International Statistical Review (1988), **56**, 2, pp. 165–180. Printed in Great Britain

Correspondent paper

Sampling and Assignment Mechanisms in Experiments, Surveys and Observational Studies

T.M.F. Smith[1] and R.A. Sugden[2]

[1] *Faculty of Mathematical Studies, University of Southampton, Southampton SO9 5NH, U.K.* [2] *Department of Mathematics, University of London, Goldsmiths' College, London SE14 6NW, UK*

Summary

A general framework is given for examining the role of mechanisms for treatment assignment and unit selection in experiments, surveys and observational studies. Conditions are established under which these mechanisms can be ignored for model-based inference. Examples are presented to show how inference can incorporate the mechanisms when the conditions do not hold.

Key words: Causal effects; Ignorability; Model-based inference.

1 Introduction

Two of the most important tasks facing an applied statistician are the design and analysis of sample surveys, and the design and analysis of experiments. Despite the many apparent links between surveys and experiments identified by Fienberg & Tanur (1987), and the contributions to both subjects by statisticians such as Yates and Cochran, the theory and methods of these two subjects appear quite different. Why should this be?

Sample survey design and analysis has grown from the need for descriptive statistics in the government and commercial stores. In contrast experiments use statistics analytically to test hypotheses for example, and are carried out most frequently in the industrial and scientific sectors. Government surveys are still used primarily to describe the properties of a finite population at a fixed point in time but many commercial surveys attempt to explain relationships between variables that will hold in the future, so they are explanatory or analytic rather than descriptive. In the analysis of an explanatory survey or of an experiment one aim is to find evidence for a law-like relationship, for causal explanations rather than descriptions.

Wold (1967) addressed the problems of reaching causal conclusions from nonexperimental data. He argued that experimental knowledge is reproducible knowledge, the reproducibility arising from the control exerted by the scientist over the assignment of treatments. It is the scientist's control over his material that is the most important distinction between surveys and experiments; in surveys there is control over sampling of units while in experiments there is control over the assignment of treatments to units. From this difference in control flows the difference in the approaches to the design and, more importantly, the analysis of surveys and experiments.

Experiments can be carried out within surveys. Brewer et al. (1977) give examples of experimental designs at the pilot stage of a survey for testing questionnaire layouts, diary

versus recall methods and for estimating variance components. Within market research it is common to find a designed experiment for comparing different formulations of a product performed within a sample survey. In agriculture, experiments are sometimes performed on a representative sample of farms to see if results from an agricultural experimental station can be reproduced in the 'real world'.

Sampling from within a controlled experiment is less common. In agriculture the yields from plots are sometimes estimated by sampling from within the plots and in variety trials the number of varieties is sometimes reduced by random sampling prior to testing. Often in social experimentation a large population is divided into subpopulations and each subpopulation is allocated some form of treatment. The responses within subpopulations are then frequently estimated by sampling. Following this theme if we enlarge the notion of an experiment to include observational studies, where 'nature', or some other agency, has allocated the treatments in an unknown manner, then in these studies sampling takes place within an uncontrolled experiment. The design, analysis and interpretation of observational studies is one of the most difficult tasks facing the applied statistician. We elaborate on these distinctions in § 2 of the paper.

In § 3 we adopt the approach of Rubin (1976, 1978) and set up a formal structure to represent the sampling and assignment mechanisms. This brings out some important distinctions between experiments and surveys and incidentally provides a scientific role for randomisation. In § 4 we apply our formal analysis to an observational study of the relationship between two binary variables. Finally in § 5 we analyse through a theoretical example the effect of incomplete covariate information on inference for an experiment within a survey.

2 Experiments, surveys and observational studies

To facilitate a unified approach to these three types of study we define a treatment to mean any stimulus, or set of stimuli, leading to an observable response even when there is no control over its assignment. Treatments are applied to units and responses are measured on these same units. Every unit should be capable of receiving every treatment and the decision as to which unit receives which treatment is determined by an *assignment mechanism.*

The set of possible units which could receive a treatment comprises a population. In practice this population is finite so that the units can be listed, although sometimes only conceptually. From a finite population a sample of units can be selected from the list, or frame, using a *sampling mechanism.*

We distinguish various studies by the control exerted by the scientist over the assignment and sampling mechanisms.

(i) An *experiment* is a study in which the scientist can control the assignment of treatments to experimental units.
(ii) A *survey* is a study in which the scientist can control the selection of units for which responses will be observed.
(iii) An *analytic survey* is a study of the effect of treatments when there is control over the selection of units but no control over the assignment of treatments.
(iv) An *uncontrolled observational study* is a study in which there is no control over either the selection of units or the assignment of treatments.

The general term observational study refers to studies of types (iii) and (iv). Since we are concentrating on the role of the assignment and sampling mechanisms we have chosen

Table 1
Classification of studies with treatments

		Treatment assignment	
		Control	No control
Sample selection	Control	Experiment within a survey, or survey within an experiment	Analytic survey
	No control	Experiment	Uncontrolled observational study

to distinguish the two cases. In Table 1 we show these studies classified by the type of control exerted by the scientist.

There are also important differences in the target populations of interest. In many surveys, particularly those in the public sector, the target population is the real finite population which exists when the survey is carried out and the targets of inference are finite population means, totals and other descriptive statistics. Such descriptive studies with no treatments have no causal objective and we do not consider this special case further in our unified approach.

When treatments are present the object of analysis is the comparison of identifiable subgroups in the finite population which have been exposed to different treatments. Whether such comparisons can be called causal inferences depends on the relation between the real finite population and the hypothetical population given by the set of all possible responses to all possible treatments measured on each unit. In an experiment the objective is overtly causal; it is to make comparisons between hypothetical, but nevertheless reproducible, populations of possible responses. The reproducibility derives from the control exerted by the scientist over the treatment assignment mechanism.

