OPINION ARTICLE
doi: 10.2306/scienceasia1513-1874.2006.32.001

ScienceAsia 32 (2006): 1-5

Understanding Data: Important for All Scientists, and Where Any Nation Might Excel

Paul Hutchinson

Centre for Automotive Safety Research, University of Adelaide, South Australia 5005, Australia.

E-mail: paul@casr.adelaide.edu.au

ABSTRACT: In some areas of science, the big rich countries are at an advantage. But in many cases, even well-resourced research projects do not make optimal use of the data collected. It is argued here that it is sensible for individuals, for institutions, and even for nations to give some degree of priority to the process of understanding data. The key point is value for money. At relatively low cost, good understanding of data gives a relatively high pay-off, as much can be done with only a desktop computer and an educated human brain. University statistics departments are central to helping a scientist develop two types of skill: a repertoire of techniques for data analysis, and an ability to create a special model tailored to a specific dataset. The example of the first discussed here is the accelerated life hypothesis. There are useful techniques in textbooks that seem not to be known by many scientists working with survival times of organisms or equipment. As an example of the second, the correlation across different studies between response to treatment and response to placebo is discussed.

KEYWORDS: data analysis, publication of research, statistical methods, survival times, accelerated life, placebo-treatment correlation.

INTRODUCTION

Thailand and Australia are not poor countries, but neither are they so rich as to readily afford billion-dollar particle accelerators or orbiting telescopes. Naturally, then, there is some selection of what subjects our scientists study. I wish to argue that the *understanding of data* is a subject that should have relatively high priority. Editorials in this journal have commented on Asian science research (especially that from Thailand) and its publishing in local and international journals.[1,2] Publishing means that gaps in research, perhaps even errors, are made public, and it is plain that all over the world, an area of weakness in science is the analysis of data and the integration of conclusions with substantive theory. Any nation might excel, says the title above; but no nation is doing that at present.

At universities, the understanding of data usually comes under the heading of statistics, but other names include research methods, quantitative methods, and biometrics. In my view, it is sensible for individuals to give some degree of priority to studying statistics, it is sensible for institutions (e.g., universities) to give some degree of priority to developing staff knowledge in statistics, and it is even sensible for nations to give some degree of priority to improving statistical analysis of scientific data. Why? Is statistics an easy subject that anyone can pick up? Far from it, that is not my point at all. Rather, there are barriers to the better appreciation of data. These barriers can be dismantled or by-passed. Details depend upon whether we are referring to an individual, an institution, or a nation, but the key point is value for money. At relatively low cost, good understanding of data gives a relatively high pay-off. Certainly much can be done with a desktop computer processing the numbers and your own (educated) brain processing what the results mean.

The dominant reason for studying statistics is to help understand your own data, and as part of a team, to help your colleagues understand their data. However, some published research has errors and omissions in the data analysis, and an additional benefit from studying statistics is that if you are experienced in analysing your own data, you are much more likely to be able to spot the limitations of someone else's analysis of their data. If you do see that someone's account of their data is incomplete, you have the opportunity to write a letter to the journal giving your criticism and alternative results. For many Asian scientists, English is not their first language, and yet it is the primary international language of science. Writing a comment giving an alternative analysis may have the welcome side-effects of improving use of English and of gaining experience with journals worldwide. Such comments are shorter than regular articles, and they are tightly focussed on the detail that is discussed; thus they may be easier to write.

The present paper will be organised as follows: choices in data analysis; the central role of university departments of statistics; examples of topics; concluding remarks.

Choices in Data Analysis

Earlier, I referred to *limitations* of an analysis rather than *mistakes*. This was because complete errors are relatively rare, but there are choices available in the processing of data. Once certain choices have been made, then certain conclusions may follow. But those conclusions are constrained by, are limited by, the choices made. Alternative choices can sometimes lead to alternative conclusions. Any of the following may be matters of choice:[3] the measurements analysed (for example, should the numbers be transformed?); how they are summarised (for example, the mean or the median?); what null hypothesis is appropriate to the scientific question; and to what alternative hypothesis should the statistical test be sensitive. Moreover, a choice is not necessarily conscious: within a particular field, analysis can easily become frozen in format, following too closely what previous workers did.

