Aula 5

OUTRAS MEDIDAS ESTATÍSTICAS

Objetivos

Nesta aula, você estudará outras características de uma distribuição de dados e verá métodos alternativos de análise que tratam de forma diferenciada os valores discrepantes. Serão apresentados os seguintes conceitos:

1. coeficiente de variação;
2. escores padronizados;
3. teorema de Chebyshev;
4. medidas de assimetria;
5. o boxplot (gráfico de caixas).

COEFICIENTE DE VARIAÇÃO

Considere a seguinte situação: uma fábrica de ervilhas comercializa seu produto em embalagens de 300 gramas e em embalagens de um quilo ou 1000 gramas. Para efeitos de controle do processo de enchimento das embalagens, sorteia-se uma amostra de 10 embalagens de cada uma das máquinas, obtendo-se os seguintes resultados:

$$300g \longrightarrow \begin{cases} \overline{x} = 295g \\ \sigma = 5g \end{cases}$$

$$1000g \longrightarrow \begin{cases} \overline{x} = 995g \\ \sigma = 5g \end{cases}$$

Vamos interpretar esses números. Na primeira máquina, as embalagens deveriam estar fornecendo peso de 300g mas, devido a erros de ajuste da máquina de enchimento, o peso médio das 10 embalagens é de apenas 295g. O desvio padrão de 5g significa que, em média, os pesos das embalagens estão cinco gramas abaixo ou acima do peso médio das 10 latas. Uma interpretação análoga vale para a segunda máquina.

Em qual das duas situações a variabilidade parece ser maior? Ou seja, em qual das duas máquinas parece haver um problema mais sério? Note que, em ambos os casos, há uma dispersão de 5g em torno da média, mas 5g em 1000g é menos preocupante que 5g em 300g.

Como um exemplo mais extremo, um desvio padrão de 10 unidades em um conjunto cuja observação típica é 100 é muito diferente de um desvio padrão de 10 unidades em um conjunto cuja observação típica é 10.000.

Surge, assim, a necessidade de uma medida de *dispersão relativa*, que permita comparar, por exemplo, esses dois conjuntos. Uma dessas medidas é o *coeficiente de variação*.

Definição 5.1.

Dado um conjunto de observações $x_1, x_2, \ldots, x_n$, o **coeficiente de variação** (CV) é definido como a razão entre o desvio padrão dos dados e sua média, ou seja,

$$CV = \frac{\sigma}{\overline{x}}. \tag{5.1}$$

Note que, como o desvio padrão e a média são ambos medidos na mesma unidade dos dados originais, o coeficiente de variação é *adimensional*. Este fato permite comparações entre conjuntos de dados diferentes, medidos em unidades diferentes. Em geral, o *CV* é apresentado em forma percentual, isto é, multiplicado por 100.

No exemplo das latas de ervilha, os coeficientes de variação para as embalagens oriundas das duas máquinas são

$$300g \longrightarrow CV = \frac{5}{300} \times 100 = 1,67\%$$
$$1000g \longrightarrow CV = \frac{5}{1000} \times 100 = 0,5\%$$

o que confirma a nossa observação anterior: a variabilidade na máquina de 300g é relativamente maior.

Exercício 5.1.

Faça uma análise comparativa do desempenho dos alunos e alunas de uma turma de Estatística, segundo as notas dadas a seguir. Para isso, calcule a média, o desvio padrão e o coeficiente de variação, comentando os resultados.

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

ESCORES PADRONIZADOS

Considere os dois conjuntos de dados abaixo, que representam as notas em Estatística e Cálculo dos alunos de uma determinada turma.

Aluno	1	2	3	4	5	6	7	8	9
Estatística	6	4	5	7	8	3	5	5	7
Cálculo	6	8	9	10	7	7	8	9	5

As notas médias nas duas disciplinas são:

$$\begin{aligned}\overline{x}_E &= \frac{6+4+5+7+8+3+5+5+7}{9} = \frac{50}{9} = 5,5556\\ \overline{x}_C &= \frac{6+8+9+10+7+7+8+9+5}{9} = \frac{69}{9} = 7,6667\end{aligned}$$

As variâncias são:

$$\begin{aligned}\sigma_E^2 &= \frac{6^2+4^2+5^2+7^2+8^2+3^2+5^2+5^2+7^2}{9} - \left(\frac{50}{9}\right)^2 =\\ &= \frac{298}{9} - \frac{2500}{81} = \frac{298\times 9-2500}{81} = \frac{182}{81} = 2,246914\end{aligned}$$

$$\begin{aligned}\sigma_C^2 &= \frac{6^2+8^2+9^2+10^2+7^2+7^2+8^2+9^2+5^2}{9} - \left(\frac{69}{9}\right)^2 =\\ &= \frac{549}{9} - \frac{4761}{81} = \frac{549\times 9-4761}{81} = \frac{180}{81} = 2,222222\end{aligned}$$

Os desvios padrões são:

$$\begin{aligned}\sigma_E &= \sqrt{\frac{182}{81}} = 1,498971\\ \sigma_C &= \sqrt{\frac{180}{81}} = 1,490712\end{aligned}$$

Analisando os dois conjuntos de notas, pode-se ver que o aluno 1 tirou 6 em Estatística e em Cálculo. No entanto, a nota média em Estatística foi 5,56, enquanto que em Cálculo a nota média foi 7,67. Assim, o 6 em Estatística "vale mais" que o 6 em Cálculo, no sentido de que ele está acima e mais próximo da média.

Uma forma de medir tal fato é considerar a posição relativa de cada aluno no grupo. Para isso, o primeiro passo consiste em comparar a nota do aluno com a média do grupo, considerando o seu desvio em torno da média. Se x_i é a nota do aluno, passamos a trabalhar com $x_i - \overline{x}$.

Dessa forma, vemos que a nota 6 em Estatística gera um desvio de 0,44, enquanto a nota 6 em Cálculo gera um desvio de -1,67, o que significa que o aluno 1 tirou nota acima da média em Estatística e nota abaixo da média em Cálculo.

Um outro problema que surge na comparação do desempenho nas duas disciplinas é o fato de o desvio padrão ser diferente nas duas matérias. A variabilidade em Estatística foi um pouco maior que em Cálculo. Assim, o segundo passo consiste em padronizar a escala. Essa padronização da escala se faz dividindo os desvios em torno da média pelo desvio padrão do conjunto, o que nos dá o escore padronizado:

$$z_i = \frac{x_i - \overline{x}}{\sigma_x}. \tag{5.2}$$

O desvio padrão das notas de Estatística é $\sigma_E = 1,49897$ e das notas de Cálculo é $\sigma_C = 1,49071$. Na tabela a seguir, temos os escores padronizados; podemos ver aí que o escore relativo à nota 6 em Estatística é maior que o escore da nota 6 em Cálculo, indicando que a primeira "vale mais" que a segunda.

