Insper

Ciência dos dados

Análise bidimensional:

Duas variáveis quantitativas

Objetivos de aprendizado

Ao final desta aula, o aluno deve ser capaz de:

- Estudar a relação existente entre duas variáveis quantitativas graficamente;
- Por meio de medidas adequadas, medir o grau de associação entre duas variáveis quantitativas;
- Descrever o comportamento médio entre duas variáveis quantitativas por meio de um ajuste linear.

http://guessthecorrelation.com/

O que compreendemos?

- Quando os pontos estão mais dispersos no gráfico de dispersão (*scatter plot*), valor da correlação (TRUE R) tende a ser próximo de zero.
- Quando a nuvem de pontos está menos dispersa, o valor da correlação (TRUE R) tende a ser mais próxima de 1.

Indicadores socioeconômicos

O arquivo **Mundo.xlsx** conta com uma amostra de **85 países**, para os quais levantou-se uma série de indicadores socioeconômicos.

Variáveis:

X_1: população em milhares de habitantes
X_2: densidade populacional
X_3: % de população urbana
X_4: expectativa de vida feminina
X_5: expectativa de vida masculina
X_6: crescimento populacional
X_7: mortalidade infantil
X_8: PIB per capita
X_9: % de mulheres alfabetizadas
X_{10}: população em 100.000 habitantes

expectativa de vida feminina
50
60
70
80
20
40
60
80
100
% de população urbana

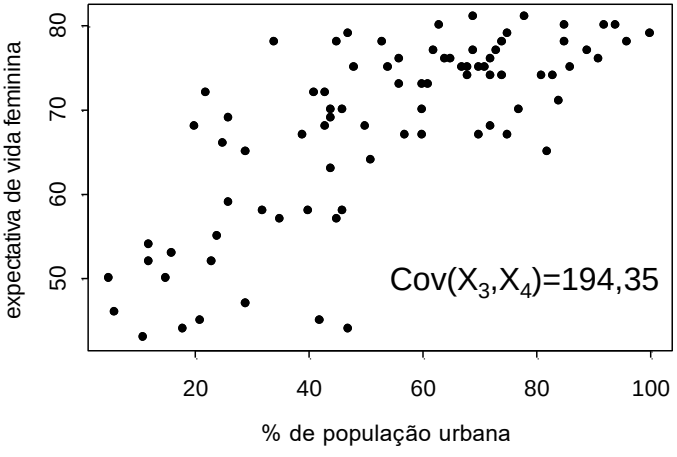
expectativa de vida masculina
40
50
60
70
20
40
60
80
100
% de população urbana

mortalidade infantil
0
50
100
150
20
40
60
80
100
% de população urbana

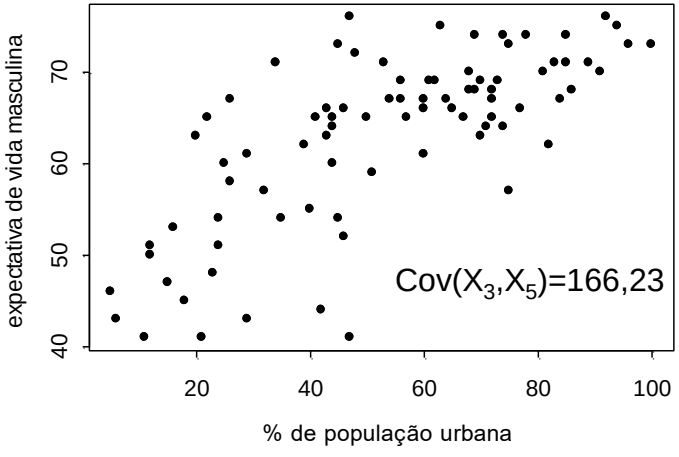
crescimento populacional
0
1
2
3
4
5
20
40
60
80
100
% de população urbana

