

Estatística Descritiva e Análise Exploratória dos Dados

Introdução às Probabilidades e Estatística

1. Conceitos Fundamentais de Estatística
2. Tabela de Frequência e Representações Gráficas
3. Medidas Descritivas
4. Análise Bivariada

1. Conceitos Fundamentais de Estatística

- A **Estatística** é um conjunto de métodos apropriados para **recolher**, **classificar**, **organizar**, **analisar** e **interpretar** conjuntos de dados, tendo em vista o conhecimento de determinado fenómeno e a possibilidade de, a partir desse conhecimento, inferir possíveis novos resultados.
- É amplamente utilizada em diversas áreas como ciência, economia, medicina, engenharia, psicologia e ciências sociais.

Conceitos Fundamentais de Estatística (cont.)

- As aplicações da estatística podem ser divididas em duas grande áreas:
- **Estatística descritiva**: conjunto de técnicas que visam a exploração dos dados para identificar padrões, resumir informações e apresentá-las de forma conveniente para o utilizador.
- **Inferência Estatística**: conjunto de técnicas que visam a utilização de um conjunto representativo de dados para fazer estimativas, tomar decisões, fazer previsões ou outras generalizações sobre um conjunto maior de dados.

Conceitos Fundamentais de Estatística (cont.)

Caracterização de um conjunto de espectadores de um programa de TV.

Espectador	Género	Idade	Habilitações Escolares	Salário Mensal	Residência
A	M	23	12º ano	900	Lisboa
B	M	19	9º ano	580	Lisboa
C	M	36	Licenciatura	3120	Porto
D	F	42	Licenciatura	2450	Leiria
E	F	21	9º ano	940	Lisboa
F	M	61	6º ano	580	Lisboa
G	F	28	12º ano	1230	Faro
H	F	52	9º ano	630	Sintra
I	M	32	12º ano	1200	Porto
J	M	46	12º ano	1200	Gaia
K	M	28	Licenciatura	2050	Braga
L	F	29	12º ano	1050	Viseu

Tabela: Dados dos espectadores: Género: M = masculino, F = feminino; Idade: em anos; Salário Mensal: em euros.

- Que possíveis padrões de audiência terá interesse procurar?
- Esses padrões são imediatamente identificáveis?
- Que representações dos dados poderiam destacar os possíveis padrões?
- O conjunto de espectadores inquirido será representativo da totalidade dos espectadores do programa?
- Poderiam as conclusões da análise deste conjunto de espectadores ser extrapoladas para a totalidade dos espectadores?

- A **análise exploratória dos dados (AED)** é uma etapa fundamental na estatística. A sua finalidade é
 - examinar os dados previamente à aplicação de qualquer técnica estatística.
 - deste modo o analista consegue compreender a estrutura dos dados e identificar padrões, anomalias, relações e resumir as suas principais características, frequentemente com a ajuda de métodos gráficos.
- Após a coleta e a inserção de dados numa base de dados apropriada, o próximo passo é a análise descritiva. Esta etapa é crucial, pois permite ao investigador familiarizar-se com os dados, organizá-los e sintetizá-los, de modo a extrair as informações necessárias para responder às questões em estudo.

- A análise exploratória dos dados (AED) foi popularizada por John Tukey, um renomado estatístico americano, no final da década de 1960. Tukey introduziu o termo “Exploratory Data Analysis” e enfatizou a importância de explorar os dados antes de aplicar modelos estatísticos formais. Ele defendia que, através da AED, os analistas poderiam desenvolver uma intuição sobre os dados e suas peculiaridades, permitindo uma análise mais robusta e informada.



Figura: Foto retirada de https://en.wikipedia.org/wiki/John_Tukey

Conceitos Fundamentais de Estatística (cont.)

Elementos fundamentais da Estatística

- **População ou universo:** Conjunto de unidades individuais, com uma ou mais características de interesse em comum, que se pretende(m) estudar. (ex: todos os alunos da FCUL).
- **Unidade estatística:** elemento da população (ex: estudante da FCUL).
- **Variável:** característica de interesse em estudo (ex: X - altura dos estudantes da FCUL).
- **Dados:** valores observados da variável (ex: x - altura observada de um estudante da FCUL).
 - **Dado bruto** $\rightarrow x_i$ (dado antes de qualquer tratamento)
 - **Dado ordenado** $\rightarrow x_{(i)}$ (dado ordenado usualmente em ordem crescente)
- **Amostra:** subconjunto da população observada ($(x_1, \dots, x_n) = (1.50, \dots, 1.80)$)
- **Dimensão da amostra:** número de unidades estatísticas da amostra.

Conceitos Fundamentais de Estatística (cont.)

Caracterização de um conjunto de espectadores de um programa de TV.

Espectador	Género	Idade	Habilitações Escolares	Salário Mensal	Residência
A	M	23	12 ^º ano	900	Lisboa
B	M	19	9 ^º ano	580	Lisboa
C	M	36	Licenciatura	3120	Porto
D	F	42	Licenciatura	2450	Leiria
E	F	21	9 ^º ano	940	Lisboa
F	M	61	6 ^º ano	580	Lisboa
G	F	28	12 ^º ano	1230	Faro
H	F	52	9 ^º ano	630	Sintra
I	M	32	12 ^º ano	1200	Porto
J	M	46	12 ^º ano	1200	Gaia
K	M	28	Licenciatura	2050	Braga
L	F	29	12 ^º ano	1050	Viseu

Tabela: Dados dos espectadores: Género: M = masculino, F = feminino; Idade: em anos; Salário Mensal: em euros.

- **Unidades estatísticas:** Os espectadores do programa.
- **Variáveis estatísticas:** Género, Idade, Habilitações Escolares, Salário Mensal, Residência.
- **Dados:** Os valores na tabela.
- **Dimensão da amostra:** 12 (espectadores inquiridos).

Tipos de Variáveis

Variável: Qualquer característica associada a uma população.

