# Introdução à análise descritiva de dados

FLG 5127 Métodos e Técnicas de Pesquisa e Redação Científica

Profa. Dra. Rúbia Gomes Morato

27 de setembro de 2019

## *Tipos de dados de acordo com a medição (Rogerson, 2012) ou tipos de escalas estatísticas*

- Qualitativos
  - Nominal: atributo sem ordenamento (tipo de solo, vegetação)
  - Ordinal: observações hierarquizadas (tamanhos de cidades)
- Quantitativos
  - Intervalar: não possui zero natural (temperatura em graus Celsius ou Fahrenheit)
  - Razão: apresenta a ideia do zero (temperatura em Kelvin)

# Medidas de tendência central
# Measures of Central Tendency

- Média aritmética
- Média geométrica
- Moda
- Mediana
- Valores mínimo e máximo
- Amplitude

- Arithmetic mean
- Geometric mean
- Mode
- Median
- Minimum and maximum values
- Range

# Medidas de dispersão
# Measures of dispersion

- Desvio em relação à média
- Variância da amostra
- Desvio padrão
- Erro padrão
- Coeficiente de variação
- Assimetria
- Curtose

- Mean deviation
- Sample variance
- Standard deviation
- Standard error
- Coefficient of variation
- Skewness
- Kurtosis

# Medidas de tendência central
# Measures of Central Tendency

# Média aritmética
# Arithmetic mean

- Somatório de todos os elementos da série divididos pelo número de elementos.

- Exemplo: 5, 3, 6, 8, 4, 5, 7, 5, 9

- $M_A = (5 + 3 + 6 + 8 + 4 + 5 + 7 + 5 + 9) / 9$
- $M_A = 52/9$
- **A média aritmética é 5,77**

# Média geométrica
# Geometric mean

- A média geométrica é definida, para números positivos, como a raiz n-ésima do produto de *n* elementos de um conjunto de dados.

$$M_G = \sqrt[n]{x_1 \,.\, x_2 \,.\, x_3 \,.\,\ldots\,.\, x_n}$$

- Exemplo: 5, 3, 6, 8, 4, 5, 7, 5, 9
- $M_G = \sqrt[9]{5 * 3 * 6 * 8 * 4 * 5 * 7 * 5 * 9}$
- $M_G = \sqrt[9]{4536000} = 5{,}49$

# Média geométrica

- No Excel: = 4536000^(1/9)
- =num^(1/n)
- "num" é o número cuja raiz se busca encontrar e "n" é a raiz (no exemplo, a nona).

- A média geométrica é usada em alguns casos, como em aplicações financeiras, por exemplo. Entretanto, em certas situações, a geométrica não faz sentido, como por exemplo, quando algum dos valores for zero.

# Moda
# Mode

- A moda é o valor que ocorre mais vezes ou com maior frequência.

- Exemplo: **5**, 3, 6, 8, 4, **5**, 7, **5**, 9

- O valor mais frequente é **5** (ocorre três vezes), portanto a moda é **5**.

# Mediana
# Median

- A mediana é determinada ordenando-se os dados de forma crescente ou decrescente e determinando o valor central da série.
- Exemplo: 3, 4, 5, 5, **5**, 6, 7, 8, 9
- Ou: 9, 8, 7, 6, **5**, 5, 5, 4, 3
- **A mediana é 5**
- Metade dos dados estão à esquerda da mediana e a outra metade à direita da mediana.
- Quando os dados são muito discrepantes, a mediana pode ser afetada por dados extremos, e a moda pode ser mais representativa.

# Valor mínimo e máximo
# Minimum and maximum values

- O menor e o maior valor da série

- Exemplo: 5, 3, 6, 8, 4, 5, 7, 5, 9
- Ordenando: 3, 4, 5, 5, 5, 6, 7, 8, 9

- **O valor mínimo é 3**
- **O valor máximo é 9**

# Amplitude
# Range

- Diferença entre o valor máximo e mínimo
- Exemplo: **3**, 4, 5, 5, 5, 6, 7, 8, **9**
- Amplitude = 9 – 3
- **A amplitude é 6**

# Separatrizes/Quantis
# Quantiles

- Qualquer separatriz que divide o intervalo de frequência de uma população, ou de uma amostra, em partes iguais:
  - Tercil: cada parte tem 33,3% dos dados
  - Quartil: cada parte tem 25% dos dados
  - Quintil: cada parte tem 20% dos dados
  - Decil: cada parte tem 10% dos dados
  - Duodecil: cada parte tem 8,33% dos dados
  - Percentil: cada parte tem 1% dos dados

