# QUESTIONS, ANSWERS AND STATISTICS

**Terry Speed**
**CSIRO Division of Mathematics and Statistics**
**Canberra, Australia**

*A major point, on which I cannot yet hope for universal agreement, is that our focus must be on questions, not models. . . . Models can – and will – get us in deep troubles if we expect them to tell us what the unique proper questions are.*

J.W. Tukey (1977)

## 1. Introduction

In my view the value of statistics, by which I mean both data and the techniques we use to analyse data, stems from its use in helping us to give answers of a special type to more or less well defined questions. This is hardly a radical view, and not one with which many would disagree violently, yet I believe that much of the teaching of statistics and not a little statistical practice goes on as if something quite different was the value of statistics. Just what the other thing is I find a little hard to say, but it seems to be something like this: to summarise, display and otherwise analyse data, or to construct, fit, test and evaluate models for data, presumably in the belief that if this is done well, all (answerable) questions involving the data can then be answered. Whether this is a fair statement or not, it is certainly true that statistics and other graduates who find themselves working with statistics in government or semi-government agencies, business or industry, in areas such as health, education, welfare, economics, science and technology, are usually called upon to answer questions, not to analyse or model data, although of course the latter will in general be part of their approach to providing the answers. The interplay between questions, answers and statistics seems to me to be something which should interest teachers of statistics, for if students have a good appreciation of this interplay, they will have learned some statistical thinking, not just some statistical methods. Furthermore, I believe that a good understanding of this interplay can help resolve many of the difficulties commonly encountered in making inferences from data.

My primary aim in this paper is quite simple. I would like to encourage you to seek out or attempt to discern the main question of interest associated with any given set of data, expressing this question in the (usually non-statistical) terminology of the subject area from whence the data came, before you even think of analysing or modelling the data. Having done this, I would also like to encourage you to view analyses, models etc. simply as means towards the end of providing an answer to the question, where again the answer should be expressed in the terminology of the subject area, although there will always be the associated statement of uncertainty which characterises statistical answers. Finally, and regrettably this last point is by no means superfluous, I would then encourage you to ask yourself whether the answer you gave really did answer the question originally posed, and not some other question.

A secondary aim, which I cannot hope to achieve in the time permitted to me, would be to show you how many common difficulties experienced in attempting to draw inferences from data can be resolved by carefully framing the question of interest and the form of answer sought. A few remarks on this aspect are made in Section 6 below.

## 2. Why speak on this topic?

Over the years I have had many experiences which have lead me to think that the interplay between questions, answers and statistics is worthy of consideration. Let me briefly mention four, each of a different type.

The first experience is a common one for me. Someone is describing an application of statistics in some area, say biology. The speaker usually begins with an outline of the background science and goes on to give an often detailed description of the data and how they were collected. This part is new and interesting to any statisticians listening, most of whom will be unfamiliar with that particular part of biology. Sometimes the biologist who collected the data is present and contributes to the explanation, but at a certain stage the statistician starts to explain what she/he did with the data, how they were "analysed". By now the biologist is quiet, deferring to the statistician on all matters statistical, and terms like main effects, regression lines, homoscedacity, interactions, and covariates fly around the room. Sooner or later I find myself thinking "Here are the answers, but what was the question?" All too frequently in such presentations neither the statistician nor the biologist has posed the main question of biological interest in <u>non-statistical terms</u>, that is, in terms which are independent of analyses or models which may or may not be appropriate for the data, and I can certainly remember occasions when the analysis presented was seen to be inappropriate once the forgotten question was formulated. Of course many scientific questions <u>can</u> be translated into statements about parameters in a statistical model, so that I am not condemning all instances of the above practice.

A similar sort of experience is surely familiar to all who have helped people with their statistical problems. This time a scientist, say a psychologist, comes to me with a set of data and one or more questions. She/he knows some statistics, or at least some of the jargon. After being briefed on the background psychology and the mode of collection of the data 1 usually say something like "What questions do you want to answer with these data?", implicitly meaning "What <u>psychological</u> questions . . . ?" Not infrequently the answer comes back "Is the difference between such and such significant?" meaning, of course, <u>statistically</u> significant. [In my perversity I often think to myself: "Well, <u>you</u> should know; it's your data and you are the psychologist!"] Another similar query might concern interactions, or regression coefficients of covariates etc. What this has in common with the previous example is the unwillingness or inability of the psychologist to state her/his questions of interest in nonstatistical terms. We should all be familiar with the idea that scientific (e.g. psychological) significance and statistical significance are not necessarily the same thing, but how many of us keep in mind the fact that the latter involves an analysis or a statistical model, and that there may be as many answers to this question as there are analyses or models? Surely much of the blame for such thinking rests with us, the teachers of statistics, who never fail to popularize the rigid formalism of Neyman-Pearson testing theory.

