



ANÁLISES INFERENCIAIS

2º Correlação

ANÁLISES INFERENCIAIS: BIVARIADA

1. O que é correlação?

A correlação é uma medida estatística que indica a força e a direção da relação entre duas variáveis. Ela varia de -1 a 1, onde:

- **1** significa uma correlação perfeita positiva: à medida que uma variável aumenta, a outra também aumenta proporcionalmente.
- **-1** significa uma correlação perfeita negativa: quando uma variável aumenta, a outra diminui proporcionalmente.
- **0** indica que não há relação linear entre as variáveis.

No entanto, a correlação não implica causalidade, ou seja, mesmo que duas variáveis estejam correlacionadas, isso não significa que uma causa a outra.

Mostrar no quadro

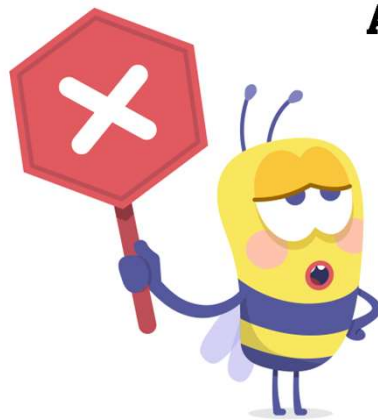
ANÁLISES INFERENCIAIS: BIVARIADA

Uma das formas de mensurar a relação entre duas variáveis são os testes de correlação

Correção de Pearson técnica para medir se duas variáveis estão relacionadas de maneira linear

Correção de Spearman é uma medida não paramétrica da dependência dos postos das variáveis

Correção de Kendall é uma medida da associação entre duas variáveis



A correlação é uma ferramenta essencial para entender relações entre variáveis, mas é importante lembrar que ela não indica causalidade. No R, temos funções simples e poderosas para calcular e visualizar essas correlações, permitindo análises rápidas e eficazes de conjuntos de dados.

ANÁLISES INFERENCIAIS: BIVARIADA

2. Tipos de correlação

- Correlação de Pearson: Mede a correlação linear entre duas variáveis numéricas contínuas.
- Correlação de Spearman: Utilizada para variáveis ordinais ou quando as variáveis não atendem aos pressupostos da correlação de Pearson (não-linearidade ou dados com outliers).
- Correlação de Kendall: Similar à correlação de Spearman, mas com um método ligeiramente diferente de calcular.




ANÁLISES INFERENCIAIS: BIVARIADA

1. Coeficiente de Correlação de Pearson

O coeficiente de Pearson é calculado como a razão entre a covariância das variáveis e o produto de seus desvios padrão:

$$r = \frac{cov(X, Y)}{\sigma_X \sigma_Y}$$

Onde:

- $cov(X, Y)$ é a covariância entre X e Y ,
 - σ_X e σ_Y são os desvios padrão de X e Y .
- 

ANÁLISES INFERENCIAIS: BIVARIADA

Exemplo

```
#Correlação de Pearson####  
# Aqui, simulamos um conjunto de dados com variáveis  
# relacionadas ao nível de educação e à renda anual,  
# temas centrais nas análises socioeconômicas.  
  
#Criar uma base de dados  
set.seed(123)  
educ ← sample(8:20, 100, replace = TRUE)  
renda ← educ * 5000 + rnorm(100, mean = 30000, sd = 10000)  
  
dados_educ_renda ← data.frame(educ, renda)  
  
# Visualizar as primeiras linhas da base de dados  
head(dados_educ_renda)
```

	educ	renda
1	10	83035.29
2	10	84482.10
3	17	115530.04
4	9	84222.67
5	13	115500.85
6	18	115089.69
7	12	66908.31
8	11	95057.39
9	13	87907.99
10	16	103119.91
11	17	125255.71
12	18	117152.27
13	12	77792.82
14	10	81813.03
15	18	118611.09
16	16	110057.64

ANÁLISES INFERENCIAIS: BIVARIADA

```
# Calcular a correlação de Pearson entre educação e renda
```

```
cor(dados_educ_renda$educ, dados_educ_renda$renda)
```

```
#OU
```

```
cor(dados_educ_renda$educ, dados_educ_renda$renda, method = "pearson")
```

```
[1] 0.8450715
```

```
#Apresentação gráfica####
```

```
# Gráfico de Dispersão (Scatter Plot) com Linha de Tendência
```

```
# O gráfico de dispersão é uma das maneiras mais comuns de
```

```
# visualizar a correlação entre duas variáveis.
```

```
# Podemos adicionar uma linha de regressão para destacar a
```

```
# relação entre educação e renda.
```

```
# Carregar pacotes necessários
```

```
library(ggplot2)
```

```
# Criar o gráfico de dispersão com linha de tendência
```

```
ggplot(dados_educ_renda, aes(x = educ, y = renda)) +
```

```
  geom_point(color = "blue") + # Adiciona pontos
```

```
  geom_smooth(method = "lm", color = "red", se = FALSE) +
```

