Fröhlich *et al.* ULAKES J Med. v.3 n.2, p.111-121, **2023**

DOI: 10.56084/ulakesjmed.v3i2.865

Procedimento Tecnológico

Metodologia científica para estatística descritiva utilizando a linguagem R

Scientific methodology for descriptive statistics using the R language

Alek Fröhlich[1,2], André Bernardes Turcato[1,3], Daniel Guimarães Tiezzi[1,3,4,5]

[1] Laboratory for Translational Data Science - Faculdade de Medicina de Ribeirão Preto - Universidade de São Paulo
[2] Departamento de Matemática - Universidade Federal de Santa Catarina
[3] Informática Biomédica - Faculdade de Medicina de Ribeirão Preto - Universidade de São Paulo
[4] Setor de Oncologia e Mastologia do Departamento de Ginecologia e Obstetrícia da Faculdade de Medicina de Ribeirão Preto - Universidade de São Paulo
[5] Centro de Pesquisa Avançada em Medicina (CEPAM) - Faculdade de Medicina da UNILAGO

[*Autor correspondente: dtiezzi@usp.br]

Data de submissão: 25 de julho de 2023
Data de aceite: 31 de agosto de 2023
Data de publicação: 06 de setembro de 2023

RESUMO

Este artigo visa descrever os procedimentos básicos de uma análise descritiva de dados clínicos utilizando a linguagem de programação R, uma ferramenta baseada no conceito de software livre, e tem como objetivo servir como guia para pesquisadores da área.

Palavras-chave: linguagem de programação; estatística descritiva; metodologia

ABSTRACT

This article aims to describe the basic procedures of a descriptive analysis of clinical data using the R programming language, a tool based on the concept of free software, and aims to serve as a guide for researchers in the area.

Keywords: programming language; descriptive analysis; methodology

INTRODUÇÃO

O desenvolvimento tecnológico recente na área da computação permitiu a geração e compartilhamento de um grande volume de dados em curto período de tempo. Atualmente, o volume de dados clínicos, epidemiológicos e biológicos depositados em repositórios de acesso público é enorme, o que permite análises inovadoras pela comunidade científica[1-3]. Existe uma série de ferramentas disponíveis para acesso, manipulação e análise desses dados. O R é uma linguagem de programação interpretada que foi desenvolvida para análises estatísticas e seu interpretador é disponível gratuitamente. Ela é uma linguagem de alto nível e a sua semelhança com a linguagem natural facilita a utilização por pessoas não especializadas. A ferramenta é baseada em softwares livres[4] e pode ser utilizada nos sistemas operacionais GNU/Linux, Windows® e Mac OS[5]. Aqui iremos descrever os passos para a utilização desta tecnologia na análise descritiva de dados clínicos.

INSTALAÇÃO E CONFIGURAÇÃO DO AMBIENTE DE TRABALHO

Para a utilização da linguagem R é necessário a instalação do interpretador. O *download* deve ser feito diretamente do repositório *The Comprehensive R Archive Network* (CRAN) em https://cran.r-project.org/. A ferramenta Rstudio é um desktop que facilita o uso do R e é recomendada a instalação no sistema. O *software* deve ser adquirido em https://posit.co/download/rstudio-desktop/#download. Exemplos de instalação, configurações básicas do sistema e um tutorial básico da utilização do R e Rstudio podem ser encontrados em alguns livros especializados ou mesmo *online*.

Banco de dados

Iremos utilizar um dos bancos de dados disponível publicamente no pacote *medicaldata* denominado de *blood_storage* para o desenvolvimento dessas análises. O banco de dados apresenta uma série de dados clínicos, bioquímicos e histológicos de 316 pacientes com diagnóstico de câncer de próstata[6].

