Statistical Science
1991, Vol. 6, No. 3, 240–268

Avoiding Statistical Pitfalls

Christopher Chatfield

Abstract. In many real-life problems, avoiding trouble can be at least as important as achieving optimality and is a necessary precondition anyway. Some guidelines are suggested for avoiding pitfalls throughout a statistical investigation. A range of real-life examples is presented to show how difficulties can arise in practice and how they may be overcome. These examples are unlike the idealized and sanitized illustrations that typically appear in the literature.

Key words and phrases: Strategy, objectives, messy data, asking questions, consulting, IDA, graphs, model-building.

1. PRELUDE

Imagine the following scenario. You have been called in as consultant to advise on a discriminant analysis of a large set of data. There are observations on 20 variables for each of several thousand customers who have been partitioned into four groups, depending on whether or not they have taken advantage of certain financial facilities. The objective is to find which variables are the best indicators for distinguishing between the groups, although the exact details need not concern us here. You have not been involved in collecting (or rather assembling) the data, which you are assured is a systematic 1 in 20 sample of all the company's customers. The latter are held in a large database organized by the computer section (rather than by statisticians). Time is pressing (they want the results yesterday!) and so (unwisely and perhaps unconsciously) you choose to cut a few corners during the initial data analysis.

Nevertheless you do look at the correlation matrix of the variables and discover, to your horror, that there is an *off-diagonal* value of 1.000000! Has one variable been erroneously duplicated? No, because a quick check of the first few values in the two data columns shows that they are not the same. Could one variable be a linear combination of the other? No, because the two suspect variables have different correlations with the other variables. Could the computer program be wrong? No, SPSS can surely be trusted to calculate a correlation correctly. So what is going on?

You next check the histogram of each variable (which you should have done in the first place!) and find to your dismay that there is a small bunch of exceptionally extreme values recorded as 999,999,999 for both variables. Subsequent enquiries show that this is the (ridiculous) default value for missing observations and that the two variables concerned are such that, when one (unusually) is missing, then so is the other. However, the "small" number of duplicated default values is enough to produce the apparently "perfect" correlation. You hastily exclude extreme values outside a sensible allowable range and then proceed with the analysis, congratulating yourself that you spotted the problem in time.

No doubt there will be some readers who do not customarily face such situations, but they are likely to be people in an academic environment who do not face deadlines with messy data. For the rest of us (including many academics), the above story may sound all too familiar. It is therefore sensible to ask what lessons can be learned from such real-life dramas, and what steps can be taken to guard against such eventualities. Clearly some general guidelines are desirable.

2. INTRODUCTION

> Most real-life statistical problems have one or more nonstandard features. There are no routine statistical questions; only questionable statistical routines [D. R. Cox]

Many statistical pitfalls lie in wait for the unwary. Indeed, statistics is perhaps more open to misuse than any other subject, particularly by the nonspecialist. The misleading average, the graph

Christopher Chatfield is Reader in Statistics and Head of the Statistics Group in the School of Mathematical Sciences, Bath University, Bath, BA2 7AY, United Kingdom.

with "fiddled" axes, the inappropriate P-value and the linear regression fitted to nonlinear data are just four examples of horror stories which are part of statistical folklore and will not be considered here (for "popular" treatments, see, for example, Huff, 1954; Reichmann, 1964; Hooke, 1983). Nowadays, with the aid of computer software, the nonspecialist can commit a much wider range of offences, which I shall not attempt to list, although I think particularly of multiple regressions with an excessive number of predictor variables and of inappropriate factor analyses. However, this article is primarily concerned with the avoidance of mistakes by the specialist statistician.

Given the importance of the topic, *avoiding trouble* has been arguably neglected in the literature. Statistical journals tend to concentrate on developing the methodology of ever-more complicated techniques, which, while important, needs to be complemented by an ability to recognize *when* and *how* to implement such techniques, and to know *why* things are done and what the results *mean*. Thus, general *strategy*, which should include the avoidance of trouble, is at least as important as knowing the details of specific *techniques*. Unfortunately statistics textbooks also tend to concentrate on techniques. The few relevant references on strategy include Cox and Snell (1981), Chatfield (1988), Preece (1987) and a series of papers by Hand (e.g., 1990). The latter are more concerned with the development of *expert systems* but are also relevant to assessing how a statistician should approach statistical problems. Understanding how a statistician thinks is at least as important as understanding how a computer can act as an expert system, and the two can learn from each other!

In addition to concentrating on techniques at the expense of strategy, textbooks also typically concentrate on *optimal* procedures for collecting and analyzing data under somewhat idealized conditions. With a single, clearly defined objective (e.g., in most sequential clinical trials), optimality can be very important. However many real-life problems fail to resemble those in textbooks for a variety of reasons, including the presence of messy data and the absence of a single clear objective. While some investigations regrettably have no clear objective, many others (e.g., most surveys) are multi-purpose and then it may not be possible to achieve simultaneous optimality for each objective. In any case, optimality properties usually depend on a parametric model that will itself depend on certain assumptions that are unlikely to be satisfied exactly and may be seriously in error. This suggests that the statistician should be more concerned with finding a safe, robust and practical solution to a given problem. In particular, biases and errors can arise in all sorts of ways and the statistician must be on the lookout for problems and take positive steps to avoid trouble. Problems include not only making mistakes of many types but also going up blind alleys and unnecessarily repeating work already done by others.

To some extent, avoiding trouble is a bit like avoiding road accidents when driving a car. As you gain experience, you are increasingly able to sense situations where danger will be lurking and take appropriate defensive action. Thus experience is the real teacher. However, some brief general guidelines could be helpful, particularly for the less experienced practitioner, and Section 3 attempts to provide these.

3. GENERAL GUIDELINES

For convenience, we divide advice into sections corresponding to the four main stages of a typical statistical investigation. Much of the advice may be regarded as "obvious commonsense," but commonsense is not as common as might be expected! Some of the points are expanded in Chatfield (1988, Part I), and some are illustrated by examples. Much of the advice comes under the heading of "good statistical practice" (Preece, 1987) and consists of making routine checks and taking reasonable precautions.

3.1 Formulating the Problem

It is sad that many investigations are carried out with no clear idea of the objective. This is a recipe for disaster or at least for an error of the third kind, namely "giving the right answer to the wrong question." Kimball (1957) gives several instructive examples of this type of error, although, as he points out, only errors of the third kind that become known can be corrected, and there may be many that we never know about. Clearly "thinking around" the problem and formulating a clear plan are essential. Problem formulation may involve *asking questions* (see below) or carrying out appropriate *desk research* in order to get the necessary background information and check the physical and/or statistical aspects of the problem. There should also be a check on any prior knowledge, particularly when similar sets of data have been analyzed before and the problem is not to fit a model from scratch but to see if the new data are compatible with earlier results. It is sad that many studies end up "re-inventing the wheel." Cost constraints also need to be considered and, if necessary, critically reviewed. Many projects are underfunded as well as underplanned, and it may be necessary to say that targets are unattainable with the resources provided. For example, some surveys have over-ambitious objectives given the possible sample size. Generally speaking, problem formulation is too often either neglected completely or the required time is grossly underestimated.

