

Capítulo 1

O material de análise multivariada

1.1 Exemplos de dados multivariados

Os métodos estatísticos descritos em textos elementares são na maioria métodos univariados porque tratam somente da análise de variação em uma única variável aleatória. Por outro lado, o ponto principal de uma análise multivariada é considerar várias variáveis relacionadas simultaneamente, sendo todas consideradas igualmente importantes, pelo menos inicialmente. O valor potencial dessa abordagem mais geral pode ser visto por meio de alguns exemplos.

Exemplo 1.1 Pardais sobreviventes de tempestade

Após uma forte tempestade em 1º de fevereiro de 1898, diversos pardais moribundos foram levados ao laboratório biológico de Hermon Bumpus na Universidade de Brown em Rhode Island. Subsequentemente, cerca de metade dos pássaros morreu, e Bumpus viu isso como uma oportunidade de encontrar suporte para a teoria da seleção natural de Charles Darwin. Para esse fim, ele fez oito medidas morfológicas em cada pássaro, e também os pesou. Os resultados de cinco das medidas são mostrados na Tabela 1.1, para fêmeas somente.

Dos dados que obteve, Bumpus (1898) concluiu que “os pássaros que morreram, morreram não por acidente, mas porque eles eram fisicamente desqualificados, e que os pássaros que sobreviveram, sobreviveram porque possuíam certas características físicas”. Especificamente, ele verificou que os sobreviventes “são mais curtos e pesam menos ... têm os ossos das asas mais longos, pernas mais longas, esternos mais longos e maior capacidade cerebral” do que os não sobreviventes. Concluiu também que “o processo de eliminação seletiva é mais severo com indivíduos extremamente variáveis, não importando em qual direção a variação possa ocorrer. É tão perigoso estar acima de um certo padrão de excelência orgânica quanto estar visivelmente abaixo do padrão”. Isso queria dizer que ocorreu seleção estabilizadora, de modo que indivíduos com medidas próximas da média sobrevivem melhor do que indivíduos com medidas distantes da média.

De fato, o desenvolvimento dos métodos de análise multivariada havia recém-iniciado em 1898, quando Bumpus estava escrevendo. O coeficiente de correlação como uma medida do relacionamento entre duas variáveis foi delineado por Francis Galton em 1877. Entretanto, decorreram outros 56

Tabela 1.1 Medidas do corpo de pardocas (em mm)

Pássaro	X_1	X_2	X_3	X_4	X_5
1	156	245	31,6	18,5	20,5
2	154	240	30,4	17,9	19,6
3	153	240	31,0	18,4	20,6
4	153	236	30,9	17,7	20,2
5	155	243	31,5	18,6	20,3
6	163	247	32,0	19,0	20,9
7	157	238	30,9	18,4	20,2
8	155	239	32,8	18,6	21,2
9	164	248	32,7	19,1	21,1
10	158	238	31,0	18,8	22,0
11	158	240	31,3	18,6	22,0
12	160	244	31,1	18,6	20,5
13	161	246	32,3	19,3	21,8
14	157	245	32,0	19,1	20,0
15	157	235	31,5	18,1	19,8
16	156	237	30,9	18,0	20,3
17	158	244	31,4	18,5	21,6
18	153	238	30,5	18,2	20,9
19	155	236	30,3	18,5	20,1
20	163	246	32,5	18,6	21,9
21	159	236	31,5	18,0	21,5
22	155	240	31,4	18,0	20,7
23	156	240	31,5	18,2	20,6
24	160	242	32,6	18,8	21,7
25	152	232	30,3	17,2	19,8
26	160	250	31,7	18,8	22,5
27	155	237	31,0	18,5	20,0
28	157	245	32,2	19,5	21,4
29	165	245	33,1	19,8	22,7
30	153	231	30,1	17,3	19,8
31	162	239	30,3	18,0	23,1
32	162	243	31,6	18,8	21,3
33	159	245	31,8	18,5	21,7
34	159	247	30,9	18,1	19,0
35	155	243	30,9	18,5	21,3
36	162	252	31,9	19,1	22,2

(continua)

Tabela 1.1 Medidas do corpo de pardocas (em mm) (*continuação*)

Pássaro	X_1	X_2	X_3	X_4	X_5
37	152	230	30,4	17,3	18,6
38	159	242	30,8	18,2	20,5
39	155	238	31,2	17,9	19,3
40	163	249	33,4	19,5	22,8
41	163	242	31,0	18,1	20,7
42	156	237	31,7	18,2	20,3
43	159	238	31,5	18,4	20,3
44	161	245	32,1	19,1	20,8
45	155	235	30,7	17,7	19,6
46	162	247	31,9	19,1	20,4
47	153	237	30,6	18,6	20,4
48	162	245	32,5	18,5	21,1
49	164	248	32,3	18,8	20,9

Nota: X_1 = comprimento total, X_2 = extensão alar, X_3 = comprimento do bico e cabeça, X_4 = comprimento do úmero, X_5 = comprimento da quilha do esterno. Pássaros de 1 a 21 sobreviveram, pássaros de 22 a 49 morreram.

Fonte: Dados de Bumpus, H.C., *Biological Lectures*, Marine Biology Laboratory, Woods Hole, MA, 1898.

anos antes de Harold Hotelling descrever um método prático para realizar uma análise de componentes principais, a qual é uma das análises multivariada mais simples que pode ser aplicada aos dados de Bumpus. Bumpus não calculou nem mesmo os desvios padrão. Apesar disso, seus métodos de análise foram sensíveis. Muitos autores têm reanalisado seus dados e, em geral, têm confirmado suas conclusões.

Tomando os dados como um exemplo para ilustrar métodos multivariados, surgem muitas questões interessantes. Em particular:

1. Como estão relacionadas as várias variáveis? Por exemplo, um valor grande para uma das variáveis tende a ocorrer com valores grandes para as outras variáveis?
2. Os sobreviventes e os não sobreviventes apresentam diferenças estatisticamente significantes para seus valores médios das variáveis?
3. Os sobreviventes e não sobreviventes mostram quantidades similares de variação para as variáveis?
4. Se os sobreviventes e não sobreviventes diferem em termos das distribuições das variáveis, é possível construir alguma função dessas variáveis que separe os dois grupos? Então seria conveniente se valores grandes da função tendessem a ocorrer com os sobreviventes enquanto que a função seria então aparentemente um índice de ajuste darwiniano dos pardais.

