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.

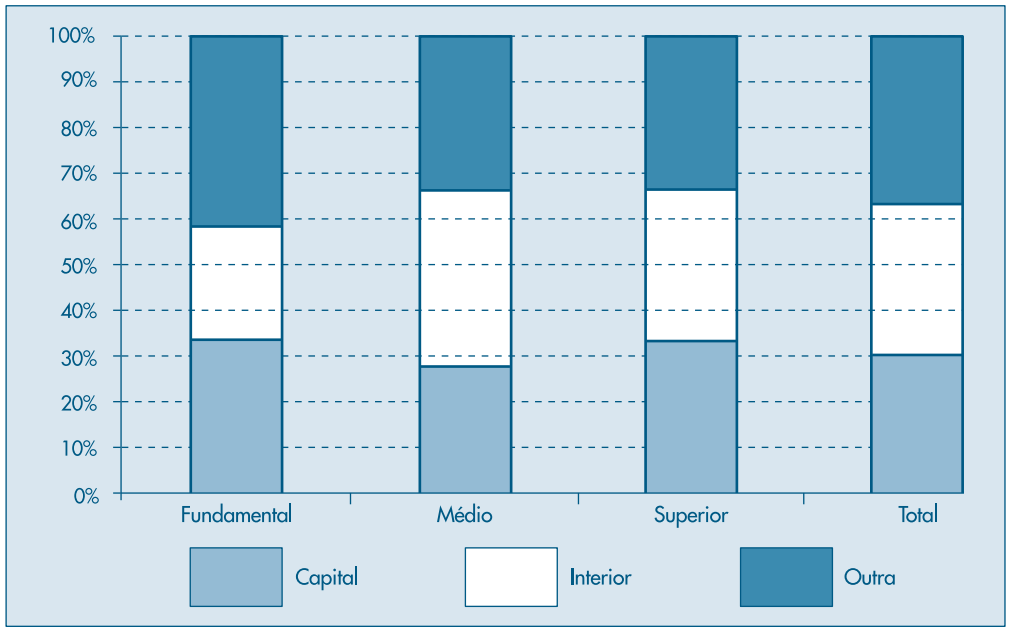
V \ Y	Fundamental	Médio	Superior	Total
Capital	33%	28%	33%	31%
Interior	25%	39%	33%	33%
Outra	42%	33%	34%	36%
Total	100%	100%	100%	100%

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 $648 \times 0{,}24 = 157$ e no Paraná seria $301 \times 0{,}24 = 73$ (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.

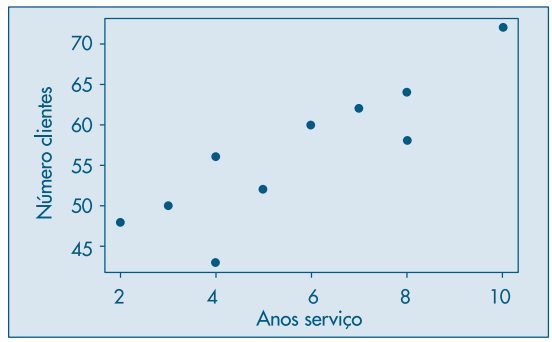
Exemplo 4.4. Na Figura 4.2 temos o gráfico de dispersão das variáveis X e Y da Tabela 4.12. Nesse tipo de gráfico temos os possíveis pares de valores (x, y), na ordem que aparecem. Para o exemplo, vemos que parece haver uma associação entre as variáveis, porque no conjunto, à medida que aumenta o tempo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

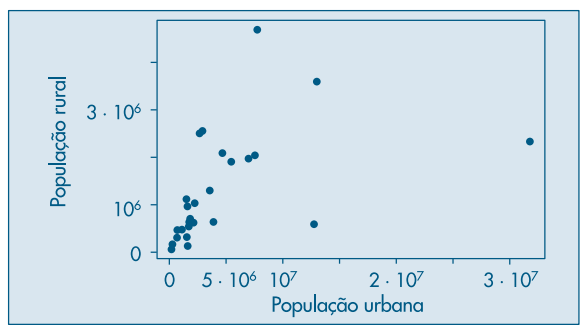
Fonte: Dados hipotéticos.

