Statistical Science
1998, Vol. 13, No. 1, 1–29

Consulting: Real Problems, Real Interactions, Real Outcomes

Richard Tweedie (Resources Appendix by Sue Taylor)

Abstract. The Pullman meeting of IMS–WNAR had, as one of its themes, "Statistical consulting." In this overview of the case studies presented there, an attempt is made to draw together some of the lessons of these papers, showing the diverse role of the statistician in collecting, analyzing and presenting the information contained in the data.

Key words and phrases: Statistical consulting, client interactions, consulting bibliography, consulting resources

1. THE PULLMAN SESSIONS

The papers in the Pullman panel discussion were presented by invitation at the IMS–WNAR meeting held at Washington State University in June 1996.

I organized this session. I have, over the years, organized and participated in a number of such sessions at conferences both in the United States and in Australia, and it is clearly a perennial topic for statistical meetings, only perhaps rivalled by sessions on the gap between "academic" and "practical" statisticians, or sessions on how to teach undergraduate service courses in our notoriously uncharismatic subject.

This time I was asked to organize the session in order to atone for, or perhaps amplify, some comments I had made on the role of consultants some years previously. In Tweedie (1992) I wrote that the comments on consulting of the Committee on New Researchers (CNR; New Researchers Committee of the IMS, 1991) were "glib": in particular, I disagreed rather strongly with the Committee statement that "... unless you need the data analysis experience your role [as a consultant] is to dispense advice. The client should be responsible for the actual analysis." This seemed to me to be very optimistic, or perhaps very pessimistic depending on one's viewpoint. In almost any interaction I have ever been involved with, a statistician, especially a new researcher, is not usually seen by clients as a guru on a mountaintop. Statisticians will not go far if they adopt that role, at least not in a real consulting situation where it is critical to understand a variety of client–consultant interactions that will influence the requirements for effective consultation: see the Appendix for numerous views on the real complexity of the consulting role.

Moreover, the CNR article advises that "... if you have put in substantial effort, in terms of time or developing new techniques, you should ask to be a co-author" on the paper that is assumed to be the end-product of consulting. I felt that this also indicated a serious misapprehension about what most consulting was about, even for those whose role was to be "new researchers" rather than full-time consulting statisticians. (Of course, one might be justifiably cynical about the reasons for these statements: given the prevailing criteria for promotion in academia in particular, the CNR may have been more realistic than one might wish in their advice.)

Even so, in the belief that consulting is valuable to academic and perhaps more particularly nonacademic statisticians, with Deb Nolan and LuAnn Johnson (the conference organizers at Pullman) a double-barrelled session was organized: the first half would describe some real consulting problems that might illustrate the range of activities that consultants face, of interest in their own right as well as illustrative of the many facets of our profession beyond mere advising and coauthorship; and the second part would be a panel discussion, giving

Richard Tweedie is Chair, Department of Statistics, Colorado State University, Fort Collins, Colorado 80523-1877 (e-mail: tweedie@stat.colostate.edu), and Director of the Center for Applied Statistical Expertise there; he was a consultant in CSIRO and the private sector in Australia (1974–1987). Sue Taylor was Consultant, Flinders University of South Australia (1990–1995); she is now a Ph.D. student, University of Colorado Health Sciences Center, Denver, Colorado, and an active consultant in epidemiology and medical statistics.

anecdotes and advice and insights, with (we hoped) strong audience participation.

This worked out surprisingly well, and the papers that follow are from the first part of the session: regrettably, the insight, experience and wit of the second part are lost to all but those who were there, although the bibliography and WWW resources appended will enable interested readers to delve further into the available literature on this topic and share at least one of the more tangible bits of advice from the panel.

2. THE PARTICIPANTS

In what follows you will see four articles from four rather different perspectives. When inviting speakers, we looked for a spectrum of statistical backgrounds: the participants represent the experiences of graduate students, of university faculty and of statisticians working full time as consultants in academia, in-house with a government instrumentality and in a private sector capacity.

Karl Broman, a Ph.D. student from U.C. Berkeley, illustrates the best of the classic "on-campus" interactions: a good scientific collaboration (even a potential co-authorship!); some new techniques needed and developed; and an outcome of academic value to both parties. This is very much in the mold of the projects envisioned in the CNR article, but even here the consultant is doing the analysis, not just assuming the client will carry it out.

Jennifer Hoeting, a new researcher now at Colorado State, describes the role of the academic statistician as funded advisor–consultant on a project. She illustrates the way in which careful investigation of the data sources is vital for any analysis to be meaningful (and lack of knowledge of the data collection, or a limited client understanding of the data under study, can cause a well-planned analysis to be inappropriate).

Sue Taylor describes the sort of rare project where the statistician is actually involved early enough to be able to influence data collection and enhance the ability to analyze. As a consultant to the Australian Longitudinal Study of Ageing, she was able to ensure (with much work) that analysis would actually be carried out on reasonably clean and reliable data, thus saving a large amount of later work in analysis.

Missing is the paper from Bob O'Brien from Battelle. From the perspective of an in-house consultant, he described a major environmental problem, and one where much of the consultant's role lay in trying to determine what the real goals and constraints were, since there were very many stakeholders in the problem: statistical analysis would provide the answers if only the questions were clear. This omission reveals again the priorities of many consultants: his ongoing duties preclude even the writing of a sole-authored paper.

Finally, wearing an oldish hat as a private sector consultant, I describe a situation where modelling does work, and an appropriate analysis yields a counterintuitive and effective solution—but yet again, only after the data are revisited and the whole context of the problem is understood.

3. CONCLUSIONS

Most statisticians will find something familiar in these case studies.

We know that understanding our data and the questions of the client are of paramount importance: it is reassuring to see an example where the statistician can control that process (Taylor). We know that close examination can reveal far more than the client originally told us (Hoeting). We know that problems are rarely standard, and that at its best statistical thinking can come up with new ways to cope with new problems (Broman), or perhaps more typically we can see the role of our assumptions and decide how well the old ways fit (Tweedie).

However, in the end, these case studies, and many other war stories that many of us tell or have heard, all illustrate two things overwhelmingly.

First, no matter what subject areas we enter, statistics can contribute something that was not there previously, and we have much to offer to almost everyone. These examples cover bench research problems, social surveys, management practices and environmental problems on a local and a global scale. Without statistical thinking, none of them would be solvable. Moreover, they show that it is often the mere fact of such thinking, rather than the specific technical input, that proves invaluable. It is hard to overestimate how powerfully our discipline trains us to think about complicated issues in ways that allow us to quickly diagnose difficulties in esoteric disciplines to which we have had only several minutes of introduction—a fact reinforced by these examples and even further by the referees of this collection.

But second, for that contribution to be truly at its best, the statistician must enter into the context of the problem, not just as an "advisor," but as someone prepared to understand the data, analyze the data, interact with those who really own the questions being asked and consider the impact of statistics within the real context of the problem. The

Pullman case studies show many of these attributes, but also illustrate vividly the problems we have in achieving such an idealistic state.

RESOURCES APPENDIX

Bibliography for Consultants

One of the most useful tools for consultants is a bibliography of information put together by other consultants. The papers below do not pretend to be exhaustive: they represent those that I have found of most use in my professional practice.

They also contain pointers to other material covering a wider range if needed. Those references marked with an asterisk contain an extensive list of statistical consulting references which are not duplicated here.

ASA AD HOC COMMITTEE ON PROFESSIONAL ETHICS (1983). Ethical guidelines for statistical practice: report of the Ad Hoc Committee on Professional Ethics. *Amer. Statist.* **37** 5–20.

*BASKERVILLE, J. C. (1981). A systematic study of the consulting literature as an integral part of applied training in statistics. *Amer. Statist.* **35** 121–123.

BOEN, J. R. (1972). The teaching of interpersonal relationships in statistical consulting. *Amer. Statist.* **26** 30–31.

BOEN, J. R. and FRYD, D. (1978). Six-state transactional analysis in statistical consulting. *Amer. Statist.* **32** 58–60.

BOEN, J. R. and ZAHN, D. A. (1982). *The Human Side of Statistical Consulting*. Lifetime Learning Publications, CA.

CHATFIELD, C. (1988). *Problem Solving: A Statistician's Guide*. Chapman and Hall, London.

CHATFIELD, C. (1991). Avoiding statistical pitfalls. *Statist. Sci.* **6** 240–268.

ELLENBERG, J. H. (1983). Ethical guidelines for statistical practice: a historical prospective. *Amer. Statist.* **37** 1–4.

HAND, D. J. and EVERITT, B. S. (1987). *The Statistical Consultant in Action*. Cambridge Univ. Press.

HUNTER, W. G. (1981). The practice of statistics: the real world is an idea whose time has come. *Amer. Statist.* **35** 72–76.

HYAMS, L. (1971). The practical psychology of biostatistical consultation. *Biometrics* **27** 201–211.

*JOINER, B. L. (1961). Consulting, statistical. In *Encyclopaedia of Statistical Sciences* 147–155. Wiley, New York.

JOWELL, R. (Chairman) (1986). International Statistical Institute declaration on professional ethics. *International Statistical Review* **54** 227–242.

KIRK, R. E. (1991). Statistical consulting in a university: dealing with people and other challenges. *Amer. Statist.* **45** 28–34.

RUSTAGI, J. S. and WOLFE, D. A. (1982). *Teaching of Statistics and Statistical Consulting*. Academic Press, New York.

SLOAN, J. A. (1992). How to consult with a statistician. *The Statistical Consultant* **9** 3–4.

*WOODWARD, W. A. and SCHUCANY, W. R. (1977). Bibliography for statistical consulting. *Biometrics* **33** 564–565.

ZAHN, D. A. and ISENBERG, D. J. (1980). Non-statistical aspects of statistical consulting. In *Proceedings of the Statistical Education Section* 67–72. Amer. Statist. Assoc., Washington, DC.

Electronic Resources

A search of the World Wide Web as of October 1996 revealed a large number of sites relevant to consultants. The following are just a few of these; we have found them to be useful entry points, although they are by no means intended to cover the growing information resource available on the World Wide Web:

- The World-Wide Web Virtual Library: Statistics,
  http://www.stat.ufl.edu/vlib/statistics.html
- Statistics on the Web,
  http://www.execpc.com/~helberg/statistics.html
- Statistics Resources on the Web,
  http://www.stats.gla.ac.uk/cti/links_stats.html
- A Guide to Statistical Computing Resources on the Internet,
  http://asa.ugl.lib.umich.edu/chdocs/statistics/stat_guide_home.html
- An "Essential Book List," and useful if a year or two older than desirable,
  http://www.stat.wisc.edu/statistics/consult/statbook.html

A very useful and up-to-date document, the "List of Statistics Lists" is also available by sending the one-line message

send minitab list-of-lists

to

mailbase@mailbase.ac.uk

or by pointing your Web browser at

http://www.mailbase.ac.uk/lists-k-o/minitab/files/list-of-lists

This document contains details of all current statistics-related e-mail lists, including subscription information. These enable consultants to share information or conduct discussions on a timely basis.

Estimation of Antigen-Responsive T Cell Frequencies in PBMC from Human Subjects

Karl Broman, Terry Speed and Michael Tigges

Abstract. We describe a consulting project in which the statisticians found that they needed to develop a new method of analysis in order to reach valid conclusions about the effect of a herpes vaccine on the immune systems of human subjects. This method estimates the frequency of blood cells responding to a viral antigen at a single, carefully chosen dilution. The traditional analysis of such data uses a cutoff to separate experimental sites ("wells") which contain no responding cells from wells which contain at least one responding cell, whereas our method uses the scintillation count to estimate the actual number of responding cells for each well. We describe the experiment in some detail, as we found that one of the major challenges facing the statisticians was the need to understand the biology in enough detail to provide a relevant model.

Key words and phrases: Antigen-responsive T cell frequency, limiting dilution assay, EM algorithm.

1. INTRODUCTION

A vaccine to protect against the herpes simplex virus type 2 has been developed at Chiron Vaccines in Emeryville, California. In order to determine immunogenicity of this vaccine (its effect on the immune system) in human subjects, an assay which measures a subject's cellular immune response to viral antigens was developed, and a statistical analysis was required to estimate the magnitude of the immune response.

The process leading to the method described herein was long, arduous and very stimulating. The statisticians were asked to become involved in this project because, though an ad hoc procedure which more or less worked (in about 80% of cases) was available, a scientifically acceptable method of analysis was required.

Initially, our task was to make sense of the biological context of the problem and to understand the existing method of analysis, and we describe these in the next two sections. The existing method, developed by an intelligent nonstatistician and embodied in a nice computer program that accepted a number of different data types and gave results in a form desired by the user, consisted of procedures which were not analytically supportable, and ultimately we developed a rather different method that we describe in Section 4.

Karl Broman was a graduate student and Terry Speed is Professor, Department of Statistics, University of California, Berkeley, California 94720. Dr. Broman's current address is Center for Medical Genetics, Marshfield Medical Research Foundation, Marshfield, Wisconsin 54449. Michael Tigges is Senior Scientist, Chiron Corporation, Emeryville, California 94608.

2. THE ASSAY

Frequent discussions between the statisticians and the scientist and visits to the laboratory at Chiron, where we were able to observe the entire assay procedure and inspect the various instruments that are used, were crucial, as we began to gain an understanding of the science of the assay. The lab visits also helped us to judge the sources of error in the assay procedure.

Essentially, the assay procedure uses the fact that the human immune response involves the recognition of an antigen by T cells whose surfaces contain receptors which are complementary to part of that specific antigen. The T cells respond by emitting chemical signals (cytokines), stimulating other cells to replicate, or replicating themselves.

