Teste para a correlação

Prof. Walmes Marques Zeviani

Departamento de Estatística
Universidade Federal do Paraná

Conteúdo

Neste vídeo

- Associação vs causalidade.
- Teste de correlação de Pearson.

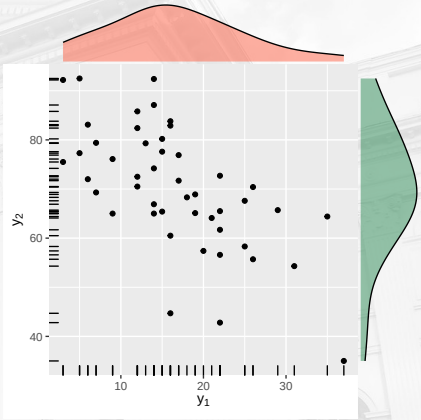


Figura 1. Diagrama de dispersão entre um par de variáveis.

Testes de correlação

- Visam determinar a **existência de associação** entre duas ou mais variáveis.
- Para **um par** de variáveis **qualitativas** (ou categóricas), pode-se usar o teste de qui-quadrado para independência.
- Para **um par** de variáveis **quantitativas** pode-se usar o teste de correlação de Pearson.
- As suposições precisam ser atendidas.
- **Associação não implica causalidade**.

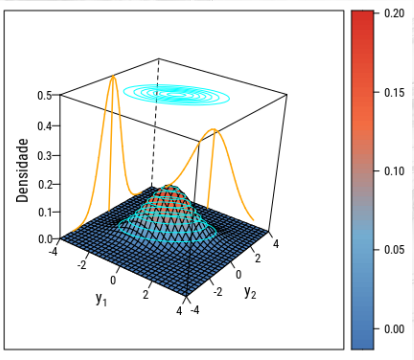


Figura 2. Exemplo da função de densidade conjunta de duas variáveis aleatórias contínuas.

Associação não implica causalidade

Figura 3. Exemplo de que associação não é causa.

Figura 4. Foto de Guy Kawasaki no Pexels.

Teste de Correlação de Pearson

Motivação

- É usado para determinar se existe relação linear entre v.a. quantitativas.
- Assume que

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} \sim \mathrm{N}_2 \left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \right).$$

- Isto é, que Y_1 e Y_2 tenham distribuição Normal Bivariada.
- O teste é para o parâmetro de correlação ρ: $-1 < \rho < 1$.
- Quando $\rho = 0$, as variáveis Y_1 e Y_2 são independentes.

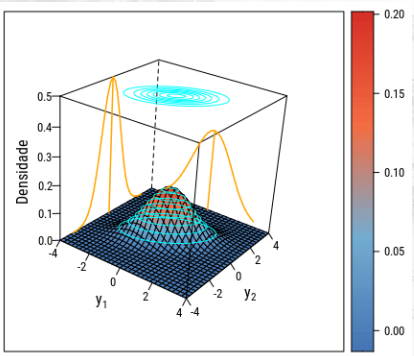


Figura 5. Exemplo da função de densidade conjunta de duas variáveis aleatórias contínuas.

Exemplos de grau de correlação

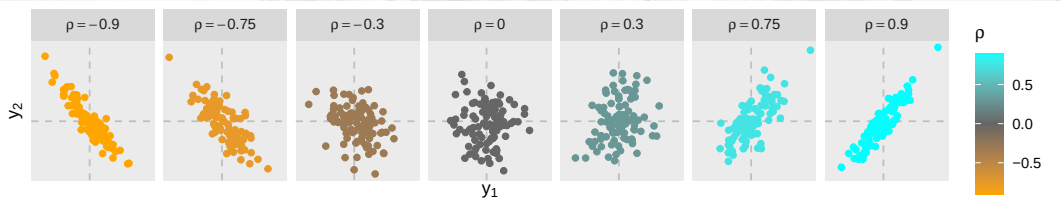


Figura 6. Correlação entre duas variáveis quantitativas.