My third type of experience concerns recent graduates in statistics, students I and my colleagues have taught and whom we believe should be able to operate independently as statisticians. Many of these graduates go into jobs in big public enterprises: railways, agriculture bureaux, mining companies, government departments and so on, and a few – far too many for comfort – get in touch with us when they meet a difficulty in their new job. It is not the <u>fact</u> that they get in touch which is discomfiting, but the questions they ask! For we then learn how little they have grasped. They have questions in abundance, often important policy questions, access to lots of data, or at least the possibility of collecting any data that they deem necessary, but they are quite unsure how to proceed, how to answer the questions. Out there in the world there are "populations" of real trains, field plots, cubic metres of ore or people, and even the simplest question relating to a mean or a proportion or a sample size can be forbidding. Perhaps they should standardize something to compare it with something else, perhaps include the variability of one factor when analysing another, or something else again, all things which we feel that a graduate of our course <u>should</u> be able to cope with unaided. But how well did we train them for this experience?

Finally, and briefly, let me castigate my professional colleagues – and myself, since I am no exception – for allowing ourselves to forget the fundamental importance of the interplay of questions, answers and statistics, for in so many of our professional interactions we act as if it is irrelevant. How many times have we presented new statistical techniques to one another, <u>illustrated</u> on sets of "real" data, drawing conclusions about those data concerning questions no one ever asked, or is ever likely to ask? And how often do we derive statistical models or demonstrate properties of models which are unrelated to any set of data collected so far, and certainly not to any questions from a substantive field of human endeavour. We are, so we tell ourselves, simply adding to the stock of statistical methods and models, for possible later use. Is it any wonder that we or our coworkers then find ourselves using these models and methods in practice, regardless of whether or not they help us to answer the main questions of interest. For a discussion of some closely related issues of great relevance to teachers of statistics, see the two excellent articles Preece (1982, 1986).

## 3. Why this audience?

I don't think I will be very wide off the mark if I assume that most of you – at least the active teachers of statistics amongst you – have come from a background of mathematics rather than statistics, and that few of you have actually been statisticians before you started teaching the subject. I would further guess that many of you still teach mathematics, and perhaps at the school level, statistics within a mathematics curriculum. It is on this assumption that I have chosen to focus on non-mathematical aspects of our subject, ones with which I feel you will generally be less familiar. As I said in the introduction, I hope that my talk will encourage you to give more attention to the non-mathematical aspects of statistics in your teaching, in particular to spend more time considering real questions of interest with real sets of data.

It is a curious thing that interest in the teaching of statistics in schools, colleges and universities has sprung up worldwide as an extension of mathematics teaching, because I certainly feel that the practice of statistics is no closer to mathematics than cooking is to chemistry. Both mathematics and chemistry are reasonably precise subjects in their own ways, and in general what goes on in them both is repeatable; perhaps they are true sciences. On the other hand, statistics and cooking are as much arts as they are science, although both have strong links to their corresponding science: mathematics in the case of statistics, and chemistry in the case of cooking. Who would recommend that a chemistry teacher with no cooking experience be appointed as cooking teacher as well? If I can convey to you some of the enjoyment and intellectual challenge that lies in my particular variety of cooking, and encourage you to try it yourself, I will have succeeded in my aims.

## 4. Two further examples

In this necessarily too brief section I offer two more concrete illustrations of interplay of the questions, answer and statistics. The first one is a very simple paraphrase of Neyman's classic illustration of hypothesis testing involving X-ray screening for tuberculosis, and I refer you to Neyman (1950, Section 5.2.1) for a fuller background and further details.

You have a single X-ray examination and, after the photograph has been read by the radiologist, you are given a clean bill of health, that is, you are told that there is no indication that you are affected by tuberculosis. With Neyman we will assume that previous experience has led to

$$\text{pr(clean bill} \mid \text{no TB)} = 0.99$$
$$\text{pr(clean bill} \mid \text{TB)} = 0.40$$

You now ask the radiologist "What are the chances that I have TB?" She says "I can't answer that question but I can say this: Of the people with TB who are examined in this way, 60% are correctly identified as having TB, and of . . . " You interrupt her. "Doctor, I know the procedure is imperfect, but you have just examined my X-ray . . . What are the chances that I have TB?"

