



Serviço Geológico do Brasil – CPRM

Correlação e Regressão Linear

Aula 01 : Correlação

Eber José de Andrade Pinto
Coordenador Executivo do DEHID
eber.andrade@cprm.gov.br
www.cprm.gov.br

Belo Horizonte, 19 de outubro de 2020

Livro Texto

HIDROLOGIA ESTATÍSTICA vem preencher significativa lacuna na literatura técnica especializada em língua portuguesa no campo dos recursos hídricos. O conhecimento das ferramentas de estatística é fundamental para a evolução e para a prática da Hidrologia, onde encontra diversificada gama de aplicações nas atividades rotineiras ligadas aos estudos e projetos de engenharia hidrológica, que necessitam das teorias probabilísticas para a sua solução.

Conhecer e investigar as variáveis do meio físico são atributos comuns entre os conceitos aqui registrados e o Serviço Geológico do Brasil – CPRM. O livro apresenta o material didático capaz de orientar a pesquisa, e, com essa iniciativa, a instituição amplia a visibilidade do seu papel de agente promotor dos levantamentos hidrológicos básicos no país.

HIDROLOGIA ESTATÍSTICA é publicação dirigida para os profissionais do setor, bem como para a formação de alunos de graduação e pós-graduação. Municia o leitor com princípios introdutórios, análise de dados, teoria das probabilidades, variáveis aleatórias discretas e contínuas, análise de frequência, correlação e regressão. Destaca também técnicas mais sofisticadas de tratamento, manipulação e representação de dados estatísticos, com exemplos práticos reais e selecionados da rede hidrometeorológica operada pela CPRM.

www.cprm.gov.br



Secretaria de
Geologia, Mineração e
Transformação Mineral

Ministério de
Minas e Energia



Período Contemporâneo



ANO INTERNACIONAL DO PLANETA TERRA - 2006



AGOSTO
DE 2007



Hidrologia Estatística

MAURO NAGHETTINI
ÉBER JOSÉ DE ANDRADE PINTO



2007

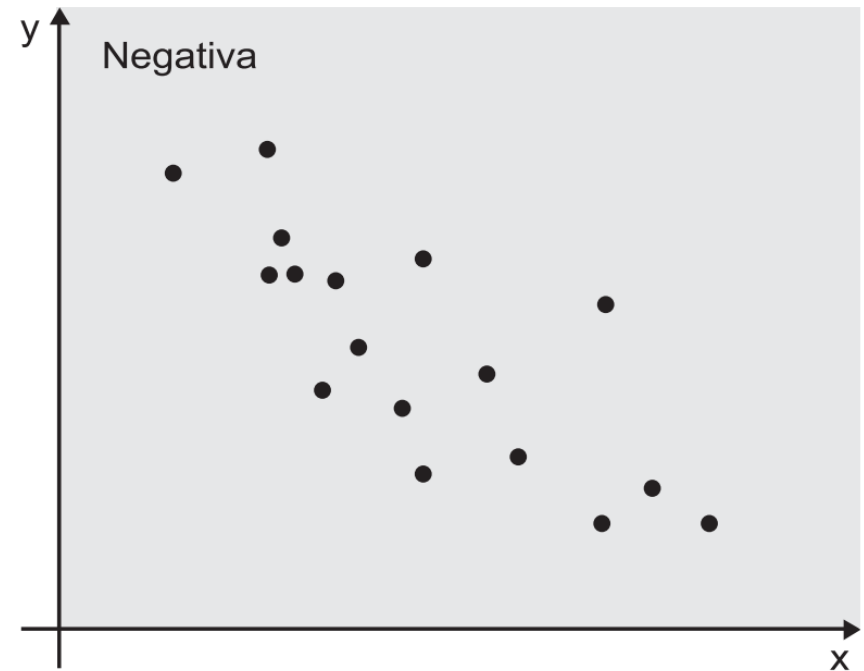
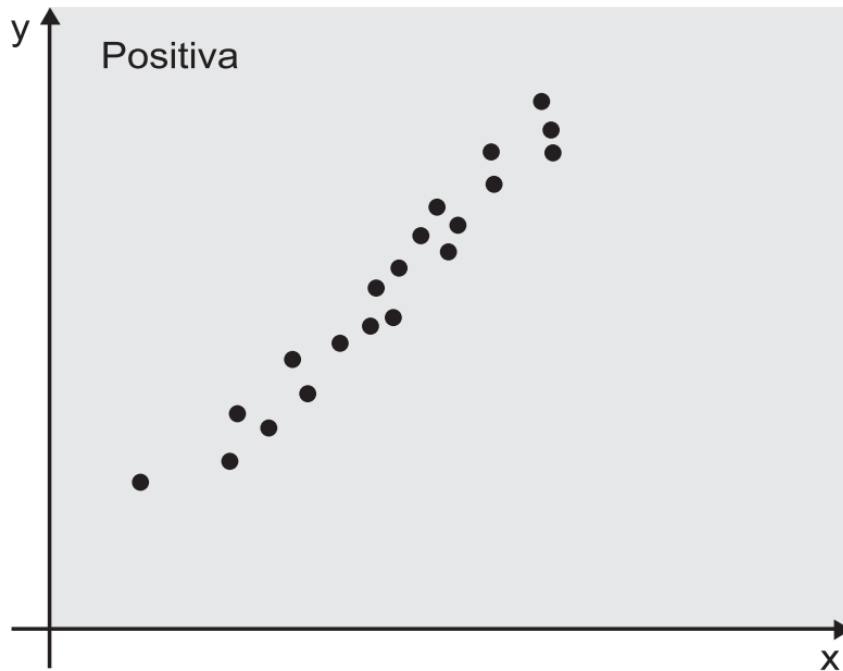
Introdução

