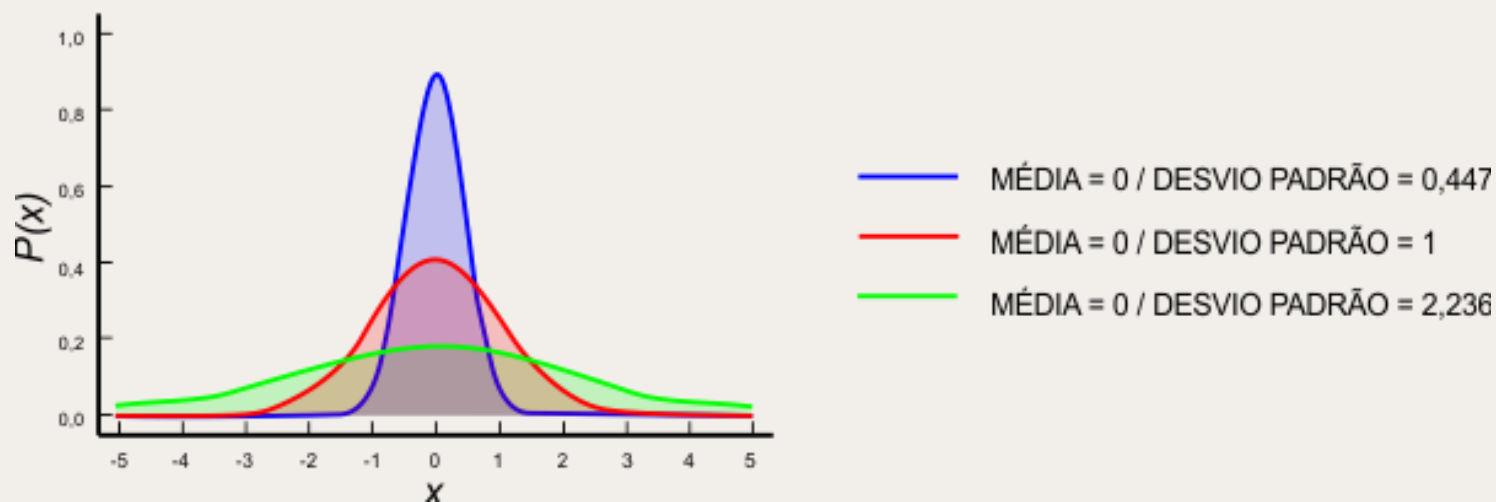
# DESVIO-PADRÃO

- Raiz quadrada da variância, fornece o resultado na mesma ordem de grandeza da variável.
- Quanto menor o desvio-padrão, maior a homogeneidade.

$$\sigma = \sqrt{\sigma^2} \quad (\text{para a população})$$

$$S = \sqrt{S^2} \quad (\text{para a amostra})$$

DIFERENÇA ENTRE DISTRIBUIÇÕES COM MESMA MÉDIA E DESVIOS PADRÃO DIFERENTES



# DESVIO-PADRÃO

## PASSO A PASSO PARA CALCULAR O DESVIO PADRÃO

Medidas	Passo 1	Passo 2	Passo 3	Passo 4	Passo 5
1 440	→ 446-500 = -54	→ $-54^2 = 2916$	→ 2916	$\begin{array}{r} 24966 \\ \div 9 \\ \hline 2774 \end{array}$ <p>10-1</p> <p>Passo 5</p> $\sqrt{2774} = 52,7''$ <p>DESVIO PADRÃO</p>	
2 450	→ 450-500 = -50	→ $-50^2 = 2500$	→ +2500		
3 554	→ 554-500 = 54	→ $54^2 = 2916$	→ +2916		
4 547	→ 547-500 = 47	→ $47^2 = 2209$	→ +2209		
5 486	→ 486-500 = -14	→ $-14^2 = 196$	→ +196		
6 498	→ 498-500 = -2	→ $2^2 = 4$	→ +4		
7 440	→ 440-500 = -60	→ $-60^2 = 3600$	→ +3600		
8 560	→ 560-500 = 60	→ $60^2 = 3600$	→ +3600		
9 451	→ 451-500 = -49	→ $-49^2 = 2401$	→ +2401		
10 508	→ 568-500 = 68	→ $68^2 = 4624$	→ +4624		
			<u>24966</u>		

