The Statistician (1999)
48, *Part* 2, *pp.* 257–269

Statistics for the Millennium

From statistics to statistical science

John A. Nelder

Imperial College of Science, Technology and Medicine, London, UK

[Received January 1998. Revised November 1998]

Summary. It is asserted that statistics must be relevant to making inferences in science and technology. The subject should be renamed statistical science and be focused on the experimental cycle, design–execute–analyse–predict. Its part in each component of the cycle is discussed. The *P*-value culture is claimed to be the main prop of non-scientific statistics, leading to the cult of the single study and the proliferation of multiple-comparison tests. The malign influence of *P*-values on protocols for the analysis of groups of experiments is discussed, and also the consequences of the formation of inferentially uninteresting linear models. Suggestions for action by statisticians include the sorting out of modes of inference, the removal of non-scientific procedures, the offering of help to editors, the promotion of good software and teaching methods built round the experimental cycle.

Keywords: Combination of information; Experimental cycle; Fiducial inference; Hierarchical generalized linear model; Mathematical statistics; Multiple-comparison test; *P*-value culture; Statistical science; Statistical software

1. Introduction

The public image of statistics is poor and may be becoming worse. Almost nobody knows what statisticians do, and we in turn have been remarkably ineffective in explaining to non-statisticians what we are good at. In the present climate of intense competition between subjects for funds, statistics will lose further influence unless we embark on a discussion of what our subject is about and how we persuade others of its usefulness. This paper constitutes a personal attempt to start such a discussion. The somewhat polemical tone is intentional.

The main theme of my argument is that the practice of statistics has become encumbered with non-scientific procedures which perceptive scientists and experimenters are increasingly finding to be irrelevant to the making of scientific inferences. As Box (1976) has pointed out

> 'The penalty for scientific irrelevance is, of course, that the statistician's work is ignored by the scientific community'.

The kernel of these non-scientific procedures is the obsession with significance tests as the end point of any analysis. It may seem surprising that most papers in applied scientific and technological journals continue to use these modes of expression in spite of the fact that they are of so little help in making scientific inferences. I believe that there are two reasons for this. The first is that many editors and referees will not accept papers unless they contain these non-scientific

Address for correspondence: John A. Nelder, Department of Mathematics, Huxley Building, Imperial College of Science, Technology and Medicine, 180 Queen's Gate, London, SW7 2BZ, UK.
E-mail: j.nelder@ic.ac.uk

 0039–0526/99/48257

modes of inference, and that authors know this and act accordingly. The second is that many scientists pass through their training without gaining any insight into the methods of science; this means that they use non-scientific statistics as a crutch to present their results in a form that they believe will be acceptable.

Section 2 deals with nomenclature and the important question of what we should call our subject; Section 3 introduces the experimental cycle and the relevant statistical processes concerned with each of its components. In Section 4, I discuss what I believe are some non-scientific procedures established in present-day statistics and give reasons for rejecting them; finally, in Section 5, I make some proposals for reform.

2. Nomenclature

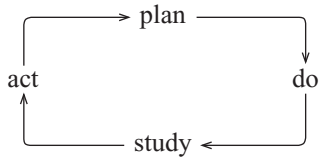
One of our biggest problems is the word 'statistics' itself. We need a new term, and that term should be, I believe, 'statistical science'. It is the name of a journal, and it also the title of the new professorship in the University of Cambridge. It shows that statistics belongs with science (and hence technology) and not with mathematics. If the new name is accepted several changes follow. First 'applied statistics' becomes a tautology, for statistics is nothing without its applications. The phrase should be abandoned. It has arisen to distinguish it from 'mathematical statistics'. However, this is also a misnomer, because it should be 'statistical mathematics', as A. C. Aitken entitled his book many years ago. To make this change does not in any way diminish the importance of mathematics. Mathematics remains the source of our tools, but statistical science is not just a branch of mathematics; it is not a purely deductive system, because it is concerned with quantitative inferences from data obtained from the real world. Bertrand Russell said

> 'mathematics is a subject in which we do not know what we are talking about, nor do we care whether what we say is true'.

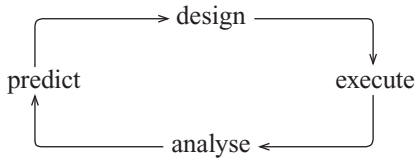
As statisticians, we *should* know what we are talking about and should care that what we say is true, in the sense of agreeing with phenomena in the real world. If we statisticians are to become statistical scientists we must become thoroughly familiar with the processes of science.

3. The experimental cycle

In the context of quality improvement, an important application of statistical science to technology, Deming (1986) defines the experimental cycle as



A possible scientific equivalent might be



We look briefly at the four components of the cycle.