Questões típicas da Engenharia de Recursos Hídricos referentes à associação correlativa entre duas ou mais VA's:

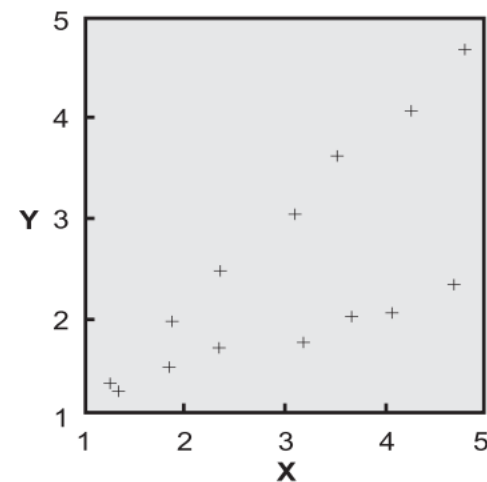
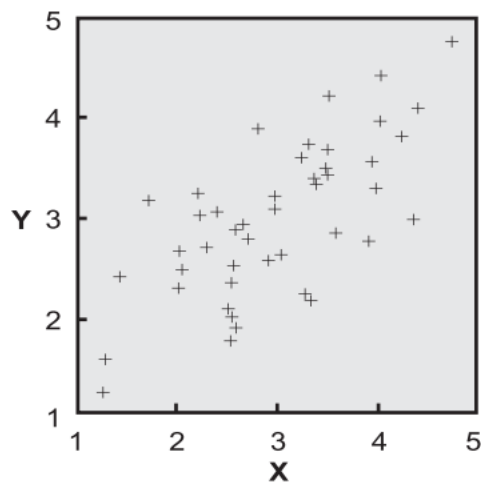
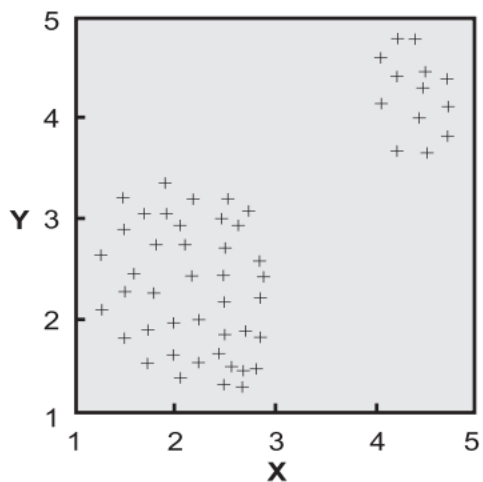
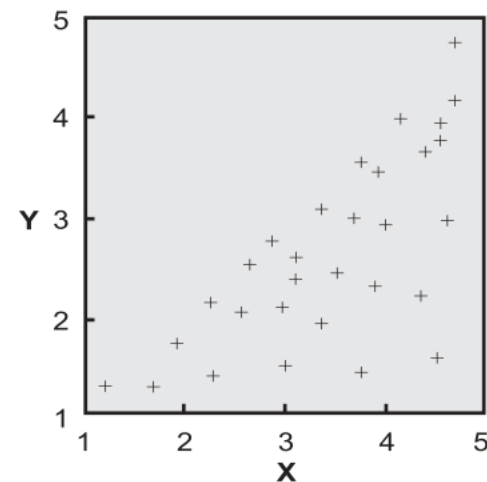
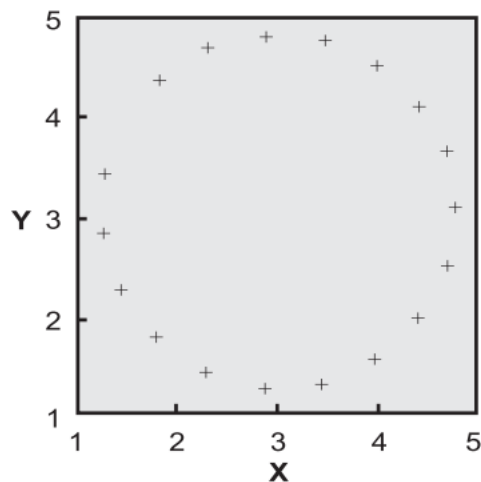
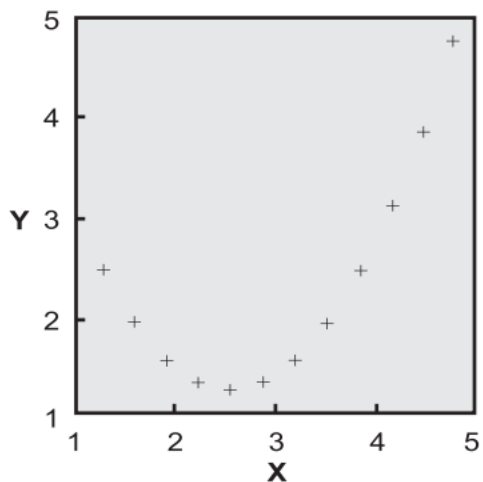
- (i) intensidades, as durações e as frequências das precipitações intensas;
 - (ii) vazões médias anuais e áreas de drenagem;
 - (iii) relação entre séries de cotas e vazões de duas estações próximas.
 - (iv) alturas anuais de precipitação e altitudes dos postos pluviométricos; ou
 - (v) definição da curva-chave de uma estação fluviométrico
-

