Designing Social research for the 21st Century.

A Professorial address at the University of Bristol; October 14th 2002.

A gold standard?

The birth of modern statistics is often credited to R. A. Fisher, who introduced two key elements. The first of these was a systematic exposition of experimental design and the second was to lay the foundations for 'classical' statistical inference through the use of the 'likelihood' approach (Fisher, 1922). In this talk, I shall look at the ideas behind what is now accepted as good statistical design practice; my particular concern is with the analysis of data from the Social sciences. I shall argue that we need to rethink some of the assumptions that ordinarily go into the design of social research; that we need to make certain clear distinctions between social, (and much of medical and other research), and research in the natural sciences. In particular, I will highlight the crucial role played by the original ideas of experimental design, implicit in the term 'experimental' itself, in the area of agriculture and later incorporated into activities such as clinical drug trials and animal experimentation.

A key central concept in Fisher's expositions was the notion of 'randomisation'. If one can assume that the measurements being studied have a truly random distribution then, with a few more assumptions about conditioning, independence and variance structures, one can apply a suitable statistical model that then allows inferences to be drawn from the data about the values of 'parameters' and in particular about 'confidence intervals' for them. Likewise, one can derive 'hypothesis tests' to make statements about the existence of group differences etc. This notion of a random distribution for measurements survives throughout modern statistics; without it, there would be no statistical theory.

The importance of randomisation is its use *as a practical device for generating measurements so that they actually do satisfy the basic statistical assumption of a random distribution*. Thus, in an agricultural field experiment, as Fisher emphasised, fertilisers should be applied to different crop plots *at random* in order to satisfy the randomisation assumption of the statistical model used to analyse the subsequent data. If allocation is truly random, then the only reasons for differences in crop yields are those due to different properties of the fertilisers together with the chance variation arising from the randomisation process. The statistical model formally incorporates both these elements and Fisher's genius was to show, in a wide variety of situations, how the model could be used to yield inferences about the fertilisers, taking account of the random variation. Note that in this scenario the emphasis is on treating the random variation as 'noise' that is as essentially of no interest having been generated solely by the exigencies of the need to take account of inherent natural variation of crop growth in real fields. I shall return to this issue later, but for now note that this model of randomisation was hugely successful in agriculture, and subsequently in medicine, introduced by e.g. Bradford Hill for drug trials (Hill, 1951). I shall refer to it as the 'randomisation principle'.

The text books in statistics and experimental design adopted the randomisation principle, and the notion of a randomised controlled trial (hereafter referred to as an RCT) is now generally accepted as a 'gold standard' in applied statistical work. If one wishes to make sound inferences about differences between, for example, public health programmes or reading schemes, it is generally taken for granted that a properly conducted RCT is the best. In the area of so called 'evidence based medicine' (see for example Chalmers, 1993) this finds strong support, as well as in the Social sciences where it is far less common, and other approaches are typically regarded as second best. This notion of a 'gold standard' is closely tied up with the idea of making causal inferences, although they are logically distinct categories. Certainly, in the case of agriculture, or drug trials, establishing that, on average, one treatment is superior to another does not necessarily tell us anything about causal mechanisms. Nor, as I shall argue, is it necessary to have an RCT to draw causal conclusions. I shall also argue that the so-called 'gold standard' may, in some circumstances, turn out to comprise a rather baser metal.

Confounders

The notion of 'confounding' factors is fundamental in much of statistical inference. To illustrate this we can consider research carried out on the link between smoking in pregnancy and perinatal mortality; a link now well recognised in health education campaigns and even featured on cigarette packets although it was not always so.

In the 1970s, a number of large-scale studies disagreed about the relationship between maternal smoking in pregnancy and perinatal mortality. Table 1 summarises these disagreements (Goldstein, 1977 gives details).

Table 1. Percentage low birthweight by smoking category with ratios of mortality rates (mortality ratios) for six large-scale perinatal mortality studies. Ordered by mortality ratio.

Study	*% <2.5kg*		*Mortality ratio: smokers/non-smokers*
	Smokers	***Non-smokers***	
Rantakallio	6.1	3.5	1.01
Yerushalmy	6.4	3.2	1.03*
Niswander & Gordon	9.5	4.3	1.12
Ontario study	8.9	4.5	1.27
Butler et al.	9.3	5.4	1.28
Comstock et al.	11.1	5.9	1.40*

* Neonatal mortality.

The most striking thing about this table is the strong positive relationship between the mortality ratio for smokers/non-smokers and the percentage of low birthweight babies in the different studies. This relationship can be understood if we study the relationship between perinatal mortality and birthweight, which is given in Figure 1, adapted from Goldstein (1997).

