Acta Scientiarum
http://www.periodicos.uem.br/ojs/
ISSN on-line: 1807-8621
Doi: 10.4025/actasciagron.v45i1.56882

REVIEW

A brief review of the classic methods of experimental statistics

André Mundstock Xavier de Carvalho[1*], Fabrícia Queiroz Mendes[1], Pedro Henrique de Castro Borges[1] and Matthew Kramer[2]

[1]Instituto de Ciências Agrarias, Universidade Federal de Viçosa, Campus Rio Paranaíba, Rodovia BR-230, km 7, Cx. Postal 22, 38810-000, Rio Paranaíba, Minas Gerais, Brazil. [2]U.S. Department of Agriculture, Agricultural Research Service, Northeast Area, Statistics Group, Beltsville, Maryland, United States of America. *Author for correspondence. E-mail: andre.carvalho@ufv.br

ABSTRACT. Experimental statistics are a key element for innovation in the agricultural sector. Commonly used statistical methods in experimentation are relatively simple, reliable, and widely used. However, the many problems in the quality of statistical analyses reported in the agricultural science literature highlight a need for continuing discussion on and updating of this topic. This article reviews critical points about classic linear models procedures commonly used in agricultural statistics, frequent procedures in publications in the agricultural sciences. Due to the evolution of statistical science some common recommendations from the past should no longer be followed.

Keywords: ANOVA assumptions; parametric statistics; agricultural experimentation.

Received on November 29, 2020.
Accepted on February 17, 2021.

Introduction

Analysis of variance (ANOVA) completes its first centenary with no signs of aging. The genius and simplicity of this analysis has allowed it to become the basic reference for analyzing data from experiments, which is important in the agricultural sciences. Even though it was developed long before the popularization of simulation studies, the ANOVA F-test has repeatedly demonstrated its good qualities.

The quality of statistical analysis methods can be assessed using many dimensions, but three deserve to be highlighted: i) type I error rates (at least family-wise) can be accurately set, ii) good power, the ability to detect real differences, even if of small magnitude, and iii) simplicity. We have included this last dimension pragmatically, although it is difficult to define (Chater & Vitanyi, 2003). Considering that most scientists use statistical methods only as tools, without simplicity a method of analysis is unlikely to be widely adopted and understood by non-statisticians unless there is no other option. Furthermore, scientific "negationism" opens up the need to popularize science and demystify its methods, corroborating the appeal of simplicity (Little, 2013). Finally, for most of the scientific community, a simple, with widely recognized reasoning research has more value than something complex and of dubious validity (Volpato, 2010). Of course, these arguments should not be used to justify the non-adoption of new methods of analysis. After all, statistical science is also dynamic and is continuously offering new alternatives and methods for analysis.

In addition to ANOVA, means tests and regression analyses comprise the majority of statistical tests most used in agronomic experimentation (Tavares, Carvalho, & Machado, 2016; Kramer, Paparozzi, & Stroup, 2016; Possatto Júnior et al., 2019). Among statistical models, the vast majority of research in the agricultural sciences uses only simple models in a completely randomized design (CRD), randomized blocks (RBD), and split-plots, usually structured on just three or fewer factors (Lúcio et al., 2003; Tavares et al., 2016; Possatto Júnior et al., 2019). Note that, except for a CRD, these designs build on a simple all fixed effects ANOVA by including one or more additional random effects (so now in the framework of mixed models, see below). Despite the apparent simplicity of these experimental designs, is it possible to advance the quality and efficiency of the analysis of our experiments? Furthermore, in addition to the main textbooks on the topic, what is new in the discussion of these procedures? This brief review discusses these issues.

ANOVA requirements

First, remember that ANOVA is an analytical procedure that has prerequisites. While data do not need to come from experimental studies to be suitable for ANOVA, ANOVA was developed to analyze experiments

where the researcher has control over all experimental conditions. This means that one must be careful using ANOVA in other frameworks, for example, for observational studies, which are common in the biological sciences and in the area of soil management, although some of these can be analyzed using ANOVA. As in a correlation analysis, the basic problems using ANOVA for observational studies are bias (systematic error) and confounding (when the outcome is affected by variables ignored by the model). In an observational study there will always be a lower level of confidence in establishing cause and effect relationships between predictors and response variables than in an experimental study because one can always hypothesize that the dependent variable was affected by unmeasured underlying variables whose effects were not removed through randomization.

Nevertheless, observational studies can usually be drawn from far larger sample sizes than is practical in an experimental study and are informative if bias and confounding can be adequately controlled. They allow for data collection in real situations, allowing for greater participation by companies, farmers, and agriculture extension personnel in research. Basic precautions for validity in these studies are given in Casler (2015) and Tavares et al. (2016), and for conditions satisfying ANOVA, see Ferreira, Cargnelutti Filho, and Lúcio (2012). For geo-based data, hierarchical models (with nested random effects, e.g. sites within regions within states) will likely yield more appropriate results than those from a classic fixed model.

There are four general prerequisites of ANOVA (assuming the design has removed biases and confounding, and the model includes all important independent variables, appropriately scaled, and their interactions): error (residual) normality, error homogeneity, model additivity, and error independence (Montgomery, 2019).

Normality

The normality of the residuals refers to the distribution pattern of the deviations of each observation from its respective average (typically the cell mean). However, when the model is not a simple CRD, the deviations must be calculated according to the model, and not just by the difference between the observation and the treatment mean. Otherwise, a large effect for a particular block may be misinterpreted as a heavy tail in a non-normal distribution. Note that the Gaussian distribution is a continuous distribution (as opposed to a discrete one, like counts; see above). Although the concept of a continuous variable is relative (Gotelli & Ellison, 2011; Mann, 2015) due to the limited sensitivity of instruments, it may not make sense to assume one is sampling from a continuous distribution.

The Gaussian pattern of error distribution is very common in nature, at least approximately (Eidous & Al-Salman, 2016). This condition was used by statisticians to create familiar tests (*t*-tests, *F*-tests) that can determine, with great sensitivity, whether one population differs from another. In statistical language, sensitivity to detect a difference is synonymous with the "power" of the test. This means that the non-parametric Kruskal–Wallis, Friedman, Mann–Whitney, Wilcoxon, and Nemenyi tests, among others, are in general less able to perceive real differences between the two treatments than the corresponding parametric ones (Thorpe & Holland, 2000). Since we want experiments with efficiency to detect differences, we choose parametric tests first, but only if their assumptions are met.

Before discussing classic ANOVA in more detail, we need to mention that the linear models framework (which includes ANOVA and regression) can be considered a subset of a larger group of models, generalized linear mixed-effect models (often abbreviated as GLMM's). In these models effects can be considered fixed (under the experimenter's control or chosen by the experimenter, such as which varieties to plant) or random (where the levels of the factor are a representative sample of all possible levels, e.g. plots in a field). The inference space differs, for fixed effects the inference space is only to those levels used in the experiment, for random effects it is to the population that the levels were drawn from. Because of the additional uncertainty arising from sampling a population of levels, estimates of standard errors of means will be larger for a model with random effects, other things being equal. Our discussion does include models with random effects, as these are common in experimental designs used in agriculture and natural resources, but is not exhaustive.

