

1) Aula 16 - Síncrona (27/09/2022)

Índice de correlação

$$r(x,y) = \frac{1}{n} \sum \left(\frac{x_i - \bar{x}}{dp(x)} \right) \cdot \left(\frac{y_i - \bar{y}}{dp(y)} \right)$$

$$r(x,y) = \frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{\sqrt{\left(\sum_{i=1}^n x_i^2 - n \bar{x}^2 \right) \left(\sum_{i=1}^n y_i^2 - n \bar{y}^2 \right)}}$$

$-1 < \rho < 1$

Tbm temo covariância: ~~população~~

$$COV(x,y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n}$$

$$\Rightarrow \left[\rho(x,y) = \frac{COV(x,y)}{dp(x) dp(y)} \right]$$

- onde:
- COV > 0 (+)
 - COV < 0 (-)
 - COV = 0 (nula) (0)

COV(x,y) → p/ Amostras (n-1)

* Lembre-se que: $Var(x) = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$ } População

σ^2

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

Amostra
graus de liberdade

veja q. p/ fazer COV(x,y) é necessário que q. as duas variáveis tenham o mesmo nº de dados.

* Tbm é afetada por valores extremos

2) Outro exemplo:

Há correlação entre tx de câmbio e retorno do ativo?

seja $C =$ câmbio
 $n =$ retorno do Ativo

anos	Câmbio	retorno
1	3,3	5,3
2	3,4	6,5
3	4,1	7,1

(3,3, 5,3)
(3,4, 6,5)
(4,1, 7,1)

$$\bar{C} = \frac{3,3 + 3,4 + 4,1}{3} = 3,6$$

$$\bar{n} = \frac{5,3 + 6,5 + 7,1}{3} = 6,3$$

$$\text{cov}(C, n) = \frac{\sum (C_i - \bar{C})(n_i - \bar{n})}{(n-1)} \quad \left\{ \text{Amostra} \right.$$

$$\text{cov}(C, n) = \frac{\overset{P_1}{(3,3-3,6) \cdot (5,3-6,3)} + \overset{P_2}{(3,4-3,6) \cdot (6,5-6,3)} + \overset{P_3}{(4,1-3,6) \cdot (7,1-6,3)}}{3-1}$$

$$\text{cov}(C, n) = \frac{(-0,3)(-1) + (-0,2)(0,2) + (0,5)(0,8)}{2} = 0,3 - 0,04 + 0,4 = 0,33$$

$$\text{Var}(C) = \frac{(3,3-3,6)^2 + (3,4-3,6)^2 + (4,1-3,6)^2}{2} = 0,19$$

$$\text{Var}(n) = \frac{(5,3-6,3)^2 + (6,5-6,3)^2 + (7,1-6,3)^2}{2} = 0,84$$

$$\Rightarrow \left. \begin{array}{l} dp(C) = 0,436 \\ dp(n) = 0,917 \end{array} \right\} \text{com } dp(C) = \sqrt{\text{Var}(C)}$$

$$\text{Como } \rho = \frac{\text{cov}(C, n)}{dp(C) dp(n)} = \frac{0,33}{(0,436)(0,917)} = 0,826$$

3) Associação entre variável qualitativa e quantitativa.

⇒ Remuneração (\bar{w}) → Quantitativa
 Escolaridade { ensino fundamental } Qualitativa
 { ensino médio } Ordinal
 { ensino superior }

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

Tabela Remuneração

Nº	Estado civil	Grau de instrução	Nº de filhos	Salário (x sal. mín.)	Idade		Região de procedência
					anos	meses	
1	solteiro	ensino fundamental	—	4,00	26	03	interior
2	casado	ensino fundamental	1	4,56	32	10	capital
3	casado	ensino fundamental	2	5,25	36	05	capital
4	solteiro	ensino médio	—	5,73	20	10	outra
5	solteiro	ensino fundamental	—	6,26	40	07	outra
6	casado	ensino fundamental	0	6,66	28	00	interior
7	solteiro	ensino fundamental	—	6,86	41	00	interior
8	solteiro	ensino fundamental	—	7,39	43	04	capital
9	casado	ensino médio	1	7,59	34	10	capital
10	solteiro	ensino médio	—	7,44	23	06	outra
11	casado	ensino médio	2	8,12	33	06	interior
12	solteiro	ensino fundamental	—	8,46	27	11	capital
13	solteiro	ensino médio	—	8,74	37	05	outra
14	casado	ensino fundamental	3	8,95	44	02	outra
15	casado	ensino médio	0	9,13	30	05	interior
16	solteiro	ensino médio	—	9,35	38	08	outra
17	casado	ensino médio	1	9,77	31	07	capital
18	casado	ensino fundamental	2	9,80	39	07	outra
19	solteiro	superior	—	10,53	25	08	interior
20	solteiro	ensino médio	—	10,76	37	04	interior

