Cahill

# RADICAL STATISTICS 8

?
Isn't that a contradiction in terms?...
This is Radical Statistics 8.
Alright then, 7.6 - 8.4...
a motor bike
GROAN!
A Statistician
A Prince, who is interested in Statistics...

This issued prepared by
Jeff Evans and John Bibby

Editors of next issues

RS9 (Feb '77) Lee Harvey, Commerce Centre,
City of Birmingham Polytechnic
Birmingham B4 7HA (021-359-6851)
** COPY BY 11 FEB. **

RS10 (May '77) Roy Carr-Hill, 16 Hameau,
Michel Ange, PARIS

CONTENTS PAGE NO.

## A.G.M. SATURDAY-SUNDAY 5th-6th MARCH 1977

We have booked the Notting Dale Studies Centre for an all-day A.G.M. on 5th March. We will probably have a social do on the Friday evening. Further details will be set in the next newsletter.

People from out of town who want to say at Notting Dale should let Lee Harvey know as soon as possible (address on front cover) - cost is about £1.

The "caretaker committee" elected at last year's meeting (Liz Atkins, Jeff Evans, John Irvine, Ian Plewis, Stephen Shenfield) has suggested the following structure for the A.G.M.

Morning - Colloquium or workshop(s).
Afternoon - business meeting (including discussion of the draft constitution).
Sunday - informal workshops and discussions.

We have suggested the Saturday morning session in order to give a stimulating focus to the discussions, besides the business meeting. The afternoon session might discuss, among other things, recent developments in the Group's activities, especially publications (see the Health Subgroup's article elsewhere in this issue). And the Sunday sessions would give people the chance to meet informally, in particular to discuss practical proposals such as the Worker's Handbook, which was first mentioned to the Group at last year's A.G.M.

We might also consider the following subjects for discussion.

1. Mechanics of pamphlet production.

2. The need for a proper bank account ("every time I pay in cheques for pamphlet in a Coop I get a lecture on how our account is meant to be a personal one").

3. Intellectual topics - thoughts may be postponed owing to our present preoccupation with items 1 & 2!

4. Reports of sub-groups.

## HEALTH GROUP

Since the last newsletter appeared the Health Group's draft critique of the DHSS'consultative document 'Priorities for Health and Personal Social Services in England' has been transformed into a printed pamphlet entitled 'Whose Priorities' (and described as Radical Statistics Pamphlet No. 1 to encourage other subgroups to follow it with $^{nos.}$2, ..., n, n>>2). This process has not been exactly effortless, but having found out what was involved we would encourage others to repeat the performance.

We started by trying to find someone to publish the pamphlet for us. Although our approaches were received sympathetically we realized early in October that the only way to get it published by the end of the month when DHSS' "consultation" period closed was to do it ourselves.

We were very lucky that one of our group lives in a house which is a nerve-centre of alternative printing operations. Her friends were able to typeset and print it for us cheaply and at short notice. They are also in the demystification business and patiently initiated us into the arts of pasting up the typeset copy and assembling the printed pages into pamphlets, as well as plying us with welcome cups of tea. Maryanne Gordon of BSSRS very kindly drew up cartoons to enliven some of our points and others in BSSRS gave advice and helped with publicity. Finally who paid the bill? Not the CIA, KGB, drug companies or other organizations injurous to health but 20 members of the group. We hope that we will all be recompensed from sales perhaps leave some money in the kitty for the next production.

'Whose Priorities' was received favourably in the 'New Scientist' (Nov. 4, pp. 228-9) (with more than a little help from our friends!) and with qualified approval in the 'Lancet' on Nov. 6 (p.1035). It has since been reviewed favourably elsewhere in a variety of publications to whom we sent review copies. Martin Bland and I were interviewed by London Broadcasting but have yet to meet anyone who heard this broadcast! As a result of the publicity we have either sold or distributed all our original 735 copies and have had to have some more printed.

This concentration on stapling and folding paper and licking envelopes has not put an end to all other activities especially for those lucky enough to live away from London and thus avoid these activities. We have started reading DHSS' latest 'consultative document', 'Sharing Resources', which is a euphemism for sharing cuts. For this document the 'consultation' period is indefinite but the policies will be implemented in January. Other projects under discussion by us are the subject of clinical trials and the production of a set of notes to help people find out for themselves about resources and policies for health in their area.

'Whose Priorities' can be obtained from Alison Macfarlane (address below) at a price of 45p + 15p postage. Its production has attracted new members who doubtless have further ideas and suggestions. To discuss these and to work out how to organize outselves to carry them out, we are holding a one day meeting on Saturday January 15 from 11 a.m. - 5.30 p.m. in London at the Notting Dale Urban Studies Centre. For further information please write to:

Alison Macfarlane,
40 Warwick Road,
St. Albans,
Herts.

Note from the Editor:

It might be useful for RSN to have a discussion of "Whose Priorities" in the next issue. Would anybody who has comments to make please send them to Lee Harvey, the next editor (address on cover).

# THE INTRODUCTION OF TESTS OF SIGNIFICANCE INTO SOCIOLOGY

by

Liz Atkins

Some initial notes on a paper for the forthcoming 'Demystifying Social Statistics' book.

The use of quantification and statistical methods of inference in social science has been the subject of controversy within the discipline lately and since its introduction. The use of tests of significance, developed at the turn of the century, took very little time to appear in journals of sociology and gradually became almost mandatory in empirical research, yet a considerable section of the discipline resisted their introduction. As an example of a struggle within a discipline for ideological dominance the case of tests of significance in sociology seemed worth examining.

I wanted to collect evidence on the various explanations which seemed to me to contribute to an understanding of this development. Looking at the origins of sociology and its objectives, according to its practitioners, to see whether these already oriented sociologists towards quantification and automated scientific inferences, one can see a strong strain of positivism, social engineering and attempts to remove political action and replace it with 'rational' decision making based on 'social facts'. This objective appeared to develop in direct response to the social upheavals and revolutions to the nineteenth century and the claims of sociologists to a scientific method of understanding social and political movements would obviously make them powerful advisors of beseiged governments.

One can find clear evidence of this role for sociologists in the early developments of social surveys.

The introduction of tests of significance, contrary to my expectations, seems to have proceeded gradually without violent opposition perhaps because the basic idea of an ultimate scientific method to solve social problems was already current. The advocates of statistical methods certainly presented them as the 'great white hope' of social science, the cookbook approach of a set of rules for inference was presented as the means of raising sociology to a science on the level of natural sciences.

Criticisms of the automatic approach to scientific inference appeared quite soon after these prescriptions, but seem to have stopped short of a fundamental philosophical critique and concentrated instead on certain technical misdemeanours and the rigid use of fixed significance levels.

The level of use as represented by articles in the American Journal of Sociology was quite low in the period 1895-1921 being about 2% of published articles. The real expansion took place in the following 30 years so that by the late forties and early 50's sociologists of the Bureau of Applied Statistics felt obliged to write an extended defense of their ommission of the tests from their survey reports.

The use of tests obviously gave sociological work an appearance of scientific status which helped to convince governments as well as the public of the validity of their conclusions. As such, they enabled sociology to establish itself as a legitimate academic discipline and as a powerful ally of government, industry and the military.

Evidence for the ideological role of tests outweighing any philosophical or methodological criticism can be found both in the comments of sociologists themselves, e.g. Carl C. Taylor in 1920.

> 'It is imperative that the social sciences win for themselves the acceptance of their generalisations as trustworthy. A faith in such trustworthiness has almost as great a part to play in converting a body of knowledge into "science" as has the established method of analysing phenomena or an adequate set of working tools.'

and also in the fact that even after the philosophical and scientific weaknesses of this method were clearly demonstrated, rather than their use diminishing, it continued to grow.

In order to set this into its historical and social context one needs to link this development to the wider economic and political background. Sociology as a discipline clearly based its case for academic acceptance on its use as a means of social control.

