Cruzeiro do Sul Virtual
Educação a Distância

CORRELAÇÃO

Prof. Ismar Frango

Correlação

Considerando as relações lineares entre duas variáveis quantitativas, dizemos que elas são **fortes** se os pontos se aproximam de uma reta imaginária, e **fracas** se os pontos estão bastante espalhados em torno dessa mesma reta. Isso pode ser visualmente verificado nos diagramas de dispersão. Porém, precisamos de medidas precisas do quão forte ou fracas são essas associações.

Partindo de um *dataset* de *n* elementos, com dados para duas variáveis *x* e *y*, a correlação *r* entre essas variáveis é dada por meio do seguinte procedimento:

1. Calcular as médias ($\bar{x}$ e $\bar{y}$) e os desvios-padrões (s_x e s_y)
2. Para todo par de valores *x* e *y*:
   a. calcular a distância entre o valor de *x* e sua média $\bar{x}$, dividindo pelo desvio-padrão s_x: $\left(\frac{x_i-\bar{x}}{s_x}\right)$
   b. fazer o mesmo para *y*: $\left(\frac{y_i-\bar{y}}{s_y}\right)$
   c. Multiplicar um resultado pelo outro: $\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right)$
3. Ao final, somar todos os resultados e dividir por *n-1*

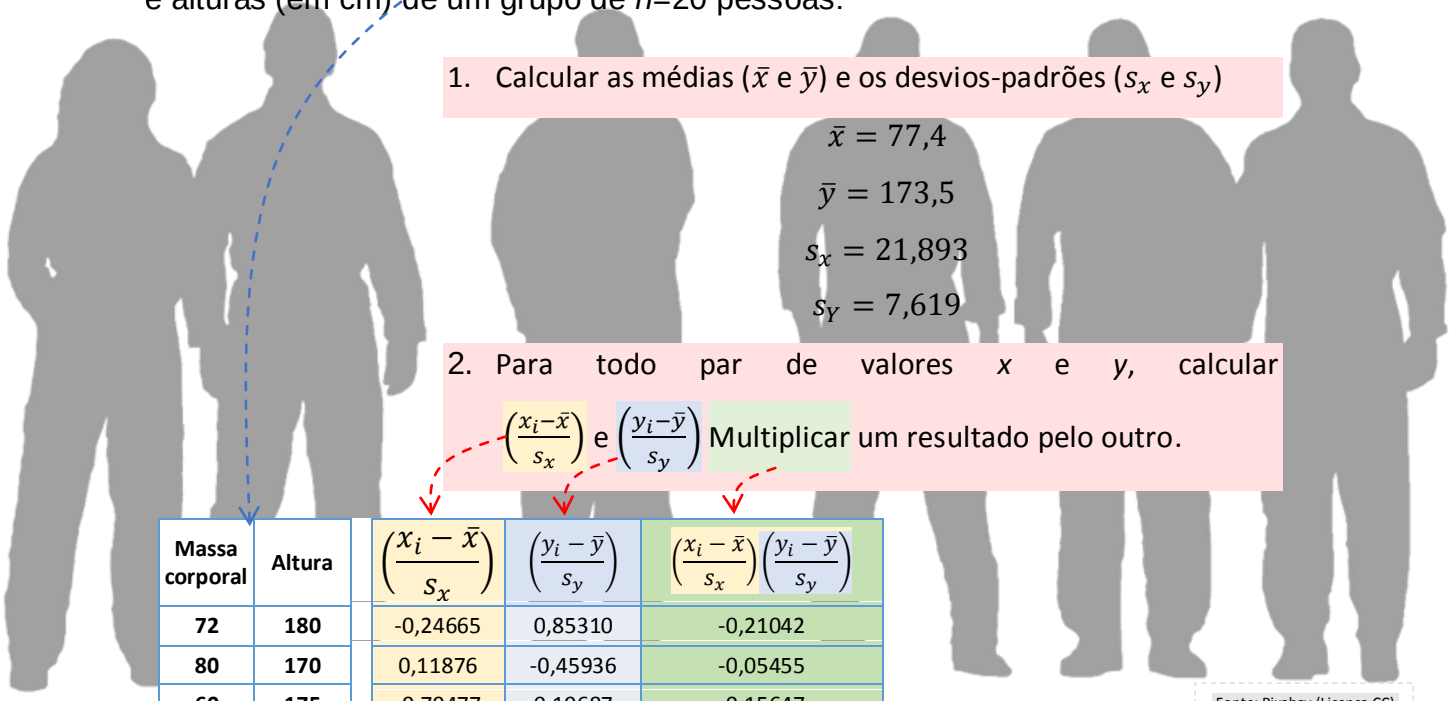
Resumindo em uma fórmula, temos:

