SOFTWARE RELEASE

RunData: an easy and intuitive online tool for statistical analyses

Artur Guerra Rosa[1*] and Maiele Leandro da Silva[1]

Crop Breeding and Applied Biotechnology
20(3): e31802032, 2020
Brazilian Society of Plant Breeding.
Printed in Brazil
http://dx.doi.org/10.1590/1984-70332020v20n3s36

Abstract: *RunData is an online tool that requires internet access. This web application is built in R language, which performs multiple statistical procedures in an intuitive, and simple manner. This tool performs Shapiro–Wilk test, analysis of variance (ANOVA), means tests, and regressions. RunData is available at www.rundata.com.br and sistema.rundata.com.br/shiny/app/.*

Keywords*: R language, experimental agriculture, online*

INTRODUCTION

Statistical tests and methods play a vital role in agricultural research, providing different and reliable perspectives regarding the data collected. Currently, these tests are mostly performed by computer software; this is a major advancement in scientific research, providing results of countless necessary calculations rapidly without the possibility of human errors.

Among the tools that were created for the execution of statistical procedures, the R language excelled globally for being an easy-to-learn, free, and fully vectorized language. These advantages, combined with an active virtual community, have rendered R an ideal tool for the creation of various applications and packages with the most varied statistical functions (Rizzo 2019).

For every new software, users face challenges with regard to proper and efficient usage of the program during the learning process. This challenge is aggravated when users have no experience with another software in the intended area, or with the use of computers in general. Therefore, RunData aims to provide easy learning by offering the necessary support to users who wish to acquire such knowledge.

RunData was created to provide an online interface that performs statistical analyses in a user-friendly manner. The websites, which comes with this R Shiny application, provides more details about each functionality and creates a cooperative environment for easy learning. RunData is available at sistema.rundata.com.br/shiny/app/ and *www.rundata.com.br* (support website).

METHODS

RunData is a software built in R language that uses Shiny and R Markdown packages to enable the creation of a web application that performs multiple statistical procedures (Xie et al. 2018). As RunData is an online application, downloading or installation is not needed; only a computer and a browser with access to the Internet are required. However, taking all the available browsers

***Corresponding author:**
E-mail: arturguerra921@hotmail.com
ORCID: 0000-0002-5013-4408

Received: 07 April 2020
Accepted: 14 June 2020
Published: 20 July 2020

[1] Universidade Estadual de Mato Grosso do Sul, Unidade Universitária de Aquidauana, Rodovia Graziela Maciel Barroso, 79.200-00, Aquidauana, MS, Brazil

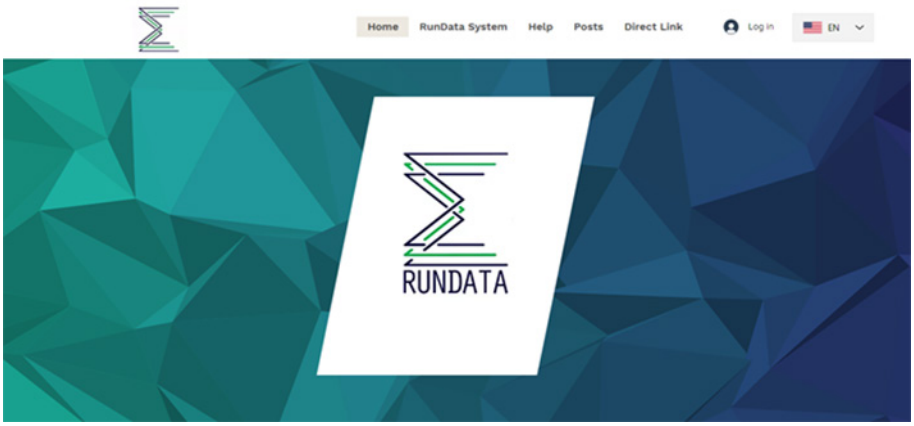

Figure 1. Rundata software homepage

into account, Mozilla Firefox (https://www.mozilla.org/en-US/firefox/new/) and Google Chrome (https://www.google.com/intl/pt-BR/chrome/) are the most recommended browsers.

The website (www.rundata.com.br) offers all the necessary support for new users; it contains a thorough guide that provides detailed information from organizing the data and loading the files to choosing the best statistical tests. The website offers a virtual community, through which users can comment, discuss problems, offer and receive suggestions, and review other related topics, thereby creating a comfortable and beneficial environment where any new user can learn how to use the program from without facing difficulties.

