Revista Edutec

ISSN 2764-9067
v. 4, n. 1, e21886 | 2024

heRcules: UM REPOSITÓRIO COM SCRIPTS PARA O APRENDIZADO DA ANÁLISE DE DADOS EM R

heRcules: A REPOSITORY WITH SCRIPTS FOR LEARNING DATA ANALYSIS IN R

heRcules: UN REPOSITORIO CON SCRIPTS PARA EL APRENDIZAJE DEL ANÁLISIS DE DATOS EN R

Hércules Rezende Freitas
Universidade Federal do Rio de Janeiro

RESUMO. *A análise de dados é uma etapa crucial no desenvolvimento de projetos científicos, desempenhando um papel central na validação e interpretação dos resultados obtidos. Antes de iniciar a coleta de dados, o pesquisador deve planejar seus experimentos e análises de maneira meticulosa e estruturada, garantindo uma abordagem robusta que minimiza a influência de vieses que possam comprometer a validade dos resultados. O presente documento relata a criação do repositório "heRcules", um recurso de acesso público que oferece modelos de scripts em linguagem R para a análise de dados científicos, com foco especial nas disciplinas das Ciências Biológicas e da Saúde. Este repositório é projetado para ser uma ferramenta valiosa para pesquisadores, fornecendo scripts prontos para executar tarefas essenciais como planejamento experimental, análise de dados, visualização de resultados e testes de hipóteses. O modelo inicial, descrito neste documento, inclui scripts para uma ampla gama de funções: cálculo de tamanho amostral, cálculo de poder estatístico, importação de planilhas, criação de vetores e data frames, estatísticas descritivas, exportação de arquivos, criação de gráficos (tanto com base R quanto com ggplot2), testes de outliers, testes de normalidade, e criação de notebooks com o R Markdown. O repositório está hospedado na plataforma GitHub (https://github.com/drhrf/heRcules.git), assegurando que os recursos estejam disponíveis de forma eficiente, gratuita e colaborativa para a comunidade científica. Esse repositório tem como objetivo não apenas facilitar o trabalho de pesquisadores individuais, mas também promover a transparência e a reprodutibilidade da pesquisa científica, oferecendo uma base sólida para a condução de análises de dados rigorosas e bem fundamentadas, tais como aquelas exemplificadas no presente modelo.*

Palavras-chave: *R. Análise de dados. Tamanho amostral. Gráficos. GitHub. Teste de hipóteses.*

ABSTRACT. *Data analysis is a crucial step in the development of scientific projects, playing a central role in the validation and interpretation of study results. Before data collection begins, the researcher must meticulously and systematically plan their experiments and analyses, ensuring a robust approach that minimizes the influence of biases that could compromise the validity of the results. This document reports the creation of the "heRcules" repository, a public resource offering script models in the R language for scientific data analysis, with a particular focus on the Biological and Health Sciences. This repository is designed to be a valuable tool for researchers, providing ready-to-use scripts for executing essential tasks such as experimental planning, data analysis, result visualization, and hypothesis testing. The initial model,*

described in this document, includes scripts for a wide range of functions: sample size calculation, statistical power calculation, spreadsheet import, creation of vectors and data frames, descriptive statistics, file export, graph creation (using both base R and ggplot2), outlier tests, normality tests, and notebook creation with R Markdown. The repository is hosted on the GitHub platform (https://github.com/drhrf/heRcules.git), ensuring that the resources are available efficiently, free of charge, and collaboratively to the scientific community. This repository aims not only to facilitate the work of individual researchers but also to promote transparency and reproducibility in scientific research, providing a solid foundation for conducting rigorous and well-founded data analyses, such as those exemplified in the current model.

Keywords: *R. Data analysis. Sample size. Plots. GitHub. Hypotheses tests.*

RESUMEN. *El análisis de datos es una etapa crucial en el desarrollo de proyectos científicos, desempeñando un papel central en la validación e interpretación de los resultados obtenidos. Antes de comenzar la recopilación de datos, el investigador debe planificar sus experimentos y análisis de manera meticulosa y estructurada, garantizando que el enfoque sea robusto y minimizando la influencia de sesgos que puedan comprometer la validez de los resultados. El presente documento informa sobre la creación del repositorio "heRcules", un recurso de acceso público que ofrece modelos de scripts en lenguaje R para el análisis de datos científicos, con un enfoque especial en las disciplinas de Ciencias Biológicas y de la Salud. Este repositorio está diseñado para ser una herramienta valiosa para los investigadores, proporcionando scripts listos para ejecutar tareas esenciales como la planificación experimental, el análisis de datos, la visualización de resultados y las pruebas de hipótesis. El modelo inicial, descrito en este documento, incluye scripts para una amplia gama de funciones: cálculo del tamaño de la muestra, cálculo del poder estadístico, importación de hojas de cálculo, creación de vectores y data frames, estadísticas descriptivas, exportación de archivos, creación de gráficos (tanto con base R como con ggplot2), pruebas de valores atípicos, pruebas de normalidad y creación de cuadernos con R Markdown. El repositorio está alojado en la plataforma GitHub (https://github.com/drhrf/heRcules.git), lo que garantiza que los recursos estén disponibles de manera eficiente, gratuita y colaborativa para la comunidad científica. Este repositorio tiene como objetivo no solo facilitar el trabajo de los investigadores individuales, sino también promover la transparencia y la reproducibilidad de la investigación científica, proporcionando una base sólida para la realización de análisis de datos rigurosos y bien fundamentados, como los ejemplificados en el modelo actual.*

Palabras clave: *R. Análisis de datos. Tamaño de muestra. Gráficos. GitHub. Prueba de hipótesis.*

1 INTRODUÇÃO

Pesquisadores modernos possuem acesso a uma ampla variedade de recursos para o planejamento, análise e representação visual de dados científicos. Muitos desses recursos, porém, são funcionalmente limitados ou exigem compra de licenças a preços frequentemente proibitivos. Como alternativa, é possível utilizar linguagens de programação como R[1] e Python[2], que são estruturadas para não somente suportar todas as tarefas executáveis nos recursos pagos, mas fornecer ao usuário uma infinitude de outras ferramentas, inclusive a possibilidade de desenvolver os próprios programas e funções.

Apesar do poder dessas linguagens, a ausência de interfaces amigáveis acaba por afastar um grande número de potenciais usuários, mantendo-os reféns de ferramentas limitadas. Nesse contexto, nota-se a emergência de um movimento voltado para facilitar o acesso de usuários não familiarizados com linguagens de programação. A grande maioria desses novos recursos, porém, é publicado em inglês, falhando em dar suporte aos não proficientes na língua. Recentemente, alguns esforços têm sido feitos para remediar a situação, a exemplo do guia "Introdução ao R"[3], elaborado por V. J. Debastiani (Debastiani, 2021).

