

# REGRESSÃO E CORRELAÇÃO

- Estudo da relação entre 2 ou mais variáveis. Em regressão uma das var. é *dependente* e a outra *independente*.

- Ex.  $Y$ : quantidade de tecido adiposo intra-abdominal (factor de risco) e  $X$ : medida em cm à volta do peito. Queremos usar o valor de  $X$  para prever  $Y$ .

- **Ex: A velocidade de uma reacção química varia com a temperatura:**

$$\log(\text{velocidade}) = A - \frac{B}{T} \quad (\text{Eq. de Arrhenius})$$

**Há portanto uma relação linear entre as var.  $Y = \log(\text{velocidade})$  e  $X = \frac{1}{T}$**

- **Velocidade de reacção é var. *dependente* e temperatura *independente*. É o aumento da temp. que causa aumento na vel. e não o contrário.**
- **Em correlação não há esta distinção entre variável dependente e independente: ambas estão ao mesmo nível.**

■  
**Ex: estudo da associação entre largura e comprimento da cabeça de seres humanos; é claro que nenhuma destas var. está dep. da outra. ■ No final procura-se normalmente calcular um coeficiente de correlação. ■**

- **Há muitos pontos em comum entre Regressão e Correlação. ■**
- **Em regressão uma das variáveis é fixa e a outra aleatória. Em correlação não; ambas são aleatórias. É possível usar regressão num contexto de correlação mas o contrário não.**

# CORRELAÇÃO

- Aqui ambas as var.  $X$  e  $Y$  são aleatórias. ■
- Seja  $f(x, y)$  a função densidade de probabilidade conjunta do vetor aleatório  $(X, Y)$ . Sejam  $f_X(x)$  e  $f_Y(y)$  as distribuições marginais e  $F(x, y)$  a função de distribuição. ■
- A correlação permite estudar a associação entre variáveis quantitativas ou ordinais. ■
- Exemplo: mediram-se e pesaram-se 20 alunos, obtendo os dados:

**altura (m): 1.74,1.83,1.68,1.89,1.72,1.72,1.75,1.78,  
1.83,1.71,1.77,1.64, 1.70,1.69,1.67,1.71,1.78,1.84,  
1.65,1.75**

**peso (kg): 68,80,62,89,68,70,71,70,78,65,72,66,65,  
67,61,73,71,72,61,63 ■**

● **Em R:**

**➤ altura=c(1.74,1.83,1.68,1.89,1.72,1.72,1.75,1.78,  
1.83,1.71,1.77,1.64, 1.70,1.69,1.67,1.71,1.78,1.84,  
1.65,1.75)■**

**> peso=c(68,80,62,89,68,70,71,70,78,65,72,66,65,  
67,61,73,71,72,61,63)**

**➤ plot(altura,peso) ■**

- **O diagrama de dispersão mostra que existe uma associação entre as duas variáveis segundo um padrão rectilíneo.**

# COEFICIENTE DE CORRELAÇÃO LINEAR DE PEARSON

- Podemos agora estudar a variabilidade conjunta, ou covariabilidade. ■
- Se tivermos um vetor aleatório  $(X, Y)$ ,  
$$\text{Cov}(X, Y) = E((X - E(X))(Y - E(Y))) = \int \int (x - \mu_x)(y - \mu_y) f(x, y) dx dy \text{ (caso contínuo)} \quad \blacksquare$$
- Este valor pode ser estimado a partir de uma amostra  $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ :  
$$\hat{\text{Cov}}(X, Y) = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) \quad \blacksquare$$
- Calcule a cov. no exemplo anterior, sabendo que

a altura,  $X$ , e o peso,  $Y$ , dos 20 alunos forneceu:  
 $\bar{x} = 1.74m$ ,  $s_x^2 = 0.0045m^2$ ,  $\bar{y} = 69.6Kg$ ,  $s_y^2 = 47.31Kg^2$  e  $\sum x_i y_i = 2429.585$ . ■

- O valor positivo indica uma associação crescente entre as duas variáveis. ■
- Se não houver associação entre as var., por exemplo QI e altura, a cov. é perto de 0. ■
- A covariância mede se há uma associação linear; nada nos diz sobre outro tipo de associações. ■
- **INCONVENIENTE** da covariância: depende das unidades de medida. ■ No exemplo anterior deu



