REVIEW

SISVAR: A COMPUTER STATISTICAL ANALYSIS SYSTEM

Sisvar: um sistema computacional de análise estatística

Daniel Furtado Ferreira[1]

ABSTRACT

Sisvar is a statistical analysis system, first released in 1996 although its development began in 1994. The first version was done in the programming language Pascal and compiled with Borland Turbo Pascal 3. Sisvar was developed to achieve some specific goals. The first objective was to obtain software that could be used directly on the statistical experimental course of the Department of Exact Science at the Federal University of Lavras. The second objective was to initiate the development of a genuinely Brazilian free software program that met the demands and peculiarities of research conducted in the country. The third goal was to present statistical analysis software for the Brazilian scientific community that would allow research results to be analyzed efficiently and reliably. All of the initial goals were achieved. Sisvar gained acceptance by the scientific community because it provides reliable, accurate, precise, simple and robust results, and allows users a greater degree of interactivity.

Index terms: Multiple comparisons, analysis of variance, regression, hypothesis tests.

RESUMO

O Sisvar é um sistema de análise estatística que foi lançado em 1996, embora o seu desenvolvimento tenha sido iniciado em 1994. A primeira versão foi desenvolvida em linguagem de programação Pascal e compilada com o Borland Turbo Pascal 3. O Sisvar foi desenvolvido em virtude de algumas razões específicas. O primeiro objetivo foi o de obter um *software* que pudesse ser usado diretamente no curso de estatística experimental do Departamento de Ciências Exatas da Universidade Federal de Lavras. O segundo objetivo foi o de iniciar o desenvolvimento de um software genuinamente brasileiro, gratuito que atendesse às demandas e peculiaridades das pesquisas realizadas no país. O terceiro objetivo foi o de apresentar um *software* de análise estatística para a comunidade científica brasileira que permitisse que os resultados da pesquisa pudessem ser analisados de forma eficiente e confiável. Todos os objetivos iniciais foram atingidos. O motivo da aceitação Sisvar pela comunidade científica é decorrente do fato de que ele é capaz de permitir uma maior interatividade com o usuário e produzir análises confiáveis, pelo fato de elas serem exatas, precisas, simples e robustas.

Termos para indexação: Comparações múltiplas, análises de variância, regressão, testes de hipóteses.

(Received in september 15, 2011 and approved in november 30, 2011)

INTRODUCTION

Statistics are admittedly an essential part of the scientific method and it is (in general) not possible that any article would be accepted by the scientific community without its results have been confirmed by statistical analysis. New statistical methods are proposed every year, and researchers from different fields can use them to develop their scientific works, testing their hypotheses and solving problems of economic importance and technological innovation. But to use these new methods, researchers need new computer programs or routines to be developed and incorporated into the existing computer *programs* for *statistical* analysis. However, many analyses are made using traditional methods. The analysis of variance (anova) of linear models resulting from data observed on experiments is the most common of them. Models with simple, hierarchical and *cross* classified effects are very common. In this context, computer *programs* for *statistical* analysis which slice the interaction or hierarchical effects into their component parts and apply multiple comparison procedures and adjust regression models are required.

Statistical analysis systems such as SAS and R, when dealing with the problem of slicing the crossover or hierarchical effect of a linear model do not properly apply multiple comparisons for the means or F tests for the parameters of fitted regression models. Multiple comparison procedures and tests for the adjusted regression models should be redone manually. Many errors are unconsciously published annually, since many researchers are unaware of the limitations related to those programs.

The statistical analysis computer program Sisvar was developed to circumvent these limitations and is currently used as a tool for teaching and research at the

[1]Universidade Federal de Lavras/UFLA – Departamento de Ciências Exatas/DEX – Lavras – MG – Brasil – danielff@dex.ufla.br

Federal University of Lavras (UFLA) and many other educational and research institutions (public and private) throughout the country. Moreover, Sisvar is free software (the same as R). Researchers can use it for free by downloading and installing directly from the address: www.dex.ufla.br/~danielff.

The aim of this paper is to present the computer statistical program Sisvar, illustrating its advantages, limitations and analysis capabilities. In addition, a secondary objective is to introduce the new Sisvar version in Java that is under development. The new features under implementation will be emphasized.

THE SISVAR SYSTEM

The Sisvar was first released in 1996, although its development was initiated in 1994. The first version was developed in the programming language Pascal and compiled with Borland Turbo Pascal 3.0 for the generation of an executable program. Initially, the database was implemented using unique features of the Pascal programming language. As the computer mouse was rarely used at that time, the navigation system across the data records was done via shortcut keys, also with implementations using only features of Pascal. The report used a similar text editor to Notepad in Windows. In this version, only the first module of an analysis of variance was developed. This is where the program's name came from, in that "Sis" refers to the abbreviation of system in Portuguese and "var", the abbreviation of variance. Thus the Sisvar system consists of an analysis of variance statistical program. It was possible to apply multiple comparison procedures and regression analyses for quantitative effects in the same way Sisvar do in the current version. Moreover, interaction effects could be sliced into their component parts and the same tests and analyses as applied to the main dataset could be re-applied in at this level.