The assay under study seeks to estimate the frequency of T cells in a blood sample which respond to each of two test antigens, called gD2 and gB2. If inoculating a subject with the vaccine results in an increase in the frequency of respond-

ing T cells, this may reflect on the efficacy of the vaccine.

The assay is performed as follows: in each well of a 96-well (8×12) microtiter plate, diluted peripheral blood mononuclear cells (PBMC) are combined with growth medium, antigen and [^{3}H]thymidine (thymidine which has been radioactively labelled with tritium). Thymidine is required to form DNA. Replicating cells must first duplicate their DNA, and so they must take up thymidine. Providing tritiated thymidine allows us to measure the extent of cell replication: we extract the DNA from the cells and measure the amount of incorporated [^{3}H]thymidine using a scintillation counter. If the cells respond to the antigen, they will take up [^{3}H]thymidine, and a higher scintillation count will be obtained. For a complete description of the assay methods, see Broman, Speed and Tigges (1996a).

This type of assay is generally performed in one of two ways. A standard proliferation assay (e.g., James, 1991) compares the response from three wells containing antigen to three wells in which no antigen was added. The ratio of the average scintillation count for the antigen-containing wells to the average count for the antigen-free wells, called the *stimulation index*, gives a rough indication of the immune response to the antigen. For a more precise estimate of immune response, a full limiting dilution assay (LDA) may be performed, in which a number of wells are used at each of several dilutions of PBMC.

For our purposes, the stimulation index is too imprecise, and a full LDA requires far too many PBMC. Thus, we study a pair of microtiter plates at a single, well-chosen dilution.

In this design, in each of two 96-well plates, we have 24 wells which contain cells alone, 24 wells which contain the test antigen gD2, 24 wells which contain the test antigen gB2, 22 wells which contain tetanus toxoid (which serves as a positive control; nearly all subjects should show a strong response to tetanus), plus two wells containing phytohemagglutinin (PHA; a chemical which stimulates all T cells to replicate, and so serves as a second positive control). When analyzing these data, we seek estimates of the frequencies of responding cells in the four groups of wells: cells alone, gD2, gB2 and tetanus toxoid. The PHA wells help to select useable data.

Data for a pair of plates are displayed in Table 1. These data are for a single dilution of PBMC from a full LDA consisting of six dilutions of cells from one subject, collected in order to develop the assay.

3. THE TRADITIONAL METHOD OF ANALYSIS

The traditional method for analyzing such data is to classify each well as either positive (contains at least one responding cell) or negative (contains no responding cells). This is usually done (see, e.g., Langhorne and Fischer-Lindahl, 1981) by selecting a threshold, often the mean plus two or three standard deviations of the counts for the "cells alone" wells, and considering a well positive if its count exceeds this threshold. One then uses a statistical model, typically a Poisson model, relating the fre-

TABLE 1
Scintillation counts for a pair of plates from subject #713 at density 11,400 cells per well

	Cells alone			**gD2**			**gB2**			**Tetox**		**PHA**
A	179	249	460	2133	2528	2700	2171	1663	6200	761	9864	12842
B	346	1540	306	8299	1886	3245	1699	2042	3374	183	7748	10331
C	117	249	1568	1174	4293	979	1222	1536	2406	6497	2492	6188
D	184	414	308	2801	2438	1776	2193	3211	1936	2492	5134	927
E	797	233	461	1076	1527	2866	2205	2278	2215	3725	3706	4050
F	305	348	480	3475	902	3654	2046	1285	1187	9899	5891	3646
G	1090	159	89	1472	90	3639	657	2393	1814	3330	4174	2389
H	280	571	329	4448	3643	881	3462	2118	1013	8793	4313	672
	1	2	3	4	5	6	7	8	9	10	11	12
A	178	111	630	4699	5546	5182	3982	3104	2496	4275	2831	9727
B	244	593	259	5622	560	1073	1479	2978	4362	5017	5074	10706
C	261	964	167	2991	3390	3986	2321	2157	3278	8216	3579	3538
D	221	544	299	1838	4368	322	1022	1554	2980	2732	6177	5212
E	533	228	615	1938	4046	333	3253	5091	2843	200	1110	5063
F	818	98	160	1032	3269	4918	1778	3810	2372	6355	1869	2695
G	234	472	243	4143	3351	1118	530	1174	1881	3447	4491	2945
H	169	481	478	3237	1565	2211	2460	2715	4793	3029	6225	4679
	1	2	3	4	5	6	7	8	9	10	11	12

quency of positive wells to the frequency of responding cells in the wells. The frequency of responding cells per well is then estimated from the data on wells using maximum likelihood.

For the data shown in Table 1, the mean + 3 SD of the 48 cells alone wells (24 on each plate) is 1,401. In the cells alone group, 46 wells out of 48 are below this threshold value. In the gD2, gB2 and tetanus groups, 12/48, 8/48 and 6/44 wells, respectively, are below this value. The maximum likelihood estimate of the frequency of responding cells per well for a group of wells, assuming that the number of responding cells in a well follows a Poisson distribution, is obtained by taking the negative of the natural log of the proportion of wells below the cutoff. Thus, the MLE's of the frequencies of responding cells per well for the cells alone, gD2, gB2 and tetanus groups are 0.04, 1.39, 1.79 and 1.99, respectively.

The traditional approach clearly works well much of the time. However, for our application, determination of the threshold separating positive and negative wells was not at all straightforward: a comparison of the counts corresponding to wells for which no responding cells were expected with those for wells to which antigen was added revealed no clear cutoff in many cases.

Moreover, in many cases the entire set of wells for a given antigen would be positive. This arose whenever the density of cells chosen for the assay was a poor guess, something that could not always be avoided. In such cases the standard analysis of the data does not yield a point estimate of the frequency of responding cells, but only a lower confidence limit. This causes difficulties later when such results are to be compared or combined with other results.

Thus we were faced with a method that involved several rather mysterious cutoff procedures and an ad hoc treatment of cases in which no wells or all wells were above the cutoff (in which case the upper or lower binomial-based confidence limit was used as the "mean" estimate). In addition, the choice of cutoff was not accounted for in the estimated SE's, though it was clearly an important source of variation.

Recognizing these problems, most notably the difficulty in choosing a cutoff and the sensitivity of the results to that choice, the scientists sought the advice of outside statisticians.

4. A NEW METHOD

In our first approach to the problem, we attempted to optimize the choice of cutoff by minimizing the residual sum of squares. This required a model using the quantitative response in a well and eventually led to the model we now describe.

Figure 1 displays a plot of the average scintillation count against the cell density for each antigen group in a six-point LDA (of which the data in Table 1 is one dilution). It can be seen that the average scintillation count scales approximately linearly with the density of cells. This indicates that the scintillation count may contain information about the actual number of responding cells in a well, and not just about whether the well contains any responding cells.

This figure led us to make direct use of the scintillation counts in our analysis. As with the traditional approach, we assume that the number of responding cells in a well follows a Poisson distribution, but we add an extra relationship connecting the number of responding cells in a well to the scintillation count. Specifically, we suppose that there are plate-specific parameters, a, b and σ, and a widely applicable power parameter p such that the pth power of the scintillation count in a well containing k responding cells is approximately normally distributed with mean $a + bk$ and standard deviation σ. Since there are four groups of wells on each plate (cells alone, gD2, gB2 and tetanus), there are 10 parameters in all: the four frequencies and three additional parameters for each plate.

Algebraically, our assumptions are as follows. Let y_{ijs} denote the transformed scintillation count for well j of class i on plate s, and let k_{ijs} denote the corresponding number of responding cells. Here $i = c, d, b, t$ corresponds to the cells alone, gD2, gB2 and tetanus toxoid classes, respectively. We assume that the (y_{ijs}, k_{ijs}) are mutually independent, that k_{ijs} follows a Poisson distribution with mean λ_i and that, given k_{ijs}, y_{ijs} follows a normal distribution with mean $a_s + b_s k_{ijs}$ and standard deviation σ_s. The aim of our analysis is to estimate the parameters λ_i.

The power parameter was selected by maximum likelihood (Box and Cox, 1964) from the values 1, 1/2, 1/4 and 0 (corresponding to log). The model itself was fitted by the method of maximum likelihood, specifically, using a form of the EM algorithm (Dempster, Laird and Rubin, 1977), although we also carried out a number of confirmatory analyses using the fully calculated likelihood. Standard errors for the parameter estimates were computed using the SEM algorithm (Meng and Rubin, 1991). A more detailed description of the statistical methods can be found in Broman, Speed and Tigges (1996b).

Table 2 displays the results of our analysis of the data in Table 1. We used the square root of the

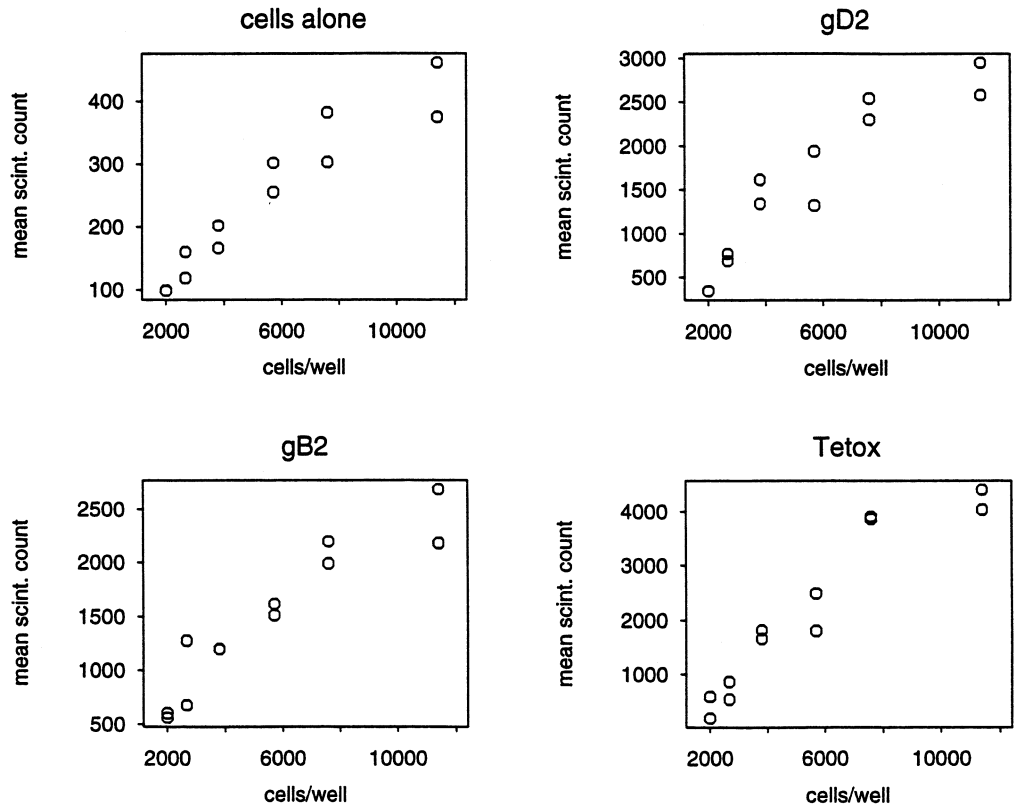

FIG. 1. *Mean scintillation counts in relation to cell density for the six-point LDA from subject #713.*

scintillation counts, as indicated by a Box–Cox analysis. Estimates were obtained by treating each plate separately, and also for the joint analysis of the pair of plates, where the frequencies of responding cells per well were constrained to be equal on the two plates.

Figure 2 displays the estimated frequencies of responding cells per well against cell density for the six-point LDA, of which the data in Table 1 is a single dilution.

5. CONCLUSIONS

The aim of the single dilution assay under study was to be a more sensitive version of the standard proliferation assay, which obtains a stimulation index as described above, and not a replacement for the standard dilution assay.

Our method for estimating the frequency of responding cells in a sample avoids many of the problems associated with the traditional analysis, which reduces the well counts to quantal responses using a cutoff. Assuming that the model we use is appropriate, our analysis will also be more efficient.

