

The background is a dark navy blue. It features abstract geometric line art in teal and purple. In the top-left corner, there are several overlapping lines forming a series of connected 'V' and inverted 'V' shapes. In the bottom-right corner, there are more lines forming a similar geometric pattern, including a large 'V' shape and some horizontal segments.

# AceLeraDev Data Science

Análise de dados exploratória

# Análise Exploratória

/

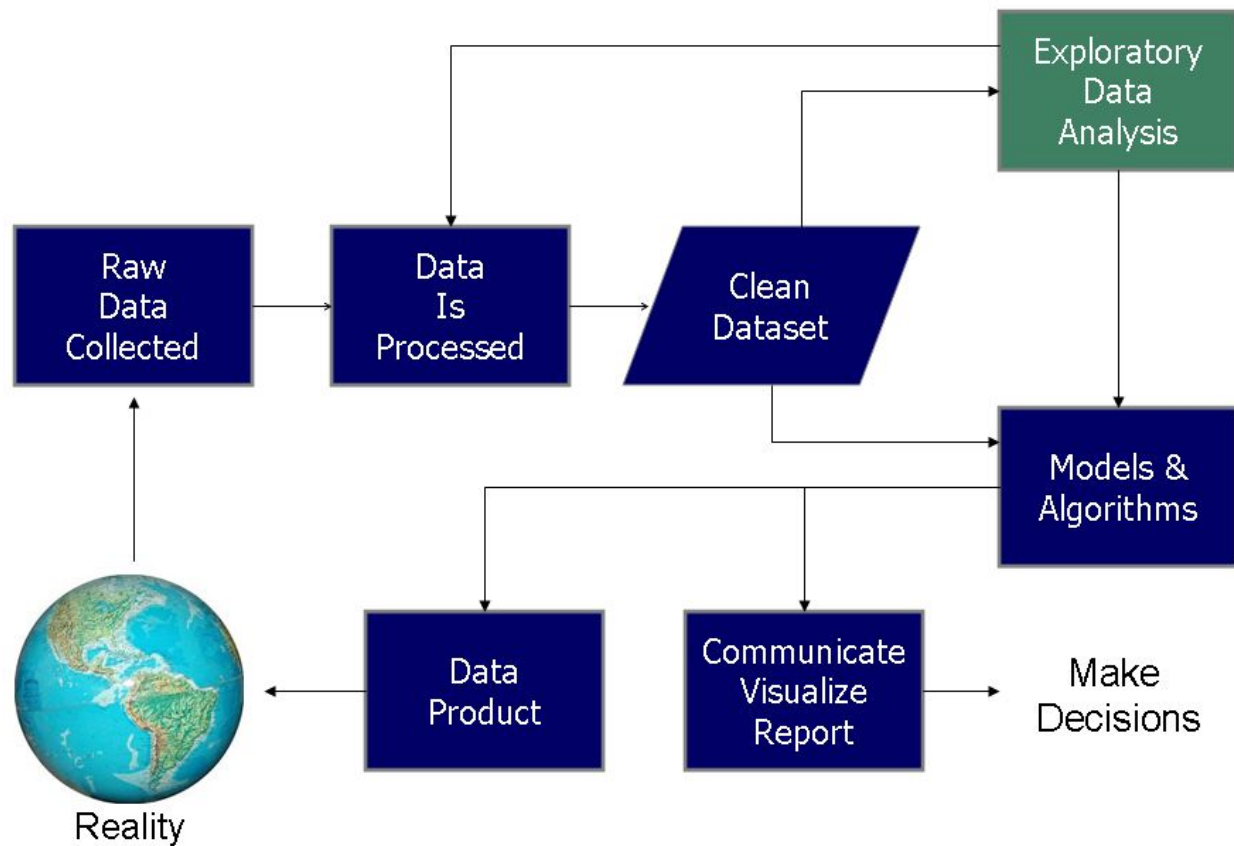
- Análise de dados exploratória (EDA)
  - *"Procedimentos para analisar dados, técnicas para interpretar os resultados de tais procedimentos, formas de planejar a reunião dos dados para tornar sua análise mais fácil, mais precisa ou mais exata e toda a maquinaria e os resultados da estatística (matemática) que se aplicam a análise de dados." - John W. Tukey*

# Análise Exploratória

/

- Análise de dados exploratória (EDA):
  - Sugerir **hipóteses** sobre as causas dos fenômenos observados;
  - Avaliar **pressupostos** sobre os quais a inferência estatística se baseará;
  - Apoiar a seleção de ferramentas e técnicas estatísticas apropriadas;
  - Oferecer uma base para coleta posterior de dados por meio de pesquisas e experimentos;

# Data Science Process



The background is a dark navy blue. It features abstract geometric line art in teal and purple. In the top-left corner, there are several interconnected lines forming a series of triangles and polygons. In the bottom-right corner, there are more geometric shapes, including a large triangle and some smaller polygons, also formed by teal and purple lines.

