Biometrika (1984), **71**, 3, *pp.* 495–506
Printed in Great Britain

Ignorable and informative designs in survey sampling inference

BY R. A. SUGDEN
Department of Mathematics, University of London, Goldsmiths' College, London, U.K.

AND T. M. F. SMITH
Faculty of Mathematical Studies, University of Southampton, Southampton, U.K.

SUMMARY

The role of the sample selection mechanism in a model-based approach to finite population inference is examined. When the data analyst has only partial information on the sample design then a design which is ignorable when known fully may become informative. Conditions under which partially known designs can be ignored are established and examined for some standard designs. The results are illustrated by an example used by Scott (1977).

Some key words: Bayesian predictive inference; Face-value likelihood; Finite population; Model-based inference; Partial design information; Regression through the origin; Secondary analysis; Selection mechanism.

1. INTRODUCTION

In a model-based approach to survey sampling inference the role played by the survey design is not completely clear. Some authors such as Godambe (1966) and Basu (1971) have asserted that the randomization distribution has no role in a purely model-based approach. Others, such as Ericson (1969), Royall & Pfeffermann (1982), Little (1982) and Smith (1983), recognize that random sampling schemes may have desirable robustness properties in a model-based approach but that other designs, such as balanced samples, may be better for some purposes.

Scott (1977) and Scott & Smith (1973) examine the conditions under which any survey design can be ignored for Bayesian inference. If these conditions are not satisfied then averages over subsets of the randomization distribution may be necessary for valid Bayesian inference. Rubin (1976), in a fundamental paper on missing values, interprets sampling as a special case of missing values and establishes conditions under which the selection method can be ignored for model-based inferences from the Bayesian, likelihood or sampling theory viewpoints. Little (1982) extends Rubin's results to nonresponse and Smith (1983) to nonrandom designs such as quota sampling which depend on response variables.

The key to understanding the role of survey design is to follow Scott (1977) and introduce the idea of design variables, known to the sampler before the sample is drawn, in addition to the response variables measured in the survey. For a finite population of N units we define these as follows.

Response variables, $y = (y_1, \ldots, y_N)^{\mathrm{T}}$, known only for those units observed in the sample, s, which is a subset of n units from N.

Design variables, $z = (z_1, \ldots, z_N)^T$, known for all units in the population, which may include label information such as cluster or stratum indicator variables which determine group membership, other group variables and quantitative variables such as measures of size.

In a model-based approach the matrix of values y is viewed as the realization of a random matrix Y. The degenerate case when $Y = y$ is fixed is used in the classical randomization approach to survey sampling inference. The design variables z can also be regarded as a realization of a random matrix, or they can be regarded as fixed in the conditional distribution of Y given z.

Drawing a sample partitions the labels into two sets, s and $\bar{s}$, defining submatrices y_s, $y_{\bar{s}}$ and z_s, $z_{\bar{s}}$. The selection of the observed sample, s, is the final part of the design stage of a sample survey and in probability sampling designs takes place according to a set of chosen selection probabilities

$$[\{s, p(s)\},\ s \in \mathscr{S}], \tag{1·1}$$

where $\mathscr{S}$ is the set of feasible samples. The choice of design is often an ill-defined process involving the determination of strata, size of cluster, measure of size of sampling units, overall sample size, sample allocation and method of unequal probability sampling. Frequently cost and administrative constraints dominate the choice rather than considerations of statistical inference. Nevertheless, we suppose that the selection process can be modelled by a sample selection scheme which depends on the design variables z and may depend also on the response variables y and a vector of parameters ψ. We write this scheme as

$$p(s \mid y, z;\ \psi), \quad s \in \mathscr{S}. \tag{1·2}$$

This is the discrete distribution of the sample indicator variable S which takes the value one for sampled units and zero otherwise. The realized value of S is also denoted by s. For surveys with nonresponse, or for quota samples, the selection may depend on y and ψ as well as on z; see Little (1982) and Smith (1983).

