

# Estatística Descritiva

# Sumário

- ❑ Análise paramétrica univariada
- ❑ Análise multivariada com gráficos
- ❑ Medidas de associação

# Análise univariada

Uma breve introdução às Funções de  
Distribuição de Probabilidade





# Análise paramétrica

- ❑ A distribuição de probabilidade é uma função que descreve todos os resultados possíveis de um experimento aleatório e atribui a cada um deles uma probabilidade, obedecendo aos axiomas da probabilidade.
  - ❑ Por exemplo, a soma das probabilidades para variáveis discretas ou a integral da densidade para variáveis contínuas é igual a 1.

- ❑ Para variáveis discretas, temos a função de massa de probabilidade (PMF), que atribui a cada valor:  $f(x) = P(X = x)$

- ❑ Para as contínuas, temos a função densidade de probabilidade (PDF)  $f(x)$ , que deve satisfazer:

$$\int_{-\infty}^{\infty} f(x) dx = 1$$

- ❑ A probabilidade de  $X$  estar em um intervalo  $[a, b]$  é dada por:

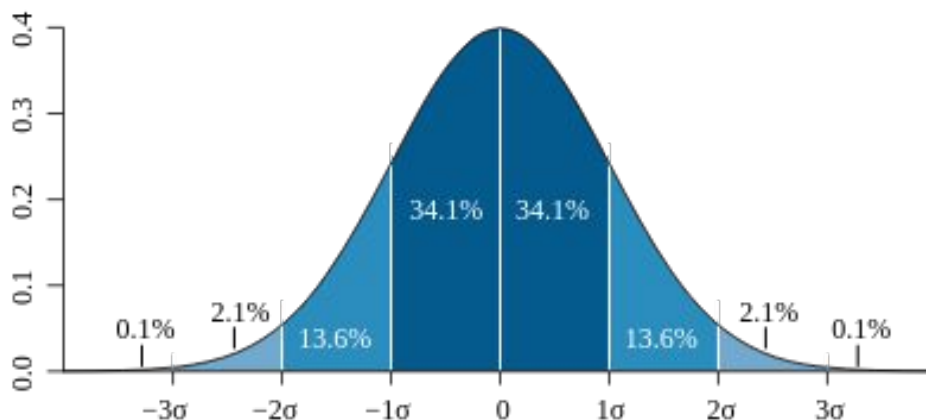
$$P(a \leq X \leq b) = \int_a^b f(x) dx$$

# Análise paramétrica

Exemplos de distribuições conhecidas:

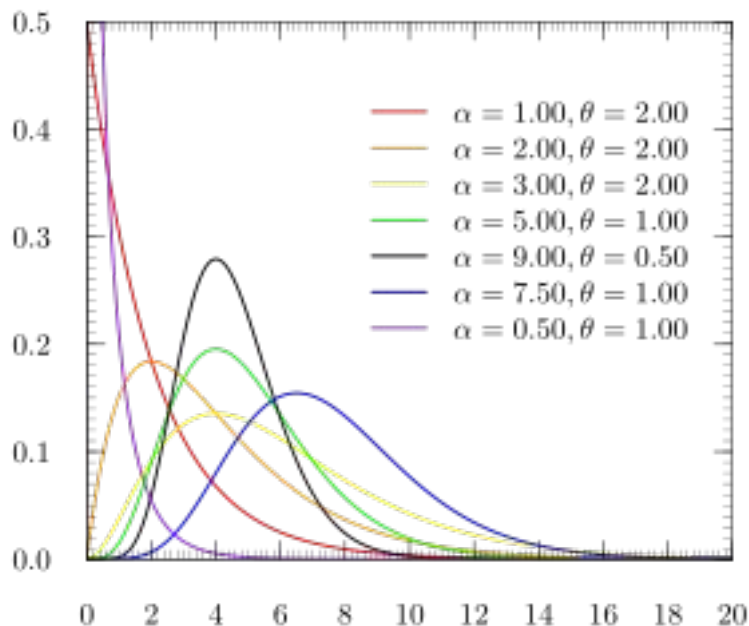
- ❑ **Distribuição normal:** muita utilizada para representar eventos naturais.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad x \in \mathbb{R},$$



# Análise paramétrica

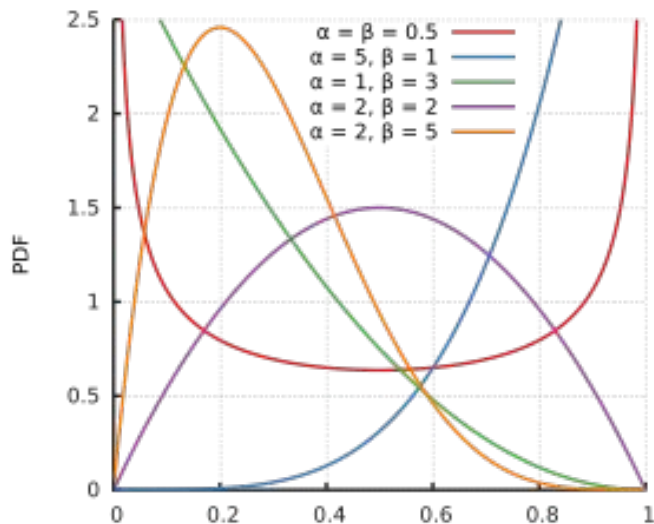
- ❑ **Distribuição Gamma:** utilizada para modelar o tempo até um evento de referência (vida útil, sobrevivência, etc).
- ❑ Utiliza um parâmetro de forma  $\alpha$ , e um parâmetro de escala  $\theta$ .



