Jour. Ind. Soc. Ag. Statistics
49 (Golden Jub. No.), 1996-97 : 21–46

Optimal Estimation in a Causal Framework

V.P. Godambe and M.E. Thompson
University of Waterloo, Ontario, Canada

SUMMARY

Consider a population consisting of units or individuals and two treatments. *In principle* or potentially, with every individual are associated responses y and y′ under the first and second treatment respectively. *In practice*, however, each individual can receive only one of the two treatments. Suppose one is given responses for a sample of individuals drawn from the population and treated with one treatment, as well as responses from another sample treated with the other treatment. From these sampled responses, one is supposed to estimate an average response difference over the entire population.

The sampling design may be known fully or partially. Such a framework is often used to investigate the measurement of causal effects in randomized experiments or observational studies. These investigations have provided deep philosophical insight into the problem of causation. Yet, by restricting themselves generally to finding *an* unbiased estimate of 'causal effect', they have mostly ignored the problem of 'optimal estimation'. In contrast, in our paper, although we do not discuss causation itself, we investigate 'optimal estimation' of quantities which may represent departures from hypotheses of 'no effect'. In Section 9, we particularly emphasize how optimality considerations are tied to the concept of 'nonconfounding' which is so basic for most causal theories.

Key words : Confounding, Estimating functions, Ignorable treatment assignment, Inclusion probabilities, Post-sampling stratification, Optimal estimation, Propensities, Superpopulation.

1. Introduction

A current view among some thinkers, particularly theoretical physicists (Bell [1]), is that reality is 'acausal'. Whatever is the ontological significance of cause-effect concepts, they are an integral part of our epistemology. Traditionally 'causation' has played an important role in all scientific development. The idea of 'manipulability of cause' seems essential for most scientific investigations. These investigations, it can however be argued, could also be carried out by restricting ourselves simply to situations of *unconfounded association.* Here, *unconfoundedness* is with reference to all *known* factors, and is thus 'provisional'; subsequent knowledge of new factors can change the situation. Such a performatory viewpoint of causation enables one to get around a very deep and often frustrating questions: What is a cause?

Even if a cause is identified only up to an unconfounded association, it can be practically fruitful to measure or estimate the 'causal effect'. In this paper a model is provided which assumes a causal framework to estimate 'causal effects' in the sense just mentioned. The context is a population consisting of units/individuals and the existence of two treatments such that with every individual are associated responses y and y′ under the first and second treatment respectively. We are given responses for a sample drawn from the population and treated with one treatment, as well as responses from another sample treated with the other treatment.

In this paper causation per se is not discussed. However, the bearing of such related concepts as covariate sufficiency (Stone [20]), randomization (Fisher ([5], [7])) unconfounded association (Cox [4]), ignorable treatment assignment (Rosenbaum and Rubin [17]) etc. on our model is explained in the paper. A generally useful reference is Holland [13].

We begin by making the model precise, and in Section 2 discuss several ways of expressing a null hypothesis of 'no effect'. Each of these is associated with a population or superpopulation quantity or quantities which may represent departures from the null hypothesis. The purpose of the rest of the paper is to discuss optimal estimation of these quantities, using the theory of estimating functions.

Our model concerns a finite population of N units or individuals i, denoted by $\mathcal{P} = \{i : i = 1, \ldots, N\}$. From the subsequent discussion it will be clear that the knowledge of the size of the population $|\mathcal{P}| = N$ is not always essential.

Suppose that there are two treatments, T and T′. Denote by y_i the response value which would be associated with unit i if it were given treatment T, and by y'_i the response value which would arise from treatment T′.

In our framework imagine a treatment to have been assigned, though not necessarily administered, to every unit in the population. Let $z_i = 1$ if individual i is assigned treatment T and $z_i = 0$ if it is assigned treatment T′. The assignment is summarized in the treatment vector $\mathbf{z} = (z_1, \ldots, z_N)$.

