Introdução à Probabilidade e Estatística

Manuela Oliveira

Ano letivo 2015/2016

Capítulo 1

Estatística Descritiva

1.1 Introdução

Neste capítulo começamos por rever conceitos de estatística descritiva. A **estatística descritiva** tem por objectivo descrever, resumir e representar a informação contida num conjunto de dados, através da construção de tabelas e gráficos ou através da determinação de medidas numéricas que adequadamente sintetizem os dados.

A dificuldade do Homem em interpretar grandes conjuntos de dados é aqui ultrapassada pela distribuição dos dados em **classes** e pelo no cálculo de **medidas resumo** que os descrevam de forma fiel.

A forma de analisar os dados depende, em primeira instância, da sua natureza. Os dados numéricos podem ser **discretos**, quando se referem a contagens ou números inteiros, ou **contínuos**, quando podem tomar qualquer valor dentro de um determinado intervalo de números.

Para além disso os dados estatísticos são ainda classificados de acordo com a sua escala de medição. Assim temos dados **qualitativos** e **quantitativos**. Os primeiros dizem respeito a dados cujos atributos de interesse são categorias e dividem-se em dados **nominais** e **ordinais**.

Os dados **nominais** não são na verdade dados numéricos, mas apenas etiquetas ou valores atribuídos que designam uma classe, não havendo uma relação de ordem entre as classes. Por exemplo, a situação em que os dados se referem à cor dos olhos de um conjunto de indivíduos (1=preto, 2=castanho, 3=azul, 4=verde, 5=cinzento).

Os dados **ordinais** referem-se a dados do tipo dos nominais, com a diferença que para estes se estabelece uma relação de ordem entre as classes. Por exemplo, as classificações de cada aluno num determinado teste dadas por "Mau", "Suficiente"e "Muito Bom".

Os dados **quantitativos** são aqueles em que a sua característica de interesse é intrinsecamente numérica. Dividem-se em dados com **escala intervalar** ou com **escala absoluta**, residindo a distinção no facto de estes últimos terem a si associado uma origem definida. Para decidir se determinado tipo de dados está em qual das escalas pergunte a si próprio se o dobro do valor do que está a estudar corresponde ao dobro de intensidade. Por exemplo, 20°C é duas vezes mais quente que 10°C? A resposta é não e, por isso, dados deste tipo são de escala intervalar. Agora um campo com 4 hectares é o dobro de outro com 2 hectares? Sim, por isso temos agora dados de escala absoluta. Notamos que as técnicas estatísticas não fazem distinção entre estes dois tipos de dados.

É exclusivamente sobre esta última classe de dados, os quantitativos, que vamos trabalhar.

1.2 Distribuições de frequência e representação gráfica de dados

Quando lidamos com grandes conjuntos de dados podemos obter uma boa ideia global dos mesmos se os agruparmos em classes ou intervalos disjuntos. Ao procedermos assim perdemos informação mas esta perda é largamente compensada pelo conhecimento que ganhamos acerca da forma dos dados.

Assuma que estamos a tratar com dados contínuos. No caso discreto os valores observados definem eles próprios as classes a considerar.

Para escolher o número de classes k a usar é usual aplicar-se a regra de Sturges:

$$k \approx 1 + \frac{\log n}{\log 2},$$

onde n é a dimensão do conjunto de dados.

Sabendo k e a amplitude total do conjunto de dados, L, dada por:

$$L = \text{máximo}\{\text{dados}\} - \text{mínimo}\{\text{dados}\},$$

obtém-se a amplitude de cada classe, l, como:

$$l = \frac{L}{k}.$$

Podemos então definir os limites de cada classe e contar o número de observações que caiem dentro de cada uma delas, obtendo assim as **frequências absolutas** de cada classe - f_i para a classe i, $i = 1, \ldots, k$. Este procedimento vem facilitado se ordenarmos os dados. Notamos que:

$$\sum_{i=1}^{k} f_i = n$$

O conjunto das frequências absolutas de todas as classes, eventualmente colocadas numa tabela, chama-se **distribuição de frequências**.

