Beyond Traditional Statistical Methods

William M. Duckworth and W. Robert Stephenson

Iowa State University, Ames, IA 50011-1210

[1]

Abstract

Today's courses in statistical methods, for the most part, focus on the same methods that were taught 30 years ago. The actual practice of statistics has moved beyond these traditional statistical methods. Modern methods including dynamic graphics, nonlinear estimation, resampling and other simulation based inference methods are being used by many scientists and engineers. However, these methods generally are not included in courses in statistical methods, especially at the undergraduate level. This paper will discuss the development of a collection of instructional modules, built around actual applications from science and engineering. Each module is self-contained and includes instructional materials such as: objectives, examples, lecture materials, computer implementation of the methodology, homework, class/discussion exercises and assignments. The modules are intended as a resource for instructors to experiment with and explore the use of modern statistical methodology in undergraduate statistics methods courses. Two of the modules will be presented in some detail. We will also discuss the use of the modules in a new course that goes beyond our traditional methods courses.

[1] William M. Duckworth is Assistant Professor, Department of Statistics, Iowa State University, Ames, IA, 50011-1210 (E-mail: wmd@iastate.edu). W. Robert Stephenson is University Professor, Department of Statistics, Iowa State University, Ames, IA 50011-1210 (E-mail:wrstephe@iastate.edu) Development of modules and purchase of computing equipment was supported by an NSF/ILI Program grant DUE 9761644. In addition to the authors of this paper several other faculty members at Iowa State University have contributed to the development of the modules and the teaching of the course. They are Di Cook, Philip Dixon, Mark Kaiser, Kenneth Koehler and William Q. Meeker.

Key Words: Instructional modules, undergraduate statistics education, computer intensive

1 The Need

In the past 20 years, computers and statistical software have had a major impact on the practice of statistics and the teaching of statistical methods. In particular, computer software is now routinely used for graphics and analytical statistical methods. Change continues as computer technology improves.

Statisticians who interact directly with scientists and engineers working on cutting edge research in industry, academia, and government research laboratories or who follow the latest developments in subject-matter research journals are beginning to see a disturbing pattern. Computers have had a huge effect on the statistical tools being used (indeed, even developed) in many areas of application. These tools include, for example, neural networks in chemical engineering and finance; spectral analysis in mechanical engineering and ecology; partial least squares in chemistry; and dynamic graphics, nonlinear estimation of various types, resampling and other simulation-based inference in many disciplines. As ASA past President Jon Kettenring (1996, p 3) said

> "We smell trouble all around us. Other disciplines and organizations have been seizing opportunities that should have been ours. Example: statistics courses are taught by mathematicians, engineers, psychologists, sociologists, and others not steeped in real data experience. ... Example: the catch-up we try to play when statistical voids are filled by others in various applications areas .... Make your own list. It's not a pretty picture."

Moreover, computer software is making it extremely easy to use these methods. Without proper grounding in the principles of statistics and the limitations and required assumptions for these sophisticated methods, the potential for serious misuse is high.

One implication of Kettenring's observations is that today's statistical methods courses have fallen behind. Although the way that students interact with statistical methods in class has changed to some degree, the statistical methods in today's text books and statistical methods courses differ little from the content of 30 years ago when use of computers in statistics courses was a novelty. Correspondingly, the scientists and engineers using the more sophisticated modern methods view our current statistical methods courses as out of date, relative to methods that are available in or that can be developed rapidly using progressive, extendible statistical or mathematical packages (e.g. Splus, Gauss, Matlab, and Mathematica) and special-purpose software. Once could ask: "Shouldn't the power of the modern computer have had more of an effect on the actual methods being taught in our statistics courses?"

There are several reasons that today's statistical methods courses have fallen behind. There is a natural conservatism in our academic departments. Radical changes in curriculum are difficult and perhaps ill-advised. Some professors who did not grow up with computers are still uncomfortable with their use. Professors teaching statistical methods at many universities do not have regular contact with the real problems that are motivating the development and application of modern statistical methods. This results in a low level of awareness, motivation, and accessibility to the new methods. There also appears to be some fear of the methods that are not well understood, but that are used widely by non-statisticians (e.g. neural networks). Most undergraduate courses need a high degree of structure afforded by a solid text book or detailed notes and ancillary materials (like exercises and lab assignments). Instructional materials have not yet been developed nor been disseminated widely for many modern statistical methods.

