Análise Bidimensional

Agenda

1. Introdução
2. Relação entre Duas Variáveis Quantitativas
3. Relação entre Duas Variáveis Qualitativas
4. Relação entre Uma Variáveis Qualitativa e Uma Quantitativa
5. Conclusões

Agenda

1. **Introdução**
2. Relação entre Duas Variáveis Quantitativas
3. Relação entre Duas Variáveis Qualitativas
4. Relação entre Uma Variáveis Qualitativa e Uma Quantitativa
5. Conclusões

Introdução

Nessa aula, mediremos o grau de relação entre:

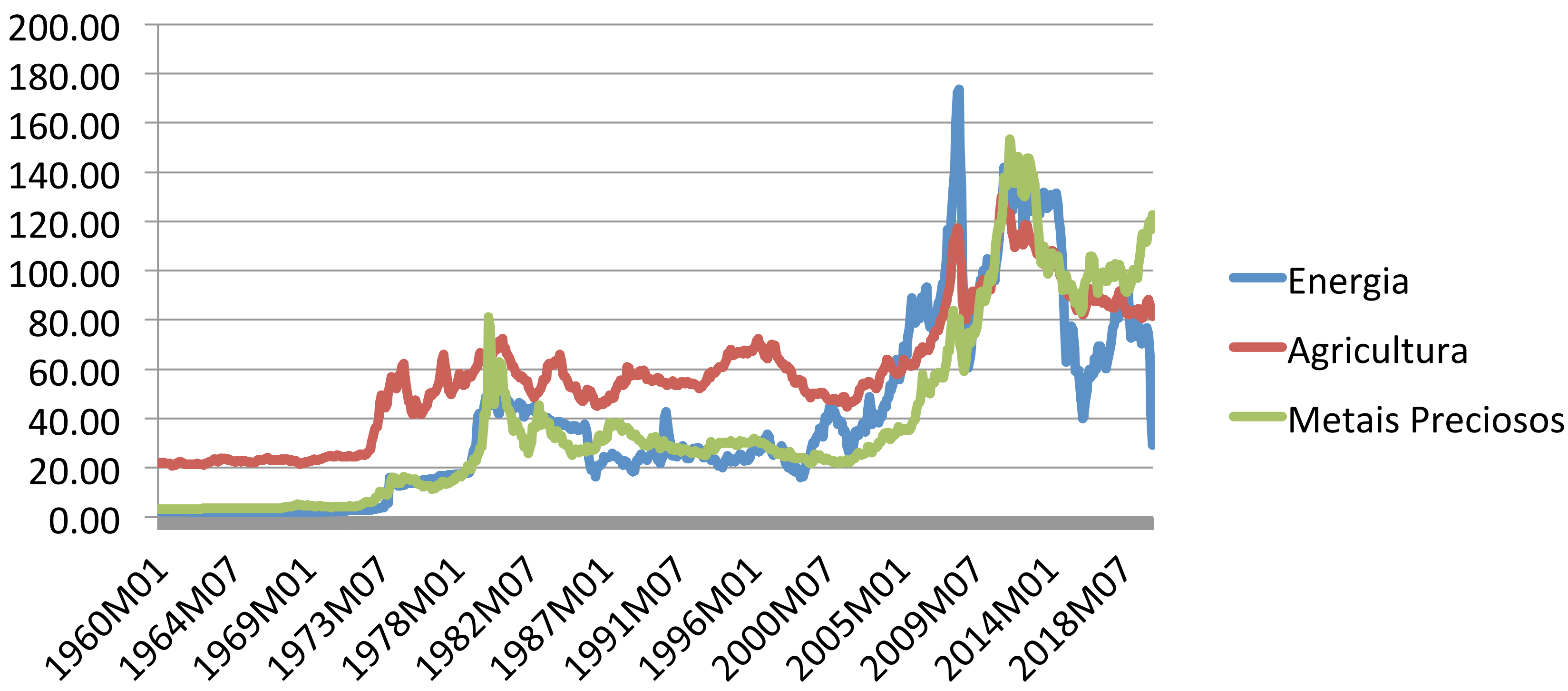
- duas variáveis qualitativas
- duas variáveis quantitativas
- uma variável qualitativa e a outra quantitativa

Agenda

1. Introdução
2. **Relação entre Duas Variáveis Quantitativas**
3. Relação entre Duas Variáveis Qualitativas
4. Relação entre Uma Variáveis Qualitativa e Uma Quantitativa
5. Conclusões