Olhando especificamente p/ os variáveis:

Grau de instrução	n	\bar{x}	dp(S)	var(S)	$s_{(1)}$	q_1	q_2	q_3	$s_{(n)}$
Fundamental	12	7,84	2,79	7,77	4,00	6,01	7,13	9,16	13,65
Médio	18	11,54	3,62	13,10	5,73	8,84	10,91	14,48	19,40
Superior	6	16,48	4,11	16,89	10,53	13,65	16,74	18,38	23,30
Todos	36	11,12	4,52	20,46	4,00	7,55	10,17	14,06	23,30

\bar{x} = salário médio ✓
 dp(S) - desvio padrão = $\sqrt{var(S)}$ ✓
 $s_{(1)}$ - Menor salário ✓
 $s_{(n)}$ - Maior salário ✓
 q_2 - Mediana ✓
 q_1 - 1º quartil ✓
 q_3 - 3º quartil ✓
 dado da variável quantitativa //

Por conceito → Educação → Salário, senão? //

4) P/ medir o grau de associação podemos construir uma medida a partir das variâncias

Tempos que verificam a variância global
 → Aquela p. independente do grupo;

$$\text{Var}(s) = \frac{\sum (s_i - \bar{s})^2}{n} = \frac{(4 - 11,12)^2 + (4,56 - 11,12)^2 + \dots}{36}$$

$$= \boxed{20,16} \quad \text{global}$$

Podemos comparar a variância de cada grupo:

$$\text{Var}(S_F) = 7,77$$

$$\text{Var}(S_M) = 13,10$$

$$\text{Var}(S_S) = 16,89$$

Intuitivamente:
 Se $\text{Var}(S_i) < \text{Var}(s)$

⇒ Associadas

∴ veja q. s associadas;

mas pto? → R²
 Precisamos transformar isso em medida:

⇒ Fazemos a Var(s)

↓
 média ponderada das variâncias

$$\overline{\text{Var}(s)} = \frac{\sum_{i=1}^k n_i \text{Var}_i(s)}{\sum_{i=1}^k n_i} \quad \text{com } k=3$$

temos assim:

$$\overline{\text{Var}(s)} = \frac{12(7,77) + 18(13,10) + 6(16,89)}{12+18+6}$$

$\text{Var}(S_F) \xrightarrow{n_F}$ $\text{Var}(S_M) \xrightarrow{n_M}$ $\text{Var}(S_S) \xrightarrow{n_S}$

$$= \boxed{11,96}$$

5) Podemos escrever:

$$R^2 = \frac{\text{Var}(s) - \overline{\text{Var}(s)}}{\text{Var}(s)}$$
 onde $\text{Var}(s) = \text{Var. global}$

$$\Rightarrow R^2 = 1 - \frac{\overline{\text{Var}(s)}}{\text{Var}(s)} = 1 - \frac{11,96}{20,46}$$

$R^2 = 0,415$ → esse valor é alto ou baixo? Vamos ser significados.

R^2 informa o qto da variável quantitativa é explicada pela variável qualitativa.

* No caso, dizemos que 41,5% da variável salário é explicada pelo nível educacional.

SERÁ Que tempo de espera na entrega está relacionado com localização do fornecedor?

ORIGEM	n	\bar{a}	$\text{Var}(a)$
Capital	11	11,46	27,27
Interior	12	11,55	25,71
Outra	13	10,95	9,13
	36	11,12	20,46

Vejamos $\text{Var}(a) = 20,46$ (global)

$\text{Var}(a_C) = 27,27$
 $\text{Var}(a_I) = 25,71$
 $\text{Var}(a_O) = 9,13$

Primeiro teste
 $\text{Var}(a_k) < \text{Var}(a)$
Duas maiores
↳ Precisamos ver o R^2

b) P/ esta relação temos:

$$\text{Var}(a) = 20,46$$

$n_c \rightarrow \text{Var}^{a_c}$ $n_I \rightarrow \text{Var}^{a_I}$ $n_o \rightarrow \text{Var}^{a_o}$

$$\overline{\text{Var}(a)} = \frac{11(27,27) + 12(25,71) + 13(9,13)}{11 + 12 + 13} = \underline{\underline{20,20}}$$

$$\Rightarrow R^2 = \frac{1 - 20,20}{20,46} = 0,013$$

Apenas 1,3% dos atrasos são explicados pela região de procedência do fornecedor.