ESTATÍSTICA DESCRITIVA – PARÂMETROS OU VALORES TÍPICOS

Das medidas, ou estatísticas que iremos definir, para caracterizar os dados, destacam-se as medidas de localização - nomeadamente as que localizam o centro da amostra (média, moda e mediana) , e as medidas de dispersão.
Será mesmo necessário utilizar os dois tipos de medidas, para caracterizar os dados?

Dois alunos do 7º ano obtiveram as seguintes notas, no 3º período:

Pedro	4	3	3	3	3	3	4	3	4	3
João	5	2	2	3	4	3	5	3	3	3

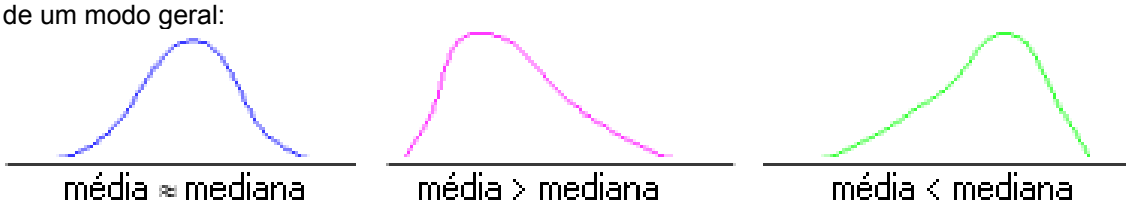
O Pedro e o João tiveram a mesma média de 3.3, mas o João não transitou de ano.
Uma característica apresentada pelas notas do João, é uma maior variabilidade do que as notas do Pedro.

A média será sempre uma medida representativa dos dados?
Ao determinar a média dos seguintes dados 12.4, 13.5, 13.6, 11.2, 15.1, 10.6, 12.4, 14.3, 113.5 obteve-se o valor = 24.1.
Embora todos os dados, menos um, estejam no intervalo [10.6, 15.1], o valor obtido para a média está "bem afastado" daquele intervalo! O que aconteceu é que a média é muito sensível a valores muito grandes ou muito pequenos. No caso do exemplo foi o valor 113.5 que inflacionou a média. Além disso temos razões para pensar que pode ter havido um erro ao digitar o valor 113.5, digitando um 1 a mais!
E se em vez de 113.5 o valor correcto fosse 13.5, qual o valor da média ?
Ao determinar a média dos seguintes dados 12.4, 13.5, 13.6, 11.2, 15.1, 10.6, 12.4, 14.3, 13.5 obteve-se o valor da média = 13.0, significativamente diferente do obtido no caso anterior!
Sendo a média uma medida tão sensível aos dados, é preciso ter cuidado com a sua utilização, pois pode dar uma imagem distorcida dos dados, que pretende representar!

Média ou Mediana?
Consideremos o seguinte exemplo:
Um aluno do 10º ano obteve as seguintes notas: 10, 10, 10, 11, 11, 11, 11, 12.
A média e a mediana da amostra anterior são respectivamente 10.75 e 11.
Admitamos que uma das notas de 10 foi substituída por uma de 18. Neste caso a mediana continuaria a ser igual a 11, enquanto que a média subiria para 11.75!

Resumindo, como a média é influenciada quer por valores muito grandes, quer por valores muito pequenos, se a distribuição dos dados:
1. For aproximadamente simétrica, a média aproxima-se da mediana
2. For enviesada para a direita (alguns valores grandes como "outliers"), a média tende a ser maior que a mediana
3. For enviesada para a esquerda (alguns valores pequenos como "outliers"), a média tende a ser inferior à mediana.