$$r = \frac{1}{n-1}\left[\left(\frac{x_1-\bar{x}}{s_x}\right)\left(\frac{y_1-\bar{y}}{s_y}\right) + \left(\frac{x_2-\bar{x}}{s_x}\right)\left(\frac{y_2-\bar{y}}{s_y}\right) + \cdots + \left(\frac{x_n-\bar{x}}{s_x}\right)\left(\frac{y_n-\bar{y}}{s_y}\right)\right]$$

Uma outra maneira, mais resumida, de escrever essa mesma fórmula, é:

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

Vamos retomar um *dataset* já trabalhado em uma outra unidade, para compreender este cálculo. Trata-se do *dataset* com massas corporais (em kg) e alturas (em cm) de um grupo de *n*=20 pessoas:



1. Calcular as médias ($\bar{x}$ e $\bar{y}$) e os desvios-padrões (s_x e s_y)

$$\bar{x} = 77{,}4$$
$$\bar{y} = 173{,}5$$
$$s_x = 21{,}893$$
$$s_Y = 7{,}619$$

2. Para todo par de valores *x* e *y*, calcular $\left(\frac{x_i-\bar{x}}{s_x}\right)$ e $\left(\frac{y_i-\bar{y}}{s_y}\right)$ Multiplicar um resultado pelo outro.

Massa corporal	Altura	$\left(\frac{x_i-\bar{x}}{s_x}\right)$	$\left(\frac{y_i-\bar{y}}{s_y}\right)$	$\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right)$
72	180	-0,24665	0,85310	-0,21042
80	170	0,11876	-0,45936	-0,05455
60	175	-0,79477	0,19687	-0,15647
90	174	0,57553	0,06562	0,03777
100	185	1,03229	1,50934	1,55808
120	190	1,94582	2,16557	4,21382
82	182	0,21011	1,11560	0,23440
79	179	0,07308	0,72186	0,05276
78	165	0,02741	-1,11560	-0,03057
55	165	-1,02316	-1,11560	1,14143
71	170	-0,29233	-0,45936	0,13429
75	169	-0,10962	-0,59061	0,06475
130	177	2,40259	0,45936	1,10366
105	173	1,26067	-0,06562	-0,08273
60	172	-0,79477	-0,19687	0,15647
54	162	-1,06883	-1,50934	1,61323
58	163	-0,88613	-1,37809	1,22116
57	167	-0,93180	-0,85310	0,79493
60	171	-0,79477	-0,32812	0,26078
62	181	-0,70342	0,98435	-0,69241

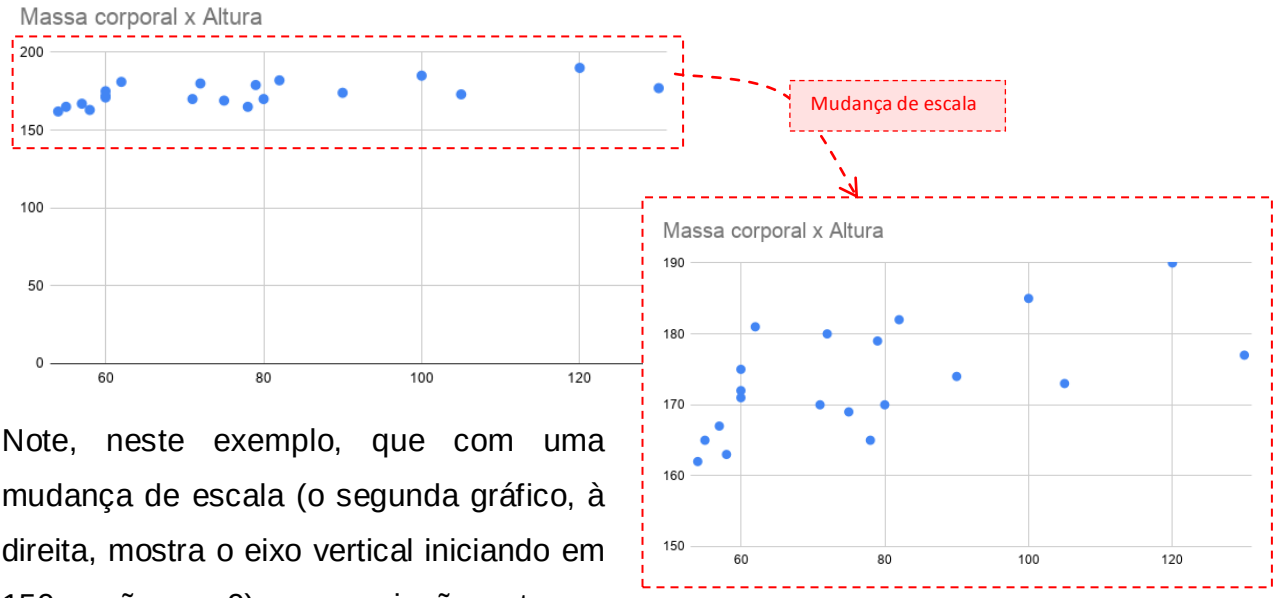
Fonte: Pixabay (Licença CC)

