MATHEMATICAL MODELING USING STATISTICAL TOOLS: THE EXPLORATION OF LIFE-RELATED DATA USING COMPUTER AND GRAPHING CALCULATOR TECHNOLOGY

Vince Geiger, Hillbrook Anglican School (Brisbane), Australia
Rex Boggs, Glenmore High School (Rockhampton), Australia

The accessibility of datasets from the Internet coupled with mathematically enabled software and technologies such as graphing calculators which include data analysis capabilities, place in the hands of students and teachers tools that have the potential to create a rich learning environment for the study of statistics. Further, the advent of graphing calculator peripherals such as Texas Instruments' Calculator Bases Laboratory (CBL) and Calculator Based Ranger (CBR), or Casio's Data Logger provide teachers and students with the opportunity to investigate data they have gathered themselves.

INTRODUCTION

The following represents a familiar style of problem used in traditional introductory statistics courses in senior secondary schools in Queensland.

> *Over the last year, Jane spent these monthly amounts on CDs - $23, $48, $32, $85, $46, $35, $51, $39, $42, $45, $29, $52. Calculate the mean of this dataset. Is the value $85 an outlier?*

This style of problem attempts to recognise the importance of making mathematics relevant to the audience by setting a mathematical situation in a life-related context. Unfortunately the question of relevance is a highly personal issue. Students may be very interested in the price of *their* favourite CD in *their* local music store, but are less likely to be interested in what someone else paid for their CDs over the last year. Further, the problem is really a masked attempt at providing students practice within a limited range of mathematical techniques. By using such problems, the not-so-subliminal message is that statistics deals with the trivial, the irrelevant and the uninteresting.

Syllabus initiatives in Queensland (Board of Senior Secondary School Studies, 1992), the availability of genuine datasets via the World Wide Web, and affordable technology that permits the collection of data in real time, provide the opportunity to offer learning experiences that are at last genuinely relevant, permit teachers and students to focus on mathematical process, and are problem driven.

A STATISTICS COURSE FOR SECONDARY SCHOOLS

A secondary school statistics program should sensibly pay attention to the following features:

- the course is about data - in particular what the data tell us about a population;
- the problems are real, and have inherent interest;
- the focus of the learning is on concepts, and not techniques and calculations.

The design of a course consistent with these aims faces the following challenges:

- datasets are often too large to be tackled using traditional paper and pencil techniques;
- data are often ‘messy’, so there is not always a nice, neat solution to the problem;
- the time required for our students to master necessary techniques leaves little time to teach concepts;
- the lack of easy availability and accessibility of datasets suitable for study in school mathematics courses;
- the lack of availability, accessibility and affordability of tools powerful enough to deal with non-trivial problems.

Information technology, especially that of graphing calculator technology, offers the opportunity to overcome a number of these challenges.

TECHNOLOGY AND THE TEACHING OF STATISTICS

Accessibility and cost of both hardware and software have been limiting factors in the uptake of statistically enabled technologies in secondary mathematics classrooms. These challenges can now be largely overcome through the increasing affordability of computer hardware, the ready availability of freeware and shareware over the Internet, and the maturation of hand-held computing technologies such as graphing calculators. There now exists a range of free and low-cost statistics software available from the Internet that can provide powerful demonstrations of statistical ideas and concepts. One example is WinStat (available from http://www.exeter.edu/public/peanut.html) which dynamically updates the least-squares line of regression and its equation as the user plots or drags points on a scatterplot. The effect of outliers and influential points on a line of regression can be demonstrated in a striking fashion in just a few minutes. A single computer with classroom display technology (such as a large screen TV) is all that is required to take advantage of such software in the classroom context.

