Estatística Aplicada a Recursos Hídricos

Docente: Rachid Muleia

(rachid.muleia@uem.mz)

Mestrado em Gestão de Recursos Hídricos - DGEO/UEM

Tema: Covariância e Correlação

Ano lectivo: 2023

Relação entre variáveis

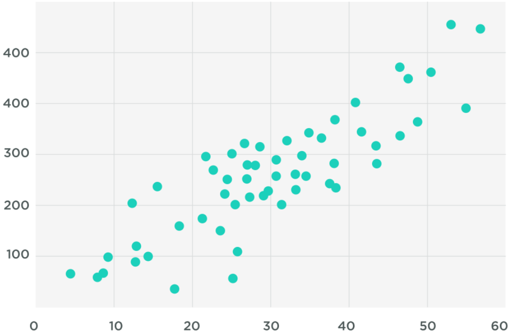
- Até agora, examinamos apenas maneiras de caracterizar a distribuição de uma única variável e testar hipóteses sobre a população com base em uma amostra.
- Por vezes, pode-se estar interssado em análise duas variávies em simultâneo
- Pode-se, por exemplo, estar interessado em :
  - ➜ Até que ponto a alteração na pressão sanguínea de um paciente está relacionada/associada ao nível de dosagem de um medicamento que eles receberam
  - ➜ Até que ponto o número de casos de malária está associado ao aumento da precipitação
  - ➜ Pode-se estar interessado em análisar o aumento da temparatura e as vendas de sorvetes
- Como medir/ estudar o (grau) de relação entre duas variáveis?

Medidas de associação

- As medidas de associação referem-se a uma ampla variedade de coeficiente que medem o grau de associação da relação entre variáveis
- As medidas de associação podem ser descritas de diversas formas, dependendo do tipo de análise
- Podem ser agrupadas em dois grupos: midadsa para variáveis quantitativas e medidas para variáveis qualitativas

Diagrama de disperão

- Uma das formas de estudar a relação entre duas variáveis é através do diagrama de dispersão
- Seja $(x_1, y_1), \quad (x_2, y_2), \ldots, (x_n, y_n)$ um conjunto de dados bivariados
- O diagrama de dispersão seria a representação dos pares $(x_i, y_i), \quad i = 1, \ldots, n$



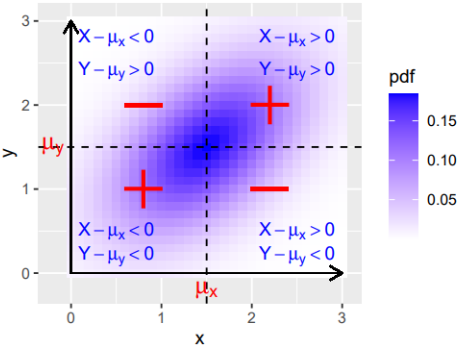
Covariância

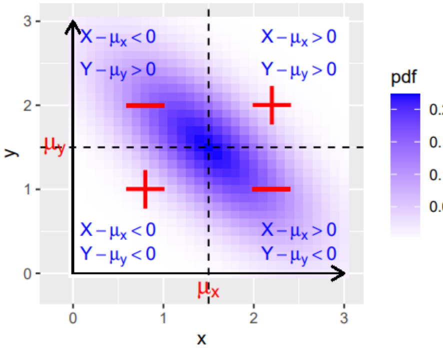
- Covariância entre duas variáveis x e y: medida de variação conjunta entre duas variáveis em relação as suas médias

$$cov(x, y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}), \quad -\infty < cov(x, y) < +\infty$$

- A covariância mostra como duas variáveis se relacionam, isto é a direcção da relação
  - ➜ $cov(x, y) > 0 \equiv$ relação positiva entre x e y, isto é, quando x aumenta, y tende a aumentar
  - ➜ $cov(x, y) < 0 \equiv$ relação negativa entre x e y, isto é, quando x aumenta, y tende a diminuir

Covariância





Com a covariância pode-se apenas avaliar a direcção da relação (se as variáveis tendem a se mover em conjunto ou mostram uma relação inversa). No entanto, não indica a força da relação, nem a dependência entre as variáveis.

