Capítulo VII – Covariância e Correlação

- Covariância na propagação de erros
- Coeficiente de Correlação Linear

135

Covariância na Propagação de Erros

- Suponhamos que, para acharmos o valor da função q(x, y), medimos duas quantidades x e y várias vezes, obtendo N pares de dados (x_1, y_1), ..., (x_N, y_N).
- A partir das N medidas x_1, ... x_N, podemos calcular a média $\bar{x}$ e o desvio padrão σ_x da forma habitual.
- De modo análogo, dos y_1, ..., y_N, podemos determinar $\bar{y}$ e σ_y .
- Por outro lado, usando os N pares de medidas, podemos calcular N valores da quantidade que nos interessa: $q_i(x_i, y_i)$, i = 1, ..., N.
- Dados os q_1, ... q_N obtidos, podemos então calcular a média $\bar{q}$, que assumimos ser a melhor estimativa para q, e o desvio padrão σ_q, que é a nossa medida da incerteza aleatória nos valores q_i.
- Assumindo que todas as incertezas são pequenas e, portanto, que todos os valores x_1, ... x_N estão perto de $\bar{x}$ e todos os y_1, ..., y_N estão perto de $\bar{y}$ podemos fazer a aproximação

$$q_i = q(x_i, y_i) \approx q(\bar{x}, \bar{y}) + \frac{\partial q}{\partial x}(x_i - \bar{x}) + \frac{\partial q}{\partial y}(y_i - \bar{y}) \qquad (7.1)$$

136

$$q_i \approx q(\bar{x},\bar{y}) + \frac{\partial q}{\partial x}(x_i - \bar{x}) + \frac{\partial q}{\partial y}(y_i - \bar{y}) \qquad (7.1)$$

- Nesta expressão, as derivadas parciais $\partial q/\partial x$ e $\partial q/\partial y$ são calculadas no pontos $x = \bar{x}$ e $y = \bar{y}$ e têm, portanto, o mesmo valor para todos os i = 1, ... N. Com esta aproximação a média vem:

$$\bar{q} = \frac{1}{N}\sum_{i=1}^{N} q_i = \frac{1}{N}\sum_{i=1}^{N}\left[q(\bar{x},\bar{y}) + \frac{\partial q}{\partial x}(x_i - \bar{x}) + \frac{\partial q}{\partial y}(y_i - \bar{y})\right]$$

- O valor médio de q corresponde então à soma de 3 termos: o 1º é apenas $q(\bar{x},\bar{y})$ e os outros 2 são exactamente zero.

- Chegamos assim ao resultado notavelmente simples:

$$\boxed{\bar{q} = q(\bar{x},\bar{y})} \qquad (7.2)$$

ou seja, para determinarmos o valor médio $\bar{q}$, temos apenas de calcular a função q(x, y) no ponto $x = \bar{x}$ e $y = \bar{y}$.

137

- O desvio padrão associado aos N valores q_1, ... q_N é dado por:

$$\sigma_q^2 = \frac{1}{N}\sum_i (q_i - \bar{q})^2$$

- Substituindo 7.1 e 7.2, vem:

$$\sigma_q^2 = \frac{1}{N}\sum_i \left[\frac{\partial q}{\partial x}(x_i - \bar{x}) + \frac{\partial q}{\partial y}(y_i - \bar{y})\right]^2$$

$$= \left(\frac{\partial q}{\partial x}\right)^2 \frac{1}{N}\sum_i (x_i - \bar{x})^2 + \left(\frac{\partial q}{\partial y}\right)^2 \frac{1}{N}\sum_i (y_i - \bar{y})^2 + 2\frac{\partial q}{\partial x}\frac{\partial q}{\partial y}\frac{1}{N}\sum_i (x_i - \bar{x})(y_i - \bar{y})$$

