

Ficha 9 — Dados bivariados: correlação e reta de regressão

10.º Ano

Matemática A

4 páginas

12 exercícios

Objetivo: Construir e ler diagramas de dispersão; interpretar o coeficiente de correlação $r \in [-1, 1]$; usar a reta de mínimos quadrados para fazer inferências; distinguir correlação de causalidade (caso Anscombe).

Revisão rápida — dois dados por observação

- **Dados bivariados:** cada observação é um par $(x, y) \rightarrow$ diagrama de dispersão (nuvem de pontos). Lê: *direção* (crescente/decrescente), *força* (nuvem apertada vs. difusa), *forma* (linear vs. curva), *outliers*.
- **Coeficiente de correlação r** (Pearson): mede só a **associação linear**: sinal = direção $\cdot |r|$ = força $\cdot r = 0 \rightarrow$ **sem relação linear** (pode haver curva!). $r = +1/-1$ = perfeita.
- **Reta de mínimos quadrados** $\hat{y} = a \cdot x + b$: a que **minimiza a soma dos quadrados das distâncias verticais** aos pontos — serve para **prever** y a partir de x (fiável *entre* os dados observados).
- **Correlação $\neq$ causalidade**: r alto pode vir de confundidor (ex.: gelo derretido $\leftrightarrow$ acidentes — ambos ligados ao calor) ou de coincidência. **Anscombe**: 4 conjuntos com mesmo $\bar{x}$, σ , $r \approx 0,816$ e mesma reta — gráficos completamente diferentes $\rightarrow$ *desenha sempre o gráfico*.

Exercícios

1. Define dados bivariados e o que é um par ordenado — dá um exemplo real de duas variáveis de uma turma (x e y) e escreve 3 observações.

2. Leitura de dispersão: para a nuvem «sobe da esquerda para a direita, apertada, com um ponto muito acima do resto»: a) direção ___ b) força ___ c) outlier: como se chama e o que fazer? ___

3. Interpretar r : a) $r = 0,85$ ___ b) $r = -0,30$ ___ c) $r = 0$ ___ — por que é que $r = 0$ não significa « x não influencia y » de todo?

4. Sinal e extremos: a) se y diminui quando x sobe, o sinal de r : ___ b) o que é $r = -0,99$? ___ c) ordena por «força de associação linear»: $0,4 / -0,9 / 0,1 / 0,7 \rightarrow$ ___

5. Mínimos quadrados: o que a reta minimiza? (1 frase) e porque é que se usam *quadrados* das distâncias em vez das distâncias absolutas?

6. Prever com $\hat{y} = 0,8x + 2$: a) y para $x = 10$ ___ b) y para $x = 0$ ___ (é o parâmetro?) c) os dados iam de $x = 2$ a $x = 15$ — prever para $x = 40$ é ___ porque ___.

7. Correlação $\neq$ causalidade: explica por que «países com mais geladeiras têm mais acidentes» não prova causalidade — nomeia o confundidor e dá mais 2 exemplos curtos.

8. Anscombe: 4 conjuntos com médias idênticas, σ idênticas e $r = 0,816$ — mas 4 gráficos distintos. a) o que cada gráfico mostra de diferente? b) lição metodológica (1 frase) c) porque é que só o r falharia aqui?

9. Algoritmo (Python):

```
x=[1,2,3,4,5]; y=[2,4,5,4,5]
n=len(x); mx=sum(x)/n; my=sum(y)/n
num=sum((a-mx)*(b-my) for a,b in zip(x,y))
print(round(num,1))
```

a) impresso __ b) sinal do num → a nuvem cresce ou decrescente? __ c) completa até ao r: $S_{xx} = 10$, $S_{yy} = 6 \rightarrow r = \text{num} / \sqrt{(S_{xx} \cdot S_{yy})} = _$ (2 casas)

10. Extrapolação: reta de notas por horas de estudo ajustada a $x \in [0, 10]$ h. a) por que prever a 25 h é arriscado? b) nomeia 2 razões (não-linearidade / sem dados / outros fatores).

11. Projeto (planificação): «Horas de sono vs. notas na turma» — escreve: pergunta estatística (1) · como recolhes (2) · que gráfico + r calculas (3) · que inferência é legítima (4) · uma ameaça de enviesamento (5).

12. Caso: redes sociais vs. notas: $r = -0,42$, reta $\hat{y} = -1,2x + 18$ ($x = \text{h/dia}$, $y = \text{nota}$). a) prever para $x = 4$ e $x = 8$: __ b) que tipo e que força de associação? c) escreve uma conclusão honesta (3 frases: número + cuidado causal).

