

Capítulo 1: Introdução à Análise Multivariada

Conteúdo

- 1.1. O que é uma análise multivariada?
- 1.2. Classificação das técnicas estatísticas
 - a. Métodos de dependência
 - b. Métodos de interdependência
- 1.3. Modelo multivariado
- 1.4. Parâmetros de um modelo multivariado: Esperança e matriz de covariância (definição e propriedades)
- 1.5. Amostra aleatória multivariada
- 1.5. Medidas amostrais multivariadas.

Bibliografia de base

- ▶ Johnson, R.A. e Wichern, D.W. (1982). *Applied multivariate statistical analysis*. 3 Edição. Prentice-Hall.
- ▶ Sharma, S. (1996). *Applied multivariate techniques*. John Wiley & Sons.

Análise Estatística Multivariada

- ▶ As técnicas de Análise Estatística são caracterizadas por métodos, exploratórios ou de inferência, destinados a extrair informação ou a modelar uma ou várias medições (**variáveis** ou **características**) de indivíduos, objetos, provas experimentais ou entidades de interesse (**unidades de amostragem** ou **observações**), de uma ou mais amostras. Tipicamente, as variáveis são medidas simultaneamente sobre cada unidade de amostragem e dentro de cada amostra.
- ▶ Uma Análise Estatística diz-se **Multivariada** quando o estudo inclui medições de mais do que uma variável por indivíduo.

Exemplo: Analisar o grau de satisfação de passageiros da TAP. São selecionados casualmente 250 passageiros não frequentes, que respondem a um inquérito de satisfação constituído por 18 questões medidos numa escala de Likert de 5 itens (1-muito insatisfeito ... 5-muito satisfeito). Os dados são multivariados com 18 variáveis (questões) e 250 unidades de amostragem (passageiros).

Escolha de Técnicas de Estatística Multivariada

Dependente dos objetivos (exploratórios e/ou de inferência) da análise:

Redução de dados ou simplificação da estrutura dos dados. Representar os dados de modo tão simples quanto possível, sem sacrificar informação relevante, facilitando a sua interpretação. (ACP, AF)

Agrupar e classificar. Identificar grupos “similares” de indivíduos ou de variáveis ou, alternativamente, definir regras de classificação dos indivíduos em grupos bem definidos. (AC, AD)

Estabelecer dependência entre variáveis. Estudar a existência de associação entre todas ou algumas variáveis. (ACorr, Tc, ACC)

Predição. Estabelecer relações entre variáveis que permitem a predição dos valores de uma ou mais variáveis com base nas observações de outras variáveis. (Reg, ANOVA)

Construção de hipóteses e testes. Testar hipóteses estatísticas específicas.

Tipo de dados

Os dados multivariados surgem quando um número $p > 1$ de variáveis são registadas por indivíduo podendo cada variável ser de tipo numérica ou não numérica (i.e., quantitativa ou qualitativa).

Tipo qualitativa (mesmo que codificada com números):

Nominal. Variável assumindo atributos não numéricos que não admitem ordenação.

Exemplos: Sexo (M / F); Tipo de sangue (A / B / AB / O).

Ordinal. Variável que apenas assume atributos não numéricos que admitem ordenação.

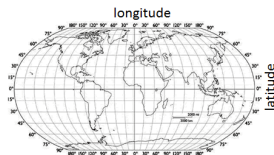
Exemplos: Grupo etário (<6 anos / 6-12 anos / >12 anos);
Nível de escolaridade (analfabeto / 1^o ciclo básico / 2^o ciclo básico / 3^o ciclo básico / 12^o ano / licenciatura / doutoramento);
Qualidade de serviço (medida na escala de Likert de 5 itens: 1-Muito mau, 2-Mau, 3-Razoável, 4-Bom, 5-Muito bom).

Tipo de dados

Tipo quantitativa:

Intervalar. Variável que assume valores numéricos com o zero numa posição arbitrária (i.e., o zero é desprovido do sentido habitual de "ausência de atributo").

Exemplos: Temperatura (em $^{\circ}\text{C}$);
Latitude; Longitude.



Escalar. Variável que assume valores numéricos de natureza discreta (finito ou infinito numerável) ou contínua.

Exemplos: Número de dias com precipitação por mês (natureza discreta, finita); Número de erros ortográficos por página (natureza discreta, infinita); Concentração diária de CO_2 na atmosfera num local pré-fixado (natureza contínua).

Matriz de dados

Os dados multivariados são representados genericamente por uma **matriz de dados** ($n \times p$),

$$\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

onde cada elemento x_{ij} representa o valor do i -ésimo indivíduo para a j -ésima variável, n representa o número de indivíduos ou unidades de amostragens em estudo e p é o número de variáveis ou características observadas sobre cada indivíduo.

Matriz de dados

Em estudos onde existem $p = 2$ variáveis de tipo nominal ou ordinal, os dados poder ser apresentados numa **tabela de contingência** (i.e., tabela de frequências absolutas) $r \times c$:

	B_1	B_2	\dots	B_c
A_1	n_{11}	n_{12}	\dots	n_{1c}
A_2	n_{21}	n_{22}	\dots	n_{2c}
	\vdots	\vdots	\ddots	\vdots
A_r	n_{r1}	n_{r2}	\dots	n_{rc}

onde n_{ij} representa o número de indivíduos observados com os atributos A_i e B_j em simultâneo, r é o número de atributos da variável A descrita em linha e c é o número de atributos da variável B descrita em coluna.

Matriz de dados -Exemplo-

A matriz de dados contém $p = 4$ variáveis: *Grupo etário* (tipo ordinal), *Sexo* (tipo nominal), *Peso antes* e *Peso depois* da dieta (ambos do tipo escalar). O tamanho da amostra é $n = 9$.

As variáveis não numéricas *Grupo etário* e *Sexo* podem ser representadas numa tabela de contingência

Age group	gender	before diet	after diet
young	F	65	60
young	M	75	73
kid	M	35	35
adult	F	85	78
adult	F	67	60
kid	F	40	38
young	M	57	55
adult	M	65	62
adult	M	71	68

	Grupo etário		
	kid	young	adult
M	1	2	2
F	1	1	2

Classificação das técnicas

As técnicas para dados multivariados podem ser classificadas em termos das relações entre as variáveis.

- ▶ **Métodos de dependência**: Técnicas para analisar conjuntos de dados de variáveis que podem ser divididas em dois grupos: um grupo de **variáveis independentes** ou **explanatórias** e outro grupo de **variáveis dependentes** ou **variáveis resposta**. (AD)

O objetivo é determinar ou testar se o conjunto dos dados das variáveis independentes afeta o conjunto dos dados das variáveis dependentes, individualmente e/ou em conjunto.

- ▶ **Métodos de interdependência**: Técnicas para analisar conjuntos de dados de variáveis em que é impossível, conceptualmente, dividi-las em dois subconjuntos, de variáveis dependentes e outro de variáveis independentes. (ACP, AF, AC)

O objetivo é identificar como as variáveis estão relacionadas entre elas e interpretar as relações.

Esquema geral das técnicas

Técnicas multivariadas em estudo em EM:

- ▶ Métodos de dependência: AD
- ▶ Métodos de interdependência: ACP, AF, AC

Partiremos de matrizes de dados com várias variáveis observadas sobre vários indivíduos e sem valores omissos

indivíduo	variáveis		
	X_1	\dots	X_p
I_1	x_{ij}		
\vdots			
I_n			

Esquema geral das técnicas

Se interessa ...

- ▶ estabelecer regras discriminantes dos indivíduos da amostra provenientes de diferentes grupos ou subpopulações, ou determinar como alocar novas observações dentro desses grupos /subpopulações, ...**poderá ser aplicada AD**

$\underbrace{Y}_{\text{não numérica}}$ depende de uma função de $(\underbrace{X_1, X_2, \dots, X_k}_{\text{numéricas ou não numéricas}})$

AD - Uma ilustração de uma aplicação

Journal of Theoretical and Applied Information Technology

15th October 2017, Vol.95, No.19

© 2005 – ongoing JATIT & LLS



ISSN: 1992-8645

www.jatit.org

E-ISSN: 1817-3195

MEASURING SMARTPHONE USAGE TIME IS NOT SUFFICIENT TO PREDICT SMARTPHONE ADDICTION

¹MYOUNGHEE SHIN, ²KANGWOO LEE

Usage time is a major criterion to determine whether a user is addicted to their smartphone, and many smartphone apps aiming to decrease smartphone addiction have been developed with this criterion in mind. However, this rule of thumb is based on an incorrect assumption that develops from studies on internet addiction. Our study tests how applicable this rule truly is, through correlation and discriminant analysis on smartphone usage patterns. Using a self-diagnosis scale for smartphone addiction (S scale for short) and

3.4 Linear Discriminant Analysis of Smartphone Addiction Using Smartphone Usage Patterns

To test whether a participant can be properly classified into a normal or addicted user using smartphone usage patterns, a linear discriminant analysis was carried out. For this purpose, we firstly divided the participants into two groups — normal and addicted, and the risky group was discarded. This was done to make the groups be more distinctive. So, according to the S scale

score, lower 132 participants were assigned to the normal group and the upper 27 participants were assigned to the addicted group. In some sense, it is assumed that the S scale value is the ground truth, and thus a statistical model is required to accurately classify smartphone usage patterns into corresponding classes.

With the factors corresponding to the 10 largest eigenvalues obtained from factor analysis, we performed Fisher's linear discriminant analysis [20] on the smartphone usage patterns. This method



Ser ou não viciado no smartphone = função de 10 fatores associados do questionário S-scale

Esquema geral das técnicas

Se interessa ...

- ▶ descrever os dados através de um conjunto menor de novas variáveis (definidas por combinações lineares das variáveis originais), que expliquem grande parte da variação observada nos dados,
...poderá ser aplicada ACP

$CP1$ = uma função linear de X_1, X_2, \dots, X_p

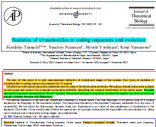
$CP2$ = outra função linear de X_1, X_2, \dots, X_p


\vdots (p componentes principais)

onde X_1, X_2, \dots, X_p são as p variáveis originais (numéricas).

- └ 1. Introdução à Análise Multivariada
 - └ 1.2. Classificação das técnicas estatísticas

ACP - Uma ilustração de uma aplicação





Available online at www.sciencedirect.com

SCIENCE @ DIRECT®

Journal of Theoretical Biology 222 (2003) 139–149

Journal of
Theoretical
Biology

www.elsevier.com/locate/jtbi

Statistics of trinucleotides in coding sequences and evolution

Fumihiko Takeuchi^{a,b,*}, Yasuhiro Futamura^a, Hiroshi Yoshikura^c, Kenji Yamamoto^a

Abstract

The aim of this paper is to give measurements indicative of evolutionary stages of the species. Two types of statistics of trinucleotides in coding regions are analysed for 27 species.

The first one is the codon space, the nucleotide ratio for each of the three codon positions. We apply principal component analysis on this space and extract two principal components faithfully describing the original distribution of the codon space. The first principal component corresponds to the G+C content. The second principal component classifies the species into three evolutionary groups, *Archaea*, *Bacteria* and *Eukaryota*.

Keywords: Statistics of Trinucleotides; Coding sequence; Codon space; Principal component analysis; Theoretical amino acid frequency; Randomness; Codon usage; tRNA abundance; Evolution

Componente Principal 1 = uma função linear das 12 variáveis originais (codon space)

Componente Principal 2 = outra função linear das 12 variáveis originais (codon space)

Esquema geral das técnicas

Se interessa ...

- ▶ expressar os dados em termos de um número menor de variáveis (não observáveis), que descrevam informação essencial contida nas variáveis originais e as suas inter-correlações,
... poderá ser aplicada AF.

X_1 = uma função linear de F_1, F_2, \dots, F_q

X_2 = outra função linear de F_1, F_2, \dots, F_q

\vdots (p relações lineares)

com $q < p$ e onde X_1, X_2, \dots, X_p são as p variáveis originais (numéricas).

- └ 1. Introdução à Análise Multivariada
 - └ 1.2. Classificação das técnicas estatísticas

AF - Uma ilustração de uma aplicação



Science competitions using technology: a study of the behavior of the participating schools in the CNC in Portugal

Alberto Oliveira da Silva¹

<https://orcid.org/0000-0002-3496-6802>

Adelaide Freitas¹

<https://orcid.org/0000-0002-4685-1615>

Maria Paula de Sousa Oliveira¹

<https://orcid.org/0000-0002-6376-1099>

Alexandre Mota da Silva¹

<https://orcid.org/0000-0002-5693-5775>

Ciênc. Educ., Bauru, v. 24, n. 3, p. 677-693, 2018

Abstract: In this work we investigate the dynamics of 143 schools of the 3rd cycle of Basic Education in Portugal, regarding the preparation and participation in online science competitions on curricular contents of Mathematics, Portuguese, Physics and Chemistry and Geology. An exploratory factorial analysis of empirical data concerning the competitions in 2015 was carried out, to analyze the characteristics inherent to schools' performance. Four latent factors describing the schools' behavior were identified: Quantitative Training, Qualitative Training, Proficiency and Users, which allowed us to verify that: (i) participation in mathematics competition is predominant; (ii) schools participating in two or three competitions present

N. de alunos em treinos nas 4 competições = uma função linear de 4 fatores F_1, F_2, F_3, F_4

N. de treinos realizados nas 4 competições = outra função linear de 4 fatores F_1, F_2, F_3, F_4

⋮
(12 variáveis originais conduzem a 12 relações lineares)

Esquema geral das técnicas

Se interessa ...

