Inferência Estatística

Correlação e Regressão

- **Coeficiente de Correlação**
- Regressão Linear Simples

Correlação

- Os testes de hipóteses vistos até agora analisam informações referentes a uma única variável, porém frequentemente estamos interessados em analisar o **comportamento conjunto** de duas variáveis.

- Com duas variáveis também pode ser de interesse conhecer se elas têm algum tipo de **associação entre si**.

  - se valores baixos (altos) de uma das variáveis implicam em valores altos (ou baixos) da outra variável.

Exemplos:

- relação entre a altura dos pais e a altura dos filhos,
- relação entre renda familiar e número de filhos.

Gráfico de Dispersão

- Uma forma bastante útil de se observar a relação entre duas variáveis é o **gráfico de dispersão**.
- Em geral vamos supor que há uma **variável dependente (Y)** que depende de outra **variável preditora (X)**.
- O diagrama de dispersão fornece uma ideia do tipo de relacionamento entre as duas variáveis.
  - pais altos (X) e filhos altos (Y),
  - renda familiar alta (X) e baixo número de filhos (Y).

Exemplo: Observe os seguintes diagramas de dispersão que dizem respeito ao número do calçado (tamanho da sapatilha) e a altura dos atletas que estão a escalar uma montanha e, no segundo caso, à relação entre a altitude e a temperatura.

Gráfico 1

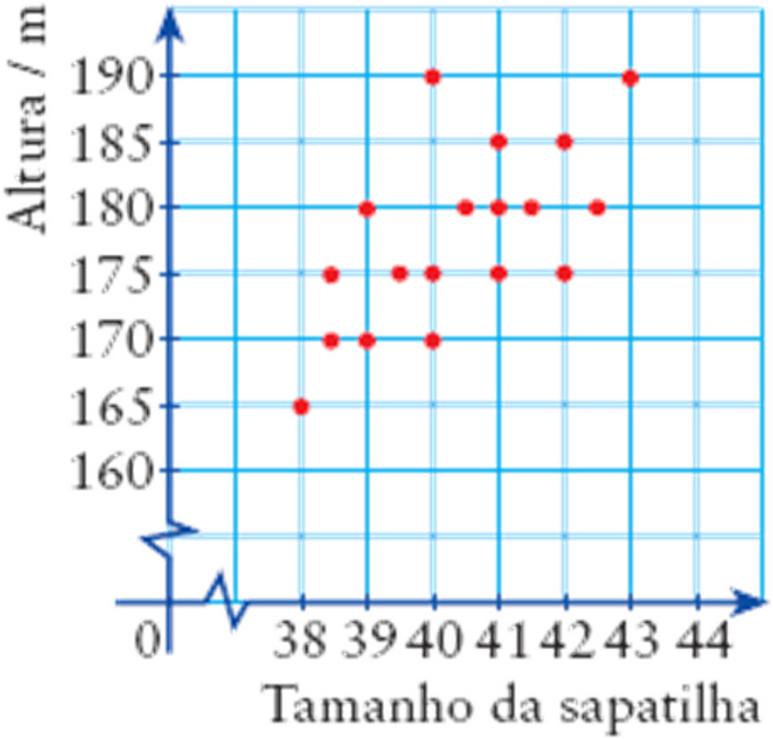
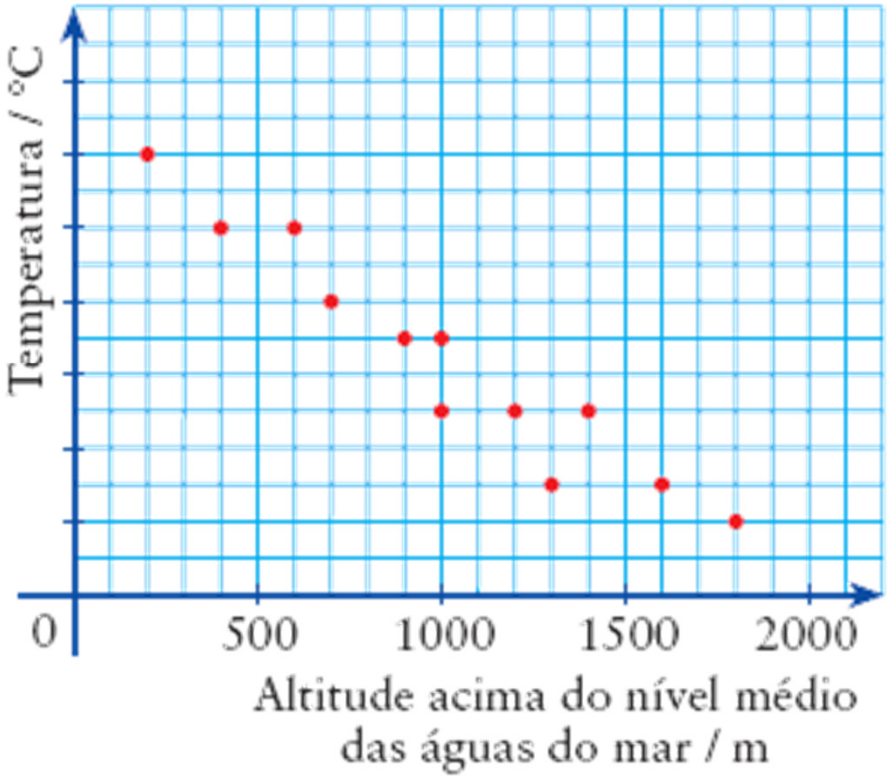


Gráfico 2



Pode concluir-se que há uma relação entre a altura de uma pessoa e o número de sapatilha que usa?