Exemplo 1.2 Crânios egípcios

Para um segundo exemplo, considere os dados mostrados na Tabela 1.2 para medidas feitas em crânios masculinos da área de Tebas, no Egito. Há cinco amostras de 30 crânios cada uma do período pré-dinástico primitivo (cerca de 4000 a.C.), do período pré-dinástico antigo (cerca de 3300 a.C.), das 12^a e 13^a dinastias (cerca de 1850 a.C.), do período Ptolemaico (cerca de 200 a.C.) e do período Romano (cerca de 150 d.C.). Quatro medidas são apresentadas para cada crânio, como ilustrado na Figura 1.1.

Para este exemplo, algumas questões interessantes são:

1. Como estão relacionadas as quatro medidas?
2. Existem diferenças estatisticamente significantes nas médias amostrais das variáveis, e, se existem, essas diferenças refletem mudanças graduais ao longo do tempo na forma e no tamanho dos crânios?
3. Existem diferenças significantes nos desvios padrão amostrais para as variáveis, e, se existem, essas diferenças refletem mudanças graduais ao longo do tempo na quantidade de variação?
4. É possível construir uma função das quatro variáveis que, em algum sentido, descreva as mudanças ao longo do tempo?

Essas questões são, claramente, bastante similares àquelas sugeridas para o Exemplo 1.1.

Veremos mais adiante que existem diferenças entre as cinco amostras que podem ser explicadas parcialmente como tendências no tempo. É preciso ser dito, entretanto, que as razões para as aparentes mudanças são desconhecidas. Migração de outras raças dentro da região pode muito bem ter sido o fator mais importante.

Exemplo 1.3 Distribuição de uma borboleta

Um estudo de 16 colônias de borboletas *Euphydryas editha* na Califórnia e em Oregon produziu os dados apresentados na Tabela 1.3. Aqui existem quatro variáveis ambientais (altitude, precipitação anual e temperaturas máxima e mínima) e seis variáveis genéticas (frequências percentuais para diferentes genes (Fósforo glucose-isomerase) como determinado pela técnica de eletroforese). Para os objetivos deste exemplo, não há necessidade de entrar em detalhes de como as frequências gênicas foram determinadas e, estritamente falando, elas não são exatamente frequências gênicas. É suficiente dizer que as frequências descrevem, de certa forma, a distribuição genética das borboletas. A Figura 1.2 mostra as localizações geográficas das colônias.

Neste exemplo, questões que podem ser feitas incluem:

1. As frequências Pgi são similares para as colônias que estão próximas no espaço?
2. O quanto, se este for o caso, as frequências Pgi estão relacionadas às variáveis ambientais?

Tabela 1.2 Medidas de crânios egípcios masculinos (em mm)

Crânios	Pré-dinástico primitivo				Pré-dinástico antigo				12ª e 13ª dinastias				Período ptolomaico				Período romano			
	X ₁	X ₂	X ₃	X ₄	X ₁	X ₂	X ₃	X ₄	X ₁	X ₂	X ₃	X ₄	X ₁	X ₂	X ₃	X ₄	X ₁	X ₂	X ₃	X ₄
1	131	138	89	49	124	138	101	48	137	141	96	52	137	134	107	54	137	123	91	50
2	125	131	92	48	133	134	97	48	129	133	93	47	141	128	95	53	136	131	95	49
3	131	132	99	50	138	134	98	45	132	138	87	48	141	130	87	49	128	126	91	57
4	119	132	96	44	148	129	104	51	130	134	106	50	135	131	99	51	130	134	92	52
5	136	143	100	54	126	124	95	45	134	134	96	45	133	120	91	46	138	127	86	47
6	138	137	89	56	135	136	98	52	140	133	98	50	131	135	90	50	126	138	101	52
7	139	130	108	48	132	145	100	54	138	138	95	47	140	137	94	60	136	138	97	58
8	125	136	93	48	133	130	102	48	136	145	99	55	139	130	90	48	126	126	92	45
9	131	134	102	51	131	134	96	50	136	131	92	46	140	134	90	51	132	132	99	55
10	134	134	99	51	133	125	94	46	126	136	95	56	138	140	100	52	139	135	92	54
11	129	138	95	50	133	136	103	53	137	129	100	53	132	133	90	53	143	120	95	51
12	134	121	95	53	131	139	98	51	137	139	97	50	134	134	97	54	141	136	101	54
13	126	129	109	51	131	136	99	56	136	126	101	50	135	135	99	50	135	135	95	56
14	132	136	100	50	138	134	98	49	137	133	90	49	133	136	95	52	137	134	93	53
15	141	140	100	51	130	136	104	53	129	142	104	47	136	130	99	55	142	135	96	52
16	131	134	97	54	131	128	98	45	135	138	102	55	134	137	93	52	139	134	95	47
17	135	137	103	50	138	129	107	53	129	135	92	50	131	141	99	55	138	125	99	51
18	132	133	93	53	123	131	101	51	134	125	90	60	129	135	95	47	137	135	96	54
19	139	136	96	50	130	129	105	47	138	134	96	51	136	128	93	54	133	125	92	50
20	132	131	101	49	134	130	93	54	136	135	94	53	131	125	88	48	145	129	89	47

(continua)

Tabela 1.2 Medidas de crânios egípcios masculinos (em mm) (continuação)

Crânios	Pré-dinástico primitivo				Pré-dinástico antigo				12ª e 13ª dinastias				Período ptolemaico				Período romano			
	X ₁	X ₂	X ₃	X ₄	X ₁	X ₂	X ₃	X ₄	X ₁	X ₂	X ₃	X ₄	X ₁	X ₂	X ₃	X ₄	X ₁	X ₂	X ₃	X ₄
21	126	133	102	51	137	136	106	49	132	130	91	52	139	130	94	53	138	136	92	46
22	135	135	103	47	126	131	100	48	133	131	100	50	144	124	86	50	131	129	97	44
23	134	124	93	53	135	136	97	52	138	137	94	51	141	131	97	53	143	126	88	54
24	128	134	103	50	129	126	91	50	130	127	99	45	130	131	98	53	134	124	91	55
25	130	130	104	49	134	139	101	49	136	133	91	49	133	128	92	51	132	127	97	52
26	138	135	100	55	131	134	90	53	134	123	95	52	138	126	97	54	137	125	85	57
27	128	132	93	53	132	130	104	50	136	137	101	54	131	142	95	53	129	128	81	52
28	127	129	106	48	130	132	93	52	133	131	96	49	136	138	94	55	140	135	103	48
29	131	136	114	54	135	132	98	54	138	133	100	55	132	136	92	52	147	129	87	48
30	124	138	101	46	130	128	101	51	138	133	91	46	135	130	100	51	136	133	97	51

Nota: X_1 = largura máxima, X_2 = altura basibregmática, X_3 = comprimento basialveolar, X_4 = altura nasal.
Fonte: Dados de Thomson, A. and Randall-Maciver, P., *Ancient Races of the Thebaid*, Oxford University Press, Oxford, London, 1905.

