International Statistical Review (1993), **61**, 2, pp. 317–337. Printed in Great Britain

The Role of Sampling Weights when Modeling Survey Data

Danny Pfeffermann

Department of Statistics, Hebrew University, Jerusalem, Israel 91905

Summary

The purpose of this paper is to provide a critical survey of the literature, directed at answering two main questions. i) Can the use of the sampling weights be justified for analytic inference about model parameters and if so, under what circumstances? ii) Can guidelines be developed for how to incorporate the weights in the analysis? The general conclusion of this study is that the weights can be used to test and protect against informative sampling designs and against misspecification of the model holding in the population. Six approaches for incorporating the weights in the inference process are considered. The first four approaches are intended to yield design consistent estimators for corresponding descriptive population quantities of the model parameters. The other two approaches attempt to incorporate the weights into the model.

Key words: Design consistency; Estimating functions; Nonignorable sampling design; Pseudo likelihood; Randomization distribution; Weighted distribution.

1 Introduction

Sampling weights weigh sample data to correct for the disproportionality of the sample with respect to the target population of interest. The weights reflect unequal sample inclusion probabilities and compensate for differential nonresponse and frame undercoverage. They are routinely included in survey data files released to analysts. The role of the sampling weights in the statistical analysis of survey data is however a subject of controversy among theorists. For descriptive inference, that is, inference about known functions of the finite population values, weighting of sample data is widely accepted although modifications to control variances are occasionally recommended. Yet, for analytical inference about model parameters, there is a wide spectrum of opinions on the role of the sampling weights, from modelers who view the weights as largely irrelevant to survey statisticians who incorporate the weights into every analysis.

In order to illustrate the controversy, consider the second National Health and Nutrition Examination Survey (NHANES) in the US (McDowell et al., 1981). The NHANES consists of a stratified four stage probability cluster sample of households. The primary sampling units (PSU's) are counties or groups of contiguous counties and the stratification is based on size, income and racial distribution. The selection of the PSU's and the three stage selection of persons within the PSU's is with unequal probabilities so as to oversample the poor, the young and the old age groups. Let $\pi_i = P(i \in s)$ define the sample inclusion probability for person i, $i = 1 \ldots N$. The probabilities π_i are products of the conditional selection probabilities at the various sampling stages. Let Y define a variable of interest with typical values Y_i, $i = 1 \ldots N$. For estimating the population mean $M = \sum_{i=1}^{N} Y_i/N$ for example, classical sampling theory advocates the use of estimators like,

$$\sum_{i \in S} (Y_i/\pi_i)/N, \qquad \sum_{i \in S} (Y_i/\pi_i)/\sum_{i \in S} (1/\pi_i), \qquad \left[\sum_{i \in S} (Y_i/\pi_i)/\sum_{i \in S} (x_i/\pi_i)\right]\sum_{i=1}^{N} (x_i/N)$$

and so forth where in the last estimator x is a concomitant variable known for all the population units. The attractive feature of these estimators is that they are unbiased, or approximately unbiased with respect to the randomization distribution, induced by the random selection of the sample, irrespective of the distribution of the Y-values in the population.

Suppose now that one is interested in regressing the Y values against a given set of regressor variables. For example, Harlan et al. (1985) fit regression models to the NHANES data with diastolic blood pressure as the dependent variable and blood lead levels, age and other health variables as the independent variables. The question which arises is whether the sampling weights should play a role when estimating the model parameters. For example, one could estimate the regression coefficients as

$$\hat{\boldsymbol{\beta}}_w = (X_s' W_s X_s)^{-1} X_s' W_s \mathbf{Y}_s = \left(\sum_{i \in S} w_i \mathbf{x}_i \mathbf{x}_i'\right)^{-1} \sum_{i \in S} w_i \mathbf{x}_i y_i \tag{1.1}$$

where $w_i = 1/\pi_i$, $\mathbf{x}_i$ is the vector of regressor values for unit i, $s = \{1 \ldots n\}$ represents the sample, $W_s = \text{diag}\,(w_1 \ldots w_n)$, $X_s = [\mathbf{x}_1 \ldots \mathbf{x}_n]'$ and $\mathbf{Y}_s = (y_1 \ldots y_n)'$. Clearly, the use of (1.1) cannot be justified in general based on optimality considerations.

The purpose of this article is to provide a critical survey of the literature, aimed at answering the following two main questions:

(1) Can the use of the sampling weights be justified for a model based inference and if so, under what circumstances?
(2) Can guidelines be developed for how to use the weights?

Put together, the two questions can be phrased as 'weighting: why, when and how?', the title of a talk by Kish (1990) which focuses on the use of the weights in descriptive analysis. The main conclusion of this study is that the sampling weights can play a vital role in two different aspects of the modeling process.

(1) The weights can be used to test and protect against nonignorable sampling designs which could cause selection bias.
(2) The weights can be used to protect against misspecification of the model holding in the population.

The robustness of inference procedures that incorporate the weights is obtained by changing the focus of the inference to finite population quantities. We discuss alternative definitions of the target parameters in Section 2. Section 3 discusses the conditions ensuring the ignorability of the design and in Section 4 we show how to use the weights to test that the design is ignorable with respect to a given model. Section 5 illustrates how the use of the sampling weights can protect against nonignorable designs and misspecified models. In Section 6 we show examples where the use of the weights is either the only possible inference tool or the optimal tool. Section 7 discusses different approaches of incorporating the weights. We conclude the article with a brief summary in Section 8.

An important aspect of the weighting issue not addressed in this article is the construction of the weights. Throughout this article we assume that the weights represent the inverse of the sample inclusion probabilities. Recent articles considering the construction of the sampling weights with rich lists of references to earlier studies are Cox (1987) and Kish (1990).

Much of the discussion of this article relies on the theoretical and empirical results included in the book *The Analysis of Complex Surveys* edited by C. Skinner, D. Holt & T.M.F. Smith (1989). This pioneering book covers a large variety of inference methods applicable to complex survey data. We use the abbreviation 'SHS' when referring to this

book. Another rich source to the present discussion is the book *Panel Surveys* edited by D. Kasprzyk, G. Duncan, G. Kalton & M. P. Singh (1989). We use the abbreviation 'KDKS' for that book.

2 Definition of the Target Population and Parameters

In descriptive inference, the target population consists of all the units in the population from which the sample is drawn. The target parameters are some known functions of the survey variables values like means, proportions, regression coefficients etc. In what follows we refer to such functions as 'descriptive population quantities' (DPQ). All other inferences are 'analytic' but the term usually refers to inference about model parameters like expected values, variances, regression coefficients or cell probabilities. Classical inference methods attempt to infer about these parameters in the form of point estimators, confidence intervals, or posterior distributions. What happens when modeling survey data? Are the parameters of interest different? More specific to our discussion.

> Is there any role for descriptive population quantities in analytic inference from sample surveys?

Positions expressed in the literature on this question range from those who see no role for DPQ's in analytic inference to those maintaining that inference should focus on only the DPQ's. A third position which, in some way, compromises between the other two considers the model parameters as the ultimate target parameters but in the same time focuses also on the DPQ's as a way to secure the robustness of the inference.

The first position represents the approach that models are used in order to draw inference on populations more general than the fixed finite population giving rise to the sample. See for example Hoem (KDKS, p. 540) and Fienberg (KDKS, p. 570). The second position reflects the concern of survey statisticians that with the heterogeneous populations encountered in practice and the complex designs used to select the sample, the fitting of models that closely approximate the behaviour of the population values is not practical. Hence, they recommend replacing the hypothetical model parameters by simple DPQ's which are interpretable and can be used to explain the relationships between the survey variables in a more robust way. See for example Kish & Frankel (1974), Jonrup & Rennermalm (1976) and Shah, Holt & Folsom (1977).

In order to illustrate the difference between the two approaches, consider the fitting of a regression model to data arising from a cluster sample. Population clusters are usually homogeneous groups with large differences between the clusters and so an analyst following the first approach would possibly allow for different regression equations to operate in different clusters in his model. When the number of observations in each cluster is small, he will have to model also the relationships between the regression coefficients operating in different clusters (see for example Pfeffermann & Lavange, SHS, Ch. 12). Alternatively, he may postulate a single regression line but allow for intracluster correlations between residual terms pertaining to the same cluster (Scott & Holt, 1982). Once the model is defined, the analyst will estimate the unknown model parameters using maximum likelihood, Bayesian, or some other optimal strategies. See Pfeffermann & Smith (1985) for a review and discussion of such models.

The analyst following the second approach is likely to define the target quantity as the least squares solution in the case of a census, that is, the DPQ

$$B = (X'X)^{-1}X'\mathbf{Y} = \left(\sum_{i=1}^{N} \mathbf{x}_i\mathbf{x}_i'\right)^{-1} \sum_{i=1}^{N} \mathbf{x}_i y_i \tag{2.1}$$

where $X = [\mathbf{x}_1 \ldots \mathbf{x}_N]'$ and estimate B using for example the estimator $\hat{\boldsymbol{\beta}}_w$ defined in (1.1). As discussed below, the DPQ B has a clear and meaningful interpretation even if the model fails to hold. Notice that the two analysts will differ also in the estimators they use for the variances of the estimated regression coefficients. Thus, the first analyst will attempt to estimate the variance under the model whereas the second analyst will attempt to estimate the variance under the randomization distribution, that is, over all possible samples from the finite population. For further discussion on the alternative approaches to regression analysis of survey data see Brewer & Mellor (1973), Dumouchel & Duncan (1983), Fuller (1975, 1984), Little (1989), Pfeffermann & Smith (1985) and Skinner (SHS, ch. 3).

Before describing the third approach to analytic inference from survey data we discuss the notions of 'corresponding descriptive population quantity' (CDPQ) and 'design consistency' (DC) which form the basis to this approach.