- ▶ encontrar grupos de indivíduos mais homogêneos entre si ou encontrar grupos de variáveis mais correlacionadas dentro dos grupos,
... **poderá ser aplicada AC** sobre os indivíduos ou sobre as variáveis, respetivamente.
- ▶ encontrar subgrupos de indivíduos mais homogêneos sobre um subgrupo de variáveis apenas (e não para todas as variáveis)
... **poderá ser aplicada ABic** (Biclustering).

- └ 1. Introdução à Análise Multivariada
 - └ 1.2. Classificação das técnicas estatísticas

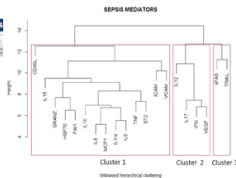
AC - Uma ilustração de uma aplicação



RESEARCH Open Access

A multiplex analysis of sepsis mediators during human septic shock: a preliminary study on myocardial depression and organ failures

Keyvan Razavi^{1,2*}, Florence Boissier^{1,2,3,4}, Mathieu Surenaud^{1,6}, Alexandre Bedet^{1,2}, Aurélien Seemann¹, Guillaume Carteaux^{1,2}, Nicolas de Prost^{1,2}, Christian Brun-Buisson^{1,2}, Sophie Hue^{1,6,7} and Armand Mekontso Dessap^{1,2}



Abstract

Background: The mechanisms of organ failure during sepsis are not fully understood. The hypothesis of circulating factors has been suggested to explain septic myocardial dysfunction. We explored the biological coherence of a large panel of sepsis mediators and their clinical relevance in septic myocardial dysfunction and organ failures during human septic shock.

Methods: Plasma concentrations of 24 mediators were assessed on the first day of septic shock using a multi-analyte cytokine kit. Septic myocardial dysfunction and organ failures were assessed using left ventricle ejection fraction (LVEF) and the Sequential Organ Failure Assessment score, respectively.

Results: Seventy-four patients with septic shock (and without immunosuppression or chronic heart failure) were prospectively included. Twenty-four patients (32%) had septic myocardial dysfunction (as defined by LVEF < 45%) and 30 (41%) died in ICU. Hierarchical clustering identified three main clusters of sepsis mediators, which were clinically meaningful. One cluster involved inflammatory cytokines of innate immunity, most of which were associated with septic myocardial dysfunction, organ failures and death; inflammatory cytokines associated with septic myocardial dysfunction had an additive effect. Another cluster involving adaptive immunity and repair (with IL-17/IFN pathway and VEGF) correlated tightly with a surrogate of early sepsis resolution (lactate clearance) and ICU survival.

Três grupos de variáveis “similares”

Vetor aleatório. Distribuição conjunta.

Um vetor $(p \times 1)$ onde cada componente $X_i, i = 1, 2, \dots, p$, representa uma variável aleatória (v.a.),

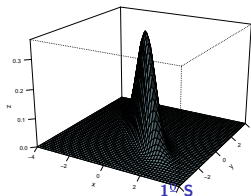
$$\mathbf{X} = [X_1 \ X_2 \ \dots \ X_p]',$$

chama-se **vetor aleatório (v.a.) p -dimensional**.

A distribuição de probabilidade do v.a. \mathbf{X} corresponde a uma função real com p variáveis.

No caso da distribuição ser contínua, a distribuição do v.a. \mathbf{X} fica bem definida pela chamada **função de densidade de probabilidade conjunta**.

Se $p = 2$, geometricamente, a função de densidade de probabilidade conjunta do v.a. $\mathbf{X} = (X_1, X_2)$ representa uma superfície em \mathbb{R}^3 .



Esperança do vetor aleatório \mathbf{X}

Sejam os v.a. $\mathbf{X} = [X_1 \ X_2 \ \dots \ X_p]'$ e $\mathbf{Y} = [Y_1 \ Y_2 \ \dots \ Y_p]'$, ambos $(p \times 1)$.

- **Definição.** A esperança de \mathbf{X} é um vetor p -dimensional $\boldsymbol{\mu} = E(\mathbf{X})$ cujas componentes são os valores esperados das componentes do vetor \mathbf{X} ; ou seja,

$$\boldsymbol{\mu} = E(\mathbf{X}) = \begin{bmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_p) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix}.$$

- **Propriedade.** A esperança é um operador linear:

$$E(\alpha\mathbf{X} + \beta\mathbf{Y} + \mathbf{c}) = \alpha E(\mathbf{X}) + \beta E(\mathbf{Y}) + \mathbf{c},$$

para todas as constantes $\alpha, \beta \in \mathbb{R}$ e vetor de números reais $\mathbf{c}_{(p \times 1)}$.

Esperança do vetor aleatório \mathbf{X}

- **Propriedade.** A esperança da combinação linear $\mathbf{a}'\mathbf{X} = a_1X_1 + \cdots + a_pX_p$ (unidimensional) é um número real dado por:

$$E(\mathbf{a}'\mathbf{X}) = \mathbf{a}'E(\mathbf{X}) = \mathbf{a}'\boldsymbol{\mu},$$

para todo vetor de números reais $\mathbf{a}_{(p \times 1)}$.

- **Propriedade.** A esperança das q combinações lineares $\mathbf{B}\mathbf{X}$ ($q \times 1$) é um vetor ($q \times 1$) dado por:

$$E(\mathbf{B}\mathbf{X}) = \mathbf{B}E(\mathbf{X}),$$

para toda a matriz de números reais $\mathbf{B}_{(q \times p)}$.

Exercício: Se $\mathbf{X} = [X_1 \ X_2]'$ é um par aleatório com esperança $\boldsymbol{\mu} = [0 \ 1]'$, calcule $E[2\mathbf{X} + \mathbf{Y}]$ onde $\mathbf{Y} = \left[X_1 - X_2 \quad \frac{X_1 + X_2}{2} \right]'$

Covariância entre duas variáveis aleatórias

A covariância entre X_i e X_j , $i \neq j$, $i, j = 1, 2, \dots, p$, mede o grau de relação linear entre as duas variáveis e é definida por:

$$\sigma_{ij} \doteq \text{Cov}(X_i, X_j) = E[(X_i - \mu_i)(X_j - \mu_j)] = E(X_i X_j) - \mu_i \mu_j.$$

- ▶ $\sigma_{ij} > 0$: as variáveis variam, em média, no mesmo sentido (se uma cresce, a outra em média também cresce; e vice-versa).
- ▶ $\sigma_{ij} < 0$: as variáveis variam em média em sentido oposto (se uma cresce, a outra em média decresce; e vice-versa).
- ▶ se X_i e X_j forem independentes, então $\sigma_{ij} = 0$.

Quando $i = j$, a covariância entre X_i e X_i reduz-se à variância de X_i

$$\sigma_{ii} = \sigma_i^2 = \text{Var}(X_i) = \text{Cov}(X_i, X_i) = E[(X_i - \mu_i)^2].$$

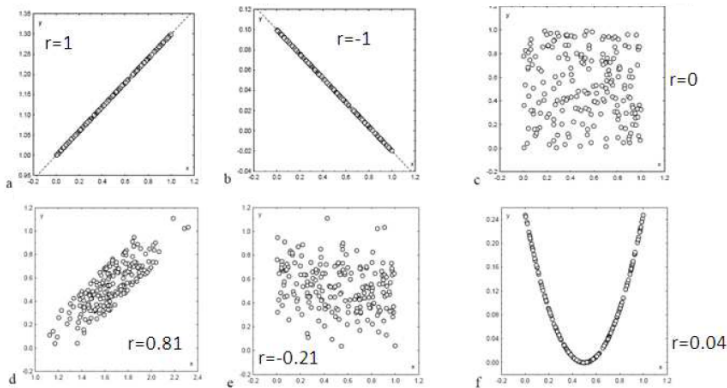
Correlação entre duas variáveis aleatórias

A correlação entre X_i e X_j , $i \neq j$, $i, j = 1, 2, \dots, p$, é uma medida relativa da covariância e é definida por:

$$\rho_{ij} \doteq \frac{\text{Cov}(X_i, X_j)}{\sqrt{\text{Var}(X_i) \text{Var}(X_j)}} = \frac{\sigma_{ij}}{\sqrt{\sigma_i^2 \sigma_j^2}}$$

- ▶ $-1 \leq \rho_{ij} \leq 1$, variando desde uma correlação forte negativa até uma correlação forte positiva. O sinal de ρ_{ij} depende do sinal de $\text{Cov}(X_i, X_j)$;
- ▶ $|\rho_{ij}| = 1$ significa que existe uma correlação linear perfeita entre X_i e X_j (ie., $X_i = aX_j + b$, com $a > 0$ se $\rho_{ij} = 1$ e $a < 0$ se $\rho_{ij} = -1$);
- ▶ $|\rho_{ij}| = 0$ significa que X_i e X_j não estão linearmente associados.

Correlação entre duas variáveis aleatórias - ilustração -



Marques de Sá, 2003 (pag 54)

(discutir o valor da covariância e da correlação para pontos sobre uma reta horizontal)

Matriz de covariâncias do vetor aleatório \mathbf{X}

- **Definição.** A variância do v.a. $\mathbf{X} = [X_1 \ X_2 \ \dots \ X_p]'$ é uma matriz $(p \times p)$ definida por:

$$Var(\mathbf{X}) = Cov(\mathbf{X}, \mathbf{X}) = E [(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})'] = E (\mathbf{X}\mathbf{X}') - \boldsymbol{\mu}\boldsymbol{\mu}'.$$

Esta matriz é habitualmente designada por *matriz de covariâncias* ou *matriz de variâncias-covariâncias* do v.a. \mathbf{X} sendo representada por $\boldsymbol{\Sigma}$. Em termos matriciais,

$$\boldsymbol{\Sigma} = Var(\mathbf{X}) = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \dots & \sigma_p^2 \end{bmatrix}.$$

Na diagonal principal estão as p variâncias, σ_i^2 , $i = 1, 2, \dots, p$.

Fazer exercício 1.2 (alínea 1)

Matriz de covariâncias do vetor aleatório \mathbf{X}

- ▶ **Propriedade.** A matriz de covariâncias Σ é
 - ▶ uma matriz simétrica: $\Sigma = \Sigma'$
 - ▶ uma matriz semidefinida positiva: $\mathbf{a}'\Sigma\mathbf{a} \geq 0, \forall \mathbf{a} \in \mathbb{R}^p \setminus \{\mathbf{0}\}$.
- ▶ **Propriedade.** A variância é um operador não linear tal que

$$\text{Var}(\alpha\mathbf{X} + \mathbf{c}) = \alpha^2 \text{Var}(\mathbf{X}),$$

para toda a constante $\alpha \in \mathbb{R}$ e vetor de números reais $\mathbf{c}_{(p \times 1)}$.

Demonstrar.

- ▶ **Propriedade.** Se \mathbf{X} ($p_1 \times 1$) e \mathbf{Y} ($p_2 \times 1$) forem independentes, então

$$\Sigma_{\mathbf{X},\mathbf{Y}} = \text{Cov}(\mathbf{X}, \mathbf{Y}) = \mathbf{0}.$$

onde $\Sigma_{\mathbf{X},\mathbf{Y}}$ e $\mathbf{0}$ são matrizes de dimensão $p_1 \times p_2$.

Matriz de covariâncias do vetor aleatório \mathbf{X}

- **Propriedade.** A variância da combinação linear $\mathbf{a}'\mathbf{X} = a_1X_1 + \cdots + a_pX_p$ (unidimensional) é um número real dado por:

$$\text{Var}(\mathbf{a}'\mathbf{X}) = \mathbf{a}' \text{Var}(\mathbf{X}) \mathbf{a},$$

para todo vetor de números reais $\mathbf{a}_{(p \times 1)}$, com \mathbf{X} v.a. em \mathbb{R}^p .

Ilustrar caso $p=2$, com $a_i = 1$. Relembrar a propriedade da variância da soma e comparar o resultado com a notação matricial acima. Associar esta propriedade com a condição da matriz de covariância ser uma matriz semidefinida positiva.