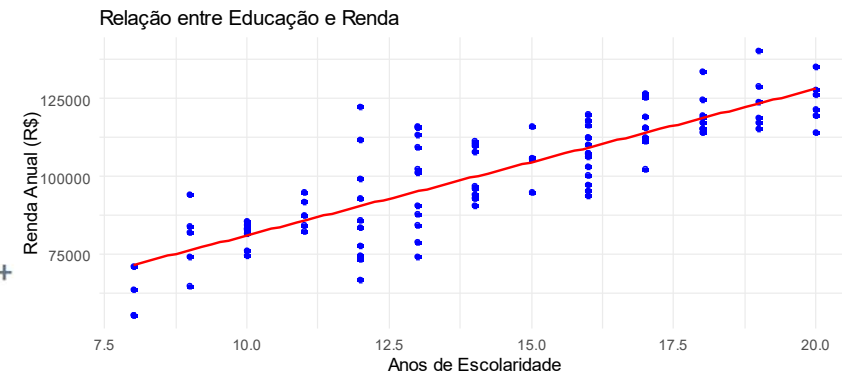
```
  # Adiciona linha de regressão linear
```

```
  labs(title = "Relação entre Educação e Renda",
```

```
        x = "Anos de Escolaridade",
```

```
        y = "Renda Anual (R$)" ) +
```

```
  theme_minimal()
```



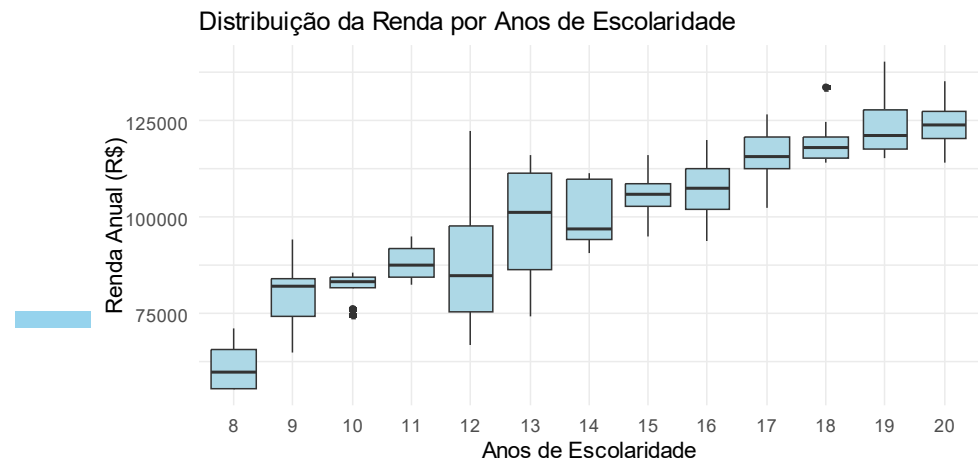
ANÁLISES INFERENCIAIS: BIVARIADA

Esse exemplo simula a relação entre o nível educacional
(anos de escolaridade) e a renda anual. Em pesquisas sociais,
esperamos encontrar uma correlação positiva, pois mais anos de educação
geralmente estão associados a maiores rendimentos.

#Boxplot: Mostra a variação da renda em cada nível educacional,
#incluindo valores atípicos.

Gráfico de Boxplot

```
ggplot(dados_educ_renda, aes(x = factor(educ), y = renda)) +  
  geom_boxplot(fill = "lightblue") +  
  labs(title = "Distribuição da Renda por Anos de Escolaridade",  
        x = "Anos de Escolaridade",  
        y = "Renda Anual (R$)") +  
  theme_minimal()
```



ANÁLISES INFERENCIAIS: BIVARIADA

4. Correlação de Spearman

A correlação de Spearman é uma medida não-paramétrica que avalia a força da associação entre duas variáveis classificando-as e calculando a correlação de Pearson entre esses postos.

A fórmula da correlação de Spearman é:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

Onde:

- d_i é a diferença entre os postos das duas variáveis para cada observação,
- n é o número de observações.

