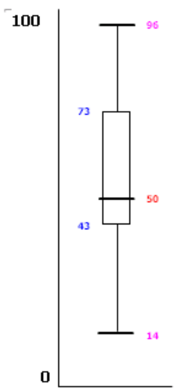
N.º 20 - Diagrama de Extremos e Quartis

Por: Maria Eugénia Graça Martins
Departamento de Estatística e Investigação Operacional da FCUL
memartins@fc.ul.pt

O diagrama de extremos e quartis é a representação gráfica que melhor se posiciona na relação custo/benefício! Efectivamente, à custa unicamente de cinco números, três dos quais calculados a partir dos dados e os outros dois resultantes de uma simples observação dos dados, obtém-se uma representação muito esclarecedora sobre a forma como os dados se distribuem, nomeadamente quanto à:

- maior ou menor concentração;
- simetria;
- existência de valores "aberrantes".



É também muito útil ainda para comparar vários conjuntos de dados.

Quais são então esses cinco números, a partir dos quais se constrói o diagrama de extremos e quartis? Quatro desses números são, como o nome indica, os extremos – **mínimo** e **máximo**, e os quartis – **1º quartil** (Q_1 ou $Q_{0,25}$) e **3º quartil** (Q_3 ou $Q_{0,75}$). O outro número é a **mediana** (m).

De seguida serão apresentados alguns exemplos de representações de conjuntos de dados em diagramas de extremos e quartis, realçando-se as principais características desta representação gráfica.

Esta ActivALEA é acompanhada de uma aplicação interactiva que possibilita a representação de um conjunto de dados num diagrama de extremos e quartis.

O diagrama de extremos e quartis pode-se construir horizontal ou verticalmente. Vamos começar por fazer uma representação horizontal e depois apresentamos uma representação vertical, que é a forma apresentada pelo Excel.

Exemplo 1 – Os dados seguintes representam as pontuações obtidas por 48 estudantes, num determinado teste. Apresente-os num diagrama de extremos e quartis.

75	98	42	75	84	87	65	59	63	86	78	37
99	66	90	79	80	89	68	57	95	55	79	88
76	60	77	49	92	83	71	78	53	81	77	58
93	85	70	62	80	74	69	90	62	84	64	73

Para obter os cinco números a partir dos quais se constrói o diagrama de extremos e quartis, deve-se começar por ordenar a amostra. Uma representação em caule-e-folhas pode ser útil para ordenar a amostra e calcular a mediana e os quartis.[1]

3	7
4	2 9
5	3 5 7 8 9
6	0 2 2 **3** **4** 5 6 8 9
7	0 1 2 4 5 5 **6** **7** 7 8 8 9 9
8	0 0 1 3 4 **4** **5** 6 7 8 9
9	0 0 2 3 5 8 9

3|7 significa 37 pontos

Como temos 48 dados, a mediana é a semi-soma dos elementos que se encontram nas posições 24 e 25 (inteiros que rodeiam o valor $\frac{48+1}{2}$=24,5), que assinalámos a azul, ou seja

Mediana = $\frac{76+77}{2}$=76,5 pontos.

Os quartis são as medianas de cada uma das partes em que ficou dividido o conjunto dos dados pela mediana, cada uma com 24 elementos.

O 1º quartil é a semi-soma dos elementos que se encontram nas posições 12 e 13 (inteiros que rodeiam $\frac{24+1}{2}$=12,5) a contar do início do caule-e-folhas, assinalados a verde, ou seja

1º quartil = $\frac{63+64}{2}$=63,5 pontos.

O 3º quartil é a semi-soma dos elementos que se encontram nas posições 12 e 13 a contar do fim do caule-e-folhas, assinalados a laranja, ou seja **3º quartil** = $\frac{84+85}{2}$ = 84,5 pontos.