Se for utilizar funções oferecidas por uma planilha de cálculo, lembre-se de usar a função de **desvio-padrão da amostra** (`DESVPAD.A` ou `STDEV.S`).
As planilhas costumam ter uma função (`CORREL`) que calcula diretamente o valor da correlação.

4. Ao final, somar todos os resultados e dividir por *n-1*

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right) = \frac{11{,}36036}{20-1} = \mathbf{0,59791}$$

Vejamos essa relação em um gráfico de dispersão:



Note, neste exemplo, que com uma mudança de escala (o segunda gráfico, à direita, mostra o eixo vertical iniciando em 150 e não em 0), a associação entre as variáveis pode ficar mais clara.

A correlação entre massa corporal e altura, neste *dataset*, é de 0,598 – positiva e não muito forte.

O que isto significa? Veremos a seguir.

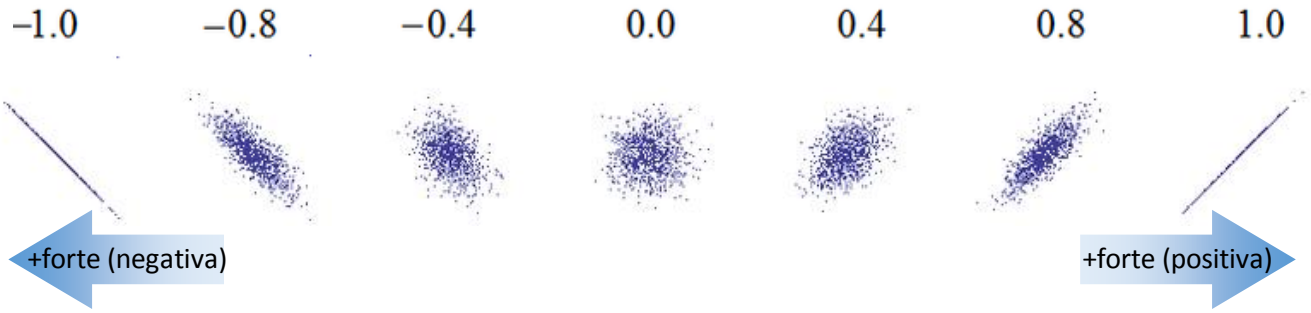
Interpretando as correlações

Os valores de correlação sempre resultam em algo entre -1 e 1, ou seja, para qualquer dataset:

$$-1 \leq r \leq 1$$

Mas o que signficam esses valores?

Veja o exemplo a seguir com os valores de correlação e os gráficos de dispersão obtidos a partir de sete *datasets* diferentes:



Fonte: Wikimedia Commons (Licença CC) + autor

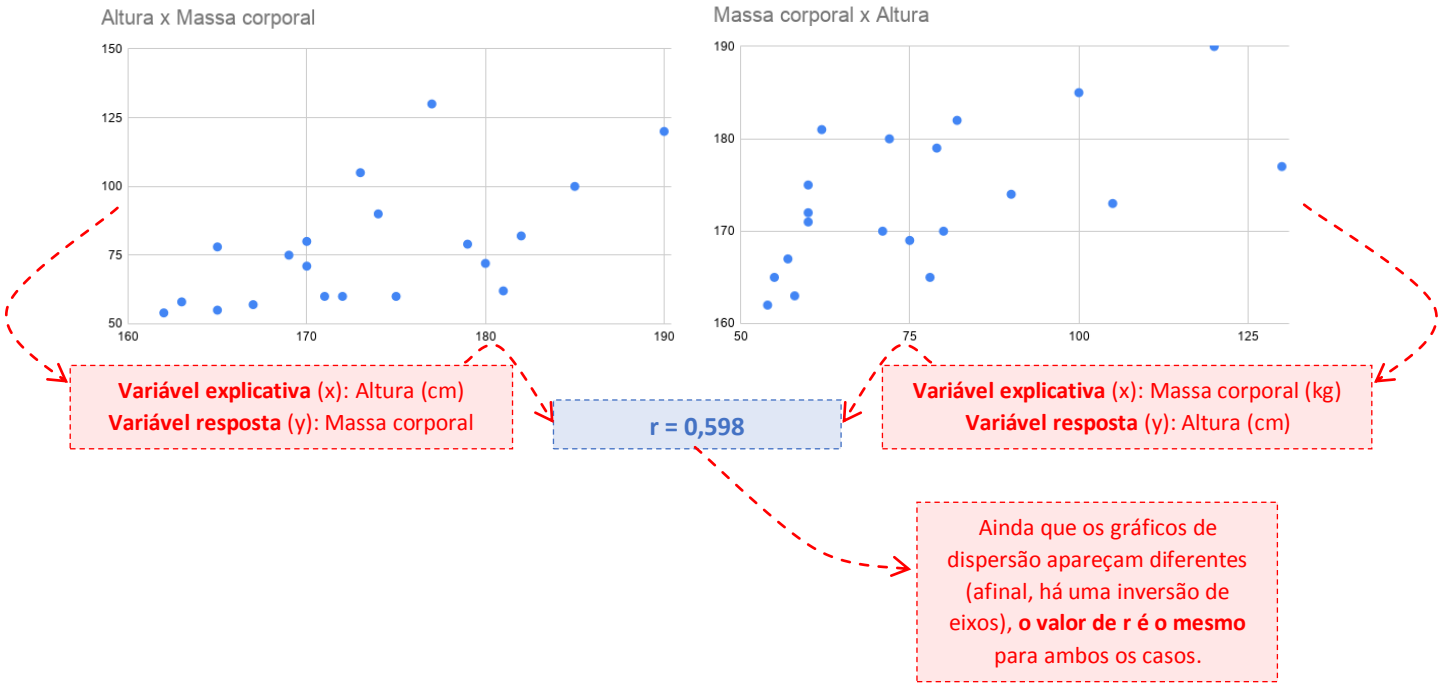
Claramente, a correlação r assume um valor positivo quando há uma associação positiva entre as variáveis x e y, e negativo quando a associação entre x e y é negativa.

Sobre o cálculo de r, é importante mencionar que:

não há distinção entre variável explicativa e variável resposta

Para o cálculo de r, é indiferente quem assume o valor de variável explicativa e variável resposta. O valor será sempre o mesmo.

Veja um exemplo com o *dataset* de massas corporais e alturas. À esquerda, o gráfico de dispersão para **Altura** como variável explicativa (x) e **Massa corporal** como variável resposta. À direita, o oposto.



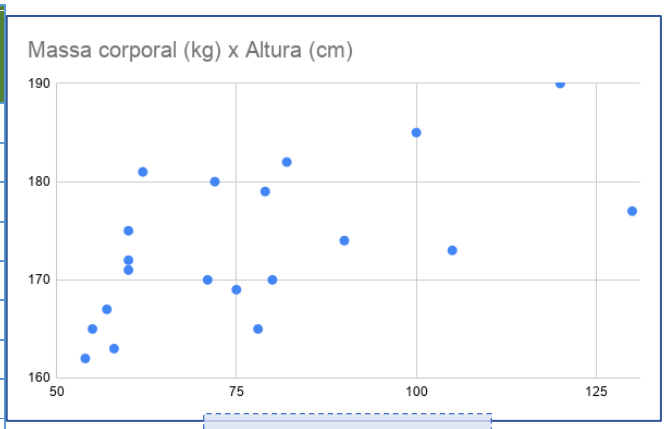
não importa a unidade de medida

Como o valor de r é calculado de maneira padronizada (ou seja, levando-se em consideração o desvio-padrão), a unidade de medida utilizada tanto para x quanto para y é irrelevante.

