III - Estatística e métodos numéricos

Somatórios

A soma dos números $x_1, x_2, \cdots, n$ pode ser representada por um **somatório** $\sum_{i=1}^{n} x_i$, isto é

$$x_1 + x_2 + \cdots + n = \sum_{i=1}^{n} x_i.$$

Exemplos:

- $1+2+3+4+5+6+7+8+9+10 = \sum_{i=1}^{10} i$
- $2+4+6+8+10+12+14+16+18+20 = \sum_{i=1}^{10} 2i$
- $4+5+6+7+8+9+10 = \sum_{i=4}^{10} i$
- $1+4+9+16+25+36 = \sum_{i=1}^{6} i^2$
- $1+3+5+7+9+11+13+15 = \sum_{i=1}^{8}(2i-1) = \sum_{i=0}^{7}(2i+1)$

Dados

Chama-se **coleção de dados** a um conjunto de observações que podem ser recolhidas de forma diferente.

Exemplo:
1) Taxa de casamento em Portugal de 1980 a 1989 (número de casamentos durante o ano referente à população por 1000 habitantes)

1980	181	1982	1983	1984	1985	1986	1987	1988	1989
7.4	7.8	7.5	7.6	7.1	6.9	7.0	7.2	7.2	7.4

2) O lançamento de uma moeda ao ar e o registo do lado que fica voltado para cima. Numa sequência de 20 lançamento otém-se

$$F, C, F, C, C, C, F, F, F, C, C, F, F, C, F, C, F, C, C, F$$

Dados retirados de "IntroduÃ§Ã£o Ã estatÃstica", B.Murteira, C.S. Silva, J.A. Silva, C. Pimenta

Medida de localização do centro da amostra - Média

Para o estudo de uma **amostra de dados quantitativos** $x_1, x_2, \cdots, x_n$ a **média**

$$\mathbf{\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i}$$

é uma medida de localização do centro da amostra, sendo uma boa indicação quando a amostra é aproximadamente simétrica.

Exemplo:
Numa escola pesaram-se 15 crianças de 7 anos, obtendo-se os seguintes valores para o peso (em Kg):

27, 32, 26, 28, 30, 30, 33, 29, 41, 27, 31, 31,29, 31, 28.

Qual é o peso médio das crianças?

$\overline{x} = \frac{1}{17}(25 + 27 + 32 + 26 + 28 + 30 + 30 + 33 + 29 + 41 + 27 + 31 + 31,29 + 31 + 28 + 42) = \frac{520}{17} = 30,58.$

O peso médio das crianças é 30,58 Kg.

Medida de localização do centro da amostra - Média

Da definiçõ de média $\overline{x}$ temos que :

$$\sum_{i=1}^{n}(x_i - \overline{x}) = 0.$$

isto é se somarmos os desvios de todos os dados relativamente à média, o resultado é igual a 0.

Observação: Não faz sentido calcular a média de dados qualitativos, mesmo quando são representados por números. Por exemplo, não faz sentido calcular a média de uma amostra de pessoas na qual os homens são representados por 1 e as mulheres por 2.

Medida de localização do centro da amostra - Mediana

Outra medida de localização do centro da amostra é a **mediana**, intuitivamente o valor que divide a amostra a meio, sendo metade dos valores da amostra menores ou iguais à mediana e os restantes superiores ou iguais.

Medida de localização do centro da amostra - Mediana

Para determinar a mediana, ordenamos a amostra e

- se o número de dados é ímpar, a mediana é o elemento do meio;
- se o número de dados é par, qualquer valor entre os dois elementos do meio podia servir, mas toma-se a semi-soma desses valores.

No exemplo anterior, ordenando os pesos obtemos:

$$25, 26, 27, 27, 28, 28, 29, 29, \textcircled{30}, 30, 31, 31, 31, 32, 33, 41, 42$$

A mediana é igual a 30.

Medida de localização do centro da amostra - Mediana

Determinação da mediana numa amostra com dados qualitativos hierarquizáveis:

Os resultados de uma turma de 10 alunos de matemática elementar 105 foram os seguintes: 4 D's; 3 C's; 2 B's e 1 A.

D D D D C C C B B A.

A mediana é C.

Quartis

A mediana é um caso particular dos **quatis**. O quartil de ordem α, $0 < \alpha < 1$, q_α é o valor da coleção que tem αn observações inferiores e $(1-\alpha)n$ observações superiores para uma amostra de tamanho n. A mediana é o quartil de ordem 0.5.

- O **primeiro quartil** é o quartil de ordem 0.25;
- O **segundo quartil** é a mediana.
- O **terceiro quartil** é o quartil de ordem 0.75.