3.1. Design

The development of the design of experiments is one of the great contributions of statistical science to science and technology. Yet, almost nobody knows anything about it! It grew up in the context of agriculture and has spread quite widely in the biological sciences. However, there are huge areas of physics, chemistry and engineering where the main ideas are unknown. In Britain (and, I suspect, elsewhere also) most professors of engineering believe that the right way to do experiments is to vary one factor at a time. The Japanese engineer Taguchi rediscovered fractional factorial designs decades after their original discovery by statisticians in the 1940s. He had no idea that this work had already been done. In spite of his hugely successful advocacy of these designs for quality improvement, and their enormous effect on Japanese industry, the engineering institutes in the UK show very little interest in making sure that these methods are taught in universities. This reflects a general ignorance of the subject, and this in turn reflects the almost total failure of statisticians to make these methods known and their power understood. Several people have said to me that they believe that the teaching of the design of experiments to statisticians is declining rather than increasing. If this is true, the position is even worse than I feared. How has this happened? I suspect that it is because of the baleful influence of fashionable mathematics, and the consequent pressure on statisticians to do work that will give them kudos in the eyes of mathematicians rather than to become of use to scientists in making quantitative inferences.

I believe that design should be taught as following from the single aim of maximizing information per run. This assumes that the idea of Fisher information has been taught first, which, of course, it should have been. It should be stressed that the choice of design, where the statistician can have a major influence, has to be preceded by the choice of experimental factors, their levels in the experiment and the choice of responses. These other choices, all of which involve difficult and complex decisions, must be decided primarily by the experimenter, although a good statistician, by acting as a Socratic teacher, can play a useful part. The practical complexity of the design process has often been grossly oversimplified to produce mathematical theorems of optimal design.

An interesting question is why so much bad design is tolerated in projects that could benefit from the introduction of fractional factorials etc. I believe that it is because a failure to generate maximum information per run does not show up explicitly in the budget and administrators do not know enough to ask awkward questions.

3.2. Execution

Although it could be argued that this component of the cycle is the responsibility of the experimenter rather than the statistician, I believe that all statisticians should have contact with experiments during their training. There are two good reasons for this. First, the statistician must understand the problems involved in obtaining reliable data from an experiment and the mistakes that experimenters can and do make in collecting such data. Secondly, many scientists are not taught as part of their training how to collect reliable data, so that the statistician who is knowledgeable may be able to help. Good execution of experiments is not easy; it requires administrative ability as well as knowledge. Many experiments fail because the data collectors have not been properly trained and many statisticians have their own horror stories to illustrate this.

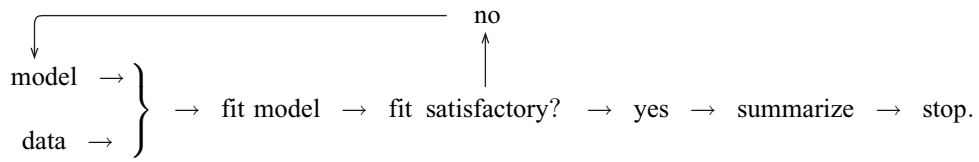
3.3. Analysis

The two main components of analysis are model selection and model checking. The old style of analysis, regrettably all too commonly illustrated in many statistical text-books, has the following unbranched form:

$$\left.\begin{array}{l}\text{model} \rightarrow \\ \\ \text{data} \rightarrow\end{array}\right\} \rightarrow \text{fit model} \rightarrow \text{summarize} \rightarrow \text{stop.}$$

I call this the *read–calculate–print–stop* mode of working; it is very much associated with the batch processing mode of computing. It assumes that the model class postulated is adequate for the analysis.

The new style incorporates a model checking loop as follows:



The introduction of this loop profoundly changes the way that we do analysis and requires fully interactive software if it is to be practised effectively. The depressing thing, to me, is that so many papers that are published in statistical journals, let alone in scientific journals, contain no model checking, and often the authors can be shown to have used a model for inference that does not fit. I do not understand why referees and editors allow this to happen. There are now model checking techniques for a wide class of models (see McCullagh and Nelder (1989) for model checking of generalized linear models (GLMs)) and there can be no justification for not using them as a matter of course. Students should know that a good model is

(a) *a priori* reasonable, which is a matter of initial formulation,
(b) parsimonious, which results from model selection, and
(c) internally consistent, which results from model checking.

3.4. Prediction

I use the word prediction to cover activities that follow analysis (the term is not ideal but I cannot think of a better one). Two major components are involved:

(a) the formation of quantities of interest from the basic results of the analysis, together with measures of their uncertainty;
(b) the combination of information from past experiments with the current experiment.

