Revista
Brasileira de
Ciência do Solo

Division – Soil in Space and Time | Commission – Pedometry

An Evaluation of the Use of Statistical Procedures in Soil Science

Laene de Fátima Tavares[1], **André Mundstock Xavier de Carvalho**[2]* **and Lucas Gonçalves Machado**[1]

[1] Universidade Federal de Viçosa, Instituto de Ciências Agrárias, Programa de Pós-graduação em Agronomia – Produção Vegetal, *Campus* Rio Paranaíba, Rio Paranaíba, Minas Gerais, Brasil.
[2] Universidade Federal de Viçosa, Instituto de Ciências Agrárias, *Campus* Rio Paranaíba, Rio Paranaíba, Minas Gerais, Brasil.

ABSTRACT: Experimental statistical procedures used in almost all scientific papers are fundamental for clearer interpretation of the results of experiments conducted in agrarian sciences. However, incorrect use of these procedures can lead the researcher to incorrect or incomplete conclusions. Therefore, the aim of this study was to evaluate the characteristics of the experiments and quality of the use of statistical procedures in soil science in order to promote better use of statistical procedures. For that purpose, 200 articles, published between 2010 and 2014, involving only experimentation and studies by sampling in the soil areas of fertility, chemistry, physics, biology, use and management were randomly selected. A questionnaire containing 28 questions was used to assess the characteristics of the experiments, the statistical procedures used, and the quality of selection and use of these procedures. Most of the articles evaluated presented data from studies conducted under field conditions and 27 % of all papers involved studies by sampling. Most studies did not mention testing to verify normality and homoscedasticity, and most used the Tukey test for mean comparisons. Among studies with a factorial structure of the treatments, many had ignored this structure, and data were compared assuming the absence of factorial structure, or the decomposition of interaction was performed without showing or mentioning the significance of the interaction. Almost none of the papers that had split-block factorial designs considered the factorial structure, or they considered it as a split-plot design. Among the articles that performed regression analysis, only a few of them tested non-polynomial fit models, and none reported verification of the lack of fit in the regressions. The articles evaluated thus reflected poor generalization and, in some cases, wrong generalization in experimental design and selection of procedures for statistical analysis.

Keywords: statistical analysis, means test, misuse of statistics.

*** Corresponding author:**
E-mail: andre.carvalho@ufv.br

Received: August 3, 2015

Approved: March 23, 2016

How to cite: Tavares LF, Carvalho AMX, Machado LG. An Evaluation of the Use of Statistical Procedures in Soil Science. Rev Bras Cienc Solo. 2016;40:e0150246.

DOI: 10.1590/18069657rbcs20150246

INTRODUCTION

Statistical analysis procedures are quantitative techniques used in experimental and observational science for assessing uncertainties and their effects on the interpretation of experiments and observations of natural phenomena (Steel et al., 1997; Zimmermann, 2004). Experimental statistical procedures are fundamental for better interpretation of the experimental results of agrarian sciences, and are used in almost all current scientific papers. However, incorrect use of statistical procedures applied to experimental data analysis may lead the researcher to incomplete or erroneous conclusions. This, in turn, can hinder review of themes in the literature and consequently delay the advancement of scientific knowledge.

Despite the importance of experimental statistics in soil science, a great difficulty persists in the selection of statistical procedures (Bertoldo et al., 2007). A few studies have assessed these difficulties from the different perspectives of agricultural sciences. For example, 35 % of the articles in the journal *Horticultura Brasileira* and 57 % of the articles in *Pesquisa Agropecuária Brasileira* were classified as “incorrect or partially correct” and “inadequate”, respectively, with respect to the use of mean comparison tests (MCTs) (Santos et al., 1998; Bezerra Neto et al., 2002).

These studies also highlight the importance of assessing and discussing other recurring difficulties, not only for better use of MCTs, but in order to support better use of statistical procedures in future papers. Therefore, more discussion is necessary on issues such as understanding the types of factors involved in research, the nature and structure of treatments, the choice of experimental design, the use of tests for detection of outliers, and the appropriate selection of regression models, among others. These and other aspects are essential theoretical decisions in experimental design that could result in articles with better methodological support.

As most academic knowledge comes from experimental data, correct interpretation of data becomes crucial. According to Alvarez V and Alvarez (2013), the validity and reliability of scientific writing are based on correct use of statistical inference. Thus, researchers’ knowledge from experimental planning to statistical analysis of the data is critical to experimental success and the credibility of findings. With this in mind, the aim of this study was to evaluate the characteristics of the experiments and the quality of the statistical procedures used in soil science as support for the correct use of statistical procedures.

MATERIALS AND METHODS

Two hundred articles, published between 2010 and 2014 in five Brazilian journals (*Acta Scientiarum – Agronomy*, Bioscience Journal, *Ciência Rural*, *Pesquisa Agropecuária Brasileira*, and *Revista Brasileira de Ciência do Solo*) were randomly selected, choosing 40 articles from each journal. They involve only experimentation and studies by sampling in the soil areas of fertility, chemistry, physics, biology, use and management. The journals selected have a Qualis-Capes B1 or higher rating and a strong tradition in publishing studies related to soil science. As the articles were chosen at random from each journal, the years of publication were not equally represented. The method followed in the evaluation was similar to that used by Lúcio et al. (2003), Bertoldo et al. (2008a), and Lucena et al. (2013) using a questionnaire. A questionnaire was prepared containing 28 questions, divided into queries related to the characteristics of the experiments, the statistical procedures used, and the quality of selection and use of these procedures. In studies where not all the variable responses were subjected to the same procedures of statistical analysis, only the most important variable response was considered according to the purpose of each study. After evaluation of the articles, the data were tabulated, involving calculation of the frequencies of answers to the questions.