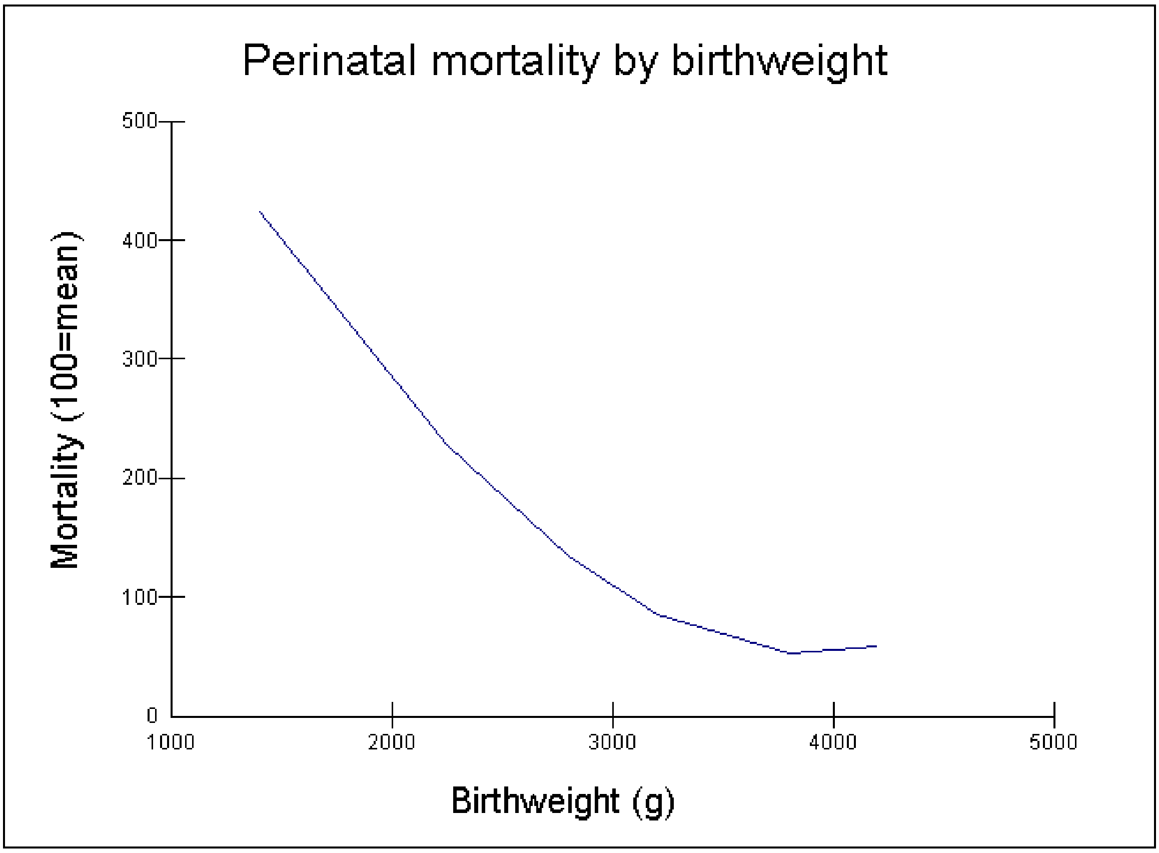

Figure 1. Relationship between perinatal mortality and birthweight (British Perinatal Mortality Survey, 1958)

The average reduction in birthweight when a mother smokes is about 200g. We see immediately from Figure 1 that for those mothers destined to have relatively heavy babies, above about 3500g, a reduction in birthweight will have little effect on the mortality risk, even improving it. For lower birthweight babies, however, a reduction of this amount will have a relatively large effect on the risk. Thus, in those studies where the percentage of low birthweight babies tends to be small, the overall mortality ratio will tend to be small, as is observed.

