GETTING THE MOST OUT OF STATISTICS

BU-631-M by November, 1977

S. R. Searle
Biometrics Unit, Cornell University, Ithaca, New York

Abstract

Support for collecting data is more easily aroused than for analyzing them. Research workers can maximize their benefits from statistical analysis of their data by having statisticians as colleagues in their work and by using their advice for the statistical part of it. But to do so, both research worker and statistician need to understand and appreciate the role of statistical consulting.

1. COLLECTING DATA

In almost any branch of scientific endeavor, collecting data is relatively easy. It may be expensive, but generally speaking there is usually not much difficulty in acquiring funds to pay for data collection. And having collected data, we most times like to squirrel it away. How many of you are like this? How many of you have data sheets stashed away here and there with data on them that have not been fully analyzed? Certainly this sort of thing occurs in universities, and I'm sure it does in industrial corporations also.

Why is this? Perhaps it's because collecting data is fun, it requires considerable skill, effort, time and people, and the outward appearance to others is obviously one of being busy at important tasks. In contrast, once data

---

Paper No. BU-631-M in the Biometrics Unit Mimeo Series.

have been collected, subjecting them to thorough statistical analysis is largely a matter of desk work, of quiet but hard thinking, of doing mathematics and arithmetic (which so many people hate) at various levels of sophistication – and finally, of writing reports and scientific papers, which many people also hate doing. All this seems so unglamorous, compared to buying expensive laboratory equipment, setting it up and hiring people to use it, or sending out thousands of survey questionnaires, or travelling to foreign lands to interview people, or plowing fields and growing crops. Furthermore, on application to the right sources of funding, money is usually forthcoming for these sorts of activities – and time is readily made available for carrying them out. But try adding to your dollar and time budgets $10,000 for computing and statistical advice and 3 months of your own time for analyzing and writing up the conclusions from your two-year data collection project – and see what sort of reception you get. In addition, note your own displeasure at having to be out of your lab for three months while you do this work.

In contrast to this ideal of having adequate time for analyzing data, many scientists are impatient with this part of their work. They take a cursory look at the data, draw some "obvious" conclusions and plan their next experiment. There is nothing wrong with this, they are indeed following the scientific method, but my contention is that in most cases nowhere near the full information available from present data has been deduced before planning collection of future data – i.e., of doing the next experiment or making the next survey.

2. STATISTICAL ANALYSIS

A contributing reason for this state of affairs is, I believe, a feeling that statistical analysis is either very easy or impossibly difficult. Either way it need not occupy much time. Maybe it is not a feeling that statistical analysis is easy but, at least in most cases of useful analysis, that it is standard, well-

defined, straightforward, and essentially no trouble at all; or that the proper analysis is so difficult that "we'll just look at the averages"! This misapprehension is, I fear, followed at great cost to scientific progress. Certainly, there are many standard statistical analyses that are well-known, well-understood and consequently easy to use. Mostly they are the analyses taught in standard statistical methods courses taken by budding research workers in their first year of graduate study. By their very nature such courses cover only a limited amount of statistical methodology because they are, after all, beginning courses, they are hard to teach and therefore often poorly taught. Furthermore, most students taking such courses are there with great reluctance, under duress almost, their prime interest being to satisfy degree requirements and maybe to "learn enough statistics to get by".

The statistical methods and ideas taught in these courses are the fundamental and basic ones. As such, they involve fairly simple arithmetic that is geared to desk calculators. But what is happening today in the real world, in the data-collecting aspect of things? The computer revolution is upon us. The impact is that we can do vast amounts of arithmetic that were never dreamed of as feasible on desk calculators; e.g., solving 100 equations in 100 unknowns. More than that, with computers and other sophisticated electronic gear we can collect and store enormous quantities of data – all at fantastic speeds; for example, physiological data from laboratory animals wired directly to computers. What then?

True, the same computer that provides storage of data can also do the arithmetic of the statistical analysis – and do it very quickly. The question is "what analysis?". This is the heart of the problem. The planning of small experiments and the analysis of them does not always carry over in simple and obvious ways to large experiments and large sets of data. There may be little to doubt about the

analysis of variance of 24 observations from a randomized complete blocks experiment of 6 treatments in 4 blocks, whereas for analyzing some 7000 observations ($7392 = 11 \times 7 \times 4 \times 4 \times 3 \times 2$) arising from 31 levels of 6 factors and having 25 lines in the analysis of variance, the appropriateness and validity of such an analysis may well be open to question. Furthermore, with large data sets such as this there is no one, universally accepted, solely right way of analyzing the data. There are usually many ways of looking at any extensive set of data, and it is often good practice to take time and energy to investigate several of them. Some analyses will confirm and reinforce others, almost all may give some information that others may not.

Regression is a good example of a statistical method that is easy to understand in small cases, but often difficult with large data sets. Studying the relationship between the weight and height of men is straightforward. But analyzing 5000 observations from an economic survey might be very difficult: e.g., household expenditure on food related to income, age of husband, and of wife, number of children, education of husband, and of wife, and maybe 10 or 20 other such factors that have been enquired about in the survey. The arithmetic, with today's computers, is achieved almost as quickly as for the simple set of data – but the interpretation, the different forms of the analysis available and their appropriateness will be much more difficult to decide upon.

