N.º 19 - DIAGRAMA DE CAULE-E-FOLHAS

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

Emília Oliveira
Escola Secundária de Tomaz Pelayo
ecmo.estp@gmail.com

De entre a grande variedade de representações que se podem utilizar para representar dados, temos o gráfico ou diagrama caule-e-folhas. É um tipo de representação que se pode considerar entre a tabela e o gráfico, uma vez que, de um modo geral, são apresentados os verdadeiros valores dos dados, mas numa representação sugestiva, que faz lembrar um histograma.

Quando comparada com o histograma, é uma representação mais simples de construir quando se trabalha com papel e lápis e tem uma vantagem imediata, que é a de facilitar a ordenação dos dados, quando não se dispõe de um computador. Por outro lado, como na maior parte das vezes preserva os dígitos dos dados, ao contrário do histograma que os agrupa, permite a reconstituição da amostra.

Nesta ActivALEA são apresentados alguns exemplos de representações dos dados num gráfico caule-e-folhas e dadas algumas orientações a ter em conta na construção de uma representação deste tipo.
Para representar um conjunto de dados num diagrama de caule-e-folhas pode ser utilizada a aplicação interactiva[1] que acompanha esta ActivALEA.

[1] Aplicação interactiva desenvolvida por Manuel Ramos, Escola Secundária de Tomaz Pelayo.

OS CAULES E AS FOLHAS

A base da construção de uma representação em caule-e-folhas está na separação de cada dado em duas partes: o "caule" e a "folha". Tendo em consideração a ordem de grandeza dos dados a representar, escolhe-se o(s) dígito(s) dominante(s) que se coloca(m) ao longo de um eixo vertical, do lado esquerdo. Estes dígitos constituem os caules. Para cada dado toma-se o digito que se segue imediatamente ao(s) digito(s) dominante(s) e coloca-se do lado direito do eixo, em frente ao respectivo caule. Estes dígitos são as folhas. As folhas são registadas à medida que vamos percorrendo o conjunto dos dados. No entanto, na representação final, ordenam-se por ordem crescente as folhas penduradas em cada caule.

Vamos exemplificar a construção do caule-e-folhas com o seguinte conjunto de dados.

Exemplo 1 – Os dados seguintes representam as pontuações obtidas por 48 estudantes, num determinado teste. Apresente-os num gráfico de caule-e-folhas.

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

Como o menor e o maior dos dados anteriores são, respectivamente, 37 e 99, vamos considerar para caules o dígito das dezenas:

3
4
5
6
7
8
9

Depois de traçar uma linha vertical do lado direito dos caules, começa-se a colocar as folhas. Após representarmos os primeiros 4 dados, o diagrama apresenta o seguinte aspecto:

Caule	Folha
3	
4	2
5	
6	
7	5 5
8	
9	8

e quando todos os dados estiverem representados, teremos:

Caule	Folha
3	7
4	2 9
5	9 7 5 3 8
6	5 3 6 8 0 2 9 2 4
7	5 5 8 9 9 6 7 1 8 7 0 4 3
8	4 7 6 0 9 8 3 1 5 0 4
9	8 9 0 5 2 3 0

Na apresentação final do gráfico caule-e-folhas, ordenamos as folhas por ordem crescente e para não haver ambiguidade na leitura dos números que representam os nossos dados, indicamos a forma de os ler:

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

Esta indicação na forma de ler os dados é bastante importante, pois podemos ter a mesma representação, para dados de outro tipo como indicamos a seguir:

Admitamos que as alturas, em cm, de um conjunto de plantas, duas semanas depois de se ter lançado a semente à terra, eram:

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

Nota: estes dados são fictícios e obtiveram-se dos dados do exemplo 1, dividindo por 10 cada uma das pontuações.

A representação destes dados num gráfico de caule e folhas é precisamente igual à dos dados das pontuações, mas com a indicação de como se deve fazer a leitura não haverá ambiguidade:

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 3,7 cm

Como aumentar o número de caules?

