

ESTATÍSTICA DESCRITIVA

Coeficiente de correlação linear de Pearson

O coeficiente de correlação Linear de Pearson mensura a relação linear entre duas variáveis quantitativas X e Y (ρ_{xy}). Não implica, necessariamente, em causalidade. A relação de causa e efeito entre as variáveis é determinada com estudo detalhados.

O coeficiente ρ_{xy} mede se existe uma relação entre as variações de aumento ou diminuição entre duas variáveis X e Y . Por exemplo, existe alguma relação entre massa e altura? Isto é, pessoas mais altas tendem a ter maior massa.

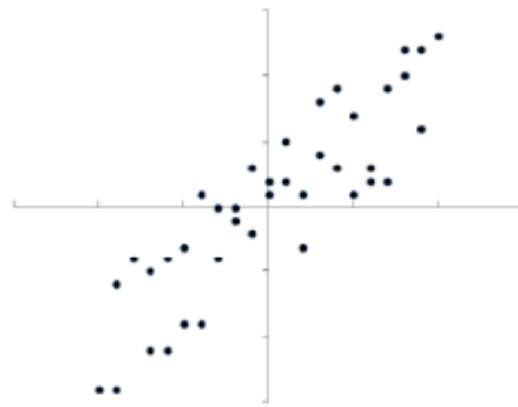
Existe alguma relação entre quantidade de cimento e resistência de um concreto?



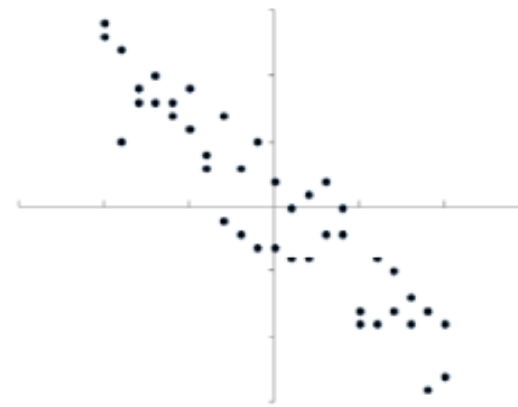
ESTATÍSTICA DESCRITIVA

Coeficiente de correlação linear de Pearson

Para o cálculo, considere $X = \{x_1, x_2, x_3, \dots, x_n\}$ e $Y = \{y_1, y_2, y_3, \dots, y_n\}$ duas variáveis que podem ser organizadas em pares $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$. A medida de correlação é calculada por:



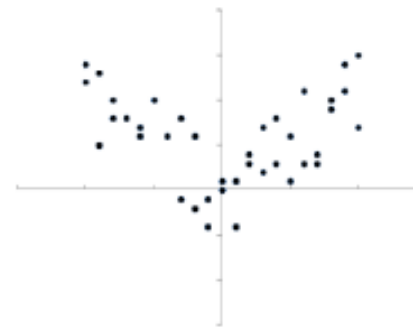
(a)



(b)



(c)



(d)



ESTATÍSTICA DESCRITIVA

Coeficiente de correlação linear de Pearson

$$\rho_{xy} = \frac{COV(x, y)}{\sigma_x \times \sigma_y} = \frac{\left(\frac{\sum_{i=1}^n (y_i x_i)}{n} - \frac{(\sum_{i=1}^n x_i)}{n} \frac{(\sum_{i=1}^n y_i)}{n} \right)}{\sqrt{\frac{\sum_{i=1}^n x_i^2}{n} - \left(\frac{\sum_{i=1}^n x_i}{n} \right)^2} \sqrt{\frac{\sum_{i=1}^n y_i^2}{n} - \left(\frac{\sum_{i=1}^n y_i}{n} \right)^2}}$$

Para calcular a correlação linear de Pearson precisamos calcular os somatórios, somatórios quadráticos e o produto dos somatórios

$$n \quad \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right) \quad \sum_{i=1}^n x_i^2 \quad \sum_{i=1}^n y_i^2 \quad \sum_{i=1}^n (y_i x_i)$$