# AceLeraDev

# Data Science

Estatística descritiva univariada 1

# Estatística descritiva univariada

/

Média (*Mean/Average*):

- Ponto central de um conjunto de informações definido pela somatória das informações de uma conjunto dividido pela quantidade de informações
- **$M = \text{SUM}(\text{valores}) / \text{COUNT}(\text{valores})$**

# Estatística descritiva univariada

Média (*Mean/Average*):

- $[10, 20, 20, 12, 13, 20, 21, 25] = \text{Média } 17,62$
- $[10, 20, 20, 12, 13, 20, 21, 240] = \text{Média } 44,5$

# Estatística descritiva univariada

/

Mediana (*Median*):

- Valor que separa a metade das informações em 2 **conjuntos de quantidade iguais**.
- Para conjuntos de quantidade par, é a **média** dos 2 valores centrais



# Estatística descritiva univariada

/

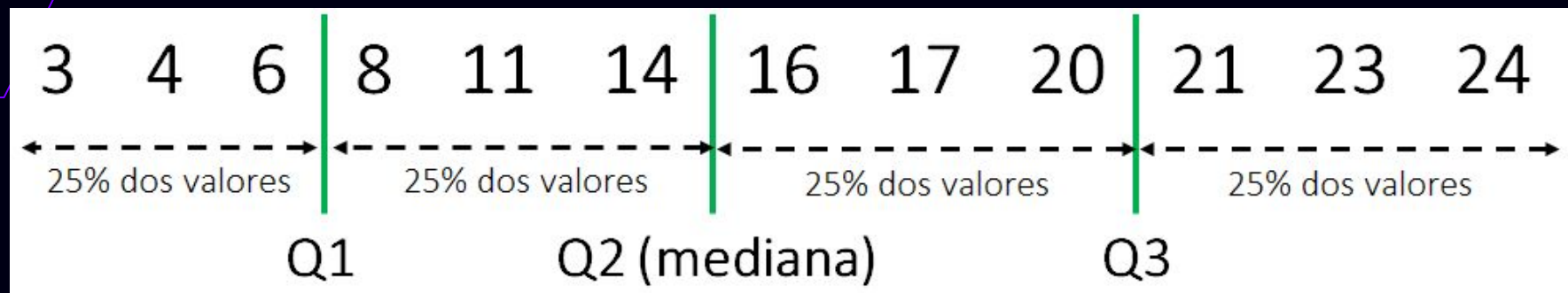
Mediana (*Median*):

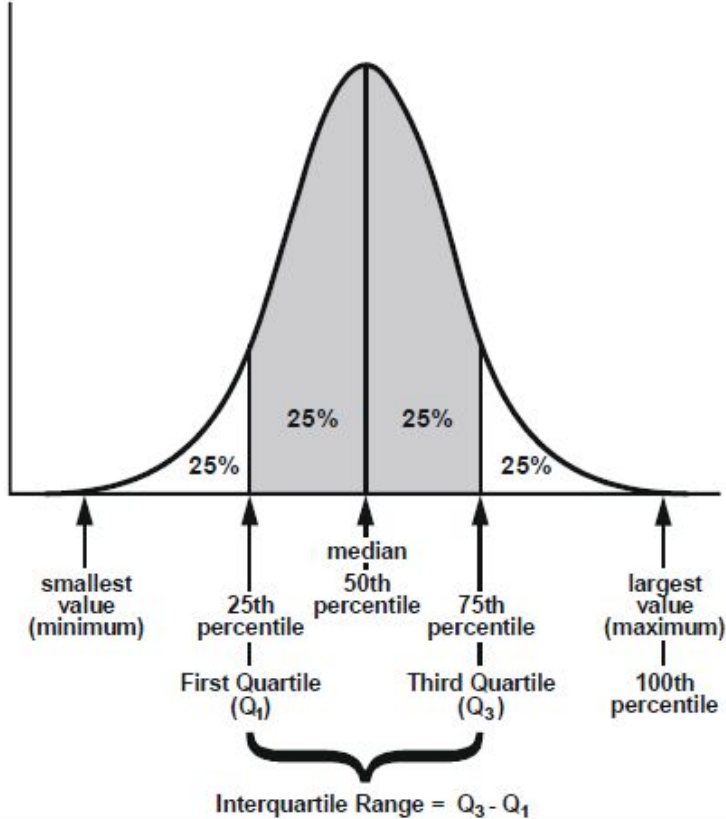
- $[1, 3, 3, 6, 7, 8, 9] = \text{Mediana } 6$
- $[1, 2, 3, 4, 5, 6, 8, 9] = \text{Mediana } 4,5$