While sometimes in charge, the statistician is more likely to be advising, collaborating with or providing consultancy advice to nonstatisticians, and there are many potential difficulties and dangers involved in this (e.g., see Hand and Everitt, 1987). In addition to the advice listed by Chatfield (1988, page 68), I would add the following. (a) If data have already been collected, always *ask to see them* (see Example 1) and find out exactly how they were collected. (b) Try to avoid answering questions over the telephone. The most important general piece of advice is to *ask lots of questions* (see Example 5 in Section 4) and to be persistent where necessary. Most statisticians are all too familiar with conversations which start:

Q: What is the purpose of your analysis?
A: I want to do a significance test.
Q: No, I mean what is the overall objective?
A (with puzzled look): I want to know if my results are significant.

And so on . . . .

EXAMPLE 1. Asking to See the Data. Asking questions is particularly important in avoiding catastrophes with "the scientist who knows what he wants." Dr. X, from our Chemistry department, asked for "5 minutes" of my time, to answer a "simple" question about regression. He had fitted a straight line to some data and wanted to know the formula for calculating the standard error of the intercept. His knowledge of regression was clearly nontrivial, but even so I asked to see the data first (I always do). The scatter plot revealed a relationship that was clearly curvilinear over at least part of the range. However Dr. X assured me that "chemists always fit straight lines to this sort of data." I pointed out that this was unwise and that if the point estimate of the intercept was biased, then there was little point in calculating its standard error under inappropriate assumptions. In a similar vein, Barnett (1987) describes three case studies that began with a "simple" request to help fit a straight line to some data but that turned out to be far more complicated.

Another colleague who "knew what he wanted" was Dr. Y from the Social Sciences department. He simply wanted to know the formula for carrying out a chi-squared goodness-of-fit test on a two-way table and did not want to bother me with doing the actual analysis. Nevertheless I asked to see the data and found that he had already committed the common cardinal sin of converting the observed frequencies to percentages!

To overcome students' natural (?) reluctance to ask questions, I have tried setting projects that are deliberately incomplete or partially wrong, as often happens in real life. Another instructive exercise is to ask students what is the average of the two numbers 10 and 350 (Hand and Everitt, 1987, page 4). Most will look suspicious, but say 180. Only a few quick-witted students will respond with the correct answer, which is not an answer at all but rather another question, namely "What are the numbers?" (you should never analyze "numbers," but rather data on known variables). On being told that the observations are phase angles, so that 350° corresponds to $-10°$, it becomes obvious that the more meaningful average is zero degrees.

The other major aspect of problem formulation is to decide how to collect the data or, if they have already been collected, to rigorously assess the procedure that has been used. This will be considered in the next section.

3.2 Collecting the Data

Statistics courses traditionally concentrate on data analysis. In practice, data collection is at least as important, because poor data cannot necessarily be rescued by a "fancy" analysis. Thus, consultants and data analysts need to pay more attention to designing a "good" collection method as an essential element of any scientific investigation. For example, a common experience in advising on clinical trials is to find that one has to start by re-designing the protocol.

In an ideal world, the statistician might hope to have complete control over the data-collection process so as to ensure that the general principles of good experimental design and good survey practice are followed. In practice, this may not be possible for a variety of good or not-so-good reasons. Data are usually collected by staff with little or no statistical expertise, and the collection process may be subject to a variety of practical constraints. If, in addition, the data have been collected without the advice of a statistician, then the data may be suspect or even worthless. Many pitfalls are associated with not finding our exactly how such data were collected and with applying statistical procedures based on incorrect assumptions about the data-collection method.

Bad data may also arise due to poor recording techniques (see Example 2). This may be revealed during the initial examination of the data (see Section 3.3).

EXAMPLE 2. A Suspect Recording Procedure. This example illustrates the difficulties that can arise in recording measurements in real life. A scientist wanted a function to predict the conductivity of a copper sulphate solution. A composite experiment (a full two-level factorial augmented with axial and central points) was designed with three control variables, namely $CuSO_4$ concentration, H_2SO_4 concentration and temperature. The results were collected, and a full quadratic model was fitted using least-squares regression analysis. An almost perfect fit was expected, but in fact only about 10% of the variation was explained by the fitted model. The scientist then complained to a statistician that there must be something wrong with the regression package.

The statistician questioned the scientist carefully and found that the observations had been collected, not by the scientist himself, but by a technician who was thought to be reliable. The experiment consisted of making up solutions in beakers, which were put onto temperature control pads. When they reached the design temperature, a conductivity probe was held in each beaker and a reading was taken from a digital panel. The statistician asked for the experiment to be repeated so that he could see exactly how the observations were taken. Two problems were soon spotted. Sometimes the probe (held by the technician) touched the side or bottom of the beaker and sometimes it was immersed by only about one centimeter. Secondly, the digital readout took about 15 seconds to settle to a stable value, (the probe required diffusion of the solution through a membrane), but the technician often didn't wait that long. The statistician therefore made two recommendations, neither of which was "statistical." First, the probe should be fixed in a clamp so that it always went to the same depth. Second, a timer should be used so that the conductivity wasn't read until 30 seconds after immersion.

The whole experiment was repeated in less than one hour, the regression analysis in 10 minutes, and this resulted in a 99% fit!

The different aims of *achieving optimality* and *avoiding trouble* are well illustrated by the problems involved in *survey design*. The literature devotes much attention to optimal allocation so as to minimize the variance of a sample estimate subject to a cost constraint. Most of this theory ignores the multipurpose, multivariate nature of most surveys and may also disregard nonsampling errors, which are usually at least as important as sampling errors. In real life, a survey must be practical, simple, flexible and robust, and factors such as administrative convenience and cost may take precedence over theoretical considerations. The term *proximum*, rather than optimum, has been suggested (O'Muircheartaigh, 1977) for describing a design that in some sense aspires to "approximate optimality" while also being practical.

The general principles of good survey design will not be repeated here (Moser and Kalton, 1971, is still as good a reference as any). There are many pitfalls associated with questionnaire design (Schuman and Presser, 1981, is helpful) and it is essential to check that questions are fair (not always easy to assess) and have been pilot tested. Checks should also be made that (a) the sample is representative and of a suitable size (see Example 3), (b) interviewer bias is reduced as far as possible and (c) the data have been coded and processed accurately. Quite apart from the nonresponse problem, many sampling procedures do not in fact give every member of a population an equal chance of being selected even when the sample does appear "random" (e.g., see Huff, 1954, Chapter 1). A final precaution worth mentioning here when *using* survey data, particularly official statistics, is that it is advisable to check the comparability of information drawn from different sources, and even from within a single source.

