SENAI

Ciência de Dados (CIDA)

Aula 3 – Gráficos e Correlação

Prof.: Hugo S. Idagawa

Introdução

- Até o momento apenas trabalhamos com os dados observando uma única variável por vez. Porém, é muito comum na análise estatística verificar a associação entre duas ou mais variáveis.

- A partir desse conhecimento, podemos dizer se o conhecimento do valor de uma variável nos permite obter (ou prever) alguma informação sobre a outra.

- Como exemplo, gostaríamos de estudar a relação existente entre o peso e a altura de pessoas.

- Da mesma forma que na análise de 1 única variável, também iremos utilizar tabelas, gráficos e medidas de resumo para representar a distribuição conjunta das variáveis de interesse. Entre os diferentes estudos possíveis de associação, podemos destacar 3 casos:

  - ✓ Duas variáveis qualitativas
  - ✓ Duas variáveis quantitativas
  - ✓ Uma variável qualitativa e outra quantitativa

Introdução

- Enquanto avaliamos as relações entre as variáveis, é importante classificar as variáveis segundo à forma de coleta dos dados:

  - **Variáveis explicativas (ou preditoras):** são aquelas cujas categorias ou valores são fixos, geralmente devido ao planejamento do experimento.

  - **Variáveis respostas:** são aquelas cujos valores são aleatórios.

- **Exemplo:** Deseja-se avaliar o efeito do tipo de aditivo adicionado ao combustível no consumo de automóveis. Nesse teste, um conjuntos de 5 automóveis (de mesmo modelo) foi observado sob o tratamento com um de 4 tipos de aditivo. O consumo (em km/L) foi avaliado após um determinado período de tempo. Qual seria a variável explicativa e qual seria a variável resposta nesse experimento?

Resposta:

- **Variável explicativa:** variável qualitativa "Tipo de aditivo" (com 4 categorias).
- **Variável resposta:** variável quantitativa "Consumo de combustível".

Gráfico de Dispersão

- Uma das principais ferramentas para avaliar a associação entre duas variáveis quantitativas é o **gráfico de dispersão**. Nesses gráficos, temos um conjunto de n pares de valores (x_i, y_i) de duas variáveis X e Y. Nesse gráfico, cada par (x, y) é desenhado em um plano cartesiano.

- **Exemplo:** na tabela abaixo, temos um conjunto de dados amostrais de 10 crianças com o seu respectivo valor de peso (kg) e altura (cm). Ao seu lado temos o gráfico de dispersão desses dados. Existe alguma correlação entre o peso e a altura?

Criança	*Peso*	*Altura*
1	30	145
2	32	150
3	24	125
4	28	140
5	26	127
6	34	145
7	25	132
8	23	112
9	35	155
10	31	145