ESTATÍSTICA DESCRITIVA

Coeficiente de correlação linear de Pearson

O coeficiente de correlação linear de Pearson assume valores entre -1 e 1, podendo ser classificado nesse intervalo de acordo com direção (positiva ou negativa) e intensidade (muito forte, forte, moderada, fraca e muito fraca). É uma medida sensível a valores extremos (*outliers*) e paramétrica que exige que os dados tenham comportamento de normalidade (Distribuição Normal). Falaremos disso nas próximas aulas quando falaremos de distribuições de probabilidade.

Intervalos	Classificação
1.0 a 0.8	Muito forte
0.8 a 0.6	Forte
0.6 a 0.4	Moderada
0.4 a 0.2	fraca
0.2 a 0.0	Muito fraca



ESTATÍSTICA DESCRITIVA

Coeficiente de correlação linear de Pearson

Dados de [UCI](#):

- 1) Resistência de concreto
- 2) Cimento
- 3) Escória
- 4) Água
- 5) Areia grossa
- 6) Areia fina
- 7) Tem de cura

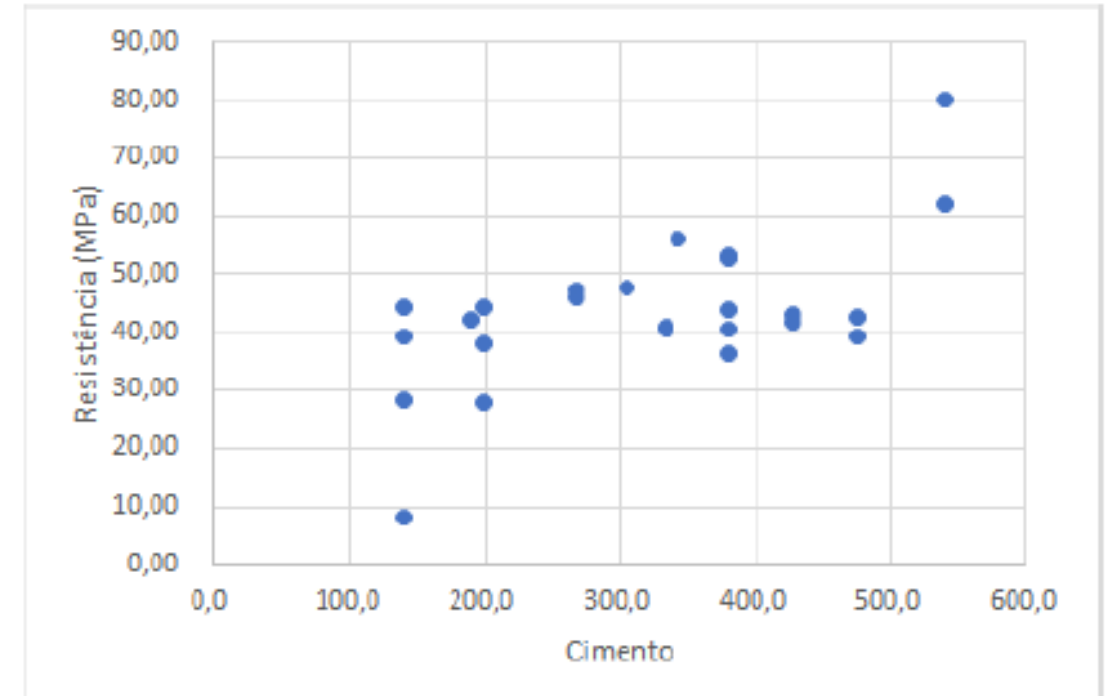
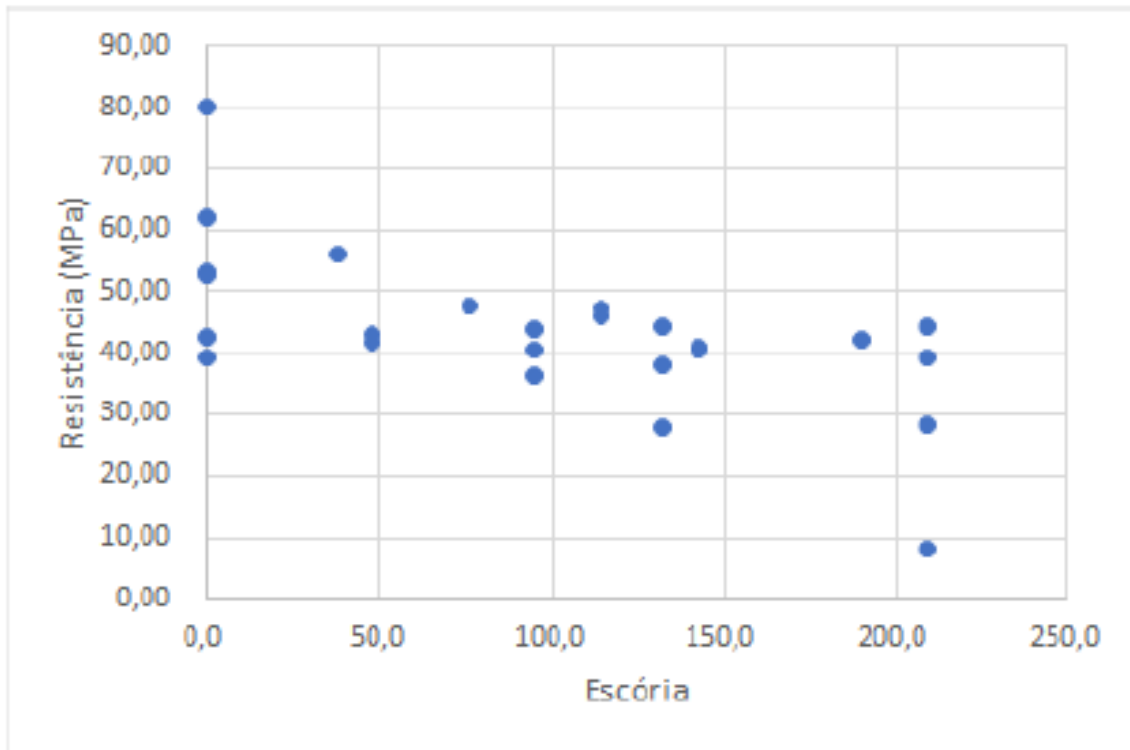
Qual mistura para melhor resistência?