We define the *apparent population* as the set of responses to treatments actually applied to units, that is, the population of responses after assignment. The corresponding *true population* is the unobservable set of responses to all treatments which might have been applied to units whether they were applied or not. It is the population of potential responses prior to assignment. Causal effects are expectations defined for the population of interest. An *apparent causal effect,* or prima facie causal effect (Holland, 1986), is thus a comparison between the means of two or more subgroups of the finite population defined by the outcome of the assignment mechanism. The *true population causal effect* is a comparison between the population means which would result if all units received the same treatment. True causal effects are unobservable as only one treatment can be applied to each unit.

There are differences too between surveys and experiments in the methods and approaches to inference. In descriptive surveys methods of inference are almost always based on the distribution generated by repeated applications of the sample selection mechanism. When sampling is random and the selection probabilities are known this generates the *randomisation distribution.* The randomisation distribution is clearly appropriate for pre-data decisions about choice of design but its use for post-data inference is less clear cut. The theoretical limitations of the randomisation distribution for conditional inference after the data are observed, and for predictive inference, have been discussed by, for example, Smith (1976, 1983), Royall (1976). Recently a model-based

theory of inference has emerged which encompasses both analytic and descriptive inference.

For experiments inferences are usually based on an underlying statistical model, usually a linear model, and not on the distribution generated by repeated applications of the treatment assignment mechanism. Although randomisation inference has its advocates and randomisation tests are sometimes employed instead of the model-based procedures, the impression from outside must be that experimental inference is model-based and that survey inference is randomisation-based. This is one of the main perceived differences between surveys and experiments. We shall not pursue the reasons for this difference here but rather we will point out that a complete stochastic description of both surveys and experiments must incorporate the selection and allocation mechanisms even when a model-based approach to inference is adopted. Within a model-based approach the issue then becomes one of identifying the conditions under which the mechanisms do not affect the inference, which we consider in § 3.

3 A general approach to sampling and assignment mechanisms

3.1 Introduction

Following Rubin (1976, 1978) we examine the role of sampling and assignment mechanisms within a model-based framework for inference. This allows us to examine both controlled studies which employ randomisation and uncontrolled observational studies. We consider inferences from a sample of units where the responses are affected by a set of stimuli, the treatments. Thus we exclude simple descriptive surveys. The ideal studies are experiments within surveys or surveys within experiments where control is exercised by the scientist over both treatment assignment and unit selection. But our main interest is in the other studies in Table 1 where nature has determined one or both of the mechanisms.

We will use the word *nature* to personify the process of uncontrolled unit selection or treatment assignment. When either of these processes is controlled we shall refer to the person or persons exercising the control, designing the study and making the measurements as the *scientist.* A person concerned solely with the analysis of data will be called an *analyst.* Thus in an observational study nature assigns the treatments, the scientist makes the measurements and the analyst produces the results. Frequently the scientist and the analyst are the same person.

The design stage of a study has two mechanisms which generate the data for analysis:

(i) the selection of a subset, or sample s, of n units from the finite population of N units;
(ii) the assignment of treatments to units.

The specification of these mechanisms and the implicit assumptions underlying them must be considered by the analyst who may have much less data available than either the scientist or nature, as, for example, in a secondary analysis of a data set.

Suppose there are N units in the population labelled $1, 2, \ldots, N$, and a set of T treatments labelled $1, 2, \ldots, T$. When treatment j is applied to unit i the response is the realised value of a random variable $Y_i^{(j)}$ which we assume to be observable without error. Only one of these T potential responses ($j = 1, \ldots, T$) can be observed for each unit i. The fact that the random variable is indexed only by i and j corresponds to an assumption of no interference between different units (Cox, 1958), which Rubin (1980) refers to as SUTVA (Stable Unit Treatment Value Assumption). The $N \times T$ matrix of all potential

responses defines the *true population*

$$\mathbf{Y} = (\mathbf{Y}^{(1)}, \ldots, \mathbf{Y}^{(T)}), \tag{3.1}$$

where $\mathbf{Y}^{(j)}$ is the $N \times 1$ vector of responses if all the units in the population were to receive treatment j.

Following Rubin (1978) we define the *true causal effect* of treatment j versus treatment j' for the ith unit as the unobservable difference

$$Y_i^{(j)} - Y_i^{(j')}, \tag{3.2}$$

regardless of whether either treatment has actually been applied or not. The target parameter for inference is the mean of this contrast. We define the *true population causal effect* to be the finite population mean

$$\bar{Y}^{(j)} - \bar{Y}^{(j')}. \tag{3.3}$$

Other definitions could be given, for example by taking means in some marginal statistical superpopulation model for the responses or in a conditional model given a set of covariates. These alternatives are not considered here. We note that the definition of the relevant population is a key step in the process of examining the validity of causal inferences.

Design depends on prior information. Let $\mathbf{z}$ be an $N \times q$ matrix of values of q background variables, known to both nature and the scientist for all the units in the population. This matrix will include labels indicating blocks, strata or clusters as well as quantitative variables such as size measures or covariates, and can be used by both nature and the scientist at the design stage; $\mathbf{z}$ can only be used at the *analysis stage* if it is known by the analyst. The additional data known to nature but not to the scientist is denoted formally by a matrix $\mathbf{W}$, which contains all the hidden and unknown covariates. We assume that the scientist can never know the values of the variables in $\mathbf{W}$; that is they are unobservable. Thus $\mathbf{W}$ excludes the response variables $\mathbf{Y}$. In the terminology of Holland (1986) $\mathbf{z}$ and $\mathbf{W}$ are pre-exposure variables and $\mathbf{Y}$ are post-exposure variables.

The specification of the two mechanisms depends on the order in which the mechanisms are used. In an observational study or a survey within an experiment it is assumed that treatments are assigned to all the N units in the population, and that the sample s is selected after the treatment assignment. For experiments within surveys the sample is selected first and then treatments are assigned only to those units in the sample.