2 The Idea

Given the current need, a team of faculty members at Iowa State University set out to develop a collection of instructional modules designed to make it easy for instructors to experiment with and explore the use of modern statistical methodology in undergraduate statistics courses. Each module is self-contained with a minimum prerequisite of a first course in statistics. Each module is motivated by, and developed around, one or more real engineering or scientific problems. The modules include, in a format standardized across the modules, all necessary instructional materials including objectives, examples, lecture materials in the form of transparency masters, handouts, computer implementation of the methodology (e.g., in Splus), homework, class/discussion exercises, assignments to test student's knowledge of basic material and reinforce important conceptual ideas, and a list of references for further study. Each module contains between 3 and 6 hours of material for presentation and discussion in class. For each module there is one or more comprehensive assignments that will require students, or groups or students, to apply the methodology in a real example, interpret the results, formulate conclusions, and write a summary report. All of these materials are being made available in electronic form at http://www.public.iastate.edu/∼stat415.

The modules place primary focus on data, concepts, models, methods, and interpretation with an emphasis on applications. When appropriate, emphasis is placed on scientific modeling of variable relationships and variability (i.e., modeling guided by subject matter knowledge and the motivating scientific question, as well as the information in the data). For real applications, computers are used to do most of the computations and create graphical displays for data analysis and model interpretation.

An understanding of intermediate level statistical methods is a prerequisite for the modules.

For example, a good understanding of simple linear regression is helpful when approaching logistic regression and nonlinear regression models. However, a detailed understanding of underlying statistical theory is *not* necessary to apply the methods in these modules. For example, likelihood methods of estimation and testing are used in many of the modules. These methods are motivated and explained without going into the underlying theoretical details.

3 The Modules

The eight modules currently in development are listed below. This list, with links to the individual modules, can be found on the Web at http://www.public.iastate.edu/∼stat415.

1. Principles of graphical methods for high-dimensional data.
2. Principles of maximum likelihood estimation and the analysis of censored data.
3. Binary response and logistic regression analysis.
4. Resampling and other simulation-based inference methods.
5. Linear mixed effect models.
6. Repeated measures data and random parameters models.
7. Modeling biological and physical mechanisms with random-parameter models.
8. Model free curve fitting.

The modules are varied in their motivation and approach. Most of the modules take advantage of one or more of the following recent developments in the modern practice of statistics:

- high-resolution dynamic graphics to view data and fitted models

- the ability to fit scientifically-specified parametric nonlinear models (instead of traditional polynomial models)

- the ability to fit nonparametric local-smoothing regression model (again instead of traditional polynomial models)

- the ability to use appropriate models for multiple sources of variability in scientific/engineering problems

- the use of simulation to replace the potentially-crude large-sample approximations.

More details on the binary response and logistic regression module and the module on resampling and other simulation-based inference are given below. Following that, two other modules are described to give the flavor of the types of examples and methods found in the modules.

3.1 Binary Response and Logistic Regression

Data involving the relationship between explanatory variables and binary responses abound in just about every discipline from engineering, to the natural sciences, to medicine, to education, etc. This module looks at three examples in order to motivate and develop logistic regression. The first involves the space shuttle *Challenger*. Data on the temperature at the time of launch and whether or not the solid rocket boosters on that launch had damage to the field joints is derived from data from the Presidential Commission on the Space Shuttle *Challenger* Accident (1986). The object of the analysis of these data is to be able to predict the chance of booster rocket field joint damage given the temperature at the time of launch. A second example involves the sex of turtles. With humans, the sex of a child is a matter of genetics. With turtles, environment during the period eggs are incubated plays a significant role in the sex of the hatchlings. Data from an experiment where turtle eggs are incubated at various temperatures and the proportion of male hatchlings is recorded are used to examine the link between incubation temperature and the chance of getting a male turtle. A third example looks at the incidence of Brocho Pulmonary Dysplasia (BPD), scars on the lungs, in children having various hours of exposure to three levels of oxygen.

All of the examples mentioned above, and many others, have several things in common. They all have a binary (or categorical) response (damage/no damage, male/female, BPD/no BPD). They all involve the idea of prediction of a chance, probability, proportion or percentage. Unlike other prediction situations, what we are trying to predict is bounded below by 0 and above by 1 (or 100%).

These common features turn out to present special problems for prediction techniques such as simple linear and multiple regression. A different statistical technique, logistic regression, is developed and used in the analysis of binary response problems. The module continues by first reviewing simple linear regression procedures and the assumptions behind them. The curvilinear response of the binary response data leads to the idea of a logit transformation and eventually to the idea of maximum likelihood estimation of the parameters in a logistic regression model. The analysis is carried further by looking at ways of assessing the fit of the logistic regression model and the significance of parameters in that model. These ideas are then extended to a multiple logistic regression model. Topics of model selection and multicollinearity of explanatory variables are discussed.