One would expect statistical methods to be of crucial value in the legitimations of these claims, and by examining the types of problems which were studied using these techniques I hope to be able to demonstrate this ideological role.

As a statistician with a very recent interest and limited knowledge of sociology I'd be grateful for any suggestions for approaching and developing this analysis from others in the group.

# STATISTICS AND IDEOLOGY: THE BRITISH SCHOOL OF STATISTICS 1865 - 1925

Paper read to the Summer Meeting of the British Society for the History of Science at Southampton, July 1976; by Donald Mackenzie.

The period 1865 - 1925 is crucial in the development of modern statistical theory, and in that time Britain was perhaps the key centre of the subject. Galton's invention of the concepts of regression and correlation, Pearson's work on the chi-square test and his elaboration of Galton's insights into an extensive system of statistical theory, the work of 'student' (W. S. Gosset) on the t-test, the beginnings of Fisher's comprehensive reformulation of statistical theory: these all belong in this period. In 1865 there was only a scattered awareness that probability theory and the numerical data of the social researcher or scientist could productively be brought together: by 1925 there was a considerable bulk of theory and applications of the theory, there was at least one institution (the Department of Applied Statistics at University College, London) devoted to teaching and research in statistics, there was in Biometrika a well established journal of statistical theory. Although full institutionalisation of the discipline hadnot as yet taken place (that had to wait until after the Second World War), the crucial stage of a cluster of scientists involved had been reached.

This perhaps explains my interest in this particular country and period, and thus one side of my title. What of the other side - ideology? My aim in this paper is to discuss one crucial factor (an external factor if you wish, though I don't like the internal/external distinction too much) that was of importance in this development of statistics: eugenics. The eugenists sought to improve the genetic characteristics of the British nation by increasing the birth-rate of some groups (the 'fit') while decreasing that of others (the 'unfit'). Eugenics I argue (elsewhere) was not simply an example of the over-rash extension of uncertain knowledge: it was an ideology. It was a set of ideas that served the interests of aspecific social group (the group one might loosely call the professional middle class) in a specific historical situation, a crucial element in this situation being the problem of the urban slum-dwelling poor, the so-called 'residuum', that formed a particular problem of social control for pre-first World War capitalism.

What was the relationship between eugenics and British statistics? One level of approaching this question is to ask why people started doing statistical theory. Once a discipline has become institutionalised, there need be no too specific answer to this question: they may do it simply because it provides them with a job. However, in the period discussed here this was not the case. Doing statistical theory was not a ticket to a job. Instead I shall argue that at least some of the key individuals in the development of British statistics got involved in it primarily because of their prior interest in eugenics.

In the case of Francis Galton this is, I think, established. Galton was both the founding father of British statistical theory - all the debelopment in this period owes a tremendous lot to him - and the founder of the eugenics movement (he coined the word). The connection between Galton's statistics and his eugenics has been clearly shown, first by his biographer Karl Pearson and more recently by Ruth Schwartz Cowan. It was his passionate commitment to eugenics, and his belief that in the statistical analysis of inheritance lay the key to developing a scientific eugenics, that was Galton's motivation in his statistics. He was not a sophisticated mathematician, yet he pursued the thread of his investigations into heredity till they led him first to regression and then correlation. These concepts were the intellectual breakthrough that Pearson was to build on.

What then of Karl Pearson? His work in statistics did not really begin until 1892 when he was 35 and established as Professor of Applied Mathematics at University College, London. His previous mathematical work was mainly on the theory of elasticity. The reason he started to work on statistics lies more in his non-mathematical involvement. He was a radical, a free-thinker, a feminist, a socialist (of the Fabian rather than revolutionary variety) and a vigorous pro-imperialist Now eugenics we think of nowadays a straightforwardly right-wing belief system: at this time it had considerable attraction also to technocratic, meritocratic socialists such as Pearson. It fitted well into his imperialistic social Darwinism. He was not believer in the capacity for self-action of the working class: they had to be led into socialism by the middle class - and the socialist utopia must also be the eugenic utopia. In eugenics and a mathematicised Darwinism Pearson saw the key science of social reconstruction. This was the context of his work in mathematical statistics: the very title of his main series of papers indicates this - 'Mathematical Contributions to the Theory of Evolution'.

Galton and Pearson are the key figures of the pre-1914 development of British statistics. But is seems clear that similar motives also operated in the case of some less important figures. (For many people of course we just don't know - the historical evidence isn't there). One of these was a man called Arthur Black.

He has received no attention from historians as he died without publishing anything. However, some notebooks that have been discovered by him show him to have started work in mathematical statistics before Karl Pearson, and to have been a mathematician of considerable ability. He too was interested in problems of heredity and evolution: his magnum opus, now unfortunately lost, was entitled 'An Algebra of Animal Evolution'. We can only guess what that contained, but surviving notebooks show him to have been the first person to reach chi-square as the limit to a multinomial distribution, and also the first Briton to rediscover (independently) the Poisson distribution.

Another man is worth mentioning, even though he was to reach fame in a field very different from statistics - H. J. Laski, the distinguished political scientist and also supposed red menace behind the Labour Party. He was converted to eugenics by a travelling lecturer in the subject, and worked for a short time with Pearson on statistical studies of heredity. He was held by Galton to show great promise, but entering Oxford University he was won away from science and moved to history and politics.

Just before the First World War another young man, who was to prove of more importance to statistics than Laski, also became interested in the subject: R. A. Fisher. Once again, Fisher's interest in eugenics seems to have been a causal factor in his getting involved in statistical work. While an undergraduate at Cambridge he was one of the driving forces behind the Cambridge University Eugenics Society. Reading what remains of the records of this society makes it clear that his eugenic interests led him to a remarkable extra-curricular study programme in statistics (it had to be extra-curricular: error theory was the only statistical area you could get much teaching in at Cambridge until Yule got a job there in the Autumn of 1912). Most interesting of the papers of this society is a paper Fisher read to it in October 1911 (when he was just starting his third year as an undergraduate in mathematics and physics) on "Heredity - comparing the Methods of Biometry and Mendelism". This little paper shows the extent of his thought and reading in the subject, and prefigures many of the ideas of his famous paper of 1918 on the "Correlation of Relatives" that was so important in the development of population genetics.

Now I don't want to argue that all British statisticians of this period were motivated by an interest in eugenics - indeed it is quite clear that some, notably Edgeworth, Yule (whose work is discussed later) and Gosset were not. But there is a crucial difference between those who were and those who weren't. Galton and Pearson were key figures, central to the network of personal and intellectual links of the statistical community. They were the organisational leaders. Pearson in particular played the key role of the intellectual entrepreneur - guiding developments at University College from a few scattered workers with productive ideas to an established research institute and teaching department. The history of this department makes clear the connections between statistics and eugenics: thus from 1911, Pearson as head of the Department of Applied Statistics held the title Galton Professor of Eugenics.

We cannot, however, stop at this level of analysis. The connection between statistics and eugenics was not simply at the level of recruitment and institutionalisation: it was also at the level of ideas. Certain methodological problems, however, appear when we attempt to work at this level. How can we attempt to show that eugenics affected how statistical theory developed? Might it not have developed in the way it did anyway, even had the eugenics movement not existed? In the mathematical sciences, especially, concepts and theories are frequently thought of as developing by an immanent process of logical debelopment. As Wittgenstein puts it "so long as one thinks it can't be otherwise, one draws logical conclusions". Could statistical theory have been 'otherwise', could it have developed differently? If so, what is the cause of the difference, of the specific characteristics of the way the theory did develop?

One way of approaching this question is to take a comparative view. Thus, in this period a vigorous school of statistics developed in Italy, led latterly by Corrado Gini, which developed a radically different approach to statistical theory. Their approach was very different to that of Pearson and his pupils. At least part of the difference can be attributed to the differeint circumstances of development, in Italy, statistics developed in the Law Faculties of Universities and was primarily concerned with problems of administrative, social and economic statistics. It developed along more intuitive, less mathematical, more descriptive lines. Interestingly, Gini argued that the approach of what he called the 'English' school was too narrowly focussed on particular indices - notably the standard dviation (and related measures) and the coefficient of correlation. Now their particular prominence in the work of Galton and Pearson, can in part be attributed to their particular significance as measures of the variability of a generation and the relatedness of two generations.