er

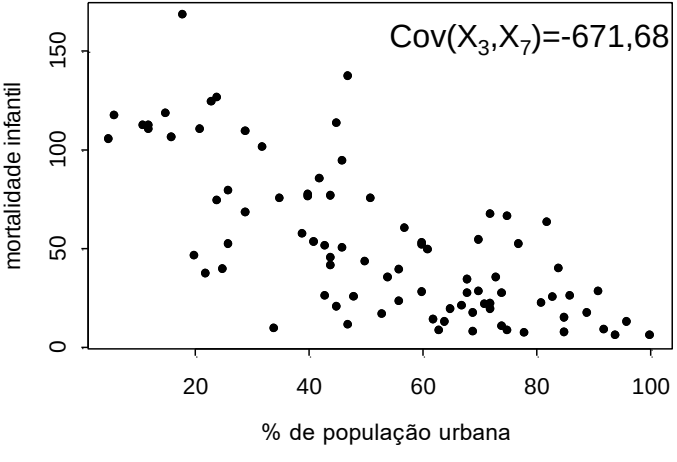
Associação Positiva



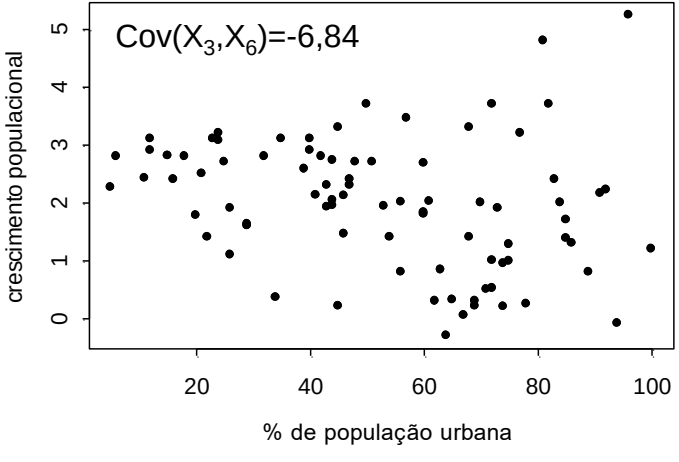
Associação Positiva



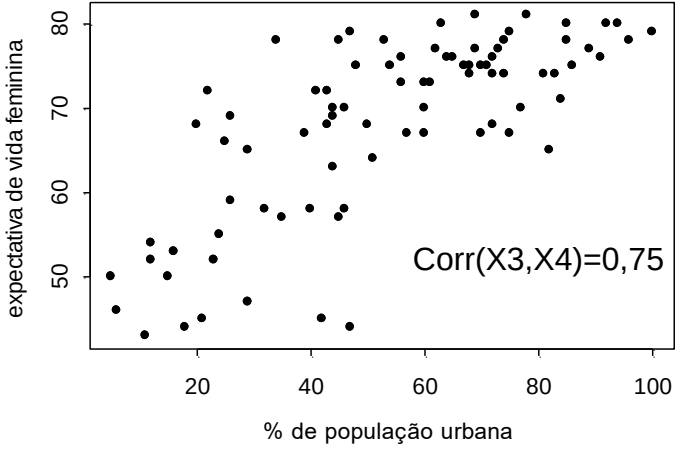
Associação Negativa



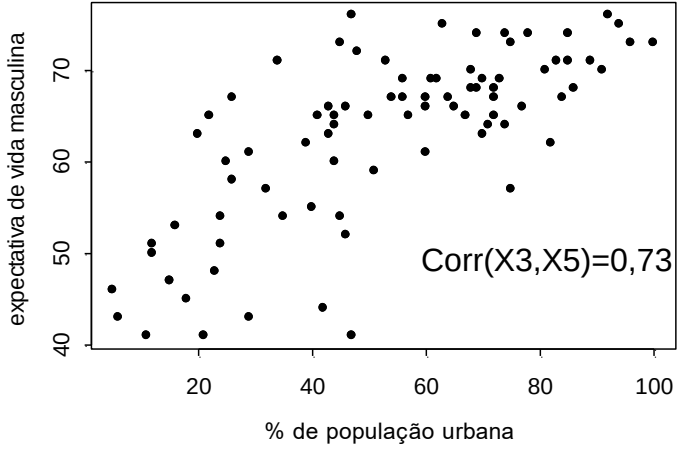
Baixo indício de associação



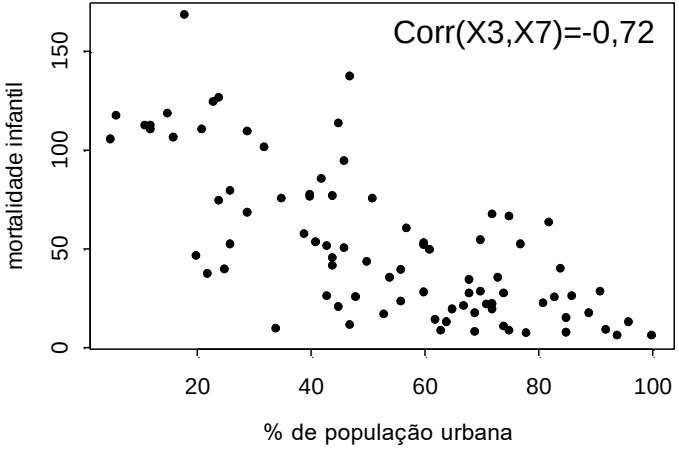
Associação Positiva



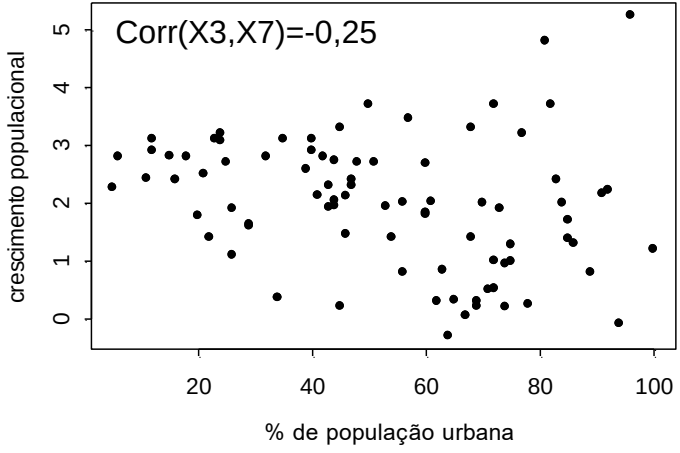
Associação Positiva



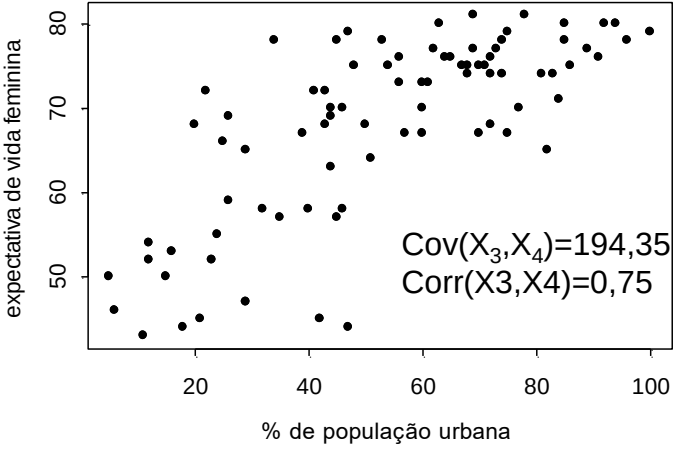
Associação Negativa



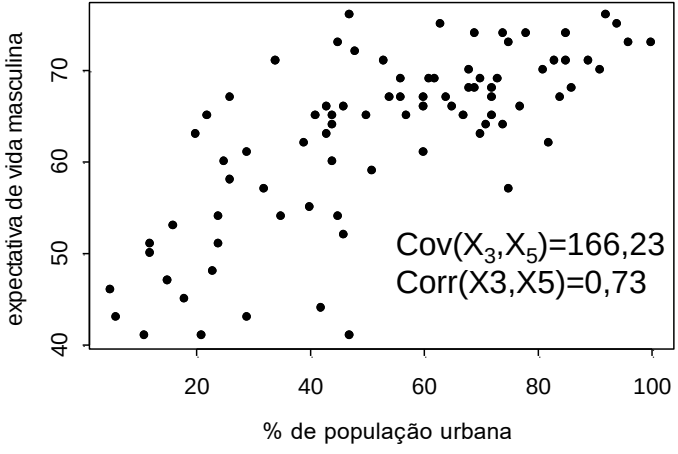
Baixo indício de associação



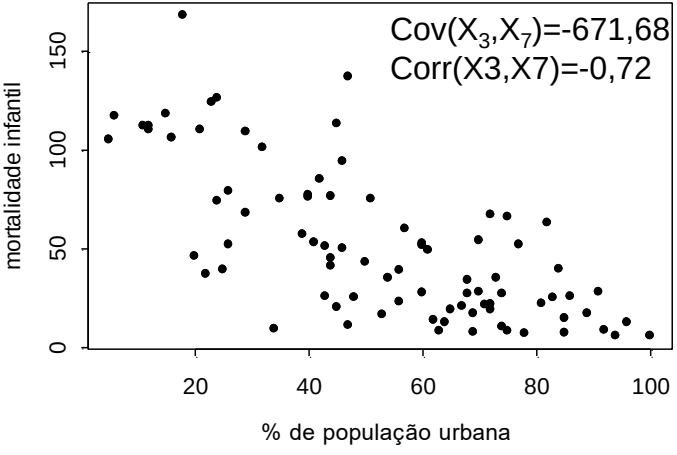
Associação Positiva



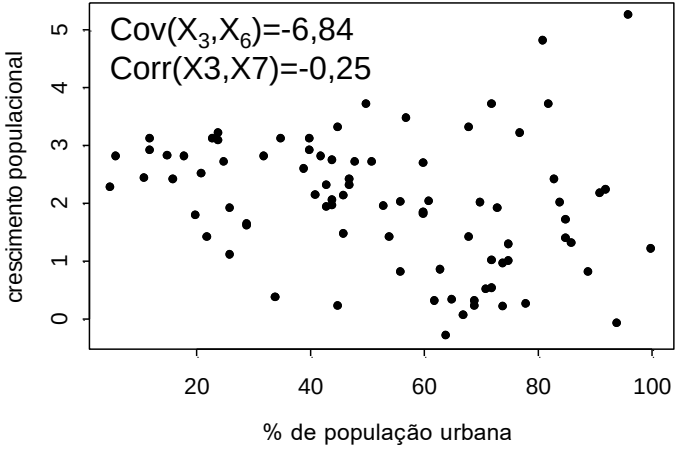
Associação Positiva



Associação Negativa



Baixo indício de associação



Exemplo:

Taxa de mortalidade infantil e taxa de analfabetismo no Brasil, segundo região.

Ano: 1997

Região	Taxa de analfabetismo	Taxa de mortalidade infantil
Norte	13	36
Nordeste	29	59
Sudeste	9	25
Sul	8	22
