ANADI

Análise de dados em Informática

Aulas T - Testes de Correlação

Ana Madureira, João Matos

Instituto Superior de Engenharia do Porto

Ano letivo 2023/2024

Teste de Correlação Linear de Pearson

- Supor duas variáveis aleatórias contínuas X, Y. O coeficiente de correlação linear de Pearson $r(X, Y) = r$ mede o grau da relação linear entre as duas variáveis X e Y.
- Tem-se,

$$r(X, Y) = r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})\sum_{i=1}^{n}(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

- A estatística teste é:

$$\mathrm{T(X, Y)} = r \times \sqrt{\frac{n-2}{1-r^2}} \sim T_{n-2}$$

- Hipóteses:

$$H_0 : r = 0 \quad \text{versus} \quad \begin{array}{l} r \neq 0 \\ r > 0 \\ r < 0 \end{array}$$

Pressupostos:

1. As variáveis devem ser contínuas e não devem existir outliers significativos
2. Deve existir uma relação linear entre as duas variáveis
3. As variáveis devem ter aproximadamente uma distribuição normal
4. Homocedasticidade (variâncias iguais)

- Tem-se, $-1 \leq r \leq 1$
  - Se r estiver próximo de 1, as variáveis X e Y estão, positivamente, fortemente correlacionadas.
  - Se r estiver próximo de -1, as variáveis X e Y estão, negativamente, fortemente correlacionadas
  - Se $r(x, y) = \pm 1$, então os pontos (x_i, y_i), $i = 1, 2, \ldots, n$ estão todos numa recta com declive positivo (declive negativo)
  - Se r estiver próximo de 0, então variáveis X e Y estão fracamente correlacionadas

☞ *Não indica relação de causalidade*

Link1 Link2

Exemplo:

Consideremos a variável X, "horas de estudo" e a variável Y "Nota do aluno". Pretende-se saber se existe associação entre estas duas variáveis supondo os dados da seguinte tabela.

Aluno	1	2	3	4	5	6
Horas de estudo	6	2	1	5	3	2
Nota do aluno	82	63	57	88	68	75

```
from scipy.stats import pearsonr
x=[6,2,1,5,3,2]
y=[82,63,57,88,68,75]
coef_corr_pearson, p_value = pearsonr(x, y)
print('Coeficiente de correlacao de Pearson:',coef_corr_pearson)
print('Valor de prova do teste:',p_value)
Coeficiente de correlacao de Pearson: 0.8601963041027428
Valor de prova do teste: 0.027951373331788983
```

Conclusão: A medida de associação entre as variáveis é alta ($r = 0.86$) e o coeficiente de correlação r é significativo (p.value < 0.05). Também se pode dizer que 74% da variabilidade das notas é explicada pelo número de horas de estudo, dado que, **coeficiente de determinação** $r^2 = 0.74$

Teste de Correlação Ordinal de Spearman

- Este teste é usado quando as variáveis são ambas ordinais ou quando uma das variáveis é contínua e a outra é ordinal
- Para se obter a estatística atribuímos a cada valor observado x_i e y_i, $i = 1, 2, \ldots, n$ os números de ordem $R(x_i)$ e $R(y_i)$ e calculamos as diferenças $d_i = R(x_i) - R(y_i)$ (Em caso de empates procede-se da mesmo modo que no teste de Wilkoxon)
- Caso não haja empates, o coeficiente de correlação ordinal de Spearman, ρ, é dado por

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}$$

- Caso haja empates o cálculo de ρ sofre uma correcção

- Tem-se, $-1 \leq \rho \leq 1$
  - Se ρ estiver próximo de 1, as variáveis X e Y estão, positivamente, fortemente correlacionadas
  - Se ρ estiver próximo de -1, as variáveis X e Y estão, negativamente, fortemente correlacionadas
  - Se ρ estiver próximo de 0, então variáveis X e Y estão fracamente correlacionadas
- Tem-se as hipóteses

$$H_0 : \rho = 0 \quad \text{versus} \quad \begin{array}{l} \rho \neq 0 \\ \rho > 0 \\ \rho < 0 \end{array}$$

- A estatística teste é conhecida e a sua distribuição também (no caso de haver empates o p-value é calculado apenas de forma aproximada)

Exemplo:

Pediu-se a dois jornalistas, A e B, para classificarem a qualidade dos cafés servidos em 5 estabelecimentos. Verifique se existe associação entre as classificações dos dois jornalistas.

Estabelecimento	1	2	3	4	5
Jornalista A	3	8	7	9	5
Jornalista B	6	7	10	8	4

```
from scipy.stats import spearmanr
A=[3,8,7,9,5]
B=[6,7,10,8,4]
rho, p_value = spearmanr(A, B)
print('Coeficiente de correlacao de Spearman :',rho)
print('P-value do teste de Spearman:' ,p_value)
Coeficiente de correlacao de Spearman : 0.6
P-value do teste de Spearman: 0.28475697986529375
```

Conclusão: Existe uma fraca associação positiva entre as classificações ($\rho = 0.6$). Contudo esta associação não é significativa (p-value =0.28).

Teste de Correlação Ordinal de Kendall

- O coeficiente de correlação de Kendall (τ) é uma alternativa ao coeficiente de correlação de Spearman (ρ) especialmente quando as amostras são pequenas e/ou há muitos empates.
- O coeficiente de correlação de Kendall baseia-se no número de pares de observações concordantes (consistentes) e no número de pares de observações discordantes (inconsistentes)
- Dados dois pares ((x_1, y_1) e (x_2, y_2)) de observações dizemos que são:
  - Concordantes: se $x_2 - x_1$ e $y_2 - y_1$ tiverem o mesmo sinal
  - Discordantes: se $x_2 - x_1$ e $y_2 - y_1$ tiverem sinais diferentes
  - Empatados: $x_2 - x_1 = 0$ ou $y_2 - y_1 = 0$
- Tem-se

$$\tau = \frac{n_c - n_d}{\frac{n(n-1)}{2}}$$

onde n_c é o número de pares concordantes e n_d o número de pares concordantes

- A estatística teste é conhecida e a sua distribuição também
- Tem-se as hipóteses

$$H_0 : \tau = 0 \quad \text{versus} \quad \tau \neq 0 \quad \tau > 0 \quad \tau < 0$$

Exemplo:

Usando o mesmo enunciado do exemplo anterior tem-se:

```
from scipy.stats import kendalltau
A=[3,8,7,9,5]
B=[6,7,10,8,4]
tau, p_value = kendalltau(A, B)
print('Coeficiente de correlacao de Kendall :',tau)
print('P-value do teste de Kendall:' ,p_value)
Coeficiente de correlacao de Kendall : 0.3999999999999997
P-value do teste de Kendall: 0.48333333333333334
```

Conclusão: Existe uma fraca associação positiva entre as classificações ($\tau = 0.4$). Contudo esta associação não é significativa (p-value =0.4833).