# Experiences in Training Students in Statistical Consulting and Data Analysis

Gordon K Smyth - Brisbane, Australia

## 1. Introduction

This paper describes some of my experiences teaching consulting and data analysis at the University of California, Santa Barbara, and at The University of Queensland, Australia. Both universities have formal data analysis courses. At the University of California, however, there is also an organised consulting service, the "StatLab", directed by a faculty member and staffed by graduate students. The graduate students discharge their Teaching Assistant obligations by consulting under the supervision of the Director. The aims of the StatLab are: (1) to give graduate students experience in consulting, (2) to provide statitical assistance to researchers and graduate students on campus, and (3) to provide a commercial consulting service to the general community. At The University of Queensland, graduate students are drastically more scarce, and there is no consulting laboratory along the above lines. It is still possible, though, to give a small number of the most interested students experience in consulting through Research Assistantships, Vacation Scholarships, and Honours projects.

How we teach consulting depends of course on how we see the consulting process itself. I find it useful to emphasise four distinct steps: (1) defining the problem, (2) initial examination of the data using tables and graphs, (3) more formal analysis as appropriate using parametric models and statistical tests, and (4) presenting conclusions. I often ask students in my data analysis courses to write reports in four sections corresponding to the above four steps. Most of the following discussion is organised under the same headings.

This paper is based entirely on my own experience. Other accounts of the consulting process are contained in Boen and Zahn (1982) and Hand and Everitt (1987), and in the references therein. Other accounts of teaching consulting are given in Rustagi and Wolfe (1982).

## 2. Defining and tackling the problem

*Interviewing the client:* Most students are surprised to discover how long it takes to grasp what a client's problem is all about. A consultation with a new client seldom takes less than an hour, and most of that time is spent uncovering the real nature of the problem. Clients, through long familiarity with their study, often take for granted and fail to mention details which are highly significant and unexpected to the statistician. In ongoing collaborative projects, features of the physical problem may keep coming to light months after discussion has begun.

It is important to have the client describe the experiment in concrete terms. One of the questions that I ask most often is "What does $x$ physically mean; what are you actually measuring?". Clients occasionally feel that they can save time for the statistician by giving him or her only a simplified abstraction of the real problem. Or they feel insecure about discussing the prosaic details of their experiment with a mathematician, and so hide them behind a mathematical pseudo-problem.

Clients are seldom accustomed to speaking in mathematical terms, nor are statisticians usually familiar with the jargon of the client's specialty. I believe that the onus is on the statistician to talk in the client's language. I usually find it useful to talk about "chi-squared statistics" rather than "deviances" or "likelihood ratio tests", and perhaps "variability" or "standard error" rather than "variance". More deeply, as one tries to assimilate the concrete problem, one tries to talk to the client in those concrete terms, and to draw all conclusions back to those terms.

To teach good "desk side manner" one must involve students in actual consulting sessions. In Santa Barbara I tried to involve at least one student in each consultation, actively if possible, but passively if necessary. In Queensland, students can sometimes have a lot of contact with a client when involved on a collaborative project.

*Identifying the questions of interest:* I feel that accurately identifying the basic questions of interest is the most fundamental issue of consulting or data analysis, and I emphasise it accordingly to students. I often ask clients early in a consultation "What is the basic question you are trying to answer with this data set?", or, less directly, "How will you use the estimates from this model?". The precision with which this question can be answered determines in good part how successful the consultation will be. It can be quite different to ask "Is treatment A better than the rest?" instead of "Which treatment is the best?". One needs to know whether the analysis is exploratory or confirmatory, which variables are blocking factors or covariates, and which are the variables of interest, which is the response, random or fixed effect, and so on.

In an academic environment, most analyses are primarily exploratory. This affects the choice of test size (if one is testing hypotheses) and formality of analysis in general. It is useful to emphasise this to students, since it gives them the confidence to loosen up their common sense - for example, to observe that "the data suggests" such and such, instead of simply giving a list of which contrasts are significant at the prescribed level.