Objetivo 1: Correlação

Analisar o **comportamento simultâneo** das variáveis, tomadas duas a duas, verificando eventuais associações positivas ou negativas entre elas.



Correlação: Abordagem Exploratória



Correlação: Abordagem Quantitativa

Coeficiente de correlação linear $\rho_{X,Y} = \frac{\sigma_{X,Y}}{\sigma_X \cdot \sigma_Y}$

$\rho = 0$ → nenhuma associação linear

$\rho = 1$ → associação positiva linear perfeita

$0 < \rho < 1$ → associação positiva parcial

$\rho = -1$ → associação negativa linear perfeita

$-1 < \rho < 0$ → associação negativa parcial

Comentários:

Forte associação correlativa \neq causa-efeito

- As evidências de relações causais devem ser obtidas a partir do **conhecimento** dos processos envolvidos
 - Exemplo: A alta correlação entre as vazões máximas de duas bacias vizinhas não significa que a mudança da vazão de uma delas é causada pela alteração da outra (**fatores comuns às duas bacias**).
-

Grau de Associação entre as VA's

Coeficiente de Correlação Linear de Pearson:
base → covariância normalizada

$$r = \frac{\text{cov}(x, y)}{s_x s_y} \Rightarrow -1 \leq r \leq 0 \text{ ou } 0 \leq r \leq +1$$