Com o objetivo de ampliar o acesso gratuito a ferramentas de computação estatística na língua portuguesa, o presente trabalho reporta a criação do repositório heRcules[4], que servirá como uma plataforma aberta para acesso a modelos de análise de dados na linguagem R. O primeiro modelo do repositório, apresentado abaixo, contém o código necessário para o planejamento amostral, o cálculo de estatística descritiva, a geração de visualizações e os testes de hipóteses. Os códigos disponíveis são completamente anotados em língua portuguesa para facilitar a compreensão e reutilização pelo usuário inexperiente.

[1] https:/www.r-project.org/
[2] https:/www.python.org/
[3] https://vanderleidebastiani.github.io/tutoriais/Introducao_ao_R.html
[4] https://github.com/drhrf/heRcules.git

2 RECOMENDAÇÕES

Ao usuário completamente inexperiente na linguagem R, recomenda-se a leitura do guia "Introdução ao R", mencionado anteriormente. Para utilização do ambiente de forma mais amigável, é recomendável que o usuário instale não somente o software R (R Core Team, 2013), mas também um ambiente de desenvolvimento integrado (IDE), que permitirá maior flexibilidade no uso da linguagem. O RStudio[5] é, de longe, a IDE mais utilizada para programação em R. Ambos os recursos são gratuitos e podem ser utilizados em diversas plataformas UNIX, Windows e MacOS.

O modelo abaixo foi construído de forma a representar um conjunto típico de dados produzidos por experimentos científicos nas áreas de Ciências Biológicas e da Saúde, mas pode ser facilmente modificado para atender a outros modelos e demandas. Adicionalmente, os blocos de código disponíveis aqui podem ser utilizados de forma independente para satisfazer alguma demanda específica. Para aplicá-los a um novo conjunto de dados, basta ao usuário formatar os próprios dados experimentais no modelo proposto e substituir o nome das variáveis dentro de cada bloco de código.

3 MODELO

3.1 Pacotes

Os pacotes abaixo foram utilizados no presente modelo (Quadro 1). Alguns pacotes importados aqui possuem funções similares e não precisam se repetir no caso de uma análise de fato. Se o pacote desejado não estiver disponível no sistema, deve-se instalá-lo com o comando "install.packages("nome_do_pacote")". O código necessário para importar os pacotes indicados se encontra no **Bloco 1** (arquivo suplementar 1).

[5] https://www.rstudio.com/products/rstudio/download/

Quadro 1 – Pacotes utilizados no presente modelo

Pacote	Referência
“psych”	Revelle, 2021
“dplyr”	Wickham *et al.*, 2021
“pastecs”	Grosjean; Ibanez, 2018
“fBasics”	Wuertz *et al.*, 2020
“skimer”	Waring *et al.*, 2021
“ggplot2”	Wickham, 2016
“xlsx”	Dragulescu; Arendt, 2020
“reshape2”	Wickham, 2007
“rstatix”	Kassambara, 2021
“kableExtra”	Zhu, 2021
“pwr”	Champely, 2020
“effsize”	Torchiano, 2020
“stats”	R Core Team, 2013

Fonte: Elaboração própria. **Legenda:** Os pacote "psych" fornece ferramentas para análises psicológicas e estatísticas, "dplyr" facilita a manipulação de dados, "pastecs" realiza análises descritivas e de séries temporais, "fBasics" oferece funções para análises financeiras, "skimer" permite a visualização interativa de séries temporais, "ggplot2" cria gráficos de alta qualidade, "xlsx" manipula arquivos Excel, "reshape2" transforma e reorganiza dados, "rstatix" simplifica testes estatísticos, "kableExtra" formata tabelas para apresentação, "pwr" calcula o tamanho de amostras e o poder estatístico, "effsize" calcula tamanhos de efeito e "stats" inclui funções básicas para análises estatísticas.

3.2 Dados do modelo

Os dados utilizados no presente modelo são adaptações de resultados obtidos em experimentos reais. A importação de objetos é um processo simples em R, e exige do usuário apenas uma linha de código. O caso mais habitual para pesquisas em Biociências é a importação de dados contidos em planilhas, normalmente salvas no formato .xlsx.

Para esse tipo de documento, basta que o pesquisador execute o seguinte código[6]: “df <- readxl::read_excel(‘caminho_do_arquivo_no_computador.xlsx’)”. Esse comando importará para o ambiente R a primeira aba da planilha e atribuirá ao objeto “df” as informações contidas nela.

A Tabela 1, criada com o pacote “kableExtra”, apresenta os dados utilizados no presente modelo. Aqui, foi simulado um experimento com grupos controle, controle

[6] Na janela “*Environment*” do RStudio existe a opção “*Import Dataset*”, que permite a importação de dados a partir de uma planilha do Excel. Usar esse recurso é menos recomendado ao usuário experiente.

positivo e dois tratamentos (I e II). O **Bloco 2** (arquivo suplementar 1) apresenta os códigos para a construção do conjunto de dados e a geração da tabela.

Tabela 1 – Dados utilizados no modelo

Controle	Positivo	Tratamento_I	Tratamento_II
95.26	0.67	97.93	76.47
92.13	0.10	88.39	75.64
97.84	0.34	96.36	76.49
98.84	1.66	92.36	75.33
90.50	0.62	88.26	75.45
87.31	0.66	97.31	78.72
90.77	1.29	87.79	76.70
89.36	1.75	95.44	76.20
95.00	0.56	87.45	76.08
97.33	0.82	91.02	75.49
90.82	1.05	92.61	76.43
89.81	0.89	91.05	78.48
97.75	1.31	91.24	75.94
98.18	1.89	92.28	76.91

Fonte: Elaboração própria. **Legenda:** Tabela contendo os valores simulados de quatro grupos experimentais: “Controle”, “Positivo”, “Tratamento_I” e “Tratamento_II”, a serem utilizados nos exemplos do presente trabalho.