Aluno	1	2	3	4	5	6	7	8	9
Estatística	0,297	−1,038	−0,371	0,964	1,631	−1,705	−0,371	−0,371	0,964
Cálculo	−1,118	0,224	0,894	1,565	−0,447	−0,447	0,224	0,894	−1,789

Da mesma forma, o 5 em Estatística do aluno 7 vale mais que o 5 em Cálculo do aluno 9: ambos estão abaixo da média, mas o 7 em Estatística está "mais próximo" da média.

Ao padronizarmos os dados, a nossa escala passa a ser definida em termos de desvio padrão. Ou seja, passamos a dizer que tal observação está abaixo (ou acima) da média por determinado número de desvios padrões. Com isso, tira-se o efeito de as médias e as variabilidades serem diferentes. Podemos escrever o escore padronizado como

$$z_i = \frac{1}{\sigma_x} x_i - \frac{\overline{x}}{\sigma_x}$$

e daí vemos que esse escore é obtido a partir dos dados originais por uma transformação linear: somamos uma constante $\left(-\frac{\overline{x}}{\sigma_x}\right)$ e multiplicamos por outra constante $\left(\frac{1}{\sigma_x}\right)$.

Das propriedades da média e do desvio padrão vistas nas aulas anteriores, resulta que a média e o desvio padrão dos escores padronizados podem ser obtidos a partir da média e do desvio padrão dos dados originais:

$$\overline{z} = \frac{1}{\sigma_x}\overline{x} - \frac{\overline{x}}{\sigma_x} = 0 \quad \text{e} \quad \sigma_z^2 = \frac{1}{\sigma_x^2}\sigma_x^2 = 1$$

Logo, os escores padronizados têm sempre média zero e desvio padrão (ou variância) 1.

TEOREMA DE CHEBYSHEV E VALORES DISCREPANTES

Os escores padronizados podem ser usados para se detectarem valores discrepantes ou muito afastados do conjunto de dados, graças ao Teorema de Chebyshev.

Teorema 5.1 (Teorema de Chebyshev)**.**

Para qualquer distribuição de dados, pelo menos $\left(1-1/z^2\right)$ dos dados estão dentro de z desvios padrões da média, onde z é qualquer valor maior que 1. Dito de outra forma, pelo menos $\left(1-1/z^2\right)$ dos dados estão no intervalo $[\overline{x}-z\sigma;\ \overline{x}+z\sigma]$.

Vamos analisar esse teorema em termos dos escores padronizados. Suponha que x' seja um valor do conjunto de dados dentro do intervalo $[\overline{x}-z\sigma;\ \overline{x}+z\sigma]$. Isso significa que

$$\overline{x}-z\sigma < x' < \overline{x}+z\sigma.$$

Subtraindo $\overline{x}$ e dividindo por σ todos os termos dessa desigualdade obtemos que

$$\begin{aligned} \frac{\overline{x}-z\sigma-\overline{x}}{\sigma} \;&<\; \frac{x'-\overline{x}}{\sigma} \;<\; \frac{\overline{x}+z\sigma-\overline{x}}{\sigma} \Rightarrow \\ -z \;&<\; \frac{x'-\overline{x}}{\sigma} \;<\; +z \end{aligned}$$

O termo do meio nada mais é que o escore padronizado da observação x'. Assim, o teorema de Chebyshev pode ser estabelecido em termos dos escores padronizados como: para pelo menos $\left(1-1/z^2\right)$ dos dados, os respectivos escores padronizados estão no intervalo $(-z,+z)$, onde z é qualquer valor maior que 1.

O fato interessante desse teorema é que ele vale para qualquer distribuição de dados. Vamos ver alguns exemplos numéricos.

- $z=2$

  Nesse caso, $1-1/z^2=3/4$, ou seja, para pelo menos 75% dos dados, os escores padronizados estão no intervalo $(-2,+2)$.

- $z = 3$

  Nesse caso, $1 - 1/z^2 = 8/9 = 0,889$, ou seja, para aproximadamente 89% dos dados, os escores padronizados estão no intervalo $(-3, +3)$.

- $z = 4$

  Nesse caso, $1 - 1/z^2 = 15/16 = 0,9375$, ou seja, para 93,75% dos dados, os escores padronizados estão no intervalo $(-4, +4)$.

Como regra de detecção de valores discrepantes, pode-se usar o Teorema de Chebyshev para se estabelecer, por exemplo, dados cujos escores padronizados estejam fora do intervalo $(-3, +3)$ são valores discrepantes e, portanto, devem ser verificados cuidadosamente para se identificar a causa de tal discrepância. Algumas vezes, tais valores podem ser resultados de erros, mas muitas vezes eles são valores legítimos e a presença deles requer alguns cuidados na análise estatística.

Exercício 5.2.

Considere os dados da **Tabela 5.1** sobre a densidade populacional das unidades da federação brasileira. Calcule os escores padronizados e determine se alguma UF pode ser considerada valor discrepante com relação a essa variável.

Tabela 5.1: Densidade populacional dos estados brasileiros:

UF	Densidade populacional (hab/km^2)	UF	Densidade populacional (hab/km^2)
RO	6	SE	81
AC	4	BA	24
AM	2	MG	31
RR	2	ES	68
PA	5	RJ	328
AP	4	SP	149
TO	5	PR	48
MA	17	SC	57
PI	12	RS	37
CE	51	MS	6
RN	53	MT	3
PB	61	GO	15
PE	81	DF	353
AL	102		

Fonte: IBGE - Censo Demográfico 2000

MEDIDAS DE ASSIMETRIA

Considere os diagramas de pontos dados nas partes (a) a (c) da **Figura 5.1**, onde a seta indica a média dos dados. Analisando-os, podemos ver que a principal e mais marcante diferença entre eles diz respeito à simetria da distribuição. A segunda distribuição é simétrica, enquanto as outras duas são assimétricas.

