SSC Annual Meeting, July 1995
Proceedings of the Survey Methods Section

SOCIAL SURVEYS AND SOCIAL SCIENCE

T.M.F. Smith[1]

ABSTRACT

Random sample surveys have become one of the key research tools in quantitative social science. Some of the claims for statistical methods in general, and for sample surveys in particular, have been exaggerated. The limitations of the survey method will be explored. Inferences from sample surveys fall into two classes, descriptive (enumerative) and analytic (explanatory). The role of these two types of inference will be examined and related to the controversy between randomisation and model-based inference. An alternative role will be proposed for sample surveys within the social sciences.

RÉSUMÉ

Les enquêtes à partir d'échantillons aléatoires sont devenues un des outils importants de recherche dans les sciences sociales qualitatives. Certaines prétentions concernant les méthodes statistiques en général, et les enquêtes à partir d'échantillons aléatoires en particulier, ont été exagérées. Les limites de la méthode d'enquête vont être examinées. Les inférences à partir d'enquêtes se partagent en deux classes, descriptive (énumérative) et analytique (explicative). Le rôle de ces deux types d'inférences va être examiné et relié à la contreverse entre la randomisation et l'inférence basée sur un modèle. On propose un rôle alternatif pour les enquêtes à partir d'échantillons à l'intérieur des sciences sociales.

1. INTRODUCTION

Statisticians sometimes argue that Statistics (capital S is used for the discipline) stands at the centre of the scientific universe, that it is a fundamental discipline for all scientific enquiry. Kish (1978), in his presidential address to the American Statistical Association said

> *"Statistics is a peculiar kind of enterprise of contradictory character because it is at the same time so special and so general. Statistics exists only at the interfaces of chance and empirical data. But it exists at every such interface, which I propose to be both necessary and sufficient for an activity to be properly called statistics. It has a special and proscribed function whenever and wherever empirical data are treated; in scientific research of any kind; in government, commerce, industry, and agriculture".*

Smith (1977) has described Statistics as a universal discipline, and more recently Morris (1995) recommended a hub and spoke model for Statistics within a university with Statistics at the hub. Stigler (1986) recognised that this all-embracing view of Statistics must have limits.

> *"If all sciences require measurement - and statistics is the logic of measurement - it follows that the history of statistics can encompass the history of all of science. As attractive as such an imperialistic proposition is to a statistician, some limits must be imposed."*

I have recently observed two examples of the consequences of the imperialist view which have frightened me. In the UK some medical statisticians have asserted that they should have a power of veto over the publication of quantitative medical research if it fails to meet their standards. On a similar theme the failure of the recent opinion polls in Italy has led to the proposal that polls based on "unsatisfactory methodologies", such as quota samples, should be banned. The statistician as censor is not a position that I can support. How can the discipline that can never say it is certain be so sure of its position?

Not all statisticians are convinced about their real value. John C Bailar III (1995), discussing what the customer needs in statisticians, states:

> *"As academic statisticians, we are missing the boat. We are barking up the wrong tree. We do not see what is plainly before us. We are kidding ourselves when we think that 'our' kind of statistics is vital to the welfare of the nation and the world. Think about the whole range of the really big problems of the day: violence, crime and criminal justice, education and industrial productivity in the broadest senses, unemployment, the balance of trade, federal deficits, the health and welfare of millions of disadvantaged persons, urban rot, racial and ethnic*

[1] T.M. Fred Smith, Dept. of Mathematics, University of Southampton, Southampton, S09 5NH, England, United Kingdom, e-mail: tmfs@maths.soton.ac.uk.

tensions, homelessness, and many others. The kinds of statistics that we teach in undergraduate and especially in graduate programs have almost nothing to contribute to anything that matters on the scale of these problems."

In the spirit of Bailar's contribution I too would like to play devil's advocate about the importance of Statistics. If Statistics is such an important discipline why is it that statisticians have so little influence in the world in general and in the scientific world in particular? What is the value-added of Statistics within scientific enquiry? Is the value-added necessarily positive? Does Statistics have a role in tackling the social problems highlighted by Bailar?

2. STATISTICS AND SCIENCE

Science existed before Statistics was invented and would continue even if Statistics were to disappear. In discussing the relevance of Statistics it is convenient to distinguish theoretical science from empirical science. Theories are attempts at explanation and useful theories lead to predictions, either quantitative or qualitative, which can then be validated. Very useful theories can be expressed mathematically which then allows for a wide range of deductions. The Social Sciences are replete with theories. For example in demand theory it is asserted that if the price of a good rises, other things remaining the same, then the demand will fall. The near impossibility of actually holding other things constant means that the quantitative measurement of the effect of a price rise is well nigh impossible; but as a qualitative statement the theory still has great value. The construction of theories which explain phenomena is one of the most creative activities of scientists, and although Probability theory, which is a branch of Pure Mathematics, has a role in some theories, I am not aware that Statistics, as such, contributes to this form of creativity.

Empirical science involves observation and measurement, both of which may introduce errors. From the point of view of statistical analysis the process generating the data combines both theory, expressed through a model, and error. The model belongs to the scientist, and it is only in the treatment of the errors that a statistician can claim any expertise. But how often do statisticians concentrate on the distribution of the errors? Too frequently we turn our attention to the model and propose model-fitting criteria, such as least squares, which lead to error distributions with variances far smaller than can be justified by the variances due to measurement errors. If the error variances are too small then we may fail in our duty to warn against the dangers of misinterpretation in the presence of error. Errors are a nuisance and concentration on error makes us the Jeremiahs of science.