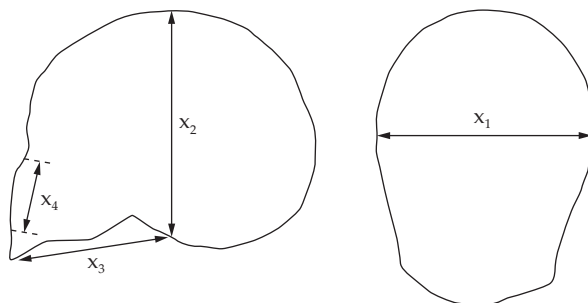


Figura 1.1 Quatro medidas feitas em crânios egípcios masculinos.

Essas são questões importantes na tentativa de decidir como as frequências P_{gi} são determinadas. Se a composição genética das colônias foi largamente determinada pelas migrações passadas e presentes, então as frequências gênicas tenderão a ser similares para colônias que estão localizadas nas proximidades, apesar de elas poderem apresentar um pequeno relacionamento com as variáveis ambientais. Por outro lado, se o meio ambiente é mais importante, então isso deve aparecer em relacionamentos entre as frequências gênicas e as variáveis ambientais (assumindo que tenham sido medidas as variáveis corretas), mas colônias próximas somente têm frequências gênicas similares se elas têm ambientes similares. Obviamente colônias que estão próximas no espaço usualmente têm ambientes similares, de modo que pode ser difícil chegar a uma conclusão sobre essa questão.

Exemplo 1.4 Cães pré-históricos da Tailândia

Escavações de locais pré-históricos no nordeste da Tailândia têm produzido uma coleção de ossos caninos cobrindo um período em torno de 3500 a.C. até o presente. Entretanto, a origem dos cães pré-históricos não é certa. Podem descender dos chacais-dourados (*Canis aureus*) ou do lobo, mas o lobo não é nativo da Tailândia. As fontes de origem mais próximas são a parte ocidental da China (*Canis lupus chanco*) ou o subcontinente indiano (*Canis lupus pallides*).

Para tentar esclarecer os ancestrais dos cães pré-históricos, foram feitas medidas da mandíbula dos espécimens disponíveis. Estas foram então comparadas com as mesmas medidas feitas no chacal-dourado, no lobo chinês e no lobo indiano. As comparações foram também estendidas para incluir o dingo, o qual tem suas origens na Índia, o cuon (*Cuon alpinus*), o qual é indígena do sudeste da Ásia, e os cães modernos de cidade da Tailândia.

A Tabela 1.4 apresenta os valores médios para as seis medidas de mandíbulas para espécimens de todos os sete grupos. A questão principal aqui é o que as medidas sugerem sobre o relacionamento entre os grupos e, em particular, como os cães pré-históricos parecem se relacionar com os outros grupos.

Tabela 1.3 Variáveis ambientais e frequências gênicas Fósforo Glucose-Isomerase (Pgi) para colônias de borboletas *Euphydryas editha* na Califórnia e em Oregon, nos EUA

Colônia	Altitude (pés)	Precipitação anual (pol.)	Temperatura (°F)		Frequências de mobilidade gênica Pgi (%) ^a						
			Máxima	Mínima	0,4	0,6	0,8	1	1,16	1,3	
SS	500	43	98	17	0	3	22	57	17	1	
SB	808	20	92	32	0	16	20	38	13	13	
WSB	570	28	98	26	0	6	28	46	17	3	
JRC	550	28	98	26	0	4	19	47	27	3	
JRH	550	28	98	26	0	1	8	50	35	6	
SJ	380	15	99	28	0	2	19	44	32	3	
CR	930	21	99	28	0	0	15	50	27	8	
UO	650	10	101	27	10	21	40	25	4	0	
LO	600	10	101	27	14	26	32	28	0	0	
DP	1.500	19	99	23	0	1	6	80	12	1	
PZ	1.750	22	101	27	1	4	34	33	22	6	
MC	2.000	58	100	18	0	7	14	66	13	0	
IF	2.500	34	102	16	0	9	15	47	21	8	
AF	2.000	21	105	20	3	7	17	32	27	14	
GH	7.850	42	84	5	0	5	7	84	4	0	
GL	10.500	50	81	-12	0	3	1	92	4	0	

Nota: Os dados originais são de 21 colônias, mas, neste presente exemplo, cinco colônias com amostras pequenas para a estimação das frequências gênicas foram excluídas para tornar todas as estimativas quase igualmente confiáveis.

^a Os números 0,40, 0,60, etc. representam diferentes tipos genéticos de Pgi, de modo que as frequências para uma colônia (somando a 100%) mostram as frequências dos diferentes tipos para a *E. editha* naquele local.

Fonte: Dados de McKechnie, S.W. et al., *Genetics*, 81, 571-594, 1975, com variáveis ambientais arredondadas para números inteiros para simplificar.

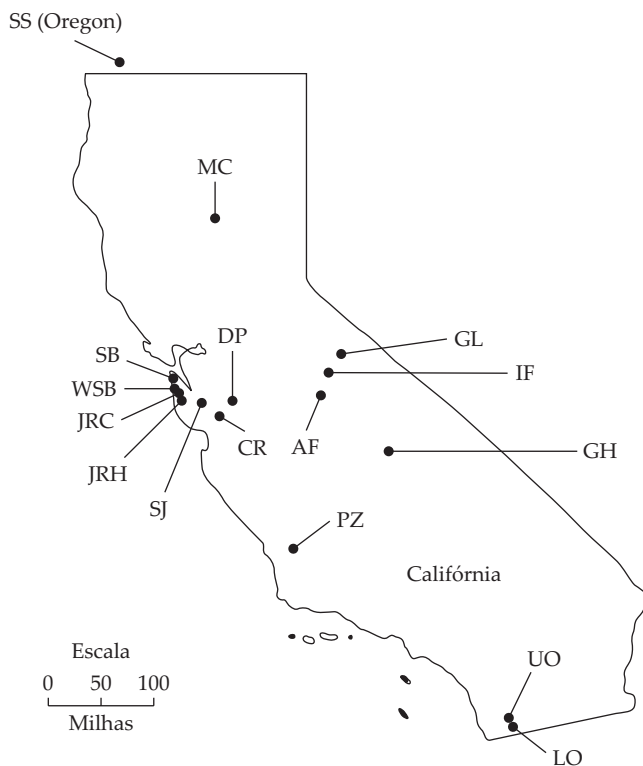


Figura 1.2 Colônias de *Euphydryas editha* na Califórnia e em Oregon.