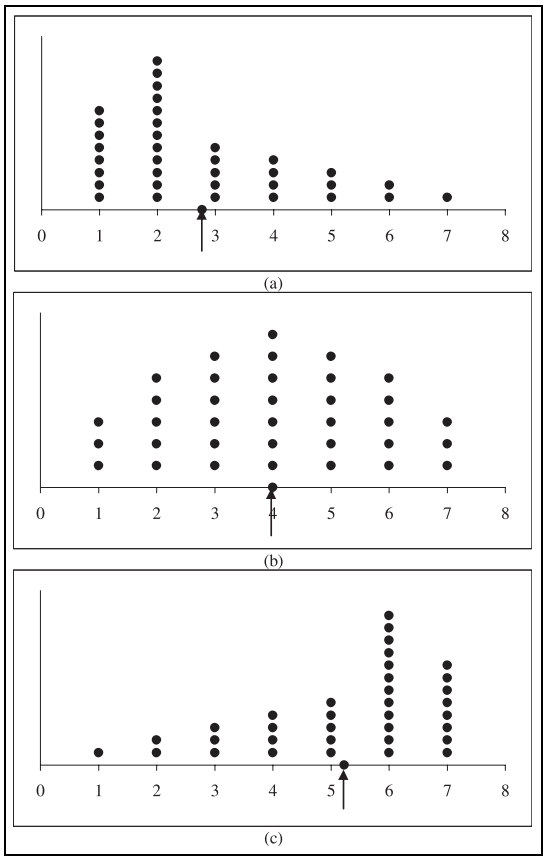

Figura 5.1: Diagramas de pontos de distribuições com diferentes tipos de assimetria.

No diagrama (a), a assimetria é tal que há maior concentração na cauda inferior, enquanto no diagrama (c), a concentração é maior na cauda superior. Visto de outra maneira, no diagrama (a), os dados se estendem para o lado positivo da escala, enquanto no diagrama (c), os dados se estendem para o lado negativo da escala. Dizemos que a distribuição ilustrada no diagrama (a) apresenta uma *assimetria à direita*, enquanto a do diagrama (c) apresenta uma *assimetria à esquerda*. No diagrama (b), temos uma *simetria* perfeita ou *assimetria nula*.

Esses três tipos de assimetria podem ser caracterizados pela posição da moda com relação à média dos dados. No primeiro

tipo, a moda tende a estar à esquerda da média, enquanto no terceiro tipo, a moda tende a estar à direita da média. (Lembre-se de que a média é o centro de gravidade ou ponto de equilíbrio da distribuição). Para distribuições simétricas, a moda coincide com a média. Definem-se, assim, os três tipos de assimetria:

- se a média é maior que a moda ($\overline{x} > x^*$), dizemos que a distribuição é *assimétrica à direita* ou tem *assimetria positiva* [diagrama (a) da **Figura 5.1**];
- se a média é igual à moda ($\overline{x} = x^*$), dizemos que a distribuição é simétrica ou tem assimetria nula [diagrama (b) da **Figura 5.1**];
- se a média é menor que a moda ($\overline{x} < x^*$), dizemos que a distribuição é assimétrica à esquerda ou tem assimetria negativa [diagrama (c) da **Figura 5.1**].

Essas definições, no entanto, não permitem "medir" diferentes graus de assimetria. Por exemplo, considere os diagramas de pontos (a) e (b) dados na **Figura 5.2**, ambos assimétricos à direita.

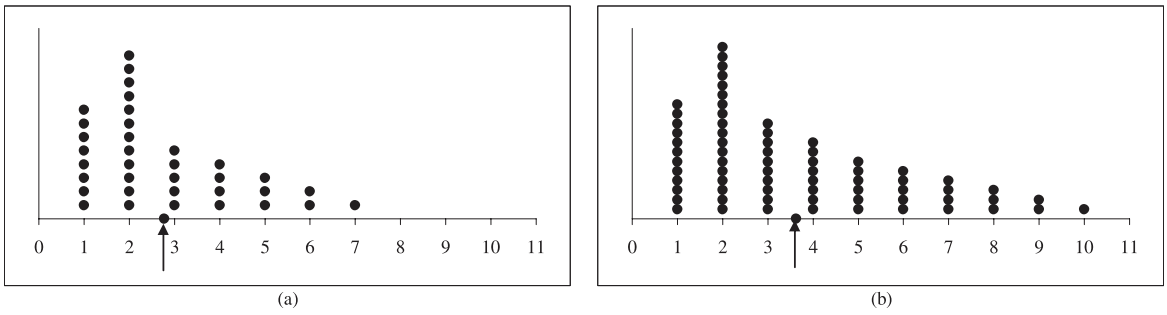

Figura 5.2: Duas distribuições assimétricas à direita.

Uma forma de medirmos essas diferentes assimetrias é através da distância $\overline{x} - x^*$ entre a média e a moda, mas como as distribuições podem ter graus de dispersão diferentes, é importante que consideremos a diferença acima na mesma escala. Assim, define-se um dos coeficientes de assimetria (definição devida a Karl Pearson) como:

$$\boxed{e = \frac{\overline{x} - x^*}{\sigma}}. \qquad (5.3)$$

Se o coeficiente é negativo, tem-se assimetria negativa; se é positivo, tem-se assimetria positiva e se é nulo, tem-se uma distribuição simétrica. Note que aqui, assim como nos escores padronizados, tiramos o efeito de escalas diferentes ao dividirmos pelo desvio padrão, o que resulta na adimensionalidade do coeficiente.

Para os dados do diagrama (a) da **Figura 5.2**, temos que $x^* = 2, \overline{x} = 2,7714$ e $\sigma = 1,6228$; logo,

$$e = \frac{2,7714 - 2}{1,6228} = 0,475351$$

Para os dados do diagrama (b) da **Figura 5.2**, $x^* = 2$, $\overline{x} = 3,6232$ e $\sigma = 2,3350$; logo,

$$e = \frac{3,6232 - 2}{2,3350} = 0,6952$$

o que indica uma assimetria mais acentuada.

É interessante observar que existem outros coeficientes de assimetria; o que apresentamos é o menos utilizado, mas é o mais intuitivo.

Exercício 5.3.

Considere novamente as notas de 50 alunos, cujo ramos e folhas é dado a seguir. Calcule o coeficiente de assimetria de Pearson para essa distribuição.

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

INTERVALO INTERQUARTIL

A mediana divide o conjunto de dados ao meio, deixando 50% das observações abaixo dela e 50% acima dela. De modo análogo, podemos definir qualquer *separatriz* como sendo um valor que deixa p% dos dados abaixo e o restante acima.

Vamos nos concentrar aqui em um caso particular das separatrizes, que são os *quartis*. O primeiro quartil, que indicaremos por Q_1, deixa 25% das observações abaixo e 75% acima. O segundo quartil é a mediana e o terceiro quartil, Q_3, deixa 75% das observações abaixo e 25% acima. Na figura a seguir, **Figura 5.3**, temos uma ilustração desses conceitos.

