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.

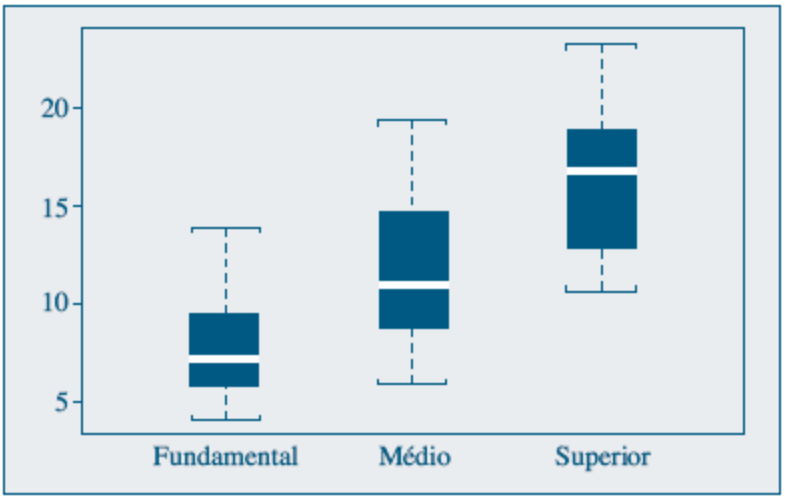
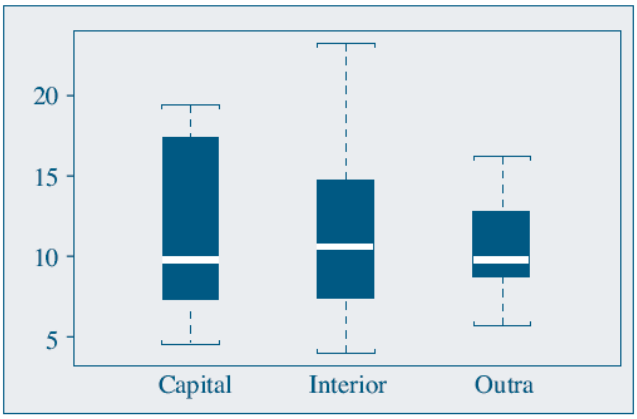


Tabela 4.17 Medidas-resumo para a variável salário segundo a região de procedência, na Companhia MB.

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*

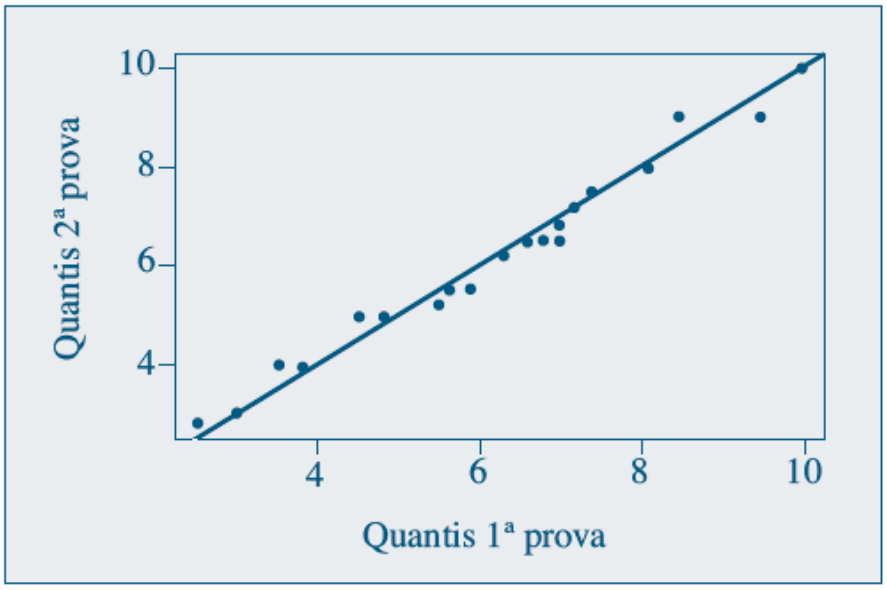
Suponha que temos valores x_1, ..., x_n da variável X e valores y_1, ..., y_m da variável Y, todos medidos pela mesma unidade. Por exemplo, temos temperaturas de duas cidades ou alturas de dois grupos de indivíduos etc. O gráfico $q \times q$ é um gráfico dos quantis de X contra os quantis de Y.

