# Package in the R environment for analysis of variance and complementary analyses

## *Pacote em ambiente R para análise de variância e análises complementares*

**Emmanuel ARNHOLD**[1]

[1] Escola de Veterinária e Zootecnia, Campus Samambaia, Goiânia – GO, Brasil

**Abstract**

The objective was to create a package in the R environment, called "easyanova" to provide functions for performing analysis of variance of qualitative treatments with several test options for mean contrasts and residual analysis. The package comprises the functions "ea1()","ea2()" and "ec ()" and 19 data sets. The "ea1()" performs analyses in 13 designs and also the tests of Kruskal-Wallis and Friedman. The "ea2()" function performs analyses in experimental schemes with interaction (13 schemes). The "ec()" function tests mean group contrasts. The data and examples available in the package are associated mainly with animal research. Some of the possible analyzable designs are widely used in animal experimentation, such as Latin squares and split plot in time. In addition, the functions of the package are easy to use and make the necessary adjustments in cases of unbalanced experiments.

**Keywords:** Software. Statistics. Data analysis.

**Resumo**

Objetivou-se criar pacote em ambiente R, denominado "easyanova", a fim de disponibilizar funções para realizar análise de variância de tratamentos qualitativos, com várias opções de testes de contrastes de médias e análise de resíduos. O pacote dispõe das funções "ea1()", "ea2()" e "ec()" e 19 conjuntos de dados. A função "ea1()" realiza análises em 13 delineamentos e também os testes de Kruskal-Wallis e Friedman. A função "ea2()" realiza análises em esquemas experimentais com interação (13 esquemas). A função "ec()" testa contrastes de grupos de médias. Os dados e exemplos disponíveis no pacote são voltados principalmente à experimentação animal. Entre delineamentos possíveis de serem analisados têm-se alguns muito utilizados na experimentação animal, como os quadrados latinos e parcelas subdivididas no tempo. As funções do pacote também são de fácil utilização e fazem ajustes necessários nos casos de experimentos desbalanceados.

**Palavras-chave:** Software. Estatística. Análise de dados.

The R environment (R Core Team, 2013) was created in 1996 by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand. It was later developed by contributors in different parts of the world. Some of the advantages are the possibility of amplifying its functions due to easy programming and the underlying "package" system, consisting of supplements with specific functions, which greatly expands analysis capacity. A set of basic packages are installed with R, but many others are available. Currently, over 7000 packages are disposable for various fields of knowledge. An example of a package of more specific use is the pedigreemm (VAZQUEZ et al., 2010), which is used in animal breeding, based on the relationship matrix in generalized linear mixed models and, as an example of broader utility, the multcomp package (HOTHORN; BRETZ; WESTFALL, 2008) can be mentioned, with more general functions for testing contrasts in linear models. Thus, R has become an important technological tool in data analysis and manipulation, performing variance analysis, statistical hypothesis testing, mathematical operations, simulation, linear and nonlinear modeling, time series analysis, survival analysis, and multivariate analysis, among others, aside from the suitability for graphic representation.

**Correspondence to:**
Emmanuel Arnhold
Escola de Veterinária e Zootecnia, Campus Samambaia
Rodovia Goiânia – Nova Veneza, Caixa Postal 131
CEP 74001-970, Goiânia – GO, Brasil
e-mail: earnhold@pq.cnpq.br

Received: 23/05/2013
Approved: 18/12/2013

In the basic version and additional packages, the R environment provides functions to perform variance analysis of qualitative treatments, residue analysis and mean contrast tests. However, the available functions performing these analyses are complicated for users that are inexperienced in the R environment and in statistics. This is particularly true when the analyses involve more complex designs and experimental units are lost. Therefore, we developed the “easyanova” package, entitled “Analysis of variance and other important complementary analyses”. The objective was to provide functions in the R environment to perform analysis of variance of qualitative treatments in different designs (Table 1) and scheme of experimental treatments (Table 2), adjusting the means in case of experimental imbalance, making analysis very complete and simple to use.

