SOFTWARE RELEASE

SPEED Stat: a free, intuitive, and minimalist spreadsheet program for statistical analyses of experiments

André Mundstock Xavier de Carvalho[1*], Felipe Queiroz Mendes[2], Fabrícia Queiroz Mendes[1] and Laene de Fátima Tavares[1]

Crop Breeding and Applied Biotechnology
20(3): e327420312, 2020
Brazilian Society of Plant Breeding.
Printed in Brazil
http://dx.doi.org/10.1590/1984-70332020v20n3s46

Abstract: *SPEED Stat is a new spreadsheet program for univariate statistical analyses, focused on the dominant profile of agricultural experimentation. The program can perform analysis of variance; tests for normality, homoscedasticity, additivity, outliers; complex contrasts; multiple comparison tests; Scott-Knott's grouping analysis; regression analysis; and others. It has available at speed-statsoftware.wordpress.com.*

Keywords: *Decision support, statistical software, biometrics*

INTRODUCTION

Classical experimentation is the basis for generating new knowledge and developing new technologies and products for agriculture, including plant breeding. Parametric statistical analysis provides the basic support for interpreting and making decisions about experimental data. Unfortunately, the statistical procedures employed to analyze experimental data in the agrarian sciences are frequently misused, as evidenced by Lúcio et al. (2003), Kramer et al. (2016), Tavares et al. (2016), and Possatto Júnior et al. (2019). Such misuse may be associated with a lack of statistical science knowledge; however, low-quality statistical analysis also can be associated with statistical software, which is difficult to use.

For example, Tavares et al. (2016) verified that in the soil sciences, the renowned SAS (Statistical Analysis System, SAS® Institute Inc., Cary, NC, USA) and R (R Project for Statistical Computing) systems are not among the most commonly used software, suggesting that they are not intuitive. In part, the difficulties using a statistical application are linked to the vast number of procedures available in such software (lack of conciseness and specific focus, both typical of non-minimalist design). This massive number of procedures hinders access to the classic procedures.

In this sense, some software has sought to transform the most powerful applications, such as R, into simpler and more accessible tools for each area of knowledge (Bhering 2017, Matias et al. 2018). As a tool, statistical applications are expected to be simple, practical, and user-friendly, and a spreadsheet program combines all these features. Spreadsheet programs allow professionals not connected to advanced programming or software development to easily create and edit algorithms and source codes to solve specific problems related

***Corresponding author:**
E-mail: andre.carvalho@ufv.br
ORCID: 0000-0002-2806-6058

Received: 25 May 2020
Accepted: 29 July 2020
Published: 21 August 2020

[1] Universidade Federal de Viçosa, Instituto de Ciências Agrárias, Campus Rio Paranaíba, 38.810-000, Rio Paranaíba, MG, Brazil
[2] Universidade Estadual de Campinas, Faculdade de Engenharia Química, 13.083-852, Campinas, SP, Brazil

to their field of study (Frownfelter-Lohrke 2017, Weber 2018).

Concise and minimalist software emphasizes the criterion of minimum actions (minimization of the number of actions and commands required to perform a task) to the detriment of the multiuser and multifunctionality criteria (Malan and Bredemeyer 2002). In other words, a minimalist project first delimits the profile of a specific audience and seeks to avoid providing procedures and options irrelevant or rarely useful to this group. With this, minimalist software loses scope to gain in agility, simplicity, and intuitiveness.

This paper aims to present a new spreadsheet program for univariate parametric statistical analyses using a minimalist concept focused on the dominant profile of agricultural experimentation. It describes the basic structure of the application developed, the programming language used, the main algorithms created, and the statistical bases of the procedures. The new program (called SPEED Stat 2.4) is available at *speedstatsoftware.wordpress.com*.

DESCRIPTION

SPEED Stat is an acronym for the expression "spreadsheet program for experimental and descriptive statistics" (in Portuguese). The first step in developing *SPEED Stat* was to clearly define the target audience experimental profile. This profile was utilized to define a list of useful analytical procedures based on the literature, especially the data obtained by Tavares et al. (2016).

