

Estatística



Dhara Avelino

Adaptado do material disponibilizado pela professora Dr^a. Letícia Raposo. Consulte <https://leticiaraposo.netlify.app/> para vídeos, scripts, aulas e mais informações.

Introdução ao processo de pesquisa

- Por que pesquisamos?
- Desejamos responder questões interessantes sobre o mundo.

- Etapas da pesquisa:



- O que é a estatística?
- Ciência que tem como objetivo a coleta, análise e interpretação de dados qualitativos e quantitativos.
- Pode ser dividida em:
- Descritiva
 - Estatística voltada para a descrição dos dados. Na estatística descritivas são exploradas técnicas de visualização, como gráficos, tabelas e cálculos de medidas resumo das variáveis quantitativas.
- Probabilística
 - Estatística que tem como objetivo utilizar a teoria da probabilidade para

lidar com a incerteza presente nos estudos, uma vez que trabalhamos com amostras, não com todas as unidades de observação.

→ Inferencial

- Estatística onde inferimos para uma população aquilo que foi observado nas amostras.

- Por que estudar estatística?

- Desenvolvimento da análise crítica.
- Independência para analisar nossos próprios dados, sem precisar da ajuda de um estatístico.

Variável

- O que é uma variável?
- Conceito que foi mensurado de alguma forma.
- As variáveis:
- São características de uma população (amostra) em estudo, possível de ser medida, contada ou categorizada.
- Assumem diferentes valores, dependendo da pessoa, situação ou tempo.
- Possuem um e apenas um resultado por respondente.
- Tipos de variáveis:
- Qualitativa
 - Representam características de um indivíduo, objeto ou elemento que não podem ser medidas ou quantitativas.
 - As respostas são dadas em categorias.
- Quantitativa
 - Representam características de um indivíduo, objeto ou elemento resultantes de uma contagem (conjunto finito de valores) ou de uma mensuração (conjunto infinito de valores).

- São em geral mais informativas do que as qualitativas.

- Escala de mensuração:

- Escala qualitativa nominal

- Classifica as unidades em classes ou categorias em relação à característica representada, não estabelecendo qualquer relação de grandeza ou de ordem.

⇒ Exemplos: cor dos olhos, sexo, fumante/não fumante e sadio/doente.

- Escala qualitativa ordinal

- Classifica as unidades em classes ou categorias em relação à característica representada, estabelecendo qualquer relação de grandeza ou de ordem, mas não há intervalos iguais entre os pontos adjacentes na escala.

⇒ Exemplos: avaliação do atendimento, faixa etária, grau de escolaridade e classe social.

- Escala quantitativa intervalar

- Ordena as unidades quanto à característica mensurada e a diferença entre pontos adjacentes é igual, mas não tem um ponto zero (origem).

⇒ Exemplos: temperatura, altitude, QI e ano censitário.

- Escala quantitativa de razão

- Ordena as unidades quanto à característica mensurada e a diferença entre pontos adjacentes é igual, tem um ponto zero (origem) e o valor zero expressa a ausência de quantidade. É possível calcular a razão.

⇒ Exemplos: nº de sintomas de uma doença, renda, idade e distância percorrida.

- Números de categorias e escalas de precisão

- Qualitativas

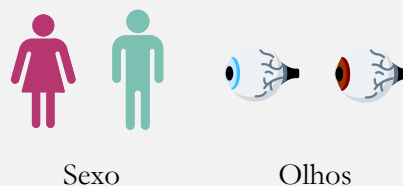
- Dicotômica ou binária- nº igual a 2
- Politômica- nº igual ou maior que 3.

- Quantitativas

- Discreta- relacionada a contagem, nº inteiros, sem casas decimais.
- Contínua- valores representados com casas decimais.