I conjecture that statisticians in their Jeremiah role, when they fail to estimate the correct contribution of errors, or when they assert that effects are not statistically significant, which is then interpreted to mean not significant, may sometimes actually inhibit scientific progress. If my conjecture is true then it is possible for the value-added of Statistics to be negative. However, statisticians acting as scientists, as members of a team, can frequently make useful scientific contributions, especially in the area of applied probability. My point is that these are scientific contributions in the sense of applied mathematics, not inferential statistical contributions.

Theorising is an expression by scientists of their personal knowledge. Such knowledge must be strongly believed by the scientist and by colleagues if it is to command support. Others may not share this belief and they may express their uncertainty by carrying out experiments aimed at refuting the theory. If there are measurement errors present then Statistics may have a role, but I find that many of my colleagues in the physical sciences choose to ignore the potential of Statistics. They prefer to design experiments with very high signal to noise ratios so that statistical inference becomes irrelevant. The design of a good refutation experiment is yet another highly creative area of Science where Statistics may have little to contribute.

Hypothesis formation and refutation are but a small area of Physical Science. Most experiments are not of the highly creative "paradigm shift" type, rather they are routine experiments aimed at making deductions on the assumption that the underlying theory is true. These deductions are frequently descriptions or classifications of physical material based upon the premise that the theory is true. Experiments of this type constitute the bulk of scientific work. They are the contributions of the artisans of science upon which the great discoveries rest, and in these experiments the role of Statistics is often limited to the problem of handling measurement errors since the models on which analyses are based are not in dispute. Unknown structures are estimated from the signals recorded by the measuring instruments, and if these signals are generated by a stochastic process then statistical methods can be applied. Measurement errors are controlled either by taking many replicated measurements or by systematically improving the measuring instruments, and this is where scientists put their effort. If sample sizes are small, with a small signal to noise ratio, then statistical methods become appropriate, but for the majority of these routine experiments there may be little need for statistical advice.

The data in the Life and Social Sciences may differ from the precisely measured data that feature in much of Physical Science. Stigler (1986, p. 309) quotes an important distinction made by Edgeworth (1885) between different classes of data:

"Observations and statistics agree in being quantities grouped about a Mean; they differ, in that the Mean of observations is real, of statistics is fictitious. The mean of observations is a cause, as it were the source from which diverging errors emanate. The mean of statistics is a description, a representative quantity put for a whole group, the best representative of the group, that quantity which, if we must in practice put one quantity for many, minimizes the error unavoidably attending such practice. Thus measurements by the reduction of which we ascertain a real time, number, distance are observations. Returns of prices, exports and imports, legitimate and illegitimate marriages or births and so forth, the averages of which constitute the premises of practical reasoning, are statistics. In short observations are different copies of one original; statistics are different originals affording one "generic portrait". Different measurements of the same man are observations; but measurements of different men, grouped round l'homme moyen, are primâ facie at least statistics."

Edgeworth's distinction between observations, or measurements, and statistics is an important one which suggests to me that different statistical approaches may be needed for the different cases. Today the choice of words seems unfortunate since modern usage would describe both observations and statistics as statistics. Edgeworth's terminology merely reflects the historical development of the subject at the end of the 19th century with the word statistics arising from work in political arithmetic and the description of the State, and observations arising from the need for the combination of observations in the realm of scientific measurement. In the 19th century statistics, in the Edgeworth sense, were based either on censuses or administrative records, which were assumed to give complete records, there was no need to consider uncertainty due to sampling and there was little consideration given to the uncertainty in the numbers themselves. Of course there could still be uncertainty in some of the inferences made from statistics and I will return to this point later.

Many others have recognized the need for statistical methods to reflect the processes that generate the data. Fisher (1956, p. 77) in his book "Statistical Methods and Scientific Inference", note the title, carried on his long-standing argument with Neyman about hypothesis testing in the following terms:

"In attempting to identify a test of significance as used in the natural sciences with a test for acceptance, one of the deepest dissimilarities lies in the population, or reference set, available for making statements of probability. Confusion under this head has on several occasions led to erroneous numerical values; for, where acceptance procedures are appropriate the population of lots of one or more items, which could be chosen for examination, is unequivocally defined. The source of supply has an objective empirical reality. Whereas, the only populations that can be referred to in a test of significance have no objective reality, being exclusively the product of the statistician's imagination through the hypotheses "

Fisher is distinguishing between a null hypothesis in a scientific context, which leads to a pure test of significance based on a hypothetical population, and a test based on sampling homogeneous material from a real population, which leads to a decision. He is arguing that the repeated sampling concept is potentially relevant to the latter problem but not to the former. Another distinction is between the processes generating the data. In a scientific context there is an unknown process, (Nature), which generates the data, and the scientist's model of this process is a product of human imagination. In industrial sampling the population is real, it actually exists, and the uncertainty is due to drawing a sample rather than carrying out a complete enumeration. This distinction will become vital to my argument when I consider Social Science.

Anscombe (1961) makes a similar distinction but expresses it in terms of the decision process.

"When the choice turns on the procedure in a decision to be used impartially in an automatic way (as, for example, a project of public-opinion sampling by a Government department), the analysis made by an orthodox statistician may be good and pertinent: the Bayesian statistician will accept such an analysis and can add very little to it. On the other hand, when the problem concerns a single intelligent decision (such as, should this new idea be actively developed or set aside?), it is not clear that the orthodox statistician's analysis can have much deciding force; often only the Bayesian can enlighten and assist the workings of common sense."