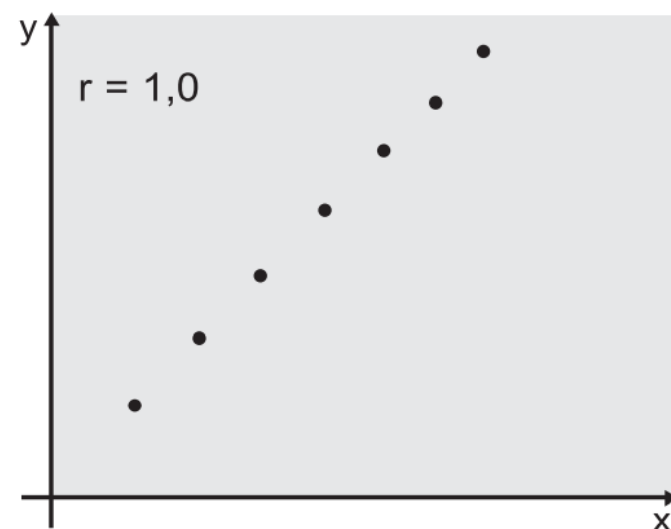
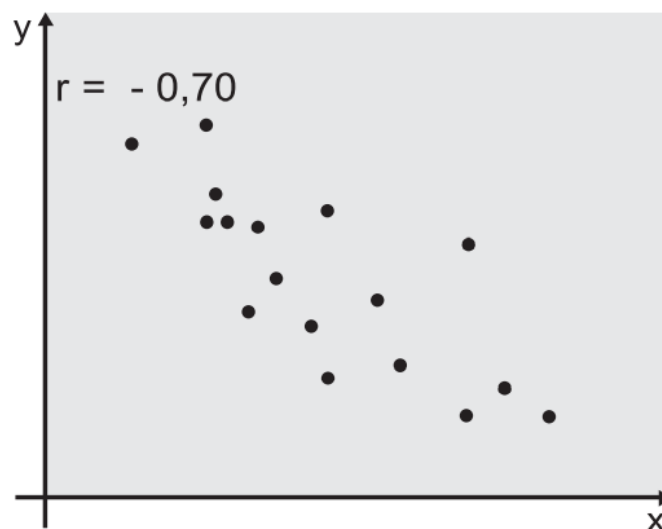
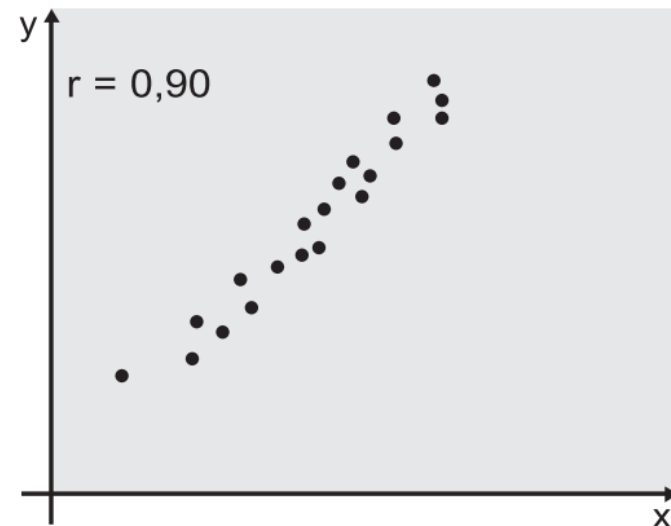
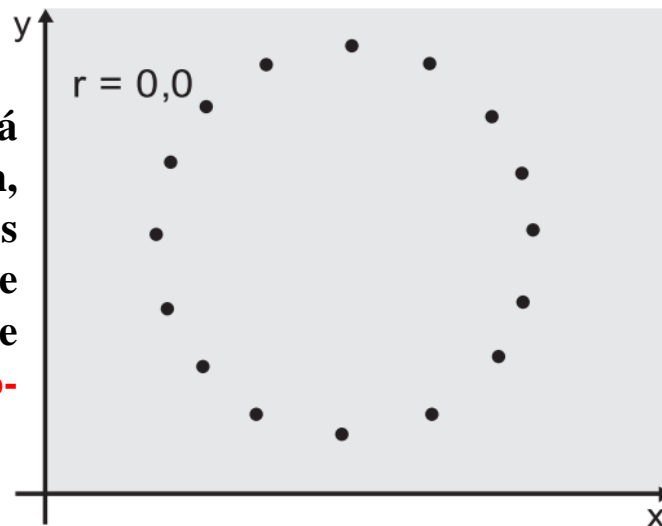
$$s_y = \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1}}$$

