

ExpDes: An R Package for ANOVA and Experimental Designs

Eric B. Ferreira¹, Pórtia P. Cavalcanti², Denismar A. Nogueira¹

¹Exact Sciences Department, Federal University of Alfenas, Alfenas, Brazil

²Faculty of São Lourenço, São Lourenço, Brazil

Email: eric.ferreira@unifal-mg.edu.br

Received 11 September 2014; revised 6 October 2014; accepted 16 October 2014

Copyright © 2014 by authors and Scientific Research Publishing Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

Abstract

Analysis of variance (ANOVA) is a usual way for analysing experiments. However, depending on the design and/or the analysis scheme, it can be a hard task. ExpDes, acronym for *Experimental Designs*, is a package that intends to turn such task easier. Devoted to fixed models and balanced experiments (no missing data), ExpDes allows user to deal with additional treatments in a single run, several experiment designs and exhibits standard and easy-to-interpret outputs. It was developed at the Exact Sciences Institute of the Federal University of Alfenas, Brazil. Stable versions of package ExpDes are available on CRAN (Comprehensive R Archive Network) since 2012. Based on users' feedback, the package was used to illustrate graduation and post-graduation classes and to carry out data analysis, in Brazil and many other countries. Package ExpDes differs from the other R tools in its easiness in use and cleanliness of output.

Keywords

ANOVA, Computer Package, R Software, Open Source

1. Introduction

An *experiment* is a planned inquiry to obtain new facts or to confirm or deny the results of previous experiments, where such inquiry will aid in a decision [1]. Usually, statistical experiments are conducted in situations in which researchers can manipulate the conditions of the experiment and can control the factors that are irrelevant to the research objectives.

According to [2], *experimental design* refers to a plan for assigning subjects to experimental conditions and the statistical analysis associated with the plan. The design of an experiment to investigate a scientific or research hypothesis involves a number of integrated activities:

- Formulation of statistical hypotheses that are germane to the scientific hypothesis;
- Determination of the experimental conditions (independent variable) to be used, the measurement (dependent variable) to be recorded, and the extraneous conditions (nuisance variables) that must be controlled;
- Specifications of the number of subjects (experimental units) required and the population from which they will be sampled;
- Specification of the procedure for assigning the subjects to the experimental conditions;
- Determination of the statistical analysis to be performed.

In short, an experimental design identifies the independent, dependent, and nuisance variables and indicates the way in which the randomization and statistical aspects of an experiment are to be carried out.

The most common experimental plans are: Completely Randomized Design (CRD), Randomized Blocks Design (RBD) and Latin Squares Design (LSD), while the most popular schemes are: single factor, factorial and split-plot experiments [3].

The analysis of variance was introduced by Sir Ronald A. Fisher and is essentially an arithmetic process for partitioning a total sum of squares into components associated with recognized sources of variation. It has been used to advantage in all fields of research where data are measured quantitatively [1].

According to [4], the two main aims of classical ANOVA are:

- to examine the relative contribution of different sources of variation (factors or combination of factors, *i.e.* the predictor variables) to the total amount of the variability in the response variable, and;
- to test the null hypothesis (H_0) that population group or treatment means are equal.

That is, according to [5], given a certain number of treatments (say I treatments) and denoting their means by $\theta_1, \theta_2, \dots, \theta_I$, the hypothesis are

$$H_0 : \theta_1 = \theta_2 = \dots = \theta_I$$

$$H_1 : \theta_i \neq \theta_j, \text{ for some } i \neq j$$

Nevertheless, for the ANOVA's F test to be valid, three basic assumptions are required [6]:

- 1) Errors are independent and normally distributed;
- 2) Errors present homogeneity of variance; and
- 3) Additivity of terms of the model.

For instance, for the completely randomized design, one assumes that

$$y_{ij} = \theta_i + e_{ij}$$

and

$$e_{ij} \stackrel{iid}{\sim} N(0, \sigma_e^2)$$

where y_{ij} is the j -th replication of the i -th treatment; $\theta_i = \theta + \tau_i$ is the mean of the i -th treatment; e_{ij} is the random error associated to y_{ij} and σ_e^2 the common error variance.