Covariância e correlação

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

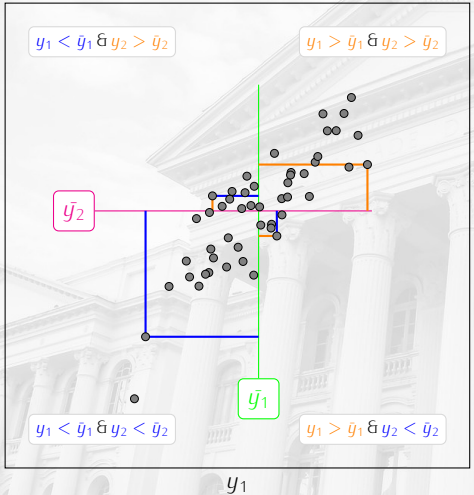
$$\text{Cov}(y_1, y_2) = \frac{1}{n-1} \sum_{i=1}^{n} (y_{1i} - \overline{y}_1) \cdot (y_{2i} - \overline{y}_2).$$

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

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

- 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).

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 7. A interpretação do coeficiente de correlação de Pearson.

O teste de correlação de Pearson

- Sejam as **hipóteses** nula e alternativa

$$H_0 : \rho = 0.$$
$$H_a : \rho \neq 0 \quad \text{ou} \quad H_a : \rho < 0 \quad \text{ou} \quad H_a : \rho > 0.$$

- A **estatística de teste** é

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \overset{\text{aprox}}{\sim} t_{n-2} \quad \text{apenas quando } H_0 : \rho = 0.$$

- Rejeita-se a hipótese nula H_0 ao nível de significância α se

$$\text{abs}(t) > t_{\alpha/2} \quad \text{quando o teste é bilateral}$$
$$\text{abs}(t) > t_{\alpha} \quad \text{quando o teste é unilateral,}$$

em que $t_{\alpha/2}$ e t_{α} são os **quantis superiores** da distribuição t-Student com $(n-2)$ graus de liberdade.

Exercício: comprimento radicular e produtividade

- Pesquisadores acreditam que, em regiões de falta de chuva, a produtividade do milho está relacionada ao comprimento das suas raízes.
- Aplique o teste de correlação para a hipótese $H_0 : \rho = 0$ ao nível de 10% de significância. Os dados estão na tabela a seguir.

Figura 8. Foto de Balázs Benjamin no Pexels.

Exercício (cont.)

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

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

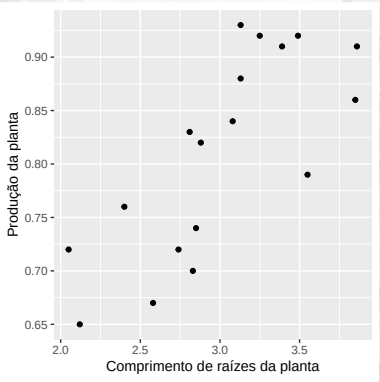


Figura 9. Diagrama de dispersão entre comprimento de raízes (y_1) e produção (y_2).

Solução

Será considerada como hipótese alternativa $H_a : \rho \neq 0$.

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.

A estatística do teste, por sua vez, é

$$t = \frac{0.7555\sqrt{18-2}}{\sqrt{1-0.7555^2}} = 4.6127.$$

O quantil da t-Student, que delimita a região de rejeição, é $t_{0.05} = 1.7459$. Portanto, rejeita-se a hipótese nula.

Teste para valor ρ_0 diferente de zero

O procedimento para as hipóteses com valor em julgamento ρ_0, ou seja,

$$H_0 : \rho = \rho_0 \quad \text{vs} \quad H_a : \rho \neq \rho_0 \quad \text{ou} \quad H_a : \rho < \rho_0 \quad \text{ou} \quad H_a : \rho > \rho_0,$$

usa outra estatística de teste.

A estatística é

$$Z = \operatorname{arctanh}(r) = \frac{1}{2} \ln \left(\frac{1+r}{1-r} \right) \overset{\text{aprox}}{\sim} \mathrm{N}(\mu_z, \sigma_z^2),$$

em que

$$\mu_z = \operatorname{arctanh}(\rho) = \frac{1}{2} \ln \left(\frac{1+\rho}{1-\rho} \right) \quad \text{e} \quad \sigma_z^2 = \frac{1}{n-3},$$

sendo que $\tanh(.)$ é a tangente hiperbólica e $\operatorname{arctanh}(.)$ é o arco da tangente hiperbólica ou $\tanh^{-1}(.)$.

