Análise Multivariada

Lupércio França Bessegato
Dep. Estatística/UFJF

Roteiro

Introdução

Análise Multivariada

- Considera várias variáveis relacionadas simultaneamente.
- Variáveis de interesse não independentes uma das outras.
- Associação entre conjuntos de medidas.

Estatística Multivariada

- Métodos estatísticos utilizados em situações nas quais as variáveis são medidas simultaneamente, em cada elemento amostral
  - √ Em geral, independem do conhecimento da forma matemática da distribuição subjacente

5

Objetos

- Entidades das quais são tomadas medidas
  - √ Itens, pessoas, organizações, etc.
  - √ São portadores de medidas
  - √ São medidos somente com respeito a certas variáveis de interesse

6

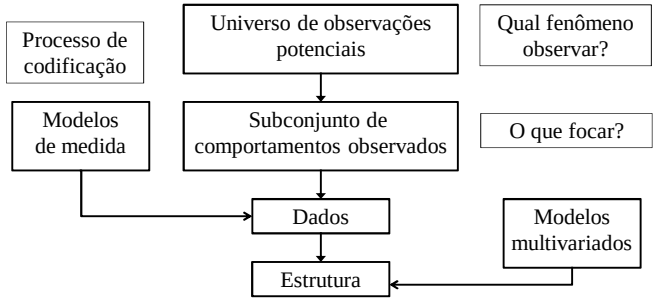
Variáveis

- Características ou propriedades
  - √ São os aspectos dos objetos que são medidos

7

Observação e Dados

- Dados:
  - √ Observações documentadas ou resultados de medição



8

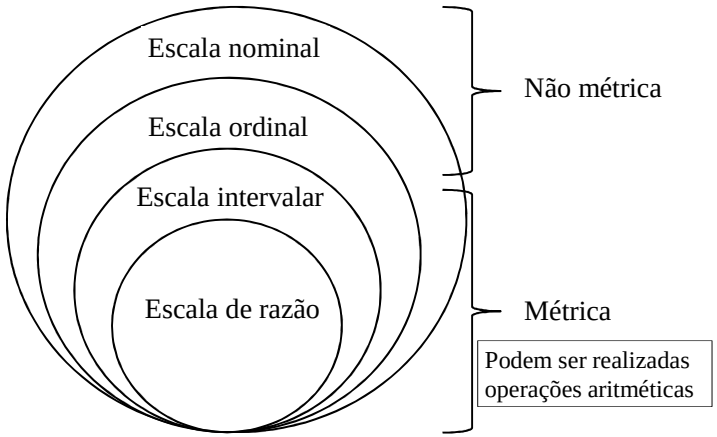
- Teoria subjacente sobre a área de interesse:
  - √ Necessária em cada etapa do processo de ir das observações aos dados
  - √ A interação do pesquisador (orientado pela teoria) com o ambiente de observações potenciais é que leva aos dados
- Última etapa:
  - √ Buscar estrutura associativa nos dados adequando os modelos multivariados

9

Medida

- Processo pelo qual atribuem-se aos números (ou, algumas vezes, outros símbolos) características ou propriedades de objetos, de acordo com um procedimento predeterminado
- Escala de medida:
  - √ Refere-se à quantidade de informação que está contida na medida e o que ela nos informa sobre a relação entre dois objetos

10



18

Organização de Dados Multivariados

$$\mathbf{X}_{n\times p} = \begin{array}{c} \\ 1 \\ 2 \\ \vdots \\ n \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & \cdots & p \end{array} \\ \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \end{array}$$

p: número de variáveis
n: número de objetos

- √ Matriz:
  - – Organização em arranjo estrutural de dados multivariados (métricos ou não métricos)
- √ Colunas: variáveis
  (características medidas de objetos)
- √ Linhas: objetos
  (lista de características medidas de um objeto)

20

Conjunto de Dados #1

- Bumpus (1898)
- Pardais sobreviventes de tempestade
  - √ Dados de 1/Fev/1898
  - √ Medidas morfológicas e peso de 49 pássaros fêmeas
  - √ 28 morreram, 21 não morreram
- Dados: *birds.csv* (*birds.txt*)

- √ Variáveis:
  - sv: 1= vivo, 2= morto.
  - ag: 1= adulto, 2 = jovem.
  - tl: comprimento total (bico à ponta da cauda), em mm.
  - ae: extensão alar, (ponta a ponta de asas), em mm.
  - wt: peso, em gramas.
  - bh: comprimento bico e cabeça, em mm.
  - hl: comprimento do úmero (osso braço), em polegadas.
  - fl: comprimento do fêmur (osso coxa), em polegadas.
  - tt: comprimento da tíbia-tarso (osso perna), em polegadas.
  - sk: amplitude do cranio, em polegadas.
  - kl: comprimento da quilha do esterno, em polegadas.