# COEFICIENTE DE VARIAÇÃO

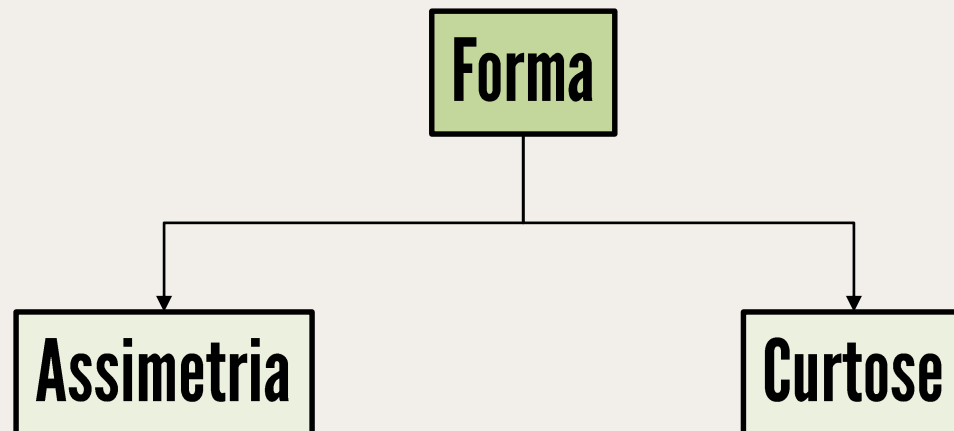
- Medida de dispersão relativa que fornece a variação dos dados em relação à média.
- Quanto menor for o seu valor, mais homogêneos serão os dados (menor a dispersão em torno da média).

$$CV = \frac{\sigma}{\mu} \times 100 (\%) \text{ (para a população)}$$

$$CV = \frac{S}{\bar{X}} \times 100 (\%) \text{ (para a amostra)}$$

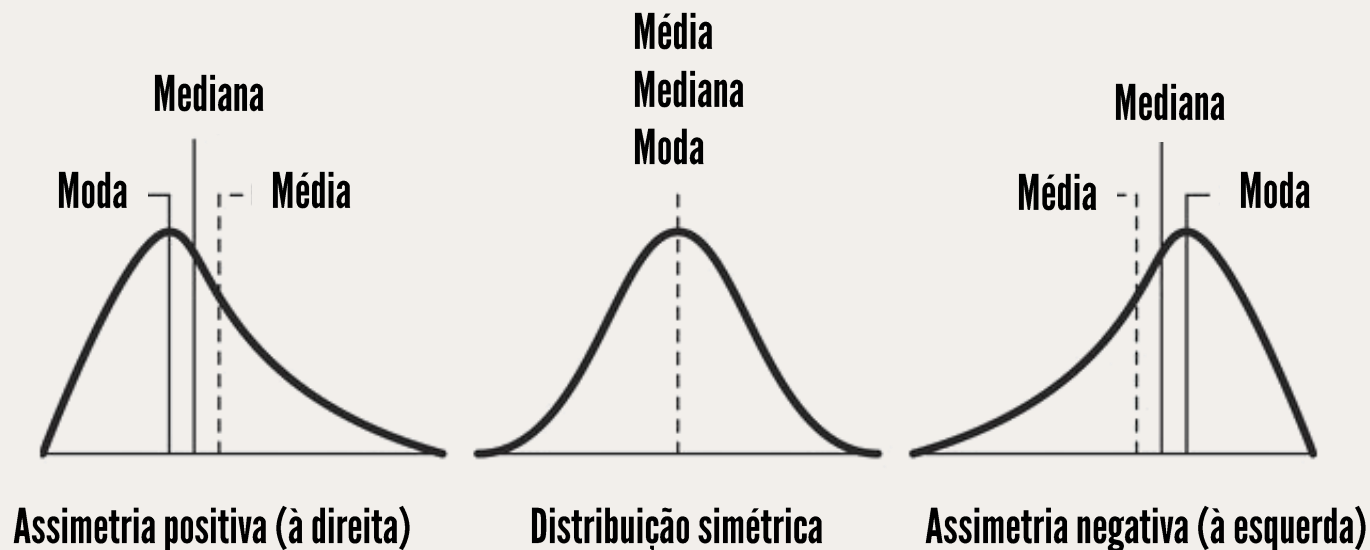
**CV < 30%: baixo, conjunto de dados razoavelmente homogêneos;**  
**CV > 30%: conjunto de dados pode ser considerado heterogêneo.**