In this paper we concentrate on designs which depend only on the design variable z. Such designs, which include all random sampling designs, can be written

$$p(s \mid z), \quad s \in \mathscr{S}. \tag{1·3}$$

All designs of the form (1·3) satisfy the basic design assumption of Scott (1977), that

$$S \perp\!\!\!\perp Y \mid Z, \tag{1·4}$$

where the symbol $\perp\!\!\!\perp$ means independent as employed by Dawid (1979). Thus the sample indicator, S, and the response variable, Y, are conditionally independent given the value of the design matrix Z for all the N units in the population. When z is known (1·4) implies both of the conditions of Rubin (1976), missing at random and observed at random, for the ignorability of the process causing missing values. Sampling is regarded as a case where units are intentionally missing, or missing by design.

Formally, let A_s denote the event $S = s$. Then missing at random corresponds to

$$A_s \perp\!\!\!\perp Y_{\bar{s}} \mid y_s, z. \tag{1·5}$$

Similarly observed at random can be written

$$A_s \perp\!\!\!\perp Y_s \mid Y_{\bar{s}}, z, \tag{1·6}$$

where the conditional independence condition must hold for all values of the missing data, $Y_{\bar{s}}$. Conditions (1·5) and (1·6) are together weaker than (1·4). This is true even if they hold for all z, $s \in \mathscr{S}$ and (1·5) holds also for all possible data sets y_s. A sufficient condition for (1·4) is that in addition $Y_s \perp\!\!\!\perp Y_{\bar{s}} | z$ for all z. For the examples considered in Table 2 in § 4 the schemes satisfy (1·4) and can all be written in the form (1·3).

The object of this paper is to consider situations in which there is only partial knowledge of the values in the matrix of design variables z. The person drawing the sample, the sampler, is assumed always to know the values in z, and so can ignore the sample selection scheme in making model-based inferences (Scott, 1977; Rubin, 1976). Frequently, however, the survey data are analysed by a research worker or statistician, the analyst, who is different from the sampler. If the sampler has not included the values of z in the data then the analyst must make inferences about Y based on the knowledge of z gleaned from the sample and its selection mechanism, together with other auxiliary information.

Our basic question is under what circumstances can the analyst ignore the sample selection scheme? In § 2 we define various types of partial information. In § 3 we develop some general model-based theory for inference based on partial information and establish conditions for the ignorability of the sample selection scheme. In § 4 we examine some standard probability designs to see to what extent they satisfy the conditions for ignorability. In § 5 we look at some models for suvey data in the light of our previous results.

2. Partial design information

When the analyst of survey data is not the sampler the values in z for all N units of the population may not be known. In this case knowledge of the selection probabilities in (1·1) may carry information about the unknown values of the design variables z and this information may be useful in a statistical analysis. For example, the selected units may suggest that the sample is unexpected in that the inclusion probabilities of the units in s are all below expectation. The data in (1·1) may then be used by the analyst to correct this possibly misleading or unrepresentative sample.

Let the data available to the analyst be

$$e_s = (s, y_s, d_s), \tag{2·1}$$

where d_s is data derived from knowledge of the selection mechanism (1·2) and from values of the selection probabilities (1·1) if these are available, and also from any known values or functions of z. When z is the realized value of a random matrix Z we write $D_s = D_s(Z)$ as the random quantity with realized value $d_s = D_s(z)$ for fixed s.

The case where labels s are not available is not covered; see Scott & Smith (1973). D. B. Rubin, in an unpublished conference paper, considers the special case where d_s consists of a single variable observed for all units of the population. His model in a Bayesian framework has joint exchangeability for both response and design variables, so that s carries no information in addition to data (y_s, z_s).

The design information d_s may take many forms depending on the knowledge of z available to the analyst. It is convenient to introduce the idea of a proxy design variable which is a function of z summarizing this information.

Definition. A proxy design variable $w = W(z)$ is a vector function of the design variables z. We write $W = W(Z)$ as the random variable with realized values w and w_s the subvector for a sample s as for z_s.

In a stratified population the formation of strata may have been based on knowledge of quantitative variables z, for example by ordering the population on the values of a component of z or by grouping the population into fixed ranges of a component variable. If the analyst knows stratum membership only, then d_s corresponds to a proxy design variable $w = W(z)$ which determines the stratum indicator variables. Many designs such as stratifed random sampling permit the determination of w from the selection probabilities (1·1). For poststratification, however, the membership of strata may be known only for sampled units so that w_s alone is known to the analyst.

For probability proportional to size sampling when all the inclusion probabilities are known, a vector w can be inferred for the whole population. If the probabilities are known only for sampled units then d_s is of the form w_s. In general the size measures z cannot be deduced from w unless the population mean $\bar{z}$ is known and the sample size n is fixed. If both w_s and z_s are known, then $\bar{z}$ can be deduced from the proportionality of w_i and z_i.