3.3 Estimativa de poder estatístico e tamanho amostral

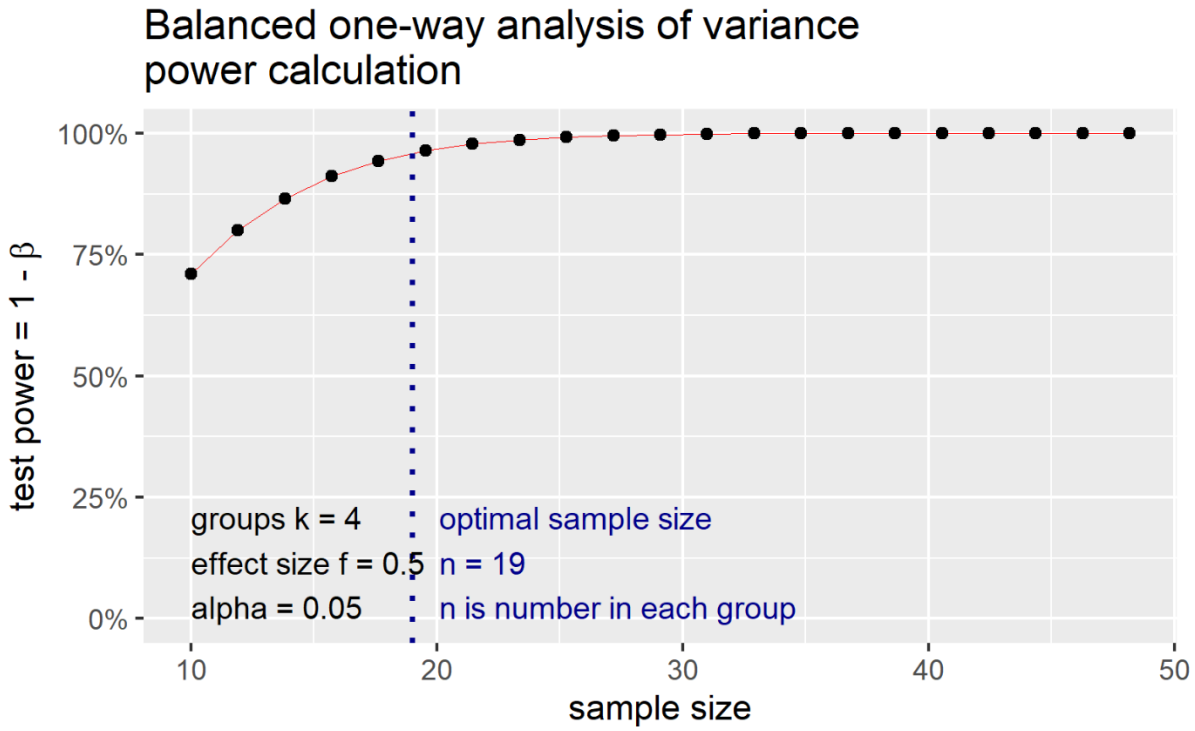
A estimativa do tamanho amostral para testes ANOVA e testes t é realizada com base em alguns parâmetros-chave que influenciam a capacidade do teste de detectar efeitos significativos. Esses parâmetros incluem o nível de significância (α), o poder estatístico desejado (1−β), a magnitude do efeito (ou tamanho do efeito) e a variabilidade dos dados.

O nível de significância (α) representa a probabilidade de cometer um erro do Tipo I (rejeitar a hipótese nula quando ela é verdadeira); o poder estatístico desejado (1−β, ou “um menos beta”), indica a probabilidade de detectar um efeito verdadeiro, ou seja, de não cometer um erro do Tipo II (não rejeitar a hipótese nula quando ela é falsa); a magnitude do efeito quantifica a diferença esperada entre as médias dos grupos para o teste t ou a variância entre grupos em relação à variância dentro dos grupos para a ANOVA. A variabilidade dos dados, geralmente expressa como o desvio padrão ou variância, que reflete a dispersão dos valores observados em relação à média. Esses

parâmetros combinados determinam o número de amostras necessário para assegurar que o teste seja sensível o suficiente para detectar diferenças estatisticamente significativas, caso estas existam.

As estimativas de tamanho amostral (**Bloco 3**, arquivo suplementar 1) e poder estatístico (**Bloco 4**, arquivo suplementar 1) para a análise de variância (ANOVA) podem ser realizadas utilizando o mesmo código, sendo apenas necessário substituir o valor desejado por "NULL". Também é possível criar um gráfico para representar a relação entre as variáveis da estimativa amostral (Figura 1, códigos disponíveis nos **Blocos 3 e 4**).

Figura 1 – Estimativa de tamanho amostral e poder estatístico para análise de variância



Fonte: Elaboração própria. **Legenda:** O gráfico apresenta o cálculo de poder estatístico (1 - β) para uma análise de variância (ANOVA) balanceada de um fator, em função do tamanho da amostra (número de participantes por grupo). O eixo x representa o tamanho da amostra por grupo, enquanto o eixo y exibe o poder do teste. A curva mostra que o poder aumenta conforme o tamanho da amostra cresce, estabilizando em 100% para tamanhos maiores. Os parâmetros utilizados incluem 4 grupos (k = 4), tamanho de efeito moderado (f = 0,5) e nível de significância (α = 0,05). A linha vertical pontilhada indica o tamanho de amostra ótimo calculado (n = 19 por grupo) para alcançar um poder adequado, sugerindo que esse é o tamanho mínimo recomendado para uma análise robusta.

O procedimento para estimar o poder estatístico e o tamanho amostral de testes t é similar ao realizado para a ANOVA (**Bloco 5**, arquivo suplementar 1). Também similar é a geração de um gráfico representando a relação entre poder estatístico e tamanho amostral.

O pacote “pwr” oferece, ainda, a possibilidade de se estimar o tamanho de efeito (Cohen D) sem a necessidade de calculadoras ou aplicativos externos. No presente modelo, pode-se observar um exemplo da análise de tamanho de efeito entre os grupos “Controle” e “Tratamento_II” (ver Tabela 1), que resulta em um d = 5.96, com intervalo de confiança entre 4.15 e 7.78 (**Bloco 6**, arquivo suplementar 1).

3.4 Estatística descritiva e exportação de dados

Uma das primeiras etapas de análise dos dados coletados é a geração de estatísticas descritivas. Dependendo do projeto, o pesquisador precisará de mais ou menos informações sobre seus resultados, mas é procedimento comum no universo das Ciências Biológicas e da Saúde obter ao menos uma medida de tendência central (e.g. média) e uma medida de dispersão (e.g. desvio padrão) do conjunto de dados. Nesse contexto, é importante que o analista desenvolva certa familiaridade com os principais conceitos da estatística descritiva.

A “mediana” é o valor central em um conjunto de dados ordenado; se o número de observações for ímpar, é o valor no meio da distribuição, enquanto se for par, é a média dos dois valores centrais. A “média” (ou média aritmética) é calculada dividindo-se a soma de todos os valores pelo número de observações, fornecendo uma medida de tendência central que representa o "valor típico" dos dados.

O “erro padrão da média” (se) é uma medida da precisão da média como estimativa do valor verdadeiro da população. Ele é calculado dividindo-se o desvio padrão pela raiz quadrada do número de observações, refletindo a variação que se esperaria entre as médias de diferentes amostras da mesma população. O “intervalo de confiança” (IC), considerando-se, por exemplo, o IC 95%, é uma faixa de valores que, com 95% de confiança, contém a média verdadeira da população; ele é calculado como

a média mais ou menos 1,96 vezes o erro padrão da média, assumindo uma distribuição normal dos dados. A “variância”, por sua vez, é uma medida da dispersão dos dados em relação à média, calculada como a média dos quadrados das diferenças entre cada valor e a média. Já o “desvio padrão” (sd) é a raiz quadrada da variância e expressa a dispersão dos dados na mesma unidade dos valores originais, sendo uma medida amplamente usada de variabilidade.