3.2 Inference and the ignorability of mechanisms

In our discussion of the sampling and assignment mechanisms we have introduced three sets of variables, the responses $\mathbf{Y}$, the known covariates $\mathbf{z}$ and the unknown covariates $\mathbf{W}$. Formally we can write down a joint distribution for these variables as

$$f(\mathbf{W} \mid \mathbf{Y}, \mathbf{z}; \boldsymbol{\psi}) g(\mathbf{Y} \mid \mathbf{z}; \boldsymbol{\theta}) h(\mathbf{z}; \boldsymbol{\phi}), \tag{3.4}$$

where we assume that the parameters $\boldsymbol{\psi}$, $\boldsymbol{\theta}$, $\boldsymbol{\phi}$ are distinct, as Rubin (1976). The mechanisms generate observed responses, $\mathbf{Y}_s$ say, and an analysis based on the joint distribution of $\mathbf{Y}_s$ and the known covariates $\mathbf{z}$ but which ignores the sampling and assignment mechanisms is called a *face value* analysis (Dawid & Dickey, 1977). The face value likelihood is given by

$$h(\mathbf{z}; \boldsymbol{\phi}) \int_C g(\mathbf{Y} \mid \mathbf{z}; \boldsymbol{\theta})\, d\mathbf{y}, \tag{3.5}$$

where C is the set of unobserved responses in two groups:

(a) $\mathbf{Y}_{\bar{s}}^{(j)}$, for $j = 1, \ldots, T$, all potential responses for unsampled units $\bar{s}$;
(b) $\mathbf{Y}_{s-s^{(j)}}^{(j)}$, for $j = 1, \ldots, T$, potential responses for units in the observed sample s which do not receive the treatment j, for $j = 1, \ldots, T$. (3.6)

Note that in (3.5) the unknown covariates $\mathbf{W}$ integrate out completely as they cannot be observed. A full analysis of the observed data $\mathbf{Y}_s$, $\mathbf{z}$ would employ the *full likelihood* including the sampling and assignment mechanisms, and we examine this in subsequent sections taking into account the order in which the mechanisms are applied. We say that the sampling and/or assignment mechanisms are *ignorable for likelihood inference* if the full likelihood and the face value likelihood lead to the same inference.

In general the ignorability of the mechanisms will depend on:

(i) the approach to inference, whether Bayes, likelihood or sampling theory;
(ii) the target for inference, whether true population causal effect, a prediction problem or the estimation of some function of the unknown parameters, $\boldsymbol{\theta}, \boldsymbol{\phi}$;
(iii) the population model (3.4) and its properties;
(iv) the nature of the sampling and assignment mechanisms;
(v) the amount of covariate information available to the analyst.

Sugden & Smith (1984) have examined many of these conditions under sampling alone, with no treatment assignment. Here we extend the approach to cover more general classes of studies within the framework of likelihood inference.

3.3 *Assignment before sampling*

These studies include surveys within experiments, analytic surveys and observational studies where the treatments can be assumed to have been assigned prior to sampling. Let $U^{(j)}$ be the set of population units assigned to treatment j ($j = 1, \ldots, T$). The T sets form a partition, or stratification, of the population. In its most general form the assignment mechanism is specified by the variables $\mathbf{Y}$, $\mathbf{z}$, $\mathbf{W}$ and can be written

$$P(U^{(j)}, \quad j = 1, \ldots, T \mid \mathbf{Y}, \mathbf{W}, \mathbf{z}). \tag{3.7}$$

We define *controlled assignment* as a mechanism used by a scientist for a survey within an experiment which depends only on $\mathbf{z}$ and can be written as

$$P(U^{(j)}, \quad j = 1, \ldots, T \mid \mathbf{z}). \tag{3.8}$$

The sampling mechanism for an observational study or a survey within an experiment frequently depends on the allocation $U^{(j)}$, for $j = 1, \ldots, T$, of treatments to units. If $S^{(j)}$ denotes the sample units which receive treatment j then the most general sampling mechanism can be written

$$P(S^{(j)}, \quad j = 1, \ldots, T \mid U^{(j)}, \quad j = 1, \ldots, T, \mathbf{Y}, \mathbf{W}, \mathbf{z}). \tag{3.9}$$

If the scientist implements a sampling mechanism, then it cannot depend on $\mathbf{W}$, and so the analyst can write the mechanism as

$$P(S^{(j)}, \quad j = 1, \ldots, T \mid U^{(j)}, \quad j = 1, \ldots, T, \mathbf{Y}, \mathbf{z}), \tag{3.10}$$

where the possible dependence on the response variables $\mathbf{Y}$ is to allow for retrospective studies as well as prospective studies (Breslow & Day, 1980). Note that in retrospective studies the sampling mechanism depends on the response variables $\mathbf{Y}$ only through those responses to treatments *actually received by the units*, defined below in (3.11).

A more restrictive form of sampling mechanism, where (3.10) additionally does not depend on $\mathbf{Y}$, is called *controlled sampling.*

For all forms of assignment, the values of $U^{(j)}$, for $j = 1, \ldots, T$, impose an implicit stratification on the population. We define the *apparent population* as

$$\mathbf{Y}_0 = (\mathbf{Y}_{U^{(j)}}, j = 1, \ldots, T). \tag{3.11}$$

Then the *apparent population causal effect* of treatment j versus j' is simply the difference between the two subpopulation means

$$\bar{Y}^{(j)}_{U^{(j)}} - \bar{Y}^{(j')}_{U^{(j')}}. \tag{3.12}$$

It is important to realise that this may be very different from the true population causal effect (3.3), regardless of the sampling mechanism. In § 4 we consider an example to bring out this point.