The list of procedures showed that, despite the need for a minimalist design, the software could not be overly limited. Excel was used as the development environment; the procedures that were programmed are listed in Table 1. All these

Table 1. Statistical analysis procedures available in SPEED Stat 2.4

Statistical analysis procedures	Details and comments
i. Assumptions for ANOVA	
Normality test	Jarque-Bera test. Residuals are previously corrected by the block effect in RBD or by plots.
Equality of variance test	Hartley, Bartlett, Levene and Levene(Med). Levene(Med) when r > 3 in CRD and with removal of structural zeros.
Non-additivity test	Tukey test for non-additivity. Only for RBD.
Scan by transformations	Algorithm to find the type that satisfies both assumptions.
ii. Analysis of Variance	
ANOVA for CRD	For balanced and unbalanced data, in single schema or in double or triple factorials. With or without up to six additional treatments to the factorial.
ANOVA for RBD	With Yates estimate for missing data when in RBD. Correction of SS for treatments when unbalanced.
- Split-plot ANOVA	With estimated values when unbalanced (limited).
- Split-block ANOVA	With estimated values when unbalanced (limited).
Analysis for repeated measures	Greenhouse-Geisser correction for DF (most conservative condition, ε=1/(p-1), editable).
Nested ANOVA (mixed model)	Hierarchical model only for two factors (with B random)
Non-parametric ANOVA	Simple rank and block rank (RT-2). Aligned rank (ART) to estimate the interaction in double factorials.
iii. Means tests	
Tukey test	For balanced and unbalanced data.
Student-Newman-Keuls test	For balanced and unbalanced data.
Dunnett test	For balanced and unbalanced data (also for factorial).
Scott-Knott test	For balanced and unbalanced data. Only approximate for unbalanced data.
t test	For balanced and unbalanced data (indicated for orthogonal contrasts and independent samples).
Bonferroni mod. by Conagin test	For balanced and unbalanced data (P(F) for treatments).
Dunn-Sidak test	For balanced and unbalanced data.
iv. Regression and other procedures	
Regression ANOVA	For balanced and unbalanced data. With or without 1 additional treatment. Only for nine select models.
Descriptive statistics	Standard error, standard deviation, margin of error.
Outliers test	ESD test (L+1=10). Residuals are previously corrected by the block effect in RBD. Chauvenet criterion.
Size effect statistics	d-Cohen.
Others	complex contrasts, graphs, tables, information from the regression models, verification of correlated errors between successive treatments and other analyzes (since it allows to insert external values of experimental error and DF of error)

procedures are available for both balanced and unbalanced data for experiments with uni-, bi-, or tri-factorial structures (with or without additional treatments) for randomized block (RBD) or completely randomized designs (CRD), and for simple experimental schemes (considering fixed models), split-plots, and split-blocks.

The software consists of a spreadsheet program and its sheets, which use various mathematical, statistical, and logical Excel functions. Each statistical procedure was planned and created in modules, which were later connected in a single file with sheets organized by topic. Macros and VBA (Visual Basic for Applications) functions were avoided due to problems with compatibility between different Excel versions. After linking the developed algorithms and elaborating the final interface, *SPEED Stat* underwent many tests to verify the consistency of the generated results, using solved exercises available in the literature and the authors' collection (data not shown).

The algorithms for the Bartlett, Levene, Hartley, and Jarque-Bera tests were developed as described in Jarque and Bera (1987), Nunes (1998), and Montgomery (2019). The Levene (Med) test programmed into *SPEED Stat 2.4* is a well-known adaptation proposed by Brown and Forsythe (1974) and consists of the use of deviations in relation to medians and not to means. However, the test is performed after removing structural zeros when they exist (Hines and O'Hara Hines 2000). The Jarque-Bera test is processed in *SPEED Stat 2.4* by considering the pure deviations from the means of each treatment and considering them as subpopulations of a total population. The additivity test programmed into *SPEED Stat* is the classic F test for non-additivity proposed by Tukey.

Algorithms for ANOVA, Tukey, Dunnett, and SNK tests were developed as described in Pimentel-Gomes (2009) and Montgomery (2019). For unbalanced data in randomized blocks, the Yates procedure (Pimentel-Gomes, 2009) was used to estimate up to eight empty cells (missing values) by developing recursive functions with three iterations. When the condition of the additivity of the assumed model is violated, absurd estimates generated by the Yates method are corrected by a maximum or minimum value considering maximum amplitude according to the critical values of the generalized ESD test for outliers.