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.

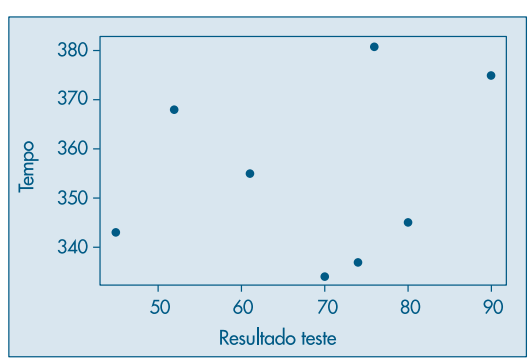


Tabela 4.14: Resultado de um teste (X) e tempo de operação de máquina (Y) para oito indivíduos.

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.

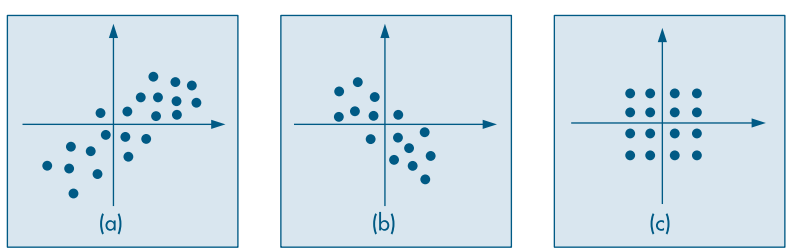
Os dados estão na Tabela 4.14. Do gráfico de dispersão (Figura 4.5) concluímos que parece não haver associação entre as duas variáveis, pois conhecer o resultado do teste não ajuda a prever o tempo gasto para aprender a operar a máquina.

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 $(X,Y) = 8{,}769/10 = 0{,}877$.

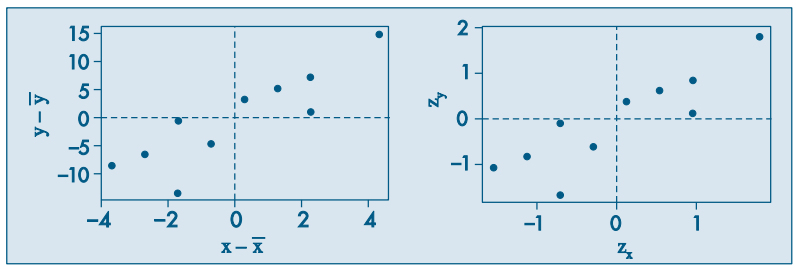
Tabela 4.15: Cálculo do coeficiente de correlação.

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.

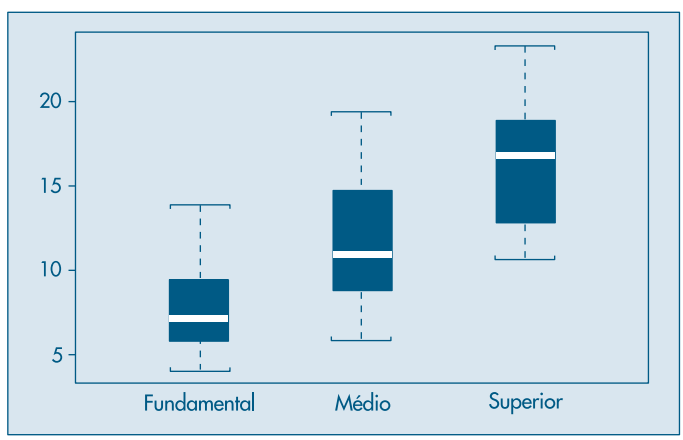
Exemplo 4.8. Retomemos os dados da Tabela 2.1, para os quais desejamos analisar agora o comportamento dos salários dentro de cada categoria de grau de instrução, ou seja, investigar o comportamento conjunto das variáveis S e Y.

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.