There are choices, too, in the basic approach to a dataset. Sometimes there is a specific hypothesis, or a few hypotheses, and the task is to test it. But sometimes the approach is much more open-minded, somewhat exploratory. Both hypothesis testing and data exploration are valid activities. However, when exploring data, one needs to be aware that conclusions from hypothesis tests may not have their full meaning: when many different variables and sets of variables are examined, it is to be expected that some comparisons will show "statistically significant" differences just by chance.

The Central Role of University Departments of Statistics

If better understanding of data is needed, how can this be achieved? Probably through a university department of statistics. I say probably because I am arguing principally for better application of relatively elementary statistical methods, and conceivably there may be expertise in this in almost any department. For some types of data, mathematicians can be as useful as statisticians — sometimes what is needed is to describe in equations what is going on, and sampling variability (the province of statisticians) is not of prime concern.

For a scientist who wants to study statistical methods in order to apply them in some other subject, what is needed is a learning environment where (a) data from the scientist's own subject is readily available, and (b) there are people who collectively have experience in applying a variety of statistical methods to a variety of datasets. This probably means studying for a research degree, perhaps within a statistics department or perhaps within another department that has some statistics staff. It should be an environment where researchers, junior and senior, all help each other, not one where instructors pass on chunks of knowledge to students. (Of course, lectures are a very efficient means for covering basic statistical methods. But it is the integration of these with the practice of scientific research that is lacking.) In such an environment, the scientist may hope to develop two types of skill. One is to acquire a repertoire of techniques for data analysis, and to be able to recognise when they are appropriate. (A lot of scientists have taken only an elementary statistics course; there are many more techniques available than are met there.) The second is an ability to create a special model for the relevant dataset, the predictions of which may be compared with the particular data at hand.

Some people might say that the task is to tailor the analysis to the research question, not tailor the analysis to a dataset. There is truth in this. But in reality, scientists often have only a vague notion of what their question is, and seeing the data leads to refinement of the research question, or to generating new questions.

Examples of Topics

I have said above that opportunities often arise to comment on published data analysis, and this may either involve applying a textbook technique or proposing a new hypothesis. This section will give examples of topics that repeatedly occur in research articles in slightly unsatisfactory or incomplete form. Specific instances could easily be traced with the aid of a search engine such as Google Scholar.

Example 1

There are many contexts in which change over time is of interest — for example, when material breaks, or equipment fails, or plants die. Testing in more harsh conditions (e.g., at higher temperatures) than those used in practice is common, because more failures are observed in a short time period. Drawing conclusions often involves extrapolation — either in time from early failures to late, or in temperature (or other condition) from hostile to relatively benign — and thus it is important that the assumed mathematical model be as nearly correct as possible. There are a number of standard techniques to be found in textbooks, and yet many published papers do not take full advantage of them. (Some of these techniques apply to quantitative change in a measured property as well as to qualitative change such as from life to death, but others become less interpretable in that context.)

In journals of technology and biotechnology, it is common to find graphs showing plots of log(proportion surviving) against time and an assumption that these relationships are linear. There may then be an attempt

to relate the rate constant to conditions (e.g., to examine over what range of temperature is the Arrhenius law valid). This is a topic where there is a real chance of inappropriate conclusions if the analysis becomes frozen in format, as a different method of plotting may reveal a systematic departure from the hypothesis of linearity. This may even occur when the plots of log(proportion surviving) versus time do appear roughly linear: this method of plotting is not very sensitive. Let s be the proportion surviving, and t be time. Instead of ln(s) versus t, plot ln[-ln(s)] versus ln(t). This will be a straight line of slope 1 if the first order (exponential degradation) model is correct. It will be a straight line of some other slope if the Weibull survival model is correct; the slope is the shape parameter of the Weibull distribution of survival time. The exponential model is a special case of the Weibull. The key feature of the "accelerated lifetime" hypothesis is that plots are parallel, in the sense of being separated by a constant horizontal distance when time is on a logarithmic scale. Being a straight line, or being a straight line of slope 1, is not the core concept, though several straight lines all having the same slope is indeed easily noticed when plotted. The patterns to be looked for in a set of Weibull plots are as follows: (a) straight lines (failure to find this would mean the Weibull hypothesis is not valid); (b) same slopes for different conditions (failure to find this would mean the accelerated lifetime hypothesis is not valid); (c) some simple relationship between conditions and acceleration of life (e.g., the Arrhenius relationship with temperature). More than one experimental parameter might be varied within the same set of experiments, and it would be of interest which had an accelerative effect and which had a more complex effect.