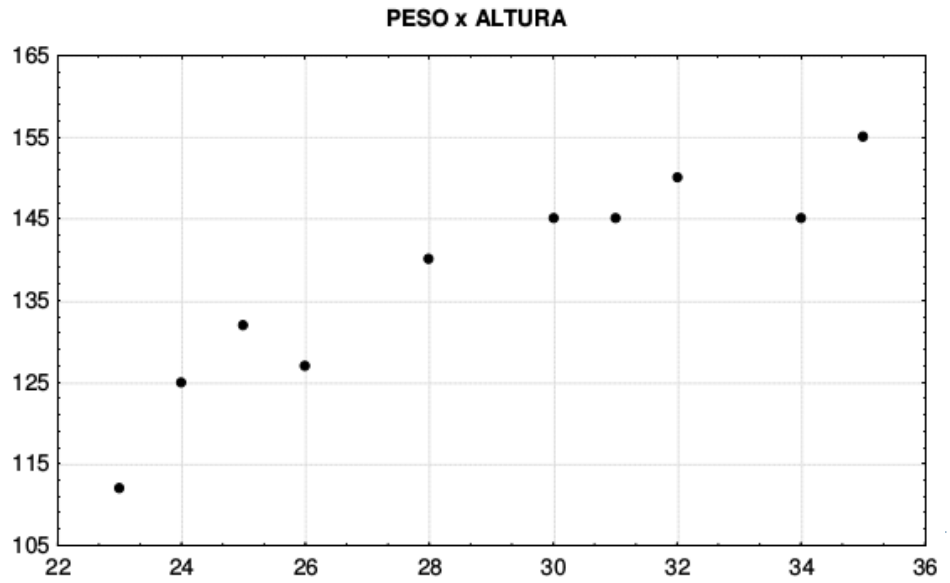


Gráfico de Dispersão

- Para desenhar essa dispersão podemos utilizar as seguintes funções do `matplotlib`:
  - `scatter(x, y, s=20, c='r', marker='o')`: produz um gráfico de dispersão. Os parâmetros, s (tamanho), c (cor) e marker (marcador), são parâmetros opcionais que controlam o desenho do gráfico.
  - `plot(x, y, color='red', marker='o', linewidth=2, markersize=12)`: produz um gráfico de linhas, onde os parâmetros color, marker, linewidth e markersize são parâmetros opcionais que controlam o desenho do gráfico.
- **Ex:** Implemente em python um programa que apresenta a dispersão do exemplo de alturas e pesos anterior. Experimento com os parâmetros opcionais para entender o funcionamento das funções scatter e plot.

Correlação

- A correlação é uma medida estatística com o objetivo de quantificar a “força” da relação existente entre duas variáveis.

- Um desafio na determinação da correlação ocorre quando as variáveis que queremos comparar geralmente não estão expressas na mesma unidade. E, mesmo quando, estão, podem apresentar distribuições diferentes.

- Assim, existem duas soluções comuns para esses problemas:

  1. Transformar cada valor para uma graduação padronizada e realizar a correlação. Nesse caso temos o coeficiente de correlação de Pearson.

  2. Transformar cada valor em uma posição ranqueada de valores. Nesse caso tempos o coeficiente de correlação de Spearman.

Correlação de Pearson

- A correlação de Pearson funciona bem quando a relação entre as variáveis em estudo é linear e as suas distribuições são aproximadamente normais.

- Para deteminar o coeficiente de correlação, inicialmente precisamos saber qual a covariância entre as duas viaráveis em estudo. A covariância é uma medida de tendência das variáveis de se modificarem em conjunto.

- Para se calcular a covariância, precisamos inicialmente determinar o desvio da média de cada dado da amostra, utilizando a fórmula abaixo:

$$dx_i = x_i - \bar{x}$$

$$dy_i = y_i - \bar{y}$$

- A partir dos resultados abaixo, podemos calcular a covariância (Cov(X,Y)):

$$Cov(X,Y) = \frac{1}{n}\sum dx_i \; dy_i$$

Correlação de Pearson

- Apesar da covariância ser útil, ela é raramente apresentada como um resultado estatístico porque é um número difícil de ser interpretado, principalmente devido às unidades das variáveis em estudo.

- Uma solução para esse problema é simplesmente dividir os desvios da média pelo desvio-padrão das amostras (Sx e Sy). Isso produz um conjunto de valores chamado de valores padrão (“standard scores”):

$$p_i = \frac{(x_i - \bar{x})}{S_X} \frac{(y_i - \bar{y})}{S_Y}$$

- Ao tirar a média dos resultados anteriores, podemos reescrever a correlação entre as variáveis da seguinte forma:

$$\rho = \frac{Cov(X,Y)}{S_X S_Y}$$

Correlação de Pearson

- O valor ***ρ*** calculado anteriormente é chamado de coeficiente de Correlação de Pearson e é sempre um número entre -1 e +1 (inclusive).

- Se esse número for positivo, significa que a correlação entre as variáveis é positiva, ou seja, se uma varia para mais a outra também varia positivamente. Caso contrário, se esse número for negativo, significa que a correlação entre as variáveis é negativa, ou seja, se uma varia para mais a outra também varia para menos.

- A magnitude do valor de ***ρ*** indica a força dessa correlação. Ou seja, se ***ρ=1***, significa uma correlação positiva perfeita, enquanto que ***ρ=-1***, significa uma correlação negativa perfeita.