Further, suppose there is a covariate x, with value x_i for unit i, and let $\mathbf{x} = (x_1, \ldots, x_N)$ denote the array of x values for the population. Assume a semiparametric superpopulation model, as follows :

(i) conditional on $\mathbf{x}$, the variates (y_i, y'_i), $i = 1, \ldots, N$ are independent;

(ii) $E(y_i \mid \mathbf{x}) = \theta(x_i)$ and $E(y'_i \mid \mathbf{x}) = \theta'(x_i)$;

(iii) $\mathrm{Var}(y_i \mid \mathbf{x}) = v(x_i)$, $\mathrm{Var}(y'_i \mid \mathbf{x}) = v'(x_i)$ and $\mathrm{Cov}(y_i, y'_i \mid \mathbf{x}) = c(x_i)$.

The above model clearly satisfies, at the level of first and second order moments, the following conditions often mentioned in relation to causal inference. (1) Non-interaction between units (Cox [3], Kempthorne [14], Rubin [18]): the joint distributions of y_i, y'_i given $\mathbf{x}$ depend exclusively on the specified unit i and the specified treatment. (2) Covariate sufficiency and the implied nonconfounding (Stone [20]): For any two units i_1 and i_2 with covariates $x_{i_1} = x_{i_2}$, the joint distributions of (y_{i_1}, y'_{i_1}) and (y_{i_2}, y'_{i_2}) given $\mathbf{x}$ are the same. That is, the joint distributions conditional on $\mathbf{x}$ are not affected by any unknown factors.

To complete the model, it is necessary to make assumptions about the treatment assignment $\mathbf{z}$. Assume the treatment assignment to be *ignorable*, in the sense of giving no information about the y and y′ values for the population, as follows.

Assumption : Given $\mathbf{x}$, the treatment assignment vector $\mathbf{z}$ and the response values $\{ y_i, y'_i : i = 1, \ldots, N \}$ are independent.

For a detailed discussion of the ignorability assumption, particularly in relation to ideas of 'causation', refer to Rosenbaum and Rubin [17] and Stone [20]. This assumption is consistent with our formulation, in which we have in fact defined (y_i, y'_i) as existing before z_i. It also plays a central role in the construction of our estimating functions for the superpopulation parameters, as in (10) of Section 3. It is satisfied in randomized experiments (Fisher [7], Sprott and Farewell [19]) but generally cannot be more than a hopeful assumption in observational studies. We will discuss in Section 9 the extent to which the ignorability assumption can be relaxed for estimation of finite population quantities, particularly Δ of Section 2.

The values $z_i, i = 1, \ldots, N$ are now thought of as random variates. The probability that individual i receives treatment T is called its *propensity*, and is denoted by α_i :

$$\alpha_i = P(z_i = 1 \mid \mathbf{x}) \tag{1}$$

Thus $1 - \alpha_i = P(z_i = 0 \mid \mathbf{x})$, and assume $0 < \alpha_i < 1$ for all i. Do not assume that the z_i are independent of each other unless it is said so specifically.

In a controlled experiment, the α_i are known and under the control of the experimenter. In an observational study, the α_i are typically unknown, and uncontrollable. With this latter context in mind, in the case where x defines strata we do not assume the α_i to be constant within strata.

2. *Hypotheses of No Effect*

In testing for treatment differences, there are several possible ways of formulating a hypothesis of "no effect". The simplest and most direct is Fisher's [6] formulation

$$H1: \ y_i = y'_i \ \text{for all i}$$

Without further assumptions, the full strength of this hypothesis cannot be tested because for any unit i only one of y_i, y'_i is observed. A weaker and more easily tested hypothesis, which might be called "no effect in expectation", would be written as

$$H2: \ \theta(\cdot) \equiv \theta'(\cdot)$$

implying $\theta(x_i) = \theta'(x_i)$ for all i. Another kind of weakening gives the hypothesis of "no effect on average over the population" :

$$H3: \Delta = \frac{1}{N} \sum_{i=1}^{N} (y_i - y'_i) = 0$$

with superpopulation counterpart

$$H4: \frac{1}{N} \sum_{i=1}^{N} (\theta(x_i) - \theta'(x_i)) = 0$$

In these latter hypotheses, "= 0" might more meaningfully be replaced by "is negligibly small".

We will confine attention mainly to hypotheses H2, H3 and H4 in what follows. However, under some models for potential departures from the null hypothesis, we might want to replace H3 by "no effect on average" in another sense, such as

$$H3': \sum_{i=1}^{N} a_i (y_i - y'_i) = 0$$

for constants $a_i > 0$ or

$$H3'': \sum_{i=1}^{N} y_i / \sum_{i=1}^{N} y'_i = 1$$

See Sections 4 and 5 for further discussion of these.