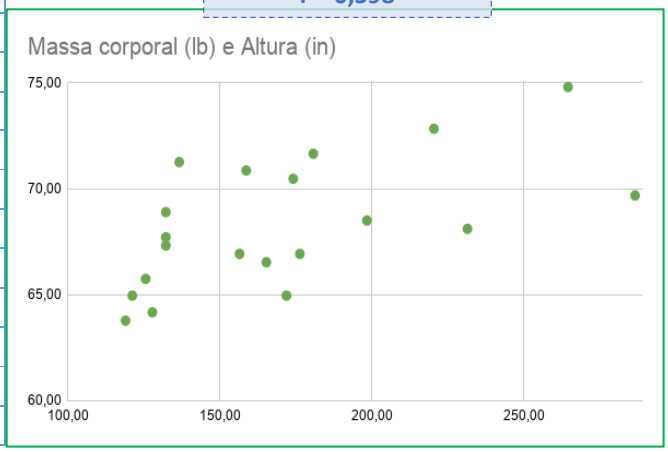
Desta forma, o coeficiente de relação r é **adimensional**, ou seja, não tem dimensão ou unidade de medida associada;

Usando o mesmo *dataset*, convertemos as massas corporais de quilogramas para libras (1kg = 2,20462lb) e as alturas, de centímetros para polegadas (1cm= 0,39370in). Em ambos os casos, r é o mesmo.

Massa corporal (kg)	Altura (cm)	Massa corporal (lb)	Altura (in)
72	180	158,73	70,87
80	170	176,37	66,93
60	175	132,28	68,90
90	174	198,42	68,50
100	185	220,46	72,83
120	190	264,55	74,80
82	182	180,78	71,65
79	179	174,16	70,47
78	165	171,96	64,96
55	165	121,25	64,96
71	170	156,53	66,93
75	169	165,35	66,54
130	177	286,60	69,69
105	173	231,49	68,11
60	172	132,28	67,72
54	162	119,05	63,78
58	163	127,87	64,17
57	167	125,66	65,75
60	171	132,28	67,32
62	181	136,69	71,26