The 'generalized' part of GLMM's refers to the allowed distributions for data (basically, they have to be members of the exponential family of distributions; the Gaussian distribution is one). Count data often do not follow a normal distribution, but rather a Poisson-like distribution, where the variance of the data increases as the mean increases. So, theoretically, data like these would not satisfy the assumption given above, error homogeneity. That is, residuals tend to increase in absolute value if they are associated with larger means. Percent data, or binomial data, such as the number of positives per total count, will also not satisfy the assumptions for normally distributed data given above. Often, data like these were transformed,

e.g. taking logs, square roots, square root of the arcsine, so that residuals would better satisfy the ANOVA assumptions necessary for valid *p*-value calculations. While these (mostly) work, and are useful as a check, a better approach for data that clearly match a known distribution (e.g. Poisson) and clearly are not Gaussian might be to use software that can directly estimate GLMM's with the appropriate underlying distribution. We do not cover GLMM's in this paper, but recommend Stroup (2013). Agricultural examples can also be found in Gbur et al. (2012).

Another important statistical framework that is becoming increasingly popular is Bayesian analysis. Like GLMM's, one can construct models based on many statistical distributions, and one can identify factors as fixed or random. Modeling is done in a somewhat different way, and results are also interpreted differently. For the same data set, results are usually similar to what one obtains using the 'frequentist' framework (traditional statistical methods), though confidence intervals on estimates tend to be a bit larger, especially for smaller data sets. Despite this, Bayesian approaches can increase the sensitivity of the tests in some cases by allowing the incorporation of a priori information. While we expect to see more analyses done in this framework in coming years, most statistical analyses in agriculture will likely continue to be done in the 'frequentist' framework, and we do not discuss Bayesian methods further, other than to note that it would be wise to monitor developments in this field and to use it when it is advantageous to do so (typically when there is good prior information that bears directly on the current experiment). For more details, we suggest Besag and Higdon (1999) and Ghosh, Delampady, and Samanta (2006).

Returning to classic ANOVA, how can we determine if the data have residuals with a Gaussian distribution? After the 1980s, with the advancement of simulation studies (Dambolena, 1986), it became evident that some tests for normality, such as the Kolmogorov–Smirnov test, do not have good sensitivity to detect non-normality in many situations (Razali & Wah, 2011; Torman, Coster, & Riboldi, 2012; Pino, 2014). However, we should not assume the existence of normality without verifying it, either formally by testing (preferable for less subjectivity) or using visual aids, such as Q-Q plots (Razali & Wah, 2011). While the F-test is considered to be relatively robust to slight violations of normality, it should be remembered that tests for normality are also not as powerful for detecting non-normality when the *n* is small (perhaps one should consider testing at 10% significance). If assumptions of normality are violated because they were not investigated, the conclusions from any subsequent *t* or *F* tests may be false (Lucena, Lopez, Pulgar, Abalos, & Valderrama, 2013) and the study may be discredited in the future. According to Ferreira (2014), the Tukey mean test, for example, can yield a Type I error rate (family-wise) of more than 50% (rather than the nominal 5%) when residuals are clearly not normal and with a large number of pairs of means being compared.

The most recommended tests for normality appear to be the Shapiro–Wilk, Anderson–Darling, and Jarque–Bera tests (Yazici & Yolacan, 2007; Torman et al., 2012). There may be other good tests, but these have been properly examined for their ability to detect non-normality in small samples. Unfortunately, all seem to have limited power when the total number of samples is less than 30 (Razali & Wah, 2011; Torman et al., 2012). A total *n* less than 15 is prohibitive to perform these tests, since they will not detect non-normality in most cases. For this reason, one should consider performing non-parametric analyses not only when non-normality is detected, but also when the total number of experimental/sample units is below 15 (Torman et al., 2012). Unfortunately, small data sets are exactly where one needs the most power to discriminate among groups. Razali and Wah (2011) argue that histograms and Q-Q plots can be subjective for assessing the distribution of residuals. We do not advocate any particular methodology for small data sets, but do insist that normality assumptions be checked as well as possible, and ANOVA not used for residues that have significant non-normality or non-homoscedasticity.

Homoscedasticity

The homogeneity of errors is the second basic requirement of ANOVA. Sometimes referred to as homoscedasticity or homogeneity of variances, it is important to be able to obtain a single, safe, and representative estimate of the experimental error as a whole (Casler, 2015). If homoscedasticity does not hold, the mean square of the ANOVA residuals will not be valid for some, or even most, comparisons. If this occurs, the results of some tests will have incorrect *p*-values attached, leading to erroneous conclusions.

This is a good place to bring up the problem of zeros in a data set, which can often lead to treatment combinations which have a very small or zero variance. This may happen if most plants die (so measures are zero) or don't grow well (perhaps given a disease without an effective treatment). Not only are tests involving *these* groups invalid because the homogeneity of variances assumption has been violated, tests on all the other

combinations are also affected because the small variance groups lower the average residual variance (and recall that outliers have an outsize effect on averages). If there are only a few treatment combinations with small variances and the rest would satisfy the homogeneity of variances assumption, we recommend deleting those combinations from the formal analysis (note that this will affect tests of interactions, etc.) so that the *p*-values that are reported are correct. The justification for removal, when applicable in a scientific paper, can be mentioned in the Methods section. If one needs to compare one of these groups with another used in the formal analysis, one can use a non-parametric test or not test at all (if all values are zero, there is no variance, and that group will differ from any group whose observations are most or all non-zero values).

If it is clear that the homogeneity of variances assumption has been violated, the *p*-values from an ANOVA will not be correct, and an alternative analysis is needed. The mixed-effect models can do variance groupings, which may help. Some transformation, defined *a priori* or not, may help. A non-parametric analysis may help, especially if the cell variances on ranked data are not too dissimilar.

Among tests for homoscedasticity, there is no consensus on which is the most suitable. Some very popular ones like the Levene test modified by Brown and Forsythe (1974) are, in fact, very liberal in some situations (Hines & O'Hara-Hines, 2000). This could be linked to the structural zeros that appear in the residuals column when calculating the deviations from the medians of a simple CRD experiment with an odd number of repetitions. On the other hand, Levene's original test is considered excessively conservative, detecting heteroscedasticity very easily (Sharma & Kibria, 2013). The Hartley or "*F*-max" test can be very permissive in some cases, in addition to not being developed for situations with more than 12 treatments or with treatments with variance equal to zero. As well as being very sensitive to the violation of normality (Sharma & Kibria, 2013), the Bartlett test is less suitable for split-plots and split-block designs because of the complexity of estimating the various variance estimates and correctly partitioning them for a valid Bartlett test, so that they consider only the residual error and ignore the uncertainty due to other random effects (plots/blocks).