For inference by the analyst we consider six cases of design information, case (i) corresponding to complete design information and cases (ii)–(vi) to partial design information. In all cases (s, y_s) is assumed to be known. The design information in d_s in each case is given in Table 1.

Table 1. *Design information in* d_s

Case	d_s	Case	d_s
(i)	z	(iv)	w_s, z_s
(ii)	w, z_s	(v)	z_s
(iii)	w	(vi)	w_s

In § 3 we show that a key condition for the ignorability of selection for inference given the design information is as follows.

Condition 1. The relation $A_s \perp\!\!\!\perp Z \mid D_s$ holds.

Various designs are examined in § 4 in the light of this condition which is equivalent to the assertion that D_s is sufficient for z in the model $p(s \mid z)$. For case (i), where $d_s = z$, Condition 1 is always satisfied.

Note that in the above we are assuming that the analyst knows the selection mechanism which is part of his statistical model. For example he knows that stratified random sampling has been employed, the rule by which strata are formed and also the sample size allocation rule, but not necessarily the strata themselves or equivalently the selection probabilities (1·1). An unknown selection mechanism, as for example in quota sampling (Smith, 1983), requires further assumptions.

3. MODEL-BASED INFERENCE

3·1. *Introduction*

We suppose that the analyst can formulate a model, or family of distributions, for the variables Y, Z in the population which we write

$$f(y, z;\, \theta, \phi) = f(y \mid z;\, \theta)\, g(z;\, \phi), \tag{3·1}$$

where θ and ϕ are distinct in the sense of Rubin (1976). The analyst may be interested in making inferences about θ or ϕ or about functions of θ and ϕ such as parameters in the

marginal distribution of Y. Predictive inferences may also be made about the unobserved random variables, $Y_{\bar{s}}$, when the target of interest is a descriptive statistic such as a finite population total or mean. Such descriptive inferences however, may be parametric rather than predictive for models based on, say, finite exchangeability (Sugden, 1979) or the 'classical model' (Royall, 1968).

If selection is ignored then model-based inferences are usually made from the joint distribution of Y_s and D_s holding s fixed (Rubin, 1976). This joint distribution is found by integrating (3·1) over the z's that could have generated d_s for the fixed s, that is

$$f_s(y_s, d_s;\, \theta, \phi) = \int_{\mathscr{D}_s} f_s(y_s \mid z;\, \theta)\, g(z;\, \phi)\, dz, \tag{3·2}$$

where $\mathscr{D}_s = \{z\colon D_s(z) = d_s\}$. We use the subscript s for notational clarity to denote densities derived from (3·1) for the fixed s. The likelihood function based on (3·2) is called the face-value likelihood by Dawid & Dickey (1977). However, the full distribution of the data e_s in (2·1) is

$$f(s, y_s, d_s;\, \theta, \phi) = \int_{\mathscr{D}_s} f_s(y_s \mid z;\, \theta)\, g(z;\, \phi)\, p(s \mid z)\, dz, \tag{3·3}$$

and the full likelihood is based on this distribution. The analyst can ignore the effects of selection if inferences based on (3·2) are equivalent to those based on (3·3), and this will depend on the type of inference, Bayesian, likelihood or sampling theory, and on the target of inference θ, ϕ or $Y_{\bar{s}}$. The choice of target may suggest the conditioning of (3·3) or (3·2) under a sampling theory or likelihood approach.

3·2. *Likelihood/Bayesian inference on θ and ϕ*

If the selection mechanism satisfies Condition 1 for the observed d_s then Condition 1′ is satisfied.

Condition 1′. The relation $A_s \perp\!\!\!\perp Z \mid D_s(Z) = d_s$ holds. This implies that $p(s \mid z) = p(s \mid d_s)$ for any z in $\mathscr{D}_s$.

Equation (3·3) now becomes

$$f(e_s;\, \theta, \phi) = p(s \mid d_s)\, f_s(y_s, d_s;\, \theta, \phi), \tag{3·4}$$

where $p(s \mid d_s)$ is not generally equal to $p(s)$ in (1·1).

We conclude that under Condition 1′, Bayesian or likelihood inference on θ and ϕ leads to identical results with either (3·2) or (3·3), that is the design is ignorable.