Resumo

Escala qualitativa nominal



Sexo

Olhos

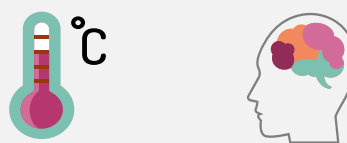
Escala qualitativa ordinal



Escolaridade

Avaliação de atendimento

Escala quantitativa intervalar



Temperatura

QI

Escala quantitativa de razão



Medida

Renda

Introdução à estatística descritiva

Organizar

Resumir

Apresentar os dados

- Nesta etapa observamos determinados aspectos relevantes e começamos a delinear as hipóteses
- Não há conclusões na estatística descritiva, para concluirmos algo devemos utilizar a estatística inferencial.

Os dados estão dizendo algo importante?

Preciso coletar mais dados?

Vale a pena fazer uma análise?



- A estatística descritiva pode ser dividida em três grupos:
 - Univariada
 - Análise independente de cada variável
 - Bivariada
 - Análise de duas variáveis ao mesmo tempo, buscando um relacionamento entre as variáveis.

→ Multivariada

- Análise de três ou mais variáveis, procurando saber como essas variáveis em conjunto podem influenciar um dado evento.

- Estatística Descritiva Univariada

Representando as variáveis

Variável Qualitativa

Tabela de distribuição de frequências

Gráficos:
Barras
Setores

Variável Quantitativa

Tabela de distribuição de frequências

Gráficos:
Histograma
Densidade
Boxplot
Linha

Medidas-resumo:
Posição
Dispersão
Forma

Estatística Descritiva Univariada Qualitativa

- Tabela de distribuição de frequências

Tipo ABO	F_i	$F_r(\%)$	F_{ac}	$F_{rac}(\%)$
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

- Dado bruto informado. Neste caso seria o número total de pessoas de cada grupo sanguíneo

→ Frequência absoluta acumulada

- Dado bruto informado somado ao dado anterior. Neste caso seria o número de pessoas com tipo sanguíneo A+ somado ao número de pessoas com o tipo sanguíneo A- e assim em diante.

→ Frequência relativa

- Dado bruto dividido pelo número total da amostra multiplicado por 100. Neste caso seria o número de pessoas com tipo sanguíneo A+ dividido pelo total 60 e multiplicado por 100, já que o valor é expresso em porcentagem

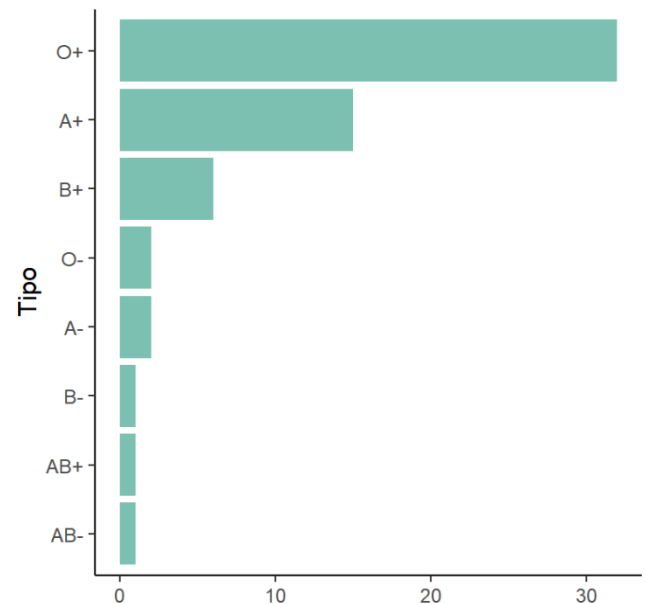
→ Frequência relativa acumulada

- Frequência relativa somado ao dado anterior. Neste caso seria a frequência relativa de pessoas com tipo sanguíneo A+ somado à frequência relativa de pessoas com o tipo sanguíneo A- e assim em diante.

