DATA ANALYSIS: AN ADJUNCT TO MATHEMATICAL STATISTICS
AT OBERLIN COLLEGE

JEFFREY WITMER
OBERLIN COLLEGE

TECHNICAL REPORT NO. 91-003

SLAW is supported by a grant from the Exxon Corporation
Statistics in the Liberal Arts Workshop (SLAW) is a group of educators concerned with the teaching of statistics. The workshop was initially funded by the Sloan Foundation. Continuing support has been provided by the Exxon Corporation.

Participants in the summer of 1991 are:

Donald L. Bentley  
Pomona College

George W. Cobb  
Mt. Holyoke College

Janice Gifford  
Mt. Holyoke College

Katherine T. Halvorsen  
Smith College

Homer T. Hayslett, Jr.  
Colby College

Gudmund Iversen  
Swarthmore College

Robin H. Lock  
St. Lawrence University

David S. Moore  
Purdue University

Thomas L. Moore  
Grinnell College

Norean Radke-Sharpe  
Bowdoin College

Rosemary Roberts  
Bowdoin College

Dexter Whittinghill III  
Colby College

Jeffrey Witmer  
Oberlin College
DATA ANALYSIS: AN ADJUNCT TO

MATHEMATICAL STATISTICS AT OBERLIN COLLEGE

Jeff Witmer

Associate Professor of Mathematics

Oberlin College

Oberlin, Ohio
Are SAT scores useful predictors of success in college? I led a group of mathematics majors at Oberlin College in an exploration of this question as the core of a one-credit course, entitled Data Analysis, last year. This course, MATH 337, is an adjunct to the standard, junior-level, two-semester sequence in probability and mathematical statistics that we offer each year at Oberlin. Unlike most statistics courses for mathematics majors, the Data Analysis course allows - indeed, it forces - students to "get their hands dirty" exploring real data and trying to answer real questions. Each year I select a set of data, such as the SAT data, to serve as the central focus of the course.

I believe that it is imperative that students learn something of how statistical theory is applied in practice and I try to show this side of statistics in the courses I teach at all levels. However, it is particularly difficult to cover much material on applied statistics while at the same time covering the many important topics in the mathematical statistics course - probability, random variables, functions of random variables, expectation, the central limit theorem, estimators, confidence intervals, hypothesis testing, and others - that are fundamental to the discipline and are an essential foundation for advanced (graduate) training in statistics. The Data Analysis course provides a workable solution to this problem.

Data Analysis, which meets once per week and is graded on a pass/fail basis, is intended to expose the students (most of whom are mathematics majors) to applied statistical methods, with an emphasis on the use of computers in statistics. For the Spring 1990 semester I obtained, with the help of the Oberlin administration, a random sample of 361 cases from the student records data base of Oberlin graduates. Each case included a case number (in place of the name of the student) and values of several variables: SATM and SATV scores, high school GPA, high school class rank and class size, gender, college major, home state (or
country), year of college graduation, and whether or not the student applied for financial aid. I gave a copy of the data set to each student and, after discussing with the class the appropriate precautions for handling a sensitive set of data, told the class "Our job is to find out how these variables are related to each other and which, if any of them, are good predictors of success at Oberlin. I welcome your ideas. Tell me where you want to begin your analysis."

We used the Macintosh program Data Desk as a fundamental tool in our work. I brought a Mac and a projection system into the classroom so that we could all look at the data together. The students immediately proposed hypotheses and suggested approaches for the analysis. We looked at histograms, boxplots, and scatterplots to get a feel for the data. At the beginning of the semester the students did not know what a boxplot was, for example; they only knew that they wanted to compare the SATV distribution for men with the SATV distribution for women. In the process of doing this, they learned about boxplots.

Similarly, the class had never studied regression modeling but found this to be a very important tool in understanding the data. They learned about residual plots, influential points, outliers, normal probability plots, transformations, interaction terms, multicollinearity, the use of dummy variables in regression, and other topics as they were needed.

One goal of Data Analysis is to get the students to think about real questions that do not have simple answers, to fit models to data, and to consider the dialog between data and models that are fit to the data. Since each student had a copy of the data and access to a Macintosh, I could assign specific homework problems related to the data, as well as the open invitation to "Spend time looking at the data between now and next week and tell me if you find anything interesting."

In addition to learning how and why certain statistical tools are used, the class discussed how the data were obtained and what questions the Oberlin
Admissions office wanted to have answered. We discussed the difference between observational studies and controlled experiments as well as limitations in the data set: we wanted to separate the grades received in science courses at Oberlin from other grades, to see if SATM scores were highly correlated with success in science, but this was not possible. We also discussed the many criteria for success other than college GPA. At the end of the semester we agreed on how to report our findings to the Dean of Enrollment Planning.

We did spend class time on topics that did not arise in our analysis of the SAT data, such as statistical quality control, including some of the ideas of Deming and Taguchi, simple control charts, experimental design, bias, randomized response sampling, capture-recapture methods, and Simpson's paradox.

In the mathematical statistics sequence students learn the mathematical theory that underlies statistical practice, but they see little of the applied side of the discipline. I find, to no surprise, that I could easily spend more than the allotted three class periods per week presenting a mathematically rigorous (by undergraduate standards) treatment of statistics. Of course, my colleagues in the Mathematics Department want me to prove theorems and to use mathematical tools freely and precisely in teaching at this level. Moreover, I feel that it is important that I offer this sequence, which provides a marked contrast to the introductory courses I teach for social science students, in which I effectively say “Here is the t-test; don’t worry about how the percentiles in this t-table were generated, just use them.” Nonetheless, mathematical statistics presents only one side of the discipline. By offering the Data Analysis course I can show students another side of statistics.

In Data Analysis the students are actively involved in the learning process; the open-ended nature of the course calls for an adjustment in teaching style.
Rather than being in charge of what is happening in the classroom, I have to react to what the students do with the data. I also have to curb my impulse to immediately tell them what I see in the data, lest they fall back into the position of being passive note-takers. This change from my traditional use of lectures is a bit of a challenge, but is very rewarding.

The reactions of students to this course have been positive. I believe, and they seem to agree, that seeing statistical methods applied to real data motivates students to want to learn more about the subject. A one-credit course in data analysis is a good way to give mathematics majors a wider view of statistics than that gained in studying mathematical statistics. (By the way, our answer to the question posed at the beginning of this article is, at least for Oberlin, Yes.)