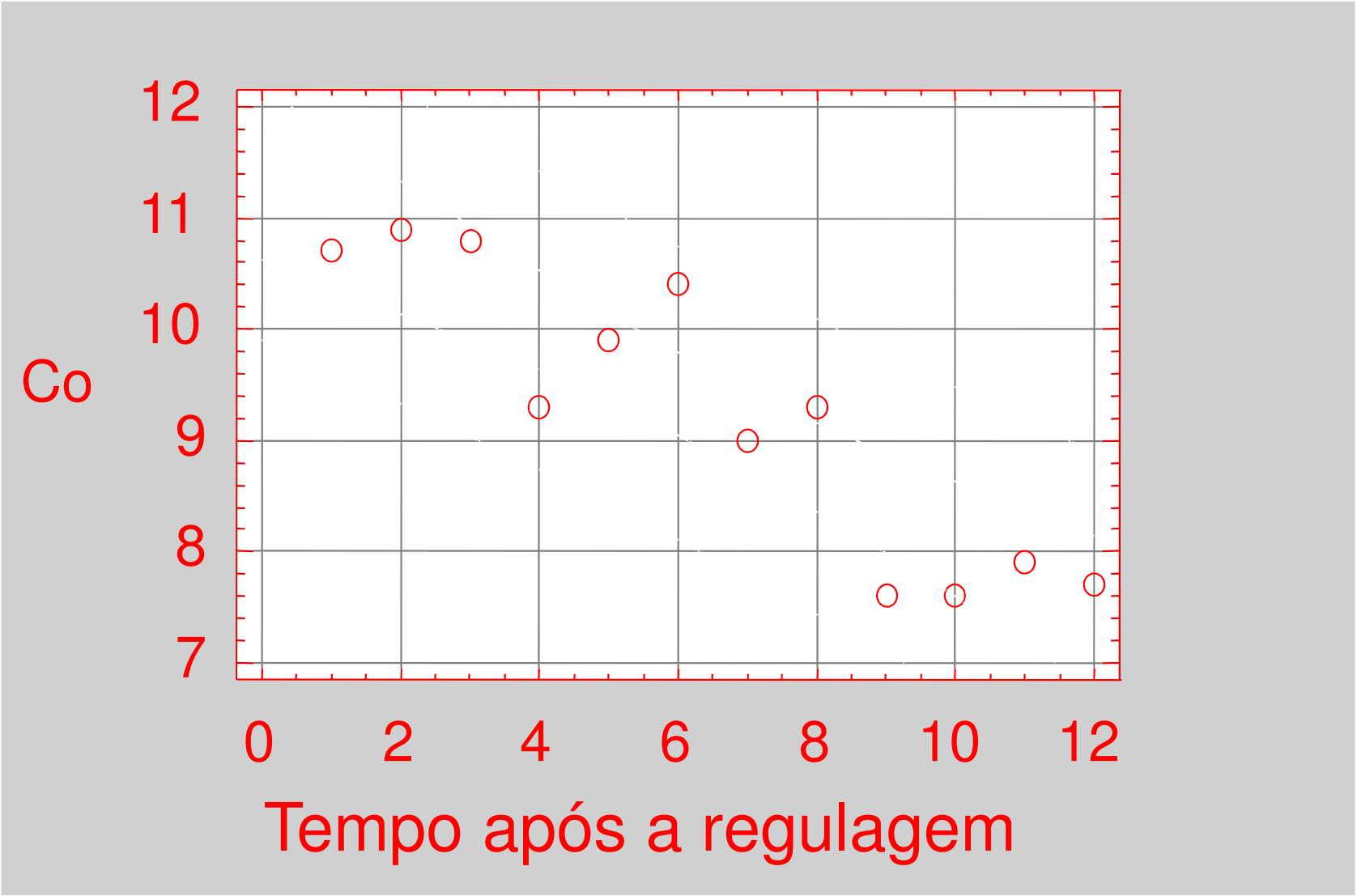
À medida que se subia a montanha a temperatura subia ou descia?

Exemplo:

Após uma regulagem eletrônica um veículo apresenta um rendimento ideal no que tange a consumo de combustível. Contudo, com o passar do tempo esse rendimento vai se degradando. Os dados a seguir representam o rendimento medido mês a mês após a regulagem.