Representando as distribuições dos dados (esta observação é válida para as representações gráficas na forma de diagramas de barras ou de histograma) na forma de uma mancha, temos, de um modo geral:



Tendo-se decidido registar os pesos dos alunos de uma determinada turma prática do 10º ano, obtiveram-se os seguintes valores (em kg):

52	56	62	54	52	51	60	61	56	55	56	54	57	67	61	49

a) Determine os quantis de ordem 1/7, 1/2 e os 1º e 3º quartis.
b) Um aluno com o peso de 61 kg, pode ser considerado "normal", isto é nem demasiado magro, nem demasiado gordo?
Ordenando a amostra anterior, cuja dimensão é 16, temos

49	51	52	52	54	54	55	56	56	56	57	60	61	61	62	67

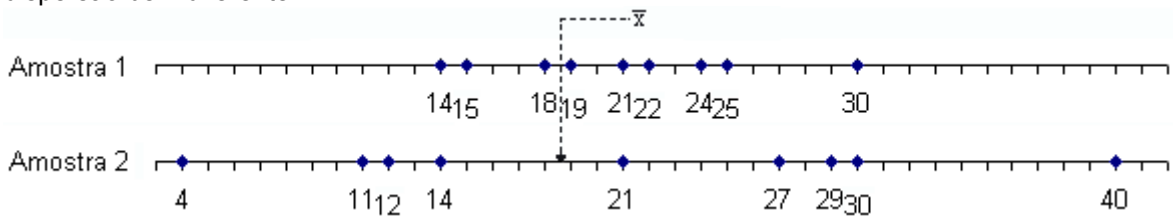
a) 16 * 1/7 = 16/7 donde [16/7] = 2 e $Q_{1/7} = x_{3:16} = 52$
16 * 1/2 = 8 donde $Q_{1/2} = [x_{8:16} + x_{9:16}]/2 = 56$
16 * 1/4 = 4 donde $Q_{1/4} = [x_{4:16} + x_{5:16}]/2 = 53$
16 * 3/4 = 12 donde $Q_{3/4} = [x_{12:16} + x_{13:16}]/2 = 60.5$

b) Um aluno com 61 kg pode ser considerado um pouco "forte", pois naquela turma só 25% dos alunos é que têm peso maior ou igual a 60.5 kg!

Medidas de dispersão…

A sua importância…

Um aspecto importante no estudo descritivo de um conjunto de dados, é o da determinação da variabilidade ou dispersão desses dados, relativamente à medida de localização do centro da amostra.
Repare-se nas duas amostras seguintes, que embora tenham a mesma média, têm uma dispersão bem diferente:



Como a medida de localização mais utilizada é a média, será relativamente a ela que se define a principal medida de dispersão - a **variância**, apresentada a seguir.

Define-se a variância, e representa-se por s^2, como sendo a medida que se obtém somando os quadrados dos desvios das observações da amostra, relativamente à sua média, e dividindo pelo número de observações da amostra menos um:

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{(n - 1)} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{(n - 1)}$$

Uma vez que a variância envolve a soma de quadrados, a unidade em que se exprime não é a mesma que a dos dados. Assim, para obter uma medida da variabilidade ou dispersão com as mesmas unidades que os dados, tomamos a raiz quadrada da variância e obtemos o desvio padrão:

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

O desvio padrão é uma medida que só pode assumir valores não negativos e quanto maior for, maior será a dispersão dos dados.
Algumas propriedades do desvio padrão, que resultam imediatamente da definição, são:
o desvio padrão é sempre não negativo e será tanto maior, quanta mais variabilidade houver entre os dados.
se s = 0, então não existe variabilidade, isto é, os dados são todos iguais.