We now return to consideration of the full likelihood. In general the assignment mechanism is given by (3.7) and the sampling mechanism by (3.9) and so the full likelihood derived from the model (3.4) is

$$\begin{aligned} h(\mathbf{z}; \boldsymbol{\phi}) \int\int_C & P(U^{(j)}, j = 1, \ldots, T \mid \mathbf{Y}, \mathbf{z}, \mathbf{W}) \\ & \times P(S^{(j)}, j = 1, \ldots, T \mid \mathbf{Y}, \mathbf{z}, \mathbf{W}, U^{(j)}, j = 1, \ldots, T) \\ & \times g(\mathbf{Y} \mid \mathbf{z}; \boldsymbol{\theta}) f(\mathbf{W} \mid \mathbf{Y}, \mathbf{z}; \boldsymbol{\psi})\, d\mathbf{y}\, d\mathbf{w}, \end{aligned} \tag{3.13}$$

assuming the partitioning given by $U^{(j)}$, for $j = 1, \ldots, T$, is known, where C is defined in (3.6) and the additional integration is over all $\mathbf{W}$. Integrating out the unobservable variables $\mathbf{W}$ gives the likelihood

$$\begin{aligned} h(\mathbf{z}; \boldsymbol{\phi}) \int_C & P(U^{(j)}, j = 1, \ldots, T \mid \mathbf{Y}, \mathbf{z}; \boldsymbol{\psi}) P(S^{(j)}, j = 1, \ldots, T \mid \mathbf{Y}, \mathbf{z}, \mathbf{U}^{(j)}, j = 1, \ldots, T; \boldsymbol{\psi}) \\ & \times g(\mathbf{Y} \mid \mathbf{z}; \boldsymbol{\theta})\, d\mathbf{y}. \end{aligned} \tag{3.14}$$

The mechanisms are ignorable for likelihood inference if they can be taken outside of the integral in (3.14) or (3.13). Sufficient conditions for ignorability are either of type A or B.

Conditions A. (i) $U^{(j)}, j = 1, \ldots, T \amalg \mathbf{Y} \mid \mathbf{z}; \boldsymbol{\psi}$;
(ii) $S^{(j)}, j = 1, \ldots, T \amalg \mathbf{Y} \mid \mathbf{z}, U^{(j)}, j = 1, \ldots, T; \boldsymbol{\psi}$.

Conditions B. (i) $U^{(j)}, j = 1, \ldots, T \amalg \mathbf{Y} \mid \mathbf{z}, \mathbf{W}$;
(ii) $S^{(j)}, j = 1, \ldots, T \amalg \mathbf{Y} \mid \mathbf{z}, \mathbf{W}, U^{(j)}, j = 1, \ldots, T$;
(iii) $\mathbf{Y} \amalg \boldsymbol{W} \mid \boldsymbol{z}; \boldsymbol{\psi}$.

The symbol $\amalg$ means 'is independent of'. The conditions in B imply those in A.

A controlled assignment mechanism (3.8) satisfies A(i), and a controlled sampling mechanism satisfies A(ii).

We are now in a position to give sufficient conditions for the mechanisms to be ignorable for likelihood inference within three cells of Table 1. For surveys within experiments condition A(i) is automatically satisfied, and condition A(ii) will be satisfied for any sampling scheme that does not depend on the response variables $\mathbf{Y}$. This excludes retrospective sampling but even these mechanisms may be ignorable if, for example, they depend only on the apparent population $\mathbf{Y}_0$ through the observed responses $\mathbf{Y}_s$. These mechanisms will be the subject of future work. Scott & Wild (1986) give some examples of ignorable choice based sampling in case-control studies. In an analytic survey there is no control over assignment but there is over sampling. For assignment to be ignorable

conditions B(i) and B(iii) must be satisfied, and sampling is ignorable since A(ii) is automatically satisfied. Finally for observational studies with no control over either assignment or sampling we require conditions B(i), B(ii) and B(iii) for ignorability of the mechanisms. We discuss the implications of those conditions in § 3.5.

3.4 Assignment after sampling

These studies include experiments, experiments within surveys, and those prospective observational studies where the treatment can be assumed to have been assigned after unit selection. The unit sampling mechanism is given in general by

$$P(S \mid \mathbf{Y}, \mathbf{z}, \mathbf{W}), \tag{3.15}$$

which can depend on the hidden covariates $\mathbf{W}$ as well as on the known covariates $\mathbf{z}$ and the potential responses $\mathbf{Y}$. We define *controlled selection* as a mechanism by the scientist of the form

$$P(S \mid \mathbf{z}), \tag{3.16}$$

which depends only on the known covariates, $\mathbf{z}$.

The treatment assignment mechanism assigns treatment j to a subsample $S^{(j)}$ of S. In general we represent this by

$$P(S^{(j)}, j = 1, \ldots, T \mid S, \mathbf{Y}, \mathbf{z}, \mathbf{W}). \tag{3.17}$$

If the scientist performs a 'laboratory-based' treatment assignment, we say there is *controlled assignment* and the mechanism can be written

$$P(S^{(j)}, j = 1, \ldots, T \mid S, \mathbf{z}). \tag{3.18}$$

In general the sampling mechanism is given by (3.15) and the assignment mechanism by (3.17) and the full likelihood is

$$h(\mathbf{z}; \boldsymbol{\phi}) \int\int_C P(S^{(j)}, j = 1, \ldots, T \mid S, \mathbf{Y}, \mathbf{z}, \mathbf{W}) P(S \mid \mathbf{Y}, \mathbf{z}, \mathbf{W}) \times g(\mathbf{Y} \mid \mathbf{z}; \boldsymbol{\theta}) f(\mathbf{W} \mid \mathbf{Y}, \mathbf{z}; \boldsymbol{\psi})\, d\mathbf{y}\, d\mathbf{w}, \tag{3.19}$$

where the additional integration is again over all possible $\mathbf{W}$, and C is given by (3.6). Again the unobservable covariates $\mathbf{W}$, may be integrated out from (3.19) if required.

The mechanisms are ignorable for likelihood inference if they can be taken outside of the integral in (3.19). Sufficient conditions for ignorability are either type C or D.

Conditions C. (i) $S \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{z}; \boldsymbol{\psi}$;
(ii) $S^{(j)}, j = 1, \ldots, T \parallel \mathbf{Y} \mid \mathbf{z}, S; \boldsymbol{\psi}$.

Conditions D. (i) $S \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{z}, \mathbf{W}$;
(ii) $S^{(j)}, j = 1, \ldots, T \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{z}, S, \mathbf{W}$;
(iii) $\mathbf{Y} \perp\!\!\!\perp \mathbf{W} \mid \mathbf{z}; \boldsymbol{\psi}$.

The conditions D imply those in C.

