
EXPLORANDO DADOS REAIS À NOSSA MANEIRA...

MEDIDAS DE ASSOCIAÇÃO E GRÁFICO DE DISPERSÃO

#Pré Aula 06

Preparo Prévio:

1. Leitura prévia necessária: Magalhães e Lima (7ª. Edição) – Seção 2.6. Dados Multivariados.

Até a próxima aula:

1. Analisar, graficamente, a associação entre duas variáveis quantitativas.
2. Compreender a medida nomeada *covariância*, no que tange ao significado da fórmula e uso do valor resultante para descrever associação entre variáveis.
3. Avaliar vantagens e desvantagens da medida de covariância e buscar alternativa que compense seu mau uso (coeficiente de correlação).
4. A leitura do livro é fundamental para aprendizado e desenvolvimento deste exercício. Outra opção de livro é: Bussab e Morettin – Capítulo 4 – Seção 4.5.
5. O objetivo principal deste exercício é que você seja capaz de construir um gráfico de dispersão útil para compreender a relação entre duas variáveis quantitativas, calcular a medida de covariância e interpretar ambos resultados (gráfico e valor resultante da covariância). **Pode ser feito no Excel ou no Python!**


Mundo

A análise bidimensional tem como objetivo encontrar associação ou relação entre as variáveis quantitativas. Essas relações podem ser identificadas por meio de gráficos ou medidas numéricas. Entende-se por associação a mudança de opinião sobre o comportamento de uma variável na presença de informação sobre a segunda variável.

Inicialmente, a base de dados <Mundo.xlsx> deverá ser explorada com intuito de entender a associação entre algumas de suas variáveis quantitativas. Esse conjunto de dados contém alguns indicadores socioeconômicos referentes ao ano de 2008 para 85 países, os quais são:

- X1: população em milhares de habitantes
- X2: densidade populacional
- X3: % de população urbana
- X4: expectativa de vida feminina
- X5: expectativa de vida masculina
- X6: crescimento populacional
- X7: mortalidade infantil
- X8: PIB per capita
- X9: % de mulheres alfabetizadas
- X10: população em 100.000 habitantes

1. Um gráfico de dispersão pode ser utilizado para compreender a relação entre duas variáveis quantitativas. Não necessariamente essa relação implica em causalidade¹ e, em alguns casos, pode se tratar de uma relação espúria².

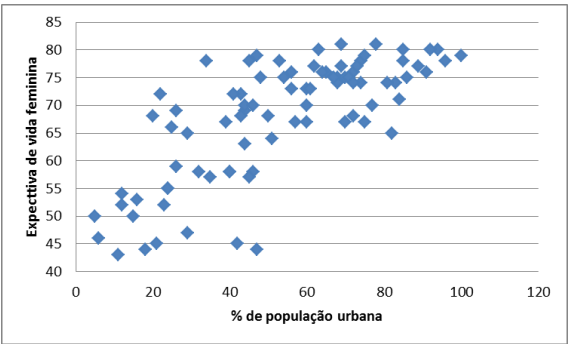



A Figura 1 apresenta um gráfico de dispersão entre as variáveis X3 (eixo das abscissas) e X4 (eixo das ordenadas), em que cada ponto  representa um país da base de dados de acordo com o par de coordenadas nessas duas variáveis quantitativas.

Construa os gráficos de dispersão considerando as variáveis indicadas na Figura 2 (variáveis X3 e X5), Figura 3 (variáveis X3 e X6) e Figura 4 (variáveis X3 e X7).

Para cada um dos quatro gráficos, descreva a relação entre as variáveis.

¹ <https://economiadependrive.wordpress.com/2014/09/25/correlacao-nao-implica-em-causalidade/>

² <http://www.tylervigen.com/spurious-correlations>

	
<p>Figura 1. Gráfico de dispersão entre as variáveis % de população urbana e expectativa de vida feminina</p>	<p>Figura 2. Gráfico de dispersão entre as variáveis % de população urbana e expectativa de vida masculina</p>
	
<p>Figura 3. Gráfico de dispersão entre as variáveis % de população urbana e crescimento populacional</p>	<p>Figura 4. Gráfico de dispersão entre as variáveis % de população urbana e mortalidade infantil</p>

2. O gráfico de dispersão é uma ferramenta descritiva simples, porém útil para examinar uma possível relação entre variáveis quantitativas. A literatura estatística apresenta uma medida, nomeada de covariância, cujo sinal pode ser um indicativo do tipo de associação linear: positiva, se maior que zero; negativa, se menor que zero; e ausente de associação linear, se igual a zero.

Essa medida de covariância é expressa por:

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

A tabela abaixo apresenta a covariância entre as variáveis X3 (% de população urbana) e X4 (expectativa de vida feminina) e seu resultado deve indicar a possível relação entre as variáveis ilustradas na Figura 1, caso essa relação seja linear.

	Expectativa de vida feminina	Expectativa de vida masculina	Crescimento populacional	Mortalidade infantil
% de população urbana	192,0634	?	?	?

Calcule a covariância entre os demais pares de variáveis conforme as Figuras 2, 3 e 4 e preencha a tabela acima. Contraste as interpretações gráficas e os resultados das covariâncias. É possível perceber que a medida expressa na equação (1) é capaz de traduzir o sinal da associação entre as variáveis quantitativas? Por quê? Justifique a sua resposta. *Observação:* faça uma explicação dizendo também porque a expressão (1) é capaz de mensurar corretamente uma possível associação linear.

3. O coeficiente de correlação entre duas variáveis é expresso por:

$$r_{XY} = \text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\text{DP}(X)\text{DP}(Y)} \quad (2),$$

sendo $\text{Cov}(X, Y)$ a covariância entre X e Y, $\text{DP}(X)$ o desvio padrão de X e $\text{DP}(Y)$ o desvio padrão de Y.

Calcule o coeficiente de correlação para os quatro casos acima e avalie o que essa medida traz a mais na sua interpretação. Responda essa pergunta consultando o livro básico e, caso ache interessante, entre no site <http://guessthecorrelation.com/> para melhorar sua compreensão sobre essa medida descrita na equação (2).