$$f(x; k, \theta) = \frac{x^{k-1} e^{-x/\theta}}{\theta^k \Gamma(k)}, \quad x \geq 0$$

# Análise paramétrica

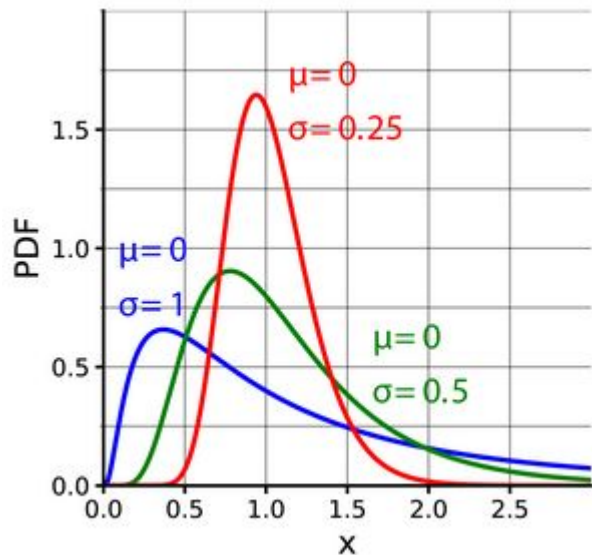
- ❑ **Distribuição Beta:** definida no intervalo  $[0,1]$  utilizada para modelar variáveis limitadas.
- ❑ Bastante flexível, utiliza o parâmetros  $\alpha$  e  $\beta$  para definir a forma da distribuição.



$$f(x; \alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}, \quad 0 \leq x \leq 1$$

# Análise paramétrica

- ❑ **Log-Normal:** curva positivamente assimétrica, modela eventos de valor estritamente e monotonicamente positivos.
- ❑ Utiliza os mesmo parâmetros da distribuição normal.



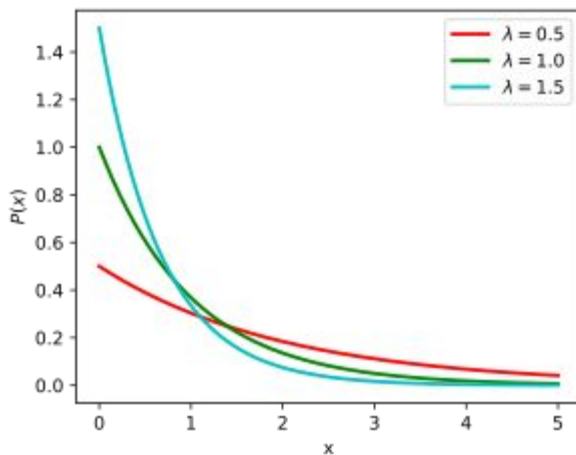
$$f(x; \mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right), \quad x > 0$$



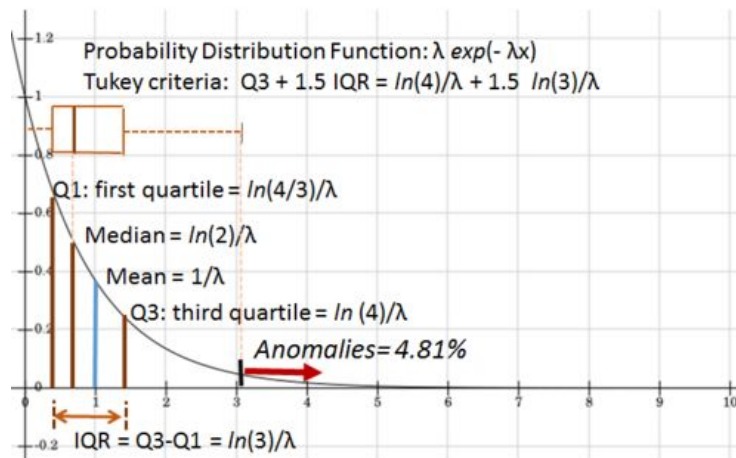
# Análise paramétrica

- ❑ **Distribuição Exponencial:** utilizada para modelar o tempo entre eventos em um processo de Poisson, com taxa  $\lambda$ :

$$f(x) = \lambda e^{-\lambda x}, \quad x \geq 0.$$



**Tukey criteria for anomalies.**  
**Exponential probability distribution function.**





# Análise paramétrica

A escolha da função de distribuição que melhor se ajusta a uma amostra é fundamental para a modelagem estatística e envolve diversas etapas e técnicas:

## Visualização dos Dados:

- ❑ Histograma: Permite comparar a forma empírica da distribuição dos dados com as curvas teóricas das distribuições candidatas.
- ❑ Gráficos Q-Q (Quantile-Quantile): Confrontam os quantis da amostra com os quantis teóricos de uma distribuição proposta. Se os pontos se alinharem aproximadamente em uma reta, isso indica uma boa aderência.
- ❑ P-P (Probability-Probability) Plot: Compara as probabilidades acumuladas empíricas com as teóricas.