Additivity

The additivity of the model is an important basic requirement since the effects need to be separated into a sum of parts to be properly compared to the magnitude of the error. Although the model can have several components, the most worrying violation of additivity is the non-additive effect between blocks and treatments since other interactions can usually be estimated in the models. If the effect of each block cannot be separated from the effect of each treatment in an additive way (that is, the treatment effect is the same in each block), there is an interaction between blocks and treatments. When this occurs, there is no way to distinguish the block effect from the treatment effect, compromising the entire analysis. In addition, lost unit estimates (for unbalanced data in RBD, split-plots, split-blocks and other) can be wrong when additivity is violated. Thus, in experiments with a random block effect, it is essential to estimate non-additivity.

Tukey's non-additivity test is the best-known procedure for the common situation of only one replication per block (Karabatsos, 2005; Alin & Kurt, 2006). It is of great concern that this requirement has been systematically ignored in agricultural research and in many other fields, such as cross-over trials in the pharmaceuticals industry. The analysis problem can also be solved with an experiment in which some (or all) of the treatments are replicated at least twice in each block, so the treatment by block interaction can be calculated in the usual way.

Independence

The fourth and final requirement, independence of errors, is the most complex requirement to be assessed. To date, there is no general test recommendation for widespread use that addresses all possible forms or patterns of violation of independence (Gotelli & Ellison, 2011). For this reason, this is the only ANOVA requirement that can be assumed without being formally tested. It can be assumed based on a theoretical assessment of randomization and "physical independence" of experimental/observational units, but this should be verified by investigating the common forms of non-independence, such as those based on geography (spatial correlation) or time (time-series dependencies). In observational studies, important missing independent variables (whether measured or not measured) can also create residuals that are not independent. Some software can relatively easily estimate and test models with many kinds of simple correlated residual error structures, and those could be used to determine if the independence assumption is violated.

Not infrequently in agricultural experimentation, independence between experimental units is not respected when there are excessively small and geographically close experimental units. Since the result of

an experimental unit, in these cases, does influence the results of the neighboring units, the analysis should be done using a model that allows for spatially correlated errors (Ferreira, 2019) or that allows a partial correction of this interference (Papadakis method for ANCOVA). Another relatively common situation of violation of independence is when there are no distinct experimental units (physically or geographically distinct) (Kramer et al., 2016), for example, in incubation experiments when comparing contents from the same flask incubated in several successive times. This is a kind of repeated measures design and needs to be analyzed in that framework, not in one where the successive times are considered levels of a factor. The result observed in the first measurement is likely to affect that of the second measurement, with no guarantee of independence between measurements (Alvarez V. & Alvarez, 2013).

The same reasoning may apply to evaluations in successive agricultural years or successive harvests. It also applies to successive soil layer assessments (Ferreira, 2019). For example, if the residual effect of a given fertilizer on the soil is to be evaluated, the result observed in layers 0 - 10 cm is not independent of those observed in layers 10 - 20 cm in the same experimental unit. Therefore, if there is not an independent experimental unit for each layer to be evaluated, these layers will have to be considered only as different response variables or the model should account for this dependency structure. Otherwise, the model artificially inflates the degrees of freedom (DF) of the residual and increases the type I error rates of the tests that will be applied to these data (Kramer et al., 2016). The confusion about the adequacy of these experiments, exemplified above, may be in part related to the misinterpretation of what split-plot and split-block experiments are, which we address below.

If any of the requirements for ANOVA are violated, what are the available paths for analyzing the data? Data with a distribution other than Gaussian, that are heteroscedastic, or with non-additive effects should not be synonymized with experiments that are "poorly conducted" or "with problems". The most recommended procedure in these cases is to assess whether a transformation in the scale can adequately transform the data to satisfy ANOVA requirements (Piepho, 2009; Ribeiro-Oliveira, Santana, Pereira, & Santos, 2018), alternatively to consider a GLMM model (Kramer, Paparozzi, & Stroup, 2019). There is not an established rule between the nature or source of the data and the ideal transformation (O'Hara & Kotze, 2010).

In addition to the classic log, root, and angular transformations, the Box–Cox transformation family (Box & Cox, 1964) can be useful. If it is still not possible to meet the requirements of ANOVA using transformation or a GLMM approach, non-parametric methods can be tried. In many cases, the simple ranking of data (rank transformation) will allow an ANOVA on ranks and subsequent tests that are adequately sensitive and safe to carry out the comparisons of interest of the research (Conover & Iman, 1981; Zimmermann, 2004; Zimmerman, 2012; Conover, 2012). In addition, the problem of estimating the interaction in factorials under rank transformation can be overcome using aligned rank transformation (ART) (Wobbrock, Findlater, Gergle, & Higgins, 2011). In some situations, there are specific non-parametric tests that are more powerful. Bootstrap versions of mean tests can also be more powerful than the same tests under ranked data. Note that inferences on ranked data are on median values, not means. Also, there is no way to back-transform results to the original scale. So, while one may want to give results of statistical testing on ranked transformed data, figures showing data should be on the original scale.

Tests for comparisons of means

Most agricultural research uses a posteriori Tukey test to compare means almost indiscriminately. Although this test has good control of real type I error rates (family-wise), there are better options. What are the "real type I error rates"? Type I error (α) is the probability that a test indicates that a difference is significant when, in reality, it is not. The frequency of error α can be estimated each time the test is applied individually (comparison-wise), for the set of times it is applied within each response variable (family-wise), or other ways. We will not go into details here about the distinction between error rates by comparison, family, or experiment (Keselman, 2015). In experimentation, there is no reason to trust tests that do not control error rates (family-wise) since we almost always perform several comparisons for each response variable and many false positives can represent a considerable waste of time and money. For example, in an experiment with eight treatments, a multiple comparison test like the Tukey test is applied 28 times ($C_{8;2}$). With so many comparisons, it is easy to understand that the risk of at least one of them resulting in a false positive is large (Keselman, 2015). Not coincidentally, the Tukey test statistic considers higher tabulated q values as the number of treatments that are compared increases.

The first basic dilemma of statistical tests is that they were developed to control type I or type II error rates (false negatives). We usually choose to control what is considered to be the most serious, which is the type I error. In other words, our usual averaging tests are almost always accompanied by larger type II errors. When one treatment does not differ statistically from another, we should state that "there was not enough evidence to say that they are different". Note that this is not the same as saying that "the averages are equal", which might be the null hypothesis in words. After all, absence of evidence is not synonymous with evidence of an absence of an effect. If all one can say is "means do not differ statistically," it may also lead to difficulty publishing results, since results are "inconclusive," especially in the face of a low *n* and a high coefficient of variation (Onofri, Carbonell, Piepho, Mortimers, & Cousens, 2010; Loureiro & Gameiro, 2011).