EXAMPLE 3. Three Unrepresentative Samples. (a) The Prelude data set (Section 1), used to highlight coding problems arising with missing observations, was later found to be faulty in another way. We were told that a 1 in 20 systematic sample had been taken from the population data base by the computer staff who were handling it. (Although nonrandom, a systematic sample can be expected to be reasonably representative here given that there is no cyclic behavior in the storage of customer details.) Subsequent checks revealed that the data base was stored on several different tapes because of its enormous size and that one tape, comprising all the customers from the north of England, had been missed completely! It would be nice to report that the "missing" tape came to light as a result of systematic checks. In fact, it was partly due to good fortune that the problem was noticed when the customers were partitioned by their home address area.

(b) A company wanted to carry out a statistical analysis of a small "random" sample of its clients. The selected sample comprised everyone whose surname began with the letter "V". This sample was selected, not by the computing staff, but by a management consultant with access to the data base. The choice was made solely on the grounds of computing convenience on the mistaken assumption that the alphabetical order of surnames is somehow random. In fact, this appalling choice effectively excludes all Scots and Welshmen, for example. Fortunately the deficiency was easily spotted by the statistician who was asked to analyze the data. As in (a) above, it emphasises the importance of checking exactly how a computer sample was selected.

(c) When Puerto Rico was hit by a recent hurricane, there were 10,000 claims by residents for hurricane damage. The U.S. government decided to base its total grant aid by finding the total of claims in the first 100 applications and then multiplying by 100. A colleague was involved in the difficult task of persuading the U.S. government that the first 100 applications need not necessarily constitute a representative sample! The grant aid was eventually increased.

There are also many pitfalls in *designing experiments*. Example 4 presents brief notes on some experimental designs that "went wrong" for one reason or another, while Andersen (1990) presents an alarmingly diverse selection of dubious clinical trials. Always check the following.

(a) *Randomization* should be incorporated correctly in the design. Randomization is the only safe way to overcome the effects of unforeseen nuisance factors, but when randomization is left to the client it is often carried out wrongly. In complicated factorial experiments, it is tempting to randomize levels of each treatment factor separately, instead of randomizing the whole design. In a regression experiment with one explanatory variable, where observations have to be taken sequentially, the order of the tests is often taken in the "natural" order determined by the increasing values of the explanatory variable. Then the effect of the explanatory variable is confounded with time.

(b) Treatments of interest should be able to be estimated in an unbiased way, and these estimates should be capable of a unique interpretation. In particular, Example 4(f) illustrates the problem of pseudoreplication.

(c) Potentially important interactions must be able to be estimated.

(d) There should be enough degrees of freedom to estimate the error term adequately. The dangers of saturated designs in Taguchi methods are highlighted by Bissell (1989).

(e) Check that the people carrying out the experiment know exactly what to do (see Example 4a). There is no point in devising a brilliant design if it is wrongly implemented. For example, every field plan should have a compass indication of North!

(f) The design must take account of the given practical situation (see Examples 4b to e).

Example 4. Some Faulty Experimental Designs. (a) In a particular field experiment, the treatments had to be applied on two separate occasions, but the field worker read the plan from the wrong end on the second occasion. The treatment combinations were therefore quite different from what had been planned.

(b) In another field experiment, rows were used for blocking even though all the fertility and management differences were between columns.

(c) In a large field experiment, the statistician was assured that all the blocks would be long and thin. A restricted-randomization scheme was therefore devised that ensured that it was impossible to have a long run of plots all with high (or all with low) levels of nitrogen. At one site, the shape of block, and the numbering of the plots within the block, was quite different from what the statistician was told. This resulted in a long run of plots all with high levels of nitrogen. These particular plots were badly affected by a fungus disease. Given the poor lay-out resulting from the above misinformation, there was no way of telling whether the high levels of nitrogen encouraged the disease.

(d) A statistician carefully devised a design for a $7 \times 3 \times 3$ factorial experiment. Only when the data had already been collected did the client reveal that one level of the first factor was a "control," for which all levels of the other two factors were equivalent. The treatment structure was therefore $1 + (6 \times 3 \times 3)$. A better design for this structure could therefore have been devised.

(e) Preece (1987) recounts an experience visiting a forestry experiment on some sloping land in Africa. When he commented that the design did not seem to match up with the terrain, he was told that the experiment was designed in Rome! While it may be reasonable to select the treatments in Rome, and perhaps the size and type of design, the details of the field plan should surely be resolved on the spot.

(f) The final example illustrates *pseudoreplication* and is adapted from the examples in Hurlbert (1984). Suppose we want to compare the decomposition rates of maple leaves on a lake bottom at depths of 1 meter and 10 meters. Two nylon bags are filled with an equal amount of leaves and one is placed at each of the two depths. One month later the amount of organic matter lost is measured. Unfortunately the difference tells us very little as we have no estimate of error. Thus we replicate the experiment by putting say eight bags at each of the two spots. This may look fine at first sight, but is actually not much better. The difference between the two group means will not necessarily indicate differences due to the depths but may simply be due to the two different spots we happen to have picked. In other words, the depth effect is confounded with the site effect. To carry out the experiment properly, we need to replicate the observations at different sites so that the estimate of the treatment difference is capable of a unique interpretation. Hurlbert (1984) defines pseudoreplication as testing for treatment effects with an error term inappropriate to the hypothesis being considered. In this case, you may think that it is obvious (at least after it has been pointed out) that the replicates are not *real* replicates, but in many situations it is easy to be fooled by pseudoreplication.

If the given data are not from a "proper" design or survey, but are simply historical observation data, then the textbooks rightly say that "extra care is needed." What does this mean? It means that important effects of interest may be confounded with nuisance factors, and there may be biases, either in the selected sample or in the way the data have been collected. Then careful thought is required to see if the data have any value. Even quite a large data set may be of limited value as size considerations are not enough. Thus although properly constructed observational studies can sometimes prove useful (e.g., Cochran, 1983), at least in providing useful pointers to check in subsequent studies, they can sometimes be worthless or even positively misleading.

3.3 Analyzing the Data

Data analysis can often usefully be thought of as having two main stages: (a) The initial examination of data (or IDA; see Chatfield, 1988, Chapter 6). This includes processing the data, checking the quality, and obtaining simple descriptive summaries, including summary statistics, graphs and tables. (b) Carrying out an appropriate inferential procedure. Selecting an appropriate method of analysis will often involve formulating and fitting a sensible parametric model, although nonparametric methods are also widely used. Finally the fit of the model and/or the appropriateness of the procedure must also be checked. We consider these two stages in turn.

