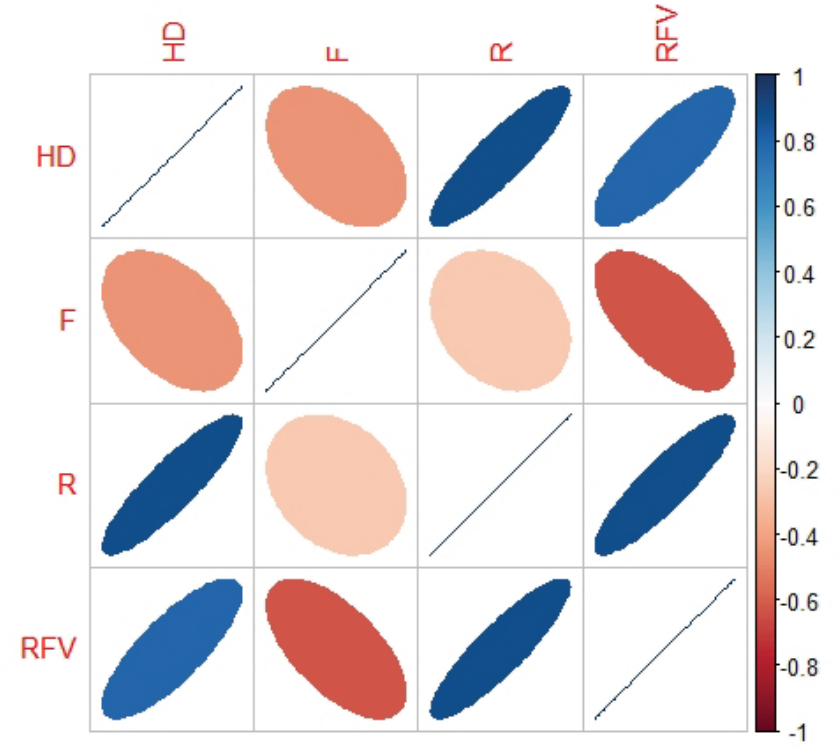
Rinaldo Artes
Lúcia Pereira Barroso

MÉTODOS MULTIVARIADOS DE ANÁLISE ESTATÍSTICA



Blucher
ABE - PROJETO FISHER

MÉTODOS MULTIVARIADOS DE ANÁLISE ESTATÍSTICA

Rinaldo Artes
Lúcia Pereira Barroso

Métodos multivariados de análise estatística

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.

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

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.

[1] Mais 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 $\mathbf{x} = (X_1, \ldots, X_p)^\top$, no qual X_j, $j = 1, \ldots, p$, 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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

em que $\rho_{ij} = \rho_{ji}$, *para* i,$j = 1, \ldots, p$, *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* $\mathrm{Cov}(\mathbf{x}) = \boldsymbol{\Sigma}_x$ *e* $\mathrm{Cov}(\mathbf{y}) = \boldsymbol{\Sigma}_y$*. 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. $\mathrm{E}\left(\mathbf{a}^\top\mathbf{x} + \mathbf{b}^\top\mathbf{y}\right) = \mathbf{a}^\top\boldsymbol{\mu}_x + \mathbf{b}^\top\boldsymbol{\mu}_y$.

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

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

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

b. $\mathrm{E}(Z_i) = 0$ *e, consequentemente,* $\mathrm{E}(\mathbf{z}) = (0, \ldots, 0)^T = \mathbf{0}_p$*, vetor nulo.*

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

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

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 $\mathbf{x} \sim N_p\left(\boldsymbol{\mu}; \boldsymbol{\Sigma}\right)$.

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*

a. $\mathbf{a}^\top \mathbf{x} \sim N\left(\mathbf{a}^\top \boldsymbol{\mu}; \mathbf{a}^\top \boldsymbol{\Sigma} \mathbf{a}\right)$.

b. $\mathbf{x} + \mathbf{a} \sim N_p(\boldsymbol{\mu} + \mathbf{a}; \boldsymbol{\Sigma})$.

c. $\mathbf{A}\mathbf{x} \sim N_m\left(\mathbf{A}\boldsymbol{\mu}; \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top\right)$.