The questions relating to diagnosis of the experimental characteristics involving the experimental environment (field, greenhouse, or laboratory), the treatment structures (factorial designs), the experimental design (completely randomized, randomized block, studies by sampling, split-plot or split-block), the number and the nature of the treatments (qualitative or quantitative), the number of replications, the use of analytical replicates, and the duration of the experiments. In this study, the expression "design" was used in the broad sense, including completely randomized, randomized block, split-plot, split-block, and other experimental arrangements. In addition, the expression "studies by sampling" was used to mean the same as "studies based on sampling" used by Lira Júnior et al. (2012) or "design by sampling" used by Alvarez V and Alvarez (2013). Analytical replicates were understand in this study as those obtained from more than one measurement for each true replication. It is therefore different from the concept of true replications. In studies with a factorial structure, the number of treatments was calculated by multiplying the number of levels of each factor under study.

Questions related to diagnosis of the statistical procedures used involved descriptions of verification of the condition of normality and homogeneity, transformation of variables, the loss of data, and performance of analysis of variance. It also involved mentioning the software used, the dispersion measures presented, the MCT applied, and the use of regression and correlation analysis.

Questions regarding diagnosis of the quality of use of the statistical procedures involved assessment of coherence in analysis of the structured treatments, presentation criteria for the deployment of factorial designs, indication of the significance level in regression analysis, evaluation of the use of MCTs, and evaluation of coherence in analysis of the experiments in split-plots and split-blocks.

RESULTS

Characteristics of the experiments

Many of the articles analyzed contained data from studies conducted under field conditions (73 %), followed by studies conducted in greenhouses and other environments (Table 1). In most studies, the authors referred to the use of a randomized complete block design (RCB), and the majority of the field studies were conducted in blocks. In greenhouses, on the other hand, most of the experiments were conducted in a completely randomized design (CRD). In the articles analyzed, 7.5 % were conducted in a greenhouse using RCB and 17.5 % were in the field using CRD (Table 1).

A significant number of the papers (27 %) were performed as field studies that assumed pre-existing situations as "treatments" (studies by sampling, studies by "systematic sampling", or "observational studies"), in which, therefore, no true replications of the treatments or randomization among them (Table 1). Furthermore, most papers (60.5 %) corresponded to short-term studies in which the treatment effects were evaluated for a maximum of 12 months. Among the long-term studies (three years or more), most corresponded to studies by sampling (Table 1).

Most of the papers evaluated included relatively small experiments, up to 12 treatments and with up to four replications (72 %). Only 6.5 % of the papers contained experiments with more than 36 treatments (Table 2). Only 1 % of the studies conducted experiments with merely two replications, and only 4.5 % of the papers mentioned the use of more than eight replications (Table 2). A significant number of the articles (6.5 %) did not report the number of replications used. Furthermore, only 2 % reported the use of analytical replicates for measuring one or more attributes evaluated (data not shown).

Half of the papers evaluated (50.5 %) did not have or consider the existence of a factorial structure (Table 2). However, in post analysis of the results presented in those papers,

Table 1. Frequency of studies, in percentages of the total number of studies assessed, classified according to the characteristics of the experiments in soil science in Brazilian journals in relation to the environment and experimental design

Technical feature	%
Experimental environment	
Field (experimental study)	46.0
Field (studies by sampling)	27.0
Greenhouse	22.0
Other	5.0
Experimental design	
Completely randomized design (CRD)	34.5
Randomized complete block design (RCB)	56.0
Other	1.5
Not indicated	8.0
Environment and experimental design	
Field in CRD	17.5
Greenhouse in CRD	14.0
Field in RCB	47.5
Greenhouse in RCB	7.5
Other environments or designs	13.5
Environment and duration of experiments	
Field (experiment): 0 to 12 months	27.0
Field (experiment): 12 to 36 months	7.5
Field (experiment): >36 months	5.5
Studies by sampling: 0 to 12 months	18.0
Studies by sampling: 12 to 36 months	1.0
Studies by sampling: >36 months	14.5
Other: 0 to 12 months	15.5
Other: 12 to 36 months	4.5
Other: >36 months	6.5

the frequency of unstructured treatment was observed to drop to 40.5 %. Among the structured experiments, most of them were considered double or triple factorial, without additional treatments, with a small portion of the factorial experiments having additional treatments (e.g., 3 × 5 + 1).

Statistical procedures used and the quality of selection and use of these procedures

Most of the articles reviewed mentioned the realization of analysis of variance (ANOVA) (Table 3), although only 19.5 % presented some ANOVA results (Table 4). Moreover, in most studies it was not mentioned or not conducted tests to verify the presuppositions of normality of residuals (92.5 %) and homoscedasticity of variance (93.5 %) (Table 3). Data transformation of the variables to meet these presuppositions (assumptions for parametric tests) was mentioned in only 4 % of the papers.

In the articles reviewed, 66 % mentioned the software used, SISVAR being the most cited, followed by SAS, SAEG, and ASSISTAT (data not shown). None of the studies reviewed reported the data loss or the use of tests for outliers (Table 3). Most studies used only

Table 2. Frequency of studies, in percentages of the total number of studies assessed, classified according to the characteristics of the experiments in soil science in Brazilian journals in relation to experimental size and the treatment structure

Technical feature	%
Number of treatments	
Up to 6 treatments	41.0
From 7 to 12 treatments	31.0
From 13 to 18 treatments	7.0
From 19 to 24 treatments	8.0
From 25 to 36 treatments	6.5
More than 36 treatments	6.5
Number of true replications	
Not indicated	6.5
Two	1.0
Three	21.5
Four	49.5
Five	10.5
Six	5.5
Seven or eight	1.0
More than eight	4.5
Nominal structure of treatments	
Unstructured	50.5
Two factors under study	36.0
Three factors under study	10.0
Four factors under study	1.0
Factorial with additional treatments	2.5
Real structure of the treatments	
Unstructured	40.5
Two factors under study	42.0
Three factors under study	10.5
Four factors under study	2.0
Factorial with additional treatments	5.0

the Tukey test for multiple comparisons of means (46 %), followed by the Scott-Knott test, Fisher's LSD test, and the Duncan test. Only 0.5 % of the studies were done using the Student-Newman-Keuls (SNK) test (Table 3). A few studies used contrasts involving more than two means (5 %) (data not shown). In 70 % of the cases, the use of the MCT was classified as appropriate, 9 % as partially appropriate, and 21 % as inappropriate (Table 4). Most of the cases classified as 'inappropriate' involved the use of Duncan or Fisher LSD tests. A small number of studies involved use of an MCT in cases where regression analysis would be more appropriate (four or more quantitative levels).