3. CONSULTING WITH STATISTICIANS

Today we still sadly hear the phrase "do a computer analysis". This is nonsense. No computer does an analysis – it only does the arithmetic of the analysis that we decide upon. In this connection I like the story of the person who left a box of cards at a computing center and two weeks later came back for the results.

There were none, and the person was mortified. "I thought the computer could do it", he said. "Yes, but what analysis did you want?", asked the computing director. "Oh, I thought the computer would decide that." So the computer, while at the same time as it has greatly aided our arithmetic problems, has also accentuated the need for experimenters to know how to make good use of statistics.

There are at least two ways of doing this. One is to be well-qualified as a statistician – but this is just about a full-time job in itself. Another way is to understand a few of the things a statistician deals with, and as a result of this know when to consult with a statistician, both for the planning of the data collection and for the analyzing of the data once obtained. Research workers can maximize their benefits from statistical analysis of their data by having statisticians as colleagues in their work and using them as consultants for the statistical parts of it. But to do so, both they and the statistician need to understand and appreciate the role of statistical consulting.

Some of the basic statistical ideas that a researcher needs to appreciate are things like population, random sample, estimation, replication, control of variability, experimental design and regression, and the more technical tools like hypothesis testing, confidence intervals and, of course, others, the more the better. Since less rather than more is usually the case, it might be thought that "a little knowledge is a dangerous thing"; but it is so, only if one attempts to put it into practice. This danger can be avoided, or at least lessened, by utilizing the help of a statistician.

Thus it is that consulting is the focal point for getting the most out of statistics. From the point of view of the statistician, consulting is the raison d'être of statistics because the ultimate objective of statistics is methodology for using data to help solve real-life problems. More specifically, as one

statistician has put it, the object of statistics is "to use mathematical theory of probability to determine what conclusions can be drawn with what confidence, and in determining the amount and type of data needed as evidence for these conclusions." Notice here, the dual role of the statistician, namely that of both planning how to get data and analyzing them once obtained. Both aspects of this role are important. Almost always, data that are obtained as the result of good statistical planning are easier to analyze, have analyses that are more readily interpreted and provide more reliable information than data gathered without such planning. And what is particularly valuable is that part of the planning procedure involves planning the analysis. Then, as soon as the data are available, the analysis can be made. No debate need occur, as to what the analysis should be. This aspect of planning is very, very valuable. Without it, as so often is the case, voluminous data are collected, at considerable expense, whereupon there is great embarrassment and guilt of the kind "what can I do with all these numbers?; they must be useful for something." These feelings are aggravated by the inherent pressure of the fact that "now, after 2 years, we've got all our data and the project is over." The implicit feeling is that, with the data in, the job is done – well, maybe not done, but almost so. "We just have to do the analysis", and with that remark goes the implication that that will take only a day or two. In point of fact, for purposes of getting information from the project, the job has only just begun. If the analyses were planned at the outset, they may not take too long to do. But if analyses now have to be planned, there may be quite a long time until completion.

One requirement for being successful in consulting with a statistician is to have a good problem, well formulated. Saying that the problem needs to be well formulated does not necessarily mean well formulated statistically. To put a well formulated problem into statistical terms, so that it is amenable to statistical

analysis, is the task of the statistician. But he cannot do this if his client (the person who has data or plans to gather some) does not know what his own problem is. He, the client, must have thought about it to the very best of his ability. This may seem obvious, but you would be surprised how often someone comes to a statistician with a question no more specific than "what do the data say?". True, the statistician can help the client formulate his problem, but mostly along the lines of the statistician asking the client if such and such a statistical formulation agrees with the real-life formulation – or is not unreasonably far removed from it. The client must offer all the knowledge of the data he can. Not like a big city medical and air pollution study that I was briefly involved in some years ago. Among the 50 million data items were records of medical symptoms from 1000 people interviewed weekly over a 2-year period. The doctors wanted to use the data to define diseases – and after spending 2 years to collect the data they wanted their analyses yesterday without contributing any of their own knowledge of medicine to those analyses.

As client and statistician look at one another across the consulting table, I'm sure they try to sum each other up. One statistician (Hyams, 1971) has recently categorized some client stereotypes as follows:

The probabilist: he just wants a P-value.

The numbers collector: he does N^N experiments, gets 8 side inches of computer output and proclaims "it's too big and too complicated to understand or explain."

The sporadic leech: he button-holes statisticians in the hallway but declines to sit down at a desk with them and discuss a problem thoroughly.

The amateur statistician: he says "statistics is easy enough but I don't have time to learn it" all.

Fortunately, most research workers do not fall into these categories.

