DATA ANALYSIS: AN ADJUNCT TO MATHEMATICAL STATISTICS
AT OBERLIN COLLEGE

JEFFREY WITMER
OBERLIN COLLEGE

TECHNICAL REPORT NO. 90-003

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
I teach a standard, junior-level, two-semester sequence in probability and mathematical statistics, MATH 335-336, at Oberlin College. This sequence covers a great many topics that statisticians have come to see as fundamental to the discipline and as an essential foundation for advanced (graduate) training in statistics, including probability, random variables, functions of random variables, expectation, the central limit theorem, estimators, confidence intervals, hypothesis testing, and regression, among others.

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 MATH 335-336. 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."

However, there are problems associated with the mathematical statistics sequence. Most of the students who take this sequence have no previous experience with statistical applications or with data. In MATH 335-336 the students learn the mathematical theory that underlies statistical practice, but they see little of the applied side of the discipline - there only so much that we can do in two semesters! Although they learn about sampling
distributions and large-sample properties of estimators, they
learn little about the concerns practicing statisticians have
about how samples are actually drawn: experimental design,
randomization, bias, etc.

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. It
is particularly difficult to cover much material on applied
statistics while at the same time covering the mathematical
statistics topics outlined above. At Oberlin, I address this
problem by offering an additional, one-credit, course at the upper
level.

Several years ago Bob Geitz, who was teaching mathematical
statistics at the time, developed MATH 337 - DATA ANALYSIS - as an
adjunct to MATH 336. This course, 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. Over the past few years MATH 337 has evolved. I made
some changes to the course when I first taught it in 1987. The
following is a brief sketch of what I covered during the spring
semester of 1988.

I started the course by discussing some of the statistical
packages that are available on computers. I used MINITAB, which I
had on-line, with the aid of television monitors, while teaching
MATH 337. I then presented some ideas from exploratory data
analysis, including stem-leaf diagrams, histograms, boxplots, and
median traces. My philosophy is that concepts should be introduced through the use of real data. Thus, for example, in presenting stem-leaf diagrams I used seasonal snowfall totals in Cleveland; when discussing boxplots I showed the students side-by-side boxplots of the heights of men and of women in a large class at Oberlin.

I presented the use of normal probability plots and of transformations of data. We then discussed compound smoothers and applied them to data such as interstate highway fatality rates in Ohio over several years. I spent almost two class sessions on statistical quality control, including some of the ideas of Deming and Taguchi, simple control charts (such as $x$-charts), and a bit of experimental design.

Unfortunately, the material on the theory underlying regression comes near the end of the course in MATH 336, so it was difficult to coordinate the applied coverage of regression in DATA ANALYSIS with the theoretical coverage in MATHEMATICAL STATISTICS. Nonetheless, I devoted roughly half of the DATA ANALYSIS course to the general topic of regression, starting with scatter diagrams and residual plots for simple linear regression. I presented the famous Anscombe collection of four identical fitted regression models from four drastically differing settings to demonstrate the power of residual plots and to get the students to think about diagnostics, the data that generate fitted equations, and the dialog between models and data (although this was one case in which I deviated from my rule of presenting only real data in class). The Anscombe plots led to a discussion of influential
points and outliers in regression. We discussed studentized residuals and the Bonferroni procedure for testing for outliers. We then considered multiple regression, t-tests for individual coefficients, and the general F-test for nested models.

I presented the use of dummy variables in regression as a means of fitting an ANOVA model and of conducting an analysis of covariance. We considered the use of interaction terms in regression and discussed multicollinearity in multiple regression. This led to a brief consideration of experimental design for regression settings. Finally, we examined the effects of various transformations, such as a log transformation, on scatter plots and normal probability plots of residuals. I relied heavily on MINITAB through the unit on regression.

We next considered the analysis of paired data and returned to the topic of analysis of variance to consider two-way ANOVA models, randomized block designs, and interaction plots. I ended the course with discussions of the difference between ignoring and controlling for a variable, Simpson's paradox and observational versus controlled experimental studies.

Unfortunately, we lost a few class meetings and did not have time to cover other topics that I wanted to present, including randomized response sampling and capture-recapture methods.

In the spring of 1989 I handled MATH 337 in a different fashion for much of the semester. Rather than present lectures to the students, I involved the entire class in a data analysis project. We analyzed the results of a survey I helped conduct of libraries at liberal arts colleges. The students helped me
explore roughly 200 variables measured on 97 colleges. In the process, we used several tools of exploratory data analysis and regression. I gave each of the students access to the computer file that contained the data and told them "Explore, generate graphs, fit models, and let me know what you learn." By looking at dotplots, boxplots, and scatter plots the students found many unusual observations (some of which were typographical errors), and interesting patterns. We spent class periods discussing the data set and the statistical methods that we were using to analyze it.

I am planning to teach MATH 337 in 1990 in a similar fashion. This time the class will explore the relationship between SAT scores and academic success at Oberlin.

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.