There are three important special conditions which allow the testing of hypothesis H2. The first, the *stratification* case, occurs when the value of x stratify the population into strata of reasonably large sizes :

$$\mathcal{P}_j = \{ i : x_i = x^j \}, \; j = 1, ..., k$$

In that case, if $\theta_j = \theta (x^j)$, $\theta'_j = \theta' (x^j)$, H2 becomes

$$H2_{st} : \theta_j = \theta'_j, j = 1, \ldots, k$$

The second, the *regression case*, has

$$\theta (x_i) = \theta f (x_i), \; \theta' (x_i) = \theta' \; f (x_i)$$

where f is a known function. Here H2 becomes

$$H2_{reg} : \; \theta = \theta'$$

The third is a *constant difference* assumption, that

$$\theta (\cdot) = \theta' (\cdot) + \gamma$$

for some constant γ. In that case, H2 is equivalent to

$$H2_{cd} : \; \gamma = 0$$

These three special cases will be considered in Sections 3, 4 and 5 respectively.

It is important to emphasize that the superpopulation given by (i) - (iii) of Section 1 is hypothetical, in contrast to actual survey or experimental population $\mathcal{P}$. In observational studies, such as epidemiological studies, the survey population is in many cases left to the imagination and interpretation of the investigator. However in our setup the superpopulation and the experimental population are clearly distinguished.

Now in relation to the survey population $\mathcal{P}$ we define a sampling design d:

$$d = (S, p) \tag{2}$$

where $\mathcal{S} = \{s : s \subset \mathcal{P}\}$, and p is a probability distribution on $\mathcal{S}$. Following the general terminology of survey sampling, s is called a 'sample'. In the conduct of a survey, a sample s is drawn using the sampling design d. For every individual i, the probability of its being included in the sample drawn is denoted by π_i :

$$\pi_i = P (s \ni i), i = 1, \ldots N \tag{3}$$

Assume the sampling design d in (2) to be such that all inclusion probabilities π_i are positive. Here it is important to distinguish the inclusion probabilities π_i in (3) from the propensities α_i in (1); the former are assumed to be under the control of the investigator while the latter need not be. Allow the possibility that π_i may depend on the z values, i.e. on the treatment allocations, as in the example of Section 7.

After the sample s is drawn, each individual $i \in s$ is seen to be given the treatment T or T′ according as $z_i = 1$ or 0; the corresponding response y_i or y'_i is observed. Thus we can denote the survey data as

$$(i, \mathbf{y}_i) : \; i \in s \tag{4}$$

where

$$\mathbf{y}_i = [\, z_i, (y_i z_i), \{ y'_i (1 - z_i) \}, x_i \,] \tag{5}$$

We note there are three sources of stochastic variation in the data, namely the superpopulation model for the y_i and y'_i, the generation of the treatment assignments z_i, and the sampling design producing s.

The problems of estimation which address the testing of the hypotheses H2 – H4 can now be stated as follows. Given the data in (4), the sampling design d in (2) and the superpopulation model in (i) - (iii) and (1), to test H2 we would want to estimate

$$\theta_j \text{ and } \theta'_j, \; j = 1, \ldots, k \tag{6}$$

in the stratification case, θ and θ' or perhaps $\theta - \theta'$ or θ / θ' in the regression case, and γ in the constant difference case. In testing the hypothesis H3 we would want to estimate

$$\Delta = \sum_{i=1}^{N} (y_i - y'_i) / N \tag{7}$$

Testing H4 in the stratification case would mean estimating

$$\Theta = \sum_{j=1}^{K} N_j \, (\theta_j - \theta'_j) / N \tag{8}$$

3. *Estimating Functions : Stratification Case*

First consider the case where x stratifies the population. To estimate the 2k parameters of interest for H2, namely $\theta_j, \theta'_j, j = 1, \ldots, k$ of the

superpopulation model, the following are the elementary estimating functions. For $i \in \mathcal{P}_j$, define

$$\phi_i = \frac{y_i - \theta_j}{\alpha_i} z_i, \quad \phi'_i = \frac{y'_i - \theta'_j}{1 - \alpha_i}(1 - z_i), \qquad j = 1, \ldots, k \tag{9}$$

