INTRODUÇÃO À ANÁLISE DESCRITIVA DE DADOS NO *SOFTWARE* R

Vanessa Ferreira Sehaber
Universidade Estadual do Paraná Campus Campo Mourão - UNESPAR/CM
vsehaber@gmail.com

Resumo: O *software* R tem ganho espaço na comunidade acadêmica nacional e internacional por ser um *software* livre, disponibilizar códigos abertos, além de permitir maior liberdade ao pesquisador no tratamento de seus dados. A proposta deste minicurso é de introduzir a linguagem ao usuário, capacitá-lo para a importação de dados e obter análises descritivas de conjuntos de dados. Este minicurso se justifica pela necessidade que muitos alunos de graduação têm em organizar e analisar dados de pesquisa, como, por exemplo, na fase de elaboração de seu trabalho de conclusão de curso. Além de saber executar as análises, os participantes sairão do minicurso com um olhar mais crítico sobre como explorar possíveis relações que os dados possam revelar por meio de métodos de representação gráfica, e terão as noções necessárias para o entendimento inicial da informação e da variabilidade presente em conjunto de dados, especialmente, antes do uso de análises estatísticas posteriores.
Palavras-chave: Análise descritiva de dados. Gráficos. Estatística. *Software* R.

OBJETIVOS DA ATIVIDADE

- Importar conjunto de dados e manipulá-los;
- Calcular medidas de posição e dispersão;
- Representar graficamente variáveis de alguns conjuntos de dados de modo a desenvolver uma descrição clara e concisa de cada problema proposto;
- Explorar possíveis relações que os dados possam revelar por meio da representação gráfica e identificar, ou pelo menos tentar identificar, os fatores importantes que afetam os dados;
- Discutir como a variabilidade afeta os dados coletados e usados para tomar decisões na pesquisa;
- Produzir uma saída rápida e dinâmica das análises feitas pelo aluno do minicurso.

DESCRIÇÃO DA PROPOSTA

Este minicurso tem como proposta introduzir estudantes de graduação a análise descritiva dos dados, que é o processo de obtenção de informações significativas a partir de conjuntos de dados, em geral demasiadamente grandes para serem trabalhados diretamente (DOWNING; CLARK, 2006).

A análise descritiva dos dados desempenha papel importante em uma pesquisa, seja esta desenvolvida na matemática, agricultura, na biologia, no comércio, na química, nas comunicações, na economia, na educação, na eletrônica, na medicina, na física, nas ciências políticas, na psicologia e em outros numerosos campos da ciência e da engenharia (SPIEGEL, 1967).

Em alguma fase de seu trabalho, o pesquisador depara-se com o problema de analisar e entender um conjunto de dados relevante ao seu particular objeto de estudos. Ele necessitará trabalhar os dados para transformá-los em informações, para compará-los com outros resultados ou, ainda, para julgar sua adequação a alguma teoria (BUSSAB; MORETTIN, 2017; ASSIS et al., 2016).

Uma análise descritiva dos dados não se limita apenas a resumos numéricos, por exemplo, o cálculo de algumas medidas de posição e varibilidade, como a média e variância, por exemplo. Há também o uso de métodos gráficos, o quais têm um forte apelo visual e permitem simplificar a informação que há em um conjunto de dados. Além isso, a variabilidade das variáveis pode ser melhor explorada (Montgomery; Runger, 2016).

Segundo Bussab e Morettin (2017), os gráficos são utilizados para diversos fins:

- Buscar padrões e relações;
- Confirmar (ou não) certas expectativas que se tinha sobre os dados;
- Descobrir novos fenômenos;
- Apresentar resultados de modo mais rápido e fácil;
- avaliar a qualidade de ajuste de métodos estatísticos aos dados.

Para a implementação de uma análise descritiva de dados, pacotes estatísticos foram desenvolvidos e atualmente são usados em larga escala, tanto no meio acadêmico como em indústrias, bancos, órgãos do governo, etc. (BUSSAB; MORETTIN, 2017).