Concrete compressive strength(MPa)	Cement (kg in a m³ mixture)	Blast Furnace Slag(kg in a m³ mixture)	Water (kg in a m³ mixture)	Coarse Aggregate(kg in a m³ mixture)	Fine Aggregate (kg in a m³ mixture)	Age (day)
79,99	540,0	0,0	162,0	1040,0	676,0	28
61,89	540,0	0,0	162,0	1055,0	676,0	28
40,27	332,5	142,5	228,0	932,0	594,0	270
41,05	332,5	142,5	228,0	932,0	594,0	365
44,30	198,6	132,4	192,0	978,4	825,5	360
47,03	266,0	114,0	228,0	932,0	670,0	90
43,70	380,0	95,0	228,0	932,0	594,0	365
36,45	380,0	95,0	228,0	932,0	594,0	28
45,85	266,0	114,0	228,0	932,0	670,0	28
39,29	475,0	0,0	228,0	932,0	594,0	28
38,07	198,6	132,4	192,0	978,4	825,5	90
28,02	198,6	132,4	192,0	978,4	825,5	28
43,01	427,5	47,5	228,0	932,0	594,0	270
42,33	190,0	190,0	228,0	932,0	670,0	90
47,81	304,0	76,0	228,0	932,0	670,0	28
52,91	380,0	0,0	228,0	932,0	670,0	90
39,36	139,6	209,4	192,0	1047,0	806,9	90
56,14	342,0	38,0	228,0	932,0	670,0	365
40,56	380,0	95,0	228,0	932,0	594,0	90
42,62	475,0	0,0	228,0	932,0	594,0	180
41,84	427,5	47,5	228,0	932,0	594,0	180
28,24	139,6	209,4	192,0	1047,0	806,9	28
8,06	139,6	209,4	192,0	1047,0	806,9	3
44,21	139,6	209,4	192,0	1047,0	806,9	180
52,52	380,0	0,0	228,0	932,0	670,0	365