Na amostra anterior com os pesos das crianças

$$25, 26, 27, 27, 28, 28, 29, 29, 30, 30, 31, 31, 31, 32, 33, 41, 42$$

o primeiro quartil é igual a 28 (na posição $\frac{n+3}{4}$) e o terceiro quartil é igual a 31 (na posição $\frac{3n+1}{4}$).

Medidas de localização do centro da amostra

Observações:

- Não se pode dizer qual das medidas de localização do centro da amostra, média ou mediana, é preferível, depende do contexto.
- Quando a amostra é aproximadamente simétrica, elas são aproximadamente iguais.
- A mediana não é tão sensível como a média aos dados que são muito maiores ou muito menores que os restantes dados (outliners).

A moda

A **moda** é o valor que surge com mais frequência na amostra. A moda não é uma medida de localização e pode ser determinada para qualquer tipo de amostra.

Nos exemplos anteriores

- na amostra dos pesos das crianças

24, 26, 27, 27, 28, 28, 29, 29, 30, 30, 31, 31, 31, 32, 33, 41, 42

a moda é 31;

- nas notas de uma turma da M105

D D D D C C C B B A.

a moda é D.
(ordenamos as notas tendo em conta que D é a pior e A a melhor, B é mellhor que C).

Medidas de dispersão - Variância

Duas amostras com a mesma média e a mesma mediana podem ser muito diferentes. Por exemplo, ambas as amostras seguintes tem média e mediana igual a 10:

A: 8, 8, 9, 9, 9, 9, 10, 10, 10, 10, 10, 11, 11, 11, 11, 12, 12
B: 5, 6, 7, 8, 8, 9, 9, 10, 10, 10, 11, 11, 12, 12, 13, 14, 15

A **variância**

$$\mathbf{s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

dá-nos informação de quão distantes os dados estão da média.

Por exemplo, na amostra A, a variância é igual a 1,5 e na amostra B é igual a 6,625.

Nota: Alguns autores definem variância de maneira ligeiramente diferente, dividindo a soma por n em lugar de por n-1

Medidas de dispersão - Desvio Padrão

O **desvio padrão**

$$\mathbf{s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}}$$

é igual à raiz quadrada da variância e é a medida de dispersão usualmente usada.

No exemplo anterior, a amostra A apresenta o desvio padrão aproximadamente igual a 1,22 e a amostra B a 2,63.

O desvio padrão é um número positivo e será tanto maior quanto maior for a variabilidade dos dados. Em particular, se $s = 0$, os dados da amostra são todos iguais.

Distribuição aproximadamente normal

Diz-se que uma amostra tem uma **distribuição aproximadamente normal** se

- aproximadamente 68% dos dados estão no intervalo $[\overline{x} - s, \overline{x} + s]$,
- aproximadamente 95% dos dados estão no intervalo $[\overline{x} - 2s, \overline{x} + 2s]$,
- aproximadamente 100% dos dados estão no intervalo $[\overline{x} - 3s, \overline{x} + 3s]$.

Distribuição aproximadamente normal

Quando a amostra apresenta uma distribuição aproximadamente normal podemos calcular os **valores normalizados ou estandardizados** da amostra

$$z_i = \frac{x_i - \overline{x}}{s}, i = 1, 2, \cdots, n.$$

Estes valores dão informação de quanto o dado se desvia da média, contando múltiplos do desvio padrão.

Considere-se a amostra A:

8, 8, 9, 9, 9, 9, 10, 10, 10, 10, 10, 11, 11, 11, 11, 12, 12

Calcule as valores normalizados.

Distribuição aproximadamente normal - Exemplo

Um curso faz a seriação para admissão dos estudantes mediante a realização de um de dois testes: teste A ou teste B. Calculou-se a média e o desvio padrão para os resultados de ambos os testes e obteve-se o seguinte:

- teste A: média 135 e desvio padrão 12
- teste B: média 105 e desvio padrão 11

O Francisco realizou o teste A e teve 150 enquanto que o Miguel obteve 130 no teste B.

$$\text{valor normalizado da classificação do Francisco} = \frac{150-135}{12} = 1,25$$

$$\text{valor normalizado da classificação do Miguel} = \frac{130-105}{11} \simeq 2,27$$

Diagrama de dispersão

O **diagrama de dispersão** é uma representação gráfica para os pares de dados (dados bivariados) em que cada par de dados (x_i, y_i) é representado por um ponto do plano com essas coordenadas.

