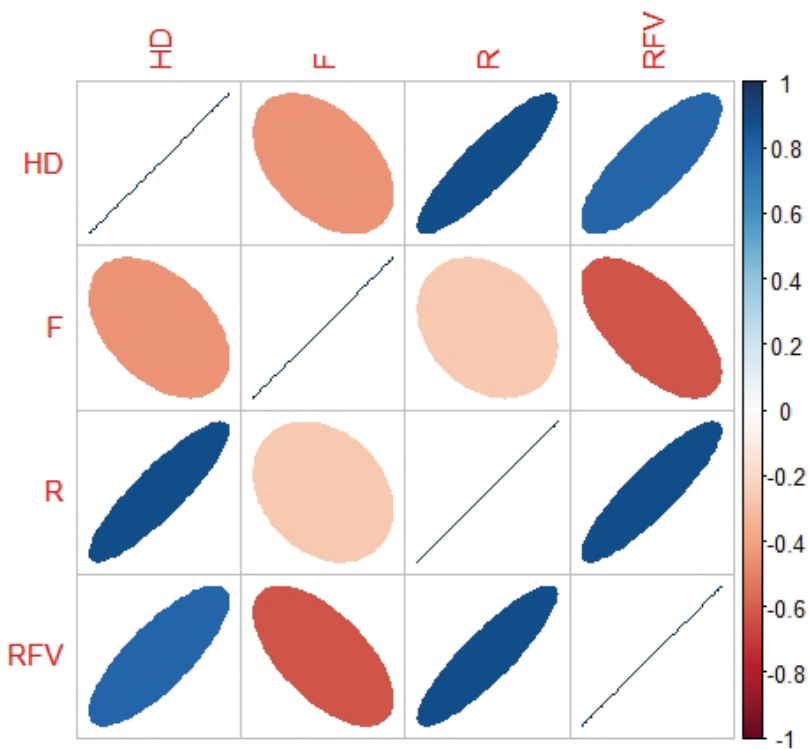
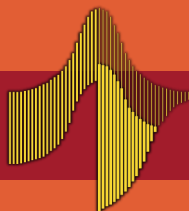


Rinaldo Artes
Lúcia Pereira Barroso

MÉTODOS MULTIVARIADOS DE ANÁLISE ESTATÍSTICA



Blucher



ABE - PROJETO FISHER

**MÉTODOS
MULTIVARIADOS DE
ANÁLISE ESTATÍSTICA**

Rinaldo Artes

Lúcia Pereira Barroso

Métodos multivariados de análise estatística
© 2023 Rinaldo Artes e Lúcia Pereira Barroso
Editora Edgard Blücher Ltda.

Publisher Edgard Blücher
Editores Eduardo Blücher e Jonatas Eliakim
Coordenação editorial Andressa Lira
Produção editorial Lidiane Pedroso Gonçalves
Revisão de texto Maurício Katayama
Imagem da capa Rinaldo Artes e Lúcia Pereira Barroso
Capa Leandro Cunha

Blucher

Rua Pedroso Alvarenga, 1245, 4º andar
CEP 04531-934 – São Paulo – SP – Brasil
Tel.: 55 11 3078-5366
contato@blucher.com.br
www.blucher.com.br

Segundo o Novo Acordo Ortográfico, conforme 6. ed. do *Vocabulário Ortográfico da Língua Portuguesa*, Academia Brasileira de Letras, julho de 2021.

É proibida a reprodução total ou parcial por quaisquer meios sem autorização escrita da editora.

Todos os direitos reservados pela Editora Edgard Blücher Ltda.

Dados Internacionais de Catalogação na
Publicação (CIP) Angélica Ilacqua
CRB-8/7057

Artes, Rinaldo
Métodos multivariados de análise estatística / Rinaldo Artes, Lúcia Pereira Barroso. São Paulo : Blucher, 2023.
534 p.

Bibliografia
ISBN 978-65-5506-702-6

1. Estatística 2. Estatística descritiva I. Título
II. Barroso, Lúcia Pereira

23-2059 CDD 519-5

Índice para catálogo sistemático:
1. Estatística

CONTEÚDO

1	NOTAÇÕES, RESULTADOS BÁSICOS E CONVENÇÕES	1
1.1	Introdução	1
1.2	Notações e resultados básicos	1
1.3	Resultados básicos da distribuição normal multivariada	5
1.3.1	Distribuição normal bivariada	7
2	ESTATÍSTICA DESCRITIVA	9
2.1	Introdução	9
2.2	Medidas descritivas	9
2.2.1	Vetor média amostral	10
2.2.2	Matriz de covariâncias amostrais	10
2.2.3	Variância total e variância generalizada	12
2.2.4	Matriz de correlações amostrais	14
2.3	Representação gráfica	15
2.3.1	Representação de variáveis	15
2.3.2	Representação de casos	19
2.4	Valores aberrantes multivariados	26
2.4.1	Valores aberrantes unidimensionais	26