Definition 1. Let $\mathbf{Y}' = (Y_1 \ldots Y_N)$ be generated from a distribution indexed by a vector $\boldsymbol{\theta}$ of unknown parameters. Let $U(\mathbf{Y}, \boldsymbol{\theta}) = 0$ define a set of estimating equations for $\boldsymbol{\theta}$ obtained by an estimation rule $R(\mathbf{Y} \rightarrow \boldsymbol{\theta})$. The solution $T(\mathbf{Y})$ such that $U[\mathbf{Y}, T(\mathbf{Y})] = 0$ is the CDPQ for $\boldsymbol{\theta}$ under the rule $R(\mathbf{Y} \rightarrow \boldsymbol{\theta})$.

The estimating equations can result from the minimization of a particular loss function, (or a Bayesian risk function) or coincide with the likelihood equations. For example, in the linear model $Y_i = \alpha + \beta x_i + \varepsilon_i$ where $E_\xi(\varepsilon_i) = 0$, $E_\xi(\varepsilon_i^2) = \sigma^2 x_i$, $E_\xi(\varepsilon_i \varepsilon_j) = 0$ for $i \neq j$, the CDPQ of (α, β) in the case of maximum likelihood inference is the solution to the likelihood equations $(X'V^{-1}X)\ (\alpha, \beta)' - X'V^{-1}\mathbf{Y} = \mathbf{0}$ where $X = [1_N, \mathbf{x}]$, $\mathbf{x}' = (x_1 \ldots x_N)$ and $V = \text{diag}\,(\mathbf{x})$. The CDPQ of (α, β) under the estimation rule

$$R[(Y \rightarrow (\alpha, \beta)] = \min_{\alpha, \beta} \sum_{i=1}^{N} (Y_i - \alpha - \beta x_i)^2$$

is (A, B) defined as,

$$A = \bar{Y} - B\bar{X}, \quad B = \sum_{i=1}^{N} (x_i - \bar{X}) Y_i / \sum_{i=1}^{N} (x_i - \bar{X})^2 \tag{2.2}$$

where $(\bar{Y}, \bar{X}) = \sum_{i=1}^{N} (y_i, x_i)/N$. The estimating equations are in this case the familiar normal equations $(X'X)\ (\alpha, \beta)' - X'\mathbf{Y} = \mathbf{0}$. The definition of the CDPQ given above is similar in essence to the definitions given in Binder (1983), Godambe & Thompson (1986), Scott & Wild (SHS, ch. 9) and Skinner (SHS, ch. 3).

We now turn to the notion of design consistency. In classical theory of statistics, consistency refers to the limiting behaviour of a sample statistic as the sample size is increased to infinity. Thus, defining the concept of consistency in finite population sampling requires that the population size will also be allowed to increase. This raises the question, however, of a suitable formulation of the manner by which the population and the sample increase such that their structure is preserved. For example, Isaki & Fuller (1982) propose a formulation which consists of constructing nested populations and sampling from each population increasing each time the sample size. The following definition assumes that the manner by which the population and the sample size increase is well defined.

Definition 2. A sample statistic $\mathbf{t}_S(n)$ is said to be design consistent for a DPQ $\mathbf{T}(N)$ if $\text{plim}_{n \rightarrow \infty, N \rightarrow \infty} [\mathbf{t}_S(n) - \mathbf{T}(N)] = 0$ where 'plim' stands for 'limit in probability' under the randomization distribution, n is the sample size and N is the population size.

Using Definitions 1 and 2, the third approach to analytic inference from sample surveys can be described as follows: A model is postulated as in the first approach and inference is directed at the model parameters. However, rather than seeking an optimal estimator under the model, the analyst seeks an estimator from the class of estimators which are DC for the CDPQ.

Why will an analyst restrict to DC estimators? The answer is robustness. If the model holds in the population and the estimation rule he uses yields a CDPQ which is consistent under the model, then as the population size increases the CDPQ will converge to the model parameter. Thus, any DC estimator of the CDPQ will be consistent for the model parameter under the mixed $D\xi$ distribution. On the other hand, if the model fails to hold, the model parameter and its optimal estimator under the model may no longer have a meaningful substantive interpretation. The CDPQ, on the other hand, is a real entity with a clear interpretation that continues to exist irrespective of the validity of the model. For example, the coefficients A and B in (2.2) define the best linear approximation to the Y-values in the finite population with respect to the least squares distance function. Moreover, assuming that the population values can be viewed as a random sample from the joint ξ distribution of (Y, X), the DPQ (A, B) are ξ consistent for the coefficients (α, β) in the linear regression of Y on X.

The consistency of DC estimators of the CDPQ as estimators of the model parameters can be established formally by writing

$$\mathbf{t}_S - \boldsymbol{\theta} = (\mathbf{t}_S - \mathbf{T}) + (\mathbf{T} - \boldsymbol{\theta}) = O_p(n^{-\frac{1}{2}}) + O_p(N^{-\frac{1}{2}}) = O_p(n^{-\frac{1}{2}}) \tag{2.3}$$

where the probability measure $O_p(n^{-\frac{1}{2}})$ applies to the randomization distribution and the probability measure $O_p(N^{-\frac{1}{2}})$ to the ξ distribution. Moreover, we may decompose the variance of $\mathbf{t}_S$ around $\boldsymbol{\theta}$ as

$$\mathrm{Var}_{D\xi}(\mathbf{t}_S) = E_\xi[\mathrm{Var}_D(\mathbf{t}_S \mid \mathbf{Y})] + \mathrm{Var}_\xi[E_D(\mathbf{t}_S \mid \mathbf{Y})] = E_\xi[\mathrm{Var}_D(\mathbf{t}_S \mid \mathbf{Y})] + O(N^{-1}) \tag{2.4}$$

Thus, in the usual case where the population is much larger than the sample, the variance of $\mathbf{t}_S$ under the $D\xi$ distribution is approximately the same as the ξ expectation of the randomization variance and it can be estimated therefore by estimating the randomization variance. Isaki & Fuller (1982) use the term 'anticipated variance' for the $D\xi$ variance in (2.4).

3 Ignorable and Informative Sampling Designs

3.1 Illustration of the Problem

When the sample is selected by simple random sampling, the model holding for the sample data is the same as the model holding in the population before sampling. With the complex sampling designs often used in practice, the two models can be very different however and failure to account for the sample selection process might bias the inference. As already mentioned in Section 2 and illustrated in Section 5, incorporating the sampling weights in the analysis is one way of dealing with the effects of the design.

In order to illustrate the problem, suppose that the population is made up of N units and that with every unit i is associated a vector (y_i, z_i) of measurements where (y_i, z_i) are independent draws from a bivariate normal distribution with mean $\boldsymbol{\mu}$ and $V - C$ matrix Σ. Suppose further that the values $\{(y_i, z_i), i = 1, \ldots, n\}$ are observed for a sample s of n units selected by a probability sampling scheme and that it is desirable to estimate $\mu_y = E_\xi(Y)$. If the sample is selected by simple random sampling with replacement, $\bar{Y}_S = \sum_{i=1}^n y_i/n$ is unbiased for μ_y under the model and it carries other optimal properties.

Clearly, the sample selection scheme can be ignored in this case in the inference process. Consider, however, the case where the sample is selected with probabilities proportional to z_i, with replacement, such that at each draw $k = 1, \ldots, n$, $P(i \in s) = (z_i / \sum_{i=1}^{N} z_i)$. If $\text{Corr}(Y, Z) = \sigma_{yz}/\sigma_y\sigma_z > 0$, $P(Y_i > \mu_y \mid i \in s) > \frac{1}{2}$ so that the distribution of the y-values in the sample is different in this case from the distribution in the population and in particular, $E_\xi(\bar{Y}_S) > \mu_y$. Clearly, ignoring the sampling scheme and estimating μ_y by $\bar{Y}_S$ can be very misleading in this case. Suppose, however, that the values z_i are known for all the population units. An unbiased (maximum likelihood) estimator of μ_y is in this case the regression estimator $\bar{Y}_{\text{reg}} = \bar{Y}_s + b(\bar{Z} - \bar{Z}_S)$ where b is the ordinary least squares (OLS) estimator of the regression coefficient (σ_{yz}/σ_z^2) and $\bar{Z}(\bar{Z}_S)$ is the population (sample) mean of Z. Thus, by utilizing all the population values z_i in the inference, the sampling design becomes ignorable.

In the next section we define the ignorability conditions more formally. In Section 4 we discuss the use of the sampling weights for testing the fulfillment of these conditions.

3.2 Definition and Conditions for the Ignorability of the Sampling Design

The definition of ignorability and the conditions under which the design is ignorable are widely discussed in the literature, see e.g., Little (1982), Rubin (1976), Scott (1977) and Sugden & Smith (1984), so we only sketch the main results.

Let $\mathbf{z}_i' = (z_{i1} \ldots z_{ik})$ represent the values of k design variables associated with unit $i \in U$ and denote $Z = (\mathbf{z}_1 \ldots \mathbf{z}_N)'$. The design variables may include strata indicator variables and quantitative variables measuring cluster and unit characteristics. Let $\mathbf{I}' = (I_1 \ldots I_N)$ be a sample indicator variable such that $I_i = 1$ for $i \in s$ and $I_i = 0$ otherwise. We denote the survey response variables values by $\mathbf{Y}_i$ and let $Y = (Y_S, Y_{\bar{S}})$ where $Y_S = \{\mathbf{Y}_i, i \in s\}$ and $Y_{\bar{S}} = \{\mathbf{Y}_i, i \notin s\}$. In general, the sample selection scheme depends on the design variables and may also depend on the response variables. Thus, $P(s) = P(\mathbf{I} \mid Y, Z; \boldsymbol{\varphi})$ where $\boldsymbol{\varphi}$ is a vector parameter. Let $f(Y \mid Z; \boldsymbol{\theta}_1)$ denote the conditional probability density function (pdf) of Y given Z, indexed by the vector parameter $\boldsymbol{\theta}_1$. The marginal pdf of Z will be denoted by $g(Z; \boldsymbol{\phi})$.

