

ANÁLISE EXPLORATÓRIA DE DADOS

Ralph dos Santos Silva

<https://www.ralphsilva.org/aed.html>

Instituto de Matemática
Centro de Ciências Matemáticas e da Natureza
Universidade Federal do Rio de Janeiro

Referências

- Freire, S. M. (2021). *Bioestatística Básica*. ([PDF](#))
- Kabacoff, R. (2015). *R in Action: Data Analysis and Graphics with R*, 2^a Ed.
- **Morettin, P. A. e Bussab, W. O. (2017). *Estatística Básica*, 9^a Ed.**
- Spector, P. (2008). *Data Manipulation with R*.
- Teetor, P. (2011). *R Cookbook*.
- Wickham, H., Çetinkaya-Rundel, M. e Grolemund, G. (2023). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*.

Introdução

O que é **Análise Exploratória de Dados**?

Tukey, 1961 – “Procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data.”

Tukey, 1977

- “It is important to understand what you can do before you learn to measure how well you seem to have done it.”
- Exploratory data analysis can never be the whole story, but nothing else can serve as the foundation stone – as the first step.”
- “To learn about data analysis, it is right that each of us try many things that do not work.”

Análise Exploratória de Dados

- Extrair informação geral da massa numérica de dados original.
- Visualização da distribuição dos dados - **gráficos**.
- **Resumos numéricos** da massa de dados (medidas resumo: média, variância, correlação, etc.).

Objetivo: entender como os dados se comportam. AED é conjunto de métodos dedutivos (descrição somente do conjunto de dados observado). Entretanto, os elementos levantados na AED podem ser utilizados, posteriormente, para se fazer inferência (processo indutivo) sobre universos maiores que o conjunto de dados observado.

Tipos de Variáveis

- Variável é cada característica registrada sobre uma unidade de observação (ou seja, sobre um indivíduo ou elemento qualquer).
- São exemplos de variáveis: sexo, idade, grau de instrução, peso, altura, nacionalidade, estado civil, IMC, nível de colesterol, nível de glicose, etc.
- Importante determinar o tipo de cada variável a ser analisada: métodos mais adequados dependem do tipo da variável.

Tipos de Variáveis: Categóricas

- **Variáveis Categóricas** são aquelas em que se registra um atributo qualitativo do elemento observado.
- Assume-se que a variável categórica é definida sobre k níveis (categorias) e, ainda, que:
 - ♦ as k categorias são mutuamente excludentes, disjuntas ou mutuamente exclusivas (não há interseção entre categorias - cada indivíduo classificado em uma única categoria);
 - ♦ as k categorias são exaustivas (a união das categorias dá conta de todas as possibilidades de classificação)

Variáveis Categóricas: Exemplos

- Sexo (Feminino/ Masculino) é variável categórica com $k = 2$ níveis.
- Grau de instrução é variável categórica que pode apresentar diferente quantidade de níveis, por exemplo:
 - ♦ ($k = 7$): sem formação, fundamental incompleto, fundamental completo, médio incompleto, médio completo, superior incompleto, superior completo;
 - ♦ níveis dicotomizados ($k = 2$), por exemplo: tem ou não nível superior completo.

Variáveis Categóricas - Nominais e Ordinais

- Variáveis Categóricas dividem-se entre os seguintes tipos:
 - ♦ **Categóricas Ordinais**: quando há alguma ordenação possível entre as categorias. Exemplo: grau de instrução, classe social;
 - ♦ **Categóricas Nominais**: quando não há qualquer ordenação possível entre as categorias. Exemplo: sexo, cor, região em que reside.

Variáveis Categóricas - Codificação

- É usual que variáveis categóricas sejam codificadas numericamente em bases de dados. Por exemplo:
 - ♦ uma variável dicotômica, como Sexo, pode ter codificação: (F-0 , M-1) ou (F-1, M-2);
 - ♦ uma variável com diversas categorias pode ser codificada numericamente, por exemplo: Região do Brasil em que mora (Norte - 1, Nordeste -2, Centro-Oeste - 3, Sudeste - 4, Sul - 5).
- A codificação numérica facilita a digitação dos dados.
- Os números usados na codificação são meros rótulos e não faz sentido realizar operações com esses rótulos. Por exemplo, não faz sentido calcular a região **média** observada.

Variáveis Quantitativas

- Variáveis quantitativas têm valores numéricos que se originam de processos de contagem ou mensuração.
- Dividem-se em:
 - ♦ **Quantitativas discretas**: originadas, tipicamente, de processos de contagem. São exemplos: número de quartos em uma residência, número de leucócitos a cada $1\mu\text{L}$;
 - ♦ **Quantitativas contínuas**: originadas, tipicamente, por processo de mensuração. São exemplos: nível pressórico, peso, altura.
- É lícito realizar operações aritméticas com variáveis quantitativas, por exemplo calcular o peso médio entre os indivíduos observados.
- Podem ser categorizadas. Exemplo: nível pressórico (baixo, normal, alto).