2.4.2	Valores aberrantes bidimensionais	27
2.4.3	Valores aberrantes multidimensionais	31
2.4.4	Comentários de ordem prática	31
2.4.5	Aplicação	32
2.5	Avaliação da normalidade multivariada	39
2.6	Biplots	41
2.6.1	Dados sobre crimes	42
2.6.2	Desenvolvimento teórico	43
2.6.3	Interpretação do biplot	46
2.7	Utilizando o R	51
2.8	Exercícios	59
3	ANÁLISE DE COMPONENTES PRINCIPAIS	69
3.1	Introdução	69
3.2	Conceitos básicos	70
3.3	Obtenção das componentes principais	72
3.4	Propriedades das componentes principais	75
3.5	Decomposição da matriz de correlações	77
3.6	Comentários gerais	81
3.6.1	Obtenção dos valores das variáveis originais a partir das componentes principais	81
3.6.2	Número de componentes principais	82
3.6.3	Interpretação das componentes principais	84
3.6.4	Multicolinearidade e componentes principais	89
3.7	Biplot	91

3.8	Utilizando o R	93
3.9	Exercícios	96
4	ANÁLISE FATORIAL	101
4.1	Introdução	101
4.2	Constructos	102
4.3	Análise fatorial ortogonal	105
4.3.1	Cargas fatoriais	107
4.3.2	Matriz de covariâncias de \mathbf{x}	107
4.3.3	Comunalidades e especificidades	108
4.3.4	Padronização das variáveis	110
4.4	Métodos de obtenção de fatores	111
4.4.1	Método das componentes principais	111
4.4.2	Método da máxima verossimilhança	117
4.5	Escolha do número de fatores	119
4.6	Rotações ortogonais	121
4.7	Escores fatoriais	123
4.7.1	Método dos mínimos quadrados ponderados	125
4.7.2	Método da regressão	125
4.8	Estudo da adequabilidade da AF	126
4.8.1	Matriz anti-imagem	126
4.8.2	MSA: measure of sampling adequacy	127
4.8.3	KMO: Kaiser-Meyer-Olkin	128
4.8.4	Teste de esfericidade de Bartlett	129
4.9	Avaliação do ajuste do modelo	130

4.10	Variáveis ordinais	132
4.11	Análise fatorial confirmatória	135
4.12	Comentários gerais	136
4.13	Utilizando o R	138
4.14	Exercícios	140
5	ESCALONAMENTO MULTIDIMENSIONAL	147
5.1	Introdução	147
5.2	Escalonamento multidimensional métrico	149
5.2.1	Desenvolvimento teórico	150
5.2.2	Aplicação	153
5.3	Escalonamento multidimensional não métrico	157
5.3.1	Desenvolvimento teórico	157
5.3.2	Aplicação	159
5.4	Utilizando o R	161
5.5	Exercícios	164
6	ANÁLISE DE CORRESPONDÊNCIA	169
6.1	Introdução	169
6.2	Análise de correspondência simples	171
6.3	Análise de correspondência para múltiplas tabelas	193
6.3.1	Análise de correspondência para tabelas justapostas . . .	193
6.3.2	Análise de correspondência interna	203
6.3.3	Análise fatorial múltipla para tabelas de contingência . .	206
6.3.4	Análise de correspondência múltipla	208

6.4	Utilizando o R	212
6.5	Exercícios	216
7	ANÁLISE DE CORRELAÇÃO CANÔNICA	223
7.1	Exemplo	223
7.2	Obtenção das correlações canônicas populacionais	228
7.3	Cargas canônicas	231
7.4	Cargas canônicas cruzadas	232
7.5	Teste de Bartlett	232
7.6	Dados padronizados	233
7.7	Análise de correlação canônica e regressão linear múltipla	235
7.8	Utilizando o R	236
7.8.1	Alternativa 1: Álgebra linear	237
7.8.2	Alternativa 2: Comando cancel	238
7.8.3	Alternativa 3: Biblioteca CCA	239
7.8.4	Teste de Bartlett	241
7.9	Exercícios	242
8	ANÁLISE DE AGRUPAMENTOS	245
8.1	Conceitos básicos	246
8.2	Notação e medidas de parença	250
8.2.1	Dados numéricos	251
8.2.2	Dados categorizados	257
8.2.3	Dados categorizados e numéricos	261
8.2.4	Outras abordagens	263

8.3	Algoritmos de agrupamentos	263
8.3.1	Métodos hierárquicos aglomerativos	263
8.3.2	Métodos de partição	275
8.4	Comparação dos métodos	281
8.4.1	Outros métodos	283
8.5	Validação e interpretação	283
8.5.1	Correlação cofenética	284
8.5.2	Gráfico da silhueta	285
8.5.3	Replicabilidade	288
8.6	Interpretação	289
8.6.1	Representação gráfica de casos	289
8.7	Aplicações	291
8.7.1	Eleição presidencial	292
8.7.2	Tipologia de agricultores familiares	295
8.7.3	Identificação da cultura organizacional	299
8.8	Comentários adicionais	303
8.9	Utilizando o R	304
8.10	Exercícios	308
9	ANÁLISE DISCRIMINANTE E CLASSIFICATÓRIA	311
9.1	Introdução	311
9.2	Análise discriminante para duas populações	317
9.2.1	O método de Fisher	318
9.2.2	O método geral de classificação	323
9.3	Análise discriminante em situações com mais de duas populações .	333