Soluções — Ficha 9: Dados bivariados

1. *Dados bivariados*: registos com **duas variáveis** por observação (par) · *Par ordenado*: (x, y) — 1.º valor = explicativa, 2.º = resposta (ex.: (3, 14), (4, 13), (5, 16) para horas de sono $\times$ nota).

2. a) **crescente (positiva)** — y aumenta com x b) **forte** — nuvem apertada em redor de uma direção c) **outlier** (ponto fora do padrão): investigar (erro de registo?) e, se for legítimo, mostrar o gráfico com e sem ele — ele pode inflacionar ou inverter r .

3. a) associação linear **positiva forte** — sobe com x , nuvem apertada b) **negativa fraca** — tendência a descer, mas muito difusa (x explica pouco de y) c) **sem relação linear** — mas pode existir relação *curva* (ex.: $y = x^2$ simétrico dá $r = 0$) ou efeito escondido por subgrupos.

4. a) sinal **negativo** b) associação linear **negativa quase perfeita** c) por $|r|$: $|-0,9|$ (0,9) > 0,7 > 0,4 > 0,1 $\rightarrow -0,9, 0,7, 0,4, 0,1$.

5. Minimiza a **soma dos quadrados das distâncias verticais** entre pontos e reta. Quadrados: (1) eliminam os sinais (distâncias para cima e para baixo não se cancelam) (2) penalizam *mais* os pontos longe da reta — evita que um desvio grande seja compensado por um desvio no sentido oposto.

6. a) $0,8 \times 10 + 2 = 10$ b) $0 + 2 = 2 \rightarrow$ é a **ordenada na origem (b)** — valor previsto quando $x = 0$ c) **extrapolação**, arriscada — não há observações ali; a relação pode não manter-se linear fora de [2, 15] (mudam fatores e a própria inclinação).

7. O **confundidor é a riqueza** (e o desenvolvimento): países ricos compram mais geladeiras e têm mais tráfego/indústria $\rightarrow$ acidentes; a geladeira não aciona nada. Outros: «sardinas $\rightarrow$ longevidade» (países piscatórios têm melhor dieta/saúde) · «acidentes de gelo derretido $\times$ nadadores» (o calor impulsiona ambos) · «nº de bombeiros $\times$ estragos do incêndio» (mais incêndio $\rightarrow$ mais bombeiros — efeito inverso do esperado!).

8. a) ex.: um é linear apertado, um tem nuvem difusa, um é uma curva (parábola), outro é dois grupos com um outlier que «faz a reta» b) **sempre visualizar os pontos antes de confiar em r /reta** c) porque r resume só a relação *linear média* — valores idênticos de r /reta escondem formas (curva, grupos) que mudam tudo a conclusão.

9. a) $mx = 3, my = 4 \rightarrow (-2)(-2) + (-1)(0) + (0)(1) + (1)(0) + (2)(1) = 4 + 0 + 0 + 0 + 2 = 6,0$ b) positivo $\rightarrow$ **crescente** c) $r = 6/\sqrt{(10 \times 6)} = 6/\sqrt{60} = 6/7,746 = 0,77$.

10. a) $x = 40$ está muito fora de [0, 10] — extrapolar assume que a reta continua para sempre, sem qualquer observação que a confirme b) (1) a relação real pode ser não-linear (achatou-se, saturou) (2) outros fatores dominam em extremos (sono de 25 h não existe; nota limitada a 20), logo a previsão sai do domínio válido da resposta.

11. (modelo) 1) «Existe associação entre horas de sono e nota na turma?» 2) censo/sorteio dos 24 alunos com consentimento, variáveis cronometradas 7 dias + nota atual 3) diagrama de dispersão + r (e reta se linear) 4) legítimo: descrever associação *nesta turma* (não «sono causa notas») 5) enviesamento de seleção (só quem respondeu) ou confundidor (horas de sono correlacionam com horas de estudo).

12. a) $x = 4: -4,8 + 18 = 13,2$ · $x = 8: -9,6 + 18 = 8,4$ b) **negativa de força moderada** ($|-0,42|$ entre 0,3 e 0,7 — conta, mas há bastante dispersão em redor) c) (modelo) «Nesta amostra, mais horas de redes estão associadas a notas mais baixas (reta sugere -1,2 pontos por hora extra; $r = -0,42$, moderada). A dispersão é grande — previsões individuais são pouco fiáveis. Como é observacional, não se conclui causalidade: talvez mais tempo à frente do ecrã roube sono/estudo (ou notas baixas levem a mais ecrã) — seria preciso controlar confundidores ou um experimento.»