The second basic dilemma of the tests is that to better control type I error, power or sensitivity is lost (Gelman, Hill, & Yajima, 2012). In most cases, we design experiments so that they are sensitive to detect real differences between treatments, even if they are small differences. It is not a function of the statistical test to judge the importance of the difference (Loureiro & Gameiro, 2011; Kramer et al., 2019) because in some cases, a 2% increase can be important, and up to a 20% increase can be considered irrelevant in others. Effect-size statistics, like the d-Cohen statistics, can help in this regard (Loureiro & Gameiro, 2011). To get an idea of the severity of this dilemma, some power estimates of the Tukey test applied with a nominal α of 5% can be consulted. Even in experiments with a coefficient of variation of only 10%, Conagin, Ambrosano, and Nagai (1997) and Conagin and Pimentel-Gomes (2004) observed that the test was able to detect real differences of 20% between means only 30% of the time. However, if a Holm-test was applied only to a few comparisons defined *a priori*, the power could reach 70%.

Among the most well-known tests that adequately control the real type I error rates per family, simulation studies show that Holm, Tukey, Student–Newman–Keuls (SNK), Bonferroni, and Dunnett satisfy this criterion (Perecin & Barbosa, 1988; Borges & Ferreira, 2003; Conagin, Barbin, & Demétrio, 2008; Girardi, Cargnelutti Filho, & Storck, 2009; Sousa, Lira Júnior, & Ferreira, 2012; Gonçalves, Ramos, & Avelar, 2015). Of these, the SNK is the most controversial, but a large volume of Monte Carlo studies corroborate that it has reasonable error control capacity (Perecin & Barbosa, 1988; Borges & Ferreira, 2003; Conagin, Barbin, & Demétrio, 2008; Girardi et al., 2009; Gonçalves et al., 2015). LSD (*t* for multiple comparisons) and Duncan tests do not adequately control false positives and should not be used (Perecin & Barbosa, 1988; Conagin et al., 2008; Girardi et al., 2009; Sousa et al., 2012). Among the less popular, the Bonferroni test modified by Conagin et al. (2008) and the Scott–Knott cluster analysis also controls α error rates relatively well (Borges & Ferreira, 2003; Conagin et al., 2008). However, the latter should be used sparingly since, in many situations, the ambiguity of the performance of treatments can be natural and inherent to the phenomenon under study. In addition, in situations of partial nullity, the Scott-Knott test (in reality a procedure to group means) may have somewhat inflated Type I error rates (Borges & Ferreira, 2003).

In terms of the power of these tests, we can order them from the most to the least sensitive: Holm (for a few comparisons) > Dunnett > Scott-Knott > Bonferroni modified by Conagin ~ SNK > Tukey > Scheffé. Although it is only an approximate order, it helps us to understand why planned comparisons should be the first option when they satisfy the research objectives. It is possible that planned contrasts are seldom used due to the difficulty of connecting scientific hypotheses and comparisons of interest (Yossa & Verdegem, 2015). In all tests, Fisher's protection criterion may be respected, although this is not mandatory. The same reasoning applies to regression analyses. In the case of factorial experiments, even the factorial split should only be carried out if there is a global effect for treatments (preliminary ANOVA) (Barbosa & Maldonado Júnior, 2015).

Univariate regression analysis under experimental conditions

Quantitative treatments or predictors are those that are defined by numbers or doses. Whenever the treatments are quantitative and there are more than three levels, a regression analysis is recommended (Piepho & Edmondson, 2018; Possatto Júnior et al., 2019). In addition to being more sensitive than treating treatments as qualitative, such analyses are more consistent with the continuous effect of treatment levels (Piepho & Edmondson, 2018). A regression analysis relates changes in the dependent variable to changes in an independent variable (quantitative treatment levels) in an 'empirical' way, since estimates of the intercept and slope of the line are calculated from the data, and not from the 'functional' or theoretical relationship (e.g. based on the underlying biology), which would likely be complicated and involve non-linear relationships of many independent variables.

In agricultural experimentation, univariate regression analysis (only one predictor or causative variable) is more common than multivariate (two or more predictors). Multivariate regression analysis is widely used in observational studies, as an exploratory technique to model correlations between *n* supposedly causal variables and a response variable of interest. For more details on this technique, we suggest Alexopoulos (2010). Despite the usefulness of regression analysis for understanding many phenomena, there are questions about how to best perform it. In experimentation, such analysis always involves two steps. The first is to fit different regression models (linear, quadratic, root, exponential, etc.) to the data. In the second stage, the quality of this model fit should be tested to determine which models are statistically adequate, and perhaps rank the models. In other words, whenever it is stated that "the data were subjected to regression analysis", it would be useful to state which models were considered.

In experimental statistics, where true replications are almost always used, the quality of the adjustment can be assessed in a simple way using two criteria: i) the quality of the fit of the model to the data (assessed by the significance of the regression in the regression's ANOVA), and ii) the level of deviation between the data and the model fit, also known as "lack of fit" (Cecon, Silva, Nascimento, & Ferreira, 2012; Dancey, Reidy, & Rowe, 2017). The regression ANOVA is a simple decomposition of the sum of squares of treatments (SSTreat) into two parts: one that can be explained by the model in question and one that cannot (the lack of fit). A regression model will be adequate if the part of SSTreat that it is able to explain (or estimate) is significant according to the *F*-test (Cecon et al., 2012; Dancey, Reidy, & Rowe, 2017). In addition, if the model is appropriate, the part of SSTreat not explained by it must be "negligible" (i.e., non-significant lack of fit). Therefore, the ANOVA of the regression performed separately for each model to be tested is sufficient for determining those models that are statistically adequate. Testing the significance of each model parameter via the t-test is redundant in experimental conditions with true replications (Dancey et al., 2017). It is important that the ANOVAs of the regressions are performed separately for each model tested so that the lack of fit is correctly estimated by the difference between the SSTreat and the sum of squares of the respective model.

There are two additional important details to consider. The model chosen must also be a parsimonious model, that is, one with relatively fewer estimated parameters. In this sense, the value of the model's "R^2 adjusted" (R^2_a) is of great value because it shows that the largest R^2 does not always correspond to the best model. Other criteria have been developed for this purpose, like AIC, BIC, etc. Finally, some authors maintain that the model should have an adequate theoretical interpretation, consistent with the phenomenon studied (Alvarez & Alvarez, 2003). For example, in a classic experiment involving levels of any resource (water, nutrient, etc.), the expected response in growth is exponential followed by stabilization. This expectation can be used as a tiebreaker between two statistically adequate models with very close R^2_a. Thus, the 'empirical' model should reasonably match the 'functional' model.

A common problem in regression analysis in agricultural research is that only linear, quadratic, and cubic models are often tested (Possatto Júnior et al., 2019). Exponential models, such as those that predict exponential growth followed by stabilization, are still underused, possibly due to the difficulty of performing calculations to estimate parameters (Mazucheli & Achcar, 2002; Kniss, Vassios, Nissen, & Ritz, 2011). These models have a good theoretical interpretation of various phenomena, particularly the growth responses of plants or animals due to the availability of resources (Freitas, 2005). Some details about common non-linear models, such as the Mitscherlich and Logistic models are addressed by Carvalho (2023). On the other hand, the cubic model makes a poor theoretical one. Although widely used, its oscillatory behavior ("S shaped" curve) rarely has a biological explanation if we assume that the treatments were properly isolated and controlled. Furthermore, it is less parsimonious because it has three parameters for shape versus only two regression parameters for the exponential models.