9.3.1	O método de Fisher	333
9.3.2	O método geral de classificação	336
9.4	Avaliação da função de classificação	339
9.5	Aplicação	341
9.6	Comentários adicionais	345
9.7	Utilizando o R	346
9.8	Exercícios	349
10	CLASSIFICAÇÃO COM REGRESSÃO LOGÍSTICA	355
10.1	Desenvolvimento do modelo logístico	357
10.1.1	Teste de Hosmer e Lemeshow	358
10.2	Ajuste do modelo do Exemplo 10.1	359
10.3	Classificação	360
10.3.1	Curva ROC	361
10.3.2	Estatística KS	365
10.3.3	Aplicação	366
10.4	Comentários finais	370
10.5	Utilizando o R	370
10.6	Exercícios	372
11	ÁRVORES DE DECISÃO	377
11.1	Exemplos	378
11.1.1	Variável resposta binária	378
11.1.2	Variável resposta contínua	383
11.2	Terminologia	387

11.3	Partições	388
11.3.1	Exemplo de partições politômicas	391
11.4	Critérios de partição	392
11.4.1	Árvores de regressão	393
11.4.2	Árvores de classificação	398
11.5	Aspectos técnicos	399
11.5.1	Algoritmo	399
11.5.2	Critérios de parada	400
11.6	Validação da classificação	402
11.7	Comentários adicionais	403
11.7.1	Uso da técnica na construção de modelos de regressão	403
11.7.2	Limitações da técnica	403
11.8	Métodos agregados	404
11.8.1	<i>Bagging – Bootstrap aggregation</i>	404
11.8.2	Floresta aleatória	405
11.9	Aplicação	405
11.9.1	Árvore de regressão	406
11.9.2	Árvore de classificação	409
11.10	Utilizando o R	411
11.10.1	Árvores de regressão	412
11.10.2	Árvores de classificação	414
11.11	Exercícios	416

APÊNDICE A CONJUNTOS DE DADOS **419**

A.1	Arquivo BancoAlemao.xlsx	419
-----	------------------------------------	-----

A.2	Arquivo BemEstarFin.xlsx	421
A.3	Arquivo CapitaisDem.xlsx	421
A.4	Arquivo Carros.xlsx	422
A.5	Arquivo Celular.xlsx	422
A.6	Arquivo Ceramica.xlsx	423
A.7	Arquivo Covid19MGSP.xlsx	424
A.8	Arquivo Covid19SP.xlsx	425
A.9	Arquivo: Diabetes.xlsx	425
A.10	Arquivo Eleições_SE_2002.xlsx	426
A.11	Arquivo ExpVida.xlsx	427
A.12	Arquivo ILE2020.xlsx	428
A.13	Arquivo Insfin2.xlsx	429
A.14	Arquivo Mochila.xlsx	430
A.15	Arquivo Notebook.xlsx	430
A.16	Arquivo Otolito.xlsx	432
A.17	Arquivo: Pizza.xlsx	433
A.18	Arquivo PublicDoutor.xlsx	433
A.19	Arquivo Stress.xlsx	434
A.20	Arquivo Tumor.xlsx	435
A.21	Arquivo VinhoTinto.xlsx	435
A.22	Arquivo: WVS6.xlsx	436
A.23	Arquivo: WVS6b.xlsx	437

B.1	Vetores	439
B.2	Matrizes	442
APÊNDICE C VETORES ALEATÓRIOS		455
APÊNDICE D TESTES DE HIPÓTESES MULTIVARIADOS		459
D.1	Testes para um vetor média de população normal multivariada . .	459
D.2	Testes para a comparação de vetores médias de duas populações normais multivariadas	460
D.2.1	Caso $\Sigma_1 = \Sigma_2$	460
D.2.2	Caso $\Sigma_1 \neq \Sigma_2$	461
D.2.3	Comentários	461
D.3	Comparação de vetores médias de várias populações normais . . .	461
D.3.1	Teste de Wilks	462
D.3.2	Teste de Lawley e Hotelling	463
D.3.3	Teste de Pillai	463
D.4	Comparação de matrizes de covariâncias	463
D.4.1	Teste da razão de verossimilhanças	464
D.4.2	Teste de M de Box	464
APÊNDICE E CONSTRUÇÃO E AVALIAÇÃO DE ESCALAS		467
E.1	O processo de mensuração	467
E.1.1	Níveis de mensuração	468
E.1.2	A natureza da medida	469
E.1.3	Operacionalização de constructo e construção de escalas .	470
E.1.4	Escalas	473

E.1.5	Avaliação de uma escala	476
E.2	Erros de medida	477
E.3	Fidedignidade de escalas	479
E.3.1	Repetição	479
E.3.2	Teoria clássica da mensuração	480
E.3.3	Escalas aditivas	483
E.3.4	Análise de itens	488
E.3.5	Coefficiente alfa estratificado	490
E.3.6	Coefficiente L_2	490
E.3.7	Comentários	491
E.4	Validade de escala	491
E.4.1	Validade de conteúdo (<i>content validity</i>)	492
E.4.2	Validade de critério (<i>criterion validity</i>)	493
E.4.3	Validade de constructo (<i>construct validity</i>)	494
E.5	Comentários finais	495
E.6	Demonstrações de resultados do capítulo	495

BIBLIOGRAFIA

CAPÍTULO 1

NOTAÇÕES, RESULTADOS BÁSICOS E CONVENÇÕES

1.1 Introdução

Apresentamos neste capítulo a notação adotada e alguns resultados básicos de teoria das probabilidades.¹ O Exemplo 1.1 será utilizado na descrição de alguns resultados.

