Pedro A. Morettin
Wilton de O. Bussab

ESTATÍSTICA BÁSICA

6ª edição
Revista e atualizada

Editora Saraiva

Capítulo 3

Medidas-Resumo

3.1 Medidas de Posição

Vimos que o resumo de dados por meio de tabelas de freqüências e ramo-e-folhas fornece muito mais informações sobre o comportamento de uma variável do que a própria tabela original de dados. Muitas vezes, queremos resumir ainda mais estes dados, apresentando um ou alguns valores que sejam *representativos* da série toda. Quando usamos um só valor, obtemos uma redução drástica dos dados. Usualmente, emprega-se uma das seguintes medidas de posição (ou localização) central: média, mediana ou moda.

A *moda* é definida como a realização mais freqüente do conjunto de valores observados. Por exemplo, considere a variável *Z*, número de filhos de cada funcionário casado, resumida na Tabela 2.5 do Capítulo 2. Vemos que a moda é 2, correspondente à realização com maior freqüência, 7. Em alguns casos, pode haver mais de uma moda, ou seja, a distribuição dos valores pode ser bimodal, trimodal etc.

A *mediana* é a realização que ocupa a posição central da série de observações, quando estão ordenadas em ordem crescente. Assim, se as cinco observações de uma variável forem 3, 4, 7, 8 e 8, a mediana é o valor 7, correspondendo à terceira observação. Quando o número de observações for par, usa-se como mediana a média aritmética das duas observações centrais. Acrescentando-se o valor 9 à série acima, a mediana será $(7 + 8)/2 = 7{,}5$.

Finalmente, a *média aritmética*, conceito familiar ao leitor, é a soma das observações dividida pelo número delas. Assim, a média aritmética de 3, 4, 7, 8 e 8 é $(3 + 4 + 7 + 8 + 8)/5 = 6$.

Exemplo 3.1. Usando os dados da Tabela 2.5, já encontramos que a moda da variável *Z* é 2. Para a mediana, constatamos que esta também é 2, média aritmética entre a décima e a décima primeira observações. Finalmente, a média aritmética será

$$\frac{4 \times 0 + 5 \times 1 + 7 \times 2 + 3 \times 3 + 5 \times 1}{20} = \frac{33}{20} = 1{,}65.$$

Neste exemplo, as três medidas têm valores próximos e qualquer uma delas pode ser usada como *representativa* da série toda. A média aritmética é, talvez, a medida mais usada. Contudo, ela pode conduzir a erros de interpretação. Em muitas situações, a mediana é uma medida mais adequada. Voltaremos a este assunto mais adiante.

Vamos formalizar os conceitos introduzidos acima. Se $x_1, \ldots, x_n$ são os n valores (distintos ou não) da variável X, a média aritmética, ou simplesmente média, de X pode ser escrita

$$\bar{x} = \frac{x_1 + \ldots + x_n}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i. \tag{3.1}$$

Agora, se tivermos n observações da variável X, das quais n_1 são iguais a x_1, n_2 são iguais a x_2 etc., n_k iguais a x_k, então a média de X pode ser escrita

$$\bar{x} = \frac{n_1 x_1 + n_2 x_2 + \ldots + n_k x_k}{n} = \frac{1}{n} \sum_{i=1}^{k} n_i x_i. \tag{3.2}$$

Se $f_i = n_i / n$ representar a freqüência relativa da observação x_i, então (3.2) também pode ser escrita

$$\bar{x} = \sum_{i=1}^{k} f_i x_i. \tag{3.3}$$

Consideremos, agora, as observações ordenadas em ordem crescente. Vamos denotar a menor observação por $x_{(1)}$, a segunda por $x_{(2)}$, e assim por diante, obtendo-se

$$x_{(1)} \leqslant x_{(2)} \leqslant \ldots \leqslant x_{(n-1)} \leqslant x_{(n)}. \tag{3.4}$$

Por exemplo, se $x_1 = 3$, $x_2 = -2$, $x_3 = 6$, $x_4 = 1$, $x_5 = 3$, então $-2 \leqslant 1 \leqslant 3 \leqslant 3 \leqslant 6$, de modo que $x_{(1)} = -2$, $x_{(2)} = 1$, $x_{(3)} = 3$, $x_{(4)} = 3$ e $x_{(5)} = 6$.

As observações ordenadas como em (3.4) são chamadas *estatísticas de ordem*.

Com esta notação, a mediana da variável X pode ser definida como

$$\text{md}(X) = \begin{cases} x_{\left(\frac{n+1}{2}\right)}, & \text{se } n \text{ ímpar;} \\ \dfrac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}, & \text{se } n \text{ par.} \end{cases} \tag{3.5}$$

Exemplo 3.2. A determinação das medidas de posição para uma variável quantitativa contínua, através de sua distribuição de freqüências, exige aproximações, pois perdemos a informação dos valores das observações. Consideremos a variável S: salário dos 36 funcionários da Companhia MB, agrupados em classes de salários, conforme a Tabela 2.6. Uma aproximação razoável é supor que todos os valores dentro de uma classe tenham seus valores iguais ao ponto médio desta classe. Este procedimento nos deixa na mesma situação do caso discreto, onde as medidas são calculadas usando-se os pares (x_i, n_i) ou (x_i, f_i), como em (3.2) e (3.3).

A moda, mediana e média para os dados da Tabela 2.6 são, respectivamente,

$$\text{mo(S)} \simeq 10{,}00,$$
$$\text{md(S)} \simeq 10{,}00,$$
$$\bar{s} \simeq \frac{10 \times 6{,}00 + 12 \times 10{,}00 + 8 \times 14{,}00 + 5 \times 18{,}00 + 1 \times 22{,}00}{36} = 11{,}22.$$

Observe que colocamos o sinal de $\simeq$ e não de igualdade, pois os valores verdadeiros não são os calculados. Por exemplo, a mediana de S é a média entre as duas observações centrais, quando os dados são ordenados, isto é, 9,80 e 10,53, portanto md(S) = 10,16. Quais são, neste exemplo, a média e moda verdadeiras?

Observe que, para calcular a moda de uma variável, precisamos apenas da distribuição de freqüências (contagem). Já para a mediana necessitamos minimamente ordenar as realizações da variável. Finalmente, a média só pode ser calculada para variáveis quantitativas.

Estas condições limitam bastante o cálculo de medidas-resumos para as variáveis qualitativas. Para as variáveis nominais somente podemos trabalhar com a moda. Para as variáveis ordinais, além da moda, podemos usar também a mediana. Devido a esse fato, iremos apresentar daqui em diante medidas-resumo para variáveis quantitativas, que permitem o uso de operações aritméticas com seus valores.

Exemplo 3.2. (continuação) Retomemos os dados da Companhia MB. A moda para a variável V: região de procedência é mo(V) = outra. Para a variável Y: grau de instrução, temos que mo(Y) = ensino médio e md(Y) = ensino médio.

Veremos, na seção 3.3, que a mediana é uma medida resistente, ao passo que a média não o é, em particular para distribuições muito assimétricas ou contendo valores atípicos. Por outro lado, a média é ótima (num sentido que será discutido no Capítulo 10) se a distribuição dos dados for aproximadamente normal.

Uma outra medida de posição também resistente é a média aparada, definida no Problema 39. Esta medida envolve calcular a média das observações centrais, desprezando-se uma porcentagem das iniciais e finais.

3.2 Medidas de Dispersão

O resumo de um conjunto de dados por uma única medida representativa de posição central esconde toda a informação sobre a variabilidade do conjunto de observações. Por exemplo, suponhamos que cinco grupos de alunos submeteram-se a um teste, obtendo-se as seguintes notas:

grupo A (variável X): 3, 4, 5, 6, 7
grupo B (variável Y): 1, 3, 5, 7, 9
grupo C (variável Z): 5, 5, 5, 5, 5
grupo D (variável W): 3, 5, 5, 7
grupo E (variável V): 3, 5, 5, 6, 6

Vemos que $\bar{x} = \bar{y} = \bar{z} = \bar{w} = \bar{v} = 5{,}0$. A identificação de cada uma destas séries por sua média (5, em todos os casos) nada informa sobre suas diferentes variabilidades. Notamos, então, a conveniência de serem criadas medidas que sumarizem a variabilidade de um conjunto de observações e que nos permita, por exemplo, comparar conjuntos diferentes de valores, como os dados acima, segundo algum critério estabelecido.

Um critério freqüentemente usado para tal fim é aquele que mede a dispersão dos dados em torno de sua média, e duas medidas são as mais usadas: desvio médio e variância. O princípio básico é analisar os desvios das observações em relação à média dessas observações.

Para o grupo A acima os desvios $x_i - \bar{x}$ são: –2, –1, 0, 1, 2. É fácil ver (Problema 14 (a)) que, para *qualquer* conjunto de dados, a soma dos desvios é igual a zero. Nestas condições, a soma dos desvios $\sum_{i=1}^{5} (x_i - \bar{x})$ não é uma boa medida de dispersão para o conjunto A. Duas opções são: (a) considerar o total dos desvios em valor absoluto; (b) considerar o total dos quadrados dos desvios. Para o grupo A teríamos, respectivamente,

$$\sum_{i=1}^{5} |x_i - \bar{x}| = 2 + 1 + 0 + 1 + 2 = 6,$$

$$\sum_{i=1}^{5} (x_i - \bar{x})^2 = 4 + 1 + 0 + 1 + 4 = 10.$$

O uso desses totais pode causar dificuldades quando comparamos conjuntos de dados com números diferentes de observações, como os conjuntos A e D acima. Desse modo, é mais conveniente exprimir as medidas como médias, isto é, o *desvio médio* e a *variância* são definidos por

$$\text{dm}(X) = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}, \tag{3.6}$$

$$\text{var}(X) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}, \tag{3.7}$$

respectivamente. Para o grupo A temos

$$\text{dm}(X) = 6/5 = 1{,}2,$$
$$\text{var}(X) = 10/5 = 2{,}0,$$

enquanto para o grupo D temos

$$\text{dm}(W) = 4/4 = 1{,}0,$$
$$\text{var}(W) = 8/4 = 2{,}0.$$

Podemos dizer, então, que, segundo o desvio médio, o grupo D é mais homogêneo que A, enquanto ambos são igualmente homogêneos, segundo a variância.

Sendo a variância uma medida de dimensão igual ao quadrado da dimensão dos dados (por exemplo, se os dados são expressos em cm, a variância será expressa em cm^2), pode

causar problemas de interpretação. Costuma-se usar, então, o *desvio padrão*, que é definido como a raiz quadrada positiva da variância. Para o grupo A o desvio padrão é

$$\text{dp}(X) = \sqrt{\text{var}\,(X)} = \sqrt{2} = 1{,}41.$$

Ambas as medidas de dispersão (dm e dp) indicam em média qual será o "erro" (desvio) cometido ao tentar substituir cada observação pela medida resumo do conjunto de dados (no caso, a média).

Exemplo 3.3. Vamos calcular as medidas de dispersão acima para a variável Z = número de filhos, resumida na Tabela 2.5. Como vimos no Exemplo 3.1, $\bar{z} = 1{,}65$. Os desvios são $z_i - \bar{z}$: $-1{,}65$; $-0{,}65$; $0{,}35$; $1{,}35$; $3{,}35$. Segue-se que

$$\text{dm}(Z) = \frac{4 \times (1{,}65) + 5 \times (0{,}65) + 7 \times (0{,}35) + 3 \times (1{,}35) + 1 \times (3{,}35)}{20} = 0{,}98.$$

Também,

$$\text{var}(Z) = \frac{4(-1{,}65)^2 + 5(-0{,}65)^2 + 7(0{,}35)^2 + 3(1{,}35)^2 + 1(3{,}35)^2}{20} = 1{,}528.$$

Conseqüentemente, o desvio padrão de Z é

$$\text{dp}(Z) = \sqrt{1{,}528} = 1{,}24.$$

Suponha que observemos n_1 vezes os valores x_1 etc., n_k vezes o valor x_k da variável X. Então,

$$\text{dm}(X) = \frac{\sum_{i=1}^{k} n_i |x_i - \bar{x}|}{n} = \sum_{i=1}^{k} f_i |x_i - \bar{x}|, \tag{3.8}$$

$$\text{var}(X) = \frac{\sum_{i=1}^{k} n_i (x_i - \bar{x})^2}{n} = \sum_{i=1}^{k} f_i (x_i - \bar{x})^2, \tag{3.9}$$

$$\text{dp}(X) = \sqrt{\text{var}\,(X)}. \tag{3.10}$$

O cálculo (aproximado) das medidas de dispersão no caso de variáveis contínuas, agrupadas em classes, pode ser feito de modo análogo àquele usado para encontrar a média no Exemplo 2.2.

Exemplo 3.4. Consideremos a variável S = salário. A média encontrada no Exemplo 3.2 foi $s = 11{,}22$. Com os dados da Tabela 2.6 e usando (3.9) encontramos

$$\text{var}(S) \simeq [10(6,\,00 - 11{,}22)^2 + 12(10{,}00 - 11{,}22)^2 + 8(14 - 11{,}22)^2 + 5(18{,}00 - 11{,}22)^2 + 1(22{,}00 - 11{,}22)^2]/36 = 19{,}40$$

e, portanto,

$$\text{dp}(S) \simeq \sqrt{19{,}40} = 4{,}40.$$

É fácil ver que $\text{dm}(S) \simeq 3{,}72$.