# Análise paramétrica

## Estimação dos Parâmetros:

- ❑ Máxima Verossimilhança (MLE): Método que busca os valores dos parâmetros que maximizam a probabilidade (ou verossimilhança) de observar os dados amostrais.
- ❑ Método dos Momentos: Utiliza os momentos amostrais para estimar os parâmetros da distribuição.
- ❑ Outros métodos podem incluir técnicas bayesianas, especialmente quando se incorpora conhecimento prévio.



# Análise paramétrica

## Testes de Aderência (Goodness-of-Fit):

- ❑ Teste de Kolmogorov-Smirnov (K-S): Compara a função de distribuição acumulada empírica com a teórica para avaliar se as diferenças são estatisticamente significativas.
- ❑ Teste de Anderson-Darling: Uma variação do K-S que dá mais peso às caudas da distribuição.
- ❑ Teste Qui-Quadrado: Compara as frequências observadas e as esperadas em classes definidas; útil quando se trabalha com dados categorizados ou quando o tamanho da amostra é grande.

## CrITÉrios de Informação:

- ❑ Akaike Information Criterion (AIC) e Bayesian Information Criterion (BIC): comparam diferentes modelos de distribuição, considerando o ajuste aos dados e a complexidade do modelo. O modelo com menor AIC ou BIC é geralmente preferido.

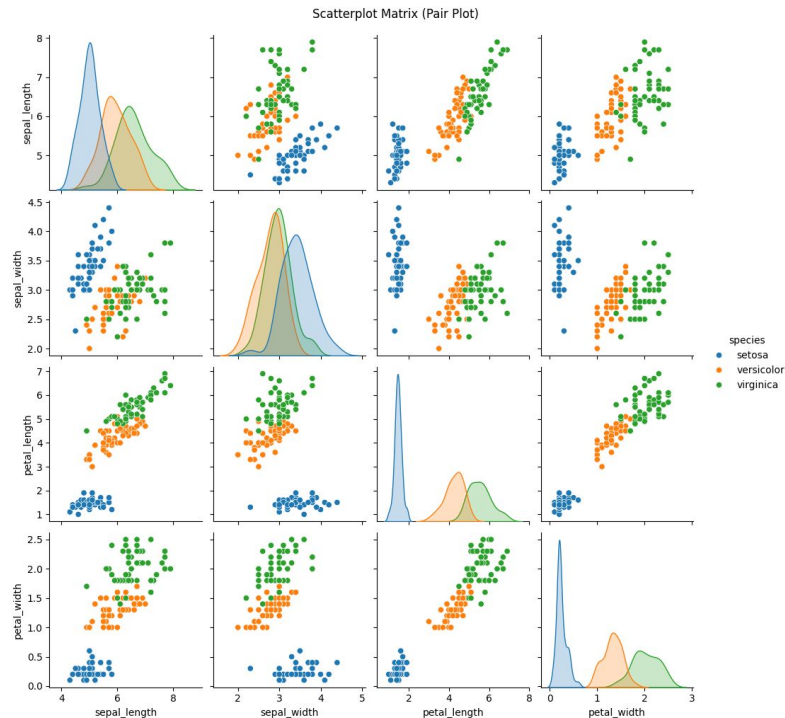
# Análise multivariada com Gráficos



# Gráfico de dispersão

Exibe gráficos de dispersão para cada par de variáveis em um conjunto de dados, permitindo a inspeção visual de correlações e relações bivariadas.

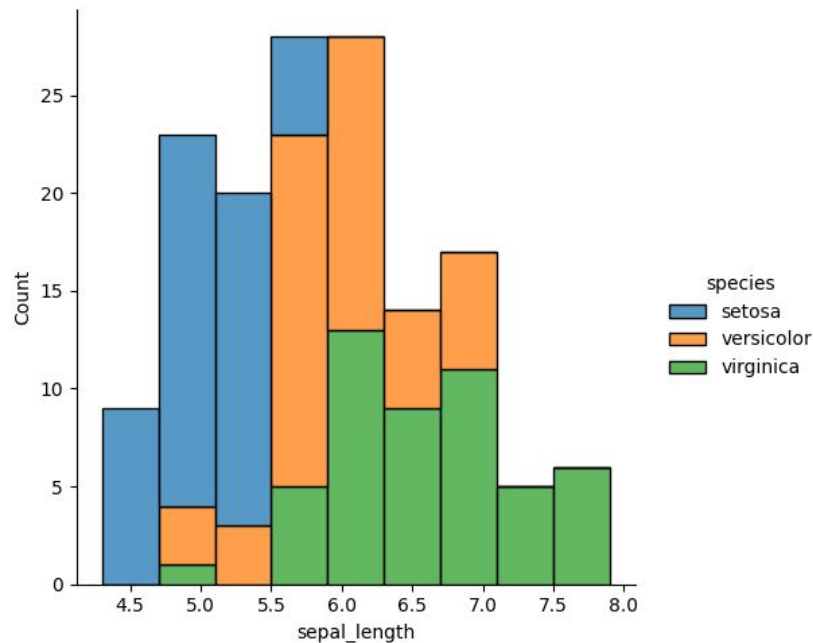
- ❑ Comumente associado à criação de matrizes de dispersão (onde vários recortes são comparados ao mesmo tempo);
- ❑ Ideal para uma análise exploratória inicial:



# Gráfico de frequência

Auxilia a visualização da contribuição de cada recorte (dado por uma variável categórica) da amostra na contagem da frequência.