Another technique, instead of plotting a transform of the proportion surviving, is to estimate the proportionate rate of failure of those surviving, i.e., (-ds/dt)(1/s). (This is often termed the hazard function.) Reference 4 has an example of this, where interest centred on the survival of bees.

Example 2

Survival time, whether of living organisms or equipment, is a familiar topic, relevant data often are published in journals, mathematical descriptions appear in textbooks, and the obvious gaps in analyses occur in one publication after another. These things are not true of this second example, which will involve developing a model (a very simple one) tailored to a specific dataset.

There are many drug trials conducted, and for some diseases, it is possible to systematically study the response to placebo, and to correlate this with the response to active treatment, the data points referring to different trials. In some such studies, it has been found that trials reporting high level of success with active treatment tend to also report high level of success with placebo. Considering this, our thoughts might turn to the disease itself. We might form the opinion that the so-called disease is a pattern of symptoms, themselves imprecisely defined, and there is not a known pathological origin or mechanism of action of successful treatment. If that is so, the idea might come to us that sufferers from the condition that has been labelled with a single name actually have one or other of two different diseases, that one of these responds to placebo and the other does not, and that their relative prevalences vary across populations. More precisely, there are two groups of patients. In one, the average response is Q_1% with placebo and with active treatment also. In the second, no response occurs with placebo, and with active treatment, the average response is Q_2%. The proportions in the two groups are p and 1 - p, and p varies from study to study. Therefore, overall the average level of response with placebo is pQ_1 and the average level of response with active treatment is $pQ_1 + (1 - p)Q_2$. As in this model Q_1 and Q_2 are constants, across trials there is a linear relationship between response to placebo and response to active treatment. The statistician's work probably finishes at this point, but the medical researcher may want to propose an interpretation for the two groups. It might be that there are really two diseases, not one, the first having quite a high spontaneous recovery rate, and the drug being partially effective against the other. (The question might arise of whether there are two diseases, or two forms of one disease.) Or the patients themselves might differ, some having quite a high spontaneous recovery rate, and the drug being partially effective for the others.

I have contrasted the survival time analysis of Example 1 with proposing a special model in Example 2, the former appearing in statistics textbooks and the latter not. I should admit that in Example 2, the essence of the model is that all points along a straight line are differently-weighted averages of the two end points, and that this is a very familiar idea in elementary mathematics. However, it is not a familiar approach to analysing data. Also in relation to Example 2, it should be noted that an alternative to the hypothesis of two groups of patients is to focus on the variation that exists (for reasons that are usually unknown) between studies in the average level of response to placebo. Given this, it might be said that the correlation is unsurprising, simply because whatever it is causing the variation, it is likely to be shared by the treatment arms of the trials. (As with the two group hypothesis, the medical researcher may want an interpretation. That is, the source of the common variation may be of importance: is it the population studied, the nonspecific features of

the therapeutic environment, what the active treatment is, or something else?)

Other common formats of data

Other common formats of data include the following.

- Comparison of two methods of measuring the same thing, or capturing the same concept: typically, two or more methods have been used on the same items, and differences between them examined, or their correlation.
- Multiple linear regression: several quantities have been measured for each of the experimental units, and the aim is to predict the dependent variable from some linear combination of the others.
- Factorial experiments: several factors have been manipulated, with data being obtained for many combinations of them, and the aim is to predict the dependent variable from the levels of the factors and possibly from their interactions.
- Analysis of ordinal grades: when a dependent variable is expressed as a grade rather than a true measurement, the question may arise of whether parametric or nonparametric statistical methods are more appropriate.