4. THE CLIENT AND HIS STATISTICIAN

One thing a client must be prepared for is to answer the many questions that will be asked by the statistician, especially when the statistician is unfamiliar with the client's subject-matter area. Indeed, the client will have to educate the statistician, and to speed up this process the statistician will ask lots of questions pertinent to the possible analyses he has in mind as he learns progressively more about the problem at hand. How did you weigh that compound? Did you use the same spring balance for every weighing? Was the same brand of test-tubes used all the time? How many lab technicians did the weighings? When? All on the same day? Was the cover put back on the balance each night? And so on and so on. In this process the questions might appear impertinent, as if the statistician is casting aspersions at the experimenter's ability to perform his work. Not at all. The statistician, in order to formulate the concepts of population and random sample pertinent to the whole discipline of statistics, must find out exactly how the data were gathered. The procedures involved usually become an intimate part of the statistical analyses. The client must therefore be prepared to be put through the third degree, so to speak.

In return, the client will want the statistician to be good at his trade. Let's hope that that goes without saying. But the astute client will want something more, as well. As Cox (1968) puts it, so well, the statistician should have "competence to understand, to criticize, to appreciate an experiment (and data) and the result that ensues". In this sense it will, as has already been said, be the responsibility of the client to help educate the statistician in the background of the data at hand. If he does, he will be surprised how often the statistician will ask a question that the client has never thought of. On many occasions a client will ask the statistician "how did you think of that?". The question arises

from genuine surprise that a mathematician can ask pertinent questions about something which perhaps he knows very little about, such as biochemistry, engineering or industrial processes or whatever the topic may be. This happens, I think, just because the statistician is indeed ignorant in this manner, and so is then unprejudiced and naive – and because of this, and the analytical and logical nature of his training and of statistical methodology, he does indeed develop an ability to think about things in a rather fundamental manner. And, after all, that is just what the client wants.

5. COMPUTING

Something else a client might want of a statistician is an ability at, and knowledge of, computing. To some degree I would take issue with this. First, the client should not think of a statistician as the computing person – and vice versa. In pre-computer days statisticians did not do a client's arithmetic – nor, maybe, did the client do it himself. A clerk was hired to do it. Today that clerk is a computer programmer. The relationship is the eternal triangle of science: experimentalist (or data gatherer), statistician and programmer, with communication problems existing between each pair of them. Certainly the statistician comes between the experimentalist and the programmer and needs to know something about computer work so as to communicate with the programmer. This is not so for the experimentalist – although he will be at no loss if he does know something about computing. But while the statistician should have some knowledge of computing, he does <u>not</u> need to be an expert, and this the client must understand. The programmer is the expert at computing – but, in turn, not at statistics. The statistician's job is, I believe, to know what to compute, not so much how to compute it. He needs to know his statistics first and foremost, and then to know enough about computing to be able to tell the computernik what is wanted. And this is true

even in today's world of canned, readily-available, computer packages for statistical analyses. In fact their presence makes it more imperative than ever that trained statisticians are given the opportunity to make recommendations about the suitability of different analyses. The existence of computing facilities does not put the hallmark of quality or appropriateness on any particular analysis. Indeed, such existence is only promoting and increasing the wrong use of statistics.

6. THE STATISTICIAN

One other thing that clients must not assume about their consulting statisticians – they cannot be available for consultations for 100% of their time. A statistician needs time to himself, to think about and work on his clients' problems. The excitement of consulting work to a statistician is two-fold: (i) that of helping a fellow scientist make inroads into scientific knowledge and (ii) the prospect of developing new statistical methodology that will be useful in wider settings than the problem at hand. It is to be remembered that many of today's statistical methods that have widespread use originated from practical, real-world problems brought, as such, to a statistician; e.g., analysis of variance started from the analysis of field crop experiments, and regression started with a genetic study of the heights of fathers and sons. The statistician needs opportunity, therefore, to attend to what is peculiarly his own contribution both to the consulting work and to his own profession. As Barnard (1972) says, "statistics has the honourable role of midwife to scientific advances. Once the baby is born she can fade into the background – or, rather, go to attend to other births." Before the next birth, though, the midwife statistician must see that the previous one is entirely complete.

So, when all is said and done and the client has been through a consulting session, what does he see in the statistician? Adapted from Hyams (1971) there

are several amusing – perhaps appropriate – stereotypes:

The one-analysis expert: subjects all data to his favorite analysis.

The hunter: does every conceivable analysis, stupid and otherwise, gets 14 side inches of computer output and has no idea what to do with it.

The nit-picker: makes mountains out of molehills.

The more-data man: always recommends more experiments.

The gong: always draws a bell-shaped curve!

Let's hope that in practice there are few of these stereotypes: that we at the universities can train people who are able to really help the user disciplines. I'm sure you sometimes have your doubts. But we are trying.

Some References on Statistical Consulting

Barnard, G. A. The unity of statistics (Presidential address). J. Roy. Statistical Soc. A, 1-14, 1972.

Cameron, J. M. The statistical consultant in a scientific laboratory. Technometrics 11, 247-254, 1969.

Cox, C. P. Some observations on the teaching of statistical consultancy. Biometrics 24, 789-802, 1968.

Daniel, C. Some general remarks on consulting in statistics. Technometrics 11, 241-246, 1969.

Hyams, Lyon. The practical psychology of biostatistical consultation. Biometrics 27, 201-211, 1971.

Sprent, P. Some problems of statistical consultancy (with discussion). J. Roy. Statistical Soc., A, 133, 139-165, 1970.