Alternative Introductions to Applied Statistics for Mathematics Students

Robin H Lock - Canton, New York, USA

1. Introduction

In a discussion of the status of statistics education at liberal arts colleges, Moore and Roberts (1989) point out the need for a course in applied statistics which is designed for mathematics students and serves to motivate such students into pursuing careers in statistics. While this need is not unique to liberal arts colleges, it is generally more critical at smaller colleges which do not have a separate statistics department or a wide range of introductory courses. The statistics offerings at such schools are often limited to an introductory service course for non-majors and the traditional mathematics major's sequence in probability and mathematical statistics. A group of statistics educators, with support from the Sloan Foundation's New Liberal Arts Programme, has been investigating alternatives to this approach. We will argue that these courses may be insufficient for sparking an interest in statistics among mathematics majors. Our goal is to capture the interest of mathematically talented students and encourage them to pursue further work in statistics. We believe that this goal can be addressed by developing an applied statistics course which exposes students more quickly to the joys of data analysis, emphasises the applications of statistics, and utilises their ability/interest in mathematics. Some general criteria for alternative courses which head towards this goal will be given, and illustrated with brief descriptions of several model courses which have been developed by participants in a series of Statistics in the Liberal Arts Workshops (SLAW).

2. The need for alternative courses

Most mathematics students first encounter statistics, if at all, as part of a two semester probability/mathematical statistics sequence. While we believe that these are important courses which should be a key part of any mathematics curriculum, we feel that they may not be so successful in *attracting* students to study applied statistics. A student might hesitate to invest a full year in a sequence or drop out after the first semester of probability, thinking that's what statistics is all about. Even after going through the entire year, a theoretical course in mathematical statistics often slights the applications and might never impart to students the excitement of working with real data. Another approach is for mathematics students to enrol in a standard service course in applied statistics. Hopefully, they would then at least see some real applications, but such courses are typically designed with a very modest mathematical content which wouldn't challenge most math majors. Also, many mathematics programmes do not allow the introductory statistics course to count towards a mathematics major.

A major effective introduction could be achieved by offering a course emphasising the applications of statistics which also assumes and uses a reasonable level of mathematical sophistication. The proper balance between solving applied problems and maintaining mathematical rigour can be very delicate. This can be complicated in some mathematics departments by a reluctance for colleagues to recognise statistics as a distinct discipline where the problems and methods may be quite different from traditional mathematics courses. Nevertheless, we feel that mathematics students should be able to take an applied statistics course at an appropriate level, preferably before taking mathematical statistics.

3. Characteristics of suitable courses

While there are a number of possible models for implementing an introduction to applied statistics for mathematics students, we feel that certain features should be common to any such course.

(i) *The course should be data-driven.* The best way to excite students about statistics is to demonstrate the usefulness for solving interesting problems in the real world. This requires lots of real data which motivate compelling questions. These questions should be relevant to students' interests and encourage the discussion of statistical concepts and techniques.

(ii) *No previous background in statistics should be required.* If the course is to serve as an introduction to applied statistics we should not expect students to have taken a previous course (such as mathematical statistics). For maximum flexibility, it would be advantageous for the course to be structured so that math students could take it before, after, or even concurrent with, a traditional mathematical statistics course.

(iii) *Fundamental concepts of statistics should be discussed.* Although it is impractical for any course to cover the entire breadth of applied statistics, the course should illuminate basic principles, concepts, questions and modes of thought which are common and unique to statistics.

(iv) *The mathematical level should be nontrivial.* To attract mathematically talented students it is important that we dispel the myth that statistics is purely "number-crunching". Students need to see that there are interesting applications of mathematics and a solid theoretical framework for statistics. Thus, typical prerequisites for the course might be two or three semesters of calculus and perhaps a course in linear algebra.

(v) *The computer should be used liberally.* Experience with a statistical computer package is essential if we want students to get an accurate feel for how a statistician works. By freeing students from doing routine computations by hand, the computer allows them to tackle nontrivial problems with substantial data and pay more attention to the interpretation of the data and the underlying statistical concepts. This does not imply that the course should include no hand calculations or be simply a course in statistical computing.