- Gráfico de barras

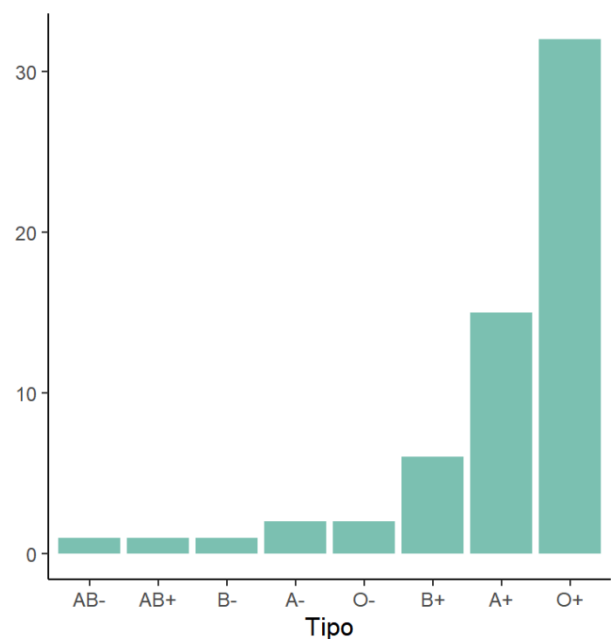
→ Barras horizontais

- Muito usadas quando os nomes das categorias são extensos.



→ Barras verticais

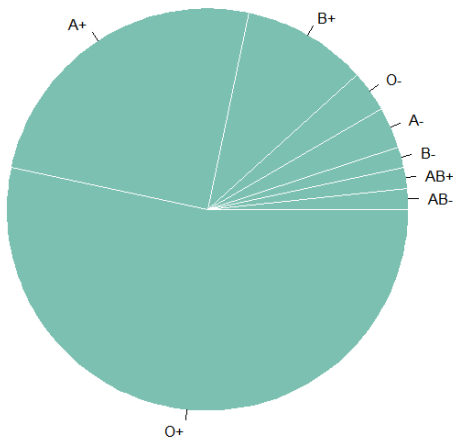
- Muito usadas com variáveis ordinais.



Dica!!!

Organizar as categorias ordinais da esquerda para a direita para serem visualizadas em sequência.

- Gráfico de setores
 - Representa as frequências relativas de cada possível categoria.
 - É frequentemente usado para mostrar porcentagem, em que a soma dos setores é igual a 100%.
 - Erros comuns
 - Usar 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.



Importante!!!

O gráfico de setores deve ser evitado e substituído pelo gráfico de barras sempre que possível.

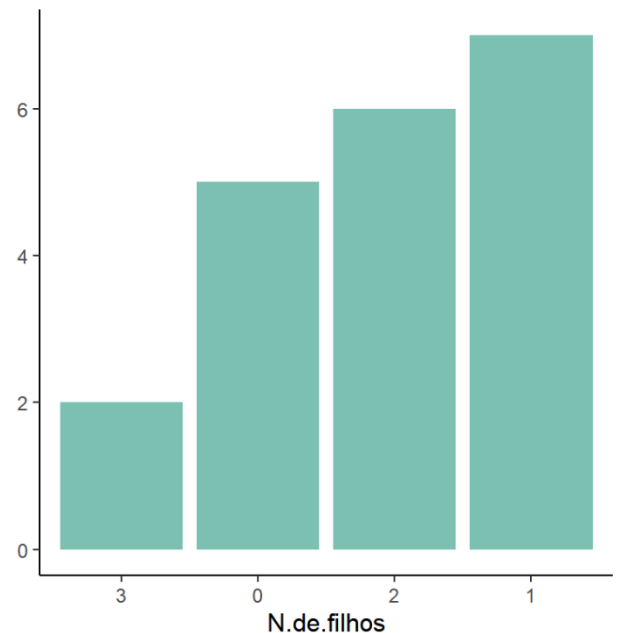
Estatística Descritiva Univariada Quantitativa