X: meses após a regulagem	1	2	3	4	5	6
Y: rendimento	10,7	10,9	10,8	9,3	9,5	10,4
X: meses após a regulagem	7	8	9	10	11	12
Y: rendimento	9,0	9,3	7,6	7,6	7,9	7,7

Rendimento de combustível



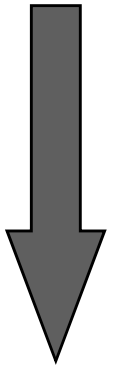
Coeficiente de Correlação de Pearson

- Para uma amostra de ***n*** pares de valores *(x,y)* o coeficiente de correlação linear de Pearson ***r*** fornece uma medida da relação linear que existe entre duas variáveis *X* e *Y*.

$$S_{xx} = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$$

$$S_{yy} = \sum y_i^2 - \frac{(\sum y_i)^2}{n}$$

$$S_{xy} = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}$$



$$r = \frac{S_{xy}}{\sqrt{S_{xx} \times S_{yy}}}$$

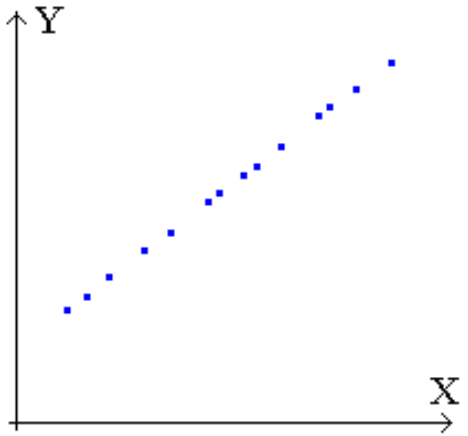
Interpretação do coeficiente

Apesar de r ser um valor adimensional, ele não é uma taxa e, portanto, o resultado **não** deve ser expresso em percentagem.

- r positivo ⇒ correlação positiva entre x e y
- r negativo ⇒ correlação negativa entre x e y
- r próximo de 0 indica ausência de correlação entre x e y

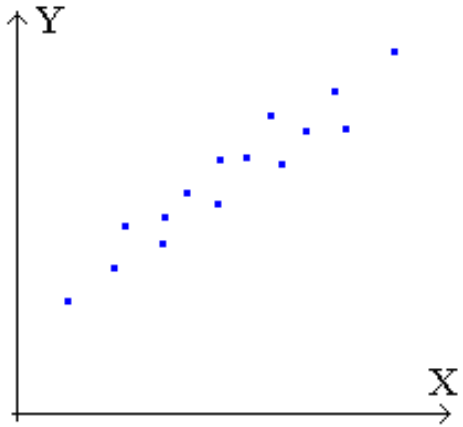
r	Interpretação da correlação
0 a 0,40	Fraca
0,40 a 0,60	Regular
0,60 a 0,80	Boa
0,80 a 0,99	Forte
1	Perfeita

