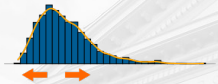
# Medidas de dispersão, forma e associação

Prof. Walmes M. Zeviani

Departamento de Estatística
Universidade Federal do Paraná

# Conteúdo

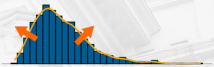
## Medidas de dispersão



Expressam:
- → O domínio observado da variável.
- → O grau de dispersão ao redor do centro.
- → O distanciamento médio dos valores.

São elas:
- → Amplitude total.
- → Variância.
- → Desvio-padrão.
- → Desvio absoluto médio e mediano.
- → Coeficiente de variação.

## Medidas de forma



Expressam:
- → Aspectos da forma.
- → Assimetria.
- → Curtose.

São elas:
- → Coeficiente de assimetria.
- → Coeficiente de curtose.

Figura 1. Medidas de dispersão e forma usadas em análise descritiva de dados.

# Medidas de dispersão

# A importância de quantificar a dispersão

- O resumo de variável observada apenas por uma medida de posição, **ignora** a informação sobre a sua variabilidade.
- Não é seguro analisar um conjunto de dados somente pelo emprego de medidas de tendência central.
- Por isso, precisamos de medidas que caracterizem a **dispersão** ou **variabilidade** dos dados em relação a um valor central.

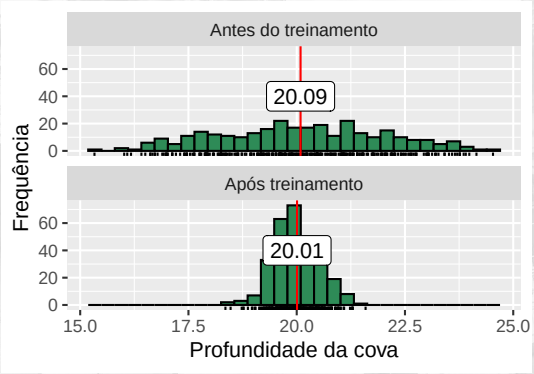



Figura 2. Histogramas exibindo a profundidade das covas para transplante de mudas antes e após ser dado treinamento sobre cultivo.

# Amplitude total

- A **amplitude** é a diferença entre o maior e o menor valor da variável:

$$A = \max(y) - \min(y) = y_{(n)} - y_{(1)}.$$

- A notação $y_{(k)}$ refere-se a **estatística de ordem**, ou seja a observação que está na $k$-ésima posição na amostra com valores ordenados de forma crescente.
- A amplitude está expressa na mesma unidade de medida da variável.
- **Apenas** usar máximo e mínimo torna **sensível** a valores extremos.
  - Melhor medida de variabilidade: considerar **todos os dados disponíveis**
  - **Desvio** de cada valor em relação à uma medida de posição central (média ou mediana).

# Desvio médio e mediano

**Desvio absoluto médio da mediana** (desvio da mediana)

- Usa a **mediana** como medida de posição central. É definido por

$$\text{desvio mediano} = \frac{1}{n}\sum_{i=1}^{n} \text{abs}(y_i - md),$$

em que abs(.) é a função que retorna o valor absoluto ou módulo. Assim, abs($y$) é o mesmo que $|y|$.

**Desvio absoluto médio da média** (desvio da média)

- Usa a **média** como medida de posição central. É defido por

$$\text{desvio médio} = \frac{1}{n}\sum_{i=1}^{n} \text{abs}(y_i - \overline{y}).$$

# A variância

- Uma alternativa melhor é usar a **soma dos quadrados dos desvios**, que dá origem à **variância** de um conjunto de dados.
- A variância é definida por

$$s^2 = \text{Var}(y) = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2$$
$$= \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n}\right)$$

- A unidade de medida do valor da variância é a unidade de medida da variável ao quadrado. Ou seja, se a variável $y$ é a altura em cm dos estudantes, então a variância será $\text{cm}^2$.
- A segunda expressão é considerada mais eficiente em termos de operações algébricas. Ela não requer calcular a média.