Resultado 1.4 *Seja* $\mathbf{x} = \left(\mathbf{x}_1^\top, \mathbf{x}_2^\top\right)^\top$, *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 $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, $\mathrm{Cov}(\mathbf{x}_1) = \boldsymbol{\Sigma}_{11}$, $\mathrm{Cov}(\mathbf{x}_2) = \boldsymbol{\Sigma}_{22}$ *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*

a. $\mathbf{x}_1 \sim N_m(\boldsymbol{\mu}_1; \boldsymbol{\Sigma}_{11})$ *e* $\mathbf{x}_2 \sim N_q(\boldsymbol{\mu}_2; \boldsymbol{\Sigma}_{22})$.

b. $\mathbf{x}_1$ *e* $\mathbf{x}_2$ *são independentes se e somente se* $\boldsymbol{\Sigma}_{12} = \mathbf{0}$.

c. *A distribuição condicional de* $\mathbf{x}_1$ *dado* $\mathbf{x}_2 = \mathbf{a}$ *é normal m-variada com*

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

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

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

a. $(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \sim \chi_p^2$.

b. *Sejam* $\mathbf{a}_i$ *e* λ_i, $i = 1, \ldots, p$, *respectivamente, os autovetores normalizados e os autovalores de* $\boldsymbol{\Sigma}$, *com* $\lambda_1 \geq \ldots \geq \lambda_p$, *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, $\rho = \mathrm{Corr}(X_1, X_2)$. Em particular,

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

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.