Exemplo 1.1 *A Secretaria de Segurança Pública do Estado de São Paulo, para fins administrativos, divide o território em regiões. A Figura 1.1 apresenta como era essa segmentação territorial em 2002. A Tabela 1.1 mostra as taxas de delitos por 100.000 habitantes por região.*

1.2 Notações e resultados básicos

Admitimos a existência de p variáveis observadas para n indivíduos. Vetores e matrizes são representados em negrito; utilizamos letras minúsculas para vetores e maiúsculas para matrizes.

¹Mais detalhes sobre os resultados podem ser encontrados em Johnson e Wichern (2007), Mardia, Kent e Bibby (1979) e Dillon e Goldstein (1984), por exemplo.

Tabela 1.1: Taxa de delitos por 100.000 habitantes por divisão territorial das polícias do estado de São Paulo em 2002

Região	Homicídio doloso	Furto	Roubo	Roubo e furto de veículos
SJRP	10,85	1.500,80	149,35	108,38
RP	14,13	1.496,07	187,99	116,66
Bauru	8,62	1.448,79	130,97	69,98
Campinas	23,04	1.277,33	424,87	435,75
Sorocaba	16,04	1.204,02	214,36	207,06
SP	43,74	1.190,94	1.139,52	909,21
SJC	25,39	1.292,91	358,39	268,24
Santos	42,86	1.590,66	721,90	275,89
GSP	42,55	797,16	520,73	602,63
Média	25,25	1.310,96	427,56	332,64
Desvio padrão	14,36	239,48	330,76	275,01

Fonte: Secretaria de Segurança Pública do Estado de São Paulo.

<http://www.ssp.sp.gov.br/estatisticas/criminais/>,

Acesso em 11 fev. 2003.

SJRP: São José do Rio Preto; Ribeirão Preto; São Paulo (capital);

SJC: São José dos Campos e GSP: Grande São Paulo, exceto SP.

Uma observação multivariada é representada por $\mathbf{x} = (X_1, \dots, X_p)^\top$, no qual X_j , $j = 1, \dots, p$, indicam as variáveis aleatórias consideradas no problema. Esse vetor é denominado vetor aleatório. Assumimos a existência de independência entre as observações de indivíduos diferentes.

Representamos uma matriz de dados por

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1^\top \\ \mathbf{x}_2^\top \\ \vdots \\ \mathbf{x}_n^\top \end{pmatrix} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)^\top,$$

sendo x_{ij} o valor assumido pela variável X_j , $j = 1, \dots, p$, para o indivíduo i , $i = 1, \dots, n$ e $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^\top$ o vetor de observações para o indivíduo i .

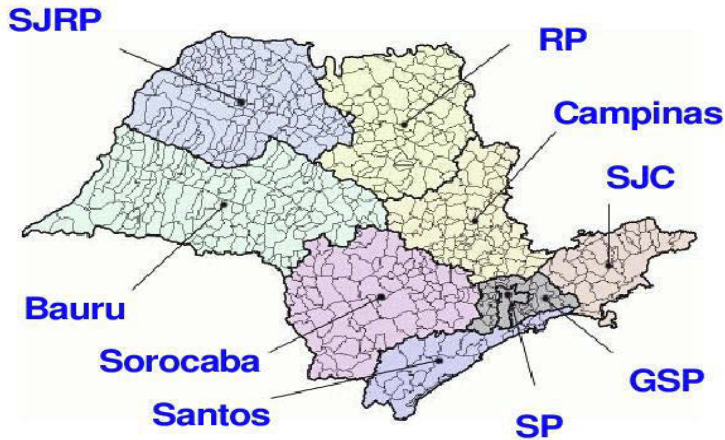


Figura 1.1: Divisão territorial das polícias do Estado de São Paulo em 2002

Fonte: <http://www.ssp.sp.gov.br/estatisticas/criminais/>. Acesso em: 11 fev. 2003.