(a) The Initial Examination of Data. A thorough IDA is important in any analysis, not only to check data quality and produce a descriptive summary, but also to help in formulating an appropriate model. In all these roles, IDA is particularly helpful in avoiding trouble. IDA has many similarities with *exploratory data analysis* (EDA), but there are also important differences. Tukey's landmark book (1977) describes a variety of important techniques for exploring data, but it can be criticized for introducing too much new jargon, for omitting standard tools such as the arithmetic mean and for failing to emphasize the importance of using the initial data analysis to formulate an appropriate model and hence choose an appropriate inferential method. Thus Tukey's approach has not always been integrated with the rest of statistics, and this explains my alternative choice of title (Chatfield, 1986).

The first step is to assess the *structure* of the data, particularly the sample size (are there enough observations to satisfactorily answer the given questions?), the number of variables (are any important variables missing?) and the type of variables. Many mistakes are caused by failing to distinguish between the different types of measurement scale (nominal, ordinal, interval and ratio), and in particular between count and measured data. For example, a company that produced y meters of a certain material each shift, of which w meters were scrap, set up a control chart scheme under the assumption that w was a binomial variate because w and y were recorded as integers. In fact, they were measured variables, albeit rounded to the nearest integer. As a result the calculated control limits were much too wide.

Data are increasingly made available in the form of a data base over whose construction the statistician may have little or no control. Rather the data base is likely to be managed by a computing department of some sort. Then it is particularly important to assess the structure and quality of the data, and better software is needed for handling large data bases and carrying out routine checks on them. An even more difficult problem arises when the statistician does not have access to the data base, but rather is given a sample and has to find ways of checking that the sample is representative. This is not easy; see the Prelude and Example 3(a).

Of course, checking data quality is important for any data set and may be effected by a variety of *data-snooping* techniques such as that described by Preece (1981) for looking at the distribution of final digits in order to check on the recording procedure. One feature which often gives trouble is the treatment of *missing observations* as already demonstrated in the Prelude. As a second example, I was asked to analyze some daily stock prices. Close inspection revealed that the values for public holidays had all been coded as "999" and these had to be removed or replaced before the data could be analyzed as an "ordinary" time series.

Another feature that can give trouble is the presence of *outliers*. Knowing when to adjust or remove an apparently extreme observation is something of an art. Do not forget that a large residual may result from a wrongly specified model rather than from an error or from a genuine extreme observation.

The next important task is to produce a descriptive summary of the data. Lack of space prevents a repetition of the many helpful guidelines available (see, e.g., Chatfield, 1988, Section 6.5). While apparently "simple," these descriptive summaries are not as easy to produce as they seem and are often done badly. For example, graphs are produced with poorly labeled axes or no title, summary statistics are tabulated with far too many significant digits and hideous computer tables are reproduced with no thought as to how they might be improved. Figure 1 shows an example of a poor time plot produced by a computer package. Note the peculiar intervals chosen to divide the vertical scale, the hideous labeling of the horizontal axis and the failure to name the dependent variable. This graph, bad as it is, is unfortunately all too typical of much computer output, particularly from PCs and emphasises the need for packages which allow the user to control the parameters of a plot. It is even more regrettable that graphs with similar characteristics to Figure 1 sometimes get published.

As another example, a published paper analyzed a time series of 349 observations and listed them in an appendix using horrendous E-format. The first four observations were listed as 0.572000 E + 03, 0.544000E + 03, 0.521000E + 03, 0.548000 E + 03, . . . and so on. All the numbers ended in "000E + 03." Why use 3 digits (e.g., 572) when you can fill more journal space with 12 digits!!? It is hard to see how tables like this, or graphs like Figure 1, are passed for publication by referees and editors.

(b) Inference. Mistakes can arise during the inferential process in a variety of ways. The analyst may use the wrong technique, or use the right technique but carry it out incorrectly, or use the right technique but adopt an inflexible approach that does not allow for suspect data or other peculiarities. Other problems arise when the data are really not good enough to answer the given question, although the statistician should nevertheless be prepared to do his best with inadequate data.

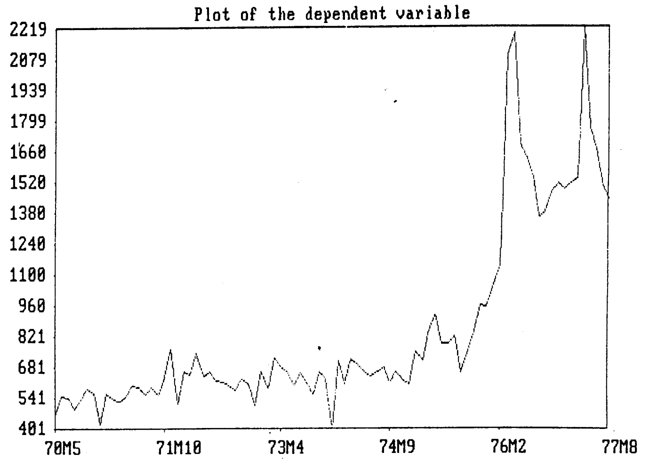


FIG. 1. *An example of a poor graph produced by a computer package.*

Choosing the wrong technique happens more often than might be expected, not only by nonstatisticians (see, e.g, Gore, Jones and Rytter, 1977), but also by statisticians. While we may be able to use a given technique correctly, the result of a technique-oriented training can make us poorly equipped to *choose* the most appropriate method. In fact, the first reaction to the "solution" of a problem may well be the wrong reaction, particularly if an IDA has been omitted.

Even when the right method is chosen, it may still be carried out incorrectly. The possibility of performing the arithmetic wrongly should become less of a problem given the spread of good computer software (but it is still advisable to check that the "answers" are of the right order of magnitude). Failure to understand and control sophisticated software is becoming a more common problem, and it is therefore important to choose software with desirable features, such as good documentation and easy data-editing.

Mistakes unassociated with the computer are also possible. In particular, it is possible to be misled by incorrect formulas in the literature (see Example 7), while the derivation and interpretation of P-values is particularly prone to error. Computer output could perhaps be better designed to avoid the latter problem and also to avoid giving undue emphasis to the results of significance tests rather than to the *estimation* of effects. The overemphasis on significance tests has, for example, led to publication bias (Begg and Berlin, 1988), wherein studies with positive results are more likely to be published than negative ones. The idea of looking for *significant sameness* deserves more emphasis as statisticians devote excessive attention to single data set problems rather than seeing whether an interesting effect will generalize to different situations.

Implementing the right technique in an inflexible way can be just as damaging. For example, the failure to identify an outlier can wreck the ensuing analysis (e.g, see Example B.1 in Chatfield, 1988), while failure to plot the data may lead to fitting an inappropriate model (e.g., in an over-rigid Box-Jenkins time-series analysis as described by Chatfield and Schimek, 1987). These examples illustrate the types of mistake that can arise from an inadequate IDA and remind us that an IDA is needed, not only to summarize data, but also to help formulate an appropriate model and hence select an appropriate method of inference.

Some further general advice is as follows: (a) Be prepared to try more than one type of analysis. (b)

Be prepared to make ad-hoc modifications to a standard analysis. (c) Be willing to seek help where necessary. (d) Be prepared to use lateral thinking.