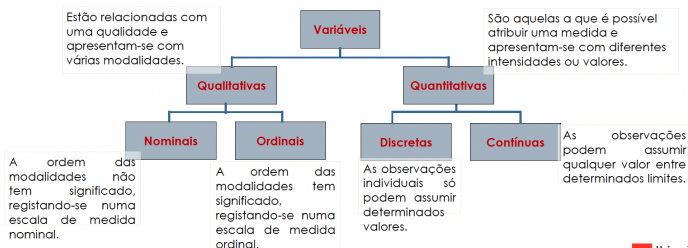
Classificação da variável:

- **Qualitativa:** é uma variável que descreve ou categoriza atributos, qualidades, ou características não numéricas de um objeto ou evento. Por exemplo,
 - a cor dos olhos,
 - o nome de um partido político,
 - o tipo de transporte usado para ir pro trabalho.
- **Quantitativa:** é uma variável que representa quantidades mensuráveis. As variáveis quantitativas fornecem valores numéricos que permitem a realização de cálculos matemáticos e análises estatísticas mais detalhadas. Os valores que essas variáveis podem assumir podem ser ordenados de forma lógica e natural. Por exemplo,
 - tamanho dos sapatos,
 - preço das casas,
 - número de semestres cursados,
 - peso de uma pessoa.

Tipos de Variáveis (cont.)

- **Variáveis discretas** são variáveis que só podem assumir um número finito ou infinito numerável de valores. Todas as variáveis qualitativas são discretas, como a cor dos olhos ou a região de um país. Mas também as variáveis quantitativas podem ser discretas: o tamanho dos sapatos ou o número de semestres estudados seriam discretos porque o número de valores que estas variáveis podem assumir é limitado.
- **Variáveis contínuas** são aquelas que podem assumir qualquer valor em um intervalo contínuo de valores. Exemplos são o tempo que leva para chegar à universidade, o comprimento de um cavalo e a distância entre dois planetas. Às vezes diz-se (de maneira informal) que variáveis contínuas são variáveis que são “medidas em vez de contadas”.

Tipos de Variáveis (cont.)



As considerações anteriores indicam que diferentes variáveis contêm diferentes quantidades de informação. Uma classificação útil destas considerações é dada pelo conceito de escala de uma variável.

Principais escalas de medida:

- 1 Escala nominal
- 2 Escala ordinal
- 3 Escala intervalar
- 4 Escala absoluta ou de razão

Tipos de escala (cont.)

Escala nominal:

- É a mais básica e simplesmente classifica dados em categorias distintas e não ordenadas. Estas categorias, embora possam ser representadas por números, não são numéricas.
- Não há ordem ou hierarquia entre as categorias.
- Não se pode realizar operações matemáticas com esses dados.

Exemplos:

- **Género:** Masculino, Feminino.
- **Cores de Olhos:** Castanho, Azul, Verde.
- **Tipos de Rochas:** Ígnea, Sedimentar, Metamórfica.
- **Tipo de construção:** Edifício, Ponte, Barragem, Reservatório, Outros.
- **Tipo de fonte primária de energia utilizada:** Eólica, Carvão, Gás, Fuel, Fios de água, Albufeiras, Saldo Importador.

Tipos de escala (cont.)

Escala ordinal:

- A escala ordinal classifica dados em categorias, mas é possível estabelecer uma ordem ou hierarquia.
- As diferenças entre essas categorias não são mensuráveis.

Exemplos:

- **Classificações de Satisfação:** Insatisfeito, Neutro, Satisfeito.
- **Grau de Escolaridade:** Ensino Básico, Ensino Secundário, Ensino Superior.
- **Níveis de Risco:** Baixo, Médio, Alto.
- **Qualidade das águas balneares:** Má, Aceitável, Boa, Excelente.
- **Dureza das rochas:** Nível 1 (Rochas muito macias), Nível 2, Nível 3, Nível 4, Nível 5, Nível 6, Nível 7, Nível 8, Nível 9, Nível 10 (Rochas muito duras)

Tipos de escala (cont.)

Exemplo variável em escala ordinal em Geologia

A **dureza das rochas** pode ser classificada em níveis que indicam a resistência de uma rocha a ser riscada. A escala de Mohs é frequentemente utilizada para descrever essa variável ordinal.

Classificação da Escala de Mohs:

- | | |
|---|---|
| 1 Talco - Nível 1 (Rochas muito macias) | 6 Ortoclásio - Nível 6 |
| 2 Gipsita - Nível 2 | 7 Quartzo - Nível 7 |
| 3 Calcita - Nível 3 | 8 Topázio - Nível 8 |
| 4 Fluorita - Nível 4 | 9 Corindo - Nível 9 |
| 5 Apatita - Nível 5 | 10 Diamante - Nível 10 (Rochas muito duras) |

- **Intervalos Não Uniformes:** Embora as rochas estejam ordenadas, a diferença de dureza entre níveis consecutivos na escala de Mohs não é uniforme. Por exemplo, a diferença entre a dureza do talco e da gipsita não é a mesma que entre o corindo e o diamante.

Tipos de escala (cont.)

Escala intervalar:

- Mede dados em intervalos iguais, onde a diferença entre os valores é significativa. Não possui um ponto zero verdadeiro (ausência de zero absoluto).
- Permite operações de adição e subtração. As diferenças entre os valores são uniformes.
- A razão entre valores não tem sentido.

Exemplos:

- **Escala de temperatura:** A diferença entre 20°C e 30°C é a mesma que entre 30°C e 40°C , mas 0°C não significa ausência de temperatura.
- **Pontuação em Testes de Inteligência (QI):** As diferenças entre pontuações são significativas, mas não existe um ponto de origem que indique ausência de inteligência. A diferença de 10 pontos de QI (por exemplo, de 100 a 110) é significativa e indica o mesmo intervalo que entre 110 e 120.

Tipos de escala (cont.)

Escala absoluta ou de razão:

- É semelhante à escala intervalar, mas possui um ponto zero verdadeiro, que indica a ausência completa da quantidade medida.
- Permite todas as operações aritméticas: adição, subtração, multiplicação e divisão.

Exemplos:

- **Altura** (em m)
- **Peso** (em Kg)
- **Idade** (em anos)
- **Distância** (em Km)
- **Densidade das rochas** (em g/cm^3)
- **Profundidade de depósitos geológicos** (em m)
- **Espessura de camadas sedimentares** (em m ou cm)

2. Tabela de Frequências e Representações Gráficas

Organização de dados

Tabela de distribuição de frequências

1ª coluna: Categorias c_i , os valores x_i , ou as classes $[c_i; c_{i+1}[$ que a variável pode assumir.

2ª coluna: Frequência Absoluta, n_i , é a contagem ou número de vezes que cada categoria ou valor da variável foi observado. Para uma amostra de tamanho n , $\sum_{i=1}^k n_i = n$ (o somatório das frequências absolutas é igual ao total de observações).

3ª coluna: Frequência relativa, f_i , é a proporção com que cada categoria ou valor da variável foi observado: $f_i = \frac{n_i}{n}$.

