CROP BREEDING AND APPLIED BIOTECHNOLOGY

SOFTWARE/DEVICE RELEASE

Rbio: A tool for biometric and statistical analysis using the R platform

Leonardo Lopes Bhering[1*]

Crop Breeding and Applied Biotechnology
17: 187-190, 2017
Brazilian Society of Plant Breeding.
Printed in Brazil
http://dx.doi.org/10.1590/1984-70332017v17n2s29

Abstract: *Rbio is a free software for data processing. It is compatible and integrated with the free R software, which is globally accepted for statistical analysis. Thus, Rbio takes advantage of all processing potential of the R software. However, this new software allows the user to perform all the analyses without knowing the programming in R language. Rbio is available in Portuguese; it can be downloaded from the Internet (http://www.biometria.ufv.br), and requires the Windows operating system. It has a set of sample files, making it easy to be used. Currently, it supplies the user with all scripts internally used to process the analyses. Rbio can perform descriptive statistics, analysis of variance, estimation of genetic parameters and means tests, multivariate analysis, nonparametric tests, regressions, correlations, biometrics, bioinformatics, and simulation.*

Key words: *Software, programming, data analysis, quantitative genetics, biometrics.*

INTRODUCTION

Data processing and subsequent interpretation is a fundamental step in scientific research. Statistics allows summarizing and interpreting a huge set of information based on parameters such as mean and variance, making them applicable to any scientific area. Genetics and breeding present several particularities, which require the knowledge of biometrics and quantitative genetics in order to assist the decision-making process with less error as possible, and consequently maximizes the gain with selection of certain materials, without exhausting the available genetic variability.

Thus, integrating the knowledge on statistics and on biometric models applicable to genetics and breeding is necessary. Only a few software packages make this information available to the user, and allow simple and quick decision-making with the correct interpretation of results. Therefore, for genetics and breeding, obtaining additional genetic parameters is fundamental, such as heritability, intraclass correlation, genotypic variance, among other parameter that are not estimated by most statistical software. In this way, Rbio provides all this information to assist the correct decision-making by the breeder, without requiring the knowledge on the programming in R language.

The Rbio software output files are easy to use, since they present all the significances of the statistical tests, and indicate how to interpret the values. In addition, mean tests already show all letters indicating the differences between the means, which makes the Rbio an excellent software for users who have just started working with data analysis. The software is also useful for users who

***Corresponding author:**
E-mail: leonardo.bhering@ufv.br

Received: 23 November 2016
Accepted: 17 January 2017

[1] Universidade Federal de Viçosa (UFV), Departamento de Biologia Geral, Campus UFV, 36.570-900, Viçosa, MG, Brazil

already have good knowledge on statistics and biometrics, but do not have time to program in the R. Thus, the breeder can use the scripts available in Rbio by running the analyses in a quick and efficient way.

DESCRIPTION

The software Rbio is compatible with IBM PCs and requires the Windows operating system. The user must have the R software (R Development Core Team 2011) installed in the computer, which can be downloaded for free from the Internet (https://www.r-project.org/).

Currently, Rbio is available in Portuguese, with 19 sample files on how the data should be arranged in the file to be analyzed, and is easy to be downloaded due to its small size (2MB). The software installation is simple: after the download, the user needs to unzip the folder, click on the "setup" and follow the instructions, and install the _Rbio file in the C: drive of the computer.

AVAILABLE PROCEDURES

Rbio has several procedures for data processing. In the main screen, the user will find the following items in the menu:

Basic statistics

Initial step of analysis, used to describe, organize and summarize the collected data. In this menu, basic analyses can be performed:

a. Descriptive statistics of data (mean, standard deviation, mean standard error, variance, coefficient of variation, and scatter plots). All these measures are extremely important for data analysis, since they allow verifying the potential of the breeding material, and the presence of typing errors, by means of the maximum and minimum values for each variable.

b. Correlation. This procedure allows the simultaneous analysis of the association between variables by calculating the Pearson correlation between all pairs of variables and their significance test.

c. Regressions: simple linear, multiple and polynomial, response surface with 3D graphic, and Probit analysis. These procedures allow the identification of the behavior of a main dependent variable with one or more explanatory variables. In addition, the user can construct regression models for variables with different types of transformation.

d. Normality test: One of the assumptions of the analysis of variance is that the data have normal distribution. This procedure performs the Kolmogorov-Smirnov and Shapiro-Wilks tests to determine whether the data of a certain variable comes from a normal distribution. Multivariate normality test can also be carried out.

e. Simulation: Allows the sampling in the field in completely randomized design, randomized blocks, Latin Square, in addition to experimental arrangements in single and triple factorial, 2x2 and 3x3 lattices, and m augmented blocks. In this item, experiments in randomized block with the mean and variance of interest can also be performed.

f. Data transformation: This procedure evaluates the assumptions of the analysis of variance for variables and some transformations, in order to assist the user in identifying the need for transformation, and which transformation would be more suitable.