ESTATÍSTICA DESCRITIVA

Coeficiente de correlação linear de Pearson



ESTATÍSTICA DESCRITIVA

Coeficiente de correlação linear de Pearson

	Concrete compressive strength(MPa)	Cement (kg in a m³ mixture)	Blast Furnace Slag(kg in a m³ mixture)	Water (kg in a m³ mixture)	Coarse Aggregate(kg in a m³ mixture)	Fine Aggregate (kg in a m³ mixture)	Age (day)
<u>n</u>	25	25	25	25	25	25	25
<u>Σx</u>	1.085,5	7.972,2	2.431,8	5.316,0	24.130,2	17.092,1	3.667,0
<u>Σx²</u>	51.042,9	2.926.345,5	369.494,7	1.142.280,0	23.349.244,7	11.880.477,2	960.381,0
<u>Σxy</u>	51.042,9	369.905,6	91.200,3	229.844,6	1.046.858,3	733.694,4	165.925,9

$$\rho_{xy} = \frac{\left(\frac{\sum_{i=1}^n (y_i x_i)}{n} - \frac{(\sum_{i=1}^n x_i)}{n} \frac{(\sum_{i=1}^n y_i)}{n} \right)}{\sqrt{\frac{\sum_{i=1}^n x_i^2}{n} - \left(\frac{\sum_{i=1}^n x_i}{n} \right)^2} \sqrt{\frac{\sum_{i=1}^n y_i^2}{n} - \left(\frac{\sum_{i=1}^n y_i}{n} \right)^2}}$$



ESTATÍSTICA DESCRITIVA

Coeficiente de correlação linear de Pearson

Escória e resistência

$$\rho_{xy} = \frac{\left(\frac{\sum_{i=1}^n (y_i x_i)}{n} - \frac{(\sum_{i=1}^n x_i)}{n} \frac{(\sum_{i=1}^n y_i)}{n} \right)}{\sqrt{\frac{\sum_{i=1}^n x_i^2}{n} - \left(\frac{\sum_{i=1}^n x_i}{n} \right)^2} \sqrt{\frac{\sum_{i=1}^n y_i^2}{n} - \left(\frac{\sum_{i=1}^n y_i}{n} \right)^2}} = \frac{\left(\frac{91200,3}{25} - \frac{(1085,5)}{25} \frac{(2431,8)}{25} \right)}{\sqrt{\frac{51042,9}{25} - \left(\frac{1085,5}{25} \right)^2} \sqrt{\frac{369494,7}{25} - \left(\frac{2431,8}{25} \right)^2}} = -0,63$$

Cimento e resistência

$$\rho_{xy} = \frac{\left(\frac{\sum_{i=1}^n (y_i x_i)}{n} - \frac{(\sum_{i=1}^n x_i)}{n} \frac{(\sum_{i=1}^n y_i)}{n} \right)}{\sqrt{\frac{\sum_{i=1}^n x_i^2}{n} - \left(\frac{\sum_{i=1}^n x_i}{n} \right)^2} \sqrt{\frac{\sum_{i=1}^n y_i^2}{n} - \left(\frac{\sum_{i=1}^n y_i}{n} \right)^2}} = \frac{\left(\frac{369905,6}{25} - \frac{(1085,5)}{25} \frac{(7972,2)}{25} \right)}{\sqrt{\frac{51042,9}{25} - \left(\frac{1085,5}{25} \right)^2} \sqrt{\frac{2926345,5}{25} - \left(\frac{7972,2}{25} \right)^2}} = +0,61$$



ESTATÍSTICA DESCRITIVA

Introdução à análise de variância

Também é possível mensurar a associação entre variáveis quantitativas e qualitativas. Para isso, comparamos utilizamos as medidas descritivas de média e variância.