Model-Building. Many analyses depend on an assumed parametric model. Thus model-building (e.g,. Gilchrist, 1984; Edwards and Hamson, 1989) is a crucial part of problem-solving and consists, not only of model-fitting, but also of *formulating* the model in the first place and then *checking* it afterwards. It is unfortunate that most textbooks concentrate on the easy estimation stage, when trouble is more likely to occur at the earlier specification stage or at the later validation stage.

A case study with realistically messy data, illustrating the difficulties involved in fitting a multiple regression model to data with as many as 20 variables, is provided by Janson (1988) and the subsequent discussion is also well worth reading. Unthinking use of robust procedures may actually increase the chance of model misspecification.

Some other general points on model-building are: (a) Many subjective choices may be involved in building a model. (b) Don't try to model without understanding the nonstatistical aspects of the system under study. (c) Don't throw out variables just because they are nearly co-linear. (d) Don't throw out observations just because they are influential or have high leverage (Belsley and Welch, 1988). (e) Don't extrapolate the model outside the range over which it has been fitted (see Example 8).

Model specification depends on a variety of inputs including the results of the IDA, a priori subject-matter knowledge and experience. Even so, certain assumptions may need to be made (e.g,. Edwards and Hamson, 1989, page 73). Subject-matter knowledge is particularly vital (e.g., which variables should be included and in what way?), but often ignored, although specialist knowledge can occasionally be prejudicial (see Example 6).

Model-checking is another vital phase. After fitting a model, a residual analysis is carried out and the model assumptions checked as far as possible. There are also various more general questions to consider when evaluating a model. Has anything important in the data been overlooked? Are there alternative models that fit nearly as well but lead to substantially different conclusions? Does the model really provide an adequate description of the data? This is the time to consider modifications to the model and *iterate* toward a satisfactory end result (e.g,. Gilchrist, 1984, Part VII). An ability and willingness to iterate is an essential attribute of a good statistician. The choice between models that fit approximately equally well should, where possible, be based on external considerations. If none are available, then some other practical, data-based criterion should be considered. For example, the choice between ARIMA time-series models with low but approximately equal values of the AIC should probably be made, not on which happens to give the minimum AIC, but on which gives the best forecasts of the most recent year's data.

The topic of forecasting provides a good illustration of the potential dangers involved in placing too much credence on a model. Model-based forecasts rely on the inbuilt model assumptions and involve extrapolation, which, while an inherent unavoidable feature of prediction, is well known to be unwise in other situations (see (e) above). Schnaars (1989) discusses many business and economic forecasts that have gone wrong in the past and suggests ways of avoiding similar mistakes in the future. One recommendation for long-term forecasting is to avoid giving a single point forecast, or even a single prediction interval, as this will depend on a single model and hence on a single set of assumptions. Instead a range of forecasts, based on different assumptions, is advisable (called multiple scenarios in the trade!). A similar point was made much earlier by Schumacher (1973) when he suggested long-term feasibility studies rather than "presumptious" long-term forecasts. Another suggestion for avoiding over-optimistic forecasts, supported by empirical results rather than theory, is to damp the trend as proposed by Gardner and McKenzie (1989). The main point is not to be blinded by a sophisticated model. An assumption is still an assumption, and extrapolation is still extrapolation!

3.4 Communicating the Results

After analyzing the data, the statistician faces the tricky task of interpreting the results and then communicating the conclusions to interested parties. This usually involves writing a report. The outcome of the project may well be judged by what is written rather than by what has actually been done. This is a disturbing thought given that many people have difficulty in expressing themselves in writing. Thus another potential pitfall is to write an inadequate, incomplete or incomprehensible report.

Many publications (e.g., Ehrenberg, 1982; Gowers, 1986) have given advice on report-writing. Here I simply make the following points in brief: (a) Write simple clear English in short sentences. (b) If you have trouble getting started, jot down all the points you wish to make, not necessarily in the right order. Revising a draft, however sketchy, is much easier than writing the very first version. (c) Do not assume your reader knows more than he actually does. (d) Give extra attention to the presentation of graphs and tables. Don't just include undigested computer output, which may be in a quite unsuitable form. Don't be afraid to use correcting fluid and hand-writing to modify computer output if necessary, although it is better to use software that allows the user to control the output as required. (e) Revise the report several times. If possible get someone else to read it and make comments. Given these simple guidelines, there is no reason why anyone should fail to produce an adequate report, however inadequate they may feel.

Occasionally you may have to summarize your results with an oral presentation. Many people find this prospect even more terrifying than writing a report, giving another way that a project may be sabotaged at a late stage. Some sort of visual aid is usually essential and the overhead projector is most commonly used. If you have not used one before, then some practice is necessary. A common mistake is to write too much material on one transparency so that it becomes visually off-putting or even illegible. Word processors now allow the easy production of good quality printed transparencies. Practice the whole lecture, including the use of visual aids, with an audience of at least one person. Time this practice session so as to ensure that you take no more and no less than the time allotted to you.

4. EXAMPLES

Through mistakes we can learn the truth [Ancient Chinese proverb]

This section presents some additional examples to illustrate practical problems and give some indication as to how these may be avoided or overcome. Most published examples are sanitized or oversimplified versions of the "real thing" and so present a false picture of real-life statistics. Textbook examples are typically small-scale and selected primarily to illustrate a particular technique, while examples in journals tend to be too brief to be helpful. The iterative nature of real statistical analysis often fails to be evident. In particular, authors customarily avoid mentioning mistakes, blind alleys and other problems, partly to save space (perhaps at the insistence of an editor), partly to avoid the author's blushes and partly because "the reader will not be interested." In fact, I suspect the reader would often be interested to hear more about problems and how they were overcome, because we learn more from mistakes (our own and other people's) than from straightforward successes. As Greenfield (1987) says, many stories about consultation are "tales of woe. But these are where the lessons are to be learned." Thus my examples are not only honest, but also inevitably somewhat personal or anecdotal in nature. I hope the reader will regard this as a plus rather than a minus.

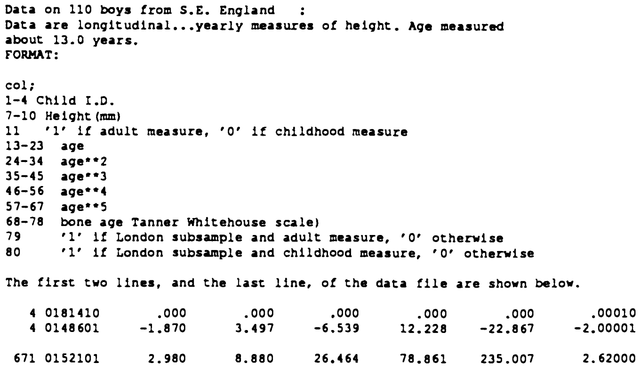
Examples 5 and 6 demonstrate the importance of asking questions and clarifying objectives. Example 7 highlights the dangers of believing formulas published in the literature, while Example 8 considers the Challenger space shuttle disaster.