Correlação = Coeficiente de correlação

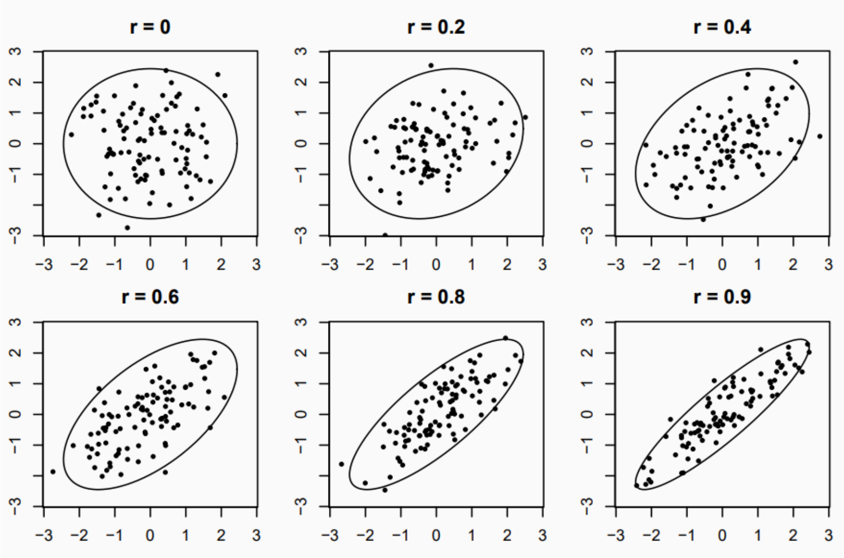
- O coeficiente correlaçao), r, também conhecido por coeficiente de correlação de pearson é uma medida númerica que mede o grau de associação linear entre duas variáveis

- O coeficiente de correlação é uma medida padronizada da covariância

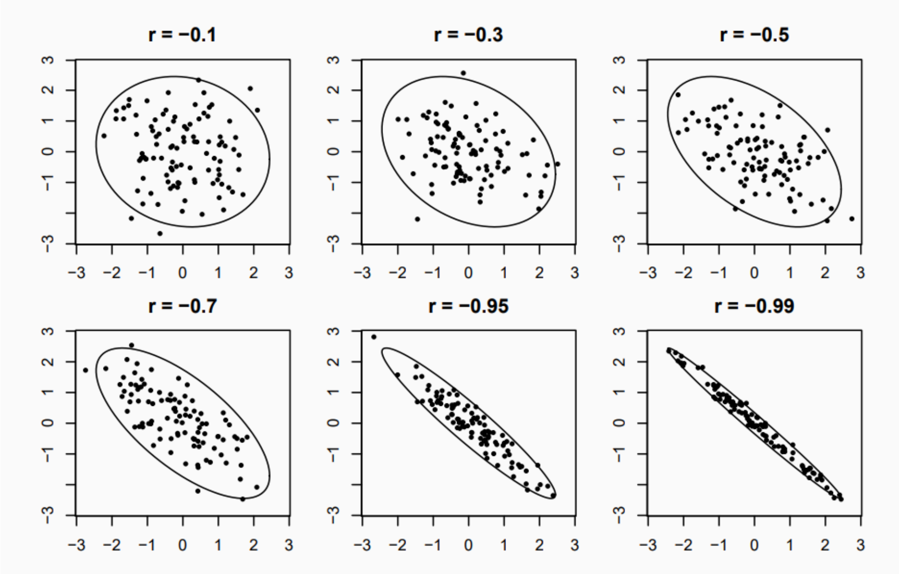
$$r = \frac{1}{n-1} \sum_{i=1}^{n} \underbrace{\left(\frac{x-\bar{x}}{s_x}\right)}_{z \text{ padrão para } x_i} \underbrace{\left(\frac{y-\bar{y}}{s_y}\right)}_{z \text{ padrão para } y_i}$$

- r varia entre -1 e $+1$; o grau de associação a medida que nos afastamos do 0 para -1 ou $+1$
  - ➜ $r > 0$: associação postiva
  - ➜ $r < 0$: associação negativa
  - ➜ $r \approx 0$: associação fraca
  - ➜ $r = -1$ ou $r = +1$, somente quando todos os pontos de dados no o gráfico de dispersão estão exatamente ao longo de uma linha recta