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.

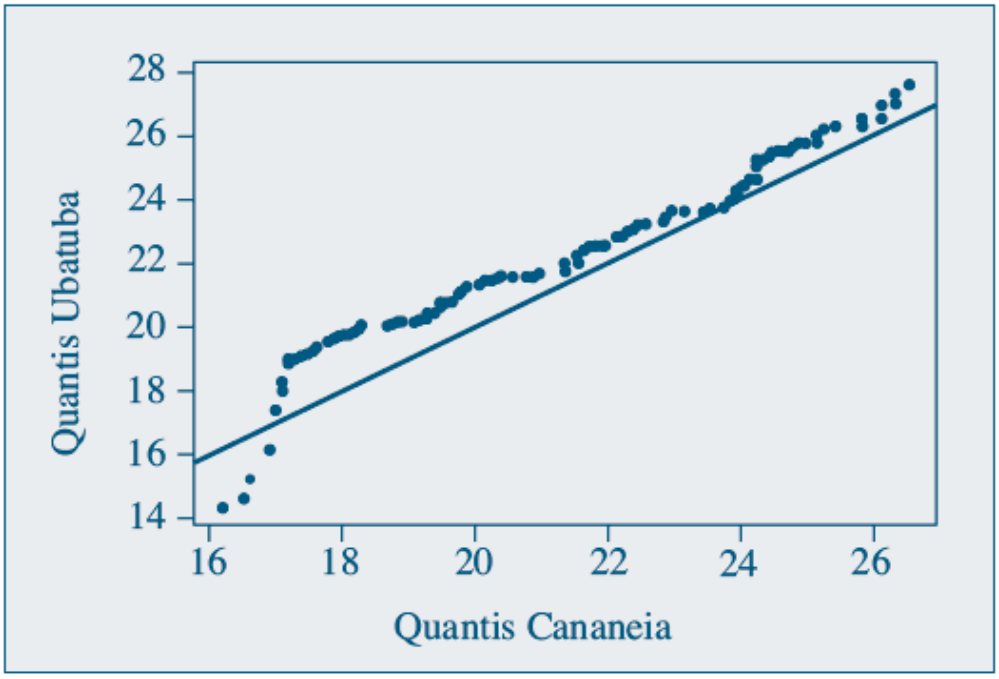
Aluno	Prova 1	Prova 2	Aluno	Prova 1	Prova 2
1	8,5	8,0	11	7,4	6,5
2	3,5	2,8	12	5,6	5,0
3	7,2	6,5	13	6,3	6,5
4	5,5	6,2	14	3,0	3,0
5	9,5	9,0	15	8,1	9,0
6	7,0	7,5	16	3,8	4,0
7	4,8	5,2	17	6,8	5,5
8	6,6	7,2	18	10,0	10,0
9	2,5	4,0	19	4,5	5,5
10	7,0	6,8	20	5,9	5,0

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.



Probabilidades

Exemplo 5.1. Queremos estudar as frequências de ocorrências das faces de um dado. Um procedimento a adotar seria lançar o dado certo número de vezes, n, e depois contar o número n_i de vezes em que ocorre a face i, $i = 1, 2, ..., 6$. As proporções n_i/n determinam a distribuição de frequências do experimento realizado. Lançando o dado um número $n'(n' \neq n)$ de vezes, teríamos outra distribuição de frequências, mas com um padrão que esperamos ser muito próximo do anterior.

Tabela 5.1 Modelo para lançamento de um dado.

Face	1	2	3	4	5	6	Total
Frequência teórica	1/6	1/6	1/6	1/6	1/6	1/6	1

Exemplo 5.2 De um grupo de duas mulheres (M) e três homens (H), uma pessoa será sorteada para presidir uma reunião. Queremos saber as probabilidades de o presidente ser do sexo masculino ou feminino. Observamos que: (i) só existem duas possibilidades: ou a pessoa sorteada é do sexo masculino (H) ou é do sexo feminino (M); (ii) supondo que o sorteio seja honesto e que cada pessoa tenha igual chance de ser sorteada, teremos o modelo probabilístico da Tabela 5.2 para o experimento.