Repeated measures on the same experimental units

There is a relatively common confusion between the concept of split-plot in time and split-plot in space as a synonym for ANOVA for repeated measures in time or space. In reality, split-plot schemes require independent experimental units to exist, even though they alter the complete randomization pattern of the design. When there are truly independent experimental units (including physically independent ones) for different evaluation periods or layers, some researchers consider it reasonable to accept these experiments as simple factorials (Gotelli & Ellison, 2011), because there is consistency between the number of DF of the error and the true number of experimental units. However, it is important to note that this analysis ignores that there were restrictions on how units were randomized to times or to the successive layers of evaluation.

Some authors argue that, instead of split-plots, it is valid to consider the split-block model (Alvarez V. & Alvarez, 2013; Steel, Torrie, & Dickey, 1997). However, it should be noted that the combined errors (Satterthwaite correction for DF (Steel et al., 1997)) are not valid in this case, and the main treatments should be compared using only the residue "a" (Snedecor & Cochran, 1989). A better option than split-blocks is to perform an ANOVA for repeated measures (Ferreira, 2019), where the correlation structure of the experimental units, due to the restrictions on randomization, is taken into account by estimating a residual covariance structure that explicitly models dependencies in a mixed models framework. Another option may be to consider geostatistics tools. For example, for observations distributed in space, one can describe the residual correlation structure using a geostatistical model, theme that goes beyond the scope of this review.

There are also multivariate strategies for analyzing repeated measures, but these are more complex and are not always more sensitive than univariate strategies. ANOVA for univariate repeated measures can be understood as a correction of the split-plot ANOVA. However, it requires the sphericity condition. Briefly, sphericity is the condition that the variances of the differences between the possible pairs of the levels of the non-independent factor are homogeneous. Despite its complexity, this condition can be verified, at least approximately, by the Mauchly test, the significance of which can be used to establish a correction for the DF of the factor whose levels are not independent (Greenhouse & Geisser, 1959; Huynh & Feldt, 1976). However, the most conservative value for this correction corresponds to 1/(p-1), allowing the analysis of repeated measures even when the sphericity condition is not known (Greenhouse & Geisser, 1959).

Alternatively, Vivaldi (1999) and Quinn and Keough (2002) highlight a simple option in cases where it is not possible to have independent experimental units over time. They suggest calculating rates, asymptotes, inflection points, or other parameters of interest separately for each group of correlated units. Subsequently, these parameters are compared as a new response variable. This works if each group of correlated units can be fit with a similar time series model, which could be as simple as a regression with time as the independent variable (in a ANOVA for repeated measures).

Some other procedures and final considerations

In most agricultural experiments, only ANOVA, means tests, and regression analyses are used, except in the area of plant breeding. Recently, some multivariate techniques have become more popular (Possatto Júnior et al., 2019) and some non-parametric analyses are occasionally performed. Among the multivariate techniques, principal component and cluster analyses are the most common, although they are traditionally applied to observational studies without predefined predictors. In addition, in controlled experimental conditions, multivariate indices can overcome some of the limitations of univariate analyses. Among these indices, the Mulamba-Mock rank sum index (Mulamba & Mock, 1978) and the Desirability index (Candioti, Zan, Camara, & Goicoechea, 2014) deserve mention. In some cases, these indices are a simple option for multivariate analysis of variance (MANOVA).

Multivariate analyses also have their requirements, but recommendations do not seem as well established. In addition to the complex assessment of what would be a normal multivariate distribution, in most multivariate analysis techniques it is also important to avoid multicollinearity (Yoo et al., 2014). There are a variety of ways one can do this, for example, by using variables resulting from a principal components decomposition instead of the original dependent variables or using the condition number of the dependent variable matrix (Manly, 1994). In cluster analysis, there still seems to be no consensus recommendation on the grouping method and the most appropriate stopping criterion in each situation (Gotelli & Ellison, 2011; Saraçli, Dogan, & Dogan, 2013).

A classic analysis that is still underutilized is the analysis of covariance (Tavares et al., 2016). The analysis of covariance is very useful when it is desired to separate the effect/interference of an unplanned variable (for example the variation in the plant stand or in the incidence of some disease between the plots in a field experiment) from another response variable of interest. Despite being a well-known procedure (Montgomery, 2019), it has some analytical complexities that seem to restrict its popularity. The estimation of the corrected variable is the most controversial step of the analysis, since there are several alternatives and there is no consensus on which correction algorithm is the most appropriate. Many researchers recognize that a simple correction by simple rule of three is not adequate. Schimildt, Cruz, Zanuncio, Pereira, and Ferrão (2001) present several methods, one of the simplest and usual being the correction by the "average value"

$$(Z_{ij} = Y_{ij} - b\,.\,(X_{ij} - \bar{X})$$

where Z_{ij} and Y_{ij} represent the corrected and original values of the response variable in the experimental unit belonging to the *i* treatment, *j* repetition; X_{ij} represents the covariate; and *b* is the linear regression coefficient as a function of X_{ij}.

More recently, outlier testing has gained more popularity. Although it is still a non-consensual subject, some tests for outliers have strong practical and theoretical support and should not be treated as a fraudulent device. Criticisms of the use of tests for outliers generally come from researchers in the field of ecology (Gotelli & Ellison, 2011), where research is not always conducted under controlled experimental conditions. Using a good test for outliers can prevent suspicious data from resulting in unrealistic significant effects for treatments (Onofri et al., 2010; Tavares et al., 2016).

The Grubbs-Beck test and ESD tests reviewed by Rosner (1983) have been the most suitable for outliers since they allow the magnitude of the experimental error to be considered as a whole (Manoj & Senthamarai-Kannan, 2013; Cohn et al., 2013). Although they are dependent on normality, they are quite conservative, which is important in this case. Even if we consider the critical values of the Rosner ESD test for many outliers simultaneously, the critical values are always superior to the Chauvenet or Grubbs tests.

Despite the analytical ease resulting from the popularization of statistical software, problems of misuse of statistical procedures in agricultural sciences are frequent (Montanhini Neto & Ostrensky, 2013; Tavares et al., 2016; Kramer et al., 2016; Possatto Júnior et al., 2019). Apparently, problems occur even with the development of free applications that are particularly simple to use, such as SISVAR (Ferreira, 2019), Assistat (Silva & Azevedo, 2016), SPEED Stat (Carvalho, Mendes, Mendes, & Tavares, 2020), AgroEstat (Barbosa & Maldonado Júnior, 2015), BioEstat (Ayres, Ayres Júnior, Ayres, & Santos, 2007), RBio (Bhering, 2017), and others. This shows that, in many cases, the problem may also be related to the lack of knowledge about statistics. Considering that statistical science and its recommendations are always evolving, it is normal to expect some problems. However, we must also try to understand the reasons for these recurring problems. Why is there still little interest in statistics on the part of agricultural science students? If we look at statistics only as one of the tools for agricultural innovations, perhaps we have poorly balanced the time invested between teaching people how to use the software and teaching statistics itself. How can we best train and update researchers from different areas of the agricultural sciences in statistical science?