Veremos, mais tarde, que a variância de uma amostra será calculada usando-se o denominador $n - 1$, em vez de n. A justificativa será dada naquele capítulo, mas para grandes amostras pouca diferença fará o uso de um ou outro denominador.

Tanto a variância como o desvio médio são medidas de dispersão calculadas em relação à média das observações. Assim como a média, a variância (ou o desvio padrão) é uma boa medida se a distribuição dos dados for aproximadamente normal. O desvio médio é mais resistente que o desvio padrão, no sentido a ser estudado na seção seguinte.

Poderíamos considerar uma medida que seja calculada em relação à mediana. O desvio absoluto mediano é um exemplo e é mais resistente que o desvio padrão. Veja o Problema 41.

Usando o Problema 14 (b), uma maneira computacionalmente mais eficiente de calcular a variância é

$$\text{var}(X) = \frac{\sum_{i=1}^{n} x_i^2}{n} - \bar{x}^2, \tag{3.11}$$

e, no caso de observações repetidas,

$$\text{var}(X) = \sum_{i=1}^{k} f_i x_i^2 - \bar{x}^2. \tag{3.12}$$

Problemas

1. Quer se estudar o número de erros de impressão de um livro. Para isso escolheu-se uma amostra de 50 páginas, encontrando-se o número de erros por página da tabela abaixo.
   (a) Qual o número médio de erros por página?
   (b) E o número mediano?
   (c) Qual é o desvio padrão?
   (d) Faça uma representação gráfica para a distribuição.
   (e) Se o livro tem 500 páginas, qual o número total de erros esperado no livro?

Erros	Freqüência
0	25
1	20
2	3
3	1
4	1

2. As taxas de juros recebidas por 10 ações durante um certo período foram (medidas em porcentagem) 2,59; 2,64; 2,60; 2,62; 2,57; 2,55; 2,61; 2,50; 2,63; 2,64. Calcule a média, a mediana e o desvio padrão.

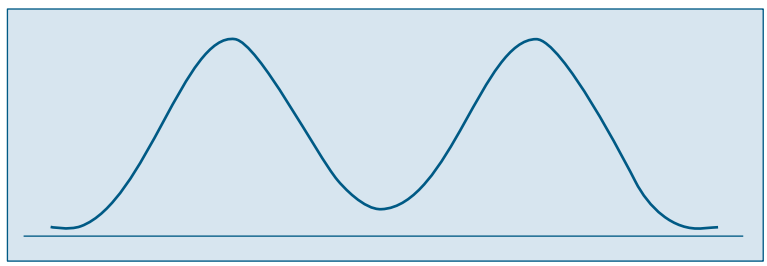
3. Para facilitar um projeto de ampliação da rede de esgoto de uma certa região de uma cidade, as autoridades tomaram uma amostra de tamanho 50 dos 270 quarteirões que compõem a região, e foram encontrados os seguintes números de casas por quarteirão:

2	2	3	10	13	14	15	15	16	16
18	18	20	21	22	22	23	24	25	25
26	27	29	29	30	32	36	42	44	45
45	46	48	52	58	59	61	61	61	65
66	66	68	75	78	80	89	90	92	97

(a) Use cinco intervalos e construa um histograma.
(b) Determine uma medida de posição central e uma medida de dispersão.

4. (a) Dê uma situação prática onde você acha que a mediana é uma medida mais apropriada do que a média.
(b) Esboce um histograma onde a média e a mediana coincidem. Existe alguma classe de histogramas onde isso sempre acontece?
(c) Esboce os histogramas de três variáveis (X, Y e Z) com a mesma média aritmética, mas com as variâncias ordenadas em ordem crescente.

5. Suponha que a variável de interesse tenha a distribuição como na figura abaixo.



Você acha que a média é uma boa medida de posição? E a mediana? Justifique.

6. Numa pesquisa realizada com 100 famílias, levantaram-se as seguintes informações:

Número de filhos	0	1	2	3	4	5	mais que 5
Freqüência de famílias	17	20	28	19	7	4	5

(a) Qual a mediana do número de filhos?
(b) E a moda?
(c) Que problemas você enfrentaria para calcular a média? Faça alguma suposição e encontre-a.

3.3 Quantis Empíricos

Tanto a média como o desvio padrão podem não ser medidas adequadas para representar um conjunto de dados, pois:

(a) são afetados, de forma exagerada, por valores extremos;
(b) apenas com estes dois valores não temos idéia da simetria ou assimetria da distribuição dos dados.

Para contornar esses fatos, outras medidas têm de ser consideradas.

Vimos que a mediana é um valor que deixa metade dos dados abaixo dela e metade acima (ver fórmula (3.5)). De modo geral, podemos definir uma medida, chamada *quantil de ordem p* ou *p-quantil*, indicada por $q(p)$, onde p é uma proporção qualquer, $0 < p < 1$, tal que $100p\%$ das observações sejam menores do que $q(p)$.

Indicamos, abaixo, alguns quantis e seus nomes particulares.

$q(0{,}25) = q_1$: 1º Quartil = 25º Percentil
$q(0{,}50) = q_2$: Mediana = 2º Quartil = 50º Percentil
$q(0{,}75) = q_3$: 3º Quartil = 75º Percentil
$q(0{,}40)$: 4º Decil
$q(0{,}95)$: 95º Percentil

Dependendo do valor de p, há dificuldades ao se calcular os quantis. Isso é ilustrado no exemplo a seguir.

Exemplo 3.5. Suponha que tenhamos os seguintes valores de uma variável X:

$$15,\ 5,\ 3,\ 8,\ 10,\ 2,\ 7,\ 11,\ 12.$$

Ordenando os valores, obtemos as estatísticas de ordem $x_{(1)} = 2$, $x_{(2)} = 3$,..., $x_{(9)} = 15$, ou seja, teremos

$$2 < 3 < 5 < 7 < 8 < 10 < 11 < 12 < 15.$$

Usando a definição de mediana dada, teremos que md $= q(0{,}5) = q_2 = x_{(5)} = 8$. Suponha que queiramos calcular os dois outros quartis, q_1 e q_3. A idéia é dividir os dados em quatro partes:

$$2 \quad \mathbf{3} \quad \mathbf{5} \quad 7 \quad \mathbf{8} \quad 10 \quad \mathbf{11} \quad \mathbf{12} \quad 15$$

Uma possibilidade razoável é, então, considerar a mediana dos primeiros quatro valores para obter q_1, ou seja,

$$q_1 = \frac{3 + 5}{2} = 4,$$

e a mediana dos últimos quatro valores para obter q_3, ou seja,

$$q_3 = \frac{11 + 12}{2} = 11{,}5.$$

Obtemos, então, a sequência

$$2 \quad 3 \quad (\mathbf{4}) \quad 5 \quad 7 \quad (\mathbf{8}) \quad 10 \quad 11 \quad (\mathbf{11{,}5}) \quad \mathbf{12} \quad 15$$

Observe que a média dos $n = 9$ valores é $\overline{x} = 8{,}1$, próximo à mediana.

Exemplo 3.5. (continuação). Acrescentemos, agora, o valor 67 à lista de nove valores do Exemplo 3.5, obtendo-se agora os $n = 10$ valores ordenados:

$$2 < 3 < 5 < 7 < 8 < 10 < 11 < 12 < 15 < 67$$

Agora, $\bar{x} = 14$, enquanto que a mediana fica

$$q_2 = \frac{x_{(5)} + x_{(6)}}{2} = 9,$$

que está próxima da mediana dos nove valores originais, mas ambas (8 e 9) relativamente longes de $\bar{x}$. Dizemos que a mediana é *resistente* (ou *robusta*), no sentido que que ela não é muito afetada pelo valor discrepante (ou atípico) 67.

Para calcular q_1 e q_3 para este novo conjunto de valores, considere-os assim dispostos:

2 3 **5** 7 8 **9** 10 11 **12** 15 67

de modo que $q_1 = 5$ e $q_3 = 12$.

Obtemos, então os dados separados em 4 partes por q_1, q_2 e q_3:

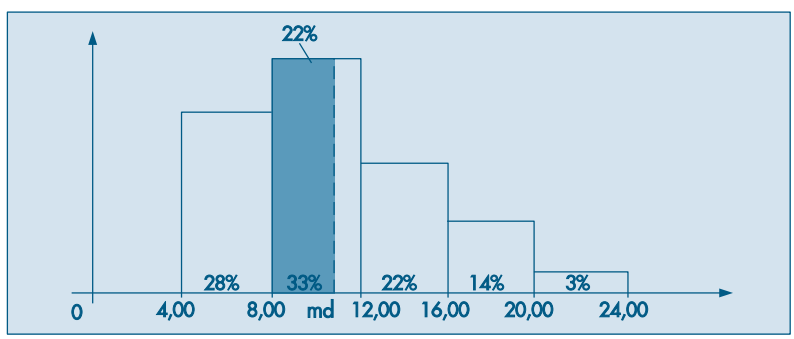
2 3 (**5**) 7 8 (**9**) 10 11 (**12**) 15 67

Suponha, agora, que queiramos calcular $q(0{,}20)$, ou seja, aquele valor que deixa 20% dos dados à sua esquerda, para o conjunto original de $n = 9$ valores de X. Como 20% das observações correspondem a 1,8 observações, qual valor devemos tomar como $q(0, 20)$? O valor 3, que é a segunda observação ordenada, ou 5, ou a média de 3 e 5? Se adotarmos esta última solução, então $q(0, 20) = q(0, 25) = q_1$, o que pode parecer não razoável.

Para responder a esta questão, temos que definir quantil de uma sequência de valores de uma variável de modo apropriado. Isto está feito no Problema 17.

Se os dados estiverem agrupados em classes, podemos obter os quantis usando o histograma. Por exemplo, para obter a mediana, sabemos que ela deve corresponder ao valor da abscissa que divide a área do histograma em duas partes iguais (50% para cada lado). Então, usando argumentos geométricos, podemos encontrar um ponto, satisfazendo essa propriedade. Vejamos como proceder através de um exemplo.

Exemplo 3.6. Vamos repetir abaixo a Figura 2.7, que é o histograma da variável S = salário dos empregados da Companhia MB.



Devemos localizar o ponto das abscissas que divide o histograma ao meio. A área do primeiro retângulo corresponde a 28% do total, os dois primeiros a 61%; portanto, a mediana md é algum número situado entre 8,00 e 12,00. Ou melhor, a mediana irá corresponder ao valor md no segundo retângulo, cuja área do retângulo de base 8,00 ⊢ md é a mesma altura que o retângulo de base 8,00 ⊢ 12,00 seja 22% (28% do primeiro retângulo mais 22% do segundo, perfazendo os 50%). Consulte a figura para melhor compreensão. Pela proporcionalidade entre a área e a base do retângulo, temos:

$$\frac{12{,}00 - 8{,}00}{33\%} = \frac{\text{md} - 8{,}00}{22\%}$$

ou

$$\text{md} - 8{,}00 = \frac{22\%}{33\%} \cdot 4{,}00,$$

logo

$$\text{md} = 8{,}00 + 2{,}67 = 10{,}67,$$

que é uma expressão mais precisa para a mediana do que a mediana bruta encontrada anteriormente.

O cálculo dos quantis pode ser feito de modo análogo ao cálculo da mediana, usando argumentos geométricos no histograma. Vejamos a determinação de alguns quantis, usando os dados do último exemplo.

(a) $q(0{,}25)$: Verificamos que $q(0{,}25)$ deve estar na primeira classe, pois a proporção no primeiro retângulo é 0,28. Logo,

$$\frac{q(0{,}25) - 4{,}00}{25\%} = \frac{8{,}00 - 4{,}00}{28\%},$$

e então

$$q(0{,}25) = 4{,}00 + \frac{25}{28}\ 4{,}00 = 7{,}57.$$

(b) $q(0{,}95)$: Analisando a soma acumulada das proporções, verificamos que este quantil deve pertencer à quarta classe, e que nesse retângulo devemos achar a parte correspondente a 12%, pois a soma acumulada até a classe anterior é 83%, faltando 12% para atingirmos os 95%. Portanto,

$$\frac{q(0{,}95) - 16{,}00}{12\%} = \frac{20{,}00 - 16{,}00}{14\%},$$

logo

$$q(0{,}95) = 16{,}00 + \frac{12}{14} \times 4 = 19{,}43.$$

(c) $q(0{,}75)$: De modo análogo, concluímos que o terceiro quantil deve pertencer ao intervalo $12{,}00 \vdash 16{,}00$, portanto

$$\frac{q(0{,}75) - 12{,}00}{14\%} = \frac{16{,}00 - 12{,}00}{22\%}$$

e

$$q(0{,}75) = 14{,}55.$$

Uma medida de dispersão alternativa ao desvio padrão é a *distância interquartil*, definida como a diferença entre o terceiro e primeiro quartis, ou seja,

$$d_q = q_3 - q_1. \tag{3.13}$$

Para o Exemplo 3.5, temos $q_1 = 4$, $q_3 = 11{,}5$, de modo que $d_q = 7{,}5$. Para um cálculo mais preciso, veja o Problema 17. Lá obtemos $q_1 = 4{,}5$, $q_3 = 11{,}25$, logo $d_q = 6{,}75$.

Os quartis $q(0{,}25) = q_1$, $q(0{,}5) = 92$ e $9(0{,}75) = 93$ são medidas de localização resistentes de uma distribuição.

