The Dire Dangers of Dabbling with Data

Neil S Barnett - Melbourne, Australia

1. Introduction

In perusing the report of The Second International Conference on Teaching Statistics held in Canada in 1986, I was "taken" with a number of comments made by invited and contributing speakers alike. "Taken" in the sense that they "echo" my own experience, having had the opportunity, over twenty-two years, to teach mathematics and statistics at high school, college and university to specialists, non-specialists and industrial groups alike. In particular, my namesake, Professor Vic Barnett, was quoted as "... seeing a consulting service as a vital component of a university or college department of statistics providing, amongst other things, ideas for research and teaching ...". This talk, and another that I've prepared to give later in the conference, use material gleaned from my own consulting experience.

Many books on statistical methods concentrate on theoretical considerations basing development on probabilistic models. Other books purport to cut through theoretical "mumbo-jumbo" and get down to usable techniques. The trouble with this latter approach is the tendency to produce what is little more than a set of recipes. Still other manuscripts take techniques out of their original context and demonstrate their use in areas where there is a growing interest and demand. There is no doubt "our" subject is today much in demand and this itself presents its own peculiar set of problems.

Much of the written material on statistical methods proceeds as if the "world of data" is normal and literally filled with random samples! In reality, however, non-normality is common place and samples seldom random. Discussion of robustness and simulations to demonstrate its existence can, to some extent, de-fuse this as a major issue. However, the term "robustness", as commonly used, is a property of a statistical procedure and as such deals with distributional assumptions of the data or lack of them. A lot of effort has gone into modifying various statistical procedures so as to handle sets of data when there are missing data values and of course there are various types of common and not so common sampling schemes. Despite all this accumulated "know how" statistically things can still go wrong! In fact things can go wrong much more fundamentally than with technique and with choosing the appropriate sampling scheme. Even getting these right the data available for statistical analysis can be grossly misleading and from dubious data one can only ever draw dubious conclusions. I'll illustrate with a number of examples that highlight, amongst other things, the importance of the human factor. It is this human factor that our teaching needs to stress more if we are to promulgate effectiveness in the use of statistical techniques.

2. The nature of the data

A colleague of mine was relating, recently, a situation he knew of, involving a statistically designed experiment being conducted in an industrial plant. The statistician, having stipulated the design and factors to be varied, decided to go onto the plant floor to see how things were going. To his "horror" he found that rather than following the stipulated procedure involving several machines and fittings, workers were merely moving around the signs that had been made to indicate the status of each machine!

As another example, I recall once discussing with an industrial group the statistical analysis to be used in order to process some plant data. Amongst the group was the individual responsible for collecting the data. He commented, during my discussion, that the way he collected his sample seriously flavoured the test result. At the time of sampling the product was being conveyed on an open conveyor having first been supposedly well mixed in a previous stage of the process. Depending from which side of the conveyor he drew his sample he could get quite discrepant results.

Yet another case involved the laboratory testing of a fluffy sample of product produced in a certain chemical process. This "fluffiness" made chemical analysis impossible without first compressing the sample. This compression was originally done by hand - depending on the compression exerted a different end test result was obtained.

The routine sampling of material produced in a furnace or reactor can be quite a difficult task. The usual aim in such samplings is to obtain a sample of product supposedly recently produced - such sample values can then help with process monitoring. One example I recall involved sampling from the heart of a furnace with a long steel tube. Investigation showed that problems of contamination were common and that samples supposedly obtained of recently produced material contained material produced several hours previously.

If a statistician asks for data he'll generally get it - one way or the other. I hope these and following examples are sufficient to illustrate the importance of the practical statistician concerning himself with data quality. His job doesn't start with the statistical analysis, with establishing test hypotheses, with specifying an appropriate experimental design, it starts with the people who collect the samples, who measure the items, who generate the data through testing. Do they understand what is required, why, and the value of it, have they got fail-safe equipment, have they been adequately trained to use the equipment they have, are they aware of the importance of environmental factors, has their equipment been recently calibrated? Perhaps most importantly, have their comments been sought on how to ensure the provision of reliable data? It is my experience that it is in these areas where the statistical machinery often comes unstuck.

Statistical process control has become a boom industry of late and there is much educational and consultancy work to be done in this area. Control charts are tools of trade and pre-occupation with process variability the norm. Much teaching material, however, is presented as if data quality were not an issue. Sometimes data is a simple measurement for which we would expect little error; however, in a chemical process, data generation is the result of regular chemical analysis of samples. Now one doesn't need to be a chemist to realise that in a chemical analysis all manner of things can occur that affect data quality. Typically, a number of separate procedures need to be followed that are subject to significant human or equipment variability. Environment, temperature, humidity etc. can also be issues. The end result can be a data value that exhibits a considerable amount of variation between testers, between environments and between equipment. I have known error variability to swamp process variability - the very thing process sampling and testing were designed to monitor. When it comes to final product specifications it is recommended by various authorities that a measurement device or system is qualified for use if specification range ≥ 5 × (testing error range).