Conceitos Fundamentais da Estatística

Problema geral da Estatística: queremos tirar conclusões sobre toda uma população, mas não conseguimos investigar todos os seus indivíduos.

- Seleccionamos uma amostra de indivíduos que represente razoavelmente bem a população.
- Com os resultados amostra, inferimos resultados para a população.

Definições

- **População:** conjunto de todos os elementos de interesse em um estudo (pessoas, objetos, cidades, etc.).
- **Amostra:** subconjunto escolhido dentre os elementos da população.
- **Parâmetro:** quantidade numérica que representa uma característica da população.
- **Estimador:** variável obtida em função da amostra para estimar um parâmetro da população.
- **Inferência:** área que trata de tirar conclusões sobre a população tendo como base uma amostra.

Análise Exploratória de Dados

- A seguir são apresentados instrumentos para exploração dos diferentes tipos de variáveis, por meio de:
 - ◆ tabelas;
 - ◆ gráficos; e
 - ◆ estatísticas sumárias.

Exemplo: Partos Prematuros

Tabela 1: Histórico de partos prematuros em 189 gestantes.

```
000000000000001000000000000000000000001100000
000000000000000000000000000000000000210021200000
01000000000000000000000030000000001000000000000
00000000000000100010010001000200000010020
01010100001001110100000010010100100000
```

Exemplo: Partos Prematuros

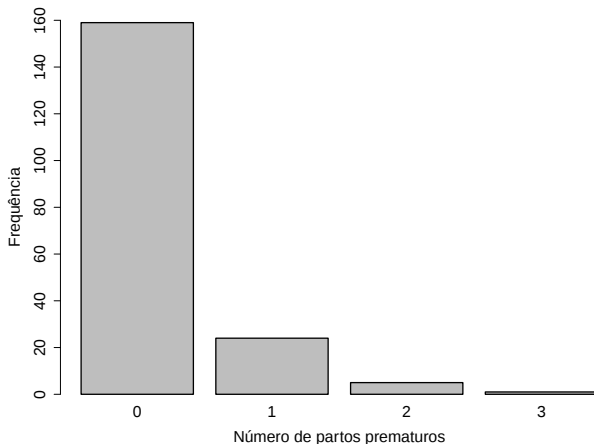


Figura 1: Frequência (absoluta) do número de partos prematuros.

Tabelas de Frequências

- Resumos **numéricos** da massa de dados.
- A construção das tabelas varia de acordo com o tipo de variável:
 - ◆ Variável Qualitativa (Ordinal ou Nominal);
 - ◆ Variável Quantitativa Discreta; ou
 - ◆ Variável Quantitativa Contínua.

Tabelas de Frequência para Variáveis Qualitativas e Quantitativas Discretas

- Contagem de quantas observações pertencem à cada valor possível da variável.
- Pode conter frequências relativas.

Tabela 2: Histórico de partos prematuros de 189 gestantes.

Nº de partos prematuros	Frequência Absoluta	Frequência Relativa (%)
0	159	84,1
1	24	12,7
2	5	2,5
3	3	0,5
Total	189	100,0

Tabelas de Frequência para Variáveis Quantitativas Discretas

Exibir a frequência de cada valor observado da variável.

- 20 domicílios de classe A no Rio de Janeiro.
- Variável: número de filhos por família.
- Observações: 0, 1, 0, 1, 2, 1, 0, 0, 1, 0, 1, 4, 2, 1, 3, 1, 2, 1, 1, 1.

Tabelas de Frequência para Variáveis Quantitativas Discretas

Tabela 3: Frequência (absoluta) e relativa do número de filhos por família.

Nº de filhos	Frequência	Frequência Relativa (%)
0	5	25
1	10	50
2	3	15
3	1	5
4	1	5
Total	20	100

Tabelas de Frequência para Variáveis Quantitativas Contínuas

- Valores em um intervalo da reta;
- Dividir em intervalos menores (classes); e
- Contar quantas observações caem em cada classe.

Perguntas:

- Como escolher as classes?
- Quantas classes? (Evitar extremos)

Observação: aplicável a variáveis discretas se as contagens são elevadas (número de pontos dos jogadores na NBA em um determinado ano).

Exemplo: Idades de Árvores em uma Floresta

Tabela 4: Idades (anos) de 30 árvores em uma floresta.

67	71	78	82	85	86	87	87	92	95
97	99	99	100	101	102	103	103	104	105
107	108	109	112	114	116	118	122	124	125

- Variável: idade da árvore (anos).
- Amostra de 30 árvores ($n = 30$).
- Número de classes: $\sqrt{n} \simeq 5,477$ (**sugestão**).

Passos para Elaboração da Tabela

Passo 1: Máximo e mínimo (67, 125).

Passo 2: Número de classes (k): 5 ou 6, por exemplo. Digamos 5.

Passo 3: Calcule a **amplitude** (A) da amostra (max-min): $125 - 67 = 58$.