It appears that smoking acts on mortality largely by reducing expected birthweight. In this case, birthweight is an intermediary 'confounding' variable and an understanding of its role is crucial in interpreting results such as those in Table 1. Even in populations such as that of the Rantakallio study with no apparent average higher risk from smoking, there will be subpopulations, those with low birthweight, for whom the mortality risk can be expected to be much higher for smokers. Even if we could have carried out a RCT in this situation (and some have tried using random allocation in prenatal health education programmes) the crucial issue is that of understanding the underlying pathways of influence. What we have here is an illustration of the overriding scientific importance of replication under different circumstances, compared to which randomisation is an issue of lesser concern. It is also worth mentioning that human beings, unlike crops, may consciously interact with the 'treatment' they are given. Thus, in a randomised smoking cessation trial, it may be the case that there is a greater probability for those women destined to have heavy babies to cooperate with the study, whatever group they are randomly assigned to. Since these are the ones for whom smoking cessation matters least, the study will tend to underestimate the effect of smoking – an example of what is termed ‘informative non response’. Again, the application of randomisation cannot compensate for such a mechanism, and indeed may make the investigator unwisely optimistic about her inferences. Such remarks, of course, also apply to conventional drug trials, which rely upon the cooperation of patients. It is also worth pointing out that power calculations for such trials typically are concerned with detecting average effects, whereas what may be more important is the ability to explore interactions involving smaller subpopulations, and this will generally demand far larger samples.

The level of randomisation

Consider again an agricultural experiment with fertilisers applied to crops and suppose that yield is a function of soil type and that soil type varies systematically across the fields used in the trials, say from North to South. If a particular treatment was applied in a systematic fashion for example always in the North part of a field, then any ‘effect’ of that treatment might be spurious, simply reflecting soil differences. The allocation of fertiliser at random to plots across the fields guards against this *on average*. If a large enough sample, or set of trials, is used then even small differences due to fertiliser would be apparent without need for the standard apparatus of statistical inference. Thus, randomisation has the additional function of helping to guard against ‘confounding’ effects.

Suppose now that a further factor begins to affect crop yield in a particular fashion when a field overall contains more than a certain amount of a particular fertiliser per square meter, say more than 90% of the field is covered by this fertiliser. Such an effect might operate, for example, through some kind of influence on the behaviour of crop predators. This effect, if present, is clearly important commercially since in practice fertiliser will be applied on a whole field basis. Yet, the experiment, with its reliance on randomisation of treatments across plots within fields, will fail to detect such an effect; the randomisation creates an experimental situation that is artificial, so that results from the experiment will not necessarily apply in other environments. Of course, if we suspected that such a ‘field composition’ factor existed, we could modify our procedure, for example, by using a sample of fields and allocating fertilisers at random to fields. In medical trials, such designs are often known as ‘cluster randomised’ trials and I will return to these later. The point here is that the *level* at which randomisation takes place may matter, and this constitutes an important modification or elaboration of the principle of randomisation.

All of this is to emphasise the need to consider carefully the role of conventional wisdom about RCTs, and I now turn to a specific example from the social sciences where these and other considerations should cause us to think carefully about the randomisation principle.

Class size and achievement

I shall take as my illustration research into the relationship between pupil class size and progress in achievement. This has been one of the most researched areas of study in education, even though out of the thousands of such studies published perhaps only about 10 satisfy minimal standards of satisfactory design. These standards include having longitudinal data (Goldstein et al., 2000). Of these there is just one large scale RCT, known as the STAR study (Word et al., 1990) and I will look at this in some detail since it illustrates most of the points I wish to make.

The STAR study was carried out in the state of Tennessee in the late 1980s. There were 65 schools with at least three entry forms that were selected, and children randomly assigned to small (about 15), regular (about 25), and regular with teacher-aide kindergarten classes. The study children were followed for up to 6 years. Despite its size and expense (some $11million) it suffers from some notable weaknesses:

1. There is 'contamination' of effects among classes of different sizes within the same schools; i.e. lack of treatment independence. Applying a single treatment to each school (cluster randomisation) would have been better.

2. It is zero blind because all the participants including children knew the treatment to which they were assigned.

3. There was a lack of entry assessment to improve precision and inference details and to allow a check on the success of the randomisation procedure.

Children were assigned to small or large classes *within schools.* Inevitably, information about the progress of children in the different size classes will be available within the school and to all the responsible teachers, and so is likely to compromise the requirement for treatments to function independently. The nature of educational systems and social systems in general, is such that the complexity of their structures typically does not allow us to assume the independent operation of units within them. When an RCT changes such a structure in a research study, this implies, in a strict sense, that its conclusions can be accepted only, if at all, for populations with a similar structure. In order to generalise beyond such a structure would require an understanding of the interactions among the units at different levels within a population. In the case of the STAR study, this would require an understanding of how the interactions among teachers of different sized classes can influence teaching and learning. In fact, randomisation at the school level would have been better, although requiring more schools and creating serious logistical and possibly ethical difficulties (parents were allowed to switch classes after a year within the school – it would not have been possible to do this across different schools). Furthermore, such a study design would not in general apply to any real world population where there is inevitable variation within schools in terms of class size and where such variation may be determined by policies designed to take advantage of any demonstrated class size effects. Thus, the important requirement of generalisability would seem to be violated.