Na representação anterior considerámos 7 caules e o intervalo entre caules sucessivos é de 10 unidades. É como se tivéssemos considerado as classes [30, 40[, [40, 50[, [50, 60[, [60, 70[, [70, 80[, [80, 90[e [90, 100[, para agrupar os dados.

Suponhamos que em vez de considerar estas classes, de amplitude 10, estávamos interessados em considerar classes de amplitude 5, a saber [30, 35[, [35, 40[, [40, 45[, [45, 50[, [50, 55[, [55, 60[, [60, 65[, [65, 70[, [70, 75[, [75, 80[, [80, 85[, [85, 90[, [90, 95[e [95, 100[.

Então a representação anterior teria o seguinte aspecto:

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

Qualquer que seja a representação considerada, qualquer caule tem sempre a possibilidade de ter penduradas o mesmo número de folhas. No exemplo anterior, cada caule foi desdobrado em dois subcaules. No primeiro subcaule aparecem penduradas as folhas 0, 1, 2, 3 e 4, enquanto que no segundo subcaule aparecem penduradas as folhas 5, 6, 7, 8 e 9. Para distinguir os dois subcaules, é costume colocar no primeiro subcaule um asterisco e no segundo um ponto:

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

No caso de a representação ainda apresentar muitas folhas em cada caule, existe a possibilidade de considerar classes de amplitude 2, fazendo cada caule dividido em 5 subcaules e cabendo a cada subcaule 2 folhas. Numa representação com os caules 0 e 1, estes seriam indicados da seguinte forma:

0*	folhas 0 e 1
t	folhas “**t**wo” e “**t**hree”
f	folhas “**f**our” e “**f**ive”
s	folhas “**s**ix” e “**s**even”
0.	folhas 8 e 9
1*	…
t	
f	
s	
1.	

Repare-se na analogia com a construção do histograma, em que também nos preocupamos em ter classes de igual amplitude. A esta amplitude de classe é usual chamar, na representação em caule-e-folhas, ***comprimento de linha***.

Das considerações anteriores concluímos que, ao contrário do histograma em que não existem restrições para a amplitude de classe, no caso do caule-e-folhas o comprimento de linha tem que ser 10, 5 ou 2 vezes uma potência de 10. Por exemplo, vários valores possíveis para o comprimento de linha são:

$2x10^0$ ou seja 2;

10^0 ou seja 1;

$5x10^{-1}$ ou seja 0,5;

$2x10^1$ ou seja 20;

Etc.

Exemplo 2 - No quadro seguinte apresenta-se o número de concelhos de cada um dos distritos de Portugal Continental e das Regiões Autónomas de Açores e Madeira (INE, 2009). Represente os dados num diagrama de caule-e-folhas[2].

Região	Nº concelhos	Região	Nº concelhos
Aveiro	19	Lisboa	16
Beja	14	Portalegre	15
Bragança	12	Porto	18
Braga	14	Santarém	21
Cast.Branco	11	Setúbal	13
Coimbra	17	Viana Cast.	10
Évora	14	Vila Real	14
Faro	16	Viseu	24
Guarda	14	Açores	19
Leiria	16	Madeira	11

[2] No Excel, embora não exista uma representação imediata para a construção de um caule-e-folhas, é possível elaborar este tipo de gráfico. No Dossier XIII do ALEA - Estatística Descritiva com Excel – Complementos, disponível em **www.alea.pt/index.php?option=com_content&view=article&id=319&Itemid=1722**, é apresentado um processo desenvolvido por Neville Hunt (Hunt, 2001) para a construção de um caule-e-folhas no Excel.

Se considerarmos para caules os algarismos das dezenas, só temos 2 caules diferentes. Utilizando unicamente esses caules, a representação ficaria muito pesada, com muitas folhas em cada caule e muito pouco elucidativa quanto à estrutura dos dados. Assim, vamos considerar para comprimento de linha o 2 e obtemos a seguinte representação:

1*	0 1 1
t	2 3
f	4 4 4 4 4 5
s	6 6 6 7
1.	8 9 9
2*	1
t	
f	4

1|0 significa 10 concelhos

Na construção de um gráfico de caule-e-folhas, nem sempre é imediata a escolha dos dígitos dominantes. Vejamos, por exemplo, a situação seguinte.