If the omnibus hypothesis of equality of means is rejected, a researcher is still faced with the problem of deciding which of the means are not equal. A significant F test indicates that something has happened in an experiment that has a small probability of happening by chance [2]. Thus, according to [1] [2] [4] [7], when treatments are qualitative, several proposals of contrasts among means are available in literature, such as the test of Tukey, Duncan, Student-Newman-Keuls, t, Bonferroni, Dunn, Dunn-Šidák, Dunnett, Gomes-Howell, Fisher-Hayter, Scheffé and others.

On the other hand, when treatments are quantitative, linear regression analysis can be applied in ANOVA [5]. According to [2], a linear model consists of two parts: a model equation and associated assumptions. The model equation relates the observable random variable to underlying parameters and random variables in a linear manner. The assumptions specify the nature of the random components and any restrictions that the parameters must satisfy. Consider the multiple regression model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_{h-1} x_{i,h-1} + e_{-i}, \quad (i = 1, 2, \dots, I).$$

where y_i is the value of the dependent variable for experimental unit i ; $\beta_0, \dots, \beta_{h-1}$ are unknown parameters; $x_{i1}, \dots, x_{i,h-1}$ are independent variables; e_i is a random error term with mean equal to zero and variance equal to σ_e^2 ; e_i and $e_{i'}$ are uncorrelated for all $i \neq i'$.

Several features of the model deserve further comment:

1) The observed value of y_i is the sum of two components: the constant predictor term, $\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_{h-1} x_{i,h-1}$, plus the random error term, e_i . The error term reflects that portion of y_i that is not accounted for by the independent variables. Because the error term is a random variable, y_i is also a random variable.

2) The expected value of the error term equals zero, $E(e_i) = 0$; it follows that the expected value of y_i is equal to the constant predictor term.

3) If the independent variables are quantitative, the unknown parameters can be interpreted as follows: The parameter β_0 is the Y intercept of the regression line. The $\beta_1, \dots, \beta_{h-1}$ are partial regression coefficients, or weights applied to the x_{ij} to optimally predict y_i . The parameter β_1 , for example, indicates the change in the mean response per unit increase in x_{i1} when $x_{i2}, \dots, x_{i,h-1}$ are held constant.

4) When there is only one independent variable, the model is a simple regression model; when there are two or more independent variables, it is a multiple regression model.

5) The error term, e_i , is assumed to have constant variance σ_e^2 . It follows that the variance of y_i is $V(y_i) = V(e_i) = \sigma_e^2$.

6) The error terms are assumed to be uncorrelated. The value of e_i is not related to the value of $e_{i'}$ for all $i \neq i'$. Because the e_i 's are uncorrelated, the y_i 's are also uncorrelated.

When planning a factorial experiment, it is often desirable to include certain extra treatments falling outside the usual factorial scheme. Reference [8] discusses the exact analysis of an experiment of this type. The author points out that, since the additional treatment is randomized in with others, one should perform the standard ANOVA considering all treatments (including the additional) without the factorial structure; and then perform the factorial analysis (without the additional treatment). Such procedure enables to contrast the additional treatment with the factorial experiment and decide if their means differ through F test.

2. R Packages to Analysis Experiments

The analysis of experimental designs already can be performed in R using some specific packages. First of all, we have the basic package `stats`, that contains standard (general) functions for analyzing data from designed experiments, such as `lm()` and `aov()`. Package `stats` also has a few functions for get and set contrast matrices, for multiple comparison and some convenience functions like `model.tables()`, `replications()` and `plot.design()` [9]. However, depending on the design and the complexity of the experiment, the analysis can be a hard task.

In this topic, we briefly describe some contributed packages for the same purpose.

AlgDesign: Algorithmic experimental Designs. According to [10], with package `ALgDesign` we may evaluate a design, blocked or not, using `eval.blockdesign()` and `eval.design()` functions respectively. It also enables to generate a full factorial design using function `gen.factorial()` and create a candidate list of mixture variables with `gen.mixture()` function. This package also allows to block experimental designs using various criteria with function `optBlock()`, calculate an exact or approximate algorithmic design for one of three criteria (D, A and I), using Federov's exchange algorithm (function `optFederov()`), and find a design using the specified criterion via Federov's algorithm applied to a random subset of all possible candidate points (function `optMonteCarlo()`).

dae: Design and ANOVA of Experiments. The package `dae` [11] provides several tools on experimental design and R factors. It includes a group of functions that aid to generate experimental designs, as `fac.nested()` that can handle various nested structures and functions for combining several factors into one (function `fac.combine()`) or dividing one factor into several factors (function `fac.divide()`). It also includes functions that facilitate diagnostic checking after an `aov`, like function `fitted.aovlist()` that extracts the fitted values for all the fitted terms in the model and `residuals.aovlist()` that extracts the residuals in the analysis.