$\sum_{i=1}^k f_i = 1$ (o somatório das frequências relativas é igual à unidade)

Colunas seguintes (apenas para variáveis quantitativas ou qualitativas ordinais):
Frequência Acumulada, N_i ou F_i .

Organização de dados (cont.)

Tabela de distribuição de frequências

Variável $c_j, x_j, [c_j, c_{j+1}[$	Freq. absoluta n_j	Freq. relativa f_j	Freq. absoluta acumulada N_j	Freq. relativa acumulada F_j
x_1	n_1	f_1	$N_1 = n_1$	$F_1 = f_1 = N_1/n$
x_2	n_2	f_2	$N_2 = n_1 + n_2$	$F_2 = f_1 + f_2 = N_2/n$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k	n_k	f_k	$N_k = \sum_{i=1}^k n_i = n$	$F_k = \sum_{i=1}^k f_i = N_k/n = 1$
Total	$\sum_{i=1}^k n_i = n$	$\sum_{i=1}^k f_i = 1$	-	-

Notação:

- X - Variável em estudo: representa determinada característica de uma população ou de uma amostra
- x_i - i -ésima observação da amostra $i = 1, \dots, n$
- $x_{(i)}$ - i -ésima observação da amostra ordenada $i = 1, \dots, n$
- n - número de elementos da amostra (dimensão da amostra)

Representações gráficas

- Variáveis Qualitativas (nominais/ordinais):
 - Gráfico circular/Gráfico de pizza
 - Gráfico de barras
 - Boxplot/Caixa de bigodes (ordinais)
- Variáveis Quantitativas (discretas/contínuas):
 - Gráfico de barras/Histograma
 - Boxplot/Caixa de bigodes

Organização de dados (cont.)

Exemplo: Estudo sobre a satisfação de um conjunto de pessoas com acesso à Internet na residência.

Estudante	Género	Idade	Salário	Agregado Familiar	Acesso à Internet	Satisfação c/ Acesso
A	F	19	1200	2	ADSL	4
B	F	20	1400	3	4G	5
C	M	21	2000	3	Fibra	3
D	M	26	1800	5	ADSL	2
E	M	30	1900	3	4G	4
F	M	37	900	3	ADSL	4
G	F	20	1300	3	ADSL	3
H	M	18	750	3	ADSL	4
I	M	25	3000	2	Fibra	5
J	F	22	2500	4	Fibra	5
K	F	20	1500	3	ADSL	3
L	M	33	2300	3	ADSL	3

Tabela: Salário em euros; Agreg. familiar: número de elementos; Satisfação com o Acesso: 1 (mín.) a 5 (máx.)

Classificação das variáveis do exemplo anterior em qualitativa nominal/ordinal e quantitativa discreta/contínua.

Variáveis qualitativas:

- Género - nominal
- Acesso à Internet - nominal
- Satisfação com o Acesso - ordinal

Variáveis quantitativas:

- Idade - contínua
- Salário - contínua
- Agregado familiar - discreta

Organização de dados (cont.)

Dados qualitativos nominais: podem ser agrupados numa **tabela de frequências**, onde cada categoria distinta que a variável pode assumir é listada junto com a contagem do número de ocorrências em cada categoria.

Tabela de frequências da variável qualitativa nominal “Género”

Género	Número de pessoas (n_i)	% de pessoas (f_i)
F	5	$5/12 \equiv 41.67\%$
M	7	$7/12 \equiv 58.33\%$
Total	12	$1 \equiv 100\%$

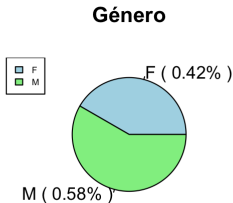
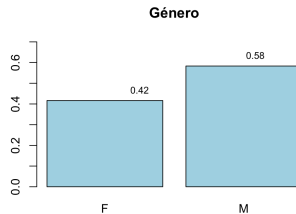
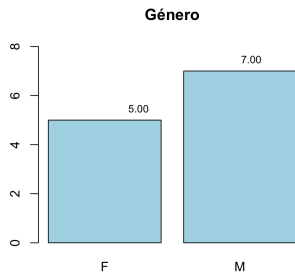
Tabela de frequências da variável qualitativa nominal “Acesso à Internet”

Acesso à Internet	Número de pessoas (n_i)	% de pessoas (f_i)
ADSL	7	$7/12 \equiv 58.33\%$
Fibra	3	$3/12 \equiv 25\%$
4G	2	$2/12 \equiv 16.67\%$
Total	12	$1 \equiv 100\%$

Dados qualitativos nominais: A representação gráfica pode ser feita com base num gráfico de barras ou num gráfico circular.

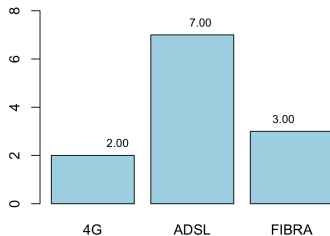
- **Gráfico de barras:** Conjunto de barras verticais ou horizontais. Cada barra representa uma categoria, e a altura da barra mostra a frequência absoluta ou relativa dessa categoria. A largura das barras não tem significado.
- **Gráfico circular (ou gráfico de pizza):** Exibe as proporções ou percentagens de diferentes categorias de dados em relação a um todo. Cada categoria é representada como uma “fatia” do círculo, e o tamanho de cada fatia é proporcional à sua contribuição para o total.

Organização de dados (cont.)

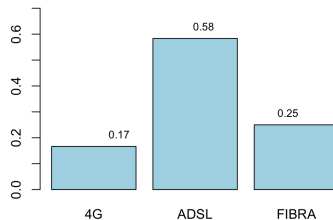


Organização de dados (cont.)

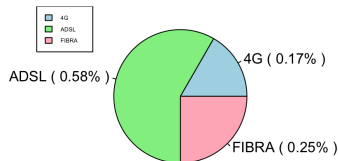
Acesso à Internet



Acesso à Internet



Acesso à Internet



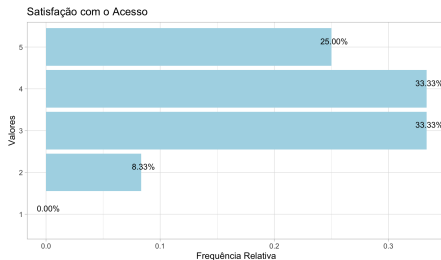
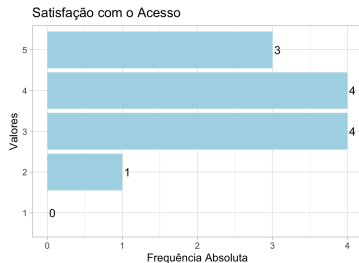
Organização de dados (cont.)