# MEDIDAS DE FORMA



# MEDIDAS DE ASSIMETRIA (*SKEWNESS*)

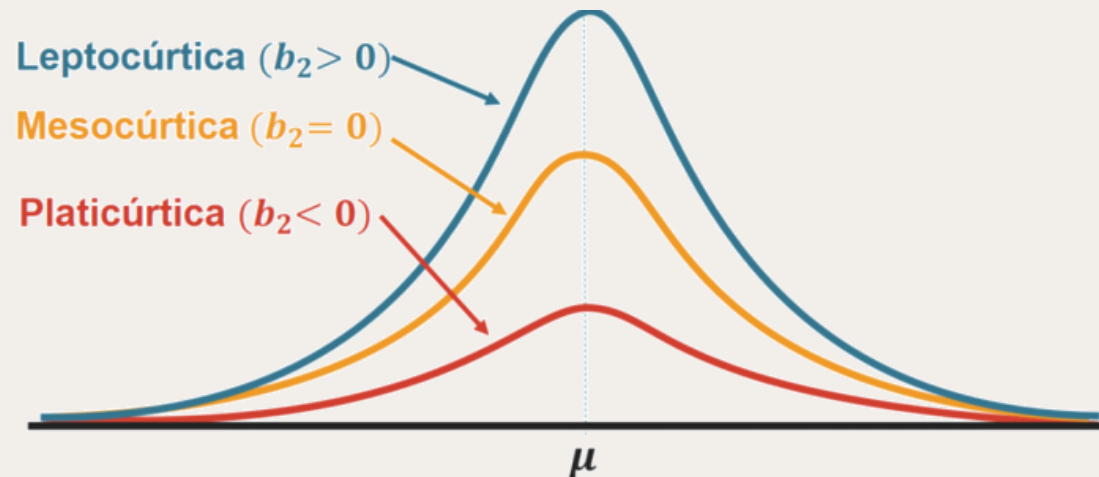
- Referem-se à forma da curva de uma distribuição de frequências.
- *Curva ou distribuição de frequências simétrica*: média, moda e mediana iguais.
- *Curva assimétrica*: média distancia-se da moda, e a mediana situa-se em uma posição intermediária.