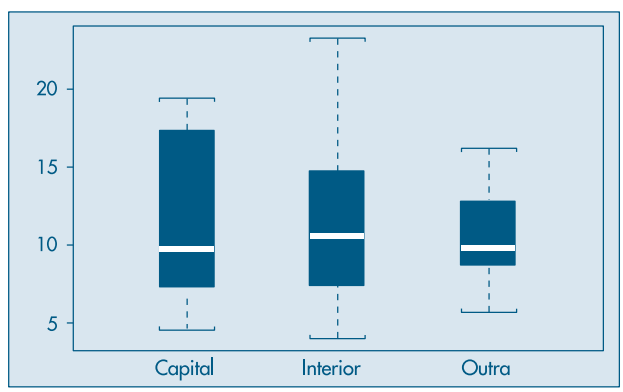
A leitura desses resultados sugere uma dependência dos salários em relação ao grau de instrução: o salário aumenta conforme aumenta o nível de educação do indivíduo. O salário médio de um funcionário é 11,12 (salários mínimos), já para um funcionário com curso superior o salário médio passa a ser 16,48, 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 $\overline{\text{var}(S)} \leq \text{var}(S)$, 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 $0 \leq R^2 \leq 1$. 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.

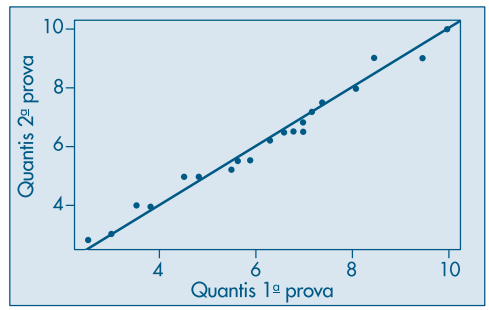


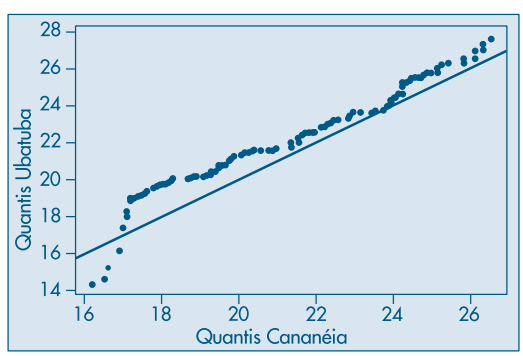
Tabela 4.18: Notas de 20 alunos em duas provas de Estatística.

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.



Problemas

16. Faça o gráfico $q \times q$ para as notas em Redação e Economia dos 25 funcionários da MB Indústria e Comércio (Problema 9 do Capítulo 2).

17. Faça o gráfico $q \times q$ para as variáveis *salário de professor secundário* e *salário de administrador* do CD-Salários. Comente.