O “enviesamento” (*skewness*) quantifica a assimetria da distribuição dos dados. Se a distribuição é simétrica, o enviesamento é zero; valores positivos indicam uma cauda à direita mais longa, enquanto valores negativos indicam uma cauda à esquerda mais longa. A “curtose” (*kurtosis*) mede a "pontualidade" ou achatamento da distribuição em comparação com uma distribuição normal. Distribuições com alta curtose apresentam caudas mais pesadas e picos mais altos, enquanto aquelas com baixa curtose têm caudas mais leves e picos mais baixos.

No presente modelo, são apresentadas quatro alternativas de código para o cálculo de estatísticas descritivas. A utilização desses códigos, novamente, fica ao critério do pesquisador.

A primeira opção de comando, “summary( )” (**Bloco 7**, arquivo suplementar 1) de análise é gerada pelo ambiente nativo do R, e produz as seguintes informações: Mínimo, 1º quartil, Mediana, Média, 3º quartil e Máximo (Tabela 2).

Tabela 2 – Estatística descritiva com o comando “summary( )”

Controle	Positivo	Tratamento_I	Tratamento_II
Min. :87.32	Min. :0.1011	Min. :87.46	Min. :75.34
1st Qu.:90.57	1st Qu.:0.6326	1st Qu.:89.05	1st Qu.:75.72
Median :93.57	Median :0.8601	Median :91.77	Median :76.32
Mean:93.64	Mean:0.9766	Mean:92.11	Mean:76.46
3rd Qu.:97.65	3rd Qu.:1.3136	3rd Qu.:94.74	3rd Qu.:76.65
Max. :98.85	Max. :1.8930	Max. :97.93	Max. :78.72

Fonte: Elaboração própria. **Legenda:** Resumo estatístico dos valores de intensidade para os quatro grupos experimentais (“Controle”, “Positivo”, “Tratamento_I” e “Tratamento_II”). Para cada grupo, são apresentados os valores mínimo (Min.), primeiro quartil (1st Qu.), mediana (*Median*), média (*Mean*), terceiro quartil (3rd Qu.) e valor máximo (*Max*.).

A segunda opção de comando, “stat.desc( )”, utiliza o pacote “pastecs” e fornece as seguintes informações: Número de valores, Número de valores nulos, Número de valores ausentes, Mínimo, Máximo, Intervalo mínimo-máximo, Soma, Mediana, Média, Erro padrão da média, Intervalo de confiança 95%, Variância, Desvio padrão e Coeficiente de variação (**Bloco 8**, arquivo suplementar 1).

O comando “skim( )”, do pacote “skimr”, possui a habilidade adicional de fornecer um histograma simples dos dados, além de outras informações relevantes, como: Número de linhas, Número de colunas, Tipo de dado nas colunas (ex.: numérico), Variáveis agrupadas (categóricos), Nome da variável, Valores faltando, Taxa de completude (indica ausência de valores), Média, Desvio padrão (sd), Percentil 0 (p0), Percentil 25 (p25), Percentil 50 (p50), Percentil 75 (p75), Percentil 100 (p100) e os Histogramas (**Bloco 9**, arquivo suplementar 1).

O pacote “psych”, por sua vez, possui o comando “describe( )”, que fornece dados importantes sobre o enviesamento e a curtose das distribuições contidas no conjunto de dados. São as informações geradas por “describe( )”: Variáveis, Número de observações, Média, Desvio padrão (sd), Mediana, Média truncada, Desvio mediano absoluto (mad), Mínimo, Máximo, Intervalo mínimo-máximo, Enviesamento (skew), Curtose (kurtosis) e Erro padrão da média (se) (**Bloco 9**, arquivo suplementar 1).

Os dados produzidos pelos comandos acima podem, ainda, ser exportados no formato .xlsx utilizando o pacote “xlsx”. Para isso, primeiro é necessário atribuir o comando a um objeto. Utilizando “describe(df)” como exemplo, tem-se: “tabela <- describe(df)”, onde tabela é o novo objeto contendo as informações geradas por “describe( )”. Depois, basta executar o comando “write.xlsx(x = tabela, file = ‘caminho_onde_o_arquivo_será_salvo.xlsx’)”. Não existe a necessidade, porém, de exportar todos os dados para um arquivo, já que eles podem ser visualizados ou copiados do console.

3.5 Geração e exportação de gráficos

Uma das principais vantagens do ambiente R é a capacidade produzir visualizações altamente personalizáveis e com qualidade de publicação. O ambiente nativo do R (base) é capaz de produzir gráficos a partir de um conjunto de dados com um simples comando "plot(df)" (lembre-se que "df" é o nome atribuído ao objeto que contém as informações do conjunto de dados do modelo). Apesar disso, o gráfico gerado pode não atender exatamente ao objetivo do pesquisador, tornando necessário o fornecimento de um código mais específico. No **Bloco 10** (arquivo suplementar 1), por exemplo, um conjunto de diagramas de caixa (*boxplots*) é gerado a partir de todas as variáveis do modelo (Figura 2).

O *boxplot*, também conhecido como diagrama de caixas, é uma representação gráfica utilizada para resumir e visualizar a distribuição de um conjunto de dados numéricos. Ele oferece uma visão clara da dispersão, da simetria e da presença de possíveis valores atípicos (*outliers*[7]) na distribuição. A estrutura do *boxplot* é composta por uma caixa retangular, que representa o intervalo interquartil (IQR), delimitado pelo primeiro quartil (Q1) e o terceiro quartil (Q3). A linha dentro da caixa indica a mediana, que é o valor central dos dados quando ordenados, dividindo o conjunto em duas metades.