The easyanova package is dedicated to the area of agricultural sciences, specifically for animal science. Of the 19 data sets available in the package (Tables 1 and 2), 14 were obtained from books and articles dealing with animal science (SANDERS; GAYNOR, 1987; KAPS; LAMBERSON, 2009; SAMPAIO, 2010) and five from papers in the areas of breeding, agronomy

Table 1 - Designs, their respective codes (design) of the function “ea1()” and available data sets for examples in the easyanova package (Goiânia, Goiás, Brazil – 2013)

| Design | design | Name of data set |
|---|---|---|
| Completely randomized | 1 | data1* |
| Randomized blocks | 2 | data2* |
| Latin squares | 3 | data3* |
| Multiple Latin squares | 4 | data4* |
| Analysis of covariance (completely randomized) | 5 | data10[,-3]* |
| Analysis of covariance (randomized blocks) | 6 | data10* |
| Incomplete blocks, types I and II | 7 | data11**; data12** |
| Incomplete blocks, type III | 8 | data13*** |
| Incomplete blocks, type III (experiments with animals) | 9 | data14**** |
| Lattice (intra-block analysis) | 10 | data15** |
| Lattice (inter-block analysis) | 11 | data15** |
| Rotating test in a completely randomized design | 12 | data16**** |
| Rotating test in a randomized block design | 13 | data17***** |
| Kruskal-Wallis test | 14 | data1* |
| Friedman test | 15 | data2* |

Data source: * Kaps and Lamberson (2009); ** Pimentel-Gomes and Garcia (2002); *** Cruz and Carneiro (2006); **** Sampaio (2010); ***** Sanders and Gaynor (1987).

Table 2 - Scheme of experimental treatment, the respective codes (design) of the function “ea2()” and available data sets for examples in the easyanova package (Goiânia, Goiás, Brazil – 2013)

| Scheme of experimental treatments | Design | Name of data set |
|---|---|---|
| Double factorial completely randomized | 1 | data5* |
| Double factorial in randomized blocks | 2 | data6** |
| Double factorial in Latin square | 3 | nd |
| Completely randomized split plots | 4 | data7* |
| Split plots in randomized blocks | 5 | data8* |
| Split plots in Latin square | 6 | nd |
| Triple factorial completely randomized | 7 | nd |
| Triple factorial in randomized blocks | 8 | nd |
| Double factorial in completely randomized split plots | 9 | nd |
| Double factorial in split plots in randomized blocks | 10 | nd |
| Combined analysis of randomized blocks (hierarchical experiment with blocks within the experiment) | 11 | data18*** |
| Combined analysis of Latin squares (hierarchical experiment with rows within the experiment) | 12 | data19**** |
| Combined analysis of randomized blocks (hierarchical experiment with rows and columns within the experiment) | 13 | data19**** |

Data source: * Kaps and Lamberson (2009); ** Pimentel-Gomes and Garcia (2002); *** Ramalho, Ferreira and Oliveira (2005); **** Sampaio (2010); nd No data available in the package.

and forestry (PIMENTEL-GOMES; GARCIA, 2002; RAMALHO; FERREIRA; OLIVEIRA, 2005; CRUZ; CARNEIRO, 2006). The data in this package can be loaded using the function "data()".

The package provides functions that perform analyses of the most common designs such as completely randomized and randomized blocks can also analyze the nonparametric options, respectively, the Kruskal-Wallis and Friedman tests. Other, more complex and widely-used designs in animal experimentation, such as Latin square, multiple Latin squares (with and without interaction between replications), analysis of covariance, rotating tests, factorial and split-split plot in time can also be analyzed with functions of the easyanova package (Tables 1 and 2).

As option of mean contrast tests, the functions of the package perform multiple mean comparisons by the Tukey, SNK, Duncan, t, and Scott-Knott tests. Contrasts of groups of means can also be evaluated. For residue analysis, functions provide tests for normality and homogeneity of residual variance, coefficient of variation, and the most discrepant values that can be evaluated in residual plots.