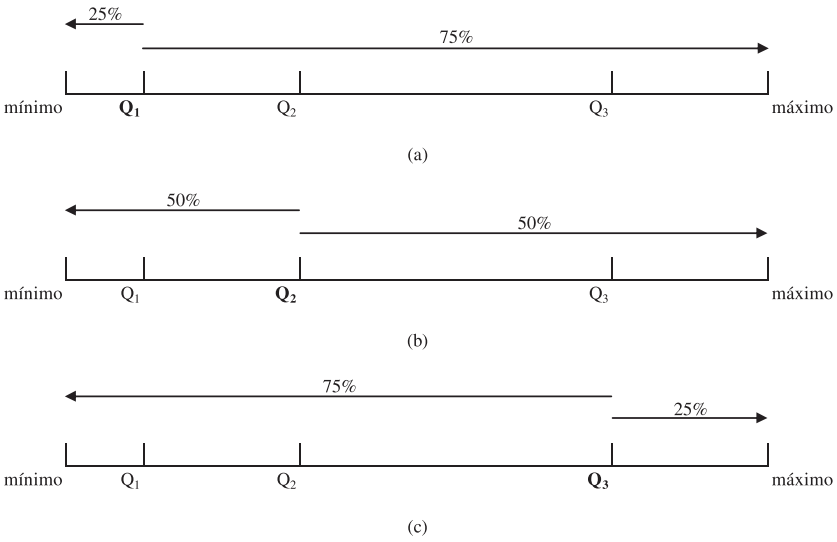

Figura 5.3: Ilustração da definição de quartis.

Analisando essa figura, podemos ver que entre Q_1 e Q_3 há sempre 50% dos dados, qualquer que seja a distribuição. Assim, quanto maior for a distância entre Q_1 e Q_3, mais dispersos serão os dados. Temos, assim, uma nova medida de dispersão, o *intervalo interquartil*.

Definição 5.2.

O **intervalo interquartil**, que denotaremos por IQ, é definido como a distância entre o primeiro e o terceiro quartis, isto é:

$$IQ = Q_3 - Q_1 \tag{5.4}$$

O intervalo interquartil tem a mesma unidade dos dados. A vantagem do intervalo interquartil sobre o desvio padrão é que, assim como a mediana, o IQ não é muito influenciado por valores discrepantes.

CÁLCULO DOS QUARTIS

O cálculo dos quartis pode ser feito da seguinte forma: depois de calculada a mediana, considere as duas partes dos dados, a parte abaixo da mediana e a parte acima da mediana, em ambos os casos excluindo a mediana. Essas duas partes têm o mesmo número de observações, pela definição de mediana.

O primeiro quartil, então, será calculado como a mediana da parte abaixo da mediana original e o terceiro quartil será calculado como a mediana da parte acima da mediana original.

Exemplo 5.1.

Vamos calcular os quartis e o intervalo interquartil para o número de dependentes dos funcionários do Departamento de Recursos Humanos, cujos valores já ordenados são:

0 0 0 0 0 1 1 **1** 2 2 2 3 3 3 4

Como há 15 observações, a mediana é a oitava observação (em negrito), isto é:

$$Q_2 = x_{\left(\frac{n+1}{2}\right)} = x_{(8)} = 1$$

Excluída essa oitava observação, a parte inferior dos dados é

0 0 0 **0** 0 1 1

cuja mediana é

$$Q_1 = x_{\left(\frac{7+1}{2}\right)} = x_{(4)} = 0$$

A parte superior dos dados, excluída a mediana, é

2 2 2 **3** 3 3 4

e, portanto,

$$Q_3 = x_{(4+8)} = x_{(12)} = 3$$

o intervalo interquartil é calculado como

$$IQ = Q_3 - Q_1 = 3 - 0 = 3.$$

MEDIDA DE ASSIMETRIA COM BASE NOS QUARTIS

É interessante observar que entre Q_1 e Q_2 e entre Q_2 e Q_3 há sempre 25% dos dados. Então, a diferença entre as distâncias $Q_2 - Q_1$ e $Q_3 - Q_2$ nos dá informação sobre a assimetria da distribuição.

Se $Q_2 - Q_1 < Q_3 - Q_2$, isso significa que "andamos mais rápido" para cobrir os 25% inferiores do que os 25% superiores, ou seja, a distribuição "se arrasta" para a direita.

Analogamente, se $Q_2 - Q_1 > Q_3 - Q_2$, isso significa que "andamos mais devagar" para cobrir os 25% inferiores do que os 25% superiores, ou seja, a distribuição "se arrasta" para a esquerda. De forma mais precisa, temos o seguinte resultado:

$$Q_2 - Q_1 < Q_3 - Q_2 \Longrightarrow \text{assimetria positiva}$$

$$Q_2 - Q_1 > Q_3 - Q_2 \Longrightarrow \text{assimetria negativa}$$

$$Q_2 - Q_1 = Q_3 - Q_2 \Longrightarrow \text{simetria ou assimetria nula}$$

Para tirar o efeito de escala, temos que dividir por uma medida de dispersão – lembre-se de que dividimos pelo desvio padrão quando trabalhamos com as diferenças $\overline{x} - x^*$. Aqui, para não termos efeito dos valores discrepantes, usaremos o intervalo interquartil para gerar a seguinte medida de assimetria, que é chamada medida de assimetria de Bowley:

$$B = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1}$$

que pode ser reescrita como

$$B = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{(Q_3 - Q_2) + (Q_2 - Q_1)}$$

Analisando essa expressão, podemos ver que quanto mais assimétrica à direita for uma distribuição, mais próximos serão Q_1 e Q_2 e, portanto, B se aproxima de $+1$. Analogamente, quanto mais assimétrica à esquerda, mais próximos serão Q_2 e Q_3 e, portanto, B se aproxima de -1.

Exercício 5.4.

Considere novamente os dados da **Tabela 2.2** sobre os funcionários do Departamento de Recursos Humanos, cujos salários (em R$) são os seguintes: 6300, 5700, 4500, 3800, 3200, 7300, 7100, 5600, 6400, 7000, 3700, 6500, 4000, 5100, 4500. Analise a assimetria da distribuição com base no coeficiente de Bowley.

O BOXPLOT

A partir dos quartis constrói-se um gráfico chamado *boxplot* ou *gráfico de caixas*, que ilustra os principais aspectos da distribuição e é também muito útil na comparação de distribuições.