The algorithm for the generalized ESD test was programmed according to Rosner (1983). Tests for complex contrasts (t, Dunn-Sidak, and Bonferroni modified by Conagin) were developed as described in Conagin (2001). The Bonferroni modified test uses a Bayesian approach that adjusts the critical value based on the size of the F statistic. The algorithms for the Scott-Knott test for balanced data are exact. However, for unbalanced data, an approximation is performed that consists of replacing the number of repetitions by an average value between the median and mean of the number of repetitions of the compared set. Thus, the Scott-Knott test in *SPEED Stat 2.4* is not accurate for unbalanced data.

SPEED Stat was programmed to automatically test three basic assumptions for analysis of variance: normality of residuals, homogeneity between treatment variances, and model additivity. This is an innovative strategy to inhibit the frequent violation of parametric analyses assumptions, which has caused recurring problems of incorrect conclusions in scientific papers (Lucena 2013). When one basic assumption is not satisfied, the software verifies if the violation occurs due to a single critical point and also suggests a transformation of the data based on the previous verification of 20 transformation options. Among the options are logarithmic, square root, cubic root, sine arc of square root, rank transformation, block rank transformation (RT-2) (Conover 2012), adapted Johnson Sb transformation, and 13 lambda (λ) options for the Box-Cox transformation. By suggesting one or more types of transformation, the application reports an index (from 0 to 10) of the degree of satisfaction of the perfect parametric conditions based on the three assumptions tested.

Before programming the algorithms for the regression analysis, we conducted a study to choose which mathematical models best combine the attributes of simplicity (a smaller number of dependent parameters), popularity (some frequency of use in scientific work in the agricultural sciences), and theoretical adequacy for the phenomena under study (data not shown). Nine models were selected based on this study (Table 2). The fit quality of the models is compared using the mean square of the models in the regression's ANOVA (F test) and the adjusted coefficient of determination. The level of data deviation in the model is evaluated by the lack of fit in the regression's ANOVA (Piepho and Edmondson 2018).

The algorithms for the regression analyses for these models were developed using the least squares and Gauss-Newton methods according to Lai et al. (2017). The least squares method aims to minimize the sum of the squares of the differences between the estimated value and observed data. In other words, considering any statistical model in

Table 2. Regression models available in SPEED Stat 2.4 and methods for estimating model parameters

Name	Model	Method
Linear	$y = y0 + ax$	Least squares
Quadratic	$y = y0 + ax + bx^2$	Least squares
Root	$y = y0 + ax + bx^{0.5}$	Least squares
Simple logarithmic	$y = y0 + a\ln(x)$	Least squares
Linear logarithmic	$y = y0 + ax + b\ln(x)$	Least squares
Simple exponential	$y = y0 + ae^{bx}$	Gauss-Newton
Mitscherlich*	$y = y0 + a(1 - e^{-bx})$	Gauss-Newton
Michaelis-Menten	$y = y0 + \frac{ax}{b+x}$	Gauss-Newton
Logistic	$y = y0 + \frac{a}{1+\left(\frac{x}{c}\right)^b}$	Gauss-Newton

The constant "e" corresponds to Euler's irrational number. "y0" corresponds to independent parameter. *Model adapted and equivalent to the model originally proposed by Mitscherlich discussed by Ware et al. (1982). This model is a reparametrization of the simple exponential model.

which *y* is the dependent variable, *xj* the independent variables, *βj* the coefficients, and *ε* the residual, the least squares method attempts to minimize the residuals according to Equation 1, where n is the total number of observations:

$$\min\left(\Sigma_{i=1}^{n}\varepsilon_i^2\right) = \min(\Sigma_{i=1}^{n}(y_i - \beta_0 - \Sigma_{j=1}^{k}\beta_j x_{ij})^2) \quad (1)$$