Using the assumptions of Section 1, we have

$$E(\phi_i \mid \mathbf{z}, \mathbf{x}) = E(\phi'_i \mid \mathbf{z}, \mathbf{x}) = 0$$

$i = 1, \ldots, N$, and the functions are all orthogonal for fixed $\mathbf{x}$, unconditionally or conditionally on $\mathbf{z}$. Clearly one could also define as elementary estimating functions, the functions of (9) with the denominators α_i and $1 - \alpha_i$ omitted. However, their inclusion will be convenient for the development at the end of this section, since the expectation of ϕ_i holding $\mathbf{y}$, $\mathbf{y}'$ fixed is simply $(y_i - \theta_j)$.

If hypothetically the variates $\mathbf{y}_i$ in (5) were known for all the individuals in the population $\mathcal{P}$, the population based optimal estimating functions conditional on $\mathbf{z}, \mathbf{x}$ would be given by

$$g_j = \sum_{i=1}^{N} \left[\phi_i \frac{E\{(\partial \phi_i / \partial \theta_j) \mid \mathbf{z}, \mathbf{x}\}}{E(\phi_i^2 \mid \mathbf{z}, \mathbf{x})} + \phi'_i \frac{E\{(\partial \phi'_i / \partial \theta_j) \mid \mathbf{z}, \mathbf{x}\}}{E\{(\phi'_i)^2 \mid \mathbf{z}, \mathbf{x}\}} \right]$$

$$g'_j = \sum_{i=1}^{N} \left[\phi_i \frac{E\{(\partial \phi_i / \partial \theta'_j) \mid \mathbf{z}, \mathbf{x}\}}{E(\phi_i^2 \mid \mathbf{z}, \mathbf{x})} + \phi'_i \frac{E\{(\partial \phi'_i / \partial \theta'_j) \mid \mathbf{z}, \mathbf{x}\}}{E\{(\phi'_i)^2 \mid \mathbf{z}, \mathbf{x}\}} \right]$$

$j = 1, \ldots, k$ (Godambe and Thompson [12]). Similar expressions are obtainable for optimal estimating functions conditional on $\mathbf{x}$ only. It is easy to see that in either case the equations $g_j = 0, g'_j = 0$ are equivalent to

$$\sum_{i \in \mathcal{P}_j} (y_i - \theta_j) z_i = 0, \quad \sum_{i \in \mathcal{P}_j} (y'_i - \theta'_j)(1 - z_i) = 0 \tag{10}$$

yielding as the population based estimates for θ_j and θ'_j the corresponding means of the y and y′ values. It is interesting to note that these estimates and the optimal estimating equations (10), being optimal conditionally on $(\mathbf{z}, \mathbf{x})$, as well as on $\mathbf{x}$, are *independent* of the propensities α_i in (1).

Now for the sampling design d in (2), the optimal estimating equations for the parameters θ_j, θ'_j, based on the survey data (4), are obtained from (10) as

$$\sum_{i \in s_j} \{ (y_i - \theta_j) z_i / \pi_i \} = 0, \quad \sum_{i \in s_j} \{ (y'_i - \theta'_j) (1 - z_i) / \pi_i \} = 0 \qquad (11)$$

where $s_j = s \cap \mathcal{P}_j, j = 1, \ldots, k$ and $\pi_i, i = 1, \ldots, N$ are the inclusion probabilities in (3) (Godambe and Thompson [11]). From (11) we get the estimates

$$\hat{\theta}_j = \left(\sum_{i \in s_j} y_i z_i / \pi_i \right) \Big/ \left(\sum_{i \in s_j} z_i / \pi_i \right)$$

$$\hat{\theta}'_j = \left\{ \sum_{i \in s_j} y'_i (1 - z_i) / \pi_i \right\} \Big/ \left\{ \sum_{i \in s_j} (1 - z_i) / \pi_i \right\} \qquad (12)$$

$j = 1, \ldots, k$. Note that because of their relationship with (10) the estimating equations (11) and the estimates in (12) are independent of the propensities α_i in (1). However, they do depend on the design inclusion probabilities π_i given by (3). For a sampling design with inclusion probabilities π_i constant in each stratum,

$$\pi_i = \pi^{(j)}, i \in \mathcal{P}_j, j = 1, \ldots, k \qquad (13)$$

the estimates $\hat{\theta}_j$ and $\hat{\theta}'_j$ reduce to the sample means

$$\bar{y}_j = \left(\sum_{i \in s_j} y_i z_i \right) \Big/ \left(\sum_{i \in s_j} z_i \right)$$

$$\bar{y}'_j = \sum_{i \in s_j} y'_i (1 - z_i) \Big/ \left\{ \sum_{i \in s_j} (1 - z_i) \right\} \qquad (14)$$

Among the sampling designs which satisfy condition (13) simple random sampling and stratified random sampling are obviously included.