Sisvar was developed to achieve specific goals. The first objective was to obtain software that could be used directly on the statistical experimental course of the Department of Exact Sciences at UFLA. It was expected that there would be an improvement of *learning* and teaching processes. This was achieved. The second objective was to initiate a development of genuinely Brazilian free software that met the demands and peculiarities of research conducted in the country. There was a fear that non-free programs were becoming increasingly inaccessible to researchers, teachers and students from Brazil. There were some Brazilian programs for statistical analysis at the time such as Sanest, SOC and Saeg. The limitations inherent in these programs were the reasons why the Sisvar project was initiated, intending to overcome them. The third goal was to introduce statistical analysis software for the Brazilian scientific community that would allow the research results to be analyzed efficiently and reliably. Because Sisvar was being developed by only one researcher, it took approximately two and a half years to release the first version.

Later, with the advent of Windows and with the release of the Borland Delphi 1.0, the first version of Sisvar for Windows was launched. In this version and in its updates several analysis modules were incorporated and the bugs were corrected as they arose. Sisvar's popularity grew more and more, with the program becoming increasingly well-known by the Brazilian scientific community and also by the Portuguese scientific community. This wide acceptance of Sisvar is due to many factors. The main thing is its great interactivity with users. Analyses are performed in Sisvar with mouse clicks only, in an almost self-explanatory sequence until results are obtained. The reliability, accuracy and robustness of the software can be highlighted in that it rarely freezes or displays error messages and by the richness and detail of results. In the first version of Sisvar, the inspiration of the construction of the interfaces was my master's supervisor, who at the time had many difficulties with the use of computers. Thus, it was thought that the interface should be accessible to users less familiar with computers to aid them in performing their statistical analysis. In fact this aspect was very successfully implemented in Sisvar. Many details and deficiencies still exist and should be improved further.

During the development of new versions for Windows, statistical analysis modules were implemented such as: descriptive statistics, hypothesis testing for various parameters, interval estimation for several population parameters, normality tests, kernel density estimation, simple and multiple linear regression adjustments, methods of model selection as stepwise, backward and forward and multiple comparisons using bootstrap. The current Sisvar release is 5.3. With these new implementations, basic statistics courses can now use Sisvar in classroom practice, which could already be done on experimental statistics courses.

Some features of Sisvar's analytical functions should be mentioned so that the user knows exactly what the advantages and disadvantages of this program are. As mentioned previously, the main analysis module is the analysis of variance (anova) of experimental linear models. Although very powerful, this option can only be used for

balanced data in Sisvar. The exception is the case of one way anova design with only one factor. Sisvar cannot perform an analysis if there are empty cells in the database. When there is missing data, the user should simply omit the line corresponding to the missing portion during the process of file creation.

Sisvar accepts a maximum of 200 variables, columns in the data. The linear regression and models selection modules can be used with a maximum of 100 regression variables. Finally, one can observe that Sisvar can only deal with fixed Gauss-Markov linear models. There are no limits on the number of records, except for limitations of space in the user's hard drive.

Among the advantages that Sisvar has over its competitor's statistical analysis systems is the ability to slice the interactions and nested effects amongst fixed factors of linear models. Besides the analysis of variance of this process, there is the possibility to apply multiple comparison procedures and contrast of the sliced means of one factor into settled levels of the other including the Scott-Knott test, absent in most of Sisvar's competitors. For quantitative effects, the sliced crossover or hierarchical effects could be done by regression analysis. Besides Sisvar, no other statistical analysis program is known to apply appropriate tests of hypothesis for the parameter of regression models when it is performed in one quantitative factor to settled levels of the other effects of interaction or hierarchical effects.

The ability to perform the Scott-Knott test is one of the great advantages of using Sisvar, as mentioned earlier. This test has no ambiguous results. In addition it has high power and good control of type I experiment error rates under partially H_0. The R and SAS statistical analysis systems have not implemented such tests in their basic analysis routines. Another major advantage is that Sisvar allows a high level of user interactivity, as pointed out before, providing a statistical analysis environment that is easier to use. Moreover, the robust, easily interpreted and highly accurate results make Sisvar one of the most attractive statistical analysis systems.