When there are such problems the statistician needs first to examine ways to reduce data variability before he can afford the luxury of pontificating on ways to reduce process variability - one of the main objectives of quality management. The time honoured ANOVA can be useful for examining tester, equipment and laboratory consistency but even here attempts to get "good" data can come dramatically unstuck. To cite one example, in some inter-laboratory tests that I was involved with, results showed a consistency never before experienced, with an accuracy exceeding the capacity of the equipment!

A recommended practice in statistical process control is the conducting of capability runs designed to pick up the degree of natural variation in a process when external interference is kept to a minimum. A common comment that I've heard is that process performance under these conditions is far better than under normal operating conditions, so what value can one place on the results?

In many situations there is a large human component that impinges on the statistical problem, or more accurately, on the data collection on which the statistical analysis depends. It all comes down to the simple matter that if the people involved don't support, are misinformed, are not taught or have inadequate facilities, the whole statistical exercise is likely to come unstuck. A practical statistician cannot be a remote boffin juggling with his formulae and making a fetish out of analytical rigour, but rather someone who is people oriented and able to relate to the very real difficulties associated with obtaining reliable data.

With non-measurement data there is even more potential for data to be unreliable. In survey work, when asking fairly involved questions with a number of acceptable responses, none of which seem to quite fit, the respondents can easily slip into giving ill-considered answers, especially if the questionnaire has been fairly lengthy already - I've done it, haven't you? Have you seen the questionnaire that forms the basis of the Australian radio and television ratings? ... But that's another story.

Data can be suspect for a host of human reasons. On one occasion when analysing some laboratory test results I discovered some values having less figure accuracy than I was led to expect. On investigation I found that one tester, because of insufficient information, was rounding his test results. Undeterred I set about trying to "dig-out" his uncorrected results from the original recordings. I discovered a 20% error rate in his rounding!

Other data problems can occur if the statistician doesn't familiarise himself with the detail of the situation he is dealing with. I remember once requesting samples to be taken quarter hourly only to find that a sample took typically twenty minutes to collect!

I would have been given the data too if it hadn't been for one concerned individual.

Emphasis on the importance of data is very necessary when teaching techniques. It has become fashionable in recent years, with the quality push hitting industry, for statistical training to occur "in-house".. This features the statistical trainer coming into the company rather than the company sending employees out on specified training courses. It has been my experience that such courses are often given with an insensitivity to the requirements and limitations of the audience. It is easy to assume that abstractions, analyses and situation similarities will automatically help in getting the message across. One quality manager (in a company producing multi-purpose synthetic fibres) recently lamented to me that a statistical trainer they had through the company was illustrating his techniques by data on bags of flour. One may be able to get away with this in the classroom with certain audiences but not in "in-house" industrial training. Most present only want to see how these "new-fangled" techniques relate to their job, their environment, their data and not to the effective working of a flour mill! In such situations it's better to discuss techniques in the context of data obtained from the plant itself, openly discussing likely problems with the data itself and soliciting opinion on its reliability and ways to improve it.

Without this awareness of the nature of the data on which techniques are to be used, some misleading information can be imparted. In continuous processes, generated data can be auto-correlated; the usual discrete item, independent samples statistical approach is then inappropriate without some qualification, or is even totally inappropriate, depending on circumstances. There is scope to use simulation to illustrate the effects of sample size on different methods of estimation but even this should be done using simulated data with parameter values akin to those obtained in reality.

3. Conclusions

So what is the punch-line, what's the message I'm wanting to convey? In simple point form:

(i) The statistician should first examine available data carefully before embarking on any analysis.

(ii) The industrial statistician should aim to make contact with those collecting or generating the data, to perceive their capacity, both through training and through available equipment, to deliver "clean" data.

(iii) Care should be taken to explain to these "key links" in the statistical chain why the analysis is desirable or necessary. If there are ultimate benefits to them, let them know.

(iv) Don't automatically assume that the data is clean; it won't be unless effort is expended to make it so.

(v) Don't teach techniques without knowing they are appropriate for the purpose to which they are to be put.

(vi) Finally, let's teach these seldom conveyed aspects of "statistics" to our students in the classroom so that they will go out to industry with a more realistic approach to problem solving and analysis.