Correlação de Pearson

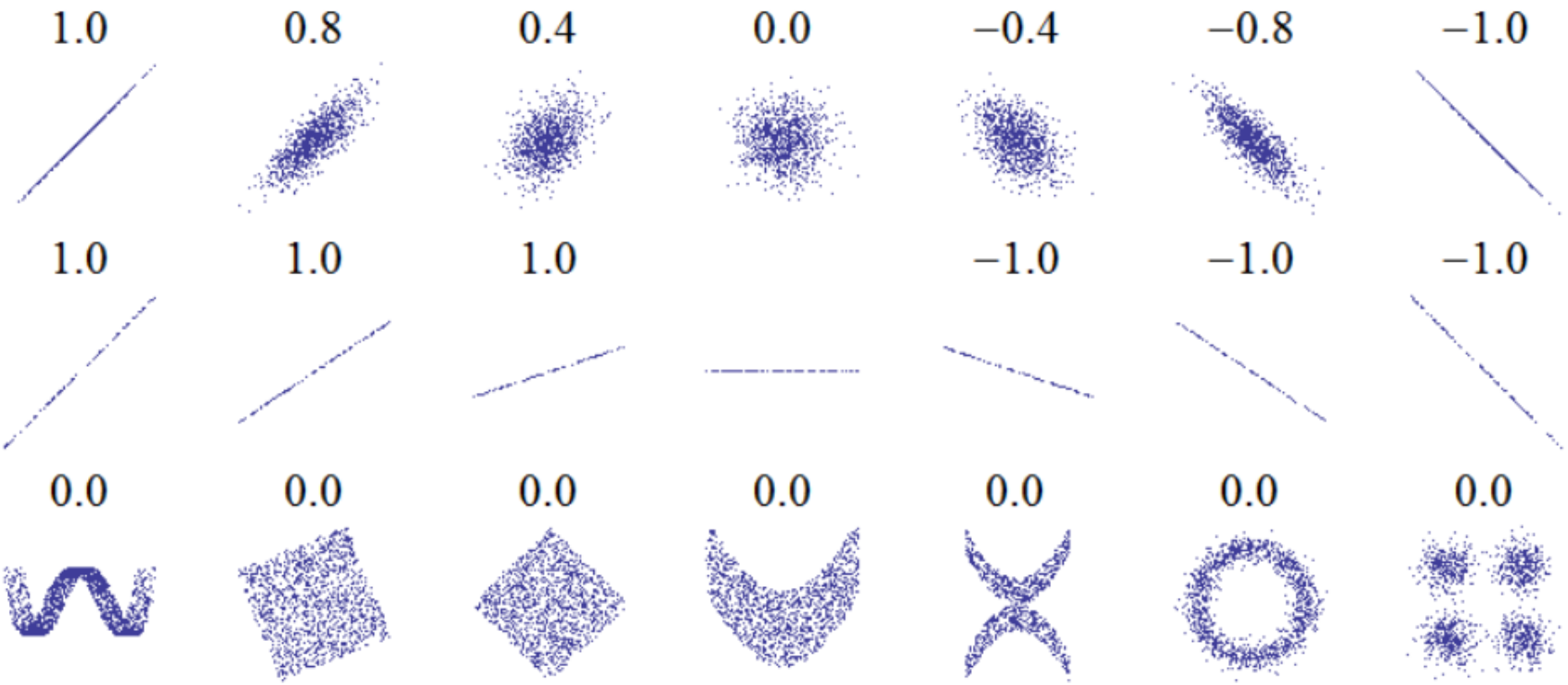
- O quadro abaixo, apresenta algumas regras que auxiliam na interpretação e no uso do coeficiente de correlação de Pearson:

Interpretação	Coeficiente de correlação linear de Pearson
Forte associação positiva	(0,9; 1]
Alta associação positiva	(0,7; 0,9]
Moderada associação positiva	(0,5; 0,7]
Baixa associação positiva	(0,3; 0,5]
Associação nula	[0; 0,3]

Interpretação	Coeficiente de correlação linear de Pearson
Forte associação negativa	[-1; -0,9)
Alta associação positiva	[-0,9; -0,7)
Moderada associação negativa	[-0,7; -0,5)
Baixa associação negativa	[-0,5; -0,3)
Associação nula	[-0,3; 0]

