Associação entre Variáveis Qualitativas e Quantitativas

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**.

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

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

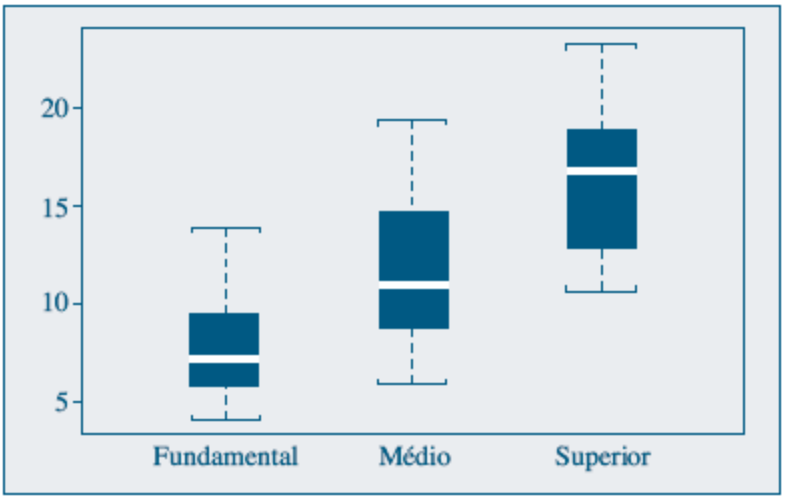
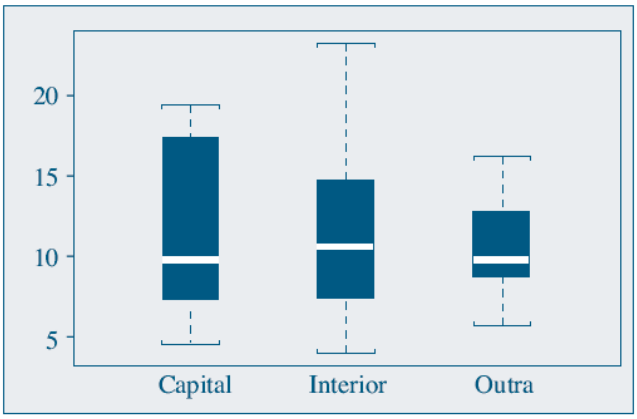


Tabela 4.17 Medidas-resumo para a variável salário segundo a região de procedência, na Companhia MB.

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.



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,

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

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.

Gráficos q × q

o *gráfico quantis × quantis*

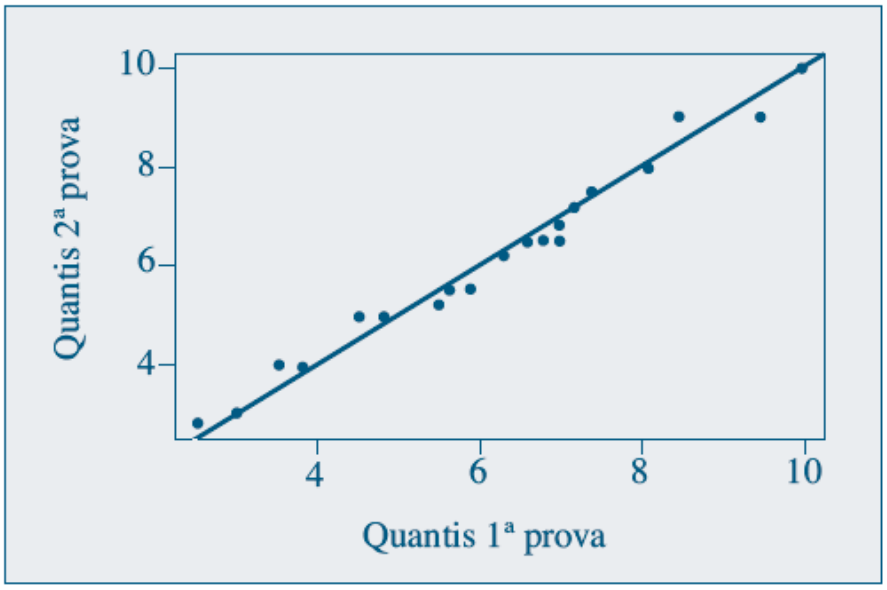
Suponha que temos valores x_1, ..., x_n da variável X e valores y_1, ..., y_m da variável Y, todos medidos pela mesma unidade. Por exemplo, temos temperaturas de duas cidades ou alturas de dois grupos de indivíduos etc. O gráfico $q \times q$ é um gráfico dos quantis de X contra os quantis de Y.