ANÁLISES INFERENCIAIS: BIVARIADA

Exemplo

```
# Correlação de Spearman####  
# Primeiro, vamos carregar o conjunto de dados  
# e calcular a correlação de Pearson entre duas  
# variáveis: mpg (milhas por galão) e  
# wt (peso do carro).
```

```
# Carregar conjunto de dados  
data(mtcars)
```

```
# Visualizar as primeiras linhas dos dados  
head(mtcars)
```


	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225	105	2.76	3.460	20.22	1	0	3	1

```
- |
```

ANÁLISES INFERENCIAIS: BIVARIADA

Exemplo

Essas variáveis vêm do conjunto de dados mtcars, que contém especificações técnicas de diferentes modelos de carros. CODEBOOK

- **mpg:** *Miles per gallon* (milhas por galão) — Representa a eficiência de combustível do carro, ou seja, quantas milhas ele pode percorrer com um galão de combustível. Quanto maior o valor, mais eficiente o carro em termos de consumo de combustível.
 - **cyl:** *Cylinders* (cilindros) — Refere-se ao número de cilindros no motor do carro. Carros com mais cilindros tendem a ter motores mais potentes, mas também consomem mais combustível.
 - **disp:** *Displacement* (deslocamento) — O volume total deslocado por todos os cilindros do motor, medido em polegadas cúbicas. Está relacionado à capacidade do motor e é uma medida de seu tamanho.
 - **hp:** *Horsepower* (potência) — A potência do motor, medida em cavalos de potência (HP). Uma medida de quão rápido e potente o carro pode ser.
 - **drat:** *Rear axle ratio* (relação do eixo traseiro) — Refere-se à relação entre as rotações do eixo de saída do motor e as rotações das rodas traseiras. Influencia a aceleração e a economia de combustível.
 - **wt:** *Weight* (peso) — O peso do carro em milhares de libras. Carros mais pesados tendem a ser menos eficientes em termos de consumo de combustível.
 - **qsec:** *1/4 mile time* (tempo de 1/4 de milha) — O tempo que o carro leva para percorrer um quarto de milha, medido em segundos. Quanto menor o valor, mais rápido o carro é.
 - **vs:** *Engine shape* (forma do motor) — Um indicador binário (0 ou 1) que refere-se ao tipo de motor. 0 representa um motor em V, e 1 representa um motor em linha.
 - **am:** *Transmission* (transmissão) — Tipo de transmissão do carro. 0 representa transmissão automática, e 1 representa transmissão manual.
 - **gear:** *Gears* (marchas) — Número de marchas na transmissão do carro. 
 - **carb:** *Carburetors* (carburadores) — Número de carburadores no carro. Um carburador mistura ar e combustível para alimentar o motor.
- Essas variáveis são usadas frequentemente em análises para estudar a eficiência, desempenho e outras características dos carros, muitas vezes em contextos de consumo, desempenho ambiental ou inovação automotiva.

ANÁLISES INFERENCIAIS: BIVARIADA

```
# Calcular correlação de Pearson
cor(mtcars$mpg, mtcars$wt)

# Neste caso, estamos calculando a correlação
# entre o consumo de combustível e o peso do carro.
# Espera-se uma correlação negativa, ou seja,
# quanto maior o peso, menor a
# eficiência do carro em termos de consumo.

# Se suspeitarmos que a relação não é linear,
# podemos utilizar a correlação de Spearman,
# que não exige linearidade entre as variáveis.

# Calcular correlação de Spearman
cor(mtcars$mpg, mtcars$wt, method = "spearman")
> cor(mtcars$mpg, mtcars$wt, method = "spearman")
[1] -0.886422
> |
```

ANÁLISES INFERENCIAIS: BIVARIADA

```
# Matriz de correlação
# Podemos calcular a correlação entre várias
# variáveis de uma vez, gerando uma matriz de
# correlação.
|
# Matriz de correlação entre algumas variáveis
#do conjunto de dados
cor(mtcars[, c("mpg", "wt", "hp", "qsec")])

# Para facilitar a interpretação, é possível visualizar a correlação
# utilizando um gráfico de calor.

# Instalar e carregar a biblioteca necessária para visualização
install.packages("corrplot")
library(corrplot)

# Criar a matriz de correlação
matriz_cor ← cor(mtcars)
```

```
> cor(mtcars[, c("mpg", "wt", "hp", "qsec")])
```

	mpg	wt	hp	qsec
mpg	1.0000000	-0.8676594	-0.7761684	0.4186840
wt	-0.8676594	1.0000000	0.6587479	-0.1747159
hp	-0.7761684	0.6587479	1.0000000	-0.7082234
qsec	0.4186840	-0.1747159	-0.7082234	1.0000000

```
└─ |
```