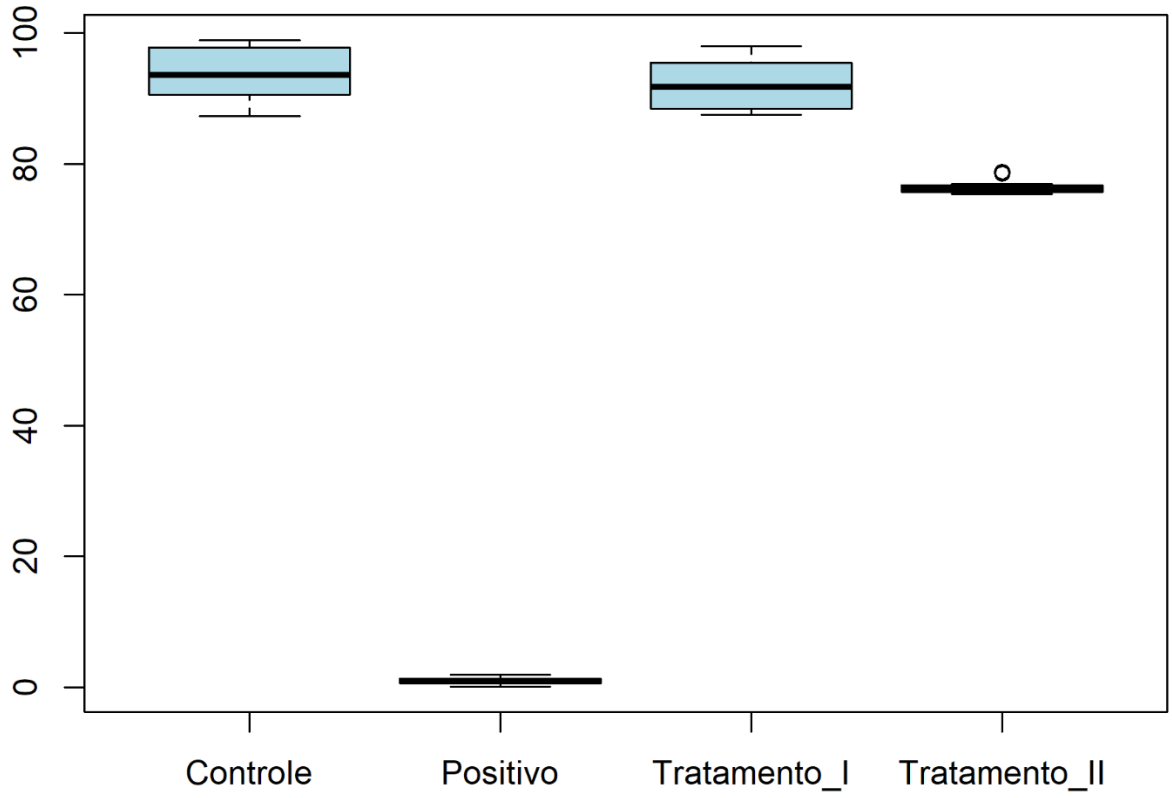
As "extensões" da caixa, conhecidas como *whiskers*, se estendem até o menor e o maior valor dentro de um intervalo específico, geralmente definido como 1,5 vezes o IQR a partir dos quartis. Valores fora desse intervalo são considerados *outliers* e são representados como pontos individuais além das extremidades dos whiskers.

Ao interpretar um *boxplot*, observa-se a posição da mediana dentro da caixa para avaliar a simetria da distribuição: uma mediana central sugere simetria, enquanto uma mediana próxima a Q1 ou Q3 indica assimetria. A altura da caixa, correspondente ao IQR, reflete a variabilidade dos dados; quanto maior a caixa, maior a dispersão dos valores em torno da mediana. *Outliers* identificados fora dos *whiskers* podem indicar a presença de valores atípicos que merecem atenção especial na análise. Dessa forma, o *boxplot* fornece

[7] Um valor atípico, ou *outlier*, é uma observação que se encontra consideravelmente distante da maioria dos dados em um conjunto, podendo ser identificado estatisticamente por diversos métodos. É crucial analisá-lo para determinar um certo valor se reflete variabilidade natural, erro de medição ou um fenômeno de interesse.

uma ferramenta eficaz para comparar distribuições entre diferentes grupos ou variáveis, oferecendo insights rápidos sobre a estrutura dos dados.

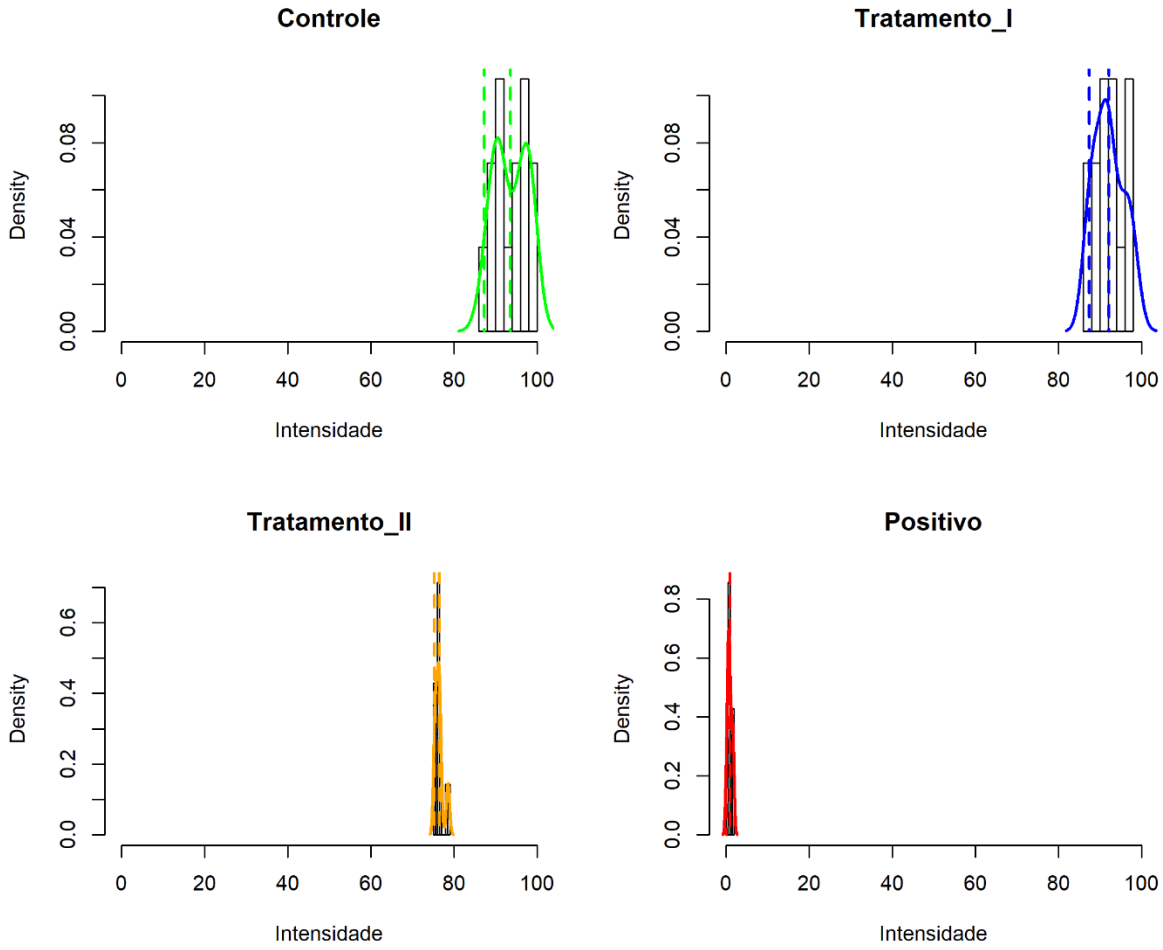
Figura 2 – Diagramas de caixa (*boxplots*) das variáveis do modelo



Fonte: Elaboração própria. **Legenda:** O gráfico apresenta diagramas de caixa (*boxplots*) comparando a variável "Intensidade" entre quatro grupos experimentais: “Controle”, “Positivo”, “Tratamento_I” e “Tratamento_II”. Cada *boxplot* (em azul) ilustra a distribuição dos dados por meio do intervalo interquartílico (caixa), a mediana (linha dentro da caixa), os valores mínimos e máximos (extensões ou "*whiskers*") e possíveis valores atípicos (pontos fora das extremidades).