Interpretação do coeficiente



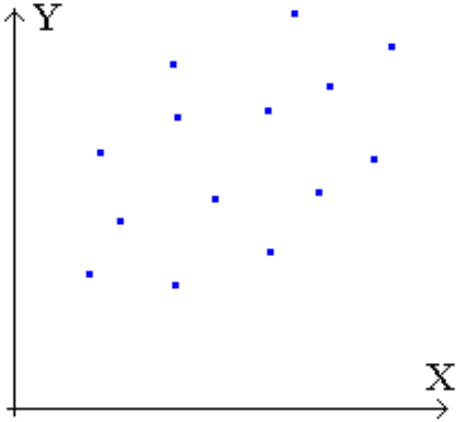
Positiva perfeita

r = +1



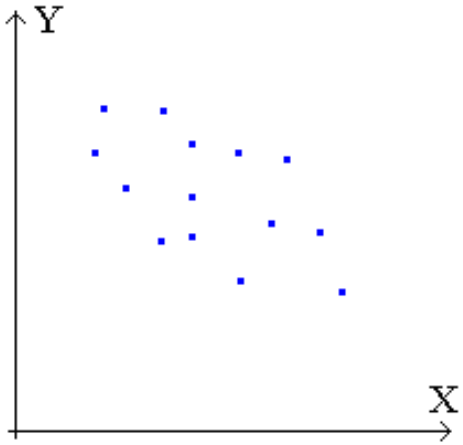
Positiva forte

r = +0,9



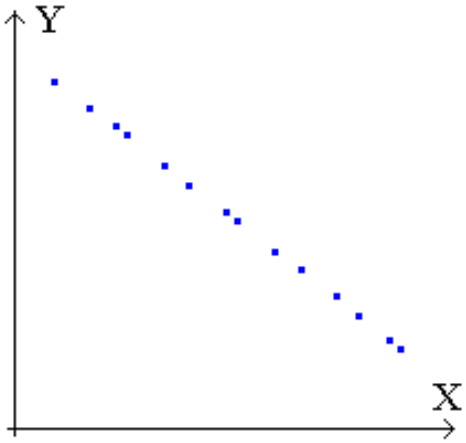
Positiva fraca

r = +0,1



Negativa regular

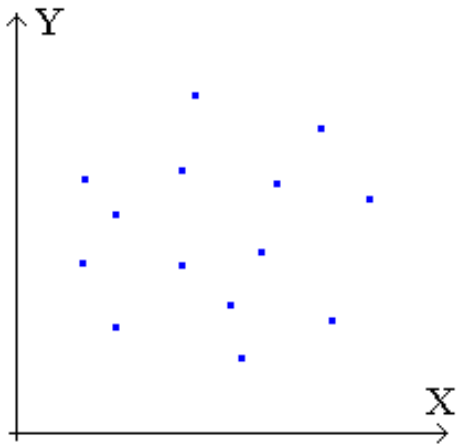
r = -0,5



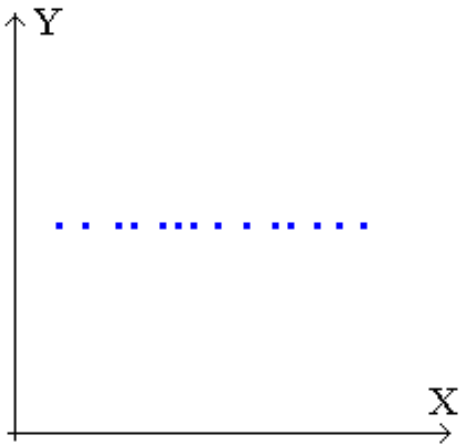
Negativa perfeita

r = -1

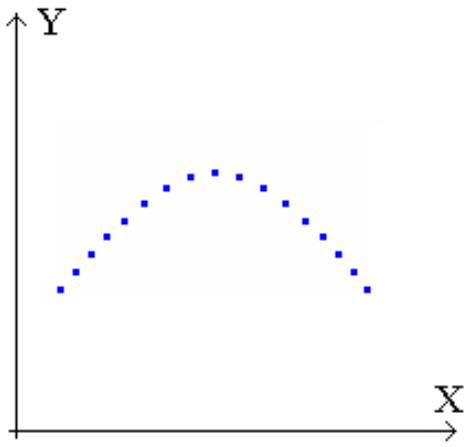
Interpretação do coeficiente



$r = 0$



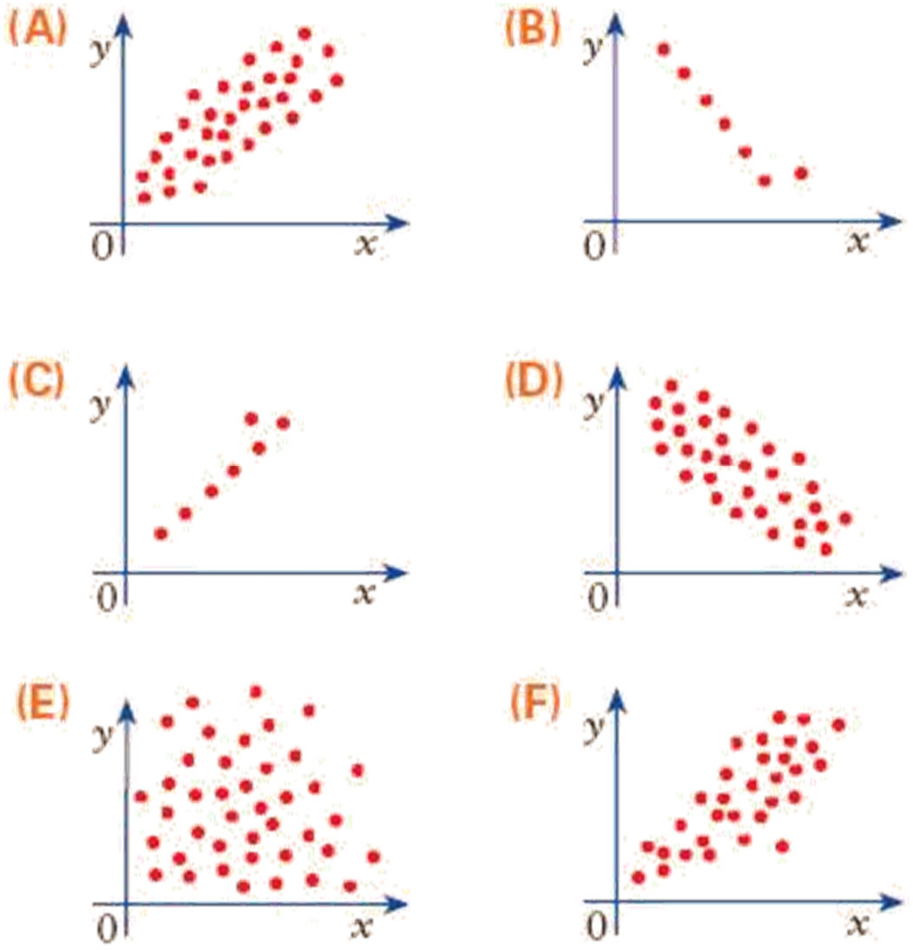
$r = 0$



$r = 0$

Exercício:

Observe os seguintes diagramas de dispersão.



1 Indique, pela letra correspondente, aqueles em que se observa:

a) uma associação positiva;

b) uma associação negativa.

2 Indique, pela letra correspondente, o diagrama em que não há uma associação clara entre as duas variáveis.

Exercício:

Observe os diagramas de dispersão.

