Análise Exploratória de Dados - Avaliação Presencial 01 (Gabarito) - 2022/01

Prof. Hugo Carvalho

28/06/2022

Questão 1:

a) Seguindo a dica e indo "de dentro para fora", temos que $x_i - \overline{x}$ mede o desvio de cada observação x_i em relação à média $\overline{x}$; como tal quantidade tem sinal, ela ainda "guarda" a informação se x_i está "acima" ou "abaixo" da média. Analogamente, vale o mesmo para a quantidade $y_i - \overline{y}$. Passemos então ao produto de tais quantidades, $(x_i - \overline{x})(y_i - \overline{y})$: sua magnitude mede conjuntamente o quanto que as observações x_i e y_i estão longe de suas respectivas médias, e o seu sinal mede se elas estão "do mesmo lado" da média (se for positivo) ou "do lado oposto" da média (se for negativo). Finalmente, a média das quantidades $(x_i i\overline{x})(y_i - \overline{y})$ mede uma espécie de "tendência conjunta média" de dispersão de ambos os conjuntos de dados de suas respectivas médias; o sinal de tal quantidade mede portanto qual é "direção preferencial" de tal tendência de dispersão das respectivas médias. O enunciado sugere que o denominador é somente uma constante normalizadora, portanto, o interpretemos somente assim. Uma interpretação alternativa é pensar em $\text{cor}(x, y)$ como $\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{\text{dp}(x)}\right)\left(\frac{y_i - \overline{y}}{\text{dp}(y)}\right)$, ou seja, a média dos desvios em relação à média multiplicados, porém devidamente padronizados. O restante da interpretação segue analogamente como feito acima.

b) Se $y_i = ax_i + b$ temos que:

$$\begin{aligned}
\text{cor}(x, y) &= \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\text{dp}(x)\text{dp}(y)} \\
&= \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(ax_i + b - (\overline{ax_i + b}))}{\text{dp}(x)\text{dp}(ax + b)} \\
&\overset{*}{=} \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(ax_i - a\overline{x_i})}{\text{dp}(x)|a|\text{dp}(x)} \\
&= \frac{a}{|a|}\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}{\text{dp}(x)^2} \\
&= \frac{a}{|a|},
\end{aligned}$$

sendo a última quantidade igual a 1 ou -1, se a é positivo ou negativo, respectivamente, e não está definida se $a = 0$. A passagem feita em $*$ foi adequadamente detalhada e explicada em aula, e ela consiste em atividades que foram passadas como exercício ao longo das aulas presenciais.

c) Pelo enunciado da questão b), temos que g não pode ser uma função linear. Temos também que todos os x_i e todos os y_i devem ser distintos entre si, para que o denominador de $\text{cor}(x, y)$ não dê zero. A fim de simplificar a fórmula, pensemos em um conjunto de dados que tenha média zero, digamos, $\{-1, 0, 1\}$, e como sabemos que a função g não pode ser linear, tentemos, digamos, $g(x) = x^2$. Nesse caso, temos que o conjunto de dados y é dado por $\{1, 0, 1\}$, cuja média é $2/3$. Calculando portanto somente o numerador de $\text{cor}(x, y)$, visto que o denominador não será nulo, temos que $\text{cor}(x, y) = 0$. Isso indica que o coeficiente de correlação linear tem dificuldades em "captar" dependências não-lineares entre conjuntos de dados, sendo mais adequada para "medir" dependências lineares, como o nome sugere.

d) Conforme mencionado no item c), uma possível desvantagem é a dificuldade em capturar dependências não-lineares; vantagens são a facilidade de seu cálculo e o fato de ser uma medida adimensional, por exemplo. Outras vantagens e desvantagens foram discutidas em sala.

e) Analisando a fórmula para o cálculo de $\text{cor}(x, y)$, façamos a expansão primeiro do numerador:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) &= \frac{1}{n}\sum_{i=1}^{n}(x_i y_i - x_i\overline{y} - y_i\overline{x} + \overline{x}\,\overline{y}) \\ &= \frac{1}{n}\sum_{i=1}^{n}(x_i y_i) - \overline{x}\,\overline{y} - \overline{x}\,\overline{y} + \overline{x}\,\overline{y} \\ &= \frac{1}{n}\left[\sum_{i=1}^{n}(x_i y_i) - n\overline{x}\,\overline{y}\right],\end{aligned}$$

algo bem próximo com o numerador da nova expressão, a menos do fator multiplicativo $1/n$. Um procedimento análogo de expansão (que foi passado como exercício nas aulas presenciais) nos dará que

$$\text{dp}(x) = \sqrt{\frac{1}{n}\left[\sum_{i=1}^{n}(x_i^2) - n\overline{x}^2\right]}.$$

Analogamente, vale resultado semelhante para y. Substituindo as novas expressões para o numerador e denominador na fórmula para $\text{cor}(x, y)$, obtemos o resultado desejado.

Questão 2:

a) Há diversas formas possíveis de argumentar que foram discutidas em sala, mas aqui ilustro uma: a pergunta que é feita no item a) é diferente da pergunta para qual a resposta é 90%. Mais especificamente, queremos saber a resposta para "sabendo que o teste deu positivo, qual a chance de estar enfermo?" enquanto que 90% é a resposta para a pergunta "sabendo que está enfermo, qual a chance do teste dar positivo?". Dessa forma, sendo perguntas diferentes, não há porque as respostas serem as mesmas, a princípio!

b) Façamos a "simulação" proposta. Assumindo o universo de 1.000 pessoas, temos 950 saudáveis e 50 enfermas. Caso todas se testem, temos que 90% das 950 saudáveis (ou seja, 760) saudáveis terão teste negativo, e os 20% restantes (ou seja, 190) terão teste positivo. Analogamente, 90% das 50 enfermas (ou seja, 45) terão teste positivo e os 10% restantes (ou seja, 5) terão teste negativo. Dessa forma, em um universo de $190 + 45 = 235$ pessoas cujo teste deu positivo somente 45 estão efetivamente enfermas, uma fração de $45/235 \approx 19\%$. Dessa forma, a resposta para a pergunta "sabendo que o teste deu positivo, qual a chance de estar enfermo?" é de aproximadamente 19%.