Para o Exemplo 1.1, temos X_1 : Taxa de homicídios dolosos, X_2 : Taxa de furtos, X_3 : Taxa de roubos e X_4 : Taxa de roubos e furtos de veículos. Os dados relativos a SJRP são denotados por

$$\mathbf{x}_1 = (10,85 \quad 1.500,80 \quad 149,35 \quad 108,38)^\top.$$

Por fim, a matriz de dados é

$$\mathbf{X} = \begin{pmatrix} 10,85 & 1.500,80 & 149,35 & 108,38 \\ 14,13 & 1.496,07 & 187,99 & 116,66 \\ 8,62 & 1.448,79 & 130,97 & 69,98 \\ 23,04 & 1.277,33 & 424,87 & 435,75 \\ 16,04 & 1.204,02 & 214,36 & 207,06 \\ 43,74 & 1.190,94 & 1.139,52 & 909,21 \\ 25,39 & 1.292,91 & 358,39 & 268,24 \\ 42,86 & 1.590,66 & 721,90 & 275,89 \\ 42,55 & 797,16 & 520,73 & 602,63 \end{pmatrix}. \quad (1.1)$$

Definição 1.1 Seja $\mathbf{A} = [a_{ij}]$ uma matriz de dimensão $(p \times p)$ e $\mathbf{b} = (b_1, \dots, b_p)^\top$, então

a.
$$\text{Diag}(\mathbf{b}) = \begin{pmatrix} b_1 & 0 & \dots & 0 \\ 0 & b_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & b_p \end{pmatrix}.$$

b.
$$\text{diag}(\mathbf{A}) = (a_{11}, a_{22}, \dots, a_{pp})^\top.$$

c. Considere $b_i \geq 0$ e defina $\mathbf{B} = \text{Diag}(\mathbf{b})$, então

$$\mathbf{B}^{1/2} = \text{Diag}(\sqrt{b_1}, \dots, \sqrt{b_p}),$$

se $b_i > 0$, então $\mathbf{B}^{-1/2} = (\mathbf{B}^{1/2})^{-1}$. O Resultado B.19, do Apêndice B, generaliza essa operação.

Definição 1.2 Seja $\mathbf{x} = (X_1, \dots, X_p)^\top$ um vetor aleatório com $E(X_i) = \mu_i$, $\text{Var}(X_i) = \sigma_i^2$, $\text{Cov}(X_i, X_j) = \sigma_{ij}$ e $\text{Corr}(X_i, X_j) = \rho_{ij}$, $i, j = 1, \dots, p$. Defina

a. Vetor média de \mathbf{x} : $\boldsymbol{\mu} = (\mu_1, \dots, \mu_p)^\top$.

b. Matriz de covariâncias de \mathbf{x} :

$$\boldsymbol{\Sigma} = E[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top] = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \dots & \sigma_p^2 \end{pmatrix},$$

em que $\sigma_{ij} = \sigma_{ji}$, para $i, j = 1, \dots, p$, ou seja, $\boldsymbol{\Sigma}$ é simétrica.

c. Seja $\mathbf{V} = \begin{pmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_p^2 \end{pmatrix}$, então a **matriz de correlações** de \mathbf{x} é dada por

$$\boldsymbol{\rho} = \mathbf{V}^{-1/2} \boldsymbol{\Sigma} \mathbf{V}^{-1/2} = \begin{pmatrix} 1 & \rho_{12} & \dots & \rho_{1p} \\ \rho_{21} & 1 & \dots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \dots & 1 \end{pmatrix},$$

em que $\rho_{ij} = \rho_{ji}$, para $i, j = 1, \dots, p$, ou seja, $\boldsymbol{\rho}$ é simétrica. Consequentemente,

$$\boldsymbol{\Sigma} = \mathbf{V}^{1/2} \boldsymbol{\rho} \mathbf{V}^{1/2}.$$

A seguir, apresentamos alguns resultados sobre esperança e covariância de vetores aleatórios.

Resultado 1.1 *Sejam \mathbf{x} e \mathbf{y} vetores aleatórios de dimensão p com vetores médias $\boldsymbol{\mu}_x$ e $\boldsymbol{\mu}_y$, respectivamente, e com $\text{Cov}(\mathbf{x}) = \boldsymbol{\Sigma}_x$ e $\text{Cov}(\mathbf{y}) = \boldsymbol{\Sigma}_y$. Sejam \mathbf{a} e \mathbf{b} vetores de constantes de dimensão p e \mathbf{A} uma matriz de constantes de dimensão $(m \times p)$. Então*

a. $E(\mathbf{a}^\top \mathbf{x} + \mathbf{b}^\top \mathbf{y}) = \mathbf{a}^\top \boldsymbol{\mu}_x + \mathbf{b}^\top \boldsymbol{\mu}_y.$

b. $\text{Cov}(\mathbf{A}\mathbf{x}) = \mathbf{A}\boldsymbol{\Sigma}_x\mathbf{A}^\top.$

Resultado 1.2 : *Seja $\mathbf{x} = (X_1, \dots, X_p)^\top$ um vetor aleatório com $E(X_i) = \mu_i$, $\text{Var}(X_i) = \sigma_i^2$, $i = 1, \dots, p$. Defina $Z_i = (X_i - \mu_i)/\sigma_i$, uma variável padronizada construída a partir de X_i , e $\mathbf{z} = (Z_1, \dots, Z_p)^\top$. Então*

a. $\mathbf{z} = \mathbf{V}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})$, sendo \mathbf{V} dada na Definição 1.2.

b. $E(Z_i) = 0$ e, conseqüentemente, $E(\mathbf{z}) = (0, \dots, 0)^\top = \mathbf{0}_p$, vetor nulo.

c. $\text{Var}(Z_i) = 1$ e $\text{Cov}(\mathbf{z}) = \boldsymbol{\rho}$.