O boxplot é formado basicamente por um retângulo vertical (ou horizontal). O comprimento do lado vertical (ou horizontal) é dado pelo intervalo interquartil (**Figura 5.4.a**, onde estamos trabalhando com um retângulo vertical). O tamanho do outro lado é indiferente, sugerindo-se apenas uma escala razoável. Na altura da mediana, traça-se uma linha, dividindo o retângulo em duas partes [**Figura 5.4.b**].

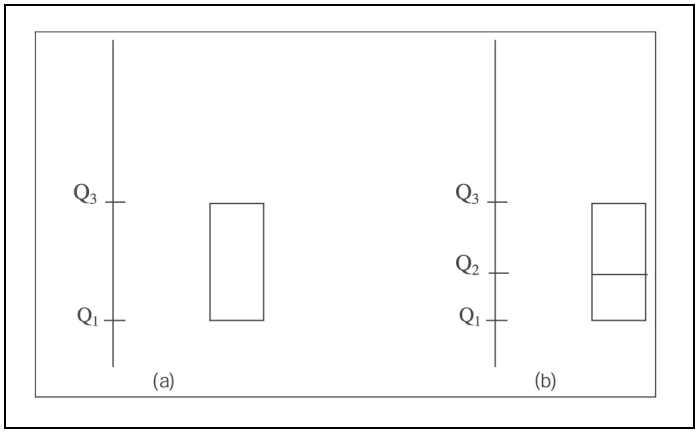

Figura 5.4: Construção do boxplot - Etapa 1.

Note que aí já temos representados 50% da distribuição e também já temos ideia da assimetria da mesma – nessa figura temos uma leve assimetria à direita, já que $Q_2 - Q_1 < Q_3 - Q_2$. Para representar os 25% restantes em cada cauda da distribuição, temos que cuidar primeiro da presença de possíveis *outliers* ou valores discrepantes, que, como já dito, são valores que se distanciam dos demais.

!

Regra de Valores Discrepantes

Um dado x será considerado valor discrepante ou *outlier* se

$$x < Q_1 - 1,5\ IQ$$

ou

$$x > Q_3 + 1,5\ IQ$$

Veja a **Figura 5.5.a**. Qualquer valor para fora das linhas pontilhadas é considerado um valor discrepante. Para representar o domínio de variação dos dados na cauda inferior que não são *outliers*, traça-se, a partir do lado do retângulo definido por Q_1, uma linha para baixo até o menor valor que não seja *outlier.*

Da mesma forma, na cauda superior, traça-se, a partir do lado do retângulo definido por Q_3, uma linha para cima até o maior valor que não seja *outlier.* [**Figura 5.5.b**]. Esses pontos são chamados *juntas*. Dito de outra forma, as juntas são os valores mínimo e máximo do conjunto de dados formado pelos valores não discrepantes.

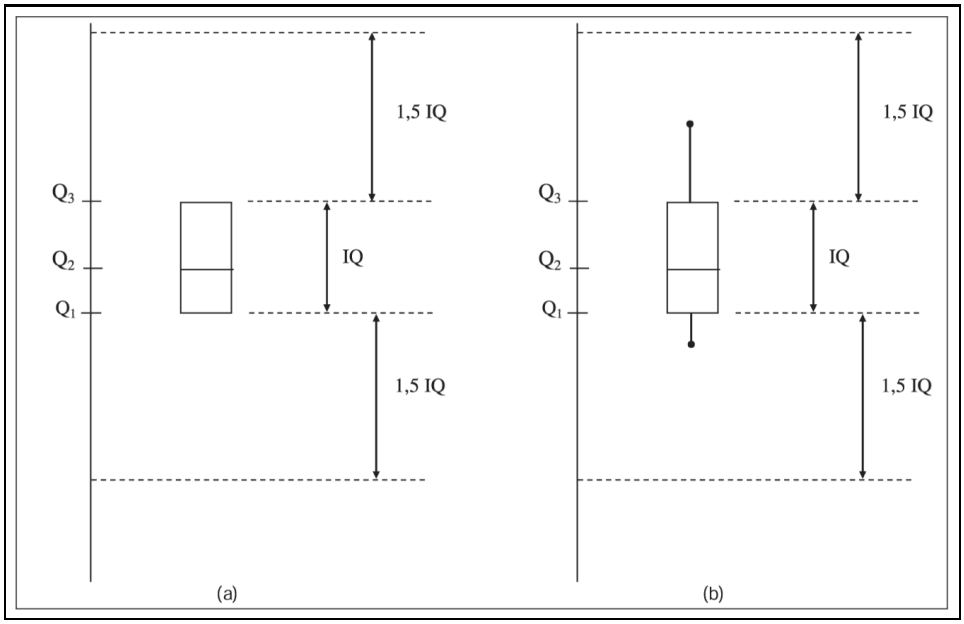

Figura 5.5: Construção do boxplot - Etapa 2.

Quanto aos *outliers*, eles são representados individualmente por um X (ou algum outro tipo de caracter), explicitando-se, de preferência, os seus valores, mas com uma possível quebra de escala no eixo (**Figura 5.6**).

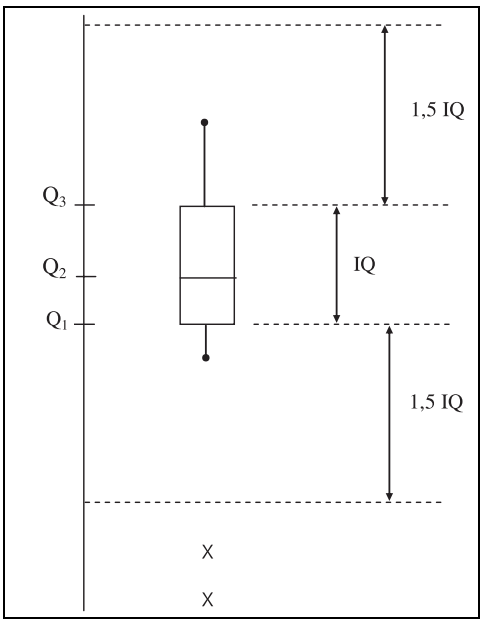

Figura 5.6: Construção do boxplot - Etapa 3.

Note que a construção do boxplot é toda baseada nos quartis, que são medidas resistentes contra valores discrepantes.

Exemplo 5.2.

Consideremos novamente as notas de 50 alunos, representadas no gráfico ramos e folhas da **Figura 5.7**.