# Estatística descritiva univariada

Quartis e Percentis:

- Quartis:
  - Divide os dados em **4 conjuntos de dados**;
  - 25% em cada cada conjunto
- Percentis:
  - Divide os dados em **100 parte do todo**
  - **1%** acumulado em cada segmento

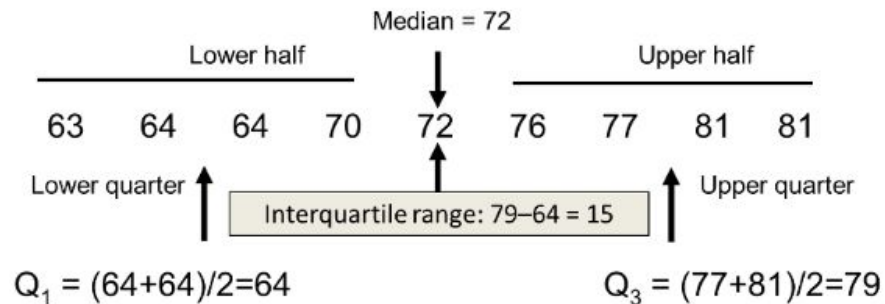
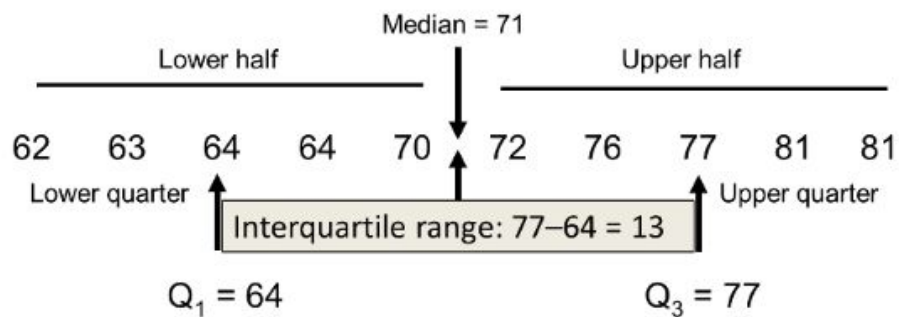




# Estatística descritiva univariada

Amplitude Interquartil (*InterQuartile Range (IQR)*)

- O 50% central dos valores quando ordenados do menor para o maior
  - Encontra-se a mediana (valor do meio) **da menor e da maior metade dos dados.**
  - São o **quartil 1 (Q1)** e o **quartil 3 (Q3)**. A amplitude interquartil é a **diferença entre Q3 e Q1.**



The background is a dark navy blue. It features several thin, light blue and purple lines that form geometric shapes, primarily hexagons, in the corners. These lines are not solid but appear as outlines or partial outlines of larger shapes.

# AceLeraDev

# Data Science

Estatística descritiva univariada 2

# Estatística descritiva univariada

Desvio Padrão (*standard deviation-std*):

- Medida que expressa o **grau de dispersão de um conjunto de dados**.
- Indica o quanto um conjunto de dados é uniforme
- Quanto mais próximo de 0 for o desvio padrão, mais homogêneo são os dados.