The inclusion of design weights $1/\pi_i$ in (11) relates those equations to equations (10). It renders the estimation based on (11) "robust" against some departures from the assumed model (i) - (iii) in the following sense: even if the solution of (10) fail to have meaning in terms of the superpopulation, the solution of (11) will estimate them as finite population parameters (Godambe and Thompson [11]). However, if this robustness is not required we might ignore the design weights and use the estimation given by (14).

Turning now to H4, according to criteria of estimating function theory (Godambe and Thompson ([11], [12])) the optimal estimation for the superpopulation parameter Θ in (8) is given by the estimate

$$\hat{\Theta} = \sum_{j=1}^{k} N_j (\hat{\theta}_j - \hat{\theta}'_j) / N \tag{15}$$

where $\hat{\theta}_j$ and $\hat{\theta}'_j$ are as in (12).

Finally we set out to estimate the finite population parameter relevant to H3, namely Δ of (7), in a manner "assisted' by the estimation of the superpopulation parameters. In this direction, construct a function

$$H = \sum_{i=1}^{N} (\theta_i - \theta'_i) / N \tag{16}$$

where ϕ_i and ϕ'_i are the same as in (9). The superpopulation expectation of H conditional on $y_i, y'_i, i = 1, \ldots, N$ is given by

$$E(H \mid y_i, y'_i : i = 1, \ldots, N; \mathbf{x}) = \Delta - \Theta \tag{17}$$

where Δ and Θ are as in (7) and (8). Under suitable conditions for a generalized law of large numbers, for large N the right hand side of (17) is approximated by the function H in (16) :

$$H \cong \Delta - \Theta \tag{18}$$

These relationships have a special significance in the context of our estimation problem. Under the superpopulation model of Section 1, for any fixed $\theta_j, \theta'_j, j = 1, \ldots, k$ and the sampling design in (2), the optimal estimate of H based on the survey data (4) is given by

$$\hat{H} = \{ \sum_{i \in s} (\phi_i - \phi'_i) / \pi_i \} / N \tag{19}$$

where $\pi_i, i = 1, \ldots, N$ are the design inclusion probabilities in (3) (Godambe and Thompson [11]). Note that the estimate $\hat{H}$ does depend essentially on the propensities α_i in (1).

The required estimate of the finite population parameter Δ in (7) can be achieved in two steps :

1. H in (16) is replaced by $\hat{H}$ in (19).

2. $\theta_j, \theta'_j, j = 1, \ldots, k$ are replaced by their optimal estimates given by (12), or (14) if appropriate.

Thus the parameter Δ is estimated *approximately optimally* because of (17) and (18) by

$$\hat{\Delta} = \hat{H}(\hat{\theta}, \hat{\theta}') + \hat{\Theta} \tag{20}$$

where $\hat{\Theta}$ is as in (15) and

$$\hat{H}(\hat{\theta}, \hat{\theta}') = \sum_{j=1}^{k} \sum_{i \in s_j} \left[\frac{\{(y_i - \hat{\theta}_j) z_i\}}{(\pi_i \alpha_i)} - \frac{\{(y'_i - \hat{\theta}'_j)(1 - z_i)\}}{\{\pi_i (1 - \alpha_i)\}} \right] / N \tag{21}$$

It is easy to see that if the propensity α_i depends exclusively on the covariate $x_i, i = 1, \ldots, N$, then because of (12)

$$\hat{H}(\hat{\theta}, \hat{\theta}') = 0$$

in (20), and the estimate $\hat{\Delta}$ would then reduce to

$$\sum_{j=1}^{k} \frac{N_j}{N} (\hat{\theta}_j - \hat{\theta}'_j) \tag{22}$$

which is independent of the α_i. The estimate (22) for a simple random sampling design is equivalent to

$$\sum_{j=1}^{k} \frac{N_j}{N} (\bar{y}_j - \bar{y}'_j) \tag{23}$$

where $\bar{y}_j$ and $\bar{y}'_j$ are given by (14). This is a kind of post-sampling stratification. If the stratum proportions N_j / N are not known, but the sample size is large, we might replace N_j / N in (23) by appropriate estimates.