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.

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

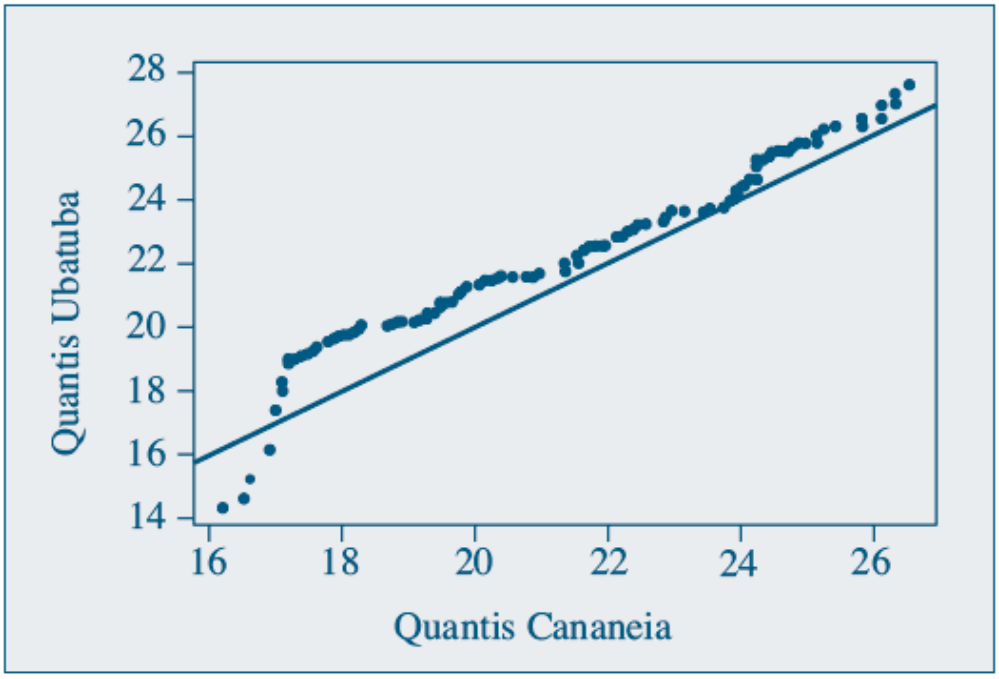
Aluno	Prova 1	Prova 2	Aluno	Prova 1	Prova 2
1	8,5	8,0	11	7,4	6,5
2	3,5	2,8	12	5,6	5,0
3	7,2	6,5	13	6,3	6,5
4	5,5	6,2	14	3,0	3,0
5	9,5	9,0	15	8,1	9,0
6	7,0	7,5	16	3,8	4,0
7	4,8	5,2	17	6,8	5,5
8	6,6	7,2	18	10,0	10,0
9	2,5	4,0	19	4,5	5,5
10	7,0	6,8	20	5,9	5,0

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



Exemplo 4.11 Consideremos, agora, as variáveis *temperatura de Ubatuba* e *temperatura de Cananeia*, 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 Cananeia, para valores maiores do que 17 graus.

Figura 4.11 Gráfico $q \times q$ para os lados de temperatu-ra de Cananeia e Ubatuba.



Variáveis aleatórias discretas

Função de probabilidade conjunta

Sejam X e Y duas variáveis aleatórias discretas originárias do mesmo fenômeno aleatório, com valores atribuídos a partir do mesmo espaço amostral. A função de probabilidade conjunta é definida, para todos os possíveis pares de valores de (X, Y), da seguinte forma:

$$p(x, y) = P(X = x, Y = y) = P((X=x) \cap (Y=y))$$

Exemplo 8.1. Suponha que estamos interessados em estudar a composição de famílias com três crianças, quanto ao sexo. Definamos:

X = número de meninos,

$$Y = \begin{cases} 1, \text{ se o primeiro filho for homem} \\ 0, \text{ se o primeiro filho for mulher,} \end{cases}$$

Z = número de vezes em que houve variação do sexo entre um nascimento e outro, dentro da mesma família.

Tabela 8.1: Composição de famílias com três crianças, quanto ao sexo.

Eventos	Probabilidade	X	Y	Z
HHH	1/8	3	1	0
HHM	1/8	2	1	1
HMH	1/8	2	1	2
MHH	1/8	2	0	1
HMM	1/8	1	1	1
MHM	1/8	1	0	2
MMH	1/8	1	0	1
MMM	1/8	0	0	0

Tabela 8.2: Distribuições de probabilidades unidimensionais.