# Quartil
# Quartile

- O primeiro quartil corresponde aos primeiros 25% dos dados (começa no menor valor até o primeiro quarto dos dados)
- O segundo quartil corresponde ao intervalo entre 25 e 50% (a mediana)
- O terceiro quartil corresponde ao intervalo entre 50 e 75%
- O quarto quartil corresponde ao intervalos entre 75 e 100% (ou o valor máximo)

# Quartis de uma amostra

- Exemplo: 5, 3, 6, 8, 4, 5, 7, 5, 9
- Ordenando: 3, 4, 5, 5, 5, 6, 7, 8, 9

- **O valor mínimo é 3, o máximo é 9 e a mediana 5**

- A identificação do quartil é determinado por:

***Número de observações (ordem do quantil/ quantil)***

# Cálculo de quartis

- ***Cálculo:***
  ***Número de observações (ordem do quantil/quantil)***

***Para quartis (1/4 ou 0,25 ou 25%):***
*Primeiro quartil: número de observações*1/4 (ou 0,25)*
*Segundo quartil: número de observações*2/4 (ou 0,5)*
*Terceiro quartil: número de observações*3/4 (ou 0,75)*
*Quarto quartil: número de observações*4/4 (máx)*

## *Amostra ordenada: 3, 4, 5, 5, 5, 6, 7, 8, 9*

- O primeiro quartil é determinado por:
- 9*(1/4) = 2,25 (que pode ser arredondado para 2), correspondendo ao segundo valor, que é 4.

- O segundo quartil é determinado por:
- 9*(2/4) = 4,5 (que pode ser arredondado para 5), correspondendo ao quinto valor, que é 5.

- O terceiro quartil é determinado por:
- 9*(3/4) = 6,75 (que pode ser arredondado para 7), correspondendo ao sétimo valor, que é 7.

## *Amostra ordenada: 3, 4, 5, 5, 5, 6, 7, 8, 9*

- Assim, temos:

- Primeiro quartil: 3 e 4
- Segundo quartil: 5, 5 e 5
- Terceiro quartil: 6 e 7
- Quarto quartil: 8 e 9

# Amplitude interquartílica
# Interquartile range (IQR)

- O intervalo interquartil é utilizado para avaliar o grau de espalhamento de dados (dispersão) em torno da medida de centralidade (mediana).
- Corresponde a diferença entre o primeiro e o terceiro quartil e concentra os 50% dos dados.

## *Amostra ordenada: 3, 4, 5, 5, 5, 6, 7, 8, 9*

- No exemplo, temos:

- Primeiro quartil: 3 e 4
- ***Segundo quartil: 5, 5 e 5***
- ***Terceiro quartil: 6 e 7***
- Quarto quartil: 8 e 9

- O intervalo interquartil corresponde aos valores entre 5 e 7, que concentram 50% dos dados centralizados na mediana.

# Quintil
# Quintile

- O primeiro quintil corresponde aos primeiros 20% dos dados (começa no menor valor até o primeiro quinto dos dados)
- O segundo quintil corresponde ao intervalo entre 20 (segundo decil) e 40% (ou quarto decil)
- O terceiro quintil corresponde ao intervalo entre 40 (quarto decil) e 60% (ou sexto decil)
- O quarto quintil corresponde ao intervalos entre 60 (sexto decil) e 80% (ou oitavo decil)
- O quinto e último quintil corresponde ao intervalos entre 80 (oitavo decil) e 100% dos dados

# Cálculo de quintis

- ***Cálculo:***
  ***Número de observações (ordem do quantil/ quantil)***

***Para quintis (1/5 ou 0,2 ou 20%):***

*Primeiro quintil : número de observações*1/5*

*Segundo quintil : número de observações*2/5*

*Terceiro quintil : número de observações*3/5*

*Quarto quintil : número de observações*4/5*

*Quinto quintil: número de observações*5/5(máx)*

# Vantagens/desvantagens dos quantis

- Definição dos intervalos para mapa coropléticos de modo equilibrado (cada classe tem aproximadamente a mesma quantidade de unidades)

- Desvantagens: pode separar unidades semelhantes e resultar em classes heterogêneas, agrupando unidades diferentes e separando unidades semelhantes