Relação entre Variáveis Quantitativas

- Suponha que temos séries quantitativas, como os índices de preços das commodities:



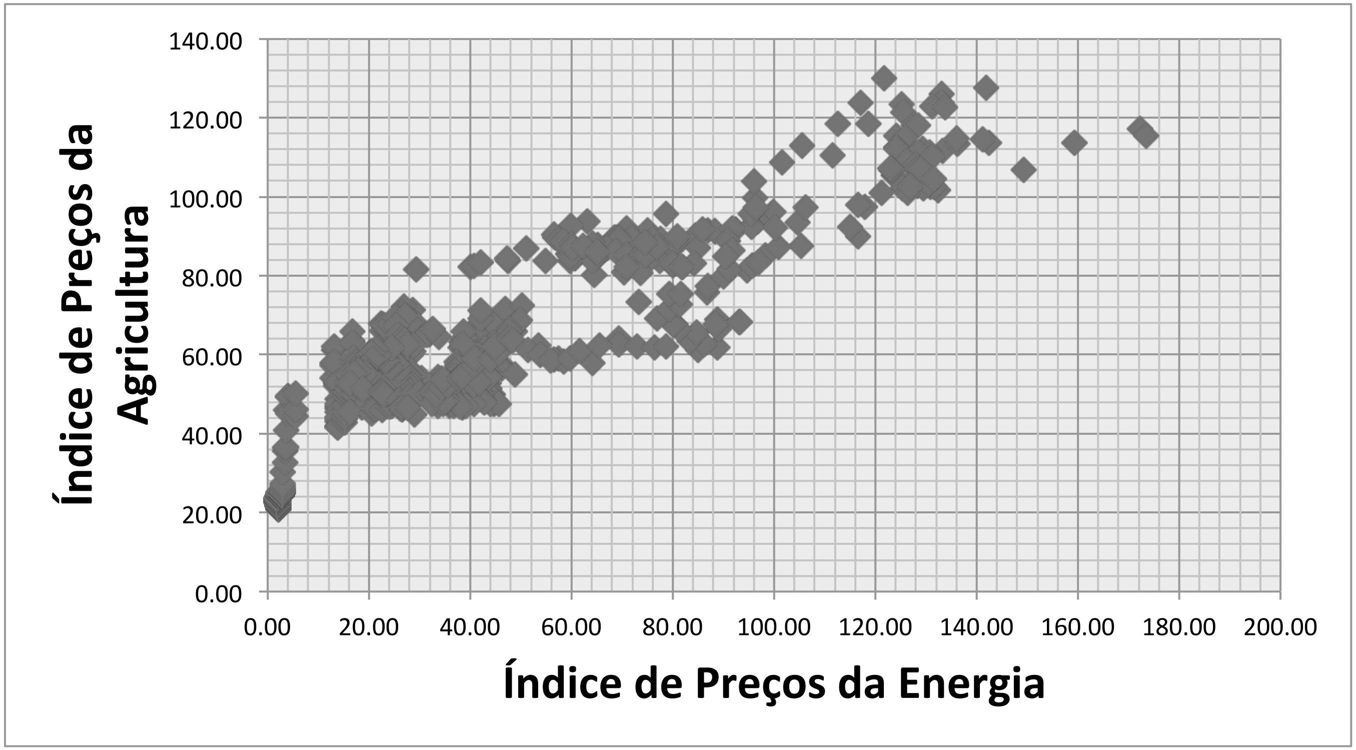
Fonte: Bloomberg, Banco Mundial

Relação entre Variáveis Quantitativas

- Queremos analisar a relação entre essas séries, duas a duas.
- A sua análise pode se iniciar através da construção do gráfico dos dados do índice da agricultura no eixo Y contra os dados do índice da energia no eixo X. Esse seria o gráfico de dispersão
- No gráfico de dispersão, cada ponto equivale a um par de dados (xi,yi) presente em cada linha do banco de dados.