Centro Oeste	12	25

Fonte: IBGE.

Taxa de analfabetismo: Percentual de pessoas com 15 ou mais anos de idade que não sabem ler e escrever pelo menos um bilhete simples, em determinado espaço geográfico, no ano considerado.

Taxa de mortalidade infantil: Número de óbitos de menores de um ano de idade, por mil nascidos vivos, em determinado espaço geográfico, no ano considerado.

Fonte: RIPSA

Exemplo:

Taxa de mortalidade infantil e taxa de analfabetismo no Brasil, segundo região.

Ano: 1997

Região	Taxa de analfabetismo	Taxa de mortalidade infantil
Norte	13	36
Nordeste	29	59
Sudeste	9	25
Sul	8	22
Centro Oeste	12	25

Fonte: IBGE.

Considere :

X: Taxa de analfabetismo

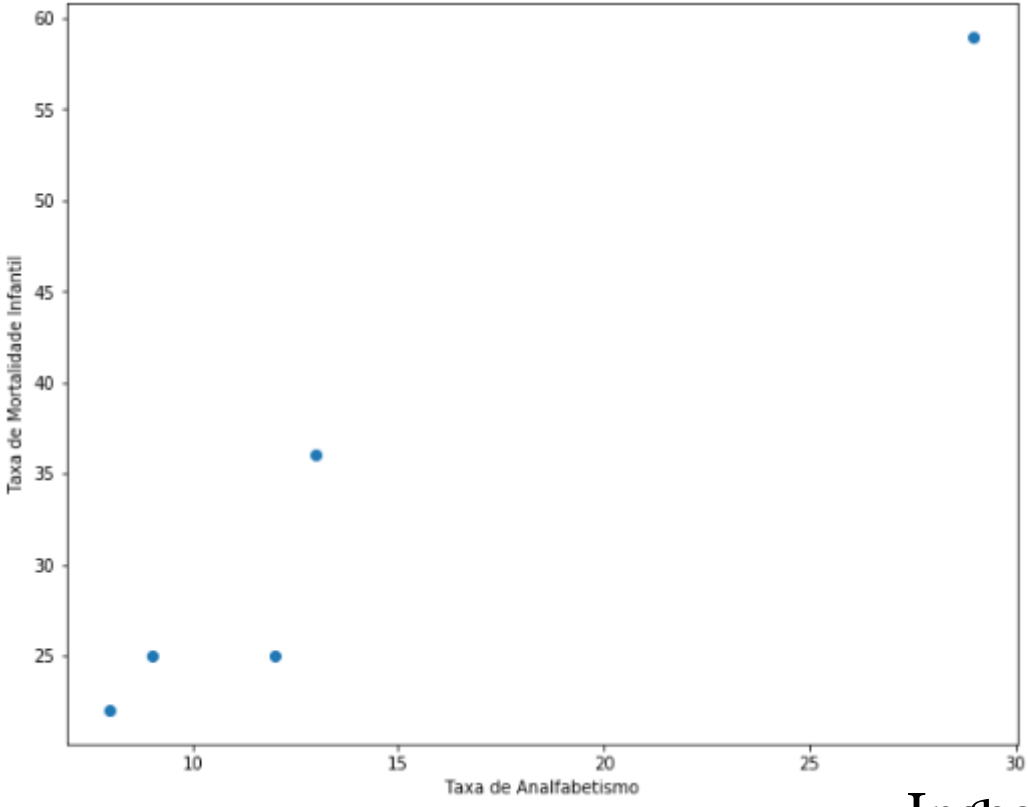
Y: Taxa de mortalidade infantil

$$\overline{x} = 14,2$$
$$\overline{y} = 33,4$$

Exemplo:

Taxa de mortalidade infantil e taxa de analfabetismo no Brasil, segundo região.

Ano: 1997



Fonte: IBGE.