- **Propriedade.** A variância das q combinações lineares $\mathbf{B}'\mathbf{X}$ ($q \times 1$) é a matriz de covariâncias ($q \times q$) dada por:

$$\text{Var}(\mathbf{B}\mathbf{X}) = \mathbf{B} \text{Var}(\mathbf{X}) \mathbf{B}',$$

para toda a matriz de números reais $\mathbf{B}_{(p \times q)}$.

Matriz de covariâncias do vetor aleatório \mathbf{X}

- **Propriedade.** A variância da soma de dois v.a.'s \mathbf{X} ($p \times 1$) e \mathbf{Y} ($p \times 1$) é uma matriz ($p \times p$) dada por:

$$\text{Var}(\mathbf{X} + \mathbf{Y}) = \text{Var}(\mathbf{X}) + \text{Var}(\mathbf{Y}) + \text{Cov}(\mathbf{X}, \mathbf{Y}) + \text{Cov}(\mathbf{Y}, \mathbf{X})$$

- **Propriedade.** Para os v.a.'s \mathbf{X} ($p \times 1$), \mathbf{Y} ($p \times 1$) e \mathbf{Z} ($q \times 1$) se tem:

$$\text{Cov}(\mathbf{X} + \mathbf{Y}, \mathbf{Z}) = \text{Cov}(\mathbf{X}, \mathbf{Z}) + \text{Cov}(\mathbf{Y}, \mathbf{Z})$$

cujo resultado é uma matriz de dimensão ($p \times q$).

- **Propriedade.** Dados os v.a.'s \mathbf{X} ($p \times 1$) e \mathbf{Z} ($q \times 1$) e as matrizes de constantes \mathbf{A} ($a \times p$) e \mathbf{B} ($b \times q$), a matriz de covariâncias das matrizes de combinações lineares \mathbf{AX} e \mathbf{BY} é dada por:

$$\text{Cov}(\mathbf{AX}, \mathbf{BZ}) = \mathbf{A} \text{Cov}(\mathbf{X}, \mathbf{Z}) \mathbf{B}'$$

cujo resultado é uma matriz de dimensão ($a \times b$).

Fazer exercício 1.2 (álíneas 2,3,4)

Matriz de correlações do vetor aleatório \mathbf{X}

- **Definição.** Se todas as componentes de um v.a. $\mathbf{X} = [X_1 \ X_2 \ \dots \ X_p]'$ admitem variância não nula, define-se matriz de correlações do v.a. \mathbf{X} à matriz dada por:

$$\begin{bmatrix} 1 & \rho_{12} & \dots & \rho_{1p} \\ \rho_{21} & 1 & \dots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \dots & 1 \end{bmatrix}.$$

Trata-se de uma matriz $(p \times p)$ simétrica com 1s na diagonal principal.

- **Propriedade.** A matriz de correlações de um v.a. \mathbf{X} , de esperança $\boldsymbol{\mu}$ e matriz de covariâncias $\boldsymbol{\Sigma}$, coincide com a matriz de covariâncias do v.a. \mathbf{Z} de componentes normalizadas dado por

$$\mathbf{Z} = \left[\frac{X_1 - \mu_1}{\sigma_1} \quad \frac{X_2 - \mu_2}{\sigma_2} \quad \dots \quad \frac{X_p - \mu_p}{\sigma_p} \right]'$$

Provar esta propriedade usando notação matricial. Começar por verificar que

$\mathbf{Z} = \boldsymbol{\Sigma}_{diag}^{-1/2}(\mathbf{X} - \boldsymbol{\mu})$. Discutir a notação $\boldsymbol{\Sigma}_{diag}^{-1/2}$.

Amostra aleatória

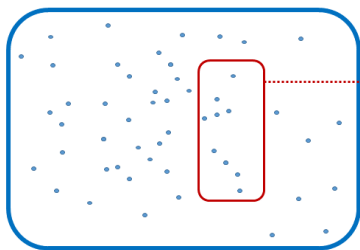
Uma amostra aleatória (a.a) de uma população p -multivariada é uma coleção de n extrações aleatórias independentes da população, a qual pode ser representada de duas formas:

- ▶ por uma sequência $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ de n vetores (coluna) p -dimensionais (i.e., $p \times 1$) e independentes; ou
- ▶ por uma matriz aleatória de dimensão $(n \times p)$, $\mathbf{X}_{n \times p}$, onde as n linhas são independentes e cada uma delas é um vetor aleatório p -dimensional $\mathbf{X}'_i, i = 1, 2, \dots, n$.

Uma realização da a.a. $\mathbf{X}_{n \times p}$ corresponde a uma matriz de dados $(n \times p)$ (ver slide 6) e é nessa forma que habitualmente se constrói o ficheiro de dados (ou seja, observações em linha e variáveis em coluna).

(Começar por relembrar o caso univariado)

Ilustração - caso multivariado -



Population: Patient registers in a health clinic

$$\mathbf{X} = [X_1 \ X_2 \ X_3 \ X_4]'$$

Sample (n=9):
(uma realização da a.a.)

Age group	Gender	Before diet	After diet
young	F	65	60
young	M	75	73
kid	M	35	35
adult	F	85	78
adult	F	67	60
kid	F	40	38
young	M	57	55
adult	M	65	62
adult	M	71	68

- ▶ Vetor aleatório: $\mathbf{X} = [X_1 \ X_2 \ X_3 \ X_4]'$, onde X_1, X_2, X_3, X_4 são as v.a.'s: X_1 = grupo etário; X_2 = género; X_3 = peso antes da dieta; X_4 = peso depois da dieta.
- ▶ $\mathbf{\tilde{X}}_{9 \times 4} = [\mathbf{\tilde{X}}_1 \ \mathbf{\tilde{X}}_2 \ \mathbf{\tilde{X}}_3 \ \mathbf{\tilde{X}}_4]$, onde $\mathbf{\tilde{X}}_j$ representa a a.a. relativa à variável X_j , $j = 1, 2, 3, 4$.

(Efetuar em paralelo a ilustração do caso multivariado)

Medidas amostrais -caso univariado-

Seja X_1, \dots, X_n uma a.a. de uma população (univariada) X .

Define-se:

- média amostral:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

- variância amostral (corrigida):

$$S_c^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

Para cada concretização x_1, \dots, x_n da a.a. X_1, \dots, X_n , uma grande variedade de outras medidas amostrais que podem ser calculadas.

Exemplo com o R:

```
> #dados
> x=c(1,1.5,1.5,2,2,1)
> mean(x); sd(x); var(x)
[1] 1.5
[1] 0.4472136
[1] 0.2
```

Exercício:

Explorar o comando `scale`

```
> ?scale
> scale(x)
> scale(x, scale=FALSE)
```


Medidas amostrais -caso multivariado-

Seja $\mathbf{X}_{n \times p} = [X_{ij}]$ uma a.a. de uma população (multivariada) \mathbf{X} .

Define-se:

► **média amostral**, $\bar{\mathbf{X}}$ ($p \times 1$):

$$\bar{\mathbf{X}} = \begin{bmatrix} \bar{X}_1 \\ \vdots \\ \bar{X}_p \end{bmatrix} = \begin{bmatrix} \frac{1}{n} \sum_{i=1}^n X_{i1} \\ \vdots \\ \frac{1}{n} \sum_{i=1}^n X_{ip} \end{bmatrix} =$$

$$= \frac{1}{n} \begin{bmatrix} X_{11} & \cdots & X_{n1} \\ \vdots & & \vdots \\ X_{1p} & \cdots & X_{np} \end{bmatrix} \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} = \frac{1}{n} \mathbf{X}' \mathbf{1}_{n \times 1}$$

Exemplo com o R:

```
> #dados
> x=matrix(x,nrow=3)
> colMeans(x)
[1] 1.333333 1.666667
```

Exercício:

Explorar o comando `apply`

```
> ?apply
> apply(x,2,mean)
> apply(x,1,mean)
```

(Alertar para as dimensões: matriz de dados, média amostral (modelo teórico e R))

Medidas amostrais -caso multivariado- (cont.)

- matriz de covariâncias amostral, \mathbf{S} ($p \times p$):

$$\mathbf{S} = \begin{bmatrix} S_1^2 & \cdots & S_{1p} \\ \vdots & & \vdots \\ S_{p1} & \cdots & S_p^2 \end{bmatrix} = \begin{bmatrix} \frac{1}{n} \sum_{i=1}^n (X_{i1} - \bar{X}_1)^2 & \cdots & \frac{1}{n} \sum_{i=1}^n (X_{i1} - \bar{X}_1)(X_{ip} - \bar{X}_p) \\ \vdots & & \vdots \\ \frac{1}{n} \sum_{i=1}^n (X_{ip} - \bar{X}_p)(X_{i1} - \bar{X}_1) & \cdots & \frac{1}{n} \sum_{i=1}^n (X_{ip} - \bar{X}_p)^2 \end{bmatrix}$$

$$= \frac{1}{n} \begin{bmatrix} \sum_{i=1}^n (X_{i1} - \bar{X}_1)^2 & \sum_{i=1}^n (X_{i1} - \bar{X}_1)(X_{i2} - \bar{X}_2) & \cdots & \sum_{i=1}^n (X_{i1} - \bar{X}_1)(X_{ip} - \bar{X}_p) \\ \sum_{i=1}^n (X_{i2} - \bar{X}_2)(X_{i1} - \bar{X}_1) & \sum_{i=1}^n (X_{i2} - \bar{X}_2)^2 & \cdots & \sum_{i=1}^n (X_{i2} - \bar{X}_2)(X_{ip} - \bar{X}_p) \\ \vdots & \vdots & & \vdots \\ \sum_{i=1}^n (X_{ip} - \bar{X}_p)(X_{i1} - \bar{X}_1) & \sum_{i=1}^n (X_{ip} - \bar{X}_p)(X_{i2} - \bar{X}_2) & \cdots & \sum_{i=1}^n (X_{ip} - \bar{X}_p)^2 \end{bmatrix}$$

Medidas amostrais -caso multivariado- (cont.)

$$\mathbf{S} = \frac{1}{n} \begin{bmatrix} \sum_{i=1}^n (x_{i1} - \bar{x}_1)^2 & \sum_{i=1}^n (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2) & \cdots & \sum_{i=1}^n (x_{i1} - \bar{x}_1)(x_{ip} - \bar{x}_p) \\ \sum_{i=1}^n (x_{i2} - \bar{x}_2)(x_{i1} - \bar{x}_1) & \sum_{i=1}^n (x_{i2} - \bar{x}_2)^2 & \cdots & \sum_{i=1}^n (x_{i2} - \bar{x}_2)(x_{ip} - \bar{x}_p) \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n (x_{ip} - \bar{x}_p)(x_{i1} - \bar{x}_1) & \sum_{i=1}^n (x_{ip} - \bar{x}_p)(x_{i2} - \bar{x}_2) & \cdots & \sum_{i=1}^n (x_{ip} - \bar{x}_p)^2 \end{bmatrix}$$

$$= \frac{1}{n} \begin{bmatrix} x_{11} - \bar{x}_1 & x_{21} - \bar{x}_1 & \cdots & x_{n1} - \bar{x}_1 \\ x_{12} - \bar{x}_2 & x_{22} - \bar{x}_2 & \cdots & x_{n2} - \bar{x}_2 \\ \vdots & \vdots & \ddots & \vdots \\ x_{1p} - \bar{x}_p & x_{2p} - \bar{x}_p & \cdots & x_{np} - \bar{x}_p \end{bmatrix} \begin{bmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & \cdots & x_{1p} - \bar{x}_p \\ x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2 & \cdots & x_{2p} - \bar{x}_p \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} - \bar{x}_1 & x_{n2} - \bar{x}_2 & \cdots & x_{np} - \bar{x}_p \end{bmatrix}$$

$$\mathbf{X}_{n \times p} - \mathbf{1}_{n \times 1} \bar{\mathbf{X}}' = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} - \begin{bmatrix} \bar{x}_1 & \bar{x}_2 & \cdots & \bar{x}_p \\ \bar{x}_1 & \bar{x}_2 & \cdots & \bar{x}_p \\ \vdots & \vdots & \ddots & \vdots \\ \bar{x}_1 & \bar{x}_2 & \cdots & \bar{x}_p \end{bmatrix}$$

Medidas amostrais -caso multivariado- (cont.)

$$\begin{aligned}
\mathbf{S} &= \frac{1}{n} (\mathbf{X}_{n \times p} - \mathbf{1}_{n \times 1} \bar{\mathbf{X}}')' (\mathbf{X}_{n \times p} - \mathbf{1}_{n \times 1} \bar{\mathbf{X}}') = \frac{1}{n} (\mathbf{X}' - \bar{\mathbf{X}} \mathbf{1}') (\mathbf{X} - \mathbf{1} \bar{\mathbf{X}}') = \\
&= \frac{1}{n} (\mathbf{X}' \mathbf{X} - \mathbf{X}' \mathbf{1} \bar{\mathbf{X}}' - \bar{\mathbf{X}} \mathbf{1}' \mathbf{X} + \bar{\mathbf{X}} \mathbf{1}' \mathbf{1} \bar{\mathbf{X}}') \\
&= \frac{1}{n} \left(\mathbf{X}' \mathbf{X} - \mathbf{X}' \mathbf{1} \frac{1}{n} \mathbf{1}' \mathbf{X} - \frac{1}{n} \mathbf{X}' \mathbf{1} \mathbf{1}' \mathbf{X} + \underbrace{\frac{1}{n} \mathbf{X}' \mathbf{1} \mathbf{1}' \mathbf{1} \frac{1}{n} \mathbf{1}' \mathbf{X}}_{\mathbf{1} \mathbf{1}' \text{ (mostrar)}} \right) \\
&= \frac{1}{n} \left(\mathbf{X}' \mathbf{X} - \frac{1}{n} \mathbf{X}' \mathbf{1} \mathbf{1}' \mathbf{X} \right) = \frac{1}{n} \mathbf{X}' \left(\mathbf{I}_n - \frac{1}{n} \mathbf{1} \mathbf{1}' \right) \mathbf{X}.
\end{aligned}$$

(Resumir as diferenças entre μ , \bar{x} , \bar{X} , μ , \bar{x} , \bar{X} ,
e entre σ^2 , s^2 , s_c^2 , S^2 , S_c^2 , Σ , s , s_c , \mathbf{S} , \mathbf{S}_c)

Medidas amostrais -caso multivariado- (cont.)

