CORRELAÇÃO LINEAR

Referência

Cap. 7 - Métodos Estatísticos para Geografia

Maria Elisa Siqueira Silva – PPGGF - USP

Correlação linear - Definição

- Permite verificar se duas variáveis independentes estão associadas uma com a outra

- Questionamentos iniciais:

  "A temperatura de superfície dos oceanos tem alguma relação com a vazão de rios?"

  "Ou, a diminuição do preço de um produto tem relação com o aumento de sua oferta? Podem, em um primeiro momento, ser observada através da correlação linear?"

COEFICIENTE DE CORRELAÇÃO r

- Uma das formas utilizadas para se encontrar essas relações é o cálculo do coeficiente de correlação linear de Pearson, r

$$r \in [-1,0; +1,0]$$

$r = 1,0 \rightarrow$ correlação positiva perfeita

$r = -1,0 \rightarrow$ correlação negativa perfeita

COEFICIENTE DE CORRELAÇÃO r

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

observações

t_i	x_i	y_i
1	x_1	y_1
2	x_2	y_2
...	....	....
t_n	x_n	y_n

$\sum_{i=1}^{N}$ → Somatória

x_i y_i → VETORES ($x_1, x_2, ..., x_n$) e ($y_1, y_2, ..., y_n$) - duas variáveis observadas em cada observação, por exemplo, a cada passo de tempo i

$\bar{x}$ $\bar{y}$ → média da amostra x e de y

σ_x σ_y → desvio padrão das amostras x e y

SOMATÓRIA

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

Numerador:

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = (x_1 - \bar{x})(y_1 - \bar{y}) + (x_2 - \bar{x})(y_2 - \bar{y}) + \cdots + (x_n - \bar{x})(y_n - \bar{y}),$$

i = 1, ..., n

Como escrever o denominador???

DESVIO PADRÃO σ s dp

- É uma medida de dispersão e indica a dispersão média de um conjunto de dados em relação à média aritmética da amostra

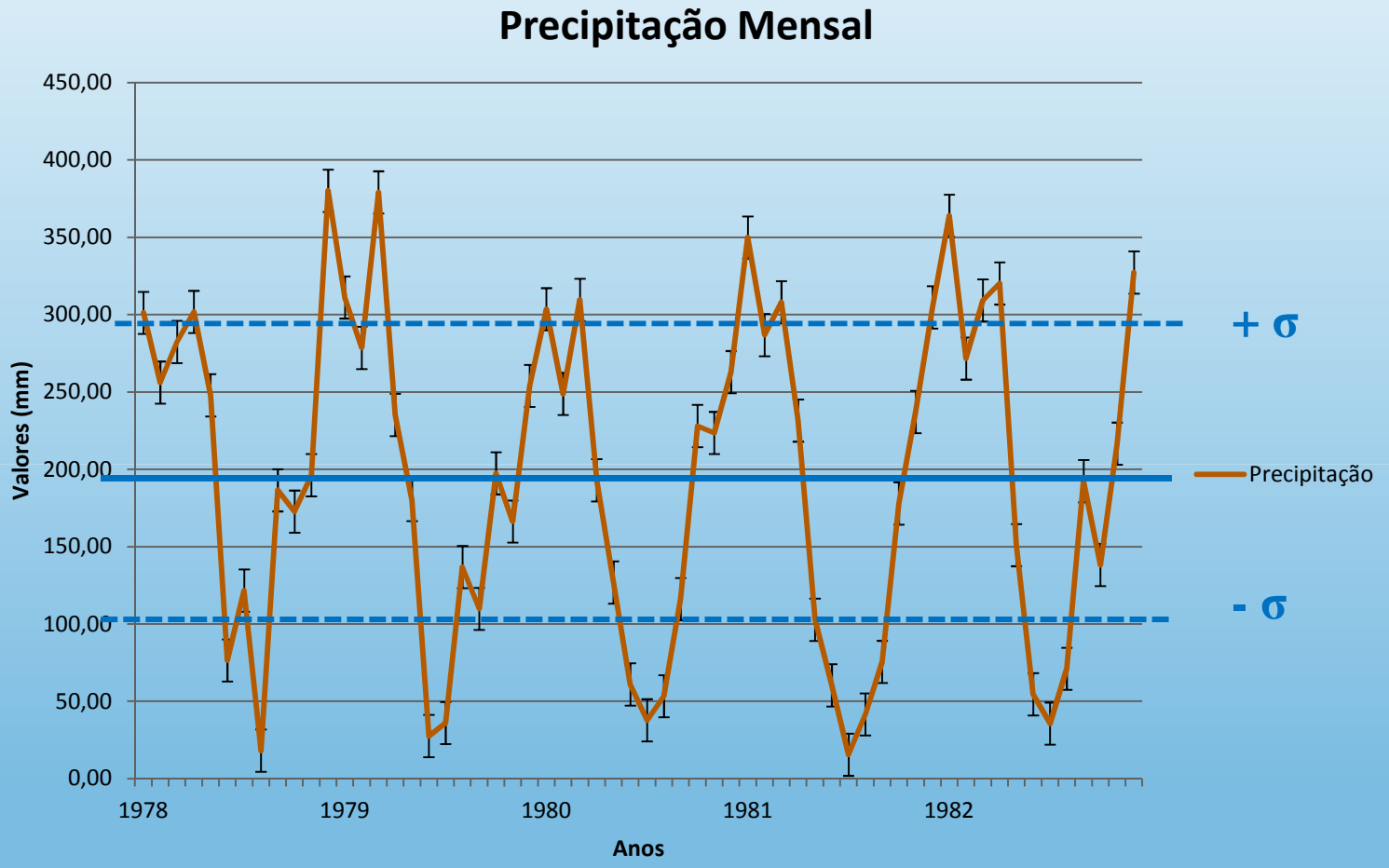
- Variância = var = s^2

  variância = desvio padrão ao quadrado

DESVIO PADRÃO

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

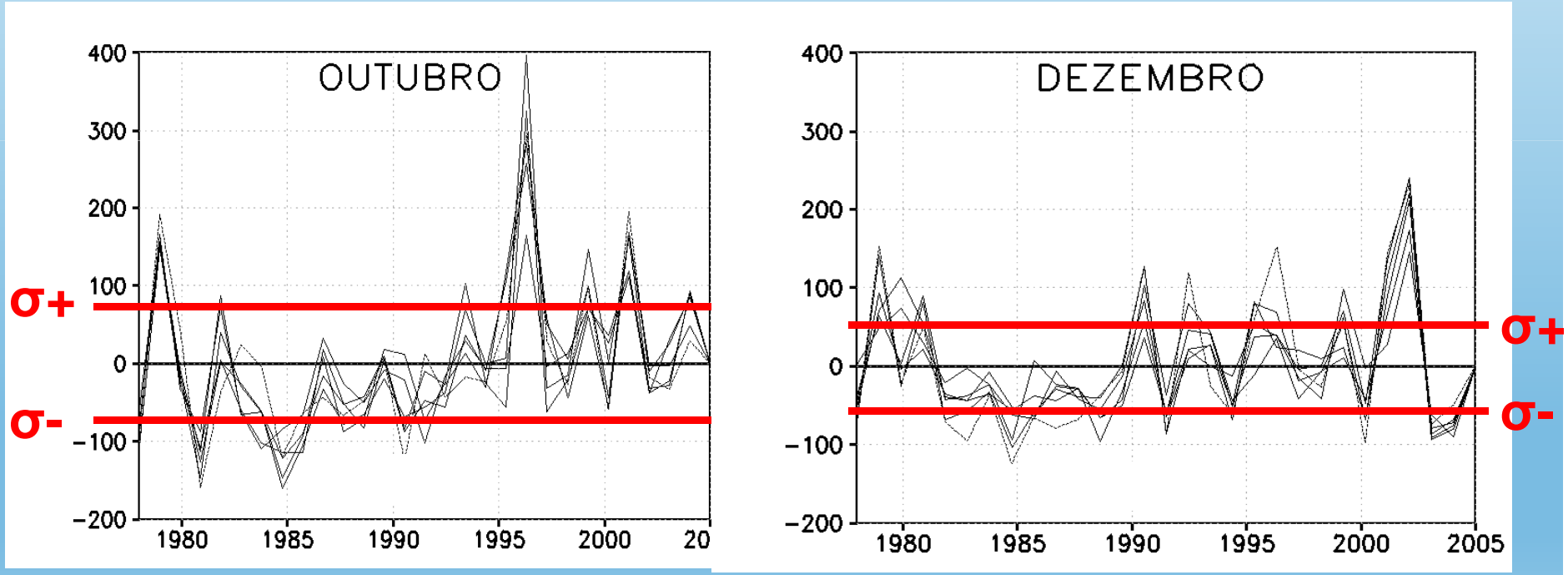
Desvio Padrão - exemplo



σ = 105,6634
pcp média= 194,36
σ^2= 11.164,77

Dada uma série temporal,
quantos valores de desvio padrão tem a série?

ANOMALIA PRECIPITAÇÃO NO NOROESTE DO RS 1978-2005



Sleiman (2005)

VARIÂNCIA σ^2

A variância mostra o quão distantes os valores estão da média, é expressa por:

$$s^2 = \text{var} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

INTERPRETAÇÃO DA CORRELAÇÃO ENTRE DUAS VARIÁVEIS

- **Correlação positiva**

  Quando uma variável aumenta (diminui), a outra também aumenta (diminui)

  → relação diretamente proporcional

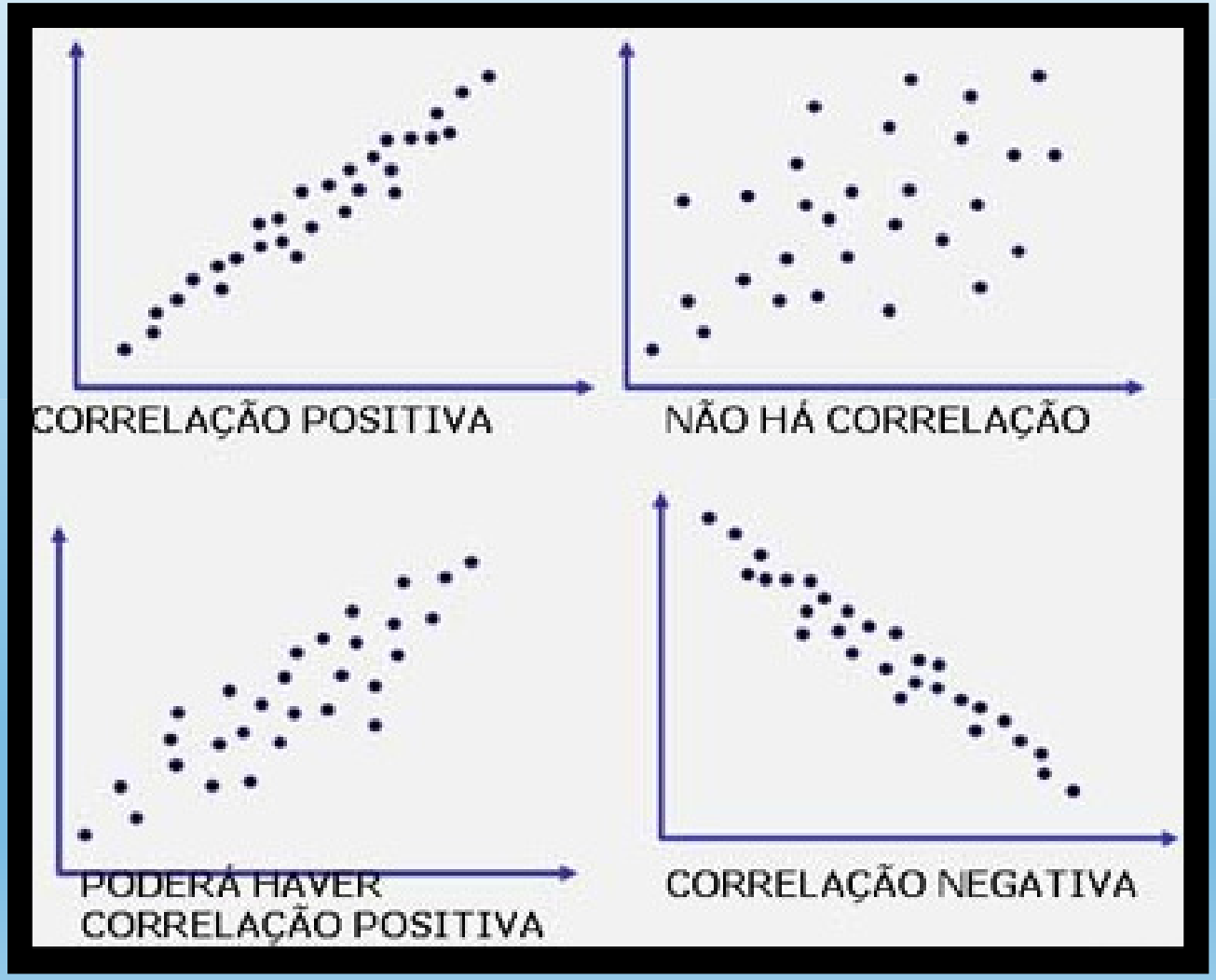
- **Correlação negativa**

  Quando uma variável aumenta (diminui), a outra diminui (aumenta)