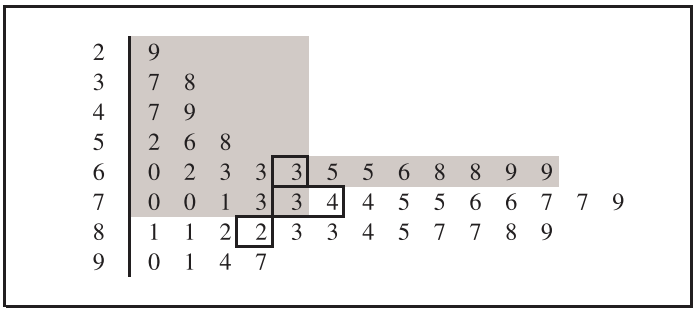

Figura 5.7: Notas de 50 alunos - Cálculo dos quartis.

A mediana divide o conjunto de dados em duas partes com 25 observações de cada lado (parte sombreada de cinza e a outra).

Como o número de observações é par, a mediana é a média dos valores centrais, que estão circundados por uma borda, um na parte inferior e outro na parte superior.

$$Q_2 = \frac{x_{\left(\frac{50}{2}\right)} + x_{\left(\frac{50}{2}+1\right)}}{2} = \frac{x_{(25)} + x_{(26)}}{2} = \frac{73+74}{2} = 73,5$$

O primeiro quartil é a mediana da parte inferior, que é o valor circundado por uma borda na parte sombreada de cinza e o terceiro quartil é a mediana da parte superior, que é o valor circundado por uma borda na parte superior, não sombreada.

$$\begin{aligned} Q_1 &= 63 \\ Q_3 &= 82 \\ IQ &= 82 - 63 = 19 \end{aligned}$$

Para estudarmos os outliers, temos que calcular

$$\begin{aligned} Q_1 - 1,5\ IQ &= 63 - 1,5 \times 19 = 34,5 \\ Q_3 + 1,5\ IQ &= 82 + 1,5 \times 19 = 110,5 \end{aligned}$$

Como a maior nota é 97, não há outliers na cauda superior, mas na cauda inferior, temos a nota 29 que é menor que 34,5 e, portanto, um outlier inferior. Excluído esse outlier, o menor valor que não é discrepante é 37 e o maior valor é 97; logo, as juntas são 37 e 97. Na **Figura 5.8**, temos o boxplot resultante.

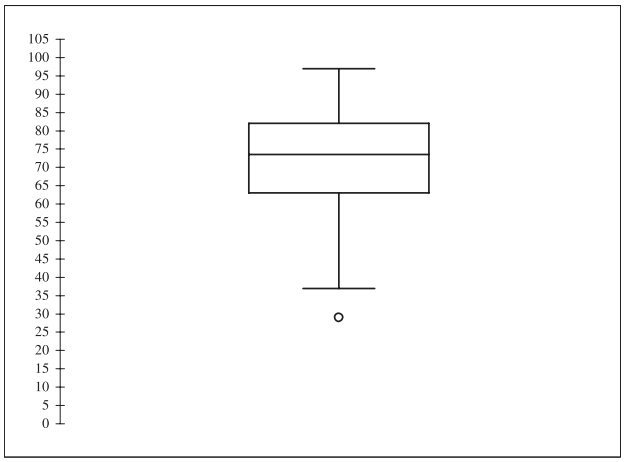

Figura 5.8: Boxplot para as 50 notas.

Note que no gráfico final não marcamos os valores 34,5 e 110,5; eles são usados apenas para delimitar os outliers. São as juntas que são exibidas no gráfico.

Exemplo 5.3.

Considere os dados apresentados na **Tabela 5.2**, onde temos as populações urbana, rural e total, em 1000 habitantes, dos estados brasileiros.

Tabela 5.2: População urbana e rural das UFs brasileiras (em 1000 hab.)

UF	População			UF	População		
	Urbana	Rural	Total		Urbana	Rural	Total
RO	885	496	1381	MG	14672	3220	17892
AC	371	188	559	ES	2464	635	3099
AM	2108	706	2814	RJ	13822	570	14392
RR	248	78	326	SP	34593	2440	37033
PA	4121	2072	6193	PR	7787	1778	9565
AP	425	53	478	SC	4218	1139	5357
TO	860	298	1158	RS	8318	1870	10188
MA	3365	2288	5653	MS	1748	331	2079
PI	1789	1055	2844	MT	1988	517	2505
CE	5316	2116	7432	GO	4397	607	5004
RN	2037	741	2778	DF	1962	90	2052
PB	2448	997	3445	PE	6059	1861	7920
AL	1920	903	2823	SE	1274	512	1786
BA	8773	4298	13071				

Fonte: IBGE - Censo Demográfico 2000

Vamos, inicialmente, construir o boxplot para a população total e, em seguida, um boxplot comparativo das populações urbana e rural. Na tabela a seguir, temos as estatísticas necessárias para a construção desses gráficos.

Estatística	Total	Urbana	Rural
Q_1	2052(DF)	1748 (MS)	496 (RO)
Q_2	3099 (ES)	2448 (PB)	741 (RN)
Q_3	7920 (PE)	6059 (PE)	1870 (RS)
IQ	5868	4311	1374
$Q_1 - 1,5\ IQ$	-6750	-4718,5	-1565
$Q_3 + 1,5\ IQ$	16722	12525,5	3931
Junta inferior	326 (RR)	248(RR)	53 (AP)
Junta superior	1439 (RJ)	8733(BA)	3220 (MG)
Outliers	17892 (MG)	13822 (RJ)	4298 (BA)
	37033 (SP)	14672 (MG)	
		34593 (SP)	

Na **Figura 5.9**, temos o boxplot para a população total; vemos aí que as populações de São Paulo e Minas Gerais são outliers e a distribuição apresenta uma forte assimetria à direita, ou seja, muitos estados têm população pequena enquanto alguns poucos têm população bem grande.

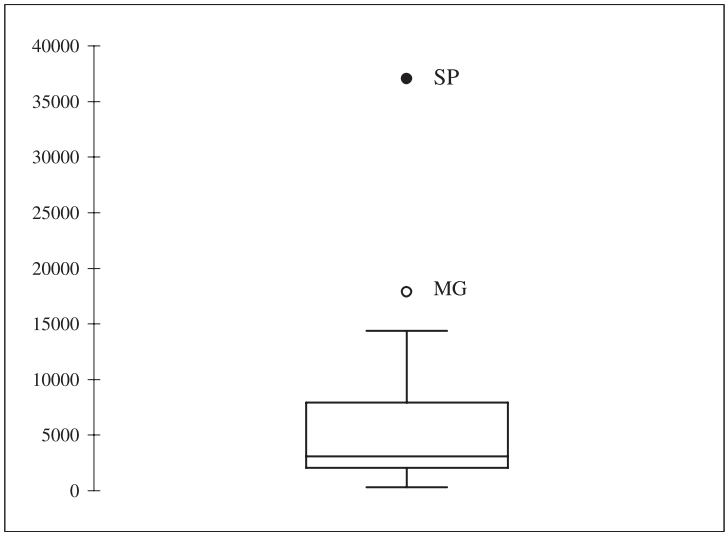

Figura 5.9: População total (em 1000 hab) das Unidades da Federação brasileiras.