PERFORMANCE CHARACTERISTICS-AVAILABLE PROCEDURES

There are two main sections:

Parametric analysis

Once the user uploads the data, the application automatically performs all of the following parametric statistical analyses immediately.

Normality and data transformation

The normality assumption is checked using the Shapiro–Wilk test, which is different for each design and experimental scheme. There are several tests that checks the adjustment of data to enable normal distribution using different assumptions and algorithms. Among these tests, the Shapiro–Wilk and Shapiro–Francia tests exhibits the best performance (Torman et al. 2012, Le Boedec 2016). To ensure normality, the RunData application is equipped with a feature that performs the Shapiro–Wilk normality test with the function "shapiro.test()" from the package stats (R Core Team 2015) and interprets the test results; while showing whether the data are normal or not; if normality is rejected, the possibility of data transformation is offered. The following transformations are available: x^2, x^3, $\sqrt{x}$, $\sqrt{x+1}$, $\sqrt[3]{x}$, $ln(x)$, $log(x)$, and $arcsin(x)$.

Analysis of variance (ANOVA)

The RunData application performs ANOVA in different experimental designs and arrangements, such as complete randomized design (CRD), randomized blocks design (RBD), factorial with two factors on CRD or RBD, split-splot and split-split splot arrangement on CRD or RBD. The function "aov()" from the package stats (R Core Teams 2015) is used. The equations for each design are shown in Table 1.

Means tests

The RunData application performs Tukey (Tukey 1963), Duncan (Duncan 1955), SNK (Student 1927, Newman 1939, Keuls 1952), and Scheffé (Scheffé 1959) using the package "agricolae" (De Mendiburu 2019); for the Scott-Knott test (Scott and Knott 1974), the package "ScottKnott" (Jelihovschi et al. 2014) is used, and finally, the Dunnett's test (Dunnett

1955) uses the package "multcomp" (Hothorn et al. 2019). All results include key values and information that can be useful to the user, including grouping by letters, indicating which averages do not differ; multiple contrasts are shown for Dunnett's test.

Additionally, two models are available for download: simplified and detailed models. The simplified model uses several tables in a clean and intuitive way that makes it easier for the user to find and interpret data. The detailed model uses the package "ExpDes.pt" (Ferreira et al. 2018), delivering more information about the results in a less intuitive way.

Regression

It is imperative to understand and predict the performance of a dependent variable in comparison to one or more independent variables (Wilcox 2016). Supported models are linear, generalized linear models (GLM), and polynomials; these models are constantly updated, offering more options in the future. There is also the possibility of adding interactions in the independent variables for different seasons or any other qualitative variables.

Once a user uploads the data, the user must choose a desired model. Concurrently, a responsive graph is shown on the side toolbar, which changes to reflect the options being selected Furthermore, the regression summary is shown on the other tab of the toolbar, providing key information such as coefficients, R multiple squares, deviation residues, among others. For linear and polynomial models, the function *lm*() from package stats is used, and the GLM uses the function *glm*() also from the R Core Teams 2015 package stats. Table 2 shows all the models supported by the software.

Table 1. Equations for each experimental design and arrangement using the function aov()

Experimental Design	Equation
CRD	Variable ~ Treatments
RBD	Variable ~ Blocks + Treatments
Factorial (A × B) CRD	Variable ~ A + B + A*B
Factorial (A × B) RBD	Variable ~ Blocks + A + B + A*B
Split-plot CRD	Variable ~ A*B + Error (Replications: A)
Split-plot RBD	Variable ~ Blocks + A*B + Error (Blocks: A)
Split-split plot CRD	Variable ~ A*B*C + Error (Replications: A/B)
Split-split plot RBD	Variable ~ Blocks + A*B*C + Error (Blocks: A/B)

Table 2. Models used for regression with the functions lm() and glm()

Regression type	Equation inside function
Linear	Y ~ X
Linear	Y ~ I(X^1.5)
Linear	Y ~ I(X^1.5) + X
Quadratic	Y ~ I(X^2) + X
Square root	Y ~ I(X^1/2)
Logarithm	Y ~ I(log10(X))
Natural logarithm	Y ~ I(log(X))
GLM (Binomial)	Y ~ X
GLM (Poisson)	Y ~ X

The download button exports a Word document containing the graph, equation, regression summary, and a special table containing 100 values obtained from the equation, which can be copy-pasted into an Excel spreadsheet. This enables exact reproducibility of the same graph found in RunData within Excel, thereby providing access to all the tools offered by Excel. This creates numerous options for graph customization.

