Estatística Multivariada aplicada a dados biológicos

Jhessica Letícia Kirch
Universidade de São Paulo

Simpósio de Microbiologia Agrícola
11 de abril de 2023

OBJETIVOS

- O objetivo do minicurso é abordar os mais importantes métodos de análise multivariada utilizando o software R.

RESULTADOS ESPERADOS

- Ao final do minicurso o aluno deverá ser apto a perceber, de acordo com cada problema real, qual o método multivariado mais adequado para tratar o problema que lhe for apresentado.

CONTEÚDO PROPOSTO

- Representação de dados multivariados;
- Análise de variância multivariada;
- Análise de agrupamento;
- Análise de componentes principais;
- Análise de fatores;
- Escalonamento multidimensional.

MATERIAL A SER UTILIZADO

Todo o material, scripts e slides estão disponíveis em:

https://github.com/jhessicakirch/Multivariada

PRÉ-REQUISITOS

- Conhecimento prático de estatística elementar;
- Conhecimento no software R.

SOBRE O SOFTWARE R

- Software é ferramenta.

RELEMBRANDO

- Instalação do R e RStudio;
- Comandos básicos no R;

ANÁLISE MULTIVARIADA

- Consiste em um conjunto de métodos que podem ser utilizados quando diversas medidas são feitas em cada indivíduo/objeto (RENCHER, 2002).

- O ponto principal de uma análise multivariada é considerar várias variáveis relacionadas simultaneamente, sendo todas consideradas igualmente importantes, pelo menos inicialmente (MANLY, 2008).

ANÁLISE MULTIVARIADA

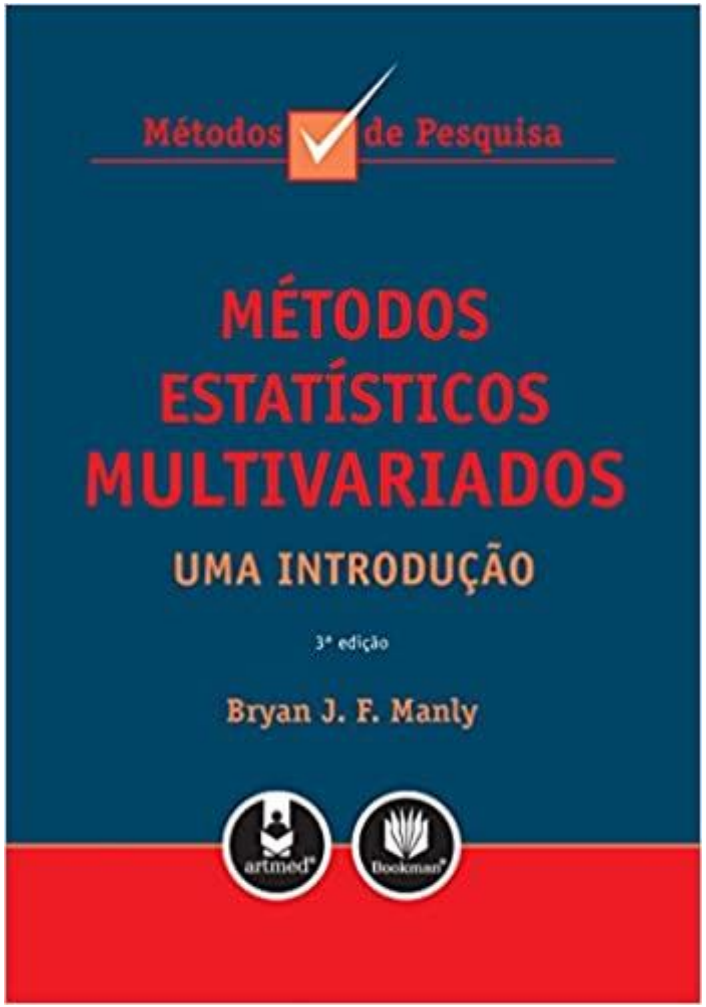
- Na prática, conjuntos de dados multivariados são comuns, embora nem sempre sejam analisados como tal.