  → relação inversamente proporcional

- **Sem correlação**

EXEMPLOS HIPOTÉTICOS DE CORRELAÇÃO ENTRE VARIÁVEIS ALEATÓRIAS



Fonte: http://arquivos.unama.br/professores/iuvb/7semestre/GQ/aula08/verprint.htm

EXEMPLOS

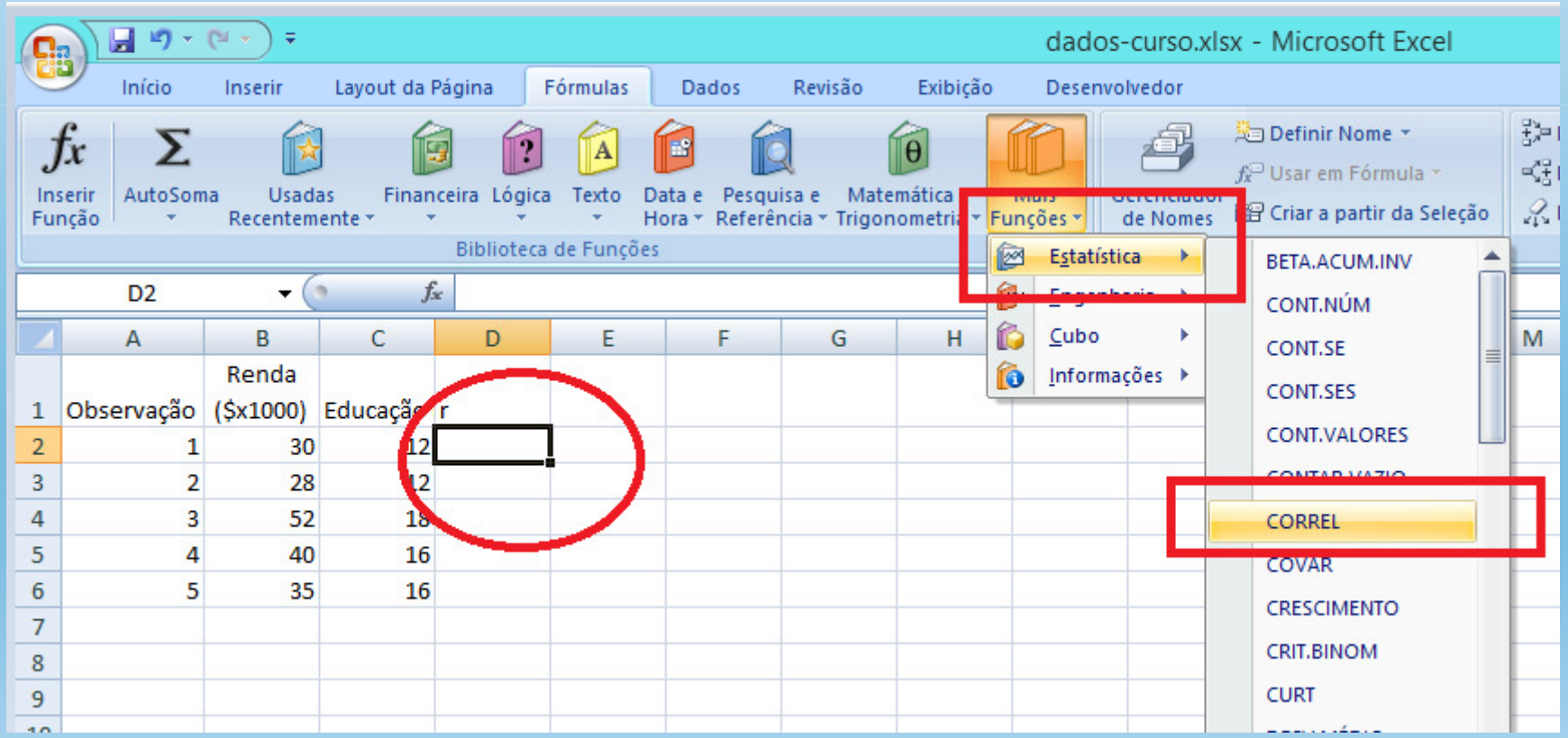
- Faremos alguns exercícios simples de correlação utilizando uma planilha eletrônica, como o Excel

Os exemplos dados a seguir foram criados a partir do Excel

EXERCÍCIO 01: Cálculo da correlação, r, para as amostras de dados de renda e educação (anos de estudos).

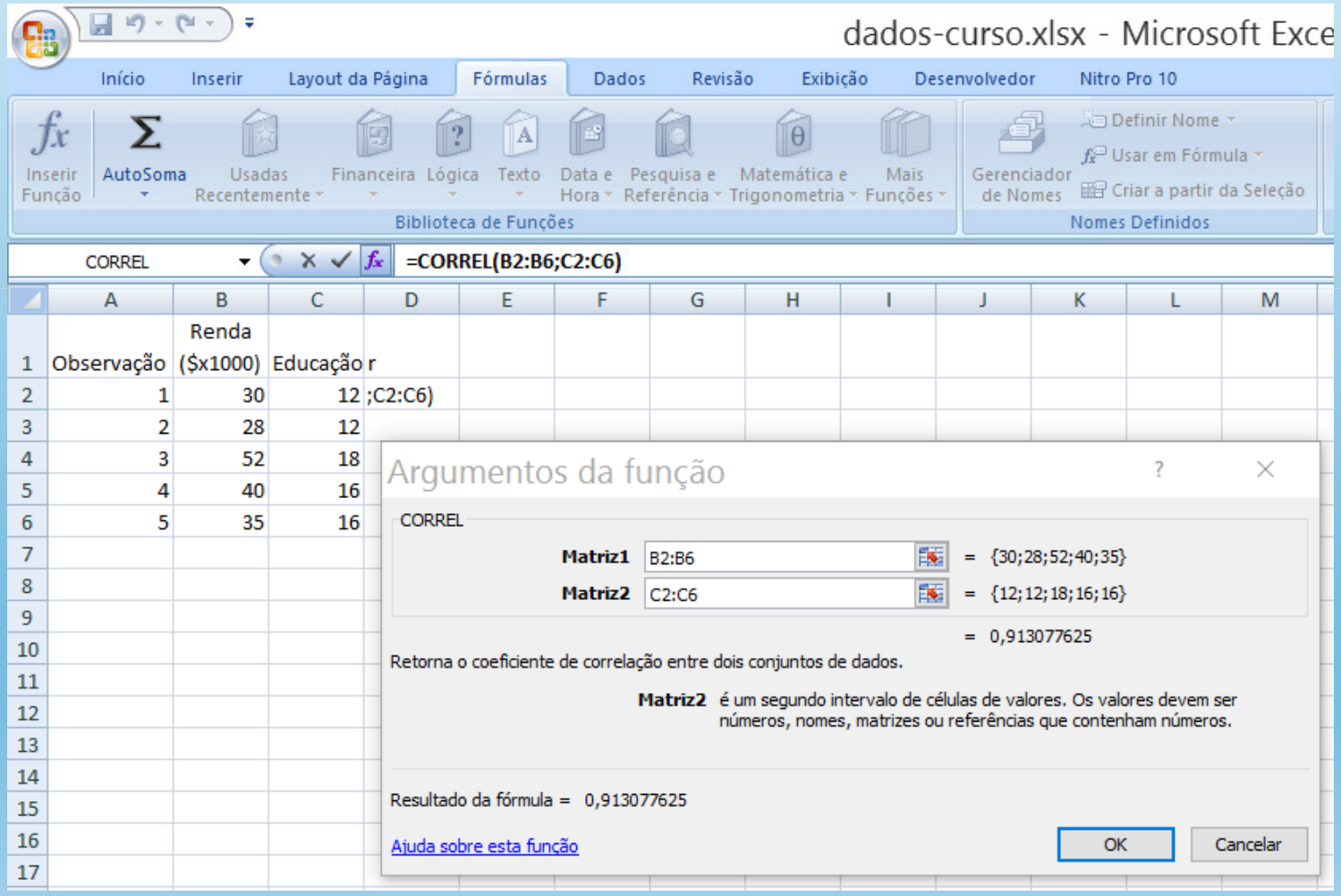
1) Clique na célula D2
2) Na barra de ferramentas, selecione:

Fórmulas – Mais Funções - Estatística - **CORREL**



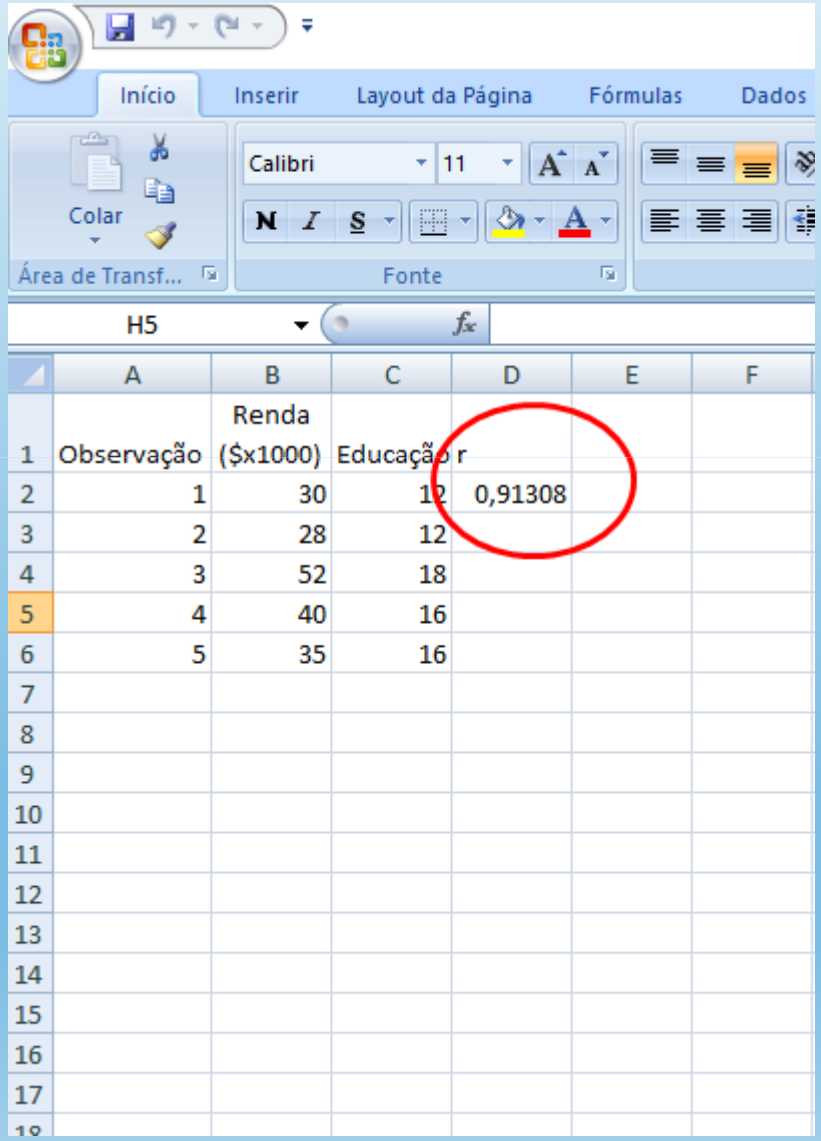
3) Na caixa que se abrirá, o campo Matriz1 deverá ser preenchido com os dados referentes à coluna com a renda, ou seja, Coluna B2:B6;

4) O mesmo procedimento deverá ser realizado para a Matriz2, porém com os dados sobre educação, Coluna C2:C6.