Exemplo com o R:

```
> #matriz de covariâncias (corrigida)
```

```
> var(x); cor(x)
```

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

```
[1,]  0.08333333 -0.08333333
```

```
[2,] -0.08333333  0.33333333
```

Exercícios:

1) Obter a matriz de correlações usando o comando `cor` e comparar com a seguinte matriz (ver slide 37 para o caso populacional)

```
> var(scale(x))
```

2) Explore o comando `eigen` e use-o para verificar que a matriz de covariâncias amostral é uma matriz semidefinida positiva (ver slide 26 para o caso populacional)

```
> eigen(scale(x))
```

Recordar que

- **Propriedade.** Uma matriz simétrica é semidefinida positiva sse todos os seus valores próprios são positivos.

Capítulo 2: Distribuição Normal Multivariada

Conteúdo

2.1. Revisão (caso univariado):

Distribuição normal univariada. Distribuições de amostragem

2.2. Normal multivariada

Definição e propriedades

Estimação de máxima verosimilhança

2.3. Distribuições de amostragem. Comportamento assintótico

2.4. Inferência sobre um vetor de médias

2.5. Validação da normalidade

Bibliografia de base

- ▶ Johnson, R.A. e Wichern, D.W. (1982). *Applied multivariate statistical analysis*. 3 Edição. Prentice-Hall. (Capítulos 4 e 5)

Distribuição normal univariada

Uma v.a. Z tem distribuição normal reduzida, $Z \sim N(0, 1)$, se a sua f.d.p. é

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}, \text{ para } z \in \mathbb{R}.$$

Considerando a transformação

$$Z = \frac{X - \mu}{\sigma}, \mu \in \mathbb{R}, \sigma > 0,$$

vem $X = \sigma Z + \mu$, $E(X) = \mu$ e $Var(X) \doteq E((X - E(X))^2) = \sigma^2$. A f.d.p. de X é obtida da f.d.p. de Z usando o Jacobiano da transformação linear:

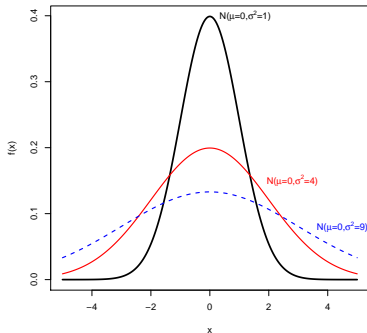
$$f_X(x) = f_Z(z) |J|, \text{ onde } J = \frac{dz}{dx} = \frac{1}{\sigma}.$$

Deste modo, denota-se $X \sim N(\mu, \sigma^2)$ e tem-se:

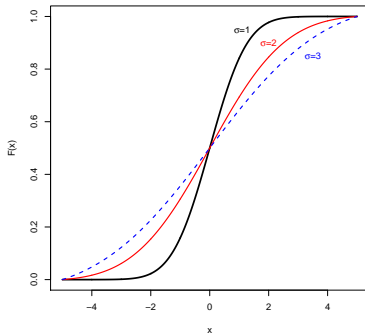
$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x - \mu}{\sigma}\right)^2\right), x, \mu \in \mathbb{R}, \sigma > 0.$$

Distribuição normal univariada, $X \sim N(\mu, \sigma^2)$

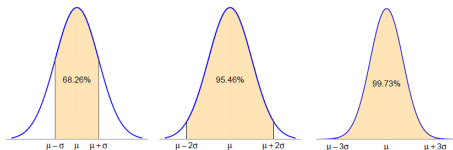
Função densidade da normal (univariada)



Função de distribuição da normal (univariada)



$$P(|X - \mu| \leq a\sigma) \simeq \begin{cases} 0.68, & \text{se } a = 1 \\ 0.95, & \text{se } a = 2 \\ 0.99, & \text{se } a = 3 \end{cases}$$



Distribuições de amostragem univariadas

Seja X_1, \dots, X_n uma amostra aleatória (a.a.) de uma população (univariada) X . Então, define-se

- ▶ média amostral:

$$\bar{X} = \sum_{i=1}^n X_i / n$$

- ▶ variância amostral (corrigida):

$$S_c^2 = \sum_{i=1}^n (X_i - \bar{X})^2 / (n - 1)$$

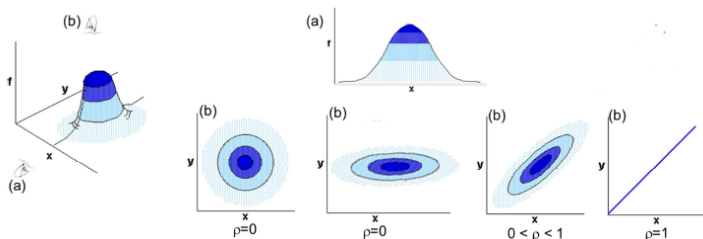
Para amostras de uma população normal,

- ▶ Distribuição da média amostral ...
- ▶ Distribuição da variância amostral ...

Distribuição normal multivariada

A f.d.p. conjunta da distribuição normal multivariada é uma generalização da f.d.p. da distribuição normal univariada.

Visualização de uma distribuição normal bivariada:



Distribuição normal multivariada

Seja $\mathbf{Z} = [Z_1 \ Z_2 \ \dots \ Z_p]' = (Z_1, Z_2, \dots, Z_p)'$ um vetor aleatório p -dimensional tal que $Z_i, i = 1, 2, \dots, p$ são variáveis aleatórias i.i.d. com $Z_i \sim N(0, 1)$. Portanto,

$$E(\mathbf{Z}) = \mathbf{0} \text{ e } \text{Var}(\mathbf{Z}) = \mathbf{I}.$$

A f.d.p. conjunta do vetor aleatório \mathbf{Z} é dada por

$$f(\mathbf{z}) = \prod_{i=1}^p f(z_i) = \prod_{i=1}^p \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z_i^2} = \frac{1}{(2\pi)^{\frac{p}{2}}} e^{-\frac{1}{2}\sum_{i=1}^p z_i^2} = \frac{1}{(2\pi)^{\frac{p}{2}}} e^{-\frac{1}{2}\mathbf{z}'\mathbf{z}}$$

Diz-se que \mathbf{Z} tem distribuição normal reduzida p -dimensional, e denota-se

$$\mathbf{Z} \sim N_p(\mathbf{0}, \mathbf{I}),$$

onde p indica a dimensão da distribuição e corresponde ao número de variáveis no vetor \mathbf{Z} .

Distribuição normal multivariada

Considerando a transformação em termos matriciais similar a

$Z = (\sigma)^{-1}(X - \mu)$, ie.,

$$\mathbf{Z} = \left(\boldsymbol{\Sigma}^{\frac{1}{2}}\right)^{-1}(\mathbf{X} - \boldsymbol{\mu}),$$

tem-se que $\mathbf{X} = \boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{Z} + \boldsymbol{\mu}$, $E(\mathbf{X}) = \boldsymbol{\mu}$ e $Var(\mathbf{X}) = \boldsymbol{\Sigma}$.

(Identificar as matrizes $\boldsymbol{\Sigma}^{1/2}$ e $\boldsymbol{\Sigma}^{-1/2}$.)

A f.d.p. de \mathbf{X} é obtida da f.d.p. de \mathbf{Z} usando o Jacobiano da transformação linear:

$$f(\mathbf{x}) = f(\mathbf{z}) \det(\mathbf{J}), \text{ onde } \mathbf{J} = \frac{\partial(z_1, z_2, \dots, z_p)}{\partial(x_1, x_2, \dots, x_p)}.$$

Como $\det(\mathbf{J}) = \det(\boldsymbol{\Sigma}^{-\frac{1}{2}})$ e $\mathbf{Z}'\mathbf{Z} = (\mathbf{X} - \boldsymbol{\mu})'(\boldsymbol{\Sigma})^{-1}(\mathbf{X} - \boldsymbol{\mu})$, então

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2} |\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}, \text{ para } \mathbf{x}, \boldsymbol{\mu} \in \mathbb{R}^p, |\boldsymbol{\Sigma}| > 0.$$

(Recorrendo a propriedades dos determinantes, mostrar que $\det(\mathbf{J}) = 1/\sqrt{\det(\boldsymbol{\Sigma})}$.)

Distribuição normal multivariada

Resumindo... o vetor aleatório $\mathbf{X} = (X_1, X_2, \dots, X_p)'$ ($p \times 1$) segue uma distribuição normal p -variada com parâmetros $\boldsymbol{\mu}$ e $\boldsymbol{\Sigma}$, e escreve-se

$$\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

sse a sua f.d.p. é da forma

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2} |\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}, \text{ para } \mathbf{x}, \boldsymbol{\mu} \in \mathbb{R}^p, |\boldsymbol{\Sigma}| > 0.$$

Neste caso, tem-se que $E(\mathbf{X}) = \boldsymbol{\mu}$ e $Var(\mathbf{X}) = \boldsymbol{\Sigma}$.

TPC: Verificar que a f.d.p. para o caso $p = 2$ com $\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$ e $\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$ é dada por:

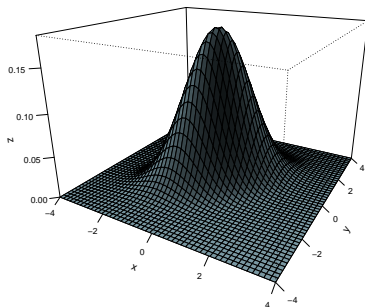
$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp \left\{ -\frac{1}{2(1-\rho^2)} \left(\frac{x_1 - \mu_1}{\sigma_1} \right)^2 - 2\rho \left(\frac{x_1 - \mu_1}{\sigma_1} \right) \left(\frac{x_2 - \mu_2}{\sigma_2} \right) + \left(\frac{x_2 - \mu_2}{\sigma_2} \right)^2 \right\}$$

onde $\rho = \sigma_{12}/(\sigma_1\sigma_2)$.

Distribuição normal multivariada

Ilustração da f.d.p. da normal bivariada, $\mathbf{Z} = (X, Y)' \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, com

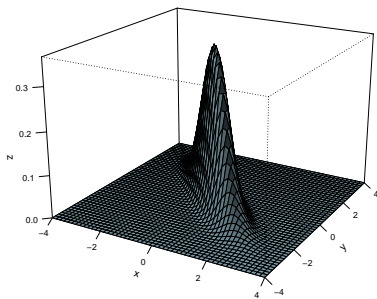
$$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \text{ e } \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$



Distribuição normal multivariada

Ilustração da f.d.p. da normal bivariada, $\mathbf{Z} = (X, Y)' \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, com

$$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \text{ e } \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix} = \begin{bmatrix} 1 & -0.9 \\ -0.9 & 1 \end{bmatrix}$$



Distribuição normal multivariada

A densidade da distribuição $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ é constante sobre superfícies onde a distância $(\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$ é constante. Essas superfícies são elipsóides da forma

$$(\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) = c^2, \quad c \in \mathbb{R} \setminus \{0\}.$$

e são chamados de **contornos de densidade constante**.

TPC: Explorar a matriz $\boldsymbol{\Sigma}$ no caso de $p = 2$. Concretamente, determinar a sua inversa ($\boldsymbol{\Sigma}^{-1}$), os valores e vetores próprios de $\boldsymbol{\Sigma}$. Particularizar para o caso de $\sigma_1 = \sigma_2$, identificando o ângulo que os vetores próprios, no plano XOY , fazem com os eixos dos xx e dos yy .