Para o conjunto das frequências absolutas obtêm-se as chamadas **frequências absolutas acumuladas** de cada classe, F_i, como a soma das frequências absolutas dessa classe e de todas as outras que a precedem:

$$F_i = \sum_{j=1}^{i} f_j$$

Repare que $F_k = n$. Ao conjunto das $\{F_i,\ i = 1, \ldots, k\}$ chama-se **distribuição de frequências absolutas**.

Obervamos que é usual identificar cada classe pelo seu **ponto médio**, calculado como a metade da soma dos seus extremos, e denotado aqui como PM_i para a classe i, $i = 1, \ldots, k$.

Definem-se ainda as chamadas **frequências relativas** de cada classe, aqui designadas por f_i^*, como:

$$f_i^* = \frac{f_i}{n}$$

Observe-se que estas frequências se encontram em $[0, 1]$ e que:

$$\sum_{i=1}^{k} f_i^* = 1$$

Associadas a f_i^* encontram-se as correspondentes **frequências relativas acumuladas**:

$$F_i^* = \sum_{j=1}^{i} f_j^*$$

Temos que $F_k^* = 1$. Ao conjunto das frequências relativas chama-se **distribuição de frequências relativas** e ao conjunto das frequências relativas acumuladas chama-se **distribuição de frequências relativas acumuladas**.

Nota: Se depois de seleccionadas as classes se verificar que, por existirem observações muito extremas, surgem "nas pontas"classes com apenas 1 ou 0 observações, é usual agregá-las, obtendo as classes abertas "menor que"e "maior que". Essas observações que se destacam por serem muito extremas, muito distantes das restantes, designam-se por ***outliers***.

Uma vez tendo as distribuições de frequências podemos construir vários dispositivos gráficos para as representar, já que uma imagem vale 1000 palavras... Assim podemos ter **histogramas**, **polígonos de frequência**, **polígonos de frequências acumuladas** representando graficamente a distribuição de frequência dos dados, ou ainda **diagramas de caixa-e-bigodes**, que apresentaremos mais tarde no texto.

O **histograma** é um gráfico de barras que se constrói escolhendo para abcissas os limites de cada uma das classes e para ordenadas, resultando na altura de cada uma das barras que o constitui, a frequência (absoluta ou relativa) dos dados na classe correspondente.

O **polígono de frequências** é obtido unindo os pontos de ordenada correspondente à altura de cada barra e abcissa dada pelo respectivo ponto médio da classe. Os polígonos de frequências são usualmente melhores que os histogramas para comparar a forma de duas ou mais distribuições de frequências.

O **polígono de frequências acumuladas** obtém-se unindo os pontos formados por ordenadas dadas pela altura das barras do histograma e respectivas abcissas que são um dos limites da classe que lhe corresponde - caso seja o superior fala-se de distribuição acumulada "acima de"; se for o inferior temos distribuição acumulada "abaixo de". A curva aqui resultante toma o nome de **ogiva**. É uma curva importante quando estamos interessados em determinar que percentagem dos dados está abaixo de um certo valor.

Exemplo 1.1 *Seguem-se as percentagens de gordura de manteiga fornecidas por 120 vacas Ayrshire, de 3 anos de idade, seleccionadas ao acaso de um livro de registos de gado canadiano:*

4.32	4.24	4.29	4.00	3.96	4.48	3.89	4.02	3.74	4.42
4.20	3.87	4.10	4.00	4.33	3.81	4.33	4.16	3.88	4.81
4.23	4.67	3.74	4.25	4.28	4.03	4.42	4.09	4.15	4.29
4.27	4.38	4.49	4.05	3.97	4.32	4.67	4.11	4.24	5.00
4.60	4.38	3.72	3.99	4.00	4.46	4.82	3.91	4.71	3.96
3.66	4.10	4.38	4.16	3.77	4.40	4.06	4.08	3.66	4.70
3.97	3.97	4.20	4.41	4.31	3.70	3.83	4.24	4.30	4.17
3.97	4.20	4.51	3.86	4.36	4.18	4.24	4.05	4.05	3.56
3.94	3.89	4.58	3.99	4.17	3.82	3.70	4.33	4.06	3.89
4.07	3.58	3.93	4.20	3.89	4.60	4.38	4.14	4.66	3.97
4.22	3.47	3.92	4.91	3.95	4.38	4.12	4.52	4.35	3.91
4.10	4.09	4.09	4.34	4.09	4.88	4.28	3.98	3.86	4.58

De Sokal & Rohlf (1995).