These quotations suggest that the context in which statistical inference takes place can affect the framework for inference. Placing statistical analysis within a social context may help to resolve some of the conflicts between different schools of Statistics. An inference in the public domain does not necessarily have to satisfy the same set of conditions as an inference in the scientific domain or in the private domain. Bayesian methods would seem to be appropriate for personal inferences, including scientific inferences where only a personal reputation is at stake, but are not necessarily appropriate in the public domain, for example in the reporting of Official Statistics. A

similar distinction exists in Medicine in the conflict between individual ethical conduct and collective ethical conduct which has been highlighted so dramatically in the discussions of the ECMO clinical trials, see for example, Royall (1991). Bayesian inferences have been adopted by those who argue for individual ethics, and classical randomization inferences for those who take a collective view. Although I favour Bayesian methods for personal inferences, the difficulties of getting collective agreement about priors, and even more importantly about likelihoods, seem to rule out these methods for collective problems. Where randomization has been used the randomization distribution can be employed for inference provided that there is collective agreement about the framework and interpretation of the resulting inference.

3. SCIENCE AND SOCIAL SCIENCE

Some people argue that Social Science is not a Science. I am not concerned here with some abstract concept of Science, rather I am interested in the human activity of scientific research and its communication. A scientific study requires the use of systematic methods which can be replicated by others and which admit the possibility of refutation. Only if a study uses reproducible techniques which are adequately recorded using a common language can it be replicated.

Science starts with the systematic classification of some phenomenon and with the naming of the classes. This provides the common language which makes consistent communication possible. Without agreed classes and an agreed language there can be no communication; terms cannot be defined unequivocally and there can be no replication and hence no chance of refutation. Defining concepts in the Social Sciences seems to me to be much more difficult than it is in the Physical Sciences. What is the rate of interest? There are different rates in different money markets and they don't necessarily change simultaneously. Failure to specify the market context is a failure to communicate. When is a person unemployed? I was a member of a working party of the Royal Statistical Society (RSS, 1995) which reported recently on the Unemployment Statistics in the U.K. We discovered that there could be many different definitions of whether a person was employed, unemployed, or not in the labour force. Considering labour as a factor of production a person working part-time who wishes to work full-time has unemployed amounts of labour, but under most definitions would be classed as employed. Within a social context a person who has been unemployed for a long period and has given up searching for work is clearly unemployed but under the widely used ILO definition is classified as not in the labour force. Again my conclusion is that the concept varies according to context, and without specifying the context there is a lack of communication. This has certainly been the case in the UK in recent years and has led to a loss of confidence in Official Statistics.

After classification the next stage of a Science is frequently the systematic observation and measurement of a phenomenon in order to describe its distribution in space and time. Description reveals relationships and relationships pose questions about their stability and about causation. Explanation of relationships is the object of theory, and this requires the modelling of the processes that generate the phenomena represented by the classes. There is general agreement about the laws of Physics and so modelling of physical phenomena is carried out within the context of these laws. In the Life Sciences there are fewer agreed laws and some concepts, such as evolution, imply change which in turn depends on the context or environment. Context dependent phenomena are much harder to model since outcomes may depend on both the process being studied and the context, historical and/or spatial, in which it is studied. Models can be used to make predictions, but a prediction made from a model fitted within one context may be invalidated by a change in context, and this makes the testing, validation and refutation of models very difficult.

The position in the Social Sciences is even worse. The term social implies that collections of individual units can combine in different ways, and that the context is essential to the understanding of social phenomena. In fact context puts the Social into Social Science. The study of the outcomes from individual units may be useful for describing social phenomena but is unlikely to lead to complete understanding. If predictions are context specific then replication may be impossible. If replication is impossible then independent studies that lead to the possibility of refutation become impossible. When the problems of social context are allied to the near impossibility of controlled experimentation the difficulties of determining the laws (if any) of Social Science are manifest. Useful theories should be robust to their context if they are to have predictive value, but if the subject matter itself is part of the context then predictions may have to be so restricted by caveats that they become empty and hence the theories become untestable.

The complexity of social processes is a challenge for both social scientists and statisticians. The generally agreed failure of macro-econometrics to provide an adequate understanding of economic phenomena and to make reliable predictions upon which successful policy could be based suggests that even complex aggregate models cannot suffice. The recent move to micro-econometric models is a reaction to this but models based on individual units assumed to be independent of context will also not suffice. The increased use of multi-level models, see for example Goldstein (1995), is a recognition of the need for social models to embrace context. In the past complexity has not lent itself to reliable quantitative

analysis and perhaps future work should concentrate more on verifiable qualitative analysis.

4. SAMPLE SURVEYS

Sample surveys are one of the main methods for collecting data in social investigations. Before analysing the role that sample surveys might play in a future Social Science research agenda it is worth putting them into their historical context. The case for sample surveys of people, social surveys, as an alternative to censuses was first put before the statistical community by the Norwegian statistician, A N Kiaer, at a meeting of the International Statistical Institute in Berne in 1895. The most accessible reference is a reprint of Kiaer (1897) produced by the Central Bureau of Statistics of Norway in 1976 to celebrate their centenary. Kiaer argued that in many fields, such as Geology and Biology, it was impossible to carry out a complete census and so sampling was generally accepted as a method for collecting information. Inferences could then be made from the sample data to the whole population or area under study. He proposed that the same principles could be employed in social research when it was either too expensive or too time consuming to carry out a complete census. He demonstrated by examples, based on his experience in Norway, that carefully selected samples could be representative of the population from which they were drawn. Nowadays such statements seem unexceptional, but at the time they drew serious criticism and it was not until 1903 that the ISI agreed to report on applications of, and the validity of, the representative method. We then have to wait until 1925 for the report to appear.