Distribuição normal multivariada

Os contornos da distribuição definidos por

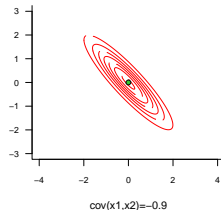
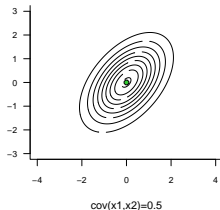
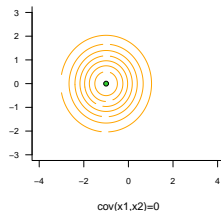
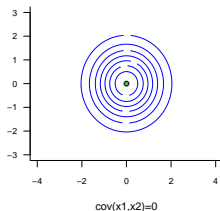
$$\{\mathbf{x} \in \mathbb{R}^p : (\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) = c^2\}$$

- ▶ têm centro em $\boldsymbol{\mu}$.
- ▶ têm semi-eixos com direções definidas pelos vetores próprios da matriz $\boldsymbol{\Sigma}$ ou $\boldsymbol{\Sigma}^{-1}$.
- ▶ têm comprimento de semi-eixos dado por $\sqrt{\frac{c^2}{\lambda_i^*}}$, onde $\lambda_i^* = \frac{1}{\lambda_i}$ são os valores próprios de $\boldsymbol{\Sigma}^{-1}$ e λ_i os valores próprios de $\boldsymbol{\Sigma}$.

Demonstrar.

Distribuição normal multivariada

Forma dos contornos da f.d.p. da normal $(X_1, X_2)' \sim N_2(\mu, \Sigma)$.



Teorema de Cramér-Wold

O Teorema de Cramér-Wold diz-nos que a distribuição multivariada de qualquer vetor aleatório p -dimensional \mathbf{X} é completamente determinada pelo conjunto de todas as distribuições univariadas de combinações lineares

$$\mathbf{a}'\mathbf{X} = a_1X_1 + a_2X_2 + \dots + a_pX_p = \sum_{i=1}^p a_iX_i,$$

onde $\mathbf{a} \in \mathbb{R}^p$ (vetor não aleatório).

Definição da normal multivariada usando Cramér-Wold

Diz-se que um v.a. \mathbf{X} normal p -variado se e só se $\mathbf{a}'\mathbf{X}$ tem distribuição normal univariada para todos os vetores $\mathbf{a} \in \mathbb{R}^p$.

Teorema de Cramér-Wold - Interpretação geométrica

Se \mathbf{x} for um ponto aleatório no espaço \mathbb{R}^p , então $\mathbf{a}'\mathbf{x}$ pode ser considerado como a projeção de \mathbf{x} sobre um sub-espço unidimensional. Portanto, a definição anterior implica que a projeção de \mathbf{x} sobre todos os sub-espços unidimensionais tem uma distribuição normal univariada.

Definição da normal multivariada usando a interpretação geométrica de Cramér-Wold

Diz-se que \mathbf{X} tem distribuição normal p -variada se e só se a sua projeção $\mathbf{a}'\mathbf{X}$ em qualquer subespaço unidimensional tem distribuição normal univariada.

Esta definição induz, sobre \mathbf{X} normal p -variada, propriedades de normalidade após ser transformado por uma qualquer translação, rotação ou projeção e assim pode deduzir-se vários resultados da normal multivariada sem recorrer à f.d.p. conjunta.

Propriedades da distribuição normal multivariada

Teorema:

Se \mathbf{X} tem distribuição normal p -variada e se $\mathbf{Y} = \mathbf{AX} + \mathbf{c}$, onde $\mathbf{A}_{(q \times p)}$ é uma matriz de números reais e $\mathbf{c}_{p \times 1} \in \mathbb{R}^p$, então \mathbf{Y} tem uma distribuição normal q -variada.

Dem.: Seja \mathbf{b} um vetor q -dimensional qualquer fixado. Então,

$$\mathbf{b}'\mathbf{Y} = \mathbf{b}'\mathbf{AX} + \mathbf{b}'\mathbf{c} = \mathbf{a}'\mathbf{X} + \mathbf{b}'\mathbf{c},$$

onde $\mathbf{a}' = \mathbf{b}'\mathbf{A}$. Uma vez que $\mathbf{a} = \mathbf{A}'\mathbf{b}$ e \mathbf{X} tem distribuição normal p -variada, então, de acordo com a definição, $\mathbf{a}'\mathbf{X}$ tem distribuição normal univariada. Consequentemente, $\mathbf{b}'\mathbf{Y}$ tem também distribuição normal univariada para qualquer vetor \mathbf{b} e, portanto, \mathbf{Y} tem distribuição normal q -variada.

Propriedades da distribuição normal multivariada

Concretamente, se $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ e

1. ... $Y = \mathbf{a}'\mathbf{X} = \sum_{i=1}^p a_i X_i$, com $\mathbf{a}_{p \times 1} \in \mathbb{R}^p \setminus \{\mathbf{0}\}$, uma combinação linear das componentes de \mathbf{X} , então

$$Y \sim N(\mathbf{a}'\boldsymbol{\mu}, \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}).$$

2. ... $\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{c}$, com $\mathbf{A}_{(q \times p)}$ matriz de números reais não singular e $\mathbf{c}_{q \times 1} \in \mathbb{R}^q$, q combinações lineares das componentes de \mathbf{X} , então

$$\mathbf{Y} \sim N_q(\mathbf{A}\boldsymbol{\mu} + \mathbf{c}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}').$$

3. ... $\boldsymbol{\Sigma} > 0$, então

$$\mathbf{Y} = \boldsymbol{\Sigma}^{-1/2}(\mathbf{X} - \boldsymbol{\mu}) \sim N_p(\mathbf{0}, \mathbf{I})$$

i.e., as componentes Y_j do v.a. \mathbf{Y} são i.i.d. com distribuição $N(0, 1)$.
A transformação \mathbf{Y} é conhecida por *transformação de Mahalanobis*.

Fazer exerc. 2.5. TPC: Demonstrar as propriedades 1 e 2 acima.

Propriedades da distribuição normal multivariada

Concretamente, se $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ e

4. ... e $|\boldsymbol{\Sigma}| > 0$, então

$$(\mathbf{X} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{X} - \boldsymbol{\mu}) \sim \chi_p^2$$

,

Esta ultima propriedade permite definir a escolha de c^2 nos contornos de densidade constante. Se $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, então

$$U = (\mathbf{X} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{X} - \boldsymbol{\mu}) \sim \chi_p^2.$$

Usando este resultado pode calcular-se a probabilidade de um ponto \mathbf{x} pertencer ao interior de um elipsóide. O elipsóide definido pelo percentil $c^2 = \chi_{p,1-\alpha}^2$ tem probabilidade $1 - \alpha$, ou seja,

$$P(U \leq c^2) = 1 - \alpha, \quad (0 < \alpha < 1).$$

Propriedades da distribuição normal multivariada

Por vezes interessa particionar o v.a. $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ em dois subvetores \mathbf{X}_1 e \mathbf{X}_2 .

Teorema:

Se $\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, com \mathbf{X}_1 ($q \times 1$) e \mathbf{X}_2 $((p - q) \times 1)$,

$\mathbf{X}_{2,1} = \mathbf{X}_2 - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\mathbf{X}_1$, com a matriz de covariâncias particionada do seguinte modo:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}$$

então:

$$\mathbf{X}_1 \sim N_q(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}) \quad , \quad \mathbf{X}_{2,1} \sim N_{p-q}(\boldsymbol{\mu}_{2,1}, \boldsymbol{\Sigma}_{22,1})$$

e \mathbf{X}_1 e $\mathbf{X}_{2,1}$ independentes com

$$\boldsymbol{\mu}_{2,1} = \boldsymbol{\mu}_2 - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\mu}_1 \quad \text{e} \quad \boldsymbol{\Sigma}_{22,1} = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$$

Distribuição da a.a. da normal multivariada

1. Se $\underline{\mathbf{X}}_{n \times p}$ é uma matriz aleatória multinormal com $\mathbf{X}'_i \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}), i = 1, 2, \dots, n$, então

$$\bar{\mathbf{X}} \sim N_p(\boldsymbol{\mu}, n^{-1}\boldsymbol{\Sigma}).$$

2. Sejam $\mathbf{A}_{m \times n}$ e $\mathbf{B}_{p \times q}$ matrizes reais, $\underline{\mathbf{X}}_{n \times p}$ uma a.a. multinormal com $\mathbf{X}'_i \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}), i = 1, 2, \dots, n$, e $\underline{\mathbf{Y}}_{m \times q} = \mathbf{A}\mathbf{X}\mathbf{B}$. Então, $\underline{\mathbf{Y}}$ é uma matriz aleatória multinormal se e só se verificar as duas condições:

- $\mathbf{A}\mathbf{1} = \alpha\mathbf{1}$ para algum escalar α , ou $\mathbf{B}'\boldsymbol{\mu} = 0$
- $\mathbf{A}\mathbf{A}' = \beta\mathbf{I}$ para algum escalar β , ou $\mathbf{B}'\boldsymbol{\Sigma}\mathbf{B} = \mathbf{0}$

Se as duas condições acima se verificarem, tem-se

$$\mathbf{Y}_i \sim N_q(\alpha\mathbf{B}'\boldsymbol{\mu}, \beta\mathbf{B}'\boldsymbol{\Sigma}\mathbf{B}), i = 1, 2, \dots, q.$$

TPC: Usando propriedades da distribuição normal multivariada, demonstrar a propriedade 1 e interpretar a propriedade 2.

Estimador e estimativa

Seja $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ uma a.a. de uma população p dimensional com parâmetro $\boldsymbol{\theta} = [\theta_1 \ \theta_2 \ \dots \ \theta_q]$ e seja $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ uma amostra de observações (i.e., uma concretização da a.a.).

- ▶ **Estimador de $\boldsymbol{\theta}$.** É qualquer função da a.a., que não depende de $\boldsymbol{\theta}$ e é definida para obter uma *aproximação* para $\boldsymbol{\theta}$.

Exemplo. $\bar{\mathbf{X}}$ é um estimador de $\boldsymbol{\mu}$; \mathbf{S} é um estimador de $\boldsymbol{\Sigma}$.

- ▶ **Estimativa de $\boldsymbol{\theta}$.** É uma concretização de um estimador de $\boldsymbol{\theta}$ calculado a partir de uma amostra observada.

Exemplo. $\bar{\mathbf{x}} = [6 \ 10]'$ é uma estimativa de $\boldsymbol{\mu}$ obtida da matriz de dados

$$\mathbf{x} = \begin{bmatrix} 2 & 8 & 6 & 8 \\ 12 & 9 & 9 & 10 \end{bmatrix}'$$

Métodos de estimação

Existem diferentes métodos para construir estimadores para θ .

- ▶ **Método dos Momentos:** uma estimativa dos momentos (e.m.) de θ é uma concretização possível do estimador que iguala os momentos populacionais aos momentos amostrais.

Exemplo. Seja $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ uma a.a. de $\mathbf{X} \sim N_p(\mu, \Sigma)$.

$$E(\mathbf{X}) = \bar{\mathbf{x}} \Leftrightarrow \mu = \bar{\mathbf{x}}$$

Logo, $\bar{\mathbf{X}}$ é E. M. e $\hat{\mu} = \bar{\mathbf{x}}$ é uma e.m. de μ .

- ▶ **Método da Máxima Verosimilhança:** a estimativa de máxima verosimilhança (e.m.v.) de θ é a solução $\hat{\theta}$ que maximiza a função de verosimilhança da amostra definida por

$$L(\theta; \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n) \doteq f(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n; \theta) = \prod_{j=1}^n f(\mathbf{x}_j; \theta)$$

E. M. V. dos parâmetros de uma $N_p(\mu, \Sigma)$

No caso de $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ ser uma a.a. de uma população $\mathbf{X} \sim N_p(\mu, \Sigma)$, a função de verosimilhança da amostra é dada por:

$$\begin{aligned} L(\mu, \Sigma; \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n) &= \prod_{j=1}^n f(\mathbf{x}_j; \mu, \Sigma) \\ &= \frac{1}{(2\pi)^{np/2} |\Sigma|^{n/2}} e^{-\frac{1}{2} \sum_{j=1}^n (\mathbf{x}_j - \mu)' \Sigma^{-1} (\mathbf{x}_j - \mu)} \end{aligned}$$

Teorema. Seja $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ uma a.a. de $\mathbf{X} \sim N_p(\mu, \Sigma)$.
As e.m.v. de μ e Σ são dadas por $\bar{\mathbf{x}}$ e \mathbf{s} , respetivamente.

E. M. V. - Propriedades

Lema. Os E. M. V. são invariantes, ou seja, se $\hat{\theta}$ é uma e.m.v. de θ e g é uma função de θ , então a e.m.v. de $g(\theta)$ é dada por $g(\hat{\theta})$.

Exemplos.

► Caso Univariado.

Se s^2 é a e.m.v. de σ^2 , então s é a e.m.v. de σ .