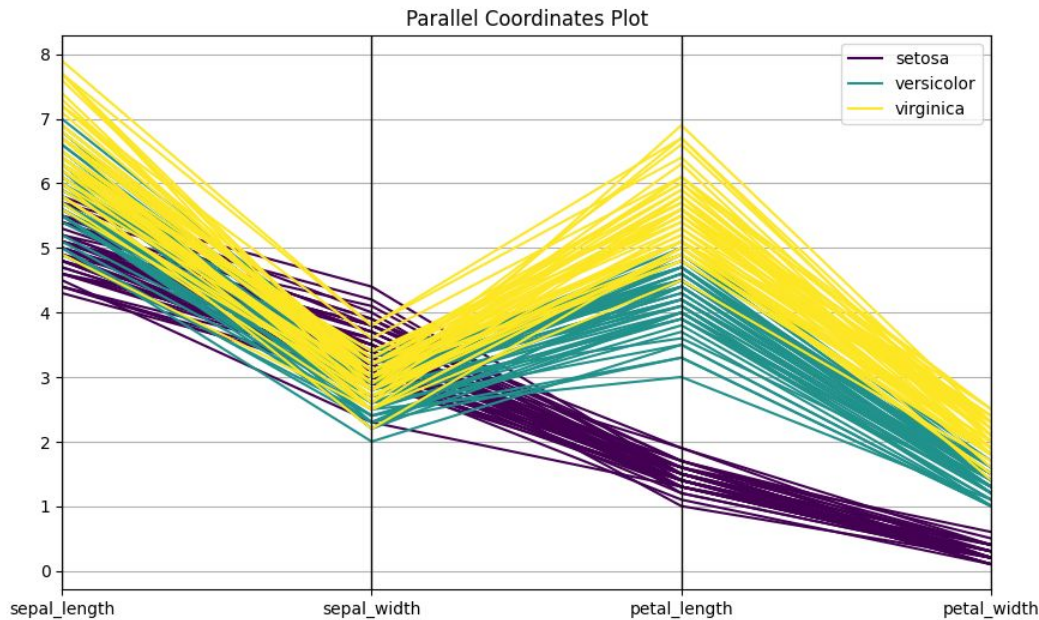
- ❑ Uma extensão do Phi para tabelas de contingência maiores que 2x2, que fornece uma medida da força da associação entre variáveis categóricas.;
- ❑ É dado pela expressão abaixo, onde  $k$  é o mínimo entre o número de linhas e colunas da tabela:



# Parallel Coordinates Plot

Exibe cada observação como uma linha que cruza vários eixos paralelos, onde cada eixo representa uma variável. Isso facilita a visualização dos padrões e a comparação de perfis entre os dados.

- ❑ Análise de dados de alta dimensionalidade.
- ❑ Comparação de padrões entre diferentes grupos.

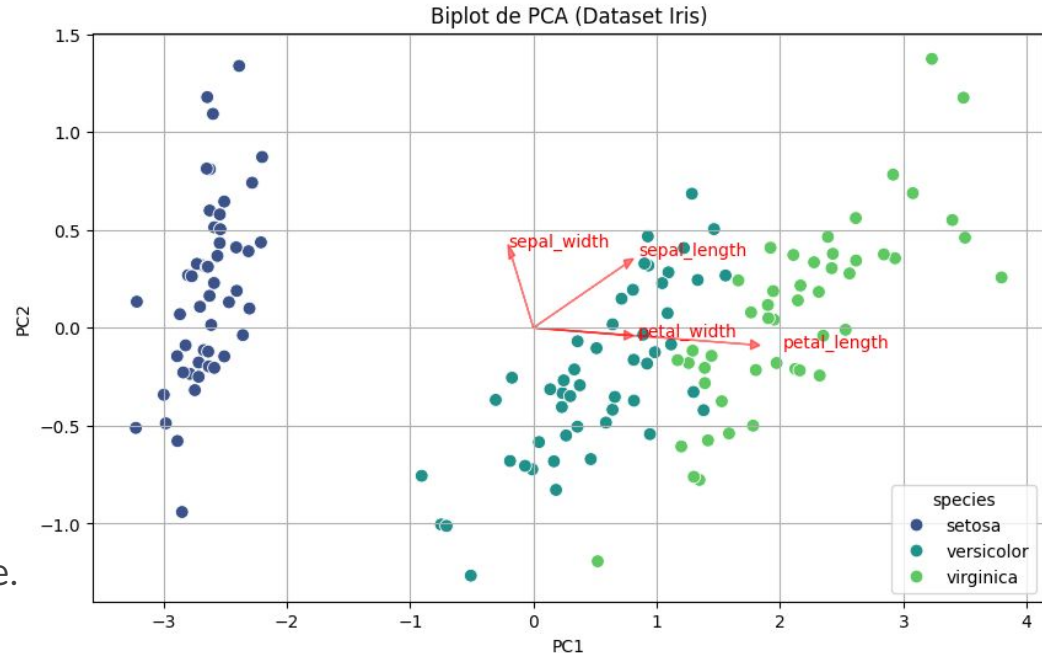




# Biplot (PCA)

Integra os resultados da Análise de Componentes Principais (PCA) para visualizar, em um espaço de menor dimensão, tanto as observações quanto as variáveis (através dos vetores de loadings).