By considering Table 2, and in particular the values of λ_c, λ_d, λ_b and λ_t, we see that the new method does give results that confirm the original results from the old method and indeed strengthens them: the two antigens give mean frequencies of responding cells of around 3–4, almost as high as

TABLE 2
Maximum likelihood estimates and estimated standard deviations of model parameters for the data in Table 1

	λ_c	λ_d	λ_b	λ_t	a	b	σ
Joint:							
plate 1	0.4 (0.1)	3.5 (0.3)	3.3 (0.3)	4.7 (0.3)	16.4 (0.9)	10.3 (0.3)	3.6 (0.5)
plate 2	0.4 (0.1)	3.5 (0.3)	3.3 (0.3)	4.7 (0.3)	14.8 (0.8)	9.4 (0.2)	2.9 (0.4)
Separate:							
plate 1	0.3 (0.1)	3.0 (0.4)	2.8 (0.4)	4.4 (0.5)	16.7 (0.9)	10.3 (0.3)	3.5 (0.4)
plate 2	0.5 (0.1)	3.9 (0.4)	3.9 (0.4)	5.0 (0.5)	14.5 (0.7)	9.3 (0.2)	2.8 (0.3)

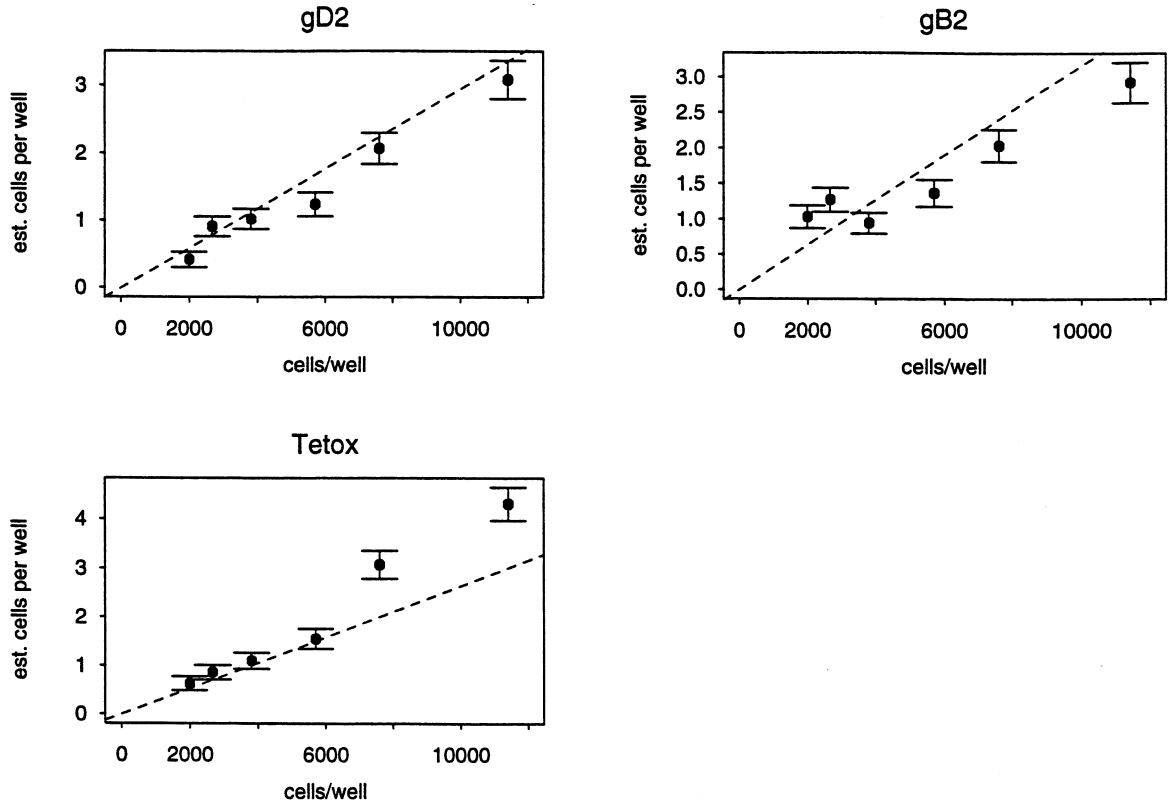

FIG. 2. *Six-point LDA for subject #713. Maximum likelihood estimates of frequencies of responding cells per well using two plates at each dilution: estimates plotted against number of cells per well. Error bars correspond to plus or minus one SD. Dotted line corresponds to estimated frequency of responding cells per well obtained using all of the data.*

for the tetanus groups and significantly above that for the cells alone group. Note that the old method gives results which are noticeably lower than those given by the new method.

In this project, by coming to an understanding of the underlying biology and developing a model which described more aspects of the data, we were able to provide the client with a valid and in fact more informative method of analysis. Its success can be judged by the fact that recently we have been involved in the analysis of a number of further clinical trials of Chiron's herpes vaccine. These trials are more complicated in that multiple subjects receive placebo or various doses of the vaccine and are assayed at multiple time points. Our analysis method is used to estimate the immune response for each subject at each time. These estimated responses are then combined, in order to estimate the effect of the vaccine, in increasing immune response to the viral antigens. This again involved a considerable amount of work, as we were not able to use off-the-shelf methods.

The chief advantage in using our new method, in the large clinical trials described above, is that it gives estimated frequencies of responding cells for all sets of data, whereas the old method suffered from the problem that if all wells in an antigen group were above the cutoff, one could not obtain an estimate of the frequency of responding cells, but only a lower confidence limit, making the combination of data from several assays very difficult.

A computer program incorporating our approach has been developed and is now available for general use.

Sandbars in the Colorado River: An Environmental Consulting Project

Jennifer A. Hoeting

Abstract. The National Park Service funded a study to determine the impact of water released from the Glen Canyon Dam on sandbars downriver through Grand Canyon National Park. The project involved considerable amounts of messy and missing data. Some of the challenges faced and lessons learned during this project are described.

Key words and phrases: Sampling interval, environmental monitoring.

1. WHY SANDBARS?

In 1990 the National Park Service (NPS) funded a project to measure sandbars in the Colorado River. The goal was to investigate the impact of water released from the Glen Canyon Dam on sandbars downriver from the dam (Figure 1).

When Glen Canyon Dam began operation in 1966, the annual flood cycle was eliminated as the dam controlled all water flow. Floods scour the river bottom, bringing up sediment deposited there. When flood waters recede, the sediment is left on the shore of the river in the form of sandbars. Surveys of the river show that sandbars have decreased in size and number since the dam opened in 1966 (Kearsley, Schmidt and Warren, 1994).

Measuring sandbar sizes may sound like another government boondoggle, but sandbars play a key role in the ecosystem of the Colorado River. For birds and insects, the sandbars offer a small strip of riparian habitat in a harsh desert environment. The sandbars also create eddies where endangered fish and other fauna feed. Finally, rafters camp on the sandbars during their trips down the Colorado River. Not only do fewer sandbars mean reduced habitat for fish and other wildlife, but reduced numbers of sandbars force all campers to use the same sandbars, thereby increasing the user impact on a fragile environment. For these reasons, the NPS wanted to investigate how patterns of water released from Glen Canyon Dam influence sandbar size.

In this paper we provide some insights on the scientific and statistical issues related to this project.

Jennifer Hoeting is Assistant Professor of Statistics, Department of Statistics, Colorado State University, Fort Collins, Colorado 80523 (e-mail: jah@stat.colostate.edu, http://www.stat.colostate.edu/~jah).

2. THE DATA

From September 1990 to July 1991, 17 helicopter flights were made above 230 miles of the Colorado River below the Glen Canyon Dam. On each flight the same 58 out of the total population of about 600 sandbars along the river were photographed. Each photograph was digitized to determine the size of the sandbar (Cluer, 1995b). The helicopter flights occurred during periods when the water was released at a constant level from the dam. Between flights, water was released from the dam in different patterns of discharge, called test flows.

The original study design called for flights every 15 days, which would result in a series of equally spaced observations over time. However, weather conditions and other difficulties resulted in a variable number of days between flights. On average, there were 20 days between flights, but flight intervals ranged from 12 to 70 days.

The original study design also specified that each of the 58 sandbars was to be photographed on every flight. Out of this sample of 58 sandbars, an average of 18 and a maximum of 40 sandbars were missed per flight. The data were missing for various reasons, primarily due to blurry photographs.

From the sandbar photographs, four numbers were recorded for each sandbar: gross area; area of erosion since the previous flight; area of deposition since the previous flight; and net change in size, where net change is the difference between sandbar size for the current flight and sandbar size for the previous flight.

In addition to sandbar size measurements, sandbar characteristics and hydrological data were recorded. The individual sandbar characteristics that were recorded included location in terms of miles from the dam, left or right river bank and type of sandbar. Nine hydrological measurements were used to characterize the test flows, including

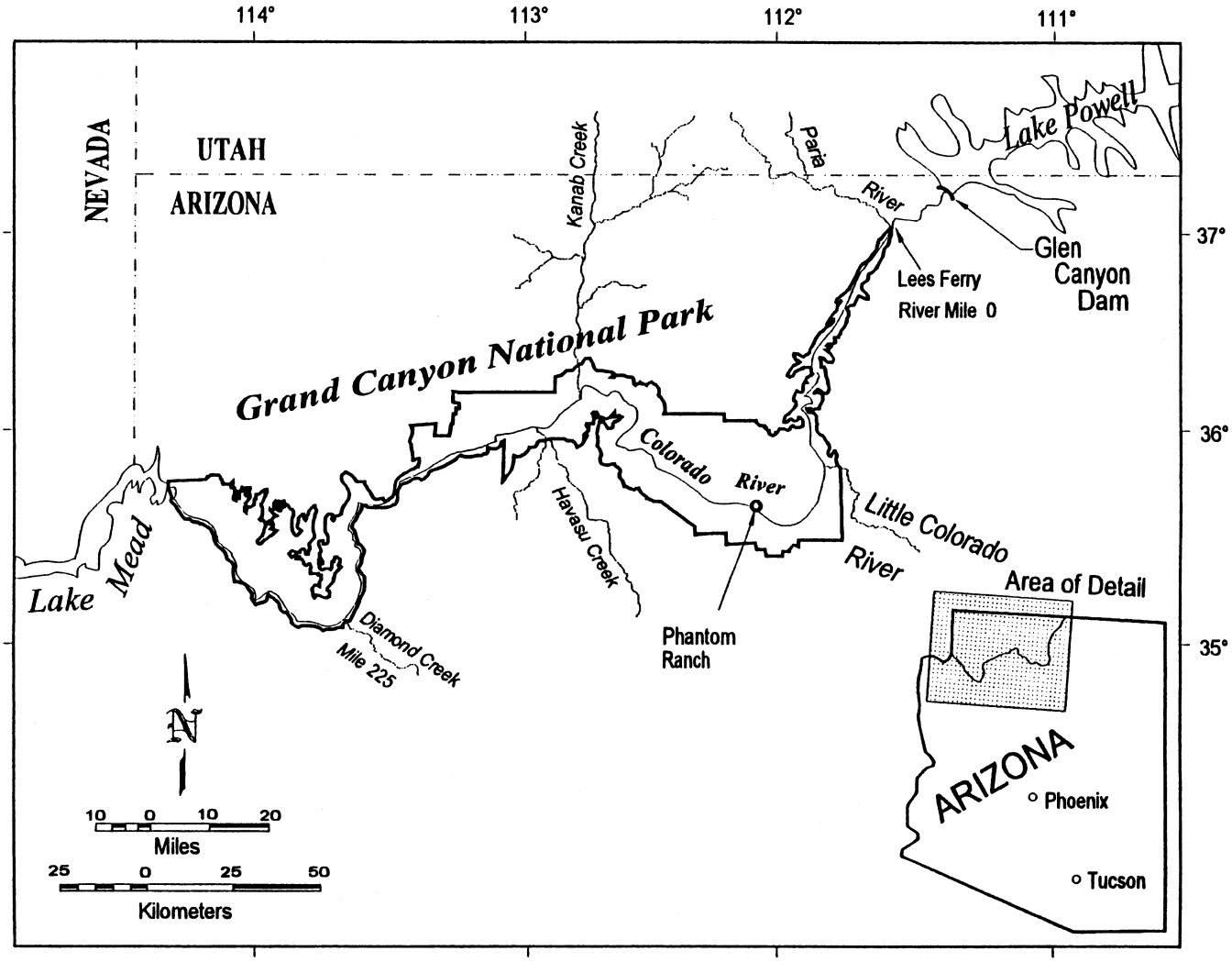

FIG. 1. *Geography of the Colorado River in the Grand Canyon. Map created by Brian Cluer, National Park Service.*

means and standard deviations of daily discharge over the flight period. "Upramp" (the increase in the level of the river at a specified point along the river) was available as mean daily maximum upramp, the average of the maximum rise in river level per day at five different gauging stations along the river. The average amount of sediment per day entering from the Little Colorado River during each inter-flight period (sediment supply) was also measured, as sediment from this tributary of the Colorado River could impact sandbar size (Figure 1).

3. PREDICTING SANDBAR SIZES: CHALLENGES IN CONSULTING

Although these data had been previously analyzed by the NPS, the NPS was interested in whether we could extend their findings using statistical models. Thus, we became involved in the project only after all the data had been collected. We addressed several important questions in this project.

3.1 Does Sediment Supply from Tributaries Influence the Sandbars?

To address this question we presented an autoregressive model to predict net change per flight averaged over all sandbars below the Little Colorado River. The auto-regressive model is of the form

$$Y = X\beta + u$$

where $u = \rho W u + \varepsilon$ and $\varepsilon \sim N(0, \sigma^2)$ (Upton and Fingleton, 1985). In this model, Y is the net change per flight averaged over all sandbars below the Little Colorado River and X is the matrix of predictors with the elements in the first column equal to 1. The parameter ρ can be interpreted as a measure of dependence between observations of the response. The weights matrix W is described below.

The auto-regressive model is sometimes called a spatial error model. The auto-regressive model takes into account previous observations of the response as well as previous observations of the predictors to improve predictions about the response for the current flight. One way to interpret this model is that it takes time for the predictors to impact the size of the sandbars. For example, results might indicate that the mean daily discharge from the previous flight is related to the response from today's flight.

In general, $W = [\omega_{i,j}]$ where $\omega_{i,j}$ is a nonnegative weight, which is representative of the "degree of possible interaction" between observation i and j and $\omega_{ii} = 0$. In this application, W was used to account for the variable number of days between flights. For example, for a lag 3 model

$$\omega_{i,j} = \begin{cases} 1/(\#\text{ of days between flight } i \text{ and } j), \\ \qquad \text{if } 0 < i - j \le 3, \\ 0, \qquad \text{otherwise.} \end{cases}$$

We considered several different lags (the number of previous flights in the weights matrix) and typically observed significant lag coefficients, but with so few observations we were reluctant to make definitive conclusions with respect to the lag component.

The results from the auto-regressive model indicated that mean daily water discharge from the dam and presence or absence of sand added to the river from the Little Colorado River were the most important predictors of net change. The estimated coefficients in these models were highly variable because the lag component was estimated using only 15 observations (2 of the 17 flights had too few observations to be included in these analyses).

Exploratory plots as well as results from the auto-regressive model indicated that there is a relationship between sand supply and changes in sandbar size in the Colorado River. One important finding was that increased sediment supply from the Little Colorado River appears to take longer than one flight period to impact sandbars. Future studies of sandbar dynamics should collect data on sediment supply from important tributaries and investigate a possible lag between sediment input and changes in sandbar size.