## Desvio Padrão (Dp)

$$Dp = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}$$

$x_i$  = valor individual

$\bar{x}$  = média dos valores

$n$  = número de valores

altura	média	altura - média	(altura - média)^2
1.55	1.68	-0.13	0.0178
1.70	1.68	0.02	0.0003
1.80	1.68	0.12	0.0136

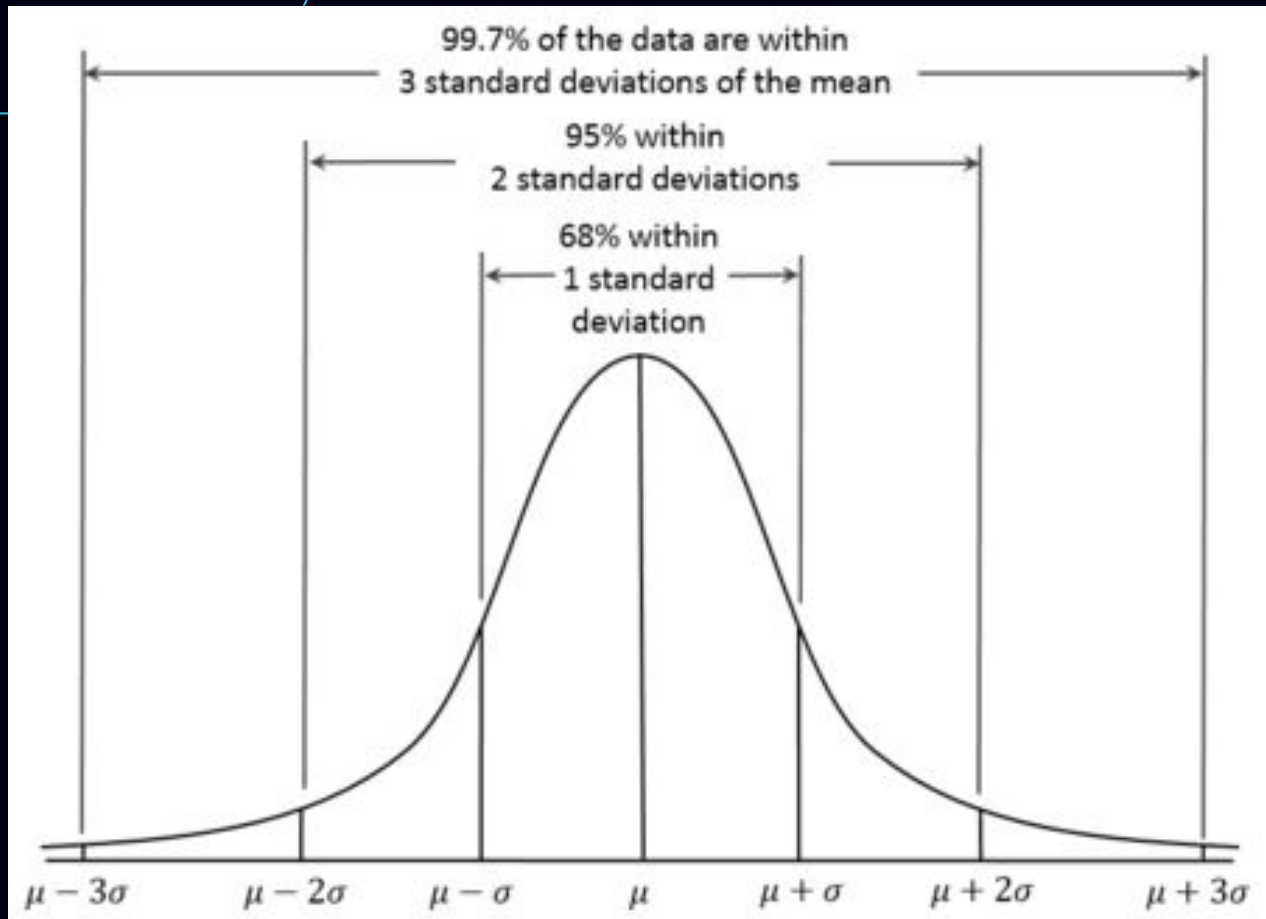
$\text{sum}(\text{altura} - \text{média})^2$
0.0317

número elementos
3

$\text{sum}(\text{altura} - \text{média})^2 / \text{número elementos}$
0.0106

$\text{raiz}(\text{sum}(\text{altura} - \text{média})^2 / \text{número elementos})$
0.103