Dizemos que uma medida de localização ou dispersão é resistente quando for pouco afetada por mudanças de uma pequena porção dos dados. A mediana é uma medida resistente, ao passo que a média não o é. Para ilustrar este fato, considere as populações dos 30 municípios do Brasil, considerados acima. Se descartarmos Rio de Janeiro e São Paulo, a média das populações dos 28 municípios restantes é 100,6 e a mediana é 82,1. Para todos os dados, a média pasa a ser 145,4, ao passo que a mediana será 84,3. Note que a média aumentou bastante, influenciada que foi pelos dois valores maiores, que são muito discrepantes da maioria dos dados. Mas a mediana variou pouco. O desvio padrão também não é uma medida resistente. Verifique como este varia para este exemplo dos municípios.

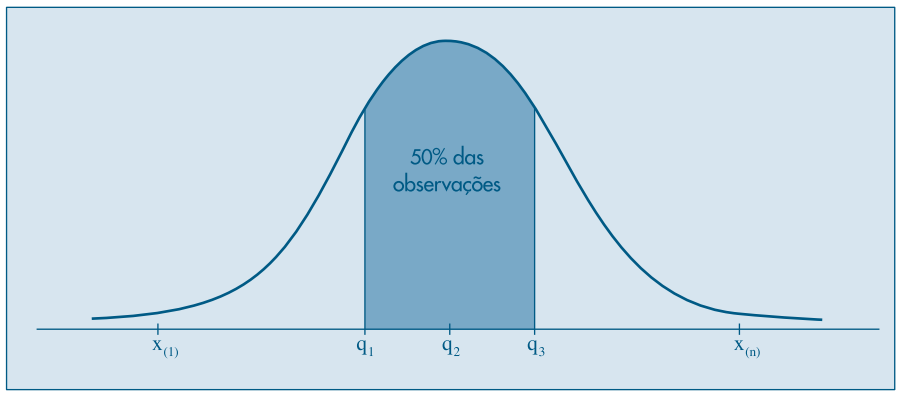
Os cinco valores, $x_{(1)}$, q_1, q_2, q_3 e $x_{(n)}$ são importantes para se ter uma boa idéia da assimetria da distribuição dos dados. Para uma distribuição simétrica ou aproximadamente simétrica, deveríamos ter:

(a) $q_2 - x_{(1)} \simeq x_{(n)} - q_2$;
(b) $q_2 - q_1 \simeq q_3 - q_2$;
(c) $q_1 - x_{(1)} \simeq x_{(n)} - q_3$;
(d) distâncias entre mediana e q_1, q_3 menores do que distâncias entre os extremos e q_1, q_3.

A diferença $q_2 - x_{(1)}$ é chamada *dispersão inferior* e $x_{(n)} - q_2$ é a *dispersão superior*. A condição (a) nos diz que estas duas dispersões devem ser aproximadamente iguais, para uma distribuição aproximadamente simétrica.

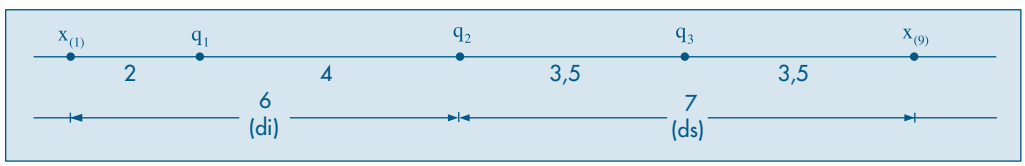
A Figura 3.1 ilustra estes fatos para a chamada *distribuição normal* ou *gaussiana*.

Figura 3.1: Uma distribuição simétrica: normal ou gaussiana.



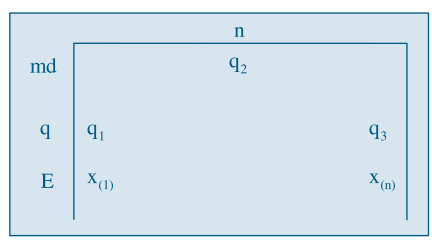
Na Figura 3.2 temos ilustradas estas cinco medidas para os $n = 9$ valores do Exemplo 3.5.

Figura 3.2: Quantis e distâncias para o Exemplo 3.5.



As cinco estatísticas de ordem consideradas acima podem ser representadas esquematicamente como na Figura 3.3, onde também incorporamos o número de observações, *n*. Representamos a mediana por *md*, os quartis por *q* e os extremos por *E*. Podemos ir além, considerando os chamados *oitavos*, ou seja, o primeiro oitavo, que corresponde a $q(0{,}125)$, o sétimo oitavo, que corresponde a $q(0{,}875)$ etc. Teríamos, então, sete números para representar a distribuição dos dados. Em geral, podemos considerar as chamadas *letras-resumos*, descendo aos *dezesseis-avos*, *trinta e dois-avos* etc. Para detalhes, ver Hoaglin, Mosteller and Tukey(1983).

Figura 3.3: Esquema dos cinco números.



Exemplo 3.7. Os aplicativos SPlus e Minitab, assim como a planilha Excel, possuem ferramentas que geram as principais medidas descritas nesse capítulo e outras. Por exemplo, o comando *describe* do Minitab, usado para as populações dos municípios brasileiros produz a saída do Quadro 3.1.

Quadro 3.1. Medidas-resumo para o CD-Municípios. Minitab.

```
MTB > Describe C1.
Descriptive Statistics
Variable    N     Mean   Median  Tr mean   StDev  SE Mean
C1         30    145.4     84.3    104.7   186.6     34.1
Variable   Min     Max       Q1       Q3
C1        46.3   988.8     63.5    139.7
```

Aqui, temos $N = 30$ dados, a média é 145,4, a mediana 84,3, o desvio padrão 186,6, o menor valor 46,3, o maior valor 988,8, o primeiro quartil 63,5 e o terceiro quartil 139,7. Além desses valores, o resumo traz a *média aparada* (*trimmed mean*) e o erro padrão da média, a ser tratado no Capítulo 11. Esse é dado por $S/\sqrt{n} = 186{,}6/\sqrt{30} = 34{,}1$.

O comando *summary* do SPlus produz a saída do Quadro 3.2 para os mesmos dados. Note a diferença no cálculo dos quantis $q(0{,}25)$ e $q(0{,}75)$. Conclui-se que é necessário saber como cada programa efetua o cálculo de determinada estatística, para poder reportá-lo.

Quadro 3.2. Medidas-resumo para o CD-Municípios. SPlus.

```
> summary (munic)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   46.3   64.48    84.3   145.4   134.3   988.8
```

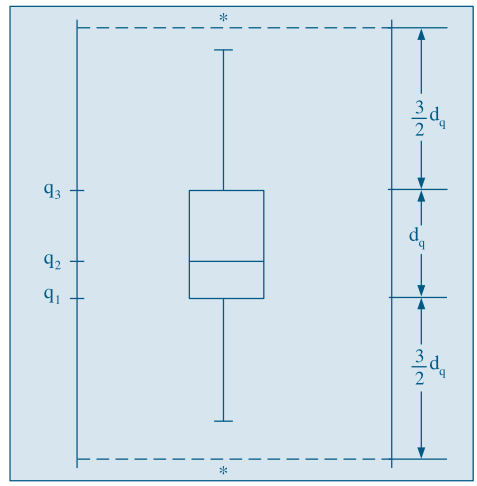
Problemas

7. Obtenha o esquema dos cinco números para os dados do Problema 3. Calcule o intervalo interquartil e as dispersões inferior e superior. Baseado nessas medidas, verifique se a forma da distribuição dos dados é normal.
8. Refaça o problema anterior, utilizando desta vez os dados do Problema 5 do Capítulo 2.
9. Obter os três quartis, $q(0{,}1)$ e $q(0{,}90)$ para os dados do Problema 3.
10. Para a variável *população urbana* do CD-Brasil, obtenha $q(0{,}10)$, $q(0{,}25)$, $q(0{,}50)$, $q(0{,}75)$, $q(0{,}80)$ e $q(0{,}95)$.

3.4 Box Plots

A informação contida no esquema dos cinco números da Figura 3.3 pode ser traduzida graficamente num diagrama, ilustrado na Figura 3.4, que chamaremos de *box plot*. Murteira (1993) usa o termo "caixa-de-bigodes".

Figura 3.4: *Box Plot.*

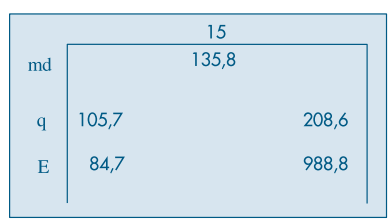


Para construir este diagrama, consideremos um retângulo onde estão representados a mediana e os quartis. A partir do retângulo, para cima, segue uma linha até o ponto mais remoto que não exceda LS = $q_3 + (1{,}5)d_q$, chamado *limite superior*. De modo similar, da parte inferior do retângulo, para baixo, segue uma linha até o ponto mais remoto que não seja menor do que LI = $q_1 - (1{,}5)d_q$, chamado *limite inferior*. Os valores compreendidos entre esses dois limites são chamados *valores adjacentes*. As observações que estiverem acima do limite superior ou abaixo do limite inferior estabelecidos serão chamadas *pontos exteriores* e representadas por asteriscos. Essas são observações destoantes das demais e podem ou não ser o que chamamos de *outliers* ou *valores atípicos*.

O *box plot* dá uma idéia da posição, dispersão, assimetria, caudas e dados discrepantes. A posição central é dada pela mediana e a dispersão por d_q. As posições relativas de q_1, q_2, q_3 dão uma noção da assimetria da distribuição. Os comprimentos das caudas são dados pelas linhas que vão do retângulo aos valores remotos e pelos valores atípicos.

Exemplo 3.8. Retomemos o exemplo dos 15 maiores municípios do Brasil, ordenados pelas populações. Usando o procedimento do Problema 17 (veja também o Problema 18), obtemos $q_1 = 105{,}7$, $q_2 = 135{,}8$, $q_3 = 208{,}6$. O diagrama para os cinco números $x_{(1)}$, q_1, $q_2 = md$, q_3, $x_{(15)}$ está na Figura 3.5 abaixo.

Figura 3.5: Esquema dos cinco números para o Exemplo 3.8.

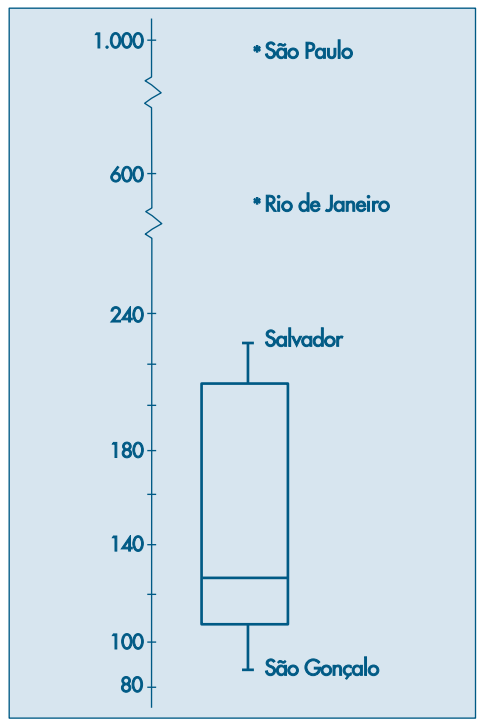


Temos que

$$\text{LI} = q_1 - (1{,}5)d_q = 105{,}7 - (1{,}5)\ (102{,}9) = -48{,}7,$$
$$\text{LS} = q_3 + (1{,}5)d_q = 208{,}6 + (1{,}5)\ (102{,}9) = 362{,}9.$$

Então, as cidades com populações acima de 3.629.000 habitantes são pontos exteriores, ou seja, Rio de Janeiro e São Paulo. O *box plot* correspondente está na Figura 3.6. Vemos que os dados têm uma distribuição assimétrica à direita, com 13 valores concentrados entre 80 e 230 e duas observações discrepantes, bastante afastadas do corpo principal dos dados.

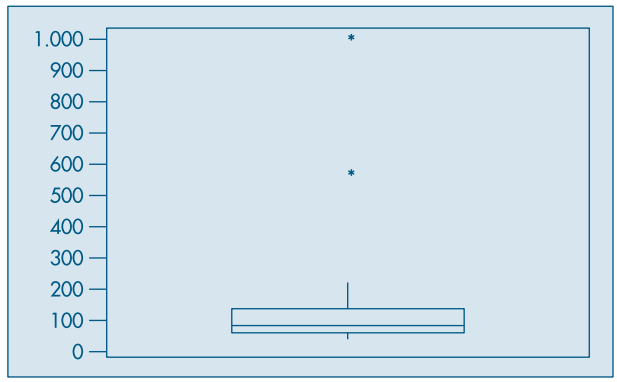
Figura 3.6: *Box plot* para os quinze maiores municípios do Brasil.



Do ponto de vista estatístico, um *outlier* pode ser produto de um erro de observação ou de arredondamento. No exemplo acima, as populações de São Paulo e Rio de Janeiro não são *outliers* neste sentido, pois elas representam dois valores realmente muito diferentes dos demais. Daí, usarmos o nome pontos (ou valores) exteriores. Contudo, na prática, estas duas denominações são freqüentemente usadas com o mesmo significado: observações fora de lugar, discrepantes ou atípicas.

