Análise Exploratória de Dados - Avaliação Presencial 01 - 2022/01

Prof. Hugo Carvalho

28/06/2022

Questão 1: *(Coeficiente de correlação linear de Pearson)* Sejam $\{x_1, \dots, x_n\}$ e $\{y_1, \dots, y_n\}$ dois conjuntos de observações de variáveis quantitativas. Defina o *coeficiente de correlação linear* entre x e y como

$$\text{cor}(x, y) = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\text{dp}(x)\text{dp}(y)},$$

onde, para relembrar,

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{e} \quad \text{dp}(x) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2},$$

sendo tais quantidades definidas analogamente para y.

a) Com base na fórmula proposta para o cálculo de $\text{cor}(x, y)$, interprete o que tal quantidade visa medir.
Dica: Olhe primeiro para o que o numerador faz, pois o denominador é somente uma constante normalizadora; depois, discuta sobre a importância do denominador. Ao olhar para o numerador, comece olhando a quantidade "de dentro para fora", ou seja, comece se perguntando o que $x_i - \overline{x}$ mede; note que tal quantidade tem um sinal, portanto, pergunte-se se tal sinal é importante; depois questione-se sobre o que mede o produto de $x_i - \overline{x}$ por $y_i - \overline{y}$, e assim sucessivamente.

b) O que podemos dizer sobre $\text{cor}(x, y)$ se os dados se relacionam de forma perfeitamente linear, ou seja, se tivermos que $y_i = ax_i + b$, para todo $i = 1, \dots, n$? Justifique matematicamente a sua resposta.

c) Construa dados artificiais $\{x_1, \dots, x_n\}$ e $\{y_1, \dots, y_n\}$, sendo $y_i = g(x_i)$ para uma função $g : \mathbb{R} \to \mathbb{R}$, de modo que $\text{cor}(x, y) = 0$. Interprete o que isso diz sobre o coeficiente de correlação linear.
Dica: Com $n = 3$ já é possível construir tal exemplo. O resultado do item b) te dá um indício de como ***não*** *pode ser a função g.*

d) Com base no que você desenvolveu nos itens anteriores, cite vantagens e desvantagens do coeficiente de correlação linear.

e) Mostre que uma outra forma de calcular $\text{cor}(x, y)$ é dada por

$$\text{cor}(x, y) = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}}{\sqrt{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}\sqrt{\sum_{i=1}^{n} y_i^2 - n\overline{y}^2}}.$$

Questão 2: Considere uma enfermidade que afeta aproximadamente 5% de determinada população. Você começa a sentir sintomas esquisitos e resolve se testar para tal doença, e a bula do teste traz as seguintes informações:

- Garantia de 90% de verdadeiros positivos;
- Garantia de 80% de verdadeiros negativos.

Tais informações devem ser interpretadas, respectivamente, como: "em 90% dos enfermos o teste dá positivo" e "em 80% dos não enfermos o teste dá negativo". Você faz o teste (corretamente) e o resultado é positivo. Seu interesse agora é saber o quanto essa informação te confirma se você tem ou não a enfermidade.

a) Argumente que a chance de você ter a enfermidade sabendo que o teste deu positivo **não** é de 90%.

b) Estude a chance de você ter a enfermidade sabendo que o teste deu positivo. Para isso, faça uma pequena "simulação": considere um universo de 1.000 pessoas, de modo que 950 estarão saudáveis e 50 estarão enfermas; veja o que acontece com cada um desses grupos ao se testarem, ou seja, qual proporção dos 950 terá teste positivo/negativo e qual proporção dos 50 terá teste positivo/negativo; finalmente, dentro da parcela da população com teste positivo veja quem de fato tem a enfermidade.