Passo 4: Calcule a amplitude das classes (h): $h = \frac{A}{k}$.

$$h = \frac{58}{5} = 11,6 \approx 12.$$

Passo 5: Calcule os limites das classes. Comece pelo valor mínimo e vá somando h :

$$[67, 79) [79, 91) [91, 103) [103, 115) [115, 127)$$

Passo 6: Conte quantas observações pertencem à cada classe:

$$x \in [a, b) \Leftrightarrow a \leq x < b$$

Tabela 5: Frequência (absoluta) e relativa das idades das árvores.

Nº de filhos	Frequência	Frequência Relativa (%)
67 ┤ 79	3	10,00
79 ┤ 91	5	16,67
91 ┤ 103	8	26,66
103 ┤ 115	9	30,00
115 ┤ 127	5	16,67
Total	30	100,00

Representações Gráficas

- Aprender a interpretar gráficos.
- Diferentes tipos $\Rightarrow$ diferentes funcionalidades.
- Todo gráfico deve “contar uma história”.
- Intencionalidade e senso crítico.
- Ferramenta exploratória que pode ser usada em outras etapas da pesquisa estatística (análise dos dados, divulgação dos resultados ou até mesmo testes formais).

Gráfico Setorial para Variáveis Qualitativas

Gráfico setorial 2D $\Rightarrow$ Bem aceito.

Gráfico setorial 3D $\Rightarrow$ **Heresia**.

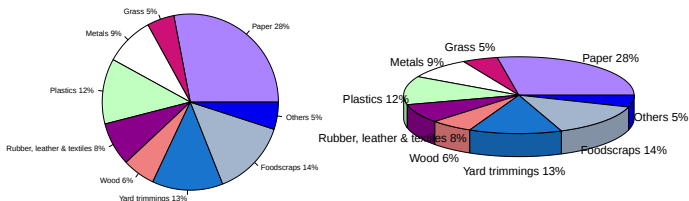


Figura 2: *Municipal Solid Generation 2010.*

Gráfico de Barras

- Representação gráfica da tabela de frequências.
- Aplicável para variáveis quantitativas ou qualitativas. As observações devem estar divididas em classes.

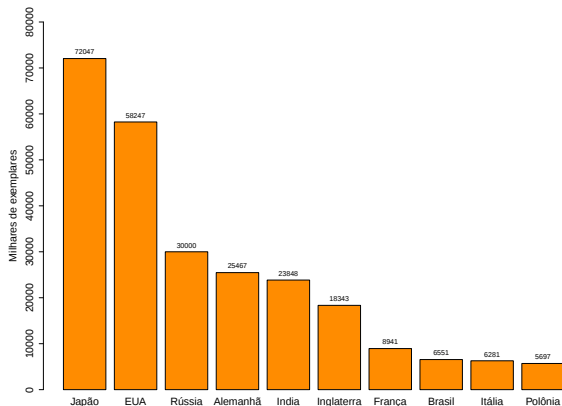


Figura 3: Número de cópias de jornais que circulam diariamente.

Gráfico de Barras

Também podem vir na horizontal.

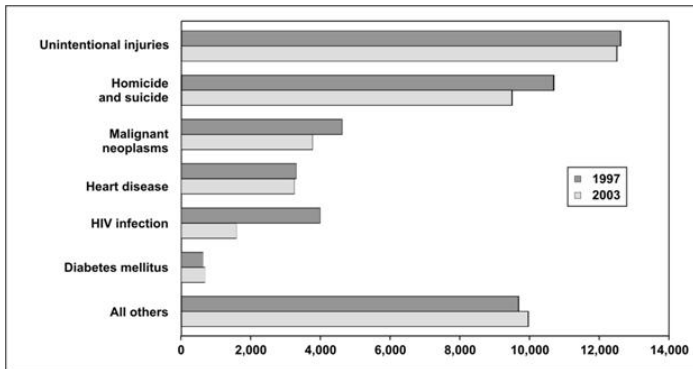


Figura 4: *Number of Deaths by Cause Among 25-34 Year Olds - EUA.*

Fonte:

https://archive.cdc.gov/www_cdc_gov/csels/dsepd/ss1978/lesson4/section4.html

Gráfico em Linha para Variáveis Quantitativas

Esse tipo de gráfico pretende mostrar a evolução temporal de uma variável quantitativa (discreta ou contínua).

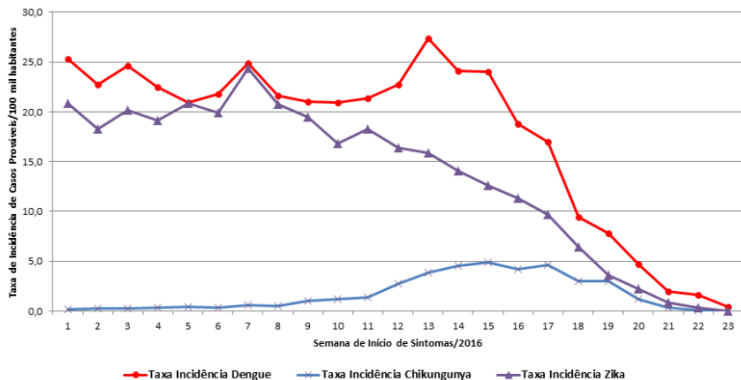


Figura 5: Taxas de incidências de casos prováveis de Dengue, Chikungunya e Zika, por semana epidemiológica de início de sintomas, Estado RJ, ano 2016. Fonte: Boletim Epidemiológico 005/2016, SES-RJ, 20/jun/2016.