Examples of quantities of interest in component (a) include the LD50 (the dose required to kill 50% of the subjects) or relative potency from a quantal bioassay (Finney, 1971), and the formation of standardized disease incidence rates for the areas, age groups, etc., of a population following the fitting of a model to a sample from that population; here external information may be used in the form of census data for the population being studied (see Lane and Nelder (1982) for a general account of this kind of prediction).

Another example is provided by Taguchi's performance measures for quality improvement experiments. Such measures have a statistical interest because they show what happens when the two stages of analysis and prediction are inverted. Taguchi proposes first to form a performance measure for each run, and then to model it. For various reasons this is not a good idea; for example one such measure involves the quantity $\Sigma 1/y^4$, which is assumed to be a good estimate of $\Sigma 1/\mu^4$. If an analysis of the relevant component quantities is done first, the performance measure (quantity of interest) can then be formed and measures of uncertainty attached to it. In the example $\Sigma 1/y^4$ will be replaced by $\Sigma 1/\hat{\mu}^4$ which may be a good estimate of $\Sigma 1/\mu^4$. It is a matter of concern to

me that many statisticians seem to have accepted uncritically Taguchi's procedure, which clearly violates good statistical practice.

The combination of information from several experiments is an old practice in science but has recently in statistics come to be called meta-analysis. I dislike the phrase, because its etymology is suspect (it ought to mean the analysis of the process of analysis) and because a new word is not needed for a well-established scientific procedure. What statistical science has added to the techniques for combining information is the development of random-effect models, whereby sites, for example, in a repeated medical trial are treated as being a sample from a distribution. This is an area in which model checking is vital, although frequently neglected, because the sites may not look like such a sample, thus forcing a rethink about a suitable model.

The combination of information is more difficult than the analysis of single experiments, because differences in the environments of the different experiments introduce uncontrolled variation of the type met in observational studies, with the consequent difficulties in making valid inferences. The use of protocols in multicentre medical trials is now well established as a measure to strengthen the inferences that may be made from the combination of information.

4. Non-scientific statistics

4.1. The P-value culture

The most important task before us in developing statistical science is to demolish the *P*-value culture, which has taken root to a frightening extent in many areas of both pure and applied science, and technology. Many experimenters who have taken a statistics course are left with the belief that the purpose of an analysis is to calculate a *P*-value. If this value is sufficiently small then an editor might look kindly on their paper. In practice many experimenters do not have access to enough resources to do an experiment that will give a clear-cut result on its own; this does not mean that the experiment is useless, but that the results will have to be combined with the results of other experiments if a clear answer is to be obtained. However, if an experiment does not produce a sufficiently small *P*-value, then in the *P*-value culture it will either be self-censored by the experimenter as being not worth submitting or submitted nevertheless, only to be rejected by the editor. The effect of this non-scientific procedure is that the experiments that survive to publication have a distribution of sizes of the measured effect across experiments censored on the left. The bias resulting from this censoring means that the average size of the effect will appear larger than it actually is. If such a bias occurred within an experiment rather than between experiments it would be greeted with horror, but between experiments it appears to be tolerated. It is not enough to try to estimate how many trials were done but not submitted because they gave non-significant results, though this has been attempted on more than one occasion; what is required is a scientific method of accumulating information, and this is obstructed by the *P*-value culture.

Another distortion of the literature can occur when an effect is sought in an experiment but the results are submitted for publication only if the measured difference carries a sufficiently small *P*-value. If the ethos of research in the subject concerned is such that few alleged effects are checked by other independent trials, because doing such trials carries little kudos for the experimenters concerned, the stage is set for the production of a literature that may be little more than a junk-yard of false positive results. What is certain is that it does not deserve the name of scientific.

A recent paper by Nester (1996) gives quotations attacking the *P*-value culture that go back to 1933. Two in particular are worth repeating here.

> 'Null hypotheses of no difference are usually known to be false before the data are collected . . . when they are, their rejection or acceptance simply reflects the size of the sample and the power of the test, and is not a contribution to science' (Savage, 1957).

> 'The stranglehold that conventional null hypothesis significance testing has clamped on publication standards must be broken' (Rozeboom, 1960).

Yet here we are, 40 years on from Savage, and hardly anything has changed. (See, however, Section 5.3.) Nester makes the interesting comment that

> 'Scientists were the first to initiate hypothesis testing, and I think that it would be wise to lay all credit and blame for hypothesis testing squarely at their feet'.