Na **Figura 5.10**, temos um boxplot comparativo das populações urbana e rural. Podemos ver que a população urbana apresenta maior variabilidade e também uma forte assimetria positiva. Há três UFs que são discrepantes: São Paulo, Minas Gerais e Rio de Janeiro. Em termos da população rural, a Bahia é o único outlier e a distribuição também é assimétrica à direita.

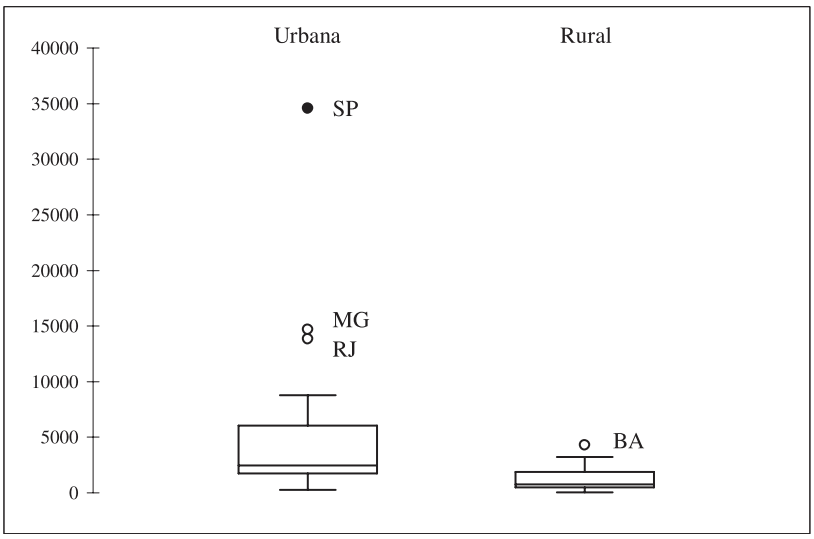

Figura 5.10: População urbana e rural das UFs brasileiras (em 1000 hab).

Exercício 5.5.

Construa o boxplot para os salários dos funcionários do Departamento de Recursos Humanos, cujos valores em reais são 6300, 5700, 4500, 3800, 3200, 7300, 7100, 5600, 6400, 7000, 3700, 6500, 4000, 5100, 4500.

Exercício 5.6.

Os dados a seguir representam o número de apólices de seguro que um corretor conseguiu vender em cada um de seus 20 primeiros dias em um emprego novo: 2, 4, 6, 3, 2, 1, 4, 3, 5, 2, 1, 1, 4, 0, 2, 2, 5, 2, 2, 1. Analise a assimetria da distribuição, utilizando os coeficientes de Pearson e de Bowley.

Exercício 5.7.

O professor Celso tem duas opções de caminho para se dirigir da sua casa até seu local de trabalho. Tentando definir qual o melhor caminho, ele anota o tempo de viagem em diferentes dias, obtendo os seguintes tempos (em minutos):

Caminho 1	12	11	10	10	8	12	15	7	20	12
Caminho 2	12	15	13	13	14	13	12	14	13	15

Faça uma análise comparativa desses dados para ajudar o professor Celso a escolher um caminho.

Exercício 5.8.

Em sua política de fidelização de clientes, determinado supermercado tem uma promoção de dar descontos especiais diferenciados no mês do aniversário do cliente. O desconto básico é de 5%, mas clientes especiais – aqueles com pontuação alta – podem receber prêmios adicionais, que variam a cada mês e de filial para filial. A seguir, você tem os pontos dos clientes aniversariantes de determinado mês em uma das filiais do supermercado.

77	69	72	73	71	75	75	74	71	72	74	73	75	71	74
73	78	77	74	75	69	76	76	80	74	85	74	73	72	74

a. Construa o gráfico ramo e folhas e comente suas principais características.

b. Calcule a mediana e o intervalo interquartil IQ.

c. Construa o boxplot e comente suas principais características.

d. Essa filial dá uma garrafa de champagne para seus clientes especiais, segundo a seguinte regra: a cada mês, os clientes com pontuação acima do terceiro quartil por 1,5 vezes o intervalo interquartil serão premiados. Algum cliente ganhará a garrafa de champagne nesse mês?

SOLUÇÃO DOS EXERCÍCIOS

Exercício 5.1.

Eis o resumo das estatísticas por sexo:

Sexo	Número Obs.	Média	Desvio padrão	Coef. variação
Masculino	11	5,273	1,884	0,357
Feminino	10	6,400	0,764	0,119

Podemos ver, então, que as mulheres, além de terem obtido uma média maior, apresentam variabilidade menor: o coeficiente de variação das mulheres é de 0,119 e o dos homens é de 0,357.

Exercício 5.2.

A densidade populacional média é 59,444 hab/km^2 e o desvio padrão das densidades é 87,253 hab/km^2. Na **Tabela 5.3**, apresentam-se os escores padronizados para cada UF, calculados pela fórmula $z_i = (x_i - \overline{x})/\sigma_x$. Por exemplo, para RO, o valor $-0,6125$ foi obtido como $(6 - 59,444)/87,253$.

Podemos ver que as únicas UFs com densidades relativamente altas, isto é, escores fora do intervalo $(-3,+3)$, são RJ e DF; não há densidade relativamente baixa.

Tabela 5.3: Escores padronizados das densidades populacionais

UF	Escores padronizados	UF	Escores padronizados
RO	-0,6125	SE	0,2470
AC	-0,6354	BA	-0,4062
AM	-0,6584	MG	-0,3260
RR	-0,6584	ES	0,0981
PA	-0,6240	RJ	3,0779
AP	-0,6354	SP	1,0264
TO	-0,6240	PR	-0,1312
MA	-0,4865	SC	-0,0280
PI	-0,5438	RS	-0,2572
CE	-0,0968	MS	-0,6125
RN	-0,0739	MT	-0,6469
PB	0,0178	GO	-0,5094
PE	0,2470	DF	3,3644
AL	0,4877		

Exercício 5.3.

Para esses dados, temos $\overline{x} = 71,42; x^* = 63; \sigma^2 = 215,2836$. Logo,

$$e = \frac{71,42 - 63}{\sqrt{215,2836}} = 0,5739$$

Exercício 5.4.