The joint distribution of Y and Z in the population is

$$f(Y, Z; \boldsymbol{\theta}, \boldsymbol{\phi}) = f(Y \mid Z; \boldsymbol{\theta}_1) g(Z; \boldsymbol{\phi}) \tag{3.1}$$

where $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ are assumed to be distinct parameters (Rubin, 1976). In analytic inference, the target parameters are $\boldsymbol{\theta}_1$, or functions of $\boldsymbol{\theta}_1$ and $\boldsymbol{\phi}$ such as the parameters of the marginal distribution of Y.

Suppose that the design variables are known for every unit in the population and that the response variables are observed for only the units in the sample. The joint distribution of Y_S, Z and $\boldsymbol{I}$ is obtained by integrating the joint distribution of Y, Z and $\boldsymbol{I}$ over $Y_{\bar{S}}$,

$$f(Y_S, \boldsymbol{I}, Z; \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\varphi}) = \int f(Y_S, Y_{\bar{S}} \mid Z; \boldsymbol{\theta}_1) g(Z; \boldsymbol{\phi}) P(\boldsymbol{I} \mid Y_S, Y_{\bar{S}}, Z; \boldsymbol{\varphi})\, dY_{\bar{S}} \tag{3.2}$$

Ignoring the sampling design in the inference process implies that the inference is based on the joint distribution of Y_S and Z obtained by integrating (3.1) over $Y_{\bar{S}}$, ignoring $P(\mathbf{I} \mid Y_S, Y_{\bar{S}}, Z; \boldsymbol{\varphi})$,

$$f(Y_S, Z; \boldsymbol{\theta}, \boldsymbol{\phi}) = \int f(Y_S, Y_{\bar{S}} \mid Z; \boldsymbol{\theta}_1) g(Z; \boldsymbol{\phi})\, dY_{\bar{S}} \tag{3.3}$$

Definition 3. The sampling design is ignorable given the set of design variables if inference based on (3.2) (the joint distribution of all the data known to the analyst) is equivalent to inference based on (3.3).

It is important to note that the ignorability of the design refers to the information provided by the selection scheme beyond what is already provided by the design variables. The exact conditions under which the use of (3.2) and (3.3) yields similar inferences are defined and illustrated in the fundamental article by Rubin (1976). The ignorability conditions are clearly satisfied in sampling schemes that depend only on the design variables since in this case $P(\mathbf{I} \mid Y_S, Y_{\bar{S}}, Z; \boldsymbol{\varphi}) = P(\mathbf{I} \mid Z; \boldsymbol{\phi})$. See Little (1982) and Smith (1984) for the implications of this property. Sugden & Smith (1984) explore the conditions under which a sampling scheme which depends only on the design variables in Z is ignorable, given partial information on the design. The authors define the set $d_S = D_S(Z = z)$ to include all the available design information from knowledge of the selection scheme, the sample selection probabilities and any known values or functions of Z. The key condition for ignorability of the sampling scheme given the design information is that $P(s \mid Z = z) = P(s \mid D_S = d_S)$ for all z such that $D_S(z) = d_S$ which implies that $f(Y_S \mid D_S, \mathbf{I}) = f(Y_S \mid D_S)$.

The results quoted so far are formulated in terms of the joint distribution of Y_S and Z but they can be translated to the case where the reference model specifies the conditional distribution of some of the response variables given another set of survey variables. Thus, the sampling design is ignorable for estimating the regression of Y on X if

$$f(Y_S \mid X_S, D_S, \mathbf{I}) = f(Y_S \mid X_S, D_S). \tag{3.4}$$

If, in addition, $f(Y_S \mid X_S, D_S) = f(Y_S \mid X_S)$ the design information entailed in D_S can likewise be ignored and classical regression analysis applies.

The distinction between conditional and unconditional inference highlights another important aspect of the ignorability problem namely, that the ignorability of the sampling design depends not only on the design and the available design information but also on the model and the parameters of interest. Thus, if the regressor variables in a regression model include all the design variables, the sampling design is ignorable for estimating the regression model. If, however, the design variables values are only known for units in the sample, the sampling design is not ignorable for estimating the unconditional mean and variance of the regression dependent variables.

4 The Role of the Sampling Weights in Testing the Ignorability Conditions

4.1 The Effects of Ignoring Informative Designs

As the discussion of the previous section suggests, satisfying the ignorability conditions can be a complicated matter particularly with complex multistage sampling designs which depend on the values of several design variables. The effects of ignoring the sample selection process when fitting models to survey data are studied extensively in SHS (1989) with the clearcut conclusion that failure to account for all the important design variables or incorrectly specifying the conditional distribution of the survey variables given the design information can have severe effects on the inference process. These effects include bias of point estimators and poor performance of test statistics and confidence intervals. The study covers a large number of statistical techniques such as regression analysis, categorical data analysis, logistic regression and principal components analysis. The SHS book contains references to other similar studies. Below we mention briefly two examples.

Example 4.1 Estimation of regression models. DeMets & Halperin (1977) and Nathan & Holt (1980) study the properties of OLS estimators when the selection to the sample is based on the values of a design variable Z which is correlated with the dependent variable

Y. Hausman & Wise (1981) and Jewell (1985), consider situations where the selection probabilities depend on the Y-values directly. These studies indicate that the OLS estimators of the regression coefficients which ignore the design altogether can be severely biased, with no real interpretation to the expected values of the estimators under the mixed, $D\xi$ distribution.

Example 4.2 *Estimation of logistic regression models.* The effect of ignoring the sampling design when estimating logistic regression models is studied empirically by Chambless & Boyle (1985). The authors use data from the lipid research clinics program prevalence study (LRCPPS). The LRCPPS uses a disproportionate stratified random sample with the strata defined by race and lipid zone. The dependent variable in the analysis is again defined by the lipid zone levels. The authors find that when the logistic model is estimated ignoring the design, (that is, assuming simple random sampling), the estimate of the intercept term has a large bias which has a deteriorating effect on the estimated prevalence probabilities. The estimates of the prevalence probabilities remain biased even when extending the model by including the strata indicators as additional covariates.

Scott & Wild (1986, SHS ch. 9) show that the effect of ignoring the design and using maximum likelihood estimation (m.l.e.) is to bias the estimate of the intercept but as long as the logistic regression model holds in the population, the estimates of the slope coefficients remain m.l.e. These results validate the empirical results of Chambless & Boyle (1985).

4.2 Testing the Ignorability Conditions

In practice, it is often the case that not all the relevant design variables are known for all the population units or that there are too many of them to be incorporated in the analysis. Not incorporating all the design variables in the model does not necessarily imply that the inference is biased and as the work of Sugden & Smith (1984) indicates, incorporating only partial design information in the model can be sufficient.

A natural question arising from this discussion is how to test that the design can indeed be ignored, given the available design information. In principle, when all the design features are known, one could verify the fulfillment of the ignorability conditions directly. Frequently, however, the statistician analyzing the data has only limited knowledge about the actual sampling process. It is here where the sampling weights come into play.

Very few studies are reported in the literature on this important aspect of the modeling process. Test statistics proposed in the literature are mostly in the area of regression analysis and they share a common feature. The ignorability of the design is tested by testing the significance of the difference between the best (optimal) estimator of the vector of regression coefficients under a particular working model which assumes that the design is ignorable and the weighted least squares estimator $\hat{\boldsymbol{\beta}}_w$ defined by (1.1). Denoting by $\hat{\boldsymbol{\beta}}$ the optimal estimator of $\boldsymbol{\beta}$ under the model, the hypothesis tested is H_0: $\text{plim}_{n\to\infty, N\to\infty}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_w) = 0$ where 'plim' stands as before for the limit in probability under the mixed $D\xi$ distribution. The test statistic is

$$\boldsymbol{\lambda} = \hat{\mathbf{D}}'[\hat{V}(\hat{\mathbf{D}})]^{-1}\hat{\mathbf{D}} \tag{4.1}$$

where $\hat{\mathbf{D}} = \hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_w$ and $\hat{V}(\hat{\mathbf{D}})$ is an estimator of the $V - C$ matrix of $\hat{\mathbf{D}}$. Dumouchel & Duncan (1983) illustrate that the test statistic in (4.1) can be constructed by augmenting the original design matrix X_S by the columns $W_S\mathbf{x}_j$, $j = 1, \ldots, p$ where $W_S = \text{diag}(w_1, \ldots, w_n)$, and fitting the unweighted regression $\hat{\mathbf{Y}}_S = X_S\hat{\boldsymbol{\beta}} + W_S X_S\hat{\boldsymbol{\gamma}}$. Testing H_0

is equivalent then to testing H_0^*: $\gamma = 0$. Assuming the regression model to hold, the ordinary regression test statistic has an F distribution with p and $(n-2p)$ degrees of freedom under H_0^*. Nordberg (1989) extends the test statistic proposed by Dumouchel & Duncan to the Generalized Linear Model (GLM).

Fuller (1984) considers the case of a cluster sample within strata and estimates the $V-C$ matrix $V(\hat{\mathbf{D}})$ by estimating the corresponding randomization $V-C$ matrix (see equation (2.4)). The resulting test statistic has an approximate F distribution under H_0 with p and $(n-2p-L)$ degrees of freedom where L is the number of strata. The use of the randomization distribution for estimating $V(\hat{\mathbf{D}})$ is more robust since it does not depend on the fulfillment of the regression model assumptions.

The rationale of the test statistic λ in (4.1) is clear. The estimator $\hat{\boldsymbol{\beta}}_w$ is design consistent for the DPQ B defined by (2.1), and so if the regression model holds in the population, $\text{plim}_{n\to\infty, N\to\infty}(\hat{\boldsymbol{\beta}}_w) = \boldsymbol{\beta}$. When the ignorability conditions are satisfied, $\hat{\boldsymbol{\beta}}$ is likewise consistent for $\boldsymbol{\beta}$. If, however, the sampling design is not ignorable, $\hat{\boldsymbol{\beta}}$ is no longer consistent for $\boldsymbol{\beta}$ and the two estimators converge to different limits. Notice that the two estimators will possibly converge to different limits also when the model holding in the population is wrongly specified. Thus, convergence to the same limit is sufficient but not necessary for ignoring the design.