Tabela 1.4 Médias de medidas de mandíbulas para sete grupos caninos

Grupo	X_1 (mm)	X_2 (mm)	X_3 (mm)	X_4 (mm)	X_5 (mm)	X_6 (mm)
Cão moderno	9,7	21,0	19,4	7,7	32,0	36,5
Chacal-dourado	8,1	16,7	18,3	7,0	30,3	32,9
Lobo chinês	13,5	27,3	26,8	10,6	41,9	48,1
Lobo indiano	11,5	24,3	24,5	9,3	40,0	44,6
Cuon	10,7	23,5	21,4	8,5	28,8	37,6
Dingo	9,6	22,6	21,1	8,3	34,4	43,1
Cão pré-histórico	10,3	22,1	19,1	8,1	32,2	35,0

Nota: X_1 = largura da mandíbula; X_2 = altura da mandíbula abaixo do primeiro molar; X_3 = comprimento do primeiro molar; X_4 = largura do primeiro molar; X_5 = comprimento do primeiro ao terceiro molar, inclusive; X_6 = comprimento do primeiro ao quarto molar, inclusive.

Fonte: Dados de Higham, C.F.W. et al., *J. Archaeological Sci.*, 7, 149-165, 1980.

Exemplo 1.5 Emprego em países europeus

Finalmente, como um contraste aos exemplos biológicos anteriores, considere os dados na Tabela 1.5. Eles mostram as porcentagens da força de trabalho em nove diferentes tipos de indústrias para 30 países europeus. Neste caso, métodos multivariados podem ser úteis para isolar grupos de países com padrões similares de empregos, e, em geral, ajudar a compreender os relacionamentos entre os países. Diferenças entre países que são relacionados a grupos políticos (UE, a União Europeia; AELC, a área europeia de livre comércio; países do leste europeu e outros países) podem ser de particular interesse.

1.2 Visão prévia dos métodos multivariados

Os cinco exemplos que acabamos de considerar são matérias brutas típicas para métodos estatísticos multivariados. Em todos os casos, existem várias variáveis de interesse e elas são claramente não independentes umas das outras. Neste momento, é útil dar uma breve visão prévia do que está por vir nos capítulos que seguem em relação a esses exemplos.

A *análise de componentes principais* é elaborada para reduzir o número de variáveis que necessitam ser consideradas a um número menor de índices (chamados de componentes principais) os quais são combinações lineares das variáveis originais. Por exemplo, grande parte da variação nas medidas do corpo dos pardais (X_1 a X_5) mostrada na Tabela 1.1 está relacionada ao tamanho geral dos pássaros, e o total

$$I_1 = X_1 + X_2 + X_3 + X_4 + X_5$$

deve medir muito bem esse aspecto dos dados. Este índice é responsável por uma dimensão dos dados. Outro índice é

$$I_2 = X_1 + X_2 + X_3 - X_4 - X_5,$$

o qual é um contraste entre as três primeiras medidas e as duas últimas. Este reflete outra dimensão dos dados. A análise de componentes principais fornece uma maneira objetiva de encontrar índices desse tipo, de modo que a variação nos dados pode ser levada em consideração tão concisamente quanto possível. Pode muito bem acontecer que dois ou mais componentes principais forneçam um bom resumo de todas as variáveis originais. A consideração dos valores dos componentes principais em vez dos valores das variáveis originais pode tornar muito mais fácil entender o que os dados têm a dizer. Em poucas palavras, a análise de componentes principais é um meio de simplificar dados pela redução do número de variáveis.

A *análise de fatores* também tem como objetivo estudar a variação em uma quantidade de variáveis originais usando um número menor de variáveis índices ou fatores. Assume-se que cada variável original possa ser expressa como

Tabela 1.5 Porcentagens da força de trabalho de empregados em nove diferentes grupos de indústrias em 30 países na Europa

País	Grupo	AGR	MIN	FAB	FEA	CON	SER	FIN	SSP	TC
Bélgica	UE	2,6	0,2	20,8	0,8	6,3	16,9	8,7	36,9	6,8
Dinamarca	UE	5,6	0,1	20,4	0,7	6,4	14,5	9,1	36,3	7,0
França	UE	5,1	0,3	20,2	0,9	7,1	16,7	10,2	33,1	6,4
Alemanha	UE	3,2	0,7	24,8	1,0	9,4	17,2	9,6	28,4	5,6
Grécia	UE	22,2	0,5	19,2	1,0	6,8	18,2	5,3	19,8	6,9
Irlanda	UE	13,8	0,6	19,8	1,2	7,1	17,8	8,4	25,5	5,8
Itália	UE	8,4	1,1	21,9	0,0	9,1	21,6	4,6	28,0	5,3
Luxemburgo	UE	3,3	0,1	19,6	0,7	9,9	21,2	8,7	29,6	6,8
Países Baixos	UE	4,2	0,1	19,2	0,7	0,6	18,5	11,5	38,3	6,8
Portugal	UE	11,5	0,5	23,6	0,7	8,2	19,8	6,3	24,6	4,8
Espanha	UE	9,9	0,5	21,1	0,6	9,5	20,1	5,9	26,7	5,8
Reino Unido	UE	2,2	0,7	21,3	1,2	7,0	20,2	12,4	28,4	6,5
Áustria	AELC	7,4	0,3	26,9	1,2	8,5	19,1	6,7	23,3	6,4
Finlândia	AELC	8,5	0,2	19,3	1,2	6,8	14,6	8,6	33,2	7,5
Islândia	AELC	10,5	0,0	18,7	0,9	10,0	14,5	8,0	30,7	6,7
Noruega	AELC	5,8	1,1	14,6	1,1	6,5	17,6	7,6	37,5	8,1
Suécia	AELC	3,2	0,3	19,0	0,8	6,4	14,2	9,4	39,5	7,2
Suíça	AELC	5,6	0,0	24,7	0,0	9,2	20,5	10,7	23,1	6,2
Albânia	Leste	55,5	19,4	0,0	0,0	3,4	3,3	15,3	0,0	3,0
Bulgária	Leste	19,0	0,0	35,0	0,0	6,7	9,4	1,5	20,9	7,5
República Tcheca/ Eslováquia	Leste	12,8	37,3	0,0	0,0	8,4	10,2	1,6	22,9	6,9
Hungria	Leste	15,3	28,9	0,0	0,0	6,4	13,3	0,0	27,3	8,8
Polônia	Leste	23,6	3,9	24,1	0,9	6,3	10,3	1,3	24,5	5,2
Romênia	Leste	22,0	2,6	37,9	2,0	5,8	6,9	0,6	15,3	6,8
URSS (antiga)	Leste	18,5	0,0	28,8	0,0	10,2	7,9	0,6	25,6	8,4
Iugoslávia (antiga)	Leste	5,0	2,2	38,7	2,2	8,1	13,8	3,1	19,1	7,8
Chipre	Outro	13,5	0,3	19,0	0,5	9,1	23,7	6,7	21,2	6,0
Gibraltar	Outro	0,0	0,0	6,8	2,0	16,9	24,5	10,8	34,0	5,0
Malta	Outro	2,6	0,6	27,9	1,5	4,6	10,2	3,9	41,6	7,2
Turquia	Outro	44,8	0,9	15,3	0,2	5,2	12,4	2,4	14,5	4,4