If your radiologist is sufficiently flexible and well informed, she will answer "Well, that depends on the prevalence of TB in your population, that is, on the proportion of people affected by TB in the (a?) population from which you may be regarded as a typical individual". Indeed a simple application of Bayes' theorem yields:

$$\text{pr(TB}\,|\,\text{clean bill)} = \frac{\text{pr(clean bill}\,|\,\text{TB)pr(TB)}}{\text{pr(clean bill)}}$$

$$= \frac{0.40\text{pr(TB)}}{0.40\text{pr(TB)} + 0.99[1-\text{pr(TB)}]}$$

At last you see how to get an answer to your question. It may not be easy to obtain a value for pr(TB): your smoking habits, the location of your home, your occupation, your ancestry . . . may all play a part in defining "your population", but this is what is needed to answer the question and it is far better to recognise this than to fob you off with the answer to another question not of interest to you.

If this example smacks of Bayesian statistics it is not entirely accidental, for there are many occasions where the Bayesian view (which is certainly not necessary in this example) helps answer the question of interest, whereas classical statistics refuses, frequently answering another, unasked, question instead. For a more complex, explicitly Bayesian example, see the very fine paper Smith and West (1983) concerning the monitoring of renal transplants.

My second example concerns the determination of the age of dingos, Australia's wild native dogs. A statistician was given a large body of data relating the age of a number of dingos to a set of physical measurements including head length. The data concerned both males and females, a number of breeds and animals from a number of locations, but for this discussion we will restrict ourselves to a single combination of sex, breed and location. The question, or at least the task, to be addressed was the following: produce an age calibration curve for dingos based upon the most suitable physical measurement, that is, produce a curve so that the age of a dingo may be predicted by reading off the curve at the value of the physical measurement. This curve was for use in the field and it was taken for granted that an estimate of the precision of any age so predicted would also be obtained.

It was found that a curve of the general form $h = a + b[1 - \exp(-ct)]$, where h and t are head length and age, respectively, and a, b and c are parameters of the curve, fitted the data from each dingo extremely well over the range of ages used. This was an exercise in non-linear regression with which the statistician took great care, special concern being given to the different possible parametrizations of the curve, the convergence of the numerical algorithm used, the residuals about the fitted line and to the validity of the resulting confidence intervals for a, b and c. The parameters estimated for different dingos naturally differed, although, not surprisingly, the values of a (head length at birth) showed less variation than those of b (ultimate head length -a) and c (a growth rate parameter).

All this seems fine, and you might wonder why I am mentioning this example at all in the present context. My answer is as follows. The statistician in question knew, or knew where to find, lots of information about the fitting of individual growth curves, and so he focussed on this aspect of the problem. To answer the original question, however, his attention should have been pointed in quite a different direction, towards: the calculation of a population or group growth curve for the calibration procedure; features of the sample of dingos measured that may affect the use of their measurements as a basis for the prediction of the age of a new dingo; properties of the parameters which are relevant to this question; and, finally, towards obtaining a realistic assessment of the prediction error inherent in the use of the curve in the field.

In summary, he was willing and able to spend a lot of time on the individual animals' curves; he was less willing and less able to focus on the issues demanded by the question, those concerning population parameters, population variability, problems of selection, unrepresentativeness, and other issues including the use of normal theory, with real but not very well defined populations.

## 5. What is the problem?

Let me oversimplify and put my message like this. In the beginning we taught mathematics and called it statistics; much of this was probability, a quite different subject. Then, with the help of computers, we started to teach data analysis and statistical modelling; this was fine apart from one feature: it was largely context-free. The real interest (for others and many statisticians), the important difficulties and the whole point of statistics lies in the interplay between the context and the statistics, that is, in the interplay between the items of my title.

Let me offer a few similar views. A.T. James (1977, p. 157) said in the discussion of a paper on statistical inference:

> *The determination of what information in the data is relevant can only be made by a precise formulation of the question which the inference is designed to answer. . . . If one wants statistical methods to prove reliable when important practical issues are at stage, the question which the inference is to answer should be formulated in relation to these issues.*

Cox (1984, p. 309) makes the following characteristically brief contribution to our discussion:

> *It is trite that specification of the purpose of a statistical analysis is important.*

Dawid (1986) is even more to the point:

> *Fitting models is one thing; interpreting and using them is another, . . . If the model is correct and we know the parameters, how ought we to compare [schools]? . . . There is in fact no unique answer; it all depends on our purpose. . . . there remains a strong need for a careful prestatistical analysis of just what is required: following which it may well be found that it is conceptually impossible to estimate it!*

Tukey and Mosteller (1977, p. 268) offer seven purposes of regression, or, as I would paraphrase it, seven types of questions which regression analysis may help answer. Summarized, these seven purposes are:

1. to get a summary;
2. to set aside the effect of a variable;
3. as a contribution to an attempt at causal analysis;
4. to measure the size of an effect;
5. to try to discover a mathematical or empirical law;
6. for prediction;
7. to get a variable out of the way.