But this is perhaps too speculative. It is also possible to look at theoretical differences within the British statistical community and attempt to account for these. The most bitter of these (at least prior to the 1920's) was the controversy between Pearson and Yule over the measurement of association of non-quantitative characters. (I exclude the biometrician-Mendelian debate, which, while involving somewhat similar political aspects, involved many non-statistical biologists). The basic issue of the Pearson/Yule controversy was this.

Suppose you have a fourfold table:

|  | $X_1$ | $X_2$ |  |
|---|---|---|---|
| $Y_1$ | a | b | a+b |
| $Y_2$ | c | d | c+d |
|  | a+c | b+d | N |

Thus $Y_1$ could be 'survived an epidemic', $Y_2$ 'dies in the epidemic', $X_1$ 'vaccinated', $X_2$ 'unvaccinated' - a is the number of those vaccinated who survived, etc. Now the problem is: what is the strength of association between the two attributes X and Y? Yule's approach was straightforward: he laid down three criteria that a coefficient of association must meet: that it should be zero if X and Y are independent (non-associated), that it should be +1 if X and Y show perfect positive association and that it should be -1 if they show perfect negative association. He put forward his coefficient $Q = \frac{ad-bc}{ad+bc}$ simply as empirically fulfilling these conditions. Clearly an unlimited number of coefficients also fulfil these conditions (for example, all the odd powers of Q), and Q has no special justification. Further, as Pearson was to point out, some of these other coefficients will rank order (according to strength of association) the same set of tables in different ways.

Pearson's approach was to produce, by a much tighter (but more precarious) theoretical argument, a uniquely justified coefficient of association: the tetrachoric coefficient of correlation. This coefficient was derived by positing a theoretical model of how the four-fold table had arisen - the assumptions of the model being that beneath the observed dicontinuous categories lay underlying variables which were continuous and which followed a bivariate normal distribution. Pearson fitted this model to the observed frequencies in the table, and deduced a value for the correlation of the underlying binormal surface. This correlation (which I repeat was a parameter of a theoretical model) he called *the* correlation of the table (and later the tetrachoric coefficient of correlation). Although Pearson was aware that this coefficient had been reached by a theoretical process and involved the positing of a model which could not generally be tested, this did not prevent him putting forward his method as *the* way to measure correlation. It was this approach that was to be criticised by Yule, who argued that Pearson was making unnecessary and contentious assumptions.

Time does not permit a full discussion of the course of the controversy between Pearson and his supporters on the one hand, and Yule on the other. I shall, however, try to explain why, in my view Pearson chose to measure association in the wya he did, and not, for example in the way Yule did, even though his very theoretical and assumption-laden approach might well be hald to contradict his own philosophical views as advanced in his *Grammar of Science*.

The tetrachoric method was put forward as a general means of finding the correlation of characters of characters not quantitatively measureable. However one particular purpose dominated Pearson's approach - that was to have aparticular tool to do a particular job. That job was laid down by his eugenic concerns. He wanted the tetrachoric coefficient as a measure of the strength of inheritance for characters for which no scale of measurement was available. Now for measureable characters such as height Pearson had a measure of the strength of inheritance: he would take a group of families, measure the heights of parents and the heights of off spring, and work out an ordinary coefficient of correlation. This for him would be the strength of heredity for height. (It is interesting to note how this method builds hereditarianism into the very process of his science all parent-child correlation is ascribed to heredity). However many characteristics of eugenic importance were not measureable in the way height was. Most important of these were the mental characteristics of humans: their intelligance (this was before the invention of the Binet scale), their temperament etc. Now Pearson felt able to convince people of the inherited nature of most characteristics in animals, and of the physical characteristics in man: he wanted to complete the eugenic argument by a convincing proof of the inheritance of human mental characteristics. This was what the tetrachoric coefficient was designed to do. Pearson knew that the ordinary correlation of brothers for height was about 0.5: he could now get teachers to classify pairs of brothers (getting parent/child data in this area was difficult) as (in effect) bright-dull, bright-dull, work out the tetrachoric coefficient of correlation, and call this the strength of inheritance for human ability. In fact doing this for a wide range of characters, mental and physical, measuring correlation by both the ordinary and the tetrachoric methods, he found values of what was for him the strength of inheritance clustered closely round 0.5. He considered the eugenic argument conclusively proven.

To review my argument: I claim that Pearson's eugenic purposes - his ideological purposes - led to his measuring association in the particular way he did. The point is that a looser method of measuring association, such as Yule's, would not have worked for Pearson. A value of Yule's Q and a value of the ordinary coefficient of correlation cannot be validly compared. If you accept Pearson's model, however, a tetrachoric coefficient and an ordinary coefficient of correlation can be validly compared: because you are assuming that your observed data are generated by underlying variables for which the ordinary product-moment coefficient of correlation is meaningful. And this comparability of the coefficients is what is needed to make the eugenic argument work.

I have only examined one particular aspect of the development of statistical theory in Britain in this period - although it should be said that this problem was the one to which Pearson devoted the largest single part of his statistical work after 1900. But I think that this instance, together with other similar instances - notably Galton's work on regression - demonstrates the influence of eugenics on the development of statistical theory as a system of knowledge.

One final note of reservation - the end of the period I consider sees the beginnings of an approach to statistics in which the dominant interest was not the demands of eigemocs but rather the need for technical control and prediction in industry and agriculture. Notable was Gosset's work - as an industrial scientist employed by the Guiness Brewery - and that of Fisher at the Agricultural Research Station at Rothamsted. This is a very different development, and needs to be analysed separately. But now is not the time to attempt to do it.

## THE INVERSE RATIO LAW OF FORECASTING by John Rowan

It would perhaps be generally agreed that economic forecasting doesn't actually work, in the sense of producing accurate pictures of the future. Galloway says that "the forecast may not be accurate for any particular year" and this view is certainly justified by the facts. So if forecasts don't tell us the future, what do they do? One pretty clear and believable answer is given by Ball, where he says:

> As matters stand in this general field of large scale model building, which is still as much an art as a science, there is much to be done - enough to keep a large body of researchers extremely busy over the next decade and more.

This idea, that the main purpose of forecasting is to keep a large body of researchers extremely busy (and supplied with food and water) is quite convincing, but does it benefit anyone except those researchers themselves?

A clue is given in the phrase "still as much an art as a science." The object of a scientific approach to forecasting would be to learn from the precise way in which a model failed as to how it could be improved next time round. The only infallible way of spoiling this possibility, and of stopping any learning taking place, would be to adjust the results in the light of commonsense, intuition, experience, general knowledge, inside information, hunch, prejudice and general zeal for improvement. This is, it seems, exactly what happens. Anxiety about being wrong far outstrips the desire for scientific purity. And so we get statements like - "The model, however complex, is generally an aid to judgement of a very specific kind, not a substitute for it."

This inability to decide whether to be scientific (and, as Berdy says, to get the benefit of an iterative series of experiments) or to lean heavily on judgement extends so far that it even leads to contradictions within the same article. At one point Macdonald is advocating a behaviourist approach which essentially involves Newtonian determinism, while at other points he denounces determinism and the Newtonian approach.

So it seems that the researchers doing economic forecasting have a nice little number going. They are like Penelope the wife of Ulysses, who sewed her tapestry all day and undid it again at night. Only it is even more brilliant, because the researcher can be ever so scientific while he is being watched, and leave it to someone else to provide the adjustments which destroy the pretence to science.

Well, supposing that we set up some kind of a self-denying ordinance, and said that scientific method must reign supreme, would that be the answer? Suppose that we gritted our teeth, and forswore subjectivity, and really kept abiding by the rules of the experimental game, would that ultimately give us more useful predictions?