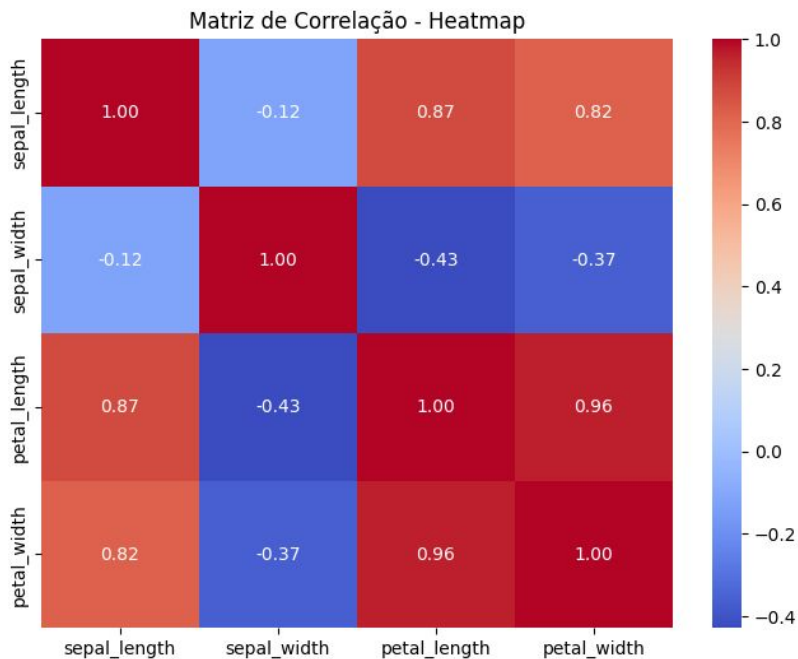
- ❑ Visualização da variabilidade dos dados e da contribuição de cada variável.
- ❑ Detecção de agrupamentos e outliers após redução de dimensionalidade.



# Heatmap

Exibe uma matriz de correlação utilizando uma escala de cores para representar a intensidade e a direção das relações entre variáveis.

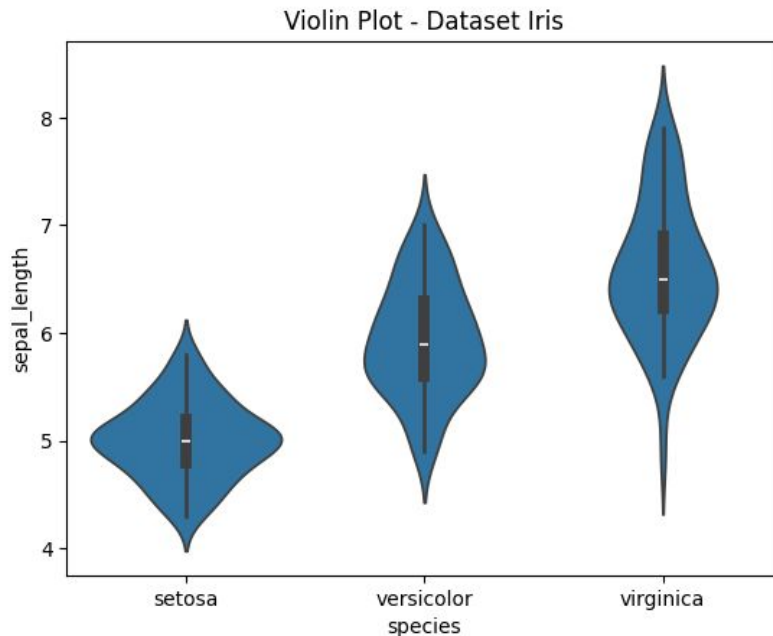
- ❑ Análise exploratória para identificar fortes associações entre variáveis.
- ❑ Auxílio na seleção de variáveis para modelagem.



# Boxplot e violin plot

Utilizado como ferramenta de estudo da estrutura da amostra, em termos de suas partições.

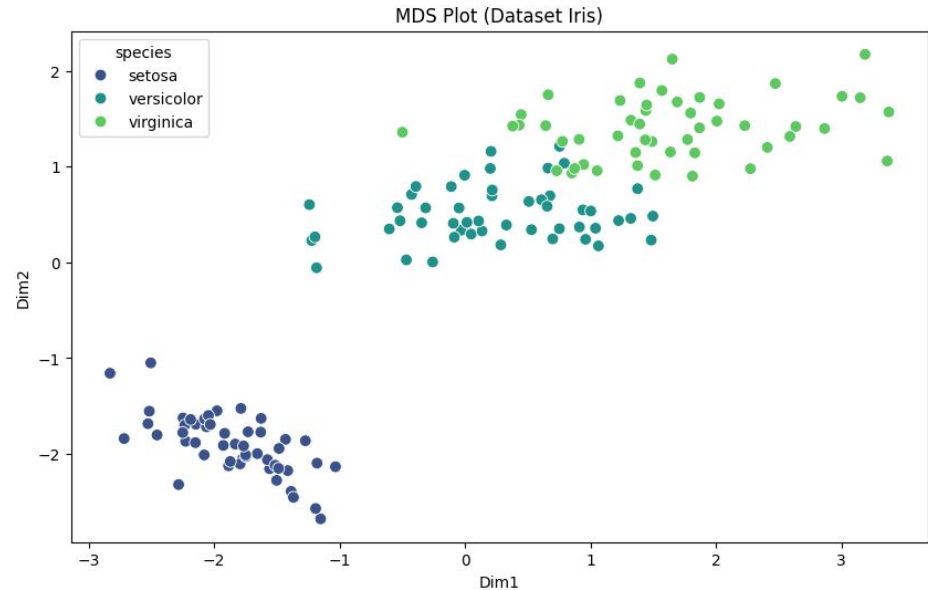
- ❑ Pode informar diferentes estruturas internas, quando subdividido por uma ou mais variáveis categóricas;
- ❑ Apresenta tanto informações de posição (quartis) quando de dispersão;