In 50.5 % of the papers, the authors did not consider the experiments as factorial designs. However, in fact, about 10 % of them had some type of structure (Table 2). The most common cases involved the bifactor structure, in situations where the evaluation periods, soil layers, and evaluation times were analyzed as treatments in the results, but were not described as such in the methods reported. Among the articles whose treatments had factorial structure, 12.5 % ignored the structure, and all of the treatment

Table 3. Frequency of studies, in percentages of the total number of studies assessed, classified according to the statistical procedures used in soil science in Brazilian journals

Technical feature	%
Analysis of variance	
Mentioned	83.5
Not mentioned	16.5
Normality test	
Kolmogorov-Smirnov	0.5
Shapiro-Wilk	1.5
Anderson-Darling	0.0
Jarque and Bera	0.0
Others	2.5
Test used not indicated	3.0
No test mentioned or performed	92.5
Homoscedasticity test	
Hartley	0.0
Bartlett	1.0
Cochran	1.5
Levene	0.5
Other	0.5
Test used not indicated	3.0
No test mentioned or performed	93.5
Data Transformation	
Mentioned	4.0
Not mentioned	96.0
Multiple Comparison Test	
More than one test was used	1.0
Only Tukey	46.0
Only Duncan	5.5
Only Dunnett	1.0
Only LSD or Fisher’s LSD	8.0
Only Student-Newman-Keuls (SNK)	0.5
Only Scott-Knott	10.0
Did not use any test	28.0
Data Loss	
Mentioned	0.0
Not mentioned	100.0

means were compared against each other (Table 4). In all these cases, the articles do not report ANOVA results, such as significance of the F values for the mean squares of the treatments or of the factors under study. In addition, 20.5 % of the articles always decomposed the interaction (unfolding of interaction) between the factors even without showing or mentioning the significance of the interaction. Divergences between nominal experimental design (which was described in the paper) and real design (which was really in the treatments) were observed. Among these differences, only 1 % of the studies are

Table 4. Frequency of studies, in percentages of the total number of studies assessed, classified according to the quality of use of statistical procedures in soil science in Brazilian journals

Technical feature	%
ANOVA results	
Some ANOVA results were presented	19.5
No ANOVA results were presented	80.5
Criteria for decomposition of interaction in factorial designs	
There was no mention of any criterion because the structure was ignored	7.5
There was no mention of any criterion but always decomposed	20.5
There was no mention of any criterion but not always decomposed	5.0
Decomposition according to the significance of the interaction	26.0
Special cases of nominal design	
None	77.0
Split-plot	19.0
Split-split-plot	3.0
Split-block	1.0
Joint analysis	0.0
Special cases of real design	
None	70.5
Split-plot	12.5
Split-split-plot	3.0
Split-block	14.0
Joint analysis	0.0
Regression and correlation analysis	
There was no regression analysis	63.5
There was, and each regression parameter was tested	10.5
There was, and the significance of the regression was indicated only in R^2	14.5
There was, but the significance of the regression parameters was not tested	11.5
There was correlation analysis between variables	17.5
Classification of the use of MCT	
Appropriate	70.0
Partially appropriate	9.0
Inappropriate	21.0

considered to have conducted experiments in a split-block design whereas 14 % of the studies showed this type of experimental design (Table 4).

Regression and correlation analyses were used in 36.5 and 17.5 % of the studies assessed, respectively (Table 4). Among the studies that conducted regression analysis, 11.5 % did not test the overall significance of the regression or the significance of the regression parameters, 14.5 % showed significance only in the R^2 (as an indication of overall significance of regression), and 10.5 % showed the significance of each parameter of the equation (Table 4). None of the articles described the observance of non-significance of the regression residuals (lack of fit of the regression) as a criterion for selection of the regression models, and most of the articles tested only the fit to the linear or quadratic models.

DISCUSSION

Characteristics of the experiments

Many field experiments conducted using the RCB are related to the greater heterogeneity of this experimental environment, mainly connected with ground slope, which usually causes differences in soil fertility, moisture, and mineral composition, among other factors. This situation was also reported by Lúcio et al. (2003) in their assessment of the studies in crop science in the *Ciência Rural* journal. Researchers, however, need to know the direction of one or more sources of variation in the experimental environment to ensure that the local control principle is adopted correctly (variability from the moisture gradients, fertility, mineral composition, and historical uses). With knowledge of the effects to be controlled, a theoretical assessment is still feasible regarding the possibility of an interaction between the blocks and treatments, situation that would forbid the installation of the experiment in that place. Therefore, it is interesting that authors report what is being controlled (commonly the slope) when using RCB, and that they do not opt for its use simply because the experiment was done under field conditions. Nevertheless, it is important to consider that, in some cases, the option for the RCB experiments in the field, even on flat landscapes, may be justified by operational issues. In such cases, planting, crop, and other activities may be performed by different workers for each block or on different days for each block.

A significant number of the studies assessed were based on studies by sampling. This type of study has generated conflicting opinions between editors and reviewers in scientific journals on soil, because, despite replications being observed within each area or sampled area, these areas are not repeated, and are considered pseudo-replications (Ferreira et al., 2012; Lira Júnior et al., 2012). However, there is no consensus on the term to be used to define such study types.