I have listed these particular formats because in each case, I would expect a scientist to get some useful results with a textbook and a software package, but that he or she would also miss other aspects of the data, that should emerge in a true collaboration with a statistician.

Scatterplots showing a relationship between one proportion and another proportion are sometimes encountered. I find that this is sometimes a signal that it may be possible to start from a vague reason for an empirical relationship and make this quantitative. (With response being a percentage, Example 2 was of that type, though this turned out not to play a part in the model proposed.) One proportion, or probability, is at least commensurate with another; one being plotted versus another suggests a connexion between them; and sometimes the reason for expecting a connexion may be turned into a theory. An example is that if it is found that the types of road crash having a high probability of death (as contrasted with survival) tend to be the same types that have a high probability of serious injury (as contrasted with slight), we might hypothesise that there is a common variable that is responsible. (This might be the violence of the impact relative to the susceptibility of the person involved.) From that, it may be possible to develop a quantitative relationship between the two proportions.[5] This example may be expressed in more general language, as follows. It is common when dealing with measurement data to assume that some experimental manipulation or variable affects only the mean of the outcome variable, not the variability or shape of the distribution. This assumption can sometimes be taken over to the context of two probabilities, and one probability interpreted as a particular region under the probability density curve and the other probability as another.

Concluding Remarks

I see much research published in American or European journals, that was conducted at great trouble and expense, where only a very limited amount of data analysis has been done. Other scientists are limited in the experimental contributions they can make by lack of money, and I cannot help thinking that it would be cost effective for them to put some of their effort into conducting a fuller analysis of existing data and creating mathematical models consistent with the empirical results.

Medicine has perhaps a stronger tradition than most fields of critical examination of data and publishing of comments. Altman[6] emphasises the importance of this: "Many readers seem to assume that articles published in peer-reviewed journals are scientifically sound, despite much evidence to the contrary. It is important, therefore, that misleading work be identified after publication". Looking on the bright side, "Disputes over analyses and interpretation can be intensely creative: they drive researchers to generate new hypotheses and devise more refined experimental protocols" (Horton[7]). Goodman[8] observes that the common errors of science are not misconduct but disputed methods / analyses and interpretive uncertainty, and notes that the content of letters of comment may be more scientifically correct than the original paper. If datasets are to be critically examined, and results checked and new analyses undertaken, an important implication is that they be available in sufficient detail. In my view, it is highly desirable for authors to do this by tabulation or listing in the paper (rather than by archiving separately), and that journal editors should cooperate in this.

Acknowledgements

The Centre for Automotive Safety Research receives core funding from the Department for Transport, Energy and Infrastructure (South Australia) and the Motor Accident Commission (South Australia). The views expressed are those of the author and do not necessarily represent those of the University of Adelaide or the sponsoring organisations.

References

1. Svasti MRJ. ScienceAsia and its role in enhancing a research culture in Thailand. http://scienceasia.tiac.or.th/message/body_message.html
2. Svasti MRJ (2005) Thirty years of ScienceAsia, Journal of the Science Society of Thailand. *ScienceAsia* **31**, 1-3.
3. Hutchinson TP (2004) Statistics - for fun and therapy. *British Journal of General Practice* **54**, 228-9.
4. Hutchinson TP (2000) Graphing the survivorship of bees. *Insectes Sociaux* **47**, 292-6.
5. Hutchinson TP (1976) Statistical aspects of injury severity. Part II: The case of several populations but only three grades of injury. *Transportation Science* **10**, 285-99.
6. Altman DG (2002) Poor-quality medical research. What can journals do? *JAMA: Journal of the American Medical Association* **287**, 2765-7.
7. Horton R (1995) Revising the research record. *The Lancet* **346**, 1610-1.
8. Goodman NW (1996) Letter: Revising the research record. *The Lancet* **347**, 474.