Some of the exercises are aimed at reproducing the analysis presented in the chapter that accompanies this module. Other exercises look at the application of logistic regression methods to other sets of binary response data. The sample exam problems give additional examples and applications of these methods.

3.2 Resampling and Other Simulation-based Inference

One of the primary activities of statisticians is estimation. When studying a population, one is often interested in estimating a population parameter based on the information contained in a single sample from the population of interest. This module begins with a review of point estimation and confidence intervals for the mean of a population. This provides a common base of knowledge familiar to students. Problems occur when the population standard deviation is not known, or the population shape is not normal and the sample size is too small to invoke the central limit theorem. Most traditional courses in statistical methods do not provide alternative methods to address these problems. In this module, the general idea of estimation is approached using simulation methods.

The first example involves the estimation of the 75th percentile of a Normal distribution based on a sample of size 25. A straight forward point estimate is given by the 19th order statistic. Finding a confidence interval is much more difficult. Simulation is used to learn about the distribution of the 19th order statistic of a random sample of 25 from a normal distributed population. This simulation is extended to introduce the idea of a parametric bootstrap and eventually the nonparametric bootstrap.

The module then discusses ways of improving the naive bootstrap confidence intervals. These include the bootstrap t-method, the variance-stabilized bootstrap t-method, the bias-corrected and accelerated (BC_a) method and the approximate bootstrap confidence (ABC) method. The bootstrap is then applied to simple linear regression.

In order to illustrate the power of the bootstrap, and simulation in general, one of the exercises looks at a set of data on the length of stay in hospital for patients diagnosed with digestive system disorders during one year. The data is used as the "population of interest" and students must use bootstrap techniques to learn about this population, and its parameters, given a random sample of 18 observations. Length of stay day is highly skewed and the sample size is too small to allow for much confidence in the usual normal theory methods. Students learn how badly normal theory methods perform as well as the effectiveness of the various bootstrap confidence intervals by examining simulated coverage probabilities.

3.3 Random Parameter Models

Many elementary statistics courses introduce the binomial and Poisson distributions for modeling count data. Although these distributions are useful in some applications, they are not adequate to describe the extra variability that is present (and often of scientific interest) when modeling some populations. In this module, students are introduced to the concept of random parameters and using mixture distributions to model scientific phenomena. The approach centers on the use of mixing distributions to model biological or physical mechanisms that have extra variability.

One motivating example is provided by a veterinary vaccine efficacy test procedure conducted by Dr. Jeffrey J. Zimmerman, of the College of Veterinary Medicine at Iowa State University. In this study, control and vaccinated groups of adult female pigs were challenged by exposure to a viral agent capable of affecting the proportion of live, normal births of piglets. Because of litter to litter variability in proportions, the standard binomial model is inadequate. The number of live births in a litter is modeled with a conditionally independent binomial distribution with proportions being independent and identically distributed from a beta distribution. Marginal distributions for data are conceptualized as weighted averages of binomial distributions so that calculus is not a pre-requisite for use of the module. In general, however, computationally intensive methods are required for fitting such mixture models. A test for efficacy of the vaccine can be based on a likelihood ratio test comparing the mixing beta distributions for the control and vaccinated groups.

Examples of other beta-binomial model applications include a study of foraging efficiency (proportion of successful attempts) in adult and immature green-backed herons described in Kaiser and Reid (1987) and a study of the effect of selenium contamination on normal birth in *Gambusia* (a fish that gives live birth) provided by Dr. Michael K. Saiki of the US Fish and Wildlife Service.

3.4 Nonlinear Regression

Meeker et al. (1998) describe the results of an experiment conducted by Dr. John Macdonald of the Martin-Lockheed Company to evaluate the reliability of a device to be installed in a communications satellite. Over the life of the satellite, the device's power output degrades. High-temperature acceleration was used to rapidly assess the degradation rate of a sample of devices. The available data consist of a sample of power-drop paths following a nonlinear function, the form of which is known from the approximate chemical model of the degradation process. There is, however, unit-to-unit variability in the asymptote and degradation rates of the devices. The unit-to-unit variability arises from material and manufacturing differences and must be modeled to answer the engineering questions concerning device reliability. An appropriate engineering model for this problem is a nonlinear mixed-effect regression model. Approximate maximum likelihood estimation for this model, given sufficient computing power, is straightforward to accomplish using the methods described in Pinheiro and Bates (1995). Their algorithm is implemented in Splus.

4 The Computer

The computer is an integral part of each module. All methods discussed require computer software to analyze the data. We have chosen to develop modules using Splus as the default package. For some of the modules, like that of the nonlinear regression discussed above, Splus is the computing package of choice. For other modules, like the binary response and logistic regression, other standard statistical packages (e.g., SAS, Minitab) can be used.