DoE.base: Package `DoE.base` [12] supplies mainly full factorial designs blocked or not and orthogonal arrays for (industrial) experiments. Some functions that allows such analysis are function `fac.design()` that creates full factorial designs with arbitrary numbers of levels, and potentially with blocking, and function `oa.design()` that access orthogonal arrays, allowing limited optimal allocation of columns.

experiment: Package `experiment`, according to [13], offers some functions for designing and analysing

randomized experiments. For instance, function `randomize()` allows to randomize the treatment assignment for complete randomized experiments, randomized-block experiments and matched-pair designs. Function `ATEcluster()` allows estimating various average treatment effect in cluster-randomized experiments without using pre-treatment covariates. Function `CACEcluster()` allows to estimate various complier average causal effect in cluster-randomized experiments without using pre-treatment covariates when unit-level noncompliance exists. Function `NoncompLI()` allows to estimate the average causal effects for randomized experiments with noncompliance and missing outcomes under the assumption of latent ignorability.

GAD: General ANOVA Designs. According to [14], package GAD handles general balanced analysis of variance models with fixed/random effects and orthogonal/nested effects (the latter can only be random). It contains functions that perform a Cochran's test of the null hypothesis that the largest variance in several sampled variances are the same (function `C.test()`), construct the mean squares estimates of an ANOVA design (function `estimates()`), considering the complications imposed by nested/orthogonal and fixed/random factors, fit a general ANOVA design (function `gad()`) and function `snk.test()` performs a SNK post-hoc test of means on the factors of a chosen term of the model.

3. Package Features

The package `ExpDes` [15] has as main objective the analysis of balanced experiments under fixed models. A special feature of this package is to allow the complete analysis in a single run. Its output summarizes what the researcher will need to interpret the most likely hypothesis to be evaluated.

`ExpDes` was created in 2009 and since then is used in classes at the Federal University of Alfenas, Brazil. During almost three years, its Portuguese and English versions were expanded, debugged and distributed on the website <https://sites.google.com/site/ericbferreira/>. Along 2011 it was released in Brazil. The first seminar took place at the Federal University of Ouro Preto, Brazil.

The main purpose of the package `ExpDes` is to analyze simple experiments under completely randomized designs (`crd()`), randomized block designs (`cbd()`) and Latin square designs (`latsd()`). Also enables the analysis of treatments in a factorial design with 2 and 3 factors (`fat2.crd()`, `fat3.crd()`, `fat2.rbd()`, `fat3.rbd()`) and also the analysis of split-plot designs (`split2.crd()`, `split2.rbd()`).

Other functionality is analyzing experiments with one additional treatments on completely randomized design and randomized blocks design with 2 or 3 factors (`fat2.ad.crd()`, `fat2.ad.rbd()`, `fat3.ad.crd()` and `fat3.ad.rbd()`).

After loading the package and reading and attaching the data, a single command is required to analyze any situation. For instance, consider a double factorial scheme under a completely randomized design plus an additional treatment:

```
fat2.ad.crd(factor1, factor2, repet, resp, respAd, quali=c(T,F), mcomp="tukey",
fac.names=c("F1", "F2"), sigT=.05, sigF=.05)
```

Besides both factors (`factor 1` and `factor 2`), repetitions (`repet`), response variable (`resp`) and response from the additional treatment (`respAd`), one must inform whether each factor is qualitative or not (`quali`), the desired multiple comparison test (`mcomp`)—to be used only for significant qualitative factors, the factor names to be used along the output report (`fac.names`) and the desired significance considered for the multiple comparison test (`sigT`, default is 5%) and F test (`sigF`, default is 5%).

The first information on the output is a legend of the factors followed by the conventional analysis of variance table (Figure 1).

Information on normality of residuals—according to the Shapiro-Wilk test [16]—are provided for almost all designs. For the specific case of experiments with additional treatment, is also displayed the contrast between such treatment and the whole factorial (Figure 2).

Considering the pre-specified significance level, the functions consider the interactions significant or not, and analyse them or the single factors (Figure 3).