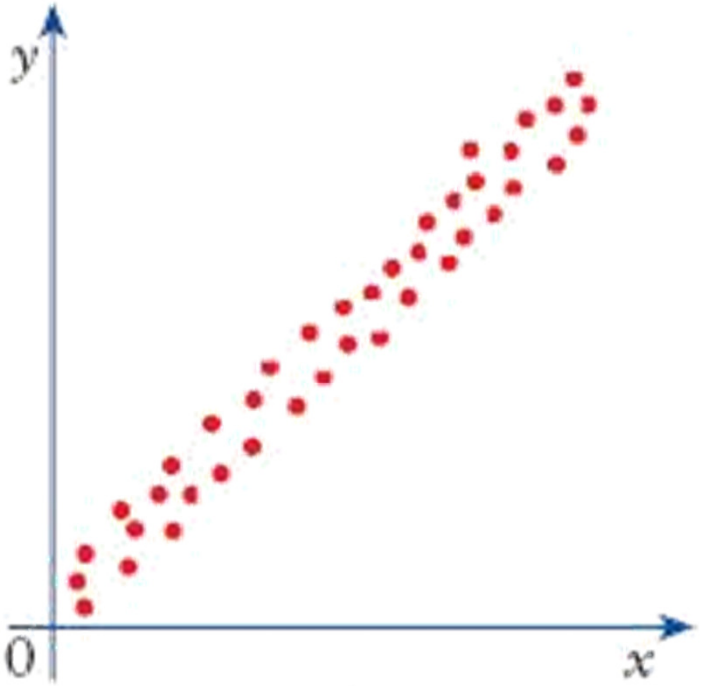


Gráfico 1

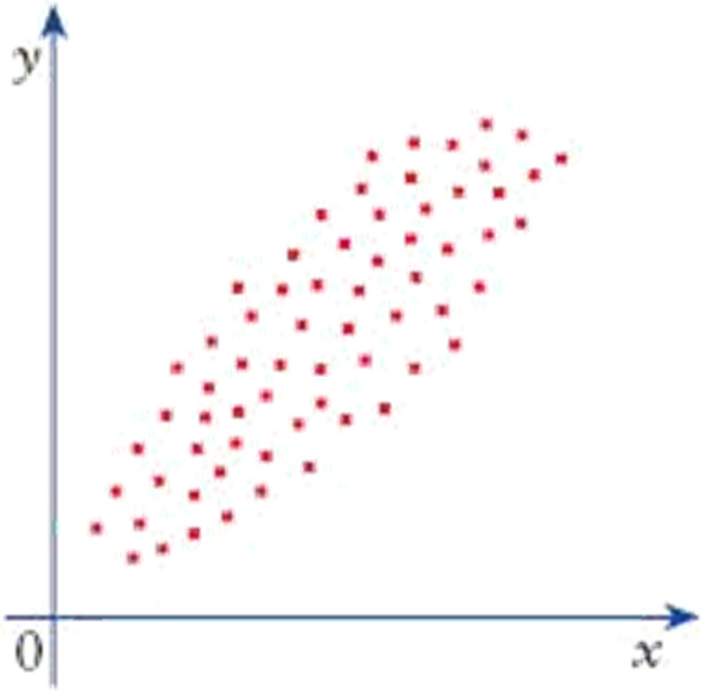
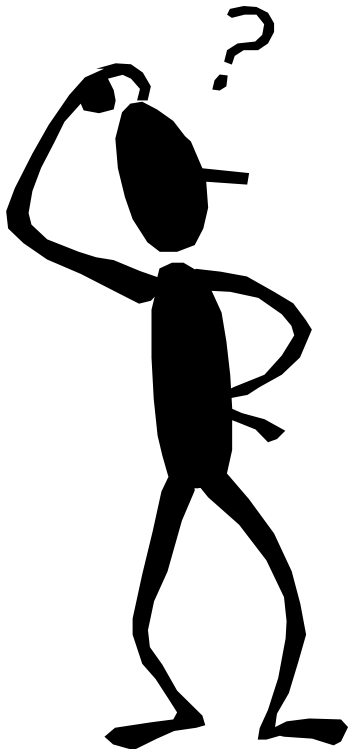


Gráfico 2

Em qual deles lhe parece haver um maior grau de associação entre as variáveis x e y ?

Explique o seu raciocínio.

Voltando ao exemplo: cálculos iniciais



Meses(X)	Rendimento(Y)	X^2	Y^2	X*Y
1	10,7	1	114,49	10,7
2	10,9	4	118,81	21,8
3	10,8	9	116,64	32,4
4	9,3	16	86,49	37,2
5	9,5	25	90,25	47,5
6	10,4	36	108,16	62,4
7	9	49	81	63
8	9,3	64	86,49	74,4
9	7,6	81	57,76	68,4
10	7,6	100	57,76	76
11	7,9	121	62,41	86,9
12	7,7	144	59,29	92,4
78	**110,7**	**650**	**1039,55**	**673,1**
6,5	**9,225**			