Nos 2 primeiros termos aparecem os desvios padrão σ_x e σ_y. O 3º termo é novo e designa-se por ***covariância de x e y***:

$$\boxed{\sigma_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})} \qquad (7.3)$$

- σ_q vem então:

$$\sigma_q^2 = \left(\frac{\partial q}{\partial x}\right)^2 \sigma_x^2 + \left(\frac{\partial q}{\partial y}\right)^2 \sigma_y^2 + 2\frac{\partial q}{\partial x}\frac{\partial q}{\partial y}\sigma_{xy} \qquad (7.4)$$

138

$$\sigma_q^2 = \left(\frac{\partial q}{\partial x}\right)^2 \sigma_x^2 + \left(\frac{\partial q}{\partial y}\right)^2 \sigma_y^2 + 2\frac{\partial q}{\partial x}\frac{\partial q}{\partial y}\sigma_{xy} \qquad (7.4)$$

Esta equação dá o desvio padrão σ_q, quer as medidas de x e y sejam independentes ou normalmente distribuídas, quer NÃO sejam.

- Se as medidas de x e y forem independentes podemos ver facilmente que, depois de muitas medidas, a covariância de x e y deve aproximar-se de zero: qualquer que seja o valor de y_i, é tão provável que a quantidade $x_i - \overline{x}$ seja positiva como negativa. Assim, depois de muitas medidas, os termos positivos e negativos em (7.3) devem compensar-se; no limite de um nº infinito de medidas, o factor 1/N assegura que σ_{xy} é nulo. (Depois de um nº finito de medidas, σ_{xy} não é exactamente zero, mas deve ser pequeno se os erros em x e y forem realmente independentes e aleatórios.)

- Com σ_{xy} nulo, a equação de σ_q vem:

$$\sigma_q^2 = \left(\frac{\partial q}{\partial x}\right)^2 \sigma_x^2 + \left(\frac{\partial q}{\partial y}\right)^2 \sigma_y^2 \qquad (7.5)$$

que é o resultado habitual proveniente da propagação de erros.

- Quando a covariância σ_{xy} não é nula, mesmo no limite de um nº infinito de medidas, dizemos que os erros em x e em y estão correlacionados.

139

Coeficiente de Correlação Linear

- A noção de covariância ajudar-nos-á a responder a uma questão que ficou do capítulo anterior: como podemos avaliar se uma série de pares de medidas (x_1, y_1), ..., (x_N, y_N) de duas variáveis, se adequa à hipótese de x e y estarem linearmente relacionados?

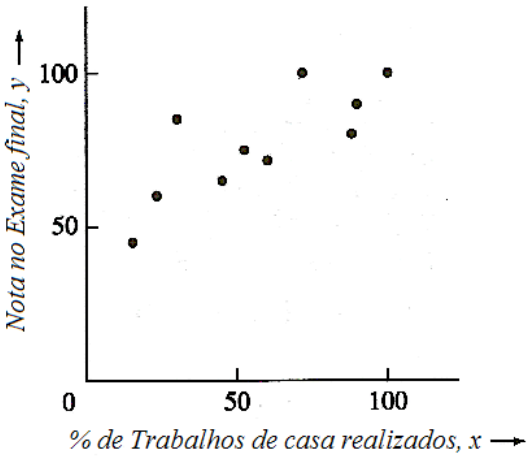
- Suponhamos esses N pares de medidas e admitamos a hipótese de existir uma relação linear entre as grandezas: y = A + Bx.

- Usando o método dos mínimos quadrados, podemos determinar os valores de A e B correspondentes à melhor recta de ajuste aos pontos experimentais.

- Se tivermos uma estimativa razoável dos erros das incertezas nas medidas e se os pontos ficarem razoavelmente perto da melhor recta, tendo em conta as incertezas experimentais, podemos concluir que as medidas apoiam a nossa hipótese da relação linear entre x e y.