Neste minicurso, usaremos programas do repositório de pacotes R (R CORE TEAM, 2019), que podem ser obtidos livremente do Comprehensive R archive Network (CRAN), no endereço: http://cran.r-project.org.

O QUE É O R, ONDE É UTILIZADO, E PORQUÊ UTILIZÁ-LO?

Segundo Motwani (2019), o s*oftware* R é um ambiente computacional e uma linguagem de programação que vem progressivamente se especializando em manipulação, análise e visualização gráfica de dados. Na atualidade é considerado o melhor ambiente computacional para essa finalidade. O ambiente está disponível para diferentes sistemas operacionais: Unix/Linux, Mac e Windows (WICHKHAM; GROLEMUND, 2017).

Em janeiro de 2009, o jornal americano The New York Times publicou um artigo sobre o aumento de aceitação do R entre os analistas de dados e apresentando uma potencial ameaça para a quota de mercado ocupada por pacotes estatísticos comerciais, como o SAS (VANCE, 2009).

O s*oftware* R é altamente expansível com o uso dos pacotes. Os pacotes são bibliotecas com dados e funções para diferentes áreas do conhecimento relacionado a estatística e áreas afins, para produzir o resultado desejado (ZUMEL; MOUNT, 2014)

De acordo com Thomas et. al. (2015), outro motivo para se usar o s*oftware* R é que muitos dos maiores cientistas e estatísticos do mundo agora usam R, porque ele é uma fonte aberta, e há frequente contribuição de novas técnicas de análises. Além disso, s*oftware* R é geralmente mais atualizado do que outros pacotes estatísticos tradicionais. Os autores acrescentam: "Eu não estou dizendo a você que será fácil aprender R - Eu estou dizendo que valerá muito a pena."

Outros autores utilizaram o s*oftware* R em suas análises. Podemos citar o trabalho de Basso e Pimentel (2015), onde estudaram sobre a eficácia do processo seletivo estendido no curso de estatística da UFPR; Coelho (2014) trabalhou sobre teoria de resposta ao item para a análise de exames multidisciplinares; Lucena (2018) pesquisou os fatores de risco associados à repetência escolar em alunos de escolas públicas do sertão de Pernambuco; e Sehaber (2018), onde desenvolveu um modelo para análise de dados com padrões geográficos ao longo do tempo.

RECURSOS NECESSÁRIOS PARA A REALIZAÇÃO DA ATIVIDADE

Serão necessários:

- Multimídia;
- Lousa e giz/caneta;

- Extensões;
- Computadores.

RECURSOS NECESSÁRIOS PARA A REALIZAÇÃO DA ATIVIDADE

O minicurso consistirá de 4 horas. Inicialmente, será apresentado os recursos computacionais que o *software* R (R CORE TEAM, 2019) dispõe para desenvolver análises estatísticas e serão dadas algumas dicas de uso.

Após, serão apresentados os conjuntos de dados a serem trabalhados durante o minicurso e as variáveis de cada conjunto de dados serão classificadas como qualitativas e quantitativas. Esse procedimento é importante para confecção dos gráficos (dispersão, barras, setores, bloxplot, histogramas, etc.) a serem produzidos.

A seguir, serão dadas as instruções para o participantes iniciarem a importação de conjuntos de dados, os quais serão passados para os alunos previamente por e-mail.

Após, a sintaxe da linguagem será explicada para que as medidas de posição e dispersão sejam obtidas e interpretadas. Na sequência, será iniciada a construção dos gráficos e a interpretação dos gráficos ocorrerá simultaneamente.

No final, será mostrado como organizar de forma prática a saída dos resultados.

OBSERVAÇÕES E OUTRAS INFORMAÇÕES

Em essência, o *software* R possui um visual não muito amigável. Para facilitar o seu uso, recomendo instalar o RStudio, que é uma interface ao R, bastante difundida na comunidade acadêmica, com mais opções de interface.

Primeiro instale o R. Após esta instalação, instale o RStudio.