Exemplo 3 – Uma empresa imobiliária pretendendo estudar os preços dos apartamentos na região de Lisboa, recolheu essa informação sobre 40 apartamentos, tendo obtido os seguintes preços (em milhares de euros):

190	138	179	162	157	357	209	138	151	255	135	235
210	281	121	189	255	170	290	126	290	186	183	122
160	147	299	182	188	149	147	208	183	185	154	236
149	185	204	208								

Organize os dados num gráfico caule-e-folhas.

Os preços das casas variam entre 121 mil euros e 357 mil euros. Se tomarmos como dígito dominante o das centenas, ficamos só com 3 caules. Se tomarmos os dígitos das centenas e das dezenas ficamos com 24 caules, o que é demasiado. Então optamos por considerar para caules os algarismos das centenas, divididos em dois subcaules, formando classes de comprimento 50:

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

A apresentação final, com as folhas ordenadas, tem o seguinte aspecto:

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

1|2 significa 120 mil euros

Nesta representação considerámos como folhas unicamente o algarismo das dezenas e desprezámos o algarismo das unidades. Esta é uma das situações em que não conseguimos recuperar exactamente os dados iniciais, mas apenas uma aproximação. Por exemplo, podemos concluir que o valor mínimo e máximo registados para os preços dos apartamentos, andam à volta de 120 mil euros e 350 mil euros, respectivamente.

Há autores que consideram como folhas os dois algarismos. Não pensamos que seja uma boa opção, sobretudo se tivermos muitos dados, já que o que se ganha com a possibilidade de recuperar os dados iniciais, pode perder-se na visualização da estrutura subjacente aos dados, que é, afinal, o nosso objectivo:

1*	21 22 22 35 38 38 47 47 49 49
1.	51 54 57 60 62 70 79 82 83 83 85 85 86 88 89 90
2*	04 08 08 09 10 35 36
2.	55 55 81 90 90 99
3*	
3.	57

1|21 significa 121 mil euros

Os caule–e–folhas paralelos

A representação em caule-e-folhas é muito sugestiva para comparar duas amostras, como se apresenta no exemplo seguinte:

Exemplo 4 – A seguir apresentam-se os tempos de sono (em horas), de dois jovens, medidos durante 30 noites seguidas. Compare-os.

Pedro			David		
8,7	9,3	8,7	7,1	9,5	7,1
9,4	5,3	7,4	8,3	7,1	7,4
6,6	7,3	6,3	7,1	7,5	7,4
6,0	6,7	5,9	7,9	7,9	7,8
6,9	5,8	10,0	7,5	6,4	6,2
9,9	4,7	6,5	6,2	6,2	8,6
6,3	5,6	8,6	8,2	7,5	8,4
8,9	5,9	7,7	8,7	7,7	6,6
10,1	9,4	9,0	8,5	7,6	8,1
9,6	7,6	7,9	7,6	8,8	7,1

Para representar os caule-e-folhas paralelos, determinamos os caules (comuns) a partir da amostra de maior amplitude, ou seja, neste caso, dos dados correspondentes ao Pedro.

Pedro					David								
			7	4.									
			3	5*									
9	9	8	6	5.									
	3	3	0	6*	2	2	2	4					
9	7	6	5	6.	6								
		4	3	7*	1	1	1	1	1	4	4		
	9	7	6	7.	5	5	5	6	6	7	8	9	9
				8*	1	2	3	4					
9	7	7	6	8.	5	6	7	8					
4	4	3	0	9*									
		9	6	9.	5								
		1	0	10*									

6|2 significa 6,2 horas

Os dados relativamente ao Pedro encontram-se para o lado esquerdo, enquanto que os referentes ao David estão para o lado direito. A representação anterior permite realçar a maior dispersão do sono do Pedro, enquanto que o David é mais regular, com uma duração de sono de um modo geral entre as 7 e as 8 horas.

Como diminuir o número de caules?