Gráfico com Múltiplas Linhas

Comparações entre tratamentos diferentes.

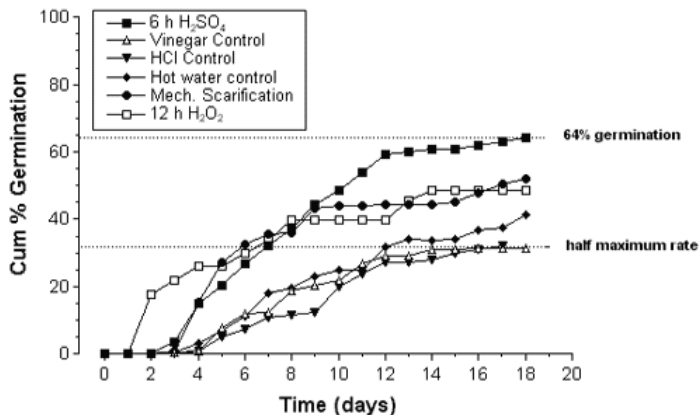


Figura 6: Cumulative germination of gourd seeds following various pregermination treatment.

Histograma

- Mimetizar um modelo de **probabilidade contínuo**.
- Se calculada em proporções, representa uma **distribuição de probabilidade** advinda de um modelo teórico subjacente (desconhecido).

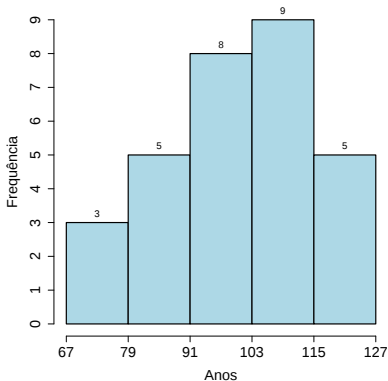


Figura 7: Idades (anos) de árvores em uma floresta.

Histograma (modelo normal; contínuo)

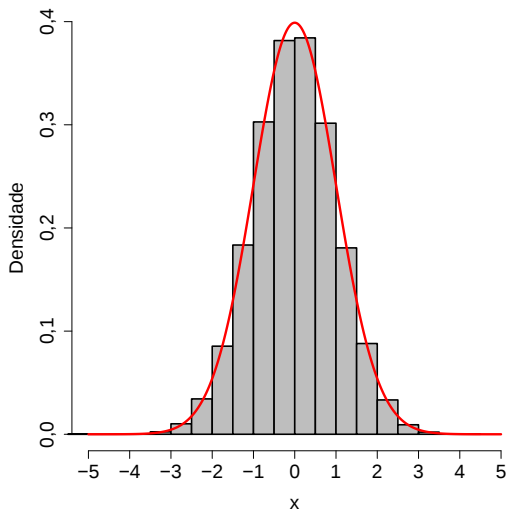


Figura 8: Dados (artificiais) provenientes de um modelo normal.

Histograma (modelo exponencial; contínuo)

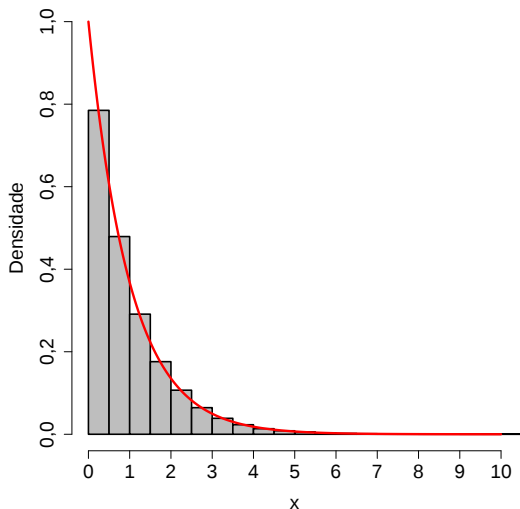


Figura 9: Dados (artificiais) provenientes de um modelo exponencial.

Histograma (modelo Poisson; discreto)

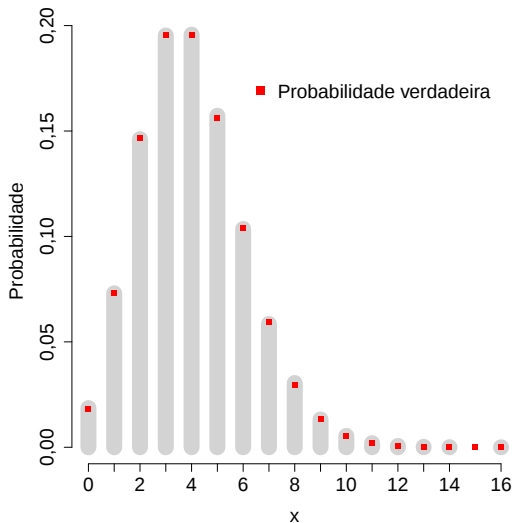


Figura 10: Dados (artificiais) provenientes de um modelo Poisson.