Os quartis para esse conjunto de dados são
$Q_2 = x_{(8)} = 5600; \quad Q_1 = x_{(4)} = 4000; \quad Q_3 = x_{(12)} = 6500.$

O intervalo interquartil é $Q_3 - Q_1 = 6500 - 4000 = 2500$. Logo,

$$\begin{aligned} B &= \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1} = \frac{(6500 - 5600) - (5600 - 4000)}{6500 - 4000} \\ &= -0,4666. \end{aligned}$$

Como B está mais próximo de -1 do que de 1, temos uma assimetria à esquerda.

Exercício 5.5.

Os quartis para esse conjunto de dados são
$Q_2 = x_{(8)} = 5600; \quad Q_1 = x_{(4)} = 4000; \quad Q_3 = x_{(12)} = 6500.$

O intervalo interquartil é $Q_3 - Q_1 = 6500 - 4000 = 2500$.

A regra para outliers é

$$\begin{aligned} x &< Q_1 - 1,5\ IQ = 4000 - 1,5 \times 2500 = 250 \\ x &> Q_3 + 1,5\ IQ = 6500 + 1,5 \times 2500 = 10250 \end{aligned}$$

Como o menor salário é 3200 e o maior salário é 7300, não há salários discrepantes. O boxplot é dado na **Figura 5.11**.

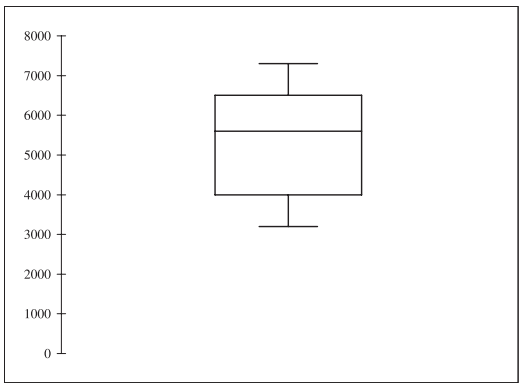

Figura 5.11: Solução do Exercício 5.5.

Exercício 5.6.

A média dos dados é $\overline{x} = 2,6$, com desvio padrão $\sigma = 1,5620$.

A moda é $x^* = 2$.

Os quartis são

$$Q_1 = \frac{x_{(5)} + x_{(6)}}{2} = 1,5;$$

$$Q_2 = \frac{x_{(10)} + x_{(11)}}{2} = 2;$$

$$Q_3 = \frac{x_{(15)} + x_{(16)}}{2} = 4.$$

Com esses valores, obtemos os coeficientes de assimetria:

$$e = \frac{\overline{x} - x^*}{\sigma} = \frac{2,6-2}{1,5620} = 0,3841$$

$$B = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1} = \frac{(4-2) - (2-1,5)}{4-1,5} = \frac{1,5}{3,5} = 0,4286$$

Existe, assim, uma assimetria positiva nos dados; veja o diagrama de pontos na **Figura 5.12**.

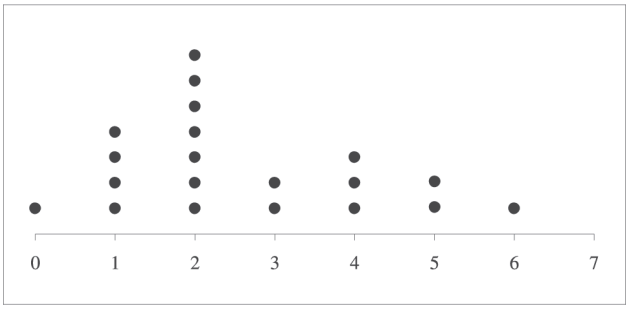

Figura 5.12: Solução do Exercício 5.6.

Exercício 5.7.

Na tabela a seguir, são apresentados os valores relevantes para a solução do exercício. Podemos concluir que o tempo pelo caminho 2 é menos variável, apesar de ser um pouco maior. Dessa forma, parece que o Prof. Celso deva optar por esse caminho, planejando-se para sair com a devida antecedência.

Caminho	Média	Desvio padrão	CV
1	11,7	3,6833	0,3148
2	13,1	0,9944	0,0759

Exercício 5.8.

a. Há uma grande concentração de folhas no ramo 7. Nesses casos, é usual "quebrar" o ramo em dois: no ramo superior ficam as folhas de 0 a 4 e no ramo inferior, as folhas de 5 a 9. Assim, fica mais saliente a maior concentração de clientes com pontos entre 70 e 74.

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

b. Temos 30 clientes. Logo,
$Q_2 = \frac{x_{(15)}+x_{(16)}}{4} = 74$,
$Q_1 = x_{(8)} = 72$,
$Q_3 = x_{(23)} = 75$,
$IQ = Q_3 - Q_1 = 75 - 72 = 3$.

c. Veja a **Figura 5.13**. É visível a presença de dois valores discrepantes. Excluindo esses dois valores, a distribuição apresenta uma leve assimetria à esquerda – note que Q_2 está mais próximo de Q_3 do que de Q_1.

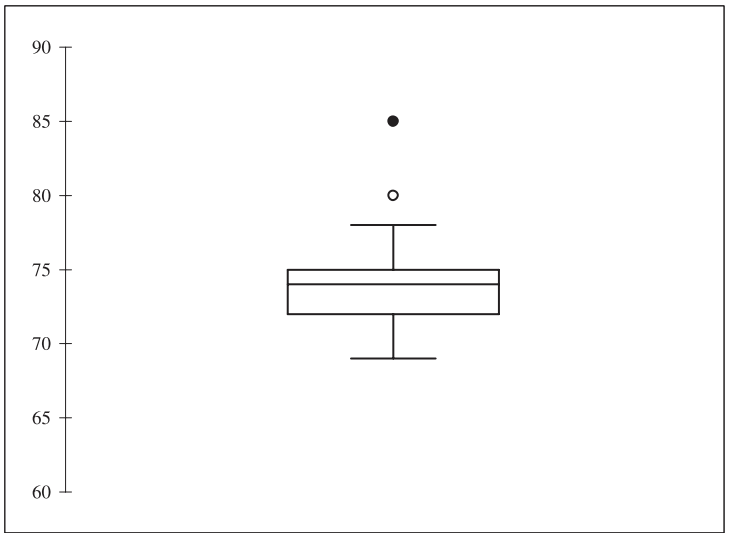

Figura 5.13: Solução do Exercício 5.8.

d. A regra para premiação especial é a regra de valores discrepantes; assim, dois clientes ganharão a garrafa de champagne.