Compreender o processo do cálculo e de interpretação da associação entre variáveis quantitativas e qualitativas te ajudará a compreender, em aulas futuras, o processo de análise de variância (ANOVA), que é amplamente utilizada nos estudos científicos. Um exemplo disso é o estudo de [MACEDO et al., 2015](#)

Para mensurar a relação entre uma variável quantitativa e qualitativa, faz-se [MORETTIN; BUSSAB, 2017](#)



ESTATÍSTICA DESCRITIVA

Introdução à análise de variância

O tempo de cura (qualitativa alto e baixo) é importante para compreender a variável resistência do concreto?

Experimento controlado: avaliar resistência com alto (28 dias) e baixo (3 dias) tempo de cura.

Os dados do experimento estão disponíveis em [Mendeley](#)

Estatística	3D	28D	Todos
n	21	21	42
Σx	3.254,3	8.589,7	11.844,0
Σx^2	558.849,3	3.633.641,9	4.192.491,2
Média	155,0	409,0	282,0
Var	2.597,8	5.721,1	20.297,4
Desvio	51,0	75,6	142,5
Q1	112,6	379,9	140,8
Q2	141,4	425,8	258,8
Q3	172,7	471,8	426,6
AS3	0,042	0,002	0,174
LS	262,9	609,7	855,4
LI	22,4	242,1	-288,0



ESTATÍSTICA DESCRITIVA

Introdução à análise de variância

$$\bar{\sigma}_{1:k}^2 = \frac{\sum_{i=1}^k n_i \times \sigma_i^2}{\sum_{i=1}^k n_i}$$

Sendo σ_i^2 a variância da variável quantitativa para a categoria i e n_i o número de observações para a categoria i . Dessa forma, $\bar{\sigma}_{1:k}^2$ representa uma variância média das categorias. Para realizar a medida associativa compara-se essa variância média com a variância caso os dados não fossem divididos por essas categorias.

A medida de associação é determinada por:

$$R^2 = 1 - \frac{\bar{\sigma}_{1:k}^2}{\sigma^2}$$

Sendo σ^2 a variância total dos dados, juntando as informações de todas as categorias. O valor de R^2 está entre 0 e 1, pois $\bar{\sigma}_{1:k}^2 < \sigma^2$.



ESTATÍSTICA DESCRITIVA

Introdução à análise de variância

O tempo de cura muda o padrão de resistência?

$$\bar{\sigma}_{1:2}^2 = \frac{21 \times 2.597,8 + 21 \times 5.721,1}{21 + 21} = 4.159,5$$