Links para instalar R versão 3.6.0

- Windows:

https://cran.r-project.org/bin/windows/base/

- Linux:

sudo apt-get install r-base r-base-core

- Vídeo com o passo a passo da instalação:

https://www.ufrgs.br/wiki-r/index.php?title=Instala%C3%A7%C3%A3o_do_R

Link para instalar RStudio (Interface do R)

- Windows:

https://www.rstudio.com/products/rstudio/download/#download. Ver o item " Installers for Supported Platforms" e escolha o arquivo de acordo com o seu Windows

- Linux:

sudo dpkg -i rstudio-0.99-amd64.deb

- Vídeo com o passo a passo da instalação:

https://www.ufrgs.br/wiki-r/index.php?title=Instala%C3%A7%C3%A3o_do_RStudio

Maiores detalhes sobre instalação, ver o tutorial:

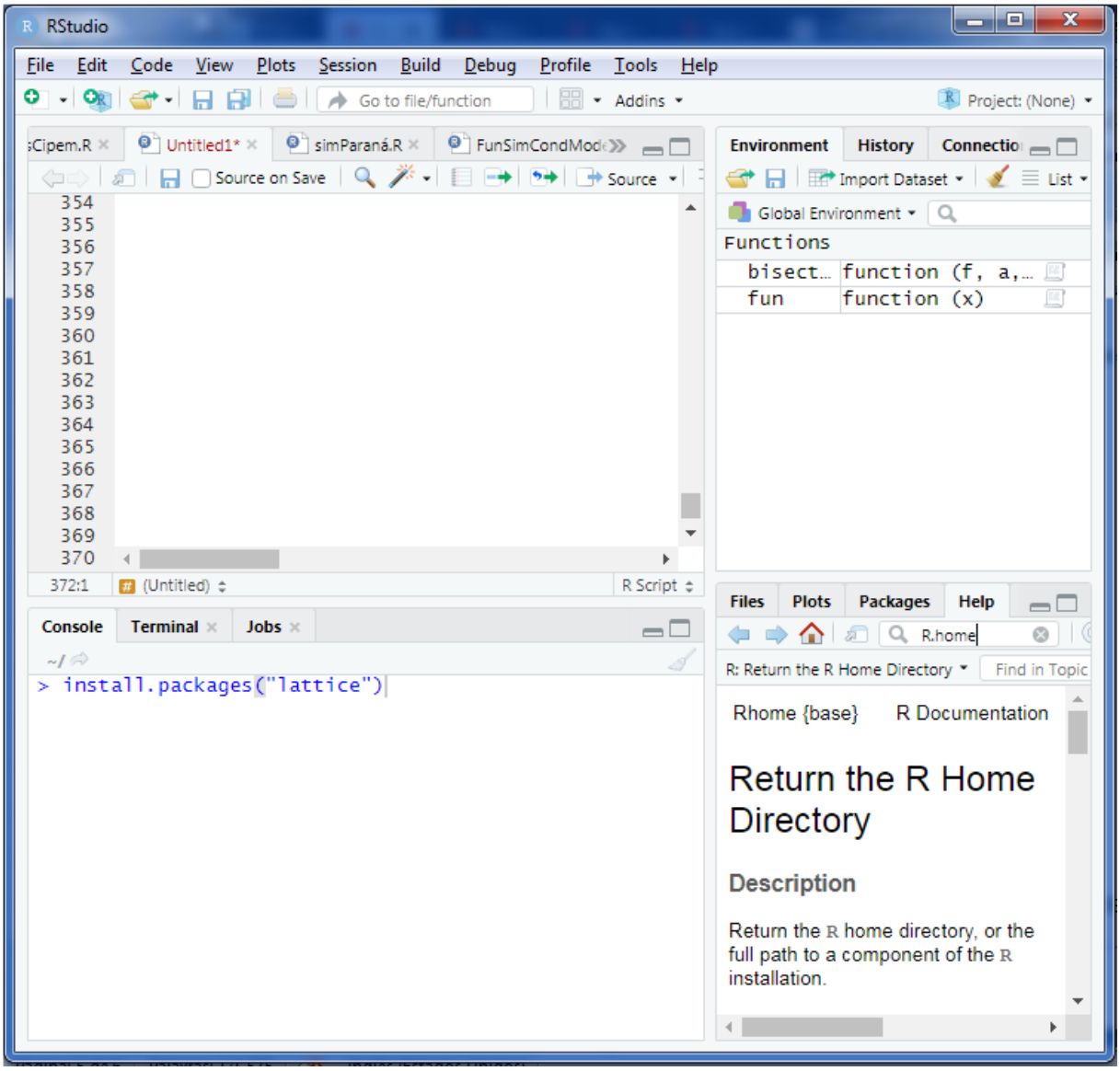
- A enciclopédia sobre Linguagem de Programação R que todos podem colaborar!