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

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

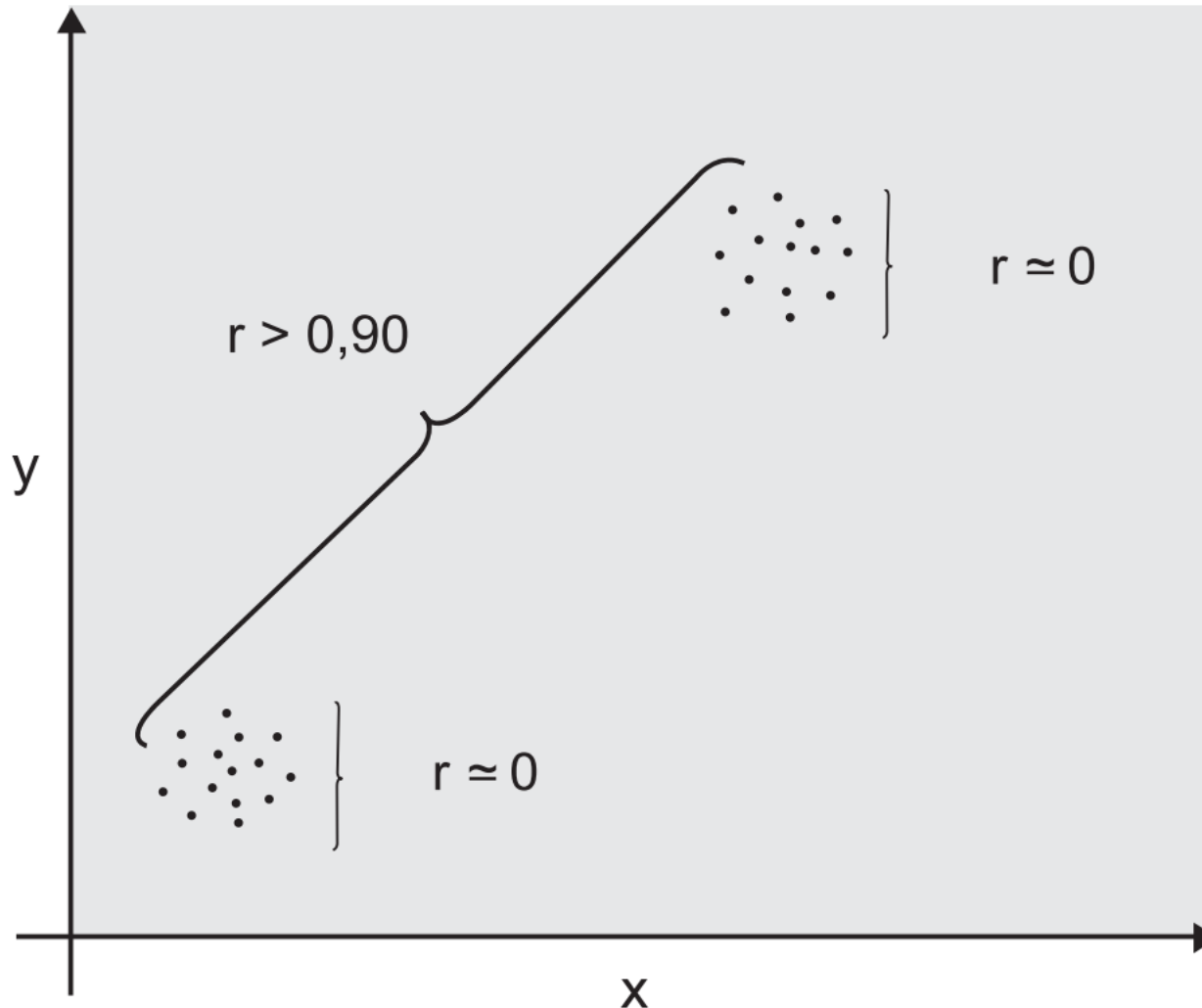
Grau de Associação entre as VA's

Atenção: se $r = 0$, não há correlação linear. Porém, isso não implica que as VA's sejam estatisticamente independentes, pois pode haver **dependência não-linear**.