Although this issue is treated as a pseudo-replication problem only (Hurlbert, 1984), the basic principle of randomization between treatments is not respected. This failure becomes critical as there is no independence between the replications of each treatment and no prior guarantee of homogeneity among the areas where the treatments are applied. Thus, in light of these limitations, these studies clearly cannot be considered as experiments. According to Ferreira et al. (2012), the reviewer in these situations should check only if the presuppositions of normality and homoscedasticity are satisfied and, if not, recommend nonparametric statistical analysis methods. Also, according to Ferreira et al. (2012) and Lira Júnior et al. (2012), such articles should not be rejected based only on this fact, although this is not consensus among reviewers. In the field of ecology, this type of study is very common, representing about 27 % of the studies conducted under field conditions (Hurlbert, 1984). It is important, however, that a correct description of the research strategy be presented in these cases, clearly indicating that it is a study by sampling. This study type does not, in principle, present an experimental design, which does not prohibit them from being analyzed with parametric statistical procedures, such as analysis of variance, means testing, regression, etc., as long as they satisfy the requirements for such (normality, homoscedasticity, etc.). These parametric statistical procedures are also used in studies in social and ecological areas where the basic principles of experimentation also cannot be fulfilled (Hurlbert, 1984; Marôco, 2011).

It is also important to consider that, based on the present data, most of the long-term studies conducted (three years or more) are studies by sampling. Thus, understanding the limitations of the conclusions drawn in these articles, which are restricted to specific study conditions, they may be considered as case studies. A certain number of case studies which point to the same fact may ultimately allow more generalized conclusions to be drawn on a specific topic, as is the case in medical sciences (An and Cuoghi, 2004). According to the frequency of studies by sampling in soil use and soil management areas, there is a tendency to accept these studies in cases restricted to situations with clear technical or economic unfeasibility in performing classical experiments.

Long-term studies enable researchers to better assess the effects of treatments on response to important variables in the study and to other complementary variables, which can be useful in evaluating unexpected effects. The high frequency of short-time studies (<12 months), however, suggests that soil researchers may be giving less importance to long-term experiments. This is probably related to the higher cost involved in such studies, and also to pressure experienced by most Brazilian researchers to increase their number of publications. The predominance of small experiments (<12 treatments) may be related to time and cost reductions. Among the articles studied, only nine mention the use of more than eight replications. However, it could also be attributed to a common understanding of better quality that these experiments permit when compared to the large experiments, especially with regard to better standardization of experimental conditions and better standardization of the activities of conducting and evaluating (Vieira, 2006).

The vast majority (72 %) included tests with up to four replications. Zimmermann (2004) argues that, in most cases, the number of replications is selected based on financial resources, the time required for the evaluations, the area available, or availability of workers. The number of replications of an experiment is extremely important because experimental errors tend to be inversely proportional to the number of replications. This relationship, however, is not linear, because the experimental error decreased increasingly smaller for each increase in the number of replications. When the researcher is required to reduce the variability of a response variable but not increase the number of replications to a great extent, one option is to make the measurements and analytical determinations with replicates of each true replication, a strategy still underused as observed in this study.

The appropriate number of replications must allow at least 15 degrees of freedom (DF) of residuals (Alvarez V and Alvarez, 2013). According to Pimentel-Gomes (1987), however, the minimal number was only 10. There is no theoretical basis for that number, only the understanding that the sensitivity of statistical tests is linked to the DF of residuals. The higher the DF of residuals, the lower the residual mean square (estimate of experimental error) tends to be, and the higher the power of the statistical tests applied. Thus, when conclusions are based on “similarity” among the treatments, it is very important that the DF of residuals be high, to enable sensitivity to the statistical tests, which results in a lower possible rate of type II error. However, according to Pimentel-Gomes (2009) and Dutcosky (2013), when conclusions are based on the differences among the treatments, this requirement is not necessarily important. When the statistical tests do not indicate differences, with high DF of residuals, the inference that the treatments do not differ will be acceptable. On the other hand, when large differences are expected among the treatments, a low DF of residuals may permit sufficient sensitivity of the statistical tests. Therefore, an experiment can be planned in accordance with technical and financial limitations with the DF of residuals less than 15 and it will still be valid from the viewpoint of its conclusions long as they are based on “differences” and not on “similarities”.

The extensive use of factorial design experiments in soil science (Table 2) must be related to the need to understand the interactions among the different factors involved in the responses to the treatments in the complex soil environment. Moreover, the structuring treatment makes it easier to discover patterns in the phenomena, evidenced by a lack of interaction between the factors studied. However, the inclusion of several factors makes the procedures of statistical analysis difficult, as well as the overview of the treatment effects and interactions among the factors; and so it is recommended not to include more than three factors in an experiment (Vieira, 2006). This recommendation appears to be followed in the articles evaluated in this study.

Statistical procedures used and the quality of selection and use of these procedures

Most of the articles reviewed did not present ANOVA results, which could reduce the reliability of statistical inferences (Table 4). Thus, some journals have required ANOVA results, especially the DF of residuals, calculated F value, or the *p* value, in order to increase the credibility of the statistical analysis (Volpato, 2010). These methods, however, do not ensure credibility, especially when restricted to the simple, and increasingly common, notation type "($F_{4, 22}$ = 0.021)", which indicates the *p* value for F corresponding to the DF of treatments and DF of residuals. It is easy to see how this notation is unsatisfactory in factorial design experiments, for example. More important than this notation is to simply and accurately present information about the structure, experimental design, number of replications, and some estimate of experimental error (such as the coefficient of variation or the residual mean square). As a result, reviewers and readers may check the statistical differences indicated.

A significant number of studies in animal sciences comparing the marginal means of a factorial experiment without mentioning the possible interactions among the factors (Cardellino and Siewerdt, 1992). This situation clearly shows the importance of presenting ANOVA results. According to Bertoldo et al. (2008b), in post evaluation of 226 scientific papers published in plant science in the *Ciência Rural* journal, most of the errors occurring in analysis of experiments in factorial designs were related to studies in which the authors did not consider interactions among the factors, testing only the marginal means. These results are only valid if the interactions are not significant because, otherwise, it becomes necessary to work within the levels of each factor (decomposition of interaction).