*Displaying and describing the data:* Most students want to push the data, without looking at it, straight into whatever the obvious parametric procedure is. I learnt myself only through the hard experience of having to re-do the parametric calculations, that it really does pay to begin a data analysis with thoughtful data display and description. Apart from the fact that the data description may settle by itself the questions of interest to everybody's satisfaction, it will make evident all sorts of salient features of the data which may be hidden in a parametric analysis with its low dimensional focus. It will make clear what the focus of the formal analysis should be, and how much trust can be placed in it.

Data description also allows one to assess the quality of the database. Medium or large data sets are usually full of gross errors that can be made obvious with appropriate displays and often corrected. Unfortunately, only compulsive obsessive personalities seem to take readily to this task. During my time as StatLab Director, I was constantly reminded of my failure to persuade students of the importance of checking the data.

Data display has been almost entirely absent from traditional mathematical statistics courses. At both Santa Barbara and Queensland, students see it for the first time in the data analysis course just before they graduate. I hope this will change soon with the movement to make statistics courses data driven from an early stage.

*Formal analysis:* Having made an initial study of the data, one will want, data willing, to fit a formal parametric statistical model. This allows one to fine-tune estimators and to be quantitative about the surety with which inferences are made. A good consultant uses whatever methodologies are required by the data and the questions to be addressed, not necessarily those suggested by his or her research pre-occupations. Students therefore need a wide overview of what techniques are available, and an understanding of when they are applicable. As a practical matter, one is generally restricted to techniques that are easily available in statistical packages or very simply programmed.

Not all data sets are suited to formal analyses, and this is a useful lesson in itself. When there is too much data, or too little, or the assumptions are too unclear, it may be best to stop with the data display. A related lesson, when one does perform a parametric analysis, is that simple and interpretable models are best.

At the stage that I see students - towards the end of their undergraduate degree in Queensland, or towards the end of their graduate programme in California - they know something about normal theory ANOVA and multiple regression and, perhaps, multivariate analysis. They don't know how to apply these methodologies, for example to choose the best model out of several covariates in a multiple regression or to decide when an ANOVA has a random rather than a fixed effect. They certainly don't know anything about unbalanced designs, residual analysis, simultaneous inference, or any sort of non-normal analysis including categorical analysis. I have found it necessary to teach many statistical methodologies, especially logistic regression and log-linear analysis, at the same time as introducing data sets to which they are applicable. Categorical, especially dichotomous, data is so common that it seems indefensible to send students out as statistics graduates who know nothing about its analysis.

Having settled on a parametric model, one will further test its assumptions by examining the residuals. Residual analysis is now very well established, and students accept it very easily.

*Presenting the conclusions:* The conclusions of an analysis must be presented in the concrete terms of the physical problem, not in terminology of mathematical statistics. If the question was "is detergent A better than detergent B", the answer is not "factor DET is significant at the 1% level", which is what students write until they have gained confidence with the idea of using ordinary language. The conclusions will be supported by plots or tables, either of the parameters of the formal model or, more often, the corresponding data display.

The presentation of conclusions emphasises Fisher's principle of data summary. Data summary is the basic process of data analysis, i.e. summarising what the data has to say about the questions of interest. Tables and descriptive statistics summarise data in an obvious way. Parametric models provide the ultimate reduction of data into a finite number of parameter estimates. I teach students that the complexity of the model determines the complexity of the data summary. If an additive ANOVA model fits, then one-way tables of means for each factor are an adequate final presentation. If interactions are present, one needs two-way or higher tables. This is in a sense the meaning of (fixed effect) ANOVA, to determine the complexity of the tables necessary to describe the salient features of the data. Log-linear analysis of contingency tables is exactly analogous, with relative frequencies replacing means.