Múltiplos Histogramas

Comparações entre as distribuições empíricas de duas populações.

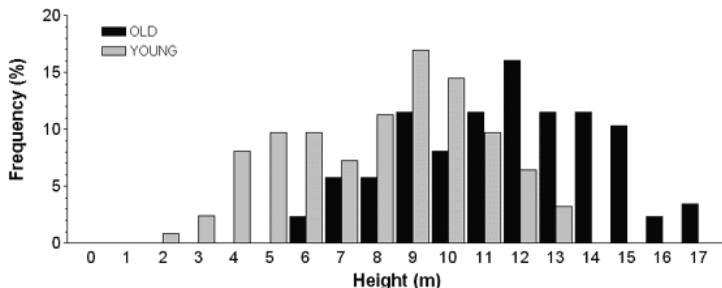


Figura 11: *Height distribution in two recruitment cohort of eastern white pine (*Pinus strobus*) near the eastern shore of Lake Auburn, Maine, in Januaray 2011. $n = 88$ OLD and 123 YOUNG trees.*

Diagrama de ramo e folhas

- Gráfico de barras sem perda de informação numérica.
- “As dezenas são como os ramos e as unidades são como as folhas”.
- Os ramos podem ser unidades e as folhas as casas decimais, etc.

6		7
7		18
8		25677
9		25799
10		0123345789
11		2468
12		245

Figura 12: Idades (anos) de árvores em uma floresta.

Boxplot

Amostra de recém nascidos abaixo do peso:

709 1021 1135 1330 1474 1588 1588 1701 1729 1790 1818 1885
1893 1899 1928 1928 1936 1970 2055 2055 2082 2084 2084
2100 2125 2126 2187 2187 2211 2225 2240 2240 2282 2296 2296
2301 2325 2353 2353 2367 2381 2381 2381 2395 2410 2410 2414
2424 2438 2442 2450 2466 2466 2466 2495 2495 2495 2495

- Quartil 1 (frequência relativa 25%): Q_1 - valor que deixa 1/4 das observações à esquerda;
- Quartil 2, ou mediana (frequência relativa 50%): Q_2 - valor que deixa 2/4 das observações à esquerda;
- Quartil 3 (frequência relativa 75%): Q_3 - valor que deixa 3/4 das observações à esquerda; e
- Amplitude interquartílica: $Q_3 - Q_1$ - medida de dispersão.

Relação Histograma versus *Boxplot*

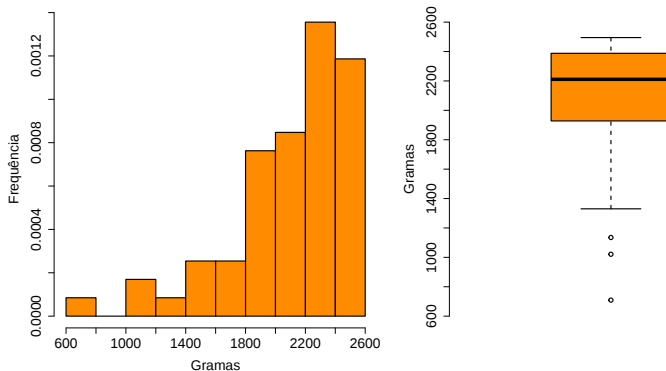


Figura 13: Recém nascidos abaixo do peso.

Construção do *Boxplot*

1. Encontre os 3 quartis.
2. Classifica-se como *outliers à direita*, pontos que se encontram a mais de $1,5 \times$ (amplitude interquartílica) unidades à direita do 3º quartil.
3. Classifica-se como *outliers à esquerda*, pontos que se encontram a mais de $1,5 \times$ (diferença interquartílica) unidades à esquerda do 1º quartil.
4. A linha se estende até os valores mais extremos do conjunto de dados **que não sejam outliers**.

Múltiplos *Boxplots*

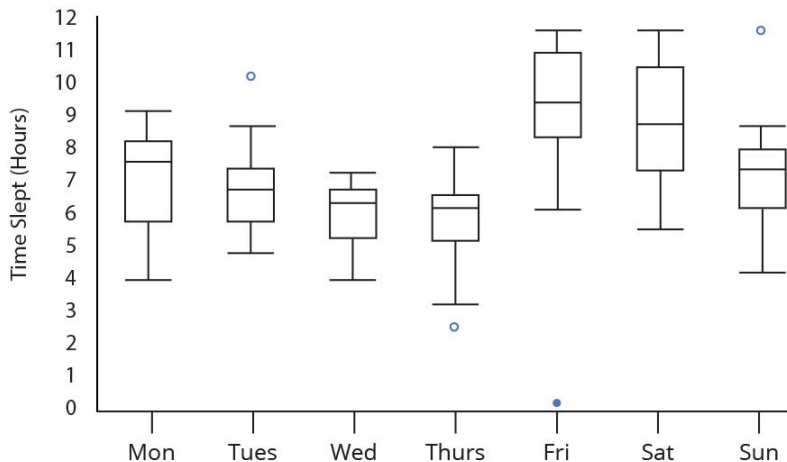


Figura 14: Número de horas dormidas por dia da semana.