The use of λ for testing the ignorability of the design suggests an important role for the sampling weights in the modeling process. They can be used to construct pivotal statistics for testing the ignorability of the design. In fact, Dumouchel & Duncan (1983) and Nordberg (1989) illustrate how an examination of the significant differences between the components of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_w$ can lead to the identification of important design variables or interactions between some of the design variables which, when added to the model, make the sampling design become ignorable. Chambless & Boyle (1985) make similar comments. See also Fuller (1984), and Pfeffermann & Smith (1985) for further discussion and applications of the test statistic λ.

Can a test statistic of the form (4.1) be constructed for different models and estimators? The answer is positive and relies on the following general result taken from Hausman (1978).

Lemma. Let $\hat{\boldsymbol{\theta}}_0$ and $\hat{\boldsymbol{\theta}}_1$ be two consistent estimators of a vector parameter $\boldsymbol{\theta}$ which are asymptotically normal with $\hat{\boldsymbol{\theta}}_0$ attaining the asymptotic Cramer–Rao bound. Thus, $\sqrt{n}(\hat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}) \overset{A}{\sim} N(0, V_0)$, $\sqrt{n}(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}) \overset{A}{\sim} N(0, V_1)$ where V_0 is the inverse of Fisher's information matrix. Let $\hat{\boldsymbol{\Delta}} = \hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_0$. Then, the limiting distributions of $\sqrt{n}(\hat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta})$ and $\sqrt{n}\hat{\boldsymbol{\Delta}}$ have zero covariance, $C(\hat{\boldsymbol{\theta}}_0, \hat{\boldsymbol{\Delta}}) = 0$ and hence in the limit

$$V(\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_0) = V(\hat{\boldsymbol{\theta}}_1) - V(\hat{\boldsymbol{\theta}}_0) \geqslant 0 \tag{4.2}$$

in the sense of being nonnegative definite.

Hausman (1978) uses the result (4.2) to construct model misspecification tests. We follow similar arguments to construct test statistics for testing the ignorability of the design. Let $\hat{\boldsymbol{\theta}}_0$ define the efficient maximum likelihood estimator of $\boldsymbol{\theta}$ under a given model assuming that the design is ignorable. Let $\tilde{\boldsymbol{\theta}}$ be the CDPQ of $\boldsymbol{\theta}$ defined as the solution of the census likelihood equations and denote by $\hat{\boldsymbol{\theta}}_w$ the design consistent estimator of $\tilde{\boldsymbol{\theta}}$. Then, by Lemma 1, if the sampling design is ignorable and $\hat{\boldsymbol{\theta}}_w$ is asymptotically normal,

$$(\hat{\boldsymbol{\theta}}_w - \hat{\boldsymbol{\theta}}_0)'[\hat{V}(\hat{\boldsymbol{\theta}}_w) - \hat{V}(\hat{\boldsymbol{\theta}}_0)]^{-1}(\hat{\boldsymbol{\theta}}_w - \hat{\boldsymbol{\theta}}_0) \overset{A}{\sim} X^2_{(p)} \tag{4.3}$$

where $p = dim\ (\boldsymbol{\theta})$. The $V-C$ matrices $\hat{V}(\hat{\boldsymbol{\theta}}_w)$ and $\hat{V}(\hat{\boldsymbol{\theta}}_0)$ can be obtained by estimating the corresponding randomization $V-C$ matrices.

The test statistic in (4.3) is quite general but notice that besides the usual risk of a type II error, retaining the null hypothesis is not sufficient to ensure that the sampling design can be ignored for other facets of the inference process like, for example, probabilistic statements. Simple residual plots of model residuals against known design variables and/or against the sample selection probabilities at the various stages of the sampling process can be useful for further assessments of the ignorability of the design. Pfeffermann & Smith (1985) use partial residual plots of the regression residuals against a size variable used for determining the sample inclusion probabilities. Pfeffermann & Nathan (1985) propose simple test statistics based on cross validation techniques. The development of ingenious residual plots and test statistics for assessing the ignorability of the sampling design is an important area for future research.

5 The Use of the Sampling Weights to Protect Against Informative Designs and Misspecified Models

5.1 Difficulties in Modeling Survey Data

As implied by the discussion of Section 4.2, testing the ignorability of the sampling design is often a complicated matter. Even when test statistics can be constructed, their use may not be conclusive or that they may indicate that the ignorability conditions are not fully satisfied. There seems to be a consensus, even among theorists who otherwise oppose the use of sampling weights in a modeling context that the sampling weights can play a vital role in situations where the ignorability of the sampling design is at stake. The following is a quotation from Fienberg (KDKS, p. 571)—'The one exception in which the use of weights may be appropriate is outcome-based sampling where the sampling plan may be informative for the model of interest. . .' See also Hoem (KDKS, p. 541). Notice also that even if all the relevant design variables are known, incorporating them in the model may become a major undertaking. As argued by Alexander (1987)—'no model will include all the relevant variables and few analysts will wish to include in the model all the geographic and operational variables which determine sampling rates. The theoretical and empirical tasks of deriving, fitting and validating such models seem formidable for many complex national demographic surveys'.

How can the sampling weights be used to protect against informative designs and the possibility of model misspecifications? The idea has already been established in Section 2. Estimators of model parameters are modified so that they are design consistent for the CDPQ in the finite population from which the sample has been drawn. Probability statements are based on the randomization distribution of sample statistics or on the mixed $D\xi$ distribution. Notice from equation (2.4) that for sufficiently large populations, the variance of sample statistics with respect to the $D\xi$ distribution can be approximated by the corresponding randomization variance. With the large samples often used in practice, the distribution of the survey estimators is approximately normal. See Binder (1983), Chambless & Boyle (1985) and Fuller (1975, 1984) for central limit theorems applicable to complex sample surveys.

Example 5.1 *Estimation of regression models.* In Sections 2 and 4.1 we considered the case of regression analysis. Focusing on the CDPQ, B, defined by (2.1) and seeking a DC estimator for B gives rise to estimators of the form $\hat{\boldsymbol{\beta}}_w$ defined by (1.1). The estimator $\hat{\boldsymbol{\beta}}_w$ is obtained as the solution to the linear equations

$$U(\mathbf{Y}_S, X_S, \boldsymbol{\beta}) = \sum_{i \in S} w_i \mathbf{x}_i (\mathbf{x}_i' \boldsymbol{\beta} - y_i) = \mathbf{0}. \tag{5.1}$$

For fixed vectors $\boldsymbol{\beta}$, $E_D[U(\mathbf{Y}_S, X_S, \boldsymbol{\beta})] = \sum_{i=1}^{N} \mathbf{x}_i(\mathbf{x}_i'\boldsymbol{\beta} - y_i) = U(\mathbf{Y}, X, \boldsymbol{\beta})$ where $U(\mathbf{Y}, X, \boldsymbol{\beta})$ are the normal equations in the case of a census, yielding the CDPQ B. Binder (1983) proposes a general method for estimating the randomization variances of estimators obtained by solving equations of the form (5.1). A notable feature of his method is that the estimators need not have explicit expressions, like for example in the case of GLM.

Example 5.2 *Fitting logistic models.* The logistic model with a dichotomous response variable Y and explanatory variables $\mathbf{X}$ postulates that $P(\mathbf{x}) = P(Y = 1 \mid \mathbf{X} = \mathbf{x}) = \exp(\mathbf{x}'\boldsymbol{\beta})/[1 + \exp(\mathbf{x}'\boldsymbol{\beta})]$. Assuming that the Y values in the population are independent, the m.l.e. of $\boldsymbol{\beta}$ in the case of a simple random sample is obtained as the solution to the likelihood equations $\sum_{i\in S}[y_i - P(\mathbf{x}_i)]\mathbf{x}_i = \mathbf{0}$. As shown by Chambless & Boyle (1985) and Scott & Wild (SHS, ch. 9), the m.l.e. becomes biased and inconsistent when the sampling design is informative or when the logistic model does not hold in the population. Hence, the authors propose to estimate $\boldsymbol{\beta}$ in such cases by solving the equations $\sum_{i\in S} w_i[y_i - P(\mathbf{x}_i)]\mathbf{x}_i = \mathbf{0}$. Examples for possible violation of the logistic model in the population are: exclusion of important covariate variables from $\mathbf{x}$, nonlinear effects of some of the covariate variables and intracluster correlations.

Even if the logistic model assumptions are violated, the use of the model can still be justified since it suggests 'useful population quantities as targets for inference' (Binder, 1983, Scott & Wild, SHS ch. 9). Indeed, Scott & Wild show that when the logistic model assumptions are violated, the coefficients $\boldsymbol{\beta}$ solving the census likelihood equations may be interpreted as 'defining the logistic model $P(\mathbf{x}, \boldsymbol{\beta})$ which 'best approximates' the true model.

Example 5.3 *Fitting models to panel survey data.* Folsom, LaVange & Williams (KDKS, p. 108–138) discuss several methods of estimation and inference based on the randomization distribution of panel survey data. Models considered include the general linear multivariate model for repeated measurements with missing observations on the dependent variable, polynomial growth curve models and models used for survival analysis like the discrete proportional hazards model. In all the analyses presented, the classical model based estimators are modified by weighting the observations with the sampling weights. A weighted version of the EM algorithm for maximum likelihood estimation is also presented. (Pfeffermann, 1988, employs a similar idea in a different context). As argued by the authors, 'the effects of sampling designs that involve unequal probabilities of selection and clustering must be taken into account when applying classical longitudinal data analysis'. Estimation of proportional hazards models from survey data is considered also by Chambless & Boyle (1985) and Binder (1992).