An important characteristic of most medical trials is that they are at least double blind, with neither the patient nor the administering physician knowing which treatment (or placebo) is being allocated to a patient. In the STAR study, we have an example of a zero blind study where everyone, including the children, was aware of the treatment to which they had been allocated. The usual reason for maintaining blindness is that knowledge of the treatment being administered, together with expectations about its possible effect, may of itself influence the outcome. For example, in one study (Shapson, Wright et al. 1980) over 90% of teachers were found to believe that larger classes produced worse results and this expectation seems to be prevalent in all educational systems. In the STAR study, this would apply to teachers particularly, but also to parents. Yet it is difficult to see how any experiment of this kind can avoid being zero blind, or at best single blind. To set up the study requires co-operation and the treatment, crucially, is at the group level; it is a social treatment rather than one applied to individuals and this is one of the important distinguishing characteristics of social research and it has important consequences, as I shall outline shortly.

The third issue is perhaps less important, yet symptomatic of the way in which researchers have sometimes relied too heavily upon the theoretical properties of randomisation to allow valid comparisons. Leaving aside the other problems, a comparison of test scores at the end of kindergarten year does allow valid comparisons to be made between the effects of small and large classes. These comparisons, however, can only be *average* ones; we cannot know for example whether initially low achieving children fared better or worse, relatively speaking, than initially high achieving ones (the answer seems to be, incidentally that the former have the most to gain from small classes (Blatchford et al., 2002)). This is a serious deficiency since it denies the possibility of certain kinds of information that may be socially important and is a good example of a naïve reliance upon the randomisation principle.

These problems represent both practical and theoretical difficulties with RCTs. There is, however, a further problem that, in some circumstances can lead to an RCT leading to a quite erroneous conclusion as a direct result of randomisation.

To see this take a hypothetical example and assume:

1. That the percentage of low achievers in the population = 10%.
2. That class size is not associated with achievement
3. That achievement is lowered where a class has at least 33% of low achievers.

With random allocation, we have the probabilities in Table 2.

Table 2. Probability of observing low achieving class for classes of different sizes with randomly selected pupils, where the overall percentage of low achievers in the population = 10%.

Class size	Prob. of class $\geq$ 33% low achievers
15	0.013
20	0.0023
25	0.0005

In a study with 200 classes size 15, and 200 classes size 25:

Expected percentage of 1 or more 'low achieving' class among large classes = 10%

Expected percentage of 1 or more 'low achieving' class among small classes = 93%

Thus, in most (93%) studies there will be at least one small class with lowered achievement but lowered achievement for one or more large classes will occur in only a small minority (10%) of the studies. The inevitable inference would be that small classes tend to lower achievement, whereas in fact this is spurious and occurs only because randomisation has made the occurrence of the triggering event (a high percentage of low achievers) extremely unlikely in large classes. For the same reason, randomisation makes it impossible to study the effect of this compositional variable in large classes because the event is so rare. By contrast, if the real population contains sufficient large classes with the compositional variable in operation then a purely observational study would be expected to allow such a comparison, and hence exhibits a theoretical advantage over a randomised study. This example of how the randomisation principle can lead to misleading inferences by altering the composition of a group is not confined to educational settings and can be found elsewhere in the social and medical sciences where compositional effects operate.

More than on average

I have already argued for the scientific and practical importance of focussing on ‘interactions’ between the ‘treatment’ of interest and other variables, so that for example we study how subgroups are affected. Another aspect of this is in the study of variability more generally. Traditional statistical analysis has concentrated largely on models that explain or predict the average value of a response, such as educational achievement, in terms of other factors, such as class size or social background. The remaining ‘unexplained’ variation is typically regarded as ‘noise’ and of little intrinsic interest. Consider, however, the STAR study again. In a reanalysis, Goldstein and Blatchford (1998) show that in the second year of schooling the difference in reading progress for Black children between those in small (15) as opposed to large (25) classes amounts to about one fifth of a standard deviation of the test score; it is effectively zero for White children. This is an interesting differential finding, but in addition, they show, using a multilevel model, that this difference is not constant, varying from school to school. Moreover, the between-school standard deviation of this difference is relatively large, about one quarter of a standard deviation, so that the effect for Black children appears to be very small or even negative in some schools and very large in others. This may of course reflect the vagaries of the study design, but might also suggest that there are other mediating variables influencing the class size differences. Either way, such information on variability needs to be taken into account when interpreting and acting upon the results of an analysis. In particular, if policy is formulated based on the average group differences we should not expect it to be effective in all circumstances.