Em R, também é possível gerar vários gráficos de forma simultânea. Esses gráficos serão organizados automaticamente na janela de visualização de acordo com a segmentação imposta pelo comando “par(mfrow = c(x, y))”, onde “x” é o número de linhas desejadas e “y” é o número de colunas. O **Bloco 11** (arquivo suplementar 1) apresenta um exemplo onde, para cada grupo do modelo, foi gerado um histograma sobreposto pela linha de densidade da distribuição, além de uma linha pontilhada indicando a média e outra, indicando o valor mínimo. Todos os gráficos foram apresentados simultaneamente em uma matriz do tipo 2x2 (Figura 3).

Figura 3 – Gráficos com sobreposição de elementos



Fonte: Elaboração própria. **Legenda:** A figura apresenta gráficos de densidade para a variável "Intensidade" em quatro grupos experimentais: “Controle”, “Positivo”, “Tratamento_I” e “Tratamento_II”. Cada painel exibe a distribuição de densidade da intensidade para o respectivo grupo, com linhas pontilhadas indicando o valor mínimo (linha à esquerda) e a média (linha à direita).

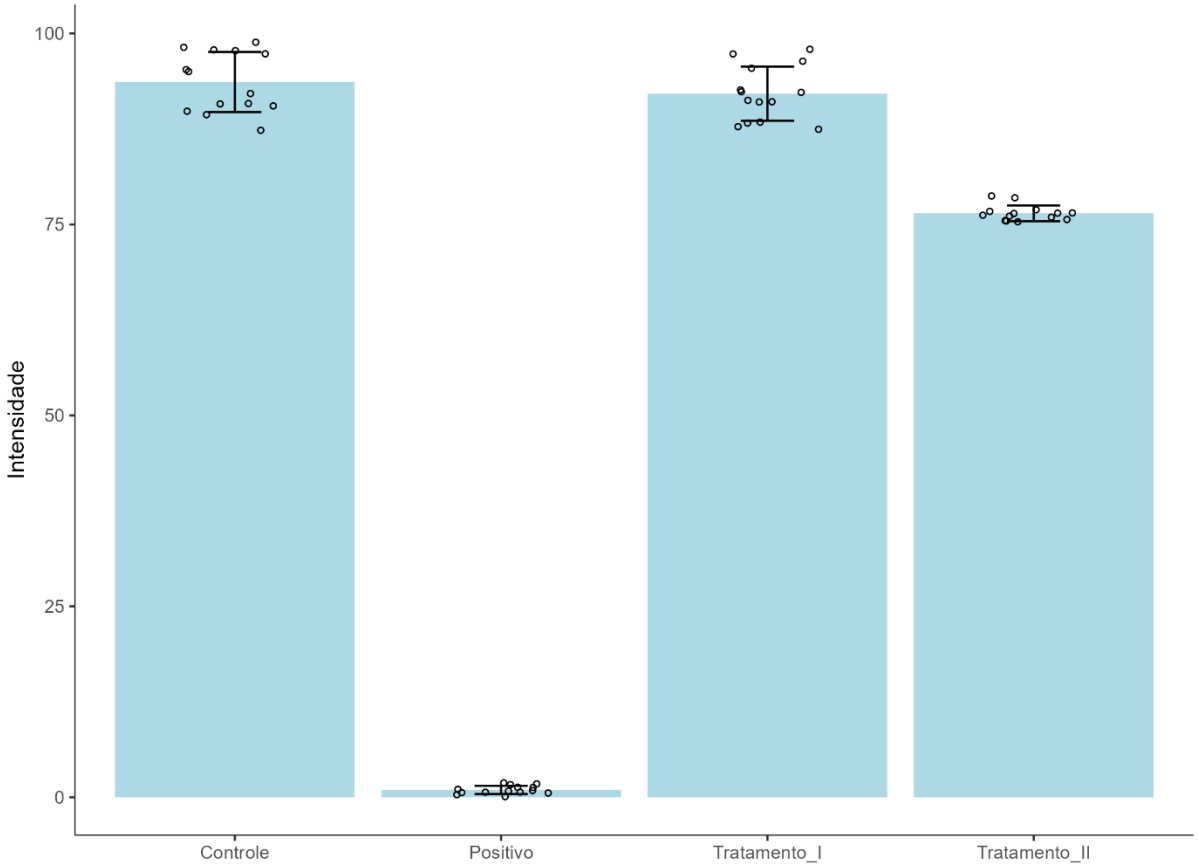
Apesar de bastante versátil, o pacote base do R possui limitações quando se trata de gráficos mais complexos. Para solucionar o problema, alguns pacotes podem ser importados para o ambiente, tornando-o uma das melhores ferramentas para geração de figuras científicas disponíveis na atualidade. O pacote mais utilizado no momento é o “ggplot2”, que constrói gráficos em um sistema de camadas, onde cada bloco de comando insere um elemento à figura.

A forma como o “ggplot2” interpreta os dados do modelo é um pouco diferente daquela utilizada nos gráficos de base R. Por isso, é necessário, em certos casos,

modificar a estrutura dos dados de maneira a convertê-la do formato “largo” (*wide*) para o formato “longo” (*long*). Essa transformação pode ser realizada com a função “melt( )” do pacote “reshape2”.

Depois de transformado, o modelo pode ser utilizado para a criação de um ggplot, que é o elemento básico de um gráfico com o pacote “ggplot2”. No **Bloco 12** (arquivo suplementar 1), calculou-se a média e desvio padrão de cada variável, que depois foram utilizados na geração de um ggplot (Figura 4) do tipo gráfico de barras (“geom_bar”), contendo os pontos individuais de cada grupo (“geom_jitter”) e as correspondentes barras de erro (“geom_errorbar”).

Figura 4 – Geração de gráfico de barras com ggplot2



Fonte: Elaboração própria. **Legenda:** O gráfico de barras apresenta a comparação da intensidade entre quatro grupos experimentais: “Controle”, “Positivo”, “Tratamento_I” e “Tratamento_II”. Cada barra azul representa a média da intensidade em cada grupo, com os desvios padrão indicados por linhas verticais sobre as barras, que revelam a média da medida de “Intensidade”. Os pontos sobrepostos às barras mostram os valores individuais dos dados, proporcionando uma visão da distribuição e variação dentro de cada grupo.