- O uso exclusivo de procedimentos univariados com tais dados não é mais desculpável, dada a disponibilidade de técnicas multivariadas e poder computacional para realizá-las.

EXEMPLOS DE DADOS MULTIVARIADOS

- Alguns conjuntos de dados multivariados serão apresentados a seguir.

- Todos os exemplos são retirados do livro Métodos Estatísticos Multivariados do Bryan J. F. Manly (2008).



EXEMPLO 1

Pardais sobreviventes da tempestade

Descrição: Após uma forte tempestade em 1˚ de fevereiro de 1898, diversos pardais moribundos foram levados ao laboratório biológico de Bumpus na universidade de Brown.

Aproximadamente metade dos pássaros morreram e Bumpus viu isso como uma oportunidade de encontrar suporte para a teoria de seleção natural. Tomou 5 medidas morfológicas em cada pássaro. Os resultados são mostrados na Tabela 1, para fêmeas somente.

EXEMPLO 1

Tabela 1: Medidas do corpo de pardocas (em mm).

Pássaro	X_1	X_2	X_3	X_4	X_5
1	156	245	31,6	18,5	20,5
2	154	240	30,4	17,9	19,6
3	153	240	31,0	18,4	20,6
⋮	⋮	⋮	⋮	⋮	⋮
49	164	248	32,3	18,8	20,9

X_1 = comprimento total;

X_2 = extensão alar;

X_3 = comprimento do bico e cabeça;

X_4 = comprimento do úmero;

X_5 = comprimento da quilha do esterno.

Aves 1–21 sobreviveram e aves 22-49 morreram.

EXEMPLO 1

QUESTÕES À RESPONDER:

1. Como estão as várias variáveis relacionadas?

Por exemplo, um valor grande para uma das variáveis tende a ocorrer com valores grandes para as outras variáveis?

EXEMPLO 1

2. Os sobreviventes e não-sobreviventes têm diferenças estatisticamente significantes para seus valores médios das variáveis?

3. 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?

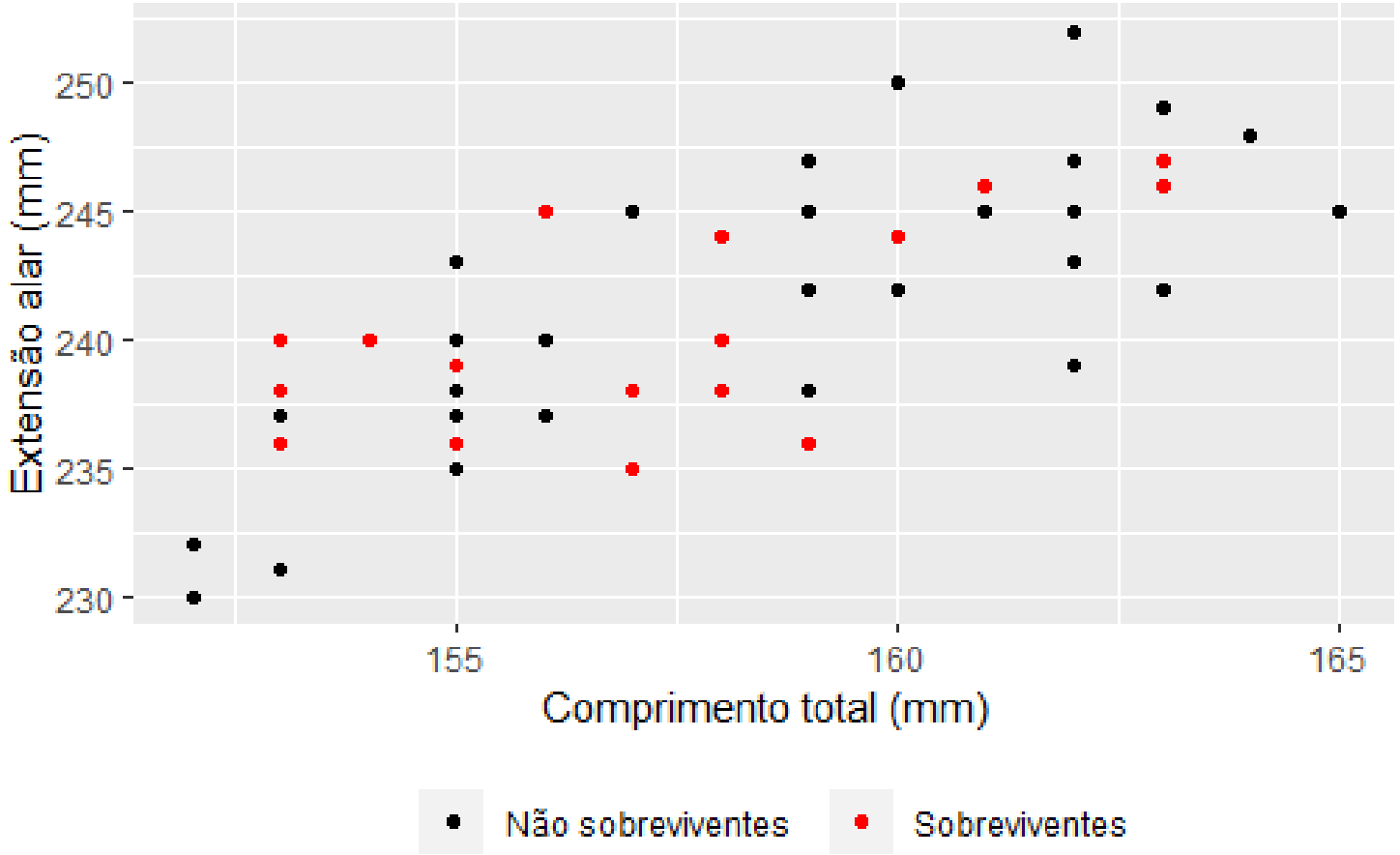
O problema da representação de muitas variáveis em duas dimensões