Further examples of real-life problems can be found for example in Cox and Snell (1981), Anderson and Loynes (1987) and Chatfield (1988), although these references do not specifically concentrate on avoiding trouble. The data sets in Andrews and Herzberg (1985) were deliberately chosen to exclude examples where there was a complete or obvious approach.

EXAMPLE 5. Getting Background Information. Tackling a statistical problem "at a distance" can be very difficult. Direct contact with the people involved in formulating the problem and collecting the relevant data is clearly desirable. This example presents an extreme, and arguably unwise, situation where the statistician (me!) was very detached from the problem. Nevertheless it has some instructive features. The 2nd Anglo-French Data-Analysis Workshop was held in September 1988 primarily to compare the English and French approaches to data analysis. All delegates, including myself, were sent four sets of data (by e-mail), together with some background information and asked to analyze at least one of them. When I came to look at the data sets, I soon realized that I had inadequate information to satisfactorily analyze any of them. This supports my view that the most important question in data analysis is usually not "What techniques should be used here?"; rather the analyst should ask questions such as: (a) What are the objectives? (b) What background information is needed to tackle the problem? (c) What background knowledge is already available (e.g., from previous data sets, or background theory)?

To illustrate the point, I focus attention on the data-set concerned with the relationship between childrens' height and age. The only information provided was that shown in Figure 2.

Three illustrative lines of data are shown and the rest of the data are available from the author by e-mail. The reader will notice that Figure 2 does not even include an objective. Before I could commence the analysis, I therefore wrote to the workshop organizer with a list of questions including: (1) What is the objective of the analysis? (2) What is the indicator variable in column 12? (3) The information in columns 11, 79 and 80 doesn't seem



```
Data on 110 boys from S.E. England   :
Data are longitudinal...yearly measures of height. Age measured
about 13.0 years.
FORMAT:

col;
1-4 Child I.D.
7-10 Height (mm)
11   '1' if adult measure, '0' if childhood measure
13-23  age
24-34  age**2
35-45  age**3
46-56  age**4
57-67  age**5
68-78  bone age Tanner Whitehouse scale)
79     '1' if London subsample and adult measure, '0' otherwise
80     '1' if London subsample and childhood measure, '0' otherwise

The first two lines, and the last line, of the data file are shown below.

   4 0181410        .000        .000        .000        .000        .000      .00010
   4 0148601      -1.870       3.497      -6.539      12.228     -22.867    -2.00001

 671 0152101       2.980       8.880      26.464      78.861     235.007     2.62000
```

FIG. 2. *The background information supplied with the height/age data.*

to make sense. For example, child 4 in line 1 is an adult but his age is 13 years. (4) There must be lots of background knowledge about the relationship between height and age. Surely we shouldn't be starting from scratch?

I subsequently received the following replies from the person who provided the data set: (1) The objectives of the analysis are several. Of particular interest is the prediction of adult (or any other age) height. One could also study the relationship between height and bone age, or between bone age and age, etc. Auxologists are interested in things such as the age of maximum height velocity. (2) Ignore column 12. (3) The data are longitudinal with the adult measure coming first. (4) There is indeed a lot known about height growth (e.g., Jolicoeur, Pontier, Pernier and Sempe, 1988).

An offer of further help was also made. This was potentially necessary as I did not know what an "auxologist" was, and I did not understand the answer to another question asking what the "bone age Tanner Whitehouse scale" is. I wondered why powers of age and column 12 were not removed from the data before circulation to delegates, while the reply to point 3 made me realize that, for adults, zero age is to be regarded as a missing value. (In fact, the adult measures do not always come first in the data.)

I was now in somewhat better shape to analyze the data. Even so, with several outstanding queries and little experience in this area, I had little idea as to what sort of formal analysis to apply. I therefore contented myself with plotting (height/adult height) against age to get a series of S-shaped curves for different children. This gave me some feel for the data. At the workshop, one delegate who was expert in the given subject area made a much more satisfying job of analyzing the data, thus confirming the need for specialist knowledge.

I also raised various queries about the other three data sets but found them equally difficult to analyze satisfactorily. One particular problem arose for some Australian migration data, where the population figures were stated to be in units of 1,000. In fact reference to an atlas revealed that the units should be 10,000s. The moral of this last point is that you should not necessarily believe what you are told.

My efforts in overcoming problems here can only be regarded as partially successful, but the experience re-emphasized (a) the importance of having direct contact with data and (b) the importance of asking questions. As regards (a), I am reminded of a Biometrics Society data workshop where several people analyzed some measurements on apple trees. Only after all the analyses had been presented did the data-author reveal that the orchard was actually in two parts with one containing older trees than the other. This late divulgence of crucial information caused some annoyance! As to (b), I was surprised to learn from the workshop organizer that no other delegate had raised any queries. Is this because other delegates did not take the analyses seriously, or does it indicate a real reluctance to ask searching questions?

EXAMPLE 6. What is the Problem? This example demonstrates how easily the analyst can "go off in the wrong direction." Plain warp-knitted fabrics are made from dyed yarn in a standard range of about 30 shades. The dye recipes were mostly mixtures of red, yellow and blue dyes, and there were about half a dozen dyestuffs of each hue in use.

Factory staff had observed a large, consistent variation in fault-rates between shades. They attributed this to one or more rogue dyestuffs that were weakening the fiber and making it less able to withstand the violent treatment that the yarn undergoes on the knitting machines. An investigation was therefore undertaken to identify the rogue dyestuff(s) that was causing the increased fault rate.

One statistician performed a multiple-regression analysis with fault rate as the dependent variable and the dyestuffs as the explanatory variables in various forms. All that emerged was that blue dyestuffs as a group were good (associated with low fault rates), while reds were bad and yellows were intermediate.

A second statistician then decided to investigate a completely different hypothesis, namely that high fault rates were associated with the darkness of the color rather than with a particular rogue dyestuff. He made paired comparisons between each shade and every other shade, using shade cards, scoring from +2 if much darker to −2 if much lighter. Summing the scores for each shade gave an overall assessment of darkness for each shade. Then a plot of fault rate against darkness score gave a very strong negative correlation. The reason for this could now readily be worked out. The assessment of fault rate is subjective. It is a fault if you can see it! A severity of damage that might be obvious in a light shade could well pass unnoticed in a dark shade. Thus there was actually no reason to suppose there were any real differences in fault rates or that there were any rogue dyestuffs. The dyestuffs only entered circumstantially in that blue dyestuffs, for example, were used predominantly in dark shades.

While the analyst should understand the relevant technology and pay due attention to what the client knows or believes about the system, the statistician must be careful not to let the analysis be constrained by the client's prejudices. The actual problem may be quite different than the one that is posed.