# MDS

Projeta dados de alta dimensão em um espaço 2D ou 3D, preservando a distância ou dissimilaridade entre as observações o máximo possível.

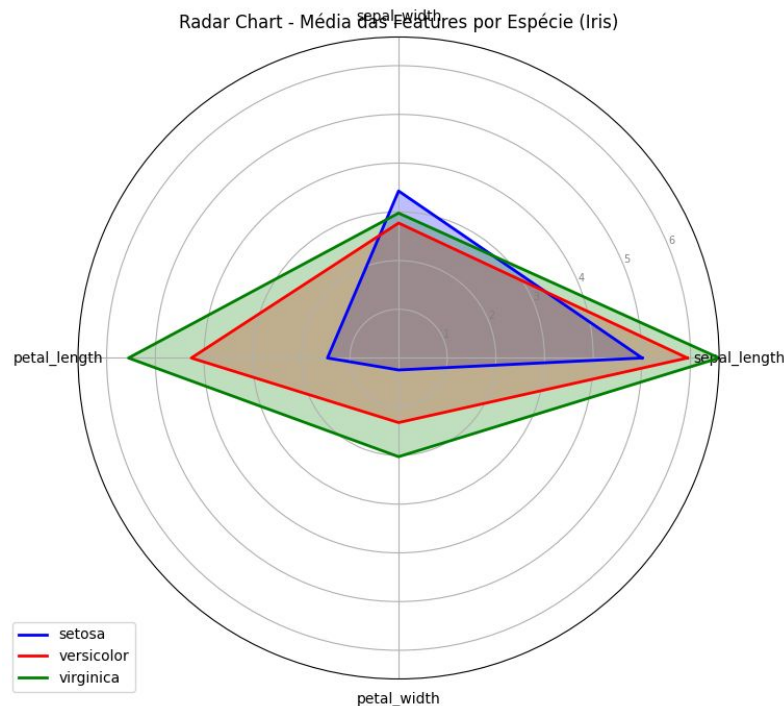
- ❑ Visualização de agrupamentos e padrões espaciais com base em medidas de similaridade ou distância.



# Radar

Exibe dados multivariados em formato circular, onde cada eixo representa uma variável. As observações são conectadas formando um polígono, facilitando a comparação de perfis..

- ❑ Comparação de atributos ou perfis entre diferentes grupos.
- ❑ Análises de desempenho ou benchmarking.

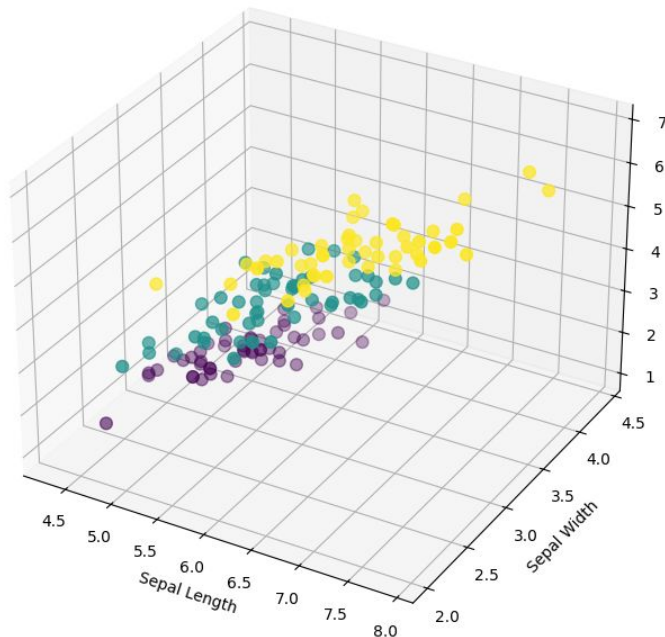


# Scatter plot 3D

Geralmente associado a séries temporais, também pode ser utilizado com ferramenta de análise de correlação e dependência entre variáveis.

3D Scatter Plot - Dataset Iris

- ❑ Análise de relações entre três variáveis principais
- ❑ Visualização exploratória quando os dados não estão muito sobrepostos.



# Prática





## Exercícios:

1. Exiba um gráfico de dispersão, do tipo KDE, da sua variável contínua, particionado por alguma variável qualitativa.
2. Exiba o gráfico de dispersão do tipo scatter plot de duas variáveis, preferencialmente contínuas, da sua amostra;
3. Exiba gráficos do tipo violin e estude a diferença de dispersão entre subclasses da sua amostra.