- Questões interessantes:
  1. Como as variáveis estão relacionadas?
     (Um valor grande de uma variável tende a ocorrer com valores grandes para as outras variáveis?)
  2. Os sobreviventes e os não sobreviventes têm diferenças estatisticamente significantes para seus valores médios das variáveis?
  3. Os sobreviventes e os não sobreviventes mostram quantidades similares de variação para a variável?
  4. Se os sobreviventes e não sobreviventes diferem em termos das distribuições das variáveis, então é possível construir alguma função dessas variáveis que separe os dois grupos?
     (Índice de ajuste Darwiano dos pardais: valores grandes da função tendem a ocorrer com os sobreviventes)

- Conclusão de Bumpus (1898):
  - √ Os sobreviventes são mais curtos e pesam menos, tem ossos das asas mais longos, pernas mais longas, esternos mais longos e maior capacidade cerebral
- O processo de eliminação seletiva é mais severo com indivíduos extremamente variáveis (independente da direção
  - √ É tão perigoso estar acima de um certo padrão de excelência orgânica como estar visivelmente abaixo desse padrão

Conjunto de Dados #2

- Thomson e Randall-Maciver (1905)
- Medidas em crânios masculinos da área de Tebas
  - √ 5 amostras de 30 crânios cada uma:
    - Período pré-dinástico primitivo (~ 4.000 aC)
    - Período pré-dinástico antigo (~3.300 aC)
    - 12ª e 13ª dinastias (~ 1.850 aC)
    - Período Ptolemaico (~ 200 aC)
    - Período romano (~150 dC)
  - √ Dados: *skulls*{*ade4*} ou *skulls2.csv* (*skulls.txt*)

25

- Variáveis:
  - √ MB (X_1): Largura máxima do crânio
  - √ BH (X_2): Altura do basibregmático do crânio
  - √ BL (X_3): Comprimento do basialveolar do crânio
  - √ NH (X_4): Altura nasal do crânio
  - √ Ano: Ano aproximado de formação do crânio

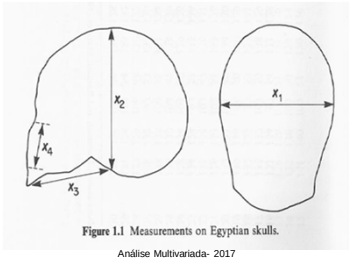


Figure 1.1 Measurements on Egyptian skulls.

26

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

27

- Conclusão de Thomson e Randall-Maciver (1905):
  - √ Há diferenças entre as cinco amostras que podem ser explicadas parcialmente como tendências no tempo.
  - √ As razões para as aparentes mudanças são desconhecidas
  - √ A migração de outras raças dentro da região pode ter sido o fator mais importante.

28

Conjunto de Dados #3

- McKechnie et al. (1975)
- Distribuição de uma borboleta:
  - √ 16 colônias de borboletas *Euphydryas editha* na Califórnia e Oregon
- Dados: *butterfly*{ade4}

29

- Variáveis ambientais (*butterfly$envir*):
  - √ Altitude, em pés
  - √ Precipitation: precipitação anual, em polegadas
  - √ Temp_max: temperatura máxima, em ºF
  - √ Temp_min: temperatura mínima, em ºF
- Variáveis genéticas: (*butterfly$genet*):
  - √ Frequências de mobilidade gênica *Fósforo Glucose-Isomerase* (*Pgi*) para colônias de borboletas
    - – Níveis: 0,4; 0,6; 0,8; 1; 1,16; 1,3

    (representa diferentes tipos genéticos de Pgi de modo que as frequências para uma colônia (somando a 100%) mostram as frequências do diferentes tipos para a *E. editha* naquele local)