Antes de iniciar a análise descritiva propriamente dita, precisamos carregar os dados para a análise. No código apresentado a seguir temos as funções para instalação do pacote e carregamento do banco de dados no ambiente.

```
# instalar o pacote para acesso ao banco de dados
> install.packages('medicaldata')
# carregar o banco de dados
> data <- medicaldata::blood_storage
# verifica a classe do objeto
> class(data)
[1] "data.frame"
# verifica as dimensões do objeto
> dim(data)
```

```
[1] 316  20
# visualizar parte da tabela (4 linhas e 4 colunas)
> data[1:4, 1:4]
  RBC.Age.Group Median.RBC.Age  Age AA
1             3             25 72.1  0
2             3             25 73.6  0
3             3             25 67.5  0
4             2             15 65.8  0
```

Este procedimento atribui o objeto da classe *data.frame* para a variável data. Os objetos desta classe em R formam um conjunto de dados estruturados em forma de tabela, uma estrutura bidimensional com 316 linhas e 20 colunas. As colunas e linhas de *data.frames* em R possuem nomes (*colnames* e *rownames*, respectivamente). Veja que a primeira coluna tem o nome de "RBC.Age.Group" e os *rownames* estão numerados de forma ordenada. Cada coluna representa uma variável do banco de dados e cada linha representa um paciente. Para fins didáticos, somente as variáveis "Age" (idade dos pacientes), "PVol" (volume da próstata em cm3, "T.Stage" (estágio do tumor), "PreopPSA" (valor da dosagem sérica da proteína PSA antes da cirurgia) e "Recurrence" (se houve ou não recorrência da doença após o tratamento) serão utilizadas para a estatística descritiva.

Estatística descritiva

A estatística descritiva é utilizada para descrever os padrões básicos dos dados de um estudo permitindo a sua visualização geral e distribuição[7]. O método a ser aplicado para descrição do dado depende do tipo de dado. Desta forma, o primeiro passo é visualizar os dados, definir qual classe que eles pertencem antes de iniciar as análises e aplicar métodos de pré-processamento de dados.

```
# Seleção de variáveis de interesse
> data <- data[, c('Age', 'PVol', 'T.Stage',
'PreopPSA', 'Recurrence')]
# visualizar os dados
> head(data)
   Age  PVol T.Stage PreopPSA Recurrence
1 72.1  54.0       1    14.08          1
2 73.6  43.2       2    10.50          1
3 67.5 102.7       1     6.98          0
4 65.8  46.0       1     4.40          0
5 63.2  60.0       1    21.40          0
6 65.4  45.9       1     5.10          0
# Definir o tipo de dado
> data$Age <- as.numeric(data$Age)
```

```
> data$PVol <- as.numeric(data$PVol)
> data$T.Stage <- as.factor(data$T.Stage)
> data$PreopPSA <- as.numeric(data$PreopPSA)
> data$Recurrence <- as.factor(data$Recurrence)
# Pré processamento dos dados com remoção de dados faltantes
any(is.na(data))
[1] TRUE
> data <- data[complete.cases(data), ]
> dim(data)
[1] 293   5
```

Podemos notar que as variáveis “Age”, “PVol” e “PreopPSA” são variáveis numéricas contínuas e as variáveis “T.Stage” e “Recurrence” são categóricas. Após a definição das classes é possível ter uma ideia geral da distribuição dos dados. Existem alguns dados faltantes para as variáveis “PVol”, “T.Stage” e “PreopPSA” e as linhas correspondentes foram removidas restando dados completos de 293 pacientes. O pré-processamento de dados sempre é necessário e consiste em uma metodologia rica em ações para identificar dados enviesados e dados faltantes, bem como o tratamento dos dados em situações de imputação de dados e normalização. Esta metodologia não faz parte do escopo do tema em questão.

Para os casos de variáveis nominais, a descrição pode ser realizada pela contagem de casos em cada categoria ou pela frequência relativa. O código a seguir extrai as contagens por categoria e gera uma tabela de contingência.

```
# Contagem de casos por categoria
> table(data$T.Stage)
  1   2
260  33
> table(data$Recurrence)
  0   1
243  50
> # Contagem cruzada por categoria - tabela de contingência
> table(data$T.Stage, data$Recurrence)
      0   1
  1 224  36
  2  19  14
```

De forma semelhante, é possível extrair as frequências relativas de cada categoria.

```
# Frequências relativas em porcentagem
> round(prop.table(table(data$T.Stage))*100,1)
   1    2
88.7 11.3
> round(prop.table(table(data$Recurrence))*100,1)
   0    1
82.9 17.1
>round(prop.table(table(data$T.Stage,
data$Recurrence))*100,1)
       0    1
  1 76.5 12.3
  2  6.5  4.8
```

Para as variáveis numéricas devemos utilizar medidas de tendência central e de dispersão para a descrição dos dados. Esses parâmetros são facilmente extraídos com o R.