The foregoing arguments supporting the use of the estimate $\hat{H}$ in (19) for $\Delta - \Theta$ in (17) also support, for each stratum $\mathcal{P}_j$, the use of the estimate

$$\sum_{i \in s_j} (\phi_i / \pi_i) \quad \text{for} \quad \left\{ \sum_{i \in \mathcal{P}_j} (y_i - \theta_j) \right\} \tag{24}$$

and the estimate

$$\sum_{i \in s_j} (\phi'_i / \pi_i) \quad \text{for} \quad \{ \sum_{i \in \mathcal{P}_j} (y'_i - \theta'_j) \}$$

assuming $|\mathcal{P}_j|$ is large. Now under the superpopulation model, based on complete observation of (y_i, y'_i), $i \in \mathcal{P}_j$, the two expression $\{\cdot\}$ in (24) are optimal estimating functions for θ_j and θ'_j respectively. This fact suggests that sample based estimates $(\tilde{\theta}, \tilde{\theta}')$, alternative to $(\hat{\theta}, \hat{\theta}')$ discussed above, could be obtained by solving the equations $\{\Sigma_{i \in s_j} (\phi_i / \pi_i)\}_{\tilde{\theta}_j} = 0$ and $\{\Sigma_{i \in s_j} (\phi'_i / \pi_i)\}_{\tilde{\theta}'_j} = 0, j = 1, \ldots, k$. Note now that in (19), $\hat{H}(\tilde{\theta}, \tilde{\theta}')$ is automatically 0, and an estimate of Δ is obtained from (22) directly by replacing in it $(\hat{\theta}, \hat{\theta}')$ by $(\tilde{\theta}, \tilde{\theta}')$:

$$\tilde{\theta}_j = \left(\sum_{i \in s_j} y_i z_i / \alpha_i \pi_i \right) / \left(\sum_{i \in s_j} z_i / \alpha_i \pi_i \right) \tag{25}$$

$$\tilde{\theta}'_j = \left\{ \sum_{i \in s_j} y'_i (i - z_i) / (1 - \alpha_i) \pi_i \right\} / \left\{ \sum_{i \in s_j} (1 - z_i) / (1 - \alpha_i) \pi_i \right\}$$

$$j = 1, \ldots, k$$

Thus applying the principles of estimating function theory in different orders may yield different results in this context. Three points are to be emphasized here. (i) Regardless of the condition that propensities α_i are functions of the covariates x_i, we have $\tilde{H}(\tilde{\theta}, \tilde{\theta}') = 0$, while as noted in the preceding paragraph, that condition is necessary for $\hat{H}(\hat{\theta}, \hat{\theta}') = 0$. The estimate of Δ based on (25) is thus simpler in general than the one in (20). (ii) However, unlike the estimates $(\tilde{\theta}, \tilde{\theta}')$, the estimates $(\hat{\theta}, \hat{\theta}')$ are free from dependence on the propensities α_i, as seems fitting under the ignorability assumption. (iii) Also, conditional on $\mathbf{z}, \mathbf{x}$ the estimating functions leading to $(\hat{\theta}, \hat{\theta}')$ are optimal, while those leading to $(\tilde{\theta}, \tilde{\theta}')$ are not.