3.2 Can Dam Release Characteristics Predict Changes in Sandbar Size?

To answer this question we considered a standard regression model to predict net change per flight averaged over all sandbars included in the study. The regression results indicated that, as mean daily discharge increased and upramp remained fixed, the sandbars tended to increase in size on average. As mean daily maximum upramp increased and mean daily discharge remained fixed, the sandbars tended to decrease in size on average. The other dam release characteristics were not significant predictors of change in sandbar size.

The results from both the auto-regressive model and regression model are somewhat suspect due to the small number of observations used to model a complex system. We described several limitations of these results in our final report to the NPS (Hoeting, Varga and Cluer, 1997). One concern is that both the response and predictors were averages, which makes interpretation of the models difficult. For example, the *mean* daily discharge may not be a good measure of the water release pattern, because two very different water release patterns could have the same mean daily discharge.

3.2 How Do Dam Release Patterns Impact Individual Sandbars?

Another goal of the project was to produce a space/time model to predict sandbar size for each sandbar based on sandbar characteristics and dam release measurements. The model was intended to provide scientists with some guidelines on how different patterns of water released from the dam impact different types of sandbars. Our analyses indicated that the large amount of missing data and, more importantly, the long time intervals between observations made this goal unattainable.

Recent data show that large time intervals between sandbar measurements can lead to erroneous conclusions. For example, it is common for large-scale rapid erosion events to occur in a matter of days or even over several hours. Figure 2 compares daily observations taken via automatic camera to results from 10 samples collected via a more traditional terrestrial survey for one sandbar in the Colorado River in 1991. (Note: a different representation of these data appears in Cluer, 1995a.)

Three substantial errors would be made if inferences were based on the 10 terrestrial survey data points. In case A daily observations show a gradual increase in area from February 2 until April 16 when the sandbar decreased from 130% to 70% of its original area over a 24-hour period. The two observations collected via terrestrial survey on February 2 and April 21 would, on the other hand, simply show a negative trend in sandbar area over the 70-day period. In case B the observations collected at 30-day intervals would completely miss a substantial erosion event and severely underestimate the variation in sandbar size. In case C the intermittent observations would both overestimate

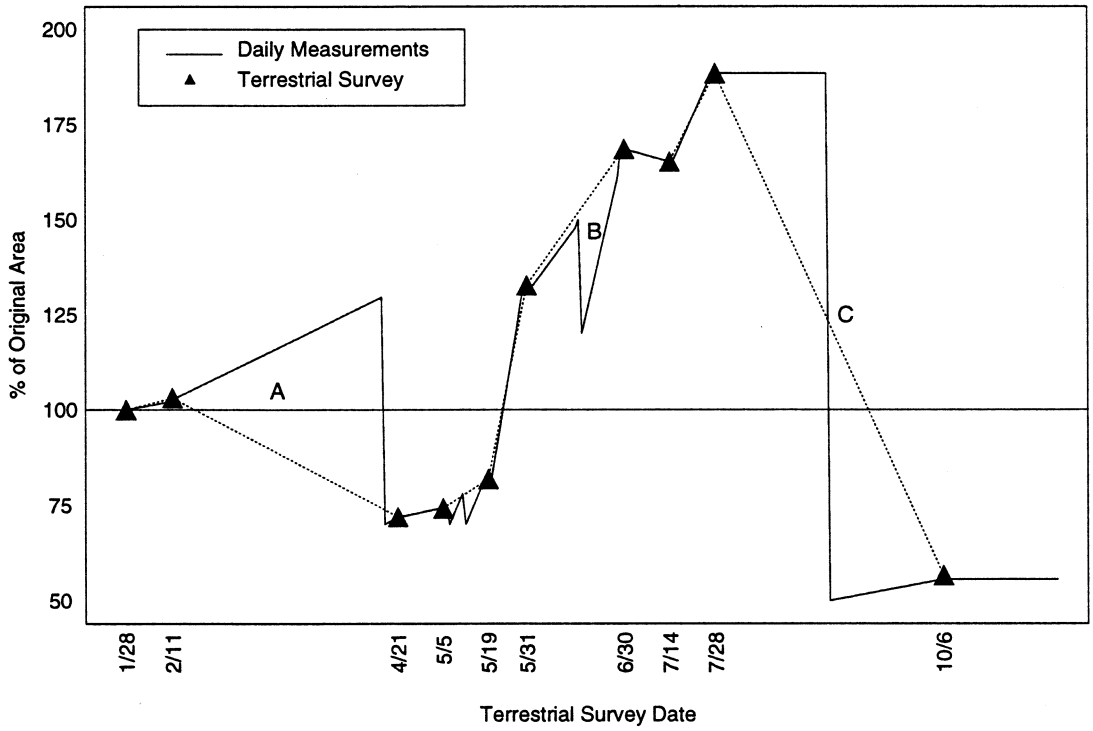

FIG. 2. 1991 *Colorado River sandbar survey, sandbar* 172. *Comparison of* 288 *daily measurements collected via automatic camera* (*solid line*) *to* 10 *intermittent measurements collected via land surveys* (*triangle symbols and dotted line*).

the time over which the erosion occurred and miss out on a portion of the erosion. These data show the danger inherent in basing inferences on sandbar data collected at sparse intervals over time. Since data analyzed in this paper were collected at intervals from 12 to 70 days, we have a very incomplete picture of what actually happened to the sandbars.

Another challenge is the large amount of missing data. With up to 40 sandbars out of the original 58 sandbars to be photographed missing for each flight, the missing data were an important concern. While we considered using data interpolation methods or likelihood-based approaches for the analysis of missing data, the high degree of uncertainty about sandbar behavior in the intervals between observations made it inappropriate to use these methods.

3.4 Suggestions for Future Studies

This is the best data set ever obtained for a sample of Grand Canyon sandbars; indeed, a large sample of sandbars was monitored over a long period of time as compared to previous studies of sandbar size. Since the data were collected via aerial photography from an airplane, it was cheaper to collect more sandbars per flight but to have fewer flights. In designing these types of studies, one must consider this tradeoff between the number of sandbars included in the study and the number of observations obtained for each sandbar. In this study there were 58 sandbars, but, with as few as 9 observations per sandbar collected over a long period of time, it was difficult to produce a credible model for individual sandbars. Our results indicate that future studies should focus on obtaining more observations of fewer sandbars which will allow scientists to understand the relationship between hydrological characteristics and changes in sandbar size more fully.

Related to this is the issue of sampling interval. In our final report to the NPS, we argued that not only will more frequent sampling result in better understanding of the underlying natural processes, but more frequent sampling of fewer sandbars can save money. Traditional sampling techniques use either aerial photography or land-based surveying. Flying at low altitude deep in the Grand Canyon is expensive, dangerous and may be ecologically unsound. Land surveying is similarly expensive and time-consuming, so it is best to budget for few flights or few surveys where many sandbars are measured. A better design would be to set up automatic cameras at a few sandbars to take photographs at specified intervals. Our analyses showed that, despite the reduction in the number of observed sandbars, more information about the questions of interest would be gained through our sug-

gested design. A formal cost model would be a good way to present these tradeoffs.

4. SOME LESSONS LEARNED

This project reinforces several basic rules for statistical consultants.

First, always check the data for errors at the start of the project. These data had been analyzed previously and thus we assumed that the database was error free. In fact, there were some serious problems still remaining. One of the most important was errors in the computation of net change, the response of main interest to our clients. We also discovered other errors, for example, in computation of the number of days between flights. Our experience on this project emphasizes the need for simple checks of data accuracy before beginning any analyses.

This project also demonstrates why statistical consultants should make every effort to obtain the raw data, if available. The original goal of the study was to relate the change in sandbar size to characteristics of the test flows for each flight, but only the summary statistics of the test flows were made available to us. While statistics such as mean and standard deviation of daily discharge over the flight period numerically characterize the test flows, the raw measurements would have provided us with further insight into the nature of each test flow.

We were also unable to obtain the raw values for daily maximum upramp from each of the five gauging stations along the river. Since we received only summary statistics averaged over the five stations and averaged over the flight period, it was impossible to relate upramp to the distance of each station from the dam, which is important because upramp increases with distance from the dam. Without doubt, increased access to raw data would have improved our ability to draw useful scientific inferences.

Finally, as consultants we must guard ourselves against standing on the "statistician's pedestal" from which we lecture scientists on the limitations of their studies. It is easy for a statistician to criticize a study after the data have been collected. We should recognize that, just as statisticians make compromises while doing analyses, investigators are under considerable constraints when designing their studies, including financial, time management and political constraints. Even with the best intentions in study design, we recognize that collecting high quality, complete data outside of a controlled laboratory environment can be an extremely difficult endeavor.

5. CONCLUSIONS

The NPS gained a considerable amount of useful information from our efforts. We provided insight into the relationship between net change in sandbar size and mean daily discharge, upramp, presence or absence of sand added to the river from the Little Colorado River and improved methods for collecting and evaluating data on sandbar sizes. Previous statistical analyses of sandbar size data have been limited, being based on simple models that ignore spatial and time correlations (Beus and Avery, 1992; Cluer, 1995b). Thus our study was a step in the right direction toward the collection of additional data and the development of useful models to predict sandbar sizes. Despite data limitations that prevented the use of highly sophisticated space–time models, we were able to identify the need for such models and the type of data and analyses that would be most useful for future studies.

The U.S. Bureau of Reclamation (USBOR), which operates the dam, faces the continual challenge of balancing the needs of the ecosystem with the needs of the power companies. In the past, discharge of water through Glen Canyon Dam has been controlled to optimize peak load hydropower production. In the spring of 1996, USBOR released a large controlled flood intended to reinvigorate the sandbars on the Colorado River. Preliminary observations show that this goal was at least partially achieved (Wegner, 1996).

Our analyses, along with analyses of data from the controlled flood, will help scientists further understand the relationship between sandbar size and dam water releases. The need to understand the impact of dams on ecosystems is continually increasing: it is predicted that by the year 2000 over 60% of the world's rivers will be regulated (Gore and Petts, 1989). Statisticians can play a key role in this research by helping scientists design good studies and by continuing to develop methodology to assist in the analyses of these and similar data.

ACKNOWLEDGMENTS

This project was supported by the NPS. The author thanks Brian Cluer of the NPS, for initiating this project and for providing scientific insights throughout the project, and Kristina Varga, who assisted on the project while a graduate student at Colorado State University.

Setting up Computer-Assisted Personal Interviewing in the Australian Longitudinal Study of Ageing

Sue Taylor

Abstract. We discuss the role of a statistical consultant in a large-scale longitudinal study of an ageing population in South Australia. Particular emphasis is given to the way in which this collaboration enhanced the accumulation of relevant data by planning the collection and analysis months before the survey began. Computer-assisted personal interviewing (CAPI) was used, providing a method of obtaining survey data which avoided expensive data entry and editing. Additional benefits were apparent in the facilitation of data management and analysis, and CAPI was also well received by both interviewers and respondents. The importance of enlisting the services of a statistician in the early stages of a large study is clear in this project.

Key words and phrases: CAPI, panel surveys, longitudinal data ageing studies, survey efficiencies.

1. INTRODUCTION

In this note, we describe one contribution of a statistician consulting on a major longitudinal study, the Australian Longitudinal Study of Ageing (ALSA). We first briefly introduce the study, then indicate the reasons for using computer-assisted personal interviewing (CAPI). This decision gives the statistician much more scope than usual to be useful. We then describe the implementation and advantages of the CAPI approach and give some descriptions of ALSA outcomes indicating the ways in which CAPI was successful.

ALSA is a multidimensional panel survey of the social, health, behavioral, economic and environmental characteristics of a random sample of people aged 70 years and over, who live in Adelaide, South Australia.

The ALSA project is the most comprehensive population-based study of ageing yet undertaken in Australia and is supported in part by the U.S. National Institute on Aging. It is a cross-national collaboration jointly undertaken by the Centre for Ageing Studies, Flinders University of South Australia and the Center for Demographic Studies, Duke University, North Carolina.

The general aim of ALSA is to gain increased understanding of how social, biomedical, behavioral, economic and environmental factors are associated with age-related changes in the health and well-being of older persons. One of the more specific research aims is to determine the health levels and functional status of a representative older population and to track these characteristics over time.

The ALSA project consists of extensive face-to-face interviews of participants covering a wide range of topics and includes over 700 individual items. The study was conducted in four phases, or "waves," at approximately one-year intervals.

Table 1 shows a selection of the interview domains. As can be seen, there was a wide variety of inputs to the study questionnaire from many diverse areas of research interest and consequently from many different researchers. In the early stages of questionnaire design, we attempted to restrict the questions to those which would answer the specific goals of the study. This was particularly important in the context of the elderly as the length of time spent with the interviewer needed to be kept as short as possible.

Consulting a statistician is invaluable to the process of questionnaire design, as typically statisticians have been exposed to many data that have

Sue Taylor is with the Department of Preventive Medicine and Biometrics, University of Colorado Health Sciences Center, Denver, Colorado 80262 (*e-mail: taylor@stat.colostate.edu*).

TABLE 1
Interview domains

Demography	Sleep
Self-rated health	Morbid conditions
Medication use	Health services
Falls and injuries	Vision and hearing
Dental	Weight
Reproductive history	Finances and income
Significant life events	Smoking and alcohol
Occupation and education	Family/social contacts

been rendered useless due to poor planning, and are therefore sensitized to the issues involved before they occur. This turned out to be particularly the case in the ALSA study, where the statistician could also play an important role in the choice of survey instrument.

The decision to use CAPI presented a much greater opportunity than anticipated to integrate proposed analyses with the collection procedure.