Example 5.4 *Estimation of distribution functions.* Chambers & Dunstan (1986) consider estimation of the distribution function of a response variable, Y, from a complex survey, when the values of an auxiliary variable, X, are known for every element in the population. The authors assume the model $Y_i = \boldsymbol{\beta}x_i + v(x_i)\varepsilon_i$ $(i = 1, \ldots, N)$ where $v(x_i) = x_i^{1/2}$ and the ε_i are iid with zero mean and derive a model based estimator for $F(t) = P(Y \leqslant t)$. Empirical results show the much better performance in small samples of the model dependent estimator as compared to the customary, design based estimator

$$\hat{F}_d(t) = \sum_{i\in S} [\Delta(t - y_i)w_i] / \sum_{i\in S} w_i, \tag{5.2}$$

where $\Delta(\,.\,)$ is the step function such that $\Delta(a) = 1$ when $a \geqslant 0$ and $\Delta(a) = 0$ otherwise. The problem with the authors' estimator is that it performs poorly when, for example, the assumption on the variance of the error terms is violated. In order to deal with this

problem, Rao, Kovar & Mantel (1990) propose a weighted difference type estimator which is asymptotically unbiased for $F(t)$ under the model and asymptotically design unbiased for the distribution $F_N(t) = \sum_{i=1}^{N} \Delta(t - y_i)/N$ in the finite population. The authors derive an estimator for the randomization variance of the proposed estimator.

5.2 Drawbacks of Randomization Based Inference

Focusing on CDPQ as the target of inference and restricting to DC estimators is not without a price. As illustrated in many studies, (see references below), if the model postulated for the sample data is correct, the use of weighted estimators can result in substantial loss of efficiency compared to the use of optimal, model dependent estimators. In general, the loss in efficiency is larger, the smaller is the sample size and the larger is the variation of the sampling weights.

It is important to emphasize also that although weighted statistics are asymptotically unbiased when averaging over all possible samples, they may exhibit serious biases under the model (ξ) distribution, with the selected sample held fixed. Smith (1984) makes the observation that the robustness of randomization based inference is only achieved 'by converting the conditional bias into a component of variance'. It is clear however that as the sample size increases, the probability of selecting 'extreme samples' which produce large conditional biases is decreased.

Articles illustrating conditional and unconditional properties of randomization based estimators under given models include: Hausman & Wise (1981), Holt, Smith & Winter (1980), Jewell (1985), Nathan & Holt (1980) and Pfeffermann & Holmes (1985) in regression analysis; Smith & Holmes (SHS, ch. 8) in more general aspects of multivariate analysis; Scott & Wild (SHS, ch. 9) in logistic regression, Nordberg (1989) in GLM and Rao, Kovar & Mantel (1990) in distribution function estimation.

Another important limitation of randomization based inference is that by focusing on CDPQ as the target quantities, the inference is restricted to populations which have a similar structure to that of the population under study. Kalton (KDKS, p. 580) discusses the following simplified example. Suppose that it is required to estimate the transition probabilities in a simple Markov chain model. If the model holds, every individual has the same transition probabilities and those probabilities can be estimated under the model by the simple unweighted sample proportions. Suppose, however, that the model is false with older persons having different transition probabilities from younger persons. Let the sample be selected by a stratified design with the strata defined by age. If older persons have higher sample inclusion probabilities than younger persons, the unweighted sample proportions depend on the actual sampling fractions within the strata and they are generally meaningless. The weighted estimates on the other hand estimate the corresponding population proportions so that they are meaningful estimates. The weighted estimators fail, however, to provide meaningful estimates for populations with a different age composition. It is clear also that the weighted estimators are biased in estimating the separate probabilities holding in the various strata. Thus, the protection offered by the use of the weights applies only in a restricted sense. This limitation is quite general and applies to other inference models.

Finally, a limitation of randomization based inference raised in the literature is that there is no clear principle in the choice of DC estimators. We come back to this issue in Section 7.

6 Weighting as the Only or Best Alternative

6.1 Weighting as the Only Alternative

When the data available to the analyst are already in the form of weighted sample estimates, the use of the weights in the inference process is inevitable. This happens for example in the case of a 'secondary analysis' from summary tables for which the researcher may not have access to the detailed data. The classical case of a secondary analysis of this kind is contingency tables analysis from cross classified tables of estimated counts. Hidiroglou & Rao (1987) describe the production of such estimates from data obtained from the Canada Health Survey (1978–1979). The sampling design used for this survey is a multistage stratified cluster sampling design which is typical to many complex surveys. For a given cell i, the count estimate $\hat{N}_i$ has the general form

$$\hat{N}_i = \sum_a (N_a/\hat{N}_a)\left[\sum_h \sum_t \sum_k w_{htk} Y_{ia(htk)}\right] = \sum_a (N_a/\hat{N}_a)\hat{N}_{ia} \tag{6.1}$$

where $Y_{ia(htk)}$ is one if sample unit k from PSU t of stratum h belongs to the ith cell and the ath age–sex group and is zero otherwise. The ratios $(N_a/\hat{N}_a)$ are poststratification adjustment factors which use the census age–sex counts N_a to decrease the variance of the estimators.

With $\hat{N}_i$ as the input data, the classical tests for homogeneity and goodness of fit of loglinear models, based on multinomial or independent Poisson sampling are no longer valid. In fact, Rao & Scott (1981, 1984, 1987) show that the classical χ^2 statistics are distributed asymptotically as weighted sums $\sum \delta_i X_i$ of independent $\chi^2_{(1)}$ variables $\{X_i\}$, where the weights δ_i are eigenvalues of a 'generalized design effects' matrix. This matrix again depends on the sampling weights through the estimated $V - C$ matrix of the cell count estimators. The authors propose first and second order corrections to the χ^2 statistics which account for the effect of the design. Other methods for analysing contingency tables obtained from complex survey data utilize the large samples Wald statistic-Grizzle et al. (1969), Koch et al. (1975), Nathan (1975) or the Jackknife Chi-squared statistic—Fay (1985). For additional references and discussions see the review articles of Binder et al. (1987), Nathan (1988) and Rao & Thomas (SHS, ch. 4).

The important implication from the discussion above is that while the use of the sampling weights is inevitable when the input data already consist of weighted statistics, classical methods of data analysis which assume simple random sampling may no longer be valid. An interesting question arising in this respect is whether in the case where individual observations are available, alternative, 'weights-free' procedures for contingency tables analysis are plausible. Following the discussion in Section 3, the answer to this question depends on whether the design variables defining the sampling scheme can be identified and incorporated in the model.

For the case of a stratified sample with simple random sampling within strata, Nathan (1975) constructs tests of overall independence between qualitative variables I and J by creating a separate layer for each stratum h. The input data consist of the sample counts $\{n_{ijh}\}$. Holt & Ewings (SHS, ch. 13) fit logistic models to data obtained from a two stage cluster sample. The input data consist in this case of the estimated logits $\hat{\lambda}_{jc} = \log[\hat{P}(Y = 1 \mid j)/\hat{P}(Y = 0 \mid j)]$ in individual clusters $c = 1, \ldots, M$ where j defines the different combinations of the explanatory variables in the contingency table. The model permits random cluster effects which vary across the combinations of the explanatory variables within the clusters. Thus, 'weights-free' procedures for contingency tables analysis are possible, provided that the ignorability conditions can be satisfied but it is not clear that

these procedures are statistically more efficient than procedures which use the aggregated weighted estimates.

6.2 Weighting as the Best Alternative

In Section 6.1 we consider situations where the use of the weights is practically imposed, either because of data availability or because alternative procedures of modeling the data are not known. There are situations, however, where the use of the sampling weights is the optimal strategy under a given model and inference rule. Below we list three examples. See the corresponding references for the models used in each case.

(a) Bayesian prediction of finite population means from a disproportionate stratified sample—Binder (1982), Little (1989).
(b) Maximum likelihood estimation of Bernoulli probabilities from poststratified samples—Alexander (1987).
(c) Maximum likelihood estimation of the transition probabilities of a Markov chain in retrospective sampling—Hoem (KDKS, p. 539).

7 Different Approaches for Incorporating the Weights

7.1 Preface

In this section we survey more systematically the approaches proposed in the literature for incorporating the sampling weights $\{w_i = (1/\pi_i),\ i \in s\}$ in the inference process. We restrict the discussion to point estimation which is where the various approaches differ mostly. The following points should be borne in mind when comparing these approaches:

–Different approaches may lead to the same estimators
–The same approach may produce different estimators
–Not all the approaches are aimed at producing DC estimators
–In some of the approaches the weights come into picture indirectly as a result of the use of particular models.

7.2 Overview

7.2.1 Modifications of Model Dependent Estimators

By this approach, estimators with explicit expressions are modified so as to make them DC for the CDPQ. Consider for example the estimation of the slope coefficient in simple regression. The OLS estimator can be expressed as

$$b_{\mathrm{OLS}} = \left[\frac{1}{n}\sum_{i\in S} y_i x_i - \left(\frac{1}{n}\sum_{i\in S} y_i\right)\left(\frac{1}{n}\sum_{i\in S} x_i\right)\right] \Big/ \left[\frac{1}{n}\sum_{i\in S} x_i^2 - \left(\frac{1}{n}\sum_{i\in S} x_i\right)^2\right].$$

A modified, DC estimator of the census slope coefficient, B, defined by (2.2) is obtained by replacing each of the simple means in the expression of b_{OLS} by the Horvitz–Thompson (H–T) estimator of the corresponding population mean so that for example $\frac{1}{n}\sum_{i\in S} y_i x_i$ is replaced by $\frac{1}{N}\sum_{i\in S} y_i x_i/\pi_i$. The resulting estimator is not in general the same as the DC estimator obtained from (1.1) for the case of simple regression. A sufficient condition for the coincidence of the two estimators is that $\sum_{i\in S}(1/\pi_i) = N$ which for arbitrary designs holds only in expectation. Examples of the use of this approach for incorporating the weights in the case of regression analysis can be found in Dumouchel & Duncan (1983), Fuller (1975, 1984), Kish & Frankel (1974), Nathan & Holt (1980),

Pfeffermann & Lavange (SHS, ch. 12), and Shah, Holt & Folson (1977). See also Koch et al. (1975) for application of the approach for general linear models, and Smith & Holmes (SHS, ch. 8) for its possible use in multivariate analysis.