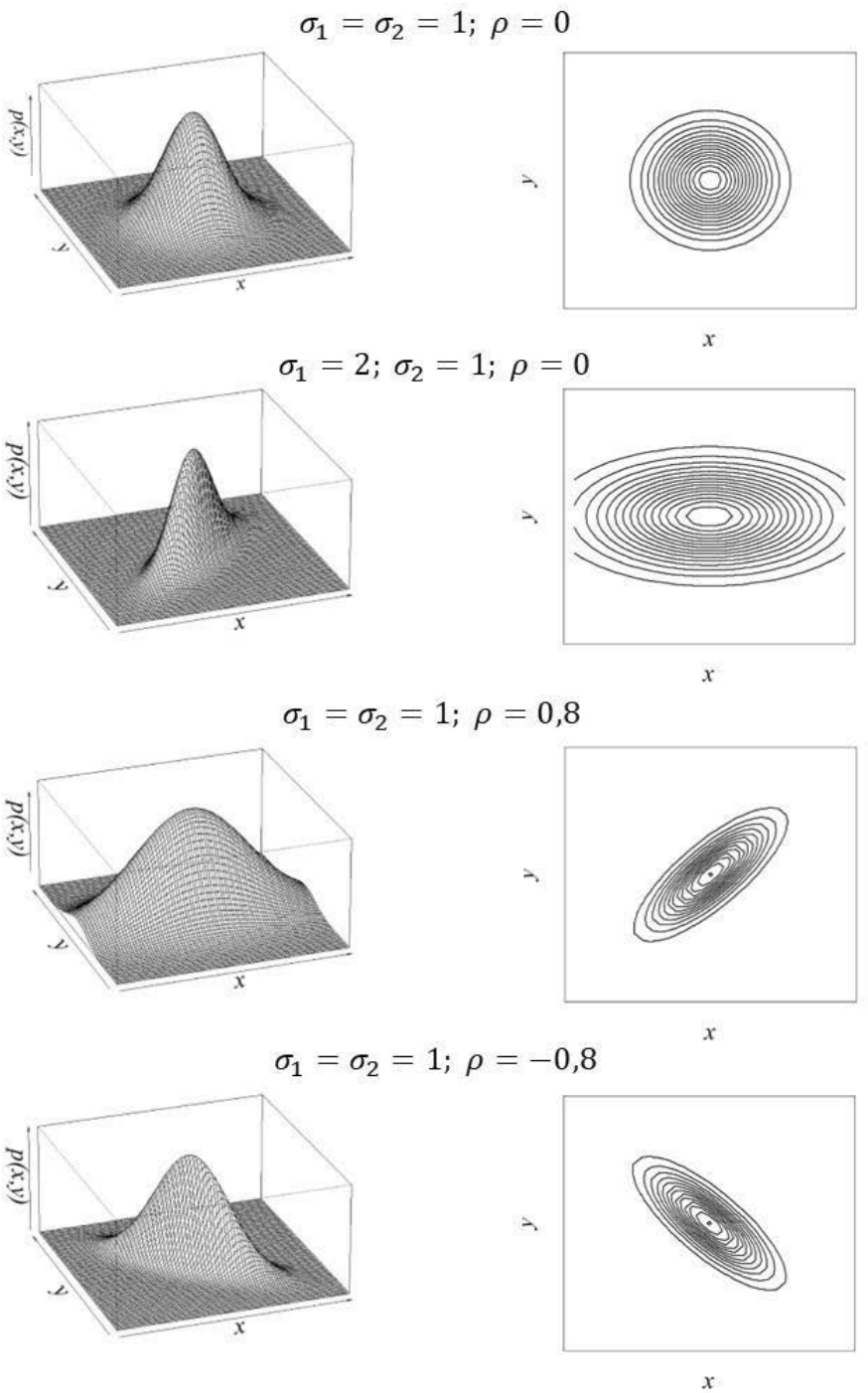


Figura 1.2: Funções densidade de probabilidade e curvas de nível de distribuições normais bivariadas com vetor média nulo

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.

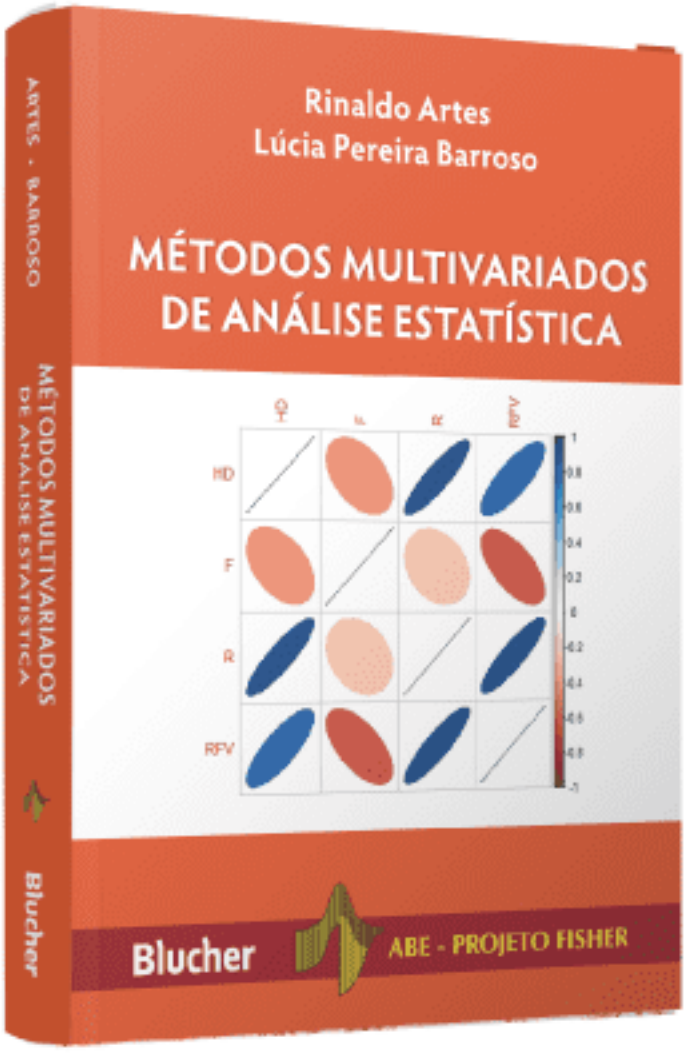
www.blucher.com.br

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