Olhando para este conjunto de 120 *números é difícil retirar algo de útil daqui, ao contrário do que acontece se os dispusermos num gráfico.*

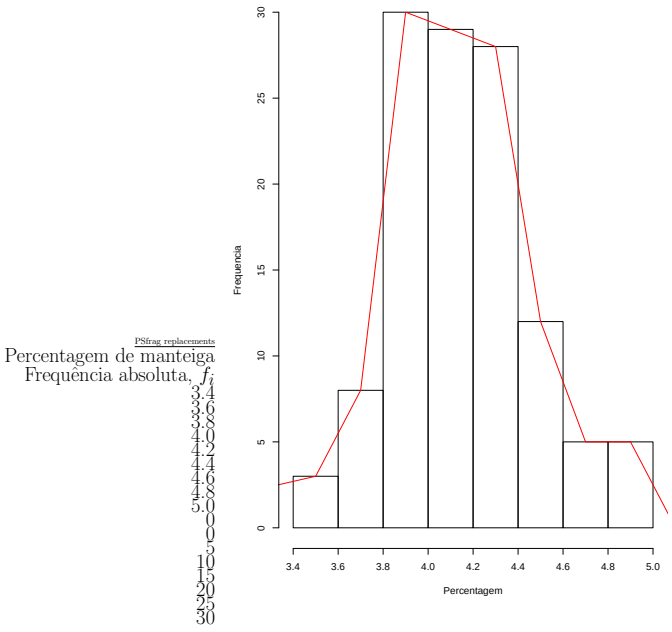
Para tal começamos por determinar o número de classes a usar para agrupar os dados, através da regra de Sturges:

$$k \approx 1 + \frac{\log n}{\log 2} = 1 + \frac{\log 120}{\log 2} \approx 1 + 6.907 = 7.907 \approx 8 \text{ classes.}$$

Notando agora que o máximo do conjunto de dados é 5.00 e o mínimo é 3.47, temos que a amplitude dos dados vale $L = 5.00 - 3.47 = 1.53$ *e, portanto, a amplitude de cada classe deve ser de* $l = \frac{L}{k} = \frac{1.53}{8} = 0.19125 \approx 0.2$. *Obtemos então as seguintes distribuições de frequências (absoluta, absoluta acumulada, relativa e relativa acumulada):*

i	Classe i	Frequência absoluta, f_i	Freq. absoluta acumulada F_i	Frequência relativa, f_i^*	Freq. relativa acumulada F_i^*
1	]3.4 ; 3.6]	3	3	0.025	0.025
2	]3.6 ; 3.8]	8	11	0.067	0.092
3	]3.8 ; 4.0]	30	41	0.250	0.342
4	]4.0 ; 4.2]	29	70	0.242	0.583
5	]4.2 ; 4.4]	28	98	0.233	0.817
6	]4.4 ; 4.6]	12	110	0.100	0.917
7	]4.6 ; 4.8]	5	115	0.042	0.958
8	]4.8 ; 5.0]	5	120	0.042	1.000

Usando agora as frequências absolutas, por exemplo, pode construir-se o seu histograma e desenhar o correspondente polígono de frequências (a vermelho):



Daqui facilmente verificamos que a grande maioria destas vacas produz percentagens de manteiga entre 3.8 e 4.4, havendo aproximadamente o mesmo número de vacas melhores e piores produtoras em termos de manteiga - simetria na distribuição das frequências.

Repare ainda nos valores das frequências relativas acumuladas de onde se pode verificar que mais de 50% das observações correspondem a uma percentagem de manteiga inferior a 4.2%.

□

1.3 Medidas descritivas

Anteriormente vimos como resumir um conjunto de dados num gráfico. Adicionalmente pode ser útil reduzir esses mesmos dados a um ou mais números que os representem, como por exemplo a uma média. Estes números vão tomar o nome de medidas descritivas.

As medidas descritivas dividem-se em 3 tipos: medidas de localização, medidas de dispersão e medidas de forma. Servem, respectivamente, para responder a questões do tipo:

1. Onde é o "meio"dos dados? Que dado ocorre mais vezes? Como se posiciona o meu valor comparado com todos os outros?
2. Quão "espalhados"se encontram os dados?
3. São os meus dados simétricos?