► Caso Bivariado.

Se $\mathbf{s} = \begin{bmatrix} s_1^2 & s_{12} \\ s_{12} & s_2^2 \end{bmatrix}$ é a e.m.v. de $\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$,

então e.m.v. do coeficiente de correlação $\rho_{12} = \frac{\sigma_{12}}{\sqrt{\sigma_1^2 \sigma_2^2}}$ é

$$\hat{\rho}_{12} = \frac{s_{12}}{\sqrt{s_1^2 s_2^2}}.$$

E. M. V. de uma população $N_p(\mu, \Sigma)$

Seja $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ uma a.a. de uma população $\mathbf{X} \sim N_p(\mu, \Sigma)$. Então

- ▶ $E(\bar{\mathbf{X}}) = \mu$, ou seja, $\bar{\mathbf{X}}$ é um estimador centrado de μ ;
- ▶ $E\left(\frac{n\mathbf{S}}{n-1}\right) = \Sigma$, ou seja, $\mathbf{S}_c = \frac{1}{n-1} \sum_{j=1}^n (\mathbf{x}_j - \bar{\mathbf{X}})(\mathbf{x}_j - \bar{\mathbf{X}})'$ é um estimador centrado da matriz de covariâncias Σ .
- ▶ a função densidade conjunta da a.a. apenas depende das observações $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ por meio de $\bar{\mathbf{x}}$ e de $n\mathbf{s}$, pelo que $\bar{\mathbf{X}}$ e \mathbf{S}_c são estatísticas suficientes, i.e., em $\bar{\mathbf{x}}$ e em \mathbf{s}_c está contida toda a informação da matriz de dados.

Distribuição amostral de $\bar{\mathbf{X}}$ e de \mathbf{S}_c

Seja $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ uma a.a. de uma população $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Então

- ▶ $\bar{\mathbf{X}} \sim N_p(\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{\Sigma})$
- ▶ $(n-1)\mathbf{S}_c \sim W_p(n-1, \boldsymbol{\Sigma})$ (diz-se que segue uma **distribuição de Wishart de ordem p , $n-1$ gl e matriz escalar $\boldsymbol{\Sigma}$**)
- ▶ $n(\bar{\mathbf{X}} - \boldsymbol{\mu})' \mathbf{S}_c^{-1} (\bar{\mathbf{X}} - \boldsymbol{\mu}) \sim T(p, n-1)$ (diz-se que segue uma **distribuição de T^2 de Hotelling de parâmetros p e $n-1$**)

- ▶ $\bar{\mathbf{X}}$ e \mathbf{S}_c são independentes.

Em particular, quando $p = 1$ (caso univariado), tem-se que

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right), \quad \frac{(n-1)S_c}{\sigma^2} \sim \chi_{n-1}^2 \quad \text{e} \quad \frac{\sqrt{n}(\bar{X} - \mu)}{S_c} \sim t_{n-1}$$

Distribuição de Wishart p dimensional

- ▶ A distribuição de Wishart, $W_p(n-1, \mathbf{\Sigma})$, é uma generalização da distribuição χ^2 para o contexto multivariado ($p > 1$);
- ▶ Em particular quando $\mathbf{\Sigma} = \mathbf{I}_p$,

$$W_p(n, \mathbf{I}) = \sum_{j=1}^n \mathbf{Z}_j \mathbf{Z}_j' , \text{ com } \mathbf{Z}_1, \mathbf{Z}_2, \dots, \mathbf{Z}_n \sim N_p(0, \mathbf{I}), \text{ independentes}$$

i.e., resulta da soma dos *quadrados* de vetores aleatórios i.i.d. com distribuição normal multivariada (analogamente à distribuição χ^2 que resulta da soma dos quadrados de variáveis aleatórias i.i.d. com distribuição normal univariada).

Nota. No caso univariado, se tem:

$$\chi_n^2 = \sum_{j=1}^n Z_j^2 , \text{ com } Z_1, Z_2, \dots, Z_n \sim N(0, 1) , \text{ independentes}$$

Distribuição T^2 de Hotelling

A distribuição T^2 de Hotelling é uma generalização da distribuição t de *Student* e é usada para testar médias de populações multivariadas.

Definição. Sejam \mathbf{Y} um vetor aleatório p -dimensional com distribuição

$N_p(\mathbf{0}, \mathbf{I})$ e \mathbf{M} uma matriz aleatória com distribuição $\mathbf{W}_p(n, \mathbf{I})$. Se \mathbf{Y} e \mathbf{M} forem independentes, diz-se que

$$n \mathbf{Y}' \mathbf{M}^{-1} \mathbf{Y},$$

tem distribuição T^2 de Hotelling de parâmetros p e n , e escreve-se $T^2(p, n)$.

Desta definição resulta então que, se $\mathbf{Y} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\mathbf{M} \sim \mathbf{W}_p(n, \boldsymbol{\Sigma})$, \mathbf{Y} e \mathbf{M} são independentes, então

$$n(\mathbf{Y} - \boldsymbol{\mu})' \mathbf{M}^{-1} (\mathbf{Y} - \boldsymbol{\mu}) \sim T^2(p, n)$$

Distribuição T^2 de Hotelling

Propriedade. Seja $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ uma a.a. de uma população com distribuição $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Então,

$$T^2 = n (\bar{\mathbf{X}} - \boldsymbol{\mu})' \mathbf{S}_c^{-1} (\bar{\mathbf{X}} - \boldsymbol{\mu}) \sim T^2(p, n-1)$$

e

$$P \left(T^2 > \frac{(n-1)p}{n-p} F_{p, n-p; 1-\alpha} \right) = \alpha$$

onde $F_{a, b; 1-\alpha}$ representa o quantil de ordem $1 - \alpha$ de uma distribuição de Fisher de g.l. (a, b) .

(Fazer Exerc 2.10 alíneas 1 e 2)

Distribuição T^2 de Hotelling

A estatística T^2 de Hotelling é uma generalização do quadrado das distâncias das observações relativamente à média para o contexto multivariado;

► Univariado:
$$t^2 = n \frac{(\bar{X} - \mu)^2}{S_c^2}$$

$$= \underbrace{\sqrt{n}(\bar{X} - \mu)}_{N(0, \sigma^2)} \underbrace{\left(\frac{(n-1)S_c^2}{n-1} \right)^{-1}}_{\substack{\sigma^2 \chi_{n-1}^2 \\ g.l.}} \sqrt{n}(\bar{X} - \mu) = t_{n-1}^2 = F_{1, n-1}$$

► Multivariado:
$$T^2 = n (\bar{\mathbf{X}} - \mu)' \mathbf{S}_c^{-1} (\bar{\mathbf{X}} - \mu)$$

$$\underbrace{\sqrt{n}(\bar{\mathbf{X}} - \mu)'}_{N_p(0, \Sigma)} \underbrace{\left(\frac{(n-1)\mathbf{S}_c}{n-1} \right)^{-1}}_{\substack{W_p(n-1, \Sigma) \\ g.l.}} \underbrace{\sqrt{n}(\bar{\mathbf{X}} - \mu)}_{N_p(0, \Sigma)} = \frac{(n-1)p}{n-p} F_{p, n-p}$$

TLC (multivariado)

Sejam $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ n v.a. i.i.d. com $E(\mathbf{X}_i) = \boldsymbol{\mu}$ e $Var(\mathbf{X}_i) = \boldsymbol{\Sigma}$. Então,

$$\sqrt{n}(\bar{\mathbf{X}} - \boldsymbol{\mu}) \xrightarrow[n \rightarrow +\infty]{d} \mathbf{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$$

Relembrar o significado de convergência em distribuição

(Fazer Exerc 2.11 alínea 1)

Transformações de estatísticas

Em alguns problemas práticos interessa analisar uma função de parâmetros que são estimados usando estatísticas assintoticamente com distribuição normal. Nesse caso, tem-se o seguinte resultado:

Teorema: Se

- ▶ $\sqrt{n}(\mathbf{Y} - \boldsymbol{\mu}) \xrightarrow[n \rightarrow +\infty]{d} \mathbf{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$
- ▶ $\mathbf{h}(\mathbf{x}) : \mathbb{R}^p \rightarrow \mathbb{R}^q$, tal que

$$\mathbf{h}(\mathbf{x}) = [h_1(\mathbf{x}) \ h_2(\mathbf{x}) \ \cdots \ h_q(\mathbf{x})]'$$

com h_1, h_2, \dots, h_q diferenciáveis em $\boldsymbol{\mu} \in \mathbb{R}^p$.

Então

$$\sqrt{n}(\mathbf{h}(\mathbf{Y}) - \mathbf{h}(\boldsymbol{\mu})) \xrightarrow[n \rightarrow +\infty]{d} \mathbf{N}_q(\mathbf{0}, \mathbf{D}'\boldsymbol{\Sigma}\mathbf{D})$$

onde $\mathbf{D} = \left[\frac{\partial h_j(\mathbf{x})}{\partial x_i} \right]_{|\mathbf{x}=\boldsymbol{\mu}}$ é a matriz $(p \times q)$ de todas as derivadas parciais.

(Fazer Exerc 2.11, alínea 3)

Inferência

Problemas de testes e intervalos de confiança para a média de uma distribuição normal

- ▶ a maioria dos procedimentos multivariados são generalizações de procedimentos univariados

(Relembrar o teste para a média de uma normal univariada, caso variância desconhecida)

Teste para a média de uma normal multivariada (parâmetros desconhecidos)

- ▶ Formulação do teste de hipóteses:

$$H_0 : \mu = \mu_0 \quad \text{vs} \quad H_1 : \mu \neq \mu_0$$

- ▶ Estatística de teste:

$$T^2|_{H_0} \doteq n (\bar{\mathbf{X}} - \mu_0)' \mathbf{S}_c^{-1} (\bar{\mathbf{X}} - \mu_0) \sim \frac{(n-1)p}{n-p} F_{p,n-p}$$

(Estatística de T^2 de Hotelling)

- ▶ Regra de teste: Rejeita-se H_0 se a distância generalizada de T^2 de Hotelling para a amostra dada é demasiado grande, ou seja, se:

$$T_{obs}^2 > \frac{(n-1)p}{n-p} F_{p,n-p;1-\alpha}$$

onde $F_{p,n-p;1-\alpha}$ é o quantil de ordem $(1-\alpha)100\%$ de uma distribuição de Fisher com $(p, n-p)$ g.l. (Fazer Exerc 2.10 alínea 3)

Regiões de confiança

Os testes de hipóteses não permitem avaliar a precisão de uma estimativa. No caso univariado, consideram-se intervalos de confiança. No caso multivariado, consideram-se regiões de confiança.

Definição. Seja $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ é uma a.a. de uma população \mathbf{X} , $\boldsymbol{\theta}$ um parâmetro desconhecido de \mathbf{X} e seja $\tilde{\mathbf{X}}$ uma matriz de dados de \mathbf{X} . Uma região $R(\tilde{\mathbf{X}})$, determinada pela matriz de dados $\tilde{\mathbf{X}}$, diz-se ser uma região de confiança com um grau de confiança de $100(1 - \alpha)\%$, $0 < \alpha < 1$, se antes da amostra ser seleccionada,

$$P(R(\tilde{\mathbf{X}}) \text{ cobrir o verdadeiro valor de } \boldsymbol{\theta}) = 1 - \alpha$$

Regiões de confiança

A estatística T^2 de Hotelling permite estabelecer o seguinte:

Propriedade. Uma região $R(\underline{\mathbf{X}})$, a $100(1 - \alpha)\%$ de confiança para a média de uma população normal p -dimensional será um elipsóide com centro na média amostral $\bar{\mathbf{x}}$ e dado pelo conjunto

$$R(\underline{\mathbf{X}}) = \left\{ \boldsymbol{\mu} \in \mathbb{R}^p : n(\bar{\mathbf{x}} - \boldsymbol{\mu})' \mathbf{S}_c^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu}) \leq \frac{(n-1)p}{n-p} F_{p, n-p; 1-\alpha} \right\}$$

onde $\bar{\mathbf{x}} = \frac{1}{n} \sum_{j=1}^n \mathbf{x}_j$, $\mathbf{S}_c = \frac{1}{n-1} \sum_{j=1}^n (\mathbf{x}_j - \bar{\mathbf{x}})(\mathbf{x}_j - \bar{\mathbf{x}})'$ e $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ são as n observações de dimensão p da amostra de tamanho n .

Região de confiança para a média de uma normal p dimensional

$$R(\mathbf{X}) = \left\{ \boldsymbol{\mu} : n(\bar{\mathbf{x}} - \boldsymbol{\mu})' \mathbf{S}_c^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu}) \leq \frac{(n-1)p}{n-p} F_{p, n-p; 1-\alpha} \right\}$$

Se para uma amostra se tem que $\boldsymbol{\mu}_0 \in R(\mathbf{X})$, tal significa que o quadrado da distância dada por $(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)' \mathbf{S}_c^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu}_0)$ não é maior do que $\frac{(n-1)p}{n-p} F_{p, n-p; 1-\alpha}$. Consequentemente, no problema de testar

$$H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0 \quad \text{vs} \quad H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$$

aquela amostra não mostra evidência para se rejeitar H_0 , ao nível de significância de $\alpha 100\%$.