The fact that more than one DC estimator is available for the slope coefficient highlights a criticism often raised in the literature against randomization based inference namely, that there is no clear principle in the choice of estimators (Little, 1989). In fact, for the case where selection to the sample is based on the known population values of a set Z of design variables, Pfeffermann & Holmes (1985), following a suggestion by W. Fuller consider another DC estimator for the slope coefficient. This estimator 'corrects' the probability weighted estimator (1.1) by utilizing the known differences between sample and population moments of the design variables. (A similar idea is employed in the estimators proposed by Rao, Kovar & Mantel, 1990, for the percentiles of distribution functions.) Another estimator for this case is proposed by Nathan & Holt (1980). The authors modify the m.l.e. of the slope coefficient as obtained under the assumption that (X, Y, Z) is multivariate normal by replacing sample means, variances and covariances by the corresponding weighted statistics.

7.2.2 Restriction to Models that Yield DC Estimators

This approach is due to Little (1983, 1989) but its application has been restricted so far to simple stratified sampling. It is based on the postulate that if DC estimators for the CDPQ are required, then inference should be based on models that yield such DC estimators. Thus, for estimating the mean of a normal population, Little proposes to consider the fixed stratum-effects model, $Y_{hi} \mid \mu_h, \sigma_h \sim_{\text{ind}} N(\mu_h, \sigma_h^2)$; $P(\mu_h, \ln \sigma_h) \propto$ *constant,* which, for large samples yields the H–T estimator $M_S = \sum_{h=1}^{L} (N_h/N) \sum_{i \in S} (y_{hi}/n_h)$ as the posterior mean where L denotes the number of strata, $\{N_h\}$ are the strata sizes and $\{n_h\}$ the corresponding strata sample sizes. Alternatively, the author considers a random stratum-effects model by which $\mu_h \mid (\mu, \delta^2) \sim_{\text{ind}} N(\mu, \delta^2)$. Assuming in addition that $P(\mu, \ln \sigma_h, \ln \delta) \propto$ *constant,* the posterior mean of $M = \sum_{h=1}^{L} \sum_{i=1}^{N_h} Y_{hi}/N$ is again DC under the randomization distribution. These results extend to the estimation of regression models yielding approximately the estimator $\hat{\boldsymbol{\beta}}_w$ of (1.1) in the fixed effects case and another DC estimator of the census vector of coefficients under a model which postulates random slopes and intercepts.

A notable characteristic of Little's approach is that the sampling design is featured in the models. As noted by the author, however, what is still lacking in this approach is guidance about the choice of such models for arbitrary designs. Indeed, simulation results presented by the author indicate that in small samples, the model dependent inference can be very sensitive to the model assumptions. Featuring the sampling design in models fitted to complex survey data (but without necessarily requiring design consistency) has been advocated and illustrated in other studies. For example, Alexander (1987), Chambless & Boyle (1985), Dumouchel & Duńcan (1983), Nathan (1975) and Scott & Wild (SHS, ch. 9), account for fixed stratum effects. Holt & Ewings (SHS, ch. 13), and Pfeffermann & Lavange (SHS, ch. 12) account for random cluster effects.

7.2.3 The Pseudo Likelihood Approach

The prominent feature of this approach is that it utilizes the sampling weights to estimate the likelihood equations that would have been obtained in the case of a census. To fix the idea, suppose that the population values $\{\mathbf{Y}_t, t = 1 \ldots N\}$ are independent with pdf's $f_t(\mathbf{y}_t; \boldsymbol{\theta})$ which depend on an unknown parameter vector $\boldsymbol{\theta}$. In the case of a

census, the m.l.e. of $\boldsymbol{\theta}$ maximizes the log likelihood $l(\boldsymbol{\theta}) = \sum_{t=1}^{N} \log f_t(\mathbf{y}_t; \boldsymbol{\theta})$ and in regular cases, it solves the likelihood equations

$$\mathbf{U}(\boldsymbol{\theta}) = dl(\boldsymbol{\theta})/d\boldsymbol{\theta} = \sum_{t=1}^{N} \mathbf{u}_t(\mathbf{y}_t, \boldsymbol{\theta}) = 0 \tag{7.1}$$

where 'd' defines the derivation operator and $\mathbf{u}_t(\mathbf{y}_t, \boldsymbol{\theta}) = d \log f_t(\mathbf{y}_t; \boldsymbol{\theta})/d\boldsymbol{\theta}$. The pseudo m.l.e. (p.m.l.e.) of $\boldsymbol{\theta}$ is defined as the solution of $\hat{\mathbf{U}}(\boldsymbol{\theta}) = 0$ where $\hat{\mathbf{U}}(\boldsymbol{\theta})$ is DC for $\mathbf{U}(\boldsymbol{\theta})$. The common estimator of $\mathbf{U}(\boldsymbol{\theta})$ in the literature is the H–T estimator so that the p.m.l.e. of $\boldsymbol{\theta}$ is the solution of

$$\sum_{i \in S} (w_i)\mathbf{u}_i(\mathbf{y}_i, \boldsymbol{\theta}) = 0 \tag{7.2}$$

The set of equations defined by (5.1) is an example for the use of this approach when estimating the regression coefficients of the linear regression model with iid normal residuals. The resulting estimator is again $\hat{\boldsymbol{\beta}}_w$ of (1.1). See Binder (1983) and Jonrup & Rennermalm (1976).

For the logistic regression model the p.m.l.e. of the vector of coefficients is obtained as the solution of the equations $\sum_{i \in S} w_i[y_i - p(\mathbf{x}_i)]\mathbf{x}_i = 0$ where $p(\mathbf{x}) = p(Y = 1 \mid X = \mathbf{x}) = \exp(\mathbf{x}'\boldsymbol{\beta})/[1 + \exp(\mathbf{x}'\boldsymbol{\beta})]$, c.f. Chambless & Boyle (1985) and Scott & Wild (SHS, ch. 9). See also Rao & Thomas (SHS, ch. 4) for the use of p.m.l.e. when estimating log linear models. Other notable references are Binder (1983) and Nordberg (1989) for the estimation of the parameters of the GLM and Binder (1992), Chambless & Boyle (1985) and Folsom et al. (KDKS, p. 108–138) for the fitting of proportional hazards models.

A general discussion of the pseudo likelihood approach is entailed in Skinner (SHS, ch. 3). Binder (1983), and Chambless & Boyle (1985), follow Fuller (1975, 1984) in establishing conditions under which the distribution of $n^{\frac{1}{2}}(\hat{\boldsymbol{\theta}}_{\text{pmle}} - \boldsymbol{\theta})$ converges to the normal law with zero mean. Binder (1983) develops a general method for estimating the randomization $V - C$ matrix of $\hat{\boldsymbol{\theta}}_{\text{pmle}}$.

7.2.4 Estimating Functions

This approach focuses on the census estimating function (EF) as the target population quantity. The key reference is Godambe & Thompson (1986). Let $\mathbf{y}' = (y_1 \ldots y_N)$ be the population values of a univariate random variable Y generated from a distribution $\xi(\theta)$ indexed by the scalar parameter θ. An unbiased EF for θ based on $\mathbf{y}$ is the function $g(\mathbf{y}, \theta)$ such that

$$E_\xi[g(\mathbf{y}, \theta)] = 0 \tag{7.3}$$

The function $g^*(\mathbf{y}, \theta)$ is optimal among the unbiased EF if it minimizes the quantity

$$E_\xi(g^2)/[E_\xi(dg/d\theta)_{\theta=\theta(\xi)}]^2 \tag{7.4}$$

If $g^*(\mathbf{y}, \theta)$ is optimal then the equation $g^*(\mathbf{y}, \theta) = 0$ is called the optimal estimating equation and the solution θ_N is the optimal estimate. Notice that θ_N is the CDPQ for θ under the estimation rule $g^*(\mathbf{y}, \theta) = 0$ as defined in Definition 1 of Section 2.

So far we considered the case of a census. Let $h(\mathbf{y}_S, \theta)$ be an EF applied to the sample data. Godambe & Thompson (1986) restrict to the case where $g^*(\mathbf{y}, \theta)$ is linear, that is, $g^*(\mathbf{y}, \theta) = \sum_{t=1}^{N} \phi_t(y_t, \theta)a_t(\theta)$ where $E_\xi[\phi_t(y_t, \theta)] = 0$, and require that $h(\mathbf{y}_S, \theta)$ is design unbiased for $g^*(\mathbf{y}, \theta)$, i.e. $E_D[h(\mathbf{y}_S, \theta)] = g^*(\mathbf{y}, \theta)$ for every population vector $\mathbf{y}$ and each θ. The optimal choice for h is defined as the function h^* minimizing

$$E_\xi E_D\{h^2[\mathbf{y}_S, \theta(\xi)]\}/[E_\xi E_D(dh/d\theta)_{\theta=\theta(\xi)}]^2 \tag{7.5}$$

(compare with 7.4). The function h^* is shown to be the H–T estimator

$$h^*(\mathbf{y}_S, \theta) = \sum_{i \in S} [\phi_i(y_i, \theta)/\pi_i]. \tag{7.6}$$

The restriction to functions h that are design unbiased for g^* implies that minimization of (7.5) is equivalent to the minimization of $A(S) = E_\xi E_D\{h[\mathbf{y}_S, \theta(\xi)] - g^*(\mathbf{y}, \theta)\}^2$. Thus, the function h^* defined by (7.6) has the smallest mean squared error as an estimator of g^* among all the functions h which are design unbiased for g^*. Here the mean squared error is taken over the mixed $D\xi$ distribution. The sample estimate $\hat{\theta}_n$ of θ is obtained by solving the equation $h^*(\mathbf{y}_S, \theta) = 0$.