The answer is in the negative. And the reason why the answer is so dusty is because of the Inverse Ratio Law of Forecasting. This law says that once the limits of commonsense have been passed, each extra increment of time and money spent on model-building will give less usable results. Let U be the Utility of the predictions made; let C be all the costs of using Commonsense in whatever way is optimal; and let M be all the costs of making and using large-scale economic models in whatevery way is optimal. If the object of the exercise is to maximise U, the equation runs:

$$U = C + \frac{1}{M}$$

An illustration of this law is given in the article by Professor Ball, where he says of the model produced by the organisation he works for that it "produced 25 pieces of information in 1966 as compared to 375 now. This has not meant that the quality of the forecasts has improved in any dramatic way ..." What it has done, of course (as well as making more work for more researchers), is to make any prediction extremely hard to understand or check. And it has been remarked more than once that figures are most sacred when least checkable.

One of the most important reasons why models get worse as they get bigger is that more and more of the figures put in are fiction. This is often not realized, though it comes through in many of the articles in this series. Estimates and guesstimates, assumptions and trends, hypotheses and insights, rules of thumb and generalizations, all come in again and again.

The other key reason why the models get worse is that the number of untenable assumptions increases. One assumption which is frequently built into economic models is that people operate as if they were examples of Rational Economic Man (or REM, for short). REM was invented to make life easier for mathematicians: you can build lovely mathematical models once granted REM. But none of these models apply in the real world because real people do not think like REM, and their behaviour does not correspond to his thoughts. The whole economic theory of consumer behaviour and household behaviour (as put forward by people like Green, Johnson and Devletoglou) is just ludicrous - it ignores all that we have found out about consumer psychology. I often feel that economists would actually prefer to use assumptions and estimates, rather than to use real data.

The fearful danger of using models is that you may begin to believe in your own magic. The Bloom and Stacey article gives a good example of this, where one can see the authors, despite all protestations to the contrary, moving further and further towards a fully automated system which is expected to become more and more believable. The reason why it can never be trusted is given in their own article. They say that causal models (like the interesting one given by David Lowe Watson in his article) are more flexible than extrapolative ones, but add that they are susceptible to the same inherent problem: "new or previously dormant factors can suddenly assume importance". This is of course true; what they do not seem to see is that this means that a dialectical logic is needed, rather than an Aristotelian one. Unfortunately computers do not at present have this facility; only human beings do, and not many even of them are able to use it in any conscious way. This is why, as Ball says - "Identification of these dynamic patterns is a crucial and difficult aspect..."

If I could finish by suggesting a modest experiment, it would be this. Take a forecast given by your favourite economic model, for a period of not less than a year ahead. Obtain an estimate, as accurately as possible, of what it cost you to get this estimate, in the form in which you are using it. Then spend one-tenth of this amount on getting as good a commonsense forecast as you can, using if possible at least one person able to think dialectically. And then compare the two, on whatever criteria seem to you to be the relevant ones. These might include general accuracy; the ability to make necessary decisions; the seriousness of any errors; success in alerting you to key things to look at further, control more tightly, etc. - and in general, any of the aims which forecasting claims to deliver. (For a true experiment to test the law,you would of course have also to order a more expensive forecast prepared by a bigger model.) If, after several trials, it turned out that my cheaper suggestion was the most cost-effective, it would presumably make sense to ditch the formal models completely.

The one thing my alternative would not give you is the ability to boast about what a big model you've got, and how much it costs you to run it. But that wouldn't worry you, would it?

Editor's Note:

The above article was a review by John Rowan of a recent series of articles on forecasting in Admap, a magazine read largely by people working in advertising and market research. This review was not accepted for publication by Admap. We felt that it raised a number of issues of interest to RS readers.

P.S. (from J.E.):

Readers may be interested in looking at the last 3 chapters of John Rowan's The Social Individual, vol.II (Davis-Poynter 1973), which I found a very stimulating introduction to the use of statistics and research methods for not-too-numerate social scientists.

# MUST MODELS MYSTIFY?

## A first look at the epistemological status of statistical models

John Bibby

1. Introduction The basic argument of this paper is that the construction of a statistical model inevitably assumes a particular construction of reality. Therefore in as much as this construction is socially and ideologically determined, statistical models must also contain an ideological component. However since this ideological component cannot be explicitly incorporated into the model, we have the basis of a central mystifying contradiction. Hence by purporting to be what they are not - pictures of the real world rather than portraits of the artist - statistical models inevitably confuse and mystify. This conclusion negates the assumptions of the use/abuse paradigm adopted by some radical statisticians.

In order to validate the above proposition we first prepare the gound by asking "What is a model"? Then we describe in more detail our understanding of the term 'statistical model'. Finally we return to describe the meaning of mystification, and consider the question posed in the title of the paper.

2. What is a model? Despite an enormous literature seeking to unravel the distinction between models and such close relations as theories, maps, metaphors and analogies (and paradigms and fairy-tales?) no general consensus on terminology has yet been found. Of course the word 'model' means various things according to contexts. A separate paper, parallel to this, will attempt to clarify these differences ("The great modelling muddle", in preparation.) That paper will distinguish amongst others, between physical models, conceptual models, mathematical models Marks I and II and statistical models. While different in many respects, these models do have one thing in common - they are all designed to extract for the investigator what seems to him to be an essence of reality, in a manner which adds to knowledge, understanding or insight. This might facetiously be presented as the 'model' of a model. Note the problematic insertion above of the idea of "essence of reality". This is to be understood as the result of a value-laden constructive activity and reflects the viewpoint and presuppositions of the investigator. Hence 'essence' means essential for the observer, and not essential for reality.

3. Statistical models In dangerously Althusserian fashion we shall view statistical models as a process which transforms a raw material (input) into a product (output). The input to the modelling process has three parts. Input I is a set of observations or data. Input II is a specification which usually takes the form of one or more mathematical equations. The specification states assumptions made concerning the relationship between the various theoretical concepts whose empirical counterparts are observed in the data. Thus Input II is a statement of assumptions concerning Input I. Input III on the otherhand represents the relationship or isomorphism between the theoretical concepts and their empirical counterparts. Thus Input III links Inputs I and II. The use here of the words like 'input' and 'data' is not meant to imply that these raw materials are unproblematic. On the contrary, neither data nor specification are value-free. They both reflect the ideological pre-suppositions of the analyst and through him of his social environment - non data sed capta (not given but captured).

Having considered the input we turn now to consider the output of the transformation process. This may take many different forms. It could be a set of point or interval estimates of unknown parameters. It could be the result of a hypothesis test - one star, two stars, or three stars, depending on the level of significance. (Not for nothing have the pages of statistical journals been likened to those of the Good Food Guide!) However, in order to paint statistical model-building in as flattering a fashion as possible, we shall use what seems to us to be the most thorough and least mystifying form of statistical model. (Nevertheless, as we shall see, it mystifies.) This is one whose output decomposes each element of observed data into a "fitted value", being that which is predicted by the model, and a "residual" which is simply "observed minus fitted". In other words observed value = fitted value + residual. There will be one equation similar to this for each element of the data. Thus if $\underset{\sim}{d}$ is a vector representing the data, then $\underset{\sim}{d} = \hat{\underset{\sim}{d}} + \hat{\underset{\sim}{u}}$, where $\hat{\underset{\sim}{d}}$ is the vector of fitted values and $\hat{\underset{\sim}{u}}$ is the vector of residuals.

The left hand side of this equation is Input I, while the right hand side is included in the output of the production process. (The output may also include such things as parameter estimates and hypothesis tests, but these seem to be less stringent requirements than those demanded by the above equation.)

So far we have discussed the input and output of the transformation process, but have said little of the process itself. This we shall now do, before passing onto consider its epistemological implications. The transformation process known as statistical modelling may be viewed as having the following stages.