Tabela 5.2 Modelo teórico para o Exemplo 5.2.

Sexo	*M*	*H*	Total
Frequência teórica	2/5	3/5	1

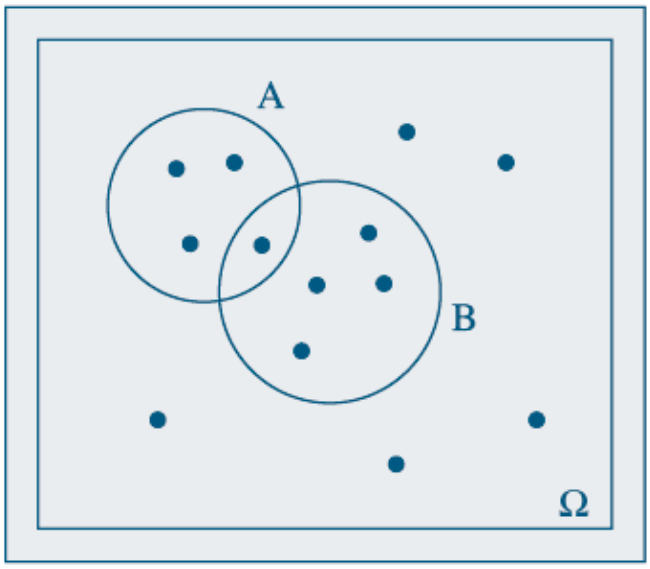
(a) um *espaço amostral*, Ω, que consiste, no caso discreto, da enumeração (finita ou infinita) de todos os resultados possíveis do experimento em questão:

$$\Omega = \{\omega_1, \omega_2, ..., \omega_n, ...\}$$

(os elementos de Ω são os *pontos amostrais* ou *eventos elementares*);

(b) uma *probabilidade*, $P(\omega)$, para cada ponto amostral, de tal sorte que seja possível encontrar a probabilidade $P(A)$ de qualquer subconjunto A de Ω, isto é, a probabilidade do que chamaremos de um *evento aleatório* ou simplesmente *evento*.

Figura 5.1 Espaço amostral e eventos aleatórios.



Exemplo 5.3 Lançamos uma moeda duas vezes. Se C indicar cara e R indicar coroa, então um espaço amostral será

$$\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$$

em que $\omega_1 = (C, C)$, $\omega_2 = (C, R)$, $\omega_3 = (R, C)$, $\omega_4 = (R, R)$. É razoável supor que cada ponto ω_i tenha probabilidade 1/4, se a moeda for perfeitamente simétrica e homogênea.

Se designarmos por A o evento que consiste na obtenção de faces iguais nos dois lançamentos, então

$$P(A) = P\{\omega_1, \omega_4\} = 1/4 + 1/4 = 1/2.$$

De modo geral, se A for qualquer evento de Ω, então

$$P(A) = \sum_j P(\omega_j), \tag{5.1}$$

em que a soma é estendida a todos os pontos amostrais $\omega_j \in A$.

Exemplo 5.4 Uma fábrica produz determinado artigo. Da linha de produção são retirados três artigos, e cada um é classificado como bom (B) ou defeituoso (D). Um espaço amostral do experimento é

$$\Omega = \{\text{BBB, BBD, BDB, DBB, DDB, DBD, BDD, DDD}\}.$$

Se A designar o evento que consiste em obter dois artigos defeituosos, então $A = \{\text{DDB, DBD, BDD}\}$.

Exemplo 5.5 Considere o experimento que consiste em retirar uma lâmpada de um lote e medir seu "tempo de vida" antes de se queimar. Um espaço amostral conveniente é

$$\Omega = \{t \in \mathbb{R} : t \geq 0\},$$

isto é, o conjunto de todos os números reais não negativos. Se A indicar o evento "o tempo de vida da lâmpada é inferior a 20 horas", então $A = \{t : 0 \leq t < 20\}$. Esse é um exemplo de um espaço amostral *contínuo*, contrastado com os anteriores, que são *discretos*.