Similarly, Tukey (1980, pp. 10-11) gives the following six aims of time series analysis;

1. Discovery of phenomena.
2. "Modelling".
3. Preparation for further inquiry.
4. Reaching conclusions.
5. Assessment of predictability.
6. Description of variability.

Similar numbers of aims, purposes, or types of questions could be given for the analysis of variance, the analysis of contingency tables, multivariate analysis, sampling and most other major areas of statistics. Yet how often do our students meet these techniques in context with even one of these aims, much less the full range? And how else are they going to learn to cope with the special difficulties which arise when questions are asked of them in context whose answers require statistics? This is the problem.

## 6. Some General Comments

In this section I will mention a few difficulties which I believe can be resolved in a given case when the relation between the questions asked, the form of the answers desired and the statistical analysis to be conducted are carefully considered. A full discussion of any one of the difficulties is out of the question, and even if that had been given, there would probably remain an element of controversy, something which would be out of place in a talk like this. The section closes with some further general comments about questions.

Some elementary difficulties which I think arise include

- What is the population?
- When are population characteristics (e.g. proportions) relevant?

- What is the "correct" variance to attach to a mean or proportion?
- When should we standardize (for comparison)?

I have found that the relations between statistical models and analyses on the one hand, and populations and samples on the other, with parameters playing a role in both, are something which puzzle many students of our subject. The former play a big role in standard statistics courses whereas the latter are prominent in applications. Just how they connect is not a trivial matter.

A few somewhat more advanced difficulties include

- Which regression: y on x, x on y or some other?
- When should we use correlation and when regression analysis?
- When can/should we adjust y for x?
- Which error terms do we compare (in anova)?
- Should we regard a given effect as fixed or random?
- Which classifications (of a multiway table) correspond to factors and which to responses?

More subtle difficulties are associated with general questions such as

- Should we do a joint, marginal or conditional analysis?

I believe that in all of the above cases the difficulties arise because insufficient attention has been given to the nonstatistical context in which the discussion is taking place, and that when the question of interest is clarified and the form of answer sought understood, the difficulty either disappears completely or is readily resolved. Of course doing so takes some experience. Note that many of the difficulties listed involve, implicitly or explicitly, the notion of conditioning, or its less probabilistic forms, standardizing or adjusting. Just what we regard as being "held fixed" and what we "average over" in any given context is crucial, and here our questions and answers determine everything. The simplest form of this issue is usually: "Are we interested in just these units (the ones we have seen), or in some population of units from which these may be regarded as a (random?) sample, or both?" Models are no help here.

A simple but easy to forget aspect of the use of a statistical method is that not all questions which could be asked and answered by that method, are necessarily appropriate in a particular context. Lord's paradox, see Cox & McCullagh (1982) and references therein, provides a good example here.

## 7. What can/should be done?

It hardly needs saying that the best way to promote interest in the interplay between questions, answers and statistics is to put trainee statisticians into situations where they are required to provide answers to clearly stated questions on the basis of real data sets. Note that this can be a very different thing from "illustrating" a statistical technique on a set of data. In particular, much more background to the data is usually required, and this is rarely available in data sets presented in statistics texts. Indeed technical journals are now so tight with their space that it is rare to find full data sets published together with analyses and conclusions in scientific articles. This means that the best sources of suitable material of the kind being discussed, that is, of questions and data, are often one's colleagues or clients: teachers and researchers in other disciplines who make use of experimental or observational data in their work. Seeking out such material can be a way of forging links with the users of statistics and of course sandwich courses are designed with this general aim in mind.

One practice which I believe is valuable is the conduct of regular practical statistics sessions where students are asked to help answer specific questions on the basis of sets of data supplied together with background material. This is much more like the situation they will meet after their training is over. Two objections which are often expressed to me when I recommend this approach are (i) Surely it is unrealistic, except with the most advanced students, for unless they have learned a wide range of techniques, they will not be able to begin attacking "real" problems with any likelihood of success?; and (ii) Surely it is unrealistic, because real problems are so complex and real data sets so large, or even ill- defined, that nothing like what happens in practice can be presented in the classroom?