Remark 1. Condition 1′ is equivalent to, in an obvious notation, the statement

$$\mathrm{pr}\,(A_s \mid z) = \mathrm{pr}\,(A_s \mid d_s).$$

D. B. Rubin, in his unpublished conference paper, would say that d_s is an 'adequate' summary of z. However his partial design information is a special case of ours.

Remark 2. The design is ignorable whatever the sample outcome if Condition 1 holds for all s.

If Condition 1′ is not satisfied, the design may still be ignorable for inferences on either θ or ϕ if one of two conditions on the model are true.

(i) For ϕ alone, suppose that D_s is sufficient for ϕ in the marginal model $g(z \mid \phi)$ for the fixed s; that is Condition 2.

Condition 2. The relation $Z \perp\!\!\!\perp \phi \mid D_s$ holds.

Then (3·3) can be written in the form

$$h(s, y_s \mid d_s;\, \theta)\, g_s(d_s;\, \phi), \tag{3·5}$$

where the first term depends only on θ and incorporates selection effects and the second term is the marginal distribution of $D_s = D_s(Z)$ for the fixed s. Similarly (3·2) becomes

$$f_s(y_s \mid d_s;\, \theta)\, g_s(d_s;\, \phi), \tag{3·6}$$

where the first term does not depend on ϕ because of Condition 2.

Comparing (3·5) and (3·6) we conclude that, under Condition 2, likelihood/Bayesian inferences on ϕ satisfy ignorability but not those on θ.

(ii) For θ alone, alternatively suppose that the conditional distribution of the responses of sampled units s depends only on z through the design information d_s; that is Condition 3.

Condition 3. For all θ, $Y_s \perp\!\!\!\perp Z \mid D_s;\, \theta$.

Then the term $f_s(y_s \mid z;\, \theta)$ is constant over z in $\mathscr{D}_s$ and (3·3) becomes

$$f_s(y_s \mid d_s;\, \theta) \int_{\mathscr{D}_s} p(s \mid z)\, g(z \mid \phi)\, dz. \tag{3·7}$$

Similarly (3·2) reduces to (3·6).

Comparing (3·6) and (3·7) we conclude that, under Condition 3, likelihood/Bayesian inferences on θ satisfy ignorability but not those on ϕ.

Remark. If both Conditions 2 and 3 hold, for example when D_s includes both a sufficient statistic for ϕ and at least z_s, then the full likelihood from (3·3) reduces to (3·6). The design is therefore ignorable for likelihood/Bayesian inference on either θ or ϕ or indeed functions of θ and ϕ. This conclusion is equivalent to that when only Condition 1′ holds. However, the factorization of (3·3) in (3·6) shows that in this case θ and ϕ are a posteriori independent.

3·3. *Predictive inference*

Under a Bayesian approach the predictive distribution of $Y_{\bar{s}}$ is the conditional distribution of $Y_{\bar{s}}$ given the full data (2·1). When Condition 1′ holds we have that

$$f(y, s, d_s;\, \theta, \phi) = p(s \mid d_s)\, f_s(y, d_s;\, \theta, \phi). \tag{3·8}$$

Integration out of θ, ϕ over their joint prior gives

$$f(y, s, d_s) = p(s \mid d_s)\, f_s(y, d_s). \tag{3·9}$$

Now

$$f(y \mid s, d_s) = \frac{f(y, s, d_s)}{f(s, d_s)} = \frac{p(s \mid d_s)\, f_s(y, d_s)}{p(s \mid d_s)\, f_s(d_s)} = f_s(y \mid d_s). \tag{3·10}$$

Thus for the predictive distribution when Condition 1′ holds

$$f(y_{\bar{s}} \mid y_s, s, d_s) = \frac{f_s(y \mid d_s)}{f_s(y_s \mid d_s)} = f_{\bar{s}|s}(y_{\bar{s}} \mid y_s, d_s), \tag{3·11}$$

and selection can be ignored. A similar conclusion has been reached by D. B. Rubin who considers the special case of exchangeable priors in the unpublished conference paper mentioned above. In this case the suffices $\bar{s}, s$ on our densities can be dropped.

Remark. Neither Condition 2 nor 3 is sufficient to give ignorability for prediction. Even if both conditions hold predictive inferences still depend on the design through a term of the form $h(s, y_{\bar{s}}, d_s; \theta)$.