Coeficiente de Covariância

$$Cov(X,Y) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{n}$$

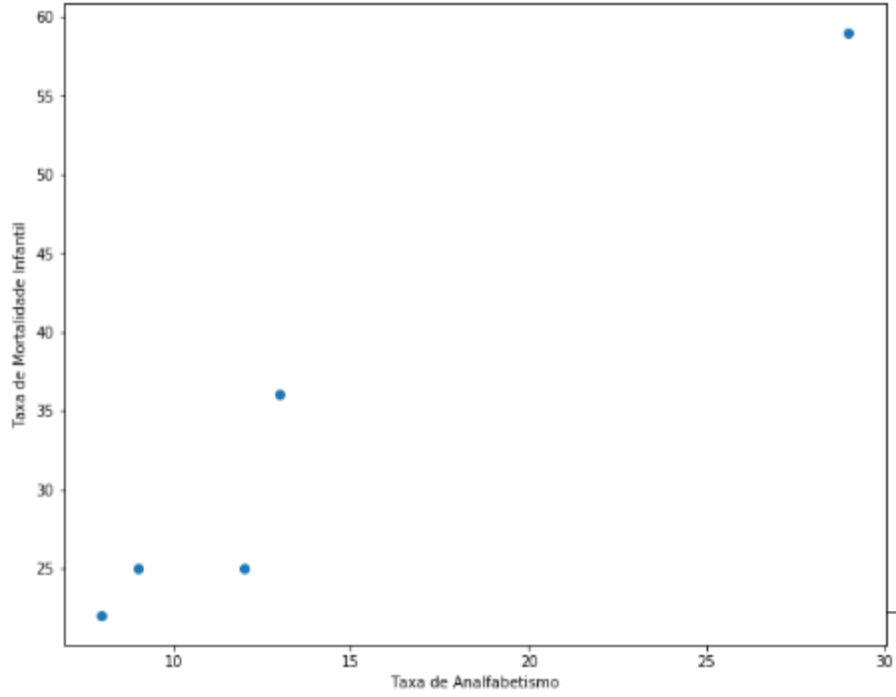
- Cov(X,Y) > 0 se a associação linear for positiva.
- Cov(X,Y) < 0 se a associação linear for negativa.
- Cov(X,Y) = 0 indica que não existe associação linear positiva, nem negativa, mas pode existir outro tipo de associação.

Região	Taxa de analfabetismo	Taxa de mortalidade infantil
Norte	13	36
Nordeste	29	59
Sudeste	9	25
Sul	8	22
Centro Oeste	12	25