# Variância amostral e variância populacional

- A explicação para o denominador $n-1$ será dada na parte de estimação. A variância calculada assim é chamada de **variância amostral** ($s^2$).
- A variância quando usa apenas $n$ no demominador é chamada de **variância populacional** ($\sigma^2$) e é aplicada quando se observam todos os elementos da população.
- Sempre considere o cálculo da variância amostral a menos que seja expressamente dito para calcular a variância populacional.
- É fácil converter de um para o outro, pois

$$s^2 = \sigma^2 \left( \frac{n}{n-1} \right).$$

- **Importante**: A variância é **sempre positiva**.

# O cálculo da variância

Considere os seguintes valores

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 4 | 8 | 9 | 10 | 10 | 11 | 12 | 13 | 15 | 18 |
| 7 | 8 | 10 | 10 | 11 | 12 | 12 | 14 | 15 | 24 |

e calcule a variância amostral sabendo que $\overline{y} = 11.65$.

Aplicando a fórmula, obtém-se

$$\begin{aligned} s^2 &= \frac{1}{19}\left[(4-11.65)^2 + (7-11.65)^2 + \cdots + (24-11.65)^2\right] \\ &= \frac{1}{19}\left[58.5225 + 21.6225 + \cdots + 152.5225\right] \\ &= 18.34. \end{aligned}$$

# O desvio-padrão

- Para ter uma medida de dispersão com a mesma unidade de medida dos dados originais, definiu-se o **desvio-padrão** como

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

# A Lei de Chebyshev

- A Lei Chebyshev estabelece a **proporção mínima dos valores** contidos em **intervalos simétricos** em relação à média.
- Tais resultados valem seja qual for a forma da distribuição.
- **Pelo menos** $3/4$ (75%) dos valores estão no intervalo $(\overline{y} - 2s, \overline{y} + 2s)$.
- **Pelo menos** $8/9$ (89%) dos valores estão no intervalo $(\overline{y} - 3s, \overline{y} + 3s)$.
- Formula geral: **pelos menos** $(1 - 1/k^2)$ dos dados estará no intervalo $(\overline{y} - ks, \overline{y} + ks)$.

# O coeficiente de variação

- O coeficiente de variação é uma medida de variabilidade relativa à média.
- É definido pelo quociente do desvio-padrão pela média, ou seja,

$$\text{CV} = 100 \cdot \frac{s}{\overline{y}}.$$

- É uma medida **adimensional**, e geralmente apresentada na forma de porcentagem, como indica a expressão.

# Cálculo do desvio-padrão e coeficiente de variação

O desvio-padrão para os dados já apresentados em slides anteriores é

$$s = \sqrt{18.34} = 4.283.$$

O coeficiente de variação é

$$\text{CV} = 100 \cdot \frac{4.283}{11.65} = 36.765\%.$$

# Quando usar cada medida de dispersão

- **Amplitude**:
  - Fácil de calcular.
  - Influenciado por valores extremos.
- **Desvios absolutos**:
  - São medidas robustas, ou seja, mais resilientes a *out-liers*.
  - Dá ideia do tamanho médio dos desvios.
- **Variância** ou **desvio-padrão**:
  - Influenciados por valores extremos.
  - Ainda assim, a Lei de Chebyshev é útil para determinar proporções dentro de intervalos simétricos.
  - Têm boas propriedades e significado que serão vistas na parte de Estimação e Inferência.
- **Coeficiente de variação**.
  - Comparar a variabilidade de variáveis de diferentes naturezas.