# Análise multivariada





# Análise multivariada

Refere-se a um conjunto de métodos estatísticos que torna possível a análise simultânea de múltiplas medidas associadas a cada elemento da amostra, com múltiplos objetivos, entre os quais:

- ❑ **Sintetizar ou simplificar** o entendimento da estrutura dos dados, no que tange às
- ❑ **Classificar ou agrupar** elementos de acordo com a multiplicidade de características registradas na amostra;
- ❑ **Inferir** a probabilidade condicional de observar certo valor em uma variável, dados os valores de outras variáveis, a partir do conhecimento das interações e padrões de associação e dependência internas à amostra;



# Análise multivariada

**Normalização:** processo de transformação dos dados que visa padronizar diferentes variáveis para melhorar sua interpretabilidade e comparação.

**Normalização Min-Max:**

$$x_{i-norm} = \frac{x_i - x_0}{x_n - x_0}$$

- ❑ Quando se deseja manter a forma original da distribuição dos dados, apenas alterando sua escala.
- ❑ Útil em algoritmos que dependem da distância entre os dados (como k-NN ou algoritmos baseados em gradiente), onde as variáveis em escalas muito diferentes podem influenciar indevidamente o modelo.
- ❑ Atenção: É sensível a outliers, pois estes definem os valores mínimo e máximo.



# Análise multivariada

Normalização por z-score:

$$z_i = \frac{x_i - \bar{x}}{S}$$

- ❑ Indicada para centralizar os dados e garantir que as diversas variáveis possam ser representadas conjuntamente, numa mesma escala;
- ❑ Muito útil em métodos estatísticos que assumem normalidade ou que são sensíveis à escala, como regressões, análise de componentes principais (PCA) e métodos baseados em distância;
- ❑ Aplicável apenas quando há forte aderência da amostra à distribuição normal;
- ❑ Sensível a outliers;



# Análise multivariada

Normalização Robusta:

$$x_{\text{robust}} = \frac{x - \text{mediana}(x)}{\text{IQR}}$$

- ❑ Indicado quando os dados contêm outliers ou apresentam uma distribuição fortemente assimétrica.
- ❑ Essa abordagem diminui a influência dos valores extremos, proporcionando uma escala mais representativa da “parte central” dos dados.



# Medidas de associação: variáveis categóricas

As medidas de associação demonstram que existe uma relação estatisticamente significativa entre duas variáveis, sem, contudo, quantificar a intensidade ou direção dessa relação.

## V de Cramer:

- ❑ Uma extensão do Phi para tabelas de contingência maiores que 2x2, que fornece uma medida da força da associação entre variáveis categóricas.;
- ❑ É dado pela expressão abaixo, onde  $k$  é o mínimo entre o número de linhas e colunas da tabela:

$$V = \sqrt{\frac{\chi^2}{n(k - 1)}}$$

# Medidas de associação: variáveis categóricas

## Coeficiente Phi:

- ❑ Utilizado para avaliar a associação entre duas variáveis dicotômicas (2x2).;
- ❑ Por ser dependente das unidades de medida das variáveis, sua interpretação direta pode ser limitada;
- ❑ Tendo como base o endereçamento na tabela de contingência ao lado, o coeficiente é dado por:

	Y=1	Y=2
X=1	A	B
X=0	C	D

$$\phi = \frac{AD - BC}{\sqrt{(A + B)(C + D)(A + C)(B + D)}}$$

# Medidas de associação: variáveis categóricas

## Teste Qui-Quadrado:

- ❑ Avalia a independência entre duas variáveis categóricas, comparando as frequências observadas com as frequências esperadas sob a hipótese nula de independência. A estatística é dada por:

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

- ❑ Onde:

- ❑  $O_{ij}$  é a frequência observada na célula correspondente à linha  $i$  e coluna  $j$ ;
- ❑  $E_{ij}$  é a frequência esperada para aquela célula

$$E_{ij} = \frac{(\sum_{i=1}^r O_i)(\sum_{j=1}^c O_j)}{n}$$

- ❑  $r$  e  $c$  são, respectivamente, o número de linhas e colunas da tabela de contingência





# Medidas de associação: variáveis categóricas

## Teste Qui-Quadrado:

- ❑ Considere o exemplo:

	Doença presente	Doença ausente	Total
Fumante	30	20	50
Não fumante	10	40	50
Total	40	60	100

- ❑ Cálculo:
  - ❑ A fórmula dá o valor de 16,6;
  - ❑ Consultando esse valor na tabela do Qui-Quadrado, para grade de liberdade 1, obtemos uma probabilidade  $p = 4,9e-5$
  - ❑ Isso significa que o  $p$ -valor é aproximadamente 0,000049, indicando uma evidência estatística extremamente significativa para rejeitar a hipótese nula de independência em um teste de qui-quadrado com 1 grau de liberdade.

# Medidas de associação: variáveis categóricas

## Coeficiente de Contingência:

- ❑ Avalia significância estatística da associação com um ajuste ao Qui-Quadrado;

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

- ❑ Seguindo o exemplo anterior, com  $\chi^2 = 16,6$  e 100 observações, teremos:

$$C = \sqrt{\frac{16,6}{16,6 + 100}} \approx \sqrt{0,143} \approx 0,378$$

- ❑ O valor de  $C$  varia entre 0 (sem associação) e um máximo que depende da dimensão da tabela (em uma tabela 2x2, esse máximo é inferior a 1). Nesse exemplo, um coeficiente de aproximadamente 0,378 sugere uma associação moderada entre o hábito de fumar e a presença de doença pulmonar.