$$R^2 = 1 - \frac{4.159,5}{20.297,4} = 0.79507$$

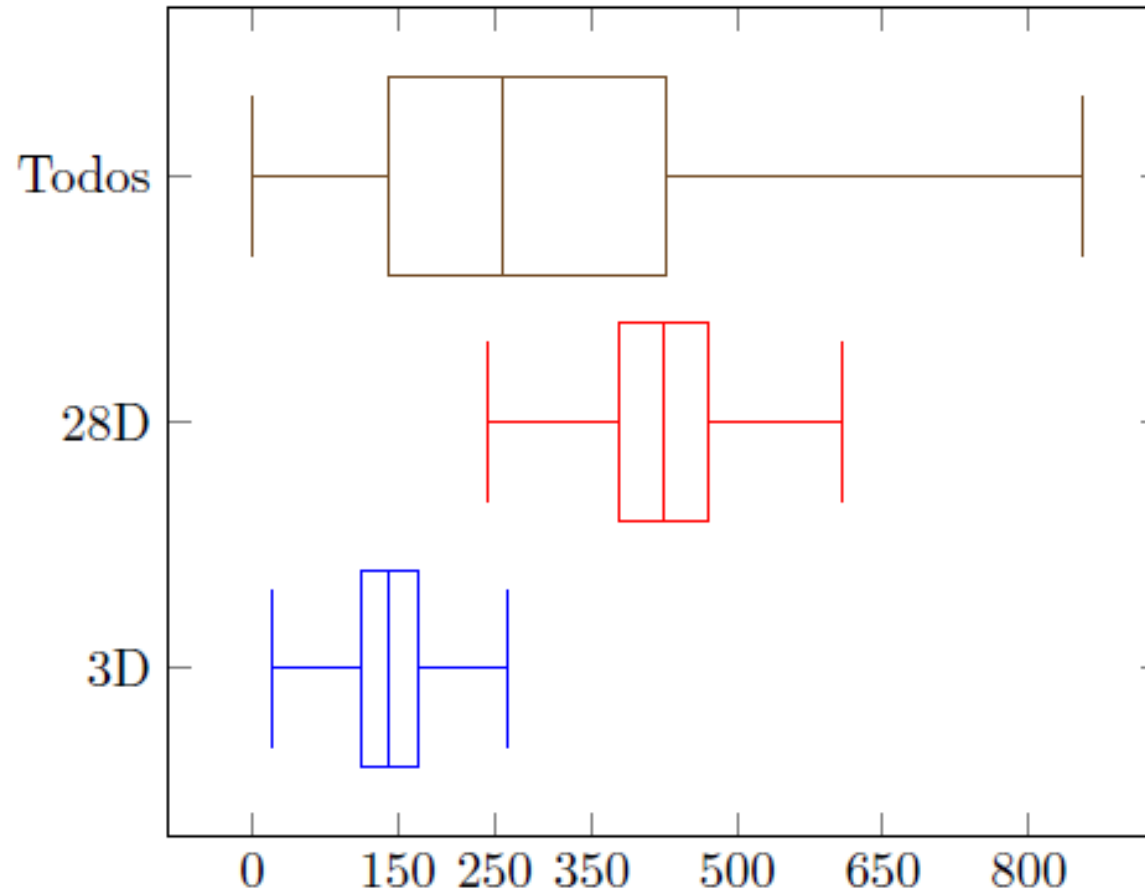
Dizemos que 79.507% da variação total da resistência à compressão é explicada pela variável “tempo de cura”. Isto é, saber o tempo de cura do concreto altera muito a resistência do concreto, pois as variâncias considerando o grupo de observações 3D (2597,8) e 28D (5721,1) separadamente **REDUZEM** drasticamente a variância total



ESTATÍSTICA DESCRITIVA

Introdução à análise de variância

O tempo de cura muda o padrão de resistência?



Referências

FILHO, D. B. F.; JÚNIOR, J. A. d. S. Desvendando os mistérios do coeficiente de correlação de pearson (r). Universidade Federal de Pernambuco, 2009.

JIAO, D. et al. Mixture design of concrete using simplex centroid design method. *Cement and Concrete Composites*, v. 89, p. 76 – 88, 2018. ISSN 0958-9465. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0958946517306467>.

MACEDO, L. d. et al. Propriedades físicas de painéis aglomerados de madeira produzidos com adição de película de polipropileno biorientado. *Revista Brasileira de Engenharia Agrícola e Ambiental*, v. 19, n. 7, p. 674–679, 2015.

MORETTIN, P. A.; BUSSAB, W. O. *Estatística básica*. [S.l.]: Editora Saraiva, 2017.

VELTEN, R. Z. et al. Caracterização mecânica de misturas solo-escória de alto-forno granulada moída para aplicações em estradas florestais. *Revista Árvore*, SciELO Brasil, v. 30, n. 2, p. 235–240, 2006.

YEH, I.-C. *Concrete Compressive Strength Data Set*. 2007. Disponível em: <https://archive.ics.uci.edu/ml/datasets/Concrete+Compressive+Strength>.