Dados qualitativos ordinais: podem ser agrupados numa **tabela de frequências** e sua representação gráfica pode ser feita com base num **gráfico de barras** ou **gráfico circular**.

Tabela de frequências da variável qualitativa ordinal “Satisfação com o Acesso”

Valores (x_i)	Freq. Abs (n_i)	Freq. Rel (f_i)	Freq. Abs Acum (N_i)	Freq. Rel Acum (F_i)
1	0	0	0	0
2	1	0.0833	1	0.0833
3	4	0.3333	5	0.4167
4	4	0.3333	9	0.7500
5	3	0.2500	12	1
Total	12	1	-	-

Organização de dados (cont.)



Organização de dados (cont.)

Dados quantitativos discretos: Quando a variável assume um número reduzido de valores observados, o tratamento desses dados, tanto no que diz respeito ao agrupamento como à representação gráfica, é semelhante ao tratamento dado a dados qualitativos ordinais.

Tabela de frequências da variável quantitativa discreta “Agregado familiar”

Agreg familiar (x_i)	Freq Abs (n_i)	Freq Rel (f_i)	Freq Abs Acum (N_i)	Freq Rel Acum (F_i)
2	2	0.1667	2	0.1667
3	8	0.6667	10	0.8334
4	1	0.0833	11	0.9167
5	1	0.0833	12	1
Total	12	1	-	-

Organização de dados (cont.)

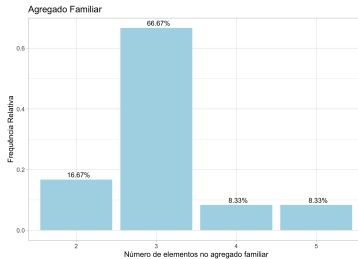
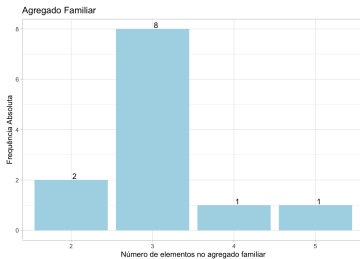
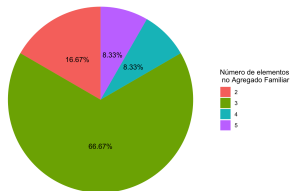


Gráfico Circular - Agregado Familiar



Organização de dados (cont.)

Dados quantitativos contínuos: Quando a variável assume um número elevado de valores observados diferentes, os dados quantitativos devem ser classificados em classes ou intervalos. A representação gráfica é feita num **histograma**.

Agrupamento em classes:

- 1 Ordenar os valores amostrais: $x_i \rightarrow x_{(i)}$
- 2 Determinar a amplitude amostral: $A = \max(x_i) - \min(x_i)$
- 3 Definição do número de classes, k :
 - Fórmula de Sturges

$$k = \text{Int}[1 + 3.321928 \cdot \log_{10}(n)]$$

- Menor inteiro k tal que $2^k \geq n$
- Regra empírica

$$n \leq 25 \quad \text{vem} \quad k = 5$$

$$n > 25 \quad \text{vem} \quad k \simeq \sqrt{n}$$

- ④ Amplitude das classes: $h = \frac{A}{k}$ (valor arredondado por excesso)
- ⑤ Metodologia para a construção das classes:

- Determinar $\varepsilon = kh - A$ e fazer $c_1 = x_{(1)} - \frac{\varepsilon}{2}$;
- Classes

$$C_1 =]c_1, c_1 + h] =]c_1, c_2]$$

$$C_2 =]c_2, c_2 + h] =]c_2, c_3]$$

...

$$C_k =]c_k, c_k + h] =]c_k, c_{k+1}] \text{ onde } c_{k+1} = x_{(n)} + \frac{\varepsilon}{2}$$

Organização de dados (cont.)

Dimensão da amostra: $n = 12$ pessoas

Valores observados (x_i): 19 20 21 26 30 37 20 18 25 22 20 33

Valores ordenados ($x_{(i)}$): 18 19 20 20 20 21 22 25 26 30 33 37

Amplitude amostral: $A = x_{(12)} - x_{(1)} = 37 - 18 = 19$

Número de classes: $\text{Int}[1 + 3.321928 \cdot \log_{10}(12)] = \text{Int}[4.585] = 4$

Amplitude das classes: $h = \frac{A}{k} = \frac{19}{4} = 4.75 \approx 5$ (valor arredondado por excesso)

$$\varepsilon = kh - A = 4 \times 5 - 19 = 1;$$

$$c_1 = x_{(1)} - \frac{\varepsilon}{2} = 18 - \frac{1}{2} = 17.5$$

$$c_2 = c_1 + h = 17.5 + 5 = 22.5 \quad \rightarrow \quad C_1 =]17.5, 22.5]$$

$$c_3 = c_2 + h = 22.5 + 5 = 27.5 \quad \rightarrow \quad C_2 =]22.5, 27.5]$$

$$c_4 = c_3 + h = 27.5 + 5 = 32.5 \quad \rightarrow \quad C_3 =]27.5, 32.5]$$

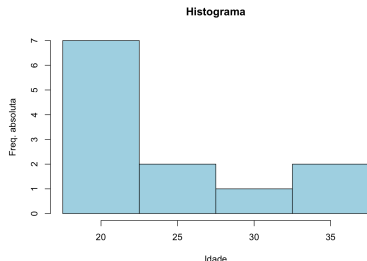
$$c_5 = c_4 + h = 32.5 + 5 = 37.5 \quad \rightarrow \quad C_4 =]32.5, 37.5]$$

Organização de dados (cont.)