Prova do item c: Do item *a*, temos $\text{Cov}(\mathbf{z}) = \text{Cov}(\mathbf{V}^{-1/2}(\mathbf{x} - \boldsymbol{\mu}))$. Aplicando o item *b* do Resultado 1.1, vem que $\text{Cov}(\mathbf{z}) = \mathbf{V}^{-1/2}\text{Cov}(\mathbf{x})\mathbf{V}^{-1/2} = \boldsymbol{\rho}$. ◻

1.3 Resultados básicos da distribuição normal multivariada

Definição 1.3 *Dizemos que um vetor aleatório p -dimensional \mathbf{x} segue uma distribuição normal multivariada com vetor média $\boldsymbol{\mu}$ e matriz de covariâncias $\boldsymbol{\Sigma}$, positiva definida, se sua função densidade de probabilidade for dada por*

$$p(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{p/2} |\boldsymbol{\Sigma}|^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\}.$$

Denota-se $\mathbf{x} \sim N_p(\boldsymbol{\mu}; \boldsymbol{\Sigma})$.

Resultado 1.3 Seja $\mathbf{x} \sim N_p(\boldsymbol{\mu}; \boldsymbol{\Sigma})$, \mathbf{a} um vetor p -dimensional de constantes e \mathbf{A} uma matriz de dimensão $(m \times p)$ de constantes, então

- $\mathbf{a}^\top \mathbf{x} \sim N(\mathbf{a}^\top \boldsymbol{\mu}; \mathbf{a}^\top \boldsymbol{\Sigma} \mathbf{a})$.
- $\mathbf{x} + \mathbf{a} \sim N_p(\boldsymbol{\mu} + \mathbf{a}; \boldsymbol{\Sigma})$.
- $\mathbf{A}\mathbf{x} \sim N_m(\mathbf{A}\boldsymbol{\mu}; \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top)$.

Resultado 1.4 Seja $\mathbf{x} = (\mathbf{x}_1^\top, \mathbf{x}_2^\top)^\top$, com $\mathbf{x}_1, \mathbf{x}_2$ de dimensão $(m \times 1)$ e $(q \times 1)$, respectivamente e $p = m + q$. Assuma que $\mathbf{x} \sim N_p(\boldsymbol{\mu}; \boldsymbol{\Sigma})$, com

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix} \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix},$$

sendo que $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \text{Cov}(\mathbf{x}_1) = \boldsymbol{\Sigma}_{11}, \text{Cov}(\mathbf{x}_2) = \boldsymbol{\Sigma}_{22}$ e $\boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{21}^\top$ são, respectivamente, de dimensão $(m \times 1), (q \times 1), (m \times m), (q \times q)$ e $(m \times q)$, então

- $\mathbf{x}_1 \sim N_m(\boldsymbol{\mu}_1; \boldsymbol{\Sigma}_{11})$ e $\mathbf{x}_2 \sim N_q(\boldsymbol{\mu}_2; \boldsymbol{\Sigma}_{22})$.
- \mathbf{x}_1 e \mathbf{x}_2 são independentes se e somente se $\boldsymbol{\Sigma}_{12} = \mathbf{0}$.
- A distribuição condicional de \mathbf{x}_1 dado $\mathbf{x}_2 = \mathbf{a}$ é normal m -variada com

$$E(\mathbf{x}_1 | \mathbf{x}_2 = \mathbf{a}) = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} (\mathbf{a} - \boldsymbol{\mu}_2)$$

$$\text{Cov}(\mathbf{x}_1 | \mathbf{x}_2 = \mathbf{a}) = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\Sigma}_{21}.$$

Resultado 1.5 Se $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, com $|\boldsymbol{\Sigma}| > 0$, então

- $(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \sim \chi_p^2$.
- Sejam \mathbf{a}_i e $\lambda_i, i = 1, \dots, p$, respectivamente, os autovetores normalizados e os autovalores de $\boldsymbol{\Sigma}$, com $\lambda_1 \geq \dots \geq \lambda_p$, então as curvas de nível da função densidade de probabilidade de \mathbf{x} são hiperelipsoides satisfazendo

$$(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) = k^2,$$

com centro em $\boldsymbol{\mu}$ e eixos dados por $\pm k \sqrt{\lambda_i} \mathbf{a}_i$.

1.3.1 Distribuição normal bivariada

Ao tomarmos $p = 2$ temos a distribuição normal bivariada, cuja função densidade de probabilidade pode ser escrita como

$$p(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp \left\{ -\frac{1}{2(1-\rho^2)} \left[\left(\frac{x_1 - \mu_1}{\sigma_1} \right)^2 - 2\rho \frac{(x_1 - \mu_1)(x_2 - \mu_2)}{\sigma_1\sigma_2} + \left(\frac{x_2 - \mu_2}{\sigma_2} \right)^2 \right] \right\},$$

na qual, $\rho = \text{Corr}(X_1, X_2)$. Em particular,

$$\begin{aligned} E(X_1|X_2 = x_2) &= \mu_1 + \beta_1(x_2 - \mu_2), & \text{Var}(X_1|X_2 = x_2) &= \sigma_1^2(1 - \rho^2), \\ E(X_2|X_1 = x_1) &= \mu_2 + \beta_2(x_1 - \mu_1), & \text{Var}(X_2|X_1 = x_1) &= \sigma_2^2(1 - \rho^2), \end{aligned} \quad (1.2)$$

com

$$\beta_1 = \rho \frac{\sigma_1}{\sigma_2} \quad \text{e} \quad \beta_2 = \rho \frac{\sigma_2}{\sigma_1}.$$