To understand the reasons for the opposition to Kiaer one has to study the lines of development of social research during the 19th Century. At the beginning of the century Laplace had proposed replacing the population census in France by a sample augmented by administrative records. Influenced by Laplace the Belgian statistician, Quetelet, proposed the same strategy for the Low Countries. He was dissuaded from this by Baron de Keverberg (Stigler, p. 165) who argued that:

> *"The law regulating mortality is composed of a large number of elements: it is different for towns and for the flatlands, for large opulent cities and for smaller and less rich villages, and depending on whether the locality is dense or sparsley populated. This law depends on the terrain (raised or depressed), on the soil (dry or marshy), on the distance to the sea (near or far), on the comfort or distress of the people, on their diet, dress, and general manner of life, and on a multitude of local circumstances that would elude any a priori enumeration."*

This argument that homogeneous groups either do not exist, or are so small that only a complete census can provide adequate estimates, pervades social research to this day. Despite accepting Keverberg's argument about the census, Quetelet appears to have devoted most of his subsequent social research to demonstrating that homogeneity did exist in population distributions because mixtures of the outcomes of many homogeneous binomial causes led to Normal distributions. Others disputed Quetelet's claims, and Lexis demonstrated through his index of dispersion that the assumption of homogeneous Bernoulli processes leading to Normal distributions, could not be upheld. Stigler (1986, p. 7) summarises the position well when he says that

> *"Quetelet found homogeneity everywhere, Lexis found it nowhere."*

Later when Pearson's goodness of fit test was developed it was discovered that few of Quetelet's distributions appeared to be Normally distributed after all, and so Lexis' position was upheld. Birth ratios are an exception to Lexis' claim since they are genetically determined.

Another line of social research which started in the 19th Century was investigation by case studies, or as it was then called, monography. This recognised the complexity of social processes by carrying out in-depth studies of a few groups of individuals or families. There were major schools of social research devoted to monography in continental Europe, with the study of family budgets being of particular interest, see Lazarsfeld (1961).

Kiaer was faced with two alternative schools of social research. Both argued that social processes were too complex to be studied reliably by a mere sample survey, but the nature of the complexity differed. The census school were concerned with coverage, with the complex spatial variation of a few simple variates, whereas monographers were concerned with the complexity of behaviour within groups such as families. One school counted all the families while the other counted all the transactions within a small group of families. Kiaer's counter arguments were persuasive for the problem of the estimation of population aggregates but did not address the issues of complexity. He argued pragmatically that it was impossible to carry out censuses for many of the problems that he had considered and so some form of sampling was essential if any information was to be obtained. Case studies were inadequate because they were not representative and so inference to a wider population was unreliable. In fact he suggested that monography should be based on representative samples. However, at no time did he deny the value of both censuses and monographs for social research.

Why did Kiaer's arguments not persuade social researchers in 1900 but were successful in 1925? One reason is that due to the pressures of World War I,

governments needed more information. It was the demand for official statistics that stimulated the change. I think that in addition that there were two weaknesses in Kiaer's position. The first is that he did not propose a general method for drawing representative samples, rather he relied on a few case studies. These cases lacked conviction, for example he chose a supposedly representative sample of families by selecting the initial letters of surnames. The second weakness is that he did not have a theoretical framework for inference, in particular he could not measure the uncertainty due to drawing a sample. These were fundamental weaknesses, but by 1925 both had been overcome. In the 1925 ISI report simple random sampling and stratified random sampling are recommended as methods for drawing representative samples, and in an Appendix to the report Bowley (1926) shows how the uncertainty due to random sampling can be measured. Bowley offers two forms of inference, both based on an expansion of the hypergeometric distribution of a proportion estimated from a simple random sample. The first asks which values of the population proportion, P, are consistent with the sample data, and answers it by evaluating the likelihood function of P. The second asks what inference can be made about P and evaluates the posterior distribution. Bowley then shows that with a flat prior you would draw much the same conclusions from both forms of analysis.

The ranking of the three main methods of social investigation with respect to coverage and complexity of relationships is shown in Table 1.

Table 1
Ranking of methods for Social Investigation

Method	Ranking for	
	Coverage	Complexity
Censuses	1	3
Representative Survey	2	2
Case Studies	3	1

This table shows that representative surveys are a compromise solution for both coverage and complexity. Compromises do not necessarily lead to the right solution and this point will be further explored in the final section.

Bowley's paper is a tour de force, but his approach did not generalize readily to the analysis of more complex sampling schemes such as those based on selecting clusters of units with possibly unequal probabilities. Neyman (1934) provided an answer with his introduction of the randomization distribution as a framework for inference. This distributional assumption, together with the related idea of confidence intervals, was rapidly adopted by survey statisticians, especially those working on official statistics. By the 1950's randomization theory had been fully worked out for a single cross-section survey with the inferences being based on the central limit theorem. The history of sampling shows the importance of a theoretical framework for the acceptance of new ideas.

5. SAMPLE SURVEY INFERENCE

Randomization inference is quite distinct from most other forms of statistical inference, as the randomization distribution is completely specified by the statistician and there are no unknown parameters to be estimated in the distribution. It is nearly devoid of inferential interest. Much of the early research was carried out by government sampling statisticians, such as Deming, Hansen, Hurwitz and Madow, working in the Bureau of the Census. Godambe (1955 and 1966) created interest in the wider statistical community when he examined the foundations of survey inference and showed that there could be no best estimators and that the likelihood function gave no information about population units not in the sample. This provided the stimulus for others, such as Ericson (1969), to develop alternatives to randomization inference, and started the randomization theory versus model-based theory controversy to which many have contributed, see for example, Smith (1976). At that time I came down firmly on the side of models and likelihood based inference, including Bayesian inference. My interest was in social research and in the search for models that would explain social phenomena, with surveys providing one of the main sources of data. I was concerned with the analytic use of surveys, not with their descriptive use.