- Contudo, em muitas experiências é difícil fazer uma estimativa correcta das incertezas, tornando-se mais difícil decidir sobre o tipo de relação entre as variáveis x e y. Nestes casos, são os próprios dados que têm que ser usados para se decidir a justeza ou não da hipótese inicial.

140



De facto, as variáveis x e y podem estar relacionadas por equações muito mais complexas do que uma equação linear.

Mas vamos restringir as nossas considerações apenas à hipótese de uma relação linear entre as grandezas x e y.

- A avaliação de até que ponto uma série de pares de pontos (x_1, y_1), ..., (x_N, y_N) apoia a hipótese de uma relação linear entre x e y, é feita através do *coeficiente de correlação linear*:

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \qquad (7.6)$$

- Substituindo pelo expressão matemática de cada parâmetro, obtemos:

141

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \qquad (7.7)$$

- r é um número entre -1 e +1. Se r estiver perto de ±1, os pontos estão perto de uma linha recta; se r está perto de zero, os pontos não estão correlacionados e têm pouca ou nenhuma tendência para serem ajustados por uma linha recta.

- Para provarmos estas afirmações, comecemos por observar que a covariância σ_{xy} satisfaz a desigualdade de Schwarz. De facto, prova-se que:

$$|\sigma_{xy}| \le \sigma_x \sigma_y \qquad (7.8)$$

Este facto implica imediatamente que $|r| \le 1$, ou seja, que:

$$-1 \le r \le 1$$

- Em seguida, suponhamos que os pares de pontos (x_i, y_i) estão todos exactamente sobre a recta y = A + Bx. Neste caso, $y_i = A + Bx_i$ para todos os ii e, portanto,

$$\bar{y} = A + B\bar{x}$$

142

• Subtraindo as duas últimas equações, vem:

$$y_i - \bar{y} = B(x_i - \bar{x})$$

para cada i. Inserindo este resultado na eq. 7.7, vem:

$$r = \frac{B\sum_i (x_i - \bar{x})^2}{\sqrt{\sum_i (x_i - \bar{x})^2 B^2 \sum_i (x_i - \bar{x})^2}} = \frac{B}{|B|} = \pm 1$$

Ou seja, se os pontos (x_i, y_i) estiverem perfeitamente sobre a recta, r dá ±1 e o sinal é determinado pelo declive da linha, o sinal de B.

• Na prática não esperamos que r seja exactamente ±1 mas sim que esteja perto de ±1 se a relação entre x e y for linear.

• Suponhamos agora que não há relação linear entre x e y. Então, qualquer que seja o valor de y_i, é tão provável que cada x_i fique acima como abaixo de $\bar{x}$.
Assim, os termos da soma $\sum_i (x_i - \bar{x})(y_i - \bar{y})$
no numerador de r, tanto podem ser positivos como negativos. O termo do denominador, contudo, é sempre positivo. No limite em que o nº de medidas N se aproxima de infinito, o coeficiente de correlação r será zero.

143

• Com um nº finito de medidas não esperamos que r seja exactamente zero mas esperamos que seja pequeno se as duas variáveis não estiverem relacionadas pela equação de uma recta.

• Se duas variáveis, x e y, forem tais que, no limite de um nº infinito de medidas, a covariância é zero (e, portanto, r = 0), dizemos que as variáveis ***não estão correlacionadas***. Se, depois de um nº finito de medidas r é pequeno, a hipótese de x e y não estarem correlacionados tem consistência.

Estudante i	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Trabalho de casa x_i	**90**	**60**	**45**	**100**	**15**	**23**	**52**	**30**	**71**	**88**
Exame y_i	**90**	**71**	**65**	**100**	**45**	**60**	**75**	**85**	**100**	**88**

A correlação encontrada é de r = 0.8. Isso permitirá ao professor dizer aos alunos do ano seguinte que é importante fazerem os trabalhos de casa, uma vez que os estudo demonstrou que os resultados no exame estão correlacionados com a % de trabalhos de casa realizados.

144