Comparações simultâneas de combinações lineares da média. Intervalos de Roy

De um resultado da Teoria da Matrices, pode ser demonstrado que:

$$\max_{\mathbf{a} \neq \mathbf{0}} \frac{(\mathbf{a}'(\bar{\mathbf{x}} - \boldsymbol{\mu}))^2}{\mathbf{a}'\mathbf{S}_c\mathbf{a}} = (\bar{\mathbf{x}} - \boldsymbol{\mu})' \mathbf{S}_c^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu})$$

Assim, a região

$$R(\underline{\mathbf{X}}) = \left\{ \boldsymbol{\mu} : n(\bar{\mathbf{x}} - \boldsymbol{\mu})' \mathbf{S}_c^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu}) \leq \frac{(n-1)p}{n-p} F_{p, n-p; 1-\alpha} \right\}$$

satisfaz:

$$1 - \alpha = P(\boldsymbol{\mu} \in R(\underline{\mathbf{X}})) = P\left(n(\bar{\mathbf{x}} - \boldsymbol{\mu})' \mathbf{S}_c^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu}) \leq \frac{(n-1)p}{n-p} F_{p, n-p; 1-\alpha}\right)$$

$$= P\left(\max_{\mathbf{a} \neq \mathbf{0}} \frac{(\mathbf{a}'(\bar{\mathbf{x}} - \boldsymbol{\mu}))^2}{\mathbf{a}'\mathbf{S}_c\mathbf{a}} \leq \frac{(n-1)p}{n(n-p)} F_{p, n-p; 1-\alpha}\right)$$

$$= P\left(\frac{(\mathbf{a}'(\bar{\mathbf{x}} - \boldsymbol{\mu}))^2}{\mathbf{a}'\mathbf{S}_c\mathbf{a}} \leq \frac{(n-1)p}{n(n-p)} F_{p, n-p; 1-\alpha}, \forall \mathbf{a} \neq \mathbf{0}\right)$$

Consequentemente, simultaneamente, para todos os $\mathbf{a} \in \mathbb{R}^p \setminus \{\mathbf{0}\}$, o intervalo

$$\left(\mathbf{a}'\bar{\mathbf{x}} - \sqrt{\frac{(n-1)p}{n(n-p)} F_{p, n-p; 1-\alpha} \mathbf{a}'\mathbf{S}_c \mathbf{a}}, \mathbf{a}'\bar{\mathbf{x}} + \sqrt{\frac{(n-1)p}{n(n-p)} F_{p, n-p; 1-\alpha} \mathbf{a}'\mathbf{S}_c \mathbf{a}} \right)$$

contem $\mathbf{a}'\boldsymbol{\mu}$ com probabilidade $1 - \alpha$.

- ▶ Os intervalos de confiança para combinações da média $\mathbf{a}'\boldsymbol{\mu}$ são conhecidas por **intervalos de confiança de Roy**, ou **intervalos de T^2** visto que a sua probabilidade de cobertura é determinada pela distribuição da estatística T^2 de Hotelling.
- ▶ Escolhas diferentes para \mathbf{a} determinam diferentes combinações lineares das componentes do vetor média. Por exemplo, se $\mathbf{a} = [1 \ 0 \ 0 \ \cdots \ 0]$ resulta um intervalo de confiança para μ_1 .

Intervalos de Roy para as componentes $\mu_1, \mu_2, \dots, \mu_p$

Para uma distribuição normal p dimensional, os intervalos de Roy com uma confiança $(1 - \alpha)100\%$ para as p médias marginais são dados pelos seguintes p intervalos reais:

$$\left(\bar{x}_i - \sqrt{\frac{(n-1)p}{n(n-p)} F_{p, n-p; 1-\alpha} S_{ii}} , \bar{x}_i + \sqrt{\frac{(n-1)p}{n(n-p)} F_{p, n-p; 1-\alpha} S_{ii}} \right)$$

onde $i = 1, 2, \dots, p$.

Exercício...

Comparações simultâneas das componentes da média. Intervalos segundo Bonferroni

Muitas vezes não pretendemos avaliar todas as possíveis combinações das componentes da média mas apenas um número restrito de combinações. Se um número pequeno m de combinações lineares são de interesse, então intervalos simultâneos para essas m combinações podem ser desenvolvidos os quais são de menor amplitude (i.e., mais preciso) do que os obtidos pelos intervalos de confiança de Roy.

Seja $\mathbf{a}'_i \boldsymbol{\mu}$, $i = 1, 2, \dots, m$, as m apenas combinações lineares de interesse. Um intervalo para $\mathbf{a}'_i \boldsymbol{\mu}$ a $(1 - \alpha_i)100\%$ de confiança é dado por:

$$C_n(i) = \left(\mathbf{a}'_i \bar{\mathbf{x}} - t_{n-1; 1-\alpha_i/2} \sqrt{\frac{\mathbf{a}'_i \mathbf{S}_c \mathbf{a}_i}{n}}, \mathbf{a}'_i \bar{\mathbf{x}} + t_{n-1; 1-\alpha_i/2} \sqrt{\frac{\mathbf{a}'_i \mathbf{S}_c \mathbf{a}_i}{n}} \right)$$

sendo que $P(\mathbf{a}'_i \boldsymbol{\mu} \in C_n(i)) = 1 - \alpha_i$. Assim,

$$\begin{aligned} P(\mathbf{a}'_i \boldsymbol{\mu} \in C_n(i), \forall i = 1, 2, \dots, m) &= 1 - P(\exists i = 1, 2, \dots, m : \mathbf{a}'_i \boldsymbol{\mu} \notin C_n(i)) \\ &\geq 1 - \sum_{i=1}^m P(\mathbf{a}'_i \boldsymbol{\mu} \notin C_n(i)) = 1 - \sum_{i=1}^m \left(1 - \underbrace{P(\mathbf{a}'_i \boldsymbol{\mu} \in C_n(i))}_{1-\alpha_i} \right) \\ &= 1 - \sum_{i=1}^m \alpha_i = 1 - \alpha \quad , \text{ para } \alpha_i = \alpha/m \end{aligned}$$

Esta escolha de $\alpha_i = \alpha/m$ é designada por **método de Bonferroni** para m comparações múltiplas.

Consequentemente, os intervalos de confiança para m combinações lineares $\mathbf{a}_i'\boldsymbol{\mu}$, $i = 1, 2, \dots, m$, a um grau de confiança global de $(1 - \alpha)100\%$ são dado por:

$$\left(\mathbf{a}_i'\bar{\mathbf{x}} - t_{n-1;1-\alpha/(2m)}\sqrt{\frac{\mathbf{a}_i'\mathbf{S}_c\mathbf{a}_i}{n}}, \mathbf{a}_i'\bar{\mathbf{x}} + t_{n-1;1-\alpha/(2m)}\sqrt{\frac{\mathbf{a}_i'\mathbf{S}_c\mathbf{a}_i}{n}} \right)$$

Em particular, quando $\mathbf{a}_i = [0 \ 0 \cdots 1 \cdots 0]$ temos que os m intervalos simultâneos para as componentes médias, a uma confiança global de $(1 - \alpha)100\%$, são dado por:

$$\left(\bar{x}_i - t_{n-1;1-\alpha/(2m)}\sqrt{\frac{S_{ii}}{n}}, \bar{x}_i + t_{n-1;1-\alpha/(2m)}\sqrt{\frac{S_{ii}}{n}} \right)$$

Métodos para verificar o pressuposto da normalidade de dados univariados

- ▶ Cálculo de estatísticas sumárias: média e mediana (devem estar próximas), coeficiente de assimetria (deve ser próximo de zero).
- ▶ Técnicas gráficas:
 - ▶ histograma, Caixa de bigodes (*boxplot*): verificar a existência de simetria da distribuição empírica, a não existência de observações atípicas (*outliers*; a existir estas deverão ser confirmadas)
 - ▶ *qq-plot* da normal: verificar o não afastamento dos pontos a uma reta.
- ▶ Testes de ajustamentos: teste de Shapiro-Wilks, teste de Lilliefors, teste de Anderson-Darling, ...

Métodos para verificar o pressuposto da normalidade de dados multivariados

- ▶ Verificar a normalidade das distribuições marginais univariadas (seguindo os procedimentos anteriores).
- ▶ Verificar a normalidade das marginais bivariadas.
Os procedimentos a realizar tem por base o seguinte resultado:

▶ **Propriedade.** Se $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, com $|\boldsymbol{\Sigma}| > 0$, então

$$(\mathbf{X} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{X} - \boldsymbol{\mu}) \sim \chi_p^2,$$

onde χ_p^2 denota a distribuição de qui-quadrado com p graus de liberdade (g.l.).

- ▶ Na prática, em geral, fica-se por analisar marginais de dimensão 2.

Verificar a normalidade das marginais bivariadas

Dada uma amostra de tamanho n , calculam-se as n distâncias quadradas:

$$d_j^2 = (\mathbf{x}_j - \bar{\mathbf{x}})' \mathbf{s}_c^{-1} (\mathbf{x}_j - \bar{\mathbf{x}}) , j = 1, 2, \dots, n$$

- **Procedimento 1.** Contabilizar o número de amostras que caem dentro da elipse.

$$d_j^2 \leq \chi_{p;a}^2 , j = 1, 2, \dots, n$$

onde $\chi_{p;a}^2$ é o quantil de ordem $a100\%$ de uma distribuição χ_p^2 .

Deve ser esperado que, por exemplo, para $a = 0.5$, 50% das observações da amostra conduzam a distâncias quadradas que tomem valores inferiores ao percentil 50 da distribuição de qui-quadrado com p g.l. Se tal não ocorrer o pressuposto da normalidade pode ser colocado em causa.

Verificar a normalidade das marginais bivariadas

Dada uma amostra de tamanho n , calculam-se as n distâncias quadradas:

$$d_j^2 = (\mathbf{x}_j - \bar{\mathbf{x}})' \mathbf{s}_c^{-1} (\mathbf{x}_j - \bar{\mathbf{x}}) , j = 1, 2, \dots, n$$

- **Procedimento 2.** Construir um gráfico do qui-quadrado (χ^2 -plot) para verificar se as distâncias d_j^2 provem de uma distribuição χ_p^2 .

Fazer um *chi-plot* corresponde a representar as distâncias d_j^2 num *qq-plot* da distribuição χ_p^2 . Para construir um *chi-plot*

1. ordenar as distâncias por ordem crescente : $d_{(1)}^2 \leq d_{(2)}^2 \leq \dots \leq d_{(n)}^2$;
2. representar os pontos $\left(d_{(j)}^2, \chi_{p; (j - \frac{1}{2})/n}^2 \right)$.

Deve ser esperado que os pontos do gráfico se ajustem a uma reta. Se se afastam da reta, indicia que as observações correspondentes se afastam do padrão da normalidade.

Exercício

Num standard de carros usados, foram registados a idade (em anos), x_1 , e o preço de venda (em milhares de euros), x_2 , de 10 carros. Os dados obtidos estão reproduzidos a seguir:

x_1	3	5	5	7	7	7	8	9	10	11
x_2	2.3	1.9	1	0.7	0.3	1	1.05	0.45	0.7	0.3

- ▶ Calcule as distâncias quadradas $d_j^2 = (\mathbf{x}_j - \bar{\mathbf{x}})' \mathbf{s}_c^{-1} (\mathbf{x}_j - \bar{\mathbf{x}})$, $j = 1, 2, \dots, 10$.
- ▶ Determine a proporção de observações que caem dentro do contorno de uma distribuição normal bivariada com probabilidade estimada de 50%.
- ▶ Construa um χ^2 -plot e comente-o.
- ▶ Com base nos resultados obtidos, os dados são aproximadamente normal bivariada?

Capítulo 3: Análise discriminante e de classificação

#.1. Introdução

#.2. Regra discriminante de Máxima Verosimilhança

#.3. Análise discriminante de Fisher

Bibliografia de base

- ▶ Johnson, R.A. e Wichern, D.W. (1982). *Applied multivariate statistical analysis*. 3 Edição. Prentice-Hall.

Introdução

- ▶ Técnicas de Análise Discriminante (AD): conjunto de métodos estatísticos para dados multivariados que, partindo do conhecimento da população, entre várias, a que pertencem n indivíduos de uma amostra, pretende descrever, graficamente ou analiticamente, traços que diferenciam os indivíduos entre essas populações.
- ▶ Dada uma matriz de dados ($n \times p$), onde está identificada a classe/população (variável categórica X_{p+1}) a que pertence cada indivíduo, pretende-se encontrar funções das p variáveis (*funções discriminantes*) que melhor distingue ou separa os conjuntos de indivíduos pelas subpopulações existentes.
- ▶ As funções discriminantes podem ser usadas para definir *Regras discriminantes* (RD) com vista a classificar futuros indivíduos (observados nas p variáveis). Ex: $2x+3$ é uma função discriminante; *Se $2x+3 > 0$, o indivíduo pertence à população Π_1 é uma RD.)*

Ilustrar graficamente com $p = 2$.