3·4. *Sampling theory inference about θ and ϕ*

In a model based approach, sampling theory inference refers to inference under the model (3·3) and not simply to either the randomization distribution $p(s \mid z)$ or the model (3·2). A problem in this approach is to determine how the inferences should be conditioned. In particular can we examine the distribution of y_s conditional on either s, d_s or both? This usually requires an appeal to some form of ancillarity.

For the data $e_s = (s, y_s, d_s)$ and under Condition 3, (3·3) reduces to (3·7) and the conditional distribution is

$$f(y_s \mid s, d_s; \theta, \phi) = f_s(y_s \mid d_s; \theta). \tag{3·12}$$

The factorization (3·7) suggests that for inferences about θ the design can be ignored and that the analyst should employ (3·12) for the observed s.

For inferences about (θ, ϕ) jointly, Condition 1 must hold for ignorability. We then find that the conditional distribution is

$$f(y_s \mid s, d_s; \theta, \phi) = f_s(y_s, d_s; \theta, \phi)/g_s(d_s; \phi), \tag{3·13}$$

which shows that the design can be ignored and that again the analyst should condition on d_s. If in addition Condition 2 holds with D_s sufficient for ϕ, then the conditional distribution further reduces to (3·12) and the design is again ignorable for θ and inference on ϕ can be made through the sufficient statistic D_s. In general however the design is not ignorable unless Condition 1 holds and in addition $p(s \mid d_s) = p(s)$ for all s.

4. Ignorable designs

In this section we consider some examples of common probability sampling designs and purposive designs and discuss whether they satisfy the main condition for ignorability, Condition 1:

$$A_s \perp\!\!\!\perp Z \mid D_s,$$

for all s in each of the cases of design information listed in § 1. The results are summarized in Table 2.

Example 1: Simple random sampling. Here

$$p(s \mid z) = 1 \Big/ \binom{N}{n}$$

for all z whenever $|s| = n$, and is zero otherwise.

This design is independent of any variable, design or response, so certainly satisfies Condition 1. Scott (1977) regards this as the only uniformly noninformative design.

Example 2: Stratified random sampling. Assume that there is a single design variable, a measure of 'size', which is used to construct size strata $U_1, U_2, \ldots, U_H$, subsets of the labels, or equivalently a stratum indicator variable $w = W(z)$, where $w_i = h$ $(i \in U_h)$.

Two methods of forming strata are considered.

Method A: fixed endpoints, random sizes. For real numbers

$$0 = a_0 < a_1 < \ldots < a_{H-1} < a_H = \infty,$$

define $w_i = h$ if $z_i \in [a_{h-1}, a_h)$.

Method B: fixed sizes, random endpoints. Let the H strata be of sizes $N_1, \ldots, N_H$. Form the strata by ranking, where $z_{(1)} < \ldots < z_{(N)}$ are the order statistics of z, and define $w_i = h$ if $z_i \in [z_{(M_h+1)}, z_{(M_h+1)}]$, where $M_h = \Sigma N_j$, with the sum over $j = 1, \ldots, h-1$. The sample s is now selected according to

$$p(s \mid z) = 1 \Big/ \prod_{h=1}^{H} \binom{N_h}{n_h},$$

whenever $|s \cap U_h| = n_h$ $(h = 1, \ldots, H)$ and zero otherwise, where the sample allocation $n = (n_1, n_2, \ldots, n_H)$ is fixed.

Note that there is a one-to-one correspondence for this design between the selection probabilities (1·1) and the indicator w. This implies that whenever w or (1·1) is included in the design information then Condition 1 holds. When the analyst knows only the strata to which sampled units belong, that is w_s alone is observed, then the fixed stratum sizes in method B imply that Condition 1 is satisfied but not for method A where the sizes are unknown. If z_s only is available then Condition 1 is not satisfied even for B.

When the known sample allocation rule is such that different sampling fractions are used in each stratum, the inclusion probabilities $\pi_i = n_h/N_h$ $(i \in U_h)$ or equivalently the propensity scores, by Rosenbaum & Rubin (1983) and by D. B. Rubin in his conference paper, are also in one-to-one correspondence with the indicator w. Thus Condition 1 is satisfied when the whole vector of propensity scores is known. In general of course knowing the inclusion probabilities for all units is not sufficient for D_s to satisfy Condition 1.