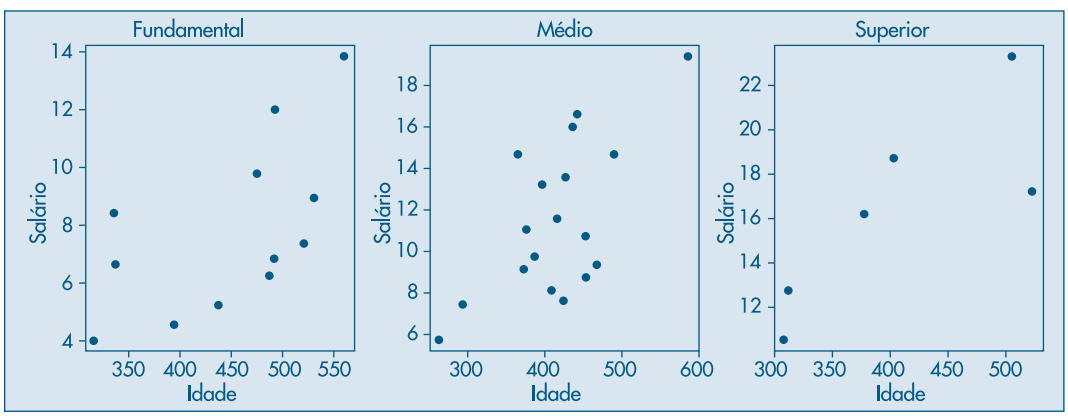
4.8 Exemplos Computacionais

Vamos considerar brevemente nesta seção o caso de mais de dois conjuntos de dados. Exemplos são os dados sobre o Brasil, de poluição e estatísticas sobre veículos, encontrados nos Conjuntos de Dados. Veremos, também, um exemplo de cálculo do coeficiente de correlação para dados reais da Bolsa de Valores de São Paulo.

Vejamos um exemplo em que temos duas variáveis quantitativas e uma qualitativa.

Exemplo 4.12. Considere as variáveis salário, idade e grau de instrução da Tabela 2.1. Separamos, agora, os salários e idades por classe de grau de instrução. Depois, podemos fazer gráficos de dispersão, como na Figura 4.12.

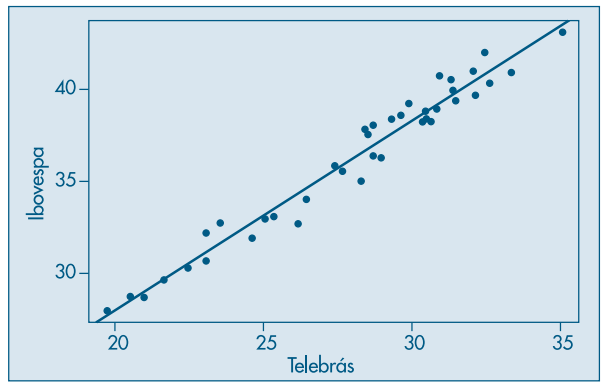
Figura 4.12: Gráficos de dispersão das variáveis salário e idade, segundo a variável grau de instrução.



Notamos que para o ensino fundamental e grau superior os salários aumentam em geral com a idade, ao passo que para o ensino médio essa relação não se verifica, havendo salários baixos e altos numa faixa entre 350 e 450 meses.

Exemplo 4.13. Considere o CD-Mercado, no qual temos os preços de fechamento diários de ações da Telebrás (X) e os índices IBOVESPA (Y), de 2 de janeiro a 24 de fevereiro de 1995, num total de $n = 39$ observações. O gráfico de dispersão está na Figura 4.13, que mostra que os pares de valores estão dispostos ao longo de uma reta com inclinação positiva. Ou seja, esse gráfico mostra que há uma forte correlação entre o preço das ações da Telebrás e o índice da Bolsa de Valores de São Paulo. No gráfico está representada a "reta de mínimos quadrados". No Capítulo 16 veremos como determiná-la.

Figura 4.13: Gráfico de dispersão para ações da Telebrás e BOVESPA.



Utilizando (4.9) obtemos que

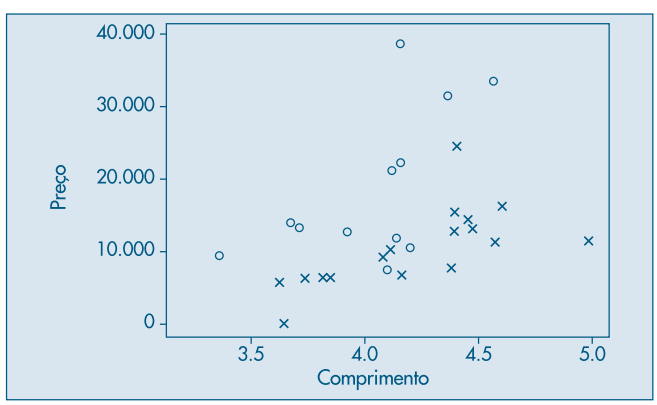
$$\text{corr}(X,Y) = \frac{40213{,}78 - (39)(27{,}99)(36{,}28)}{\sqrt{(31135{,}93 - (39)(27{,}99)^2)(51999{,}68 - (39)(36{,}28)^2)}} = 0{,}98,$$

o que mostra a forte associação linear entre X e Y.