**0.395 m.kg; se tivéssemos usado polegadas para a altura e libras para o peso dava 34.192 p.l. ■**

- **Precisamos portanto de uma medida imperturbável a transformações lineares dos dados. ■**
- **Basta para isso tomar sempre as observações padronizadas:  $(\frac{X_i - \bar{X}}{s_X}, \frac{Y_i - \bar{Y}}{s_Y})$ . ■ Se calcularmos agora a covariância obtemos: ■**
- **Coeficiente de correlação amostral ou coeficiente de correlação linear de Pearson:■**

$$r = \frac{\text{cov}(X, Y)}{S_X \cdot S_Y} =$$

$$\frac{\sum x_i y_i - \frac{1}{n} \sum x_i \sum y_i}{\sqrt{[\sum x_i^2 - \frac{1}{n} (\sum x_i)^2] [\sum y_i^2 - \frac{1}{n} (\sum y_i)^2]}} = \frac{SS_{xy}}{\sqrt{SS_{xx} \cdot SS_{yy}}}$$

- Prova-se que  $-1 \leq r_{xy} \leq 1$ . ■
- Uma correl. alta indica apenas associação estatística e não causalidade. ■ Para valores absolutos próximos de 1 a primeira coisa a fazer é ver se não se trata de uma associação espúria. Uma correlação forte entre  $X$  e  $Y$  pode significar:

■

1.  $X$  causa  $Y$ ; ■
2.  $Y$  causa  $X$ ; ■

3. um terceiro factor, directa ou indirectamente, causa  $X$  e  $Y$ ; ■
4. um acontecimento improvável ocorreu. ■
5. A correlação não faz sentido, por exemplo, porque foi obtida em indivíduos diferentes.



- Ex. Existe uma correlação forte entre cancro do pulmão e o consumo de tabaco e já se provou que de facto há uma relação de causa-efeito. ■
- Ex. Existe uma correlação forte entre doença do coração e ver televisão. No entanto não se pode concluir que ver televisão causa a doença.

**A correlação deve-se ao facto de ambas terem crescido nos últimos anos, embora por razões diferentes.**

# INFERÊNCIA SOBRE $\rho$ , COEFICIENTE DE CORRELAÇÃO DA POPULAÇÃO

- Agora temos de assumir que  $(X, Y)$  é um vector aleatório com dist. binormal, que tem 5 parâmetros:

$$(X, Y) \sim N_2(\mu, \Sigma);$$

$$\mu = (\mu_x, \mu_y) \text{ e } \Sigma = \begin{bmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix}$$

- ■
$$f(x, y) = \frac{1}{2\pi |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}d^2((x, y), \mu)}$$

- $$d^2((x, y), \mu) = ((x, y) - \mu)^T \Sigma^{-1} ((x, y) - \mu).$$

**( $d$  dist. de Mahalanobis) ■**

- **Seja agora**

$$H_0 : \rho = 0$$

$$H_1 : \rho \neq 0$$

-

- Sob  $H_0$  pode-se demonstrar que

$$t = r \sqrt{\frac{n-2}{1-r^2}} \sim t_{n-2}$$

- A  $100(1 - \alpha)\%$  rejeitamos  $H_0$  se  $|t| \geq t_{n-2, 1-\frac{\alpha}{2}}$
- Suponhamos agora que, para  $\rho_0 \neq 0$ , queremos testar

$$H_0 : \rho = \rho_0$$

$$H_1 : \rho \neq \rho_0$$

- Quando  $\rho \neq 0$  a distribuição amostral de  $r$  pode ser obtida mas é bastante complicada.

- **Transformação de Fisher:**

**Fisher (1921) demonstrou que a variável**

$$Z_r = \frac{1}{2} \log \frac{1+r}{1-r}$$

**segue uma dist. aproximadamente normal:**

$$Z_r \dot{\sim} N\left(Z_\rho, \frac{1}{n-3}\right)$$

■

- **Exercicio: suponha que numa amostra de tamanho 26 se obteve  $r = 0.74$ . Teste a hipótese de que o coef. de correlação na pop. é 0.80. ■**



- Se  $n < 25$  esta aproximação não é recomendada. Hotelling em 1953 sugeriu um outro procedimento que dá bons resultados para  $n > 10$ : ■
- Transformação de Hotelling: ■ Seja

$$Z_r^* = Z_r - \frac{3Z_r + r}{4n};$$

então, ■

$$Z^* = \frac{Z_r^* - Z_\rho^*}{1/\sqrt{n-1}} \sim N(0, 1)$$

## CORRELAÇÃO ORDINAL

- Spearman (1904) desenvolveu um coef. de correlação ordinal que usa os ranks (postos, ordens) em vez dos valores exactos. Por um lado por vezes os nossos dados já são desta forma; por outro, usar as ordens tem algumas vantagens...