1.3.1 Medidas de localização

As medidas de localização servem para determinar o "meio"dos dados ou o seu valor "mais típico"ou ainda para determinar como determinado valor se posiciona em relação aos restantes. As medidas mais usuais são a **média**, a **mediana**, **moda**, os **quartis** e os **percentis**.

Dado um conjunto de dados $D = \{x_1, \ldots, x_n\}$ temos as seguintes definições:

Média amostral:

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

Mediana:

$$M_e = \begin{cases} \left(\frac{n+1}{2}\right)^{\text{ésimo}} \text{ valor do conjunto D ordenado,} & \text{n é ímpar} \\ \text{Média dos 2 valores centrais do conjunto D ordenado,} & \text{n é par} \end{cases}$$

Moda:

$$M_o = \text{Valor em D que ocorre mais vezes}$$

Percentil de ordem p:

$$q_p = \left\lceil n \times \frac{p}{100} \right\rceil^{\text{ésimo}} \text{ valor do conjunto D ordenado,} \quad p \in [0, 100]$$

1º Quartil:

$$Q_1 = \lceil 0.25n \rceil^{\text{ésimo}} \text{ valor do conjunto D ordenado}$$

3º Quartil:

$$Q_3 = \lceil 0.75n \rceil^{\text{ésimo}} \text{ valor do conjunto D ordenado}$$

Note-se que os quartis não são mais do que percentis - o 1º quartil é o percentil 25 e o 3º quartil é o percentil 75. O 2º quartil não é mais do que o percentil 50, que por sua vez não é mais do que a mediana.

Quando os dados se encontram agrupados, as medidas anteriores não podem assim ser determinadas, tendo de se recorrer a uma interpolação linear. Notamos que a moda deverá encontrar-se contida na classe com maior frequência absoluta - dita **classe modal** - e a mediana deverá estar contida na primeira classe cuja correspondente frequência relativa acumulada ultrapasse 0.5 - dita **classe mediana**.

Denotando Li e Ls os limites inferior e superior, respectivamente, das classes onde se encontram as medidas de localização a serem determinadas, PM_i o ponto médio da classe i, me o número da classe mediana, mo o número da classe modal, $mq1$ o número da classe do 1º quartil, $mq3$ o número da classe do 3º quartil, mpp o número da classe do percentil p e l a amplitude das classes, temos que:

Média amostral:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{k} f_i PM_i$$

Mediana:

$$M_e = Li + \frac{\frac{n+1}{2} - F_{me-1}}{F_{me+1} - F_{me-1}} \times l$$

Moda:

$$M_o = Li + \frac{f_{mo+1}}{f_{mo-1} + f_{mo+1}} \times l$$

1º Quartil:

$$Q_1 = Li + \frac{\frac{n+1}{4} - F_{mq1-1}}{F_{mq1+1} - F_{mq1-1}} \times l$$

3º Quartil:

$$Q_3 = Li + \frac{\frac{3(n+1)}{4} - F_{mq3-1}}{F_{mq3+1} - F_{mq3-1}} \times l$$

Percentil de ordem p:

$$q_p = Li + \frac{\frac{p(n+1)}{100} - F_{mpp-1}}{F_{mpp+1} - F_{mpp-1}} \times l$$

Notas: Quando tratamos com dados susceptíveis de conter *outliers* a mediana verifica-se ser uma medida de localização melhor que a média, uma vez que é menos sensível a esse tipo de valores extremos. Notamos ainda que a moda não tem de ser única.

1.3.2 Medidas de dispersão

A dispersão é a tendência dos dados se espalharem em torno da média. As medidas mais habituais são a amplitude dos dados, a variância, o desvio padrão e o coeficiente de variação, que se passam a definir, relativamente ao conjunto de dados $D = \{x_1, \ldots, x_n\}$.

Amplitude:

$$L = \max D - \min D$$

Variância amostral:

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

Desvio padrão amostral:

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

Coeficiente de variação:

$$cv = \frac{s}{\bar{x}} \times 100$$

Note-se que o coeficiente de variação representa a percentagem da média amostral a que corresponde o desvio padrão amostral.

