Teaching an Introduction to Data Analysis Using Computer Assisted Learning Programs Written in cT

Glenys Bishop - Gold Coast, Australia

1. Background and subject outline

Bond University, Australia's first private university, has developed a course structure based on semesters. A full load in an undergraduate course consists of four subjects per semester and an ordinary degree can be completed in six semesters (except for a Law degree which takes eight semesters).

Every undergraduate student must take four core subjects, regardless of the course being undertaken. These core subjects are Computing and Data Skills, Communications Skills, Management, and Cultural and Ethical Values. In this paper I am going to concentrate on the core subject, Computing and Data Skills, which is taught by the School of Information and Computing Sciences. This subject aims to:

(i) provide an understanding of the uses and impact of computers;

(ii) provide initial training in the basic skills needed to make effective use of a personal computer;

(iii) show how computers can be used for the analysis, storage, and presentation of information for use in planning, prediction, and decision-making.

Students attend one one-hour lecture and two two-hour workshops per week. Since the workshops are held in computer laboratories in which each student has the sole use of a computer, they cannot all attend the same workshop. There are typically between five and ten workshop groups, each taught by a different staff member, usually with the help of one or two student assistants. The lectures cover the underlying concepts and are related to the workshop material. The workshops teach the necessary skills and reinforce the concepts discussed in the lectures.

The subject matter is divided into three modules. The first lasts three weeks and

teaches wordprocessing and document design. The second takes five weeks and teaches data skills and data presentation. The third module, lasting four weeks, splits the students into two streams. Law and Humanities students study research methods and presentation, while Commerce and Information Technology students study information analysis. Each module is assessed by an examination, an assignment, and attendance at workshops.

The information analysis module is the one on which I wish to concentrate for the rest of this paper.

2.   Information analysis

The emphasis in this module is on the analysis of data and on drawing conclusions from the data. The common thread running through the module is based on a government document, which contains a number of assertions requiring investigation. Accompanying this document is a data set consisting of details about 100 randomly selected companies.

Whenever a concept of data analysis is taught, the students then apply it to the data set as part of the investigation.

Concepts are taught intuitively and by example. No mathematical basis is given for any of the techniques used. No mention is made of t-tests. Hypotheses are tested by finding the statistic plus and minus twice the standard error.

The first workshop concentrates on histograms and measures of location and dispersion. The second is about probability; the third relates to the binomial distribution, while the fourth is about the normal distribution and taking samples. The fifth workshop is about framing hypotheses and modelling a process. The last two workshops concentrate on relationships between two variables and cover simple linear regressions and simple time series.

By the time the students study this module they have already learnt how to use the spreadsheet, Excel, and so the practical exercises involving calculation and charts of the data are performed using Excel.

Many statistical concepts can be taught intuitively by letting the student experiment. It has been our experience that when students are encouraged to experiment using spreadsheets and statistical packages, they often become bogged down in the details of how to get the package to work and lose sight of the concepts. Well-written computer assisted learning programs, on the other hand, enable students to focus on the concepts.

3.   Computer assisted learning

In the School of Information and Computing Sciences at Bond University, we write all of our computer assisted learning programs in the language cT, developed at Carnegie-Mellon University by Sherwood and Sherwood.

The goal of cT is "to facilitate writing modern, graphics-oriented computer programs" (Sherwood and Sherwood, 1988). Special features of the language include tools for the display of text and graphics, menu handling, mouse-inputs, tools for

analysing text and mathematical inputs, data management, loop and if structures, and the ability to perform mathematical calculations. Its star attraction is that one does not have to be a professional programmer to write cT programs, utilising all of the above features. For instance, a menu of four items can be created as follows:

    menu   Example;Introduction: regression_example
    menu   Example;Hypothesis: hypothesis_example
    menu   Example;Residuals: residuals_example
    menu   Example;Outliers: outliers_example

This set of commands will create a pull-down menu called Example. If the student clicks on Introduction, program control will pass to the procedure, regression_example. If the student clicks on Residuals, control will pass to the procedure, residuals_example, and so on.

Mathematical input can be accepted in several different ways. For instance, in the next example, a question requiring a numeric input is asked. Then an answering construct is set up, i.e. an arrow is drawn at specified screen coordinates. The correct answer has the value, lbound, but an estimate within 10 percent of this would be acceptable.

    write   What is the lower bound for the slope?
    arrow   310,80
    ansv   lbound,10%
    endarrow

The answer would be judged correct by cT if a value within 10 percent of lbound were input or if a formula were typed which, when calculated by cT, had a value within 10 percent of lbound.

Another of the features of cT is that a program written on a Macintosh will also run on an IBM PC.

Computer assisted learning programs have gradually been introduced to the information analysis module and there are now six such programs which allow the student to experiment with different concepts. These programs deal with probability, the binomial distribution, the normal distribution, taking samples from populations, modelling supermarket queues, and simple linear regression.

By using cT programs, students can watch dynamic examples and interact with them by inputting parameters. So, for instance, in the binomial distribution program, the student can watch a ball drop down through several levels of branching. The student provides the number of levels, n, the probability of a left branch at each level, p, and the number of trials. As the ball drops a binomial histogram is built up.

In the regression program, the student uses the mouse to draw a line by eye through some data points and then compares this with the line fitted by least squares. The student can remove an outlier by clicking on it and the regression line will be recalculated without this point.

The normal distribution program allows the student to provide values of μ and σ to see how they influence the shape of the normal distribution. Experiments with a skewness parameter can also be used to allow deviations from the normal. After

students have run this program they run another program, called sampler, which allows them to take samples of different sizes from normal and skew distributions and compare the frequency distributions of the samples with the population distribution curve.

The supermarket queue model program is interactive in the same way and also dynamic; but in addition it generates values which are saved for later analysis in an Excel spreadsheet. Like the binomial distribution program it uses animations to illustrate what is happening.

The way in which the programs are incorporated into the workshops is best shown using the simple linear regression program. The examples of code in the previous section are from this program.

The program, regress, has four data sets, which illustrate straightforward simple linear regressions with positive and negative slopes, outliers, and non-random residuals. The first data set is used as an example to take the student through the steps of fitting a line, framing hypotheses, testing the null hypothesis, and inspecting the residuals. They carry out each of these steps by choosing the relevant item from the Example menu. This means that any item can be repeated if required.

The workshop notes instruct the student to work through the example and then read a value from the fitted line. After that, the students work with the other data sets on which they can perform a variety of operations. They can fit a line by eye using the mouse, obtain the least-squares regression line, and compare the two. They can also remove outliers by clicking on them whereupon the regression line is recalculated. They can test the hypothesis of a zero slope without having to calculate a t value. The whole process is interactive and immediate without the student having to worry about programming calculations or performing t-tests. To ensure that students actually work through the data sets, they must record in their workshop notes the results for each set.

The concepts are reinforced by getting the student to perform regression calculations on the company data set using Excel.

4.   Summary

In summary, first semester students at Bond University are being taught the fundamentals of data analysis by a mixture of computer assisted learning programs and practical reinforcement using a spreadsheet.

The student interacts with cT programs by pointing and clicking at the screen, by typing text, values and formulae, and by choosing items from a menu. In response to this input, icons move around the screen, graphs are plotted or changed, histograms are constructed, calculations are performed, messages are printed, and results are saved.

The cT language is reasonably easy to learn, inexpensive to buy, and the manuals are clearly written. Programs developed on a Macintosh will run without change on an IBM personal computer.

Reference

Sherwood, B A and Sherwood, J N (1988) *The cT Language.* Stipes Publishing Company, Champaaign, Illinois 61820.