2. DESCRIPTION AND CHOICE OF CAPI

CAPI was introduced in the 1980s as part of the revolution in the measurement of public opinion using computer technology. Not surprisingly, its popularity increased soon after the price, weight and quality of portable computers began to improve significantly.

In CAPI, the interviewer takes the computer, preloaded with the questionnaire, to the respondents' homes. There, the interviewer reads the questions as they appear on the screen and then types the respondents' answers immediately into the computer.

The interview conducted in this way is little different from the traditional paper-and-pencil interview, as far as the respondent is concerned. However, there are differences between the traditional methods and CAPI with varying levels of relevance to overall study quality. Empirical evidence does not support the suggestion (Couper and Groves, 1989) that the mere presence of the computer may affect the outcome, at least with respect to refusals or partial nonresponse, even for sensitive issues. Major documented differences are that CAPI interviews take longer, and that the notes made by interviewers on the computer are shorter (Couper and Groves, 1989).

In setting up ALSA, a pilot study was conducted using conventional paper-and-pencil survey methods, but an early decision was made to use CAPI in the main survey. In the United States, most national surveys are conducted with the use of CAPI (Saris, 1991), but this is the first time such a method has been used in a survey of the elderly in Australia. The decision was based partly on the works of Groves and Mathiowetz (1984) and Harlow (1985), who both found that, in telephone surveys, computer-assisted data collection using computer-assisted telephone interviewing (CATI) was less expensive and yielded better data more quickly than traditional techniques. There was also some preliminary evidence that CAPI was likely to show even greater improvements in quality (Birkett, 1988).

The anticipated advantages of the CAPI method were similar to those that have been demonstrated previously for the CATI technique (Nicholls, 1988). These include the following:

1. the integration of several survey steps into a single activity that includes editing, coding, data entry, checking and cleaning;
2. immediate detection and resolution of errors made by interviewers and respondents;
3. reduced costs;
4. the capability to design more complex questionnaire instruments and skip patterns.

Expensive data entry and editing can be avoided, so the overall cost when compared with conventional paper questionnaire use is lower. Additional benefits are those associated with the facilitation of data management and analysis through an interface provision with statistical software, such as SPSS. Development of a data file, complete with variable information needed for immediate analysis, could typically take up to a week without this facility. In an ideal situation, this task is undertaken by the statistician who will analyze the data. Using CAPI, this task becomes part of the questionnaire construction, and generation of the data file occurs immediately after the last piece of data is entered.

In the context of a longitudinal study, this technique proves invaluable in that data from previous waves can be recalled and utilized in the subsequent interview. As an example, it is a useful way of avoiding redundant questions such as "How many natural teeth do you have left?" when it has been established previously that the respondent has none! Also it allows tailored questioning such as "Two years ago you told us that you had shingles. Do you still suffer from this condition?"

3. THE ALSA SAMPLE

The main study sample was randomly selected from within the Adelaide Statistical Division (ASD), which is essentially the greater metropolitan area of Adelaide. The sample, stratified by age and sex in five-year age groups to 85 years and older, was

TABLE 2
Population characteristics of Australia and South Australia

	Australia	South Australia
Total population (millions of people)	17.7	1.5
% living in urban areas	85.3	73.0
Median age	32.7	33.9
% aged 65 or older	11.7	13.4

obtained from the State Electoral Database. In Australia, compulsory voting ensures that this is a virtually complete listing of all adults 18 years and over. This gave the potential group of primary respondents. Spouses of this group (aged 65 and over) as well as other household members aged 70 years or older were also invited to participate. Table 2 gives some of the relevant characteristics of the Australian and South Australian populations.

The sample selection was carried out by the Australian Bureau of Statistics, using charts describing the structure of households where elderly people live. Since males in these age groups have a higher mortality rate than females over a five-year period, they were deliberately oversampled to allow sufficient numbers for longitudinal tracking. One of the unique features of the sampling plan was the provision for interviewing elderly couples and 565 couples were recruited to the study at baseline, increasing the efficiency of sampling resources.

The breakdown of the baseline sample is given in Table 3.

The study consisted of four waves, with Wave 1 being the baseline survey. Waves 1 and 3 used CAPI to interview eligible persons in their normal place of residence, and Waves 2 and 4 were short telephone interviews (approximately 15 minutes). A separate proxy instrument was developed and used very successfully for those respondents too ill or frail to respond personally.

Table 4 shows the response rates for Waves 2 and 3 (results of Wave 4 are not available at this time). The baseline data were collected from a total of 2,087 respondents (see Table 3); 1,477 of these were primary respondents, 597 were spouses or secondary respondents and 13 were other household members who agreed to take part.

TABLE 3
Participants in the baseline sample by age and sex

Age	Males	%	Females	%
65–69	17	1.6	123	11.9
70–74	279	26.4	283	27.4
75–79	283	26.8	241	23.3
80–84	235	22.3	194	18.8
85+	242	22.9	190	18.4
Total	**1,056**		**1,031**	

TABLE 4
Response rates for Waves 2 and 3

	Wave 2		Wave 3	
	n	**%**	***n***	**%**
Deceased	112	5.4	241	11.5
Moved from ASD	13	0.6	33	1.6
Unable to contact	13	0.6	9	0.4
Refusal	170	8.1	125	6.0
Responders	1,779	85.2	1,679	80.5
Total	**2,087**		**2,087**	

Excellent retention of the cohort was achieved in Waves 2 and 3 with more than 90% of the surviving respondents remaining in the study. This was largely due to the dedication of the team involved in the study who worked hard to involve the participants and maintain their interest.

4. IMPLEMENTING CAPI IN ALSA

To support the CAPI approach in this survey, we chose an integrated package for survey management known as BLAISE, which was developed by the Netherlands Central Bureau of Statistics (Bethlehem et al., 1989). The programming language is essentially a modified version of the PASCAL language.

The BLAISE system makes provision for CAPI and includes questionnaire design and administration, checking, data editing, tabulation and analysis.

The generation of a BLAISE program proceeds in a number of defined steps:

1. The questionnaire is specified using a text editor. The routing specification is then applied to provide for skips and subquestionnaires. Range checks and consistency checks are then included and can be differentiated as "hard" or "soft" errors. Hard errors, specified in terms of relational expressions, must be satisfied before the response is accepted as valid, and require correction before an interview can proceed. On the other hand, soft errors result in a warning message that can be overridden by the interviewer.
2. Raw BLAISE statements are turned into an executable program which is copied to the interviewer's laptop computer ready for use.
3. The data are collected into files which may then be converted into standard ASCII files or other formats, including SPSS system files with accompanying syntax files to describe the data.

The use of CAPI made it theoretically viable to start data analysis as soon as the last piece of data was collected. Essentially no data cleaning was necessary and all relevant data transformation programs could be written beforehand and executed immediately. Of course, this all relies on the fact that the questionnaire items are correctly specified in the initial stages and require little manipulation after collation. One of the critical benefits of early involvement of a statistician is in ensuring that this specification is indeed accurate.

There were a number of specific tools which were found to be of great value. These included:

1. electronic notepad to allow notes written by the interviewer during the course of the interview to be stored on a separate but related file;
2. interrupt facility to allow the storage of incomplete interviews and subsequent return to the incomplete record—this was particularly important in the context of the elderly and especially since (despite our best efforts to shorten interview times) the average length of interview was 132 minutes(!);
3. a "ditto" facility which copied the response from the previous questionnaire for the corresponding question; this was very useful as we were interviewing elderly couples in the same household and responses to certain questions were often identical;
4. a built-in clock facility, which allowed timing of the interview length for later analysis;
5. provision for the generation of a paper version of the questionnaire.

Data were returned on a weekly basis on floppy disks, although it is also possible to download data to a central computer with the use of modems. A program was written to backup these data so that at least two copies of all data were available at all times.

5. WHY THE STATISTICIAN WAS VALUABLE

Automatic coding of categorical variables is done by the CAPI program, rather than by the interviewers or respondents. In order for this facility to be fully utilized, it proved invaluable to have the statistician devise the coding scheme in the planning stages to allow for later statistical analysis. This was especially important since techniques such as regression were intended, and the statistical packages used required the data in a specific, nonintuitive format.

In this study, the longitudinal nature of the data made it absolutely necessary for an experienced statistician to be involved who had worked previously with data of this type. Specification of the correct data structure was an integral part of solving the data storage problem alone, as there were many megabytes of data generated by the design we used. Apart from this potential difficulty, it was also imperative to consider the inevitable missing data problems and how best to flag these with respect to future analyses. The statistician is able to design possible imputation procedures, if appropriate, or at the very least ensure that the occurrence of missing data does not render that entire record useless.

The inclusion of a statistician on the research team also proved invaluable during the "covariate brainstorming" sessions. Not only do statisticians realize the importance of collecting data on all potential covariates, but they also know the type of data to collect. We managed to persuade the collaborators to elicit covariate data as continuous measurements wherever possible, with subsequent categorization, rather than risk the loss of information by collapsing variables at the collection stage. Unlike paper questionnaires, this would result in a total loss of potentially useful data, since the computer may only store one realization of a variable, with the other lost forever.

Overall, statistical thinking at the outset helped shape the whole procedure, and it was rewarding for both the statistician and the researchers to begin the collaboration at a point where the impact was greatest.

6. CONCLUSIONS

In ALSA, it was notable that CAPI was well received by interviewers and respondents alike, and the high response rates in Table 4 reflect this. Some interviewers, initially hesitant about the use of computers in the interview process, became positively excited about their use and many of the older people in the survey showed a genuine interest in the technology.

In this study, the time spent in the initial stages was definitely worthwhile, especially that spent on defining range and consistency checks. Range checks are of course only the first step in the data cleaning process, as they simply allow a broad check on the data value for one particular variable. Although they were not completely infallible in the multivariate sense, and we still saw a few incongruous combinations of data values, we found that in addition to the consistency checks, a large percentage of the errors which typically occur in data of these types were eliminated in the ALSA implementation.

One of the most important features of CAPI is the validation facility, a task which an interviewer typically does not have the time to carry out during the interview. Data can be checked against other information to assess its quality, a critical feature particularly in a panel survey. If there are discrepancies, the respondent can be asked for immediate clarification, so that the data are cleaned while the respondent is still available. This is a clear advantage of CAPI since traditional methods may force the incorrect answer to a missing value. In general, if the questionnaire is well constructed at the outset, data can be considered optimal and later validation omitted.

In our experience, CAPI proved to be an efficient, accurate, cost-effective and acceptable method for collecting data from older people in a community survey. Data analysis was also made less of a chore by the absence of the many hours which would normally be spent on tedious and time-consuming data cleaning.

The added features of the system described in the previous section enhanced acceptability markedly. The future in this area will certainly see enhancements in the software, and hardware improvements including acceptance of voice and handwritten entry, which will make this mode of data collection an even more attractive option.

Data collection in such a large study can be daunting. However, with a modest investment of time in the planning stages, jointly between the statistical consultant and the rest of the research team, the rewards can be great.

Queueing at the Tax Office

Richard Tweedie and Nell Hall

Abstract. This paper discusses a consulting project where, by focussing on the basic parameters of a probabilistic model, advice was given that could result in real improvement in the service at an Australian tax office, without raising the costs of the operation. The results are not intuitive and illustrate that nonlinear behavior in models can be hard for nonmathematicians to follow or even believe.

Key words and phrases: Waiting times, server numbers, M/M/c queues, delays, loss of customers.

1. THE PROBLEM

Applied probability problems are often of a scale that does not lend itself to "consulting." There are of course many outstanding examples of the use of applied probability techniques in major collaborative efforts, in, for example, teletraffic and networking, epidemiology, spatial pattern recognition and the like, and these often result in the type of collaboration that is commended in the advice given by the New Researchers Committee of the IMS (1991; hereafter CNR), but they rarely look like the sort of consulting that most clients arrive with, as noted in Tweedie (1986). That is, they are usually harder or deeper, and do not have the type of constraints, benefits and rewards that one might expect if coming from a nonacademic environment.

This paper describes an anomaly in this pattern: an applied probability problem that really is pure consulting, with no new methodological research, but with the rewards and problems of difficult data, of client interactions, of approximations to reality and of time and funding constraints, and with a happier than often ending, since valuable advice could actually be provided and implemented within the client's budget.

The problem is simple to describe and will strike a chord with all who have been put on hold in automated telephone enquiry lines everywhere.

In the mid-1980s, both of us were working in a medium-sized private sector consultancy, SIROMATH Pty Ltd, in Australia. We were consulted by

Richard Tweedie is with the Department of Statistics, Colorado State University, Fort Collins, Colorado 80523 (*e-mail: tweedie@stat.colostate.edu*). *Nell Hall is with the New South Wales Department of Health, North Sydney NSW 2060, Australia.*

TABLE 1
Effectiveness of the calling system

Measure of effectiveness	Site A	Site B	Site C
Average time on hold (sec) $\hat{w}$	140	200	753
Percentage $\widehat{p_3}$ waiting > 3 min	30%	45%	92%

an officer of the Australian Tax Office who was concerned that the behaviors of the "dial-in" enquiry lines at different offices were inexplicably different. The tax office had recently installed management software to track characteristics of their enquiry system and had looked at several measures of effectiveness, including the following:

1. the average length of time $\hat{w}$ waiting on hold before questions were addressed;
2. the percentage of customers $\hat{p}_3$ who waited longer than three minutes before being answered.

In particular, the client was concerned that one office seemed to have remarkably poor behavior compared with others. Table 1 captures most of the critical material: we have (for reasons of confidentiality) labelled three of the offices we studied as Sites A, B and C, with Site C clearly being the one in distress. One can only admire the patience of those calling to clarify their tax return status and queries: note that at Site C nearly every caller waited more than three minutes and the *average* call was on hold for over 12 minutes!