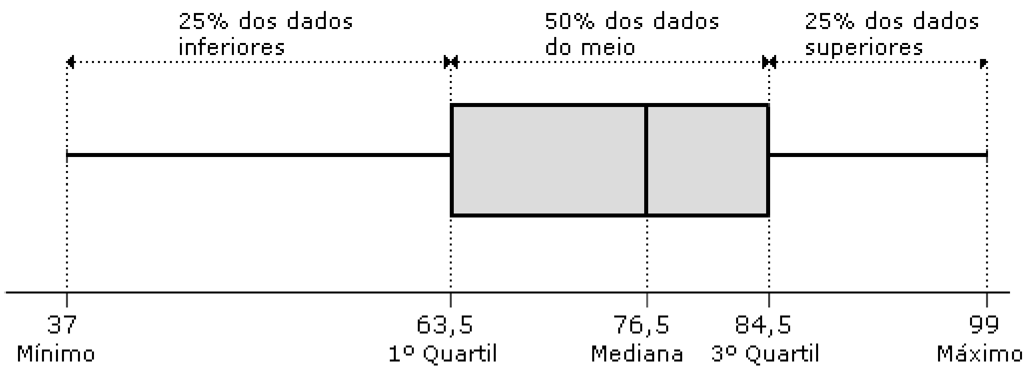
[1] ver ActivAlea n.º 19 – Diagrama de Caule-e-folhas http://www.alea.pt/html/statofic/html/dossier/html/activalea19.html

Quanto ao mínimo e ao máximo, são, respectivamente, 37 e 99 pontos.

Para construir o diagrama de extremos e quartis, desenha-se um rectângulo com comprimento igual à amplitude entre os quartis e com altura qualquer (a altura do rectângulo não tem qualquer significado).

Dentro do rectângulo desenha-se um segmento de recta que assinala a posição da mediana.

Dos lados do rectângulo determinados pelo 1º quartil e pelo 3º quartil saem dois segmentos de recta, até ao mínimo e até ao máximo, respectivamente:



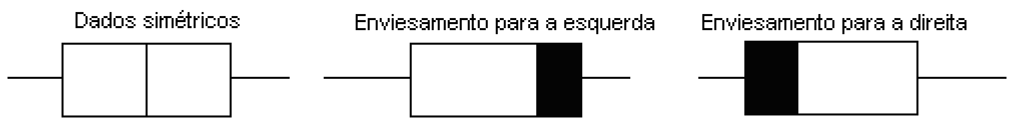
Da representação anterior ressalta imediatamente que:

- 25% das notas são menores ou iguais a 63 pontos (as notas estão dadas em números inteiros);
- 25% das notas são superiores ou iguais a 85 pontos;
- 50% das notas estão compreendidas entre 64 e 84 pontos;
- Existe algum enviesamento do lado esquerdo, isto é, os dados estão mais dispersos, ou seja, menos concentrados na parte inferior do que na parte superior; também na parte central dos dados existe algum enviesamento para a esquerda.

Existem fundamentalmente duas características do diagrama de extremos e quartis que nos dão ideia da simetria ou enviesamento dos dados e que são:

- Distância entre a linha indicadora da mediana e os lados do rectângulo;
- Comprimento das linhas que saem dos lados do rectângulo.

Apresentamos de seguida 3 exemplos de diagramas de extremos e quartis correspondentes a tipos diferentes de distribuição dos dados.



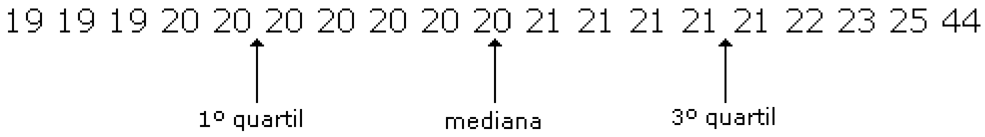
Por vezes, existem alguns conjuntos de dados cuja representação num diagrama de extremos e quartis é inesperada! Ora vejamos o seguinte exemplo.

Exemplo 2 – Recolheu-se a idade de 19 alunos que frequentam a disciplina de Estatística de um curso da Faculdade de Ciências, tendo-se obtido os seguintes resultados:

19, 20, 20, 21, 21, 20, 19, 20, 22, 23, 21, 21, 20, 25, 20, 19, 21, 44, 20

Represente os dados num diagrama de extremos e quartis e tire conclusões sobre a estrutura dos dados.