As illustrated in Figure 4, given the rejection of the null hypothesis, the comparison of the treatment means becomes necessary. For situations of qualitative treatments there are several options of multiple comparison tests available for the argument `mcomp`, such as Tukey ("`tukey`", default), Scott-Knott ("`sk`"), Duncan ("`Dun-`

```

-----
Legend:
FACTOR 1:  F1
FACTOR 2:  F2
-----

Analysis of Variance Table
-----
              DF      SS      MS      Fc  Pr>Fc
F1              1 0.13054 0.13054 46.4363    0
F2              3 0.01575 0.00525  1.8671 0.1714
F1*F2           3 0.02808 0.00936  3.3295 0.0429
Ad vs Factorial 1 0.07370 0.07370 26.2189 1e-04
Residuals      18 0.05060 0.00281
Total          26 0.29867 0.01149
-----

```

Figure 1. Legend and analysis variance table for a double factorial scheme under completely randomized design with an additional treatment.

```

-----
Shapiro-Wilk normality test
p-value: 0.02718233
WARNING: at 5% of significance, residuals can not
be considered normal!
-----
Contrast of the additional treatment with the factorial
-----
              Means
Additional 0.11000 a
Factorial 0.27625 b
-----

```

Figure 2. Shapiro-Wilk normality test and contrast between the addition treatment and the whole factorial.

```

-----
Significant interaction: analyzing the interaction
-----

Analyzing  F1  inside of each level of  F2
-----

Analysis of Variance Table
-----
              DF      SS      MS      Fc  Pr>Fc
F2              3 0.01575 0.00525  1.8671 0.1714
F2:F1 5         1 0.02042 0.02042  7.2628 0.0148
F2:F1 10        1 0.09627 0.09627 34.2451    0
F2:F1 15        1 0.03527 0.03527 12.5455 0.0023
F2:F1 20        1 0.00667 0.00667  2.3715 0.141
Ad vs Factorial 1 0.07370 0.07370 26.2189 1e-04
Residuals      18 0.05060 0.00281
Total          26 0.29867 0.01149
-----

```

Figure 3. Analysis of the significant interaction.

```

-----
F1  inside of the level  5  of  F2
-----
Tukey's test
-----
Groups Treatments Means
a         2         0.3133333
b         1         0.1966667
-----

```

Figure 4. Tukey's multiple comparison test for qualitative factors.

can"), Student-Newman-Keuls ("snk"), Student's t test ("lsd"), Bonferroni ("lsdb") and a Bootstrap multiple comparison test ("ccboot").

```

-----
$`Cubic Model
-----
      Estimate Standard.Error      tc p.value
b0 -0.36666667      0.25427 -1.44201 0.16647
b1  0.21111111      0.07778  2.71435 0.01421
b2 -0.01706667      0.00687 -2.48345 0.02309
b3  0.00040889      0.00018  2.24013 0.03794

$`R2 of cubic model`
[1] 1

$`Analysis of Variance of cubic model`
      DF      SS      MS      Fc p.value
Linear Effect      1 0.00323 0.00323 1.15 0.29816
Quadratic Effect   1 0.02253 0.02253 8.02 0.01107
Cubic Effect       1 0.01411 0.01411 5.02 0.03794
Lack of fit        0 0.00000 0.00000  0      1
Residuals          18 0.05060 0.00281

-----
$`Quadratic Model
-----
      Estimate Standard.Error      tc p.value
b0  0.17000000      0.08522  1.99489 0.06142
b1  0.04040000      0.01555  2.59832 0.01816
b2 -0.00173333      0.00061 -2.83122 0.01107

$`R2 of quadratic model`
[1] 0.6460998

$`Analysis of Variance of quadratic model`
      DF      SS      MS      Fc p.value
Linear Effect      1 0.00323 0.00323 1.15 0.29816
Quadratic Effect   1 0.02253 0.02253 8.02 0.01107
Lack of fit        1 0.01411 0.01411 5.02 0.03792
Residuals          18 0.05060 0.00281

-----
$`Cubic Model
-----
      Estimate Standard.Error      tc p.value
b0 -0.36666667      0.25427 -1.44201 0.16647
b1  0.21111111      0.07778  2.71435 0.01421
b2 -0.01706667      0.00687 -2.48345 0.02309
b3  0.00040889      0.00018  2.24013 0.03794

$`R2 of cubic model`
[1] 1

$`Analysis of Variance of cubic model`
      DF      SS      MS      Fc p.value
Linear Effect      1 0.00323 0.00323 1.15 0.29816
Quadratic Effect   1 0.02253 0.02253 8.02 0.01107
Cubic Effect       1 0.01411 0.01411 5.02 0.03794
Lack of fit        0 0.00000 0.00000  0      1
Residuals          18 0.05060 0.00281

-----
      Levels      Observed Means
1           10      0.4466667
2           15      0.3400000
3           20      0.3000000
4            5      0.3133333
-----

```

Figure 5. Linear regression models fitted up to the third power for quantitative factors.