Correlação Espúria

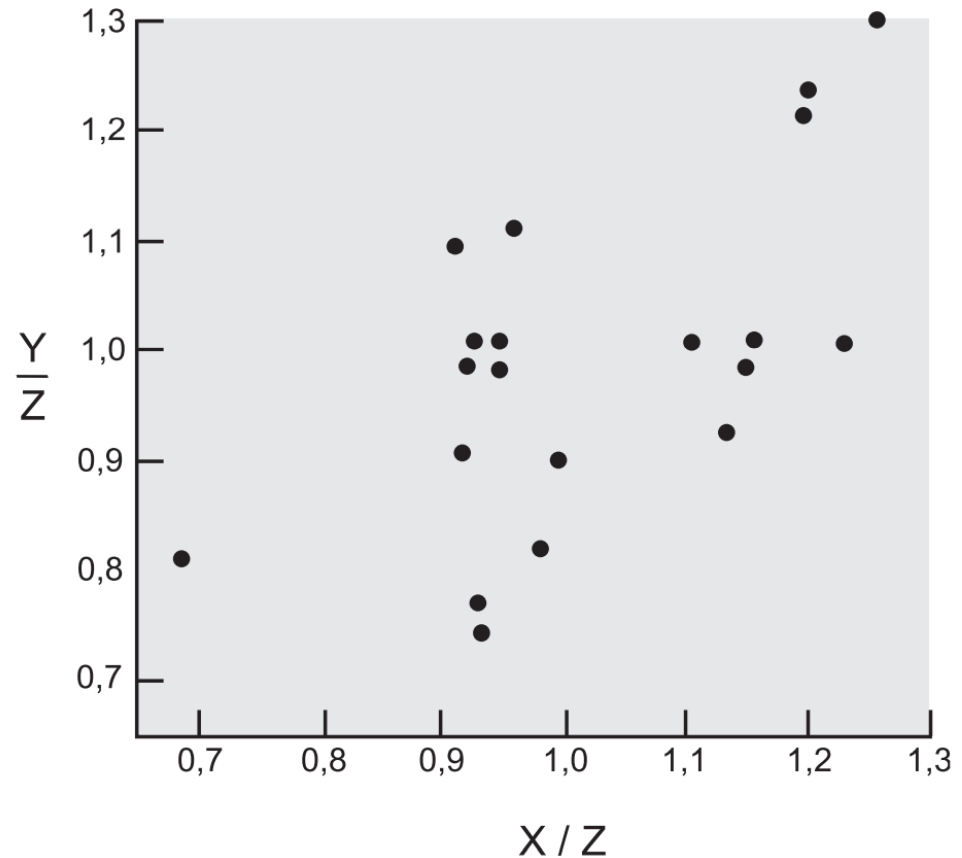
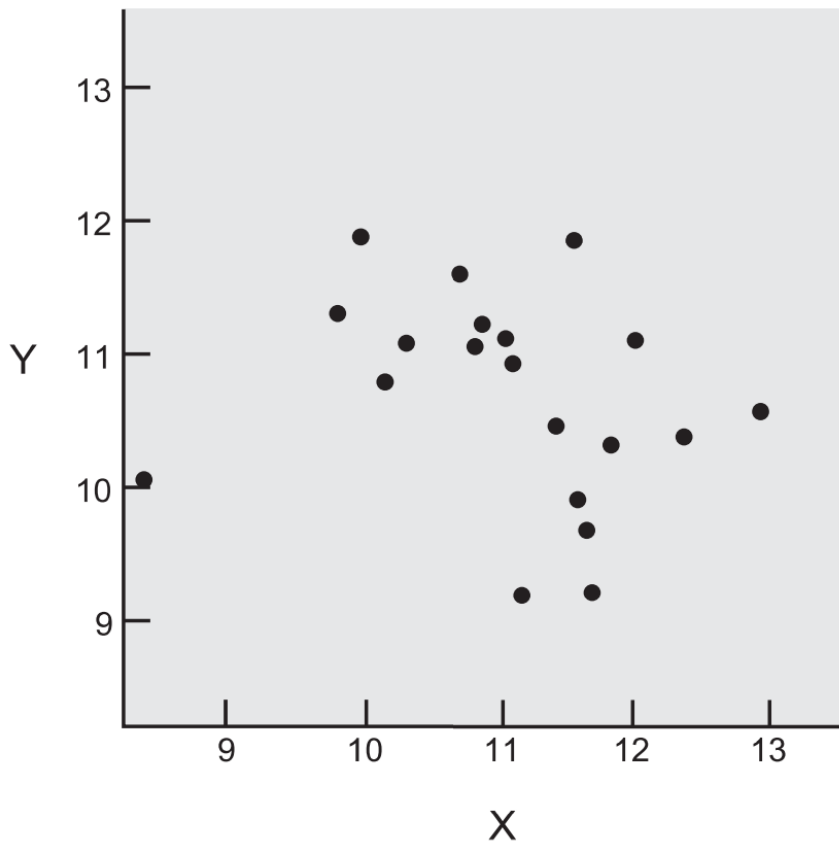
Coeficiente de correlação alto para VA's não correlacionadas