Although there were some other, well posed, questions about the possibility of networking the system, like many consultancies this one had a regrettably inexplicit main question: it was of the order of "what is going on here and can we do anything about it?" We will see that there are some options, at least, for using standard queueing theory to address this satisfactorily, but that such an application requires (as do so many consultancies) particular care in acquiring appropriate data before carrying out the analysis.

2. THE QUEUEING MODEL

In this project we had three skills to offer: the first was indeed the ability to advise on the types of models that might fit the situation, as discussed in CNR, but in contrast to CNR, the second was to help the client identify real data that might enable the model to be assessed and the third was to carry out the analysis for him since this was rather outside his capabilities.

As in all statistics applications, the modelling should come first, or we do not know what data will be relevant. In this case we turned to the simplest queueing model: a multiserver queue with c servers (the staff in the enquiry room of the tax office), the customers arriving in a Poisson process (a relatively standard assumption, and one which implies we needed to know only the average rate λ of calls per minute) and with the time to serve customers taken as i.i.d. random variables with exponential service times of mean length $1/\mu$.

This last was, as we will describe, a rather rough assumption in this case: but with it we are able to use simple known results to assess the type of effectiveness measures being proposed. [For a good if rather old-fashioned description of how this might work, see Lee (1968), who still has sound advice on how to live in a real world with the distributional assumptions involved.]

These exponential assumptions are not always appropriate. One of the truly beautiful results of queueing theory is, however, the "critical threshold" result that says that the system will, *regardless of such distributional assumptions*, have stability or instability properties according as the traffic intensity

$$\rho = \lambda/c\mu$$

is less than 1 or otherwise. In the stable situation the queue will not get too lengthy, and in the unstable situation it will grow beyond bounds. From Table 1, Sites A and B look stable and Site C is rather like an unstable situation, and so we first sought to see what the values of ρ in our system might be. For these we only need the mean interarrival and service times λ, μ and the number of servers c. The data we were given are in Table 2: as described in the next section, the system really is close to or above critical, especially when c is smaller than it is reputed to be, and this does help explain some of the longer waiting times observed.

We can then use the exponential distributional assumptions to enable us to consider the effective-

TABLE 2
Input data

Parameter	Site A	Site B	Site C
Input rate per minute (λ)	4.18	2.27	1.60
Mean call length (sec)	180	187	200
Wrap-up time (%)	10%	18%	40%
Mean service time including wrap-up ($1/\mu$)	198	221	280
Minimum number of servers $c_{\min}$	10	8	6
Average number of servers c_a	13	12	8
Maximum number of servers $c_{\max}$	15	14	9

ness parameters and predict their values, at least in the stable situation. We find, in particular, that the analytic forms are given by Lee (1968),

$$p_3 = p_0 \frac{(c\rho)^c}{c!} e^{3(\lambda - c\mu)},$$

$$w = \frac{(c\rho)^c}{c!} \frac{p_0}{c\mu(1-\rho)^2},$$

where the probability of an empty queue is

$$p_0 = \left[\sum_{r=0}^{c-1} \frac{(c\rho)^r}{r!} + \frac{(c\rho)^c}{c!} \frac{1}{(1-\rho)} \right]^{-1}.$$

In this case the key question was not to predict these (or other similar quantities) with great accuracy, but rather to decide why the input parameter combinations might be leading to the particular combinations that were being observed. Note again that only λ, μ and c are relevant to these results, and so we did not need more information than this on the system.

3. DATA COLLECTION

Our initial set of data was provided from the then-new computerized telephone system. Table 2 shows that the average rate of calls at the biggest of the sites was around 4–5 per minute: this appeared relatively stable over the day, with the exception of the first hour when, not surprisingly, the rate was usually closer to the maximum of 5 per minute. The rates at the other sites seemed acceptably constant over the whole of the day.

The system also provided average call lengths. These were relatively constant across all sites. Regrettably the system did not collect the actual distribution of calls: in principle this might have been possible but the resources to reconfigure the system for better data were not available from the client. Thus we were not able to verify if an exponential distribution was reasonable.

The most difficult information to collect was the number of servers. Internal staffing sheets showed the number of servers on an hourly basis, and these varied widely within the day. In particular the maximum number (which was possibly the number the client felt to be available) was actually 150% of the number often really working. Given the role of c in ρ or in the waiting times, this is of very considerable concern.

Note that the number of servers actually assigned to each site appears in principle to be in line with the observed input rate: indeed, if anything Sites B and C seem to have pro-rata more servers than they should have in comparison to Site A, if we judge by the input rate. This helps illustrate the client's understandable concerns about the poor behavior at Site C, since in principle the model says that this should be well under control.

Following the initial data collection, in this project we had the very real benefit of supplementing the paper data with one site visit, to Site A. There we learned rather more, as one so often does, and in particular we discovered two extra pieces of information:

1. On every call there was a "wrap-up" period after the call, when the server made notes, shifted files and so on. Once we learned of this, we found that data existed to estimate the extent of wrap-up activities in each office, and although these probably had at least some minimum length, we modelled them as a percentage of the service time and added them to the observed service time; this is used in Table 2 to give the μ we used, and then in Table 3 to give the predicted values of w and p_3. Note in particular that at Site C this wrap-up time adds much more to the actual call length than at the other sites.
2. There was also a period of "idle time" for each server, some of which was time absent from the room and which might of course be reflected in the server counts, but some of which was at the desk and in principle should be added to the service time; in Table 4 we take account of this.

These extra service times would not have been picked up without the detailed information collected on site. Some clients are insistent that statisticians visit the scene of their operations, and this can be time consuming for the statistician: others of course prefer to keep the statistician as far from the facts as possible! But whatever the attitude of the client, in order to give sensible advice within context, it is a step that one should always try to take, as this case study illustrates.

TABLE 3
Predicted behavior excluding idle time

Parameter	Site A	Site B	Site C
Minimum traffic rate $\rho_{\min} = \lambda/c_{\min}\mu$	1.38	1.04	1.24
Average traffic rate $\rho_a = \lambda/c_a\mu$	1.06	0.69	0.93
Maximum traffic rate $\rho_{\max} = \lambda/c_{\max}\mu$	0.92	0.59	0.82
Assumed c for prediction	15	9	8
Predicted (observed) mean wait W	110 (140)	262 (200)	423 (753)
Predicted (observed) p_3	22 (30)	45 (45)	57 (92)

TABLE 4
Predicted behavior including idle time

Parameter	Site A	Site B	Site C
Idle time (%)	8%	8%	14%
Mean service time including wrap-up and idle time $1/\mu^*$	213	238	319
Minimum traffic rate $\rho_{\min} = \lambda/c_{\min}\,\mu^*$	1.48	1.12	1.41
Average traffic rate $\rho_a = \lambda/c_a\,\mu^*$	1.14	0.75	1.06
Maximum traffic rate $\rho_{\max} = \lambda/c_{\max}\,\mu^*$	0.99	0.64	0.94
Assumed c for prediction	15	10	9
Predicted (observed) mean wait w	1259 (140)	162 (200)	537 (753)
Predicted (observed) p_3	90 (30)	31 (45)	63 (92)

4. RECOMMENDATIONS

Tables 3 and 4 show that the observed behavior of Sites A and B is reasonably consistent with the model predictions, especially if we assume Site A is using all servers effectively, and if (in contrast) Site B is using rather close to its minimum number of servers. Site A is also very close to critical ($\rho = 1$) even when the full complement of servers is present.

If we do not incorporate the wrap-up time in Site C then in fact that site is far from critical: even with the *minimum* observed of 6 servers, they still have $\rho = 0.89$ and a predicted mean waiting time of just 180 seconds. However, the poor behavior can be far better explained if we take into account the 40% increase to service time following the addition of the wrap-up time. Indeed, their behavior is even more consistent with the model where we also add in some percentage of the idle time as well.

In no cases were the fits of the data perfect for the model, of course. In particular, if we assume the value 5.4 minutes for the mean service time for Site C, then we get a value of around 750 seconds for the mean waiting time (consistent with reality), but we find that we only have $p_3 = 75\%$; conversely we get close to the observed value of $p_3 = 92\%$ by assuming a mean service time of just 5.57 minutes, but the expected waiting time is at a (noticeably theoretical!) 62 hours or so. This might perhaps be explained by a distribution of service times with a "lump" of probability near the origin, corresponding perhaps to part of the wrap-up times being of fairly fixed length, but we were not in a position to look further at this. Nonetheless, the general operation seemed well described by this simple model, and it was possible to give some rational advice based on it.

In Figure 1 we illustrate the prime recommendations we gave to the client:

1. At Site C the average predicted waiting time (w on Fig. 1) could be reduced to around 1.5 minutes (from the current 12.5 minutes) by adding just one extra server (to have an effective group

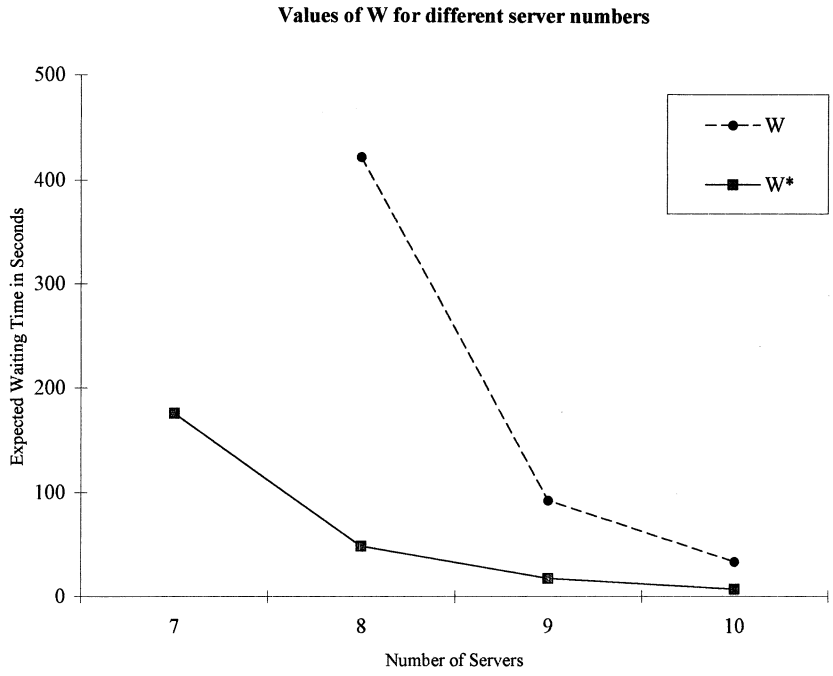

FIG. 1. *Changes in waiting time as number of servers varies at site C: W is predicted waiting time, W* is waiting time with reduced wrap-up time.*

of 9), and indeed w could be virtually halved merely by ensuring that all eight current servers were constantly available.

2. even more usefully, training should be instituted at Site C to reduce the percentage of time spent on wrap-up to at most 15% of the length of the call, as was achievable at all other sites; the resulting waiting time, given as w^* on Figure 1, is less than three minutes even with only seven servers. It is well under a minute if all eight are working. Considerable further reductions are achieved if wrap-up is only 10%, as at Site A.

One of the effects that the client found hardest to believe was that, as just illustrated, the whole system was so sensitive to very minor improvements in the parameters when close to critical. It is not intuitive that just one extra server, or, more dramatically, just saving some seconds in mean service times, could have such a powerful effect.

Various other recommendations were made, especially as the client was seriously investigating the possibility of routing calls between the sites, so the effective server pool would suddenly become around 35–50. We were able to predict that such an action would reduce the average waiting time to well under a minute and ensure no more than 10–15% of customers would be waiting for a 3-minute period: this would give far better service than at any other single site we discussed with the client.

Did this consultancy improve the service to the taxpayers? Sadly, I have no idea. And this is the last of the lessons in this article for the new consultant: do not always expect to make a great difference and be grateful if you get any level of recognition. This project led to no paper (except, a decade later, this one), even though it involved much time, so there was no reward in an academic sense; it potentially helped many people at almost no cost to the client, since it clearly identified simple management changes that would give the desired result; but as so often is the case, the client did not feel the statistical consultant was relevant to implementing these, and we heard no more of it.

So why bother with such consulting? For many reasons: first, and not to be overlooked, we were in this instance being paid to be professionals, and, like lawyers and doctors and other professionals, we should provide our skills and not necessarily expect to be further involved and not on center stage; second, statistics is designed to solve real problems, and here we did just that, and this can be its own reward; and third, the project *did* achieve one of the values noted in CNR, namely, it was fascinating and gave (at least to us) a real knowledge of yet another area in which statisticians can play a part that cannot be played by anyone else.

REFERENCES

BARLOW, R. E., BARTHOLOMEW, D. J., BREMNER, J. M. and BRUNK, H. D. (1972). *Statistical Inference under Order Restrictions*; *the Theory and Application of Isotonic Regression*. Wiley, New York.

BETHLEHEM, J. G., HUNDEPOOL, A J., SCHUERHOFF, M. H. and VERMEULEN, L. F. M. (1989). BLAISE 2.0/an introduction. CBS report, Netherlands Central Bureau of Statistics, Voorburg, The Netherlands.

BEUS, S. S. and AVERY, C. C. (1992). Final report: the influence of variable discharge on Colorado River sand bars below Glen Canyon Dam. Technical report, National Park Service.

BIRKETT, N. J. (1988). Computer-aided personal interviewing: a new technique for data collection in epidemiologic surveys. *American Journal of Epidemiology* **127** 684–690.

BOX, G. E. P. and COX, D. R. (1964). The analysis of transformations. *J. Roy. Statist. Soc. Ser. B* **26** 211–252.