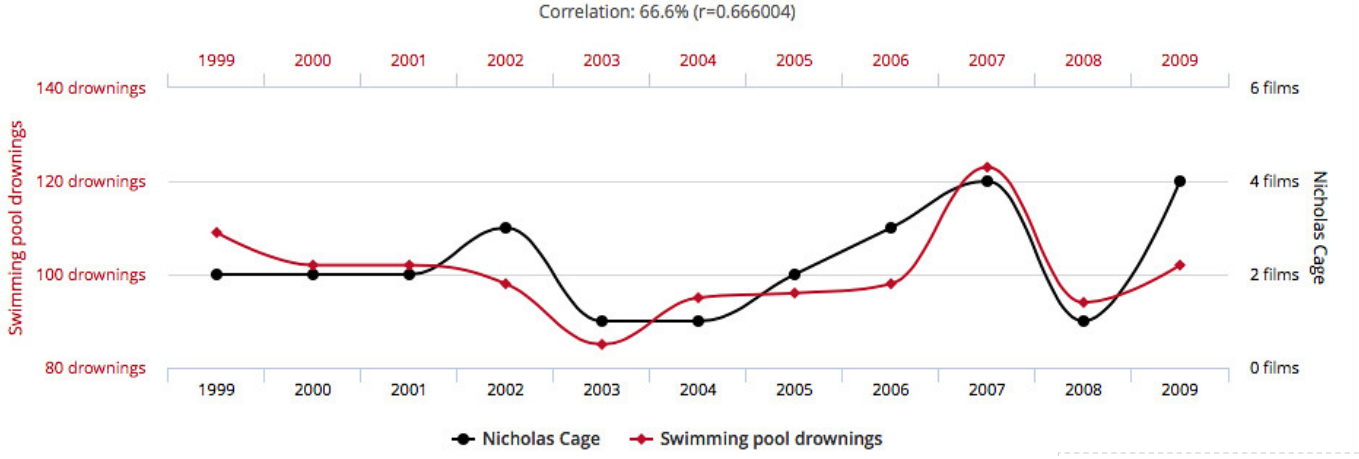


r = 0,598



correlação não é causalidade

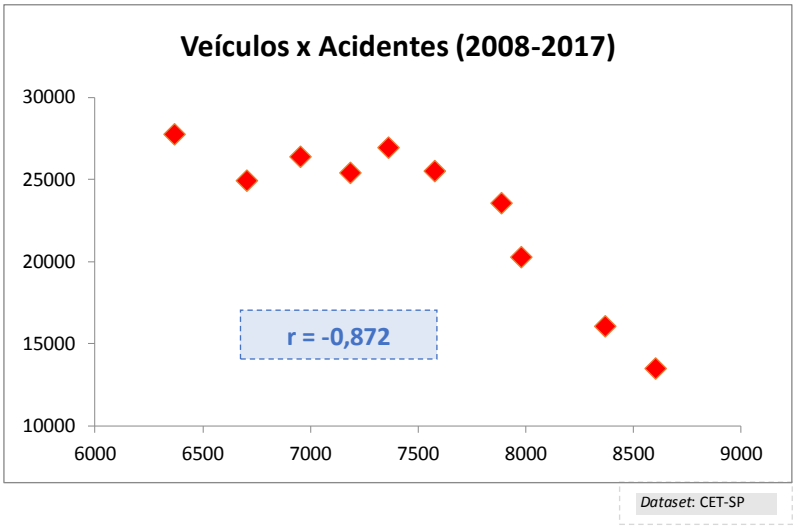
Não é porque o valor de r entre x e y é forte que podemos dizer que x é a **causa** de y, ou vice-versa. O exemplo a seguir mostra a correlação do número de pessoas afogadas em piscinas com os filmes em que o ator Nicolas Cage apareceu (r = 0,666) – uma coisa não tem a ver com outra.



Fonte: Tyler Vigen – Licença CC-BY
https://www.tylervigen.com/about

Veja o exemplo com os veículos da cidade de São Paulo e o número de acidentes registrado por ano (dados de 2008 a 2017).

Ano	Veículos	Acidentes
2008	6369	27739
2009	6705	24918
2010	6954	26371
2011	7186	25391
2012	7363	26928
2013	7578	25501
2014	7888	23547
2015	7980	20260
2016	8370	16052
2017	8604	13483



Note que a correlação aqui é forte e negativa (**r = - 0,872**). Entretanto, é um contrasenso querer afirmar que a diminuiçao do número de acidentes é explicada pelo aumento do número de carros – e não se pode afirmar isso. A única coisa que pode ser afirmada é que há uma correlação negativa e forte entre essas duas variáveis.

correlação sozinha não basta para descrever uma distribuição

À exceção das distribuições Normais, que podem ser completamente descritas por sua média e seu desvio-padrão, as demais distribuições (que não seguem o comportamento das distribuições Normais) precisam de outras medidas estatísticas para serem descritas, como o resumo dos cinco números e a correlação, por exemplo. Ainda assim, nem sempre os valores abtidos pela Estatística Básica são suficientes, sendo recomendável, na maioria das vezes. uma representação gráfica do *dataset*.

Observe, a título de exemplo, os quatro *datasets* conhecidos como Quarteto de Ascombe. Note que os quatro *datasets* têm os mesmos valores para as principais medidas estatísticas.

Cruzeiro do Sul Virtual
Educação a Distância

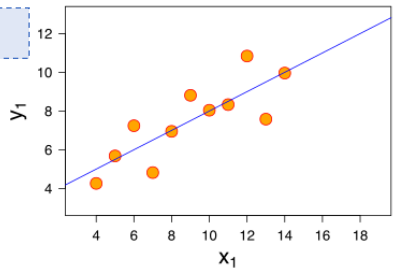
Quarteto 1		Quarteto 2		Quarteto 3		Quarteto 4	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Os quatro *datasets* têm as mesmas médias para seus valores de *x* e *y*, bem como as mesmas variâncias para seus *x* e *y*. A correlação de todos eles também é a mesma

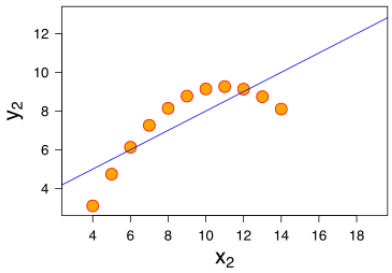
Propriedade estatística	Valor
Média de *x* ($\bar{x}$)	**9**
Variância de *x* (s_x)	**11**
Média de *y* ($\bar{y}$)	**7.50**
Variância de *x* (s_y)	**4.12**
Correlação entre *x* e *y* (r)	**0.816**

Aqui vemos a importância das representações gráficas na caracterização e compreensão das distribuições!