The theory of EF extends to the case of multivariate data and a vector parameter with minor modifications. The following observations are important when comparing this approach to the pseudo likelihood approach:

—Under certain regularity conditions on the class of estimating functions and the class of density functions, the likelihood equations are the optimal estimating equations;

—The use of estimating functions does not require a full specification of the joint distribution of the population values as required for the pseudo likelihood approach;

—Under some regularity conditions the function $h^*(\mathbf{y}_S, \theta)$ of (7.6) satisfies the inequality

$$E_\xi E_D[h^* - U(\theta)]^2 \leq E_\xi E_D[h - U(\theta)]^2 \tag{7.7}$$

where $U(\theta)$ is the log likelihood defined by (7.1) and $h(\mathbf{y}_S, \theta)$ is any other linear design unbiased EF. The relationship (7.7) provides a theoretical justification for the use of the pseudo likelihood approach in situations where it yields the optimal estimating equation.

7.2.5 The Use of the Sampling Weights as Surrogates of the Design Variables

In the approaches considered so far, the sampling weights are not included as part of the model and they are only brought in at the inference stage. In this and the next section we review studies which utilize the weights as part of the model.

Rubin (1985) proposes to use the vector $\boldsymbol{\pi}'$ of the first order inclusion probabilities as surrogates of the set of design variables in situations where the information available on the design variables is not sufficient to secure the ignorability conditions (see Section 3.2), or when modelling the conditional distribution of the response variables, given the design variables, is too complicated. Let Z denote as before the matrix of the design variables values. Rubin defines the column vector $\mathbf{a}' = (a_1 \ldots a_N) = a(Z)$ to be an adequate summary of Z if $P(\mathbf{I} \mid Z) = P(\mathbf{I} \mid \mathbf{a})$ where $\mathbf{I}$ is the sample indicator variable defined in Section 3.2, and shows that the vector $\boldsymbol{\pi}$ of inclusion probabilities ('propensity scores' in the author's terminology) is the 'coarsest' possible adequate summary of Z. Clearly, if $\boldsymbol{\pi}$ is an adequate summary of Z, $P(Y_{\bar{S}} \mid Y_S, \boldsymbol{\pi}, \mathbf{I}) = P(Y_{\bar{S}} \mid Y_S, \boldsymbol{\pi})$ so that given $\boldsymbol{\pi}$ the sampling design is ignorable (see Section 3.2) and specification of the conditional distribution of Y given $\boldsymbol{\pi}$ is all that is needed for a valid inference.

Rubin's approach offers a principled method for incorporating the weights but it requires the knowledge of the inclusion probabilities for all the population units and not just for the sample units. More crucial, and as illustrated by Rubin (1985) and Sugden & Smith (1984), the vector $\boldsymbol{\pi}$ can be too coarse and hence not be an adequate summary. See Smith (1988) for possible expansions of the vector $\boldsymbol{\pi}$ in such cases.

7.2.6 *MLE Derived from Weighted Distributions*

The other approach considered in the literature for incorporating the weights as part of the model is the use of weighted distributions (Patil & Rao, 1978; Rao, 1985). In a way, this approach is the converse of Rubin's approach since it focuses on the probabilities $P(\mathbf{y}_i; \boldsymbol{\alpha}) = P(i \in S \mid \mathbf{y}_i)$ where $\boldsymbol{\alpha}$ is a vector parameter rather than on the densities $f(\mathbf{y}_i \mid \pi_i)$.

Weighted distributions are obtained by modifying the distribution of Y in the population to account for the probability of actually observing Y given that Y has been realized. Thus,

$$f^w(\mathbf{y}_i; \boldsymbol{\lambda}) = [P(\mathbf{y}_i; \boldsymbol{\alpha}) f(\mathbf{y}_i; \boldsymbol{\theta}) / P(i \in s)] = f(\mathbf{y}_i \mid i \in s) \tag{7.8}$$

where

$$P(i \in s) = \int P(\mathbf{y}_i; \boldsymbol{\alpha}) f(\mathbf{y}_i; \boldsymbol{\theta})\, dy_i.$$

Assuming that the observations $\mathbf{y}_i$ can be considered as independent, the likelihood for $\boldsymbol{\theta}$ is

$$L(\boldsymbol{\theta}; Y_S) = \text{Const} \times \prod_{i=1}^{n} f(\mathbf{y}_i; \boldsymbol{\theta}) \Big/ \left[\int P(\mathbf{y}; \boldsymbol{\alpha}) f(\mathbf{y}; \boldsymbol{\theta})\, d\mathbf{y} \right]^n. \tag{7.9}$$

The likelihood (7.9) is seen to depend on the conditional selection probabilities $P(\mathbf{y}; \boldsymbol{\alpha})$ that enter into the denominator. Thus, the use of this likelihood requires in addition to the definition of the pdf $f(\mathbf{y}; \boldsymbol{\theta})$ a specification of the relationship $P(\mathbf{y}; \boldsymbol{\alpha})$ between the sample selection probabilities and the variables observed in the sample. This can be accomplished by modeling the empirical relationship in the sample between the sample inclusion probabilities and the observed measurements. Having identified a suitable model, the vector parameter $\boldsymbol{\alpha}$ can be included as part of the unknown parameters over which the likelihood is maximized or it can be estimated externally in which case it may be fixed at its estimated value when maximizing the likelihood (7.9).

Krieger & Pfeffermann (1992) illustrate the use of this approach for estimating the parameters of a bivariate normal distribution. The authors consider two different sampling designs—PPS sampling and disproportionate stratified sampling, and distinguish between cases where the sampling designs are ignorable and where they are informative. Simulation results illustrate the good performance of estimators obtained by this approach. An earlier use of these ideas for the estimation of regression coefficients is reported in Hausman & Wise (1981). See also Smith (1988) for a formulation applicable to PPS sampling under which the method of moments estimators, derived from the joint weighted pdf $f^w(\mathbf{y}, \boldsymbol{\pi})$ reduce to the corresponding H–T estimators.

8 Summary

In Section 7 we survey six approaches for incorporating the sampling weights in the inference process. In the first four approaches the weights are not part of the model and they are used to produce DC estimators for CDPQ of the model parameters. In the other two approaches the weights are incorporated as part of the model but the resulting estimators are not necessarily DC for the CDPQ.

The first approach described in 7.2.1 is restricted to estimators with known explicit expressions. Little's approach offers a model based theory for incorporating the weights but a more general application of this approach requires the development of guidelines for the choice of such models. The two approaches entitled 'pseudo likelihood' and

'estimating functions' are similar in the sense that inference is directed at the optimal estimating equations that would be obtained in the case of a census. These optimal equations are frequently the same under the two approaches. A critical drawback in the reported applications of these approaches (but not in principle in the philosophy behind them) is the restriction to simple weighting of the individual functions $\mathbf{u}_i(\mathbf{y}_i, \boldsymbol{\theta})$ or $\boldsymbol{\phi}_i(\mathbf{y}_i, \boldsymbol{\theta})$ to achieve exact design unbiasedness (see equations (7.2) and (7.6)). In other words, valuable information on the design variables or some other concomitant variables, possibly known to the analyst, is not exploited in the estimation of the population optimal estimating equations. The estimators proposed by Fuller (see Pfeffermann & Holmes, 1985) for regression coefficients and by Rao, Kovar & Mantel (1990) for percentiles of distribution functions are examples for the use of design variables or other concomitant information to obtain more efficient DC estimators than the simple weighted estimators. We survey these studies in previous sections. Similar procedures can be employed for estimating the population likelihood or estimating functions.

The use of 'weighted distributions' described in 7.2.6 provides a principled method for incorporating the weights in the inference process. The application of this approach requires however to model the relationship between the sample selection probabilities and the observed data. The key question to the use of this approach is therefore whether this relationship can be identified and estimated from the sample data. It would seem that this question can only be answered by analysing actual data obtained from commonly used sample surveys. Research in this direction would be a valuable contribution.

Acknowledgements

This research was supported by grant USPHS MH 37188 from the National Institute of Mental Health. Some of the work was done while I was at Statistics Canada under its Research Fellowship Program. I am grateful to Rod Little for motivating this research and for many stimulating discussions. I wish to thank also David Binder and Gad Nathan for a thorough reading of a first draft and for many valuable comments. Special thanks are due to Linda Lafontaine for her dedicated technical assistance.

References

Alexander, C.H. (1987). A Model Based Justification for Survey Weights, *Proceedings of the Section on Survey Research Methods, American Statistical Association,* pp. 183–188.

Binder, D.A. (1982). Non-Parametric Bayesian Models for Samples from Finite Populations. *Journal of the Royal Statistical Society, Ser. B* **44,** 388–393.

Binder, D.A. (1983). On the Variances of Asymptotically Normal Estimators from Complex Surveys. *International Statistical Review* **51,** 279–292.

Binder, D.A. (1992). Fitting Cox's Proportional Hazards Models from Survey Data. *Biometrika* **79,** 139–147.

Binder, D.A., Kovar, J.G., Kumar, S., Paton, D. & Van Baaran, A. (1987). Analytic Uses of Survey Data: A Review. *Applied Probability, Stochastic Processes and Sampling Theory,* eds. I.B. MacNeil and G.J. Umphrey, D. Reidel Publishing Company, pp. 243–264.

Brewer, K.R.W. & Mellor, R.W. (1973). The Effect of Sample Structure on Analytical Surveys. *Australian Journal of Statistics,* **15,** 145–152.

Chambers, R.L. & Dunstan, R. (1986). Estimating Distribution Functions from Survey Data. *Biometrika* **73,** 597–604.

Chambless, L.E. & Boyle, K.E. (1985). Maximum Likelihood Methods for Complex Sample Data: Logistic Regression and Discrete Proportional Hazards Models. *Communications in Statistics—Theory and Methods* **14,** 1377–1392.