Aperte “OK” para finalizar

O resultado aparecerá na célula D2



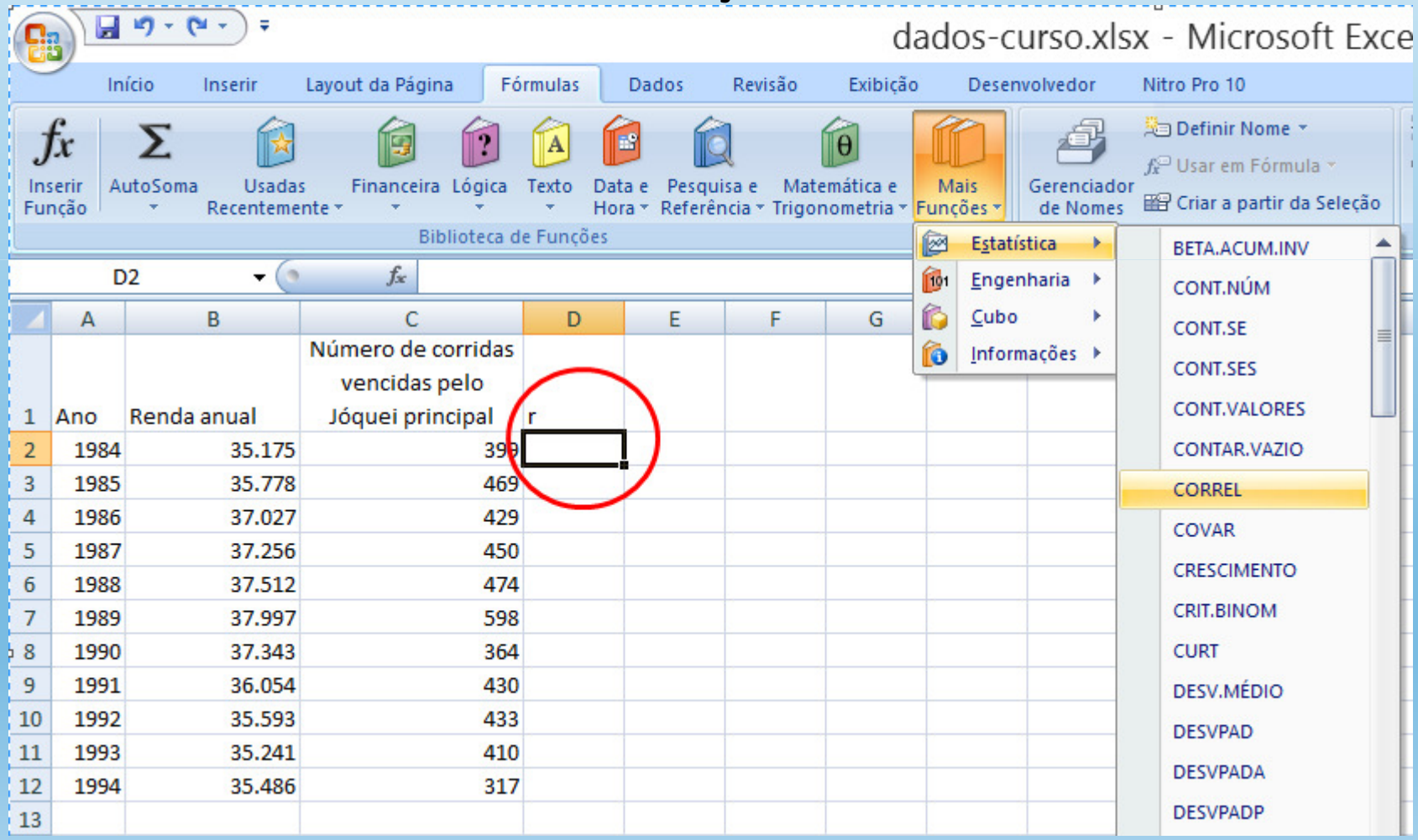
	A	B	C	D	E	F
1	Observação	Renda ($x1000)	Educação	r		
2	1	30	12	0,91308		
3	2	28	12			
4	3	52	18			
5	4	40	16			
6	5	35	16			
7						
8						
9						
10						
11						
12						
13						
14						
15						
16						
17						

EXERCÍCIO 02: Cálculo da correlação, r, para as amostras de dados de renda e número de corridas vencidas.

1) Clique na célula D2;

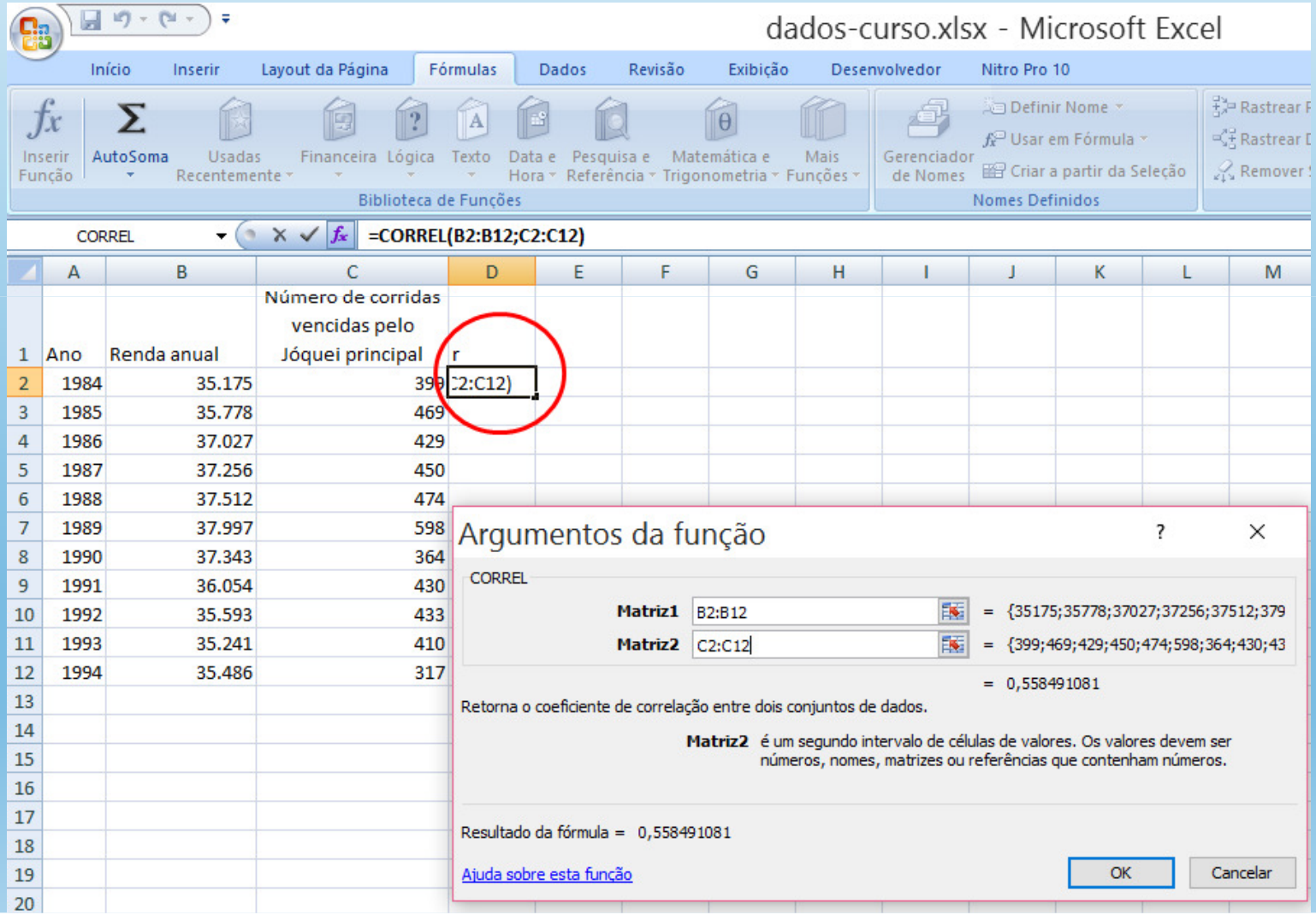
2)Na barra de ferramentas, selecione:

Fórmulas – Mais Funções - Estatística - **CORREL**

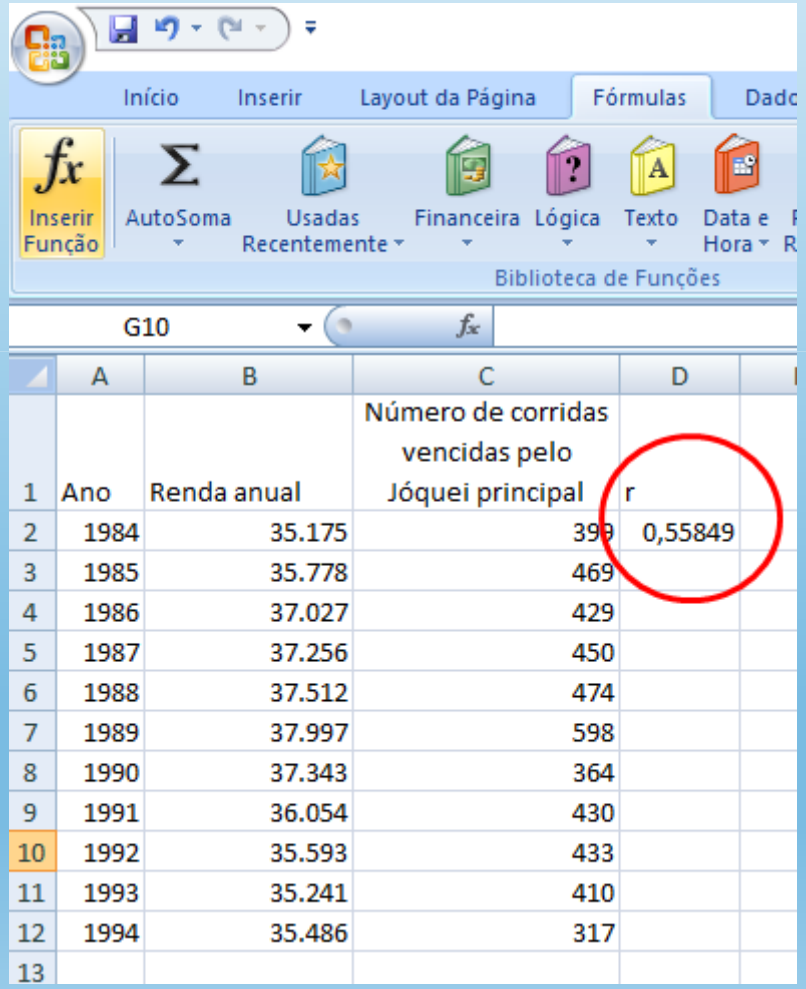


3) Na caixa que se abrirá, o campo Matriz1 deverá ser preenchido com os dados referentes à coluna com a renda, ou seja, Coluna B2:B12;

4) O mesmo procedimento deverá ser realizado para a Matriz2, porém com os dados do número de corridas, Coluna C2:C12.