Relação entre Variáveis Quantitativas

Gráfico de Dispersão



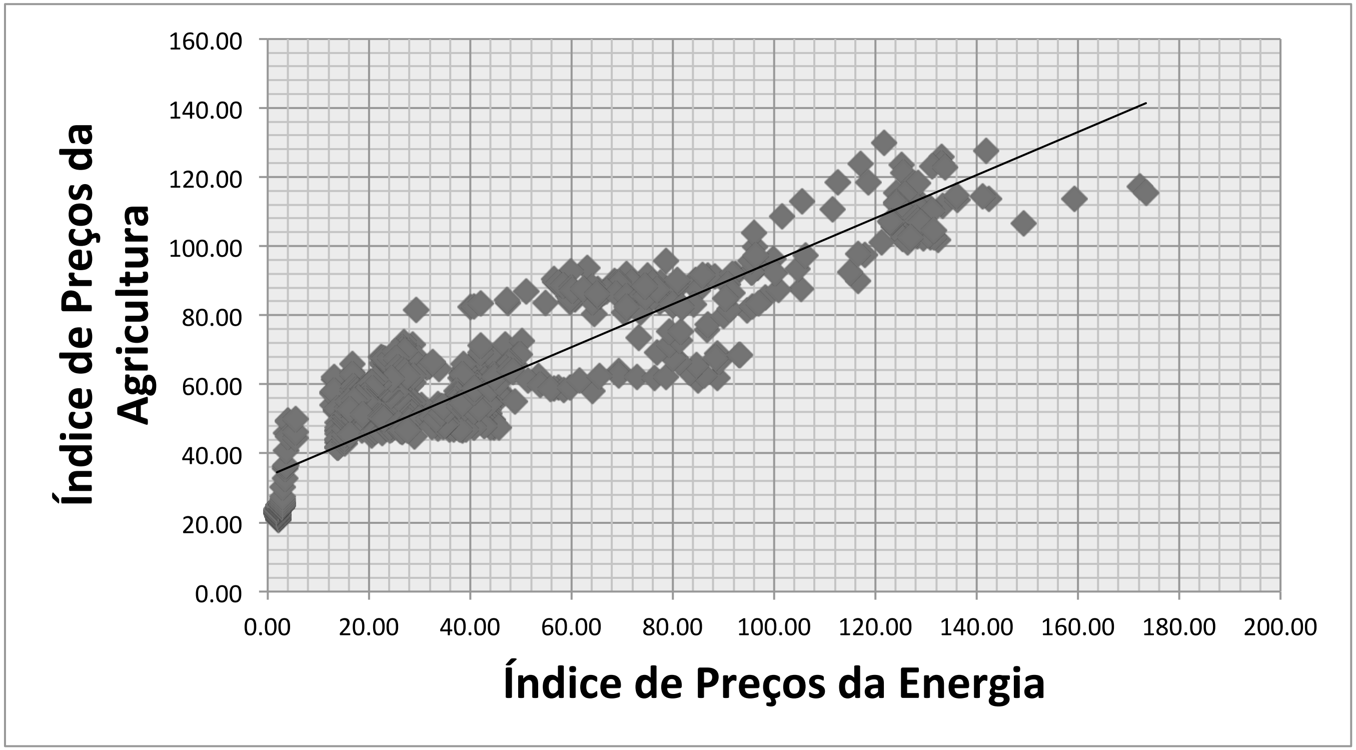
Dados mensais de 1960M1 a 2020M4, em US$ de 2010, Fonte: Bloomberg, Banco Mundial

Relação entre Variáveis Quantitativas

- Notamos uma relação linear positiva entre esses valores: quando os valores do índice da energia estão relativamente altos em relação à sua média, os valores do índice da agricultura também estão.
- A declividade da reta que melhor se aproxima ao conjunto de dados é positiva.