VARIÁVEIS DISCRETAS

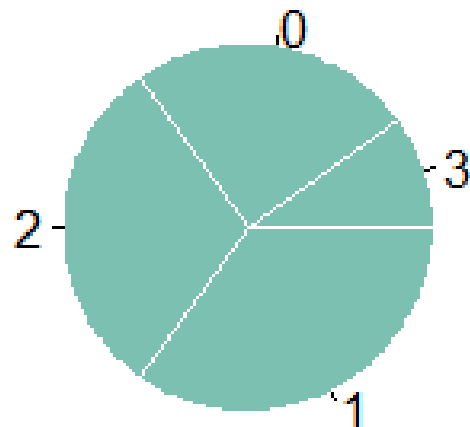
- Tabela de distribuição de frequências

Nº de filhos	F_i	$F_n(\%)$	F_{ac}	$F_{rac}(\%)$
0	5	25	5	25
1	7	35	12	60
2	6	30	18	90
3	2	10	20	100
Total	20	100		

- Gráfico de barras



- Gráfico de setores



VARIÁVEIS CONTÍNUAS

- Tabela de distribuição de frequências
- 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.
- 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.

Classe	F_i	$F_n(\%)$	F_{ac}	$F_{ac}(\%)$
[3,5;4,5)	5	16,67	5	16,67
[4,5;5,5)	9	30	14	46,67
[5,5;6,5)	7	23,33	21	70
[6,5;7,5)	7	23,33	28	93,33
[7,5;8,5)	1	3,33	29	97,67
[8,5;9,5)	1	3,33	30	100
Soma	30	100		

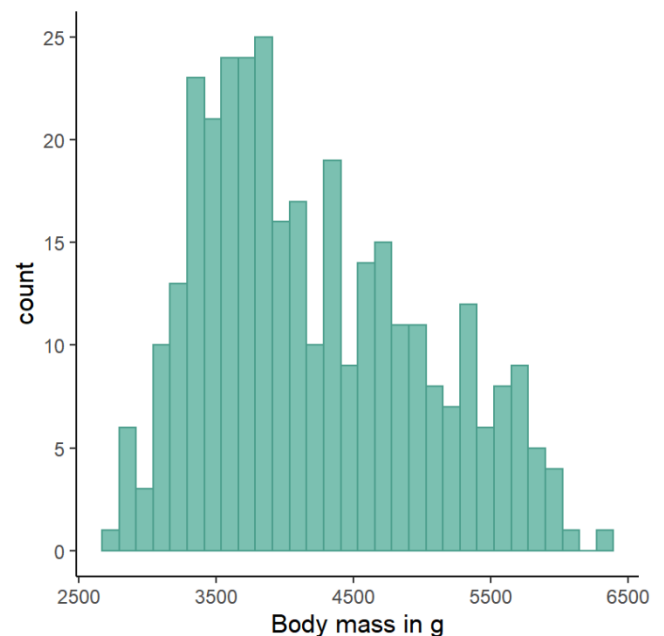
Dica!!!

Usar, aproximadamente, raiz quadrada (n) classes, em que n é a quantidade de valores.

→ Permite identificar a distribuição e a frequência dos dados.

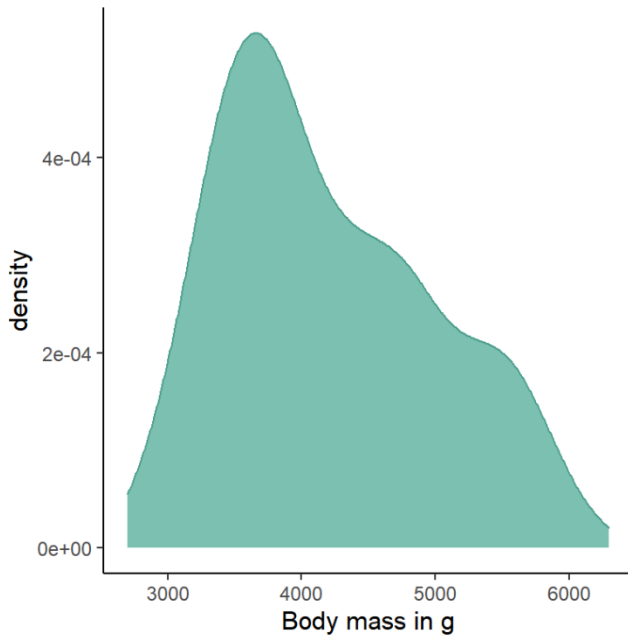
Importante!!!