We now elaborate briefly on the 'approximate optimality' of the estimate $\hat{\Delta}$ in (20). For any fixed $\theta_j, \theta'_j, j = 1, \ldots, k$ the estimate $\hat{H}$ in (19) is 'optimal' for H in (16), in the sense of its having minimum expected variance in the class of all unbiased estimates of H. Here, 'unbiasedness' and 'variance' are with respect to the sampling design (2), and the 'expectation' is with respect to the superpopulation model. The estimate $\hat{\Delta}$ is arrived at using two approximations to the just stated exact 'optimality'. The first approximation (18) is derived from (17) with a law of large numbers. The second approximation consists in replacing θ_j and θ'_j, $j = 1, \ldots, k$ by their estimates in $\Delta - \Theta$. This second step is justified as follows. Let

$$g = \Delta - \Theta - \hat{H}$$

Given (θ, θ') the estimate of Δ is obtained by solving the estimating equation $g = 0$. Now if E denotes the expectation with respect to the sampling design d in (2), and E as before the superpopulation expectation, then $EE(g \mid y_i, y'_i : i = 1, \ldots, N; \mathbf{x}) = 0$. Further since

$$(\partial g / \partial \theta_j) = -N_j + \sum_{i \in s_j} \{ z_i / (\alpha_i \pi_i) \}$$

$EE \{ (\partial g / \partial \theta_j) \} = 0$. Similarly $E E \{(\partial g / \partial \theta'_j)\} = 0$. Actually for many sampling designs, including simple random sampling and stratified random sampling with sufficiently large N_j and n_j, the quantities $(\partial g / \partial \theta_j)$ and $(\partial g / \partial \theta'_j)$, $j = 1, \ldots, k$ would themselves be close to zero. This implies (Godambe, [9]) that asymptotically the efficiencies of the estimating function g, and the one obtained from g by replacing θ_j and θ'_j by their estimates, will be nearly the same.

4. *Estimating Function : Regression Case*

In this section we suppose it is reasonable to assume

$$\theta(x_i) = \theta f(x_i), \quad \theta'(x_i) = \theta' f(x_i) \tag{26}$$

$i = 1, \ldots, N$, where θ, θ' are unknown parameters, and f is a completely specified function. In the context of the previous section, this would reduce the number of unknown parameters to be estimated from 2k to 2.

In general, the population based optimal estimating equations for θ and θ' are given by

$$\sum_{i=1}^{N} \{ y_i - \theta f(x_i) \} \frac{f(x_i)}{v(x_i)} z_i = 0, \quad \sum_{i=1}^{N} \{ y'_i - \theta' f(x_i) \} \frac{f(x_i)}{v'(x_i)} (1 - z_i) = 0 \tag{27}$$

where v and v' are the model variance functions defined in (iii) of Section 1. Further, the sample based optimum estimating equations are given by

$$\sum_{i \in s} \{ y_i - \theta f(x_i) \} \frac{f(x_i) z_i}{v(x_i) \pi_i} = 0$$

$$\sum_{i \in s} \{ y'_i - \theta' f(x_i) \} \frac{f(x_i)(1 - z_i)}{v'(x_i) \pi_i} = 0 \tag{28}$$

As in the case of (10) and (11), the optimality of (27) and (28) is both unconditional and conditional on z, the value of x being fixed in both cases.

Note that unlike the equations (10) and (11), equations (27) and (28) depend essentially on the variance functions v and v′. Thus their applicability is restricted to situations where these are known up to proportionality constants.

Here again, as in the previous section, the propensities α_i do not appear in the optimal estimating equations for superpopulation parameters. If estimating θ, θ' is the only objective, the assumption of Section 1 that $1 > \alpha_i > 0, i = 1, \ldots, N$ may be dropped. In some situations where the propensities α_i are under the control of the experimenter, it might seem reasonable to have $\alpha_i = 1$ when $f(x_i) > c$ and $\alpha_i = 0$ otherwise, c being a specified constant. (Robbins and Zhang [16], Godambe and Kunte [10]). These are situations where the experimenter is sure a priori that one treatment is at least as good as the other, but does not know how much better it is. The results from this "biased allocation" scheme yield optimal estimation through (28).

The hypothesis H2 (or $H2_{reg}$) can be tested through the estimation of $\delta = \theta / \theta'$. It is interesting to consider the case of an independent Poisson model, where if y_i, y'_i were observed for all i, the optimal population estimating equations for θ, θ' would be

$$\sum_{i=1}^{N} \{ y_i - \theta f(x_i) \} = 0, \quad \sum_{i=1}^{N} \{ y'_i - \theta' f(x_i) \} = 0$$