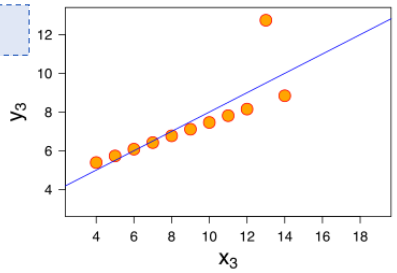
Quarteto 1



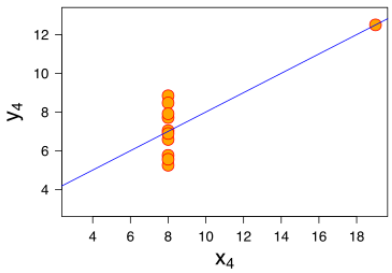
Quarteto 2



Quarteto 3



Quarteto 4

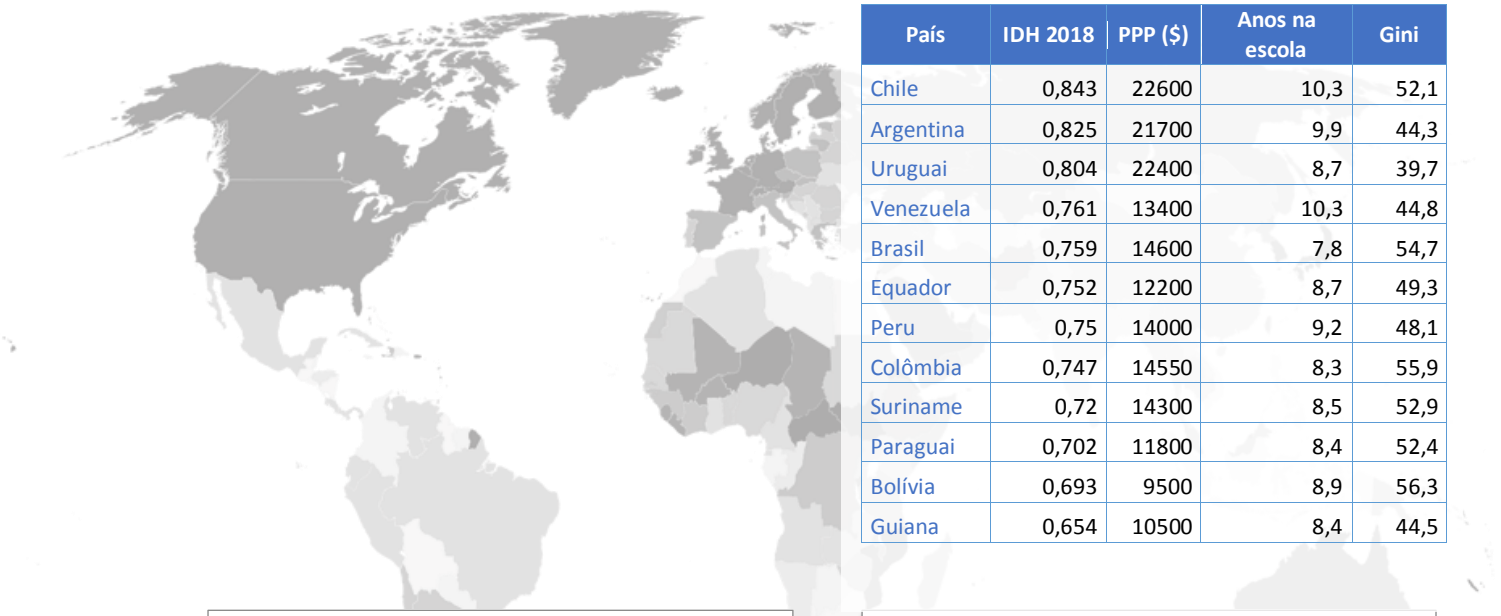


Esses quatro *datasets* são conhecidos como "**O quarteto de Anscombe**", em homenagem ao estatístico inglês **Francis Anscombe**.