COMPARISON WITH OTHER SOFTWARE

To ensure reliability of the results, all data examples offered by RunData were tested with two commonly used software: Sisvar (Ferreira 2011) and Rbio (Bhering 2017). From the parametric section, the following parameters were compared: degrees of freedom, sum of squares, mean of squares, F-statistic, P-value, coefficient of variation, mean, and Shapiro–Wilk P-value. The values of RunData analysis were equal to those of Rbio for all the experimental designs, although with minor rounding differences. With regard to Sisvar, all values were equal, with the exception of the Shapiro–Wilk P-value. This discrepancy may have occurred because this software does not consider the experimental designs from the data, unlike RunData and Rbio, which utilize the residual data of ANOVA for the normality test.

From the regression section, the results of RunData were compared with those of the Rbio regression tab. The results showed that both programs had in common, the exact regression summary values for all models. However, the results could not be compared to Sisvar regression because this program uses ANOVA followed by regression, thus offering a different and viable method to perform regression analysis.

CONCLUSION

RunData offers a convenient set of statistical analyses to evaluate data from different experimental designs and arrangements, in an easy, intuitive, and clean design to students, professors, and professionals. The RunData support website shows a step-by-step process for each functionality, providing an easy and accessible cooperative learning environment.

REFERENCES

Bhering LL (2017) Rbio: A tool for biometric and statistical analysis using the R platform. **Crop Breeding and Applied Biotechnology 17**: 187-190.

De Mendiburu F (2019) Agricolae: Statistical procedures for agricultural research. R package version 1.3. Available at: <http://cran.r-project.org/package=agricolae>. Accessed on May 07, 2020.

Duncan DB (1955) Multiple range and multiple F tests. **Biometrics 11**: 1-42.

Dunnett CW (1955) A multiple comparison procedure for comparing several treatments with a control. **Journal of the American Statistical Association 50:** 1096-1121.

Ferreira DF (2011) Sisvar: A computer statistical analysis system. **Ciência e Agrotecnologia 35**: 1039-1042.

Ferreira EB, Cavalcanti PP and Nogueira DA (2018) ExpDes.pt: PackageExperimental Designs (Portuguese). R package version 1.2. Available at <https://cran.r-project.org/package=ExpDes.pt>. Accessed on May 07, 2020.

Hothorn T, Bretz F, Westfall P, Heiberger RM, Schuetzenmeister A and Scheibe S (2019) Multcomp: Simultaneous inference in general parametric models. R package version 1.4–10. Available at: <http://ftp5.gwdg.de/pub/misc/cran/web/packages/multcomp/multcomp.pdf>. Accessed on May 07, 2020.

Jelihovschi EG, Faria JC and Allaman IB (2014) ScottKnott: a package for performing the Scott-Knott clustering algorithm in R. **Trends in Applied and Computational Mathematics 15**: 3-17.

Keuls M (1952) The use of the "studentized range" in connection with an analysis of variance. **Euphytica 1**: 112-122.

Le Boedec K (2016) Sensitivity and specificity of normality tests and consequences on reference interval accuracy at small sample size: a computer-simulation study. **Veterinary Clinical Pathology 45**: 648-656.

Newman D (1939) The distribution of range in samples from a normal population, expressed in terms of an independent estimate of standart deviation. **Biometrika 31**: 20-30.

R Core Team (2015) The R Stats Package. R package version 3.5. Available at: <https://stat.ethz.ch/R-manual/R-devel/library/stats/html/stats-package.html>. Accessed on May 06, 2020.

Rizzo ML (2019) **Statistical computing with R**. CRC Press, Florida, 488p.

Scheffé H (1959) **The analysis of variance**. John Wiley & Sons, New York, 477p.

Scott AJ and Knott MA (1974) Cluster analysis method for grouping means in the analysis of variance. **Biometrics 30**: 507-512.

Student (1927) Errors of routine analysis. **Biometrika 19**: 151-164.

Torman VBL, Coster R and Riboldi J (2012) Normality of variables: diagnostic methods and comparison of some nonparametric tests by simulation. **Magazine Clinics Hospital of Porto Alegre 32**: 227-234.

Tukey JW (1963) **The problem of multiple comparisons**. Mimeograph Princeton University, Princeton, 396p.

Wilcox RR (2016) **Understanding and applying basic statistical methods using R**. John Wiley & Sons, New Jersey, 504p.

Xie Y, Allaire JJ and Grolemund G (2018) **R Markdown: The definitive guide**. CRC Press, Florida, 304p.