Intervalo de confiança para a correlação

Logo, para **testar a hipótese** $H_0 : \rho = \rho_0$, pode-se usar a estatística

$$z = (\arctan(r) - \arctan(\rho))\sqrt{n-3},$$

e rejeitar H_0 se $\text{abs}(z) > z_{\text{crt}}$ para o nível de significância α.

A partir da distribuição amostral acima definida, um **intervalo de confiança** $100(1-\alpha)\%$ (aproximado) para o parâmetro de correlação é obtido por

$$\text{IC}_{1-\alpha}(\rho) = \left(\tanh\left(\text{arctanh}(r) - \frac{z_{\alpha/2}}{\sqrt{n-3}} \right), \tanh\left(\text{arctanh}(r) + \frac{z_{\alpha/2}}{\sqrt{n-3}} \right) \right),$$

em que $z_{\alpha/2}$ é o quantil superior da distribuição Normal Padrão com $\alpha/2$ de área a sua direita.

Atenção: a aproximação Normal para a distribuição amostral de Z é considerada satisfatória para $n \geq 25$.

Aplicação

Para os dados de produção e comprimento de raízes, o intervalo de confiança para a correlação é

$$\begin{aligned} \mathrm{IC}_{1-\alpha}(\rho) &= \tanh\left(\operatorname{arctanh}(r) \pm \frac{z_{\alpha/2}}{\sqrt{n-3}}\right) \\ &= \tanh\left(\operatorname{arctanh}(0.756) \pm \frac{1.96}{\sqrt{18-3}}\right) \\ &= \tanh\left(0.986 \pm 0.5061\right) \\ &= (0.4459, 0.9036)\,, \end{aligned}$$

em que se usou $\pm$ para salvar espaço.

Pela **conexão** entre intervalo de confiança e teste de hipótese, não rejeita-se, por exemplo, a hipótese $H_0 : \rho = \rho_0 = 0.5$ para um teste bilateral ao nível 95%.

Outros tipos de correlação

- Correlação de Pearson assume que as variáveis tem distribuição Normal Bivariada.
- Se isso for satisfeito, então as distribuições marginais também serão Normais.
- Associações diferentes da linear são descritas impropriamente pelo coeficiente de correlação de Pearson.
- Existem outros tipos de correlação.
  - Correlação de Spearman.
  - Correlação de Kendall.

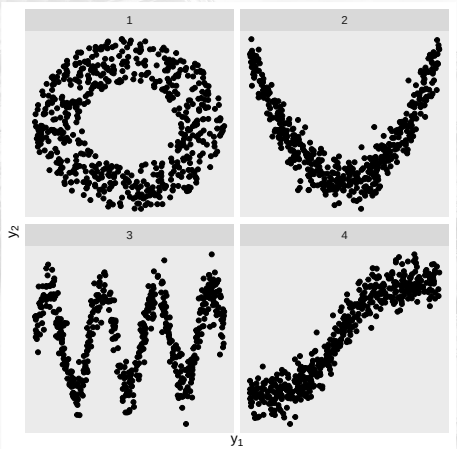


Figura 10. Tipos de associação não lineares entre variáveis.

Considerações finais

Considerações finais

Neste vídeo

- Associação vs causalidade.
- Teste de correlação de Pearson.

Correlação

- Correlação não implica causalidade.
- Ausência de correlação não indica ausência de relação.
- Os pressupostos são importantes.
- Havendo fuga dos pressupostos:
  - Procurar **remediar** via transformação.
  - Procurar um teste mais **adequado** para o caso.

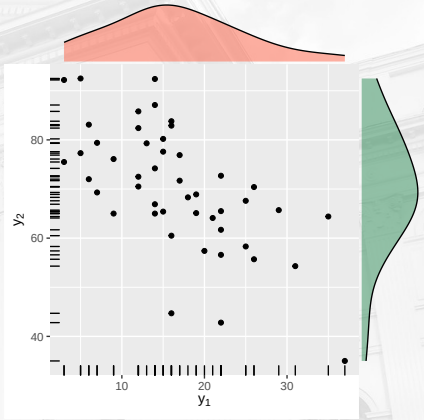


Figura 11. Diagrama de dispersão entre um par de variáveis.