The significance of the interaction gives valuable information because it enables and validates generalizations regarding the effect of factors under study (Perecin and Cargnelutti Filho, 2008). Such generalizations are especially useful in understanding the phenomena and are "general standards", as opposed to the concept of always looking for decomposed interaction, which makes perception of these standards difficult. Generalizations are obtained from comparisons among the marginal means. However, it is easy to understand that, in some situations, even without significant interaction at 5 %, the decomposed interaction demonstrates the differential effects of levels of factor B for each level of factor A, or vice versa. In order to avoid this problem, the criterion for considering the significance of the interaction (commonly $p<0.05$) can be moved to a higher value, such as 0.25, as suggested by Perecin and Cargnelutti Filho (2008). By adopting this criterion, the decomposition of interaction will be suspended only in situations where there is a great deal of evidence for the lack of interaction, increasing the reliability of the generalizations.

The high proportion of papers that do not mention verification of the presuppositions of normality and homocedasticity is alarming because erroneous conclusions may be accepted if these conditions are not met. Often, the independence and additivity conditions are already assumed when the basic principles of experimentation are respected, which becomes critical in studies by sampling. However, there is a consensus that parametric tests are valid only when the data meet the basic presuppositions of independence of errors, additivity of effects accepted in the model, normality and homoscedasticity. In fact, when these presuppositions are not fulfilled, further tests may produce results different from those that would have been generated had the data been transformed earlier to meet such presuppositions or if the data had been subjected to non-parametric tests. According to Lucena et al. (2013), in an evaluation of studies in dentistry, the use of nonparametric tests for data, in which residues did not follow the normal distribution but had been analyzed in the studies as normal, altered the conclusions in the articles in 19 % of the cases.

A possible explanation for the high number of papers that did not verify the presuppositions is that the statistical software tools do not test these presuppositions automatically when performing ANOVA (Vieira, 2006). It is also important to consider that there are differences among the tests used to evaluate these presuppositions (Jarque and Bera, 1987; Lim and Loh, 1996; Santos and Ferreira, 2003). These differences, including differences in power, robustness and adequacy of the experimental design adopted, may result in some degree of subjectivity in selection of these tests. In this context of lingering uncertainties, graphical tools for analysis of presuppositions can be useful and enable correct decisions despite the most probable levels of subjectivity.

Failure to check that ANOVA presuppositions are met may also be linked to the frequency with which outliers make it difficult for fitting the data to these conditions. The presence of outliers was not described or tested in any of the studies assessed, indicating a clear trend of omission of such information. According to Barnett and Lewis (1996), lack of criteria for detection of outliers is relatively frequent, which may lead to a biased selection of outliers.

Although there are various tests for detection of outliers, the common recommendation is that an outlier should be deleted only when there is a known reason to do so, that is, if the cause of discrepancy can be confirmed (Vieira, 2006; Pimentel-Gomes, 2009). There are situations, however, where such a check is impracticable, and the use of an impartial and rigorous statistical test is a very useful, yet unexplored, tool. Notable tests for outliers are the Cook distance (most appropriate for paired data in correlation analysis), Grubbs test, Dixon test, and the Chauvenet criterion and derivations of this criterion, which highlight the ESD (generalized Extreme Studentized Deviate) criterion (Rosner, 1983).

The Chauvenet criterion, one of the first criteria developed for this purpose, sometimes referred to as “criterion of maximum standardized standard deviation” (Vieira, 2006), is a simple criterion with good qualities when applied, considering the standard deviation calculated with the residual mean square and not with the residue for each treatment. Its derivations, however, revised the tabulated critical values, making this a more rigorous test and enabling detection of more than one outlier in one group (Rosner, 1983). This procedure (termed “generalized ESD”), although quite rigorous, is considered one of the best procedures for this purpose by Walfish (2006) and Manoj and Senthamarai-Kannan (2013) and can be used even in analyzing deviations from the adjusted regression models (Paul and Fung, 1991).

Finally, it is also possible that tests for detection of outliers are underused due to the unavailability of these tests in the most popular statistical software or because of the exclusion of the outliers results in unbalanced data. Unbalanced data produce several complications in statistical analysis, especially in factorial experiments and in RCB (Wechsler, 1998). Statistical software packages like SISVAR and ASSISTAT do not analyze unbalanced data in their routine procedures (Ferreira, 2008), which may induce erroneous replacement of lost data for average values.

A significant number of the papers assessed cited the software used without, however, correctly describing the procedures performed. It is important to emphasize that the statistical tests used need to be mentioned and not the tools used to perform them (Volpato, 2010). In addition, the frequent choice for SISVAR, SAEG, or ASSISTAT can be an indication of the unfriendly interface of the famous SAS and R. Most statistical applications provide an extensive list of useful procedures for many different scientific fields. This provides an overload of options and commands, which contributes to a more complex and less intuitive interface. With this array of options, specific procedures applicable to few areas blend into the general procedures and hinder access to the most common classical procedures of experimental statistics (such as normality tests, homoscedasticity, analysis of variance, multiple comparison tests, contrast and regression analysis). Additionally, procedures are often presented in language barely accessible to users from non-statistical fields (such as “PROC GLM” in SAS) and so may result in a complex interface that cannot be deduced by graduate students (Volpato, 2010). In this regard, knowledge of the experimental characteristics of papers on soil science may be useful in development of simpler applications considering domain specific knowledge, contributing to a friendlier interface and a more accessible statistical language for users.

The extensive use of the Tukey test confirms the results presented by Santos et al. (1998), Bezerra Neto et al. (2002) and Lúcio et al. (2003). This is a rigorous test, but with less power (sensitivity) than the other MCT (Vieira, 2006). In most situations, greater rigor is not advantageous because the more conservative a test is, the lower will be its sensitivity and ability to detect differences, resulting in the type II error. In fact, employing an MCT due to its popularity and not because of its ability to adapt to the kind of hypothesis to be tested can cause simplistic or incomplete analyses, thereby leading to the loss of relevant information.