In this brief review, we have not exhausted the discussions on improvements in the use of classic statistical procedures applicable to agricultural experimentation. Statistics should not be seen as a set of rigid recommendations, and each experimental situation should be approached individually. As general suggestions, we strongly recommend reading Onofri et al. (2010), Casler (2015), and Kramer et al. (2019) and the good sense of seeking help from a statistician in the planning phase of more complex experimental situations.

Conclusion

The classic procedures of experimental statistics continue to be very useful for most agricultural research. However, they have requirements that are often ignored or poorly understood, resulting in incorrect conclusions. With the advancement of statistical science, the properties of some tests have been better understood, resulting in changes in recommendations. For example, Duncan, t-tests for multiple comparisons (LSD), and Kolmogorov–Smirnov tests for small samples are no longer recommended. It is also not recommended to perform ANOVA without checking its assumptions under the justification that the F-test is robust. In addition, there is a growing understanding that repeated measures in time or space over the same experimental units should not be analyzed as simple split-plots. Other procedures and experimental models are more accessible and have become increasingly recommended, such as nested mixed models, models for repeated measures, regression for non-polynomial models, tests for outliers, and multivariate analysis.

References

Alexopoulos, E. C. (2010). Introduction to Multivariate Regression Analysis. *Hippokratia*, *14*(Suppl. 1), 23-28.

Alin, A., & Kurt, S. (2006). Testing non-additivity (interaction) in two-way ANOVA tables with no replication. *Statistical Methods in Medical Research*, *15*(1), 63-85. DOI: https://doi.org/10.1191/0962280206sm426oa

Alvarez V., V. H., & Alvarez, G. A. M. (2003). Apresentação de equações de regressão e suas interpretações. *Boletim Informativo da Sociedade Brasileira de Ciência do Solo*, *28*(3), 28-32.

Alvarez V., V. H., & Alvarez, G. A. M. (2013). Reflexões sobre a utilização de estatística para pesquisa em ciência do solo. *Boletim Informativo da Sociedade Brasileira de Ciência do Solo*, *38*(1), 28-35.

Ayres, M., Ayres Júnior, M., Ayres, D. L., & Santos, A. A. S. (2007). *BioEstat - Aplicações estatísticas nas áreas das ciências bio-médicas*. Belém, PA: Instituto Mamirauá.

Barbosa, J. C., & Maldonado Júnior, W. (2015). *Experimentação agronômica e AgroEstat: sistema para análises estatísticas de ensaios agronômicos*. Jaboticabal, SP: Multipress.

Besag, J., & Higdon, D. (1999). Bayesian analysis of agricultural field experiments. *Journal of the Royal Statistical Society - Statistical Methodology Series B*, *61*(4), 691-746. DOI: https://doi.org/10.1111/1467-9868.00201

Bhering, L. L. (2017). Rbio: A tool for biometric and statistical analysis using the R platform. *Crop Breeding and Applied Biotechnology, 17*(2), 187-190. DOI: https://doi.org/10.1590/1984-70332017v17n2s29

Borges, L. C., & Ferreira, D. F. (2003). Poder e taxas de erro tipo I dos testes Scott-Knott, Tukey e Student-Newman-Keuls sob distribuições normal e não normais dos resíduos. *Revista de Matemática e Estatística, 21*(1), 67-83.

Box, G. E. P., & Cox, D. R. (1964). An Analysis of Transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, *26*(2), 211-252.

Brown, M. B., & Forsythe, A. B. (1974). Robust tests for the equality of variances. *Journal of the American Statistical Association*, *69*(346), 364-367. DOI: https://doi.org/10.1080/01621459.1974.10482955

Carvalho, A. M. X. (2023). *Estatística Experimental e Observacional - uma nova abordagem sobre os métodos clássicos*. Rio Paranaíba, MG: Conselho Editorial da UFV-CRP.

Carvalho, A. M. X., Mendes, F. Q., Mendes, F. Q., & Tavares, L. F. (2020) SPEED Stat: a free, intuitive, and minimalist spreadsheet program for statistical analyses of experiments. *Crop Breending and Applied Biotechnology, 20*(3), 1-6. DOI: https://doi.org/10.1590/1984-70332020v20n3s46

Casler, M. D. (2015). Fundamentals of experimental design: Guidelines for designing successful experiments. *Agronomy Journal*, *107*(2), 692-705. DOI: https://doi.org/10.2134/agronj2013.0114

Candioti, L. V., De Zan, M. M., Cámara, M. S., & Goicoechea, H. C. (2014). Experimental design and multiple response optimization. Using the desirability function in analytical methods development. *Talanta*, *124*, 123-138. DOI: https://doi.org/10.1016/j.talanta.2014.01.034

Cecon, P. R., Silva, A. R., Nascimento, M., & Ferreira, A. (2012). *Métodos estatísticos* (Série Didática). Viçosa, MG: Editora da UFV.

Chater, N., & Vitányi, P. (2003). Simplicity: A unifying principle in cognitive science? *Trends in Cognitive Sciences*, *7*(1), 19-22. DOI: https://doi.org/10.1016/S1364-6613(02)00005-0

Cohn, T. A., England, J. F., Berenbrock, C. E., Mason, R. R., Stedinger, J. R., & Lamontagne, J. R. (2013). A generalized Grubbs-Beck test statistic for detecting multiple potentially influential low outliers in flood series. *Water Resources Research*, *49*(8), 5047-5058. DOI: https://doi.org/10.1002/wrcr.20392

Conagin, A., Ambrosano, G. M. B., & Nagai, V. (1997). Poder discriminativo da posição de classificação e dos testes estatísticos na seleção de genótipos. *Bragantia*, *56*(2), 403-417. DOI: https://doi.org/10.1590/S0006-87051997000200019

Conagin, A., & PimenteL-Gomes, F. (2004). Escolha adequada dos testes estatísticos para comparações múltiplas. *Brazilian Journal of Agriculture*, *79*(3), 288-295. DOI: https://doi.org/10.37856/bja.v79i3.1392

Conagin, A., Barbin, D., & Demétrio, C. G. B. (2008) Modifications for the Tukey test procedure and evaluation of the power and efficiency of multiple comparison procedures. *Scientia Agricola, 65*(4), 428-432. DOI: https://doi.org/10.1590/S0103-90162008000400016

Conover, W. J., & Iman, R. L. (1981). Rank transformations as a bridge between parametric and nonparametric statistics. *The American Statistician*, *35*(3), 124-129. DOI: https://doi.org/10.1080/00031305.1981.10479327

Conover, W. J. (2012). The rank transformation—an easy and intuitive way to connect many nonparametric methods to their parametric counterparts for seamless teaching introductory statistics courses. *WIREs Computational Statistics*, *4*(5), 432-438. DOI: https://doi.org/10.1002/wics.1216

Dambolena, I. G. (1986). Using simulation in statistics courses. *Collegiate Microcomputer*, *4*(4), 339-344.

Dancey, C. P., Reidy, J. G., & Rowe, R. (2017). *Estatística sem matemática para as ciências da saúde*. Porto Alegre, RS: Penso.