# Medidas de dispersão para variáveis qualitativas

- Existem várias métricas ou índices para representar a dispersão em variáveis qualitativas.
- Os índices são funções das frequências das classes.
- A **entropia** (de Shannon) é definida por

$$H = \sum_{i=1}^{k} p_i \log(1/p_i) = \sum_{i=1}^{k} p_i(-\log p_i) = -\sum_{i=1}^{k} p_i \log(p_i),$$

em que $p_i = f_{r_i}$ é a frequência relativa da classe $i$ ($i = 1, \ldots, k$).

- Quanto mais próximo $H$ estiver de 0, mais concentrada é a distribuição de frequências.
- Para mais sobre o assunto, procure sobre **índices de diversidade**.
- São usados em ecologia para caracterizar a biodiversidade.

# Medidas de forma

# Medidas de forma

- Servem para descrever características adicionais da distribuição.
  - Coeficiente de **assimetria**.
  - Coeficiente de **curtose**.
- Calculados com a variável **padronizada** pela média e desvio-padrão

$$z = \frac{y_i - \overline{y}}{s}, \text{ que resulta em } \overline{z} = 0 \text{ e } s_z = 1.$$

- São baseados em **momentos** de ordem $k$ superior a 2

$$m_k = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{y_i - \overline{y}}{s} \right)^k.$$

# Coeficiente de assimetria

- Indica um 3º aspecto da forma da distribuição: **a assimetria**.
- É a média do cubo dos desvios, ou seja

$$b_1 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \overline{y}}{s}\right)^3 .$$

- Assimetria **à esquerda** quando $b_1 < 0$ e assimetria **à direita** quando $b_1 > 0$.

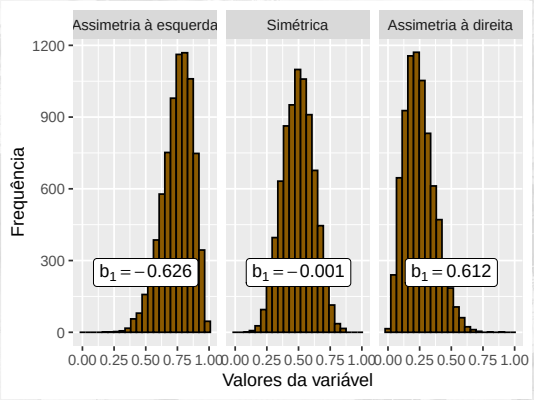



Figura 3. Histogramas com distribuições de diferentes assimetrias indicando o valor do coeficiente de assimetria.

# Coeficiente de curtose

- Indica um 4º aspecto da forma da distribuição: **a curtose**.
- É definido por

$$b_2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \overline{y}}{s}\right)^4 - 3,$$

em que a subtração por 3 serve para usar a distribuição normal como referência.

- **Platicúrtica** quando $b_2 < 0$, **mesocúrtica** quando $b_2 = 0$ e **leptocúrtica** quando $b_2 > 0$.

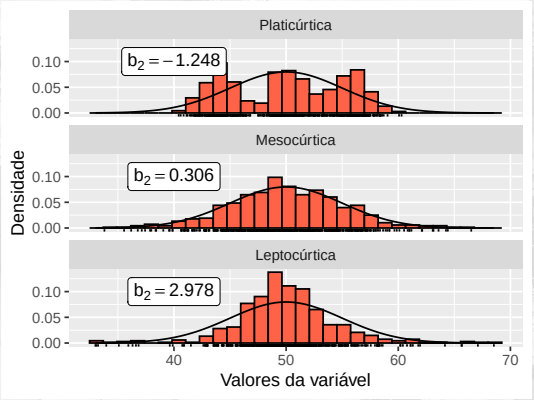



Figura 4. Histogramas com distribuições de diferentes curtoses indicando o valor do coeficiente de curtose comparada com a distribuição normal representada pela linha contínua.

# Correlação de Pearson

# Exemplos de grau de correlação

- É usado para determinar se existe relação linear entre v.a. **quantitativas**.
- A correlação $r$ assume valores entre $-1$ e 1.
  - Quando $r > 0$, então existe uma associação (linear) **positiva**.
  - Quando $r < 0$, então existe uma associação (linear) **negativa**.
  - Quando $r = 0$, então **não existe** uma associação (linear).