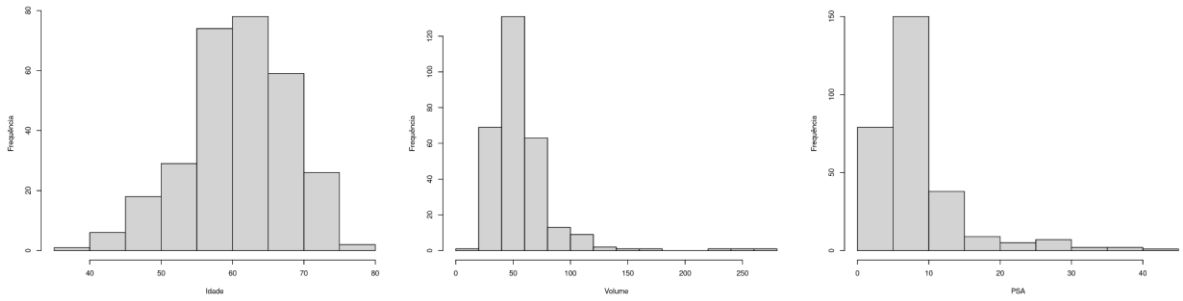
```
# Medidas de tendência central e de dispersão
> summary(data$Age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  38.40   56.00   61.80   60.97   66.10   79.00
> sd(data$Age)
[1] 7.285216
> IQR(data$Age)
[1] 10.1
> summary(data$PVol)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  19.40   41.00   48.50   56.06   64.00  274.00
> sd(data$PVol)
[1] 29.39968
> IQR(data$PVol)
[1] 23
> summary(data$PreopPSA)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.300   5.000   6.200   8.359   9.100  40.100
> sd(data$PreopPSA)
[1] 6.153984
> IQR(data$PreopPSA)
[1] 4.1
```

A estatística descritiva utiliza tabelas e representações gráficas que facilitam a compreensão da distribuição geral dos dados. A tabela é uma boa opção para representar a descrição da casuística. No caso de variáveis numéricas, antes da descrição das variáveis, precisamos definir quais são as medidas de tendência central e dispersão que melhor representam aquela variável. Para variáveis contínuas com distribuição normal ou próxima da normalidade, o uso da média e do desvio padrão é o mais adequado. Já os casos com distribuição fora da normalidade e com grandes desvios, a mediana e a distância interquartil ou o desvio absoluto da média devem ser utilizados. Existem diferentes maneiras de testar o tipo de distribuição dos dados. Vamos utilizar algumas métricas simples para checar as variáveis do banco de dados.

Um método clássico é a avaliação visual do histograma de frequências. O histograma pode ser facilmente ilustrado no R (Figura 1).

```
# Histogramas de frequência
     > png('histogramas.png', height = 5, width = 18, units =
     'in', res = 300)
     > par(mfrow=c(1,3))
     > hist(data$Age)
     > hist(data$PVol)
     > hist(data$PreopPSA)
     > dev.off()
```

Figura 1. Histogramas de frequência para as variáveis



Fonte: Próprio autor

De acordo com o histograma, a variável "Age" apresenta uma distribuição próxima da distribuição normal. Os valores próximos entre a média e a mediana também sugerem a normalidade. Por outro lado, as outras variáveis têm uma distribuição nitidamente assimétrica enviesadas para a direita. Aqui vamos tratar somente a variável "Age" como tendo uma distribuição normal.

Para criar a tabela vamos utilizar as medidas de tendência central e dispersão para as variáveis contínuas e as frequências absolutas e relativas de

cada variável nominal. Um exemplo de representação das características da população está demonstrado na Tabela 1.