Analysis of Variance (ANOVA)

In this procedure, analyses of variances of the different experimental designs and arrangements can be carried out, such as: completely randomized, randomized blocks, factorial (AxB, AxBxC, AxBxCxD), simple and triple lattice, hierarchical, split-plot, and split-split plot, Latin square and augmented blocks. In addition, within each of the procedures, normality test (Shapiro-Wilk) and test for homogeneity of variances (Bartlett) can be carried out, with the option of showing or not the genetic parameters of the experiments. In addition, the F-test for each variation source is correctly displayed, after correction, when necessary, in case of random and mixed models. When the option "estimate genetic parameters" is checked, the parameters genetic variance, residual variance, block variance, variance of the GxE interaction, heritability, intraclass correlation, and coefficient of genetic variation are calculated.

For the aforementioned models, the following means tests can be processed: Tukey (Tukey 1953), SNK (Student 1927, Newman 1939, Keuls 1952), Scheffe (Scheffé 1959), Duncan (Duncan 1955), Dunnet (Dunnet 1955). The Scott-Knott clustering can also be performed (Scott and Knott 1974).

In this module, analysis of variance from unbalanced data can be performed in several experimental designs and arrangements. There is also the possibility of using the analysis via mixed models and analyses with qualitative and quantitative factors at the same time by performing the ANOVA of the data and mean test for the qualitative factors; and regression for the quantitative factors.

Multivariate statistics

Multivariate analyses use of all variables evaluated in the experiment, unlike ANOVA, which is univariate and analyzes each variable at once. Measures of distance and clustering of genetic material, principal component analysis, factor analysis, multivariate analysis of variance (Manova), canonical correlations and discriminant analysis are available to the user.

Non-parametric statistics

The non-parametric statistical module addresses variables without normal distribution. Friedman, Kruskall-Wallis, and Wilcoxon tests are available for this type of analysis for one or several samples, paired or not, besides the Spearman and Kendall correlation.

Biometrics

The biometrics module has one of the most widely used adaptability and stability methods in the world to investigate genotype x environment interactions, the Additive Main effects and Multiplicative Interaction - AMMI (Gauch 1988), which enables graphical projection of experiments carried out in different environments (years and/or sites).

Other possibilities for the research of genotypes x environment interaction are available, such as methodologies of environmental stratification and analysis of stability and adaptability.

Multicollinearity diagnosis of the variables under study, path analysis and canonical correlations can be carried out. Estimates of direct and indirect selection gains between variables can be obtained. Selection index methodologies for gain with several variables at the same time are also available.

The software also allows diallel analyses, which are useful at the early stages of the breeding program for parent identification.

In addition, matrices of phenotypic, genotypic and environmental correlations and partial correlation analyses can be obtained.

In the biometrics module, genomic selection analyses in several methods, such as rrBLUP, Bayesian lasso, and others are also available. In addition, molecular markers can be identified and selected.

Bioinformatics

This procedure allows reading DNA data in .fasta format. DNA sequence statistics, such as length, composition of bases, variation of GC content, among others, are available in the software, as well as graphs with these statistics.

Teaching

This module contains the procedures “Mean degree of dominance” and “Analysis of generations”, which allow professors and students of Basic Genetics, Populations Genetics, and Quantitative Genetics to obtain genetic parameters, such as allele frequency, mean and genotypic variance in function of the number of individuals.

USING THE PROGRAM

The user is given a set of data to be used as an example on how the data should be arranged in Rbio. In this way,

the user can see the example and arrange the data file according to the analysis to be performed. The only information required is whether the first line of the file to be analyzed presents information such as the name of the variables, or it constitutes the data set itself.

After selecting the input file, the program will name an output file in .txt format, with the same name of the input file, followed by the word "output". This file will be allocated in the same folder as the input file, making it easier for the user to keep the data together. Several procedures have some graphic files saved in .pdf.

Some procedures require additional R packages. If the user does not have them installed in the computer, there will be the "install package" button; by clicking on this button, Rbio will proceed with the correct installation. Once it has been completed, the user will click on "process", so that all analyses are performed. Alternatively, the user can go to the "Tools" menu, option "R: install packages", and download all packages necessary to perform the procedures available on Rbio.

The user can access the scripts used in the analyses after the processing by enabling the option "see script", which will allow saving or changing the script used. This saved file can be directly used in the R software, or in the option "Tools - run R scripts" available in Rbio. This makes Rbio a unique software, since it assists the the understanding of programming in R language.

CONCLUSION

Rbio is a software composed of several routines to be processed in the R software. However, it does not require the knowledge on the programming in R language. It has great diversity of procedures for analysis and processing of different genetic-statistical models applicable mainly to plant breeding. It is free and easy to use, and is suitable for universities and public and private companies, since it assists in the decision-making inherent to all statistical experiments.

ACKNOWLEDGEMENTS

The author gratefully acknowledges FAPEMIG, CAPES and CNPq for their financial support.

REFERENCES

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.

Gauch HG (1988) Model selection and validation for yield trials with interaction. **Biometrics 44**: 705-715.

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

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

R Development Core Team (2011) R: A language and environment for statistical computing. Vienna, Austria.

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

Scott AJ and Knott M (1974) A 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.

Tukey JW (1953) **The problem of multiple comparisons**. Mimeograph Princeton University, Princeton, 396p.