Idade da mulher	19	24	20	28	26	25	27	23	32	31	35	32	34	40
Idade do marido	28	29	27	26	31	24	39	33	37	34	35	42	40	41

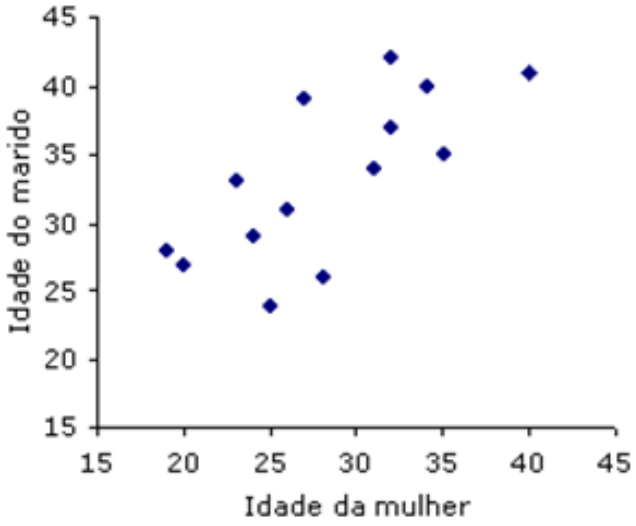


Diagrama de dispersão

Diz-se que os dados têm uma **associação (linear) positiva** se aos maiores valores de uma variável correspondem, de um modo geral, os maiores valores da outra.

Diz-se que os dados têm uma **associação (linear) negativa** se aos maiores valores de uma variável correspondem, de um modo geral, os menores valores da outra.

Os dados podem não ter qualquer associação linear.

Coeficiente de correlação

O **coeficiente de correlação** é uma medida do grau de associação linear.

Representa-se por r e calcula-se da seguinte forma:

- Calculam-se os valores normalizados de $x_1, x_2, \cdots, x_n$, isto é

$$\frac{x_i - \overline{x}}{s_x}, i = 1, 2, \cdots, n;$$

- Calculam-se os valores normalizados de $y_1, y_2, \cdots, y_n$, isto é

$$\frac{y_i - \overline{y}}{s_y}, i = 1, 2, \cdots, n;$$

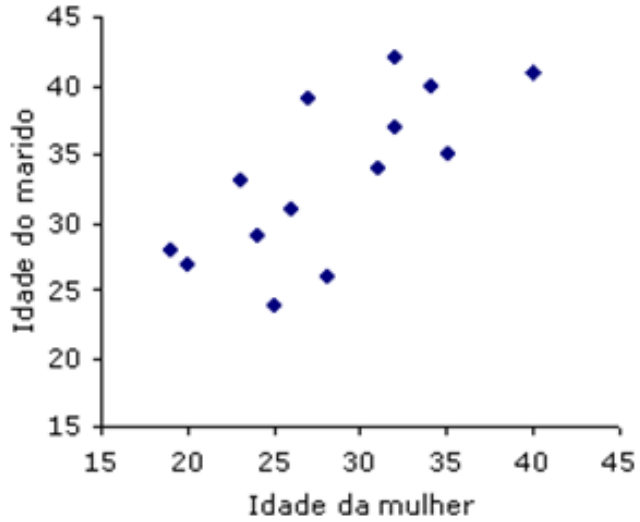
- O coeficiente de correlação é a média dos produtos dos valores normalizados, isto é

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \frac{x_i - \overline{x}}{s_x} \frac{y_i - \overline{y}}{s_y}.$$

Exemplo

A associação entre os dados é tanto maior quanto mais próximo $|r|$ for de 1.
Calcule o coefciente de correlação do exemplo

Idade da mulher	19	24	20	28	26	25	27	23	32	31	35	32	34	40
Idade do marido	28	29	27	26	31	24	39	33	37	34	35	42	40	41



Reta de regressão ou dos mínimos quadrados

Quando $|r|$ é próximo de 1, os dados seguem um padrão linear e pode ter interesse **ajustar** uma reta $y = a_0 + a_1 x$ que dê informação sobre como se refletem em y as mudanças em x.
Esta reta chama-se **reta de regressão** ou **reta dos mínimos quadrados**.

A reta chama-se reta dos mínimos quadrados porque os valores a_0 e a_1 tornam mínima a soma $\sum_{i=1}^{n}[y_i - (a_0 + a_1 x)]^2$.

A reta passa no ponto $(\overline{x}, \overline{y})$, onde $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ e $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

Mostra-se que

$$a_1 = \frac{\overline{xy} - \overline{x}\ \overline{y}}{\overline{x^2} - \overline{x}^2},$$

onde $\overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i$ e $\overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2$.
Como a reta passa em $(\overline{x}, \overline{y})$, a equação da reta é

$$y - \overline{y} = a_1(x - \overline{x}).$$

Reta de regressão ou dos mínimos quadrados

No exemplo anterior, a reta tem equação

$$y - 32,5714285714286 = 0,752285191956126(x - 28,2857142857143)$$