Tabela 1. Características clínicas de pacientes com diagnóstico de adenocarcinoma de próstata.

Variável	**n = 293**
Idade - média; desvio padrão	61; 7,3
Volume da próstata - mediana; distância interquartil	48,5; 23
Estágio Clínico - n (%)	
I	260 (87,7%)
II	33 (11,3%)
PSA - mediana; distância interquartil	6,2; 4,1
Recorrência - n (%)	
Não	243 (82,9%)
Sim	50 (17,1%)

Fonte: Próprio autor

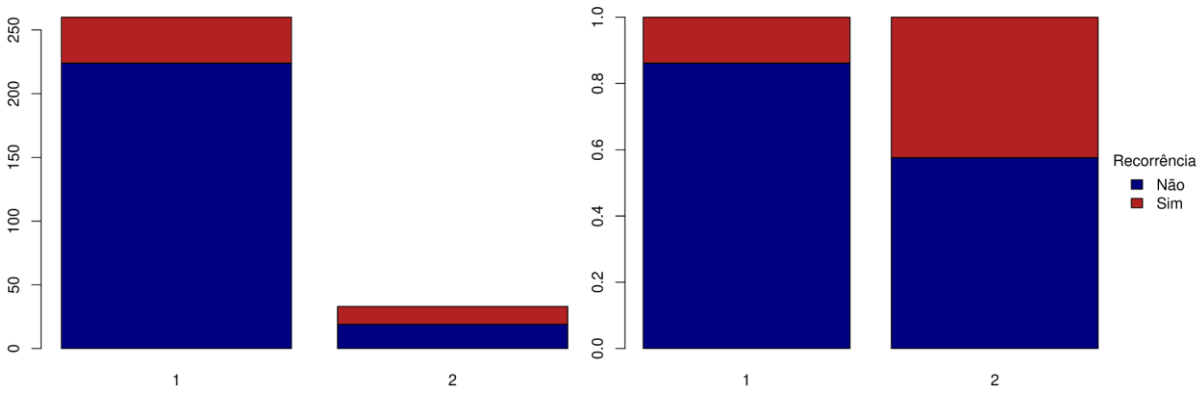
A visualização gráfica dos dados é um componente importante da estatística descritiva. Vários aspectos intuitivos para a inferência estatística vêm da visualização gráfica dos dados. A utilização de gráficos específicos para cada situação é recomendada e, no caso em questão, é possível apresentar três situações distintas: *i)* representação gráfica de duas variáveis categóricas; *ii)* representação gráfica de duas variáveis contínuas e *iii)* representação gráfica de uma variável categórica e uma contínua.

i) Duas variáveis categóricas: geralmente esta situação é muito bem descrita em uma tabela. No entanto, existem situações que uma ou mais variáveis apresentam múltiplas categorias e torna interessante a utilização de gráficos para melhor representar a distribuição dos dados. Neste caso, os gráficos de barra são os de escolha (Figura 2).

```
> png('barras.png', height = 5, width = 13, units = 'in',
res = 300)
> par(mfrow=c(1,2), mar = c(5,2,2,1))
> barplot(table(data$Recurrence, data$T.Stage), col =
c('navy', 'firebrick'))
> par(mar = c(5,1,2,5), xpd=TRUE)
```

```
>barplot(prop.table(table(data$Recurrence,data$T.Stage),2)
, col = c('navy', 'firebrick'))
> legend('right', legend = c('Não','Sim'), fill = c('navy',
'firebrick')   ,   inset=c(-0.17,0),   bty='n',   title=
'Recorrência')
> dev.off()
```

Figura 2. Gráficos de barras. Na esquerda temos as frequências absolutas e na direita a frequência relativa para cada categoria.