Diagrama de Dispersão

- Pares ordenados no plano cartesiano.
- Eixo x: variável 1. Eixo y: variável 2.
- **Objetivo:** investigar relações entre as variáveis 1 e 2.

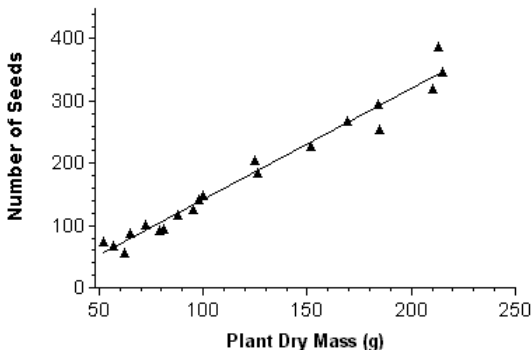


Figura 15: Seed production as a function of plant biomass in waterlilies (*Nuphar luteum*) harvested from Great Works Pound in Northern Maine in August, 2011.

Estatísticas Sumárias

- **Propósito:** resumir características relevantes sobre o comportamento de uma variável.
- Trataremos dos seguintes grupos de medidas:
 - ◆ Medidas de Posição (ou Tendência Central);
 - ◆ Medidas de Dispersão; e
 - ◆ Quantis.

Estatísticas Sumárias: Medidas de Posição

- Medidas de posição (ou de tendência central) objetivam informar em torno de qual ponto da reta real um conjunto de observações se distribui.
- São medidas de posição:
 - ◆ média aritmética (ou, simplesmente, média);
 - ◆ mediana; e
 - ◆ moda.

Medidas de Posição - Média Aritmética

- Sejam $x_1, x_2, \dots, x_n$ os n valores observados para uma variável X . Por exemplo, os níveis de glicose de n indivíduos.
- A média aritmética de X é dada por

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

Exemplo

Considere uma variável X que contém $n = 15$ valores.

10	9	8	12	11	7	10	8,5
9,5	6	14	13	11	12	9	

Daí, podemos calcular a média por

$$\bar{x} = \frac{10 + 9 + 8 + \cdots + 12 + 9}{15} = \frac{150}{15} = 10.$$

Exemplo

Suponha que o primeiro valor seja igual a 100.

100	9	8	12	11	7	10	8,5
9,5	6	14	13	11	12	9	

Daí, podemos calcular a nova média por

$$\bar{x} = \frac{100 + 9 + 8 + \dots + 12 + 9}{15} = \frac{240}{15} = 16.$$

A média pode ser bastante influenciada por **valores extremos**.

A MÉDIA COMO CENTRO DE MASSA DOS DADOS

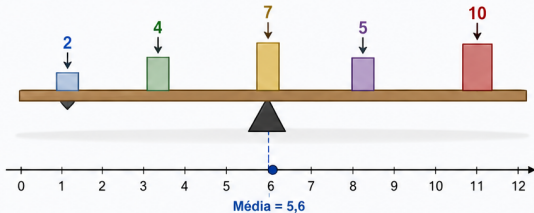


Assim como o centro de massa (equilíbrio) de uma barra depende da posição e do peso de cada ponto, a **média** é o ponto de equilíbrio do conjunto de dados.

1) Analogia física

Barras de comprimentos iguais representam os dados.

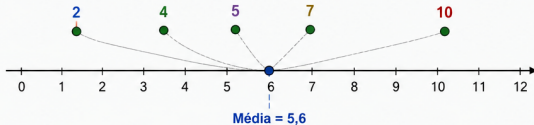
O **peso** de cada barra representa o valor de cada dado.



A **média (5,6)** é o ponto de equilíbrio: o sistema não tende nem para a esquerda nem para a direita.

2) No conjunto de dados

Cada ponto é um dado sobre a reta numérica.



A **média** é o ponto de equilíbrio: a soma das "forças" (distâncias) ponderadas até ela é zero.



Propriedade de equilíbrio da média:

A soma das distâncias (ponderadas) de cada valor até a média é zero: $\sum_{i=1}^n (x_i - \bar{x}) = 0$

x_i : cada dado

$\bar{x}$: média dos dados

n : número de dados

Medidas de Posição - Mediana

Sejam $x_{(1)}, x_{(2)}, \dots, x_{(n)}$ os n valores **ordenados** da variável X .

A mediana $\tilde{x}$ é um ponto tal que o número de observações abaixo e acima de $\tilde{x}$ é o mesmo.