Nos exemplos anteriores os caules foram divididos em subcaules, para aumentar o número de caules ou linhas na representação do caule-e-folhas. Uma situação menos evidente é aquela em que o número de caules é demasiado para o número de dados a representar, como no exemplo seguinte:

Exemplo 5 – Registaram-se os tempos, em segundos, que 28 alunos conseguiam estar sem respirar:

112	41	26	35	98	56	87	38	125	64	78	82
80	48	92	106	105	115	79	83	131	87	94	111
45	123	57	29								

Considerando como caules os algarismos das dezenas e das centenas, temos

2	6 9
3	5 8
4	1 5 8
5	6 7
6	4
7	8 9
8	0 2 3 7 7
9	2 4 8
10	5 6
11	1 2 5
12	3 5
13	1

2|6 significa 26 segundos

Os dados estão um pouco dispersos, pelo que seria conveniente considerarmos menos caules. Uma forma de resolver este problema é juntar os caules dois a dois. Em vez de termos as classes [20, 30[,[30, 40[,[40, 50[,[50, 60[, ..., teremos [20, 40[,[40, 60[, ... Para distinguir as folhas que originalmente estavam penduradas nos caules 3, 5, ..., das que estavam penduradas nos caules 2, 4,..., sublinhamos as primeiras. Esta precaução tem como único objectivo a posterior reconstituição do conjunto dos dados.

2	6 9 <u>5</u> <u>8</u>
4	1 5 8 <u>6</u> <u>7</u>
6	4 <u>8</u> <u>9</u>
8	0 2 3 7 7 <u>2</u> <u>4</u> <u>8</u>
10	5 6 <u>1</u> <u>2</u> <u>5</u>
12	3 5 <u>1</u>

2|6 significa 26 segundos

Qual o número de caules ou linhas adequado para a construção dum caule-e-folhas?[3]

A escolha do número de caules ou linhas, tal como acontece com o número de classes do histograma, depende em grande parte da experiência e da habilidade do estatístico. Os problemas que se levantam são análogos aos da construção do histograma. No entanto, dado o facto de se utilizar a notação decimal, é necessário considerar uma outra metodologia para o comprimento do intervalo correspondente a cada linha. Assim, utiliza-se normalmente o seguinte procedimento:

Considera-se para número de linhas L um valor que não exceda

$$L = \text{Parte inteira de } (10 \log_{10} n)$$

onde *n* é o número de observações.

Esta regra costuma fornecer valores de L convenientes para as dimensões das amostras usuais num tratamento estatístico. É evidente que, se n for muito grande, esta representação torna-se muito pesada e pouco maleável e tal como para o histograma, aconselha-se a não utilizar mais do que 15 classes.

Usando L como limite para o número de linhas, levanta-se o problema da determinação dos comprimentos dos intervalos correspondentes a cada linha. O processo mais simples é usar uma potência de 10 como comprimento do intervalo. Assim, dividimos R, a amplitude da amostra, por L e arredondamos por excesso (se necessário) o quociente obtido, até à potência de 10 mais próxima.

Pode acontecer que a técnica descrita anteriormente para a construção da representação de caule-e-folhas apresente demasiadas folhas por linha. Então, o processo de resolver este problema é considerar duas linhas por caule, repetindo os seus valores no caule. Neste caso o comprimento do intervalo será 5 vezes uma potência de 10.

Pode ainda acontecer que, mesmo considerando 2 linhas por caule, a representação ainda continue muito pesada, mas que se arredondássemos para a potência de 10 imediatamente abaixo do valor obtido para R/L, também ficasse muito esparsa. Então, resolve-se o problema considerando 5 linhas por caule e neste caso o comprimento do intervalo é 2 vezes uma potência de 10.

Vejamos a aplicação desta metodologia para a obtenção do número de caules convenientes para a representação dos dados do exemplo 2. Uma vez que temos 20 dados

$$L = \text{parte inteira } (10 \times \log_{10} 20)$$
$$L = \text{parte inteira } (10 \times 1{,}301)$$
$$L = 13$$

Considerando 13 o limite para o número de linhas, vejamos qual o comprimento de linha:

[3] Introduz-se este tópico a título de curiosidade, pois não se aconselha ao nível do ensino básico e secundário, a preocupação com este tipo de regras.