Example 7. Incorrect Formulas. An important type of mistake arises from getting mathematical formulas wrong. Such mistakes may result from a typographical error, or from copying somebody else's mistake, or from some more fundamental misunderstanding. Typographical errors necessitate eternal vigilance both in checking your own manuscripts (errors are your fault and not your secretary's!) and in not necessarily believing formulas in other people's work. You should: check the *dimensions* of a formula (Edwards and Hamson, 1989, page 60); check the result given by the formula in a (simple) special case where you know the correct answer; check formulas by referring to standard texts or to an experienced statistician; and check *limiting behavior* where appropriate (Edwards and Hamson, 1989, page 74).

Virtually every textbook contains at least one misprint and it would be invidious to pick one out as an example, especially as they are not always the fault of the author. The first galley proofs will typically contain hundreds of errors and even a 99% success rate in spotting them will still leave several errors. In any case, the publishers/printers may fail to correct all the marked errors or may introduce new errors while correcting the old. Nevertheless some textbooks undoubtedly contain more misprints than could reasonably be expected under any system.

(a) I recently reviewed a text which gave the linear regression model relating a response variable, y, to an 'independent' variable, x, as

$$y = b_0 + b_1 + \varepsilon.$$

It repeated the omission of the x variable and then produced incorrect formulae for the estimates of b_0 and b_1. The poor novice reader will have difficulty avoiding trouble with this guidance!

It would also be invidious to pick out specimen examples of mistakes in published papers, since we are all prone to typographical errors. However I will give two more examples, which primarily illustrate the potential dangers of copying.

(b) In time-series forecasting, the construction of prediction intervals is very important. Makridakis, Hibon, Lusk and Belhadjali (1987, Equation 1 and the Appendix) purport to show that the variance of the k-steps-ahead error is equal to k times the variance of the one-step-ahead error under certain assumptions, notably that the forecasting method is optimal so that one-step-ahead errors are independent with constant variance. In fact the above result, while plausible at first sight, holds only when the underlying process is a random walk and the apparent "proof" is unsound. What is really worrying is that this incorrect formula has already been cited in several subsequent publications (e.g., Lefrancois, 1989) and seems set fair to become part of time-series folklore.

(c) This cautionary tale concerns stock control. Let Y, L, D denote the demand rate per day, the lead time, and the lead time demand respectively. Then a well-known result, usually ascribed to Clark (1957), is that Var(D) equals $\mu_L \sigma_Y^2 + \mu_Y^2 \sigma_L^2$ in an obvious notation. A correspondent sent me a paper (Parker, 1986) that pointed out that if $D = YL$ then the dimensions of the above formula are not correct, and the alternative well-known formula for the variance of a product should be used. I thought this would be an excellent example for me to use until I realized that Clark's result was meant to refer to the random sum of L independent random variables, namely $D = Y_1 + Y_2 + \cdots + Y_L$ and that the variance formula is in fact correct. The dimension counter-argument is invalid as L, the integer number of days, is a dimensionless quantity. Since Parker (1986) criticizes "clever chaps who write in journals," this is a case of the biter bit and I nearly joined him. Looking at Clark's (1957) proof, and at subsequent papers that misleadingly describe $\{Y_i\}$ as sales rates, the confusion is understandable as the notation and assumptions are unclear. The proof is unnecessary anyway as the result is well known in the distribution literature when expressed as a random sum. The two lessons of this last story are: (1) Be clear what is assumed in any proof; (2) an apparent mistake may not be a mistake.

Example 8. The Space Shuttle Catastrophe. The catastrophic accident to the space shuttle Challenger in 1986 was caused by a combustion gas leak through a joint in one of the booster rockets, sealed by a device called an O-ring. It was subsequently realized that O-rings do not seal properly at low temperatures. The night before the fatal launch, a 3-hour meeting reviewed concern about the effect of low temperature on O-ring performance given that the forecast launch temperature was much lower than on previous occasions. Figure 3a shows the data available at the meeting and plots the number of O-rings showing thermal distress against temperature at different past launches. The meeting concluded that there was no evidence of a temperature effect and the launch went ahead with fatal results.

This simple analysis can be faulted in several ways. The flights giving zero incidents were omitted from the graph because these flights were (wrongly) thought not to contribute any information about the temperature effect. In fact, a glance at Figure 3b, which includes the omitted data, suggests that the latter *do* contribute extra information. Dalal, Fowlkes and Hoadley (1989) go on to estimate the probability of catastrophic field joint failure at 31°F and show that it is indeed much larger than at the higher temperatures of previous launches. This latter finding depends on fitting a logistic regression model. In fact, my first assessment of Figure 3b noted the absence of data under 50°F and the consequential extrapolation involved in assessing what will happen at 31°F. It may be better to say that there does seem to be a temperature effect but that any assessment of its effect at 31°F would be dangerous when human life is involved.

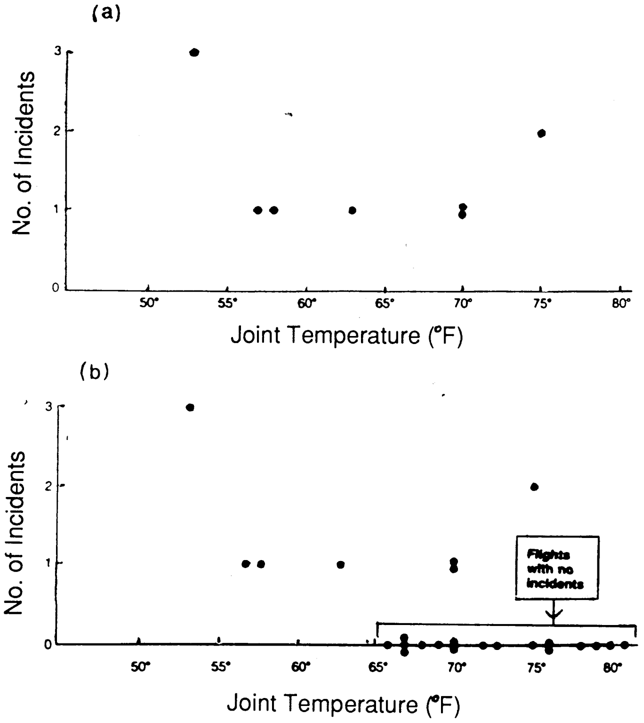


FIG. 3. *(a) The number of O-rings showing some thermal distress plotted against the temperature of the field joint at various launches prior to* 1986. *The lower graph (b) includes flights with no incidents.*

The obvious morals of this story are that any analysis should use *all* the data (and especially *not* a nonrandom sample), that a clear statement is needed when a model is extended outside the region over which it has been fitted and that extrapolation is always dangerous.

5. DISCUSSION

The two main themes of this article are that: (1) *understanding strategy* is as important as *knowing techniques* and (2) *avoiding trouble* is complementary to, and a prequisite for, *achieving optimality*, and therefore deserves at least equal attention.