- Se n é ímpar: $\tilde{x}$ é o valor central no conjunto de dados ordenados:

$$\tilde{x} = x_{(\frac{n+1}{2})}$$

- Se n é par: tome $\tilde{x}$ como a média entre os dois valores centrais no conjunto de dados ordenados:

$$\tilde{x} = \frac{1}{2} \left[x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)} \right].$$

Exemplo (continuação)

Para o cálculo da mediana, precisamos ordenar os dados.

6	7	8	8,5	9	9	9,5	10
10	11	11	12	12	13	14	

Note que $n = 15$ e $\frac{n+1}{2} = 8$.

Daí, podemos calcular a mediana por

$$\tilde{x} = x_{(8)} = 10.$$

Ao contrário da média, a mediana não sofre influência de valores extremos. Seu valor permaneceria inalterado se substituíssemos a primeira observação por 10 vezes o valor da média observada.

Medidas de Posição - Moda

A moda é o valor da variável de interesse que é observado com frequência máxima.

Pode haver mais de um valor mais frequente: dados bimodais (duas modas) ou multimodais (três ou mais modas).

Um conjunto de observações sobre uma variável também pode ser amodal (sem moda).

Exemplo

Considere idades de pessoas que fazem parte de um grupo de pais que ensinam os seus filhos bem novos a nadar.

1	1	1	2	2	2	2	3	3
31	31	32	32	32	32	33	33	33

Daí, podemos calcular a **média** por

$$\bar{x} = \frac{1 + 1 + 1 + \cdots + 33 + 33}{18} = \frac{306}{18} = 17,$$

e a **mediana** por

$$\tilde{x} = \frac{x_{(9)} + x_{(10)}}{2} = \frac{3 + 31}{2} = 17.$$

Exemplo

A média e a mediana não refletem o valor típico das idades (grupos de pais e grupo de filhos).

Os passos para se obter a moda são:

1. Encontrar todos os valores distintos da variável;
2. Obter a frequência de cada valor distinto; e
3. Selecione o valor (ou valores) com maior frequência para obter a moda.

Valor	1	2	3	31	32	33
Frequência	3	4	2	2	4	3

Os dois valores com a maior frequência são **2** e **32**.

Esta variável é **bimodal**.

Ordenação entre Medidas de Posição

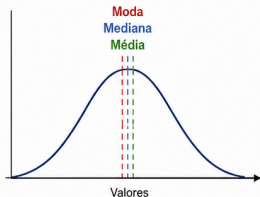
Ao ordenar os valores calculados de média, mediana e moda, podemos ter ideia do formato da distribuição dos dados.

Observação: Para dados empíricos, dificilmente média, mediana e moda serão idênticas, mas se o formato da distribuição for simétrico, seus valores devem ser próximos.

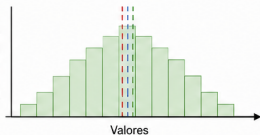
MÉDIA, MEDIANA E MODA SEGUNDO O TIPO DE DISTRIBUIÇÃO

DISTRIBUIÇÃO SIMÉTRICA

DISTRIBUIÇÃO
CONTÍNUA



HISTOGRAMA
(EXEMPLO)

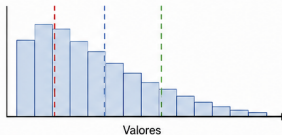
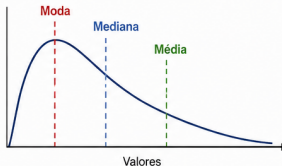


RELAÇÃO ENTRE MÉDIA, MEDIANA E MODA

$$\text{Média} = \text{Mediana} = \text{Moda}$$

A distribuição é equilibrada.
Os valores de tendência central coincidem.

DISTRIBUIÇÃO ASSIMÉTRICA À DIREITA (Assimetria Positiva)

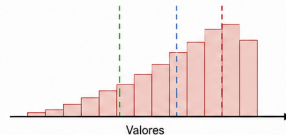
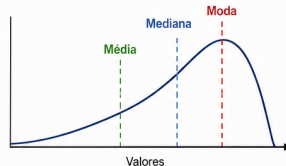


RELAÇÃO ENTRE MÉDIA, MEDIANA E MODA

$$\text{Moda} < \text{Mediana} < \text{Média}$$

A cauda à direita (valores grandes)
puxa a média para a direita.
Logo, Média > Mediana > Moda.

DISTRIBUIÇÃO ASSIMÉTRICA À ESQUERDA (Assimetria Negativa)



RELAÇÃO ENTRE MÉDIA, MEDIANA E MODA

$$\text{Média} < \text{Mediana} < \text{Moda}$$

A cauda à esquerda (valores pequenos)
puxa a média para a esquerda.
Logo, Média < Mediana < Moda.



Resumo: Em distribuições simétricas, média, mediana e moda são iguais.
Em distribuições assimétricas, a média é puxada na direção da cauda.