R = 24-10= 14 $\frac{R}{L} = \frac{14}{13} = 1{,}08$

O comprimento de linha sugerido é 2.

A utilização do caule-e-folhas para obter a mediana e os quartis

No caule-e-folhas os dados estão representados de uma forma ordenada, pelo que é possível utilizá-lo para calcular a mediana e os quartis.

Representando por n o número de dados, começa por se calcular $\frac{n+1}{2}$ para obter a posição da mediana. Se n é ímpar, então a mediana é o elemento que se encontra na posição dada pelo quociente anterior. Se n é par, então a mediana será a semi-soma dos elementos que se encontram nas posições dos dois inteiros que rodeiam o valor $\frac{n+1}{2}$.

O cálculo dos quartis resume-se a calcular a mediana de cada uma das partes em que fica dividido o conjunto dos dados pela mediana. No caso de a mediana ser um dos elementos do conjunto dos dados, situação que se verifica sempre que o número de dados é ímpar, então consideramos que a mediana pertence às duas partes. Por exemplo, se o número de dados for igual a 21, a mediana é o elemento que está na posição 11, pelo que consideramos que o elemento nesta posição pertence às duas partes em que fica dividido o conjunto dos dados. Assim, o 1.º quartil será o elemento na posição 6 a contar da parte de inicial do caule-e-folhas enquanto que o 3.º quartil será o elemento na posição 6 a contar da parte final do caule-e-folhas.

Exemplo 6 - Num determinado teste realizado a 51 estudantes, obtiveram-se as seguintes pontuações:

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 48 72 38

Faça uma representação em caule-e-folhas dos dados e determine a mediana e os quartis.

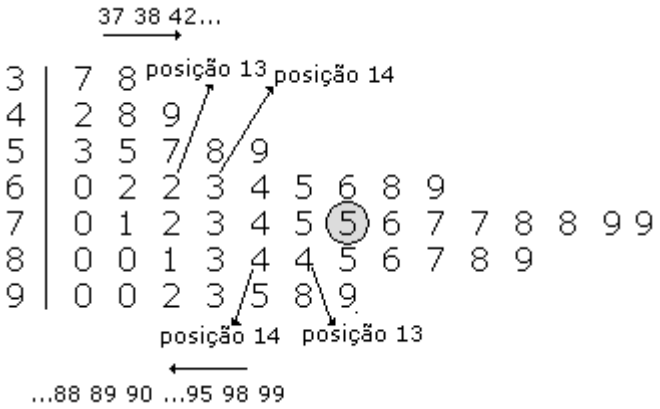
Considerando o algarismo das dezenas para caule, facilmente se obtém a seguinte representação:

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

que depois de ordenada vem:

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

A mediana encontra-se na posição 26 ($=\frac{51+1}{2}$) pelo que ficamos com dois conjuntos de dados com 26 elementos cada um (como foi dito anteriormente, tomámos a opção de considerar a mediana do total dos dados nos dois conjuntos de dados, para a seguir calcular a mediana de cada um destes conjuntos). A mediana de cada um destes conjuntos encontra-se na posição 13,5 ($=\frac{26+1}{2}$), pelo que será a semi-soma dos elementos das posições 13 e 14:



Mediana = 75 pontos
1.º quartil = 62,5 pontos
3.º quartil = 84 pontos

Do que ficou dito anteriormente sobre o gráfico caule-e-folhas, podemos resumir algumas das vantagens, relativamente a outras construções gráficas[4]:

- É, em geral muito simples de fazer, tornando-se acessível a alunos de qualquer grau (desde que ajudados na escolha dos dígitos dominantes que servem de caules);
- Dá uma informação visual sobre a forma como os dados estão distribuídos;
- Permite ordenar rapidamente a amostra;
- Facilita o cálculo da mediana e dos quartis.

[4] Graça Martins, M. E.; Loura, Luísa Canto e Castro; Mendes, Maria de Fátima, 2007 – Análise de Dados, Texto de Apoio para os Professores do 1º ciclo – Ministério da Educação, Direcção-Geral da Inovação e do Desenvolvimento Curricular.