Both these objections have some validity, but let me make a few observations concerning them. Firstly, it is not necessarily a bad thing for a student (or anyone!) to attempt to answer a particular question (solve a particular problem) without knowing of the tools or techniques that may have been developed to answer just that type of question (or problem). This goes on all the time in the real world: parts of the wheel are rediscovered time and time again, and locomotion is even found to be possible without the wheel! And of course there is very seldom a single "correct" way to answer a question; an approach using less knowledge of techniques may well be better than one which uses greater knowledge. In the hands of a good teacher, such experiences can provide valuable object lessons, and, at the very least, valuable motivation for techniques not yet learned. Surely nothing could be more satisfying than hearing a student say: "What I need (to answer this question) is a way of doing such and such, under the following circumstances (e.g. errors in this variable, that factor misclassified, these observations missing or censored, that parameter chosen in a particular way, etc.)? Group discussions, where ideas are shared and knowledge pooled, are also most appropriate for this sort of work, and most enjoyable. The teacher can then play a subsidiary role, at times focussing the discussion back on the questions, perhaps at other times supplying a sought-for technique.

It would seem to me that this is just the sort of statistics which should be taught in secondary schools, not the watered-down and frequently sterile mathematical material which is often found at that level.

The second objection, that real problems are often very complex and rarely amenable to the sort of trimming that would be necessary before they could be used in a classroom, is harder to dismiss. It is certainly true that many (most?) problems are like this, but surely this highlights even more the difference between "illustrative" data sets, taken out of context, with no realistic questions or idea what would be satisfactory answers, and what we expect students to be able to cope with upon graduation. There is certainly a big gap here – between "pseudo-applied" statistics involving context-free sets of numbers, to illustrate arithmetic, and fully-fledged "warts and all" consulting problems – and I can only state that in my experience it is possible to find problem data sets which can be presented in the way I am suggesting. It certainly takes a little effort to find such material, particularly if you are not in the habit of meeting people with data and statistical problems. But as teachers of the subject, that is not such an unreasonable thing for me to expect of you is it?

A teaching strategy which could provide a means of putting these ideas into practice might be the following: pair yourself (the statistics teacher) with a teacher in an empirical field of enquiry, e.g. biology, agriculture or medicine, and also pair your statistics students with students in the corresponding class, requiring them to work together on a practical project which will enrich their understanding of both disciplines, and how statistics helps to answer questions. Many variants on this suggestion could be devised; the important thing is try something along these lines. Statistics students must meet more than mathematics and sets of numbers in their training, and it is the teachers of statistics who must arrange for this to happen.

**Acknowledgement**

I am very grateful for the discussions and comments on this topic offered to me by my CSIRO colleagues Peter Diggle, Geoff Eagleson and Emlyn Williams.

## 8. References

Cox, D.R. (1984). Present position and potential developments: Some personal views. Design of experiments and regression. J.R. Statist. Soc. Ser. A., 147, 306-315.

Cox D.R. & McCullagh P. (1982). Some aspects of the analysis of covariance. Biometrics 38, 541-553.

Dawid, A.P. (1986). Contribution to the Discussion of: "Statistical modelling issues in school effectiveness studies" by M. Aitkin and N. Longford. J. Roy. Statist. Soc. Ser. A, 149, 1-43.

James, A.T. (1977). Contribution to the Discussion of: "On resolving the controversy in statistical inference" by G.N. Wilkinson. J. Roy. Statist. Soc. Ser. B, 39, 157.

Mosteller, Frederick & Tukey, John W. (1977). Data analysis and regression. Sydney: Addison-Wesley Publishing Company.

Neyman, J. (1950). First Course in Probability and Statistics. New York: Henry Holt & Company.

Preece, D.A. (1982). t is for trouble (and textbooks): a critique of some examples of the paired samples t-test. The Statistician, 31, 169-195.

Preece, D.A. (1986). Illustrative examples: illustrative of what? The Statistician, 35, 33-44.

Smith, A.F.M. & West, M. (1983). Monitoring renal transplants: An application of the multiprocess Kalman filter. Biometrics, 39, 867-878.

Tukey, J.W. (1977). Contribution to the Discussions of "A reformulation of linear models" by J.A. Nelder. J. Roy. Statist. Soc. Ser. A, 140, 72.

Tukey, John W. (1980). Can we predict where "Time Series" should go next? Directions in Time Series Eds D.K. Brillinger & G.C. Tiao. IMS. pp.1-31.