The R package named ExpDes (FERREIRA; CAVALCANTI; NOGUEIRA, 2013) focuses on the analysis of variance of data from qualitative and quantitative approaches. It has been widely used due to ease of use and to perform more comprehensive analyses than those available in the basic R installation. However, the functions of the ExpDes package cannot analyze experiments in Latin squares nor perform nonparametric tests alternatively to analysis of variance or analysis of covariance and other important designs. Data with any kind of imbalance cannot be analyzed either. The ExpDes performs only one test of residual normality, calculates the coefficient of variation, and cannot automatically analyze the different response variables. A special advantage of ExpDes is the possibility of regression analysis (polynomial regression) for quantitative treatments.

To perform the analyses cited in the preceding paragraphs, the easyanova package has three functions, called "ea1()", "ea2()" and "ec()". The "ea1()" function performs analysis of various designs, without considering interaction between factors (Table 1). The function "ea2()" performs analysis of experimental treatment schemes, evaluating interaction between factors. And the "ec()" function tests means contrasts, preferably groups of averages. The following is a brief description of the three functions.

The "ea1()" function has six arguments "ea1(data, design=1, alpha=0.05, list=FALSE, p.adjust=1, plot=2)". The argument "data" should contain the name of the data set in the class of R objects called "data.frame". The argument "design" which by default is equal to "1" (completely randomized design) determines the design (Table 1). The argument "alpha" is the standard significance level of 0.05 and "list" provides the opportunity to examine, in a single time, several response variables. This is done by simply changing the standard from "list=FALSE" to "list=TRUE". The argument "p.adjust" offers the possibility of adjusting probability values of the t test and the standard = 1 without any adjustment. The setting options are those of the function "p.adjust()", available in the basic R installation, and this function is used internally in the function "ea1()". The argument "plot" generates residual plots; option 1 generates a "boxplot" of the residuals, option 2 (standard) a plot of the standardized residuals in functions of the data (observing the z scores) and option 3, a plot of the standardized residuals in function of the theoretical quantiles under normality.

The "ea2()" function has seven arguments "ea2 (data, design=1, alpha=0.05, cov=4, list=FALSE, p.adjust=1, plot=2)". Except for argument "cov", the same arguments are assigned as to the function "ea1()", discussed above. The argument "cov" specifies the structure of the variance and covariance matrix that can be used in a split-plot scheme when the treatment is applied to the split plots and time ("designs" = 4, 5, 6, 9, and 10). The option of covariance “1” defines an autoregressive matrix, “2” an autoregressive matrix with heterogeneity of variances, “3” a continuous autoregressive matrix, “4” (standard function) a symmetric composite matrix and “5” defines an unstructured matrix. The best structure to be used can be identified by comparing the value of AIC (Akaike information criterion) or the BIC (Bayesian Information Criterion). The function returns values of AIC and BIC and the lower the AIC or BIC values; the more appropriate is the structure of variance and covariance.

Fixed models are used in most experimental designs and layouts. The exception is for incomplete block designs with recovery of interblock information (“design = 8”). In this model, the block effect is considered random. In split-split plot schemes (“designs = 4, 5, 6, 9, and 10”), where the effect of “subject” is considered random, the model is mixed.

Both functions “ea1()” and “ea2()” perform the analysis  of variance with type I sum of squares (sequential) or type III (partial), as defined and recommended by Wechsler (1998). They also produce means with standard errors of the factor levels (adjusted means in the case of covariance analysis or experiments with some form of imbalanced effects), mean tests of Tukey, SNK, Duncan, t, and Scott-Knott tests and residue analysis with normality test of the residuals (Shapiro-Wilk test), homogeneity of variance of residuals (Bartlett), and coefficient of variation (fixed model), AIC and BIC (mixed model). Finally, they identify the first, second and third most discrepant residue that may be less accurate than evaluated in boxplots, z scores or compared with theoretical quantiles considering normality.