Nota: AGR, agricultura, florestal e pesca; MIN, mineração e exploração de pedreiras; FAB, fabricação; FEA, fornecimento de energia e água; CON, construção; SER, serviços; FIN, finanças; SSP, serviços sociais e pessoais; TC, transportes e comunicações.

Fonte: Dados do Euromonitor (1995), exceto para Alemanha e Reino Unido, onde valores mais recentes foram obtidos do United Nations Statistical Yearbook (2000).

uma combinação linear desses fatores, mais um termo residual que reflete o quanto a variável é independente das outras variáveis. Por exemplo, um modelo de dois fatores para os dados dos pardais assume que

$$X_1 = a_{11}F_1 + a_{12}F_2 + e_1$$

$$X_2 = a_{21}F_1 + a_{22}F_2 + e_2$$

$$X_3 = a_{31}F_1 + a_{32}F_2 + e_3$$

$$X_4 = a_{41}F_1 + a_{42}F_2 + e_4$$

e

$$X_5 = a_{51}F_1 + a_{52}F_2 + e_5$$

em que

os valores a_{ij} são constantes,

F_1 e F_2 são fatores e

e_i representa a variação em X_i que é independente da variação nas outras variáveis X .

Aqui F_1 pode ser o fator tamanho. Nesse caso, os coeficientes a_{11} , a_{21} , a_{31} , a_{41} e a_{51} seriam todos positivos, refletindo o fato de que alguns pássaros tendem a ser grandes e alguns pássaros tendem a ser pequenos em todas as medidas do corpo. O segundo fator F_2 poderia então medir um aspecto da forma dos pássaros, com alguns coeficientes positivos e alguns negativos. Se esse modelo de dois fatores ajustasse bem os dados, então ele forneceria uma descrição relativamente direta do relacionamento entre as cinco medidas do corpo que estão sendo consideradas.

Um tipo de análise de fatores começa tomando alguns poucos primeiros componentes principais como os fatores nos dados a serem considerados. Esses fatores iniciais são então modificados por um processo especial de transformação chamado *rotação fatorial*, a fim de torná-los mais fáceis de serem interpretados. Outros métodos para encontrar fatores iniciais também são usados. Uma rotação para simplificar fatores é quase sempre feita.

A *análise de função discriminante* refere-se à possibilidade de separar diferentes grupos com base nas medidas disponíveis. Isso pode ser usado, por exemplo, para ver quão bem pardais sobreviventes e não sobreviventes podem ser separados usando suas medidas do corpo (Exemplo 1.1), ou como crânios de diferentes épocas podem ser separados, novamente usando medidas de tamanho (Exemplo 1.2). Assim como a análise de componentes principais, a análise de função discriminante é baseada na ideia de encontrar combinações lineares convenientes das variáveis originais para atingir o objetivo desejado.

A *análise de agrupamento* diz respeito à identificação de grupos de objetos similares. Não há muito sentido em fazer esse tipo de análise com dados como

os dos Exemplos 1.1 e 1.2, pois os grupos (sobreviventes/não sobreviventes e épocas) já são conhecidos. No entanto, no Exemplo 1.3 poderá haver algum interesse em agrupar colônias com base nas variáveis ambientais ou frequências Pgi, enquanto que no Exemplo 1.4 o principal ponto de interesse está na similaridade entre cães pré-históricos tailandeses e outros animais. Da mesma forma, no Exemplo 1.5 os países europeus podem possivelmente ser agrupados em termos de suas similaridades no padrão de empregos.

Com *correlação canônica*, as variáveis (não os objetos) são divididas em dois grupos, e o interesse está centrado no relacionamento entre elas. Então, no Exemplo 1.3, as primeiras quatro variáveis estão relacionadas ao ambiente, enquanto que as restantes seis variáveis refletem a distribuição genética nas diferentes colônias de *Euphydryas editha*. Encontrar quais relacionamentos, se houver algum, existem entre esses dois grupos de variáveis é de considerável interesse biológico.

O *escalonamento multidimensional* começa com dados sobre algumas medidas de distâncias entre um certo número de objetos. Destas distâncias, é então construído um mapa mostrando como estes objetos estão relacionados. Essa é uma técnica útil, pois muitas vezes é possível medir o quão distante estão pares de objetos sem ter nenhuma ideia de como estes objetos estão relacionados em um sentido geométrico. Assim, no Exemplo 1.4, existem maneiras de medir as distâncias entre cães modernos e chacais-dourados, cães modernos e lobos chineses, etc. Considerando cada par de grupos de animais, resultam 21 distâncias ao todo, e destas distâncias o escalonamento multidimensional pode ser usado para produzir um tipo de mapa do relacionamento entre os grupos. Com um mapa unidimensional, os grupos são colocados ao longo de uma linha reta. Com um mapa bidimensional, eles são representados por pontos em um plano. Com um mapa tridimensional, eles são representados por pontos dentro de um cubo. Soluções de quarta dimensão ou mais alta também são possíveis, apesar de terem uso limitado porque não podem ser visualizadas de uma maneira simples. O valor de um mapa de uma, duas ou três dimensões está claro para o Exemplo 1.4, pois tal mapa mostraria imediatamente quais grupos de cães pré-históricos são mais similares. Portanto, o escalonamento multidimensional pode ser uma alternativa útil para análise de agrupamento nesse caso. Um mapa de países europeus baseado em seus padrões de empregos também pode ser de interesse no Exemplo 1.5.