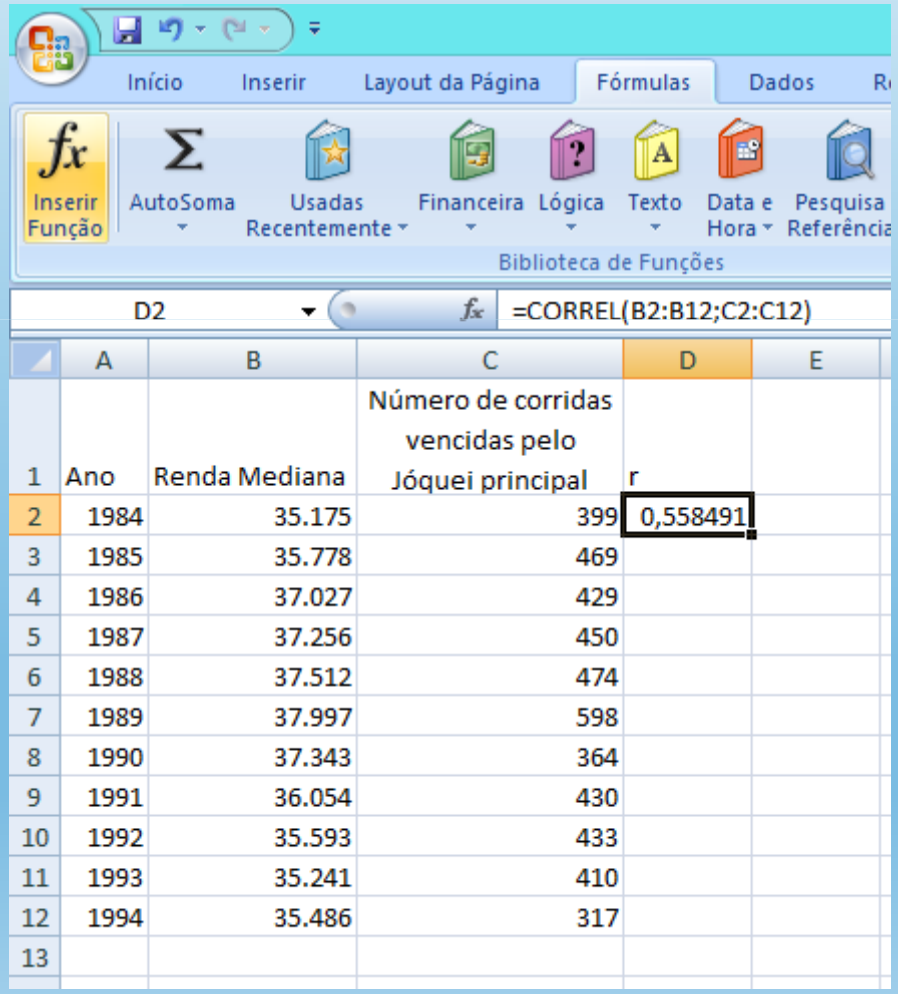
Aperte “OK” para finalizar
O resultado aparecerá na célula D2



	A	B	C	D
1	Ano	Renda anual	Número de corridas vencidas pelo Jóquei principal	r
2	1984	35.175	399	0,55849
3	1985	35.778	469	
4	1986	37.027	429	
5	1987	37.256	450	
6	1988	37.512	474	
7	1989	37.997	598	
8	1990	37.343	364	
9	1991	36.054	430	
10	1992	35.593	433	
11	1993	35.241	410	
12	1994	35.486	317	
13				

INTERPRETAÇÃO DO VALOR GERADO

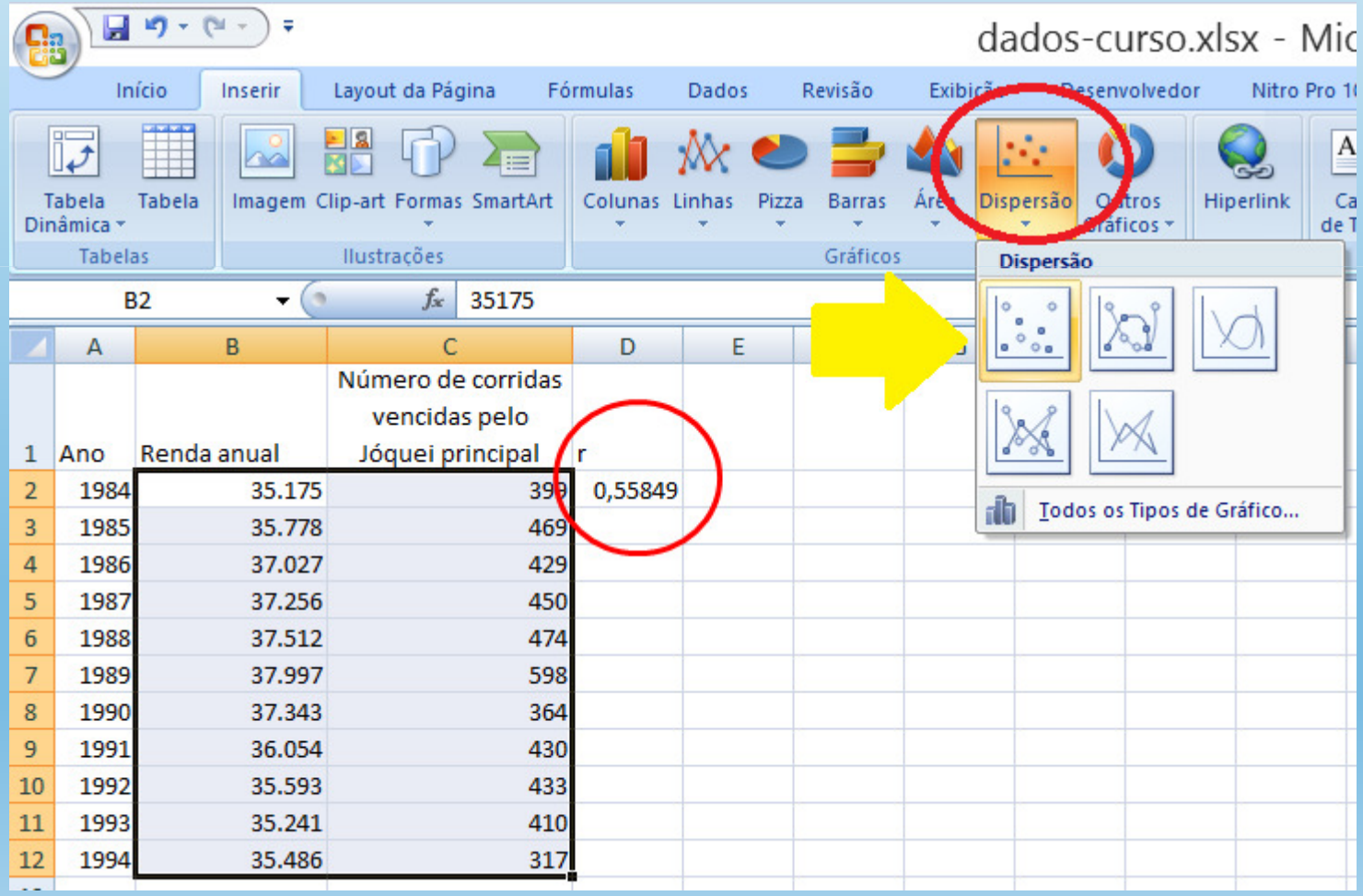
Para a série aleatória gerada nos exemplos, o valor de correlação retornado foi 0,558491



	A	B	C	D	E
1	Ano	Renda Mediana	Número de corridas vencidas pelo Jóquei principal	r	
2	1984	35.175	399	0,558491	
3	1985	35.778	469		
4	1986	37.027	429		
5	1987	37.256	450		
6	1988	37.512	474		
7	1989	37.997	598		
8	1990	37.343	364		
9	1991	36.054	430		
10	1992	35.593	433		
11	1993	35.241	410		
12	1994	35.486	317		
13					

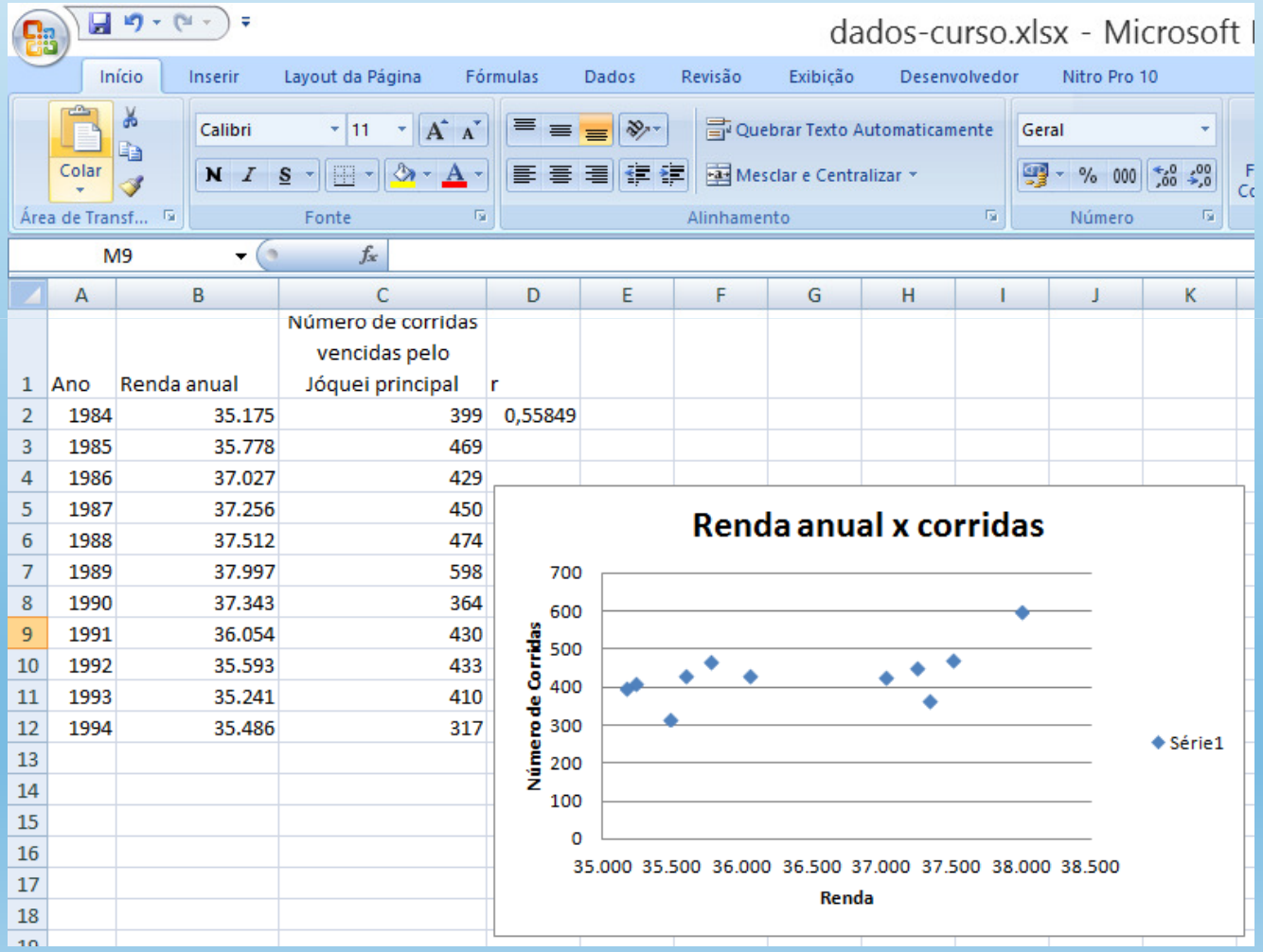
Se retornarmos à explicação anterior sobre o coeficiente de correlação, verificamos que as séries possuem alguma correlação linear positiva.