(a)

x	0	1	2	3
$p(x)$	1/8	3/8	3/8	1/8

(b)

y	0	1
$p(y)$	1/2	1/2

(c)

z	0	1	2
$p(z)$	1/4	1/2	1/4

A Tabela 8.3 apresenta as probabilidades associadas aos pares de valores nas variáveis X e Y. Nessa tabela, $p(x, y) = P(X = x, Y = y)$ denota a probabilidade do evento $\{X = x \text{ e } Y = y\} = \{X = x\} \cap \{Y = y\}$. Essa tabela é denominada *distribuição conjunta* de X e Y.

Tabela 8.3: Distribuição bidimensional da v.a. (X, Y).

(x, y)	$p(x, y)$
(0, 0)	1/8
(1, 0)	2/8
(1, 1)	1/8
(2, 0)	1/8
(2, 1)	2/8
(3, 1)	1/8

Tabela 8.5: Distribuição conjunta de X e Y, como uma tabela de dupla entrada.

Y \ X	0	1	2	3	$p(y)$
0	1/8	2/8	1/8	0	1/2
1	0	1/8	2/8	1/8	1/2
$p(x)$	1/8	3/8	3/8	1/8	1

Distribuições Marginais e Condicionais

Da Tabela 8.5 podemos obter facilmente as distribuições de X e Y. A primeira e última colunas da tabela dão a distribuição de Y, $(y, p(y))$, enquanto a primeira e última linhas da tabela dão a distribuição de X, $(x, p(x))$. Essas distribuições são chamadas *distribuições marginais*.

Observamos, por exemplo, que

$$P(X = 1) = P(X = 1, Y = 0) + P(X = 1, Y = 1) = 2/8 + 1/8 = 3/8$$

e

$$P(Y = 0) = P(X = 0, Y = 0) + P(X = 1, Y = 0) + P(X = 2, Y = 0) + P(X = 3, Y = 0)$$
$$= 1/8 + 2/8 + 1/8 + 0 = 1/2.$$

Portanto, para obter as probabilidades marginais basta somar linhas e colunas.

Quando estudamos os aspectos descritivos das distribuições com mais de uma variável, vimos que, às vezes, é conveniente calcular proporções em relação a uma linha ou coluna, e não em relação ao total. Isso é equivalente aqui ao conceito de distribuição condicional.

Por exemplo, qual seria a distribuição do número de meninos, sabendo-se que o primeiro filho é do sexo masculino?

$$P(X = x|Y = 1) = \frac{P(X = x, Y = 1)}{P(Y = 1)} = p(x|Y = 1), \qquad (8.1)$$

para $x = 0, 1, 2, 3$. Pela Tabela 8.5 obtemos, por exemplo,

$$p(2|Y = 1) = P(X = 2|Y = 1) = \frac{P(X = 2, Y = 1)}{P(Y = 1)} = \frac{2/8}{1/2} = 1/2.$$

Do mesmo modo, obtemos as demais probabilidades, e a *distribuição condicional* de X, *dado que* $Y = 1$, está na Tabela 8.6.

Tabela 8.6: Distribuição condicional de X, dado que $Y = 1$.

x	1	2	3
$p(x\|Y = 1)$	1/4	1/2	1/4

Observe que $\sum_x p(x|Y = 1) = p(0|Y = 1) + \ldots + p(3|Y = 1) = 1$.

Do mesmo modo, podemos obter a distribuição condicional de Y, dado que $X = 2$, que está na Tabela 8.7.

Tabela 8.7: Distribuição condicional de Y, dado que $X = 2$.

y	0	1
$p(y\|X = 2)$	1/3	2/3

Podemos generalizar o que foi dito acima para duas v.a. X e Y quaisquer, assumindo os valores $x_1, x_2, \ldots, x_n$ e $y_1, y_2, \ldots, y_m$, respectivamente.

Definição. Seja x_i, um valor de X, tal que $P(X = x_i) = p(x_i) > 0$. A probabilidade

$$P(Y = y_j|X = x_i) = \frac{P(X = x_i, Y = y_j)}{P(X = x_i)}, \quad j = 1, ..., m, \tag{8.2}$$

é denominada *probabilidade condicional* de $Y = y_j$, *dado que* $X = x_i$.