Distribuição Não-Equilibrada dos Pontos Amostrais

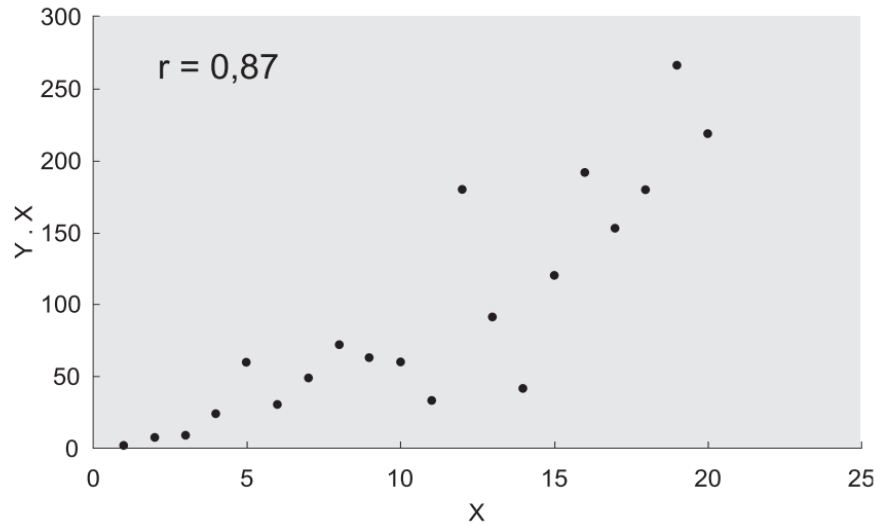
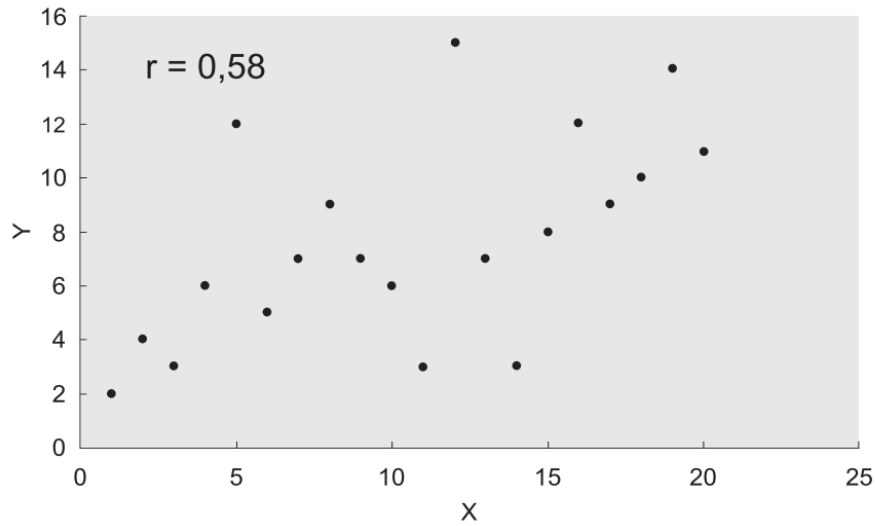
Correlação Espúria

Quocientes de VA's (mesmo denominador)



Correlação Espúria

Produtos de VA's (fator dependente)



- Benson, M. A. _1965_. "Spurious correlations in hydraulics and hydrology." *J. Hydr. Div.*, 91_4_, 35–42.
- Berges, J. A. _1997_. "Ratios, regression statistics, and 'spurious' correlations." *Limnol. Oceanogr.*, 42_5_, 1006–1007.
- Brett, M. T. _2004_. "When is a correlation between non-independent variables 'spurious?'" *Oikos*, 105, 647–656
- Hao, O. J., and Neethling, J. B. _1987_. "Effect of ratio correlation on data interpretation." *J. Environ. Eng.*, 113_1_, 205–211.
- Kenney, B. C. _1982_. "Beware of spurious self-correlations!" *Water Resour. Res.*, 18_4_, 1041–1048.
- Shivers, D. E. and Moglen, G. E (2008). Spurious Correlation in the USEPA Rating Curve Method for Estimating Pollutant Loads. *J. Environ. Eng.*, 2008, 134(8): 610-618

Testes de Hipóteses sobre o Coeficiente de Correlação

Premissa:

os resíduos $\varepsilon_i = [Y_i - (\alpha + \beta X_i)] \sim N(0, \sigma_\varepsilon)$ com σ_ε constante

Teste 1: (bilateral)