# Medidas de dispersão
# Measures of dispersion

# Desvio em relação à média
# Mean deviation

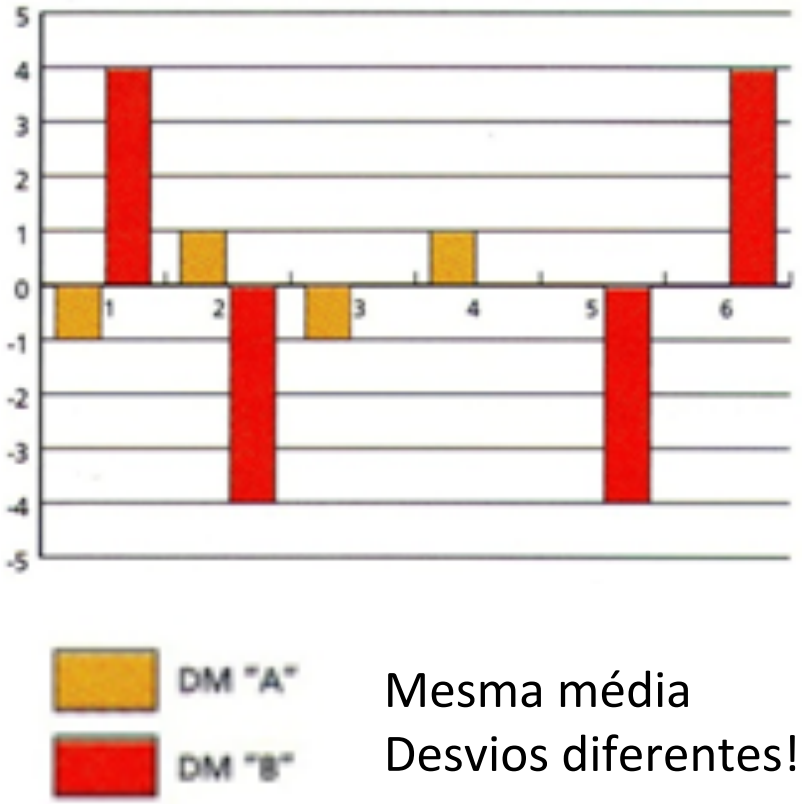
- Diferença entre o valor observado e a média

- Fornece uma ideia da variabilidade dos dados em torno da média.

$$DM = x_i - \overline{x}$$

# Desvio em relação à média (Galvani, 2011)

| A | B | DM "A" | DM "B" |
|---|---|---|---|
| 4 | 9 | -1 | 4 |
| 6 | 1 | 1 | -4 |
| 4 | 5 | -1 | 0 |
| 6 | 5 | 1 | 0 |
| 5 | 1 | 0 | -4 |
| 5 | 9 | 0 | 4 |
| $\bar{x}=5$ | $\bar{x}=5$ | $\sum=0$ | $\sum=0$ |




Mesma média
Desvios diferentes!

# Variância da amostra
# Sample variance

- Somatória do quadrado do desvio em relação à média, dividida pela quantidade de elementos da série menos 1.

$$S^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$$

# Variância da amostra (Galvani, 2011)

| $x$ | $x - \bar{x}$ | $(x - \bar{x})^2$ |
|---|---|---|
| 4 | -1 | 1 |
| 6 | 1 | 1 |
| 4 | -1 | 1 |
| 6 | 1 | 1 |
| 5 | 0 | 0 |
| 5 | 0 | 0 |
| $\bar{x} = 5$ | | $\sum (x - \bar{x})^2 = 4$ |

$$S^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$$

$$S^2 = \frac{4}{6-1}$$

$$S^2 = 0,8$$

# Encontrar a variância da amostra

- ***Amostra: 5, 10, 15, 5, 25***
- Média ($\overline{x}$): 60/5 = 12
- n = 5

$$S^2 = \frac{\sum (x_i - \overline{x})^2}{n - 1}$$

- ***Variância:***
- $[(-7)^2 + (-2)^2 + (3)^2 + (-7)^2 + (13)^2]/ (5-1)$
- [49 + 4 + 9 + 49 + 169]/4
- 280/4 = 70

# Desvio padrão
# Standard deviation

- Raiz da variância

$$S = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

- Medida do grau de dispersão em relação à média.