O próprio RStudio possui botão “Export” disponível em sua interface, facilitando a rápida obtenção de figuras, .pdfs e até de cópias de qualquer item apresentado na subjanela “Plots”. O mesmo procedimento pode ser feito para um ggplot, que também traz em seu pacote a possibilidade de personalizar as características da figura sendo salva (geralmente para qualidade de publicação). Exportar um gráfico do ggplot pode ser feito com o comando “ggsave(filename = ”nome_do_arquivo.tiff”, path = (‘caminho_onde_o_arquivo_será_salvo.xlsx’), dpi = 300)”, onde “.tiff” é o formato da figura (que poderia ser .pdf, .png, .jpeg ou outros), “path” é o local no computador onde a figura será salva e “dpi” (*dots per inch*) é a resolução da figura.

3.6 Valores extremos, normalidade e testes de hipótese

Valores extremos e *outliers* são observações que se destacam significativamente dos demais dados em um conjunto, localizando-se longe da maioria dos valores. Um valor extremo é um ponto de dados que, embora incomum, pode ainda assim ser parte da distribuição natural dos dados. Já um *outlier* é um valor que se afasta drasticamente das demais observações e pode ser resultado de diferentes fatores, como erros de medição, erros de entrada de dados, ou pode refletir variabilidade intrínseca ao processo em estudo.

A identificação de *outliers* é crucial na análise de dados, pois eles podem influenciar desproporcionalmente os resultados estatísticos, como a média, variância e correlação, levando a interpretações equivocadas. Por exemplo, em regressões lineares, a presença de *outliers* pode distorcer a inclinação da linha de melhor ajuste, afetando a precisão das previsões. Além disso, *outliers* podem indicar a presença de fenômenos subjacentes que merecem investigação adicional, como a existência de subgrupos distintos ou a ocorrência de eventos raros.

No entanto, nem todos os *outliers* devem ser removidos ou corrigidos sem uma análise cuidadosa. Em alguns casos, eles podem fornecer informações valiosas sobre o comportamento do sistema ou processo estudado. Portanto, na análise de dados, é

essencial identificar, investigar e interpretar corretamente os valores extremos e *outliers* para garantir que as conclusões sejam robustas e fiéis à realidade observada.

Um dos grandes problemas que permeiam as Biociências modernas é a escolha inadequada de testes de hipótese na comparação das variáveis de um determinado modelo. Isso provoca baixa reprodutibilidade e robustez nas conclusões produzidas pelos estudos. Uma das maneiras de contornar o problema usando o R é tirar proveito da vasta documentação e pacotes disponíveis, que permitem identificar a melhor abordagem possível na estimativa de tamanho amostral e poder estatístico, na identificação de valores extremos (e *outliers*) e na escolha de testes de hipótese. Infelizmente, muitos desses recursos só estão disponíveis em inglês, sendo pouco acessíveis ao usuário inexperiente. Contribuir para a solução desse problema é um dos objetivos da criação do repositório heRcules[8].

O pacote "rstatix", através do comando "identify_outliers( )", permite a identificação de *outliers* e valores extremos (conotações diferentes no comando). O procedimento é similar ao realizado no **Bloco 12** para a determinação de média e desvio padrão do modelo em formato *long* (**Bloco 13**, arquivo suplementar 1). Nota-se que, no presente modelo, só foram econtrados *outliers* no grupo "Tratamento_II" (Tabela 3, é possível vê-los no *boxplot* da Figura 2). Cabe ressaltar que o operador *pipe* (i.e., %>%), utilizado aqui, depende do pacote "dplyr", também importado no presente modelo (veja o Quadro 1).

Tabela 3 – Identificação de *outliers* e valores extremos em R

variable <fct>	**value** <dbl>	**is.outlier** <lgl>	**is.extreme** <lgl>
Tratamento_II	78.72481	TRUE	FALSE
Tratamento_II	78.48522	TRUE	FALSE
2 rows			

Fonte: Elaboração própria. **Legenda:** Identificação de valores atípicos e extremos no grupo "Tratamento_II". A tabela apresenta as variáveis analisadas, incluindo os valores observados ("*value*"), a classificação como valor atípico ("*is.outlier*") e como valor extremo ("*is.extreme*").

[8] https://github.com/drhrf/heRcules.git

A avaliação de normalidade também é outro componente importante para a realização de testes de hipótese que a assumem *a priori*. Uma possibilidade é a realização do teste de Shapiro-Wilk (Shapiro; Wilk, 1965), que no R usa o comando "shapiroTest( )", do pacote "fBasics". Nota-se, no **Bloco 14** (arquivo suplementar 1), que o Tratamento_II não aparenta ser distribuído normalmente no modelo (P = 0.02081). Cabe ao pesquisador tomar a decisão quanto à possibilidade de prosseguir ou não com os testes de hipótese após essa informação.

Avaliar a normalidade dos dados é um passo fundamental em muitas análises estatísticas, pois muitos testes estatísticos clássicos, como o teste *t*, ANOVA, e regressões lineares, assumem que os dados seguem uma distribuição normal. A normalidade implica que os dados são simetricamente distribuídos em torno da média, com a maioria das observações concentradas nas proximidades da média e a frequência das observações diminuindo à medida que se afastam da média.

Quando os dados são normalmente distribuídos, esses testes estatísticos podem ser aplicados com maior confiança, pois as inferências realizadas — como intervalos de confiança e valores-p — são válidas e mais precisas. A normalidade também facilita a interpretação dos resultados, uma vez que muitas propriedades estatísticas, como a centralidade da média e a simetria da variância, são mais bem compreendidas em uma distribuição normal.

No entanto, se os dados não seguem uma distribuição normal, a aplicação direta desses testes pode levar a resultados incorretos ou a interpretações errôneas, como a subestimação da variabilidade ou a superestimação da significância dos efeitos. Nessas situações, o pesquisador pode precisar transformar os dados, utilizar testes não paramétricos, ou adotar outras abordagens que não dependam da suposição de normalidade.

Existem diversas opções para a realização de análise de variância (ANOVA) em R. Uma delas é feita através da geração de um modelo ANOVA com o comando "aov( )", que depende do pacote "stats". O procedimento de criação do modelo segue o formato "aov(variável dependente ~ variável independente)", cujos resultados podem

ser apresentados utilizando o comando “summary( )”, tal como indicado no **Bloco 15** (arquivo suplementar 1).

É comum ser necessário, após a realização da ANOVA, a utilização de algum teste *post-hoc*. O teste de significância honesta de Tukey (Tukey, 1949) é um dos recursos disponíveis para essa demanda (é importante ressaltar que o pesquisador sempre deve se certificar da adequação do teste aos seus resultados experimentais). O **Bloco 16** (arquivo suplementar 1), abaixo, ilustra a utilização desse teste na resultante do ANOVA gerado pelo **Bloco 15**. O comando “TukeyHSD( )” depende do pacote “stats”.