A correlação linear calculada para o exemplo anterior também pode ser expressa através de um gráfico de dispersão. Para gerá-lo, clique na Barra de ferramentas – Inserir – Dispersão (**EXEMPLO 02**)



	A	B	C	D	E
1	Ano	Renda anual	Número de corridas vencidas pelo Jóquei principal	r	
2	1984	35.175	399	0,55849	
3	1985	35.778	469		
4	1986	37.027	429		
5	1987	37.256	450		
6	1988	37.512	474		
7	1989	37.997	598		
8	1990	37.343	364		
9	1991	36.054	430		
10	1992	35.593	433		
11	1993	35.241	410		
12	1994	35.486	317		

O gráfico de dispersão é bastante útil para demonstrar a existência ou não de relações entre duas variáveis. Quanto mais alinhados estiverem os pontos à reta, maior deve ser a correlação linear entre as duas variáveis. No exemplo utilizado, as duas séries aleatórias mostram o seguinte padrão:

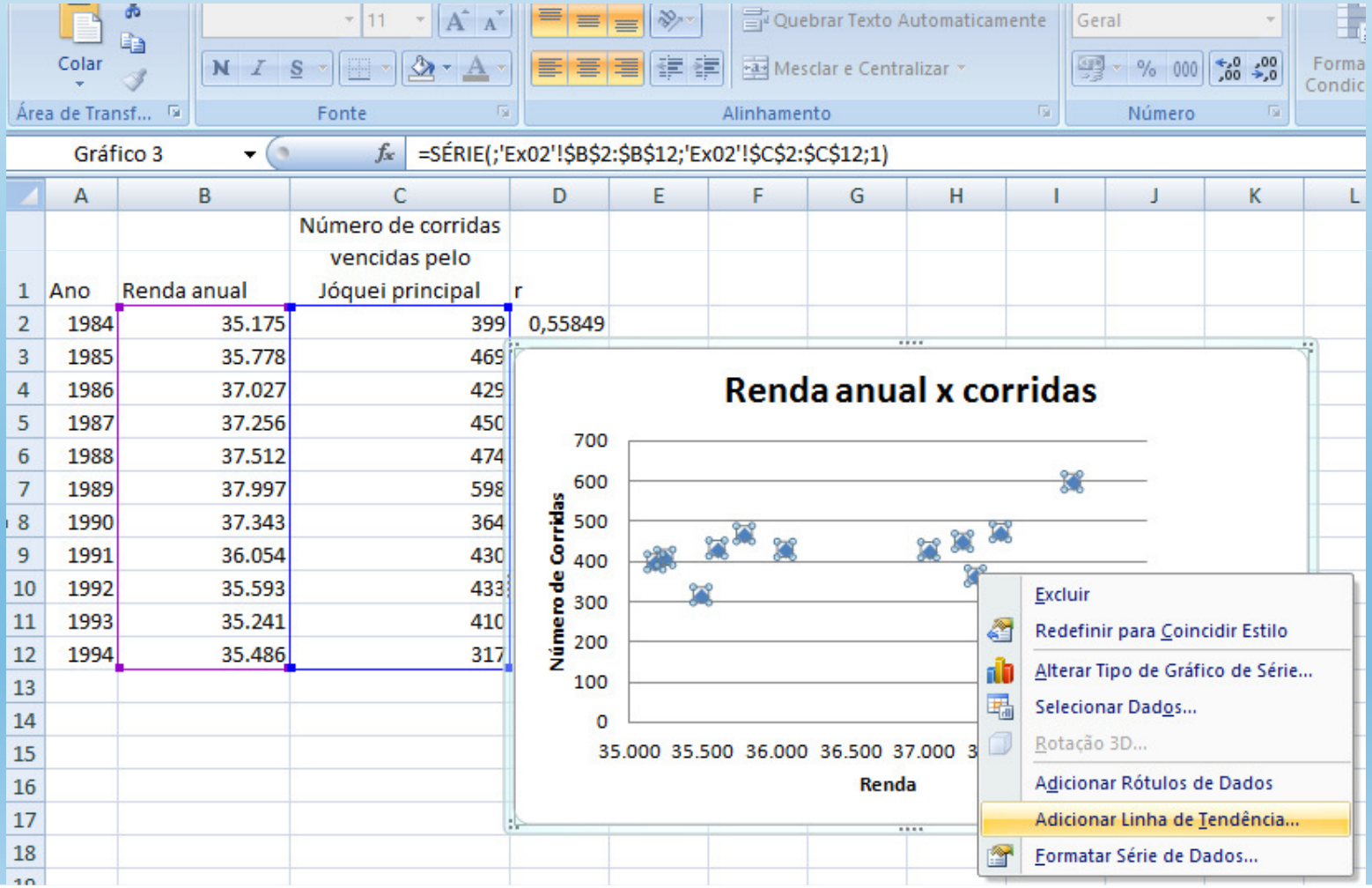


	A	B	C	D
1	Ano	Renda anual	Número de corridas vencidas pelo Jóquei principal	r
2	1984	35.175	399	0,55849
3	1985	35.778	469	
4	1986	37.027	429	
5	1987	37.256	450	
6	1988	37.512	474	
7	1989	37.997	598	
8	1990	37.343	364	
9	1991	36.054	430	
10	1992	35.593	433	
11	1993	35.241	410	
12	1994	35.486	317	

É possível, no mesmo gráfico de dispersão, inserir a reta de regressão de uma variável em relação à outra

1) Clique com o botão direito sobre um dos pontos azuis do gráfico

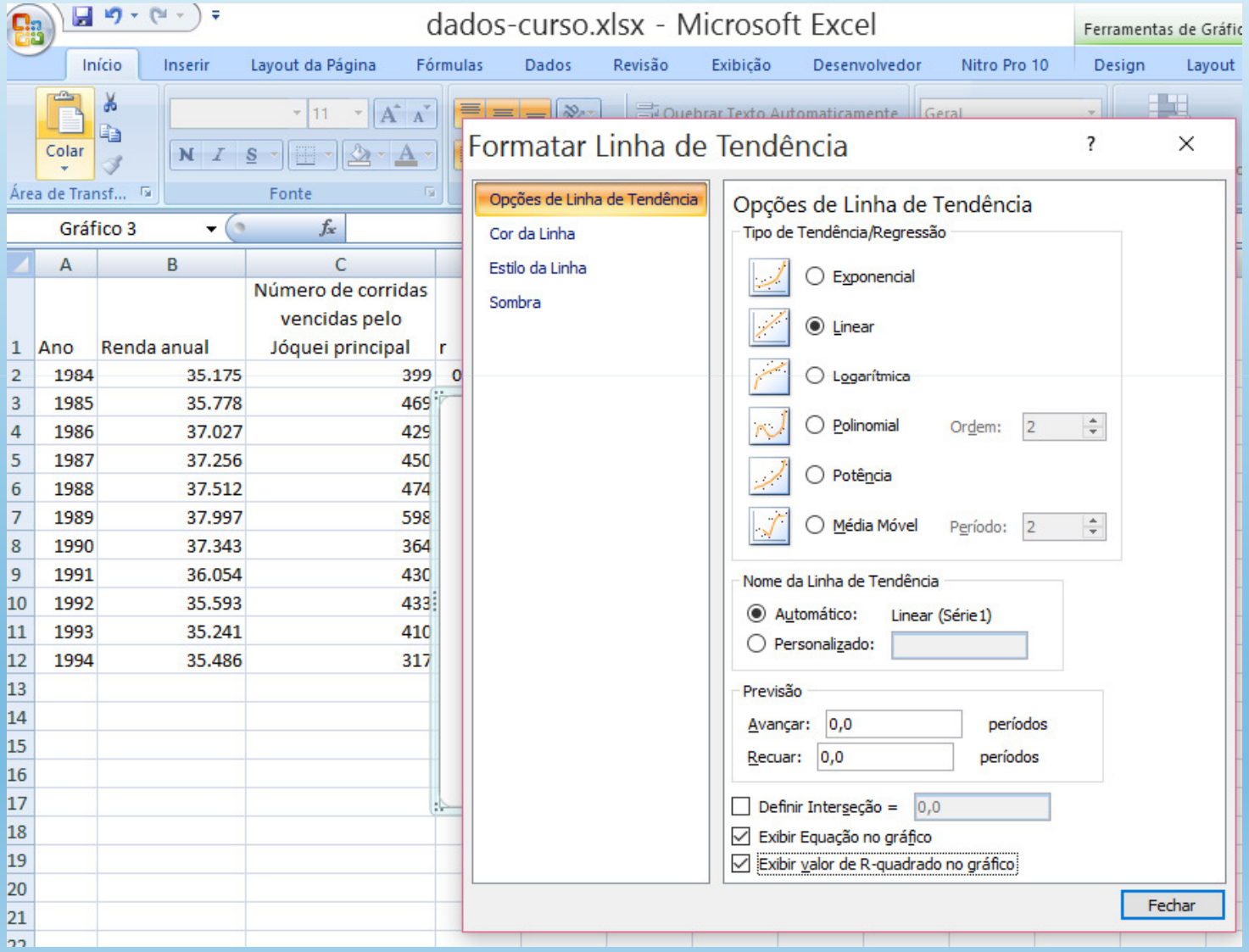
2) Selecione “Adicionar linha de tendência”