- Gráficos precisam ser apresentados em duas dimensões ou sobre papel ou na tela de um computador;

- Eixos horizontais e verticais representam variáveis.

REPRESENTAÇÃO GRÁFICA

Figura 1. Extensão alar representada contra o comprimento total para as 49 pardocas.



REPRESENTAÇÃO GRÁFICA

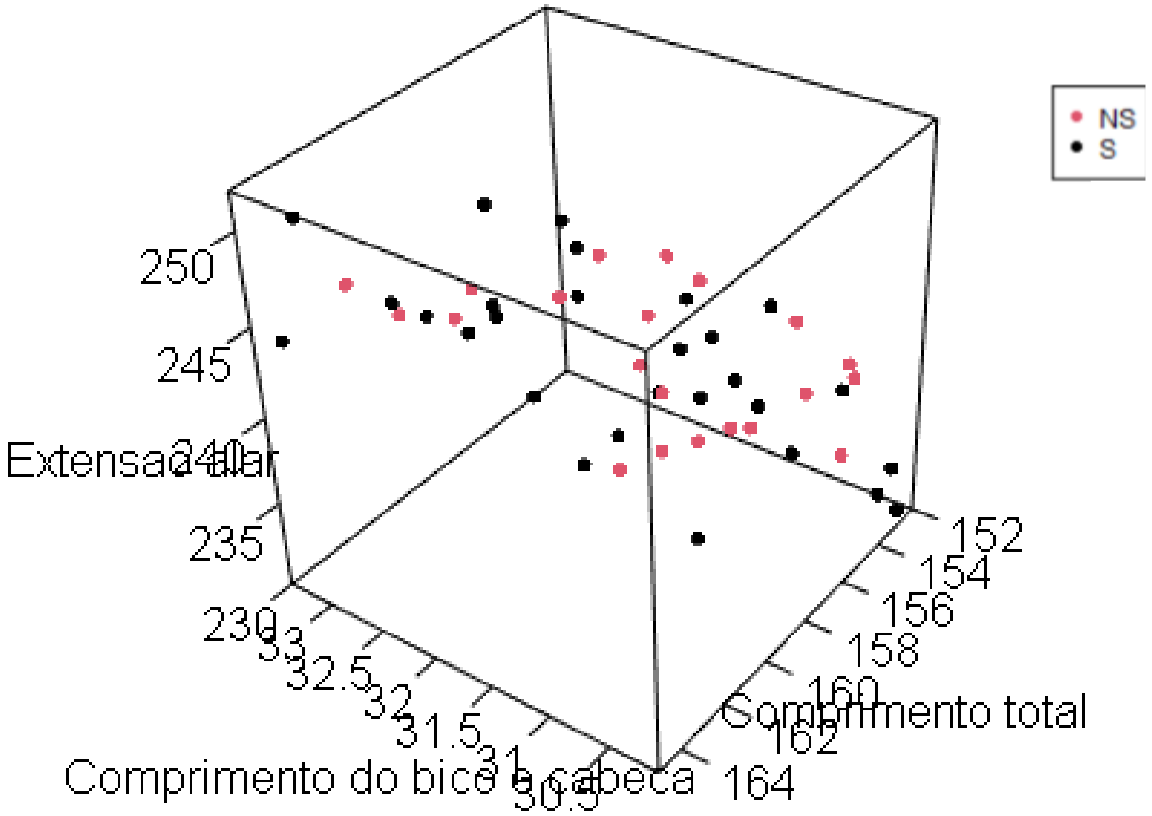
- Permite mostrar uma ou mais características dos objetos como sobreviventes e não-sobreviventes.