Hipótese Nula: $H_0 : \rho = 0$

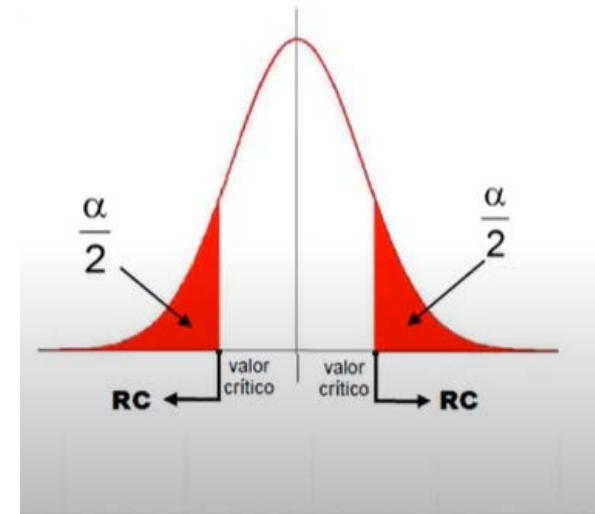
Hipótese Alternativa: $H_1 : \rho \neq 0$

Estatística de Teste: $t_0 = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$

a qual, sob H_0 , segue t de Student com $(n-2)$ gl

Decisão: Rejeitar H_0 se $|t_0| > t_{\alpha/2, n-2}$

onde, $t_{\alpha/2, n-2}$ é o valor crítico para a estatística do teste para um nível de significância, com $(n-2)$ graus de liberdade.



Testes de Hipóteses sobre o Coeficiente de Correlação

Teste 2: (bilateral)

Hipótese Nula: $H_0 : \rho = \rho_0$

Hipótese Alternativa: $H_1 : \rho \neq \rho_0$

Estatística de Teste:

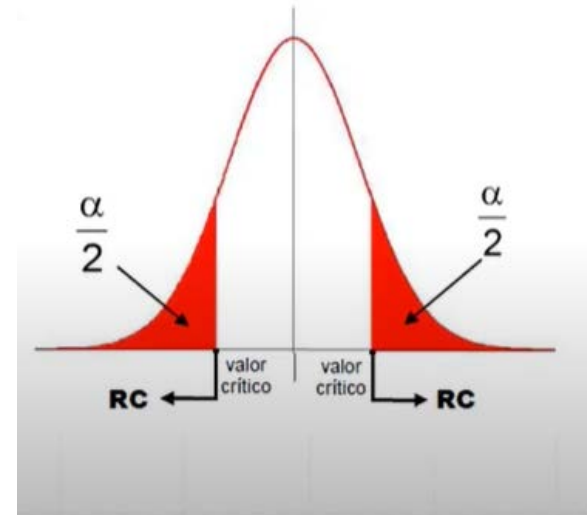
$$Z_0 = [\arctan h(r) - \arctan h(\rho_0)] \cdot (n - 3)^{1/2}$$

a qual, sob H_0 e $n > 25$, segue uma distribuição Normal com

$$\mu_Z = \arctan h(\rho) = \frac{1}{2} \ln \left(\frac{1 + \rho}{1 - \rho} \right) \quad \sigma_Z^2 = (n - 3)^{-1}$$

Decisão: Rejeitar H_0 se $|Z_0| > Z_{\alpha/2}$

onde, $Z_{\alpha/2}$ é o valor crítico para a estatística do teste dado pela distribuição normal padrão associada a um nível de significância α



Intervalo de Confiança

É possível construir um intervalo de confiança, $100.(1 - \alpha)$, para ρ dado por

$$\tanh\left[\arctan h(r) - \frac{Z_{\alpha/2}}{\sqrt{n-3}}\right] \leq \rho \leq \tanh\left[\arctan h(r) + \frac{Z_{\alpha/2}}{\sqrt{n-3}}\right]$$

onde r é o coeficiente de correlação estimado, $Z_{\alpha/2}$ é o quantil da distribuição normal padronizada com um nível de significância α , n é tamanho da amostra e

$$\tanh(u) = \frac{(e^u - e^{-u})}{(e^u + e^{-u})}$$

Comentários Finais:

O procedimento para análise da correlação:

Etapa 1: Determine se existe uma relação de causa e efeito para todos os pares de variáveis a serem testados.

Etapa 2: Plote todas as combinações de uma variável em relação a outra para examinar as relações dos dados.

Etapa 3: Faça ajustes, como transformação de dados, se necessário. Esta etapa é opcional.

Etapa 4: Calcule os coeficientes de correlação linear entre cada par de variáveis. No EXCEL: Função CORREL() e na ferramenta Análise de Dados/Correlação



Serviço Geológico do Brasil – CPRM

Departamento de Hidrologia da CPRM

Eber José de Andrade Pinto
Coordenador Executivo do DEHID
eber.andrade@cprm.gov.br
www.cprm.gov.br

Belo Horizonte, 19 de outubro de 2020