Important guidelines for avoiding trouble include the following. (a) Clarify objectives and background by asking questions. (b) Ensure that "good" data are collected and that they are processed satisfactorily. (c) If data have already been collected, find out how. Check their quality carefully. Is there enough data to answer the questions satisfactorily? Have any data been removed? (d) IDA is important in any analysis, particularly for avoiding trouble. (e) Don't trust published formulas to be necessarily correct. (f) Don't be afraid to ask for help and advice.

Despite all our best endeavors, wrong results may still be published. Other people's analyses should be assessed both technically (e.g., Are data adequate for what is needed? Has the right technique been used?), and also on broader questions such as "Why was the research carried out?" and "Have the objectives been fulfilled?". Statistical analyses are based on assumptions and choices that are often implicit, subjective and arbitrary, and it is essential to find out exactly what has been done. It is also important to check results against intuition. Disagreement may reveal (a) mathematical or arithmetical mistakes, (b) implausible or oversimplified model assumptions, (c) inadequate intuition or (more rarely!) (d) exciting new findings.

ACKNOWLEDGMENTS

I thank Tony Greenfield for Example 2, Tony Hayter for Example 3(c), Rosemary Bailey for Examples 4(a) to 4(d), Eric Whyte for Example 6 and Derek Bissell for the quality control example in Section 3.3(a). I also thank Donald Preece and a referee for detailed comments on a draft version, and I thank several other correspondents for drawing my attention to appropriate references and making helpful comments.

REFERENCES

Andersen, B. (1990). *Methodological Errors in Medical Research*. Blackwell Scientific, Oxford.

Anderson, C. W. and Loynes, R. M. (1987). *The Teaching of Practical Statistics*. Wiley, Chichester.

Andrews, D. F. and Herzberg, A. M. (1985). *Data*. Springer, New York.

Barnett, V. (1987). Straight consulting. In *The Statistical Consultant in Action* (D. J. Hand and B. S. Everitt, eds.) 26-41. Cambridge Univ. Press.

Begg, C. B. and Berlin, J. A. (1988). Publication bias: A problem in interpreting medical data. *J. Roy. Statist. Soc. Ser. A* **151** 419-463.

Belsley, D. A. and Welch, R. E. (1988). Comment on "Combining robust and traditional least squares methods: A critical evaluation" by M. A. Janson. *Journal of Business and Economic Statistics* **6** 442-447.

Bissell, A. F. (1989). Interpreting mean squares in saturated fractional designs. *Journal of Applied Statistics* **16** 7-18.

Brown, R. G. (1967). *Decision Rules for Inventory Management*. Holt, Rinehart and Winston, New York.

Chatfield, C. (1986). Exploratory data analysis. *European J. Oper. Res.* **23** 5-13.

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

Chatfield, C. and Schimek, M. G. (1987). An example of model-formulation using IDA. *The Statistician* **36** 357-363.

Clark, C. E. (1957). Mathematical analysis of an inventory case. *Oper. Res.* **5** 627-643.

Cochran, W. G. (1983). *Planning and Analysis of Observational Studies*. Wiley, New York.

Cox, D. R. and Snell, E. J. (1981). *Applied Statistics*. Chapman and Hall, London.

Dalal, S. R., Fowlkes, E. B. and Hoadley, B. (1989). Risk analysis of the space shuttle: Pre-Challenger prediction of failure. *J. Amer. Statist. Assoc.* **84** 945-957.

Edwards, D. and Hamson, M. (1989). *Guide to Mathematical Modelling*. MacMillan, London.

Ehrenberg, A. S. C. (1982). Writing technical papers or reports. *Amer. Statist.* **36** 326-329.

Gardner, E. S., Jr., and Mckenzie, E. (1989). Seasonal exponential smoothing with damped trends. *Management Sci.* **35** 372-376.

Gilchrist, W. (1984). *Statistical Modelling*. Wiley, Chichester.

Gore, S. M., Jones, I. G. and Rytter, E. C. (1977). Misuse of statistical methods: Critical assessment of articles in BMJ from January to March 1976. *British Medical Jounal* **1** 85-87.

Gowers, E. (1986). *The Complete Plain Words*, 3rd ed. Pelican, Harmondsworth.

Greenfield, A. A. (1987). Consultant's cameos: A chapter of encounters. In *The Statistical Consultant in Action* (D. J. Hand and B. S. Everitt, eds.) 11-25. Cambridge Univ. Press.

Hand, D. J. (1990). Emergent themes in statistical expert systems. In *Knowledge, Data and Computer-Assisted Decisions* (M. Schrader and W. Gaul, eds.) 279-288. Springer, Berlin.

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

Hooke, R. (1983). *How to Tell the Liars from the Statisticians*. Dekker, New York.

Huff, D. (1954). *How to Lie with Statistics*. Penguin, Harmondsworth.

Hurlbert, S. H. (1984). Pseudoreplication and the design of ecological field experiments. *Ecological Monographs* **54** 187-211.

Janson, M. A. (1988). Combining robust and traditional least squares methods: A critical evaluation (with discussion). *Journal of Business and Ecomomic Statistics* **6** 415-451.

Jolicoeur, P., Pontier, J., Pernier, M-O. and Sempe, M. (1988). A lifetime asymptotic growth curve for human height. *Biometrics* **44** 995-1003.

Kimball, A. W. (1957). Errors of the third kind in statistical consulting. *J. Amer. Statist. Assoc.* **52** 133-142.

Lefrancois, P. (1989). Confidence intervals for non-stationary forecast errors: Some empirical results for the series in the M-competition. *International Journal of Forecasting* **5** 553-557.

Makridakis, S., Hibon, M., Lusk, E. and Belhadjali, M. (1987). Confidence intervals: An empirical investigation of the series in the M-competition. *International Journal of Forecasting* **3** 489-508.

Moser, C. A. and Kalton, G. (1971). *Survey Methods in Social Investigation*, 2nd ed. Heinemann, London.

O'Muircheartaigh, C. A. (1977). Proximum designs for crude sampling frames. *Bull. Inst. Internat. Statist.* **47** 82-100.

Parker, J. B. (1986). A story about statistics. *Math. Sci.* **11** 73.

Preece, D. A. (1981). Distributions of final digits in data. *The Statistician* **30** 31-60.

Preece, D. A. (1987). Good statistical practice. *The Statistician* **36** 397-408.

Reichmann, W. J. (1964). *Use and Abuse of Statistics*. Pelican, Harmondsworth.

Schnaars, S. P. (1989). *Megamistakes: Forecasting and the Myth of Rapid Technological Change*. Free Press, New York.

Schumacher, E. F. (1973). *Small is Beautiful*. Sphere, London.

Schuman, H. and Presser, S. (1981). *Questions and Answers in Attitude Surveys*. Academic, New York.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley, Reading, Mass.