1. One assumes that the elements of Input I (data $\underset{\sim}{d}$) are observed realizations of random variables. For instance in the simple linear regression model we may have data $\underset{\sim}{d} = (x_1,\ldots, x_n, y_1,\ldots,y_n)'$. Stage 1 of the transformation process then assumes that this $\underset{\sim}{d}$ is an observed realization of a random vector $\underset{\sim}{D} = (X_1,\ldots,X_n, Y_1,\ldots,Y_n)'$.

2. One assumes that the relationship between the random variables in $\underset{\sim}{D}$ accords with Input II (the specification equation). Thus in the simple linear regression model we have

$$E[Y_i] = bE[X_i] \quad \text{or} \quad E[Y_i] = a + bE[X_i].$$

In each of these examples the specification includes n equations, together perhaps with further assertions concerning variances, covariances, normality, and a statement concerning whether the design variables $X_1,\ldots,X_n$ are random or fixed (the latter being viewed as a degenerate special case of the former). The specification may take various forms. In the above examples the specification takes the form $F(\underset{\sim}{D},\underset{\sim}{\theta}) = \underset{\sim}{0}$ where F is a mathematical function, and $\underset{\sim}{\theta}$ is a vector of unknown parameters. However it could be completely nonparametric, and it need not involve an equation. Consider for example the specification $E[Y_i] > E[X_i]$.

3. Any unknown parameters are estimated. This stage will be omitted if the specification is completely nonparametric. It may appear to be omitted e.g. in hypothesis testing, but is usually there beneath the surface (e.g. likelihood ratio hypothesis testing implies the use of maximum likelihood parameter estimates). This stage is the first one where questions of statistical expertise could possibly become relevant, although even here they should not dominate. Whatever estimation procedure is used, let us call $\hat{\underset{\sim}{\theta}}$ the estimated values of the unknown parameters.

4. The 'fitted values' of the input are calculated. If the specification takes the form $F(\underset{\sim}{D},\underset{\sim}{\theta}) = \underset{\sim}{0}$ then the fitted values $\underset{\sim}{d}$ will usually satisfy $F(\underset{\sim}{d},\underset{\sim}{\theta}) = \underset{\sim}{0}$, although this need not necessarily be the case e.g. the method of monotone regression.

5. The residuals are calculated. The residuals are defined by $\hat{\underset{\sim}{u}} = \underset{\sim}{d} - \hat{\underset{\sim}{d}}$, and the residuals together with the fitted values are part of the product of the transformation process.

4. What is meant by mystification? It is commonly argued that a major weakness of mathematical models is their tendency to oversimplify the complexity of natural events (e.g. J. Gani, Model-building in Probability and Statistics, in T. Shanin, ed., The Rules of the Game). This would perhaps not be a weakness if simplification led to greater ease of understanding. However this is by no means the case. Firstly of course, the mathematical model is expressed in a tersely recondite language, inaccessible to the vast bulk of humanity. More importantly however, the mathematical model cannot abstract the historical conflict implicit in any situation. Just as poetry is lost in translation, dialectical reality tends to be mislaid and obscured by the process of mathematical formalisation. At the same time, the veils of obscurity tend to reflect the hegemonic ideology.

This process of obscurantisation is what we mean by "mystification". Of course this is not peculiar to mathematical models. It pervades the whole of culture, art as well as science.

John Berger has shown how mystifying conventional artistic criteria can be if applied to Hals' painting of The Regents of the Old Men's Alms House in Haarlem (Ways of Seeing, pp.11-16). A conventional art critic concentrates on topics such as the "human condition", "harmonious fusion", "personal vision", and "life's vital forces". Yet, Berger argues, all this merely evades the central historical fact i.e. Hals' masterpiece was painted by a destitute old painter who was forced to live off public charity, and the painter's subjects personified the affluence which necessitated that charity. Hence the mystification was achieved by evading conflict, rendering a-historical, and "explaining away what might otherwise be evident".

A similar situation exists in model-building. Instead of aesthetic criteria we have the antiseptic conventionalities of mathematical formalism. These are ideological in the sense that they tend to obscure the real condition of society, and thereby stabilise it.

Critical path analysis is another pertinent example - it has aptly been called the science which tells you to put on your socks before your shoes, rather than vice versa. This description capures the essence of mathematical mystification, which uses complicated language to state (and confuse) the obvious, thus making things appear much more difficult than they really are.

As Berger points out, the important question to ask is "who benefits from this mystification". The answer is not difficult to see. For model-building

(a) necessitates obscure language, which is a luxury available only to those who are able to obtain initiation into the knowledge elite; (b) it thereby validates the privileged position of the expert, and (c) disenfranchies the lay man. Finally (d) it abstracts from the socio-historical setting. This and other mystificatory functions have been discussed in the context of the general linear model in Bibby (1977) ("The general linear model: a cautionary tale", to appear in C. Payne and C. O'Muircheataigh, The Analysis of Survey Data).

Significance Tests - a discussion BOB GILCHRIST November 1976

This is a personal account of a meeting of the Radstats teaching subgroup and so, although it attempts to cover the remarks of other contributors, it is naturally enough a biased account. Dave Jarrett is adding his ideas in an accompanying article.

As our discussions developed it became clear that there was a general feeling that significance tests had achieved an overemphasis in journals. (A recent example of this phenomenon is given by Chatfield, 1976). It was felt that the over-importance of significance tests is particularly prevalent in the fields of medicine and sociology, where editors have come to regard significance tests as the measure of the respectability of an article. (It has even been suggested by Bross, 1971, that in fields where there is a serious chance of publishing effects which are not clearly established by data, the attainment of a significance level of, say, 0.05 is a reasonable requirement for the publication of results). Indeed, it was noted by the discussants that some journals were often not interested in non-significant evidence; a position complemented by the failure of many such journals to accept purely descriptive statistics. The historical development of this position was later discussed, particularly with respect to sociology, and it was noted that researchers might feel tempted to carry out significance tests as a precaution against others doing so - a sort of self-preservation which had developed into editorial convention. At the same time it was suggested that researchers had found in significance tests an automatic form of induction, and indeed that statistical tests had offered the respectability of scientific methodology to the emerging social sciences.

In considering the methodology of significance tests, it was noted that abuse was rife. Applied workers sometimes looked to statistics to 'prove' the validity of their subjective judgements, whereas significance tests were often no match for the experience and knowledge of the researcher. Indeed, in a real-life situation, informal techniques could often be so clearcut that calculations of any probability level should not be necessary, although applied workers seem reluctant to accept this position. Some researchers were also not fully aware of the limitations of the statistical techniques which they used. Perhaps of greater importance was the wider misunderstanding of tests; for example, their use to 'prove' a substantive difference in a census. Or the uncritical use of tests without regard to underlying models, illustrated by the lack of data transformations. At the same time, it can be argued that applied workers might not regard significance tests in the rigorous framework of a pure statistician, but would perhaps merely see them as a descriptive measure, perhaps to test the 'randomness' of the data. (an idea due to Fisher?). As so used, significance tests seemed less objectionable, even when used in doubtful circumstances. However, most applications of significance tests would probably be within a framework of testing a null-hypothesis against some alternative. (As an aside, it may be worth reflecting at this point that the original development of the $\chi^2$ significance test by Karl Pearson, 1900, was for the case of <u>no</u> alternative hypothesis - he used an absolute test with sample points ordered in decreasing probability under the null-hypothesis: see also Martin-Lof, 1975 . However, difficulties arise if we try to apply such tests to a continuous situation. In any case, it would seem that the occasions when an absolute test is sensible are very rare in applications: see Cox, 1976).