- Gráficos *tridimensionais* permitem mostrar três variáveis.

- Representação para mais de três variáveis são discutidos a seguir.

REPRESENTAÇÃO GRÁFICA

Figura 2. O comprimento do bico e da cabeça representados contra o comprimento total e a extensão alar (todos em mm) para as 49 pardocas.



REPRESENTANDO VARIÁVEIS ÍNDICES

- Variáveis índices são variáveis não observadas obtidas por combinação linear das variáveis originais.

- O principal objetivo de muitos métodos multivariados é a geração de variáveis índices.

- Com os valores dos CP1, CP2, CP3 pode-se representar graficamente as relações entre objetos.

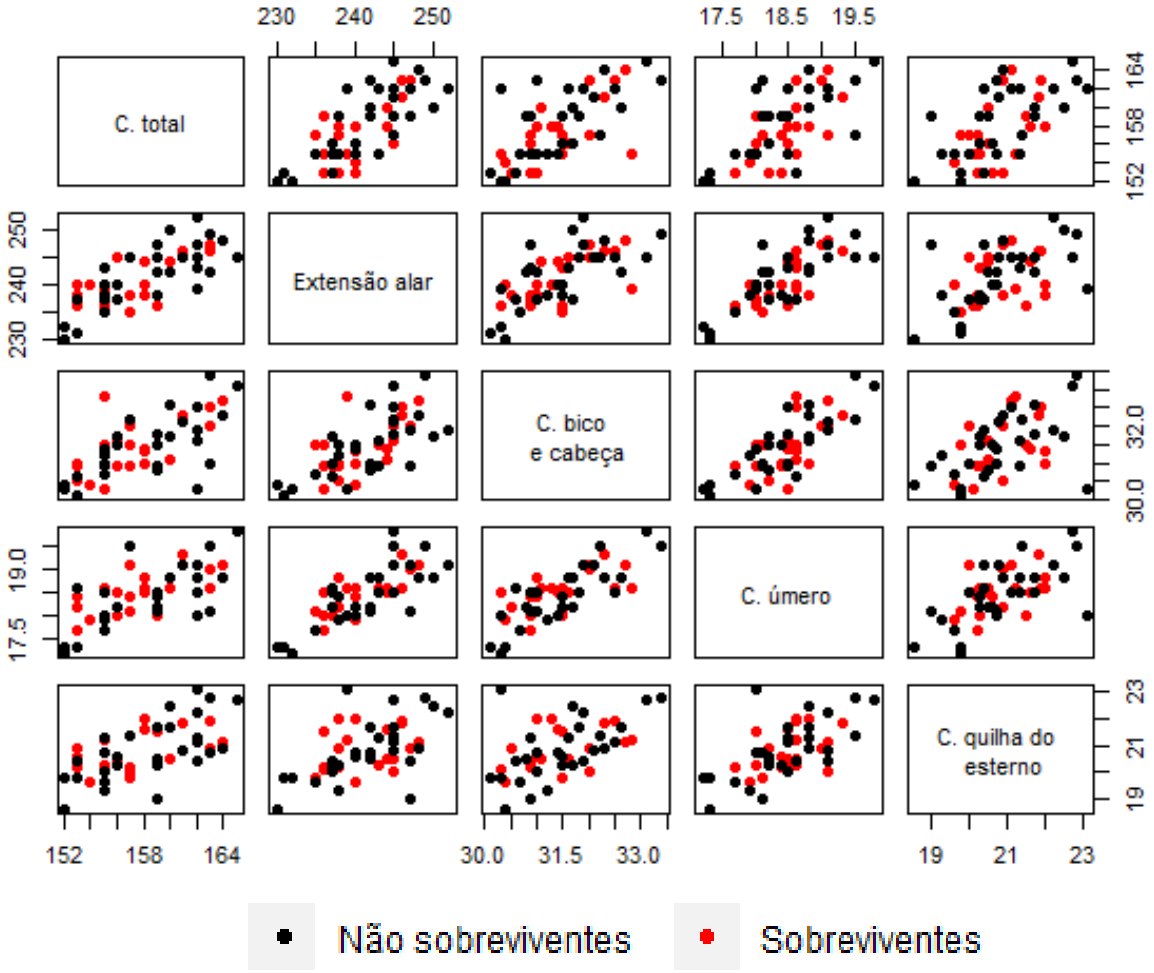
Obs.: A desvantagem de representar muitas variáveis para duas ou três dimensões é que alguma diferença-chave entre os objetos possa ser perdida na redução.

MATRIZ DE DISPERSÃO

- Consiste na representação simultânea de todos os pares de variáveis;
- **VANTAGEM:** são necessárias apenas representações bidimensionais;
- **DESVANTAGEM:** não mostram aspectos dos dados que somente seriam aparentes quando três ou mais variáveis são consideradas em conjunto.

MATRIZ DE DISPERSÃO

Figura 3. Matriz de dispersão do número de pássaros e cinco variáveis medidas (mm) em 49 pardocas.



MATRIZ DE DISPERSÃO

As variáveis são: comprimento total, extensão alar, comprimento do bico e cabeça, comprimento do úmero e da quilha do esterno, com uma variável adicional sendo a sobrevivência.

- Retas de regressão são incluídas nesses gráficos algumas vezes;

- Servem para mostrar existência entre quaisquer objetos com valores estranhos (DADOS DISCREPANTES).

PERFIS DE VARIÁVEIS

- É outra forma de representar objetos agora por linhas que mostram o perfil dos valores das variáveis.