For the input data one must consider the order of columns. In the completely randomized design, the first column must contain the treatment codes and the remaining columns the response variables. For the randomized block design, the first column must contain the treatment codes, the second the block codes and the others, the response variables. Examples with data for several designs or treatment schemes are available in the package. These data can exemplify the appropriate form of entry. In all cases, the first column refers to the effect of the explanatory variables and the others should contain response variables. The information on the number of a particular design can be found in the support files of the functions “ea1()” and “ea2()” (Tables 1 and 2).

The “ec()” function was created to test contrasts of treatment means, preferably groups of means, because the contrast tests applied in the functions “ea1()” and “ea2()” are contrasts of pairs of means (multiple comparisons). The function has five arguments “c (mg1, mg2, sdg1, sdg2, gl)”, where “mg1” and “sdg1” indicate the mean (means) and standard error (standard errors) of group 1 and “mg2” and “sdg2” refer to the mean (means) and standard error (standard errors) of group 2. The argument “gl” indicates the number of degrees of freedom of the residual analysis of variance. The function returns the contrast estimate, its variance and the probability value by the t test.

From the foregoing, it can be concluded that the functions of the easyanova package are excellent computer tools for the analysis of data from various designs and schemes of qualitative treatments. Its functions are easy to use, produce complete results and are particularly useful in the case of unbalanced data.

## References

CRUZ, C. D.; CARNEIRO, P. C. S. **Modelos biométricos aplicados ao melhoramento genético**. Viçosa-MG: Universidade Federal de Viçosa, 2006. v. 2, 585 p.

FERREIRA, E. B.; CAVALCANTI, P. P.; NOGUEIRA, D. A. **ExpDes:** experimental designs pacakge. R package version 1.1.2. 2013. Disponível em: <http://CRAN.R-project.org/package=ExpDes>. Acesso em: 17 out. 2013.

HOTHORN, T.; BRETZ, F.; WESTFALL, P. Simultaneous inference in general parametric models. **Biometrical Journal**, v. 50, n. 3, p. 346-363, 2008.

KAPS, M.; LAMBERSON, W. R. **Biostatistics for animal science:** an introductory text. Wallingford: CABI, 2009. 504 p.

PIMENTEL-GOMES, F.; GARCIA, C. H. **Estatística aplicada a experimentos agronômicos e florestais: exposição com exemplos e orientações para uso de aplicativos**. Piracicaba: Escola Superior de Agricultura "Luiz de Queiroz", 2002. v. 11, 309 p.

R CORE TEAM. **R**: a language and environment for statistical computing. Vienna: R foundation for statistical computing, 2013. Disponível em: <http://www.R-project.org/>. Acesso em: 17 out. 2013.

RAMALHO, M. A. P.; FERREIRA, D. F.; OLIVEIRA, A. C. **Experimentação em genética e melhoramento de plantas**. Lavras: UFLA, 2005. 322 p.

SAMPAIO, I. B. M. **Estatística aplicada à experimentação animal**. Belo Horizonte: Fundação de Ensino e Pesquisa em Medicina Veterinária e Zootecnia, 2010. 264 p.

SANDERS, W. L.; GAYNOR, P. J. Analysis of switchback data using statistical analysis system. **Journal of Dairy Science**, v. 70, n. 10, p. 2186-2191, 1987.

VAZQUEZ, A. I.; BATES, D. M.; ROSA, G. J. M.; GIANOLA, D.; WEIGEL, K. A. Technical note: an R package for fitting generalized linear mixed models in animal breeding. **Journal of Animal Science**, v. 88, n. 2, p. 497-504, 2010.

WECHSLER, F. S. Fatoriais fixos desbalanceados: uma análise mal compreendida. **Pesquisa Agropecuária Brasileira**, v. 33, n. 3, p. 231-262, 1998.