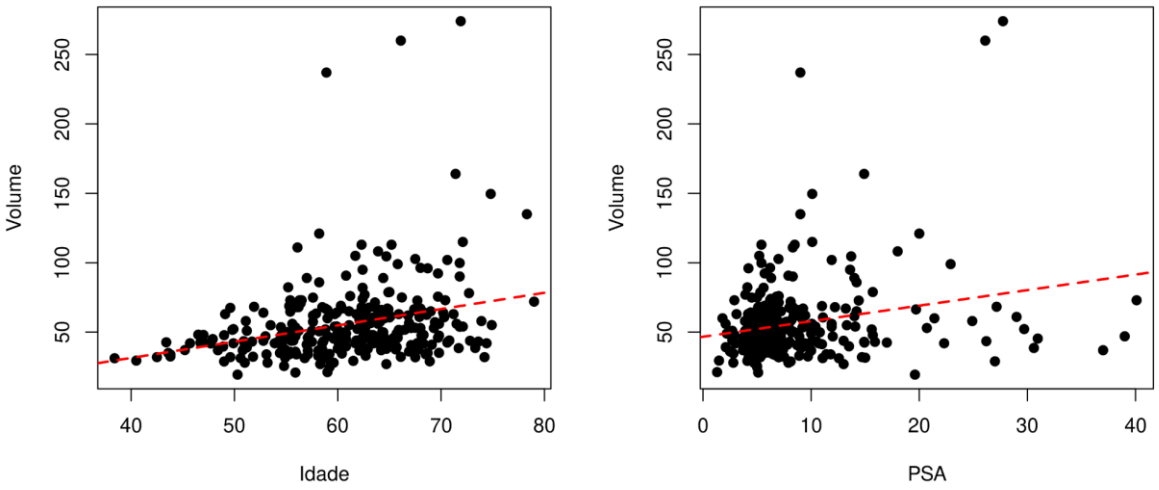


Fonte: Próprio autor

ii) Duas variáveis numéricas contínuas: o gráfico de dispersão é o ideal nesta situação. Ele permite que o leitor tenha uma ideia muito aproximada da correlação entre as duas variáveis e fazer inferências se trata de uma correlação linear positiva ou negativa bem como se é um tipo específico de correlação não linear (Figura 3).

```
> png('dispersao.png', height = 5, width = 10, units =
'in', res = 300)
> par(mfrow=c(1,2))
> plot(data$PVol ~ data$Age, pch = 19, ylab = 'Volume',
xlab = 'Idade')
> abline(lm(data$PVol ~ data$Age), col = 'red', lty = 2,
lwd = 2)
>  plot(data$PVol  ~  data$PreopPSA,  pch  =  19,  ylab  =
'Volume', xlab = 'PSA')
> abline(lm(data$PVol ~ data$PreopPSA), col = 'red', lty
= 2, lwd = 2)
> dev.off()
```

Figura 3. Gráficos de dispersão do volume da próstata em relação à idade (esquerda) e o valor do PSA (direita). As linhas tracejadas em vermelho representam a regressão linear.