Tabela de frequências da variável “Idade”

Faixa etária $]c_i, c_{i+1}]$	Freq Absoluta (n_i)	Freq Relativa (f_i)	Freq Absoluta Acum (N_i)	Freq Relativa Acum (F_i)	Ponto médio $x_i^* = \frac{c_i + c_{i+1}}{2}$
]17.5, 22.5]	7	$7/12 \approx 58.33\%$	7	$7/12 \approx 58.33\%$	20
]22.5, 27.5]	2	$2/12 \approx 16.67\%$	9	$9/12 \approx 75\%$	25
]27.5, 32.5]	1	$1/12 \approx 8.33\%$	10	$10/12 \approx 83.33\%$	30
]32.5, 37.5]	2	$2/12 \approx 16.67\%$	12	$1 \approx 100\%$	35
Total	12	$1 \approx 100\%$	-	-	-

- A representação gráfica é feita num **histograma**. Histograma é uma representação gráfica dos dados em que se marcam as classes (intervalos) no eixo horizontal e as frequências (absoluta ou relativa) no eixo vertical.
- Cada retângulo corresponde a uma classe.
- A largura de cada retângulo é igual à amplitude da classe.
- Se as classes tiverem todas a mesma amplitude, a altura do retângulo é proporcional à frequência.



3. Medidas Descritivas

Uma forma de resumir a informação contida nos dados é através da obtenção de medidas específicas chamadas estatísticas, que são calculadas diretamente a partir dos dados. Estas estatísticas fornecem uma descrição quantitativa das características principais dos dados e podem incluir:

- **Medidas de localização:** localizam o centro da amostra.
- **Medidas de dispersão:** medem a variabilidade dos dados.
- **Medidas de assimetria:** determinam o tipo de assimetria dos dados.
- **Medidas de achatamento:** indicam o tipo de achatamento dos dados.

- **Medidas de localização**

- Medidas de tendência central: média, moda, mediana
- Medidas de tendência não central: quantis

- **Medidas de dispersão**

- Medidas de dispersão absoluta: amplitude total, variância/desvio padrão, amplitude inter-quartis
- Medidas de dispersão relativa: coeficiente de variação

- **Medidas de assimetria**

- coeficiente de assimetria

- **Medidas de achatamento**

- coeficiente de achatamento

Medidas de localização

Seja $(x_1, x_2, \dots, x_n)$ uma amostra de dimensão n , n_i a frequência absoluta, f_i a frequência relativa e x_i^* o ponto médio da classe.

Média (dados não agrupados)

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{x_1 + x_2 + \dots + x_n}{n}.$$

Média (dados agrupados)

$$\bar{x} = \frac{\sum_{i=1}^k x_i n_i}{n} = \sum_{i=1}^k x_i f_i.$$

Média (dados agrupados em classes)

$$\bar{x} = \frac{\sum_{i=1}^k x_i^* n_i}{n} = \sum_{i=1}^k x_i^* f_i$$

Medidas de localização (cont.)

Exemplo (média dados não agrupados): Calcule a idade média das pessoas.

x = “idade da pessoa”

Valores observados (x_i): 19 20 21 26 30 37 20 18 25 22 20 33

$$\bar{x} = \frac{\sum_{i=1}^{12} x_i}{12} = \frac{19+20+21+26+30+37+20+18+25+22+20+33}{12} = 24.25$$

A idade média amostral é de 24.25 anos.

Média no R

```
# Criar um vetor com os valores da variável
x <- c(19,20,21,26,30,37,20,18,25,22,20,33)
# Calculando a média de x
media <- mean(x)
print(media)
[1] 24.25
```

Exemplo (média dados agrupados)

Agreg familiar (x_i)	Freq Abs (n_i)	Freq Rel (f_i)
2	2	0.1667
3	8	0.6667
4	1	0.0833
5	1	0.0833
Total	12	1

$$\begin{aligned}\bar{x} &= \frac{\sum_{i=1}^4 n_i x_i}{12} \\ &= \frac{2 \times 2 + 8 \times 3 + 1 \times 4 + 1 \times 5}{12} \\ &= 3.083\end{aligned}$$

$$\begin{aligned}\bar{x} &= \sum_{i=1}^4 f_i x_i \\ &= 0.1667 \times 2 + 0.6667 \times 3 + 0.0833 \times 4 \\ &\quad + 0.0833 \times 5 \\ &= 3.083\end{aligned}$$

Exemplo (média dados agrupados em classes)

Faixa etária $]c_i, c_{i+1}]$	Freq Absoluta (n_i)	Freq Relativa (f_i)	Ponto médio $x_i^* = \frac{c_i + c_{i+1}}{2}$
$]17.5, 22.5]$	7	$7/12 \approx 58.33\%$	20
$]22.5, 27.5]$	2	$2/12 \approx 16.67\%$	25
$]27.5, 32.5]$	1	$1/12 \approx 8.33\%$	30
$]32.5, 37.5]$	2	$2/12 \approx 16.67\%$	35
Total	12	$1 \approx 100\%$	-

$$\bar{x} = \frac{\sum_{i=1}^4 n_i x_i^*}{12} = \frac{7 \times 20 + 2 \times 25 + 1 \times 30 + 2 \times 35}{12} = 24.17$$

Moda

É o valor que ocorre com maior frequência.

- É a única medida descritiva utilizada para dados qualitativos nominais.
- Uma amostra pode possuir mais do que uma ou não ter moda.

Medidas de localização (cont.)

Exemplo: Calcule a moda do conjunto de dados $\{7, 9, 7, 5, 9, 8, 7\}$.

Vamos ordenar para ajudar na visualização.

$$\{5, 7, 7, 7, 8, 9, 9\}$$

Portanto, a **moda é 7**. O conjunto de dados é unimodal.

Exemplo: Calcule a moda do conjunto de dados $\{7, 7, 9, 6, 9, 7, 4, 9, 9, 7, 1, 2\}$.

Vamos ordenar para ajudar na visualização.

$$\{1, 2, 3, 4, 6, 7, 7, 7, 7, 9, 9, 9, 9\}$$

Portanto, a **moda é 7 e 9**. O conjunto de dados é bimodal.

Mediana

É o valor da variável que **divide os dados ordenados em duas partes de igual frequência**, ou seja, 50% das observações são menores ou iguais à mediana e 50% são maiores ou iguais à mediana.

$$me = \begin{cases} x_{(\frac{n+1}{2})}, & \text{para } n \text{ ímpar} \\ \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2}, & \text{para } n \text{ par} \end{cases}$$

Medidas de localização (cont.)

Exemplo: Calcule a mediana das observações (12, 4, 1, 3, 2, 1, 7).

- **Tamanho do conjunto:** $n = 7$
- **Dados ordenados:** (1,1,2,3,4,7,12)
- Como $n = 7$ é ímpar, temos $\frac{n+1}{2} = \frac{8}{2} = 4$. Portanto,
 $me = x_{(4)} = 3$.

Exemplo: Calcule a mediana das observações: (1, 1, 2, 1, 3, 5, 4, 2).