No caso de dados agrupados devemos reformular as nossas definições. Sendo PM_i o ponto médio da classe i e f_i a correspondente frequência absoluta:

Variância amostral:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{k}f_i(PM_i - \bar{x})^2$$

Desvio padrão amostral:

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

1.3.3 Medidas de forma

Servem para estudar a simetria dos dados. Vamos aqui apenas considerar o coeficiente de enviesamento de Pearson:

Coeficiente de enviesamento de Pearson:

$$Sk = \frac{3(\bar{x} - Me)}{s}$$

Os valores de Sk variam entre -3 e 3. Se os dados forem perfeitamente **simétricos** então $Sk = 0$, já que a mediana e a média dos dados coincidem. Se $Sk > 0$ (respectivamente, $Sk < 0$) tal significa que a média é maior (respectivamente menor) que a mediana, sendo os dados **enviesados para a direita** (respectivamente, **enviesados para a esquerda**).

Exemplo 1.2 *Retomemos o exemplo 1.1. Uma vez que dispomos dos dados desagregados podemos calcular:*

Média amostral: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{120}\sum_{i=1}^{n} x_i = 4.166$;

Mediana: como n=120 é par, Me=Média dos 2 valores centrais do conjunto ordenado de dados,

$\{3.47, 3.56, 3.58, 3.66, 3.66, 3.70, 3.70, 3.72, 3.74, 3.74, 3.77, 3.81, 3.82, 3.83, 3.86, 3.86, 3.87, 3.88,$
$3.89, 3.89, 3.89, 3.89, 3.91, 3.91, 3.92, 3.93, 3.94, 3.95, 3.96, 3.96, 3.97, 3.97, 3.97, 3.97, 3.97, 3.98,$
$3.99, 3.99, 4.00, 4.00, 4.00, 4.02, 4.03, 4.05, 4.05, 4.05, 4.06, 4.06, 4.07, 4.08, 4.09, 4.09, 4.09, 4.09,$
$4.10, 4.10, 4.10, 4.11, 4.12, \mathbf{4.14}, \mathbf{4.15}, 4.16, 4.16, 4.17, 4.17, 4.18, 4.20, 4.20, 4.20, 4.20, 4.22, 4.23,$
$4.24, 4.24, 4.24, 4.24, 4.25, 4.27, 4.28, 4.28, 4.29, 4.29, 4.30, 4.31, 4.32, 4.32, 4.33, 4.33, 4.33, 4.34,$
$4.35, 4.36, 4.38, 4.38, 4.38, 4.38, 4.38, 4.40, 4.41, 4.42, 4.42, 4.46, 4.48, 4.49, 4.51, 4.52, 4.58, 4.58,$
$4.60, 4.60, 4.66, 4.67, 4.67, 4.70, 4.71, 4.81, 4.82, 4.88, 4.91, 5.00\}$

Logo, $Me = \frac{4.14+4.15}{2} = 4.145$.

Moda: $M_o =$ *Valor que ocorre mais vezes* $= 3.97$ *e* 4.38 *(aparecem ambos 5 vezes, 2 modas).*

1º Quartil: $Q_1 = \lceil 0.25n \rceil = \lceil 0.25 \times 120 \rceil = 30^{\acute{e}simo}$ *valor do conjunto de dados ordenado =3.96*

3º Quartil: $Q_3 = \lceil 0.75n \rceil = \lceil 0.75 \times 120 \rceil = 90^{\acute{e}simo}$ *valor do conjunto dados ordenado =4.34*

Amplitude: L=5.00-3.47=1.53

Variância amostral: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = 0.091$

Desvio padrão amostral: $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} = 0.302$

Coeficiente de variação: $cv = \frac{s}{\bar{x}} \times 100 = 7.258\%$

Coeficiente de enviesamento de Pearson: $Sk = \frac{3(\bar{x}-Me)}{s} = 0.209$.

Confirma ligeiro enviesamento direito verificado no histograma. A distribuição é pois apenas ligeiramente assimétrica, o que é corroborado pelo facto de a média amostral, a mediana e a moda estarem relativamente próximas.