Relação entre Variáveis Quantitativas

Gráfico de Dispersão e Reta de Ajuste



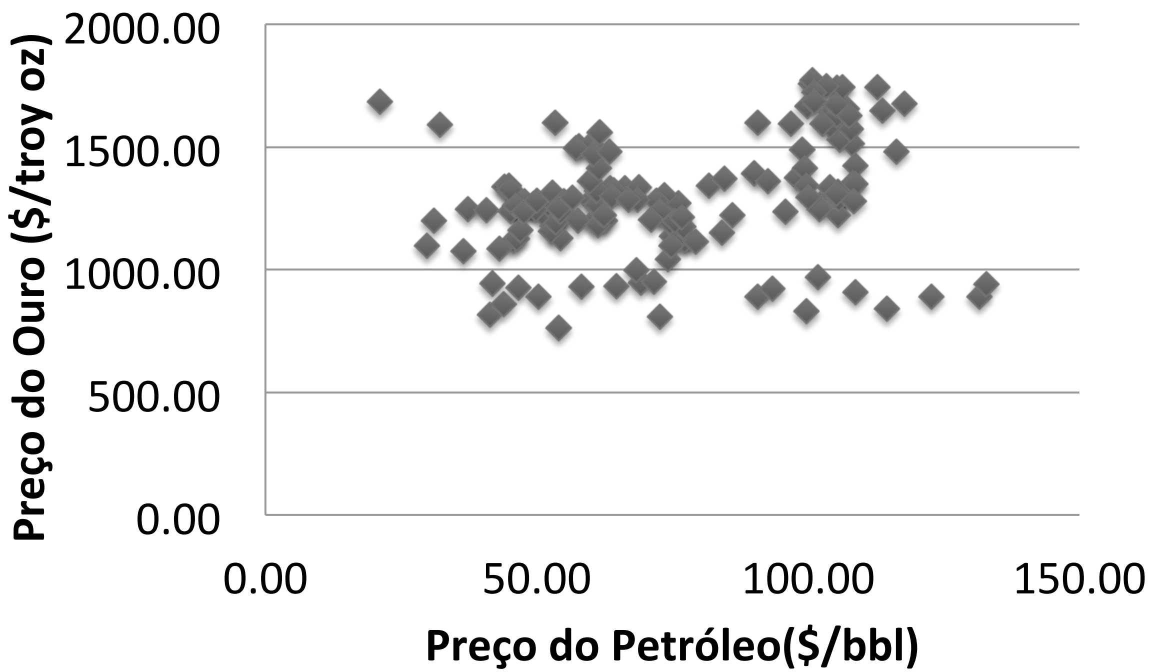
Dados mensais de 1960M1 a 2020M4, em US$ de 2010, Fonte: Bloomberg, Banco Mundial

Relação entre Variáveis Quantitativas

- Porém, a relação linear entre duas séries quantitativas pode ser negativa ou próxima de zero.
- A seguir, mostramos um exemplo de relação linear fraca ou próxima de zero entre duas séries quantitativas:

Relação entre Variáveis Quantitativas

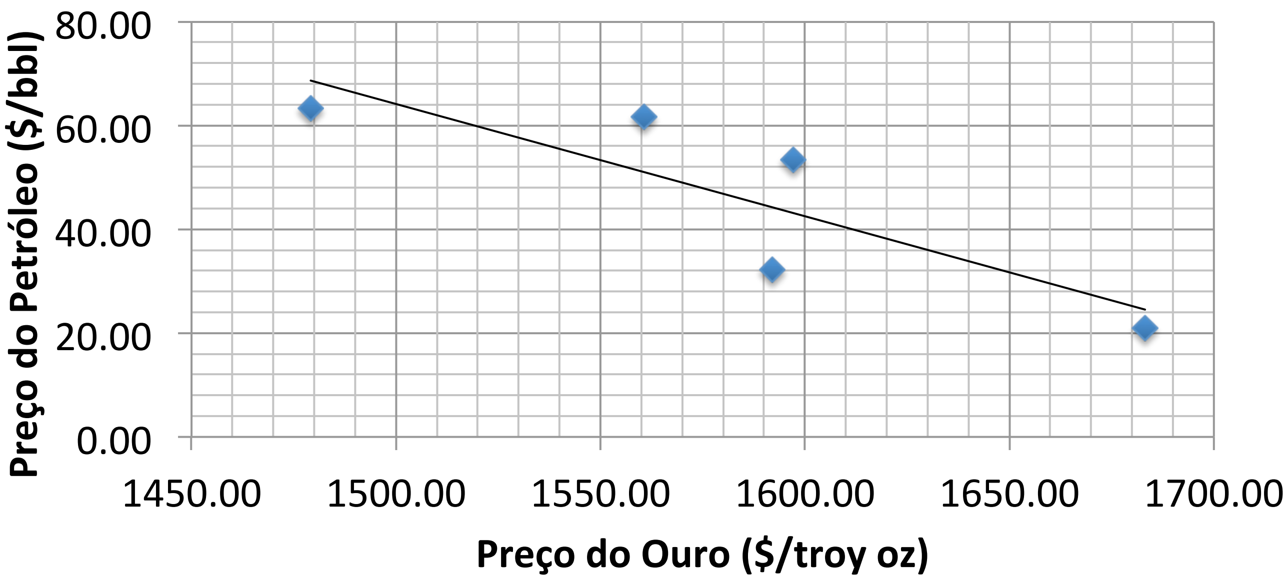
Gráfico de dispersão do preço do ouro em relação ao preço do petróleo de 1960M1 a 2020M4