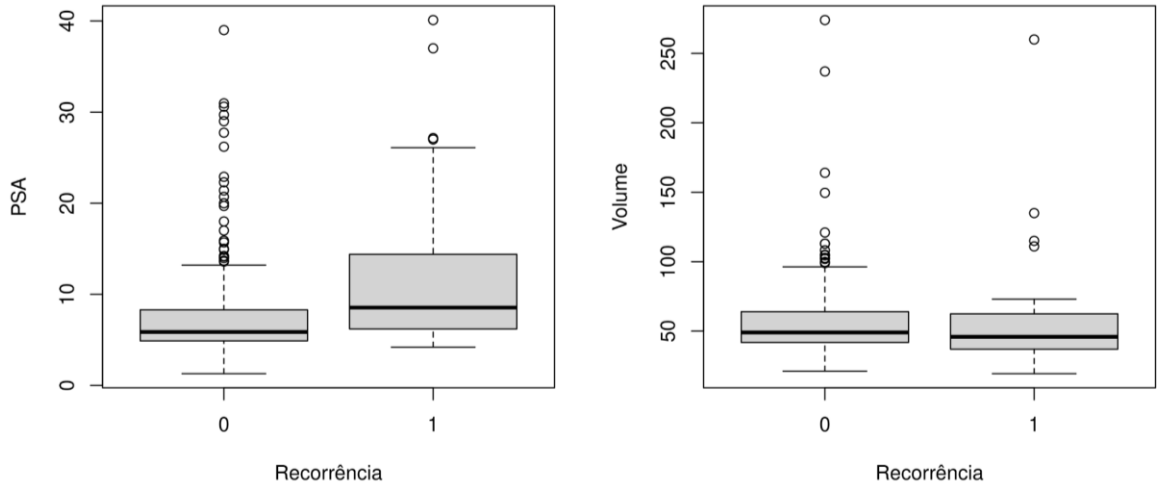
Fonte: Próprio autor

Note que, visualmente, existe uma melhor correlação linear entre a idade e o volume da próstata. Com o valor do PSA a correlação linear não é evidente e podemos notar que existem potenciais valores discrepantes (*outliers*) na distribuição dos valores de volume e PSA, o que já sugeria os histogramas correspondentes.

iii) Uma variável contínua e uma categórica: o diagrama ou gráfico de caixa ou *box plot* é o mais adequado para esta representação. A caixa destaca a mediana, o primeiro e o terceiro quartis bem como os limites inferior e superior e os *outliers* (Figura 4).

```
> png('boxplot.png', height = 5, width = 10, units = 'in', res
= 300)
> par(mfrow=c(1,2))
> boxplot(data$PreopPSA ~ data$Recurrence, ylab = 'PSA', xlab
= 'Recorrência')
> boxplot(data$PVol ~ data$Recurrence, ylab = 'Volume', xlab
= 'Recorrência')
> dev.off()
```

Figura 4. Diagrama de caixa demonstrando a distribuição dos valores do PSA (esquerda) e do volume da próstata (direita) em pacientes com (1) e sem (0) recorrência



Fonte: Próprio autor

Note no diagrama acima que os pontos discrepantes estão representados com pequenos círculos. A análise visual do diagrama permite sugerir diferenças na distribuição dos valores entre os grupos.

CONCLUSÃO

A estatística descritiva é uma ferramenta poderosa para a realização de trabalhos científicos na área da saúde. Com a alta disponibilidade de dados digitais, torna-se necessário o entendimento básico de ambientes e linguagens de programação para manipulação e interpretação dos dados digitalizados. Neste artigo, demonstramos o passo a passo do uso da linguagem R, uma plataforma de programação que pode ser adquirida de forma gratuita, no contexto da estatística descritiva desde a instalação e configuração do *software* até seu emprego no pré-processamento e visualização dos dados. Esperamos que esse artigo possa servir de guia para impulsionar ainda mais a transição da análise manual para a análise computacional de dados médicos e biológicos. O código utilizado neste tutorial pode ser adquirido diretamente no *github* em https://github.com/lab-tds/data_science_course.

REFERÊNCIAS BIBLIOGRÁFICAS

1. DATASUS – Ministério da Saúde [Internet]. [cited 2023 Jul 24]. Available from: https://datasus.saude.gov.br/

2. The Cancer Genome Atlas Program (TCGA) - NCI [Internet]. 2022 [cited 2023 Jul 24]. Available from: https://www.cancer.gov/ccg/research/genome-sequencing/tcga

3. The Cancer Genome Atlas Research Network. Integrated genomic and molecular characterization of cervical cancer. Nature. 2017 Mar 16;543(7645):378–84.

4. Essays and Articles - GNU Project - Free Software Foundation [Internet]. [cited 2023 Jul 24]. Available from: https://www.gnu.org/philosophy/essays-and-articles.en.html

5. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2021. Available from: https://www.R-project.org/

6. Cata JP, Klein EA, Hoeltge GA, Dalton JE, Mascha E, O'Hara J, et al. Blood Storage Duration and Biochemical Recurrence of Cancer After Radical Prostatectomy. Mayo Clin Proc. 2011 Feb;86(2):120–7.

7. Wickham H, Grolemund G. R for Data Science [Book] [Internet]. [cited 2023 Jul 24]. Available from: https://www.oreilly.com/library/view/r-for-data/9781491910382/