- As barras dos histogramas são normalmente chamadas de “bins”.
- Tenta vários tamanhos de bins, isso pode levar a conclusões diferentes.
- Não use larguras de bins diferentes em um mesmo gráfico.



- Gráfico 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.

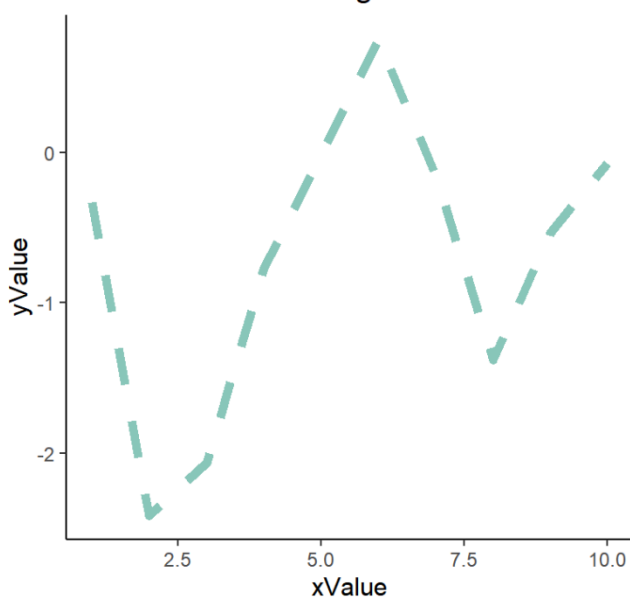
- Gráfico de Densidade
- Representação da distribuição de uma variável numérica.
- É uma versão suavizada do histograma e é usada no mesmo conceito.



O boxplot será abordado separadamente em outro momento!

- Gráfico de Linha
 - Mostra a evolução ou tendência dos dados de uma variável quantitativa, geralmente contínua, em intervalos regulares.
 - Muito comum em análises de séries temporais

Evolution of something



MEDIDAS DE POSIÇÃO

- As medidas de posição podem ser divididas em:
 - Tendência central:
 - Média
 - Moda
 - Mediana
 - Separatrizes:
 - Quartis
 - Decis
 - Percentis
- Medidas de tendência central
 - Média Simples: Soma dos valores dividida pelo número de valores observados (n).

Média Simples

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

- A média é um resumo dos dados e, por isso, pode esconder informações relevantes.
- Média Ponderada: A ponderação é feita sempre que precisamos dar mais importância a um caso do que a outro (atribuir pesos diferentes).

Média Ponderada

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

- A média resume o conjunto de dados em termos de uma posição central, 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.

→ Mediana

- 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.

Mediana



Ímpar

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

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

Par

5, 13, 9, 7, 1, 9, 2, 9, 11, 6

1, 2, 5, 6, 7, **9**, 9, 9, 11, 13

$$7 + 9 = 16/2 = 8$$

→ Moda

- Correspondente à observação que ocorre com maior frequência.
- A moda é a única medida de posição que pode ser utilizada para variáveis qualitativas.

Resumo



Média

Indicada quando não há valores extremos nos dados.



Mediana

Ótima quando há valores extremos nos dados.



Moda

Mais útil quando há dados categóricos.

Importante!!!