A possible reason for the ready acceptance of significance tests is that they go with yes–no binary reasoning, and many people are happier with the qualitative conclusions that such reasoning seems to imply than with the quantitative assertions and inferences that are associated with confidence intervals, likelihood intervals or posterior distributions.

I have suggested a possible positive use for significance tests (Nelder, 1986) as tests of significant sameness rather than of differences. Such tests are relevant when we are trying to simplify a complex model by showing, for example, that a set of slopes can be replaced by a common slope. Here we hope to find a non-significant value of the test statistic, indicating a way of simplifying the model. This is consonant with the scientific search for invariance, i.e. finding things that stay constant when other things vary.

4.2. The cult of the single study

An infuriating by-product of the *P*-value culture is the presence in tables of results of the letters NS ('not significant') in place of the value of a regression or correlation coefficient. Because the value did not meet some arbitrary test level, a scientist reading the paper, who may have measured a similar quantity in his or her own work, is denied the possibility of comparing results. The withholding of such information is totally unscientific. I contend that the *P*-value culture has encouraged the non-scientific cult of the single study, considered in isolation. It is ironic that the development of methods for dealing with small amounts of information began with Student and Fisher, two men who were fully aware of the need to combine information across studies. Others seem to have had a more restricted vision. For further discussion see Nelder (1986).

4.3. Multiple-comparison tests

Multiple-comparison tests represent an attempt to make significance statements about unstructured sets of treatment means which allow for the fact that the set contains enough elements for there to be a substantial chance of naïve significance differences occurring by accident. The tests result in the ordered set of means being grouped into subsets, possibly overlapping, which are asserted to be non-significantly different. The word 'unstructured' here is important; one might think that if the treatments had a structure, e.g. arose from a factorial experiment, nobody would ignore that structure, order the means and apply multiple-comparison tests to them. However, I know of at least two instances where this has been done, and in both cases the experimenters clearly thought that this was the fashionable thing to do. Although most, if not all, statisticians would regard treating a factorial experiment in this way as absurd, there remains a belief among some that for unstructured treatment sets the procedure has some value. I assert that to apply these procedures does not contribute to any scientific inference of value. For an example, consider a trial of lines in a plant-breeding programme: let us suppose, for simplicity, that a single measure of value for a line can be defined. The plant breeder must decide how many lines to take on to the next round of selection and will pick the best from the ranked set; the decision that the breeder must make is how many to take. This may be influenced by theoretical work on the optimum selection pressure per trial in a sequence of trials; there are also arguments against picking just

one line, however superior it looks to the rest, because it may later turn out to be unusually susceptible to a disease that has not been encountered in the trials so far. (I know of an instance of this.) Multiple-comparison tests do not help at all in tackling such scientific concerns. Considerations such as this led me to say that

> 'multiple-comparison methods have no place at all in the interpretation of data' (Nelder, 1971).

4.4. The malign influence of P-values on protocols for analysis

Protocols, i.e. standardization of procedures for the conduct of experiments, are important, particularly at the design stage, because, without them, the combination of information will be put at risk by unrecorded variation in procedures between sites. Some agreement on methods of analysis is important, but writing effective protocols for analysis is a difficult business, and it is here that the *P*-value culture, combined with mistaken notions of objectivity, has frequently led to a gross loss of information from what are often expensive trials. I give two examples: in the first a trial of a drug had four doses 0, 1, 2 and 3, of which 0 denotes a zero dose. An analysis of such a trial would normally involve establishing a suitable dose–response curve and from that deriving relevant quantities of interest as a summary of the effects. Note that there will usually have been previous preliminary trials from which a model for the form of the response can be developed, and the information from those trials should be combined with that from the current trial. The actual protocol laid down that multiple-comparison tests should be done between the four responses, thus ignoring both the ordering and the scale of the doses. This procedure throws away information unnecessarily in the interests of a spurious objectivity.

In the second example, trials were conducted on several farms to measure the effect on the growth of young cattle of adding an antibiotic to their feed. The protocol laid it down that treatment differences should be assessed by using the treatment $*$ farm interaction as an error term. Because this interaction was large, little showed up. Among the measurements made was the initial weight of the animals; using this as a covariate produced a spectacular improvement in accuracy. Adjusting for initial weight accounted for all the treatment $*$ farm interaction, reducing its mean square to the level of that within farms. The slope for the covariate was negative, implying that the effect of the antibiotic was larger the smaller the initial weight, which is not unreasonable if small weight were associated with a high bacterial load. From this analysis it was possible to give useful quantitative advice to the farmer, something that would have been impossible if the protocol had been followed. The alleged objectivity that the protocol was supposed to guarantee led instead to a wrong analysis.