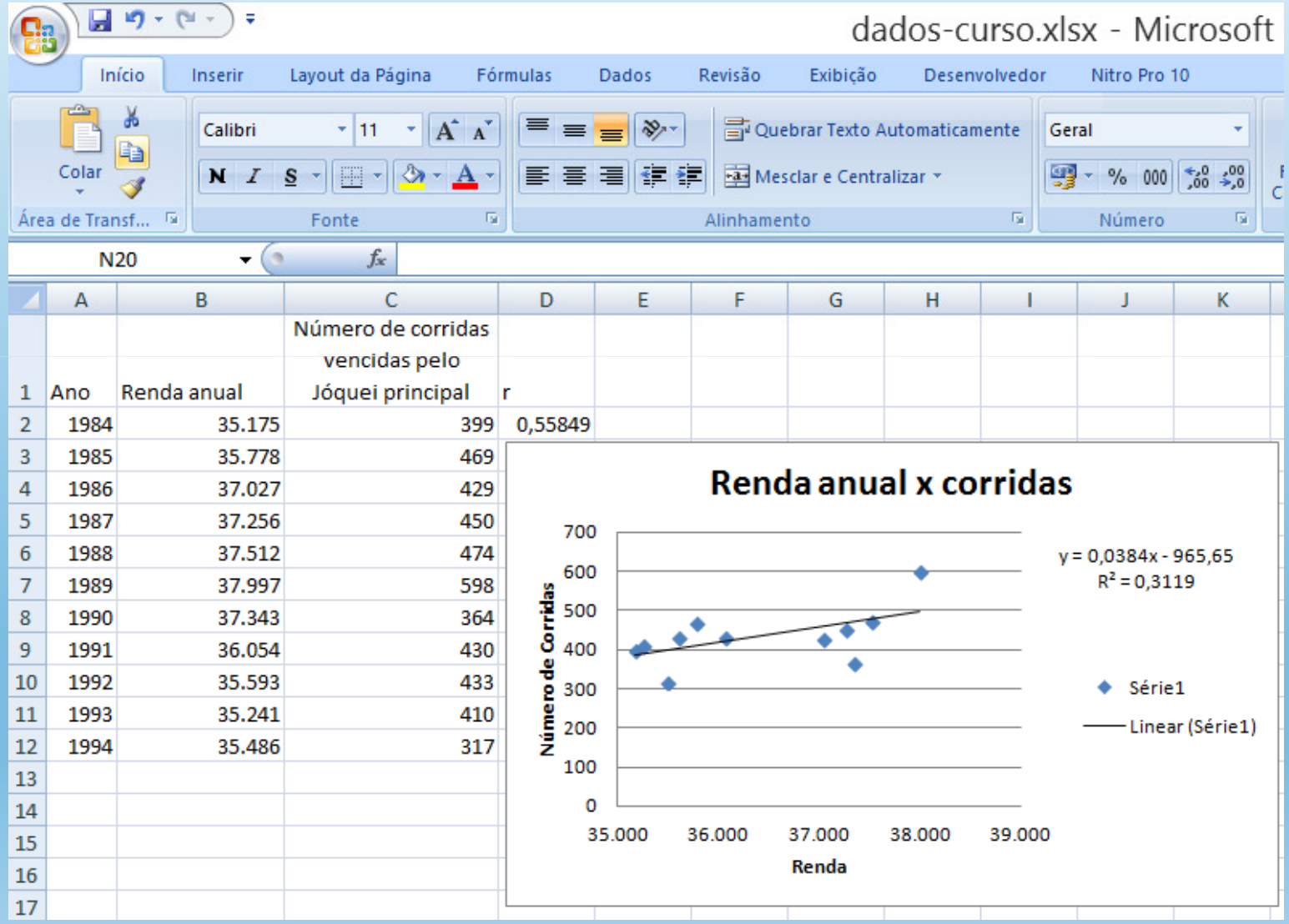
	A	B	C	D
1	Ano	Renda anual	Número de corridas vencidas pelo Jóquei principal	r
2	1984	35.175	399	0,55849
3	1985	35.778	469	
4	1986	37.027	429	
5	1987	37.256	450	
6	1988	37.512	474	
7	1989	37.997	598	
8	1990	37.343	364	
9	1991	36.054	430	
10	1992	35.593	433	
11	1993	35.241	410	
12	1994	35.486	317	

3) Escolher o tipo de ajuste, p. ex., linear

4) É possível exibir a equação da reta linear e o valor de R^2



5) Ao terminar de selecionar as opções de formato, clique em fechar;
Os resultados serão exibidos como o modelo abaixo



	A	B	C	D
1	Ano	Renda anual	Número de corridas vencidas pelo Jóquei principal	r
2	1984	35.175	399	0,55849
3	1985	35.778	469	
4	1986	37.027	429	
5	1987	37.256	450	
6	1988	37.512	474	
7	1989	37.997	598	
8	1990	37.343	364	
9	1991	36.054	430	
10	1992	35.593	433	
11	1993	35.241	410	
12	1994	35.486	317	
13				
14				
15				
16				
17				

COEFICIENTE DE DETERMINAÇÃO R^2

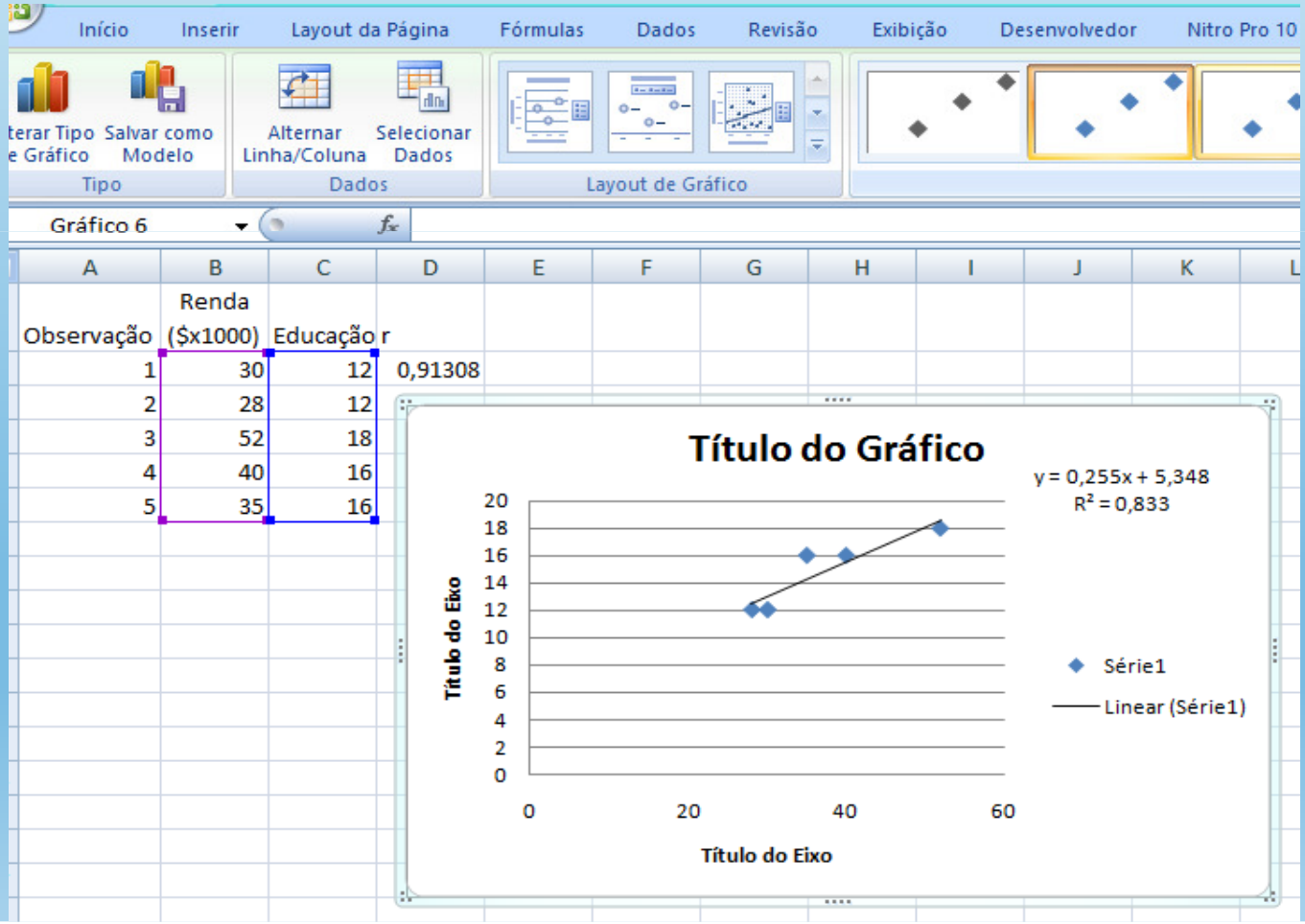
- Indica o grau do ajuste linear entre duas variáveis
- Indica o grau de dependência linear entre duas variáveis
- Se uma variável pode ser considerada como preditora em relação a outra

O que é uma variável preditora?

RETOMAR O EXEMPLO 02: Seguir os mesmos passos do exercício anterior

1) Escolha o formato do gráfico
2) Escreva o nome do gráfico
3) Coloque nome nos eixos X e Y

O Resultado final será o seguinte:



EXERCÍCIO

Obtenha dados mensais de OLR precipitação e TSM, para o período de 1980 a 2018, do CDC-NOAA, e faça os seguintes exercícios no Excel:

1) **A correlação entre a série de precipitação de OLR sobre o norte da América do Sul.**

2) **Gráfico de dispersão para as variáveis precipitação e OLR, referentes à questão 1.**

3) **Correlação linear entre a precipitação e a TSM. Precipitação na América do Sul e TSM no Pacífico Equatorial.**

4) **Gráfico de dispersão entre as variáveis precipitação e TSM referentes à questão 3.**

5) **Interprete os resultados obtidos.**

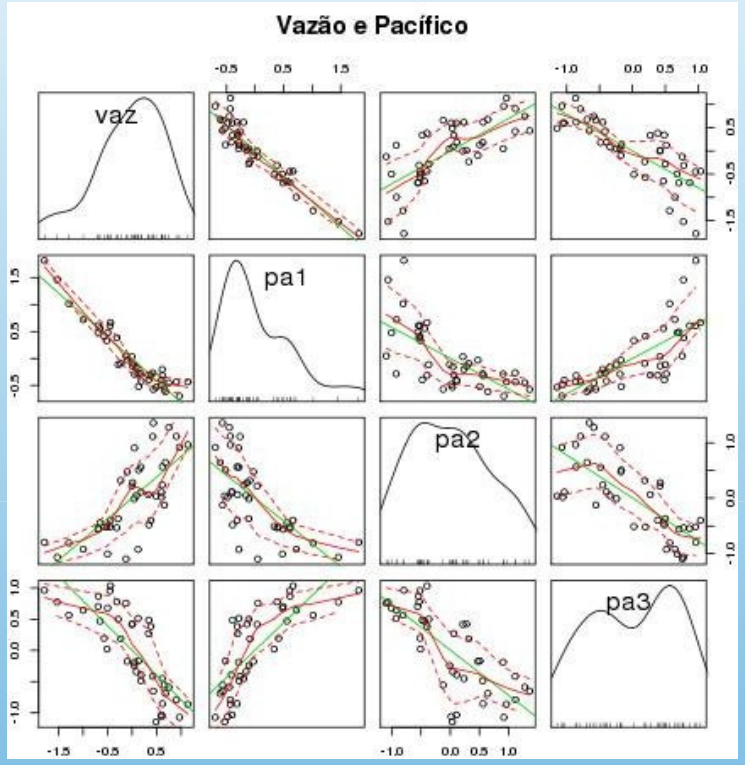
USO DE OUTROS SOFTWARES ESTATÍSTICOS

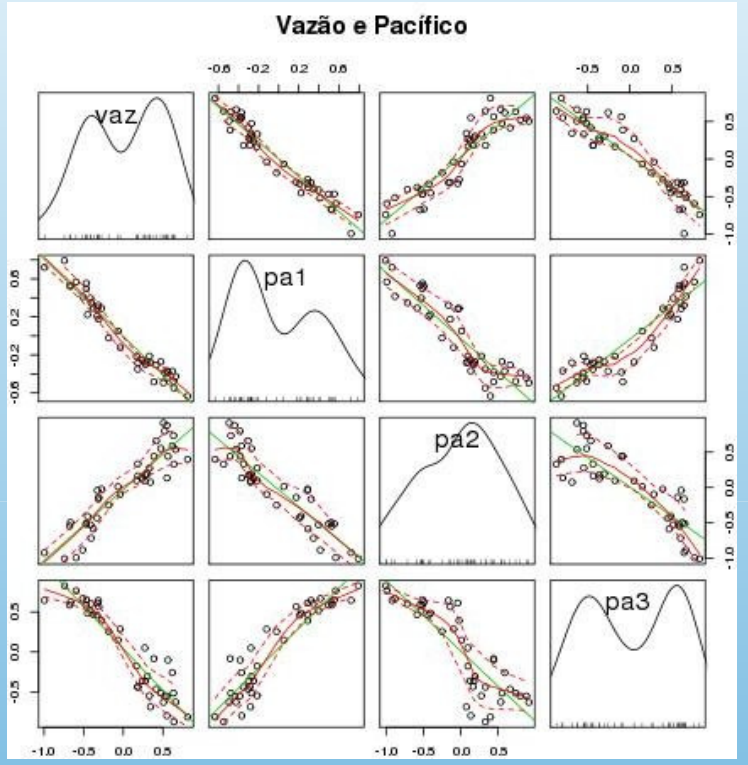
CORRELAÇÃO LINEAR

Outros softwares estatísticos, e gratuitos, tais como o **R** e o **GrADS**, são capazes de tratar séries temporais, mas também dados distribuídos espacialmente. Trazem uma série de recursos gráficos que facilitam a visualização e a geração de saídas mais elaboradas.