Finalmente, alguns modelos experimentais podem exigir a comparação entre apenas dois grupos, situação onde um modelo de ANOVA não é apropriado. Nesse caso, é comum ser apropriada a realização de um teste da família de testes *t*. Abaixo, encontram-se exemplos de comparações múltiplas entre pares de grupos do modelo, ilustrando a utilização do teste *t* no ambiente do R (**Bloco 17**, arquivo suplementar 1). Tal como o teste de Tukey e a ANOVA, o comando t.test( ) depende do pacote “stats”.

Os testes ANOVA e *t* são amplamente utilizados em estatística para comparar médias entre grupos e determinar se há diferenças estatisticamente significativas. O teste *t* é usado para comparar a média entre dois grupos, enquanto o ANOVA (Análise de Variância) é utilizado quando se deseja comparar três ou mais grupos. Ambos os testes assumem que os dados são provenientes de distribuições normais e que as variâncias dos grupos são homogêneas (no caso do ANOVA). Além disso, os dados devem ser independentes e a amostragem aleatória.

Verificar as premissas de normalidade e homogeneidade de variância é crucial antes de aplicar esses testes, pois a violação dessas condições pode comprometer a validade dos resultados. Se os dados não forem normalmente distribuídos ou se as variâncias forem muito diferentes entre os grupos, os testes *t* e ANOVA podem levar a conclusões incorretas, como a rejeição ou aceitação indevida da hipótese nula. Testes de normalidade, como o de Shapiro-Wilk, e de homogeneidade de variâncias, como o de Levene, são comumente usados para verificar essas premissas.

Caso as premissas não sejam atendidas, existem alternativas mais robustas que podem ser utilizadas. Para dados não normais, pode-se aplicar testes não paramétricos, como o teste de Mann-Whitney para dois grupos ou o teste de Kruskal-Wallis para três ou mais grupos. Da mesma forma, se as variâncias forem heterogêneas, versões robustas do ANOVA, como o Welch's ANOVA, podem ser aplicadas. Dessa forma, a escolha do teste apropriado e o cumprimento das premissas garantem a validade e confiabilidade dos resultados, permitindo que as conclusões sobre as diferenças entre os grupos sejam precisas e fundamentadas.

4 CONCLUSÃO

O aumento no número de usuários que busca linguagens como R e Python para análise de dados científicos, em substituição a ferramentas mais limitadas ou pagas, é um sinal de progresso a favor da reprodutibilidade na ciência. O objetivo do presente trabalho, marcar o início de um esforço direcionado a fornecer recursos estruturados, e em língua portuguesa, a pesquisadores em processo de transição para a linguagem R, foi satisfeito através da criação de um repositório aberto e da apresentação de um modelo para a análise de dados em R. Esses recursos servirão o importante propósito de introduzir o uso de R para a análise de dados em ciências biológicas e da saúde.

REFERÊNCIAS

CHAMPELY, S. **pwr**: Basic Functions for Power Analysis. R package version 1.3-0, 2020. Disponível em: https://link.ufms.br/1gVny. Acesso em: 4 mar. 2004.

DEBASTIANI, V. J. **Introdução ao R**. [*S. l.*], 2020. Disponível em: https://link.ufms.br/jrVkK. Acesso em: 21 dez. 2021.

DRAGULESCU, A.; ARENDT, C. **xlsx**: Read, Write, Format Excel 2007 and Excel 97/2000/XP/2003 Files. R package version 0.6.5, 2020. Disponível em: https://link.ufms.br/50ihv. Acesso em: 4 mar. 2004.

GROSJEAN, P.; IBANEZ, F. **pastecs**: Package for Analysis of Space-Time Ecological Series. R package version 1.3.21, 2018. Disponível em: https://link.ufms.br/RC3TO. Acesso em: 4 mar. 2004.

KASSAMBARA, A. **rstatix**: Pipe-Friendly Framework for Basic Statistical Tests. R package version 0.7.0, 2021. Disponível em: https://link.ufms.br/aOTIi. Acesso em: 4 mar. 2004.

R CORE TEAM. **R**: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2013. Disponível em: https://link.ufms.br/U0dqv. Acesso em: 4 mar. 2004.

REVELLE, W. **psych**: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA, 2021. Versão 2.1.9. Disponível em: https://link.ufms.br/R179A. Acesso em: 4 mar. 2004.

SHAPIRO, A. S. S.; WILK, M. B. An Analysis of Variance Test for Normality (Complete Samples). **Biometrika**, v. 52, n. 3/4, p. 591–611, 1965. Disponível em: https://doi.org/10.2307/2333709. Acesso em: 4 mar. 2004.

TORCHIANO, M. **effsize**: Efficient Effect Size Computation. R package version 0.8.1, 2020. Disponível em: https://doi.org/10.5281/zenodo.1480624. Acesso em: 4 mar. 2004.

TUKEY, J. W. Comparing individual means in the analysis of variance. **Biometrics**, v. 5, n. 2, p. 99-114, 1949. Disponível em: https://doi.org/10.2307/3001913. Acesso em: 4 mar. 2004.

WARING, E.; QUINN, M.; MCNAMARA, A.; LA RUBIA, E. A.; ZHU, H.; ELLIS, S. **skimr**: Compact and Flexible Summaries of Data. R package version 2.1.3, 2021. Disponível em: https://link.ufms.br/g9Atv. Acesso em: 4 mar. 2004.

WICKHAM, H. Reshaping Data with the reshape Package. **Journal of Statistical Software**, v. 21, n. 12, p. 1-20, 2007. Disponível em: https://doi.org/10.18637/jss.v021.i12. Acesso em: 4 mar. 2004.

WICKHAM, H. **ggplot2**: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

WICKHAM, H.; FRANÇOIS, R.; HENRY, L.; MÜLLER, K. **dplyr**: A Grammar of Data Manipulation. R package version 1.0.7, 2021. Disponível em: https://link.ufms.br/udQwn. Acesso em: 4 mar. 2004.

WUERTZ, D.; SETZ, T.; CHALABI, Y. **fBasics**: Rmetrics - Markets and Basic Statistics. R package version 3042.89.1, 2020. Disponível em: https://link.ufms.br/HOaQj. Acesso em: 4 mar. 2004.

ZHU, H. **kableExtra**: Construct Complex Table with 'kable' and Pipe Syntax. R package version 1.3.4, 2021. Disponível em: https://link.ufms.br/UUuNg. Acesso em: 4 mar. 2004.

Sobre o autor

Hércules Rezende Freitas
Hércules é Biólogo e Matemático, com especializações em Fitoterapia, Farmacologia e Big Data. Também é Mestre e Doutor em Biofísica (UFRJ/Universidade de Coimbra), com pós-doutorado em Neuropatologia na Universidade da Califórnia. Atua como Bioestatístico para o Instituto Nacional de Traumatologia e Ortopedia e é professor universitário na Universidade do Grande Rio.
E-mail: hercules.freitas@bioqmed.ufrj.br

Submetido em 4 de setembro de 2024.
Aceito para publicação em 12 de dezembro de 2024.

Licença de acesso livre