4.5. Distribution-free methods

Distribution-free methods make a virtue out of making no assumptions about the error distribution in deriving test statistics, and they depend heavily on the assumption that the purpose of an analysis is to calculate a *P*-value. If instead we look to an analysis to measure something and to give a measure of uncertainty of that measure, we shall need to find early in the scientific programme a suitable scale of measurement and to accumulate information about the form of the errors, so that these can be modelled as well. There is no point in delaying such an accumulation; if we have so little data that there is virtually no information about the form of the errors, then our scientific programme has not gone sufficiently far to allow us to make any useful statements about the size of relevant effects. To the extent that the use of distribution-free methods delays the process of modelling errors, it hinders the development of scientific procedures. Box (1976) used

the development of distribution-free tests as an example of what he calls 'mathematistry', the development of theory for theory's sake.

4.6. Uninteresting models

The exposition of linear models, a fundamental tool of statistics, is currently in a mess. I have tried to expose the reasons for this in several papers (see, particularly, Nelder (1977, 1994)). Two false steps have contributed to the formation of models, that, although well defined mathematically, make no inferential sense. The two false steps are

(a) the imposition of constraints on parameters because constraints must be put on their estimates to obtain unique solutions to the least squares equations and
(b) the neglect of marginality relations, e.g. between interactions such as $A.B$ and the corresponding marginal main effects A and B.

The result is the formation of models in which, for example, a marginal effect A is assumed null when the interaction $A.B$ is not. For random effects this is an obvious nonsense, and given that fixed effects can be regarded as random effects whose variance component is infinite it is *a fortiori* nonsense for fixed effects.

More recently a new type of uninteresting model has arisen as the result of the application of the weak heredity principle in choosing quantitative terms in linear models (see Nelder (1998)). This principle asserts that if a product term such as $x_1.x_2$ occurs in a model it is sufficient to include just one of the simple terms x_1 or x_2. Again the models are well defined mathematically but rarely make any inferential sense, because the surfaces that they define are *a priori* of no scientific interest. For example, in a recent paper (Breiman, 1995) such a model with a temperature covariate t led to different fits for t in degrees Fahrenheit and degrees Celsius.

5. What should we do?

The statistical community has several tasks before it if we are to replace statistics by statistical science.

5.1. Sort out modes of inference

At least once a year I hear someone at a meeting say that there are two modes of inference: frequentist and Bayesian. That this sort of nonsense should be so regularly propagated shows how much we have to do. To begin with there is a flourishing school of likelihood inference, to which I belong. It is relevant to the model selection stage, and it differs from Bayesian analysis in not incorporating prior distributions into the model. The consequence is that inferences can be concerned only with ratios of ordinates of the likelihood surface and not with the integration of posterior distributions as in Bayesian analysis. There are problems to be sorted out with the use of the word Bayesian to describe classes of models. Some writers apply Bayesian to any model that contains random effects; if we interpret Bayesian to mean incorporating a subjective element, as I do, this usage cannot be justified. A simple regression model contains a random effect, but nobody calls it Bayesian; similarly the model for a split-plot design contains random effects for main plots and subplots, but I have never heard this described as Bayesian. The basic distinction should be between models that have terms that are checkable, given sufficient data, and those that do not. Thus the two-stage model in which the distribution of the response y given parameters β and random effects u is combined with a distribution of u having parameters α is a non-Bayesian

model whose extended likelihood (the h-likelihood of Lee and Nelder (1996)) is derived from the joint probability

$$p_1(y|\beta, u)\, p_2(u|\alpha).$$

For this model the assumption made about the form of the distribution of u is checkable, for estimates of u can be obtained by maximizing the h-likelihood and standard model checking techniques can be used to see whether the estimates look like a sample from the assumed distribution. If, however, prior distributions are now assigned to the linear parameters β and to the dispersion parameters of the distributions of β and u, the model becomes a Bayesian model, and the distributional assumptions of these additional components are not checkable, however many data are collected. This lack of checkability is one thing that deters me from becoming a Bayesian. The important thing, however, is to be clear about the boundaries between likelihood and Bayesian inference in respect of model selection. In many practical cases differences in the inferences that are made will be small.

With model checking the situation is quite different, as noted by Box (1980). In checking a model we need to ask how our data set looks with regard to the model when compared with other data sets that we might have generated. This is an argument of the frequentist kind and needs the deductive mode of working, involving simulation and the like.