Análise de componentes principais e escalonamento multidimensional são algumas vezes referidos como métodos de *ordenação*. Isso quer dizer que são métodos para produzir eixos nos quais um conjunto de objetos de interesse pode ser representado. Outros métodos de ordenação estão também disponíveis.

A *análise de coordenadas principais* é como um tipo de análise de componentes principais que inicia com informações sobre o quanto os pares de objetos são diferentes em um conjunto de objetos em vez dos valores das medidas dos objetos. Como tal, ela pretende fazer o mesmo que o escalonamento multidimensional. Entretanto, as suposições feitas e os métodos numéricos usados não são os mesmos.

A *análise de correspondência* começa com dados sobre a abundância de cada uma das várias características para cada um de um conjunto de objetos. Isso é útil em ecologia, por exemplo, em que os objetos de interesse são muitas vezes diferentes locais, as características são diferentes espécies e os dados consistem em abundâncias de espécies em amostras tomadas dos locais. O propósito da análise de correspondência seria então o de tornar claro os relacionamentos entre os locais, expressos por distribuições das espécies, e os relacionamentos entre as espécies, expressos por distribuições dos locais.

1.3 A distribuição normal multivariada

A distribuição normal para uma única variável deve ser familiar para os leitores deste livro. Ela tem a curva de frequências na bem-conhecida forma de sino, e muitos métodos estatísticos univariados padrão são baseados na suposição de que os dados são normalmente distribuídos.

Sabendo da proeminência da distribuição normal com métodos estatísticos univariados, não será surpresa descobrir que a distribuição normal multivariada tem uma posição central nos métodos estatísticos multivariados. Muitos desses métodos requerem a suposição de que os dados que estão sendo analisados têm uma distribuição normal multivariada.

A exata definição de uma distribuição normal multivariada não é tão importante. A abordagem de muitas pessoas, para melhor ou pior, parece ser a de pensar os dados como sendo normalmente distribuídos, a menos que exista alguma razão para acreditar que isso não é verdadeiro. Em particular, se todas as variáveis individuais que estão sendo estudadas parecem ser normalmente distribuídas, então assume-se que a distribuição conjunta é normal multivariada. Esta é, de fato, uma exigência mínima, porque a definição de normalidade multivariada requer mais do que isso.

Casos surgem em que a suposição de normalidade multivariada é claramente inválida. Por exemplo, uma ou mais das variáveis que estão sendo estudadas pode ter uma distribuição altamente assimétrica com vários valores muito altos (ou baixos); pode haver muitos valores repetidos; etc. Esse tipo de problema pode ser algumas vezes superado por uma transformação de dados apropriada, como discutido nos textos elementares em estatística. Se isso não funcionar, então uma forma bastante especial de análise poderá ser necessária.

Um aspecto importante da distribuição normal multivariada é que ela é completamente especificada por um vetor de médias e uma matriz de covariâncias. As definições de um vetor de médias e uma matriz de covariâncias são dadas na Seção 2.7. Basicamente, o vetor de médias contém os valores médios para todas as variáveis que estão sendo consideradas, enquanto que a matriz de covariâncias contém as variâncias para todas as variáveis mais as covariâncias, as quais medem o quanto todos os pares de variáveis estão relacionados.

1.4 Programas computacionais

Métodos práticos para executar os cálculos para análises multivariadas têm sido desenvolvidos ao longo dos últimos 80 anos. Entretanto, a aplicação desses métodos para mais do que um pequeno número de variáveis teve de esperar até que os computadores se tornassem disponíveis. Portanto, foi somente nos últimos 40 anos, aproximadamente, que os métodos se tornaram razoavelmente fáceis de serem executados pelo pesquisador médio.

Hoje existem muitos pacotes estatísticos padrão e programas computacionais disponíveis para cálculos em computadores de todos os tipos. A intenção é que este livro forneça aos leitores informação suficiente para usar de forma inteligente qualquer um desses pacotes e programas, sem falar muito sobre qualquer um deles. Entretanto, dada a ampla acessibilidade da linguagem de programação R, enfatizamos seu uso. Portanto, o apêndice deste capítulo fornece uma breve revisão do ambiente R necessária para rodar os comandos R incluídos ao final de cada capítulo. Os códigos em R para muitos dos exemplos também estão disponíveis no endereço <http://www.manlybio.myob.net/> e no site da Bookman Editora (loja.grupo.com.br). Para acessar todo o conteúdo disponível no site da Bookman Editora, encontre a página do livro por meio do campo de busca e localize a área de Material Complementar. Um tratamento mais amplo da linguagem R pode ser encontrado em muitos livros e manuais, incluindo aqueles dos autores Adler (2012), Crawley (2013), Logan (2010), Teetor (2011) e Venables et al. (2015). Referências adicionais no uso do R para análise multivariada também são fornecidas nos capítulos seguintes.