30

- Dados agregados:
  - √ Variáveis ambientais e genéticas

```
> borboleta <- cbind(butterfly$envir, butterfly$genet)
> borboleta
   Altitude Precipitation Temp_Max Temp_Min 0.4 0.6 0.8  1 1.16 1.3
SS      500            43       98       17   0   3  22 57   17   1
SB      800            20       92       32   0  16  20 38   13  13
WSB     570            28       98       26   0   6  28 46   17   3
JRC     550            28       98       26   0   4  19 47   27   3
JRH     550            28       98       26   0   1   8 50   35   6
SJ      380            15       99       28   0   2  19 44   32   3
CR      930            21       99       28   0   0  15 50   27   8
UO      650            10      101       27  10  21  40 25    4   0
LO      600            10      101       27  14  26  32 28    0   0
DP     1500            19       99       23   0   1   6 80   12   1
PZ     1750            22      101       27   1   4  34 33   22   6
MC     2000            58      100       18   0   7  14 66   13   0
IF     2500            34      102       16   0   9  15 47   21   8
AF     2000            21      105       20   3   7  17 32   27  14
GH     7850            42       84        5   0   5   7 84    4   0
GL    10500            50       81      -12   0   3   1 92    4   0
```

31

- Rótulos das linhas: Local das colônias
  - √ SS (Oregon), SB, WSB, JRC, JRH, SJ, CR, UO, LO, DP, PZ, MC, IF, AF, GH, GL.

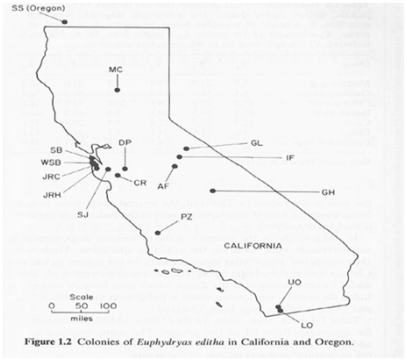


Figure 1.2 Colonies of *Euphydryas editha* in California and Oregon.

Fonte: Manly, 2005

32

- Questões interessantes:
  - √ As frequências Pgi são similares para as colônias espacialmente próximas?
  - √ O quanto, se algum, as frequências Pgi estão relacionadas às variáveis ambientais?

33

- Análise das respostas:
  - √ Se a composição genética das colônias foi fortemente determinada pelas migrações passadas e presentes
    - Frequências gênicas tenderão a ser similares para colônias que estão localizadas nas proximidades
  - √ Se o meio ambiente é mais importante
    - Deve aparecer relacionamento entre as frequências gênicas e as variáveis ambientais
    - Colônias próximas têm frequências gênicas similares somente se têm ambientes similares)
    - Pode ser difícil chegar a essa conclusão

34

Conjunto de Dados #4

- Coleção de ossos caninos do período em torno de 3.500 aC até o presente
  - √ Local: Tailândia
- A origem dos cães pré-históricos é incerta
  - √ Podem descender dos chacais dourados (*canis aureus*) ou do lobo (não é nativo da Tailândia)
- Fontes de origem mais próximas:
  - √ Parte ocidental da China: *canis lupus chanco*
  - √ Subcontinente indiano: *canis lupus pallides*
- Dados: *canine.txt*

35

- Higham et al. (1980)
- Objetivo do estudo:
  - √ Comparar medidas da mandíbula dos espécimes pré-históricos disponíveis com as mesmas medidas no chacal dourado, lobo chinês, lobo indiano, dingo (originário da Índia), cuon: *cuon alpinus* (originário do Sudeste da Ásia) e cães modernos da Tailândia.

36

- Variáveis:
  - √ X_1 : comprimento da mandíbula, em mm.
  - √ X_2: largura da mandíbula, em mm.
  - √ X_3: largura do côndilo da mandíbula, em mm.
  - √ X_4: altura da mandíbula abaixo do 1º molar, em mm.
  - √ X_5: comprimento do 1º molar, em mm.
  - √ X_6: largura do 1º molar, em mm.
  - √ X_7: comprimento do 1º ao 3º molar, em mm.
  - √ X_8: comprimento do 1º ao 4º molar, em mm.
  - √ X_9: largura do maxilar inferior, em mm.