The position of fiducial inference, for so long the subject of controversy, seems at last to have been resolved by Barnard's work (Barnard, 1995). It is interesting to note that the fiducial argument as expounded by Barnard allows for statements of probability in respect of an unknown parameter of the same kind that Bayes employed in his original billiard-table example. Note also that Bayes's problem can be expressed as a binomial–beta hierarchical GLM (Lee and Nelder, 1996) with known dispersion parameters and is therefore an objective model with no subjective (Bayesian) component! Clearly we need to decide what we mean by Bayesian. What Barnard's work shows is that the conditions for there to be a fiducial solution to an inference problem are in practical terms highly restrictive and will not be met by a model of any complexity. They are not even met in the original Behrens–Fisher problem.

5.2. Remove non-scientific procedures

We need to remove non-scientific procedures from our activities. As discussed in Section 4 this will mean replacing the calculation of P-values by measures of effects and their uncertainties, discarding multiple-comparison tests, playing down distribution-free methods and replacing them by the modelling of errors, and clearing up muddles in exposition, such as the great mixed model muddle (Nelder, 1994).

5.3. Offer editors help

Many editors of journals of applied science or technology accept non-scientific statistical methods, or, indeed, even demand them because they have been told that that is what statistics is about. If challenged on this some will say that they have received their advice from statisticians, and so, by implication, who are they to argue? One successful attempt by an editor in collaboration with statisticians to improve the statistical standard of a journal concerns the *British Medical Journal*. In an initial survey of 103 papers referred for statistical assessment, it was found that in only 35 papers was the conclusion drawn from the statistical analysis thought to be justified, while only 17 were regarded as statistically acceptable for publication. The journal published a book, *Statistics with Confidence* (Gardner and Altman, 1989), which was primarily concerned to replace significance levels with estimates and confidence limits. See also Finney and Harper (1993). We need to

build on their experience. The general task will not be easy because it involves an implicit admission that non-scientific advice has been offered in the past. I leave it to the politicians among us to suggest how this may best be done.

5.4. Promote good software

We must promote good statistical software and, even more importantly, actively oppose bad software. Good software is interactive, supports model selection and model checking for a wide class of useful models (GLMs being an absolute minimum class), does not propagate the mixed model muddle and is generally oriented to scientific procedures. Much widely distributed software does not meet these conditions, and statisticians have tolerated this state of affairs for far too long.

5.5. Teach statistical science

We need to embed courses for both statisticians and experimenters in the experimental cycle of science, so that students have a clear idea where the course subject fits into the scientific framework. In particular, classes of models should be linked throughout to classes of data, with data structures emphasized from the beginning. The power of the experimental method must be stressed, and students must be clear about the limitations of observational data for making useful inferences. The execute phase of the experimental cycle should not be neglected, and students should carry out an experiment, however simple in structure, as part of their training.

Students should not be exposed to non-scientific models or methods of procedure. The systematic parts of models must be given equal prominence with the random parts, so that if students meet, for example, a response curve tending to an asymptote they will know at least one mathematical function that has that shape.

A recent perusal of an agricultural journal, where we might have expected statistical techniques, for historical reasons, to be above average, showed that nothing statistical in the papers was less than about 40 years old. If a scientist wrote a paper that took no account of any work done in the last 25 years, it would be surprising if an editor could be found to take it seriously. Yet a paper in which the statistical methods are even more out of date is apparently acceptable. I conclude that many of today's scientists are not being taught any of the useful modern ideas of statistical science, or that, if they are, they have not understood why they might be useful. We need to take a critical look at both what scientists are being taught and how they are being taught.

6. Conclusion

As Deming said, in another context,

> 'You do not have to do any of these things; survival is not compulsory'.

I believe that the situation that we find ourselves in is as serious as these words imply.

Acknowledgements

I have benefited greatly from comments on early drafts made by David Finney, Leonard Lefkovitch, Peter McCullagh, Adrian Smith and others. Any errrors, misunderstandings or omissions that remain are, of course, entirely my responsibility.