Ordenando o conjunto de dados anteriores, tem-se

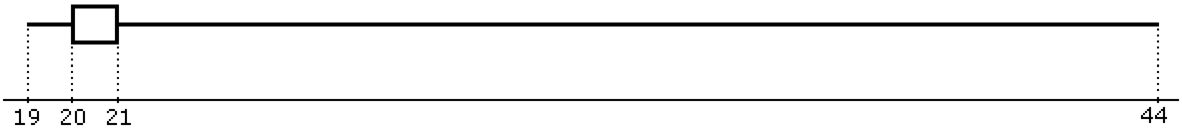


Mínimo = 19 Máximo = 44

1º quartil = $\frac{20+20}{2}$ =20

Mediana = 20

3º quartil = $\frac{21+21}{2}$ =21



A representação anterior mostra que estamos perante uma situação de extremo enviesamento, em que o diagrama de extremos e quartis é pouco claro! Efectivamente, não aparece a linha que indica a mediana, o que significa que esta se confunde com o 1º ou o 3º quartil. Esta é uma situação em que o diagrama de extremos e quartis apresenta alguma ambiguidade. Se nos tivessem apresentado esta representação sem estar acompanhada dos dados, devido ao grande enviesamento dos 25% de dados superiores, seria natural *presumir* que também existiria enviesamento para a direita nos 50% dos dados centrais e que, portanto, a mediana se confundia com o 1º quartil. Como temos os dados, podemos confirmar que, na verdade, o 1º quartil se confunde com a mediana.

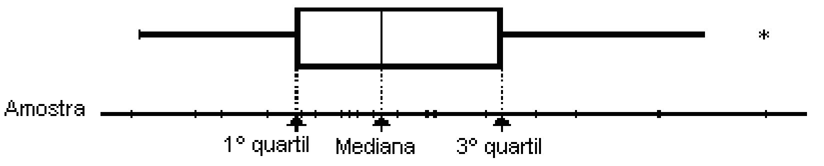
Por outro lado, o valor 44 que aparece nos dados anteriores é um pouco "aberrante", quando comparado com os restantes, pois é muito maior que todos os outros dados.

Existe uma versão do diagrama de extremos e quartis que permite visualizar os valores que se consideram "aberrantes", por saírem do contexto dos restantes, e a que é usual dar o nome de *outliers*. Essa versão designada em inglês por *box-and-whiskers*, tem a tradução portuguesa de caixa-com-bigodes.

Caixa-com-bigodes

Considera-se uma caixa que é um rectângulo tal como foi desenhado para o diagrama de extremos e quartis.

Consideram-se seguidamente duas linhas que unem os meios dos lados dos rectângulos com o menor e maior elementos da amostra que estão dentro das **barreiras de outliers**, definidas a seguir. Os outros elementos que não estão no intervalo constituído pelas barreiras de outliers são assinalados com o símbolo *.



O QUE SÃO AS BARREIRAS DE OUTLIERS?

Define-se **barreira inferior** como sendo o valor

$$Q_{.25} - 1{,}5 \times (Q_{.75} - Q_{.25})$$

Define-se **barreira superior** como sendo o valor

$$Q_{.75} + 1{,}5 \times (Q_{.75} - Q_{.25})$$

Quando é que consideramos um valor como *outlier?*

Dizemos que um valor é *outlier* quando não está compreendido no intervalo [barreira inferior, barreira superior]. Numa representação em *caixa-com-bigodes,* os outliers assinalam-se com o símbolo "*".

No caso do exemplo anterior, obter-se-ia a seguinte representação em caixa-com-bigodes:

Barreira inferior $= 20 - 1{,}5 \times (21-20)$
$= 18{,}5$

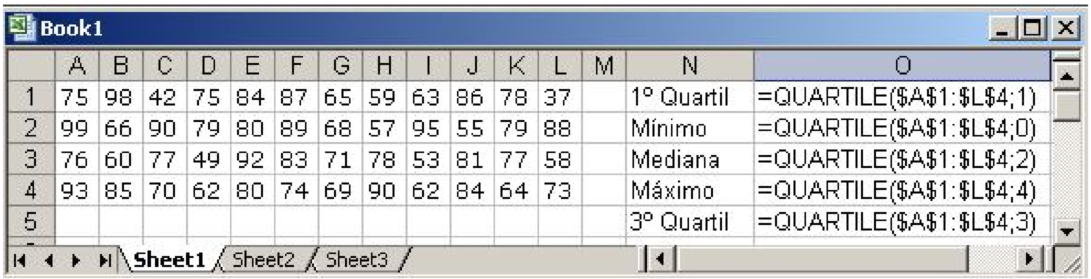
Barreira superior $= 21 + 1{,}5 \times (21-20)$
$= 22{,}5$

Da representação anterior, concluímos que existem 3 *outliers*, concretamente o 23, 25 e 44.