Referências

- Adler, J. (2012). *R in a Nutshell*. 2nd Edn. Sebastopol, CA: O'Reilly Media.
- Bumpus, H.C. (1898). The elimination of the unfit as illustrated by the introduced sparrow, *Passer domesticus*. *Biological Lectures*, 11th Lecture. Marine Biology Laboratory, Woods Hole, MA, 209–26.
- Crawley, M. (2013). *The R Book*. 2nd Edn. Chichester: Wiley.
- Euromonitor (1995). *European Marketing Data and Statistics*. London: Euromonitor Publications.
- Higham, C.F.W., Kijngam, A., and Manly, B.F.J. (1980). An analysis of prehistoric canid remains from Thailand. *Journal of Archaeological Science* 7: 149–65.
- Logan, M. (2010). *Biostatistical Design and Analysis Using R: A Practical Guide*. Chichester: Wiley.
- McKechnie, S.W., Ehrlich, P.R., and White, R.R. (1975). Population genetics of *Euphydryas* butterflies. I. Genetic variation and the neutrality hypothesis. *Genetics* 81: 571–94.
- Teetor, P. (2011). *R Cookbook*. Sebastopol, CA: O'Reilly Media.
- Thomson, A. and Randall-Maciver, R. (1905). *Ancient Races of the Thebaid*. Oxford, London: Oxford University Press.
- United Nations (2000). *Statistical Yearbook*, 44th Issue. New York: Department of Social Affairs.
- Venables, W.N., Smith, D.M. and the R Core Team. (2015). *An Introduction to R: Notes on R, A Programming Environment for Data Analysis and Graphics, Version 3.1.3* (2015-03-09). <http://cran.r-project.org/doc/manuals/R-intro.pdf>

Apêndice: Uma introdução ao R

No site do Comprehensive R Archive Network (CRAN, <http://www.r-project.org>), o R é descrito como “um software ambiente *gratuito* para cálculos e gráficos estatísticos... mantido pelo R Development Core Team”. Hoje, o R lidera um papel na ciência, nos negócios e na tecnologia como um sistema computacional por excelência com manuseio, cálculos de procedimentos estatísticos e telas gráficas. Pode-se instalar o R em qualquer sistema operacional (Windows, Mac ou Linux). Se você tem um sistema Mac ou Windows, pode querer executar o instalador, baixando do site espelho de sua preferência. Você apenas precisa seguir as instruções dadas em www.r-project.org. Habitualmente, uma nova versão do R é publicada em março e setembro de cada ano, mas você deve ficar ciente da última versão anunciada na seção News do site r-project.

A.1 A interface gráfica do usuário R

O R tem uma interface gráfica do usuário (GUI) que é adequada à entrada de códigos de programação, para obter resultados numéricos mostrados em um console e gerar novas janelas que surgem repentinamente (os *dispositivos gráficos*) onde gráficos são produzidos. Uma vez que o R seja iniciado, você verá a janela R Console contendo informações sobre a versão do R que você está executando, além de outros aspectos do R. Então, imediatamente o símbolo > aparece, indicando que o R está no modo interativo. Isso significa que o R está esperando para que um comando seja digitado e executado. Outra forma de trabalhar no R, a alternativa mais usual, é por meio do editor de rotina, acessível no menu File. Sempre que o editor de rotina for solicitado, um texto simples surge no editor, tal que você está apto a escrever sequências de comandos R, com a vantagem de que ele pode ser salvo em um arquivo no padrão ASCII, fácil de ser lido por qualquer editor de texto. Usando a linha de procedimentos executados Run ou selecionados ou todos os Run, encontrados no menu Edit, conjuntos de códigos R ou rotinas inteiras podem ser passados para o R.

É evidente que a funcionalidade rudimentar do padrão GUI do R o torna pouco atraente para aqueles usuários familiarizados com pacotes estatísticos comuns caracterizados por suas interfaces amigáveis. Como melhorias substitutas ao ambiente GUI do R e seu editor de rotina, várias aplicações estão disponíveis para facilitar o acesso aos menus e tornar a tarefa de escrever rotinas no R mais fácil. Entre esses projetos, as duas aplicações RStudio (RStudio Team 2015) e TinnR (Faria et al. 2015) sobressaem. O R é sensível a maiúsculas e minúsculas, e programas em R requerem o frequente uso de parênteses () e colchetes [] e assim por diante. TinnR e RStudio fazem uso de cores para indicar correspondentes parênteses abertos e fechados, colchetes e assim por diante, bem como diferenciar entre sentenças da linguagem R, argumentos de funções e comentários. Nenhum desses recursos está disponível no editor de rotina do GUI do R. Você deve, dessa forma, provavelmente escolher uma dessas aplicações amigáveis para melhorar sua interação com o R.

A.2 Arquivos de ajuda do R

O R inclui um sistema de ajuda, que é útil em fornecer informações básicas sobre cada comando do R, como sua sintaxe, o resultado/objeto produzido, exemplos de uso e referências de onde o comando é baseado. É essencial contar com esse sistema de ajuda padrão sempre que existir dúvida sobre a sintaxe correta dos comandos do R. No GUI do R, você pode acessar o sistema de ajuda como uma opção da barra de menu, escolhendo diferentes estágios de pesquisa. Na realidade, o sistema de ajuda fornece mais informações do que apenas a sintaxe de comando. Você pode também acessar a documentação do R e manuais, aprender sobre o desenvolvimento do R e obter respostas às questões frequentemente feitas (FAQ).

A.3 Pacotes do R

Um *pacote* do R é um conjunto de funções R relacionadas/integradas, arquivos de ajuda e conjunto de dados que o usuário ou o próprio R chama e disponibiliza para um propósito particular. Com a exceção de um grupo de pacotes já instalados com o R, o usuário necessitará baixar e instalar um pacote de interesse para o R. A instalação é realizada apenas uma vez, e cada vez que um pacote desse tipo é requerido, o usuário somente necessita carregá-lo no ambiente R para a corrente sessão. O principal repositório de pacote público está disponível nos sites espelhos no mundo todo, mas é sensato escolher o local mais próximo para baixar e instalar um pacote. Acessar o menu Packages no GUI do R é a melhor forma para selecionar repositórios e sites espelhos e para rodar automaticamente a instalação de pacotes e suas assim chamadas dependências, isto é, pacotes adicionais necessários para a funcionalidade de um pacote no qual você está interessado. Um pacote é, na realidade, um arquivo zip que é salvo em uma mídia de armazenamento interna ou externa e contém procedimentos necessários, documentação e conectividade para o R. É também possível instalar um pacote localmente (uma opção presente no menu Packages), mas sua funcionalidade pode ser afetada se alguma dependência estiver ausente no conjunto de pacotes instalados. O R oferece uma imensa variedade de pacotes. Por exemplo, na versão 3.1.3, o repositório de pacotes CRAN apresenta 6415 pacotes disponíveis. Entretanto, neste livro, iremos apenas indicar os pacotes que são necessários para conseguir os resultados para os exemplos apresentados.