References

Barnard, G. A. (1995) Pivotal models and the fiducial argument. *Int. Statist. Rev.*, **63**, 309–323.
Box, G. E. P. (1976) Science and statistics. *J. Am. Statist. Ass.*, **71**, 791–799.
———(1980) Sampling and Bayes' inference in scientific modelling and robustness. *J. R. Statist. Soc.* A, **143**, 383–430.
Breiman, L. (1995) Better subset regression using the nonnegative garrote. *Technometrics*, **37**, 373–384.
Deming, W. E. (1986) *Out of the Crisis*. Cambridge: Massachusetts Institute of Technology Press.
Finney, D. J. (1971) *Probit Analysis*, 3rd edn. Cambridge: Cambridge University Press.
Finney, D. J. and Harper, J. L. (1993) Editorial code for presentation of statistical analyses. *Proc. R. Soc. Lond.* B, **254**, 287–288.
Gardner, M. J. and Altman, D. G. (1989) *Statistics with Confidence*. London: British Medical Journal.
Lane, P. W. and Nelder, J. A. (1982) Analysis of covariance and standardization as instances of prediction. *Biometrics*, **38**, 613–621.
Lee, Y. and Nelder, J. A. (1996) Hierarchical generalized linear models. *J. R. Statist. Soc.* B, **58**, 619–656.
McCullagh, P. and Nelder, J. A. (1989) *Generalized Linear Models*, 2nd edn. London: Chapman and Hall.
Nelder, J. A. (1971) Discussion on the papers by Wynn and Bloomfield, and O'Neill and Wetherill. *J. R. Statist. Soc.* B, **33**, 244–246.
———(1977) A reformulation of linear models (with discussion). *J. R. Statist. Soc.* A, **140**, 48–76.
———(1986) Statistics, science and technology. *J. R. Statist. Soc.* A, **149**, 109–121.
———(1994) The statistics of linear models: back to basics. *Statist. Comput.*, **4**, 221–234.
———(1998) The selection of terms in response-surface models—how strong is the weak-heredity principle? *Am. Statistn*, **52**, 315–318.
Nester, M. R. (1996) An applied statistician's creed. *Appl. Statist.*, **45**, 401–410.
Rozeboom, W. W. (1960) The fallacy of the null hypothesis significance test. *Psychol. Bull.*, **57**, 416–428.
Savage, I. R. (1957) Nonparametric statistics. *J. Am. Statist. Ass.*, **52**, 331–344.

Comment on the paper by Nelder

I. M. Wilson (*University of Reading*)

The points made by Professor Nelder in his first two sections and from Section 5.2 onwards are those which directly address the general issue of our profession's image amd impact. These are the points followed up on below. He writes in quite general terms here, but in focusing the central part of his paper on experimentation is clearly most concerned with one particular area of statistical work, where the statistician's client is relatively scientifically minded and trained.

Thinking of the wider clientele, and given present-day perceptions, it is not easy to be convinced that rebadging our courses, departments or professional bodies in terms of 'statistical science' rather than 'statistics' will have much positive effect on those who should be entering our profession or those who should be engaged in dialogue with statisticians. If the numbers of unfilled university places are any guide, relatively 'hard' sciences—like chemistry—are in the same boat as statistics in being seen as difficult and uninviting: in public perception terms, 'statistical science' is two wrongs which will not make a right.

Is it through ignorance of their statistical footing, or because of awareness of the negative overtones of the word, that newer developments are known by titles that omit reference to statistics? Data visualization, mining and warehousing or management information systems—and geographic ones—are examples.

Of course many of these computer-based approaches are concerned with the management of large amounts of data. We may not like the rather crude, limited and quite often misguided use of statistics in some of these settings, but the last decade has certainly seen a huge increase in the scale of data collection and recording; something must be done with the resulting morass! Most of those with this responsibility perhaps do not see statisticians as having the key skills to do the primary data management work; if so we will not be their first port of call and the majority will get no further than the data cruncher who produces simple summaries for them, as a relatively small component of the whole job. Despite the sophistication of the Roman Empire of statistics, we may be overrun by the vandal hordes! In this context the statisticians' predicament may be like that of post offices in certain countries which have huge distribution networks but are confined to handling small lightweight items in a freight market where most of the business by volume and by value is now at the 'heavy end': international businesses operating at the heavy end may be able to offer integrated contracts and to cut out traditional providers.

The producers of statistical software may to some degree be responsive to calls to make their products interactive, and better structured to encourage good statistical practice: it will be for the best if statisticians can send clear, insistent and repeated messages about what should be done. The marketing

of such packages seems to lead to rather greater pressure to incorporate special features, some of which produce excruciating interference with the message, e.g. those which offer an automated interpretation of the results of some procedure. At least, unlike some other classes of software, there is a fair variety of competing statistical products rather than a market dominated by one or two huge companies, so some may be responsive if statisticians can produce a coherent message. We should treat these as allies!

Few experienced statisticians would take issue with Professor Nelder's remarks about the way that *P*-values, multiple comparisons and distribution-free methods, after being widely seeded by statisticians, are now well established in the wild, and often appear as pernicious weeds rather than a useful crop. Sprouting in statistical packages, as well as in journals (which might be even worse if they did still less statistics), these procedures are perhaps popular because they are short and simple to implement and report. Better practice certainly is more sophisticated, less routinized, harder to teach and more demanding of the statistical practitioner and the user of the results. It is not just in relation to statistics that most people including students, authors and editors operate as satisfiers rather than perfectionists. To work towards a better situation requires a big effort from statisticians, not least because of the huge spread of applications where our discipline plays some role.