## 3. Teaching aspects

*Course structure:* In formal courses I hand out data sets, with a couple of paragraphs describing the background and purpose of the study, and have the students prepare written reports of their analyses. I encourage them to address their report to the experimenter. Student reactions range from "Collating computer output is boring" to "This is great, it [statistical theory] all comes together at last." In the StatLab, students work on "live" consulting problems. They spend a majority of their time performing the calculations for steps two and three of the analysis using statistical packages, but are also involved in the initial interviews and the final report writing.

When choosing data sets for data analysis courses, I try to avoid problems which are too large and complicated, too full of missing data, or too nonstandard, to be done justice to in a reasonable time at the level of the course. Usually also I avoid data sets which don't lend themselves to a formal analysis.

I have found it valuable to have a regular weekly meeting with the consulting students, usually for one hour at the same time each week. We meet even if there has been no particular new development during the week. This avoids the project which gets "lost" for a month or so. It also encourages students to show me partly completed work, giving me the opportunity to pick up problems at an early stage, encourages the students to ask questions, and provides regular non-traumatic deadlines.

*Software and hardware:* Where possible, StatLab tries to help a client to perform his or her own analysis rather than taking over the data. This promotes statistical education and conserves the consultant's time, which seems very scarce compared with the volume of data requiring analysis! This approach, of course, determines both hardware and software - one uses what the client has - and requires the consultant to acquire some knowledge of a very wide variety of software.

One takes over an analysis if the calculations are non-standard or too difficult, or if the client is paying! The diverse strengths of the different software packages have encouraged us to take an eclectic approach. In the StatLab, we used mainly Systat, followed by SAS, GLIM, S-Plus, Gauss and occasionally SPSS. We used mainly PCs, sometimes Sun workstations (S-Plus) or IBM mainframe (SAS with large data sets). When teaching the consulting course in Santa Barbara, I allowed students to use their own preferred packages. They responded with SAS, Systat, GLIM, Statistix, BMDP and S-Plus, in roughly that order.

In Queensland, we teach with a combination of Minitab and GLIM on a Unix super-mini computer. When supervising students on collaborative projects, I have used GLIM, followed by Minitab, S, Gauss and PLUM, mainly on the super-mini but sometimes on a PC or Sun workstation.

*Commercial versus academic consulting:* In Santa Barbara about a third of the total consulting effort of the StatLab is for off-campus clients, for which payment is requested. Commercial consulting broadens the range of problems that academics and students are exposed to. If well done, it promotes the university and the discipline of statistics in the community, and makes it easier to place graduates. The income earned provides jobs for graduates and financial flexibility for the host department.

The distinguishing feature of commercial consulting is the over-riding need for timeliness, and this can be difficult to achieve when supervising students who are learning as they go. Commercial consulting is often quite tedious, perhaps involving very low level analysis of a large body of data. An occasional problem is the paying customer who believes he is buying a service, like dry cleaning, to which his data can be subjected to without any thought on his own part. (A more accurate simile is that statistical consultants are like therapists: we help clients to think more clearly about and to solve their own problems.) Most clients though are well aware of the value for money that they receive from the university service.

## 4. Conclusion

I don't think there is any question that teaching consulting is worthwhile and rewarding. Consulting and data analysis require many skills not taught in theoretical courses. For students who are not planning to become mathematical researchers, consulting and data analysis amount to the actual practice of their craft - theoretical studies are just background. To understand how to go about it, students need exposure to real data, and this should begin sufficiently early in their courses for them to absorb the lessons it brings. Consideration of complete problems, from a client's point of view, provides students with a coherent overview of statistics, bringing together and demonstrating the purpose of otherwise unmotivated results.

## References

Boen, J R and Zahn, D A (1982) *The Human Side of Statistical Consulting.* Lifetime Learning Publications, Belmont, California.

Hand, D J and Everitt, B S (eds) (1987) *The Statistical Consultant in Action.* Cambridge University Press, Cambridge.

Rustagi, J S and Wolfe, D A (eds) (1982) *Teaching Statistics and Statistical Consulting.* Academic Press, New York.