- Na realidade o coef. de Spearman coincide com o coef. de Pearson dos ranks:

$$r_S = \frac{\sum_{i=1}^n R_i Q_i - \frac{1}{n} \sum_{i=1}^n R_i \sum_{i=1}^n Q_i}{\sqrt{\left[ \sum_{i=1}^n R_i^2 - \frac{1}{n} \left( \sum_{i=1}^n R_i \right)^2 \right] \left[ \sum_{i=1}^n Q_i^2 - \frac{1}{n} \left( \sum_{i=1}^n Q_i \right)^2 \right]}}$$

- Quando não há empates fica

$$r_S = 1 - \frac{6 \sum_{i=1}^n (R_i - Q_i)^2}{n(n^2 - 1)}$$



- Se forem poucos os empates esta fórmula pode ser aplicada na mesma usando ranks médios. Os resultados diferem muito pouco; só se houver mesmo muitos empates (ver livro do Wayne Daniel, Biostatistics) ■
- Teorema: No caso de independência, se  $n > 30$ ,

$$\sqrt{n-1} r_S \dot{\sim} N(0, 1)$$

- Podemos usar este resultado para testar a ind. entre duas variáveis. Para  $n \leq 30$  usamos a tabela para  $r_S$ .
- No caso de empates procede-se como de costume e o resultado assintótico continua válido.
- Exemplo: Calcule o coef. de correlação de Spearman para os dados seguintes que medem os valores de excreção (mg/l) de coproporfirina urinária de 12 mulheres grávidas, e veja se há independência entre os valores diurnos ( $X$ ) e nocturnos ( $Y$ ).

X	32.31	34.97	26.51	33.16	36.15	30.58	32.19	37.44	52.09	26.23	27.32	44.14
Y	36.22	28.18	38.91	51.25	35.64	49.34	32.19	47.17	34.98	50.32	42.66	49.54

## CORRELAÇÃO PESADA

- O coef. de Spearman (1904) é da forma:
- $r_S = A + B \sum_{i=1}^n (R_i - Q_i)^2$ ; é uma função afim da distância entre os dois vetores de ranks.
- $A$  e  $B$  são tais que  $r_S$  toma os valores 1 no caso de  $R_i = Q_i$  e -1 no caso de  $Q_i$  ser o inverso de  $R_i$ .
- É óbvio que  $r_S$  só tem em conta a diferença entre os ranks.

- **Todavia, em algumas situações, os primeiros ranks, por exemplo, são mais importantes. ■**
- **Considere por exemplo o conjunto de dados seguinte que revela as preferências, obtidas num inquérito, de homens e mulheres, por um conjunto de 10 cores (Azul, Castanho, Verde, Laranja, Roxo, Vermelho, Branco, Amarelo, Preto, Cinzento), desde 1 (mais preferida) até 10 (menos preferida):**

	Az	Ca	Vd	L	R	Vm	B	Am	P	Ci
M	1	7.5	3	6	2	4	9.5	7.5	5	9.5
H	1	7.5	2	5	10	4	7.5	9	3	6

- **Pergunta: que coeficiente usar para medir a correlação entre as preferências? ■**

- Em 2005 introduziu-se um novo coeficiente de correlação, pesada, que dá mais peso aos primeiros ranks (Pinto Costa & al, ANZJS): ■



$$r_W = 1 - \frac{6 \sum_{i=1}^n (R_i - Q_i)^2 (2n + 2 - R_i - Q_i)}{n^4 + n^3 - n^2 - n},$$



- A distribuição estatística deste coeficiente, que foi estudada em 2005 e 2006, é assintoticamente normal. ■


- Em 2011 publicaram outro coef. de corr. pesada,

$$r_{W2} = 1 - \frac{90 \sum_{i=1}^n (R_i - Q_i)^2 (2n + 2 - R_i - Q_i)^2}{n(n-1)(n+1)(2n+1)(8n+11)},$$



- Na realidade  $r_{W2}$  é o coeficiente de correlação de Pearson dos ranks pesados:  $R'_i = R_i (2n + 2 - R_i)$  e  $Q'_i = Q_i (2n + 2 - Q_i)$  ■
- Assim, a aplicação de  $r_{W2}$  é muito simples; basta primeiro fazer uma transformação aos dados e depois usar Pearson.



- 
- Os coeficientes de correlação pesada têm nos últimos tempos sido objeto de aplicação em várias áreas, incluindo bioinformática e finanças. ■
  - Calcule o valor dos três coeficientes de correlação para os dados das cores.