A controlled selection mechanism (3.16) satisfies C(i) and a controlled assignment mechanism (3.18) satisfies C(ii). If we now consider Table 1 conditions C(i) and C(ii) are automatically satisfied for experiments within surveys and so the mechanisms are ignorable. For an experiment condition C(ii) is satisfied and this mechanism is ignorable if there is no sampling, that is the experimental units represent the universe of interest. This is rarely the case and for most experiments sampling is ignored and conditions D(i) and

D(iii) are implicitly assumed to be satisfied. For prospective observational studies the mechanisms can be ignored if conditions D(i), D(ii) and D(iii) are all satisfied.

3.5 Discussion

What conclusions can we draw from the sufficient conditions that we have established for the ignorability of sampling and assignment mechanisms? The most obvious one is that controlled sampling and controlled assignment carried out by a scientist in such a way that they do not depend on the response variable **Y** will *always guarantee* ignorability provided the mechanisms are known, as we have assumed throughout. Any mechanisms which depend only on the design covariates **z** will suffice but the most acceptable are likely to be those based on randomisation, since they have the additional property of 'fairness', in the sense that every unit has some known nonzero chance of being selected for a sample and of being assigned a given treatment. Thus classical random sampling and random assignment schemes are ignorable and enable model-based inference to be carried out using face-value analyses.

In practice even when random mechanisms are employed nature may intervene by introducing nonresponse and/or measurement errors. For a nonresponse mechanism to be ignorable it would have to satisfy conditions such as those in B or D, and these are not verifiable and so become a statement of belief. See Rubin (1976) and Little (1982) for a full discussion of the problems of missing values. Measurement errors which are related only to the units in the sample will affect the specification of the model, and hence the likelihood, but not the ignorability of the mechanisms.

In observational studies, where nature applies the mechanisms, then conditions such as B or D should be assumed by the analyst for ignorability to hold. Neither of these sets of conditions is verifiable, since they depend on unobservable covariates, and so the assertion that they hold is a statement of belief. However, most analyses of observational studies are based on the face value distribution and implicitly assume that the ignorability assumptions are satisfied. We think that these conditions should be made explicitly, rather than remaining implicit; and we hope that the framework that we have developed will help in this process. This framework should provide a focus for the critical discussions which usually surround the analysis of any observational study.

When assignment has been carried out before sampling then one assumption that sometimes may be verifiable on theoretical or practical grounds is that the response variable **Y** is posterior to the assignment variables **W** and **z** employed by nature. Holland (1986) calls this a post-exposure variable and if this is the case then Conditions B(i) and B(ii) are automatically satisfied and discussion can concentrate on the validity, or otherwise, of B(iii).

For analytic surveys, controlled sampling means that attention centres only on the assignment mechanism. If **Y** is a post-exposure variable then again discussion will concentrate on B(iii) for justification of face-value analysis. For experiments the situation is different as controlled assignment enables valid inferences to be made provided that sampling is ignorable. However, the usual practice is simply to ignore sampling and not to discuss ignorability; thus it is implicitly assumed that Conditions D(i) and D(iii) are satisfied. Again we would argue that these assumptions should be made explicit so that the general validity of experimental results can be discussed openly. The common scientific practice of always trying to replicate new experimental results in different environments is clearly an attempt to establish D(i) and D(iii). It is to be regretted that this practice is not more widely employed in the social sciences.

The above discussion points to B(iii) for studies with assignment before sampling and

D(iii) for studies with assignment after sampling as playing a vital role in the absence of controlled mechanisms. Yet these assumptions simply say that the unobserved covariates have no effect in the conditional distribution of the responses given the observed covariates. Such assumptions are basic to modelling by the analyst. If the effect of the unobserved covariates is small, then face-value analysis will still be approximately valid. This is one of the broad conclusions of Little (1982) and Smith (1983) who examine these issues.

4 An example of an observational study

In § 3 we distinguished the apparent population causal effect (3.12) from the true population causal effect (3.3) and noted that these two effects could be different. We now illustrate this by considering a simple example with a binary response, 0 (failure) or 1 (success), and two treatments, T_1 and T_2. The true population of all potential responses (3.1) is an $N \times 2$ matrix where each element $Y_i^{(j)}$ is either 0 or 1. The N pairs of potential responses $(Y_i^{(1)}, Y_i^{(2)})$ can take the values $(0,0)$, $(0,1)$, $(1,0)$, $(1,1)$, and the true population frequencies of these potential responses can be summarised as in Table 2.

Table 2

The true population

Response	Frequency
(0, 0)	a
(0, 1)	b
(1, 0)	c
(1, 1)	d
	N

Thus b of the N pairs take the value $(0,1)$ and for any unit in this set the causal effect (3.2) would be $Y_i^{(1)} - Y_i^{(2)} = 0-1 = -1$. For the whole population the *true population causal effect* is

$$\bar{Y}^{(1)} - \bar{Y}^{(2)} = \frac{c+d}{N} - \frac{b+d}{N} = \frac{c-b}{N}. \tag{4.1}$$

In an observational study all N units are assigned treatments by nature according to the mechanism (3.7). If N_1 units are assigned to T_1 and $N_2 = N - N_1$ to T_2 then the N responses which give the apparent population can be summarised in a 2×2 contingency table such as Table 3.

The *apparent population causal effect* (3.12) is

$$\bar{Y}_{U^{(1)}}^{(1)} - \bar{Y}_{U^{(2)}}^{(2)} = \frac{M_1}{N_1} - \frac{M_2}{N_2} = P_1 - P_2, \text{ say.} \tag{4.2}$$

Table 3

The apparent population

		Response 0	Response 1	
Treatment	T_1	$N_1 - M_1$	M_1	N_1
	T_2	$N_2 - M_2$	M_2	N_2
		$N - M$	M	N

Table 4

Probability of assigning T_1 to a unit

Potential response	Pr(gets T_1)	Pr(gets T_2)
(0, 0)	p	$1-p$
(0, 1)	q	$1-q$
(1, 0)	r	$1-r$
(1, 1)	s	$1-s$

Table 5

Expected apparent population

		Response 0	Response 1
Treatment	T_1	$pa+qb$	$rc+sd$
	T_2	$(1-p)a+(1-r)c$	$(1-q)b+(1-s)d$
			N