- **Tamanho do conjunto:** $n = 8$
- **Dados ordenados:** (1,1,1,2,2,3,4,5)
- Como $n = 8$ é par, temos $\frac{n}{2} = \frac{8}{2} = 4$ e $\frac{n}{2} + 1 = 5$. Portanto,

$$me = \frac{x_{(4)} + x_{(5)}}{2} = \frac{2 + 2}{2} = 2.$$

Mediana no R

```
> x <- c(12,4,1,3,2,1,7)
> median(x)
[1] 3
```

Quantil de ordem α (Q_α , $0 < \alpha < 1$)

Chama-se **quantil de ordem α** ($0 < \alpha < 1$), o valor Q_α tal que

- pelo menos $\alpha \times 100\%$ das observações ordenadas são $\leq Q_\alpha$
- pelo menos $(1 - \alpha) \times 100\%$ das observações ordenadas são $\geq Q_\alpha$.

Mais precisamente,

$$(x_1, x_2, \dots, x_n) \rightarrow (x_{(1)}, x_{(2)}, \dots, x_{(n)})$$

$$Q_\alpha = \begin{cases} \frac{x_{(n\alpha)} + x_{(n\alpha+1)}}{2}, & \text{se } n\alpha \text{ é inteiro} \\ x_{(\lfloor n\alpha \rfloor + 1)}, & \text{se } n\alpha \text{ é não inteiro.} \end{cases}$$

$\lfloor * \rfloor$ — parte inteira do número.

Quartis

São valores numéricos que dividem o conjunto ordenado das observações em quatro partes iguais.

- $Q_{1/4}$ é o 1º quartil
 - $Q_{2/4}$ é o 2º quartil (mediana)
 - $Q_{3/4}$ é o 3º quartil
-
- Os **decis** correspondem aos quantis de ordem $\frac{1}{10}, \frac{2}{10}, \dots, \frac{9}{10}$.
 - Os **percentis** correspondem aos quantis de ordem $\frac{1}{100}, \frac{2}{100}, \dots, \frac{99}{100}$.

Medidas de localização (cont.)

Exemplo: Calcule o 1º quartil das observações (12, 4, 1, 3, 2, 1, 7).

- **Tamanho do conjunto:** $n = 7$
- **Dados ordenados:** (1, 1, 2, 3, 4, 7, 12)
- Como $n\alpha = 7 \times \frac{1}{4} = \frac{7}{4} = 1.75$ é não inteiro, temos

$$Q_{1/4} = x_{(\lfloor 1.75 \rfloor + 1)} = x_{(2)} = 1$$

Exemplo: Calcule o 3º quartil das observações acima.

- **Tamanho do conjunto:** $n = 7$
- **Dados ordenados:** (1, 1, 2, 3, 4, 7, 12)
- Como $n\alpha = 7 \times \frac{3}{4} = \frac{21}{4} = 5.25$ é não inteiro, temos

$$Q_{3/4} = x_{(\lfloor 5.25 \rfloor + 1)} = x_{(6)} = 7$$

Quantis no R

```
# Criar vetor com os valores da variável
```

```
x <- c(12, 4, 1, 3, 2, 1, 7)
```

```
# Calcular quantis
```

```
quantis <- quantile(x, type=2)
```

```
print(quantis)
```

```
0% 25% 50% 75% 100%
```

```
1    1    3    7   12
```

```
# Somente o primeiro quartil
```

```
quantile(x, probs=0.25, type=2)
```

```
25%
```

```
1
```

Amplitude (R)

Diferença entre o valor máximo e o valor mínimo da amostra

$$R = \max(x_i) - \min(x_i) = x_{(n)} - x_{(1)}$$

Amplitude inter-quartis (IQR)

Diferença entre o terceiro e o primeiro quartil

$$IQR = Q_{3/4} - Q_{1/4}$$

Medidas de dispersão (cont.)

Variância (dados não agrupados)

Média dos quadrados dos desvios em relação à média

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1} = \frac{1}{n - 1} \left(\sum_{i=1}^n x_i^2 - n\bar{x}^2 \right)$$

Variância (dados agrupados)

$$s^2 = \frac{\sum_{i=1}^k (x_i - \bar{x})^2 n_i}{n - 1} = \frac{1}{n - 1} \left(\sum_{i=1}^k x_i^2 n_i - n\bar{x}^2 \right)$$

Variância (dados agrupados em classes)

$$s^2 = \frac{\sum_{i=1}^k (x_i^* - \bar{x})^2 n_i}{n - 1} = \frac{1}{n - 1} \left(\sum_{i=1}^k (x_i^*)^2 n_i - n\bar{x}^2 \right)$$

Medidas de dispersão (cont.)

Desvio padrão

$$s = +\sqrt{s^2}$$

Está na mesma unidade de medida da média.

Coeficiente de variação

É a proporção da média representada pelo desvio padrão.

$$cv = \frac{s}{\bar{x}} \times 100\%$$

- Se $cv \leq 15\%$, os dados apresentam uma **variabilidade fraca**;
- Se $15\% < cv < 30\%$, os dados apresentam uma **variabilidade média**;
- Se $cv \geq 30\%$, os dados apresentam uma **variabilidade elevada**.

Medidas de dispersão (cont.)

Exemplo: Consideremos que x_{1i} e x_{2i} são conjuntos de valores referentes aos salários (\$) de mulheres de homens, para os quais foram obtidas as seguintes medidas:

Mulheres (X_1):

$$\bar{x}_1 = 1300$$

$$s_1 = 340$$

Homens(X_2):

$$\bar{x}_2 = 2500$$

$$s_2 = 420$$

Qual grupo varia mais em relação aos salários?

$$\bar{x}_1 = 1300$$

$$s_1 = 340$$

$$cv_1 = 26.2\%$$

$$\bar{x}_2 = 2500$$

$$s_2 = 420$$

$$cv_2 = 16.8\%$$

- O maior desvio padrão, quando comparado à sua média, representou menor variação.

Medidas de dispersão no R

Amplitude no R

```
# Criar um vetor com os valores da variável
x <- c(12, 4, 1, 3, 2, 1, 7)
# Calcular a amplitude
amplitude <- max(x)-min(x)
print(amplitude)
[1] 11
```

Amplitude inter-quartis no R

```
# Calcular a amplitude inter-quartis
ampli_iqr <- IQR(x, type=2)
print(ampli_iqr)
[1] 6
```

Medidas de dispersão no R (cont.)