37
Análise Multivariada- 2017

- Questões interessantes:
  1. O que as medidas sugerem sobre o relacionamento entre os grupos?
  2. Como os cães pré-históricos parecem se relacionar com os outros grupos?

38
Análise Multivariada- 2017

Conjunto de Dados #5

- Porcentagem da força de trabalho em nove diferentes tipos de indústrias para 30 países europeus
  - √ De 1980 a 1995
- Dados:

39
Análise Multivariada- 2017

- Objetivo do estudo:
  - √ Isolar grupos de países com padrões similares
  - √ Auxiliar o entendimento dos relacionamentos entre países
- Cada linha do conjunto de dados soma 100%

40
Análise Multivariada- 2017

- Variáveis:
  - √ Group: Eastern= Leste europeu; EFTA: área européia de livre comércio; EU: União Européia; Other: outros países
  - √ AGR: agricultura, floresta e pesca
  - √ MIN: mineração e exploração de pedreiras
  - √ MAN: fabricação
  - √ PS: fornecimento de energia elétrica
  - √ CON: construção
  - √ SER: serviços
  - √ FIN: finanças
  - √ SPS: serviços sociais e pessoais
  - √ TC: transportes e comunicações

41

- Questões:
  1. É possível isolar grupos de países com padrões similares de emprego?
  2. Qual o relacionamento entre os países com relação aos níveis de emprego?
     - Podem ser de interesse diferenças entre países que estejam relacionadas a grupos políticos (União Europeia; área europeia de livre comércio; países do leste europeu e outros países)

42

Conjunto de Dados #6

- Estudo de poluição do ar em 41 cidades dos EUA
  - √ Ano: 1970
- Dados: *Usairpollution*{*MVA*}

43

- Variáveis:
  - √ SO2: conteúdo de dióxido de enxofre no ar, em $\mu g/m^3$.
  - √ temp: temperatura média anual (ºF)
  - √ manu: quantidade de empresas manufatureiras empregando pelo menos 20 empregados.
  - √ popul: população (censo 1970), em milhares.
  - √ wind: velocidade média anual de vento, em milhas/h
  - √ precip: precipitação média anual, em polegadas
  - √ predays: número médio anual de dias com precipitação

44

- Questões:
  - √ Como o nível de poluição, medido pela concentração de dióxido de enxolfre está relacionado com as outras seis variáveis?
  - √ Como reduzir a quantidade de variáveis que descrevem as cidades, simplificando os dados?
  - √ É possível encontrar variáveis latentes que descrevam as cidades?

Técnicas em Estatística Multivariada

- Técnicas Exploratórias:
  - √ Sintetização da estrutura de variabilidade dos dados
    - Análise de componentes principais, análise fatorial, análise de correlações canônicas, análise de agrupamentos, análise discriminante
  - √ Técnicas de Inferência Estatística:
    - Métodos de estimação de parâmetros, testes de hipóteses, análise de variância, análise de covariância, análise de regeressão multivariada

Visão dos Métodos Multivariados

- Análise de componentes principais:
  - √ Redução do número de variáveis a um número menor de índices (componentes principais)
    - Componentes principais: combinações lineares das variáveis originais
  - √ Pode acontecer de que duas ou mais componentes principais forneçam um bom resumo de todas as variáveis originais

- Exemplo – Pardais sobreviventes
  - √ Muito da variação nas medidas do corpo dos pardais (X_1 a X_5) está relacionada com:
    - Tamanho geral dos pássaros (I_1)

$$I_1 = k_1X_1 + k_2X_2 + k_3X_3 + k_4X_4 + k_5X_5$$

    - Contraste entre as três primeiras medidas e as duas últimas

$$I_2 = k^*_1X_1 + k^*_2X_2 + k^*_3X_3 - k^*_4X_4 - k^*_5X_5$$

  - √ Análise de componentes principais:
    - Maneira de simplificar os dados, reduzindo o número de variáveis

- Análise fatorial:
  - √ Estudar a variação em uma quantidade de variáveis originais, usando um número menor de fatores (variáveis índices)
  - √ Assume-se que cada variável original pode ser expressa como uma combinação linear dos fatores mais um termo residual (reflete quanto a variável é independente das outras variáveis)