Correlação positiva

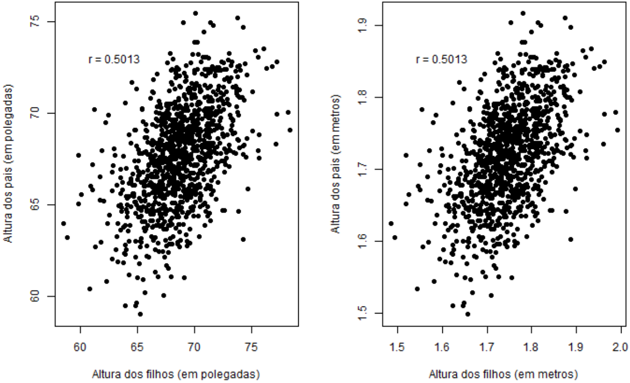


Correlação negativa



O coeficiente de correlação é admensional

- Ao alterar as unidades de medida das variáveis, não há alteração do coeficiente de correlação, visto que estamos a trabalhar com variáveis padronizadas
  - ➜ Ex: Altura dos filhos adultos vs Altura dos pais



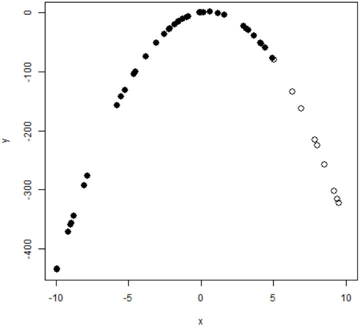
O coeficiente de correlação não distingue x e y

- Às vezes, usamos a variável X para prever a variável Y. Neste caso, X é chamado de variável explicativa e Y de variável resposta. O coeficiente de correlação r não distingue as varávies.
- Correlação de x e y $\equiv$ Correlação de y e x

$$Corr(x,y) = \frac{1}{n-1}\sum_{i=1}^{n} \underbrace{\left(\frac{x_i - \bar{x}}{s_x}\right)}_{z \text{ padrão para } x_i} \underbrace{\left(\frac{y_i - \bar{y}}{s_y}\right)}_{z \text{ padrão para } y_i} \equiv Corr(y,x) = \frac{1}{n-1}\sum_{i=1}^{n} \underbrace{\left(\frac{y_i - \bar{y}}{s_y}\right)}_{z \text{ padrão para } y_i} \underbrace{\left(\frac{x_i - \bar{x}}{s_x}\right)}_{z \text{ padrão para } x_i}$$

O coeficiente de correlação descreve uma relação linear entre duas variáveis

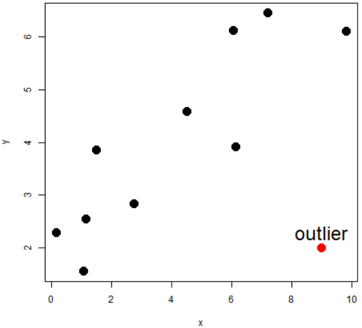
O diagrama de dispersão abaixo descreve uma relaçao perfeita entre x e y, porém não linear. Todos os pontos estão sobre a curva da função quatrática $y = 1 - 4(x - 0.5)^2$



- r para todos os pontos pretos é 0.843
- r para os pontos brancos é -0.995
- r para todos os pontos (pretos + brancos) é 0.333

O coeficiente de correlação é BASTANTE sensível a valores extremos

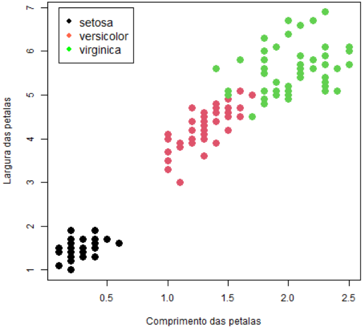
O coeficiente de correlação, por vezes, pode ser bastante influenciado por valores atípicos



$$r = \begin{cases} 0.585 & \text{na pesença do outlier} \\ 0.876 & \text{na ausência do outlier} \end{cases}$$

Outliers que podem alterar notavelmente a forma de associações quando removidos são chamados de pontos influentes

Coeficiente de correlação na presença de agrupamentos . . .