# Cálculo do desvio-padrão

| *Série* | *Desvio da média* | *Quadrado do desvio da média* |
|---|---|---|
| **9** | 4 | 16 |
| **1** | -4 | 16 |
| **1** | -4 | 16 |
| **2** | -3 | 9 |
| **8** | 3 | 9 |
| **9** | 4 | 16 |
| | | |
| ***Média*** | ***(9+1+1+2+8+9)/6 =>*** | ***30/6 = 5*** |
| | | |
| ***Soma do quadrado dos desvios (16+16+16+9+9+16)*** | | ***82*** |
| ***Variância (Soma do quadrado dos desvios/n-1) = 82/5*** | | ***16,4*** |
| ***Desvio-padrão (Raiz da variância)*** | | ***4,049691*** |

# Erro padrão
# Standard error

- O erro padrão é uma medida de variação de uma média amostral em relação à média da população.
- A aplicada para verificar a confiabilidade da média amostral calculada.
- É obtida pela divisão do desvio padrão pela raiz quadrada do tamanho amostral.
- Quanto menor o erro padrão, menor a dispersão e mais provável que qualquer média de amostra esteja próxima à média da população.

# Erro padrão

- $S_x$: é o erro padrão
- s: é o desvio padrão
- n: é o tamanho da amostra

$$S_x = \frac{s}{\sqrt{n}}$$

- Exemplo: 9, 1, 1, 2, 8, 9
- Média: 5
- Desvio padrão: 4,049691
- N = 6

- $S_x$ = 4,049691 / √6
- $S_x$ = 4,049691 / 2,4494
- ***$S_x$ = 1,6533***

# Coeficiente de variação
# Coefficient of variation

- Expresso em porcentagem, permite comparar variáveis diferentes.

$$CV = \frac{100 \cdot S}{\overline{x}}$$

- Multiplica-se o desvio padrão por 100 e divide-se pela média.

# Coeficiente de variação (Galvani, 2011)

| A | B | C |
|---|---|---|
| 4 | 9 | 9 |
| 6 | 1 | 1 |
| 4 | 5 | 1 |
| 6 | 5 | 2 |
| 5 | 1 | 8 |
| 5 | 9 | 9 |

$$CV = \frac{100 \cdot S}{\overline{x}}$$

Desvio-padrão de A = 0,9
Desvio-padrão de B = 3,6
Desvio-padrão de C = 4,0

***Série C***
Média = 5
Desvio-padrão = 4,0

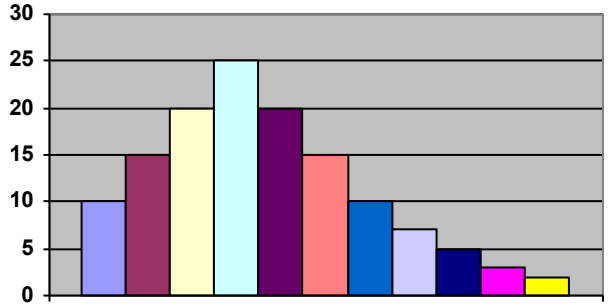
CV = (100 * 4) / 5
CV = 400/5 **= 80%**

$CV_A = 18,0\%$
$CV_B = 72,0\%$
$CV_C = 80,0\%$

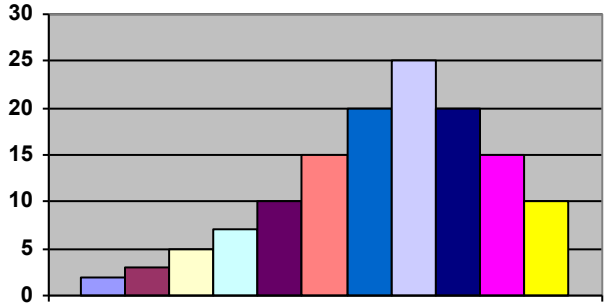
# Assimetria
# Skewness

- Mede o grau de assimetria exibido pelos dados e pelo histograma.
- Quando há mais observações abaixo da média a assimetria é positiva.
- Quando há mais observações acima da média a assimetria é negativa.

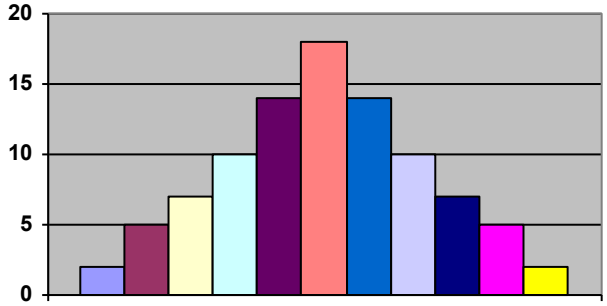
# Assimetria (GUIMARÃES, 1997)




assimetria positiva




assimetria negativa




Distribuição simétrica

# Assimetria

- A assimetria é calculada primeiro somando os cubos dos desvios da média, e então, dividindo desvio padrão.