Example 3: Probability proportional to size sampling. Let the design variable again be a measure of size and let the proxy design variable $w_i = (nz_i)/(N\bar{z})$ be the inclusion probability for unit i, where $\bar{z}$ is the population mean of the design variable and n is the sample size.

Two methods of achieving these probabilities are considered.

Scheme 1. For conditional Poisson sampling (Hájek, 1981, p. 132),

$$p(s \mid z) \propto \prod_{i \in s} w_i \prod_{i \notin s} (1 - w_i),$$

if s contains n distinct units, zero otherwise.

Scheme 2. For the Sampford–Durbin rejective scheme:

(i) select first unit with drawing probabilities $\{w_i/n,\ i = 1, \ldots, N\}$, and then replace it;

(ii) select subsequent units with probability proportional to $w_j/(1-w_i)$ with replacement, $j \neq i$;

(iii) if achieved s contains n distinct units accept it, otherwise start again.

Knowledge of the w_i's for all units is sufficient to satisfy Condition 1 using both methods as the inclusion probabilities determine the joint inclusion probabilities and also $p(s)$. This is not true in general for all probability proportional to size schemes so the conclusion of D. B. Rubin in his conference paper that the inclusion probabilities are propensity scores which form an adequate summary must be tempered by the condition that the scheme is of a similar type to scheme A or B.

These propensity scores are often used as inverse probability weights attached to sampled units to obtain approximately unbiased estimators with respect to the randomization distribution. When they are known only for sampled units, that is w_s is known, then Condition 1 is no longer satisfied for either A or B. The value of $p(s \mid z)$ depends also on the inclusion probabilities for unsampled units.

Example 4: Purposive sampling. The simplest form is a nonrandomized design

$$p(s \mid z) = \begin{cases} 1 & s = s_0(z), \\ 0 & \text{otherwise}, \end{cases} \tag{4·1}$$

where $s_0(.)$ is a known indicator function.

(a) Consider for example the population of patients arriving in a doctor's waiting-room. Let z_i be the time of arrival for the ith patient and let W be a proxy design variate denoting order of arrival so that w is a vector of ranks. The sampling scheme selecting the first n patients to arrive is purposive with

$$s_0(z) = \{i:\ 1 \leqslant w_i(z) \leqslant n\}.$$

A similar and mathematically identical example occurs when the n largest units are selected with probability one in order to achieve optimality for estimation under a regression model through the origin (Royall, 1970).

(b) Other forms of purposive sampling, such as balanced sampling (Royall & Herson, 1973), may sometimes involve some element of randomization. A particularly simple case is where the sample s is selected at random from a set of feasible samples $\mathscr{S}_0(z)$, assumed to be nonempty for all z, which satisfy the conditions of balance, thus

$$p(s \mid z) = \begin{cases} 1/k(z) & s \in \mathscr{S}_0(z), \\ 0 & \text{otherwise}, \end{cases} \tag{4·2}$$

where $k(z)$ is the number of feasible samples.

As an example, consider again a regression model where z_i is the 'size' of the ith unit. A simple balanced sampling scheme which may be adopted for robustness (Royall & Herson, 1973) is given by

$$\mathscr{S}_0(z) = \{s:\ |s| = n,\ \bar{z}_s = \bar{z}\},$$

where $\bar{z}_s$ is the sample mean. If there is a unique balanced sample, so that $k(z) = 1$ for

all z, then the scheme is of type (4·1), otherwise of type (4·2). Note that the condition $k(z)$ positive for all z may place a severe restriction on the space of design populations.

Whether Condition 1 is satisfied or not depends on the amount of design information D_s available for each observed s.

We require for all s in $\mathscr{S}$, whenever $D_s(z) = D_s(z')$,

$$p(s \mid z) = p(s \mid z'). \tag{4·3}$$

For example (a) this is clearly satisfied when w or even only w_s is available but not when z_s only is observed, case (v).

Example (b) is considerably less straightforward. If only ranking information is available, the condition is clearly not satisfied.