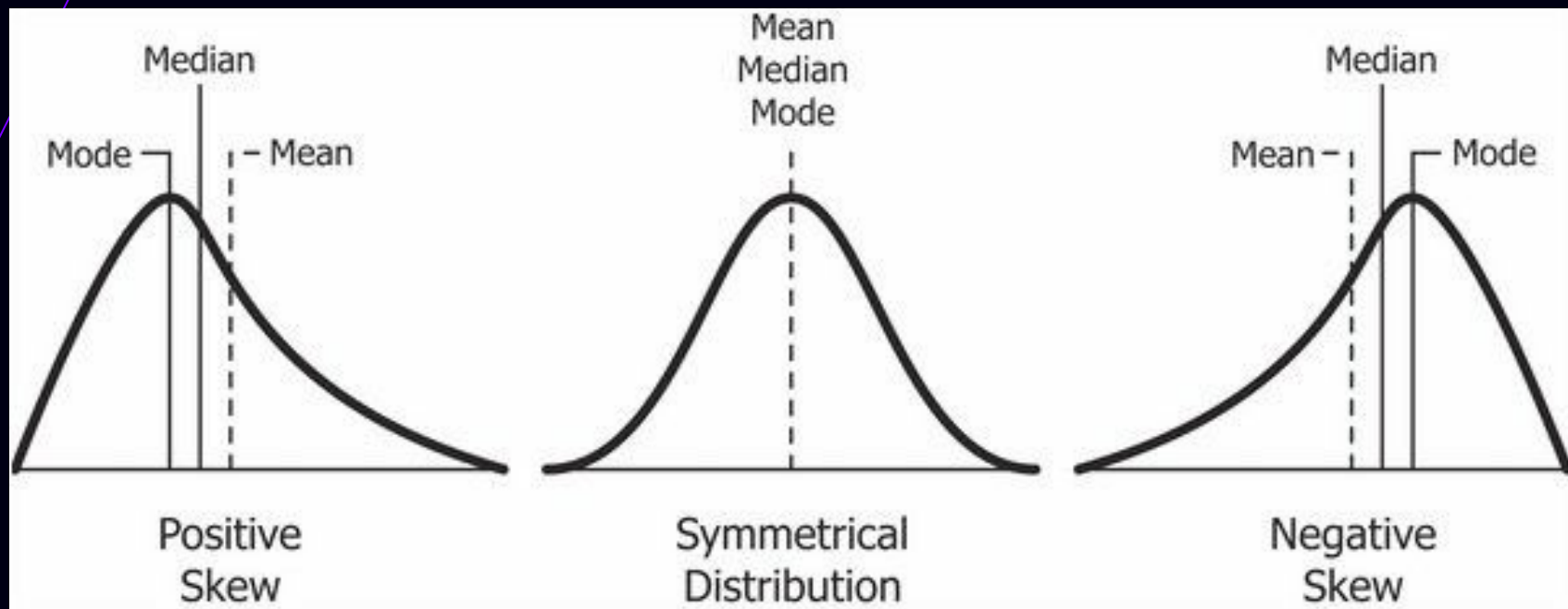
Desvio Padrão



# Estatística descritiva / univariada

Assimetria (*skewness*):

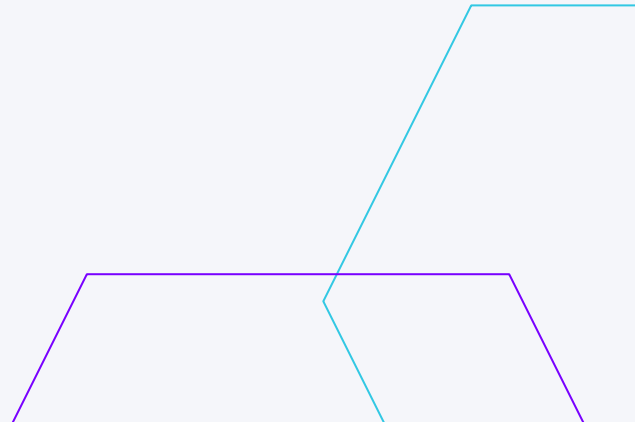
- É o grau de distorção da curva simétrica a distribuição normal
- Ele mede a falta de simetria na distribuição de dados.
- Diferencia valores extremos em uma cauda versus na outra
- Uma distribuição simétrica terá uma assimetria de 0.