A general treatment assignment mechanism, such as (3.7), may produce an apparent effect (4.2) which is very different from the true effect (4.1). So inferences about (4.2) may be misleading if the true target for inference is (4.1). It is also true that the controlled assignment mechanism, (3.8), which depends only on the covariates $\mathbf{z}$, can produce large differences. For example, if a covariate $\mathbf{z}$ is such that when $\mathbf{z}$ is large T_1 is likely to give a 0 response and T_2 a 1 response, then nature can allocate according to $\mathbf{z}$ to produce apparent causal effects vastly different from the true effects. However, since the conditional model relating the responses to the covariate values is known and the covariate values themselves are known for all units, as is assumed in (3.13), then the apparent results can be adjusted to give predictions of the true results. These predictions will still differ from the true target (4.1) because of variation in the realisations of the assignment. On the average over all realisations we would obtain (4.1). Frequently it may be better to condition only on functions of $\mathbf{z}$ such as *propensity scores* (Rosenbaum & Rubin, 1983), which still leave the mechanism ignorable.

Suppose now that the assignment is not ignorable and that the probability of assigning T_1 to a unit depends on the potential response, Y, according to Table 4. Here the probabilities p, q, r, s are unrestricted. Under this mechanism we can calculate the *expected apparent population* for which Table 5 is a realisation.

The response 0 is observed when T_1 is applied either with probability p to the potential response (0, 0) which has frequency a, or with probability q to (0, 1) which has frequency b. Similar calculations lead to the other entries.

If N is large we can equate Tables 3 and 5 disregarding the chance variation in treatment assignment. In the special case of a noninformative mechanism with no covariates, then $p=q=r=s$, and we find that (4.1) and (4.2) are equal. In general we have four equations, such as $N_1 - M_1 = pa + qb$, which constrain the possible values for the true population frequencies a, b, c, d, and hence constrain the true population causal effect (4.1). For fixed numbers in Tables 2 and 3 the necessary conditions for a solution to the four equations in p, q, r, s are

$$N_1 + (N_2 - M_2) \geq a + c \geq N_2 - M_2, \quad (N_1 - M_1) + M_2 \geq b + d \geq M_2,$$
$$M_1 + (N_2 - M_2) \geq c + d \geq M_1, \quad N_2 + (N_1 - M_1) \geq a + b \geq N_1 - M_1.$$

Table 6

Annual apparent deaths (all causes)

	Survive (0)	Die (1)	Total
Smoking (T_1)	13044	204	13248
Non-smoking (T_2)	17744	216	17960
Total	30788	420	31208

Table 7

Propensities to smoke by response

Response		
Smoking	Non-smoking	pr[smoking]
Survive	Survive	0·4227 = 12961/30665
Survive	Die	0·5533 = 83/150
Die	Survive	0·0909 = 4/44
Die	Die	0·5731 = 200/349

Table 8

Population distribution by response

Response		
Smoking	Non-smoking	Frequency
Survive	Survive	30665
Survive	Die	150
Die	Survive	44
Die	Die	349
		31208

These bounds can produce large differences between apparent and true population causal effects. For further discussion see Sugden (1988).

As an example consider a highly simplified table adapted from a study on doctors and smoking reported by Doll & Hill (1964). Details are omitted because the example is only illustrative. Table 6 shows the observed data for deaths from all causes against the treatment T_1, smoking, and T_2, non-smoking. The data show a significantly higher death rate for smokers (0·0154) than for non-smokers (0·0120) with an apparent causal effect of $0{\cdot}0154 - 0{\cdot}0120 = 0{\cdot}0034$. However if we assume probabilities of smoking varying with potential response (see Table 4) as given in Table 7, and we also assume a corresponding true population distribution of potential response (see Table 2) as given in Table 8, then the expected numbers of smokers and non-smokers with their response under the actual 'treatment' received (see Table 5) agree exactly with Table 6. The true causal effect is $(44 - 150)/31208 = -0{\cdot}0034$, a reversal of sign implying that smoking increases the chance of survival. Thus there exists an allocation by nature which changes the sign of the apparent causal effect if the true population is as in Table 8.

A similar reversal of the sign of the true causal effect compared with the apparent causal effect can be shown for the data on deaths due to lung cancer and the presence or absence of smoking. The implication of these results is that any causal conclusions based on sampling from an apparent population in an observational study are open to serious doubt. Only if severe constraints are placed on the true population (Table 2) and in the assignment probabilities (Table 4) can the sign of the causal effects remain the same. This

raises the question whether causal conclusions can ever be inferred from an observational study without making strong subjective assumptions.

5 Inference with incomplete covariate information

The analysis of ignorability in § 3 was based on the assumption that the entire covariate matrix $\mathbf{z}$ was known to the analyst. Frequently the analyst differs from the scientist who has designed the investigation, for example in a secondary analysis of survey data, and the analyst will only know the covariate values $\mathbf{z}_s$ for the sample units. In this section we study this situation using an example considered by Sugden & Smith (1984) and Scott (1977).

Consider a trivariate superpopulation with two treatments such that the potential responses $Y_i^{(1)}$, $Y_i^{(2)}$ and the covariate value z_i form independent vectors $(Y_i^{(1)}, Y_i^{(2)}, z_i)$, for $i = 1, \ldots, N$, with known distributions

$$f_1^*(y_i^{(1)} \mid z_i; \theta^{(1)}) f_2^*(y_i^{(2)} \mid z_i; \theta^{(2)}) g^*(z_i; \lambda), \tag{5.1}$$

where the * distinguishes the distribution of unit values from that of the entire data matrix in (3.4). This model assumes that the treatment responses are conditionally independent given the covariate values and parameters which is an assumption commonly made in models for the analysis of experiments.