Na Figura 1.2 são apresentados os gráficos da função densidade de probabilidade e respectivas curvas de nível, de distribuições normais bivariadas com vetor média nulo. As curvas de nível são figuras concêntricas com centro em $\boldsymbol{\mu} = \mathbf{0}$. Além disso,

- a. Os dois primeiros conjuntos de gráficos trazem situações em que a correlação entre as variáveis é nula; no primeiro caso, as curvas de nível são círculos e, no segundo, pelo fato de as variâncias serem diferentes, essas curvas são elipses, cujos eixos coincidem com os eixos cartesianos. Caso o vetor média não fosse nulo, esses eixos seriam paralelos aos cartesianos com intersecção no vetor média.
- b. Nos dois últimos conjuntos de gráficos, a correlação entre as variáveis é diferente de zero. As curvas de nível são elipses cujos eixos coincidem com as esperanças condicionais definidas no Resultado C.3 do Apêndice C. À medida que a correlação se afasta de zero, as elipses tendem a ficar mais estreitas.

Resultados adicionais sobre álgebra matricial e vetores aleatórios são apresentados no Apêndice B.

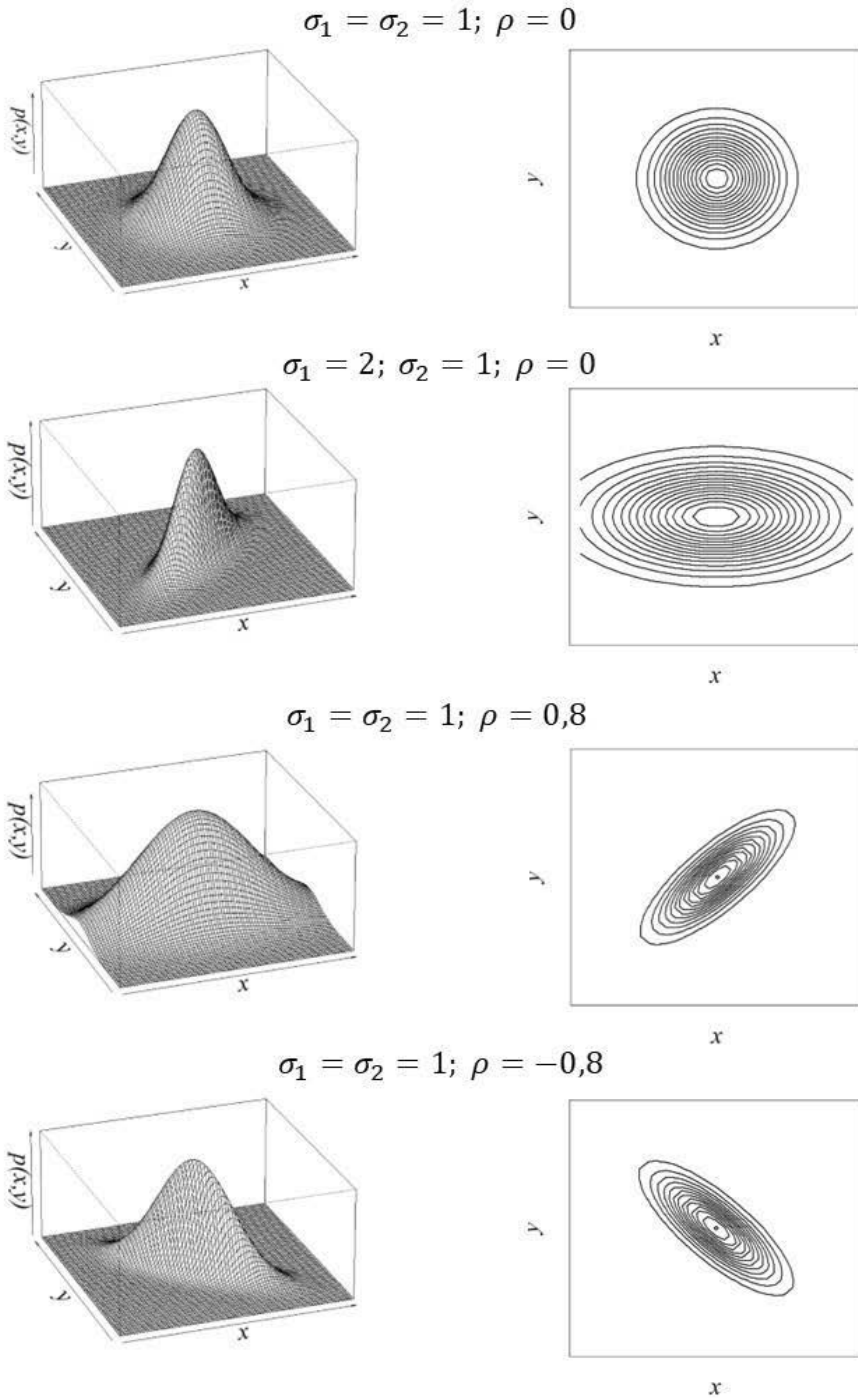


Figura 1.2: Funções densidade de probabilidade e curvas de nível de distribuições normais bivariadas com vetor média nulo