A.4 Objetos do R

A linguagem R manuseia *objetos*, os quais são definidos como entidades que podem ser representadas em um computador: números, variáveis, matrizes, funções do usuário, base de dados (conhecido como *data frame* no R) ou uma combinação de todos esses (como listas). Um nome pode ser dado para algum objeto,

desde que a sequência de caracteres que formam o nome não contenha algum dos caracteres: espaço(s), -, +, *, /, #, %, &, [,], {, }, (,) ou ~. Em adição, nomes de objetos não podem iniciar com um número e eles são sensíveis a maiúsculas e minúsculas. Assim, X e x são diferentes e referem-se a diferentes objetos. Para evitar subseqüentes erros tipográficos, é recomendado que você use nomes de objetos curtos e mnemônicos. A atribuição de um nome ao objeto significa que o objeto é alocado ao corrente espaço de memória do R via o operador de atribuição composto pelo caractere < e um hífen, sem espaço entre eles, isto é, <-. Como um exemplo, você pode querer definir uma variável chamada S carregando o valor 12. O comando é então `S <- 12`.

Se S não existe, ele será criado. Caso contrário, seus conteúdos prévios serão substituídos. Então, S será conservado na memória de trabalho (na memória do computador), mas outros comandos são necessários (p. ex., *save*) para armazená-lo em disco. Há também formas de *remover* um objeto na sessão corrente por meio do comando `rm`.

A.5 Vetores no R

O principal objeto no R é o vetor. Ele é uma n-tupla de objetos de mesma classe. O vetor familiar de um número real é o melhor exemplo de um vetor no R que pode ser criado usando a função de concatenação `c()`. Como um exemplo, considere os dados mostrados na Tabela 1.3. Assuma que você gostaria de analisar a precipitação anual dos sites em que colônias de borboletas *Euphydryas editha* foram amostradas. Para fazer isso, você pediria ao R para armazenar os valores no vetor numérico

```
precip <- c(43, 20, 28, 28, 15, 21, 10, 10, 19, 22,
           58, 34, 21, 42, 50)
```

Aqui, `precip` é um vetor, um arranjo de uma só linha cujo comprimento é igual ao seu número de elementos, neste caso, 16. Você pode verificar este comprimento por escrever

```
length(precip)
[1] 16
```

em que a segunda linha dá a resposta da primeira linha.

Cada elemento em um vetor tem seu próprio índice ou número de coluna, e isso pode ser referido por colocar entre colchetes este número. Como um exemplo, o 13º elemento de `precip` é a precipitação 34 mm. Você pode mostrar este valor particular no R como

```
precip[13]
[1] 34
```

Nomes podem ser dados a cada elemento de um vetor numérico como `precip`. Estes nomes podem ser tomados de um vetor de caracteres (isto é, um vetor contendo caracteres alfanuméricos entre aspas). Como um exemplo, para alocar nomes aos elementos de `precip` usando as abreviações das colônias como identificadores, os seguintes três comandos podem ser usados:

```
COLONY <- c("SS", "SB", "WSB", "JRC", "JRH", "SJ", "CR",
+ "UO", "LO", "DP", "PZ", "MC", "IF", "AF", "GH", "GL")
names(precip) <- COLONY
precip
  SS SB WSB JRC JRH SJ CR UO LO DP PZ MC IF AF GH GL
43 20 28 28 28 15 21 10 10 19 22 58 34 21 42 50
```

A característica mais importante dos vetores é que eles somente podem ser mantidos em uma classe de objetos (ou números, séries de caracteres, níveis de um fator ou valores lógicos), mas não em uma mistura deles. Em adição, vetores podem ser operados usando expressões aritméticas aplicadas a elemento por elemento. Consulte o Capítulo 2 do manual *Introduction to R* (Venables et al. 2015) para mais exemplos e propriedades de vetores).

A.6 Matrizes do R

Matrizes são casos particulares do que são conhecidos no R como arranjos de dados de mesma classe, o que é uma generalização multidimensional com múltiplos deles. Existem várias formas para construir uma matriz. Uma possibilidade é convertê-la a partir de um vetor usando a função de matrizes. Como um exemplo, assuma que temos um vetor oito-dimensional contendo o número de espécies (riqueza de espécies) em quatro locais, com os primeiros quatros elementos do vetor sendo os valores de riqueza para os quatro locais no Ano 1 e os últimos quatro elementos sendo os valores de riqueza correspondentes para Ano 2, como segue:

```
RICHNESS <- c(2, 2, 3, 2, 2, 1, 5, 1)
```

Uma forma melhor de arranjar esses dados é pela definição de uma matriz 4×2 com linhas indicando os locais e as correspondentes colunas correspondendo os anos:

```
RICHMAT <- matrix(RICHNESS, nrow = 4)
RICHMAT
[,1] [,2]
[1,]  2  2
[2,]  2  1
[3,]  3  5
[4,]  2  1
```

Por padrão, uma matriz é preenchida em coluna. Aqui, foi somente necessário escrever o argumento `nrow = 4` para indicar o número de linhas, mas você pode ser mais específico por escrever os números de linhas (`nrow =`) e número de colunas (`ncol =`) como

```
RICHMAT <- matrix(RICHNESS, nrow = 4, ncol=2).
```

Neste exemplo, o R endereça cada elemento da matriz `RICHMAT` pelas suas correspondentes posições linha e coluna. Então, a riqueza observada para o Local 3 no Ano 2 pode ser encontrada como

```
RICHMAT[3,2]  
[1] 5
```

Os elementos de uma particular linha ou coluna podem ser referidos por omitir coluna e linha correspondente, após ou antes da vírgula, respectivamente. Portanto, a riqueza para Anos 1 e 2 no Local 2 é

```
RICHMAT[2,]  
[1] 2 1
```

Similarmente, a riqueza observada no Ano 1 para todos os locais é

```
> RICHMAT[,1]  
[1] 2 2 3 2
```

É possível também construir uma matriz por combinar dois ou mais vetores de mesmo comprimento (e classe). Seguindo uma ideia similar do exemplo anterior, assuma que existam dois vetores de quatro entradas cada, com o primeiro vetor correspondendo à riqueza encontrada em quatro locais para o Ano 1 e o segundo contendo a riqueza correspondente para o Ano 2, definida como

```
richy1 <- c(2, 2, 3, 2)  
richy2 <- c(2, 1, 5, 1)
```

Os vetores podem então ser combinados (ligados) em uma matriz usando `cbind()` (combinada por colunas) ou `rbind()` (combinada por linhas). Então, a combinação pelas colunas é dada por

```
RICHMAT1 <- cbind(richy1, richy2)
```