Finalizamos esta seção com um tipo de gráfico que também é útil quando temos duas variáveis quantitativas e uma qualitativa.

Exemplo 4.14. Considere o CD-Veículos, no qual temos o preço, o comprimento e a capacidade do motor de veículos vendidos no Brasil, classificados em duas categorias: N (nacionais) e I (importados). Podemos fazer um *gráfico de dispersão simbólico* de preços e comprimentos, indicando por um x se o carro for N e por um ◦, se for I. Veja a Figura 4.14. Observamos, pela figura, que os preços dos veículos importados são, em geral, maiores do que os nacionais e que o preço aumenta com o comprimento.

Figura 4.14: Gráfico de dispersão simbólico das variáveis preço e comprimento de veículos, categorizadas pela variável procedência: nacional (x) e importado (o).



4.9 Problemas e Complementos

18. No estudo de uma certa comunidade, verificou-se que:
    (I) A proporção de indivíduos solteiros é de 0,4.
    (II) A proporção de indivíduos que recebem até 10 salários mínimos é de 0,2.
    (III) A proporção de indivíduos que recebem até 20 salários mínimos é de 0,7.
    (IV) A proporção de indivíduos casados entre os que recebem mais de 20 salários mínimos é de 0,7.
    (V) A proporção de indivíduos que recebem até 10 salários mínimos entre os solteiros é de 0,3.
    (a) Construa a distribuição conjunta das variáveis estado civil e faixa salarial e as respectivas distribuições marginais.
    (b) Você diria que existe relação entre as duas variáveis consideradas?

19. Uma amostra de 200 habitantes de uma cidade foi escolhida para declarar sua opinião sobre um certo projeto governamental. O resultado foi o seguinte:

Opinião	Local de residência			Total
	Urbano	Suburbano	Rural	
A favor	30	35	35	100
Contra	60	25	15	100
Total	90	60	50	200

    (a) Calcule as proporções em relação ao total das colunas.
    (b) Você diria que a opinião independe do local de residência?
    (c) Encontre uma medida de dependência entre as variações.

20. Com base na tabela abaixo, você concluiria que o tipo de atividade está relacionado ao fato de as embarcações serem de propriedade estatal ou particular? Encontre uma medida de dependência entre as variáveis.

Propriedade	Atividade			Total
	Costeira	Fluvial	Internacional	
Estatal	5	141	51	197
Particular	92	231	48	371
Total	97	372	99	568

Fonte: Sinopse Estatística do Brasil — IBGE — 1975.

21. Uma pesquisa sobre a participação em atividades esportivas de adultos moradores nas proximidades de centros esportivos construídos pelo estado de São Paulo mostrou os resultados da tabela abaixo. Baseado nesses resultados você diria que a participação em atividades esportivas depende da cidade?

Participam	Cidade			
	São Paulo	Campinas	Rib. Preto	Santos
Sim	50	65	105	120
Não	150	185	195	180

22. Uma pesquisa para verificar a tendência dos alunos a prosseguir os estudos, segundo a classe social do respondente, mostrou o seguinte quadro:

Pretende continuar?	Classe social			Total
	Alta	Média	Baixa	
Sim	200	220	380	800
Não	200	280	720	1.200

(a) Você diria que a distribuição de respostas afirmativas é igual à de respostas negativas?

(b) Existe dependência entre os dois fatores? Dê uma medida quantificadora da dependência.

(c) Se dos 400 alunos da classe alta 160 escolhessem continuar e 240 não, você mudaria sua conclusão? Justifique.

23. Refaça os cálculos do Problema 19 usando as fórmulas derivadas em (4.2) — (4.3).

24. Prove que $\frac{1}{n}\sum_i\left(\frac{x_i-\bar{x}}{dp(x)}\right)\left(\frac{y_i-\bar{y}}{dp(y)}\right) = \frac{\sum x_iy_i - n\bar{x}\bar{y}}{\sqrt{(\sum x_i^2 - n\bar{x}^2)(\sum y_i^2 - n\bar{y}^2)}}$.