Construção do diagrama de extremos e quartis com o Excel

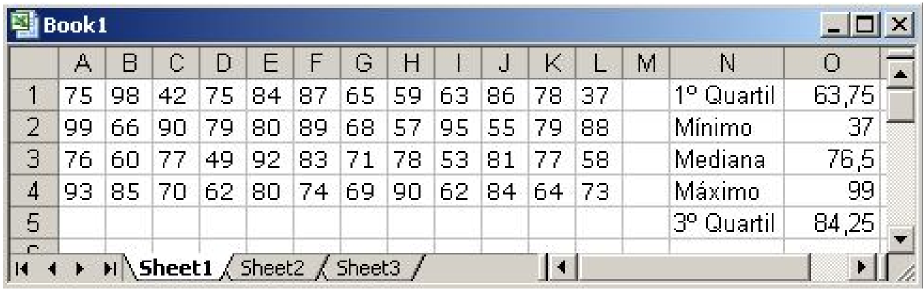
Como já temos afirmado várias vezes, embora o Excel não seja um software de Estatística, permite a construção de algumas representações gráficas, uma das quais é o diagrama de extremos e quartis. Esta construção pode ser consultada no dossiê XIII – "Estatística Descritiva com Excel – Complementos", páginas 63-66, na área "Dossiês e recursos" do ALEA[2].

Para obter a representação em diagrama de extremos e quartis dos dados do Exemplo 1, começámos por inserir os dados num ficheiro Excel e de seguida calculámos os 5 números necessários para construir o diagrama, como se apresenta a seguir:



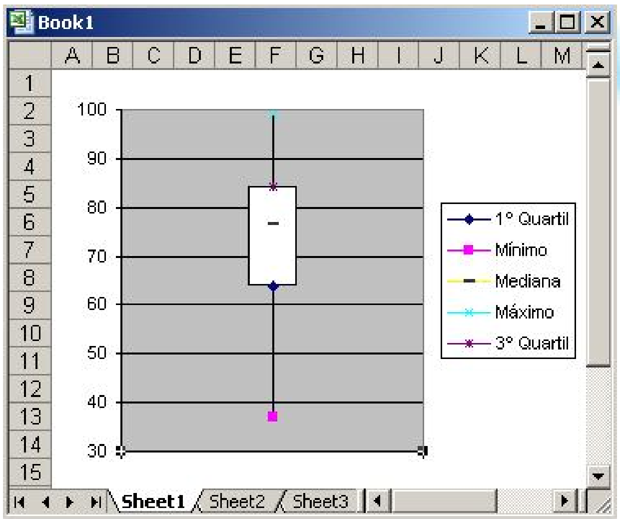
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	75	98	42	75	84	87	65	59	63	86	78	37		1º Quartil	=QUARTILE($A$1:$L$4;1)
2	99	66	90	79	80	89	68	57	95	55	79	88		Mínimo	=QUARTILE($A$1:$L$4;0)
3	76	60	77	49	92	83	71	78	53	81	77	58		Mediana	=QUARTILE($A$1:$L$4;2)
4	93	85	70	62	80	74	69	90	62	84	64	73		Máximo	=QUARTILE($A$1:$L$4;4)
5														3º Quartil	=QUARTILE($A$1:$L$4;3)

com os seguintes resultados:



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	75	98	42	75	84	87	65	59	63	86	78	37		1º Quartil	63,75
2	99	66	90	79	80	89	68	57	95	55	79	88		Mínimo	37
3	76	60	77	49	92	83	71	78	53	81	77	58		Mediana	76,5
4	93	85	70	62	80	74	69	90	62	84	64	73		Máximo	99
5														3º Quartil	84,25

Repare-se que os valores obtidos para os quartis, utilizando o Excel, não são iguais aos obtidos anteriormente, seguindo as indicações do Exemplo 1. Efectivamente, a regra utilizada pelo Excel nem sempre coincide com a regra que se utiliza para calcular os quartis, em que estes são obtidos como *a mediana de cada uma das partes em que fica dividida a amostra pela mediana* (se a dimensão da amostra for ímpar, ou seja, se a mediana for um dos elementos da amostra, considera-se como pertencente às duas partes). Esta situação não é grave, pois os valores obtidos pelas duas regras são aproximados.



[2] http://www.alea.pt/html/statofic/html/dossier/html/meio_dossier13.html