TEACHING WITH S-PLUS IN FIRST YEAR

Ewa M. Sztendur and Neil T. Diamond
Victoria University of Technology
Australia

This paper describes our experience teaching S-Plus to first year computer and mathematical science students. We explain the reason for selecting S-Plus as the statistical package for our students and describe the material prepared to introduce the package and how it was presented. We also outline the problems that students experienced, analyse the reasons for these problems and the ways we are attempting to overcome these.

INTRODUCTION

The ability to select the appropriate statistical techniques and interpret the results from computer packages is an essential part of the statistical learning process. Not only does the use of the computers in statistical education enhance students' understanding of the material presented in lectures but also stimulates interest in the subject. In this paper we describe how we have introduced S-Plus into the first year teaching of statistics for the Bachelor of Computer Science/Bachelor of Computer and Mathematical Sciences degree at Victoria University of Technology.

OUR COURSE

Victoria University of Technology is based in the multicultural western suburbs of Melbourne. Graduates of the 3-year Bachelor of Computer Science and Bachelor of Computer and Mathematical Sciences find employment in computing (including programming, software development, systems design and analysis, applications development, technical support), statistics (data analysis, quality improvement, market research, forecasting, econometrics), operations research, financial modelling and secondary teaching.

The first year of the Bachelor of Computer Science and Bachelor of Computer and Mathematical Sciences is currently common. Students compulsorily study a number of computer subjects, statistics and mathematics. Some of the students undertaking the Bachelor of Computer Science do not choose to do statistics in further years, although many of them do.

The aim of the first semester statistical unit is to get students interested in statistics. We emphasise statistical thinking rather than statistical calculation. Topics covered include: Framework of statistics, basic tools, fitting straight lines, statistical process control, time series models, and experimental design. The text used is Wood et al (2000). The second semester material is rather more traditional than first semester. We start with sample survey design and then discuss probability, random variables and distributions, one and two-sample inference and inference for linear regression.

Second year subjects offered include linear statistical models, statistical data mining and forecasting. Third year subjects include regression, experimental design, time series etc. During third year all students undertake at least one industry-based project as described in Diamond and Hallett (1998). These projects tend to be related to problems encountered in the manufacturing industry, banking or finance, government statutory authorities or services such as hospitals and local councils.

USE OF COMPUTER PACKAGES

There are a number of reasons why the proper use of statistical packages in teaching statistics is important. As outlined by Dallal (1990) "Using a statistical package forces students to come to grips with whether they really understand how to apply the techniques that they have studied". In another article Makuch et al (1990) stress the importance of computer software in industrial statistical practice: "The goal of our proposed curriculum is simple: to enable graduates to take the fullest possible advantage, in solving statistical problems in industry, of the computing environment that they will likely encounter in their careers" and "Statisticians in industry, as well

as in other applications areas, need to respond to a wide variety of problems, most of which require computers in their solution".

Our aim in introducing statistical computer packages in first year is to assist the students to understand and become confident with the theory covered in lectures and to prepare the students so that they can confidently use the package in second and third year subjects as well as the very important industry project.

In previous years the statistical package MINITAB has been used throughout the course and to a lesser extent SPSS. Although MINITAB proved to be helpful and adequate as a statistical package and the students found the program user-friendly there were a number of disadvantages. The major drawback was the cost of the licence to the School and the fact that the students were unable to access the package at a reasonable price. They could only get the package free for a thirty-day trial. Despite these disadvantages we were relatively comfortable in continuing to use MINITAB until a new license offer came from the Suppliers of S-Plus. Some of the features of this offer were so attractive that we seriously examined whether it was feasible to adopt S-Plus as the statistics package to be adopted in our course.

S-plus could have been chosen for a number of reasons, including its extensive features, its high quality graphics, because of its extendibility or because it is object-oriented. However the main reasons for selecting S-Plus was because the license fee of the package was considerably less than that for MINITAB and the fact that the School could provide each student with a copy of the software for home use free of charge.

One of the major considerations was whether we could teach with S-Plus in first year. From our own experience S-Plus was more difficult to use then MINITAB or S-Plus and it appeared that not many Universities had adopted it at the undergraduate, let alone first year level. In addition, our first year students tend to have weaker backgrounds when compared to students at other universities in Melbourne. Even at the University of Melbourne, MINITAB is used in the first two years while S-Plus is only introduced in third year. In his review, Ripley regards "S-Plus the natural choice for graduate courses and the specialist years of undergraduate courses".