The discussion of the role of the hypothesis brought forward the suggestion that the hypothesis formulation could be particularly unsuitable for applied research, because of the lack of a direct indication of the importance of any departure from a null-hypothesis. Thus, in sociology, the null-hypothesis could be one in which the researcher did not believe; e.g. a zero correlation. Collection of sufficient data would then tend to reject the null-hypothesis, through observing a non-zero (although substantively insignificant) correlation. A second possibility is that substantive differences might be so delicate that the null-hypothesis could be perpetually accepted as being true - although the data may also be consistent with departures from $H_0$ which are of great practical significance. The latter possibility obviously leads to an acceptance of the status-quo through 'statistical methods'. It is also well known that large significance levels do not necessarily coincide with large values of similar measures of association (Duggan and Dean, 1968), although such measures are not without many methodological problems of their own. In further discussions of the role of such tests it may prove useful to apply the classification given by Cox, 1976. (This paper was unfortunately not available before our discussions). He considers that there are two main categories of null-hypothesis; the plausible and the dividing. The plausible hypothesis can come in two forms - primary or simplifying. The primary plausible null-hypothesis is of intrinsic interest; he quotes the example of whether data is consistent with the hypothesis of a random walk. A simplifying plausible hypothesis can be of simple primary structure - e.g. does a variance - covariance matrix offer a convenient, complete summary of some multivariate data; or the hypothesis can be of simple secondary structure - e.g. is some two sample multivariate data sufficiently normally distributed to allow the use of Hotelling's $T^2$ to test equality of mean vectors. Rejection of a simple primary structure would lead to a whole new model whereas rejection of simple secondary structure would leave the hypothesis of equal means unchanged, but would perhaps necessitate an alternative statistic. Dividing hypothesis divide the range of possibilites into qualitatively different types; thus $\mu_1 = \mu_2$ divides the situations with $\mu_1 > \mu_2$ and $\mu_1 < \mu_2$. Cox considers that this sort of hypothesis is of legitimate interest, even if not plausible, saying that if the data are reasonably consistent with the null-hypothesis then the data by themselves give no clear indication of the sign of $\mu_1 - \mu_2$. For example, the hypothesis that a point process is a Poisson process is used for dividing 'over-dispersion' from 'under-dispersion' even if the null-hypothesis is not plausible. In the context of our earlier remarks, acceptance of this null-hypothesis would merely lead to the conclusion that the data is inadequate to notice the direction of the departure, in the sense tested.

During the discussion of possible legitimate uses of significance tests, it was concluded that there were legitimate cases, although some might argue that many (all?) such cases should really be couched in a decision theory framework; e.g. the testing of, say, the hardness of two physical compounds. The sequential fitting of linear models was also put forward as an acceptable example of testing, although the choice of 'appropriate' significance levels could present problems. Nevertheless, there seemed few examples where the main summary of an analysis should be a significance test. In his recent paper, Cox suggests that tests can form the main summary of data when (1) there is a plausible null-hypothesis of intrinsic interest and (2) there is so little data that it can be assumed both that (a) evidence of inconsistency corresponds to a departure of scientific importance and (b) that even if the data agree very well with the null-hypothesis they are also consistent with departures of scientific importance. As an alternative to (2) he suggests a situation where instead, (2)' the data are so extensive that it is reasonable to assume that consistency with the null-hypothesis implies an absence of any effect of practical importance and (3) a reasonably high observed significance level is obtained, (i.e. accepting $H_0$). He considers that significance tests are also acceptable as a central conclusion where (1) there is a dividing null-hypothesis of the absence of structure and (2) there is such a limited amount of data that it can be assumed that data consistent with the null-hypothesis are consistent also with departures of scientific importance and (3) a reasonably large observed significance level is obtained. (again accepting $H_0$). Cox further suggests that in complex situations, with plausible hypotheses of simple primary structure, simplifying assumptions are essential for incisive interpretation. Thus he considers that tests can be used, for instance, to examine data for linearity, absence of interaction, or parallel regression lines where these are not the primary objects of the analysis, so that the object of these tests is to find a simple formulation for final analysis. At the same time he also states that significance tests of simple secondary structure are generally best avoided, if possible. To the author it appeared that our discussions had much in common with this sort of approach, although a lot of additional questions were also raised.

Coming back more specifically to the discussions it seems worthwhile to mention that confidence intervals were not covered explicitly. However, although the use of interval estimation is evidently preferable to point estimation, the basis of confidence intervals seems intrinsically linked to the use of significance tests with fixed levels. The spirit of any acceptance of significance tests by the author is not to regard any fixed conventional probability level $\alpha$ , nor indeed to deal quite differently with cases for which the observed $p \leq \alpha$ or for which $p > \alpha$. Thus, even if significance levels are used, then 0.1, 0.05, 0.01, etc., are seen merely as convenient points of reference. This position seems almost to rule out the use of confidence intervals as anything but a crude rule of thumb.

The theme of repeated experiments also arose during our discussions. Even if a replicated experiment is costly and takes a long time, it is nevertheless possible to form independent exploratory and confirmatory experiments by splitting the data at random into two parts, (apparently to the disapproval of Fisher), using the second part to assess the significance of results suggested after exploratory analysis of the first part. Such a procedure might be advocated where the researcher did indeed wish to give importance to a test of significance, although there is a problem that different splits can of course give different conclusions. This matter also leads naturally to the consideration of other problems caused by modifying analysis in the light of data. One school of thought advocates always carrying both an exploratory and confirmatory analysis. In his recent paper, Cox also discusses this matter and the related need for making 'an allowance for selection' when defining a significance level whose hypothetical physical interpretation is directly related to the analysis which was carried out. This seems an interesting topic for our future considerations.

Bearing in mind the many issues raised and the complexity of the matter, it might be reasonably concluded from our discussions that, although it did not seem that significance tests are totally too dangerous to use, they are perhaps too dangerous to teach, at least without a warning of their pitfalls. But, unfortunately, a series of earlier discussions led to many of us having the unhappy feeling that, with the limited time available in many non-statistics degree courses, it is just not possible to find time to teach both methodology and an understanding of the underlying philosophy. And it seems almost impossible to teach the understanding without the methodology. So we are left in a quandary.

In conclusion, we should remark that our discussants did not put forward any strong Bayesian views so our coverage of this was scanty. Of course, adherents of the coherence school (de Finetti, 1975) could not accept significance tests; but it would be interesting to know their reactions to the provocative view which was put forward in our discussions, namely that Bayesian analysis often appeared to offer more than a significance test but that, in reality, Bayesian analysis offered another form of automatic inference and, consequentially, could have many of the problems inherent in significance testing. (see Bakan, 1967)
Perhaps we can discuss this further in a later edition of RSJ?

## REFERENCES

BAKAN, D. (1967). On Method. San Francisco: Josey-Bass.

BROSS, I.D.J. (1971). Critical levels, statistical language and scientific inference (with discussion). In: Foundations of statistical inference, pp 500-519, eds. Godambe, V.P. and Sprott, D.A. Toronto: Holt, Rinehart and Winston.

CHATFIELD, C. (1976). A statistical true story. BIAS, 3, 2, 1-18.

COX, D.R. (1967). The role of significance tests. Paper presented to the European Meeting of Statisticians. Grenoble.

de FINETTI, B. (1975). Theory of Probability, A Critical Introductory Treatment. Vol 1 and 2. (translated from the Italian by Antonio Machi and Adrian F.M. Smith), London and New York, Wiley.

DUGGAN, T.J. and DEAN, C.W. (1968). Common misinterpretations of significance levels in sociological journals. The American Sociologist, 3, 45-46.

MARTIN-LOF, P. (1976). Reply to Sverdurp's polemical article Tests without power. Scand. J. Statist. 2, 161-165.

PEARSON, K. (1900). On the criteria that a given system of deviations from the probable in a system of variables is such that it can reasonably be supposed to have arisen from random sampling. Phil. Mag., Series 5, 50, 157-175.

Note: The above papers by Duggan and Dean, and by Bakan appear in HENKEL, R.E. (1970). The Significance Test Controversy, London: Butterworths.