Variância no R

```
# Criar vetor com os valores da variável
x <- c(12, 4, 1, 3, 2, 1, 7)
# Calculando a variância
variancia <- var(x)
print(variancia)
[1] 15.90476
```

Desvio padrão no R

```
# Calculando o desvio padrão
desvio_padrao <- sd(x)
print(desvio_padrao)
[1] 3.988077
```

Coeficiente de variação no R

```
# Calculando o coeficiente de variação
coef_variacao <- sd(x)/mean(x)
print(coef_variacao)
[1] 0.9305514
```

Outlier

Outliers são pontos de dados que diferem significativamente da maioria dos dados em um conjunto de dados. Eles são valores extremos que podem indicar uma variabilidade significativa, um erro de medição ou um fenómeno diferente do restante do conjunto de dados. Em análise estatística, os outliers podem distorcer a interpretação dos dados, afetando médias e variâncias, e, por isso, é importante identificá-los e decidir como tratá-los.

Uma observação x_i é classificada como **outlier moderado** se

$$x_i \notin [Q_{1/4} - 1.5 \times IQR, Q_{3/4} + 1.5 \times IQR]$$

ou como **outlier severo** se

$$x_i \notin [Q_{1/4} - 3 \times IQR, Q_{3/4} + 3 \times IQR]$$

Outliers (cont.)

Ou seja, um outlier será **moderado** se

$$Q_{3/4} + 1.5 \times IQR < x_i \leq Q_{3/4} + 3 \times IQR$$

ou

$$Q_{3/4} - 3 \times IQR \leq x_i < Q_{1/4} - 1.5 \times IQR$$

e será **severo** se

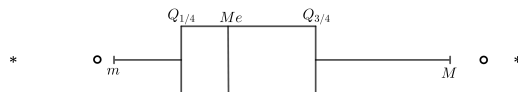
$$x_i > Q_{3/4} + 3 \times IQR$$

ou

$$x_i < Q_{1/4} - 3 \times IQR$$

Outliers e Caixa de bigodes (cont.)

Boxplot com outliers



- m — menor dos valores das observações que não é um outlier;
- M — maior dos valores das observações que não é um outlier;
- $\circ$ — outlier moderado;
- $*$ — outlier severo.

Outliers e Caixa de bigodes (cont.)

Exemplo: Num teste registaram-se as seguintes classificações

5 1 2 4 4 3 3 5 10 2 4 3 18 1 6

Verifique a existência de outliers na amostra acima.

Resolução:

Dimensão da amostra: $n = 15$

Amostra ordenada:

1 1 2 2 3 3 3 4 4 4 5 5 6 10 18

Mediana: $me = x_{(\lfloor n/2 \rfloor + 1)} = x_{(\lfloor 7.5 \rfloor + 1)} = x_{(8)} = 4$

Primeiro Quartil: $Q_{1/4} = x_{(\lfloor n/4 \rfloor + 1)} = x_{(\lfloor 3.75 \rfloor + 1)} = x_{(4)} = 2$

Terceiro quartil: $Q_{3/4} = x_{(\lfloor 3n/4 \rfloor + 1)} = x_{(\lfloor 11.25 \rfloor + 1)} = x_{(12)} = 5$

Intervalo inter-quartis: $IQR = Q_{3/4} - Q_{1/4} = 5 - 2 = 3$

Outliers e Caixa de bigodes (cont.)

Outlier moderado

$$x_i \notin [Q_{1/4} - 1.5 \times IQR, Q_{3/4} + 1.5 \times IQR] = [2 - 1.5 \times 3, 5 - 1.5 \times 3] = [-2.5, 9.5]$$

Como $x_{(14)}, x_{(15)} \notin [2.5, 9.5]$, então $x_{(14)} = 10$ e $x_{(15)} = 18$ são a princípio outliers moderados.

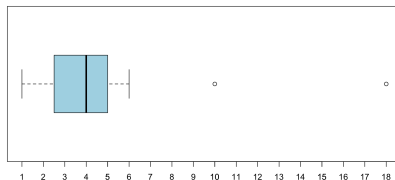
Outliers severos

$$x_i \notin [Q_{1/4} - 3 \times IQR, Q_{3/4} + 3 \times IQR] = [2 - 3 \times 3, 5 + 3 \times 3] = [-7, 14]$$

Como $x_{(15)} = 18 \notin [-7, 14]$ então $x_{(15)}$ é um outlier severo.

Portanto, temos dois outliers, 10 moderado e 18 severo.

Outliers e Caixa de bigodes (cont.)



No R

```
x <- c(1,1,2,2,3,3,3,4,4,4,5,5,6,10,18)
boxplot(x, col="lightblue", horizontal = T)
axis(1, at=1:20, labels = 1:20)
```

4. Análise Bivariada

A **análise bivariada** é o estudo da relação entre duas variáveis quantitativas. No contexto da Engenharia Informática, um exemplo seria analisar como o **tempo de resposta de um servidor** varia com o **número de utilizadores simultâneos**, ou como a **quantidade de memória utilizada** está relacionada com o **tamanho dos ficheiros processados**.

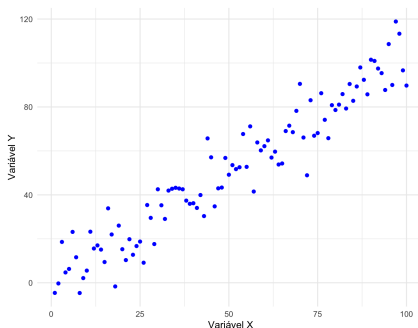
Análise bivariada (cont.)

Análise da associação entre duas variáveis

- 1 Começa pela recolha de uma amostra bivariada, ou seja, n pares de observações

$$(x_1, y_1); (x_2, y_2); \dots; (x_n, y_n)$$

- 2 Representação gráfica usando um **diagrama de dispersão**. Gráfico que mostra a relação entre duas variáveis. Cada ponto no gráfico representa uma observação.



Análise bivariada (cont.)

- 3 Verificar se a relação entre X e Y é linear. Isso é traduzido em verificar se parece existir ou não **correlação entre as variáveis** e em caso afirmativo, se essa correlação é:
- positiva ou negativa;
 - forte ou fraca;
 - tipo de correlação.
- 4 Se este for o caso, poderemos estudar a relação linear entre as duas variáveis usando um **modelo de regressão linear simples**, cuja equação é dada por

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- Y – variável explicada ou dependente;
- X – variável explicativa ou independente;
- ϵ – variável de tipo residual que inclui outros fatores explicativos de Y não incluídos em X e possíveis erros de medição;
- β_0 e β_1 são os parâmetros da reta a ajustar (constantes);
- β_0 é a intersecção da reta com o eixo vertical;
- β_1 é o declive da reta.