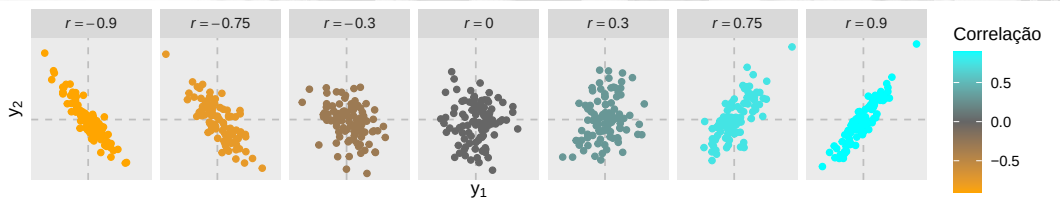



Figura 5. Correlação entre duas variáveis quantitativas.

# Covariância e correlação

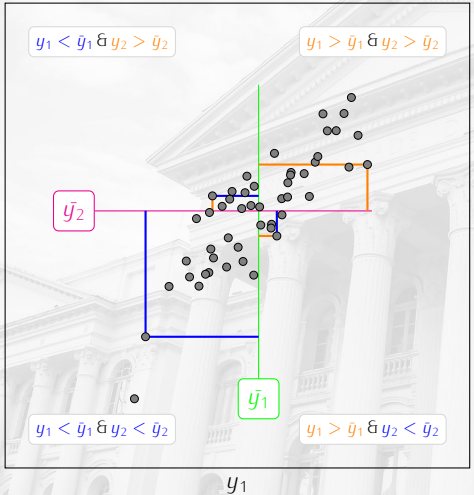
- A **covariância** amostral entre duas variáveis $Y_1$ e $Y_2$ é

$$\text{Cov}(y_1, y_2) = \frac{1}{n-1} \sum_{i=1}^{n} (y_{1i} - \overline{y}_1) \cdot (y_{2i} - \overline{y}_2).$$

- A **correlação** amostral entre duas variáveis $Y_1$ e $Y_2$ é

$$r = \frac{\sum_{i=1}^{n} (y_{1i} - \overline{y}_1) \cdot (y_{2i} - \overline{y}_2)}{\sqrt{\sum_{i=1}^{n} (y_{1i} - \overline{y}_1)^2} \cdot \sqrt{\sum_{i=1}^{n} (y_{2i} - \overline{y}_2)^2}} = \frac{\text{Cov}(y_1, y_2)}{\sqrt{\text{V}(y_1) \cdot \text{V}(y_2)}}.$$

# Interpretação gráfica




O coeficiente de correlação é

$$r = \frac{\sum_{i=1}^{n}(y_{1i} - \overline{y}_1) \cdot (y_{2i} - \overline{y}_2)}{\sqrt{\sum_{i=1}^{n}(y_{1i} - \overline{y}_1)^2} \cdot \sqrt{\sum_{i=1}^{n}(y_{2i} - \overline{y}_2)^2}}.$$

Figura 6. A interpretação do coeficiente de correlação de Pearson.

# Exemplo: comprimento radicular e produtividade

Tabela 1. Valores de produtividade e comprimento de raízes de plantas de milho.