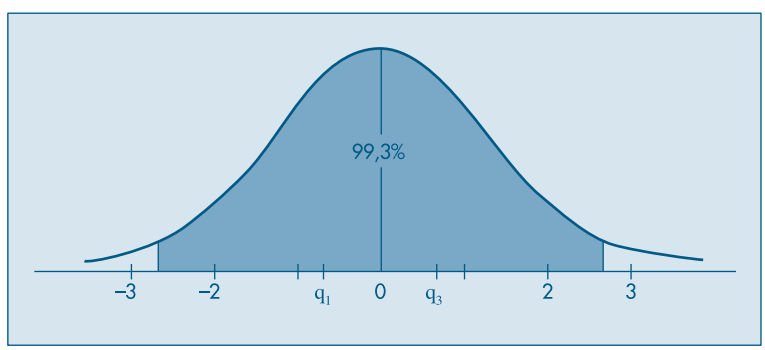
A Figura 3.7 mostra o *box plot* para as populações dos trinta municípios brasileiros, feito com o Minitab.

Figura 3.7: *Box plot* com Minitab.



A justificativa para usarmos os limites acima, LI = $q_1 - (1{,}5)d_q$ e LS = $q_3 + (1{,}5)d_q$, para definir as observações atípicas é a seguinte: considere uma curva normal com média zero e, portanto, com mediana zero. É fácil verificar (veja o Capítulo 7 e Tabela III) que $q_1 = -0{,}6745$, $q_2 = 0$, $q_3 = 0{,}6745$ e portanto $d_q = 1{,}349$. Segue-se que os limites são LI = –2,698 e LS = 2,698. A área entre estes dois valores, embaixo da curva normal, é 0,993, ou seja, 99,3% da distribuição está entre estes dois valores. Isto é, para dados com uma distribuição normal, os pontos exteriores constituirão cerca de 0,7% da distribuição. Veja a Figura 3.8.

Figura 3.8: Área sob a curva normal entre LI e LS.



Problemas

11. Construa o *box plot* para os dados do Exemplo 2.3, Capítulo 2. O que você pode concluir a respeito da distribuição?
12. Refaça a questão anterior com os dados do Problema 3 deste capítulo.
13. Faça um *box plot* para o Problema 10. Comente sobre a simetria, caudas e presença de valores atípicos.

3.5 Gráficos de Simetria

Os quantis podem ser úteis para se verificar se a distribuição dos dados é simétrica (ou aproximadamente simétrica).

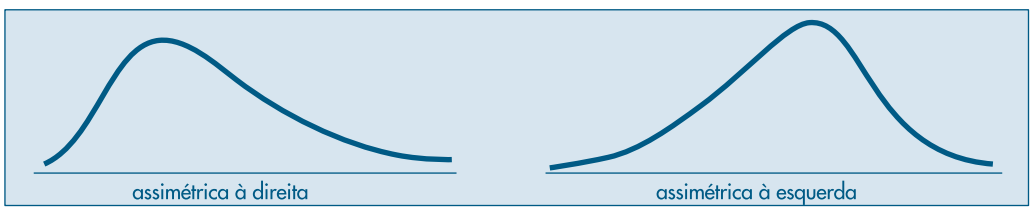
Se um conjunto de observações for perfeitamente simétrico devemos ter

$$q(0{,}5) - x_{(i)} = x_{(n+1-i)} - q(0{,}5), \tag{3.14}$$

onde $i = 1, 2, ..., n/2$, se n for par e $i = 1, 2, ..., (n + 1)/2$, se n for ímpar.

Pela relação (3.14), vemos que, se os quantis da direita estão mais afastados da mediana, do que os da esquerda, os dados serão *assimétricos à direita*. Se ocorrer o contrário, os dados serão *assimétricos à esquerda*. A Figura 3.9 ilustra essas duas situações.

Figura 3.9: Distribuições assimétricas.

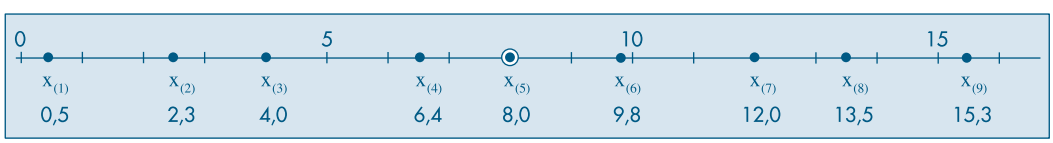


Para os dados do Exemplo 3.8, vemos que as observações são assimétricas à direita. Em geral, esse tipo de situação ocorre com dados positivos.

Podemos fazer um *gráfico de simetria*, usando a identidade (3.14). Chamando de u_i o primeiro membro e de v_i o segundo membro, fazendo-se um gráfico cartesiano, com os u_i's como abscissas e os v_i's como ordenadas, se os dados forem aproximadamente simétricos, os pares (u_i, v_i) estarão dispersos ao redor da reta $v = u$.

Exemplo 3.9. Considere os dados que, dispostos em ordem crescente, ficam representados no eixo real como na Figura 3.10.

Figura 3.10: Dados aproximadamente simétricos.

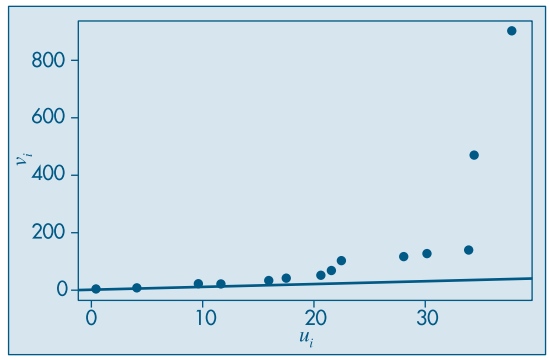


Esses dados são aproximadamente simétricos, pois como $q_2 = 8$, $u_i = q_2 - x_{(i)}$, $v_i = x_{(n+1-i)} - q_2$, teremos:

$$u_1 = 8{,}0 - 0{,}5 = 7{,}5,\ v_1 = 15{,}3 - 8{,}0 = 7{,}3,$$
$$u_2 = 8{,}0 - 2{,}3 = 5{,}7,\ v_2 = 13{,}5 - 8{,}0 = 5{,}5,$$
$$u_3 = 8{,}0 - 4{,}0 = 4{,}0,\ v_3 = 12{,}0 - 8{,}0 = 4{,}0,$$
$$u_4 = 8{,}0 - 6{,}4 = 1{,}6,\ v_4 = 9{,}8 - 8{,}0 = 1{,}8.$$

A Figura 3.11 mostra o gráfico de simetria para as populações dos trinta municípios do Brasil. Vemos que a maioria dos pontos estão acima da reta $v = u$, mostrando a assimetria à direita da distribuição dos valores. Nessa figura, vemos destacados os pontos correspondentes a Rio de Janeiro e São Paulo.

Figura 3.11: Gráfico de simetria para o CD-Municípios.



3.6 Transformações

Vários procedimentos estatísticos são baseados na suposição de que os dados provêm de uma distribuição normal (em forma de sino) ou então mais ou menos simétrica. Mas, em muitas situações de interesse prático, a distribuição dos dados da amostra é assimétrica e pode conter valores atípicos, como vimos em exemplos anteriores.

Se quisermos utilizar tais procedimentos, o que se propõe é efetuar uma transformação das observações, de modo a se obter uma distribuição mais simétrica e próxima da normal. Uma família de transformações freqüentemente utilizada é

$$x^{(p)} = \begin{cases} x^p, & \text{se } p > 0 \\ \ell n(x), & \text{se } p = 0 \\ -x^p, & \text{se } p < 0. \end{cases} \qquad (3.15)$$

Normalmente, o que se faz é experimentar valores de p na seqüência

$$\ldots, -3, -2, -1, -1/2, -1/3, -1/4, 0, 1/4, 1/3, 1/2, 1, 2, 3, \ldots$$

e para cada valor de p obtemos gráficos apropriados (histogramas, desenhos esquemáticos etc.) para os dados originais e transformados, de modo a escolhermos o valor mais adequado de p.

Vimos que, para dados positivos, a distribuição dos dados é usualmente assimétrica à direita. Para essas distribuições, a transformação acima com $0 < p < 1$ é apropriada, pois valores grandes de x decrescem mais, relativamente a valores pequenos. Para distribuições assimétricas à esquerda, tome $p > 1$.

Exemplo 3.10. Consideremos os dados das populações do CD-Municípios e tomemos alguns valores de p: 0, 1/4, 1/3, 1/2. Na Figura 3.12 temos os histogramas para os dados transformados e, na Figura 3.13, os respectivos *box plots*. Vemos que $p = 0$ (transformação logarítmica) e $p = 1/3$ (transformação raiz cúbica) fornecem distribuições mais próximas de uma distribuição simétrica.

Figura 3.12: Histogramas para os dados transformados. CD-Municípios.

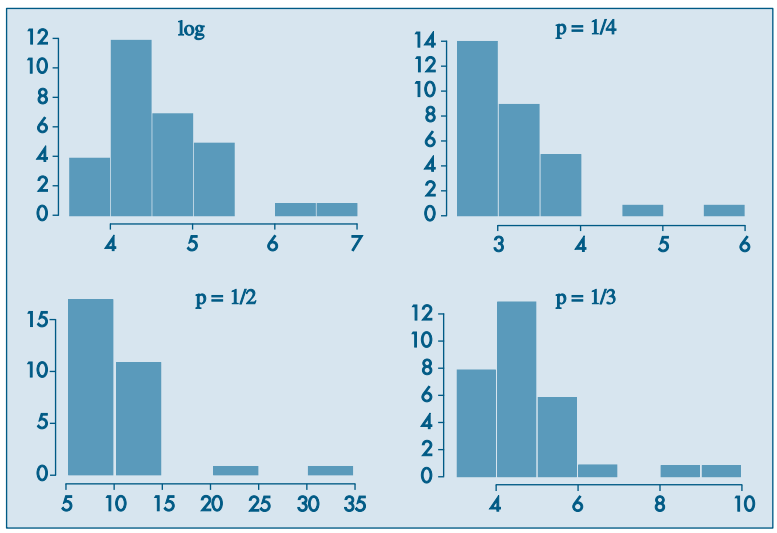
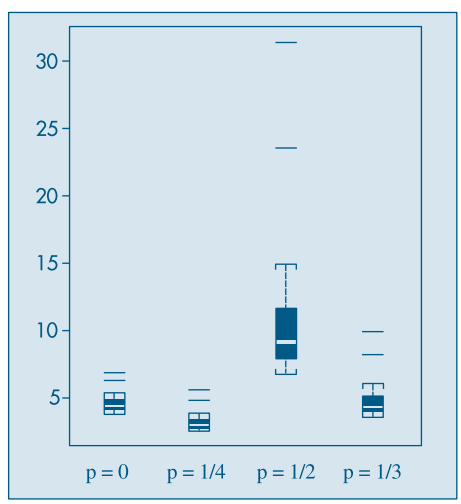


Figura 3.13: *Box plots* para os dados transformados. CD-Municípios. SPlus.



Capítulo 4

Análise Bidimensional

4.1 Introdução

Até agora vimos como organizar e resumir informações pertinentes a uma única variável (ou a um conjunto de dados), mas freqüentemente estamos interessados em analisar o comportamento conjunto de duas ou mais variáveis aleatórias. Os dados aparecem na forma de uma matriz, usualmente com as colunas indicando as variáveis e as linhas os indivíduos (ou elementos). A Tabela 4.1 mostra a notação de uma matriz com p variáveis X_1, X_2, ..., X_p e n indivíduos, totalizando np dados. A Tabela 2.1, com os dados hipotéticos da Companhia MB, é uma ilustração numérica de uma matriz 36 × 7.

O principal objetivo das análises nessa situação é explorar relações (similaridades) entre as colunas, ou algumas vezes entre as linhas. Como no caso de apenas uma variável que estudamos, a *distribuição conjunta* das freqüências será um instrumento poderoso para a compreensão do comportamento dos dados.

Neste capítulo iremos nos deter no caso de duas variáveis ou dois conjuntos de dados. Na seção 4.8 daremos dois exemplos do caso de três variáveis.

Tabela 4.1: Tabela de dados.

Indivíduo	Variável					
	X_1	X_2	...	X_j	...	X_p
1	x_{11}	x_{12}	...	x_{1j}	...	x_{1p}
2	x_{21}	x_{22}	...	x_{2j}	...	x_{2p}
⋮	⋮	⋮		⋮		⋮
i	x_{i1}	x_{i2}	...	x_{ij}	...	x_{ip}
⋮	⋮	⋮		⋮		⋮
n	x_{n1}	x_{n2}	...	x_{nj}	...	x_{np}

Em algumas situações, podemos ter dois (ou mais) conjuntos de dados provenientes da observação da mesma variável. Por exemplo, podemos ter um conjunto de dados $\{x_1, ..., x_n\}$, que são as temperaturas na cidade A, durante n meses, e outro conjunto de dados $\{y_1, ..., y_n\}$,

que são as temperaturas da cidade B, nos mesmos meses. Para efeito de análise, podemos considerar que o primeiro conjunto são observações da variável *X*: temperatura na cidade A, enquanto o segundo conjunto são observações da variável *Y*: temperatura na cidade B. Este é o caso do CD-Temperaturas. Também poderíamos usar uma variável *X* para indicar a temperatura e outra variável, *L*, para indicar se a observação pertence à região A ou B. Na Tabela 2.1 podemos estar interessados em comparar os salários dos casados e solteiros. Uma reordenação dos dados poderia colocar os casados nas primeiras posições e os solteiros nas últimas, e nosso objetivo passaria a ser comparar, na coluna de salários (variável *S*), o comportamento de *S* na parte superior com a inferior. A escolha da apresentação de um ou outro modo será ditada principalmente pelo interesse e técnicas de análise à disposição do pesquisador.