A survey inference is descriptive if the object is to describe a given finite population. The defining criterion is a perfect census. If the results of a perfect census were to leave one with no inferential uncertainty then an inference from a sample to the census population would be descriptive. Any inference to a different population, either a real population at a different time or place, or a hypothetical population, such as a superpopulation that might have generated the data, is called analytic. Social researchers are usually interested in analytic inferences, they wish to make general statements, while an important role for official statisticians is the production of a trustworthy descriptive data base which can be used by both governments and the public for aiding decisions. Governments determine policy and electorates judge the policies partly on the basis of official statistics. If the official data base is not trusted then the democratic process is eroded.

This difference in interest may help to explain aspects of the model-based versus randomization controversy. Those working closely with official statisticians tend to

favour randomization inference, whilst those working in social research tend to favour models. If, as Fisher and others seem to imply, different forms of inference are applicable to different classes of problems, if there is no single correct form of inference that solves the problem of induction, if it is accepted that inference is only a product of human imagination and that any such product must therefore be fallible, then it may not be unreasonable to propose that the randomization distribution has a role to play in descriptive inference while at the same time admitting that model-based methods are essential for analytic inference. This is the conclusion I reached in Smith (1994) and I now expand on the arguments I employed in that paper.

The positive arguments in favour of randomization inference are that the randomization distribution has objective reality, that it depends on very few assumptions and that it is robust. None of these arguments is overwhelmingly strong. Although the finite population is real, and it is possible to conceive of all the samples that might have been drawn, the reality is that only one sample will be drawn. Isn't it the inference from this single sample that is relevant? The assumptions of randomization inference are that the uncertainty arises from random sampling but in practice the greatest uncertainties arise from non-sampling errors, which may include biases at least as great as the errors due to sampling. Non-sampling errors cannot be analysed using the randomization distribution and so model-based methods must be employed. Although the randomization distribution conceivably has objective reality, in practice randomization inferences depend on an assumption that the distribution of estimators can be approximated by a normal distribution. The usual arguments for this approximation are asymptotic, being special versions of the central limit theorem, and break the link with the reality of the finite population. Finite population versions of the normal approximation, see for example Robinson (1978), do not appear to give very useful results. It is very easy to construct finite population counter-examples to the normal approximation by introducing rare events into the set of population values which ensure that normal distribution confidence intervals have very poor coverage properties. Finally Brewer and Särndal (1983) expose the emptiness of the robustness argument when they say:

> *"Probability sampling methods are robust by definition; since they do not appeal to a model, there is no need to discuss what happens under model breakdown."*

Some arguments for model-based inference are that the inference should be based on the sample actually drawn and should be made conditional on that sample; that inferences should obey the likelihood principle; that randomization inferences do not obey this principle or any other commonly agreed principle; that all the non-sampling errors depend on model-based assumptions; that only a model can provide the link between the values in the sample and those in the rest of the population, and that analytic inference must be model-based. The principal argument against model-based inference is that it depends on models. Hansen, Madow and Tepping (1983) showed how small model misspecifications could have a significant effect upon model-based inferences. They demonstrated, as have many others before and since, that model-based methods are not robust to model misspecification. The claims for the optimality of some model-based inferences depend either explicitly or implicitly on the assumption that the model is true. In the area of Social Science it is rather difficult to believe that any model is actually true and so claims of optimality are not really relevant.

My own position has been strongly affected by my growing unease about the use of models within the Social Sciences. My reading of the history of social research has convinced me that Lexis was right and that social processes are not homogeneous. If social processes are generated by stochastic systems then they would have to operate at high levels of disaggregation, possibly at the level of the individual unit. Homogeneity is the exception, not the rule. The social element also requires that the context within which individual units operate would have to be modelled. My pessimism about models within the social sciences does not mean that I oppose modelling, models are essential for analytic survey inference. It is only in the area of descriptive survey inference, in the provision of generally accepted databases, that I think a case for randomization inference can usefully be made. However, many models are essentially descriptive; they are parsimonious ways of summarizing complex sets of data rather than explanations, and so the application of descriptive inference may be wider than sometimes appears.

We are all influenced by the context of our own lives. My increased involvement with official statistics during the last decade, in particular observing the erosion during the 1980's of public confidence in official statistics within the UK, has influenced my views about statistical inference. It is rarely possible to validate publicly produced official statistics, or any other statistics, but if a democracy is to operate efficiently, if as electors we are to be able to make our judgements on the performance of governments on the basis of publicly produced data, then it is essential that we have confidence in those who produce the data. Fellegi (1989) addresses this issue:

> *"Statistical information is a product with peculiar attributes. One of them is that users are seldom in a position to check its quality directly. Yet data that are not trusted are clearly of little utility, whatever their intrinsic quality. Short of direct*

quality checking, the degree of confidence that users attached to the product is necessarily a direct function of their confidence in the producer."

Confidence in the producer is the only route to confidence in the statistical product and this confidence is enhanced if we can trust the procedures used to produce the published data. These procedures include the selection of samples, the handling of non-sampling errors and the rules for making inferences. In Smith (1994) I use the term procedural inference to describe the repetition of sampling and estimation that leads to the randomization distribution, and the term scientific inference to describe the use of model-based methods of analytic inference. Procedural inferences have little scientific content, they merely describe what is there in the population. So it is not necessary to employ the methods of scientific inference for this descriptive task, procedural inference will suffice.