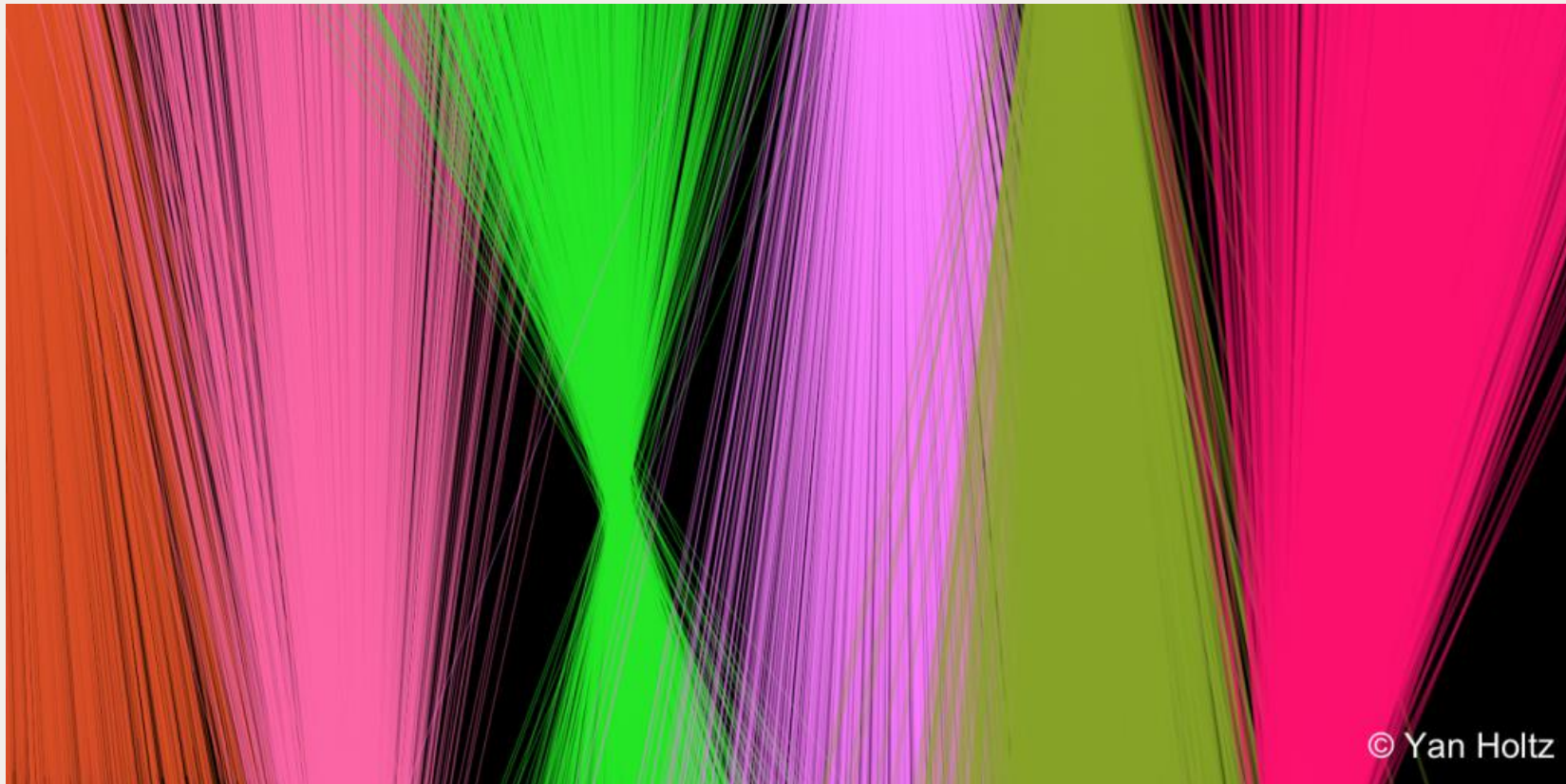


# MEDIDA DE CURTOSE (KURTOSIS)

Grau de achatamento de uma distribuição de frequências (altura do pico da curva) em relação a uma distribuição teórica que geralmente corresponde à distribuição normal.



# ARTE DO DIA FEITA EM R



© Yan Holtz

<https://www.data-to-art.com//>

# REFERÊNCIAS BIBLIOGRÁFICAS

- BARBETTA, Pedro Alberto. Estatística aplicada às ciências sociais. Ed. UFSC, 2008.
- DANCEY, Christine P.; REIDY, John G.; ROWE, Richard. Estatística Sem Matemática para as Ciências da Saúde. Penso Editora, 2017.
- MAGNUSSON, Willian E. Estatística [sem] matemática: a ligação entre as questões e a análise. Planta, 2003.