The data available to the analyst is

$$d_s = (\mathbf{y}_{s^{(1)}}^{(1)}, \mathbf{y}_{s^{(2)}}^{(2)}, \mathbf{z}_s, s, s^{(1)}, s^{(2)}). \tag{5.2}$$

We also assume that both the sampling mechanism and the treatment assignment mechanism are of the forms

$$p(s \mid \mathbf{z}), \quad p(s^{(i)}, i = 1, 2 \mid s, \mathbf{z}), \tag{5.3}$$

respectively, which include random selection and allocation as special cases. If $\mathbf{z}$ is known this leads to the full likelihood

$$\begin{aligned} L &\propto p(s \mid \mathbf{z}) p(s^{(i)}, i = 1, 2 \mid s, \mathbf{z}) \int f(\mathbf{y} \mid \mathbf{z}; \boldsymbol{\theta}) g(\mathbf{z}; \lambda)\, d\mathbf{y}_{\bar{s}} \\ &\propto p(s \mid \mathbf{z}) p(s^{(i)}, i = 1, 2 \mid s, \mathbf{z}) g(\mathbf{z}; \lambda) \prod_{i \in s^{(1)}} f_1^*(y_i^{(1)} \mid z_i; \theta^{(1)}) \\ &\quad \times \prod_{i \in s^{(2)}} f_2^*(y_i^{(2)} \mid z_i; \theta^{(2)}) \end{aligned} \tag{5.4}$$

under the model (5.1), and the sampling and assignment mechanisms are ignorable for all model-based inferences.

If the analyst only has the data d_s given by (5.2), then the likelihood, to the analyst, is

$$L_a \propto \prod_{i \in s^{(1)}} f_1^*(\,) \prod_{i \in s^{(2)}} f_2^*(\,) \int p(s \mid \mathbf{z}) p(s^{(i)}, i = 1, 2 \mid \mathbf{z}, s) g(\mathbf{z}; \lambda)\, d\mathbf{z}_{\bar{s}}, \tag{5.5}$$

where $\mathbf{z}_{\bar{s}}$ is the vector $(z_i; i \notin s)$. From (5.5) we see that $\theta^{(1)}$ and $\theta^{(2)}$ appear only in the face value likelihood

$$L_f \propto \prod_{i \in s^{(1)}} f_1^*(y_i^{(1)} \mid z_i; \theta^{(1)}) \prod_{i \in s^{(2)}} f_2^*(y_i^{(2)} \mid z_i; \theta^{(2)}), \tag{5.6}$$

and thus the sampling and assignment mechanisms can be ignored for inferences about functions of $\theta^{(1)}$ and $\theta^{(2)}$ such as the superpopulation causal effect $\theta^{(1)} - \theta^{(2)}$.

In this paper we have followed Rubin (1978) and defined the true population causal

effect as $\bar{Y}^{(1)} - \bar{Y}^{(2)}$. Inferences about this effect require predictions of the unobserved values $Y_i^{(j)}$, which in turn will depend on predictions of the unobserved values in the vector $\mathbf{z}_{\bar{s}}$. In general the predictive distribution of $\bar{Y}^{(1)} - \bar{Y}^{(2)}$ depends on the sampling and assignment mechanisms, unless condition 1′ of Sugden & Smith (1984) holds, in which case $p(s \mid \mathbf{z}) = p(s \mid \mathbf{z}_s)$ with a similar condition for the assignment mechanism.

In order to demonstrate the effect of selection on predictive inferences we assume that $f_j^*(y_i^{(j)} \mid z_i; \theta^{(j)})$ is $N(\beta^{(j)} z_i, \sigma^{(j)} z_i)$, for $j = 1, 2$, and

$$g^*(z_i; \lambda) = \frac{\lambda^k}{\Gamma(k)} z_i^{k-1} e^{-\lambda z_i} \quad (i = 1, \ldots, N). \tag{5.7}$$

We compare the effects for samples of size $n = 2$ under two sampling mechanisms.

Scheme I. Select two units at random without replacement.

Scheme II. Select the sample with probability proportional to

$$\sum_{i \in s} z_i \Big/ \sum_{i=1}^{N} z_i,$$

the Lahiri sampling scheme for unbiased ratio estimation.

We assume that the treatment assignment scheme depends only on s and $\mathbf{z}_s$ and hence is of the form

$$p(s^{(j)}, j = 1, 2 \mid \mathbf{z}_s, s),$$

which is true for any form of controlled assignment including random assignment. This form of assignment mechanism is ignorable for predictive inference.

The predictive distribution $Y_i^{(j)}$ is given by $f(y_i^{(j)} \mid d_s)$, where d_s is the data (5.2). The mean of this distribution under the distribution (5.7) is

$$E(Y_i^{(j)} \mid d_s) = E_{z_i}\{E(Y_i^{(j)} \mid d_s, z_i)\} = \tilde{\beta}^{(j)} E(z_i \mid d_s), \tag{5.8}$$

where $\tilde{\beta}^{(j)}$ is the mean of the posterior distribution of $\beta^{(j)}$ and $E(z_i \mid d_s)$ is the mean of the predictive distribution of z_i. Thus the predictive distribution of $\bar{Y}_1 - \bar{Y}_2$ has mean

$$E(\bar{Y}_1 - \bar{Y}_2 \mid d_s) = (\tilde{\beta}^{(1)} - \tilde{\beta}^{(2)}) Z^*, \tag{5.9}$$

where

$$Z^* = \frac{1}{N} \sum_{i=1}^{N} E(z_i \mid d_s), \tag{5.10}$$

is the predictive mean of the population mean $\bar{Z}$.

If we assume that the priors on $\beta^{(1)}$, $\beta^{(2)}$, $\log \sigma^{(1)}$, $\log \sigma^{(2)}$, $\log \lambda$ are independent and uniform then it follows that the posterior distributions of $\beta^{(j)}$ and $\sigma^{(j)}$, for $j = 1, 2$, are independent of the sampling mechanism and

$$f(\beta^{(j)} \mid d_s, \sigma^{(j)}) \sim N\left(\frac{y_i^{(j)}}{z_i}, \frac{\sigma^{(j)2}}{z_i}\right) \quad (i \in s^{(j)}; j = 1, 2)$$

$$f(\sigma^{(j)} \mid d_s) = \frac{d\sigma^{(j)}}{\sigma^{(j)}} \quad (j = 1, 2). \tag{5.11}$$

Note that since the effective sample size for each treatment is one the posterior distribution of $\sigma^{(j)}$ is unchanged.