Exemplo: Na 2ª classe de certa escola o professor deu uma tarefa constituída por um certo número de contas para os alunos resolverem. Pretendendo determinar a dispersão dos tempos de cálculo, observam-se 10 alunos durante a realização da tarefa, tendo-se obtido os seguintes valores:

Aluno	Tempo (minutos)		
i	xi	$xi-\overline{x}$	$(xi-\overline{x})^2$
1	13	-3.9	15.21
2	15	-1.9	3.61
3	14	-2.9	8.41
4	18	1.1	1.21
5	25	8.1	65.61
6	14	-2.9	8.41
7	16	-0.9	0.81
8	17	0.1	0.01
9	20	3.1	9.61
10	17	0.1	0.01
Soma	169	0.0	112.90

Na tabela anterior juntámos duas colunas auxiliares, uma para colocar os desvios das observações em relação à média e a outra para escrever os quadrados destes desvios. A partir da coluna das observações calculámos a soma dessas observações, que nos permitiu calcular a média = 16.9. Uma vez calculada a média foi possível calcular a coluna dos desvios. Repare-se que, como seria de esperar, a soma dos desvios é igual a zero.
A soma dos quadrados dos desvios permite-nos calcular a variância,
$s^2 = 112.9/9 = 12.54$,
donde s = 3.54.
O tempo médio de realização da tarefa foi de aproximadamente 17 minutos com uma variabilidade medida pelo desvio padrão de aproximadamente 3.5 minutos

Do mesmo modo que a média, também o desvio padrão é uma medida pouco resistente, pois é influenciado por valores ou muito grandes ou muito pequenos (o que seria de esperar já que na sua definição entra a média que é não resistente). Assim, se a distribuição dos dados for bastante enviesada, não é conveniente utilizar a média como medida de localização, nem o desvio padrão como medida de variabilidade. Estas medidas só dão informação útil, respectivamente sobre a localização do centro da distribuição dos dados e sobre a variabilidade, se as distribuições dos dados forem aproximadamente simétricas.

Propriedades para dados com distribuição aproximadamente normal:
Uma propriedade que se verifica se os dados se distribuem de forma aproximadamente normal, ou seja, quando o histograma apresenta uma forma característica com uma classe média predominante e as outras classes se distribuem à volta desta de forma aproximadamente simétrica e com frequências a decrescer à medida que se afastam da classe média, é a seguinte:
Aproximadamente 68% dos dados estão no intervalo]x-s,x+s[
Aproximadamente 68% dos dados estão no intervalo]x-2s,x+2s[
Aproximadamente 95% dos dados estão no intervalo]x-3s,x+3s[
Aproximadamente 100% dos dados estão no intervalo]x-4s,x+4s[

A informação que o desvio padrão dá sobre a variabilidade deve ser entendida como a variabilidade que é apresentada relativamente a um ponto de referência - a média, e não propriamente a variabilidade dos dados, uns relativamente aos outros.

A medida anterior tem a grande desvantagem de ser muito sensível à existência, na amostra, de uma observação muito grande ou muito pequena. Assim, define-se uma outra medida, a amplitude inter-quartil, que é, em certa medida, uma solução de compromisso, pois não é afectada, de um modo geral, pela existência de um número pequeno de observações demasiado grandes ou demasiado pequenas. Esta medida é definida como sendo a diferença entre os 1º e 3º quartis
Do modo como se define a amplitude inter-quartil, concluímos que 50% dos elementos do meio da amostra, estão contidos num intervalo com aquela amplitude.
Esta medida é não negativa e será tanto maior quanto maior for a variabilidade nos dados.
Atenção: Mas, ao contrário do que acontece com o desvio padrão, uma amplitude inter-quartil nula, não significa necessariamente, que os dados não apresentem variabilidade.

Amplitude inter-quartil ou desvio padrão?
Do mesmo modo que a questão foi posta relativamente às duas medidas de localização mais utilizadas - média e mediana, também aqui se pode por o problema de comparar aquelas duas medidas de dispersão.

1. A amplitude inter-quartil é mais robusta, relativamente à presença de "outliers", do que o desvio padrão, que é mais sensível aos dados.
2. Para uma distribuição dos dados aproximadamente normal, verifica-se a seguinte relação amplitude inter-quartil 1.3 x desvio padrão

3. Se a distribuição é enviesada, já não se pode estabelecer uma relação análoga à anterior, mas pode acontecer que o desvio padrão seja muito superior à amplitude inter-quartil, sobretudo se se verificar a existência de "outliers".