25. Numa amostra de cinco operários de uma dada empresa foram observadas duas variáveis: X: anos de experiência num dado cargo e Y: tempo, em minutos, gasto na execução de uma certa tarefa relacionada com esse cargo.
As observações são apresentadas na tabela abaixo:

X	1	2	4	4	5
Y	7	8	3	2	2

$\sum x = 16$ $\quad \sum x^2 = 62$
$\sum y = 22$ $\quad \sum y^2 = 130$
$\sum xy = 53$

Você diria que a variável X pode ser usada para explicar a variação de Y? Justifique.

26. Muitas vezes a determinação da capacidade de produção instalada para certo tipo de indústria em certas regiões é um processo difícil e custoso. Como alternativa, pode-se estimar a capacidade de produção através da escolha de uma outra variável de medida mais fácil e que esteja linearmente relacionada com ela.
Suponha que foram observados os valores para as variáveis: capacidade de produção instalada, potência instalada e área construída. Com base num critério estatístico, qual das variáveis você escolheria para estimar a capacidade de produção instalada?

X: cap. prod. inst. (ton.)	4	5	4	5	8	9	10	11	12	12
Y: potência inst. (1.000 kW)	1	1	2	3	3	5	5	6	6	6
Z: área construída (100 m)	6	7	10	10	11	9	12	10	11	14

$\sum x = 80,$ $\quad \sum y = 38,$ $\quad \sum z = 100,$
$\sum x^2 = 736,$ $\quad \sum y^2 = 182,$ $\quad \sum z^2 = 1.048,$
$\sum xy = 361,$ $\quad \sum xz = 848,$ $\quad \sum yz = 411.$

27. Usando os dados da Tabela 2.1, Capítulo 2:
    (a) Construa a tabela de distribuições de freqüências conjunta para as variáveis salário e idade, mas divida cada uma delas num certo número de intervalos de classe.
    (b) Como poderia ser calculado o coeficiente de correlação baseado nessa tabela?
    (c) Você conseguiria "escrever" a fórmula da correlação para dados agrupados?

28. Lançam-se, simultaneamente, uma moeda de um real e uma de um quarto de dólar. Em cada tentativa anotou-se o resultado, cujos dados estão resumidos na tabela abaixo.

1 Real / 1/4 dólar	Cara	Coroa	Total
Cara	24	22	46
Coroa	28	26	54
Total	52	48	100

Fonte: Experimento conduzido pelos autores.

    (a) Esses dados sugerem que os resultados da moeda de um real e as de um quarto de dólar estão associados?
    (b) Atribua para ocorrência cara o valor 0 e para a ocorrência de coroa o valor 1. Chamando de X_1 o resultado do real e de X_2 o resultado do quarto de dólar, calcule a correlação entre X_1 e X_2. Essa medida está de acordo com a resposta que você deu anteriormente?

29. Uma amostra de dez casais e seus respectivos salários anuais (em s.m.) foi colhida num certo bairro conforme vemos na tabela abaixo.

	Casal nº	1	2	3	4	5	6	7	8	9	10
Salário	Homem (X)	10	10	10	15	15	15	15	20	20	20
	Mulher (Y)	5	10	10	5	10	10	15	10	10	15

Sabe-se que:

$\sum_{i=1}^{10} X_i = 150,$ $\sum_{i=1}^{10} X_i^2 = 2.400,$

$\sum_{i=1}^{10} X_i Y_i = 1.550,$ $\sum_{i=1}^{10} Y_i = 100,$

$\sum_{i=1}^{10} Y_i^2 = 1.100.$

    (a) Encontre o salário anual médio dos homens e o seu desvio padrão.
    (b) Encontre o salário anual médio das mulheres e o seu desvio padrão.
    (c) Construa o diagrama de dispersão.
    (d) Encontre a correlação entre o salário anual dos homens e o das mulheres.
    (e) Qual o salário médio familiar? E a variância do salário familiar?
    (f) Se o homem é descontado em 8% e a mulher em 6%, qual o salário líquido anual médio familiar? E a variância?