Under Scheme I, simple random sampling, the joint predictive distribution of

$$x_i = z_i \Big/ \sum_{j \in s} z_j \quad (i \in \bar{s})$$

is a symmetric inverse Dirichlet (Johnson & Kotz, 1970, p. 238) with marginal mean $E(x_i \mid d_s) = k/(2k-1)$. Thus

$$E(z_i \mid d_s) = 2k\bar{z}_s/(2k-1),$$

and hence

$$NZ_I^* = 2\bar{z}_s + (N-2)2k\bar{z}_s/(2k-1) = 2\bar{z}_s(Nk-1)/(2k-1). \tag{5.12}$$

Under Scheme II, with

$$p(s \mid \mathbf{z}) \propto \sum_{j \in s} z_j \Big/ \sum_{j=1}^{N} z_j,$$

the joint predictive density is again a symmetric inverse Dirichlet but now the marginal mean is $E(z_i \mid d_s) = \bar{z}_s$. Hence the prediction of the population mean is

$$NZ_{II}^* = N\bar{z}_s. \tag{5.13}$$

Substituting (5.13) and (5.12) in (5.9) gives the predictive means of the true casual effect under the two mechanisms. The message is clear in that the predictive means differ despite the fact that the sampling mechanisms are of the form $p(s \mid \mathbf{z})$, which is often termed uninformative. These mechanisms are only ignorable for predictive inferences if $\mathbf{z}$ is known. In our example the analyst only knows $\mathbf{z}_s$, that part of $\mathbf{z}$ which appears in the sample, and so the sampling mechanism is not ignorable in general and the inferences will differ for different sampling mechanisms.

The overall conclusion of this paper is that even in a model-based analysis of data the analyst should always write down the full likelihood (3.13) or (3.19), which represents the joint distribution of the observations *and* of the sampling and assignment indicator variables. Failure to do this leads to inferences based on the face value likelihood (3.5) which may have no validity. In particular for observational studies where the treatment assignment mechanism is unknown, causal inference based on the apparent population has doubtful validity. When both mechanisms are under the scientist's control, then randomisation accompanied by complete reporting of the covariates $\mathbf{z}$ guarantees that the mechanisms can be ignored for statistical inferences based on the model (3.4). Clearly the scientific validity of the inferences will still depend on the validity of the underlying model.

Acknowledgement

An earlier, and shorter, version of this paper was presented at the I.S.I. Session in Amsterdam in 1985. The authors gratefully acknowledge the helpful comments of the referees.

References

Breslow, N.E. & Day, N.E. (1980). *Statistical Methods in Cancer Research I: The Analysis of Case-Control Studies*. Lyon: International Agency for Research on Cancer.

Brewer, K.R.W., Foreman, E.K., Mellor, R.W. & Trewin, D.J. (1977). Use of experimental design and population modelling in survey sampling. *Bull. Int. Statist. Inst.*, 41st Sess., Bk. 3, 173–190.

Cox, D.R. (1958). *The Planning of Experiments*. New York: Wiley.

Dawid, A.P. & Dickey, J.M. (1977). Likelihood and Bayesian inference from selectively reported data. *J. Am. Statist. Assoc.* **72**, 845–850.

Dawid, A.P. (1979). Conditional independence in statistical theory (with discussion). *J. R. Statist. Soc.* B **41,** 1–31.

Doll, R. & Hill, A.B. (1964). Mortality in relation to smoking: Ten years' observations of British doctors. *Brit. Med. J.* 1964, 1399–1410 and 1460–1467.

Fienberg, S.E. & Tanur, J.M. (1987). Experimental and sampling structures: parallels diverging and meeting. *Int. Statist. Rev.* **55,** 75–96.

Holland, P.W. (1986). Statistics and causal inference. *J. Am. Statist. Assoc.* **81,** 945–970.

Johnson, N.L. & Kotz, S. (1970). *Distributions in Statistics, 3, Continuous Multivariate Distributions.* New York: Wiley.

Little, R.J.A. (1982). Models for non-response in sample survey. *J. Am. Statist. Assoc.* **77,** 237–250.

Rosenbaum, P.R. & Rubin, D.R. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70,** 41–55.

Royall, R.M. (1976). Current advances in sampling theory: implications for human observational studies. *Am. J. Epidemiol.* **104,** 463–473.

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

Rubin, D.B. (1978). Bayesian inference for causal effects: the role of randomisation. *Am. Statistician* **6,** 34–58.

Rubin, D.B. (1980). Discussion of paper by D. Basu. *J. Am. Statist. Assoc.* **75,** 591–593.

Scott, A. J. (1977). On the problem of randomization in survey sampling. *Sankhyā* C **39,** 1–9.

Scott, A.J. & Wild, C.J. (1986). Fitting logistic models under case control or choice-based sampling. *J. R. Statist. Soc. B* **48,** 170–182.

Smith, T.M.F. (1976). The foundations of survey sampling: a review. *J. R. Statist. Soc.* A **139,** 183–204.

Smith, T.M.F. (1983). On the validity of inferences from non-random samples. *J. R. Statist. Soc.* A **146,** 394–403.

Sugden, R. A. & Smith, T. M. F. (1984). Ignorable and informative designs in survey sampling inference. *Biometrika* **71,** 495–506.

Sugden, R.A. (1988). The 2×2 table in observational studies. In *Bayesian Statistics 3,* Proceedings of the 3rd international meeting on Bayesian statistics, Valencia, Spain. To appear.

Wold, H.O. (1967). Nonexperimental statistical analysis from the general point of view of scientific method. *Bull. Int. Statist. Inst.*, 36th Sess., Bk 1, 391–424.

Résumé

Un cadre général est donné pour l'examen de rôle des mécanismes de selection d'unités et d'allocations de traitements dans les experiences, enquêtes et études d'observation. Nous donnons des conditions impliquant que ces mécanismes peuvent être ignorés pour des inférences fondées sur les modèls. Nous montrons dans les examples comment des inférences peuvent tenir compte des mécanismes quand ces conditions ne sont pas satisfaites.

[*Received April* 1987, *revised November* 1987]