# SIGNIFICANCE TESTS

David Jarrett, Middlesex Polytechnic

I want to consider criticisms of the use of significance tests in the social sciences and, in particular, question the value of the conventional introductory statistics course for social science students. The kind of course I have in mind has as its main aim an introduction to statistical inference; after a fairly brief treatment of descriptive statistics nearly all the methods included are significance tests, comparatively little attention being paid to estimation. Statistical inference is justified in terms of scientific induction and opposed to "merely descriptive" statistics, which is considered trivial and unimportant, with only those topics essential for understanding the inference part of the course being included. This emphasis on significance testing reflects that of much of the social science literature - the practice of some journals of publishing only those papers which report results attaining a certain level of significance is well known (Sterling, 1959) - but many authors have been critical of the use of significance tests. In sociology and psychology this has led to a controversy which has been collected and summarised by Henkel and Morrison (1970). Doubts about the value of significance testing have also been raised in geography (Gould, 1970); economists are usually more interested in estimation and do not use significance tests to the same extent as other social scientists.

Statisticians themselves, of course, are not in agreement about significance testing. I do not want to get involved in controversies about the foundations of statistical inference, but I think it is fair to say that in most introductory courses (and in most applications in the behavioural sciences) the approach is somewhere between that of Fisher (testing a null hypothesis without a specific alternative: if the probability of getting the observed or a more extreme result is less than a certain amount, then "*either* an exceptionally rare chance has occurred, *or* the theory of random distribution (the null hypothesis in his example) is not true" - Fisher, 1956) and that of Neyman, Pearson and Wald (a decision is made between a null hypothesis and a specified alternative hypothesis, the test used being justified by its optimal long run properties). Both approaches are criticised by statisticians of the Bayesian and likelihood schools. Some of the papers in the Henkel-Morrison collection echo the Fisher-Neyman controversy of the thirties, while others (e.g. Bakan, 1967) recommend increased use of Bayesian methods.

Henkel and Morrison draw a distinction between *statistical* and *philosophy of science* issues in the use of significance tests. Some of the statistical issues are technical or concern researchers' misunderstanding of the tests, but I think that others are difficult to separate from the philosophy of science issues; these are concerned with the problem of whether the use of significance tests is appropriate in the creation of scientific knowledge. Therefore, I will not attempt to discuss these issues in a general way but will concentrate on the role of significance testing in assessing theories.

Suppose we have a theory that two variables X and Y are related (for instance there is a non-zero correlation $\rho$ between X and Y), and wish to test this theory. A good textbook (such as Blalock, 1972) instructs us to proceed as follows: we set up the null hypothesis that X and Y are *not* related ($H_o: \rho = 0$), collect data (a random sample from the population) and carry out the test; if the null hypothesis is rejected at a sufficiently small significance level we can conclude that in fact X and Y *are* related. The criticism of this procedure is simple: there are usually *a priori* reasons for believing a (point) null hypothesis to be false - only rarely would we seriously consider the hypothesis that the correlation between X and Y was *exactly* zero - so, since most standard tests have asymptotic power one, the null hypothesis will always be rejected (at any significance level) if we take a large enough sample; in other words, our theory will always be *corroborated*. (One-sided tests - $H_o:\rho \leq 0$ against $H_1:\rho > 0$ - are a little more complicated; perhaps we could invoke the principle of indifference and conclude that the probability of corroborating our theory is 50% - see Meehl, 1967.)

The near certainty of rejecting the null hypothesis is a well known danger of such tests and it will be argued that the significance level should be chosen carefully, taking account of the power function of the test. There is little evidence, however, that many social scientists do this - standard levels of 5% and 1% are often preferred, and details of the power functions of standard tests are not taught in elementary courses, presumably because of technical difficulties. We can also ask for more precisely formulated hypotheses or for greater emphasis on estimation rather than testing - the *existence* of a deviation from the null hypothesis is rarely in doubt; what is important is its *size*. However, the testing of *a priori* false null hypotheses against vague alternatives is often justified on the grounds that the theory being tested is imprecisely formulated (Henkel and Morrison state that most hypotheses in behavioural science are "atheoretical in several ways") so perhaps the greatest need is for more developed theories.

Not all tests fit into the above framework - sometimes the null hypothesis itself is of interest and is being tested with the hope of acceptance rather than rejection. I think that at least two cases can be distinguished here: the hypothesis may be precise quantitative prediction (e.g. of the velocity of light in beer), or it may offer a useful approximation, perhaps stating that the relationship between two variables is linear, or that a series of events occur in a Poisson process. More credence might be attached to the hypothesis in the first case than in the second, but in neither instance would we seriously consider it to be exactly true (Rubin, 1971, notes as possible exceptions the constancy of the velocity of light in a vacuum and the non-existence of extra-sensory perception) so again the null hypothesis is certain to be rejected if enough data is available; but now our theory is always refuted. Rubin concludes that the use of tests at a fixed level of signficance is not appropriate. The second case is better regarded as an estimation, rather than a testing, problem - if the null hypothesis of linearity is accepted then we can use a simpler estimation procedure than if we decide that non-liniear regression is necessary - and introduces the problem of preliminary testing, currently a fashionable topic in the econometrics literature (e.g. Judge, Bock and Yancey, 1974). Clearly the choice of signficance level is of crucial importance here, and recent work of Leonard (1976) has shown that Bayesian methods can lead to the simpler estimation procedure (acceptance of the null hypothesis in classical terms) even though a significance test would reject the null hypothesis at any sensible significance level.

A further objection to the use of tests of significance concerns the problem of defining the population to which the inference applies. The theory of statistical inference is founded on probability theory, and classical procedures require a probability model for the data; in special cases the probability model is introduced artificially by selecting a sample at random from a real, finite population. Scientific theories are not usually concerned with finite populations, yet in the social sciences (except econometrics and other areas where more sophisticated stochastic models are used) the probability model is rarely stated explicitly; the problem becomes particularly acute when inferences are made from official statistics or from census data. Many elementary courses justify the procedures of statistical inference in the context of sampling from a finite population, with a hand-waving extension to a hypothetical, infinite population. Bakan (1967) makes the additional point that theories in psychology usually concern individuals, whereas significance tests only enable us to make conclusions about aggregates; a further induction from the aggregate to the general is necessary in order to obtain meaningful scientific propositions. The failure to appreciate this is part of a general confusion between statistical and scientific inference; major developments in the physical sciences took place without the use of significance tests, but to the behavioural scientist the tests appear to offer an automatic form of scientific induction.

Bakan concludes:

"What we have indicated in this paper in connection with the test of significance in psychological research may be taken as an instance of a kind of essential mindlessness in the conduct of research which may be related to the presumption of the non-existence of mind in the subjects of psychological research."

I believe that much of what we teach is of limited value if we want to encourage meaningful research in the social sciences. Moreover, Ehrenberg (1976) claims that what is taught is irrelevant for the everyday application of statistics. Perhaps we succeed only in giving our students the impression that statistics is a substitute for thinking. I do not know a universal solution to the problem but can suggest some improvements. Within the framework of the conventional course we can stress the limitations of the theory and attempt to make sure that the probability basis for inference is understood (though the latter may be difficult with students who do no mathematics). Certainly the treatment of descriptive statistics (understood as a body of techniques for exploring and interpreting data) can be extended. Some multivariate data-analytic techniques such as cluster analysis can be taught in an elementary course, though it might be objected that the unthinking use of such methods is as dangerous as the unthinking use of significance tests. Social scientists have been persuaded that statistical inference is important - it is taught to virtually all social science students, although they are not necessarily taught mathematics or the philosophy of science; I think that the philosophy of science in particular is at least as important as statistics.