30. O departamento de vendas de certa companhia foi formado há um ano com a admissão de 15 vendedores.

Nessa época, foram observados para cada um dos vendedores os valores de três variáveis:

T: resultado em um teste apropriado para vendedores;

E: anos de experiência de vendas;

G: conceito do gerente de venda, quanto ao currículo do candidato.

O diretor da companhia resolveu agora ampliar o quadro de vendedores e pede sua colaboração para responder a algumas perguntas. Para isso, ele lhe dá informações adicionais sobre duas variáveis:

V: volume médio mensal de vendas em s.m.;

Z: zona da capital para a qual o vendedor foi designado.

O quadro de resultados é o seguinte:

Vendedor	T: teste	E: experiência	G: conceito do gerente	V: vendas	Z: zona
1	8	5	Bom	54	Norte
2	9	2	Bom	50	Sul
3	7	2	Mau	48	Sul
4	8	1	Mau	32	Oeste
5	6	4	Bom	30	Sul
6	8	4	Bom	30	Oeste
7	5	3	Bom	29	Norte
8	5	3	Bom	27	Norte
9	6	1	Mau	24	Oeste
10	7	3	Mau	24	Oeste
11	4	4	Bom	24	Sul
12	7	2	Mau	23	Norte
13	3	3	Mau	21	Sul
14	5	1	Mau	21	Oeste
15	3	2	Bom	16	Norte

Dados: $\sum T = 91$ $\quad \sum T^2 = 601$ $\quad \sum TV = 2.959$

$\sum E = 40$ $\quad \sum E^2 = 128$ $\quad \sum EV = 1.260$

$\sum V = 453$ $\quad \sum V^2 = 15.509$

Mais especificamente, o diretor lhe pede que responda aos sete itens seguintes:

(a) Faça o histograma da variável V em classes de 10, tendo por limite inferior da primeira classe o valor 15.

(b) Encontre a média e a variância da variável V. Suponha que um vendedor seja considerado excepcional se seu volume de vendas é dois desvios padrões superior à média geral. Quantos vendedores excepcionais existem na amostra?

(c) O diretor de vendas anunciou que transferirá para outra praça todos os vendedores cujo volume de vendas for inferior ao 1º quartil da distribuição. Qual o volume mínimo de vendas que um vendedor deve realizar para não ser transferido?

(d) Os vendedores argumentam com o diretor que esse critério não é justo, pois há zonas de venda privilegiadas. A quem você daria razão?

(e) Qual das três variáveis observadas na admissão do pessoal é mais importante para julgar um futuro candidato ao emprego?

(f) Qual o grau de associabilidade entre o conceito do gerente e a zona a que o vendedor foi designado? Você tem explicação para esse resultado?

(g) Qual o grau de associação entre o conceito do gerente e o resultado do teste? E entre zona e vendas?

31. A seção de assistência técnica da Companhia MB tem cinco funcionários: A, B, C, D e E, cujos tempos de serviço na companhia são, respectivamente, um, três, cinco, cinco e sete anos.

(a) Faça um gráfico representando a distribuição de freqüência dos tempos de serviço X.

(b) Calcule a média $me(X)$, a variância $var(X)$ e a mediana $md(X)$.

Duas novas firmas, a Verde e a Azul, solicitaram o serviço de assistência técnica da Milsa. Um mesmo funcionário pode ser designado para atender a ambos os pedidos, ou dois funcionários podem fazê-lo. Assim, o par (A, B) significa que o funcionário A atenderá à firma Verde e o funcionário B, à firma Azul.

(c) Escreva os 25 possíveis pares de funcionários para atender a ambos os pedidos.

(d) Para cada par, calcule o tempo médio de serviço $\overline{X}$, faça a distribuição de freqüência e uma representação gráfica. Compare com o resultado de *(a)*.

(e) Calcule para os 25 valores de $\overline{X}$ os parâmetros $me(\overline{X})$, $var(\overline{X})$ e $md(\overline{X})$. Compare com os resultados obtidos em *(b)*. Que tipo de conclusão você poderia tirar?