For comparison of quantitative treatments is available the routine of linear regression (by ordinary least squares) incorporated the analysis of variance. This routine fits polynomial models up to the third power, presenting individual analysis of variance and lack of fit (Figure 5).

The routines have certain autonomy because with only the information of the level of significance adopted by the researcher and the type of treatment the analysis will be performed automatically. Tests for the single factors or interactions are held only when necessary.

4. Final Remarks

Experimental Designs are meant to be a valuable package to help researchers analyze experimental data without a lot of work or complication. It is available for free and there is no intention to charge for its use, even after new features added to it.

The `ExpDes` package must be maintained and upgraded by the authors. For next steps, authors are concerned with homogeneity of variances tests, other multiple comparison tests, analysis of unbalanced experiments and the enhancement of the regression procedures.

Acknowledgements

Special acknowledgements to FAPEMIG for the financial support.

References

- [1] Stell, R.G.D. and Torrie, J.H. (1996) Principles and Procedures of Statistics: A Biometrical Approach. McGraw-Hill, New York.
- [2] Kirk, R.E. (1995) Experimental Design: Procedures for the Behavioral Sciences. 3rd Edition, Brooks/Cole Publishing Company, New York.
- [3] Cochran, W.G. and Cox, G.M. (1992) Experimental Designs. 2nd Edition, John Wiley & Sons, New York.
- [4] Quinn, G.P. and Keough, M.J. (2002) Experimental Design and Data Analysis for Biologists. Cambridge University Press, Cambridge. <http://dx.doi.org/10.1017/CBO9780511806384>
- [5] Casella, G. and Berger, L. (2002) Statistical Inference. 2nd Edition, Duxbury, Pacific Grove.
- [6] Mead, R. and Curnow, N. (1986) Statistical Methods in Agriculture and Experimental Biology. Chapman and Hall, New York.
- [7] Bretz, F., Hothorn, T. and Westfall, P.H. (2002) Multiple Comparisons Using R. Chapman & Hall/CRC, New York.
- [8] Healy, M.J.R. (1956) The Analysis of a Factorial Experiment with Additional Treatments. *Journal of Agricultural Science*, **47**, 205-206. <http://dx.doi.org/10.1017/S0021859600040120>
- [9] Groemping, U. (2014) CRAN Task View: Design of Experiments (DoE) & Analysis of Experimental Data. Version 2011-08-10. <http://CRAN.R-project.org/view=ExperimentalDesign>
- [10] Wheeler, B. (2014) AlgDesign: Algorithmic Experimental Design. R Package Version 1.1-7. <http://CRAN.R-project.org/package=AlgDesign>
- [11] Brien, C. (2011) Dae: Functions Useful in the Design and ANOVA of Experiments. R Package Version 2.1-7. <http://CRAN.R-project.org/package=dae>
- [12] Groemping, U. (2014) DoE.base: Full Factorials, Orthogonal Arrays and Base Utilities for DoE Packages. R Package Version 0.22-8. <http://CRAN.R-project.org/package=DoE.base>
- [13] Imai, K. (2013) Experiment: R Package for Designing and Analyzing Randomized Experiments. R Package Version 1.1-1. <http://CRAN.R-project.org/package=experiment>
- [14] Sandrini-Neto, L. and Camargo, M.G. (2012) GAD: Analysis of Variance from General Principles. R Package Version 1.1.1. <http://CRAN.R-project.org/package=GAD>
- [15] Ferreira, E.B., Cavalcanti, P.P. and Nogueira, D.A. (2013) ExpDes: Experimental Designs Package. R Package Version 1.1.2. <http://CRAN.R-project.org/package=ExpDes>
- [16] Shapiro, S.S. and Wilk, M.B. (1965) An Analysis of Variance Test for Normality (Complete Samples). *Biometrika*, **52**, 591-611. <http://dx.doi.org/10.1093/biomet/52.3-4.591>