$\Sigma x_i = 78$ $\Sigma x_i^2 = 650$

$\Sigma y_i = 110,7$ $\Sigma y_i^2 = 1039,55$ $\Sigma x_i y_i = 673,1$

Cálculos

$\Sigma x_i = 78$ $\quad \Sigma x_i^2 = 650$

$\Sigma y_i = 110,7$ $\quad \Sigma y_i^2 = 1039,55$ $\quad \Sigma x_i y_i = 673,1$

$$S_{xx} = \sum x_i^2 - \left(\sum x_i\right)^2 / n = 650 - (78)^2/12 = 143$$

$$S_{yy} = \sum y_i^2 - \left(\sum y_i\right)^2 / n = 1039,55 - (110,7)^2/12 = 18,34$$

$$S_{xy} = \sum x_i y_i - \left(\sum x_i\right)\left(\sum y_i\right)/n = 673,1 - (78 \times 110,7)/12 = -46,45$$

$$r = \frac{S_{xy}}{\sqrt{S_{xx} \times S_{yy}}} = \frac{-46,45}{\sqrt{143 \text{ x } 18,34}} = -0,907$$

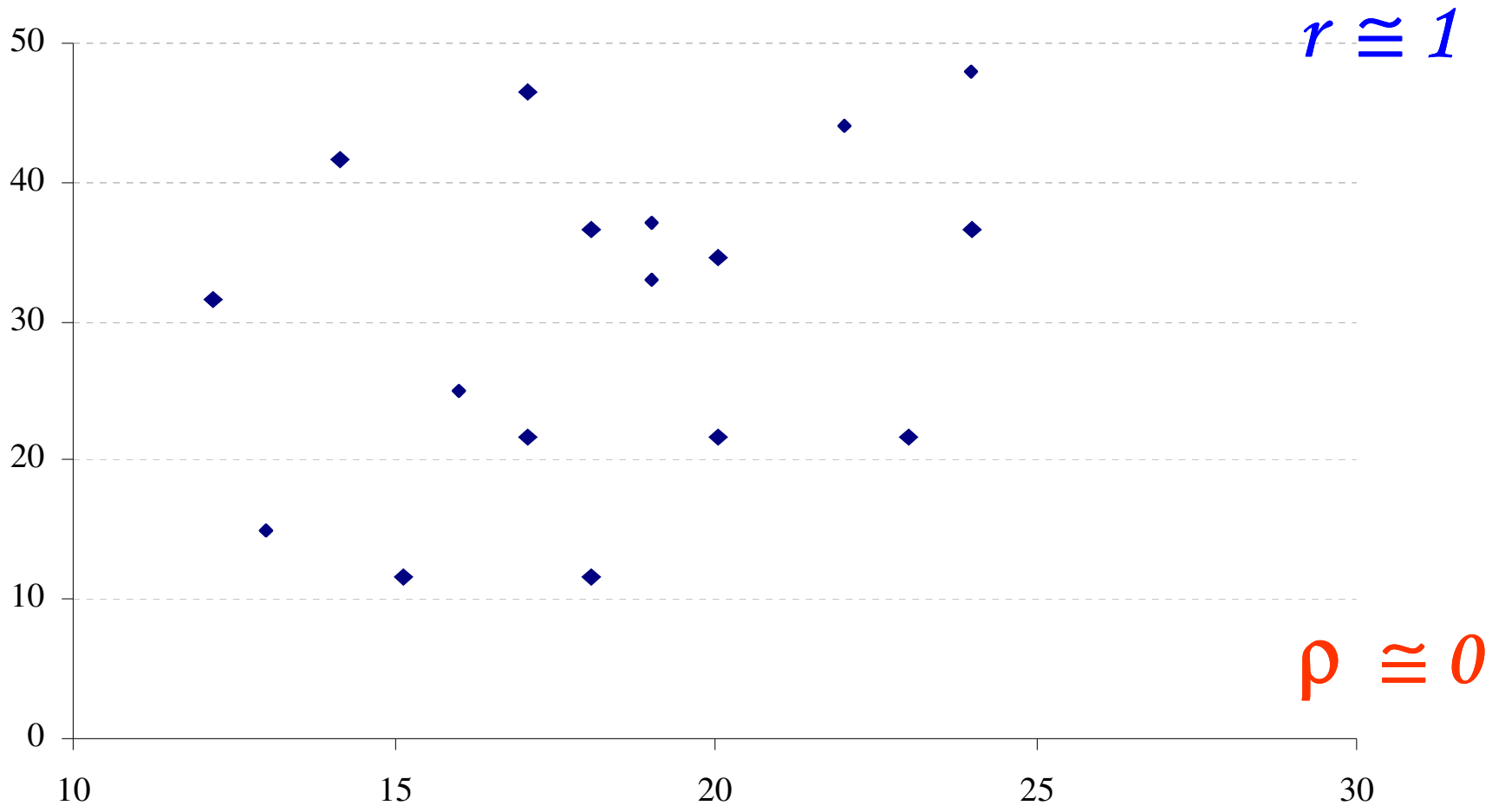
Interpretação: Existe uma correlação linear inversa na amostra entre tempo após a regulagem e rendimento; passa o tempo e diminui o rendimento do combustível. A intensidade desta correlação é forte.