Whither?

I have argued against the automatic adoption of randomisation as a gold standard; so, what is its status and what could we put in its place? One of the attractions of randomisation is that it appears to be a conceptually simple method of coming to replicable conclusions. It also often provides a means of presenting results in a relatively simple form. I have also argued that, even where randomisation is relevant or even crucial, this does not relieve the data analyst from the modelling of any real underlying complexity. I want to conclude by saying a little more on the issue of complexity.

There is an important distinction between the complexity of the system being studied, for example the factors associated with class size and progress in achievement, and the complexity with which the results of an investigation are summarised and reported. Thus, it is unnecessary to understand the complexities associated with selecting a representative survey sample, adjusting for non response bias and adjusting standard errors in order to appreciate the results in terms of estimated percentages having, say, particular voting intentions. Likewise, one does not need to follow the intricacies of a multilevel analysis to appreciate findings on the effects of class size. In both cases, however, it is necessary to perform the technicalities in order to ensure that any results are as robust as possible and that statements can have good estimates of statistical uncertainty attached to them, for example in the form of confidence intervals.

Some have argued (see e.g. FitzGibbon, 1995) that complex modelling is unnecessary since for many purposes much simpler approaches produce similar results. While this view has some force, particularly where there is a lack of adequate software or expertise to carry out more complex analysis, it is essentially misguided. Encouraging people to apply simple models to complex systems is likely to encourage a view that the systems *really are* simple, with all the dangers which that brings. The use of simple models, for example ordinary regression, when, say, complex multilevel structures are present, will tend to hide subtleties which, as I have suggested, may be among the most interesting aspects of the data. Rather, what is needed is the development and especially the application of statistical techniques to data at a level of sophistication that attempts to capture the key elements of the complexity that exists in the real world. At the same time, those who carry out such analyses have a responsibility to communicate them in ways that are intelligible, without being oversimplified.

Implementing such a perspective is easier said than done, and one might ask who would be able and who willing to do this? We certainly cannot look to politicians in power who, overall, seem more concerned to obfuscate than to illuminate issues. The media, at times, will recognise the issues, but their own general lack of understanding of quantitative matters often hinders their reporting. This leaves the universities, the learned societies and similar bodies, and this is, in my view, one of the major challenges facing such institutions today. Most especially, those who are entrusted with education, at all levels, should be prepared to do two things. The first is to be clear about their commitment to the uncovering of social complexity and the avoidance of facile oversimplification. The second is to struggle to maintain such a commitment in the face of indifference from the powers that be, as well as the outright hostility of those for whom the height of intellectual achievement is a well-spun sound bite.

Thank you for listening.

Harvey Goldstein

References

Blatchford, P., Goldstein, H., Martin, C. and Browne, W. (2002). A study of class size effects in English school reception year classes. *British Educational Research Journal* **28**: 169-185.

Chalmers I. (1993). The Cochrane Collaboration: preparing, maintaining and disseminating systematic reviews of the effects of health care. In: Warren KS, Mosteller F, eds. Doing more good than harm: the evaluation of health care interventions. Ann NY Acad. Sci.; 703:156-63.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Phil. Trans., A,* 222, 309-368

FitzGibbon, C. (1995). *The value added national project (general report)*. London, School Curriculum and Assessment Authority.

Goldstein, H. (1977). Smoking in pregnancy: some notes on the statistical controversy. *British Journal of Preventive and Social Medicine* **31**: 13-17.

Goldstein, H. and Blatchford, P. (1998). Class size and educational achievement: a review of methodology with particular reference to study design. *British Educational Research Journal* **24**: 255-268.

Goldstein, H., Yang, M., Omar, R., Turner, R., et al. (2000). Meta analysis using multilevel models with an application to the study of class size effects. *Journal of the Royal Statistical Society, Series C* **49**: 399-412.

Hill, A. B. (1951). The clinical trial. *British Medical Bulletin*, 7, 278-282

Shapson, S. M., Wright, E. N., Eason, G. and Fitzgerald, J. (1980). An experimental study of the effects of class size. *American Educational Research Journal* **17**: 144-52.

Word, E. R., Johnston, J., Bain, H. P., Fulton, B. D., et al. (1990). *The state of Tennessee's student/teacher achievement ratio (STAR) project: Technical report 1985-90.* Nashville, Tennessee State University.