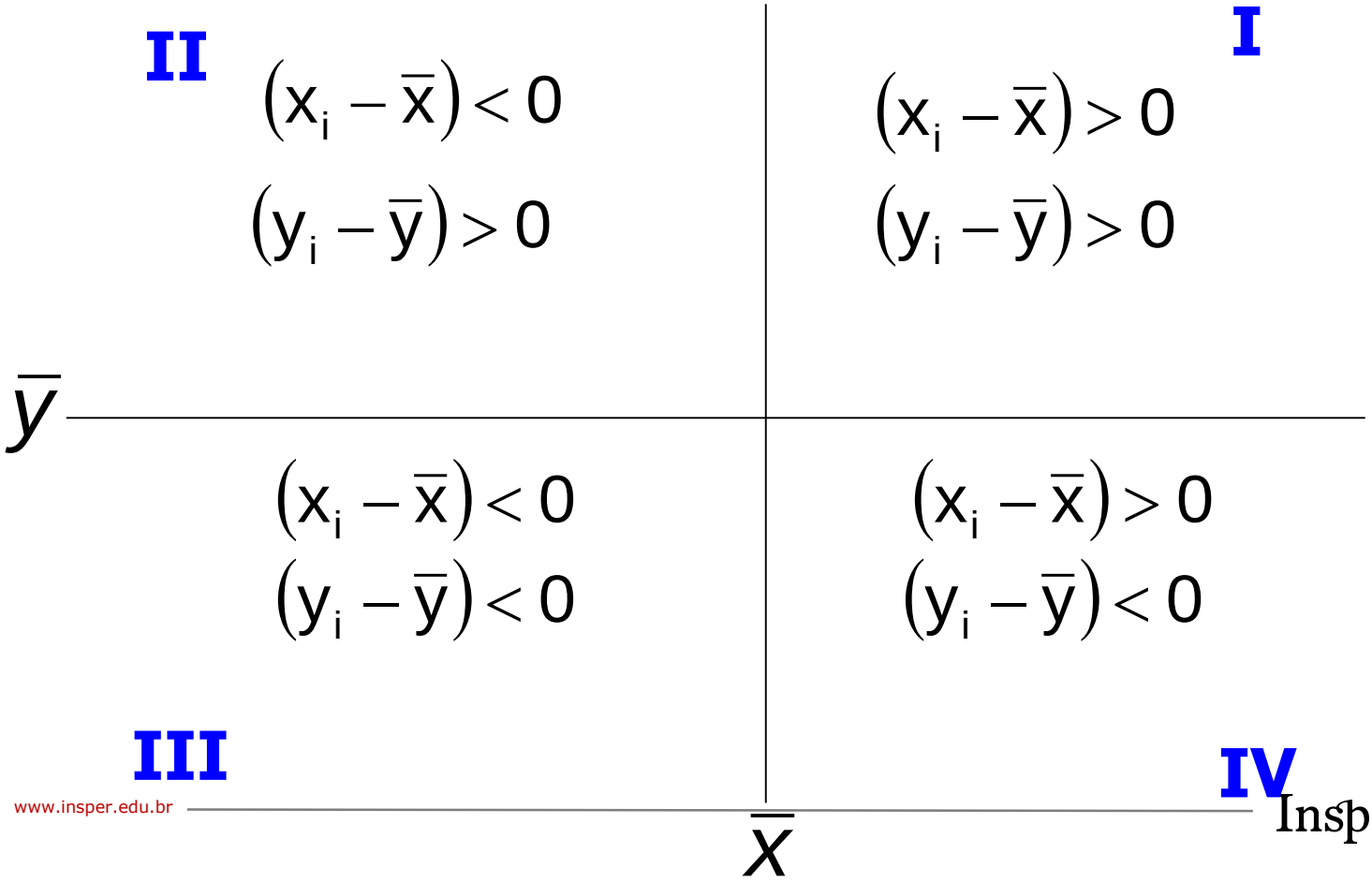
$$\bar{x} = 14,2$$
$$\bar{y} = 33,4$$

$$Cov(X,Y) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{n}$$

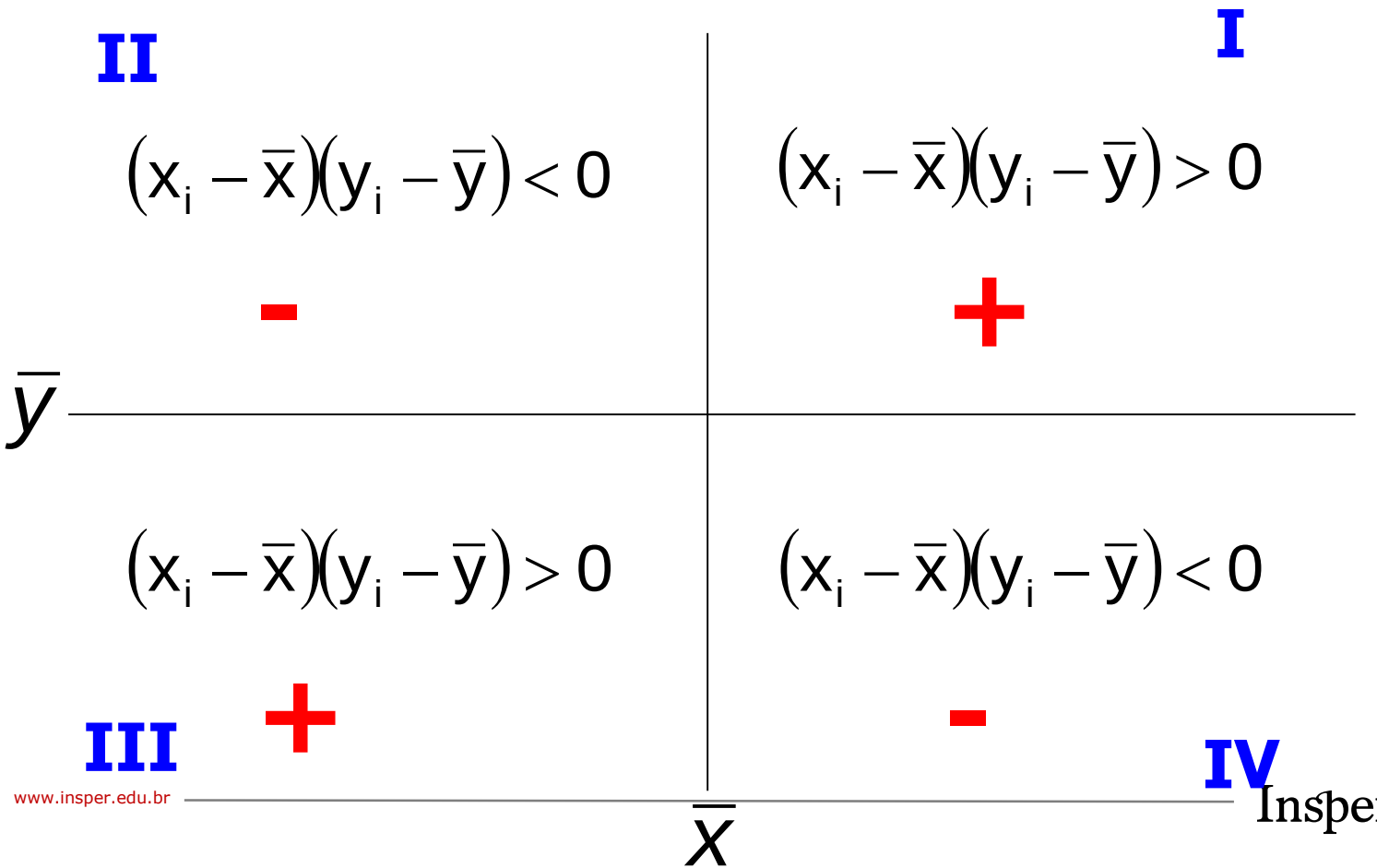
Calcule a covariância entre essas duas variáveis quantitativas:



Estudo de Sinal

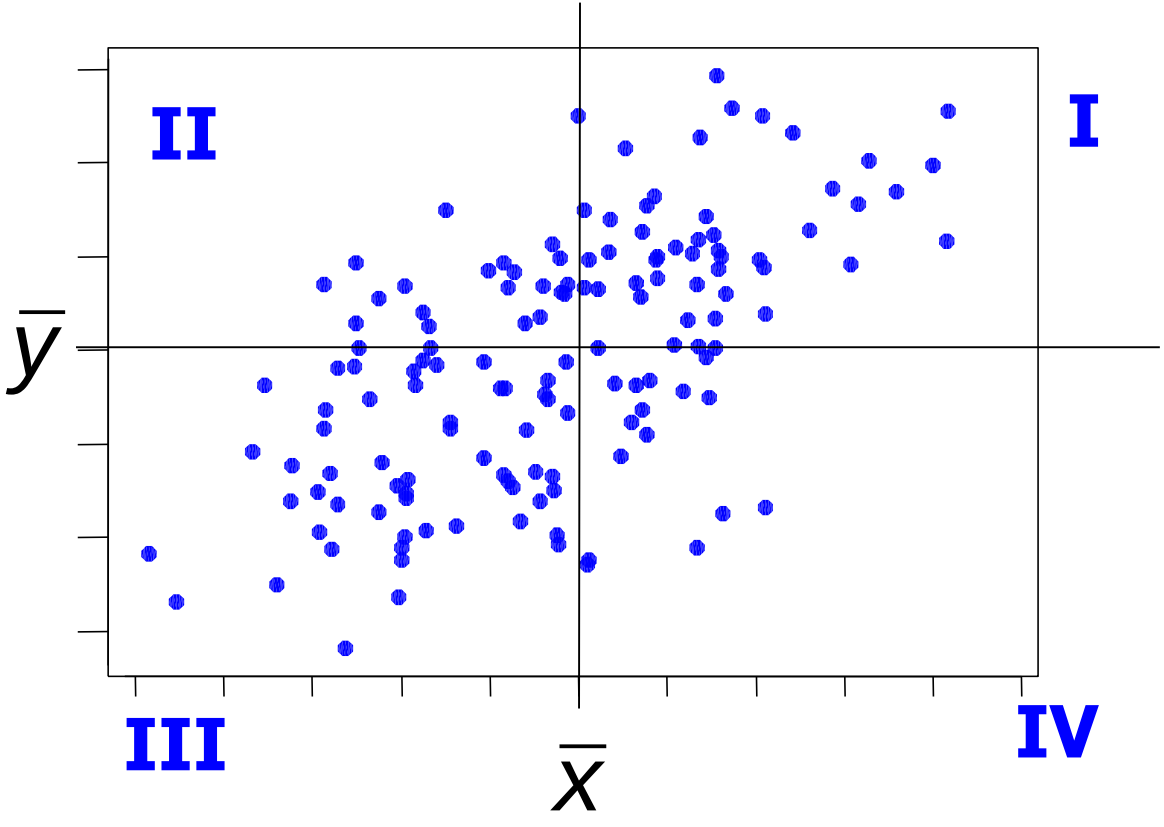


Estudo de Sinal



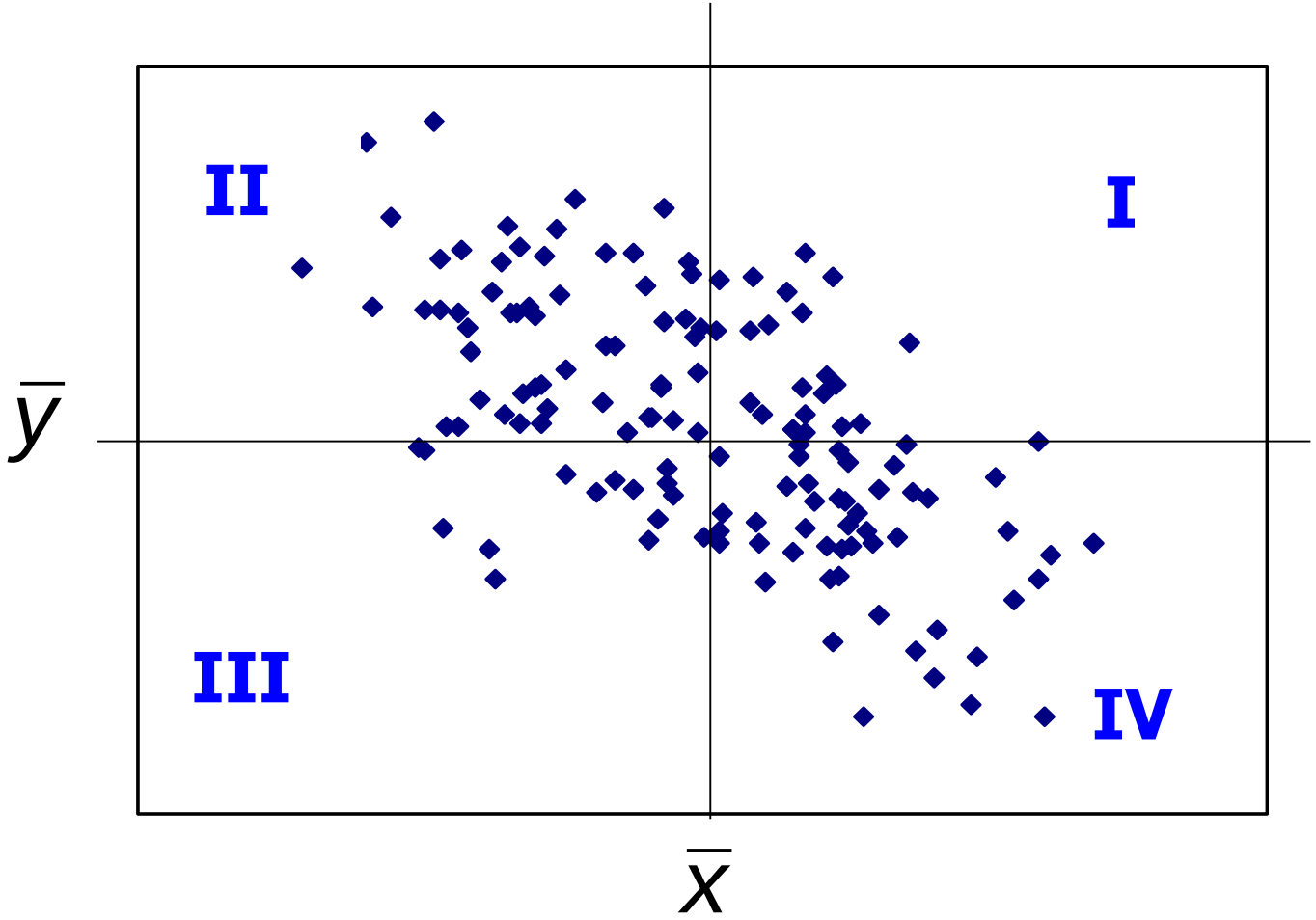
Associação Positiva

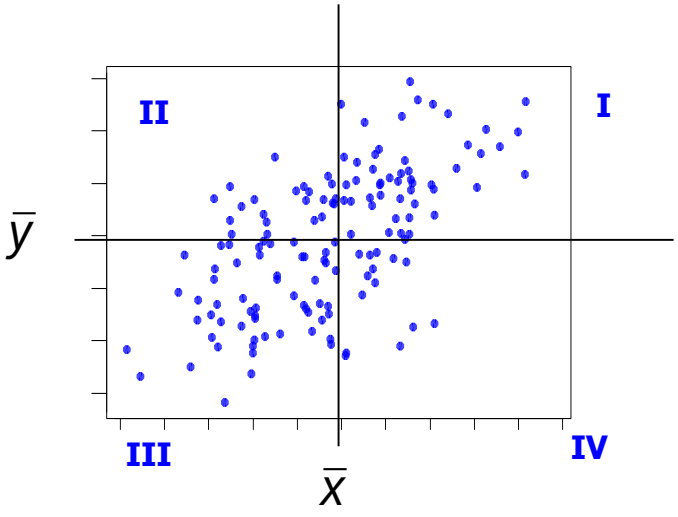
Percebe-se um acúmulo de pontos nos quadrantes ímpares.