(f) Para cada par obtido em *(c)*, calcule a variância do par e indique-a por S^2. Faça a representação gráfica da distribuição dos valores de S^2.

(g) Calcule $me(S^2)$ e $var(S^2)$.

(h) Indicando por X_1 a variável que expressa o tempo de serviço do funcionário que irá atender à firma Verde e X_2 o que irá atender à firma Azul, faça a distribuição conjunta da variável bidimensional (X_1, X_2).

(i) As duas variáveis X_1 e X_2 são independentes?

(j) O que você pode falar sobre as distribuições "marginais" de X_1 e X_2?

(l) Suponha agora que três firmas solicitem o serviço de assistência técnica. Quantas triplas podem ser formadas?

(m) Sem calcular todas as possibilidades, como você acha que ficaria o histograma de $\overline{X}$? E $me(\overline{X})$? e $var(\overline{X})$?

(n) E sobre a variável S^2?

(o) A variável tridimensional (X_1, X_2, X_3) teria alguma propriedade especial para as suas distribuições "marginais"?

32. Refaça o problema anterior, admitindo agora que um mesmo funcionário não pode atender a duas firmas.

33. Gráficos quantis × quantis. Na seção 4.5 vimos como construir um gráfico $q \times q$ quando $m = n$. Suponha $n > m$, isto é, temos um número maior de observações de X. Então, usamos as observações ordenadas $y_{(1)} \leqslant \ldots \leqslant y_{(m)}$ e interpolamos um conjunto correspondente de quantis para o conjunto dos x_i ordenados. O valor ordenado $y_{(i)}$ corresponde a $p_i = \dfrac{i-0{,}5}{m}$. Para X, queremos um valor j tal que

$$\frac{j-0{,}5}{n} = \frac{i-0{,}5}{m},$$

logo

$$j = \frac{n}{m}(i-0{,}5) + 0{,}5.$$

Se j for inteiro, fazemos o gráfico de $y_{(i)}$ *versus* $x_{(j)}$.
Se $j = k + r$, onde k é inteiro e $0 < r < 1$, então

$$q_x\left(\frac{i-0{,}5}{m}\right) = (1-r)x_{(k)} + r \cdot x_{(k+1)}.$$

Exemplo: Se $m = 20$ e $n = 40$,

$$j = \frac{40}{20}(i-0{,}5) + 0{,}5 = 2i - 0{,}5,$$

logo $k = 2i - 1$, $r = 0{,}5$, e fazemos o gráfico de

$$y_{(1)} \quad \textit{versus} \quad [0{,}5x_{(1)} + 0{,}5x_{(2)}],$$
$$y_{(2)} \quad \textit{versus} \quad [0{,}5x_{(3)} + 0{,}5x_{(4)}] \text{ etc.}$$

34. Faça o gráfico $q \times q$ para os dois conjuntos de dados em A e B a seguir.

A	65	54	49	60	70	25	87	100	70	102	40	47
B	48	35	45	50	52	20	72	102	46	82	—	—

35. Faça gráficos de dispersão unidimensionais e *box plots* para a variável salário da Tabela 2.1, segundo a região de procedência. Analise os resultados.

36. Analise as variáveis salário e idade da Tabela 2.1, segundo o estado civil de cada indivíduo. Quais conclusões você pode obter?

37. Analise a população total do CD-Brasil, segundo as regiões geográficas.

38. Considere os dados do Exemplo 4.14 e o seguinte critério: valores abaixo da média indicam mercado em BAIXA e valores maiores ou iguais à média indicam mercado em ALTA. Categorize os dados segundo esse critério e apresente os resultados numa tabela de dupla entrada. Calcule uma medida de associação. O valor obtido corrobora ou não o resultado obtido no Exemplo 4.14? Comente.

39. Considere o CD-Poluição e as variáveis CO, temperatura e umidade. Faça gráficos de dispersão para pares de variáveis. Quais conclusões você pode obter?