Trust in procedures will engender trust in descriptive inferences. Wherever samples are to be drawn the procedures for drawing the samples and making the inferences should be determined in advance and written down. The possibility of manipulation by the survey analyst, after the sample is drawn, must be avoided. Model-based methods often involve data dependent model choices which may depend on the subjective decisions of a statistician after the sample has been obtained. Such methods could be abused and so are less trustworthy than pre-specified objective methods based on procedures. Model-based distributions have no objective reality whereas for randomization inference both the random selection rule and the estimators can be specified in advance and the resulting randomization distribution has some objective reality which has a possibility of comprehension. The issue is not that of statistical efficiency it is that of trust. Alexander (1994) describes how the US Bureau of the Census explains its procedures in a Source and Accuracy Statement:

"Our final step in practical inference was explaining the confidence interval to a general audience. This is the sort of thing we say in our Source and Accuracy Statement:
'Keep in mind that the particular sample we selected was one of many possible samples ... Different samples would give different results'.
'If many samples were selected, then for approximately 95% of the samples the confidence interval which would be calculated would contain the result which would be obtained by surveying the entire population'

It sounds like someone's worst nightmare from Freshman Statistics, but it does two things:

(a) It gives the readers a concrete image to reinforce the notion that there is uncertainty because of sampling error.
(b) It precisely and completely states the fact upon which we expect the readers to base their statistical inferences about sampling error."

Alexander then challenges someone to draft a Source and Accuracy Statement for a general audience from a model-based perspective. The difficulty of doing this is one of the reasons for using randomization inference within the public sector of statistics.

The main problem with randomization inference is that of conditional inference. In principle the sampling distribution is determined by the sampling rule and the only valid form of randomization inference is to use the unconditional distribution generated by samples from the whole population. Conditional randomization inference requires disaggregation and the consideration of samples conditional on the achieved sample size in sub-groups. As argued above it is quite plausible that the only homogeneous groups in the sample are of size one. This level of conditioning then nullifies the randomization distribution. Because of this argument in Smith (1994) I proposed using only unconditional inferences based on the random sampling procedure. Royall (1994), in his discussion, exposed the weakness of this position by describing a sampling procedure which involved choosing either a sample of size 100 or a sample of size 1000 each with probability 0.5. An unconditional inference over this whole procedure would clearly be nonsensical. Royall argued that this negated unconditional inference, but to me what it does is to highlight the difficulty of defining a procedure. Where does a procedure begin and end? Royall's example convinced me that the case for unrestricted unconditional inference was not well made, and that the nature of what constitutes a procedure needs more thought.

I have argued above that procedures should be specified in advance of sampling if they are to be trusted. But this specification could include elements of conditioning. In Royall's example the procedure could include the statement that if the sample of size 100 is chosen then the inference will be based on samples of size 100. Similarly if it is proposed to use post-stratification on a particular variable before the sample is drawn then this can be included within the procedure and the inference conditioned accordingly. What I am trying to avoid is post-stratifying and reweighting after looking at the data. Changing numbers because you don't like the look of them is dangerous statistics and bad science. If results really are out of line then they should still be published but with a footnote saying why they are suspect and even suggesting alternatives. Statisticians acting as censors of data are as bad as statisticians acting as censors of scientific publications. Overall I stand by my conclusion

that procedural inferences based on the randomization distribution are justifiable for descriptive inference in the public domain, but I now think that these procedures can embrace pre-specified conditions.

A useful by-product of the controversy between model-based and randomization inference has been the recognition that both the sample selection rule and the model of the process assumed to generate the finite population values need to be included in a complete specification of the process generating the sample data, see, for example, Rubin (1976), Smith and Sugden (1988). Let y_U denote the matrix of values of the survey variables in the finite population, functions of y_U are the target parameters which are to be estimated in a descriptive inference. Let x_U denote a matrix of the known values of covariates which can be employed in the design of the surveys and let w_U denote a matrix of unknown covariates. A general selection mechanism is a function that selects a subset s of U, where s denotes the labels which identify the sample units. Let the selection mechanism be

$$f(s|x_U, y_U, w_U). \tag{5.1}$$

This rule can include censoring depending on values of y_U, quota sampling which may depend on unknown covariates w_U, as well as on x_U, and non-response, which may depend on x_U, y_U and w_U. Interest focusses on functions of y_U or of the conditional distribution of y_U given x_U. Let

$$f(x_U, y_U, w_U) \tag{5.2}$$

denote a superpopulation model, or process, which is assumed to have generated the values in the finite population. The joint distribution of s, x_U, y_U, ignoring parameters, is

$$f(s, x_U, y_U) = \int f(s|x_U, y_U, w_U) f(x_U, y_U, w_U)\, dw_U. \tag{5.3}$$

If w_U is unknown this integration cannot be performed. The unknown values of w_U are sometimes said to cause unobserved heterogeneity and ignoring w_U is a form of model misspecification. If the selection mechanism (5.1) depends only on the known auxiliary variables x_U, so that

$$f(s|x_U, y_U, w_U) = f(s|x_U), \tag{5.4}$$

then the sampling mechanism is said to be uninformative. For uninformative selection schemes, (5.3) simplifies to

$$\begin{aligned} f(s, x_U, y_U) &= f(s|x_U) \int f(x_U, y_U, w_U)\, dw_U \\ &= f(s|x_U) f(x_U, y_U). \end{aligned} \tag{5.5}$$