$$\frac{1}{n}\sum\left[\frac{X_i - \overline{X}}{s}\right]^3$$

# Curtose
# Kurtosis

- Mede o alongamento/achatamento do histograma. A fórmula é semelhante à da assimetria, com a ressalva de que a quarta potência é usada em vez da terceira:

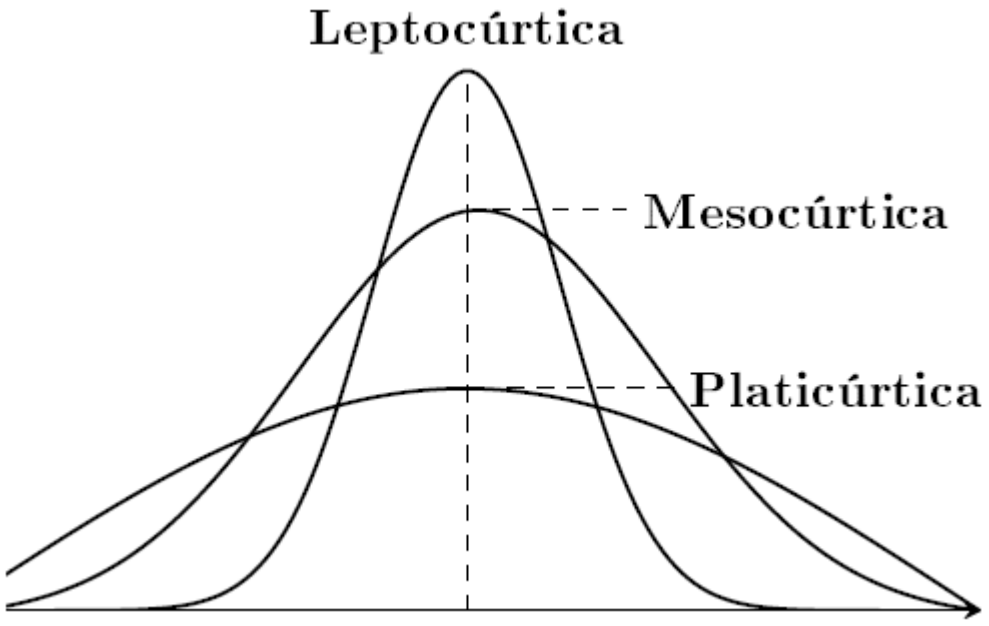
$$\frac{1}{n}\sum\left[\frac{x_i - \overline{x}}{s}\right]^4$$

- Soma-se as quartas potências dos desvios da média, e então, divide-se pelo desvio padrão.

# Curtose (Previdelli, 2018)

- De acordo com esta medida temos a seguinte classificação:

- k < 0, Platicúrtica
- k = 0, Mesocúrtica
- k > 0, Leptocúrtica



# Referências


- BUGNI, R. P., JACOB, M. S. 2017. Índice de vulnerabilidade social: uma análise da cidade de São Paulo. Disponível em: http://www.ipea.gov.br/agencia/images/stories/PDFs/livros/livros/170828_livro_territorios_numeros_insumos_politicas_publicas_2_cap04.pdf
- GALVANI, E. Estatística descritiva em sala de aula. In: VENTURI, L. A. B. Geografia: Práticas de campo, laboratório e sala de aula. São Paulo: Editora Sarandi, 2011.
- GUIMARÃES, I. A. Estatística I (Notas de aulas). 1997. Disponível em: http://www.cin.ufpe.br/~rosf/public_html/Notas%20de%20Aula%20de%20Estat%EDstica%20I.doc
- MARTINELLI, M. Gráficos e Mapas: Construa-os Você Mesmo. São Paulo: Moderna, 1998. 120 p.
- PREVIDELLI, I. Bioestatística, 2018. Disponível em: https://biostatistics-uem.github.io/Bio/descritiva.html
- ROGERSON, P. A. Métodos estatísticos para Geografia: um guia para o estudante. Porto Alegre: Bookman, 2012.

# Exercício prático

- Fazer o download dos arquivos do Dropbox pelo link encurtador.com.br/kwAM0 ou pelo QR code.

- Salvar os arquivos baixados numa pasta de localização conhecida.

# Exercício prático

- Abrir o Past.
- No menu Past3, selecionar open, o diretório onde os arquivos foram salvos, Dados_V_Andrade.xls e abrir.
- Em import setting manter rows contain Names, data, Columns contain Names, data e ok.

# Exercício prático

- Clicar na coluna arborização para seleciona-la.
- No menu Univariate, selecionar Summary statistics.
- Repetir o procedimento com as demais colunas para responder as questões propostas na terceira página do exercício.