Graphing calculator technology is now providing a genuine alternative to the use of computers in mathematics instruction. Devices such as the TI-80 and the Casio fx7400 are now available at prices that make them arguably affordable for the majority of students. Putting such technology into the hands of students for homework, classwork and assessment has the potential to transform how statistics is taught at the secondary level. The pen and paper techniques and calculations that dominate the traditional teaching of statistics can now be done with much less effort using a calculator, freeing the students to concentrate on the analysis and interpretation of data. Mid-range graphing calculators are now available with substantial statistical capabilities including inferential statistics functions. The functionality of a graphing calculator can be further enhanced by programs, developed either by the user or available from the Internet.

DATA

There are two sources of real data within the context of this paper - data gathered by others, and data gathered by students themselves. Sources of data gathered by others are now freely available on the Internet, e.g. the Exploring Data website (http://curriculum.qed.qld.gov.au/kla/eda/datasets.htm) has a Datasets page which contains a collection of datasets available in Excel 4.0, tab-delimited and NCSS Jr. 6.0 formats, along with the story behind the data. The Resources page of this site (http://curriculum.qed.qld.gov.au/kla/eda/resource.htm#datasets) has links to other sources of datasets on the Internet. Peripherals such as the Texas Instruments Calculator Based Laboratory (CBL) and the Calculator Based Ranger (CBR), or Casio’s Data Logger provide a method of gathering data in real time using a range of probes to measure temperature, light, motion, pressure and many other quantities. They offer the advantage of gathering data over which students can feel ownership. The issue of relevance is also addressed because students have gathered the data for themselves and the data gathering process is dynamic rather than static, enhancing the prospect of student engagement in whatever they are investigating.

UNI-VARIATE DATA

An example of a problem that is best tackled with access to technology is *Bradmanesque*, from the AToMIC project (Geiger, McKinley and O’Brien 1997).

Much was made of the outstanding batting performance of Steve Waugh for the Australian cricket team over the seasons 1993-1995. Some commentators believe that over this time his performance ranked with the best players of all time and point to his batting average during this period to support this view. This is of course a difficult comparison to make because different players in different eras competed under a great range of conditions. It should be noted that a number of other players during the 1993-1995 seasons also recorded very high batting averages, so it might well be, for whatever reason, batting was relatively easy during this time. How then can a comparison be made between Steve Waugh's performance and those of players in other eras?

One approach to solving this problem is to examine the averages of batsmen who played at least 10 matches for their country and averaged 40 or more for the periods 1993-1995, 1980-1989, and 1970-1979. Based on averages alone the best batsmen in their respective periods are S R Waugh, M A Taylor, and C A Davis. To determine the best player regardless of era the relative performance of players measured against their peers, i.e. their relative positions within a distribution, must be investigated. A good way to get a visual impression of these distributions is to look at the comparative box plots for the three periods (Figure 1).

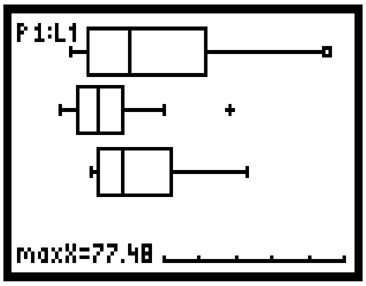

Figure 1

The boxplots (produced using a TI-83 calculator) show there are clear differences between the distributions. These appear, from top to bottom, in the order 1993-1995, 1980-1989, 1970-1979. Steve Waugh's average is better than any other player but the matter is a little more complex. Both of the other distributions are more compact. Further, the median of the period 1980-1989 is lower than any of the others and the player with the highest average for this period, M. Taylor, is a long way ahead of the rest of the players of this time. In fact his average is a genuine outlier in this distribution. The possibility exists that although Taylor's average is lower than Waugh's, he might be an outstanding player in a period where runs were difficult to score and thus, in comparison, a better batsman. Further information can be gathered by looking at the *z*-scores of the players' averages within their distributions. These scores are 2.1, 3.13 and 2.41 for Waugh, Taylor and Davis respectively, indicating that Taylor has produced the outstanding effort when compared to his peers within their designated time periods.

The advantages offered by the graphing calculator in this case include: the facility to display graphical representations of data sets for the purposes of comparison;

immediacy of the graphical analysis; the ability to use manipulate and analyse a data set that is relevant and large enough to be considered significant within the time frame of a typical school mathematics class.