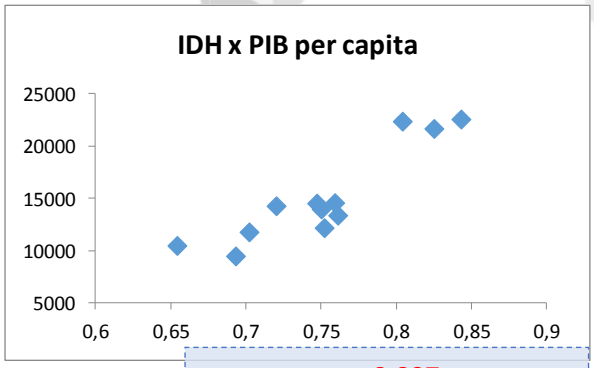
Saiba mais, sobre esses *datasets* e suas curiosas propriedades em:

https://pt.wikipedia.org/wiki/Quarteto_de_Anscombe

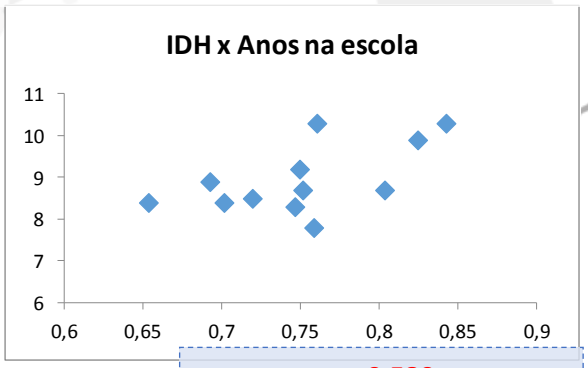
Vejamos mais um exemplo, com dados sobre os países da América do Sul, trabalhados também na unidade sobre Gráficos de Dispersão. No *dataset*, são apresentados os valores para PPP (PIB – Produto Interno Bruto Per Capita, em dólares estadunidenses), anos na escola e o índice Gini (que calcula o coeficiente de desigualdade socioeconômica).



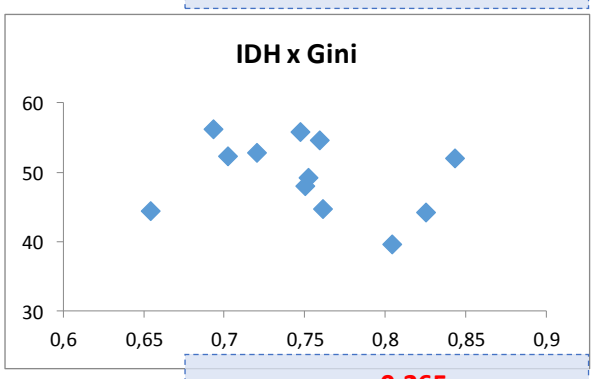
País	IDH 2018	PPP ($)	Anos na escola	Gini
Chile	0,843	22600	10,3	52,1
Argentina	0,825	21700	9,9	44,3
Uruguai	0,804	22400	8,7	39,7
Venezuela	0,761	13400	10,3	44,8
Brasil	0,759	14600	7,8	54,7
Equador	0,752	12200	8,7	49,3
Peru	0,75	14000	9,2	48,1
Colômbia	0,747	14550	8,3	55,9
Suriname	0,72	14300	8,5	52,9
Paraguai	0,702	11800	8,4	52,4
Bolívia	0,693	9500	8,9	56,3
Guiana	0,654	10500	8,4	44,5



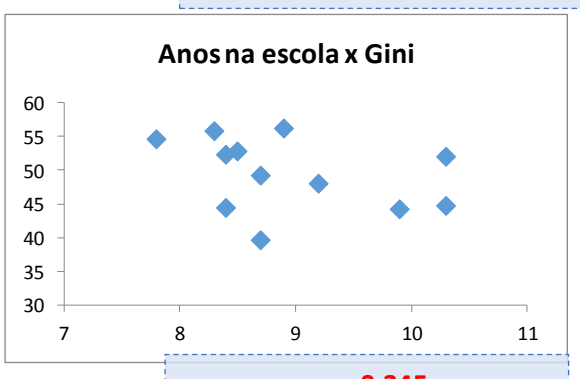
r = 0,907
(correlação forte e positiva)



r = 0,580
(correlação poucoforte e positiva)



r = - 0,265
(correlação fraca e negativa)



r = - 0,345
(correlação fraca e negativa)

Cruzeiro do Sul Virtual
Educação a Distância

Para saber mais, leia o capítulo 4 do e-book:

MOORE, David S.; NOTZ, William I.; FLINGER, Michael A. A Estatística Básica e sua Prática. 7 ed. Rio de Janeiro: LTC, 2017 – Capítulo 4