The analyst can now focus his or her attention on the joint distribution of x_U and y_U. Let y_s be the sample values of y_U and $y_{\bar{s}}$ be the complement of y_s, then for the sample data

$$d_s = (x_U, y_s, s), \tag{5.6}$$

we have

$$\begin{aligned} f(d_s) &= f(s|x_U) \int f(y_U|x_U) f(x_U)\, dy_{\bar{s}} \\ &= f(s|x_U) f(y_s|x_U) f(x_U). \end{aligned} \tag{5.7}$$

This important simplification shows the importance of an uninformative selection scheme for model-based inference. Since all random sampling schemes depend only on x_U they lead to uninformative sampling, and this gives a model-based justification for random sampling. If there is a process (x_U, y_U) generating the finite population values, and if that process can be modelled accurately, then the sampling mechanism $f(s|x_U)$ can be ignored and descriptive inferences can be based solely on predictions from the model.

On the other hand if there is no agreed model generating the values, so that y_U, x_U are assumed to be constants, then the model distribution is degenerate and no model-based form of inference is possible. Then the only source of random variation is the sampling mechanism $f(s|x_U)$, but this sampling mechanism does not depend on y_U, it is uninformative. One solution, possibly the only solution, is to adopt randomization inference and to consider not only the sample drawn but also other samples which might have been drawn.

The simple structure represented by (5.1) to (5.7) concentrates a statistician's mind on the two aspects of data generation, the model and the selection mechanism. Too frequently the selection mechanism is ignored. In the Social Sciences and Life Sciences the dangers of ignoring some forms of informative selection, such as censoring, are now widely recognized.

In my view too little attention has been given to the model $f(x_U, y_U)$ or $f(y_U|x_U)$ in (5.5). What does a model mean? Statistical models represent a stochastic process which is assumed to have generated the data. In the Social Sciences what are these processes? How do you model the joint distribution of $f(y_U|x_U)$ when it may be a 25 million x 250 matrix? In what sense is there a probability mechanism generating employment status, income or social class? Most models do not attempt to model the generating process rather they start with the fact that unemployment status has a distribution in the population with various proportions falling in the classes. Randomness arises from statements like 'if an individual was drawn at random from this population then the probability they would be from the class E is P_E, where P_E

is the population proportion'. The probability arises not from P_E, a proportion, but from the random sampling process. Proportions are not probabilities and acting as if population distributions are probability distributions seems to me to be a common fallacy. The real processes generating employment status are very complex, but it is these processes that social scientists should be modelling.

6. SOCIAL SURVEYS AND SOCIAL SCIENCE

In the 19th century social scientists relied on censuses, administrative records, monography or on their personal observations for statistical data. World Wars I and II expanded the range of official statistics but at the cost of using sample surveys rather than complete population records. Kiaer's argument was that samples could replace censuses, that in some sense they were a poor man's census, but social scientists must soon have realized that social survey questionnaires could be much longer and more complex than census questionnaires and that as a result social surveys were a much richer source of data than a census. The case studies of monography, however, still provided the richest source of data because they could include both quantitative measurements and qualitative judgements as well as being longitudinal. The widespread use of panel surveys for contemporary social research into the dynamics of social events shows that the importance of study of the same units over time is now recognised. Despite this growth of the use of sample surveys, including panel studies, in social research I detect dissatisfaction among many social scientists in the outcome of this explosion in quantification. There have been few startling discoveries from all this effort and no generally accepted laws have emerged.

There are many possible reasons for this unease but I wish to focus on two. The first is that social scientists have tried to make progress too rapidly. Too much effort has been expended on macro-analysis and too little on the micro-processes that generated the macro data. This is changing but micro analysis is now highlighting the fact that social scientists have not yet defined many of their terms precisely enough for scientific analysis. Too many social scientists seem to be prepared to work with the definitions chosen by official statisticians rather than defining concepts directly relevant to the problem at hand. Far more effort is needed at this primitive level of Social Science if real progress is to be made. On the same theme the complexity of social processes and the need to put them into social contexts means that far more research effort should be put into in-depth case studies in small areas. Generalisation should come later, not at the outset of social research.

The second point is the complexity of social processes. It is possible that the failure to find any general quantitative laws is that there are no quantitative laws to be found. The best that can be hoped for is some qualitative statements. In scientific research it is usually suggested that one of the principles governing model choice should be the principle of parsimony; make the model as simple as possible. In the Life and Social Sciences this may not be possible and R.A. Fisher and L. J. Savage, and I am sure many others, have argued for making models as complex as possible. Savage said that all models should be as big as a house, in Social Science they may need to be as big as a mountain! Complexity arises not only from the hierarchy of social institutions but also from the networks that link individuals within each level of a hierarchy. Random samples are poor instruments for studying networks, monography is better. In order to apply probability models a scientist needs some form of homogeneity, whether it be Bayesian exchangeability, the stationarity of time series or Fisher's reference set. If the reference set is of size one then there is no homogeneity and modelling and model validation become very difficult. Should we expect homogeneity? Don't we all relish our individuality, don't we cherish our freedom of thought? Has too much attention been paid to modelling means when what is really important in social processes is variation? The fact that least squares works for observations (in Edgeworth's sense) does not mean that it is appropriate for the variances of statistics. Shouldn't we be fitting to match variances rather than fitting to minimize variances?