Correlação ≠ Causalidade

- O coeficiente de correlação não mede a relação causa e efeito entre as variáveis, apesar de que essa relação possa estar presente.
- Um exemplo é a forte correlação positiva entre as vendas anuais de chicletes e a taxa de criminalidade nos EUA.
- Obviamente, não podemos concluir que haja a relação de causa e efeito e que para reduzir a taxa de criminalidade bastaria proibir a venda de chicletes.
- O que se observa é que as duas variáveis são dependentes do tamanho da população, e é essa relação mútua com a terceira variável (tamanho da população) que produz a correlação forte e positiva entre a venda de chicletes e a incidência de crimes nos EUA.

Teste de hipótese para coeficiente de correlação

- Observada uma amostra de seis pares, pode-se perceber que a correlação é quase um, isto é, r ≅ 1. No entanto, observe o que ocorre quando mais pontos são acrescentados, isto é, quando se observa a população!



Teste de hipótese para coeficiente de correlação

- Uma correlação amostral não significa necessariamente uma correlação populacional. É necessário testar o coeficiente de correlação para verificar se a correlação amostral é também populacional.

- A hipótese da existência de uma relação entre *X* e *Y*, pode ser formulada usando-se:

$H_0 : \rho = 0$ (não existe correlação)

$H_A : \rho \neq 0$ (existe correlação)

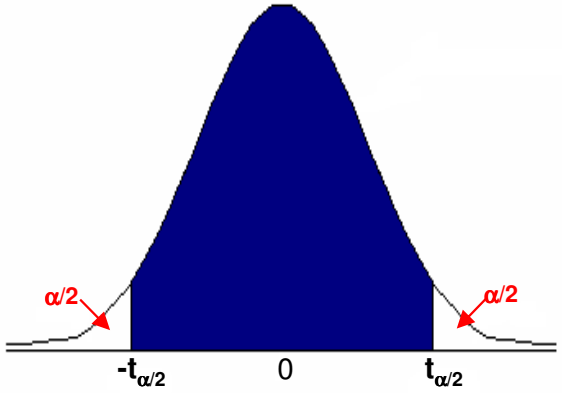
onde a letra ρ é usada para representar o valor populacional do coeficiente de correlação. Pode ser demonstrado que o valor da estatística *T* pode ser calculado usando:

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2}$$

Teste de hipótese para coeficiente de correlação

- Assim a hipótese da existência de uma relação entre *X* e *Y* pode ser verificada diretamente a partir do valor amostral do coeficiente de correlação. Assim, a hipótese nula será rejeitada se o valor t calculado for maior que o tabelado:

$$|t| > t_{\alpha/2,\, n-2}$$



- Para o exemplo em estudo tem-se:

$$t = \frac{-0{,}907\sqrt{12-2}}{\sqrt{1-(-0{,}907)^2}} = |-6{,}82| > t_{0{,}025;10} = 2{,}228 \Rightarrow \text{rejeita-se } H_0,$$

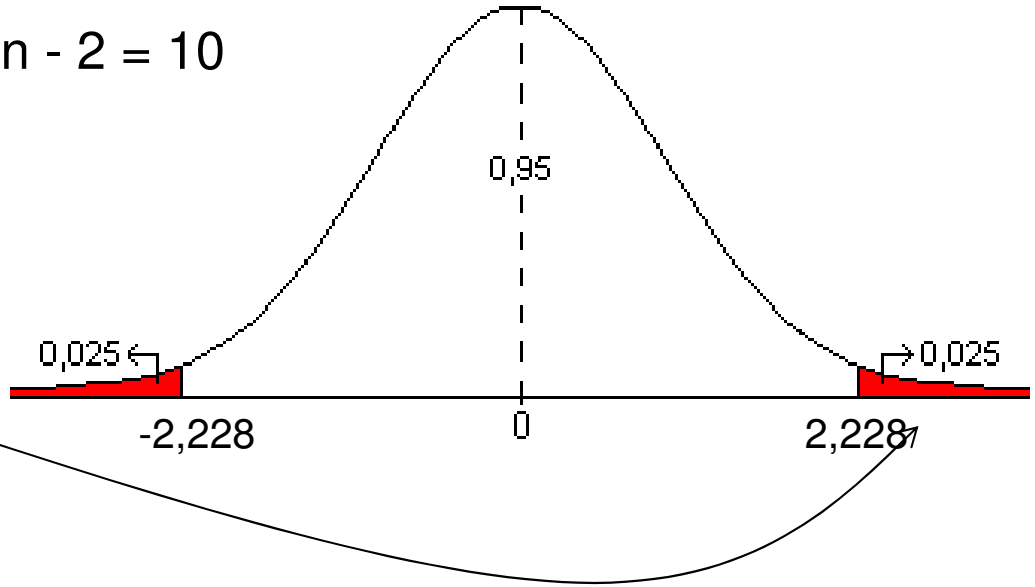
ou seja, descarta-se a hipótese nula e conclui-se que deve existir correlação entre as variáveis estudadas.

Outro exemplo: Suponha que uma amostra de n = 12 alunos forneceu um coeficiente de correlação amostral de r = 0,66 entre X = "nota em cálculo" e Y = "nota em estatística". Verifique se é possível afirmar que uma nota boa em cálculo está relacionada com uma nota boa em estatística a 5% de significância.