Como observamos acima, para x_i fixado, os pares $(y_j, P(Y = y_j|X = x_i))$, $j = 1, ..., m$, definem a distribuição condicional de Y, dado que $X = x_i$, pois

$$\sum_{j=1}^{m} P(Y = y_j|X = x_i) = \sum_{j=1}^{m} \frac{P(Y = y_j, X = x_i)}{P(X = x_i)} = \frac{P(X = x_i)}{P(X = x_i)} = 1.$$

Considere a distribuição condicional de X, dado que $Y = 1$, da Tabela 8.6. Podemc calcular a média dessa distribuição, a saber

$$E(X|Y = 1) = 1 \times \frac{1}{4} + 2 \times \frac{1}{2} + 3 \times \frac{1}{4} = 2.$$

Observe que $E(X) = 1{,}5$, ao passo que $E(X|Y = 1) = 2$.

Definição. A esperança condicional de X, dado que $Y = y_j$, é definida por

$$E(X|Y = y_j) = \sum_{i=1}^{n} x_i P(X = x_i|Y = y_j).$$

Uma definição análoga vale para $E(Y|X = x_i)$.

Exemplo 8.2. Para a distribuição condicional de Y, dado que $X = 2$, da Tabela 8.7, temos

$$E(Y|X = 2) = 0 \times \frac{1}{3} + 1 \times \frac{2}{3} = \frac{2}{3}.$$

Exemplo 8.3. Considere, agora, a distribuição conjunta das variáveis Y e Z, definidas no Exemplo 8.1. Da Tabela 8.1 obtemos a Tabela 8.8. Aqui, observamos que

Tabela 8.1: Composição de famílias com três crianças, quanto ao sexo.

Eventos	Probabilidade	X	Y	Z
HHH	1/8	3	1	0
HHM	1/8	2	1	1
HMH	1/8	2	1	2
MHH	1/8	2	0	1
HMM	1/8	1	1	1
MHM	1/8	1	0	2
MMH	1/8	1	0	1
MMM	1/8	0	0	0

Tabela 8.8: Distribuição conjunta de Y e Z.

Y \ Z	0	1	2	$p(y)$
0	1/8	2/8	1/8	1/2
1	1/8	2/8	1/8	1/2
$p(z)$	1/4	2/4	1/4	1

$$P(Z = z|Y = y) = \frac{P(Z = z, Y = y)}{P(Y = y)} = P(Z = z)$$

para quaisquer $z = 0, 1, 2$ e $y = 0, 1$. O que significa dizer que

$$P(Z = z, Y = y) = P(Z = z)\, P(Y = y),$$

isto é, a probabilidade de cada casela é igual ao produto das respectivas probabilidades marginais. Por exemplo,

$$P(Z = 1, Y = 1) = \frac{2}{8} = \frac{2}{4} \times \frac{1}{2} = P(Z = 1)P(Y = 1).$$

Também é verdade que

$$P(Y = y|Z = z) = P(Y = y)$$

para todos os valores de y e z. Dizemos que Y e Z são independentes.

Definição. As variáveis aleatórias X e Y, assumindo os valores $x_1, x_2, \ldots$ e $y_1, y_2, \ldots$, respectivamente, são independentes se, e somente se, para todo par de valores (x_i, y_j) de X e Y, tivermos que

$$P(X = x_i, Y = y_j) = P(X = x_i)\, P(Y = y_j). \qquad (8.3)$$

Basta que (8.3) não se verifique para *um* par (x_i, y_j), para que X e Y *não* sejam independentes. Nesse caso diremos que X e Y são *dependentes*.

Essa definição pode ser estendida para mais de duas variáveis aleatórias.

Exemplos:

2. A tabela abaixo dá a distribuição conjunta de X e Y.
   (a) Determine as distribuições marginais de X e Y.
   (b) Obtenha as esperanças e variâncias de X e Y.
   (c) Verifique se X e Y são independentes.
   (d) Calcule $P(X = 1|Y = 0)$ e $P(Y = 2|X = 3)$.
   (e) Calcule $P(X \leqslant 2)$ e $P(X = 2, Y \leqslant 1)$.

Y \ X	1	2	3
0	0,1	0,1	0,1
1	0,2	0	0,3
2	0	0,1	0,1

3. Considere a distribuição conjunta de X e Y, parcialmente conhecida, dada na tabela abaixo.
   (a) Complete a tabela, considerando X e Y independentes.
   (b) Calcule as médias e variâncias de X e Y.
   (c) Obtenha as distribuições condicionais de X, dado que $Y = 0$, e de Y, dado que $X = 1$.

Y \ X	−1	0	1	$P(Y = y)$
−1	1/12			
0				1/3
1	1/4		1/4	
$P(X = x)$				1