To return to Bailar's challenge about Statistics and social problems, I believe that Statistics in general and social surveys in particular do have an important role to play, but that it is an immediate need to define classes and agree terminology so that some form of rational debate can take place. This provides the data base for challenging existing prejudices and for determining future action. Actions can be of two types; the first is rapid (knee-jerk) political action, the second is considered action following research into the problem. To be useful both types of action should lead to predictions. The outcomes of predictions should be monitored, and this quality assurance exercise can be performed by a representative social survey. Both roles that I envisage for sample surveys are essentially descriptive. Only when quality assurance is expressed as a test does it become analytic.

Social scientists should engage in the search for either qualitative or quantitative models that will help to explain social phenomena and guide future actions. For this work to be scientific it must be based on agreed classifications and terminology. Once the language has been agreed then my view is that modelling is best developed from a detailed study of the social processes underlying a social phenomenon. Sample surveys are not suitable for this type of study. Only in depth case studies that incorporate social hierarchies and networks will lead to real understanding. Models of social processes should lead to

predictions which can then be validated by representative surveys. Predictions depend on estimates of parameters and as in the experimental sciences data from extreme cases may lead to more efficient parameter estimates, if the model is true, than those obtained from representative data. Social scientists should follow the pseudo-experimental methods being developed in the Life Sciences if they wish Social Science to receive the recognition that is merited by the importance of the problems studies.

My conclusion is that sample surveys are not well suited for learning about the complexities of social processes; monography is better. Sample surveys have two important roles; the first is their classical descriptive role; the second is a monitoring and refutation role. Refutation surveys should be designed to play the same role as refutation experiments in the physical sciences calling upon the same creative skills. If, as may be the case, no social theories can stand the rigours of refutation then we must accept the conclusion that Social Science is not yet a fully fledged science and concentrate our energies on effective description.

ACKNOWLEDGEMENT

This research was supported by a grant from the Economic and Social Research Council of the U.K. under its Analysis of Large and Complex Datasets programme.

REFERENCES

Alexander, C.H. (1994). Discussion of Smith (1994).

Anscombe (1961). "Bayesian statistics". *The American Statistician*, 15, February 1961.

Bailar III, J.C. (1995). "A larger perspective", *The American Statistician*, 49, 1, 10-11.

Bowley, A.L. (1926). "Measurement of the precision attained in sampling", *Bulletin of the International Statistical Institute*, 22, Livre I.

Brewer, K.R.W., and Särndal, C.E. (1983). "Six approaches to enumerative survey sampling" in *Incomplete Data in Sample Surveys*, Vol. 3, Eds. Madow, W.G. and Olkin, I., Academic Press, 363-368.

Edgeworth, F.Y. (1885). "Observations and statistics: an essay on the theory of errors of observation and the first principles of statistics", *Transaction of the Cambridge Philosophical Society*, 14, 138-169.

Ericson, W.A. (1969). "Subjective Bayesian models in sampling finite populations", *Journal of the Royal Statistical Society, B*, 31, 195-233.

Fellegi, I.P. (1989). "Challenge to statistics and statisticians", *Bulletin of the International Statistical Institute*, 1989.

Fisher, R.A. (1956). *Statistical Methods and Scientific Inference*, Oliver and Boyd, Edinburgh and London.

Godambe, V.P. (1955). "A unified theory of sampling from finite populations", *Journal of the Royal Statistical Society, B*, 17, 369-378.

Godambe, V.P. (1966). "A new approach to sampling from finite populations", *Journal of the Royal Statistical Society, B*, 28, 310-328.

Goldstein, H. (1995). *Multi-level statistical models*, Edward Arnold, London; Halsted, New York.

Hansen, M.H., Madow, W.G., and Tepping, B.J. (1983). "An evaluation of model-dependent and probability-sampling inferences in sample surveys", *Journal of the American Statistical Association*, 47, 663-685.

Kiaer, A.N. (1897). "The representative method of statistical surveys", Norwegian Academy of Science and Letters, The Historical, Philosophical Section, 1897, No. 4.

Kish, L. (1978). "Chance, Statistics, and Statisticians", *Journal of the American Statistical Association*, 73, 1-6.

Lazarsfeld, P.F. (1961). "Notes on the history of quantification in sociology - trends, sources and problems", *Isis*, 52, 277-333.

Morris, C.N. (1995). "Respondent: Symposium on modern interdisciplinary university statistics education", *The American Statistician*, 49, 1, 21-23.

Neyman, J. (1934). "On the two different aspect of the representative method: the method of stratified sampling and the method of purposive selection", *Journal of the Royal Statistical Society*, 97, 558-625.

Robinson, J. (1978). "An Asymptotic expansion for samples from a finite population", *The Annals of Statistics*, 1978, Vol. 6, No. 5, 1005-1011.

Royall, R.M. (1991). "Ethics and statisticians in randomised clinical trials", *Statistical Science*, 6, 1, 52-88.

Royall, R.M. (1994). Discussion of Smith (1994).

RSS (1995). "Report of the working party on the measurement of unemployment in the U.K.", Royal Statistical Society (1995).

Rubin, D.B. (1976). "Inference and missing data", *Biometrika*, 63, 593-604.

Smith, T.M.F. (1976). "The foundations of survey sampling: a review", *Journal of the Royal Statistical Society, A*, 139, 183-204.

Smith, T.M.F. (1977). "Statistics: the art of conjecture", *The Statistician*, 27, 65-86.

Smith, T.M.F., and Sugden, R.A. (1988). "Sampling and assignment mechanisms in experiments, surveys and observational studies", *International Statistical Review*, 56, 165-180.

Smith, T.M.F. (1994). "Sample surveys 1975-1990; an age of reconciliation?", *International Statistical Review*, 62, 1, 3-34.

Stigler, S.M. (1986). *The History of Statistics*, Harvard University Press, Cambridge, Mass.