From mathematical deduction, Equation 2 is obtained:

$$\hat{\beta} = (X^T X)^{-1} X^T Y \quad (2)$$

$$\text{In which } X^T X = \begin{bmatrix} \sum_{i=1}^{n} 1 & \sum_{i=1}^{n} x_{i1} & \cdots & \sum_{i=1}^{n} x_{ik} \\ \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i1}^2 & \cdots & \sum_{i=1}^{n} x_{i1}\, x_{ik} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{n} x_{ik} & \sum_{i=1}^{n} x_{i1}\, x_{ik} & \cdots & \sum_{i=1}^{n} x_{ik}^2 \end{bmatrix}, \text{ and } X^T Y = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{i1} y_i \\ \vdots \\ \sum_{i=1}^{n} x_{ik} y_i \end{bmatrix}.$$

The Gauss-Newton numerical method, on the other hand, requires an initial estimate for the parameters $\hat{\beta}^{(o)}$ to calculate the next iterations. Equation 3 is used to calculate $\hat{\beta}^{(s+1)}$ from $\hat{\beta}^{(s)}$.

$$\hat{\beta}^{(s+1)} = \hat{\beta}^{(s)} - (J_\varepsilon^T J_\varepsilon)^{-1} J_\varepsilon^T \hat{\varepsilon}(\beta^{(s)}) \quad (3)$$

$$\text{In which } J_\varepsilon = \begin{bmatrix} \frac{\partial \varepsilon_1}{\partial \beta_0} & \frac{\partial \varepsilon_1}{\partial \beta_1} & \cdots & \frac{\partial \varepsilon_1}{\partial \beta_k} \\ \frac{\partial \varepsilon_2}{\partial \beta_0} & \frac{\partial \varepsilon_2}{\partial \beta_1} & \cdots & \frac{\partial \varepsilon_2}{\partial \beta_k} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \varepsilon_n}{\partial \beta_0} & \frac{\partial \varepsilon_n}{\partial \beta_1} & \cdots & \frac{\partial \varepsilon_n}{\partial k} \end{bmatrix}, \text{ and } \hat{\varepsilon}\left(\beta^{(s)}\right) = \begin{bmatrix} y_1 - \beta_0^{(s)} - \sum_{j=1}^{k} \beta_j^{(s)} x_{1j} \\ y_2 - \beta_0^{(s)} - \sum_{j=1}^{k} \beta_j^{(s)} x_{2j} \\ \vdots \\ y_n - \beta_0^{(s)} - \sum_{j=1}^{k} \beta_j^{(s)} x_{nj} \end{bmatrix}.$$

Then, for each model, an algorithm was created with seven iterations and more than one initial guess of the parameters to analyze any type of data. Finally, *SPEED Stat* classifies up to four statistically suitable models for the data provided, generating graphs and calculating other useful information such as root, maximum, minimum, and inflection points.

FEATURES

The software developed has shown stability and analytical capacity that is adequate for what was proposed. However, as minimalist design software, it has limited scope. The analysis capacity of *SPEED Stat 2.4* is limited to up to 40 treatments

(uni-factorial), 86 treatments (bi-factorials with up to 10x8 + 6 treatments) or 486 treatments (tri-factorials with up to 10x8x6 + 6 treatments). The maximum number of replicates supported per treatment is only eight, which, according to Tavares et al. (2016), is sufficient for the vast majority of experiments in soil science. Although a little restricted, this capacity is sufficient for most of the tests of value for cultivation and use, such as those necessary for registering cultivars.

The application structure consists of 18 sheets, 15 for programming algorithms, and three for the user interface, totaling about 80 Mb. The calculation sheets have been hidden, but they are open source and can be modified by experienced users. Although the application developed is stable, it has a relatively slow response time (approximately five seconds) compared to other spreadsheet programs. The file loading time is about one minute (approximate performance considering 8 GB RAM and 1.6 GHz (Intel i5-6cores processor)). Nevertheless, considering multiple analyses are performed simultaneously, and the results output is fully customizable, the work and time necessary for the user to run routine procedures is reduced compared to most other software.

The three user interfaces are "About," "Input," and "Output." "About" presents the credits and describes the basic steps for using the program; "Input" allows the user to identify the structure of the data and select the desired analysis options; and "Output" provides access to the results generated where the user then formats them to save or print. These interfaces are available in Portuguese, English, and Spanish. Due to the minimalist design and calculation procedures that reduce application performance, user interfaces have been simplified to the maximum, removing images, macros, and animation.