| Comp. | Prod. | Comp. | Prod. | Comp. | Prod. |
|---|---|---|---|---|---|
| 2.85 | 0.74 | 3.08 | 0.84 | 2.12 | 0.65 |
| 3.13 | 0.93 | 3.85 | 0.86 | 3.13 | 0.88 |
| 3.86 | 0.91 | 2.05 | 0.72 | 3.55 | 0.79 |
| 2.40 | 0.76 | 2.81 | 0.83 | 2.88 | 0.82 |
| 2.74 | 0.72 | 2.83 | 0.70 | 3.49 | 0.92 |
| 3.25 | 0.92 | 2.58 | 0.67 | 3.39 | 0.91 |

$$\text{Cov}(y_1, y_2) = 0.0369, \quad s_1^2 = 0.2731, \quad s_2^2 = 0.0087.$$

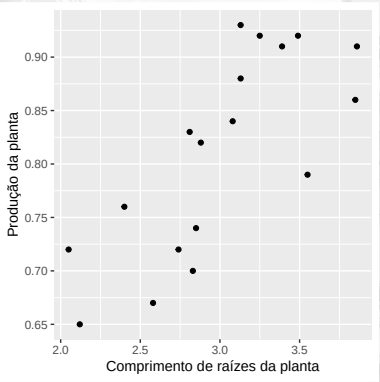



Figura 7. Diagrama de dispersão entre comprimento de raízes ($y_1$) e produção ($y_2$).

# Solução

A correlação é obtida por

$$r = \frac{0.0369}{\sqrt{0.2731 \cdot 0.0087}} = 0.7555,$$

que indica uma **associação positiva** entre as variáveis.

# Outros tipos de correlação

- A correlação de Pearson descreve o grau de associação **linear** entre variáveis.
- Associações diferentes da linear são descritas **impropriamente** pelo coeficiente de correlação de Pearson.
- Existem outros tipos de correlação.
  - Correlação de Spearman.
  - Correlação de Kendall.
- Teste de hipótese para a correlação será visto na parte de Inferência Estatística.

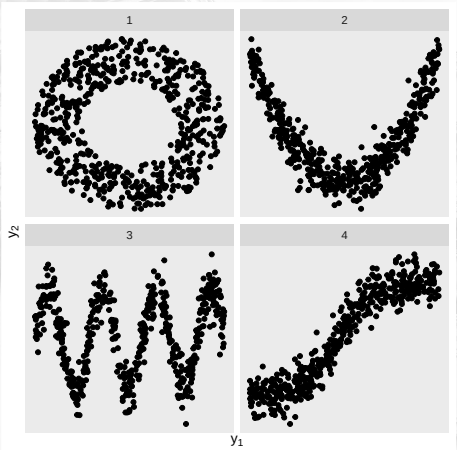



Figura 8. Tipos de associação não lineares entre variáveis.

# Mais medidas de associação

**Tipo das variáveis e medida de associação**

1. Numérica $\times$ numérica $\rightarrow$ Coeficiente de correlação de Pearson.
2. Numérica $\times$ ordinal $\rightarrow$ Coeficiente de correlação de Kendall.
3. Numérica $\times$ nominal $\rightarrow$ Coeficiente de correlação ponto-bisserial.
4. Ordinal $\times$ ordinal $\rightarrow$ Coeficiente de correlação de Kendall.
5. Ordinal $\times$ nominal $\rightarrow$ Coeficiente de correlação rank-bisserial.
6. Nominal $\times$ nominal $\rightarrow$ Coeficiente $\phi$.

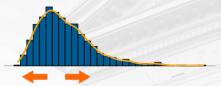
https://journals.sagepub.com/doi/pdf/10.1177/8756479308317006

# Considerações finais

# Considerações finais

## Revisão

### Medidas de dispersão



Expressam:
- → O domínio observado da variável.
- → O grau de dispersão ao redor do centro.
- → O distanciamento médio dos valores.

São elas:
- → Amplitude total.
- → Variância.
- → Desvio-padrão.
- → Desvio absoluto médio e mediano.
- → Coeficiente de variação.

### Medidas de forma

Expressam:
- → Aspectos da forma.
- → Assimetria.
- → Curtose.

São elas:
- → Coeficiente de assimetria.
- → Coeficiente de curtose.

Figura 9. Medidas de dispersão e forma usadas em análise descritiva de dados.