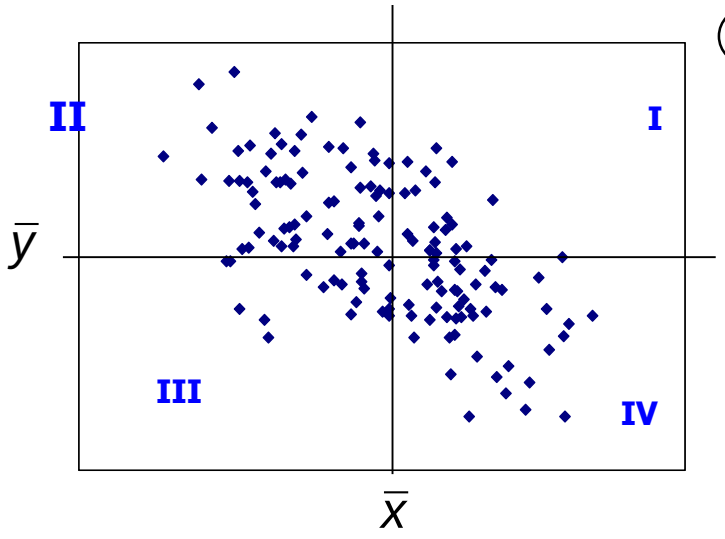


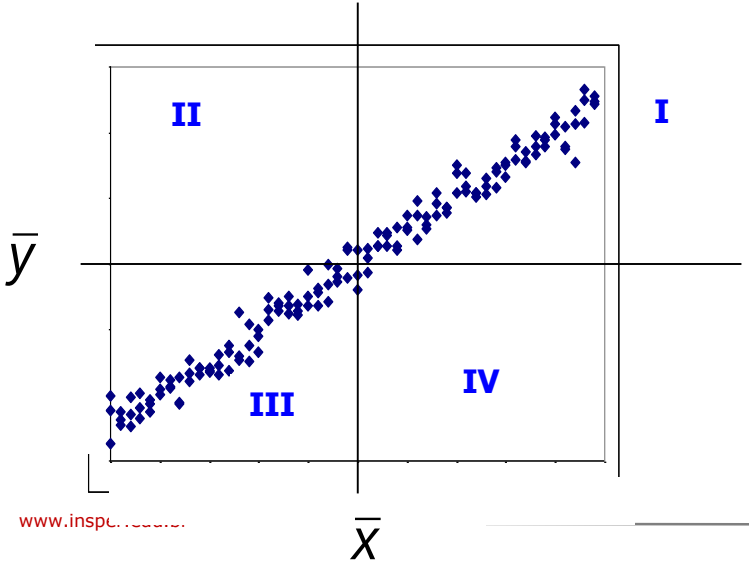
Associação Negativa

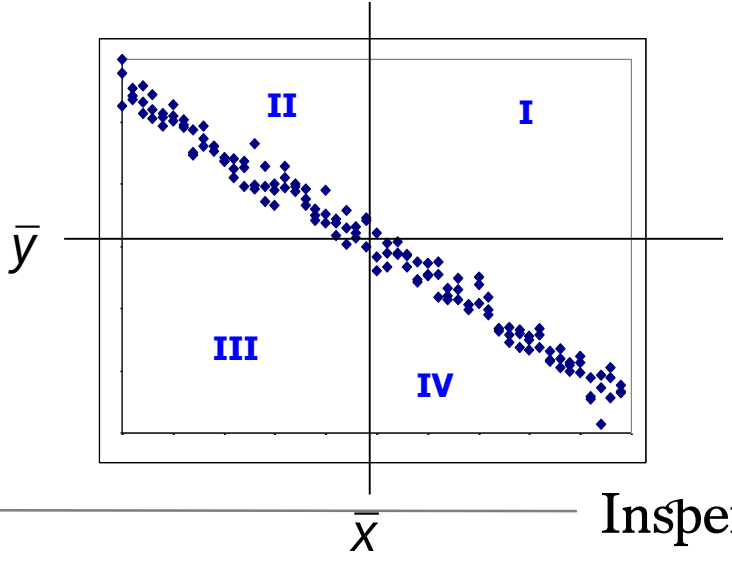
Percebe-se um acúmulo de pontos nos quadrantes pares.




II
I
III
IV
$\bar{y}$
$\bar{x}$


II
I
III
IV
$\bar{y}$
$\bar{x}$


II
I
III
IV
$\bar{y}$
$\bar{x}$


II
I
III
IV
$\bar{y}$
$\bar{x}$

Comportamento Geral

- Quando existe uma associação positiva (crescente) entre as variáveis, há um predomínio de pontos nos quadrantes ímpares.
- Quando existe uma associação negativa (decrescente) entre as variáveis, há um predomínio de pontos nos quadrantes pares.
- Quanto mais próxima de uma reta estiverem os pontos, maior é o predomínio nos quadrantes ímpares (se crescente) ou pares (se decrescente).

Exemplo:

Taxa de mortalidade infantil e taxa de analfabetismo no Brasil, segundo região.

Ano: 1997

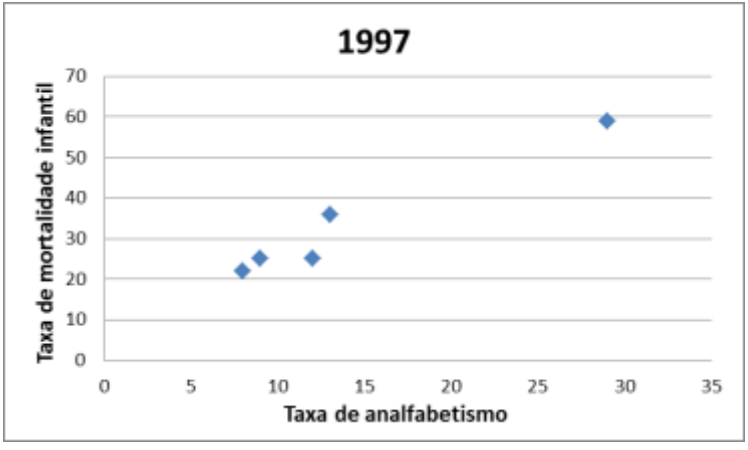
Região	Taxa de analfabetismo	Taxa de mortalidade infantil
Norte	13	36
Nordeste	29	59
Sudeste	9	25
Sul	8	22
Centro Oeste	12	25

Ano: 2009

Região	Taxa de analfabetismo	Taxa de mortalidade infantil
Norte	10,45	23,5
Nordeste	18,53	33,2
Sudeste	5,84	16,6
Sul	4,62	15,1
Centro Oeste	5,09	17,8

Fonte: IBGE.