- Modelo:

$$X_1 = a_{11}F_1 + a_{12}F_2 + \epsilon_1\,,$$
$$X_2 = a_{21}F_1 + a_{22}F_2 + \epsilon_2\,,$$
$$X_3 = a_{31}F_1 + a_{32}F_3 + \epsilon_3\,,$$
$$X_4 = a_{41}F_1 + a_{42}F_4 + \epsilon_4\,,$$
$$X_5 = a_{51}F_1 + a_{52}F_5 + \epsilon_5\,.$$

a_{ij}: constantes
F_1; F_2: fatores
ε_i: variação em X_i que é independente da variação nas outras variáveis

  - √ F_1: fator tamanho
    - Alguns pássaros tendem a ser grandes, alguns pássaros tendem a ser pequenos, em todas as medidas do corpo
  - √ F_2: mede algum aspecto da forma dos pássaros
    - Alguns coeficientes positivos e alguns negativos

- Análise discriminante:
  - √ Separação em diferentes grupos com base nas medidas disponíveis
  - √ Encontrar combinações lineares convenientes das variáveis originais para atingir o objetivo desejado
  - √ Exemplo:
    - Quão bem os pardais sobreviventes e não sobreviventes podem ser separados usando suas medidas do corpo?

- Análise de agrupamento:
  - √ Identificação de grupos de objetos similares
  - √ Exemplos:
    1. Encontrar similaridades entre cães pré-históricos tailandeses e outros animais
    2. Agrupar os países europeus em termos de suas similaridade no padrão de empregos.

- Escalonamento multidimensional:
  - √ Mapeamento, com distância, entre objetos, mostrando como eles estão relacionados
  - √ Visualização possível em até três dimensões
  - √ Alternativa para a análise de agrupamentos
  - √ Exemplos:
    1. Encontrar similaridades entre cães pré-históricos tailandeses e outros animais
    2. Agrupar os países europeus em termos de suas similaridade no padrão de empregos.

- Métodos de ordenação:
  - √ Produção de eixos nos quais um conjunto de objetos de interesse pode ser representado
  - √ Algumas técnicas de ordenação:
    - Análise de componentes principais
    - Escalonamento multidimensional

- Análise de correlação canônica:
  - √ As variáveis (não os objetos) são divididos em dois grupos e o interesse está centrado no relacionamento entre elas
  - √ Exemplo – Colônias de borboleta:
    - Encontrar relacionamento entre as variáveis ambientais e as variáveis genéticas.

- Análise de correspondência:
  - √ Dados sobre a abundância de cada uma das várias características para cada elemento de um conjunto de objetos
  - √ Exemplo – Ecologia:
    - Objetos de interesse: diferentes locais
    - Características: diferentes espécies
    - Dados: abundância de espécies em locais
    - Objetivo: tornar claro os relacionamentos entre locais (expressos por distribuições das espécies) e os relacionamentos entre as espécies (expressos por distribuições dos locais).

Usos das Técnicas Multivariadas

- Construção de índices:
  - √ Sintetizar em uma única variável a informação de todas as variáveis que foram medidas sobre o fenômeno
    - Análise de componentes principais
    - Análise fatorial
    - Análise de correlação canônica

57

- Classificação e discriminação:
  - √ Busca-se a divisão de conjunto de dados em grupos, de modo que os grupos tenham coesão interna e sejam heterogêneos entre sin
    - Análise de agrupamentos (segmentação de mercado)
    - Análise discriminante (classificação de crédito)

58

- Inferência estatística:
  - √ Comparação de grupos em relação às médias de variáveis medidas conjuntamente
  - √ Regressão multivariada:
    - Análise do efeito de fatores externos não controlados nas variáveis que são monitoras regularmente

59

Referências

Bibliografia Recomendada

- MANLY, B. J. F. *Métodos Estatísticos Multivariados: uma Introdução.* Bookman, 2008.
- JOHNSON, R. A.; WINCHERN, D. W. *Applied Multivariate Statistical Analysis*. Prentice Hall, 2007
- MINGOTI, D.C. *Análise de Dados através de Métodos de Estatística Multivariada*. Ed. UFMG, 2005.
- EVERITT, B.; HOTHORN, T. *An Introduction to Applied Multivariate Analysis with R*. Springer, 2011.

75

Análise Multivariada- 2017