No CD-Brasil temos cinco variáveis: superfície, população urbana, rural e total e densidade populacional. No CD-Poluição temos quatro variáveis: quantidade de monóxido de carbono, ozônio, temperatura do ar e umidade relativa do ar.

Quando consideramos duas variáveis (ou dois conjuntos de dados), podemos ter três situações:

(a) as duas variáveis são qualitativas;
(b) as duas variáveis são quantitativas; e
(c) uma variável é qualitativa e outra é quantitativa.

As técnicas de análise de dados nas três situações são diferentes. Quando as variáveis são qualitativas, os dados são resumidos em *tabelas de dupla entrada* (*ou de contingência*), onde aparecerão as freqüências absolutas ou contagens de indivíduos que pertencem simultaneamente a categorias de uma e outra variável. Quando as duas variáveis são quantitativas, as observações são provenientes de mensurações, e técnicas como gráficos de dispersão ou de quantis são apropriadas. Quando temos uma variável qualitativa e outra quantitativa, em geral analisamos o que acontece com a variável quantitativa quando os dados são categorizados de acordo com os diversos atributos da variável qualitativa. Mas podemos ter também o caso de duas variáveis quantitativas agrupadas em classes. Por exemplo, podemos querer analisar a associação entre renda e consumo de certo número de famílias e, para isso, agrupamos as famílias em classes de rendas e classes de consumo. Desse modo, recaímos novamente numa tabela de dupla entrada.

Contudo, em todas as situações, o objetivo é encontrar as possíveis relações ou associações entre as duas variáveis. Essas relações podem ser detectadas por meio de métodos gráficos e medidas numéricas. Para efeitos práticos (e a razão ficará mais clara após o estudo de probabilidades), iremos entender a existência de associação como a *mudança* de opinião sobre o comportamento de uma variável na presença ou não de informação sobre a segunda variável. Ilustrando: existe relação entre a altura de pessoas e o sexo (homem ou mulher) em dada comunidade? Pode-se fazer uma primeira pergunta: qual a freqüência esperada de uma pessoa dessa população ter, digamos, mais de 170 cm

de altura? E também uma segunda: qual a freqüência esperada de uma mulher (ou homem) ter mais de 170 cm de altura? Se a resposta para as duas perguntas for a mesma, diríamos que *não há* associação entre as variáveis altura e sexo. Porém, se as respostas forem diferentes, isso significa uma provável associação, e devemos incorporar esse conhecimento para melhorar o entendimento sobre os comportamentos das variáveis. No exemplo em questão, você acha que existe associação entre as variáveis?

4.2 Variáveis Qualitativas

Para ilustrar o tipo de análise, consideremos o exemplo a seguir.

Exemplo 4.1. Suponha que queiramos analisar o comportamento conjunto das variáveis *Y*: grau de instrução e *V*: região de procedência, cujas observações estão contidas na Tabela 2.1. A distribuição de freqüências é representada por uma tabela de dupla entrada e está na Tabela 4.2.

Cada elemento do corpo da tabela dá a freqüência observada das realizações simultâneas de *Y* e *V*. Assim, observamos quatro indivíduos da capital com ensino fundamental, sete do interior com ensino médio etc.

A *linha* dos totais fornece a distribuição da variável *Y*, ao passo que a *coluna* dos totais fornece a distribuição da variável *V*. As distribuições assim obtidas são chamadas tecnicamente de *distribuições marginais*, enquanto a Tabela 4.2 constitui a *distribuição conjunta de Y* e *V*.

Tabela 4.2: Distribuição conjunta das freqüências das variáveis grau de instrução (*Y*) e região de procedência (*V*).

V \ *Y*	Ensino Fundamental	Ensino Médio	Superior	Total
Capital	4	5	2	11
Interior	3	7	2	12
Outra	5	6	2	13
Total	12	18	6	36

Fonte: Tabela 2.1.

Em vez de trabalharmos com as freqüências absolutas, podemos construir tabelas com as freqüências relativas (proporções), como foi feito no caso unidimensional. Mas aqui existem três possibilidades de expressarmos a proporção de cada casela:

(a) em relação ao total geral;

(b) em relação ao total de cada linha;

(c) ou em relação ao total de cada coluna.

De acordo com o objetivo do problema em estudo, uma delas será a mais conveniente.

A Tabela 4.3 apresenta a distribuição conjunta das freqüências relativas, expressas como proporções do total geral. Podemos, então, afirmar que 11% dos empregados vêm da capital e têm o ensino fundamental. Os totais nas margens fornecem as distribuições unidimensionais de cada uma das variáveis. Por exemplo, 31% dos indivíduos vêm da capital, 33% do interior e 36% de outras regiões. Observe que, devido ao problema de aproximação das divisões, a distribuição das proporções introduz algumas diferenças não existentes. Compare, por exemplo, as colunas de instrução superior nas Tabelas 4.2 e 4.3.

A Tabela 4.4 apresenta a distribuição das proporções em relação ao total das colunas. Podemos dizer que, entre os empregados com instrução até o ensino fundamental, 33% vêm da capital, ao passo que entre os empregados com ensino médio, 28% vêm da capital. Esse tipo de tabela serve para comparar a distribuição da procedência dos indivíduos conforme o grau de instrução.

Tabela 4.3: Distribuição conjunta das proporções (em porcentagem) em relação ao total geral das variáveis Y e V definidas no texto.

V \ Y	Fundamental	Médio	Superior	Total
Capital	11%	14%	6%	31%
Interior	8%	19%	6%	33%
Outra	14%	17%	5%	36%
Total	33%	50%	17%	100%

Fonte: Tabela 4.2.

Tabela 4.4: Distribuição conjunta das proporções (em porcentagem) em relação aos totais de cada coluna das variáveis Y e V definidas no texto.

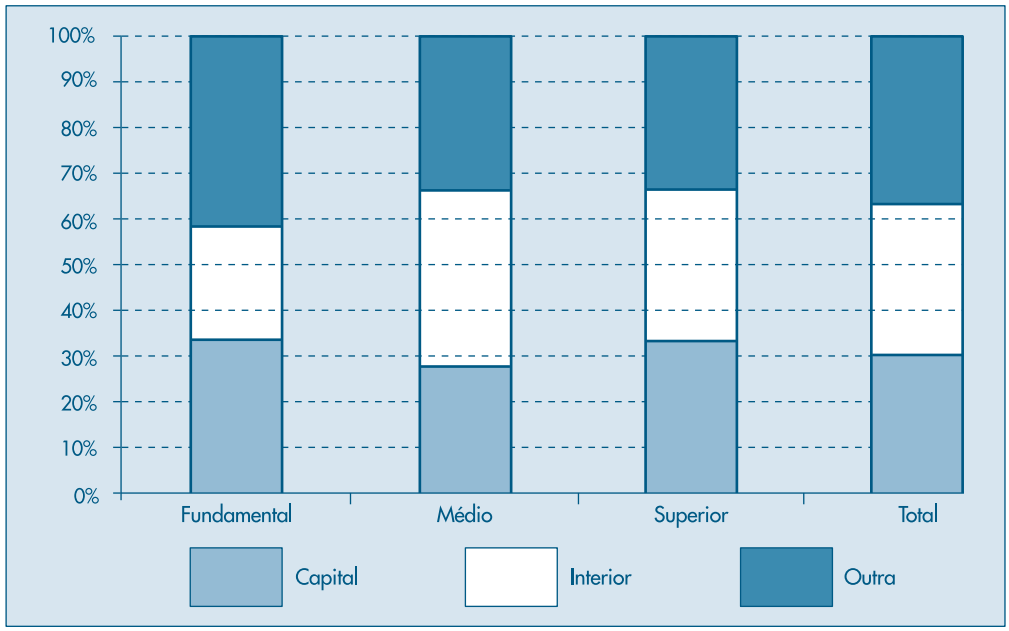
V \ Y	Fundamental	Médio	Superior	Total
Capital	33%	28%	33%	31%
Interior	25%	39%	33%	33%
Outra	42%	33%	34%	36%
Total	100%	100%	100%	100%

Fonte: Tabela 4.2.

De modo análogo, podemos construir a distribuição das proporções em relação ao total das linhas. Aconselhamos o leitor a construir essa tabela.

A comparação entre as duas variáveis também pode ser feita utilizando-se representações gráficas. Na Figura 4.1 apresentamos uma possível representação para os dados da Tabela 4.4.

Figura 4.1: Distribuição da região de procedência por grau de instrução.



Problemas

1. Usando os dados da Tabela 2.1, Capítulo 2:
   (a) Construa a distribuição de freqüência conjunta para as variáveis grau de instrução e região de procedência.
   (b) Qual a porcentagem de funcionários que têm o ensino médio?
   (c) Qual a porcentagem daqueles que têm o ensino médio e são do interior?
   (d) Dentre os funcionários do interior, quantos por cento têm o ensino médio?

2. No problema anterior, sorteando um funcionário ao acaso entre os 36:
   (a) Qual será provavelmente o seu grau de instrução?
   (b) E sua região de procedência?
   (c) Qual a probabilidade do sorteado ter nível superior?
   (d) Sabendo que o sorteado é do interior, qual a probabilidade de ele possuir nível superior?
   (e) Sabendo que o escolhido é da capital, qual a probabilidade de ele possuir nível superior?

3. Numa pesquisa sobre rotatividade de mão-de-obra, para uma amostra de 40 pessoas foram observadas duas variáveis: número de empregos nos últimos dois anos (X) e salário mais recente, em número de salários mínimos (Y). Os resultados foram:

Indivíduo	X	Y	Indivíduo	X	Y
1	1	6	21	2	4
2	3	2	22	3	2
3	2	4	23	4	1
4	3	1	24	1	5
5	2	4	25	2	4
6	2	1	26	3	2
7	3	3	27	4	1
8	1	5	28	1	5
9	2	2	29	4	4
10	3	2	30	3	3
11	2	5	31	2	2
12	3	2	32	1	1
13	1	6	33	4	1
14	2	6	34	2	6
15	3	2	35	4	2
16	4	2	36	3	1
17	1	5	37	1	4
18	2	5	38	3	2
19	2	1	39	2	3
20	2	1	40	2	5

(a) Usando a mediana, classifique os indivíduos em dois níveis, alto e baixo, para cada uma das variáveis, e construa a distribuição de freqüências conjunta das duas classificações.

(b) Qual a porcentagem das pessoas com baixa rotatividade e ganhando pouco?

(c) Qual a porcentagem das pessoas que ganham pouco?

(d) Entre as pessoas com baixa rotatividade, qual a porcentagem das que ganham pouco?

(e) A informação adicional dada em *(d)* mudou muito a porcentagem observada em *(c)*? O que isso significa?

4.3 Associação entre Variáveis Qualitativas

Um dos principais objetivos de se construir uma distribuição conjunta de duas variáveis qualitativas é descrever a associação entre elas, isto é, queremos conhecer o grau de *dependência* entre elas, de modo que possamos prever melhor o resultado de uma delas quando conhecermos a realização da outra.

Por exemplo, se quisermos estimar qual a renda média de uma família moradora da cidade de São Paulo, a informação adicional sobre a classe social a que ela pertence nos permite estimar com maior precisão essa renda, pois sabemos que existe uma dependência entre as duas variáveis: renda familiar e classe social. Ou, ainda, suponhamos que uma pessoa seja sorteada ao acaso na população da cidade de São Paulo e devamos adivinhar o sexo dessa pessoa. Como a proporção de pessoas de cada sexo

é aproximadamente a mesma, o resultado desse exercício de adivinhação poderia ser qualquer um dos sexos: masculino ou feminino. Mas se a mesma pergunta fosse feita e também fosse dito que a pessoa sorteada trabalha na indústria siderúrgica, então nossa resposta mais provável seria que a pessoa sorteada é do sexo masculino. Ou seja, há um grau de dependência grande entre as variáveis sexo e ramo de atividade.

Vejamos como podemos identificar a associação entre duas variáveis da distribuição conjunta.

Exemplo 4.2. Queremos verificar se existe ou não associação entre o sexo e a carreira escolhida por 200 alunos de Economia e Administração. Esses dados estão na Tabela 4.5.

Tabela 4.5: Distribuição conjunta de alunos segundo o sexo (X) e o curso escolhido (Y).

Y \ X	Masculino	Feminino	Total
Economia	85	35	120
Administração	55	25	80
Total	140	60	200

Fonte: Dados hipotéticos.

Inicialmente, verificamos que fica muito difícil tirar alguma conclusão, devido à diferença entre os totais marginais. Devemos, pois, construir as proporções segundo as linhas ou as colunas para podermos fazer comparações. Fixemos os totais das colunas; a distribuição está na Tabela 4.6.

Tabela 4.6: Distribuição conjunta das proporções (em porcentagem) de alunos segundo o sexo (X) e o curso escolhido (Y).

Y \ X	Masculino	Feminino	Total
Economia	61%	58%	60%
Administração	39%	42%	40%
Total	100%	100%	100%

Fonte: Tabela 4.5.

A partir dessa tabela podemos observar que, *independentemente do sexo*, 60% das pessoas preferem Economia e 40% preferem Administração (observe na coluna de total). Não havendo dependência entre as variáveis, esperaríamos essas mesmas proporções para cada sexo. Observando a tabela, vemos que as proporções do sexo masculino (61% e 39%) e do sexo feminino (58% e 42%) são próximas das marginais (60% e 40%). Esses resultados parecem indicar não haver dependência entre as duas variáveis, para o conjunto de alunos considerado. Concluímos então que, neste caso, as variáveis sexo e escolha do curso parecem ser *não associadas*.