Quantis

- Quantis são pontos que dividem o conjunto de dados em partes com igual número de observações. São quantis:
 - ♦ Quartis: 3 pontos que dividem o conjunto de dados ordenados em 4 partes com igual número de observações;
 - ♦ Decis: 9 pontos que dividem o conjunto de dados ordenados em 10 partes com igual número de observações; e
 - ♦ Percentis: 99 pontos que dividem o conjunto de dados ordenados em 100 partes com igual número de observações.
- Por exemplo, o primeiro quartil deixa 25% das observações abaixo de seu valor; o quinto percentil deixa 5% das observações abaixo de seu valor.

Estatísticas Sumárias: Medidas de Dispersão

- Medidas de dispersão objetivam informar quão disperso (alternativamente, quão concentrado) é um conjunto de dados.
- São medidas de dispersão:
 - ♦ amplitude amostral;
 - ♦ amplitude interquartílica;
 - ♦ variância amostral;
 - ♦ desvio-padrão amostral; e
 - ♦ coeficiente de variação amostral.

Medidas de Dispersão: Amplitude Amostral

- A medida mais básica possível de variabilidade em um conjunto de dados é a chamada amplitude amostral, dada pela diferença entre os valores máximo e mínimo observados.
- Algumas questões, ao usar a amplitude amostral:
 - ♦ toda a informação entre os extremos é desprezada: pontos entre os extremos são concentrados ou dispersos?
 - ♦ os pontos extremos (máximo e mínimo) destoam dos demais (são *outliers*)?

Exemplos (continuação)

Os dados ordenados provenientes do **primeiro exemplo**:

6	7	8	8,5	9	9	9,5	10
10	11	11	12	12	13	14	

Logo, a amplitude é dada por: $A_1 = 14 - 6 = 8$.

Os dados ordenados provenientes do **segundo exemplo**:

1	1	1	2	2	2	2	3	3
31	31	32	32	32	32	33	33	33

Logo, a amplitude é dada por: $A_2 = 33 - 1 = 32$.

Medidas de Dispersão: Diferença Interquartílica

Lembrando:

- Quartil 1 (frequência relativa 25%): Q_1 - valor que deixa 1/4 das observações à esquerda;
- Quartil 2, ou mediana (frequência relativa 50%): Q_2 - valor que deixa 2/4 das observações à esquerda; e
- Quartil 3 (frequência relativa 75%): Q_3 - valor que deixa 3/4 das observações à esquerda.
- Diferença interquartílica: $DI = Q_3 - Q_1$ - medida de dispersão.
- DI avalia a dispersão na porção central dos dados (entre as 50% observações centrais).

Uso da DI para Avaliação de Pontos Discrepantes

- São consideradas discrepantes observações cujos valores são:
 - ♦ superiores a $Q_3 + 1,5 \times DI$;
 - ♦ inferiores a $Q_1 - 1,5 \times DI$.
- São consideradas extremamente discrepantes observações cujos valores são:
 - ♦ superiores a $Q_3 + 3 \times DI$;
 - ♦ inferiores a $Q_1 - 3 \times DI$.

Medidas de Dispersão: Variância Amostral

- A variância amostral é dada por:

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

- É uma medida da distância “quadrática média” das observações $x_1, x_2, \dots, x_n$ de uma variável X , em relação à sua média.
- Quanto maior a distância quadrática média dos observados em relação à sua média, maior a dispersão na variável X .
- Problema: difícil avaliar se dispersão “grande” ou “pequena”, pois a variância tem unidade igual ao quadrado da unidade de X . Se X é registrada em kg , por exemplo, a variância s^2 tem unidade kg^2 .

Medidas de Dispersão: Desvio-padrão Amostral

- O desvio-padrão amostral é dado pela raiz quadrada positiva da variância amostral:

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

- Tem unidade igual à dos dados observados.
- Quanto maior a distância quadrática média dos observados em relação à sua média, maior a dispersão na variável X .
- Questão: $s = 10kg$ é desvio-padrão “grande” ou “pequeno”? Depende: se $s = 10kg$ se originou das medidas de peso de filhotes de elefantes, provavelmente se trata de elefantinhos com aspectos bastante similares. Já $10kg$ de desvio-padrão entre os pesos de crianças pode indicar pesos bastante heterogêneos.

Coeficiente de Variação

- O coeficiente de variação de uma variável X é dado por:

$$CV = \frac{s}{\bar{x}}$$

- As unidades de s e $\bar{x}$ são idênticas, se cancelam e, portanto, o CV é medida adimensional de dispersão.
- Avalia-se s relativizado pela magnitude média de X .
- Como é adimensional, pode ser usado para comparar dispersão de variáveis com magnitudes e até unidades distintas.

Exemplo

Considere idades de pessoas que fazem parte de um grupo de pais que ensinam os seus filhos bem novos a nadar.

1	1	1	2	2	2	2	3	3
31	31	32	32	32	32	33	33	33

Considere também os dados resumidos pela frequência:

Valor	1	2	3	31	32	33
Frequência	3	4	2	2	4	3

Observação: será apresentado na aula (no quadro) uma forma sistemática de fazer as contas utilizando um formato de tabela. Será apresentado também fórmulas alternativas dos cálculos da média e da variância utilizando frequências absolutas e relativas.