The use of Duncan and Fisher's LSD tests was responsible for most cases in which the MCT was classified as inappropriate. These tests do not have minimal control over the real type I error (experiment wise) as shown by Carmer and Swanson (1973), Perecin and Barbosa (1988), and Sousa et al. (2012) and should be in disuse (Pimentel-Gomes, 2009). The Student-Newman-Keuls test (SNK), although criticized by Einot and Gabriel (1975) for its greater complexity, balances high power with good control of the real type I error (Carmer and Swanson, 1973; Perecin and Barbosa, 1988; Borges and Ferreira, 2003). For these reasons, its use should be encouraged (Perecin and Barbosa, 1988), as has already occurred in other fields of science (Curran-Everett, 2000). The popularity of the Scott-Knott grouping test (Table 3) may be linked to its robustness and lack of ambiguity in the results it generates, which greatly facilitates interpretation of the results (Borges and Ferreira, 2003). Furthermore, this test has higher power than Tukey and SNK when a large number of treatments are to be compared (Silva et al., 1999). However, few studies on this test in the literature and high type I error rates in the partial nullity condition, observed by Borges and Ferreira (2003), still leave doubt in regard to its features.

A higher power of any MCT and with good real type I error control is achieved through the use of orthogonal contrasts, tested by the F test or t test (Gill, 1973). These tests can be especially useful in incomplete factorial designs or when comparisons of interest are few or involve more than two means (Baker, 1980; Alvarez V and Alvarez, 2006). Their limitations, however, involve a limited number of orthogonal comparisons among means and difficulty in manually performing the calculations because most of the software tools do not offer support to test them. In addition to these limitations, there is greater difficulty in interpreting the results, which is intensified by the lower popularity of this procedure.

A few studies classified as inappropriate correspond to quantitative factors of four or more levels to which an MCT was applied when regression analysis would have been more appropriate to compare means. These results contrasted with those presented by Cardellino and Siewerdt (1992), Santos et al. (1998), Bezerra Neto et al. (2002), and Bertoldo et al. (2008b), in which this situation was reported as being very common. This divergence suggests an observed improvement in the use of MCTs in experiments related to quantitative treatments. This improvement can be attributed to the support given by these and other studies on the same subject to the discussion on the use of statistical procedures in agricultural experimentation. In this respect, the importance of these studies is evident not only for agricultural sciences but for other scientific fields as well, as it can contribute to improvement in the quality of selection of statistical procedures used in analysis of experiments.

A significant number (10 %) of studies, in presenting experimental techniques, described an experiment devoid of structure, although, in the results, data were compared as if they had factorial structure. In this situation, the reader is surprised to find in the results section that it is a factorial design. This outcome may be linked to the difficulty in fully understanding the factors involved in the study (uni- or multifactorial) or even the purpose of the study because the structure is designed based on the objectives proposed.

If, on the one hand, there were experiments without structure being analyzed as factorial in the results, on the other hand, there were studies that showed factorial structure but a significant percentage (7.5 %) of them ignored this structure. In these cases, the MCTs lose sensitivity since the DFs of the treatments stop being decomposed by each factor level under study. In the studies whose treatments have factorial structure, we also found evidence with little careful observation of the interaction among the factors studied, given that over 30 % of these studies decomposed the interaction without even a presentation or mention of the significance of the interaction. This situation also reveals the importance of presenting certain ANOVA results, as discussed earlier.

The differences between the nominal and real experimental designs mostly occurred in situations in which the experiment was in a split-block design, but considered by the authors as a simple factorial or as a split-plot factorial. According to various statistical manuals, factors under study such as soil layers, time, or successive years of appraisal can be analyzed as a split-plot (Cochran and Cox, 1957; Steel et al., 1997; Banzatto and Kronka, 2008; Dias and Barros, 2009; Barbin, 2013), which are exemplified as split-plot in time and split-plot in space. Other authors, however, emphasize that the existence of restrictions on randomization of the treatments in the subplot implies the need for analysis as split-blocks (Pimentel-Gomes, 2009; Alvarez V and Alvarez, 2013). Therefore, the inconsistencies shown in table 4 on how to analyze such experimental situations can be linked to disagreements on the subject published in the main statistical manuals.

Two common situations in soil science are notable in this regard: factors whose levels involve hours, years, or successive appraisal cycles of production; and factors whose levels involve layers, depths, or soil sampling positions (row and inter-row or near and far, for example). According to Vivaldi (1999) and Alvarez V and Alvarez (2013), in both situations, the levels cannot be perfectly randomized to the subplots, as the first year or cycle was always preceded by the second, the second by the third, and so on. In the case of the layers evaluated by soil sampling, the situation would be quite similar, with the surface layer clearly always arranged above the subsurface layers. Furthermore, the presupposition of the independence of the errors between the levels must be ignored in these cases, because successive times and the successive layers are strongly correlated, and often evaluated in the same experimental units (Vieira, 2006). The independence problem thus becomes even more severe when no independent experimental units are presented for the different times or different layers evaluated. Vivaldi (1999) and Alvarez V and Alvarez (2013), however, make no reference to the rare situations in which the data are collected over time or space in independent experimental units.

The split-plot design is only suited to repeated measures on the same experimental unit over time or space where the conditions for non-sphericity are satisfied (Vivaldi, 1999). Otherwise, multivariate techniques, less sensitive and more complex than the univariate, need to be used (Vivaldi, 1999). Therefore, the simplest solution in these cases is exclusion of treatments of this nature in experiments, considering the successive times and the different soil layers as different response variables and not as levels of a factor under study. Comparisons among them must be restricted to descriptive statistics.

In situations where growth rate or maximum or minimum points need to be compared, which would justify inclusion of the time factor as treatments, these rates or points could simply be obtained for each replication (over time) and compared as a new variable response (Vivaldi, 1999). In a few cases, however, depending on the objectives of the researcher, these levels may not be removed from the structure of the treatment. In such cases, a better option than using a split-plot would be, according to Alvarez V and Alvarez (2013), an analysis by split-block. In a split-block analysis, the sensitivity of comparisons among the main treatments is reduced to levels similar to those that would be present if the variables were treated as different response variables.