BROMAN, K. W., SPEED, T. P. and TIGGES, M. (1996a). Estimation of antigen-responsive T cell frequencies in PBMC from human subjects. *J. Immunol. Methods* **198** 119–132.

BROMAN, K. W., SPEED, T. P. and TIGGES, M. (1996b). Estimation of antigen-responsive T cell frequencies in PBMC from human subjects. Technical Report 454, Dept. Statistics, Univ. California, Berkeley.

CLUER, B. (1995a). Cyclic fluvial processes and bias in environmental monitoring, Colorado River in Grand Canyon. *J. Geology* **103** 411–421.

CLUER, B. (1995b). Final report: aerial photography of the GCES-II test flows and interim flows. Technical report, National Park Service.

COUPER, M. and GROVES, R. (1989). Interviewer expectations regarding CAPI: results of laboratory tests II. Bureau of Labor Statistics, Washington, DC.

DEMPSTER, A., LAIRD, N. and RUBIN, D. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39** 1–38.

GORE, J. A. and PETTS, G. E. (1989). *Alternatives in Regulated River Management*. CRC Press, Boca Raton, FL.

GROVES, R. M. and MATHIOWETZ, N. A. (1984). Computer assisted telephone interviewing: effects on interviewers and respondents. *Public Opinion Quarterly* 356–369.

HARLOW, B. L. (1985). A comparison of computer-assisted and hard copy telephone interviewing. *American Journal of Epidemiology* **122** 335–340.

HOETING, J. A., VARGA, K. and CLUER, B. (1997). Predicting Colorado River sandbar size using Glen Canyon Dam release characteristics. Final report, National Park Service.

JAMES, S. P. (1991). Measurement of basic immunologic characteristics of human mononuclear cells. In *Current Protocols in Immunology* (J. E. Coligan, A. M. Kruisbeek, D. H. Margulies, E. M. Shevech and W. Stroger, eds.) Section 7.10. Green Publishing and Wiley–Interscience, New York.

KEARSLEY, L. H., SCHMIDT, J. C. and WARREN, K. D. (1994). Effects of Glen Canyon Dam on Colorado River sand deposits used as campsites in Grand Canyon National Park, USA. *Regulated Rivers*: *Research and Management* **9** 137–149.

LANGHORNE, J. and FISCHER-LINDAHL, K. (1981). Limiting dilution analysis of precursors of cytotoxic T lymphocytes. In *Immunological Methods* (I. Lefkovits and B. Pernis, eds.) **2** Section 12. Academic Press, New York.

LEE, A. M. (1968). *Applied Queueing Theory*. MacMillan, London.

MENG, X.-L. and RUBIN, D. B. (1991). Using EM to obtain asymptotic variance–covariance matrices: the SEM algorithm. *J. Amer. Statist. Assoc.* **86** 899–909.

NEW RESEARCHERS COMMITTEE OF THE IMS (CNR) (1991). Meeting the needs of new statistical researchers. *Statist. Sci.* **6** 163–174.

NICHOLLS, W. L., II. (1988). Computer-assisted telephone interviewing: a general introduction. In *Telephone Survey Methodology* (R. M. Groves, ed.) 377–387. Wiley, New York.

SARIS, W. E. (1991). Computer-assisted interviewing. Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-080. Sage: Newbury Park, CA.

TWEEDIE, R. L. (1986). In and out of applied probability in Australia. In *The Craft of Probabilistic Modelling* (J. M. Gani, ed.) 291–308. Springer, New York.

TWEEDIE, R. L. (1992). New researchers report: comments. *Statist. Sci.* **7** 263–264.

UPTON, G. and FINGLETON, B. (1985). *Spatial Data Analysis by Example* **1**. *Point Pattern and Quantitative Data*. Wiley, New York.

WEGNER, D. (1996). Personal communication (U.S. Bureau of Reclamation).

Comment

Michael W. Trosset

Richard Tweedie and his collaborators are to be congratulated for providing an interesting and enlightening forum on an exciting part of our profession—one that too many academic statisticians (are encouraged to) neglect. I will organize my contribution to this forum by elaborating on two of Tweedie's conclusions, with both of which I enthusiastically concur.

1. "It is often the mere fact of such thinking, rather than the specific technical input, that proves invaluable. It is hard to overestimate how powerfully our discipline trains us to think about complicated issues in ways that allow us to quickly diagnose difficulties in esoteric disciplines to which we have had only several minutes of introduction."

Statisticians who have not done much consulting may take for granted what I regard as the most important of the services that we offer. Of course we are sometimes asked to develop original methods for novel situations and of course we are often asked to ensure that standard methods are used appropriately in standard situations. Yet when I began consulting, I was struck by how often I provided a service without doing anything that an academic researcher would recognize as statistics. Time and again I was thanked (and paid) for asking questions and suggesting perspectives that seemed to me to be little more than common sense. This highly developed common sense is an easily overlooked, but extraordinarily valuable commodity.

Michael W. Trosset is Visiting Associate Professor, Department of Mathematics, University of Arizona, Tucson, Arizona 85721.

Several weeks into the consulting seminar that I taught in the fall of 1995, one frustrated student observed that none of our clients seemed to know exactly what they wanted. This, I believe, is the rule rather than the exception, and in such circumstances the fundamental contribution of the statistician is to help the client formulate appropriate questions. Statisticians know what kinds of questions can be answered and they excel at abstracting the essential features of an investigation without becoming distracted by the (often fascinating) details of the particular application. In my experience, despite their limited knowledge of the application, they often discover confounding factors and suggest alternative causal explanations that had not occurred to the investigator(s)—not because statisticians are more clever than scientists, but because statisticians are trained to look for such things.

Perhaps it is not surprising that statisticians take for granted the general character of statistical reasoning and tend to emphasize the technical procedures that they study and employ. Unfortunately, one consequence of this emphasis is the corresponding perception by clients that this is all that statisticians have to offer. Many—if not most—of my consultations begin with the client asking technical questions about the procedures that he or she has been using or contemplating. I invariably respond by asking the client to tell me a little about the application. Because the answer is usually too esoteric for me to understand, I follow up by asking the client to explain the project to me as though he or she was explaining it to his or her parents. (A former colleague with considerable consulting experience substitutes "grandparents" for "parents.") Not only have I found this to be a necessary prelude to more sophisticated discourse, it is really quite

remarkable how much progress can be made at this very unsophisticated level.

The fact that so much of what statistics has to offer resides in its way of thinking rather than in its technical input has important implications for how statistics should be taught, especially in service courses. Perhaps inevitably, most statistics courses are organized by procedure. Procedures are illustrated by sanitized examples that carefully avoid the complications and ambiguities that compromise their use in the real world. This practice makes it easier for students to learn the procedures, but it is apt to mislead them into identifying statistics as a collection of mathematical recipes. In fact, it is to resolve the messy issues that are carefully hidden in the typical service course that the *judgment* of a statistician is most needed.

I was a self-employed statistical consultant from 1989 through 1992. When I returned to academia in 1993, I found that my consulting experiences had affected my pedagogical priorities. For example, I regularly teach an introductory statistics course for graduate students from other departments who will require statistical guidance in their dissertation research. Because no one becomes statistically self-sufficient after one semester of study, I try to prepare students to become intelligent consumers of the assistance that they will inevitably seek. Service courses train future *clients*, not future statisticians.

2. "The statistician must enter into the context of the problem, not just as an "advisor," but as someone prepared to understand the data, analyze the data, interact with those who really own the questions being asked and consider the impact of statistics within the real context of the problem."

I have always advised my students (and anyone else who inquired) that, in selecting a statistician with whom to work, one should seek an individual who wants to learn about the application and avoid individuals who merely want to be handed a data set and to return an answer. In virtually every application, there is a gap (often a vast gulf) between the details of the application and standard statistical theory. For various reasons, it seems to me that this gap is more easily bridged by the statistician than by the client.

First, the mathematical theory of statistics is likely to seem far more esoteric to the client than is the client's discipline to the statistician. Second, the statistician will usually have a good sense of what he or she needs to understand about the client's discipline and can ask pointed questions toward that end, whereas the client will often not know what statistical issues are relevant to the problem. Third, and most important, gaps between application and theory require that compromises be made. Nature does not compromise—if natural phenomena are to be studied, then it is incumbent on the statistician either to devise relevant theory or to make informed judgments about the propriety of using standard procedures in situations not addressed by extant theory.

There are other reasons for statisticians to become aggressively involved in their consultations. Jennifer Hoeting emphasized that "statistical consultants should make every effort to obtain the raw data, if available." I definitely agree, but I submit that they should also make every effort to observe (some of) the data collection. Not only can this be enormously interesting and entertaining (as the statistical consultant to a U.S. Bureau of Reclamation study of the effect of fluctuating flows from Glen Canyon Dam on riparian bird nesting, I spent 18 days rafting the Colorado River!), but it is often essential for proper analysis and interpretation of the data.

For example, I was the statistical consultant to several longitudinal studies of the effects of Alzheimer's and Parkinson's diseases on memory and language. In one study, we administered a computerized serial reaction time task. The task comprised 6 sequences of 8 blocks of 10 items. It was well known that, within each sequence, the block mean reaction time decreased as subjects acquired greater proficiency in responding. Our data exhibited this pattern for the first five blocks and for the last three blocks, but there was a discontinuity between them. Indeed, the mean reaction time for the sixth block was dramatically greater than for the fifth block. The scientists were baffled, so I did my own detective work and asked one of the staff to administer the test to me. It turned out that the computer program attempted to store a sequence of responses in active memory, but we were using a computer with insufficient memory to store the entire sequence. In the midst of the sixth block, there was a "hiccough" as the computer wrote the contents of active memory to disk. The hiccough did not last long—the interviewers had not noticed it—but it was more than enough to break a subject's rhythm and invalidate the experiment.

In conclusion, these articles evoked fond recollections of my own consulting experiences and caused me to reflect upon their role in my professional development. Had they been available, I would have asked the students in my consulting seminar to read them. I hope that they will encourage other statisticians to broaden the scope of their professional activities.

Comment

Karen Kafadar and Max D. Morris

These articles provide further evidence of the role that statistics can play in science, as has been noted previously by eminent statisticians. We review some of these earlier references to consulting in the literature and emphasize that applications and the advancement of theory, both scientific and statistical, go hand in hand, and thus that consulting should be actively encouraged in graduate programs and valued by university faculties.

"Statistics depends for its raison d'être and continuing vitality on continued contact with substantive disciplines that actually generate data and make inferences in the face of uncertainty" (Moore, 1990, p. 268). This collection of papers provides evidence for Moore's assertion; in fact, some of them even confirm his later statement that "the direction of statistical research *is* affected by real problems, and the resulting new methods *are* used by practitioners." It is a pleasure to read a series of articles such as these, particularly when they lead to new approaches to analyzing data and new insights about the processes that generated them. The authors of these articles join other statisticians who have previously outlined important principles of statistical consulting and demonstrate the important role that statistics can play in advancing both the science of the application and the research in our own field. Curiously, despite statistics's dependence on such consultancies for its survival and continued growth, consulting problems as the basis for motivating the methodology are not fashionable in some circles, perhaps because consulting and applied statistics in general is sometimes regarded as a required function rather than as a vehicle for advancing research. This series of articles reminds us that useful research comes from useful problems and that, even in those instances where no new methodology was developed, the statistician nonetheless contributed insight into the problem.

Early exposure to consulting problems can foster an appreciation for their role in useful research. Many graduate programs in applied statistics encourage participation in consulting. One example known to us is the Stanford biostatistics workshop, a weekly two-hour seminar attended by both medical scientists and statisticians. Generally, the floor is given in the first hour to the client, who describes the problem and then relinquishes the floor in the second hour to the consulting statistician. The interaction is very stimulating, and often the problems have led to advancements in statistical practice, such as those later published by Mosteller and Parunak (1985), Efron and Feldman (1991) and Bacchetti (1990). Besides generating enthusiasm, respect for consulting and opportunities for applied research, these types of interactions can provide important motivation and guidance for basic research in statistics (but, as Tweedie mentions, the demanding pace of jobs—be they in government, industry, or academe—sometimes prevents one from following through).

The case studies here add to the many illustrations of the valuable and unique role that statistical thinking can provide and how much a statistician can contribute despite relatively cursory knowledge of the problem. As Tweedie says, "It is hard to overestimate how powerfully our discipline trains us to think about complicated issues in ways that allow us to quickly diagnose difficulties in esoteric disciplines to which we have had only several minutes of introduction." Examples of the statistician's potential contributions to the problem include the many essays in Tanur et al. (1989), as well as the missed opportunity in connection with the Challenger Space Shuttle described by Hoadley and Kettenring (1990). Many of our collective early consulting experiences had little to do with statistical analysis per se, but involved broader considerations which nonetheless required statistical reasoning and guidance. We, like other young investigators, learned at this stage that a statistical consultation is not the conversion of a "real-world problem" into a statistical exercise, but a true collaborative experience in which the statistician brings one of the critical components. Technical expertise and creative ability are, of course, extremely important here, but of equal importance is the ability and willingness to see major problems from a broader perspective. Our early consulting experiences at Hewlett Packard (HP) and the University of Texas Health Sciences Center–San Antonio (UTHSCSA), respectively, were in environments where the number of statisticians was small relative to other research staff, but where opportunities to contribute to the planning and execution of important research programs were many and varied. One of the research and development labora-

Karen Kafadar is Professor, Department of Mathematics, University of Colorado, Denver, Colorado 80217-3364. Max D. Morris is Senior Research Staff Member with the Mathematical Sciences Section, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831-6367.

tory directors at HP once spoke to the company's 50 or so statisticians and started out by joking that he'd never known a field that based its entire existence on admitting what they *don't* know—but then expressed his admiration by saying, "I don't think there has ever been a group of professionals at HP who has had such a major impact on the company in such a short period of time."