Análise bivariada (cont.)

Covariância

A **covariância** entre duas variáveis x e y é uma **medida da associação linear** entre as duas variáveis:

$$\text{cov}[x, y] = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n - 1} = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{n - 1}$$

O valor da covariância depende das unidades de medida utilizadas.

Análise da covariância

- Se $\text{cov}[x, y] > 0$, existe uma **associação linear positiva**, isto é, ambas as variáveis tendem a variar no mesmo sentido;
- Se $\text{cov}[x, y] < 0$, existe uma **associação linear negativa**, isto é, as variáveis tendem a variar em sentidos opostos;
- Se $\text{cov}[x, y] = 0$, **não existe associação linear** entre as variáveis.

A informação contida na covariância é principalmente sobre o sinal da associação entre x e y e não sobre a intensidade.

Coeficiente de correlação linear amostral

$$r = \frac{\text{cov}[x, y]}{s_x s_y}, \quad \text{com} \quad -1 \leq r \leq 1.$$

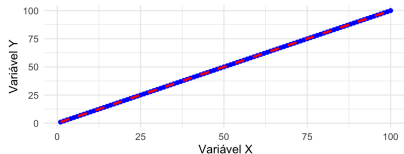
- Permite averiguar o grau de associação linear entre as duas variáveis.
- O seu valor não depende das unidades de medida que as variáveis estão expressas.

Análise do coeficiente de correlação

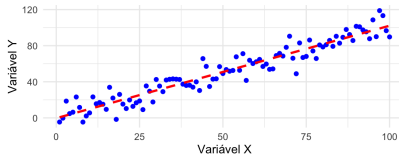
- $r = -1 \Rightarrow$ existe uma **correlação linear negativa perfeita** entre x e y ;
- $r = 1 \Rightarrow$ existe uma **correlação linear positiva perfeita** entre x e y ;
- $r = 0 \Rightarrow$ **não existe correlação linear** entre x e y ;
- $-1 < r < 0 \Rightarrow$ existe uma **correlação linear negativa** entre x e y (menos forte do que quando $r = -1$);
- $0 < r < 1 \Rightarrow$ existe uma **correlação linear positiva** entre x e y (menos forte do que quando $r = 1$).

Análise bivariada (cont.)

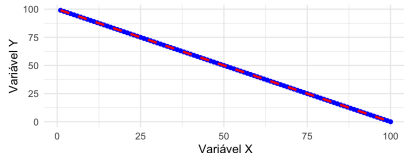
Correlação Positiva Perfeita



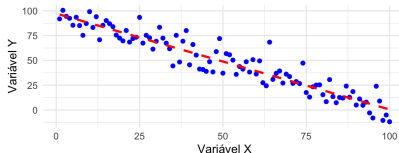
Correlação Positiva



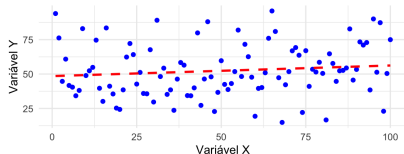
Correlação Negativa Perfeita



Correlação Negativa



Correlação Zero



Exemplo: Engenheiros de software e administradores de sistemas estão interessados em estudar a relação entre o **número de utilizadores ativos** e o **tempo de resposta de um servidor**.

Compreender essa relação é essencial para otimizar a performance dos sistemas, garantindo que os servidores consigam processar pedidos de forma eficiente sem comprometer a experiência do utilizador. Um aumento no número de utilizadores pode levar a tempos de resposta mais elevados devido ao maior consumo de recursos computacionais, mas o impacto exato depende da arquitetura do sistema e da capacidade do hardware.

Análise bivariada (cont.)

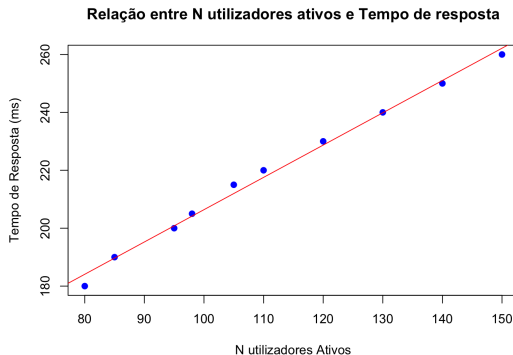
Os dados abaixo representam o **número de utilizadores ativos** e o **tempo médio de resposta do servidor** (em milissegundos) num determinado período de tempo. Com base nestes dados, calcule a covariância e o coeficiente de correlação de Pearson entre as duas variáveis para avaliar o grau de associação entre elas.

Amostra	Nº de Utilizadores Ativos	Tempo de Resposta (ms)
1	120	230
2	85	190
3	150	260
4	110	220
5	95	200
6	105	215
7	130	240
8	80	180
9	140	250
10	98	205

Análise bivariada (cont.)

Covariância entre N de utilizadores ativos e tempo de resposta = 607.5556

Coeficiente de Correlação entre N de utilizadores ativos e tempo de resposta = 0.9966742



Análise bivariada no R

```
# Dados
```

```
util_ativ <- c(120,85,150,110,95,105,130,80,140,98)
```

```
temp_resp <- c(230,190,260,220,200,215,240,180,250,205)
```

```
# Calcular a covariância
```

```
covariancia <- cov(util_ativ, temp_resp)
```

```
correlacao <- cor(util_ativ, temp_resp)
```

```
# Exibir resultados
```

```
print(paste("Covariância entre N utilizadores ativos e Tempo de  
resposta:", covariancia))
```

```
print(paste("Coeficiente de Correlação entre N utilizadores ativos  
e Tempo de resposta:", correlacao))
```

```
# Gráfico de dispersão
```

```
plot(util_ativ, temp_resp, main = "Relação entre N utilizadores ativos e  
Tempo de resposta", xlab = "N utilizadores Ativos", ylab = "Tempo de  
Resposta (ms)", pch = 19, col = "blue")
```

```
# Adicionar linha de tendência
```

```
abline(lm(temp_resp ~ util_ativ), col = "red")
```

- Manuel Cabral Morais (2020): Probabilidade e Estatística: Teoria, Exemplos & Exercícios. IST Press, 1a edição.