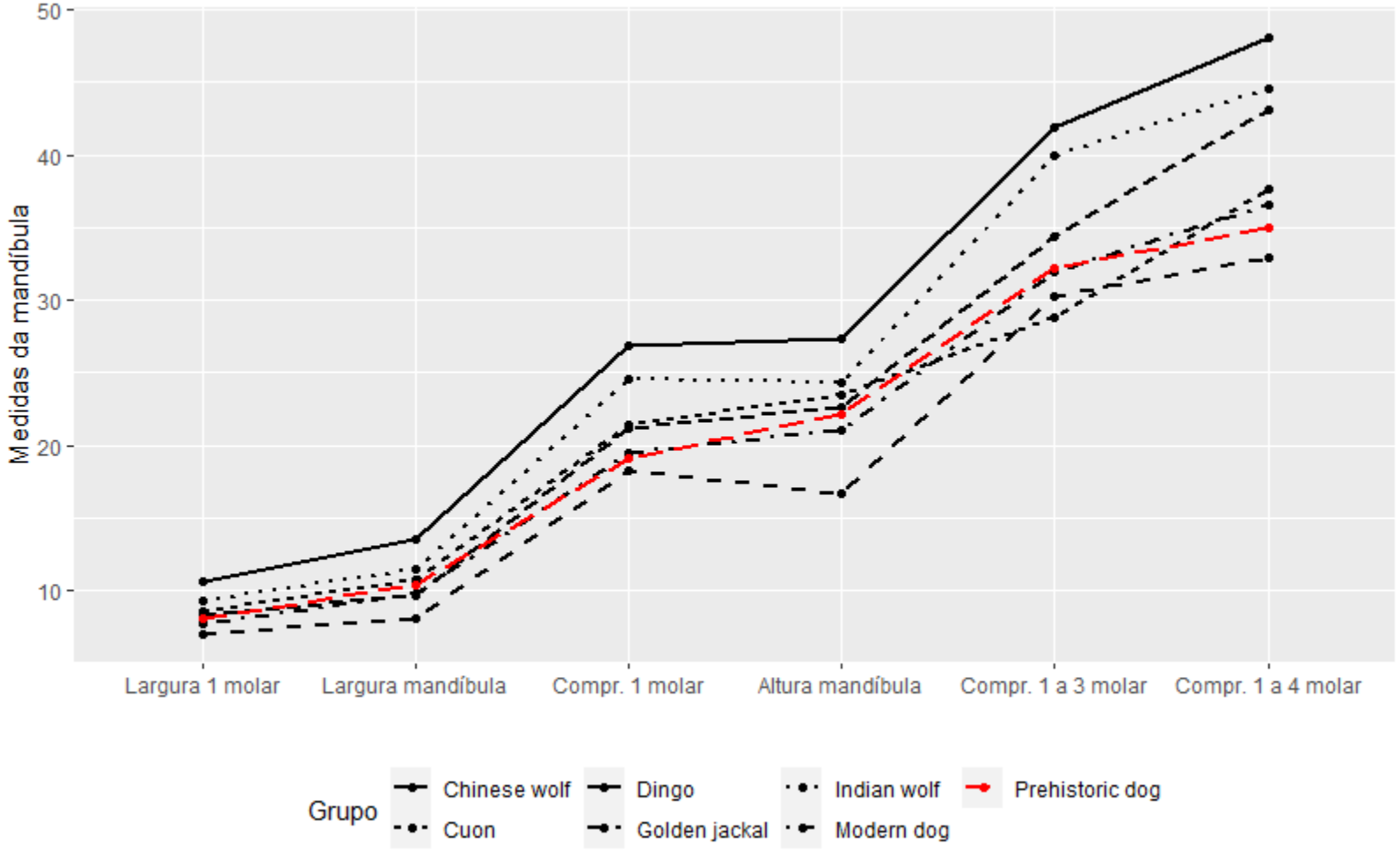
- **Exemplo:** Escavações de locais pré-históricos na Tailândia têm produzido uma coleção de ossos caninos, entretanto a origem dos cães pré-históricos não é certa.

Para tentar estabelecer os ancestrais dos cães pré-históricos, foram feitas medidas da mandíbula dos espécimes disponíveis. Foram acrescentados medições dos cães modernos da Tailândia, o dingo e o cuon.

A questão principal aqui é verificar o que as medidas sugerem sobre o relacionamento entre os grupos e, em particular, como os cães pré-históricos parecem se relacionar com os outros grupos.

PERFIS DE VARIÁVEIS

Figura 4. Perfis de variáveis em ordem crescente de valores médios para as medidas da mandíbula para sete grupos caninos



REFERÊNCIAS

MANLY, B. J. F. Métodos estatísticos multivariados: Uma introdução. Porto Alegre: Bookman, 2008.

RENCHER, A. C. Methods of multivariate analysis. Segunda edição. New York: John Wiley & Sons, 2002.

SILVA, A. R. Métodos de Análise Multivariada em R. Piracicaba: FEALQ, 2016.