Introdução

Dada r populações: Π_1, \dots, Π_r , uma **regra discriminante** é uma separação do espaço amostral em subconjuntos $\mathcal{R}_1, \dots, \mathcal{R}_r$ tal que:

se $x_i \in \mathcal{R}_j$ então *Individuo* $_i \in$ População Π_j

Objetivo da AD:

- ▶ Encontrar as funções discriminantes que definem as "melhores" regiões $\mathcal{R}_1, \dots, \mathcal{R}_r$.

A noção de "melhor" depende do critério de optimalidade considerado. Existem diversos critérios para derivar regras de classificação "óptimas". Tais critérios dependem se a RD se baseia em modelos probabilísticos (ex: **RD de Máxima Verosimilhança**) ou é construída em contexto descritivo (ex: **RD de Fisher**).

RD de Máxima Verosimilhança (MV)

É estabelecida se as distribuições das populações são conhecidas a menos de parâmetros.

Denotemos por f_j a função densidade de probabilidade da população Π_j . Então, a **RD de MV** é aquela que determina a alocação de uma observação na população sobre a qual o valor da verosimilhança é máxima; ou seja,

$$\mathbf{x} \in \Pi_j \text{ se } f_j(\mathbf{x}) = \max_{i=1,2,\dots,r} f_i(\mathbf{x}),$$

e, nesse caso, as regiões \mathcal{R}_j , $j = 1, 2, \dots, r$ são definidas por:

$$\mathcal{R}_j = \{\mathbf{x} : f_j(\mathbf{x}) > f_i(\mathbf{x}), i = 1, 2, \dots, r, i \neq j\}$$

Se existem várias populações dando o mesmo valor de máximo, então qualquer uma delas pode ser selecionada.

RD de MV - caso $r = 2$ -

$$\mathbf{x} \in \Pi_1 \text{ se } f_1(\mathbf{x}) > f_2(\mathbf{x})$$

Nesse caso, as regiões são definidas por:

$$\mathcal{R}_1 = \{\mathbf{x} : f_1(\mathbf{x}) > f_2(\mathbf{x})\} \text{ e } \mathcal{R}_2 = \text{complementar de } \mathcal{R}_1$$

Exemplo ($p = 1$): Considere o espaço amostral $\{0, 1\}$ e as populações definidas por:

$$\Pi_1 : \begin{array}{c|cc} x & 0 & 1 \\ \hline P(X=x) & 1/2 & 1/2 \end{array} \quad \Pi_2 : \begin{array}{c|cc} x & 0 & 1 \\ \hline P(X=x) & 1/4 & 3/4 \end{array}$$

RD de MV:

$$x \in \Pi_1 \text{ se } P_{\Pi_1}(X = x) > P_{\Pi_2}(X = x),$$

sendo

$$\mathcal{R}_1 = \{0\} \text{ e } \mathcal{R}_2 = \{1\}$$

Probabilidades de erros de má classificação

Quando classificamos uma nova observação \mathbf{x} numa população (Π_1 ou Π_2) podemos cometer erros de má classificação:

- ▶ Probabilidade de classificar um objecto \mathbf{x} dentro da região \mathcal{R}_2 quando ele é da população Π_1 :

$$p_{2|1} = P(X \in \mathcal{R}_2 | \Pi_1) = \int_{\mathcal{R}_2} f_1(\mathbf{x}) d\mathbf{x}$$

- ▶ Probabilidade de classificar um objecto \mathbf{x} dentro da região \mathcal{R}_1 quando ele é da população Π_2 :

$$p_{1|2} = P(X \in \mathcal{R}_1 | \Pi_2) = \int_{\mathcal{R}_1} f_2(\mathbf{x}) d\mathbf{x}$$

Custos de má classificação

As observações mal classificadas produzem um custo $C(i|j)$ quando uma observação de Π_j é assinalada para \mathcal{R}_i . A estrutura do custo pode ser esquematizado numa matriz de custos:

		Classificação	
		Π_1	Π_2
População Verdadeira	Π_1	0	$C_{2 1}$
	Π_2	$C_{1 2}$	0

RD com CEMC mínimo

O Custo Esperado de má classificação (CEMC) é dado por

$$CEMC = C_{2|1} p_{2|1} \pi_1 + C_{1|2} p_{1|2} \pi_2,$$

onde π_j representa a probabilidade à priori que um indivíduo seleccionado aleatoriamente pertença à população Π_j .

Propriedade: Para $r = 2$ populações, a RD que minimiza o CEMC é dada por:

$$\mathcal{R}_1 = \left\{ x : \frac{f_1(x)}{f_2(x)} \geq \frac{C_{1|2} \pi_2}{C_{2|1} \pi_1} \right\} \quad \text{e} \quad \mathcal{R}_2 = \left\{ x : \frac{f_1(x)}{f_2(x)} < \frac{C_{1|2} \pi_2}{C_{2|1} \pi_1} \right\}$$

$\frac{C_{1|2}}{C_{2|1}}$ é a razão dos custos e $\frac{\pi_2}{\pi_1}$ a razão das probabilidades à priori.

Na prática a RD é construída caso a caso de modo a minimizar o CEMC.

Fazer exerc. 3.2.

RD com CEMC mínimo -Exemplo em normais multivariadas com matrizes de covariâncias todas iguais-

Seja $\Pi_1 : X_1 \sim N_p(\mu_1, \Sigma)$ e $\Pi_2 : X_2 \sim N_p(\mu_2, \Sigma)$. Como a função de verosimilhança da população Π_i , $i = 1, 2$ é dada por:

$$f_i(\mathbf{x}) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} \exp \left(-\frac{1}{2} (\mathbf{x} - \mu_i)' \Sigma^{-1} (\mathbf{x} - \mu_i) \right),$$

a RD de MV que minimiza o CEMC traduz-se por:

$$\mathcal{R}_1 = \left\{ \mathbf{x} \in R^p : -\frac{1}{2} (\mathbf{x} - \mu_1)' \Sigma^{-1} (\mathbf{x} - \mu_1) + \frac{1}{2} (\mathbf{x} - \mu_2)' \Sigma^{-1} (\mathbf{x} - \mu_2) \geq \ln k \right\}$$

$$\text{com } k = \frac{C_{1|2} \pi_2}{C_{2|1} \pi_1}.$$

RD com CEMC mínimo -Exemplo em normais multivariadas com matrizes de covariâncias todas iguais-

Em particular, quando os CEMC e as probabilidades nas duas populações, Π_1 e Π_2 , são iguais, tem-se $k = 1$ pelo que a RD de MV simplifica-se para:

$$\mathcal{R}_1 = \{ \mathbf{x} \in R^p : (\mathbf{x} - \boldsymbol{\mu}_1)' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_1) \leq (\mathbf{x} - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_2) \}$$

Equivalentemente, a futura observação \mathbf{x} deve alocar-se na população que exibe menor distância de Mahalanobis.

A região discriminante pode ainda expressar-se na forma:

$$\mathcal{R}_1 = \left\{ \mathbf{x} \in R^p : (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} \mathbf{x} \geq (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} \frac{\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2}{2} \right\}$$

(Demonstrar)

RDMV –na prática, para populações normais multivariadas com matrizes de covariâncias todas iguais

Na regra anterior, caso se desconheçam as médias μ_i e matriz de covariâncias Σ populacional comum, estas são substituídas pelas seguintes estimativas calculadas à partir da matriz de dados

- ▶ médias amostrais das duas populações: $\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2$
- ▶ matriz de covariâncias amostral combinada: $\mathbf{S}_p = \frac{(n_1-1)\mathbf{S}_{c1} + (n_2-1)\mathbf{S}_{c2}}{n_1 + n_2 - 2}$

Assim, na prática, define-se:

$$\mathcal{R}_1 = \left\{ \mathbf{x} \in R^p : (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{S}_p^{-1} \mathbf{x} \geq (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{S}_p^{-1} \frac{(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)}{2} \right\}$$

Observe-se que esta RD é definida em termos de:

$$\mathbf{a}' \mathbf{x} \geq \mathbf{a}' \frac{(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)}{2} \quad \text{com } \mathbf{a}' = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{S}_p^{-1}$$

a qual será obtida na AD em contexto descritivo!

Análise discriminante de Fisher: Objetivo

Dados n indivíduos, extraídos de duas ($r=2$) populações distintas e descritos por p v.a.'s, pretende-se

- ▶ em termos analíticos, obter a combinação linear $\mathbf{a}'\mathbf{X}$
- ▶ em termos geométrico, encontrar a projeção $\mathbf{a}'\mathbf{X}$ (chamada de **eixo discriminante ou variável canónica**), que "melhor" separe as duas populações;

A qualidade da separação é definida por algum *critério de separabilidade*.

Critério de separabilidade da AD de Fisher: Definir o eixo discriminante $\mathbf{a}'\mathbf{X}$ de modo que os valores $Y = \mathbf{a}'\mathbf{X}$ dos indivíduos nesse eixo sejam mais homogêneos possíveis dentro da mesma população e claramente distintos entre populações diferentes.

(Ilustrar)

Eixo discriminante de Fisher, $y = \mathbf{a}'\mathbf{x}$

- A separação das duas populações é definida em termos da diferença entre as médias das observações y_{i1} , $i = 1, \dots, n_1$, e y_{j2} , $j = 1, \dots, n_2$, medidas em unidades do desvio padrão amostral (o qual é assumido que é comum nas duas populações), ou seja, a separação das duas populações é definida em termos de

$$\frac{|\bar{y}_1 - \bar{y}_2|}{s_y} \quad (1)$$

Da Teoria das matrizes sabe-se que $\max_{\mathbf{a} \neq 0} \frac{(\mathbf{a}'\mathbf{d})^2}{\mathbf{a}'\mathbf{S}\mathbf{a}} = \mathbf{d}'\mathbf{S}^{-1}\mathbf{d}$, sendo esse valor máximo $\mathbf{d}'\mathbf{S}^{-1}\mathbf{d}$ atingido quando $\mathbf{a}' = \mathbf{d}'\mathbf{S}^{-1}$.

Fazendo, em (1), $y = \mathbf{a}'\mathbf{x}$, tem-se:

$$\begin{aligned} \frac{(\bar{y}_1 - \bar{y}_2)^2}{s_y^2} &= \frac{(\mathbf{a}'\bar{\mathbf{x}}_1 - \mathbf{a}'\bar{\mathbf{x}}_2)^2}{\widehat{Var}(\mathbf{a}'\mathbf{X})} = \frac{(\mathbf{a}'(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2))^2}{\mathbf{a}'\widehat{Var}(\mathbf{X})\mathbf{a}} = \frac{(\mathbf{a}'(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2))^2}{\mathbf{a}'\mathbf{S}\mathbf{a}} \\ &= (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \quad \text{com } \mathbf{a} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}^{-1} \end{aligned}$$

Função discriminante de Fisher

Definição. A função discriminante linear de Fisher é dada por:

$$f(\mathbf{x}) = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{s}_p^{-1} \mathbf{x}$$

a qual separa as duas populações Π_1 e Π_2 o máximo possível, sendo essa separação máxima igual a

$$(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{s}_p^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$$

onde \mathbf{s}_p é uma estimativa da matriz de covariâncias comum às populações Π_1 e Π_2 , dada pela matriz de covariâncias combinada:

$$\mathbf{s}_p = \frac{(n_1 - 1)\mathbf{s}_{c_1} + (n_2 - 1)\mathbf{s}_{c_2}}{n_1 + n_2 - 2}$$

RD de Fisher

A RD de Fisher é dada por: $\mathcal{R}_1 = \left\{ y \in R : y \geq \frac{\bar{y}_1 + \bar{y}_2}{2} \right\}$ onde as populações Π_1 e Π_2 são tais que $\bar{y}_1 > \bar{y}_2$.

Tendo em conta que $y = \mathbf{a}'\mathbf{x}$, resulta que:

Definição. A RD linear de Fisher é dada por:

$$\mathcal{R}_1 = \left\{ \mathbf{x} \in R^p : (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{s}_p^{-1} \mathbf{x} \geq (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{s}_p^{-1} \frac{(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)}{2} \right\},$$

$$\mathcal{R}_2 = R^p \setminus \mathcal{R}_1$$

a qual separa as duas populações Π_1 e Π_2 o máximo possível, sendo essa separação máxima (entre médias normalizadas pelo desvio padrão) igual a

$$\sqrt{(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{s}_p^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)}$$

Fazer exerc. 3.4.