Vamos considerar, agora, um problema semelhante, mas envolvendo alunos de Física e Ciências Sociais, cuja distribuição conjunta está na Tabela 4.7.

Tabela 4.7: Distribuição conjunta das freqüências e proporções (em porcentagem), segundo o sexo (X) e o curso escolhido (Y).

Y \ X	Masculino	Feminino	Total
Física	100 (71%)	20 (33%)	120 (60%)
Ciências Sociais	40 (29%)	40 (67%)	80 (40%)
Total	140 (100%)	60 (100%)	200 (100%)

Fonte: Dados hipotéticos.

Inicialmente, convém observar que, para economizar espaço, resumimos duas tabelas numa única, indicando as proporções em relação aos totais das colunas entre parênteses. Comparando agora a distribuição das proporções pelos cursos, independentemente do sexo (coluna de totais), com as distribuições diferenciadas por sexo (colunas de masculino e feminino), observamos uma disparidade bem acentuada nas proporções. Parece, pois, haver maior concentração de homens no curso de Física e de mulheres no de Ciências Sociais. Portanto, nesse caso, as variáveis sexo e curso escolhido parecem ser *associadas*.

Quando existe associação entre variáveis, sempre é interessante quantificar essa associação, e isso será objeto da próxima seção. Antes de passarmos a discutir esse aspecto, convém observar que teríamos obtido as mesmas conclusões do Exemplo 4.2 se tivéssemos calculado as proporções, mantendo constantes os totais das linhas.

Problemas

4. Usando os dados do Problema 1, responda:
   (a) Qual a distribuição das proporções do grau de educação segundo cada uma das regiões de procedência?
   (b) Baseado no resultado anterior e no Problema 2, você diria que existe dependência entre a região de procedência e o nível de educação do funcionário?

5. Usando o Problema 3, verifique se há relações entre as variáveis rotatividade e salário.

6. Uma companhia de seguros analisou a freqüência com que 2.000 segurados (1.000 homens e 1.000 mulheres) usaram o hospital. Os resultados foram:

	Homens	Mulheres
Usaram o hospital	100	150
Não usaram o hospital	900	850

   (a) Calcule a proporção de homens entre os indivíduos que usaram o hospital.
   (b) Calcule a proporção de homens entre os indivíduos que não usaram o hospital.
   (c) O uso do hospital independe do sexo do segurado?

4.4 Medidas de Associação entre Variáveis Qualitativas

De modo geral, a quantificação do grau de associação entre duas variáveis é feita pelos chamados *coeficientes de associação* ou *correlação*. Essas são medidas que descrevem, por meio de um único número, a associação (ou dependência) entre duas variáveis. Para maior facilidade de compreensão, esses coeficientes usualmente variam entre 0 e 1, ou entre –1 e +1, e a proximidade de zero indica falta de associação.

Existem muitas medidas que quantificam a associação entre variáveis qualitativas, apresentaremos apenas duas delas: o chamado *coeficiente de contingência*, devido a K. Pearson e uma modificação desse.

Exemplo 4.3. Queremos verificar se a criação de determinado tipo de cooperativa está associada com algum fator regional. Coletados os dados relevantes, obtemos a Tabela 4.8.

Tabela 4.8: Cooperativas autorizadas a funcionar por tipo e estado, junho de 1974.

Estado	Tipo de Cooperativa				Total
	Consumidor	Produtor	Escola	Outras	
São Paulo	214 (33%)	237 (37%)	78 (12%)	119 (18%)	648 (100%)
Paraná	51 (17%)	102 (34%)	126 (42%)	22 (7%)	301 (100%)
Rio G. do Sul	111 (18%)	304 (51%)	139 (23%)	48 (8%)	602 (100%)
Total	376 (24%)	643 (42%)	343 (22%)	189 (12%)	1.551 (100%)

Fonte: Sinopse Estatística da Brasil — IBGE, 1977.

A análise da tabela mostra a existência de certa dependência entre as variáveis. Caso não houvesse associação, esperaríamos que em cada estado tivéssemos 24% de cooperativas de consumidores, 42% de cooperativas de produtores, 22% de escolas e 12% de outros tipos. Então, por exemplo, o número esperado de cooperativas de consumidores no Estado de São Paulo seria 648 × 0,24 = 157 e no Paraná seria 301 × 0,24 = 73 (ver Tabela 4.9).

Tabela 4.9: Valores esperados na Tabela 4.8 assumindo a independência entre as duas variáveis.

Estado	Tipo de Cooperativa				Total
	Consumidor	Produtor	Escola	Outras	
São Paulo	157 (24%)	269 (42%)	143 (22%)	79 (12%)	648 (100%)
Paraná	73 (24%)	124 (42%)	67 (22%)	37 (12%)	301 (100%)
Rio G. do Sul	146 (24%)	250 (42%)	133 (22%)	73 (12%)	602 (100%)
Total	376 (24%)	643 (42%)	343 (22%)	189 (12%)	1.551 (100%)

Fonte: Tabela 4.8.

Tabela 4.10: Desvios entre observados e esperados.

Estado	Tipo de Cooperativa			
	Consumidor	Produtor	Escola	Outras
São Paulo	57 (20,69)	–32 (3,81)	–65 (29,55)	40 (20,25)
Paraná	–22 (6,63)	–22 (3,90)	59 (51,96)	–15 (6,08)
Rio G. do Sul	–35 (8,39)	54 (11,66)	6 (0,27)	–25 (8,56)

Fonte: Tabelas 4.8 e 4.9.

Comparando as duas tabelas, podemos verificar as discrepâncias existentes entre os valores observados (Tabela 4.8) e os valores esperados (Tabela 4.9), caso as variáveis não fossem associadas. Na Tabela 4.10 resumimos os desvios: valores observados menos valores esperados. Observando essa tabela podemos tirar algumas conclusões:

(i) A soma total dos resíduos é nula. Isso pode ser verificado facilmente somando-se cada linha.

(ii) A casela Escola-São Paulo é aquela que apresenta o maior desvio da suposição de não-associação (–65). Nessa casela esperávamos 143 casos. A casela Escola-Paraná também tem um desvio alto (59), mas o valor esperado é bem menor (67). Portanto, se fôssemos considerar os desvios relativos, aquele correspondente ao segundo caso seria bem maior. Uma maneira de observar esse fato é construir, para cada casela, a medida

$$\frac{(o_i - e_i)^2}{e_i}, \tag{4.1}$$

no qual o_i é o valor observado e e_i é o valor esperado.

Usando (4.1) para a casela Escola-São Paulo obtemos $(-65)^2/143 = 29{,}55$ e para a casela Escola-Paraná obtemos $(59)^2/67 = 51{,}96$, o que é uma indicação de que o desvio devido a essa última casela é "maior" do que aquele da primeira. Na Tabela 4.10 indicamos entre parênteses esses valores para todas as caselas.

Uma medida do afastamento global pode ser dada pela soma de todas as medidas (4.1). Essa medida é denominada χ^2 (qui-quadrado) de Pearson, e no nosso exemplo teríamos

$$\chi^2 = 20{,}69 + 6{,}63 + \ldots + 8{,}56 = 171{,}76.$$

Um valor grande de χ^2 indica associação entre as variáveis, o que parece ser o caso.

Antes de dar uma fórmula geral para essa medida de associação, vamos introduzir, na Tabela 4.11, uma notação geral para tabelas de dupla entrada.

Tabela 4.11: Notação para tabelas de contingência.

X \ Y	B_1	B_2	...	B_j	...	B_s	Total
A_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1s}	$n_{1.}$
A_2	n_{21}	n_{22}	...	n_{2j}	...	n_{2s}	$n_{2.}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
A_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{is}	$n_{i.}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
A_r	n_{r1}	n_{r2}	...	n_{rj}	...	n_{rs}	$n_{r.}$
Total	$n_{.1}$	$n_{.2}$	...	$n_{.j}$	...	$n_{.s}$	$n..$

Suponha que temos duas variáveis qualitativas X e Y, classificadas em r categorias $A_1, A_2, ..., A_r$ para X e s categorias $B_1, B_2, ..., B_s$, para Y.

Na tabela, temos:

n_{ij} = número de elementos pertencentes à i-ésima categoria de X e j-ésima categoria de Y;

$n_{i.} = \sum_{j=1}^{s} n_{ij}$ = número de elementos da i-ésima categoria de X;

$n_{.j} = \sum_{i=1}^{r} n_{ij}$ = número de elementos da j-ésima categoria de Y;

$n.. = n = \sum_{i=1}^{r} \sum_{j=1}^{s} n_{ij}$ = número total de elementos.

Sob a hipótese de que as variáveis X e Y não sejam associadas (comumente dizemos independentes), temos que

$$\frac{n_{i1}}{n_{.1}} = \frac{n_{i2}}{n_{.2}} = ... = \frac{n_{is}}{n_{.s}}, \; i = 1, 2, ..., r \tag{4.2}$$

ou ainda

$$\frac{n_{ij}}{n_{.j}} = \frac{n_{i.}}{n}, \quad i = 1, ..., r, \; j = 1, ..., s$$

de onde se deduz, finalmente, que

$$n_{ij} = \frac{n_{i.}\, n_{.j}}{n}, \qquad i = 1, ..., r, \; j = 1, ..., s. \tag{4.3}$$

Portanto, sob a hipótese de independência, de (4.3) segue que, em termos de freqüências relativas, podemos escrever $f_{ij} = f_{i.}\, f_{.j}$.

Chamando de freqüências esperadas os valores dados pelos segundos membros de (4.3), e denotando-as por n^*_{ij}, temos que o qui-quadrado de Pearson pode ser escrito

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(n_{ij} - n^*_{ij})^2}{n^*_{ij}}, \tag{4.4}$$

onde n_{ij} são os valores efetivamente observados. Se a hipótese de não-associação for verdadeira, o valor calculado de (4.4) deve estar próximo de zero. Se as variáveis forem associadas, o valor de χ^2 deve ser grande.

Podemos escrever a fórmula (4.4) em termos de freqüências relativas, como

$$\chi^2 = n\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(f_{ij} - f_{ij}^*)^2}{f_{ij}^*},$$

para a qual as notações são similares.

Pearson definiu uma medida de associação, baseada em (4.4), chamada *coeficiente de contingência*, dado por

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}. \tag{4.5}$$

Contudo, o coeficiente acima não varia entre 0 e 1. O valor máximo de C depende de r e s. Para evitar esse inconveniente, costuma-se definir um outro coeficiente, dado por

$$T = \sqrt{\frac{\chi^2/n}{(r-1)(s-1)}}, \tag{4.6}$$

que atinge o máximo igual a 1 se $r = s$.

Para o Exemplo 4.3 temos que $C = 0{,}32$ e $T = 0{,}14$. Voltaremos a falar do uso do χ^2 no Capítulo 14.

Problemas

7. Usando os dados do Problema 1, calcule o valor de χ^2 e o coeficiente de contingência C. Esses valores estão de acordo com as conclusões obtidas anteriormente?

8. Qual o valor de χ^2 e de C para os dados do Problema 3? E para o Problema 6? Calcule T.

9. A Companhia A de dedetização afirma que o processo por ela utilizado garante um efeito mais prolongado do que aquele obtido por seus concorrentes mais diretos. Uma amostra de vários ambientes dedetizados foi colhida e anotou-se a duração do efeito de dedetização. Os resultados estão na tabela abaixo. Você acha que existe alguma evidência a favor ou contra a afirmação feita pela Companhia A?

	Duração do efeito de dedetização		
Companhia	Menos de 4 meses	De 4 a 8 meses	Mais de 8 meses
A	64	120	16
B	104	175	21
C	27	48	5

4.5 Associação entre Variáveis Quantitativas

Quando as variáveis envolvidas são ambas do tipo quantitativo, pode-se usar o mesmo tipo de análise apresentado nas seções anteriores e exemplificado com variáveis qualitativas. De modo análogo, a distribuição conjunta pode ser resumida em tabelas de dupla entrada e, por meio das distribuições marginais, é possível estudar a associação das variáveis. Algumas vezes, para evitar um grande número de entradas, agrupamos os dados marginais em intervalos de classes, de modo semelhante ao resumo feito no caso unidimensional. Mas, além desse tipo de análise, as variáveis quantitativas são passíveis de procedimentos analíticos e gráficos mais refinados.

Um dispositivo bastante útil para se verificar a associação entre duas variáveis quantitativas, ou entre dois conjuntos de dados, é o *gráfico de dispersão*, que vamos introduzir por meio de exemplos.

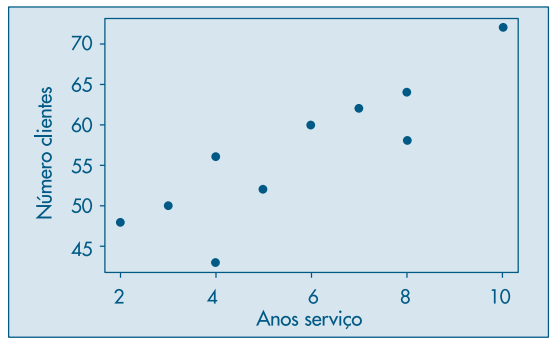
Exemplo 4.4. Na Figura 4.2 temos o gráfico de dispersão das variáveis X e Y da Tabela 4.12. Nesse tipo de gráfico temos os possíveis pares de valores (x, y), na ordem que aparecem. Para o exemplo, vemos que parece haver uma associação entre as variáveis, porque no conjunto, à medida que aumenta o tempo de serviço, aumenta o número de clientes.

