Análise Exploratória de Dados - Avaliação Presencial - Gabarito - 2023/01

Prof. Hugo Carvalho

25/05/2023

Questão 1:

a) Seguindo a dica, interpretamos $\text{med}(x)$ como a observação central do conjunto de dados, ou seja, abaixo dela temos 50% dos dados e acima dela temos os outros 50%. Os y_i são definidos como $y_i = |x_i - \text{med}(x)|$, ou seja, são os desvios de cada observação x_i em relação à mediana de x, em valor absoluto. Finalmente, a mediana dos y_i divide esse conjunto pela metade: abaixo dela temos metade dos $|x_i - \text{med}(x)|$ e acima dela temos a outra metade dos $|x_i - \text{med}(x)|$. Ou seja, $\text{dma}(x)$ é um sumário numérico dos desvios de x em relação à uma medida de centralidade.

b) Em relação ao conjunto x, temos que $\text{dma}(x)$ é uma medida de dispersão, pois ela sumariza numericamente distâncias em relação à uma medida de centralidade (a mediana de x). Compare, por exemplo, com a variância, definida como

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

ou seja, uma média dos desvios quadráticos em relação à média. O espírito de ambas as métricas é o mesmo, porém trocando-se a média pela mediana e o desvio quadrático pelo desvio absoluto.

c) Pensemos, novamente, de dentro para fora: a mediana de x é pouco sensível em relação à valores discrepantes, e mesmo caso tenhamos algum $|x_i - \text{med}(x)|$ grande, esse valor possivelmente será "filtrado" pela mediana de fora. Dessa forma, $\text{dma}(x)$ é uma medida de dispersão que é *pouco* sensível a valores discrepantes. Para se convencer, faça as contas com o conjunto $x = \{1, 2, 3, 4, 1000\}$ e verifique que $\text{med}(x) = 3$ e consequentemente $\text{dma}(x) = \text{med}(\{2, 1, 0, 1, 997\}) = 1$.

d) O desvio médio absoluto em relação à mediana difere do desvio mediano absoluto por calcular não a mediana de $|x_i - \text{med}(x)|$, mas sim a sua *média*. Sabemos que a média é mais sensível a observações discrepantes do que a mediana, portanto podemos esperar que o desvio médio absoluto em relação à mediana seja mais sensível a valores discrepantes. Note que isso não é uma desvantagem nem uma vantagem dessa métrica, mas sim uma característica dela: enquanto que o desvio mediano absoluto "filtra" grandes desvios da mediana de x, o desvio médio absoluto em relação à mediana "alerta" de sua existência. Exemplificando com o conjunto de dados anterior, temos que $\text{dmarm}(x) = 200{,}2$.

e) Seguindo a dica, temos que se $\text{dma}(x) = 0$ então a mediana das quantidades $|x_i - \text{med}(x)|$ é zero. Isso significa que metade de tais valores são menores que ou iguais a zero e a outra metade será maior que ou igual a zero. Porém, temos que $|x_i - \text{med}(x)|$ não pode ser negativo. Dessa forma, temos que metade dos valores $|x_i - \text{med}(x)|$ é *igual* a zero e a outra metade é estritamente positiva. Daí, podemos concluir que ao menos metade dos valores x_i coincidem exatamente com a mediana de x, e a outra metade, não. Caso isso pareça muito esotérico para você, façamos um exemplo: considere $x = \{1, 1, 1, 2, 2\}$. Temos que $\text{med}(x) = 1$ e $\text{dma}(x) = \text{med}(\{0, 0, 0, 1, 1\}) = 0$.

Obs.: Se você chegou a outras conclusões, por exemplo, "todas as observações x_i devem ser iguais", ou "todos os x_i devem ser iguais à zero", note que não está errado! Você apenas não capturou o cenário mais geral possível, mas foi por um caminho correto. Caso você tenha justificado seu raciocínio direitinho, sua questão não estará totalmente incorreta!

Questão 2: Aqui eu preferi falar sobre pontos interessantes no video em vez de escrever muito longamente. Vejam lá os meus pontos levantados!