Cov(X,Y)= 101,72



Cov(X,Y)= 34,45

O gráfico azul possui um coeficiente de covariância maior, mas a associação não parece ser mais forte do que a observada no gráfico vermelho.

Como resolver isso?

Coeficiente de Correlação Linear

$$Corr(X,Y) = \frac{Cov(X,Y)}{DP(X)DP(Y)}$$

Vantagem em relação à covariância:

$$-1 \leq Corr(X,Y) \leq 1$$

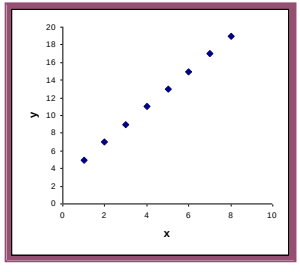
Quando Corr(X,Y)=1 ou Corr(X,Y)=-1 os pontos estarão perfeitamente alinhados sobre uma reta.

Propriedades do coeficiente de correlação

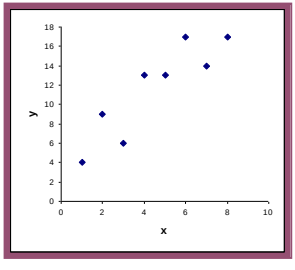
- Medida de associação linear entre duas variáveis quantitativas (varia entre –1 e +1).
- Valores próximos a +1 : indicam forte relação linear positiva
- Valores próximos a -1 : indicam forte relação linear negativa
- Valores próximos a zero: indicam ausência de relação linear.

Interpretação do Coeficiente de Correlação

Relação perfeita

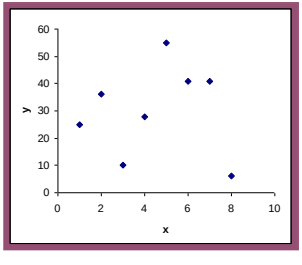


r = +1

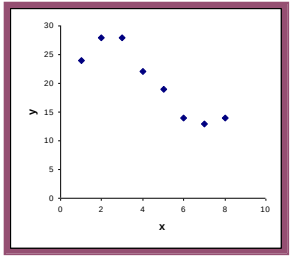


r ≈ 0,80

Ausência de relação

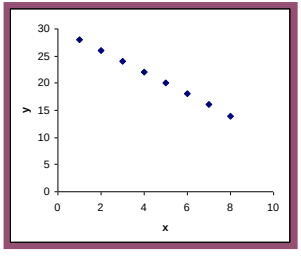


r ≈ 0



r ≈ -0,80

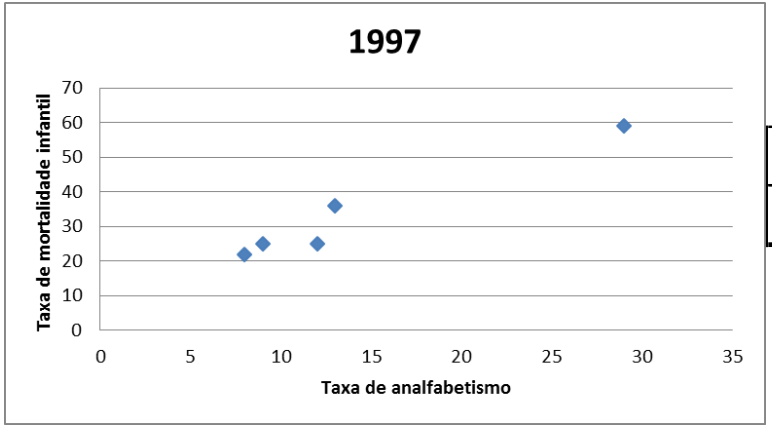
Relação perfeita



r = -1

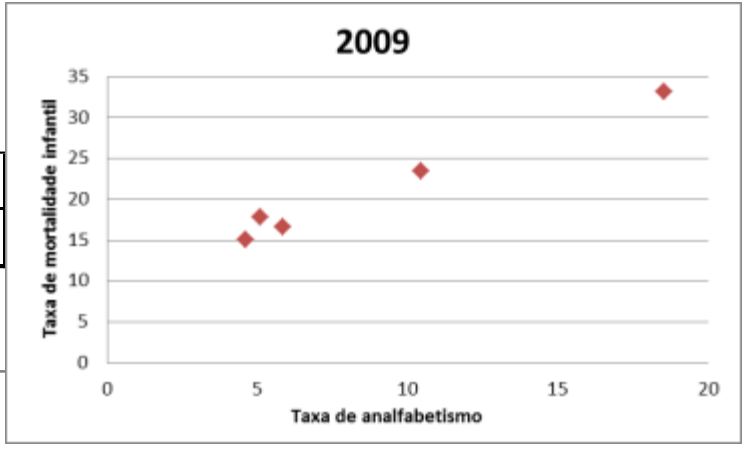
Exemplo:

Taxa de mortalidade infantil e taxa de analfabetismo no Brasil, segundo região.



Covariância	101,72
Correlação	0,976

Covariância	34,45
Correlação	0,993



Fonte: IBGE.

Associação não é causalidade

Suponha que encontremos alta correlação entre duas variáveis A e B. Podem existir diversas explicações do porque elas variam conjuntamente, incluindo:

- Mudanças em outras variáveis causam mudanças tanto em A quanto em B.
- Mudanças em A causam mudanças em B.
- Mudanças em B causam mudanças em A.
- A relação observada é somente uma coincidência (correlação espúria).

A primeira explicação é frequentemente a mais apropriada. Isto indica que existe algum processo de conexão atuando.

Fonte: http://leg.ufpr.br/~silvia/CE003/node77.html

Jupyter

Exploratória de base de dados real:

- Download pelo Github

  https://github.com/Insper/CD22-2

- Fazer individual:

  1) Notebook nomeado “Atividade”
  2) Notebook nomeado “Exercício” com APS2 no Blackboard