Tabela 4.12: Número de anos de serviço (X) por número de clientes (Y) de agentes de uma companhia de seguros.

Agente	Anos de serviço (X)	Número de clientes (Y)
A	2	48
B	3	50
C	4	56
D	5	52
E	4	43
F	6	60
G	7	62
H	8	58
I	8	64
J	10	72

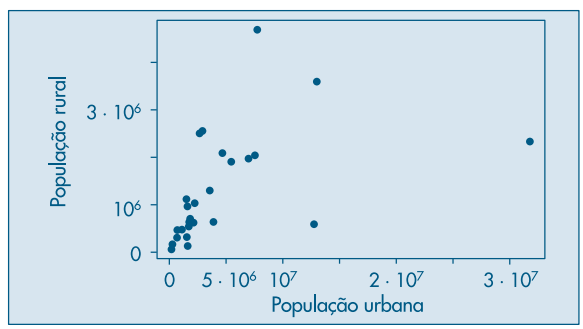
Fonte: Dados hipotéticos.

Figura 4.2: Gráfico de dispersão para as variáveis X: anos de serviço e Y: número de clientes.



Exemplo 4.5. Consideremos os dados das variáveis X: população urbana e Y: população rural, do CD-Brasil. O gráfico de dispersão está na Figura 4.3. Vemos que parece não haver associação entre as variáveis, pois os pontos não apresentam nenhuma tendência particular.

Figura 4.3: Gráfico de dispersão para as variáveis X: população urbana e Y: população rural.



Exemplo 4.6. Consideremos agora as duas situações abaixo e os respectivos gráficos de dispersão.

Tabela 4.13: Renda bruta mensal (X) e porcentagem da renda gasta em saúde (Y) para um conjunto de famílias.

Família	X	Y
A	12	7,2
B	16	7,4
C	18	7,0
D	20	6,5
E	28	6,6
F	30	6,7
G	40	6,0
H	48	5,6
I	50	6,0
J	54	5,5

Fonte: Dados hipotéticos.

(a) Numa pesquisa feita com dez famílias com renda bruta mensal entre 10 e 60 salários mínimos, mediram-se:

X: renda bruta mensal (expressa em número de salários mínimos).

Y: a porcentagem da renda bruta anual gasta com assistência médica; os dados estão na Tabela 4.13. Observando o gráfico de dispersão (Figura 4.4), vemos que existe uma associação "inversa", isto é, aumentando a renda bruta, diminui a porcentagem sobre ela gasta em assistência médica.

Figura 4.4: Gráfico de dispersão para as variáveis *X*: renda bruta e *Y*: % renda gasta com saúde.

Antes de passarmos ao exemplo seguinte, convém observar que a disposição dos dados da Tabela 4.13 numa tabela de dupla entrada não iria melhorar a compreensão dos dados, visto que, devido ao pequeno número de observações, teríamos caselas cheias apenas na diagonal.

(b) Oito indivíduos foram submetidos a um teste sobre conhecimento de língua estrangeira e, em seguida, mediu-se o tempo gasto para cada um aprender a operar uma determinada máquina. As variáveis medidas foram:

X: resultado obtido no teste (máximo = 100 pontos);

Y: tempo, em minutos, necessário para operar a máquina satisfatoriamente.

Figura 4.5: Gráfico de dispersão para as variáveis *X*: resultado no teste e *Y*: tempo de operação.

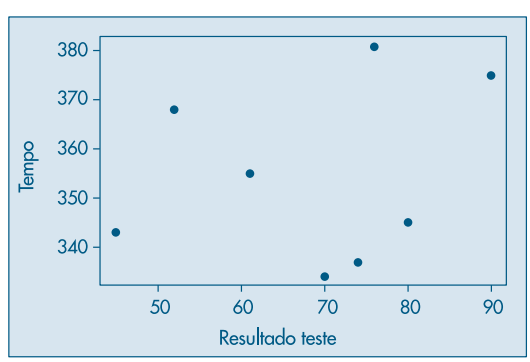


Tabela 4.14: Resultado de um teste (X) e tempo de operação de máquina (Y) para oito indivíduos.

Indivíduo	X	Y
A	45	343
B	52	368
C	61	355
D	70	334
E	74	337
F	76	381
G	80	345
H	90	375

Fonte: Dados hipotéticos.

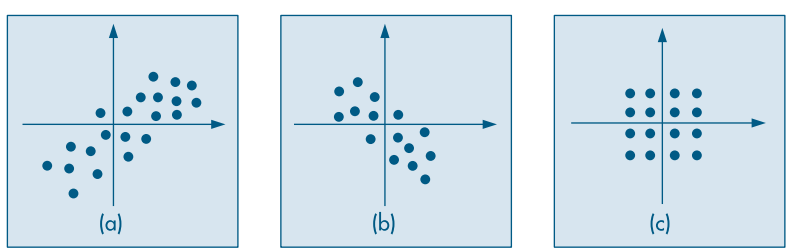
Os dados estão na Tabela 4.14. Do gráfico de dispersão (Figura 4.5) concluímos que parece não haver associação entre as duas variáveis, pois conhecer o resultado do teste não ajuda a prever o tempo gasto para aprender a operar a máquina.

A partir dos gráficos apresentados, verificamos que a representação gráfica das variáveis quantitativas ajuda muito a compreender o comportamento conjunto das duas variáveis quanto à existência ou não de associação entre elas.

Contudo, é muito útil quantificar esta associação. Existem muitos tipos de associações possíveis, e aqui iremos apresentar o tipo de relação mais simples, que é a linear. Isto é, iremos definir uma medida que avalia o quanto a nuvem de pontos no gráfico de dispersão aproxima-se de uma reta. Esta medida será definida de modo a variar num intervalo finito, especificamente, de –1 a +1.

Consideremos um gráfico de dispersão como o da Figura 4.6 (a) no qual, por meio de uma transformação conveniente, a origem foi colocada no centro da nuvem de dispersão. Aqueles dados possuem uma associação linear direta (ou positiva) e notamos que a grande maioria dos pontos está situada no primeiro e terceiro quadrantes. Nesses quadrantes as coordenadas dos pontos têm o mesmo sinal, e, portanto, o produto delas será sempre positivo. Somando-se o produto das coordenadas dos pontos, o resultado será um número positivo, pois existem mais produtos positivos do que negativos.

Figura 4.6: Tipos de associações entre duas variáveis.



Para a dispersão da Figura 4.6 (b), observamos uma dependência linear inversa (ou negativa) e, procedendo-se como anteriormente, a soma dos produtos das coordenadas será negativa.

Finalmente, para a Figura 4.6 (c), a soma dos produtos das coordenadas será zero, pois cada resultado positivo tem um resultado negativo simétrico, anulando-se na soma. Nesse caso não há associação linear entre as duas variáveis. Em casos semelhantes, quando a distribuição dos pontos for mais ou menos circular, a soma dos produtos será aproximadamente zero.

Baseando-se nesses fatos é que iremos definir o coeficiente de correlação (linear) entre duas variáveis, que é uma medida do grau de associação entre elas e também da proximidade dos dados a uma reta. Antes, cabe uma observação. A soma dos produtos das coordenadas depende, e muito, do número de pontos. Considere o caso de associação positiva: a soma acima tende a aumentar com o número de pares (x, y) e ficaria difícil comparar essa medida para dois conjuntos com números diferentes de pontos. Por isso, costuma-se usar a média da soma dos produtos das coordenadas.

Exemplo 4.7. Voltemos aos dados da Tabela 4.12. O primeiro problema que devemos resolver é o da mudança da origem do sistema para o centro da nuvem de dispersão. Um ponto conveniente é $(\bar{x}, \bar{y})$, ou seja, as coordenadas da origem serão as médias dos valores de X e Y. As novas coordenadas estão mostradas na quarta e quinta colunas da Tabela 4.15.

Observando esses valores centrados, verificamos que ainda existe um problema quanto à escala usada. A variável Y tem variabilidade muito maior do que X, e o produto ficaria muito mais afetado pelos resultados de Y do que pelos de X. Para corrigirmos isso, podemos reduzir as duas variáveis a uma mesma escala, dividindo-se os desvios pelos respectivos desvios padrões. Esses novos valores estão nas colunas 6 e 7. Observe as mudanças (escalas dos eixos) de variáveis realizadas, acompanhando a Figura 4.7. Finalmente, na coluna 8, indicamos os produtos das coordenadas reduzidas e sua soma, 8,769, que, como esperávamos, é positiva. Para completar a definição dessa medida de associação, basta calcular a média dos produtos das coordenadas reduzidas, isto é, correlação $(X,Y) = 8{,}769/10 = 0{,}877$.

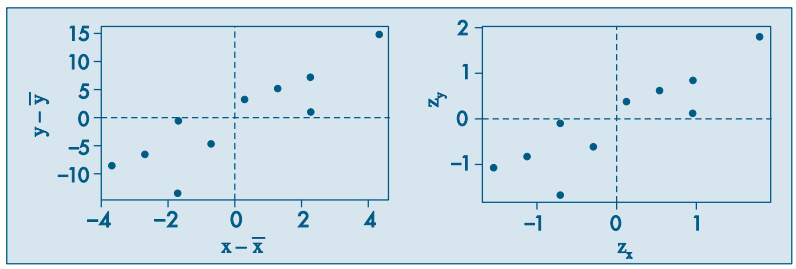
Tabela 4.15: Cálculo do coeficiente de correlação.

Agente	Anos x	Clientes y	$x-\bar{x}$	$y-\bar{y}$	$\frac{x-\bar{x}}{dp(x)} = z_x$	$\frac{y-\bar{y}}{dp(y)} = z_y$	$z_x \cdot z_y$
A	2	48	–3,7	–8,5	–1,54	–1,05	1,617
B	3	50	–2,7	–6,5	–1,12	–0,80	0,846
C	4	56	–1,7	–0,5	–0,71	–0,06	0,043
D	5	52	–0,7	–4,5	–0,29	–0,55	0,160
E	4	43	–1,7	–13,5	–0,71	–1,66	1,179
F	6	60	0,3	3,5	0,12	0,43	0,052
G	7	62	1,3	5,5	0,54	0,68	0,367
H	8	58	2,3	1,5	0,95	0,19	0,181
I	8	64	2,3	7,5	0,95	0,92	0,874
J	10	72	4,3	15,5	1,78	1,91	3,400
Total	57	565	0	0			8,769

$\bar{x} = 5{,}7$, $dp(X) = 2{,}41$, $\bar{y} = 56{,}5$, $dp(Y) = 8{,}11$

Portanto, para esse exemplo, o grau de associação linear está quantificado por 87,7%.

Figura 4.7: Mudança de escalas para o cálculo do coeficiente de correlação.



Da discussão feita até aqui, podemos definir o coeficiente de correlação do seguinte modo.

Definição. Dados n pares de valores (x_1, y_1), (x_2, y_2), ..., (x_n, y_n), chamaremos de coeficiente de correlação entre as duas variáveis X e Y a

$$\text{corr}(X, Y) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{dp(X)}\right)\left(\frac{y_i - \bar{y}}{dp(Y)}\right), \tag{4.7}$$

ou seja, a média dos produtos dos valores padronizados das variáveis.

Não é difícil provar que o coeficiente de correlação satisfaz

$$-1 \leqslant \text{corr}(X, Y) \leqslant 1. \tag{4.8}$$

A definição acima pode ser operacionalizada de modo mais conveniente pelas seguintes fórmulas:

$$\text{corr}(X, Y) = \frac{1}{n}\sum\left(\frac{x_i - \bar{x}}{dp(X)}\right)\left(\frac{y_i - \bar{y}}{dp(Y)}\right) = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sqrt{\left(\sum x_i^2 - n\bar{x}^2\right)\left(\sum y_i^2 - n\bar{y}^2\right)}}. \tag{4.9}$$

O numerador da expressão acima, que mede o total da concentração dos pontos pelos quatro quadrantes, dá origem a uma medida bastante usada e que definimos a seguir.

Definição. Dados n pares de valores (x_1, y_1), ..., (x_n, y_n), chamaremos de *covariância* entre as duas variáveis X e Y a

$$\text{cov}(X, Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}, \tag{4.10}$$

ou seja, a média dos produtos dos valores centrados das variáveis.

Com essa definição, o coeficiente de correlação pode ser escrito como

$$\text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{dp(X)\ .\ dp(Y)}. \tag{4.11}$$

Para analisar dois conjuntos de dados podemos recorrer, também, aos métodos utilizados anteriormente para analisar um conjunto de dados, exibindo as análises feitas separadamente, para efeito de comparação. Por exemplo, podemos exibir os desenhos esquemáticos, ou os ramos-e-folhas para os dois conjuntos de observações.

4.6 Associação entre Variáveis Qualitativas e Quantitativas

Como mencionado na introdução deste capítulo, é comum nessas situações analisar o que acontece com a variável quantitativa dentro de cada categoria da variável qualitativa. Essa análise pode ser conduzida por meio de medidas-resumo, histogramas, *box plots* ou ramo-e-folhas. Vamos ilustrar com um exemplo.