$H_0 : \rho = 0$
$H_A : \rho \neq 0$

⇨ $\alpha = 5\%$ e $\nu = n - 2 = 10$

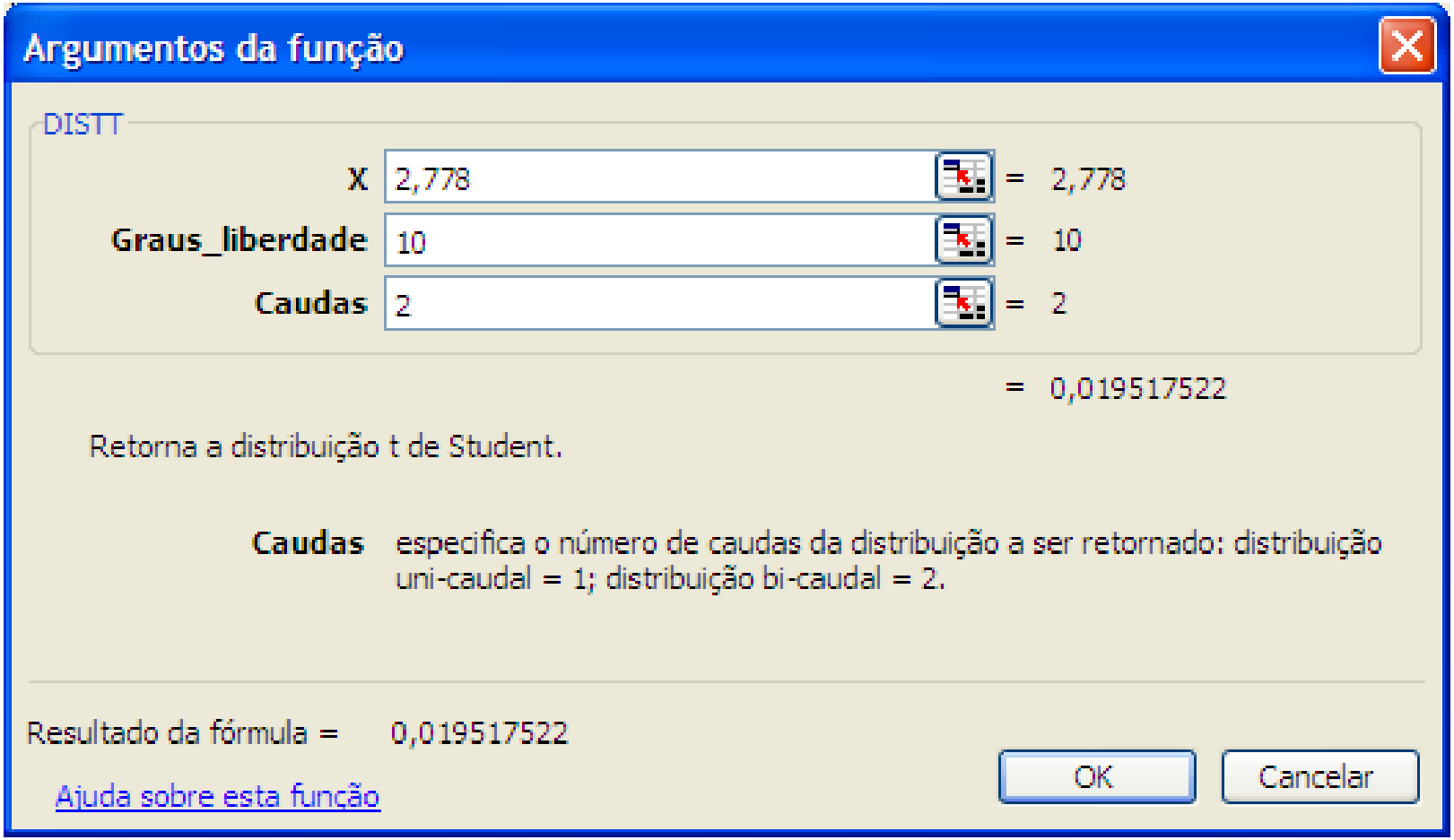
$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0,66\sqrt{12-2}}{\sqrt{1-0,66^2}} = 2,778$$



Conclusão: Rejeita-se H_0, isto é, a 5% de significância, pode-se afirmar que a nota de cálculo deve estar relacionada com a de estatística.

19

Outra opção de análise é trabalhar com a significância do resultado obtido (2,778), isto é, o valor p. Para isto, deve-se calcular $P(|t_{10}| > 2,778)$. Utilizando o Excel, tem-se:



Conclusão: Como a significância do resultado (1,95%) é menor que a significância do teste (5%) é possível rejeitar a hipótese nula.

Exercício:

Considere os dados abaixo, referentes às variáveis vendas e espaço nas prateleiras (em cm^2) para produtos.

Calcule o valor do coeficiente de correlação, interprete e teste sua significância a 5%.

Espaço (X):	340	230	405	325	280	195	265	300	350	410
Vendas (Y):	71	65	83	74	67	56	57	78	84	65

r = 0,6420

t_c = 2,368 e t_{tab} = 2,306