Dados mensais de 1960M1 a 2020M4, em US$ corrente, Fonte: Bloomberg, Banco Mundial

Relação entre Variáveis Quantitativas

Gráfico de dispersão do preço do ouro em relação ao preço do petróleo de 2019M12 a 2020M4



Dados mensais de 2019M12 a 2020M4, em US$ corrente, Fonte: Bloomberg, Banco Mundial

Relação entre Variáveis Quantitativas

- Nesse caso, a relação linear entre as duas séries é negativa, o que pode ser visto pela declividade negativa da reta de melhor ajuste aos dados.
- O gráfico de dispersão nos dá uma noção visual da relação linear entre duas séries.
- Entretanto, gostaríamos de ter uma medida estatística do grau de relação linear entre essas séries.

Relação entre Variáveis Quantitativas

- Dados n pares de valores $(x_1, y_1)..., (x_n, y_n)$, chamaremos de covariância entre as variáveis X e Y, na população:

$$\operatorname{cov}(X,Y) = \frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{n}$$

- É a soma dos produtos dos desvios em relação às médias das variáveis, dividido por n

Relação entre Variáveis Quantitativas

- Para a amostra, assim como no caso da variância, dividimos por n-1:

$$\operatorname{cov}(X,Y)=\frac{\sum_{i=1}^{n}\left(x_i-\overline{x}\right)\left(y_i-\overline{y}\right)}{n-1}$$

- A covariância é a medida do afastamento simultâneo das respectivas médias.
- Se as ambas variáveis aleatórias tendem a estar simultaneamente acima, ou abaixo, de suas respectivas médias, então a covariância tenderá a ser positiva.

Relação entre Variáveis Quantitativas

- A covariância de uma variável e ela mesma é a própria variância da variável. Como *Y* = *X*, tem-se:

$$\operatorname{cov}(X,X) = \frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\left(x_i - \overline{x}\right)}{n}$$

$$= \operatorname{var}(X)$$

Relação entre Variáveis Quantitativas

- A permutação das variáveis X e Y não altera o resultado da covariância, se os pares de valores não forem alterados.

$$\operatorname{cov}(X,Y) = \frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{n}$$

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

$$= \operatorname{cov}(Y,X)$$

Relação entre Variáveis Quantitativas

- A unidade de medida da covariância é o resultado do produto das unidades das variáveis X e Y.
- Para facilitar o entendimento da relação entre duas variáveis e evitar a unidade de medida da covariância, foi definido o coeficiente de correlação r_{XY}.

$$corr(X,Y) = \frac{\mathrm{cov}(X,Y)}{dp(X).dp(Y)}$$

Relação entre Variáveis Quantitativas

- Vimos que o coeficiente de correlação r_{XY} é definido por:

$$corr(X,Y) = \frac{\operatorname{cov}(X,Y)}{dp(X).dp(Y)}$$

- Substituindo-se pela definição de covariância para a população, obtemos:

$$corr(X,Y) = \frac{1}{n}\sum\left(\frac{x_i - \overline{x}}{dp(X)}\right)\left(\frac{y_i - \overline{y}}{dp(Y)}\right)$$

Relação entre Variáveis Quantitativas

- O coeficiente de correlação busca auferir a direção e intensidade da relação linear entre duas variáveis, dentro de um intervalo determinado entre -1 e 1:
  - valores próximos de zero indicam ausência de relação entre as variáveis
  - valores próximos de 1 indicam forte relação positiva
  - valores próximos de -1 indicam forte relação negativa

Relação entre Variáveis Quantitativas

- Se a variável *Y* é a mesma variável *X*, então o coeficiente de correlação é igual a 1:

$$r_{XX} = \frac{\sigma_{XX}}{\sigma_X \times \sigma_X} = \frac{\sigma_X^2}{\sigma_X^2} = 1$$

- Vimos que a permutação das variáveis não altera o resultado da covariância, , se os mesmos pares de valores forem mantidos. O mesmo vale para o coeficiente de correlação.