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
 - Quartis
 - Divide os dados em 4 partes, com 25% dos dados em cada uma delas.
 - Decis
 - Divide os dados em 10 partes, com 10% dos dados em cada uma delas.
 - Percentis
 - Divide os dados em 100 partes, com 1% dos dados em cada uma delas.

L	Quartis	
E	Dados	\div 4 $=$ 25 %
M	Centis	
B	Dados	\div 10 $=$ 10 %
R	Percentis	
E	Dados	\div 100 $=$ 1 %
T		
E		

- Desvio: Diferença entre cada valor observado e a média da variável.
- Desvio-médio absoluto: Média aritmética dos desvios absolutos.

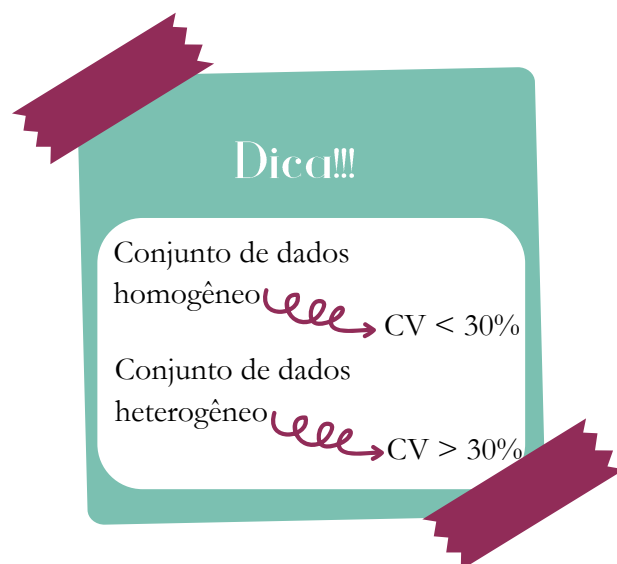
- Variância
 - Avalia o quanto os dados estão dispersos em relação à média aritmética.
 - Quanto maior a variância, maior a dispersão dos dados.
 - O valor tende a ser muito grande e de difícil interpretação.

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

- 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).
 - Por ser adimensional, permite a comparação de variáveis com unidades diferentes.

MEDIDAS DE DISPERSÃO

- As medidas de dispersão podem ser divididas em:
 - Amplitude
 - Desvio-médio absoluto
 - Variância
 - Desvio-padrão
 - Coeficiente de variação
- 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.
- Desvio-médio absoluto



Resumo

Conjunto de dados:

0; 6; 7; 7; 7; 7,5; 7,5

Amplitude

$$A = X_{\text{máx}} - X_{\text{mín}}$$

$$A = 7,5 - 0 = 7,5$$

Desvio médio absoluto

$$D_m = \frac{\sum_{i=1}^n |X_i - X|}{n}$$

$$\text{Média} = 6$$

$$|0-6| = 6 \quad |6-6| = 0$$

$$|7-6| = 1 \quad |7,5-6| = 1,5$$

$$D_m = 6 + 0 + 1 + 1 + 1 + 1,5 + 1,5 = 12/7 = 1,71$$

Variância

$$S^2 = \frac{\sum_{i=1}^n (X_i - X)^2}{n - 1}$$

$$\text{Média} = 6$$

$$|0-6| = 6 \quad |6-6| = 0$$

$$|7-6| = 1 \quad |7,5-6| = 1,5$$

$$D_m = 6^2 + 0^2 + 1^2 + 1^2 + 1^2 + 1,5^2 + 1,5^2 = 43,5/(7-1) = 7,25$$

Desvio padrão

$$S = \sqrt{S^2}$$

$$S = \sqrt{7,25^2} = 2,69$$

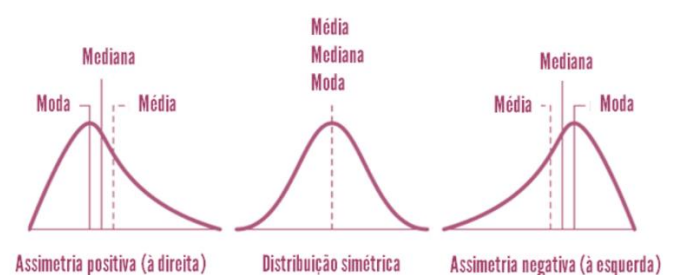
Coefficiente de variação

$$CV = \frac{S}{X} \cdot 100\%$$

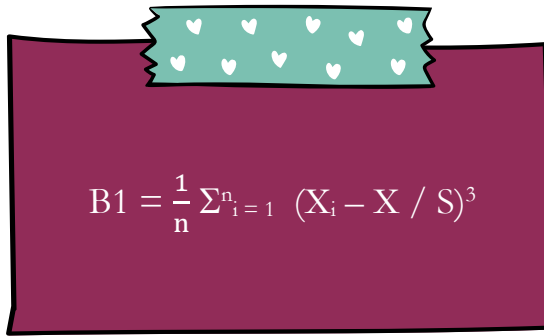
$$CV = \frac{2,69}{6} \cdot 100\% = 44,83\%$$

MEDIDAS DE FORMA

- As medidas de forma podem ser divididas em:
 - Assimetria
 - Curtose
- Assimetria
 - Refere-se à forma da curva de uma distribuição de frequências
 - Curva simétrica: média, moda e mediana iguais.
 - Curva assimétrica: média distancia-se da moda, e a mediana situa-se em uma posição intermediária.



→ Primeiro coeficiente de assimetria de Pearson

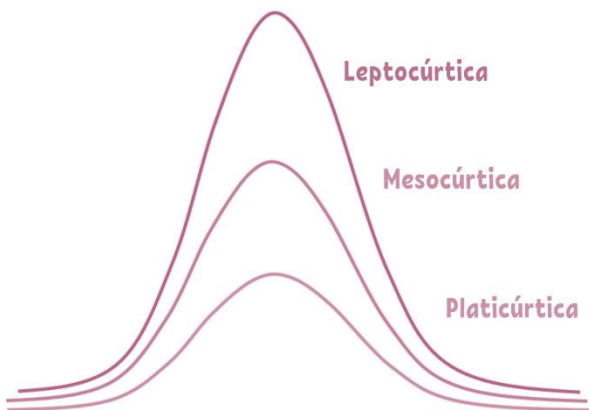


$$B1 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X} / S)^3$$

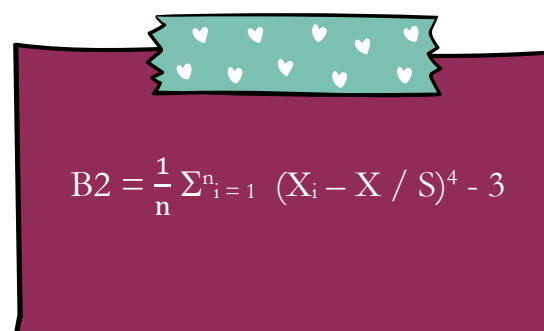
- Se $b1 = 0$: A distribuição é simétrica.
- Se $b1 > 0$: A distribuição é assimétrica positiva (à direita).
- Se $b1 < 0$: A distribuição é assimétrica negativa (à esquerda).

• Curtose

→ 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.



→ Coeficiente de Curtose



$$B2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X} / S)^4 - 3$$

- Se $b2 = 0$: A distribuição é mesocúrtica.
- Se $b2 > 0$: A distribuição é leptocúrtica
- Se $b2 < 0$: A distribuição é platicúrtica.

BOXPLOT

• Boxplot

→ Representação gráfica de cinco medidas de posição ou localização de determinada variável.

→ Permite avaliar a simetria e distribuição dos dados, e também propicia a perspectiva visual da presença ou não de dados discrepantes (outliers univariados).