- Pode-se notar que alguns dos clusters apresentam uma correlação fraca
- Para a espécie setosa a correlação é 0.332
- Na espécie versicolor a correlação é 0.787
- Para a espécie virginica a correlaçao é 0.322
  0.963

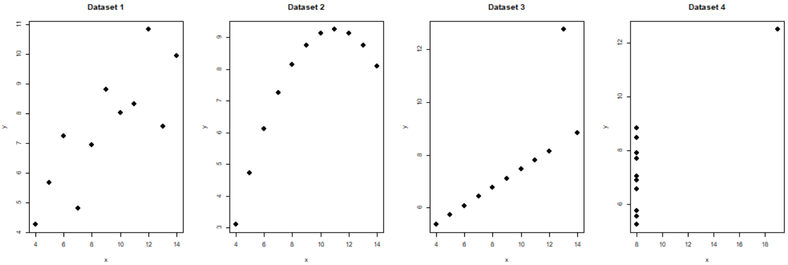
- Observe que, em conjunto, os dados apresenta uma correlação bastante forte 0.962
- Este fenómeno, onde a associação muda em função do agrupamento, designa-se de paradoxo de Simpons

Sempre olhe para o diagrama de dispersão (1)

Todos os conjuntos de dados tem a mesma média e desvio-padrão $\Longrightarrow$ o r é igual

	Dataset		Dataset 2		Dataset 2		Dataset 4	
	x	y	x	y	x	y	x	y
	10	8.04	10	9.14	10	7.46	8	6.58
	8	6.96	8	8.14	8	6.77	8	5.76
	13	7.58	13	8.75	13	12.76	8	7.71
	9	8.81	9	8.77	9	7.11	8	8.84
	11	8.33	11	9.26	11	7.81	8	8.47
	14	9.96	14	8.1	14	8.84	8	7.04
	6	7.24	6	6.13	6	6.08	8	5.25
	4	4.26	4	3.1	4	5.36	19	12.5
	12	10.84	12	9.13	12	8.15	8	5.56
	7	4.82	7	7.26	7	6.42	8	7.91
	5	5.68	5	4.74	5	5.73	8	6.89
Média	9	7.5	9	7.5	9	7.5	9	7.5
DP	3.16	1.94	3.16	1.94	3.16	1.94	3.16	1.94

Sempre olhe para o diagrama de dispersão (1)



- No conjunto de dados 2, temos uma relação perfeita. Porém, $r < 1$, pois r apenas mede o grau de associação linear

- No conjunto de dados 3, r devia ser igual à 1 no lugar de 0.82 se não tivéssemos um valor atípico

O coeficiente de correlação, por vezes, pode ser enganoso na presença de outliers, clusters múltiplos ou associação não linear

Coeficiente de correlação de Spearman, ρ

- Vimos que o r pode ser bastante afectado por valores extremos. Como alternativa, pode-se usanr o coeficiente de correlação de Spearman, ρ
- O ρ é uma medida mais abrangente, pois pode ser usada para variáveis quantitativas e de natureza ordinal
- ρ mede o grau de associação de uma relação monótona, enquanto que r mede o grau de associação de uma relação linear
- O ceoficiente de correlação de Spearman é o coeficiente de correlação de Pearson calculado sobre os postos

Como achar os postos?

X	Y	$Posto_X$	$Posto_Y$
56	66	9	4
75	70	3	2
45	40	10	10
71	60	4	7
61	65	6.5	5
64	56	5	9
58	59	8	8
80	77	1	1
76	67	2	3
61	63	6.5	6

- Observe que temos duas observações com o mesmo posto (empate)
- Em caso de empate, o posto é a média artimética $(6 + 7)/2 = 6.5$

Coeficiente de correlação de Spearman/Correlação de postos

- O coeficiente de correlação de Spearman pode ser calculado usando a formula:

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}, \quad \text{onde } d_i \text{ é a diferença entre os postos para o } i-\text{ésimo par}$$

- Alernativamente, pode-se calcular o coeficiente de correlação de Spearman, calculando o coeficiente de correlação de pearson usando os postos:

$$\rho = \frac{1}{n-1}\sum\left(\frac{r_x - \bar{r}_x}{s_{r_x}}\right)\left(\frac{r_y - \bar{r}_y}{s_{r_x}}\right)$$

Coeficiente de correlação de Spearman/Correlação de postos

- O coeficiente de correlação de Spearman pode ser calculado usando a formula:

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}, \quad \text{onde } d_i \text{ é a diferença entre os postos para o } i-\text{ésimo par}$$