Professor Nelder proposes that we offer editors help. In the same sense that quality control of a finished product has an effect on what leaves the factory gates, this is useful, but unless the causes of defective product can be traced back and remedied effectively from research proposal onwards the offending author will merely have been chastised, not reformed. Very likely the offending paper will leave by the back gate and appear in a different journal! If statisticians can acquire greater *regulatory* authority, they may have more effect: the presence of regulatory statisticians in the pharmaceutical field has surely been the major reason that the industry has become such a major employer of statisticians—companies know that they will suffer a severe disadvantage if the regulatory statisticians raise difficulties, cause delay or demand more data collection.

How is work to be done, so that better statistical practice is promoted effectively? The Royal Statistical Society holds an important position in Britain, from which it could lead, e.g. in an effort to produce, distribute and popularize guideline pamphlets for the authors and editors of non-statistical journals, not to mention the research assessors in other disciplines. Professor Nelder's paper is for a statistical audience, but it lays out several points that could be expressed in suitable terms for such readers. Of course there are many other published papers making related points; their material also could be collected, sifted, targeted and nicely packaged. This would entail a large amount of work; anyone willing to undertake it without substantial recompense may need restraint and should certainly have their work vetted by a committee with the power of veto!

The work of making good statistical practice more widespread, more sophisticated and more valued by non-statisticians faces another type of constraint. This work bears a relationship to innovative research that is similar to that which agricultural extension bears to agricultural research. Unfortunately it has generally been accorded low esteem by university and even some research institute departments of statistics, by statistical societies and by leaders of the statistical community, in comparison with technical innovation. According to some, UK Universities' Research Assessment Exercise assessors in our cost centre appear not to have sought, recognized or rewarded the kind of wisdom that applied statisticians contribute to high quality science. The argument deserves to be widely promoted that *statistical* 'extension' work is unique in its importance to the quality of work done in other disciplines!

Universities' local difficulties aside, the statistical community and the Society could ponder how to create an enabling environment for working consulting statisticians to have influence where they are working in many cases almost alone, and given little effective support by their own kind. One of the quite unreasonable pressures that they often suffer results from being heavily outnumbered by mean-spirited client groups, who are also their competitors for funding within their institution.

Looking at publications addressed to members of other professional bodies gives the impression that rather more could be done to recognize and respond to consulting statisticians' needs for community, for training and for support. For instance, many working statisticians have had less formal training than their job now requires for example in relation to sampling, or are too young to have picked up for example on still helpful discursive papers published 20 years ago. They may not have the self-confidence, the desire or the freedom to talk on the sometimes adversarial seminar circuit but do benefit from more supportive and positive contacts with their peers. Applied statistical work is often rather 'self-effacing' in the sense that, when it has been done well, the clients have been convinced that it is 'obvious' that they should use some sensible, though not novel, technique. Consulting statisticians are often better at achieving this than at making clients realize and acknowledge that the conclusion would never have become obvious to

them without the statistician's intervention! Arguably our profession does less than would be justified to market itself in this way.

Of course these same working statisticians are often those who are best equipped to help non-statistical journal editors, but they will in general receive very little credit from employers, and no money, for time spent giving low profile editorial assistance to journals outside their own field. All too often, those who are best able to have an influence as reviewers probably have all too little incentive to do so in post-Thatcherite, privatized, cross-charging Britain.

By way of conclusion, I can do no better than to concur with Professor Nelder's final remarks!

Author's reply

Much of what Mr Wilson says fully supports (alas!) my first two sentences, especially the phrase 'almost nobody knows what statisticians do', and gives chapter and verse to support this allegation. While I am grateful to him for spelling it out in such detail, I am depressed at the size of the problem that he exposes.

He is quite right to note that jargon like 'data visualization', 'data mining' etc. have caught on and not been associated with statistics. I think that there are two reasons for this: the first is the (wrong) identification of statistics with statistical mathematics, and the second is the failure of statisticians to respond adequately to new topics like neural nets and the modelling of very large data sets. Both these reasons reflect a massive failure by the statistical community to decide what its job is, and how to respond to new challenges.

I am glad to have Mr Wilson's support for improving statistical software, and I hope that the Royal Statistical Society will take this seriously.

Mr Wilson is absolutely right to be concerned about the low esteem attached to the statisticians' contributions to published research. In my view no statistician should allow his advice on design, analysis and inference to result in anything less than a joint authorship. An acknowledgement is not enough.

There is much to be done, and no time to be lost, if we want to see our subject survive and have the influence that it ought to have. I repeat Deming's words

'You do not have to do any of these things; survival is not compulsory'.