# REFERENCES

BAKAN, D. (1967), The test of significance in psychological research, from On Method, San Francisco: Jossey-Bass, 1-29. (Reprinted in Henkel and Morrison, 1970)

BLALOCK, H.M. (1972), Social Statistics, McGraw-Hill

EHRENBERG, A.S.C. (1976), We must teach what is practised, The Statistician, Vol. 25, No. 2

FISHER, R.A. (1956), Statistical Methods and Scientific Inference, Edinburgh: Oliver and Boyd

GOULD, P. (1970), Is statistix inferens the geographical name for a wild goose? Economic Geography, Vol. 46 (2), 439-50

HENKEL, R.E. and MORRISON, D.E. (1970), The Significance Test Controversy, London: Butterworths

JUDGE, G.G., BOCK, M.E. and YANCEY, T.A. (1974), Post data model evaluation, Review of Economics and Statistics, Vol. LVI, 245-53

LEONARD, T. (1976), The Bayesian analysis of categorical data, Paper presented to the University of London Joint Statistics Seminar, 29 October, 1976

MEEHL, P.E. (1967), Theory testing in psychology and physics: a methodological paradox, Philosophy of Science, Vol. 34, 103-15 (Reprinted in Henkel and Morrison, 1970)

RUBIN, H. (1971), Occam's razor needs new blades, in Godambe, V.P. and Sprott, D.A. (eds.), Foundations of Statistical Inference, Toronto: Holt, Rinehart and Winston

STERLING, T.D. (1959), Publication decisions and their possible effects on inferences drawn from tests of significance - or vice versa, Journal of the American Statistical Association, Vol. 54, 30-34 (Reprinted in Henkel and Morrison, 1970)

## Who are these Radical Statisticians?

Well. There are 167 of us as I write. 57 in London, 13 in the North, 10 in the North Midlands, 16 in the South Midlands, 6 in the East, 1 in the West, 37 in the South East, 4 in the South, (144 so far), 5 in Wales, 6 in Scotland, 2 in Ireland and 10 in the rest of the world. Not a very uniform distribution, but perhaps typical of statisticians. It is impossible to find out how many of us would call ourselves statisticians; and how many just statistics users. Four or five are self-declared semi-numerate. A large proportion are still students, and an equally large proportion are teachers or lecturers. Of those who are not, the majority are in public service or in research establishments, but again I think this reflects statisticians as a whole.

In an attempt to prune out from the mailing list people who had got on it from half-interest and who had no desire to participate in the group, we asked everyone to confirm their interest in RS6 and RS7. As a result numbers dropped from over 180 this September to 120. However, simultaneously, health group produced 'Whose Priorities?' and numbers soared back to 167 and continue to rise. Many Area Health Authorities have asked for the newsletter, but what interest it produces there remain to be seen. We now have two doctors on the list (one who says she's a radical community physician!!).

When I took over the role of "Master of the RadStats Address List" from Liz Atkins, I was astounded at the volume of mail she has dealt with in the past 22 months. Now that the list has begun to settle down we have created a card index which has become the basis of the circulation list used by newsletter editors. As much as we know of interests and abilities have been put on each person's card, so I should be pleased to help people who may be of use with particular problems, and invite you to leaf through the index if we're both at the same meeting. Compiling the list was easy because of the good organisation of Liz and John Bibby. May I put in print our thanks for all Liz's secretarying through our birth and early infancy.

Finally, please address all changes of address, interests, etc., to me for prompt delivery of YOUR RadStats newsletter.

Nic Wright
Top Flat
65 Tower St day: 01-633-7532
WINCHESTER night: 0962-66971
Hants

## MISCELLANEA

Garth Allen writes,

"The Political Education Research Unit is located in the Department of Education at the University of York. It was established in 1974 as part of the Nuffield funded Programme for Political Education. Its basic functions are:

(a) to delineate the various aspects of political literacy, and to discover appropriate means of assessment of political learning;

(b) to identify the possibilities, problem areas, and limitations of formal programmes of political education available on an open access basis to students in six case study institutions.

The Unit also runs a documentation service. Any articles, discussion papers, etc. which the Unit feels might be of interest to people concerned with political eucation are sent free of charge of people on this mailing list. If anybody wants to join the mailing list and is interested in further details about the work of the Unit, please contact:

Jane Nelson
Secretary, Political Education Research Unit
Department of Education
University of York
Heslington
York YO1 5DD "

John Utting writes:-

I note that you are editing the next RadStats newsletter and it occurs to me that you might like to include news of the Survey Unit research staff, as RadStats has shown so much interest in our closure. In any case, you may like to know where people can be contacted.

If you put anything in the newsletter, you could say that none of us had any help from SSRC in finding the new jobs and/or (more important) that we hope to maintain an informal network for advice on survey research.

Mark Abrams — Director of Research, Age Concern, 60 Pitcairn Road, Mitcham, Surrey CR4 3LL

John Utting — Assistant Director (Research), National Children's Bureau, 8 Wakley St.,
London EC1V 8QE

Colin Brown — Centre for Studies in Social Policy, 62 Doughty St., London WC1

John Hall — Principal Lecturer in Sociology, Dept. of Applied Social Studies, Polytechnic
of North London, Ladbroke House, Highbury Grove, London N5 2AD

Alan Marsh — Senior Social Survey Officer, OPCS Social Survey Division, St. Catherine's
House, Kingsway, London WC2B 6JP

Cathie Marsh — Assistant Lecturer, Social and Political Sciences, University of Cambridge,
Syndics Building, Mill Lane, Cambridge CB2 1RX

Jim Ring — Research Fellow, National Institute for Social Work Training,
9 Tavistock Place, London WC1H 9SS

Jeff Evans writes:-

Geoffrey Randle's letter to RS$_7$ raises a number of issues. One of these is whether or not the "political attitudes of the statistician could affect his design or analysis of experiments" in "small-scale, well defined, 'problem-solving' research" on e.g. the nourishment yield of various plants. He contrasts this sort of study with larger scale, more diffuse projects which could be criticised both on political and on "methodological" grounds.

If he is saying that we need to be sensitive to a wide range of possible methodological criticisms of any study, then I agree. If he is trying to demarcate studies which are problematic in terms of "political attitudes", from those which are not, then I think this is difficult.

One can often demonstrate, or invoke, a "consensus", within an audience, a social class, or a society, that seems to make a certain research study, or a decision within the strategy used, "politically" unproblematic. For example, who would disagree that we should investigate variations in nourishment yields among food-plants?

What I think is important to acknowledge, however, is that every study will involve certain conceptual and practical "commitments". Thus, in the nourishment study, we need to ask: for whose use (i.e. for which countries and which classes) are these foods being considered? How labour-intensive is the production process of the foods being evaluated meant to be? And so on: the "values of the independent variable" are never unproblematically given. Nor is the list of dependent variables "given": (as Ian Miles pointed out) a decision has to be made as to whether or not to take into account the different extents to which these different plants impoverish the soil. You could say that the decision about how to measure nourishment yield would exhibit conceptual or practical commitments too(there is no point in rigidly distinguishing the two, since all human "conceptual"/theoretical activity arguably has a "practical" aim).

What I am suggesting here is that these commitments, whether or not you call them 'political', are part of every research study.

John Bibby has heard from Patricia Holland, Editor of Society Today which she says "is a new sociology magazine for schools. It is an offshoot of New Society and mostly reprints articles that have already appeared there. However, I am hoping than an increasingly important section of the magazine will be its "Feedback" pages, where teachers and students of sociology and allied subjects can write their own pieces. I heard of your interest in the "Radical Statistics" group, and wondered if any sociologist teachers among them would be interested in contributing. Perhaps a good idea would be to look at a few issues and them to comment on them?"
Anyone interested please contact Patricia Holland direct, on 01-261-5000.

John Bibby writes:-

POSE is the Schools Council Project on Statistical Education. Based on the University of Sheffield, they are writing material for the 11-16 age group. Their first ten project papers are now available at 10p each, and they might send you some draft materials if you ask them nicely. I'm going up to Sheffield early next year to help write material on inequality and/or housing. If anyone has ideas on this I'd really like to hear from them. P.S. Spent a day last week at the Maypole School, Birmingham. The POSE material seems to go down quite well, perhaps a bit on the heavy side.