When $D_s(z) = z_s$, case (v), we require $\bar{z}$ to be known and fixed before sampling, or else $\bar{z}'$ can differ from $\bar{z}$ even though z and z' agree on the subset s. Alternatively if D_s includes the observed value of $\bar{z}$, such that $\bar{z} = \bar{z}_s$ for an observed s, then (4·3) is satisfied. This continues to hold if $\bar{z}_s$ rather than z_s is available. These statements are however only true if the assumption that there is a unique 'balanced' sample holds so that the scheme is of form (4·1). Otherwise the value of $k(z)$ may carry information. If $k(z)$ is constant over all z, for example if the possible design populations are permutations of each other so the complete set of order statistics are known, then the selection scheme carries no information in the sense that (4·3) is satisfied. If $k(z)$ is not constant then the design information D_s should include $\bar{z}_s$, $\bar{z}$ and the value of $k(z)$ itself or equivalently $p(s \mid z)$ for the observed s, in order to satisfy (4·3) and hence Condition 1. Face-value likelihood inferences using (3·2) may become quite complex.

In summary, Table 2 shows whether the ignorability Condition 1 is satisfied or not.

Table 2. *Satisfaction of ignorability Condition* 1

		Design						
							4: purposive	
	Data	1	2: STRS		3: PPS		a	b
Case	set	SRS	a	b	a	b		($\bar{z}$ known)
(i)	s, y_s, z	√	√	√	√	√	√	√
(ii)	s, y_s, w, z_s	√	√	√	√	√	×	√
(iii)	s, y_s, w	√	√	√	√	√	×	×
(iv)	s, y_s, w_s, z_s	√	×	√	×	√	×	√
(v)	s, y_s, z_s	√	×	×	×	×	×	√
(vi)	s, y_s, w_s	√	×	√	×	√	×	×

SRS, simple random sampling; STRS, stratified random sampling; PPS, probability proportional to size sampling.

Before embarking on an analysis of survey data collected by others an analyst must examine his data set and his knowledge of the selection mechanism carefully to see if Condition 1 is satisfied or not. If not, then the design forms an explicit part of the model-based inference, and is 'informative' (Scott, 1977).

5. AN EXAMPLE: BIVARIATE SUPERPOPULATION

Suppose (Y_i, Z_i) $(i = 1, \ldots, N)$ are independent, each with known distribution

$$f^*(y_i \mid z_i;\, \theta)\, g^*(z_i;\, \phi). \tag{5·1}$$

This model applies when all the information about the structure of the finite population as it relates to Y is contained in the values of Z and interest centres on the regression

relationship between Y and Z. We assume that data are available on sampled units only so that $d_s = z_s$, case (v). It follows immediately from the model that Condition 3 is satisfied and the design is ignorable for inferences about θ which therefore use the face-value likelihood

$$\prod_{i \in s} f^*(y_i \mid z_i; \theta). \tag{5·2}$$

This likelihood is appropriate for examining the regression relationship between Y and Z with linear regression being a special case. The model extends to multivariate Y, Z in an obvious way.

In general Bayes/likelihood inferences about ϕ in the marginal distribution of z will depend on the design. However, if the strong Condition 1′ is satisfied, for example by simple random sampling, then the design can be ignored and inferences made using

$$\prod_{i \in s} g^*(z_i; \phi). \tag{5·3}$$

Alternatively if the data are augmented by a sufficient statistic for ϕ then Condition 2 is satisfied and the design can again be ignored.

When both Conditions 1′ and 2 hold predictive inferences about $Y_{\bar{s}}$ can be made from the posterior distribution $f(y_{\bar{s}}, z_{\bar{s}} \mid e_s)$ ignoring the design. If in addition θ, ϕ are *a priori* independent then the posterior distribution factorizes and predictions can be made in a two-stage process by predicting z_i $(i \notin s)$ if we use the posterior distribution of ϕ derived from (5·3), and then predict y_i for the predicted z_i using the posterior distribution of θ derived from (5·2).

Consider a very simple example (Scott, 1977), with a single design variate measuring size related to the response variate by a regression through the origin. Specifically

$$f^*(y_i \mid z_i; \theta) = N(\beta z_i; \sigma^2 z_i), \quad \theta = (\beta, \sigma)^{\mathrm{T}},$$
$$g^*(z_i; \phi) = \lambda^k z_i^{k-1} e^{-\lambda z_i} / \Gamma(k).$$

Thus the sizes of the units in the finite population are a random sample of size N from a gamma distribution, $\gamma(k, \lambda)$ with known index k and unknown scale parameter λ.