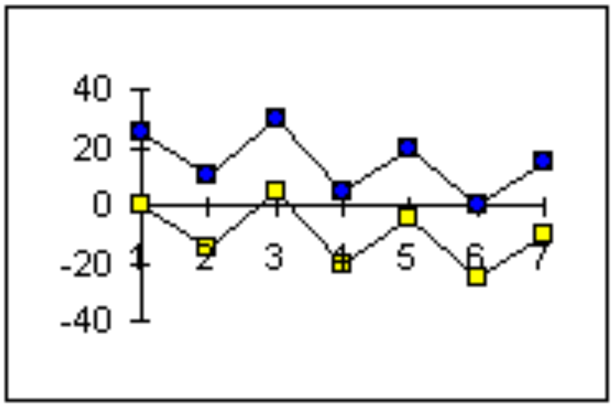
$$r_{XY} = r_{YX}$$

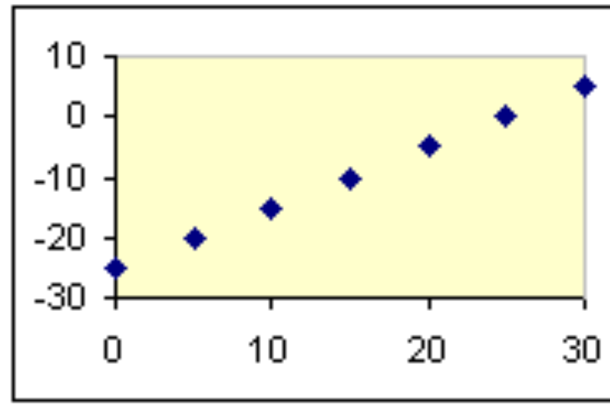
$r = +1$

Amostras perfeitamente correlacionadas em sentido positivo

X	Y
25	0
10	-15
30	5
5	-20
20	-5
0	-25
15	-10

r	1,00



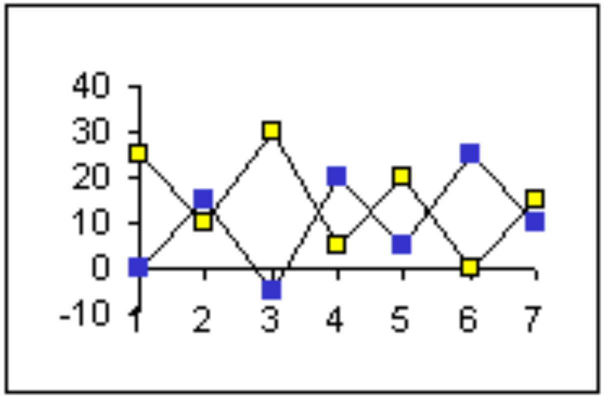


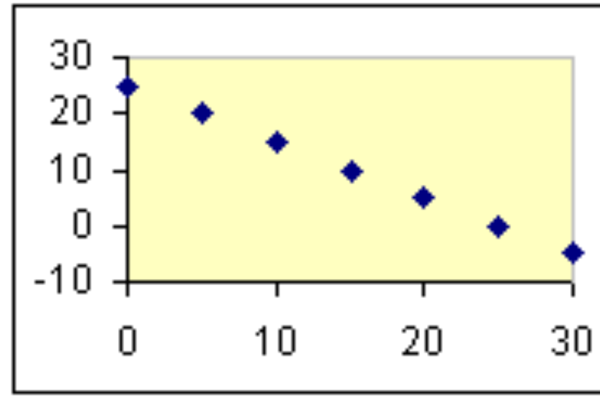
r = -1

Amostras perfeitamente correlacionadas em sentido negativo

X	Y
25	0
10	15
30	-5
5	20
20	5
0	25
15	10

r	-1,00



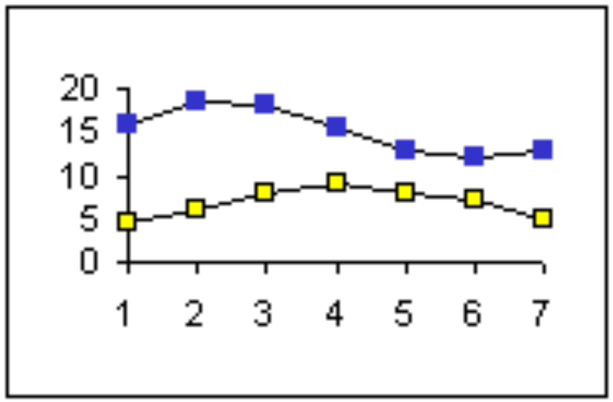


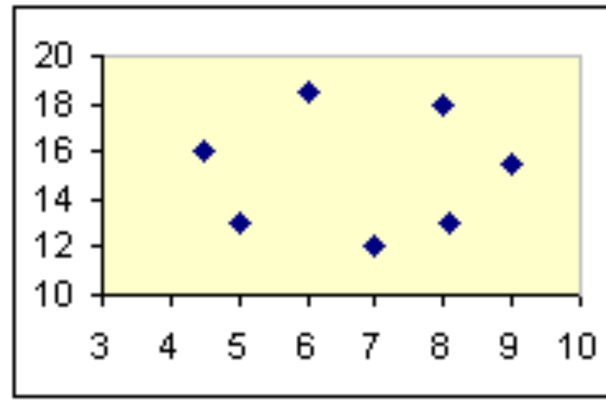
$r = 0$

Amostras não correlacionadas

X	Y
4,5	16
6	18,5
8	18
9	15,5
8,1	13
7	12
5	13

r **0,00**

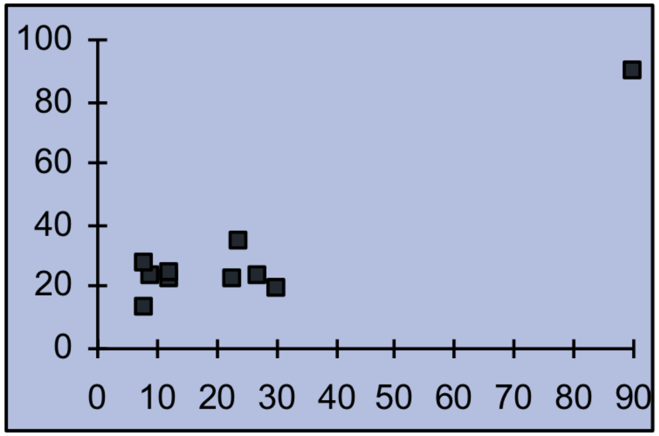


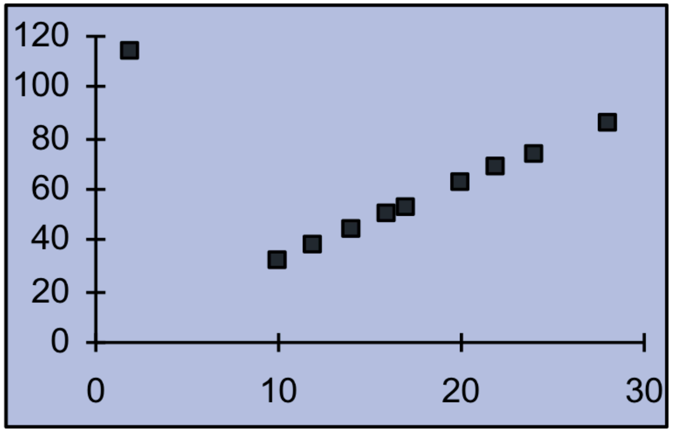


Relação entre Variáveis Quantitativas

- Cuidado!

1. Valores Extremos podem impactar significativamente o valor do coeficiente de correlação.





Relação entre Variáveis Quantitativas

- Cuidado!

2. O coeficiente de correlação **não é uma medida da relação causa-efeito entre duas variáveis. Ele apenas prove evidência de presença ou não de relação linear entre as variáveis X e Y. A presença de relação linear pode ser porque:**

    - ***A variável X de fato explica a variável Y.***
    - ***A variável Y de fato explica a variável X.***
    - ***Há uma terceira variável Z que impacta as variáveis X e Y simultaneamente.***
    - ***Ocasional e não aplicável em outros contextos.***

Agenda

1. Introdução
2. Relação entre Duas Variáveis Quantitativas
3. **Relação entre Duas Variáveis Qualitativas**
4. Relação entre Uma Variáveis Qualitativa e Uma Quantitativa
5. Conclusões

Relação entre Variáveis Qualitativas

- O objetivo de estabelecer a distribuição conjunta de duas variáveis qualitativas é o de compreender a existência de alguma associação entre elas, ou o grau de dependência entre elas.
- Exemplo: Distribuição conjunta de alunos formados segundo sexo (X) e área de formação no ensino superior, dados hipotéticos