ex.03.r

DIAGRAMAS DE DISPERSÃO NO R



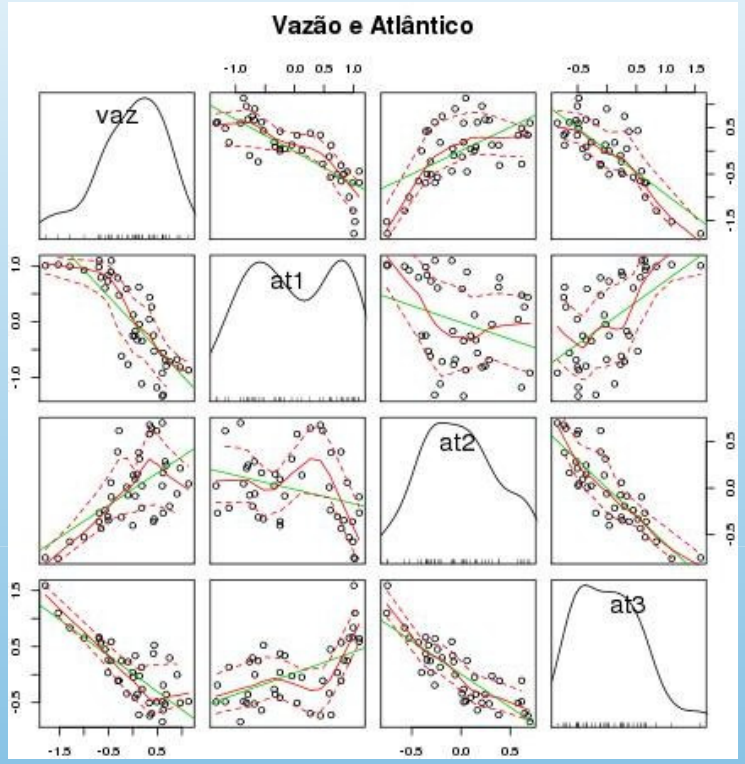


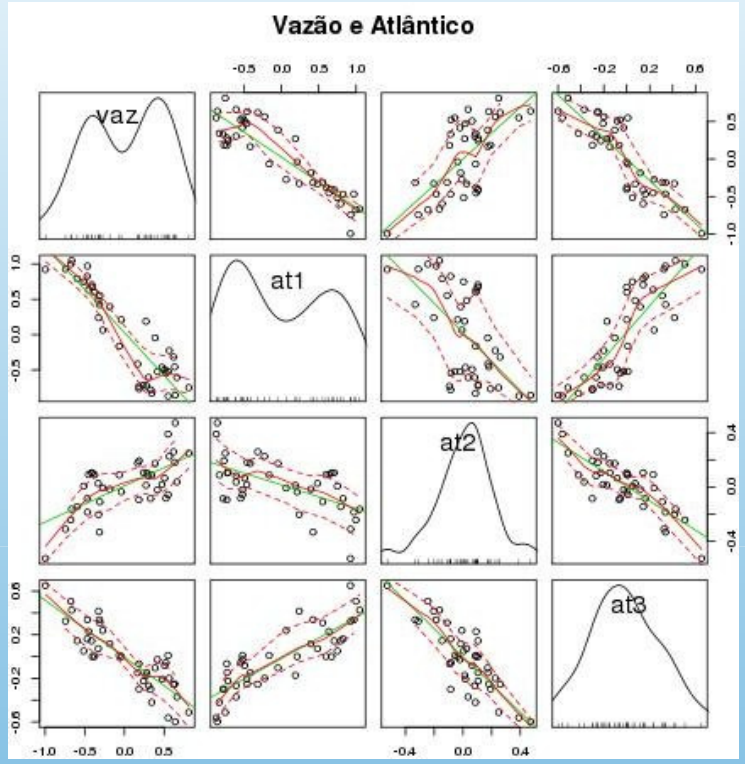
Diagramas de dispersão entre a vazão anual do rio Madeira e a TSM média nas áreas PA1, PA2 e PA3, suavizadas com média móvel (a) 6 e (b) 12 anos.
PA1 PA2 PA3 – áreas oceânicas no Pacífico

Fonte: SILVA, E.R.L.D.G. **Associação da variabilidade climática dos oceanos com a vazão de rios da Região Norte do Brasil**. Dissertação de Mestrado. São Paulo: Universidade de São Paulo. Faculdade de Filosofia, Letras e Ciências Humanas. Departamento de Geografia, 2013. 182p.

DIAGRAMAS DE DISPERSÃO NO R

ex.04.r



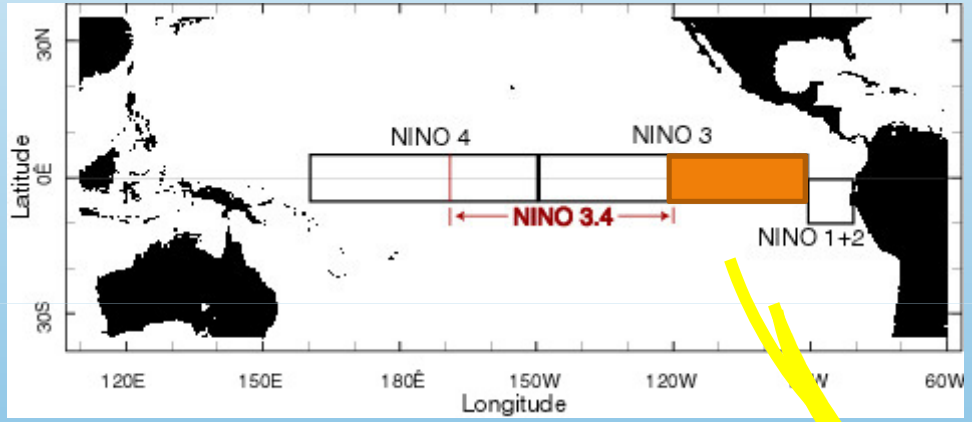


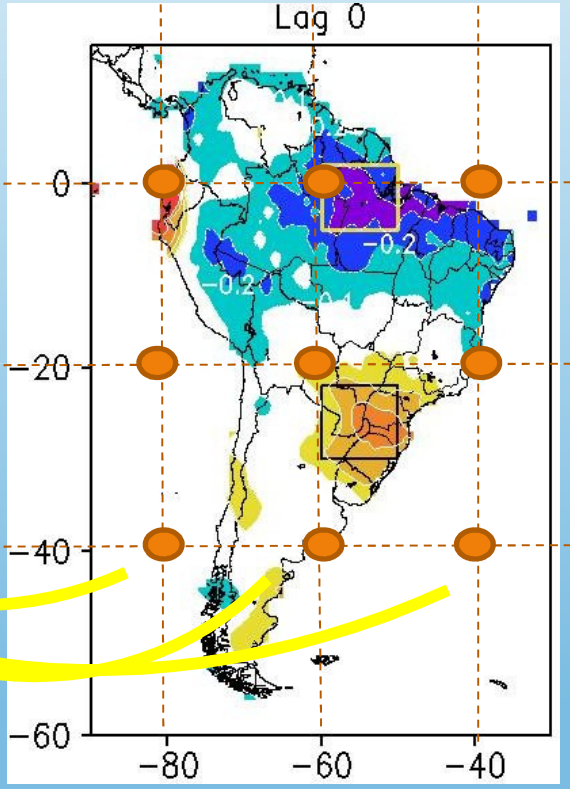
Diagramas de dispersão entre a vazão anual do rio Madeira e a TSM média nas áreas AT1, AT2 e AT3, suavizadas com média móvel (a) 6 e (b) 12 anos AT1 AT2 AT3 áreas oceânicas no Atlântico.

Fonte: SILVA, E.R.L.D.G. **Associação da variabilidade climática dos oceanos com a vazão de rios da Região Norte do Brasil**. Dissertação de Mestrado. São Paulo: Universidade de São Paulo. Faculdade de Filosofia, Letras e Ciências Humanas. Departamento de Geografia, 2013. 182p.

CORRELAÇÃO LINEAR
TSM DA REGIÃO DE NIÑO 1+2 x PRECIPITAÇÃO NA AMÉRICA DO SUL

ex.06.gs





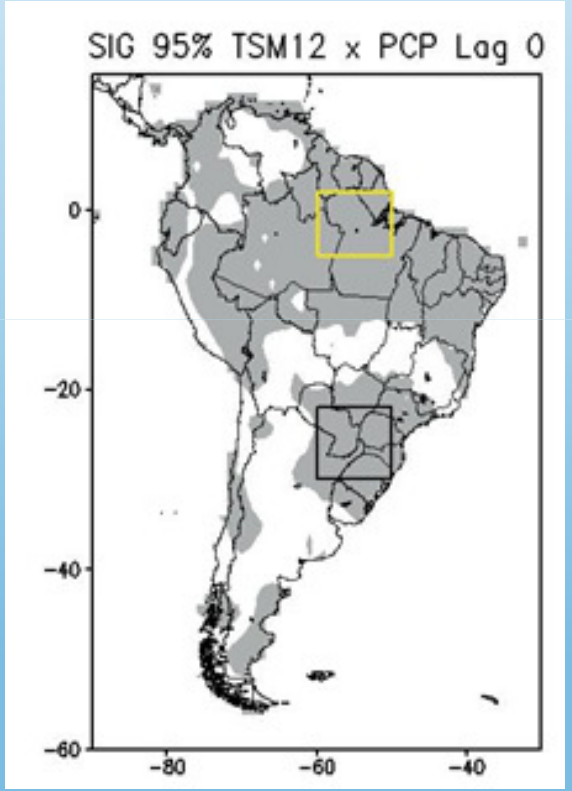
Os valores mensais de TSM das regiões de *Niño* foram correlacionados com os valores da precipitação na América do Sul

Fonte: SILVA, ERLD e SILVA, MES (2015) Memória de eventos ENOS na precipitação da América do Sul. Revista do Departamento de Geografia

SIGNIFICÂNCIA ESTATÍSTICA

A significância estatística do cálculo do coeficiente de correlação foi avaliada com a aplicação do teste *t-Student*, cujo valor limite para se considerar o cálculo significativo é definido, segundo Costa Neto (1977), por:

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



SIGNIFICÂNCIA ESTATÍSTICA

É um valor que expressa a confiabilidade estatística referente a uma estatística.

média
correlação
tendência linear

Por exemplo, r = 0,6 é um valor estatisticamente confiável de correlação linear para os dados usados?