Since computing is such an important part of each of the modules we provide Splus code to accompany each of the modules. This code is found in the chapters that accompany the module and in the exercises and their solutions. Including the code allows instructors, and students, to focus on the results of the computations and not on getting the output one needs. Still, for some exercises, students must be comfortable enough with the computing to adapt the Splus code to a new application.

5 The Course

The modules are designed to be self-contained so that any one of them can be inserted into an existing course in statistical methods. For example, in a course in regression analysis, standard regression techniques could be followed by modules on binary response and logistic regression, model free curve fitting and nonlinear regression. Alternatively, standard methods of inference could be augmented by including the module on resampling and simulation-based inference methods.

Each module will have a chapter, a set of overhead masters, a set of handouts with six transparencies per page, sample exam questions and solutions, and sample exercises and solutions. We are including solutions to the sample exam questions and sample exercises on the Web site.

This allows both students and instructors to see examples of what might be expected. In using a module, an instructor would need to come up with her/his own exam questions.

Alternatively, several modules can be put together into a full semester course in advanced statistical methods. At Iowa State University we have taught, on an experimental basis, Stat 415: Advanced Statistical Methods for Research Workers in Spring 1999, Spring 2000 and Spring 2001. The pre-requisite for this new course is Stat 401: Statistical Methods for Research Workers. Stat 401 is a traditional methods course taken by undergraduate statistics majors and minors and graduate students outside of statistics. Stat 415 is a 3 credit semester long course. The course meets each week for lecture on Monday and Friday for 50 minutes each day. The lecture is used to present material, motivate the new methods and discuss the context of the data and the interpretation of results from the analysis of those data. On Wednesday there is a one hour and fifty minute laboratory period. The laboratory is held in a computing lab and the focus is on the use of the computer to explore and analyze data. The students have the opportunity in lab to actually do the computing to reproduce the results discussed in lecture. There is also time for students to work on homework and group exercises.

In Spring 1999 there were six modules presented. These included graphical methods for high-dimensional data, an introduction to maximum likelihood estimation and the analysis of censored time-to-event data, binary response and logistic regression, resampling and simulation-based inference techniques, random parameter models for biological and physical mechanisms, and nonlinear regression and analysis of repeated measures. In Spring 2000, new modules on linear mixed effect models incorporating multiple sources of variation, and model free curve fitting were added. No new modules were included in Spring 2001.

The students in the course are evaluated on the basis of their performance on assignments that accompanied each module and on three 2-hour exams during the course of the semester.

The assignments require each student to be able to do computing but also to interpret the results within the context of the stated problem Each 2-hour exam covered 2 to 3 modules. Although most of the exam questions are done in-class, some modules lend themselves more to examination at the computer.

The enrollment in the course has been low. This may be attributable to the fact that it has been an experimental course and so has not been listed in the catalog. It is now a regular course and is listed in the upcoming catalog. The reaction of the students to the course has been fairly consistent. Students indicate that they learn a great deal. They also indicate that there is a lot to learn. In fact, many have expressed the opinion that there is too much material. Because of the modular nature of the materials, the students have indicated that there is a lack of smooth transitions from topic to topic.

In the future, we plan to continue to develop the modules as stand-alone items. Many statistics departments do not have the resources to add a whole course to their offerings. However, individual instructors may wish to use one module, or several modules, in existing courses. For the course at Iowa State University we want to establish more links between the modules so that they do not appear so disjointed. For example, students now get a preview of binary response in the module on graphics. As another example, one of the applications of the bootstrap techniques might be to get confidence intervals for parameters or predictions in logistic regression. Another link would be between the model free curve fitting and the nonlinear modeling. Eventually, we will move from having seven instructors for the course to having only one. This would be a truer test of the ease with which someone else could use the modules developed for this NSF/ILI project.

References

Kaiser, M.S. and Reid, F.A. (1987), "A comparison of green-backed heron nesting in two fresh-water ecosystems," *Colonial Waterbirds*, 10, 78-83.

Kettenring, J. (1996), "Strategic Planning - The Future of Statistics," *AMSTAT NEWS*, No. 231, May 1996, Alexandria VA: The American Statistical Association.

Meeker, W.Q., Escobar, L.A. and Lu, C.J. (1998), "Accelerated Degradation Tests: Modeling and Analysis," *Technometrics*, 40, 89-99.

Pinheiro, J.C., and Bates, D.M. (1995), "Mixed effects models, methods, and classes for S and Splus." Department of Statistics, University of Wisconsin. Available from Statlib.