At the other extreme, some of these articles remind us how much *more* a statistician can contribute with deeper knowledge of the problem, particularly the article by Broman, Speed and Tigges. David Byar once told his Biometry Branch that, for every statistical methodology article, one needs to read about 10 substantive ones related to the subject matter in the field. Wangen (1990) warns, "When we become remote from those who need our help, we cannot maximize the contributions of our professionals." Gnanadesikan is quoted as saying, "Most statisticians do not seem to become involved *deeply enough* in subject matter areas to understand the scientific problems in their contexts" (Hoadley and Kettenring, 1990, p. 245). Tweedie and Hall indirectly touch on one of the central reasons why sufficient background knowledge is important: "As in all statistics applications, the modelling should come first . . . ." Of course, a truly appropriate model and set of assumptions can be determined only after at least some, and usually considerable, understanding of the reality being modeled. The eventual value of any consultancy depends upon the statistician's ability to identify a model appropriate to the situation and a correspondingly appropriate analysis, neither of which is possible without substantial understanding. A valuable step in acquiring that understanding is a visit to the site where the data were or will be collected. Most of us have stories similar to Tweedie and Hall and to Broman, Speed and Tigges about visiting a laboratory or office and discovering a key aspect of the problem that either answered the client's question of interest or else revealed particular aspects that needed to be taken into account in any statistical recommendation or advice.

The open discussion at the WNAR meeting in Pullman highlighted further essential ingredients for successful consulting. Taylor demonstrates the value of nailing down the objectives of the study right from the start so that the study can be designed to achieve them as efficiently as possible. Some experienced investigators arrive at the statistician's office with carefully thought out, detailed questions and study goals in mind—but many do not, like the tax office clients of Tweedie and Hall. William G. Hunter taught generations of statistical consultants at the University of Wisconsin–Madison the importance of asking the right questions: "At the outset the most important question for the statistician to ask is: What is the objective of this investigation?" He went on to describe a session with two investigators who spent 45 minutes discussing this question, and, "when it ended, they agreed on what it was they were about. They thereupon said that I had been most helpful, and we said goodbye" (Hunter, 1981, p. 73).

Similarly, collecting the right data is important, as Taylor describes in her article, to answer the main questions of interest. As a related point about data, David Byar once said, "Better an imprecise measure of something important than a precise measure of something unimportant." Often, a statistician's primary contribution is to bring clients to recognize that their elaborate measurements were only tangentially related to the question of interest, and they might be better off investing a little time in collecting the relevant data. Hoeting acknowledges the possibility of this problem in her study: "The *mean* daily discharge may not be a good measure of the water release pattern, because two very different water release patterns could have the same mean daily discharge." The first sentence in Tukey's *Exploratory Data Analysis* is "It is important to understand what you CAN DO before you learn to measure how WELL you seem to have DONE it" (Tukey, 1977, p. 6). The reality of what you really CAN DO within the context of available data can occasionally frustrate the investigator, especially when the data have been expensive or difficult to collect. Hoeting's clients had generated "the best data set ever obtained," but it could not provide a complete answer. Even here (or perhaps especially), the statistician can play a critical role, by explaining why this is so, participating in planning activities and helping the investigator avoid the strong temptation to claim more than can be objectively justified based on the limited data in hand. Even when the data are more complete with respect to the questions being asked, this last point is often operative. One principal investigator of a large study at UTHSCSA insisted that the consulting statistician had to function as the "scientist's conscience" when it was time to report the results of a study, because good scientists must continually think about potential interpretations of their data which are well beyond what can be honestly called "current results."

One of Hoeting's "basic rules for statistical consultants," namely, "[A]lways check the data for errors at the start of the project," is well known. Jim Filliben at the National Institute of Standards and Technology analyzed data from the Department of Transportation's Daylight Savings Time Study of

the number of traffic and pedestrian accidents during Nixon's extended daylight savings time edict aimed at saving energy in 1973. The congressionally mandated study resulted in data with so many errors that conclusive evidence concerning changes in the number of accidents could not be confirmed. Again, quoting Tukey (1977, p. 10): "One thing we regretfully learn about work with numbers is the need for checking. Late-caught errors make for painful repetition of steps we thought finished. Checking is inevitable; yet, if too extensive, we spend all our time getting the errors out of the checks. Our need is for enough checks but not too many."

Other consultants have also noted before Hoeting that "we must guard ourselves against standing on the 'statistician's pedestal' from which we lecture scientists on the limitations of their studies. ... We should recognize that, just as statisticians make compromises while doing analyses, investigators are under considerable constraints when designing their studies." Many years ago, Lincoln Moses told the students in his experimental design course, "Now, it is always nice to have a balanced design. But if your experiment isn't *balanced*, don't throw up your hands and go home! There are ways to analyze it. And I'm going to show you how." As a client of statisticians, Wangen (1990, p. 273) suggests, "[Statisticians] should do what they can to help, regardless of personal opinions about what could have been done, and refrain from commenting negatively on aspects of the work performed before their involvement unless invited to do so. Good statisticians can nearly always provide useful assistance at any stage of a project." Compare Tweedie's comment: "Statistics can contribute something that was not there previously, and we have much to offer to almost everyone."

Another important principle for successful consulting is the statistician's use of the client's language, as do Broman, Speed and Tigges, rather than the other way around: "To be successful, we must learn to serve. That, of course, requires *statisticians* to get to know the language and problems of [the clients]" (Hunter, 1990, p. 261). (As a graduate student, during a spring picnic, I (Kafadar) once asked John Tukey a philosophical question about statistical practice. The statistical concepts in his courses used to challenge both students and professors alike, so the simplicity of the reply made a real impression on me: he once had a client who was a medical doctor at Sloan Kettering, and he told me, "It was nearly a year after our first meeting that I even came so close as to mentioning a t-test to him.") But probably the most important lesson for statisticians is to avoid the proverbial "error of the third kind," that is, providing the right answer to the wrong question: "Far better an approximate answer to the *right* question, which is often vague, than an *exact* answer to the wrong question, which can always be made precise" (Tukey, 1962, pp. 13–14).

The primary benefit to statisticians of serious creative participation in consulting, in addition to the sense of satisfaction by contributing a solution to a real problem, is the potential for advancing statistical research. This potential has been recognized by many, including past ASA President Jerome Cornfield: "Application requires understanding, and the search for understanding often leads to, and cannot be distinguished from, research. The true joy is to see the breadth of application and the breadth of understanding grow together, with the unplanned fallout—the pure gravy, so to speak—being the new research finding" (Cornfield, 1975, p. 11). Box (1984) lists several important interactions between practice and theory where practical problems stimulated theoretical development of whole new areas: small samples → Student's t (Gosset); rainfall data at Rothamstead → distributed lag models and orthogonal polynomials (Fisher); agriculture experiments → factorial designs and confounding (Yates); balancing several factors at once → Youden squares (Youden); large data sets on telephone usage → exploratory data analysis (Tukey). Of course, major new statistical ideas and methodologies such as these are often not born of one or a few consulting problems, but generally develop gradually in response to experience gained by repeatedly thinking about applications in a statistical context. For example, computerized data collection systems such as the one Taylor describes have revolutionized the kind and quantity of data available in many applications areas, and consulting on such problems can stimulate research in the critically important area of analysis of large and complex data structures. The connection between important improvements in methodology and participation in applications is clear and is documented by, for example, the IMS Panel on Cross-Disciplinary Research in the Statistical Sciences, which identified many areas where "statisticians have achieved signal advances in theory and methods as they worked on applications in many fields, and, in turn, statistical thinking and methodology have greatly influenced the development of virtually all areas of science" (IMS Panel, 1990, p. 121), including agriculture, health, military operations and transportation and communication systems (pp. 137–138). They identify "Type A" (rather routine consultancies) and "Type B" (full-fledged colloborations) interactions, with the hope that Type A ones

evolve into Type B ones, with the help of needed resources and increased supporting infrastructure (such as the development of the National Institute for Statistical Sciences that followed the publication of this Panel Report).

Despite this opportunity for advancing research through consulting, many academic programs do not actively encourage it. Tribus noted years ago, with probably little change since then, "students believe that professional statisticians are presented with well-formulated problems, which appear over the transom and for which they are to provide clever solutions that are exchanged for tokens of appreciation that have great value in the marketplace. They have been brought up on a diet of 'given this, find that'—usually with the understanding that the method to be used is the one taught in the last class. Unfortunately, the world does not 'give' problems—you have to go and get them" (Tribus, 1990, p. 271). Hogg (1991) also notes that "We do not encourage enough teamwork, with students working on projects" (p. 342); "How often do our Ph.D. students understand the importance of these ideas and develop their communication skills so as to be effective in the classroom or in consulting?" (p. 343). In an earlier article, Tweedie (1992) likewise lamented that the report of the New Researchers Committee of the IMS (1991) did little to direct new researchers into important areas, advising instead that " 'unless you need the data analysis experience, your role is to dispense advice' in a consulting context" (Tweedie, 1992, p. 264). Even as recently as August 1996, several statisticians at a meeting in Halifax admitted that, while collaboration and joint authorship can be enormously beneficial to science, they as faculty members would discourage untenured faculty from anything other than independent, sole-authored papers in prestigious statistics journals. Fortunately, the Editors of this journal have chosen to encourage statistical consulting by publishing this set of articles that serve to illustrate the important role that statistics can play both in advancing the science in the field of application, as well as providing background and motivation for future statistical research.

This discouragement of collaboration might be reduced if journals were to publish more articles demonstrating creative methodology motivated by challenging consulting problems. A few journals already emphasize applications in their editorial policies: for example, *Technometrics* ("adequate justification of the application of the technique, preferably by means of an actual application to a problem"), *Statistics in Medicine* ("The journal will publish papers on practical applications of statistics and other quantitative methods to medicine and its applied science"), *Biometrics* ("describing and exemplifying developments in these methods and their applications in a form readily assimilable by experimental scientists"), *JASA Applications and Case Studies* ("For real data sets, present analyses that are statistically innovative as well as scientifically and practically relevant. ... Using empirical tests, examine or illustrate for an important application the utility of a valuable statistical technique") and *JASA Theory and Methods* ("The research reported should be motivated by a scientific or practical problem and, ideally, illustrated by application of the proposed methodology to that problem. Illustration of techniques with real data is especially welcomed and strongly encouraged"). While a routine statistical analysis does not fall into these categories, a creative analysis that offers a novel approach to the problem often would. The recognition of the value of consulting as a vehicle for advancing statistical research should be widened through the publication of such articles. The Editors of this journal have taken some further steps toward this end.

ADDITIONAL REFERENCES

BACCHETTI, P. (1990). Estimating the incubation period of AIDS by comparing population infection and diagnosis patterns. *J. Amer. Statist. Assoc.* **85** 1002–1008.

BOX, G. E. P. (1984). The importance of practice in the development of statistics. *Technometrics* **26** 1–8.

CORNFIELD, J. (1975). A Statistician's Apology. *J. Amer. Statist. Assoc.* **70** 7–14.

EFRON, B. and FELDMAN, D. (1991). Compliance as an explanatory variable in clinical trials (with discussion). *J. Amer. Statist. Assoc.* **86** 9–26.

HAHN, G. J. (1990). Commentary on "Communications between statisticians and engineers/physical scientists." *Technometrics* **32** 257–258.

HOADLEY, A. B. and KETTENRING, J. R. (1990). Communications between statisticians and engineers/physical scientists (with discussion). *Technometrics* **32** 243–270.

HOGG, R. V. (1991). Statistical education: improvements are badly needed. *Amer. Statist.* **45** 342–343.

HUNTER, J. S. (1990). Commentary on "Communications between statisticians and engineers/physical scientists." *Technometrics* **32** 261.

HUNTER, W. G. (1981). The practice of statistics: the real world is an idea whose time has come. *Amer. Statist.* **35** 72–75.

IMS PANEL ON CROSS-DISCIPLINARY RESEARCH IN THE STATISTICAL SCIENCES (1990), Cross-disciplinary research in the statistical sciences, *Statist. Sci.* **5** 121–146.

MOORE, D. S. (1990). Commentary on "Communications between statisticians and engineers/physical scientists." *Technometrics* **32** 265–266.

MOSTELLER, F. and PARUNAK, A. (1985). Identifying extreme cells in a sizable contingency table: probabilistic and exploratory approaches. In *Exploring Data Tables, Trends, and Shapes* 189–224. Wiley, New York.

NEW RESEARCHERS COMMITTEE OF THE IMS (1991). Meeting the needs of new statistical researchers, *Statist. Sci.* **6** 163–174.

TANUR, J., MOSTELLER, F., KRUSKAL, W. H., LEHMANN, E. L.,

LINK, R. F., PIETERS, R. S. and RISING, G. R. (1989). *Statistics*: *A Guide to the Unknown*, 3rd ed. Wadsworth, Belmont, CA.

TRIBUS, M. (1990). Comment on "Communications between statisticians and engineers/physical scientists." *Technometrics* **32** 271–272.

TWEEDIE, R. L. (1992). Comment on IMS New Researchers Report. *Statist. Sci.* **6** 263–264.

TUKEY, J. W. (1962). The future of data analysis. *Ann. Math. Statist.* **33** 1–67. [Reprinted in *The Collected Works of John W. Tukey* **4**. *Philosophy and Principles of Data Analysis, 1949–1964* (L. V. Jones, ed.) 391–484. Wadsworth, Belmont, CA.]

TUKEY, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley, Reading, MA.

WANGEN, L. E. (1990). Comment on "Communications between statisticians and engineers/physical scientists." *Technometrics* **32** 273–274.