Eidous, O., & Al-Salman, S. (2016). One-term approximation for normal distribution function. *Mathematics and Statistics*, *4*(1), 15-18. DOI: https://doi.org/10.13189/ms.2016.040102

Ferreira, D. F. (2019). Sisvar: a computer analysis system to fixed effects split plot type designs. *Revista Brasileira de Biometria*, *37*(4), 529-535. DOI: https://doi.org/10.28951/rbb.v37i4.450

Ferreira, D. F. (2014). Sisvar: a guide for its bootstrap procedures in multiple comparisons. *Ciência e Agrotecnologia*, *38*(2), 109-112. DOI: https://doi.org/10.1590/S1413-70542014000200001

Ferreira, D. F., Cargnelutti Filho, A., & Lúcio, A. D. (2012). Procedimentos estatísticos em planejamentos experimentais com restrições na casualização. *Boletim Informativo da Sociedade Brasileira de Ciência do Solo*, *37*, 1-35.

Freitas, A. R. (2005). Curvas de crescimento na produção animal. *Revista Brasileira de Zootecnia*, *34*(3), 786-795. DOI: https://doi.org/10.1590/S1516-35982005000300010

Gauch H. G. (1992). *Statistical analysis of regional yield trials: AMMI analysis of factorial designs*. Amsterdam, NT: Elsevier.

Gbur, E. E., Stroup, W. W., McCarter, K. S., Durham, S., Young, L. J., Christman, M., ... Kramer, M (2012) *Analysis of Generalized Linear Mixed Models in the Agricultural and Natural Resources Sciences*. Madison, US: American Society of Agronomy, Crop and Soil Science Society of America, Inc.

Gelman, A., Hill, J., & Yajima, M. (2012) Why we (usually) don't hve to worry about multiple comparisons. *Journal of Research on Educational Effectiveness, 5*(2), 189-211. DOI: https://doi.org/10.1080/19345747.2011.618213

Ghosh, J. K., Delampady, M., & Samanta, T. (2006). *An introdution to Bayesian analysis - Theory and methods*. New York, NY: Springer.

Girardi, L. H., Cargnelutti Filho, A., & Storck, L. (2009). Erro tipo I e poder de cinco testes de comparação múltipla de médias. *Revista Brasileira de Biometria*, *27*(1), 23-36.

Gonçalves, B. A., Ramos, P. S., & Avelar, F. G. (2015). Teste de Student-Newman-Keuls bootstrap: proposta, avaliação e aplicação em dados de produtividade da graviola. *Revista Brasileira de Biometria*, *33*(4), 445-470.

Gotelli, N. J., & Ellison, A. M. (2011). *Princípios de estatística em ecologia*. Porto Alegre, RS: Artmed.

Greenhouse, S. W., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika*, *24*(2), 95-112. DOI: https://doi.org/10.1007/BF02289823

Hines, W. G. S., & O'Hara-Hines, R. (2000). Increased power with modified forms of the Levene (Med) test for heterogeneity of variance. *Biometrics*, *56*(2), 451-454. DOI: https://doi.org/10.1111/j.0006-341X.2000.00451.x

Huynh, H., & Feldt, L. S. (1976). Estimation of the Box correction for degrees of freedom from sample data in randomized block and split-plot designs. *Journal of Educational Statistics*, *1*(1), 69-82. DOI: https://doi.org/10.3102/10769986001001069

Keselman, H. J. (2015). Per family or familywise type i error control: "Eether, eyether, neether, nyther, let's call the whole thing off!". *Journal of Modern Applied Statistical Methods*, *14*(1), 24-37. DOI: https://doi.org/10.22237/jmasm/1430453100

Karabatsos, G. (2005). Additivity tests. *Encyclopedia of Statistics in Behavioral Science, 1*(2), 25-29. DOI: https://doi.org/10.1002/0470013192.bsa009

Kniss, A. R., Vassios, J. D., Nissen, S. J., & Ritz, C. (2011) Nonlinear regression analysis of herbicide absorption studies. *Weed Science, 59*(4), 601-610. DOI: https://doi.org/10.1614/WS-D-11-00034.1

Kramer, M. H., Paparozzi, E. T., & Stroup, W. W. (2016). Statistics in a horticultural journal: Problems and solutions. *Journal of the American Society for Horticultural Science*, *141*(5), 400-406. DOI: https://doi.org/10.21273/JASHS03747-16

Kramer, M. H., Paparozzi, E. T., & Stroup, W. W. (2019). Best practices for presenting statistical information in a research article. *HortScience*, *54*(9), 1605-1609. DOI: https://doi.org/10.21273/HORTSCI13952-19

Little, R. J. (2013). In praise of simplicity not mathematistry! Ten simple powerful ideas for the statistical scientist. *Journal of the American Statistical Association*, *108*(502), 359-369. DOI: https://doi.org/10.1080/01621459.2013.787932

Loureiro, L. M. J., & Gameiro, M. G. H. (2011). Interpretação crítica dos resultados estatísticos: para lá da significância estatística. *Revista de Enfermagem Referência*, *3*(3), 151-162. DOI: https://doi.org/10.12707/RIII1009

Lucena, C., Lopez, J. M., Pulgar, R., Abalos, C., & Valderrama, M. J. (2013). Potential errors and misuse of statistics in studies on leakage in endodontics. *International Endodontic Journal*, *46*(4), 323-331. DOI: https://doi.org/10.1111/j.1365-2591.2012.02118.x

Lúcio, A. D. C., Lopes, S. J., Storck, L., Carpes, R. H., Lieberknecht, D., & Nicola, M. C. (2003). Características experimentais das publicações da Ciência Rural de 1971 a 2000. *Ciência Rural*, *33*(1), 161-164. DOI: https://doi.org/10.1590/S0103-84782003000100026

Manly, B. F. J. (1994). *Multivariate statistical methods – A primer*. London, UK: Chapman & Hall.

Mann, P. S. (2015). *Introdução à estatística* (8. ed.). Rio de Janeiro, RJ: LTC.

Manoj, K., & Senthamarai-Kannan, K. (2013). Comparison of methods for detecting outliers. *International Journal of Scientific & Engineering Research*, *4*(9), 709-714.

Mazucheli, J., & Achcar, J. A. (2002). Algumas considerações em regressão não linear. *Acta Scientiarum. Technology*, *24*(6), 1761-1770. DOI: https://doi.org/10.4025/actascitechnol.v24i0.2551

Montanhini Neto, R., & Ostrensky, A. (2013). Assessment of the use of statistical methods in articles published in a journal of veterinary science from 2000 to 2010. *Acta Scientiarum. Technology, 35*(1), 97-102. DOI: https://doi.org/10.4025/actascitechnol.v35i1.13753

Montgomery, D. C. (2019). *Design and analysis of experiments*. Danvers, US: Wiley.

Mulamba, N. N., & Mock, J. J. (1978). Improvement of yield potential of the ETO blanco maize (*Zea mays* L.) population by breeding for plant traits [Mexico]. *Egyptian Journal of Genetics and Cytology*, *7*(1), 40-57.