ANÁLISES INFERENCIAIS: BIVARIADA

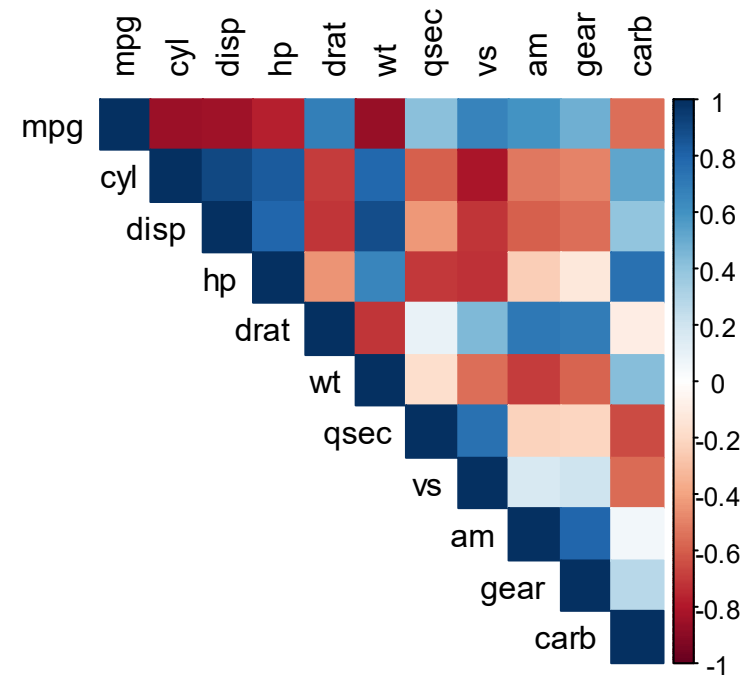
```
# Para facilitar a interpretação, é possível visualizar a correlação
# utilizando um gráfico de calor.
```

```
# Instalar e carregar a biblioteca necessária para visualização
install.packages("corrplot")
library(corrplot)
```

```
# Criar a matriz de correlação
matriz_cor <- cor(mtcars)
```

```
# Plotar o gráfico de correlação
corrplot(matriz_cor, method = "color",
         type = "upper", tl.col = "black")
```

```
# Esse gráfico mostra a força e a direção
# da correlação entre as variáveis.
# As cores indicam a intensidade da correlação,
# com azul representando correlação positiva e
# vermelho correlação negativa.
```



ANÁLISES INFERENCIAIS: BIVARIADA

5. Correlação de Kendall (Tau-b)

A correlação de Kendall mede a associação entre duas variáveis ao comparar o número de pares concordantes e discordantes. É uma outra medida não-paramétrica.

A fórmula é:

$$\tau = \frac{n_c - n_d}{\sqrt{(n_0 - n_t)(n_0 - n_u)}}$$

Onde:

- n_c é o número de pares concordantes,
- n_d é o número de pares discordantes,
- n_0 é o número total de pares possíveis,
- n_t e n_u são correções para empates nas variáveis.

Essas fórmulas são especialmente úteis quando os dados não seguem uma relação linear ou apresentam valores discrepantes (outliers). A correlação de Spearman e a de Kendall são mais robustas para esses cenários.

ANÁLISES INFERENCIAIS: BIVARIADA

Exemplo

```
#Correlação de Kendall####  
#  
# (Confiança Social e Participação Religiosa)  
# Vamos agora criar um exemplo onde avaliamos a correlação  
# entre a confiança nas instituições sociais e a frequência à igreja,  
# medidos em escalas ordinais.
```

```
# Criar uma base de dados fictícia  
set.seed(456)  
dados_conf_religiao ← data.frame(  
  confiança_instituições = sample(1:5, 100, replace = TRUE),  
  # 1 = Nenhuma confiança, 5 = Muita confiança  
  freq_igreja = sample(1:5, 100, replace = TRUE)  
  # 1 = Nunca, 5 = Todo fim de semana  
)
```

```
# Visualizar as primeiras linhas da base de dados  
head(dados_conf_religiao)
```

	confiança_instituições	freq_igreja
1	5	5
2	5	5
3	3	2
4	5	1
5	4	2
6	3	4

> |

ANÁLISES INFERENCIAIS: BIVARIADA

```
# Calcular a correlação de Kendall entre confiança nas
# instituições e frequência à igreja
cor(dados_conf_religiao$confiança_instituições,
    dados_conf_religiao$freq_igreja, method = "kendall")
> cor(dados_conf_religiao$confiança_instituições,
+     dados_conf_religiao$freq_igreja, method = "kendall")
[1] -0.05724389
```

Neste exemplo fictício, estamos analisando a correlação
entre a confiança nas instituições sociais e a frequência
à igreja. Esses dados são frequentemente utilizados para
investigar a relação
entre crenças religiosas e confiança na sociedade.



**OBRIGADO E ASSISTA A
PRÓXIMA AULA**

NAIARA ALCANTARA
NAYARASANDY@UFPA.BR