Despite these obstacles we decided to try to introduce S-Plus throughout our course, beginning with the second semester of first year in 2001. One of the factors that we took into account was that our students were studying Computer Science and were at the same time learning Java, and so we believed and hoped that the command based version of S-Plus should not be too difficult.

THE LABS

Labs were held every second week and 5 lab sheets were prepared. They covered the following topics: introduction to the menu system, commands, probability distributions, checking for normality and hypothesis tests. An outline of the labs is given in Table 1.

Although our motivation for only having labs every second week was due to resource constraints, as it turned out the frequency seemed to be quite appropriate for introducing the package to students such as ours.

Table 1
Outline of the Material Covered in the S-Plus Labs

Lab 1	At the completion of the lab students could import some sets of data, make histograms, calculate summary statistics, use the commands window to make stem and leaf displays and set up a X-bar and R-chart (covered in first semester).
Lab 2	Students were exposed to the indexing of matrices and vectors and some commands for generating plots.
Lab 3	Students used S-Plus commands to illustrate the Central Limit Theorem. They also used commands for discrete distributions.
Lab 4	Students were introduced to commands for continuous distributions and used normal probability plots to check for normality and equivalent plots to check for exponential and uniform distributions.
Lab 5	Students were introduced to the hypothesis testing and confidence interval facilities in S-Plus. These included one-sample t-test, paired t-test, two-sample t-test, tests for proportions and power and sample size calculations for normal means and binomial proportions

PROBLEMS ENCOUNTERED

In this section some of the problems we experienced will be discussed and analysed. In the next section we will outline our attempts to overcome the problems. Many of the students were having difficulty with statistics. Although some of them had done some probability and statistics at secondary level they found the inference part of the course more difficult. These students tended to slavishly follow the lab sheets without any attempt to understand what was going on and were unable to see a general pattern. In addition since many of our students are from overseas backgrounds they have some difficulty with English. Somewhat surprisingly many of the students also had difficulty using the computer.

S-Plus is a package with many features. There is a great temptation to be too challenging. Although we tried to resist this, in retrospect some of the worksheets were too demanding and were not graded appropriately. Some of our attempts to use discovery learning were only successful for the better students.

S-Plus is a very comprehensive package that can be overwhelming to a new user, almost like giving a learner driver the keys to a jumbo jet. There are some limitations to the help system. For example, many of our students while searching for help on the standard deviation ended up with information on the two-sample t-test. Only when the correct command (stdev) was known could the appropriate help be obtained. Another disconcerting feature for the novice user is that when the result of a command is assigned to an object, no output is produced.

There are generally two ways to do something in S-Plus, either using the commands or the graphical user interface. Although students find it easier to use menus rather than commands we want to emphasise the importance and usefulness of commands for later years. Commands are generally much more powerful and attractive, allowing the user to effectively use the S-plus programming language which includes loops, user-defined functions and input and output facilities. Somewhat paradoxically it is much easier to explain how to use the commands rather than the menus.

WHAT WE ARE DOING TO CHANGE THE SITUATION

Many improvements to the sheets were made over the semester since each lab was given over a period of two weeks. For example, we added boxed explanations for the commands to explain what was going on. In addition we endeavoured to split up compound expressions so that clear but long series of commands were preferred to clever but potentially confusing uses of the language. The subject is currently being offered again over the summer semester. Changes we have made are giving more information on the help system, the introduction of a cheat sheet to assist the students, some reorganisation of the material so that the lectures and the labs are more complementary rather than the lectures being given priority and the labs being just add-ons. In addition, some material is going to be relinquished from the labs, such as indexing which although very important is a bit challenging for first years. The reorganisation means that we can add material on linear regression.

CONCLUSION

Introduction of S-Plus to students in the first year of Bachelor of Computer/Computer and Mathematical Sciences proved to be an interesting experience. We are confident that with the changes we have made and foreshadowed that S-Plus can be very successfully used for our first year students.

REFERENCES

Dallal, G.E. (1990). Statistical computing packages: Dare we abandon their teaching to others? *The American Statistician*, *44*, 265-266.

Diamond, N.T., & Hallet, R.F. (1998). Industry projects in statistics. In L. Perera-Mendoza, L.S. Kea, T.W. Kee, and W.-K. Wong (Eds.), *Proceedings of the 5th International Conference on Teaching Statistics, Volume 3* (pp. 1135-1139).

Ripley, B.D. (1997). *S-Plus for Windows 4.0*. Maths&Stats, November 1997, CTI publications.

Wood, G., Dunning, D.B, Noble, A.D.L., & Wasemi, S.A. (2000). *Statistics - thinking and practice*. N.Z.: Massey University.