Most of these characteristics, with the exception of the need for mathematical sophistication, should apply to any introductory course in applied statistics. However, it is precisely that mathematical ability which opens up a new range of possibilities for an introductory course. Several courses which have been developed by SLAW members are described in the next section.

4. Specific implementations

A course on Multivariate Descriptive Statistics (Frank Wolf - Carleton College). Describing multivariate data using the geometry of n-dimensional space is emphasised in this course. Linear algebra is an essential prerequisite and used extensively throughout. After reviewing basic concepts such as linear transformations, eigenvalues, and the geometry of R^n, data are introduced as points in n-space. Measures of location and variability are introduced and interpreted geometrically. Topics continue through multiple linear regression, principal components, factor analysis, clustering, multi-dimensional scaling, and discriminant analysis. In each case, the emphasis is on the descriptive nature of these techniques. Formal methods of inference and the requisite probability structure are not included. In addition to seeing a wide variety of statistical techniques for handling multivariate data, this course provides mathematics students with an excellent reinforcement of linear algebra concepts in a setting where they may really see and appreciate the applications.

A Data Analysis course (Jeff Witmer - Oberlin College). The goal of introducing mathematics students to truly applied statistics can also be achieved by offering a supplement to the probability/mathematical statistics sequence. This course is a one-credit option which may be taken concurrent with the traditional sequence. Students are exposed to real data from the start with exploratory data analysis techniques. An effort can be made to coordinate topics in the data analysis supplement with the main course. Another option, which has worked well, in teaching the course is to feature one large data set, used to motivate discussion and illustrate concepts throughout the semester. Undertaking this sort of extensive data analysis project imparts a strong sense of the wide-ranging utility of statistical techniques in a way that cannot be accomplished within the constraints of a standard course in mathematical statistics.

A course on Time Series Analysis (Robin Lock - St Lawrence University). Time series data provide a unique way for introducing students to applied statistics. The data structures encountered are typically fairly simple - often just a univariate series of historical values - yet the techniques needed for the analysis can be very sophisticated. Students are exposed to general concepts such as the selection of an underlying model, estimation of model parameters, and diagnostic checking of assumptions. Graphical techniques are used for displaying the original series, distributions of residuals, or sample autocorrelations. Methods include linear regression on t, exponential smoothing, Box-Jenkins ARIMA models, and multiple regression on concurrent series. Ample opportunities are present for tapping mathematical interests. For example, a discussion of the duality between autoregressive and moving average models leads to questions of invertibility and stationarity which depend on locations in the complex plane of roots of a characteristic polynomial. This may seem to be *only* of mathematical interest until one encounters a complicated autoregressive model which can be replaced by a moving average model with significantly fewer parameters. Real data on relevant historical series, especially economic data, are readily available.

5. Conclusion

Of course, one could envisage other models for an applied statistics course at a level appropriate to engage the interest of a typical mathematics student. Our hope is that these examples will inspire the development of courses which satisfy the criteria listed earlier. If such courses can gain acceptance and popularity among students, the number of mathematically-able students who opt to take additional coursework in statistics and seriously consider a career in the discipline should be enhanced.

Acknowledgement

As mentioned earlier, this work is a result of discussions in a series of Statistics in the Liberal Arts Workshops which have been funded for three years by the Sloan Foundation. Other members of the SLAW include: Don Bentley (Pomona College), George Cobb (Mt Holyoke College), Pete Hayslett (Colby College), David Moore (Purdue University), Tom Moore (Grinnell College), Rosemary Roberts (Bowdoin College), Jeff Witmer (Oberlin College), Frank Wolf (Carleton College).

References

Lock, R H (1990) Forecasting/Time Series : an introduction to applied statistics for mathematics students. *SLAW Technical Report No 90-1.*

Moore, T L and Roberts, R A (1989) Statistics at liberal arts colleges. *The American Statistician* **40**, 80-85.

Witmer, J (1990) Data Analysis : an adjunct to mathematical statistics at Oberlin College. *SLAW Technical Report No 90-3.*

Wolf, F L (1990) Multivariate Descriptive Statistics : an alternative introduction to statistics. *SLAW Technical Report No 90-4.*

Note: SLAW Technical Reports may be obtained from the author of this paper.