X / Y	Feminino	Masculino	Total
Ciências Humanas e Artes	15000	5000	20000
Outras Áreas	55045	40000	95045
Total	70045	45000	115045

Relação entre Variáveis Qualitativas

- Distribuição conjunta das proporções (em %) de formados segundo sexo (X) e área (Y):

X/Y	Feminino	Masculino	Total
Ciências Humanas e Artes	21,41%	11,11%	17,38%
Outras Áreas	78,59%	88,89%	82,62%
Total	100%	100%	100%

- Número esperado de formados, se a formação não guardasse relação com o sexo:

X/Y	Feminino	Masculino
Ciências Humanas e Artes	17,38%*70045	17,38%*45000
Outras Áreas	82,62%*70045	82,62%*45000

Relação entre Variáveis Qualitativas

- Efetuando-se a multiplicação de cada célula, obtém-se os números esperados:

X/Y	Feminino	Masculino
Ciências Humanas e Artes	12176,97	7823,03
Outras Áreas	57868,03	37176,97

- Há, portanto, um desvio entre os números observados e os esperados se não houvesse relação entre as duas variáveis: sexo e formação

Relação entre Variáveis Qualitativas

- Os desvios entre os valores esperados e observados podem ser chamados resíduos
- Para calcular os desvios relativos:

  (v observado_i-v $\text{esperado}_i)^2$/v esperado_i

ou:

$$\frac{(o_i - e_i)^2}{e_i}$$

Relação entre Variáveis Qualitativas

- O quadro abaixo mostra os desvios relativos para cada célula:

X/Y	Feminino	Masculino
Ciências Humanas e Artes	654,47	1018,72
Outras Áreas	137,72	214,37

- Somando todos os valores dos desvios relativos, temos o valor de 2025,28

Relação entre Variáveis Qualitativas

- A soma de todas as medidas de afastamento é uma medida do afastamento global e é chamada qui-quadrado de Pearson e denotada por χ^2
- Quanto maior o χ^2, maior o grau de associação entre as variáveis
- Em nosso exemplo, temos:

$$\chi^2 = 2025{,}28$$

Relação entre Variáveis Qualitativas

- O coeficiente de contingência é uma medida de associação definida por:

$$\boxed{C = \sqrt{\frac{\chi^2}{\chi^2 + n}}}$$

em que n é o número de observações, 115045 em nosso exemplo.

- Contudo, o valor desse coeficiente depende do número de linhas e colunas da tabela.

Relação entre Variáveis Qualitativas

- Por isso, foi definido o coeficiente T que penaliza pelo número total de linhas r e colunas s:

$$\boxed{T = \sqrt{\frac{\chi^2 / n}{(r-1)(s-1)}}}$$

- Em nosso exemplo, r=2 e s=2.

Agenda

1. Introdução
2. Relação entre Duas Variáveis Quantitativas
3. Relação entre Duas Variáveis Qualitativas
4. **Relação entre Uma Variável Qualitativa e Uma Quantitativa**
5. Conclusões

Relação entre Var Quanti e Qualitativa

- É comum, neste caso, ver o que ocorre com a variável quantitativa dentro de cada categoria da variável qualitativa
- Pode-se usar gráficos e tabelas para ver o que acontece
- Para verificar o grau de dependência entre as variáveis, precisamos de um indicador
- As variâncias das variáveis é um instrumento:
  - A variância da var quantitativa mede a dispersão globalmente
  - Se a variância dentro de cada categoria for pequena e menor do que a global, significa que a var qualitativa melhora a capacidade de previsão da quantitativa e, portanto, existe uma relação entre as variáveis

Relação entre Var Quanti e Qualitativa

- Tomemos o exemplo do comportamento dos salários por grau de instrução (Bussab & Morettin)

Grau de instrução	**n**	**Média**	**Variância**
Fundamental	12	7.84	7.77
Médio	18	11.54	13.1
Superior	6	16.48	16.89
Todos	**36**	**11.12**	**20.46**

Relação entre Var Quanti e Qualitativa

- Definimos a média das variâncias, ponderada pelo número de observações em cada categoria:

$$\mathrm{var}^*(X) = \sum\left(\frac{n_i \, \mathrm{var}_i(X)}{n}\right)$$

- O grau de associação entre as variáveis é dado pela redução relativa na variância da variável quantitativa através da introdução da variável qualitativa:

$$R^2 = \frac{\left(var(X) - var^*(X)\right)}{var(X)}$$

Obrigada!