Correlação de Pearson

- O esquema abaixo apresenta diferentes dispersões, com os seus respectivos valores de correlação de Pearson. Observe que esse coeficiente é muito útil para os casos de correlação linear:

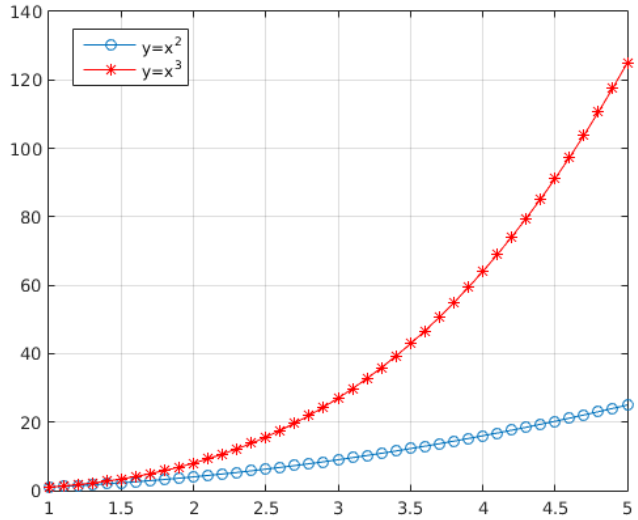


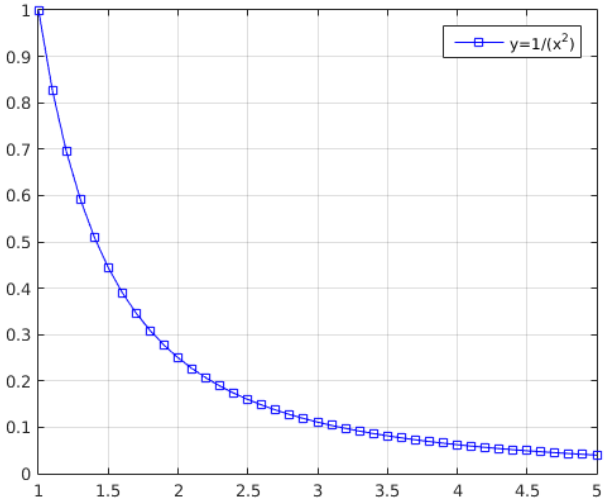
Correlação de Pearson

- Conforme pode ser visto no slide anterior, o coeficiente de Pearson é útil quando temos uma correlação linear entre as variáveis. Como devemos proceder se as variáveis em estudo não tiverem uma correlação linear?

- **Exemplos:**

  - $y = x^2$
  - $y = x^3$
  - $y = 1/x^2$





Correlação de Pearson

- Nos casos do exemplo anterior, é possível realizar uma transformação que lineariza as relações de dependência entre as variáveis
- Uma transformação bastante comum é a utilização do logarítmo:
- **Exemplos:**

- $y = x^2 \rightarrow \log(y) = \log(x^2) \rightarrow \log(y) = 2 * \log(x) \rightarrow Y = 2 * X \ (relação\ linear)$
- $y = x^3 \rightarrow \log(y) = \log(x^3) \rightarrow \log(y) = 3 * \log(x) \rightarrow Y = 3 * X \ (relação\ linear)$
- $y = \frac{1}{x^2} \rightarrow y = x^{-1} \rightarrow \log(y) = \log(x^{-1}) \rightarrow \log(y) = -1 * \log(x) \ (relação\ linear)$

Correlação de Pearson

- **Exercício:** um determinado estudo sobre a massa de diferentes peças, mapeou 3 variáveis diferentes (altura, lado e processo) para verificar a influência de cada uma dessas variáveis na massa da peça final. Utilizando o arquivo de dados de entrada “dados_corr.csv” localizado no repositório, faça um estudo de correlação entre essas variáveis, apontando qual variável tem influência na massa da peça e apresente os gráficos de dispersão relevantes. Além disso, verifique se existe alguma relação não linear entre as variáveis e faça as transformações necessárias para analisar.