BI-VARIATE DATA

In many maths curricula, statistics and algebra are taught in separate courses. As a result of this separation, students studying life-related applications of linear and nonlinear functions in an algebra class, rarely extend their study to include the statistical analysis which helps validate (or invalidate) their choice of mathematical model. Conversely, it is often the case that not all students doing an introductory statistics course have a strong grounding in algebra, which limits the subject matter to only the simplest examples of nonlinear regression. Ideally students should have both algebraic and statistical tools in their mathematical kit bag, and be able to bring both sets of tools to bear on a problem. *Striding Out* (Geiger, McKinlay and O'Brien 1997) is a problem where both algebra and statistics are needed to provide a convincing solution.

One measure of form for a runner is stride rate, defined as the number of steps per second. A runner is considered to be efficient if the stride rate is close to optimum. The stride rate is related to speed; the greater the speed, the greater the stride rate.
In a study of 21 top female runners, researchers measured the stride rate for different speeds (Nelson, R.C., Brooks, C.M., Pike, N.L., 1977). The following table gives the speed (in metres per second) versus the average stride rate of these women.

Speed	*4.83*	*5.15*	*5.33*	*5.68*	*6.09*	*6.42*	*6.74*
Stride Rate	*3.05*	*3.12*	*3.17*	*3.25*	*3.36*	*3.46*	*3.55*

The data appear to be nicely linear (Figure 2), but a residual plot (Figure 3) shows a pattern which suggests a quadratic model may be more appropriate. There is a decision to be made, however, as the quadratic model is more complex, and its predictive ability is only marginally better. Judgments about the quality of a student response would be based not just on the model chosen but also on how well the decision was supported. Because measurement error and sampling error are inherent in such situations, statistical analysis which includes residual plots should be used when fitting a mathematical model to data.

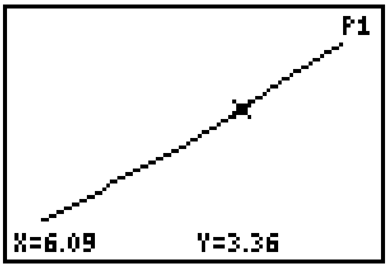

Figure 2.

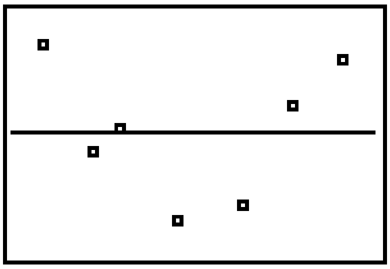

Figure 3.

The residual plots often show systematic deviations from the predicted values that are not obvious from the plot of the data. Having to interpret such a residual plot requires the student to think about the validity of the underlying mathematical model, and the process used in gathering the data.

CONCLUSION

The use of information technology in the teaching and learning of statistics is a rich field of endeavour, offering the opportunity to provide students with learning experiences that are relevant and which focus on interpretation and analysis rather than the mastery of traditional pen and paper skills and techniques. Peripherals such as the CBL unit provide contexts in which links between algebra and statistics can be investigated and explored as part of the modelling process. Further, they can give students a degree of ownership over data which is more likely to inspire interest in its analysis than data merely made up on the fly or drawn from contexts foreign to the students' realm of experience.

REFERENCES

Board of Senior Secondary School Studies (1992). *Senior Syllabus in Mathematics ABC.* Board of Senior Secondary School Studies, Brisbane, Queensland.

Geiger, V., McKinley, J., and O'Brien, G. (1997). *The AToMIC project: application to mathematics incorporating calculators*. Brisbane: Queensland Association of Mathematics Teachers.

Nelson R. C., Brooks, C. M., Pike, N. L. (1977). Biomechanical comparison of male and female runners. In *The marathon: physiological, medical, epistemiological, and psychological studies* (pp. 793-807). New York: New York Academy of Sciences.