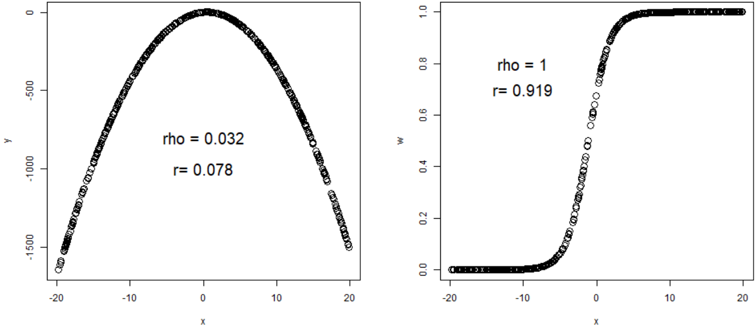
- Alernativamente, pode-se calcular o coeficiente de correlação de Spearman, calculando o coeficiente de correlação de pearson usando os postos:

$$\rho = \frac{1}{n-1}\sum\left(\frac{r_x - \bar{r}_x}{s_{r_x}}\right)\left(\frac{r_y - \bar{r}_y}{s_{r_x}}\right)$$

- Tal como o coeficiente de correlação de pearson, o ρ varia entre -1 e $+1$

Correlação de Pearson vs Spearman

- r mede uma associação linear e ρ mede uma relação monótona
- Para relação não monótona, espera-se que o $\rho \approx 0$



- Quando a relação é monótona (e não linear), o coeficiente de correlação de Pearson pode sub-estimar a relação

Teste de hipótese para R (coeficiente de correlação populacional)

- Suponha que deseja testar $H_0 : R = 0$ vs $H_1 : R \neq 0$, onde R representa o coeficiente de correlação populacional

- Se assumirmos que a hipótese nula é verdadeira, a estatística do teste é dada por

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

  - Se $|t_s| \geqslant t_{1-\alpha/2,n-2}$, rejeita-se h_0
  - Se $|t_s| \leqslant t_{1-\alpha/2,n-2}$, não rejeita-se h_0

- Para a validade do teste é important que as variáveis X e Y tenham distribuição normal bivariada. Isso significa que todas as combinações lineares $aX + bY$ têm distribuição normal

Interpretação

Turkish Journal of Emergency Medicine 18 (2018) 91–93

Contents lists available at ScienceDirect

Turkish Journal of Emergency Medicine

journal homepage: www.elsevier.com/locate/tjem

Review Article

User's guide to correlation coefficients

Haldun Akoglu*

Marmara University School of Medicine, Department of Emergency Medicine, Istanbul, Turkey

A R T I C L E I N F O

Keywords:
Correlation coefficient
Interpretation
Pearson's
Spearman's
Lin's
Cramer's

A B S T R A C T

When writing a manuscript, we often use words such as perfect, strong, good or weak to name the strength of the relationship between variables. However, it is unclear where a good relationship turns into a strong one. The same strength of r is named differently by several researchers. Therefore, there is an absolute necessity to explicitly report the strength and direction of r while reporting correlation coefficients in manuscripts. This article aims to familiarize medical readers with several different correlation coefficients reported in medical manuscripts, clarify confounding aspects and summarize the naming practices for the strength of correlation coefficients.

Interpretação

Interpretation of the Pearson's and Spearman's correlation coefficients.

Correlation Coefficient		Dancey & Reidy (Psychology)	Quinnipiac University (Politics)	Chan YH (Medicine)
+1	−1	Perfect	Perfect	Perfect
+0.9	−0.9	Strong	Very Strong	Very Strong
+0.8	−0.8	Strong	Very Strong	Very Strong
+0.7	−0.7	Strong	Very Strong	Moderate
+0.6	−0.6	Moderate	Strong	Moderate
+0.5	−0.5	Moderate	Strong	Fair
+0.4	−0.4	Moderate	Strong	Fair
+0.3	−0.3	Weak	Moderate	Fair
+0.2	−0.2	Weak	Weak	Poor
+0.1	−0.1	Weak	Negligible	Poor
0	0	Zero	None	None

The naming on the 1) Left: Dancey & Reidy.,[4] 2) Middle: The Political Science Department at Quinnipiac University, 3) Right: Chan et al.[5].