# Covariância

**Covariância:** Mede a direção de associação entre duas variáveis.

- ❑ A covariância entre duas variáveis  $X$  e  $Y$  é dada por:

$$\text{Cov}(X, Y) = \frac{1}{n - 1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

- ❑ Um valor positivo indica que ambas variam na mesma direção;
- ❑ Por ser dependente das unidades de medida das variáveis, sua interpretação direta pode ser limitada;



# Correlação

As medidas de correlação têm como objetivo quantificar a associação entre duas ou mais variáveis, permitindo identificar tanto a direção quanto a intensidade dessa relação.

## **Coeficiente de Correlação de Pearson:**

- ❑ Avalia a força e a direção de uma relação linear entre duas variáveis quantitativas;
- ❑ Pressupõe relação de linearidade entre as variáveis;
- ❑ Definido no intervalo  $[-1, 1]$ , onde:
  - ❑ 1 indica correlação linear positiva perfeita,
  - ❑ -1 indica correlação linear negativa perfeita,
  - ❑ 0 indica ausência de correlação linear.;

$$r = \frac{\text{Cov}(X, Y)}{S_x S_y}$$

# Correlação

## Coefficiente de Correlação de Spearman:

- ❑ Medida não paramétrica que utiliza os rankings dos dados para avaliar a relação monotônica entre duas variáveis;
- ❑ É útil quando os dados não atendem aos pressupostos de normalidade ou quando a relação não é estritamente linear;

- ❑ Se  $R(x_i)$  e  $R(y_i)$  representam os postos de  $x_i$  e  $y_i$ :

$$\rho = \frac{\sum_{i=1}^n (R(x_i) - R(\bar{x}))(R(y_i) - R(\bar{y}))}{\sqrt{\sum_{i=1}^n (R(x_i) - R(\bar{x}))^2} \sqrt{\sum_{i=1}^n (R(y_i) - R(\bar{y}))^2}}$$

- ❑ Se não houver repetições na amostra, também é possível utilizar a forma simplificada

$$\rho = 1 - \frac{6 \sum_{i=1}^n (R(x_i) - R(y_i))^2}{n(n^2 - 1)}$$



# Medidas de associação: variáveis contínuas

## Coeficiente de Correlação de Kendall (Tau de Kendall):

- ❑ Outra medida não paramétrica baseada na comparação dos rankings dos dados;
- ❑ Especialmente indicada para amostras pequenas ou quando há muitos empates nos dados;

- ❑ Dado pela expressão:

$$\tau = \frac{2(n_c - n_d)}{n(n - 1)}$$

- ❑ Onde:
  - ❑  $n_c$  é o número de pares concordantes;
  - ❑  $n_d$  o número de discordantes;

# Medidas de associação: variáveis contínuas

## Coeficiente de Correlação de Kendall (Tau de Kendall):

- ❑ Em casos onde haja repetições na amostra, faz-se necessário utilizar a versão modificada do coeficiente de Kendall, dada por:

$$\tau_B = \frac{n_c - n_d}{\sqrt{(n_t - n_x)(n_t - n_y)}}$$

- ❑ Onde:

- ❑  $n_c$  é o número de pares concordantes
- ❑  $n_d$  o número de discordantes

- ❑  $n_t$  é o total de possíveis pares:  $n_t = \frac{n(n-1)}{2}$

- ❑  $n_x$  é o ajuste para empates na variável X  $n_x = \sum_i^n \frac{e_{x,i}(e_{x,i} - 1)}{2}$

- ❑  $n_y$  é o ajuste para empates na variável Y  $n_y = \sum_i^n \frac{e_{y,i}(e_{y,i} - 1)}{2}$



# Correlação

## Spearman vs Pearson:

- ❑ Aplicabilidade
  - ❑ Pearson é limitado a variáveis contínuas;
  - ❑ Spearman é mais versátil, podendo ser usado tanto para variáveis contínuas quanto para variáveis discretas e ordinais;
  - ❑ Spearman é mais robusto contra outliers, pois substitui a magnitude pelo posto.
- ❑ Interpretação dos Resultados
  - ❑ Quando os coeficientes apresentam valores semelhantes, a relação entre as variáveis é provavelmente linear;
  - ❑ Quando Spearman for maior que Pearson, indica provável relação não linear monotônica;
  - ❑ Pearson maior que Spearman: pode indicar presença de outliers;





# Determinação

Em regressão linear, é possível avaliar o poder de explicação do modelo

## **Coeficiente de Determinação ( $R^2$ ):**

- ❑ Indica a proporção da variação na variável dependente que pode ser explicada pelas variáveis independentes;
- ❑ É uma medida da qualidade do ajuste do modelo;

$$R^2 = \frac{\sum_{i=1}^n (y_i - \hat{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

# Prática





## Exercícios:

1. Selecione duas variáveis categóricas e calcule as medidas de associação.
2. Selecione duas variáveis contínuas e calcule as medidas de covariância e correlação;

# Obrigado

**Stefano Mozart**

[linkedin.com/in/stefano-mozart/](https://linkedin.com/in/stefano-mozart/)

[github.com/stefanomozart](https://github.com/stefanomozart)