We compare the effects of three schemes for selecting one unit:

Scheme 1: select the unit at random,

Scheme 2: select the largest size unit with probability one,

Scheme 3: select the unit with probability proportional to size.

If we assume great prior uncertainty, with independent constant prior densities for β, $\log \sigma$ and $\log \lambda$, the posterior distribution of β is

$$p(\beta \mid \sigma, y_s, z_s) \sim N(y_s/z_s;\ \sigma^2/z_s),$$

for any selection scheme since $d_s = z_s$ and Condition 3 is satisfied. Similarly, the posterior distribution of σ is unchanged since there is no scale information in a sample of size one.

The joint predictive density of $z_{\bar{s}}$ is most simply expressed in terms of $x_j = z_j z_s^{-1}$ $(j \notin s)$ as follows.

Scheme 1. Here

$$p(x_{\bar{s}} \mid z_s) = \frac{\Gamma(Nk)}{\{\Gamma(k)\}^N} \prod_{j \notin s} x_j^{k-1} \Big/ (1 + \sum_{j \notin s} x_j)^{Nk} \quad (x_j > 0),$$

an inverted Dirichlet distribution (Johnson & Kotz, 1970, p. 238).

Scheme 2. Here

$$p(x_{\bar{s}} \mid z_s) = \frac{N\Gamma(Nk)}{\{\Gamma(k)\}^N} \prod_{j \notin s} x_j^{k-1} \Big/ (1 + \sum_{j \notin s} x_j)^{Nk} \quad (0 < x_j < 1);$$

a truncated inverted Dirichlet distribution.

Scheme 3. Here

$$p(x_{\bar{s}} \mid z_s) = \frac{\Gamma(Nk+1)}{\{\Gamma(k)\}^N \Gamma(k+1)} \prod_{j \notin s} x_j^{k-1} \Big/ (1 + \sum_{j \notin s} x_j)^{Nk+1} \quad (x_j > 0),$$

also an inverted Dirichlet distribution.

The distributions are different and so the conclusion is clear; although the sampling schemes can be ignored for inferences about β they cannot be ignored for predictive inferences about the finite population total. This shows that those who adopt a superpopulation approach to inference from sample surveys cannot in general ignore the way in which the sample has been selected.

The work of T. M. F. Smith was supported by a grant from the Social Science Research Council.

REFERENCES

BASU, D. (1971). An essay on the logical foundations of survey sampling, Part I. In *Foundations of Statistical Inference*, Ed. V. P. Godambe and D. A. Sprott, pp. 203–42. Toronto: Holt, Rinehart and Winston.

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

DAWID, A. P. & DICKEY, J. M. (1977). Likelihood and Bayesian inference from selectively reported data. *J. Am. Statist. Assoc.* **72**, 845–50.

ERICSON, W. A. (1969). Subjective Bayesian models in sampling finite populations (with discussion). *J. R. Statist. Soc.* B **31**, 195–224.

GODAMBE, V. P. (1966). A new approach to sampling from finite populations. *J. R. Statist. Soc.* B **28**, 310–28.

HÁJEK, J. (1981). *Sampling from a Finite Population.* New York: Marcel Dekker.

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 surveys. *J. Am. Statist. Assoc.* **77**, 237–50.

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

ROYALL, R. M. (1968). An old approach to finite population sampling theory. *J. Am. Statist. Assoc.* **63**, 1269–79.

ROYALL, R. M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika* **57**, 377–87.

ROYALL, R. M. & HERSON, J. (1973). Robust estimation in finite populations, I. *J. Am. Statist. Assoc.* **68**, 880–9.

ROYALL, R. M. & PFEFFERMANN, D. (1982). Balanced samples and robust Bayesian inference in finite population sampling. *Biometrika* **69**, 401–10.

RUBIN, D. B. (1976). Inference and missing data. *Biometrika* **53**, 581–92.

SCOTT, A. J. (1977). Some comments on the problem of randomisation in surveys. *Sankhyā* C **39**, 1–9.

SCOTT, A. J. & SMITH, T. M. F. (1973). Survey design, symmetry and posterior distributions. *J. R. Statist. Soc.* B **35**, 57–60.

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. (1979). Inference on symmetric functions of exchangeable populations. *J. R. Statist. Soc.* B **41**, 269–73.

[*Received February* 1984. *Revised May* 1984]