Regression analysis presented in the articles reviewed reveals the lack of a standard representation of the significance of equations, with little consensus regarding how and what needs to be tested for fitting appropriate models. Regression analysis is used for several purposes; however, in experimental statistics, it involves not only a check on which mathematical models are appropriate for the data but also includes an assessment of the explanatory quality of this fit (theoretical sense), the significance of the fitted model, and the insignificance of the fraction unexplained by regression (regardless of the regression term or residual of regression or lack of fit of regression analysis) (Alvarez V and Alvarez, 2003).

Renowned experimental statistics handbooks do not mention the need to test and indicate the significance of each parameter of the equation by the t test (Pimentel-Gomes, 1987; Zimmermann, 2004; Banzatto and Kronka, 2008; Pimentel-Gomes, 2009; Barbin, 2013). Other authors, however, support this procedure (Nunes, 1998; Alvarez V and Alvarez, 2003). It is important to remember that the first degree term of a polynomial equation of the second degree, for example, although it has significance only above 5 %, should not be excluded from the model (Pimentel-Gomes, 1987). Alvarez V and Alvarez (2003) also support this recommendation, although these authors argue that the significance of this lower degree term should also be indicated, even when it is above 5 %. This creates uncertainties regarding the real need to indicate the significance of each regression parameter because the mere indication of the significance of the model as a whole and the verification of the non-significance of the lack of fit (tested by the F test in the ANOVA regression) resulted in the same decision regarding model selection.

In general, the smaller the number of parameters is, the better the balance between the simplicity and quality of the fitted model (most parsimonious model). Rarely does an explainable behavior, isolated by experimental conditions, require complex mathematical models involving more than two dependent parameters. Nevertheless, it is also important to remember that often polynomial models are not adequate for natural phenomena (Pimentel-Gomes, 2009). Evaluation of the fitted regression models assessed in the papers revealed that several non-linear or non-quadratic phenomena were being ignored. Exponential, Mitscherlich, sigmoidal, and other models, which are relatively simple patterns involving mathematical models with only two parameters (Pimentel-Gomes and Conagin, 1991), are underutilized. The difficulty in performing a regression ANOVA with these models in most software may be contributing to this situation.

The studies reflect poor generalization and, in some situations, errors in both experimental planning and the choice of procedures for statistical analysis. In part, this situation is related to the lack of consensus on the use of certain procedures. In some cases, however, extremely useful statistical procedures have been poorly utilized, whereas in other cases the most popular applications do not offer certain procedures or offer them in a complex and unintuitive interface. In addition to the statistical procedures discussed in this study, several others also require further discussion, such as the issue of dispersion measures to be presented, the use of non-continuous variables responses, criteria for transforming data, nonlinear regression models, and the adequacy of the various tests for normality and homoscedasticity, among others.

CONCLUSIONS

Scientific papers in soil science involve a wide frequency of studies by sampling, and also typically small and short duration experiments.

The statistical procedures that most often compromise the quality of papers in soil science are linked to the practice of not checking the presuppositions for ANOVA, omission of

ANOVA results in factorial experiments, selection of regression models and presentation of their significance, incorrect description of the experimental design of studies by sampling, and misuse of experiments in split-plot.

REFERENCES

Alvarez V VH, Alvarez GAM. Apresentação de equações de regressão e suas interpretações. Bol Inf SBCS. 2003;28:28-32.

Alvarez V VH, Alvarez GAM. Comparações de médias ou testes de hipóteses? Contrastes! Bol Inf SBCS. 2006;31:24-34.

Alvarez V VH, Alvarez GAM. Reflexões sobre a utilização de estatística para pesquisa em ciência do solo. Bol Inf SBCS. 2013;38:28-35.

An TL, Cuoghi AO. A utilização da estatística na ortodontia. Rev Dent Press Ortodon Ortop Facial. 2004;9:97-108. doi:10.1590/S1415-54192004000600014

Baker RJ. Multiple comparison tests. Can J Plant Sci. 1980;60:325-7. doi:10.4141/cjps80-053

Banzatto DA, Kronka SN. Experimentação agrícola. 3a ed. Jaboticabal: FUNEP; 2008.

Barbin D. Planejamento e análise estatística de experimentos agronômicos. 2a ed. Londrina: Mecenas; 2013.

Barnett V, Lewis T. Outliers in statistical data. 3rd ed. New York: John Wiley & Sons; 1996.

Bertoldo JG, Coimbra JLM, Guidolin AF, Mantovani A, Vale NM. Problemas relacionados com o uso de testes de comparação de médias em artigos científicos. Biotemas. 2008a;21:145-53. doi:10.5007/2175-7925.2008v21n2p145

Bertoldo JG, Coimbra JLM, Guidolin AF, Toaldo AMD. Uso ou abuso de testes de comparações de médias: conhecimento científico ou empírico? Cienc Rural. 2008b;38:1145-8. doi:10.1590/S0103-84782008000400039

Bertoldo JG, Rocha F, Coimbra JLM, Zitterell D, Grah VF. Teste de comparação de médias: dificuldades e acertos em artigos científicos. Rev Bras Agrocienc. 2007;13:441-7. doi:10.18539/CAST.V13I4.1409

Bezerra Neto F, Nunes GHS, Negreiros MZ. Avaliação de procedimentos de comparações múltiplas em trabalhos publicados na revista Horticultura Brasileira de 1983 a 2000. Hortic Bras. 2002;20:5-9. doi:10.1590/S0102-05362002000100001

Borges LC, Ferreira DF. Power and type I error rates of Scott-Knott, Tukey and Student-Newman-Keuls's tests under residual normal and non normal distributions. Rev Mat Estat. 2003;21:67-83.

Cardellino RA, Siewerdt F. Utilização correta e incorreta de testes de comparação de médias. Rev Bras Zootec. 1992;21:985-95.

Carmer SG, Swanson MR. An evaluation of ten pairwise multiple comparison procedures by Monte Carlo methods. J Am Stat Assoc. 1973;68:66-74. doi:10.1080/01621459.1973.10481335

Cochran WG, Cox, GM. Experimental designs. 2nd ed. London: John Wiley; 1957.