SIGNIFICÂNCIA ESTATÍSTICA

- Esta pergunta deve ser feita para fornecer alguma garantia relativa ao valor obtido para determinada estatística, que indique que o valor não advém da aleatoriedade.

- Esta garantia pode ser expressa através de níveis de confiança:

90%, 95%, 99%

são níveis de confiança usados corriqueiramente.

SIGNIFICÂNCIA ESTATÍSTICA

- Existem alguns testes de significância mais usados:
  teste t-Student (supõe a distribuição normal dos dados)

- Para tanto, precisamos saber qual é a quantidade de valores usados no cálculo da estatística (n) e qual é o valor obtido da estatística (r, no caso do coeficiente de correlação)

para coeficiente de correlação

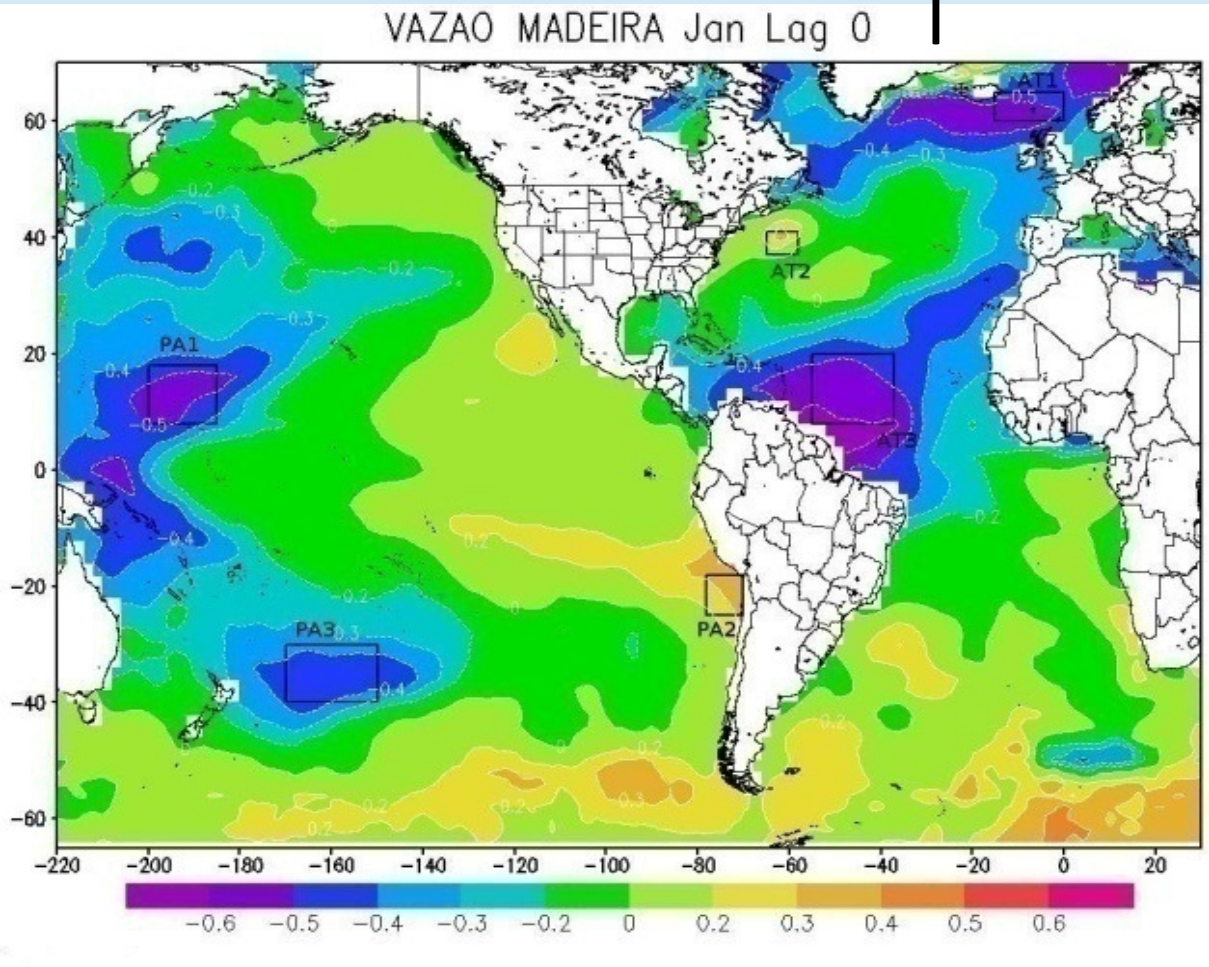
tc = $$t_{n-2} = r\frac{\sqrt{n-2}}{\sqrt{1-r^2}}$$

t > tc → cálculo estatisticamente significativo

t < tc → cálculo não é estatisticamente signif.

ex.06.gs

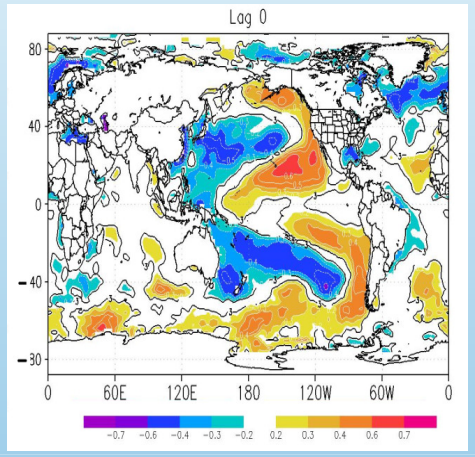
CORRELAÇÃO LINEAR ESPACIAL
TSM GLOBAL X VAZÃO DO RIO MADEIRA

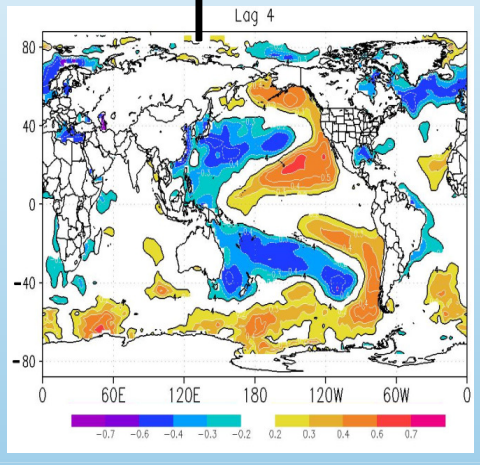


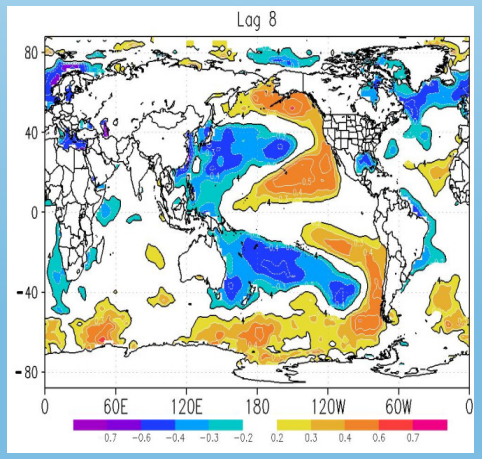
Qual a interpretação que pode ser feita do mapa ao lado?

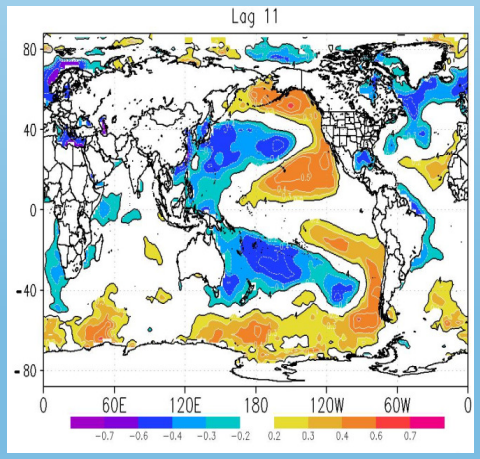
Fonte: SILVA, E.R.L.D.G. **Associação da variabilidade climática dos oceanos com a vazão de rios da Região Norte do Brasil**. Dissertação de Mestrado. São Paulo: Universidade de São Paulo. Faculdade de Filosofia, Letras e Ciências Humanas. Departamento de Geografia, 2013. 182p.

CORRELAÇÃO LINEAR ESPACIAL
TSM GLOBAL X VAZÃO NO PANTANAL





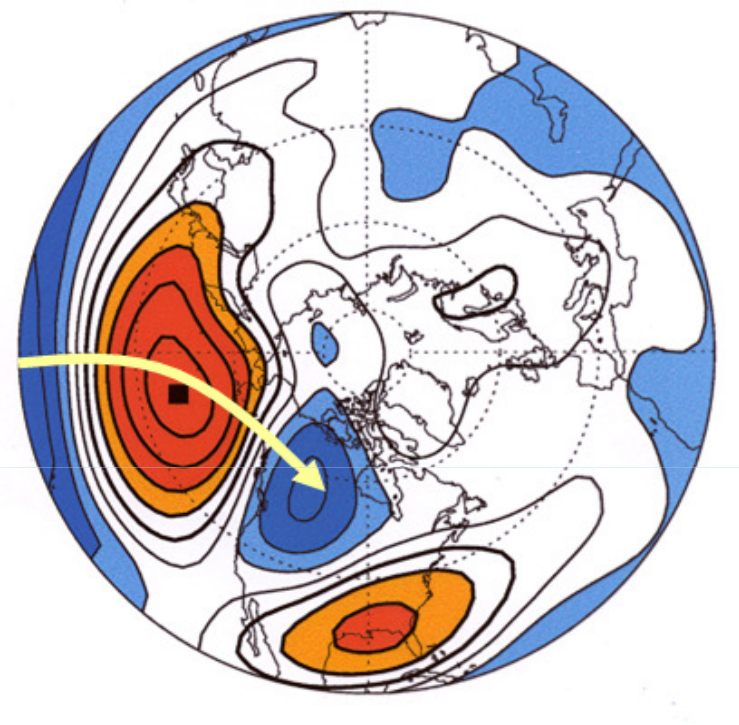




Qual a interpretação que pode ser feita dos mapas ao lado?

Lagged linear correlation between Pantanal discharge and SST monthly data for the period 1970-2003, for (a) lag=0, (b), lag=4 (c) lag=8 and (d) lag=11 months. The first month in SST time series is always January. The statistical significant areas at 99% (t-Student test) are given by the black lines. (Silva et al., 2016 TAAC)

ALTURA GEOPOTENCIAL 500 mb



Qual o padrão que pode ser observado através da correlação da altura geopotencial em 500 mb com o valor no Pacífico Norte? (R. PNA)

Spatial distribution of correlation of the 500 mb geopotential height anomaly time series (Seasonal JFM) at all points on the Northern hemisphere with the time series at a specified “base point” - North Pacific. Red colors positive correlation, blue colors negative correlation. Yellow arrow indicate meridional orientation of spatial structure existing in the correlation pattern. Picture courtesy of Prashant Sardeshmukh, CDC/OAR