# Estatística descritiva / univariada

Curtose (kurtosis):

- Curtose é uma medida de dispersão que caracteriza o "achatamento" da curva da função de distribuição.
- 

# Estatística descritiva univariada

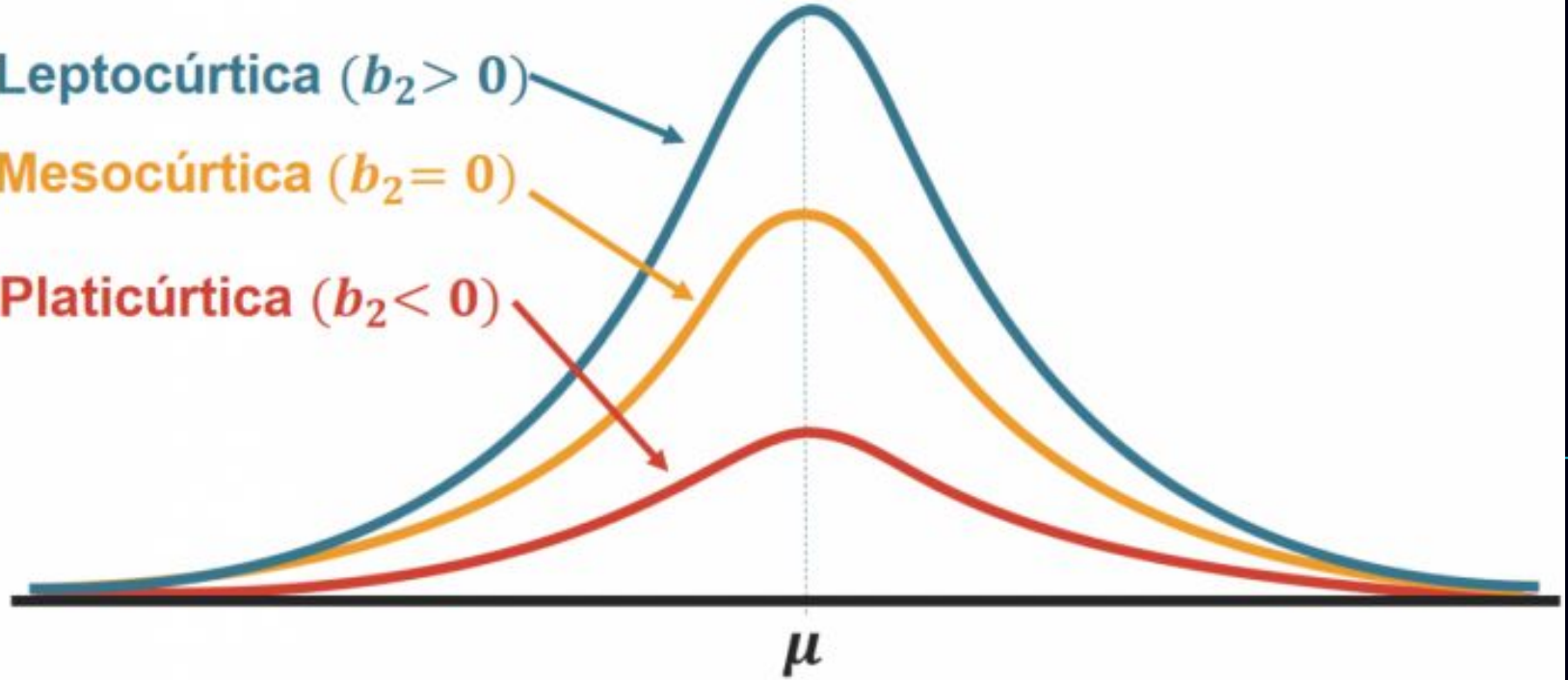
Curtose (kurtosis):

- **mesócurticas(0):**
  - achatamento da distribuição normal;
- **leptocúrtica(> 0):**
  - possui a curva da função de distribuição mais afunilada
  - pico mais alto do que a distribuição normal
  - possui caudas pesadas.
- **platicúrtica (<0):**
  - função de distribuição é mais achatada do que a distribuição normal .

Leptocúrtica ( $b_2 > 0$ )

Mesocúrtica ( $b_2 = 0$ )

Platicúrtica ( $b_2 < 0$ )





# Normalizar e Padronizar

/

- Transformar todas as variáveis na mesma ordem de grandeza.
- Padronizar as variáveis irá resultar em uma média igual a 0 e um desvio padrão igual a 1.
- Normalizar tem como objetivo colocar as variáveis dentro do intervalo de 0 e 1, caso tenha resultado negativo -1 e 1.