Curran-Everett D. Multiple comparisons: philosophies and illustrations. Am J Physiol Reg Integr Comp Physiol. 2000;279:R1-R8.

Dias LAS, Barros, WS. Biometria experimental. Viçosa, MG: Suprema; 2009.

Dutcosky SD. Análise sensorial de alimentos. 4a ed. Curitiba: Champagnat – PUCPress; 2013.

Einot I, Gabriel KR. A study of the powers of several methods of multiple comparisons. J Am Stat Assoc. 1975;70:574-83. doi:10.1080/01621459.1975.10482474

Ferreira DF, Cargnelutti Filho A, Lúcio AD. Procedimentos estatísticos em planejamentos experimentais com restrições na casualização. Bol Inf SBCS. 2012;37:1-35.

Ferreira DF. SISVAR: um programa para análises e ensino de estatística. Rev Symposium. 2008;6:36-41.

Gill JL. Current status of multiple comparisons of means in designed experiments. J Dairy Sci. 1973;56:973-7. doi:10.3168/jds.S0022-0302(73)85291-9

Hurlbert SH. Pseudoreplication and the design of ecological field experiments. Ecol Monogr. 1984;54:187-211. doi:10.2307/1942661

Jarque CM, Bera AK. A test for normality of observations and regression residuals. Int Stat Rev. 1987;55:163-172.

Lim TS, Loh WY. A comparison of tests of equality of variances. Comp Stat Data Anal. 1996;22:287-301. doi:10.1016/0167-9473(95)00054-2

Lira Júnior AM, Ferreira RLC, Sousa ER. Uso da estatística em trabalhos baseados em amostragem na ciência do solo. Bol Inf SBCS. 2012;37:1-35.

Lucena C, Lopez JM, Pulgar R, Abalos C, Valderrama MJ. Potential errors and misuse of statistics in studies on leakage in endodontics. Int Endodon J. 2013;46:323-31. doi:10.1111/j.1365-2591.2012.02118.x

Lúcio AD, Lopes SJ, Storck L, Carpes RH, Lieberknecht D, Nicola MC. Características experimentais das publicações da Ciência Rural de 1971 a 2000. Cienc Rural. 2003;33:161-4. doi:10.1590/S0103-84782003000100026

Manoj K, Senthamarai-Kannan K. Comparison of methods for detecting outliers. Int J Sci Eng Res. 2013;4:709-14.

Marôco J. Análise estatística com o SPSS Statistics. 5a ed. Pêro Pinheiro: ReportNumber; 2011.

Nunes RP. Métodos para a pesquisa agronômica. Fortaleza: Universidade Federal do Ceará; 1998.

Paul SR, Fung KY. A generalized extreme studentized residual multiple outlier detection procedure in linear regression. Technometrics. 1991;33:339-48. doi:10.2307/1268785

Perecin D, Barbosa JC. Uma avaliação de seis procedimentos para comparações múltiplas. Rev Mat Estat. 1988;6:95-103.

Perecin D, Cargnelutti Filho A. Efeitos por comparações e por experimento em interações de experimentos fatoriais. Cienc Agrotec. 2008;32:68-72. doi:10.1590/S1413-70542008000100010

Pimentel-Gomes F. A estatística moderna na pesquisa agropecuária. 3a ed. Piracicaba: Potafós; 1987.

Pimentel-Gomes F. Curso de estatística experimental. 15a ed. Piracicaba: FEALQ; 2009.

Pimentel-Gomes F, Conagin A. Experimentos de adubação: planejamento e análise estatística. In: Oliveira AJ, Garrido WE, Araújo JD, Lourenço S, coordenadores. Métodos de pesquisa em fertilidade do solo. Brasília, DF: Embrapa-SEA; 1991. p.103-88.

Rosner B. Percentage points for a generalized ESD many-outlier procedure. Technometrics. 1983;25:165-72.

Santos AC, Ferreira DF. Definição do tamanho amostral usando simulação Monte Carlo para o teste de normalidade baseado em assimetria e curtose. I. Abordagem univariada. Cienc Agrotec. 2003;27:432-7. doi:10.1590/S1413-70542003000200025

Santos JW, Moreira JAN, Beltrão NEM. Avaliação do emprego dos testes de comparação de médias na revista Pesquisa Agropecuária Brasileira (PAB) de 1980 a 1994. Pesq Agropec Bras. 1998;33:225-30.

Silva EC, Ferreira DF, Bearzotti E. Avaliação do poder e taxas de erro tipo I do teste de Scott-Knott por meio do método de Monte Carlo. Cienc Agrotec. 1999;23:687-96.

Sousa CA, Lira Júnior MA, Ferreira RLC. Avaliação de testes estatísticos de comparações múltiplas de médias. Rev Ceres. 2012;59:350-4. doi:10.1590/S0034-737X2012000300008

Steel RGD, Torrie JH, Dickey DA. Principles and procedures of statistics: a biometrical approach. 3rd ed. New York: MacGraw Hill; 1997.

Vieira S. Análise de variância: anova. São Paulo: Atlas; 2006.

Vivaldi LJ. Análise de experimentos com dados repetidos ao longo do tempo ou espaço. Planaltina: Embrapa Cerrados; 1999. (Série documentos, 8).

Volpato GL. Dicas para redação científica. 3a ed. São Paulo: Cultura Acadêmica; 2010.

Zimmermann FJP. Estatística aplicada à experimentação agrícola. 2a ed. Santo Antônio de Goiás: Embrapa Arroz e Feijão; 2004.

Walfish S. A review of statistical outlier methods. Pharm Technol. 2006. [accessed on 01 Mar 2015]. Available at: http://www.pharmtech.com/pharmtech/content/printContentPopup.jsp?id=384716.

Wechsler FS. Fatoriais fixos desbalanceados: uma análise mal compreendida. Pesq Agropec Bras. 1998;33:231-62.