THE SISVAR SYSTEM IN JAVA

A new version in Java is under development. This Java version resulted from the need to surpass the limitations of the current release, enhancing the program's interactivity and solving incompatibility issues with the current Sisvar database. The current version of Sisvar uses Paradox and dBase databases, which are considered obsolete. Microsoft, since the release of Windows Vista, failed to support these databases. Numerous cases of such problems were reported. Using Sisvar 5.3 with compatibility module of Windows XP Service Pack 2 solves the problem. Anticipating the possibility of this issue getting worse with new releases of Windows, the idea of implementing Sisvar in a cross-platform language with wide support for the database came up. Hence, the first Sisvar Java version which will be release 6.0 is under development. The database db4o (which is implemented in Java) was chosen, because it occupies little space on the hard drive, is efficient and does not require installation by the user. The database module of the new Sisvar is ready and is much friendlier than the 5.3 version file manager. CSV files can be used for passing data from different statistical analysis systems, database or spreadsheet into Sisvar. This represents a big improvement and makes the program more attractive and offers more interactivity with other files system or databases available nowadays. Many other improvements have been implemented.

The great advantage of having a Java version is to make Sisvar multiplatform, i.e., the software will run on Windows, Solaris, Mac OS X and Unix/Linux and others that run Java Virtual Machine. The current version is implemented in Pascal/Delphi and can run specifically in Windows environment. Still, with installation through Wine, the user can run Sisvar on Linux platforms. The first priority is to maintain and expand the interactive features of the existing version. The Java Programming Language is very powerful and allows the creation of computer systems capable of performing different tasks feasible to a computer. Java uses the powerful tools of object-oriented programming, OOP. It was developed by Sun Microsystems and is currently one of the languages most popular for software development (DEITEL; DEITEL, 2007). The reasons for such popularity are the portability and free compilers. Portability is the ability of a developed program to run on several platforms. The Java compilers and other tools for developing computer programs are freely available on the Oracle website. Another advantage is the ability to develop systems for the Internet and technologies such as mobile devices, as for instance mobile phones.

During the implementation of the Java version of Sisvar, the multiplatform release, several specific problems that need solutions or generalizations related to increased efficiency and improvement of the quality of the existing techniques are addressed. Among them, the multivariate statistical methods will have a special attention. The new release will focus on multivariate methods (JOHNSON; WICHERN, 1998; FERREIRA, 2008) involving the Behrens-Fisher problem, normality multivariate tests, and computational methods for calculating non central

distributions. Moreover, several classical univariate (FERREIRA, 2009) and multivariate (FERREIRA, 2008) statistical analysis techniques that are not available in the current release are being implemented in the new version of Sisvar. Many methods have been proposed during this process that are partially or fully subjects of dissertations and PhD theses of the experimental statistics program of the Department of Exact Sciences at UFLA.

The major barrier preventing the use of computational intensive methods in agriculture and plant breeding is the lack of user-friendly software associated with the small diffusion of methods of this nature among researchers from agricultural sciences or other areas. Because of these difficulties and the need that the major computer-intensive statistical methods are available to scientists from various fields of knowledge, the idea to include of such innovation in Sisvar was now real. Constant updates are still held in the Pascal/Delphi releases of Sisvar and they are available for download at www.dex.ufla.br/~danielff.

Sisvar in Java will be available online for download at the same address above. The greatest innovation will be making available the source code, beside the executable program. It will be necessary to create a developers core. All interested persons may be able to contribute to improve and develop new routines. This core, the Sisvar core team, will be responsible to approve the new routines and for releasing new versions. This is the first Brazilian seed for developing of a statistical analysis system totally free, in a similar manner to that performed by the R program (R DEVELOPMENT CORE TEAM 2011).

CONCLUSION

Sisvar's acceptance by the scientific community is due to the fact that this statistical analysis system produces reliable analyses, in that they are accurate, precise, simple and robust and allow greater interactivity for the user. Many tests carried out by Sisvar are difficult to reproduce directly when using other software of statistical analysis, requiring artifice to be obtained. It is expected that the new version will minimized limitations. The main courses of action to achieve these goals might be to solve compatibility issues, to expand the analytical capabilities of the program, to increase the efficiency of their routines, and to enhance the portability of the program to other operating systems besides Windows.

ACKNOWLEDGEMENTS

The author would like to thank CNPq, CAPES and FAPEMIG for the financial support during these 15 years of development of Sisvar.

REFERENCES

DEITEL, H. M.; DEITEL, P. J. **Java:** como programar. Pearson Prentice Hall, São Paulo, 6th edition, 2007. 1110 p.

FERREIRA, D. F. **Estatística multivariada**. Lavras: Editora Ufla, 2008. 662 p.

FERREIRA, D. F. **Estatística básica**. Lavras: Editora Ufla, 2ª ed. ampliada e revisada. 2009. 664 p.

JOHNSON, R. A.; WICHERN, D. W. **Applied multivariate statistical analysis**. Prentice Hall, New Jersey, 4th edition, 1998.

R Development Core Team. R: **A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria, 2011. Available in: <http://www.r-project.org>. Access in: 31 Jan. 2011.