In addition to the simple and timesaving operation and analytical results formatted in Excel achieved by adopting a minimalist design in a spreadsheet program, *SPEED Stat* stands out in four other innovative aspects. First, it automatically checks for the three basic parametric assumptions; second, it automatically scans the dataset for outliers; third, it indicates, when necessary, which type of transformation is most appropriate for the data considering the simultaneous satisfaction of the three assumptions; and fourth, it performs regression analyses of factorial experiments (with or without additional treatments) for selected linear and intrinsically nonlinear models. All this is achieved more quickly and simply in comparison to other applications.

CONCLUSION

The spreadsheet program developed presents itself as a simple, timesaving, and intuitive alternative for researchers in the academic and business sectors who work in agronomic experimentation and plant breeding. The application for univariate parametric statistical analyses was validated by comparing the results generated with several solved exercises available in the literature. *SPEED Stat 2.4* is a free/open source spreadsheet program and available for download at *speedstatsoftware.wordpress.com.*

REFERENCES

Bhering LL (2017) Rbio: A tool for biometric and statistical analysis using the R platform. **Crop Breeding and Applied Biotechnology 17**: 187-190.

Brown MB and Forsythe AB (1974) Robust tests for the equality of variances. **Journal of the American Statistical Association 69:** 364-367.

Conagin A (2001) Tables for the calculation of the probability to be used in the modified bonferroni's test. **Brazilian Journal of Agriculture 76:** 71-83.

Conover WJ (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:** 432-438.

Frownfelter-Lohrke C (2017) Teaching good Excel design and skills: A three spreadsheet assignment project. **Journal of Accounting Education 39**: 68-83.

Hines WGS and O'Hara Hines RJ (2000) Increased power with modified forms of the Levene (Med) test for heterogeneity of variance. **Biometrics 56:** 451-454.

Jarque CM and Bera AK (1987) A test for normality of observations and regression residuals. **International Statistical Review 55:** 163-172.

Kramer MH, Paparozzi ET and Stroup WW (2016) Statistics in a Horticultural Journal: problems and solutions. **Journal of the American Society for Horticultural Science 26:** 558-564.

Lai WH, Kek SL and Tay KG (2017) Solving nonlinear least squares problem using Gauss-Newton method. **International Journal of Innovative Science, Engineering & Technology 4**: 258-262.

Lucena C, Lopez JM, Pulgar R, Abalos C and Valderrama MJ (2013) Potential errors and misuse of statistics in studies on leakage in endodontics. **International Endodontic Journal 46:** 323-331.

Lúcio AD, Lopes SJ, Storck L, Carpes RH, Lieberknecht D and Nicola MC (2003) Características experimentais das publicações da Ciência Rural de 1971 a 2000. **Ciência Rural 33:** 161-164.

Malan R and Bredemeyer D (2002) Less is more with minimalist architecture. **IT Professional 4:** 47-48.

Matias FI, Granato I and Fritsche-Neto R (2018) Be-Breeder: an R/Shiny application for phenotypic data analyses in plant breeding. **Crop Breeding and Applied Biotechnology 18:** 241-243.

Montgomery DC (2019) **Design and analysis of experiments**. Wiley, Danvers, 688p.

Nunes RP (1998) **Métodos para a pesquisa agronômica**. UFC, Fortaleza, 564p.

Piepho HP and Edmondson RN (2018) A tutorial on the statistical analysis of factorial experiments with qualitative and quantitative treatment factor levels. **Journal of Agronomy and Crop Science 204**: 429-455.

Pimentel-Gomes F (2009) **Curso de estatística experimental**. FEALQ, Piracicaba, 451p.

Possatto Júnior O, Bertagna FAB, Peterlini E, Baleroni AG, Rossi RM and Zeni Neto H (2019) Survey of statistical methods applied in articles published in Acta Scientiarum. Agronomy from 1998 to 2016. **Acta Scientiarum. Agronomy 41:** e42641.

Rosner B (1983) Percentage points for a generalized ESD many-outlier procedure. **Technometrics 25:** 165-172.

Tavares LF, Carvalho AMX and Machado LG (2016) An evaluation of the use of statistical procedures in soil science. **Revista Brasileira de Ciência do Solo 40:** e0150246.

Weber EV (2018) **Spreadsheet fundamentals**. Kendall Hunt Publishing, Iowa, 277p.