Apesar de neste exemplo concreto termos os dados desagregados, vamos usar as classes definidas no exemplo 1.1 para calcular algumas das medidas atrás e comparar resultados. Assim:

Média amostral: $\bar{x} = \frac{1}{n}\sum_{i=1}^{k} f_i PM_i = \frac{3\times 3.5+8\times 3.7+\ldots}{120} = 4.153$;

Mediana: A classe 4, $]4.0; 4.2]$*, é a primeira cuja frequência relativa acumulada ultrapassa os* 50% *dos dados, pelo que é esta a classe mediana.*

$$M_e = Li + \frac{\frac{n+1}{2} - F_{me-1}}{F_{me+1} - F_{me-1}} \times l = 4.0 + \frac{\frac{120+1}{2} - 41}{98-41} \times 0.2 = 4.068$$

Moda: A classe modal é a classe 3, $]3.8; 4.0]$*, já que é aquela a que corresponde maior frequência absoluta. Assim:*

$$M_o = Li + \frac{f_{mo+1}}{f_{mo-1} + f_{mo+1}} \times l = 3.8 + \frac{29}{8 + 29} \times 0.2 = 3.957$$

1º Quartil: A classe 3, $]3.8; 4.0]$*, é a primeira cuja frequência relativa acumulada ultrapassa os* 25% *dos dados, pelo que é esta a classe do 1º quartil:*

$$Q_1 = Li + \frac{\frac{n+1}{4} - F_{mq1-1}}{F_{mq1+1} - F_{mq1-1}} \times l = 3.8 + \frac{\frac{120+1}{4} - 11}{70-11} \times 0.2 = 3.865$$

3º Quartil: A classe 5, $]4.2; 4.4]$*, é a primeira cuja frequência relativa acumulada ultrapassa os* 75% *dos dados, pelo que é esta a classe do 3º quartil:*

$$Q_3 = Li + \frac{\frac{3(n+1)}{4} - F_{mq3-1}}{F_{mq3+1} - F_{mq3-1}} \times l = 4.2 + \frac{\frac{3(120+1)}{4} - 70}{110-70} \times 0.2 = 4.304$$

Naturalmente que tanto a mediana, como os quartis e a moda devem estar contidos nas respectivas classes, o que constitui uma forma de confirmarmos se os nossos cálculos estão correctos.

Variância amostral: $s^2 = \frac{1}{n-1}\sum_{i=1}^{k} f_i(PM_i - \bar{x})^2 = \frac{3\times(3.5-4.153)^2+8\times(3.7-4.153)^2+\ldots}{119} = 0.095$

Desvio padrão amostral: $s = \sqrt{s^2} = 0.308$

Coeficiente de variação: $cv = \frac{s}{\bar{x}} \times 100 = 7.406\%$

Vemos pois que as aproximações obtidas a partir dos dados agrupados estão próximas dos verdadeiros valores. Quanto mais distantes estiverem os verdadeiros valores dos obtidos através dos dados agrupados, maior é a perda de informação devida ao agrupamento.

□

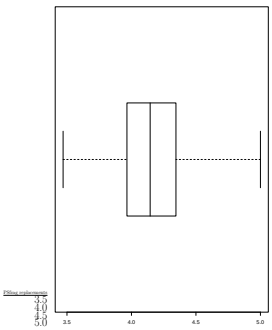
1.4 Diagrama de caixa-e-bigodes

Apresentamos por último um outro dispositivo gráfico bastante útil, o chamado **diagrama de caixa-e-bigodes**.

Para construir este diagrama temos de conhecer quanto valem os máximo e mínimo dos dados, a sua mediana e os 1º e 3º quartis. Com estes desenha-se uma caixa rectangular em que o topo inferior é dado pelo 1º quartil e o superior pelo 3º quartil. A caixa é dividida em duas partes pelo valor da mediana dos dados. Acrescentam-se-lhe então 2 bigodes que partem, respectivamente, um do extremo inferior da caixa até ao mínimo dos dados e o outro do extremo superior para o máximo - ver exemplo 1.3.

Este diagrama é muito útil para identificar assimetrias nos dados, caso a caixa esteja partida em dois pedaços muito diferentes, e para identificar *outliers*, no caso de os bigodes serem, relativamente à caixa, muito grandes.

Exemplo 1.3 *Construamos o diagrama de caixa-e-bigodes para dos dados do exemplo 1.1, lembrando que* $Me = 4.145$, $Q1 = 3.96$, $Q3 = 4.34$, *mínimo dos dados é 3.47 e máximo dos dados é 5.00:*



Confirma-se ligeira assimetria direita dos dados.

□