https://www.ufrgs.br/wiki-r/index.php?title=Bem-vindo_%C3%A0_Wiki_R

Após esses procedimentos, instalar os seguintes pacotes ao executar no console os seguintes comandos, um de cada vez:

```
install.packages("lattice")
install.packages("latticeExtra")
install.packages("knitr")
install.packages("markdown")
install.packages("reshape")
install.packages("ggplot")
install.packages("corplot")
install.packages("ellipse")
install.packages("knitrBootstrap")
install.packages("evaluate")
install.packages("stringr")
install.packages("htmltools")
install.packages("isonlite")
install.packages("base64enc")
install.packages("rprojroot")
install.packages("mime")
```

Por exemplo, para instalar o pacote "lattice", escreva install.packages("lattice") e pressione a tecla Enter.



A instalação será feira mediante conexão com internet. Caso contrário, os pacotes não poderão ser utilizados durante o minicurso.

REFERÊNCIAS

ASSIS, J. P. et al. **Estatística Descritiva**. Piracicaba: FEALQ, 2016.

BASSO, D. R. PIMENTEL, K. M. R. **Estudo Sobre a Eficácia do Processo Seletivo Estendido no Curso de Estatística da UFPR**. Monografia (Graduação em Bacharelado em Estatística) – Universidade Federal do Paraná, UFPR, Curitiba, 2015.

BUSSAB, W. O.; MORETTIN, P. A. **Estatística Básica**. ed. 9. São Paulo: Saraiva, 2017.

COELHO, E. C. **Teoria de Resposta ao Item - Desafios e Perspectivas em Exames Multidiciplinares**. Tese (Doutorado em Métodos Numéricos em Engenharia) – Programa de

Pós-Graduação em Métodos Numéricos em Engenharia. Universidade Federal do Paraná, UFPR, Curitiba, 2014.

DOWNING, D.; CLARK, J. **Estatística Aplicada**. ed. 2. São Paulo: Saraiva, 2006.

LUCENA, L. R. Fatores de Risco Associados à Repetência Escolar em Alunos de Escolas Públicas do Serão de Pernambuco. **Matemática e Estatística em Foco**. vol. 6, n 2, p. 44-51, 2018.

MONTGOMERY, D. C.; RUNGER, G. C. **Estatística Aplicada e Probabilidade para Engenheiros**. ed. 6. São Paulo: LTC, 2016.

MOTWANI, B. **Data Analytics With R**. Ontario: Wiley, 2019.

R CORE TEAM. **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

SEHABER, V. F. **A Conditional Geostatistical Spatio-Temporal Model to Non-fixed Point Locations**. Tese (Doutorado em Métodos Numéricos em Engenharia) – Programa de Pós-Graduação em Métodos Numéricos em Engenharia. Universidade Federal do Paraná, UFPR, Curitiba, 2018.

SPIEGEL, M. R. **Estatística**. Coleção Schaum. Rio de Janeiro: Livro Técnico, 1967.

THOMAS, R. et al. **Data Analysis with R statistical Software: A Guidebook for Scientists.** London: Eco-Explore, 2015.

VANCE, A. **Data Analysts Captivated by R's Power**. The New York Times, 2009.

ZUMEL, N.; MOUNT, J. **Practical Data Science with R**. New York: Manning Publications, 2014.

WICKHAM, H.; GROLEMUND, G. **R for Data Science: Import, Tidy, Transform, Visualize, and Model Data**. New York: O'Reilly Media, 2017.