MÉTODOS MULTIVARIADOS DE ANÁLISE ESTATÍSTICA



Rinaldo Artes

Professor titular do Insper Instituto de Ensino e Pesquisa, doutor em Estatística pelo Instituto de Matemática e Estatística da Universidade de São Paulo (IME-USP), na modalidade sanduíche (IME-USP/ Universidade de British Columbia). Foi secretário da Associação Brasileira de Estatística no biênio 2006–2008 e diretor de pós-graduação stricto sensu e pesquisa do Insper (2009–2012). Possui interesses na área de modelagem estatística, análise multivariada, dados circulares e metodologia de pesquisa quantitativa.

Lúcia Pereira Barroso

Professora associada do Instituto de Matemática e Estatística da Universidade de São Paulo (IME-USP), possui mestrado em Estatística pelo IME-USP e doutorado sanduíche pelo IME-USP (London School of Economics). Foi presidente da Associação Brasileira de Estatística (2004–2006) e vice-presidente do IASS International Association of Survey Statisticians (2019–2021). Atuou como editora responsável na Revista Brasileira de Estatística e também como editora executiva do Brazilian Journal of Probability and Statistics. Foi presidente da comissão de graduação e coordenadora do bacharelado em Estatística do IME-USP. Dentre as suas áreas de interesse estão a análise multivariada, modelos de regressão e inferência estatística.

Sobre o livro

O livro reúne técnicas que permitem a segmentação de dados, o entendimento da estrutura de dependência de um conjunto de variáveis, a classificação de indivíduos em diferentes populações e a representação de um grande volume de variáveis em um espaço de menor dimensão. As técnicas são apresentadas por meio de exemplos, buscando conciliar aspectos teóricos e aplicados. Leitores sem forte formação matemática podem se beneficiar desse farto material, de forma que evitem as demonstrações, privilegiando a compreensão sobre como cada técnica opera e em que condições pode ser utilizada. Ao final de cada capítulo são apresentados códigos em linguagem R para a obtenção das análises demonstradas.

www.blucher.com.br

ISBN 978-65-5506-702-6

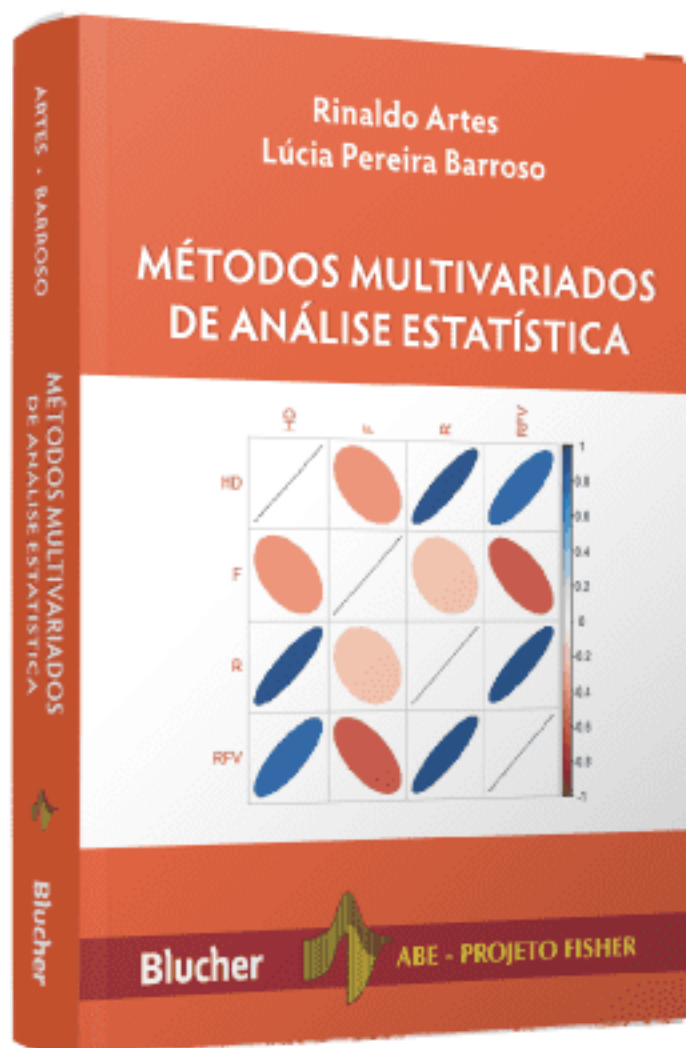


9 786555 067026



ABE - PROJETO FISHER

Blucher



Clique aqui e:

VEJA NA LOJA

Métodos multivariados de análise estatística

Rinaldo Artes, Lúcia Pereira Barroso

ISBN: 9786555067026

Páginas: 534

Formato: 17 x 24 cm

Ano de Publicação: 2023