O'Hara, R. B., & Kotze, D. J. (2010). Do not log-transform count data. *Methods in Ecology and Evolution*, *1*(2), 118-122. DOI: https://doi.org/10.1111/j.2041-210X.2010.00021.x

Onofri, A., Carbonell, E. A., Piepho, H. P., Mortimer, A. M., & Cousens, R. D. (2010). Current statistical issues in Weed Research. *Weed Research*, *50*(1), 5-24. DOI: https://doi.org/10.1111/j.1365-3180.2009.00758.x

Perecin, D., & Barbosa, J. C. (1988). Uma avaliação de seis procedimentos para comparações múltiplas. *Revista de Matemática e Estatística*, *6*(1), 95-104.

Piepho, H. P. (2009). Data transformation in statistical analysis of field trials with changing treatment variance. *Agronomy Journal*, *101*(4), 865-869. DOI: https://doi.org/10.2134/agronj2008.0226x

Piepho, H. P., & Edmondson, R. N. (2018). A tutorial on the statistical analysis of factorial experiments with qualitative and quantitative treatment factor levels. *Journal of Agronomy and Crop Science*, *204*(5), 429-455. DOI: https://doi.org/10.1111/jac.12267

Pino, F. A. (2014). A questão da não normalidade: uma revisão. *Revista de Economia Agrícola*, *61*(2), 17-33.

Possatto Júnior, O., Bertagna, F. A. B., Peterlini, E., Baleroni, A. G., Rossi, R. M., & Zeni Neto, H. (2019). Survey of statistical methods applied in articles published in Acta Scientiarum. Agronomy from 1998 to 2016. *Acta Scientiarum. Agronomy*, *41*(1), 1-10. DOI: https://doi.org/10.4025/actasciagron.v41i1.42641

Quinn, G. P., & Keough, M. J. (2002). *Experimental design and data analysis for biologists*. New York, NY: Cambridge University Press.

Razali, N. M., & Wah, Y. B. (2011). Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *Journal of Statistical Modeling and Analytics*, *2*(1), 21-33.

Ribeiro-Oliveira, J. P., Santana, D. G. D., Pereira, V. J., & Santos, C. M. D. (2018). Data transformation: an underestimated tool by inappropriate use. *Acta Scientiarum. Agronomy*, *40*(1), 1-11. DOI: https://doi.org/10.4025/actasciagron.v40i1.35300

Rosner, B. (1983). Percentage points for a generalized ESD many-outlier procedure. *Technometrics*, *25*(2), 165-172. DOI: https://doi.org/10.1080/00401706.1983.10487848

Saraçli, S., Doğan, N., & Doğan, İ. (2013). Comparison of hierarchical cluster analysis methods by cophenetic correlation. *Journal of Inequalities and Applications*, *203*(1), 1-8. DOI: https://doi.org/10.1186/1029-242X-2013-203

Schmildt, E. R., Cruz, C. D., Zanuncio, J. C., Pereira, P. R. G., & Ferrão, R. G. (2001). Avaliação de métodos de correção do estande para estimar a produtividade em milho. *Pesquisa Agropecuária Brasileira*, *36*(8), 1011-1018. DOI: https://doi.org/10.1590/S0100-204X2001000800002

Silva, F. A. S., & de Azevedo, C. A. V. (2016). The Assistat Software Version 7.7 and its use in the analysis of experimental data. *African Journal of Agricultural Research*, *11*(39), 3733-3740. DOI: https://doi.org/10.5897/AJAR2016.11522

Sharma, D., & Kibria, B. M. G. (2013). On some test statistics for testing homogeneity of variances: a comparative study. *Journal of Statistical Computation and Simulation*, *83*(10), 1944-1963. DOI: https://doi.org/10.1080/00949655.2012.675336

Snedecor, G. W., & Cochran, W. G. (1989) *Statistical methods* (8th ed.). Iowa, US: ISU Press.

Sousa, C. A. D., Lira Junior, M. A., & Ferreira, R. L. C. (2012). Avaliação de testes estatísticos de comparações múltiplas de médias. *Revista Ceres*, *59*(3), 350-354. DOI: https://doi.org/10.1590/S0034-737X2012000300008

Steel, R. G. D., Torrie, J. H., & Dickey, D. A. (1997). *Principles and procedures of statistics: a biometrical approach* (3rd ed.). New York, NY: MacGraw Hill.

Stroup, W.W. (2013) *Generalized linear mixed models: modern concepts, methods and applications.* New York, NY: CRC Press.

Tavares, L. F., Carvalho, A. M. X., & Machado, L. G. (2016). An evaluation of the use of statistical procedures in soil science. *Revista Brasileira de Ciência do Solo, 40*, 1-17. DOI: https://doi.org/10.1590/18069657rbcs20150246

Thorpe, D. P., & Holland, B. (2000). Some multiple comparison procedures for variances from non-normal populations. *Computational Statistics & Data Analysis*, *35*(2), 171-199. DOI: https://doi.org/10.1016/S0167-9473(00)00008-6

Torman, V. B. L., Coster, R., & Riboldi, J. (2012). Normalidade de variáveis: métodos de verificação e comparação de alguns testes não-paramétricos por simulação. *Revista HCPA, 32*(2), 227-234.

Vivaldi, L. J. (1999). *Análise de experimentos com dados repetidos ao longo do tempo ou espaço* (Série Documentos, 8). Planaltina, DF: Embrapa Cerrados.

Volpato, G. L. (2010). *Dicas para redação científica* (3. ed.). São Paulo, SP: Cultura Acadêmica.

Wobbrock, J. O., Findlater, L., Gergle, D., & Higgins, J. J. (2011). The aligned rank transform for nonparametric factorial analyses using only ANOVA procedures. In J. J. Higgins (Ed.), *Proceedings of the ACM Conference on Human Factors in computing systems* (p. 143–146). Vancouver, CA; New York, NY: ACM Press.

Yazici, B., & Yolacan, S. (2007). A comparison of various tests of normality. *Journal of Statistical Computation and Simulation*, *77*(2), 175-183. DOI: https://doi.org/10.1080/10629360600678310

Yossa, R., & Verdegem, M. (2015). Misuse of multiple comparison tests and underuse of contrast procedures in aquaculture publications. *Aquaculture*, *437*, 344-350. DOI: https://doi.org/10.1016/j.aquaculture.2014.12.023

Yoo, W., Mayberry, R., Bae, S., Singh, K., He, Q., & Lillard, J. W. (2014). A study of effects of multicollinearity in the multivariable analysis. *International Journal of Applied Science and Technology*, *4*(5), 9-19.

Zimmermann, F. J. P. (2004). *Estatística aplicada à pesquisa agrícola*. Santo Antônio de Goiás, GO: Embrapa Arroz e Feijão.

Zimmerman, D. W. (2012). A note on consistency of non-parametric rank tests and related rank transformations. *British Journal of Mathematical and Statistical Psychology*, *65*(1), 122-144. DOI: https://doi.org/10.1111/j.2044-8317.2011.02