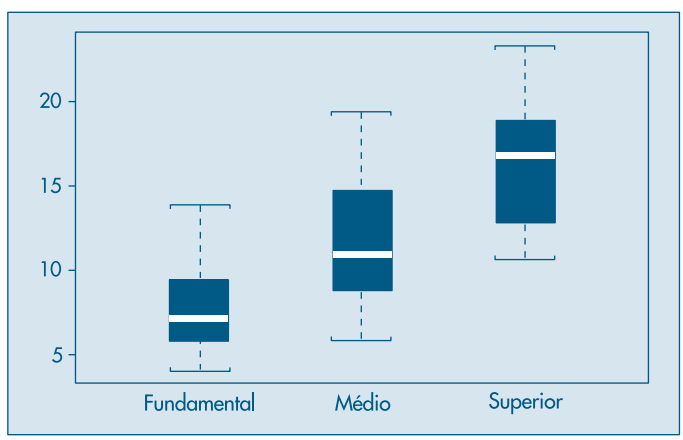
Exemplo 4.8. Retomemos os dados da Tabela 2.1, para os quais desejamos analisar agora o comportamento dos salários dentro de cada categoria de grau de instrução, ou seja, investigar o comportamento conjunto das variáveis S e Y.

Tabela 4.16: Medidas-resumo para a variável salário, segundo o grau de instrução, na Companhia MB.

Grau de instrução	n	$\bar{s}$	dp(S)	var(S)	$s_{(1)}$	q_1	q_2	q_3	$s_{(n)}$
Fundamental	12	7,84	2,79	7,77	4,00	6,01	7,13	9,16	13,65
Médio	18	11,54	3,62	13,10	5,73	8,84	10,91	14,48	19,40
Superior	6	16,48	4,11	16,89	10,53	13,65	16,74	18,38	23,30
Todos	36	11,12	4,52	20,46	4,00	7,55	10,17	14,06	23,30

Comecemos a análise construindo a Tabela 4.16, que contém medidas-resumo da variável S para cada categoria de Y. A seguir, na Figura 4.8, apresentamos uma visualização gráfica por meio de *box plots*.

Figura 4.8: *Box plots* de salário segundo grau de instrução.



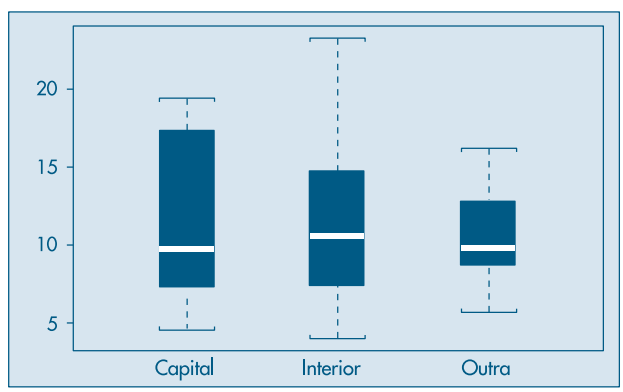
A leitura desses resultados sugere uma dependência dos salários em relação ao grau de instrução: o salário aumenta conforme aumenta o nível de educação do indivíduo. O salário médio de um funcionário é 11,12 (salários mínimos), já para um funcionário com curso superior o salário médio passa a ser 16,48, enquanto funcionários com o ensino fundamental completo recebem, em média, 7,84.

Na Tabela 4.17 e Figura 4.9 temos os resultados da análise dos salários em função da região de procedência (V), que mostram a inexistência de uma relação melhor definida entre essas duas variáveis. Ou, ainda, os salários estão mais relacionados com o grau de instrução do que com a região de procedência.

Tabela 4.17: Medidas-resumo para a variável salário segundo a região de procedência, na Companhia MB.

Região de procedência	n	$\bar{s}$	dp(S)	var(S)	$s_{(1)}$	q_1	q_2	q_3	$s_{(n)}$
Capital	11	11,46	5,22	27,27	4,56	7,49	9,77	16,63	19,40
Interior	12	11,55	5,07	25,71	4,00	7,81	10,64	14,70	23,30
Outra	13	10,45	3,02	9,13	5,73	8,74	9,80	12,79	16,22
Todos	36	11,12	4,52	20,46	4,00	7,55	10,17	14,06	23,30

Figura 4.9: *Box plots* de salário segundo região de procedência.



Como nos casos anteriores, é conveniente poder contar com uma medida que quantifique o grau de dependência entre as variáveis. Com esse intuito, convém observar que as variâncias podem ser usadas como insumos para construir essa medida. Sem usar a informação da variável categorizada, a variância calculada para a variável quantitativa para todos os dados mede a dispersão dos dados globalmente. Se a variância dentro de cada categoria for pequena e menor do que a global, significa que a variável qualitativa melhora a capacidade de previsão da quantitativa e portanto existe uma relação entre as duas variáveis.

Observe que, para as variáveis S e Y, as variâncias de S dentro das três categorias são menores do que a global. Já para as variáveis S e V, temos duas variâncias de S maiores e uma menor do que a global, o que corrobora a afirmação acima.

Necessita-se, então, de uma medida-resumo da variância entre as categorias da variável qualitativa. Vamos usar a média das variâncias, porém ponderada pelo número de observações em cada categoria, ou seja,

$$\overline{\text{var}(S)} = \frac{\sum_{i=1}^{k} n_i \text{var}_i(S)}{\sum_{i=1}^{k} n_i}, \tag{4.12}$$

no qual k é o número de categorias ($k = 3$ nos dois exemplos acima) e $\text{var}_i(S)$ denota a variância de S dentro da categoria i, $i = 1, 2, \ldots, k$.

Pode-se mostrar que $\overline{\text{var}(S)} \leqslant \text{var}(S)$, de modo que podemos definir o grau de associação entre as duas variáveis como o ganho relativo na variância, obtido pela introdução da variável qualitativa. Explicitamente,

$$R^2 = \frac{\text{var}(S) - \overline{\text{var}(S)}}{\text{var}(S)} = 1 - \frac{\overline{\text{var}(S)}}{\text{var}(S)}. \tag{4.13}$$

Note que $0 \leqslant R^2 \leqslant 1$. O símbolo R^2 é usual em análise de variância e regressão, tópicos a serem abordados nos Capítulos 15 e 16, respectivamente.

Exemplo 4.9. Voltando aos dados do Exemplo 4.8, vemos que para a variável S na presença de grau de instrução, tem-se

$$\overline{\text{var}(S)} = \frac{12(7{,}77) + 18(13{,}10) + 6(16{,}89)}{12 + 18 + 6} = 11{,}96,$$

$$\text{var}(S) = 20{,}46,$$

de modo que

$$R^2 = 1 - \frac{11{,}96}{20{,}46} = 0{,}415,$$

e dizemos que 41,5% da variação total do salário é *explicada* pela variável grau de instrução.

Para S e região de procedência temos

$$\overline{\text{var}(S)} = \frac{11(27{,}27) + 12(25{,}71) + 13(9{,}13)}{11 + 12 + 13} = 20{,}20,$$

e, portanto,

$$R^2 = 1 - \frac{20{,}20}{20{,}46} = 0{,}013,$$

de modo que apenas 1,3% da variabilidade dos salários é explicada pela região de procedência. A comparação desses dois números mostra maior relação entre S e Y do que entre S e V.

Problemas

10. Para cada par de variáveis abaixo, esboce o diagrama de dispersão. Diga se você espera uma dependência linear e nos casos afirmativos avalie o coeficiente de correlação.
    (a) Peso e altura dos alunos do primeiro ano de um curso de Administração.
    (b) Peso e altura dos funcionários de um escritório.
    (c) Quantidade de trigo produzida e quantidade de água recebida por canteiros numa estação experimental.
    (d) Notas de Cálculo e Estatística de uma classe onde as duas disciplinas são lecionadas.
    (e) Acuidade visual e idade de um grupo de pessoas.
    (f) Renda familiar e porcentagem dela gasta em alimentação.
    (g) Número de peças montadas e resultado de um teste de inglês por operário.

11. Abaixo estão os dados referentes à porcentagem da população economicamente ativa empregada no setor primário e o respectivo índice de analfabetismo para algumas regiões metropolitanas brasileiras.

Regiões metropolitanas	Setor primário	Índice de analfabetismo
São Paulo	2,0	17,5
Rio de Janeiro	2,5	18,5
Belém	2,9	19,5
Belo Horizonte	3,3	22,2
Salvador	4,1	26,5
Porto Alegre	4,3	16,6
Recife	7,0	36,6
Fortaleza	13,0	38,4

Fonte: Indicadores Sociais para Áreas Urbanas — IBGE — 1977.

    (a) Faça o diagrama de dispersão.
    (b) Você acha que existe uma dependência linear entre as duas variáveis?
    (c) Calcule o coeficiente de correlação.
    (d) Existe alguma região com comportamento diferente das demais? Se existe, elimine o valor correspondente e recalcule o coeficiente de correlação.

12. Usando os dados do Problema 3:
    (a) Construa a tabela de freqüências conjuntas para as variáveis X (número de empregos nos dois últimos anos) e Y (salário mais recente).
    (b) Como poderia ser feito o gráfico de dispersão desses dados?
    (c) Calcule o coeficiente de correlação. Baseado nesse número você diria que existe dependência entre as duas variáveis?

13. Quer se verificar a relação entre o tempo de reação e o número de alternativas apresentadas a indivíduos acostumados a tomadas de decisão. Planejou-se um experimento em que se pedia ao participante para classificar objetos segundo um critério previamente discutido. Participaram do experimento 15 executivos divididos aleatoriamente em grupos de cinco. Pediu-se, então, a cada grupo para classificar dois, três e quatro objetos, respectivamente. Os dados estão abaixo.

Nº de objetos	2	3	4
Tempo de reação	1, 2, 3, 3, 4	2, 3, 4, 4, 5	4, 5, 5, 6, 7

(a) Faça o gráfico de dispersão das duas variáveis.
(b) Qual o coeficiente de correlação entre elas?

14. Calcule o grau de associação entre as variáveis estado civil e idade, na Tabela 2.1.

15. Usando os dados do Problema 9 do Capítulo 2, calcule o grau de associação entre seção e notas em Estatística.

4.7 Gráficos $q \times q$

Outro tipo de representação gráfica que podemos utilizar para duas variáveis é o *gráfico quantis × quantis*, que passamos a discutir.

Suponha que temos valores $x_1, ..., x_n$ da variável X e valores $y_1, ..., y_m$ da variável Y, todos medidos pela mesma unidade. Por exemplo, temos temperaturas de duas cidades ou alturas de dois grupos de indivíduos etc. O gráfico $q \times q$ é um gráfico dos quantis de X contra os quantis de Y.

Pelo que vimos no Capítulo 3, se $m = n$ o gráfico $q \times q$ é um gráfico dos dados ordenados de X contra os dados ordenados de Y. Se as distribuições dos dois conjuntos de dados fossem idênticas, os pontos estariam sobre a reta $y = x$.

Enquanto um gráfico de dispersão fornece uma possível relação *global* entre as variáveis, o gráfico $q \times q$ mostra se valores pequenos de X estão relacionados com valores pequenos de Y, se valores intermediários de X estão relacionados com valores intermediários de Y e se valores grandes de X estão relacionados com valores grandes de Y. Num gráfico de dispersão podemos ter $x_1 < x_2$ e $y_1 > y_2$, o que não pode acontecer num gráfico $q \times q$, pois os valores em ambos os eixos estão ordenados, do menor para o maior.

Exemplo 4.10. Na Tabela 4.18 temos as notas de 20 alunos em duas provas de Estatística e, na Figura 4.10, temos o correspondente gráfico $q \times q$. Os pontos estão razoavelmente dispersos ao redor da reta $x = y$, mostrando que as notas dos alunos nas duas provas não são muito diferentes. Mas podemos notar que, para notas abaixo de cinco, os alunos tiveram notas maiores na segunda prova, ao passo que, para notas de cinco a oito, os alunos tiveram notas melhores na primeira prova. A maioria das notas estão concentradas entre cinco e oito.

Figura 4.10: Gráfico $q \times q$ para as notas em duas provas de Estatística.

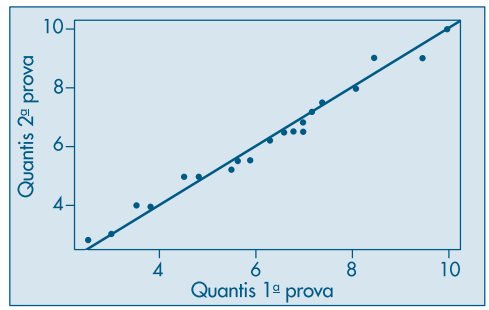


Tabela 4.18: Notas de 20 alunos em duas provas de Estatística.

Aluno	Prova 1	Prova 2	Aluno	Prova 1	Prova 2
1	8,5	8,0	11	7,4	6,5
2	3,5	2,8	12	5,6	5,0
3	7,2	6,5	13	6,3	6,5
4	5,5	6,2	14	3,0	3,0
5	9,5	9,0	15	8,1	9,0
6	7,0	7,5	16	3,8	4,0
7	4,8	5,2	17	6,8	5,5
8	6,6	7,2	18	10,0	10,0
9	2,5	4,0	19	4,5	5,5
10	7,0	6,8	20	5,9	5,0

Exemplo 4.11. Consideremos, agora, as variáveis *temperatura de Ubatuba* e *temperatura de Cananéia*, do CD-Temperaturas. O gráfico $q \times q$ está na Figura 4.11. Observamos que a maioria dos pontos está acima da reta $y = x$, mostrando que as temperaturas de Ubatuba são, em geral, maiores do que as de Cananéia, para valores maiores do que 17 graus.

Quando $m \neq n$, é necessário modificar os valores de p para os quantis da variável com maior número de pontos. Ver o Problema 33 para a solução desse caso.

Figura 4.11: Gráfico $q \times q$ para os lados de temperatura de Cananéia e Ubatuba.