Cox, B.G. (1987). Weighting Survey Data for Analysis. Internal Document, Division of Statistical Sciences, Research Triangle Institute, Research Triangle Park, NC. 27709-2194 U.S.A.

DeMets, D. & Halperin, M. (1977). Estimation of Simple Regression Coefficients in Samples Arising from Sub-Sampling Procedures. *Biometrics* **33,** 47–56.

DuMouchel, W.H. & Duncan, G.J. (1983). Using Sample Survey Weights in Multiple Regression Analysis of Stratified Samples. *Journal of the American Statistical Association* **78**, 535–543.

Fay, R.E. (1985). A Jackknifed Chi-Squared Test for Complex Samples. *Journal of the American Statistical Association* **80,** 148–157.

Fuller, W.A. (1975). Regression Analysis for Sample Surveys. *Sankhya, Ser. C* **37,** 117–132.
Fuller, W.A. (1984). Least Squares and Related Analyses for Complex Survey Designs. *Survey Methodology* **10,** 97–118.
Godambe, V.P. & Thompson, M.E. (1986). Parameters of Superpopulation and Survey Population: Their Relationships and Estimation. *International Statistical Review* **54,** 127–138.
Grizzle, J.E., Starmer, C.F. & Koch, G.G. (1969). Analysis of Categorical Data by Linear Models. *Biometrics* **25,** 489–504.
Harlan, W.R., Landis, J.R., Schmouder, R.L., Goldstein, N.G. & Harlan, L.C. (1985). Blood Lead and Blood Pressure. *Journal of the American Medical Association* **253,** 530–534.
Hausman, J.A. (1978). Specification Tests in Econometrics. *Econometrica* **46,** 1251–1271.
Hausman, J.A. & Wise, D.A. (1981). Stratification on Endogenous Variables and Estimation: The Gary Income Maintenance Experiment. *Structural Analysis of Discrete Data with Econometric Applications,* eds. C.F. Mansky and D. McFadden, Cambridge, Mass: MIT Press, pp. 366–391.
Hidiroglou, M.A. & Rao, J.N.K. (1987). Chi-Squared Tests with Categorical Data from Complex Surveys: Part I—Simple Goodness of Fit, Homogeneity and Independence in a Two-Way Table with Application to the Canada Health Survey (1978–1979). *Journal of Official Statistics* **3,** 117–132.
Holt, D., Smith, T.M.F. & Winter, P.D. (1980). Regression Analysis of Data from Complex Surveys. *Journal of the Royal Statistical Society, Ser. A* **143,** 474–487.
Isaki, C.T. & Fuller, W.A. (1982). Survey Design Under a Regression Superpopulation Model. *Journal of the American Statistical Association* **77,** 89–96.
Jewell, N.P. (1985). Least Squares Regression With Data Arising from Stratified Samples of the Dependent Variable. *Biometrika* **72,** 11–21.
Jonrup, H. & Rennermalm, B. (1976). Regression Analysis in Samples from Finite Populations. *Scandinavian Journal of Statistics* **3,** 33–37.
Kasprzyk, D., Duncan, G., Kalton, G. & Singh, M.P. eds. (1989). *Panel Surveys.* New York: John Wiley.
Kish, L. (1990). Weighting: Why, When and How? A Survey for Surveys. *Proceedings of the section on Survey Research Methods, American Statistical Association,* pp. 121–130.
Kish, L. & Frankel, M.P. (1974). Inference from Complex Samples. *Journal of the Royal Statistical Society, Ser. B* **36,** 1–37.
Koch, G.G., Freeman, D.H. Jr. & Freeman, J.L. (1975). Strategies in the Multivariate Analysis of Data from Complex Surveys. *International Statistical Review* **43,** 59–78.
Krieger, A.M. & Pfeffermann, D. (1992). Maximum Likelihood Estimation from Complex Sample Surveys. *Survey Methodology* **18,** 225–239.
Little, R.J.A. (1982). Models for Nonresponse in Sample Surveys. *Journal of the American Statistical Association* **77,** 237–250.
Little, R.J.A. (1983). Estimating a Finite Population Mean from Unequal Probability Samples. *Journal of the American Statistical Association* **78,** 596–604.
Little, R.J.A. (1989). Survey Inference With Weights for Differential Sample Selection or Nonresponse. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* pp. 62–69.
McDowell, A., Engel, A., Massey, J.T. & Maurer, K. (1981). Plan and Operation of the Second National Health and Nutrition Examination Survey, 1976–1980. DHHS publication No. (DHS) 81-1317, US Department of Health and Human Services, National Center for Health Statistics, Hyattsville, MD.
Nathan, G. (1975). Tests of Independence in Contingency Tables from Stratified Proportional Samples. *Sankhya, Ser. C* **37,** 77–87.
Nathan, G. (1988). Inference Based on Data from Complex Sample Designs. *Handbook of Statistics,* eds. P.R. Krishnaiah and C.R. Rao, Elsevier Science Publishers B.V. pp. 247–266.
Nathan, G. & Holt, D. (1980). The Effect of Survey Design on Regression Analysis. *Journal of the Royal Statistical Society, Ser. B* **42,** 377–386.
Nordberg, L. (1989). Generalized Linear Modeling of Sample Survey Data. *Journal of Official Statistics* **5,** 223–239.
Patil, G.P. & Rao, C.R. (1978). Weighted Distributions and Size Biased Sampling With Applications to Wildlife Populations and Human Families. *Biometrics* **34,** 179–189.
Pfeffermann, D. (1988). The Effect of Sampling Design and Response Mechanism on Multivariate Regression-Based Predictors. *Journal of the American Statistical Association* **83,** 824–833.
Pfeffermann, D. & Holmes, D.J. (1985). Robustness Considerations in the Choice of Method of Inference for Regression Analysis of Survey Data. *Journal of the Royal Statistical Society, Ser. A* **148,** 268–278.
Pfeffermann, D. & Nathan, G. (1985). Problems in Model Identification Based on Data from Complex Sample Surveys. *Bulletin of the International Statistical Institute* **51,** pp. 12.2.1–12.2.17.
Pfeffermann, D. & Smith, T.M.F. (1985). Regression Models for Grouped Populations in Cross-Section Surveys. *International Statistical Review* **53,** 37–59.
Rao, C.R. (1985). Weighted Distributions Arising Out of Methods of Ascertainment: What Population Does a Sample Represent? *A Celebration in Statistics,* ISI Centenary Volume, A.C. Atkinson and S.E. Fienberg, eds., New York: Springer-Verlag, pp. 543–569.
Rao, J.N.K., Kovar, J.G. & Mantel, H.J. (1990). On Estimating Distribution Functions and Quantiles from Survey Data Using Auxiliary Information. *Biometrika* **77,** 365–375.
Rao, J.N.K. & Scott, A.J. (1981). The Analysis of Categorical Data from Complex Sample Surveys: Chi-Squared Tests for Goodness of Fit and Independence in Two-Way Tables. *Journal of the American Statistical Association* **76,** 221–230.

Rao, J.N.K. & Scott, A.J. (1984). On Chi-Squared Tests for Multiway Contingency Tables With Cell Proportions Estimated from Survey Data. *Annals of Statistics* **12,** 46–60.

Rao, J.N.K. & Scott, A.J. (1987). On Simple Adjustments to Chi-Squared Tests With Sample Survey Data. *Annals of Statistics* **15,** 385–397.

Rubin, D.B. (1976). Inference and Missing Data. *Biometrika* **53,** 581–592.

Rubin, D.B. (1985). The Use of Propensity Scores in Applied Bayesian Inference. *Bayesian Statistics 2,* eds. J.M. Bernardo, M.H. Degroot, D.V. Lindley and A.F.M. Smith, Elsevier Science Publishers B.V. pp. 463–472.

Scott, A.J. (1977). Some Comments on the Problem of Randomization in Surveys. *Sankhya, Ser. C* **39,** 1–9.

Scott, A.J. & Holt, D. (1982). The Effect of Two Stage Sampling on Ordinary Least Squared Methods. *Journal of the American Statistical Association* **77,** 848–854.

Shah, B.V., Holt, M.M. & Folsom, R.F. (1977). Inference About Regression Models from Sample Survey Data. *Bulletin of the International Statistical Institute* **47,** pp. 43–57.

Skinner, C.J., Holt, D. & Smith, T.M.F. eds. (1989). *Analysis of Complex Surveys.* New York: John Wiley.

Smith, T.M.F. (1984). Present Position and Potential Developments: Some Personal Views—Sample Surveys. *Journal of the Royal Statistical Society, Ser. A* **147,** 208–221.

Smith, T.M.F. (1988). To Weight or Not to Weight, That is the Question. *Bayesian Statistics 3,* eds. J.M. Bernardo, M.H. Degroot, D.V. Lindley and A.F.M. Smith, Oxford University Press, pp. 437–451.

Sugden, R.A. & Smith, T.M.F. (1984). Ignorable and Informative Designs in Survey Sampling Inference. *Biometrika* **71,** 495–506.

Résumé

Le but de cet exposé est de fournir un examen critique des recherches, pour réspondre à deux questions: (i) Est-ce que l'emploi des poids de sondage peut être justifié pour l'inférence analytique sur les paramètres d'un modèle et, dans ce cas, dans quelles circonstances? (ii) Peut-on développer des lignes directrices pour l'introduction des poids dans l'analyse? La conclusion générale de cette étude est que les poids peuvent être utilisés pour tester et pour protéger contre des plans de sondage informatifs et contre la spécification fausse du modèle de la population. On condière six approaches differentes pour introduire les poids dans le processus d'inférence. Les quatre premières ont pour but de produire des estimateurs qui sont consistants selon le plan de sondage pour les quantités déscriptives de la population qui correspondent aux paramètres du modèle. Les deux autres approches essaient d'introduire les poids dans le modèle.

[*Received June* 1991, *accepted April* 1992]