The implied population estimate for δ is a solution of the estimating equation

$$\kappa(\delta) = \sum_{i=1}^{N} (y_i - \delta y'_i) = 0 \tag{29}$$

Then $\kappa(\delta)$ is an analogue of $\Delta - \Theta$ of Section 3, and $\sum_{i=1}^{N} y_i / \sum_{i=1}^{N} y'_i$ of hypothesis H3" is an analogue of Δ. If one defines

$$K = K(y_1, \ldots, y_N; \delta) = \sum_{i=1}^{N} \left[\{ y_i - \theta f(x_i) \} \frac{z_i}{\alpha_i} - \delta \{ y'_i - \theta' f(x_i) \} \frac{(1 - z_i)}{1 - \alpha_i} \right] \tag{30}$$

then analogously to (17) we have

$$E(K \mid y_i, y'_i, i = 1, \ldots, N, \mathbf{x}) = \kappa(\delta)$$

The optimal estimate of K is

$$\sum_{i \in s} \{ y_i - \theta f(x_i) \} \frac{z_i}{\alpha_i \pi_i} - \delta \sum_{i \in s} \{ y'_i - \theta' f(x_i) \} \frac{(1 - z_i)}{\pi_i (1 - \alpha_i)} \qquad (31)$$

Suppose we are in the special *unbiased allocation* case where population units are paired so that $f(x_i)/\pi_i$ is constant within each pair, and T and T′ are assigned at random within each pair. Then (31) reduces to

$$\sum_{T} y_i / \pi_i - \delta \sum_{T'} y'_i / \pi_i \qquad (32)$$

where Σ_T and $\Sigma_{T'}$ are sums over sampled units receiving treatments T and T′ respectively. The estimating function (32) is approximately optimal (in the sense explained at the end of Section 3) for $\kappa(\delta)$ in (29). It also yields an estimate for δ, the ratio estimate

$$\hat{\delta} = \left(\sum_{T} y_i / \pi_i \right) \Big/ \left(\sum_{T'} y'_i / \pi_i \right)$$

This estimate $\hat{\delta}$ coincides with $\hat{\theta} / \hat{\theta}'$, but the validity of (32) as unbiased for $\kappa(\delta)$ holds even when the assumptions (26) do not. Note that for large samples, the estimating function in (31) would approximate to that in (32) even regardless of the assumption of 'unbiased allocation' mentioned above.

In the regression model (26) covariates **x** are assumed to be fixed. Suppose we extend the model by letting **x** have a distribution. Then in the extended model, the 'approximate optimality' of the estimating function (32) would hold both conditionally on **x** and unconditionally. Alternatively one can arrive at the estimating function (32) directly from the *unconditional model* in which, unconditionally, the expectations $E(y_i - \theta f(x_i)) = 0$ and $E(y'_i - \theta' f(x_i)) = 0$, with (y_i, y'_i, x_i), $i = 1, \ldots, N$ being assumed iid variates. This model implies that for all individuals i, $E(y_i - \delta y'_i) = 0$. Now again assuming unbiased allocation of treatments, that is that the propensities $\alpha_i = 1/2$, the elementary estimating functions for δ are $\{ y_i z_i - \delta y'_i (1 - z_i) \}$, $i = 1, \ldots, N$. These are unbiased, that is have '0' expectation, unconditionally (with respect to the joint distribution of all variates including **z**). Hence the *optimum* estimating function (unconditionally in the same sense) based on the data in (4) is given by (32). Of course this 'optimality' is in a more restricted class of estimating functions than the 'approximate optimality' of (32), mentioned earlier. (Godambe and Thompson ([11], [12])).

5. *Estimating Functions : Constant Difference Case*

Now consider the third special situation, where the superpopulation model is such that if θ_i and θ'_i denote $\theta(x_i)$ and $\theta'(x_i)$ respectively, then

$$\theta_i - \theta'_i = \gamma, \; i = 1, \ldots, N$$

Here there is a single parameter of interest γ, and a large number of unknown nuisance parameters contained in $\{\theta_1, \ldots, \theta_N\}$. The development at the end of the last section gives some clues to a possible approach.

Suppose that if (y_i, y'_i), $i = 1, \ldots, N$, were observed, the population estimating function for γ suggested by the superpopulation model would be

$$j(\gamma) = \sum_{i=1}^{N} c_i \{ (y_i - y'_i) - \gamma \} \tag{33}$$

Typically, c_i would be a function of the variances $v(x_i), v'(x_i)$ and the covariance $c(x_i)$. We will concentrate on the estimation of γ; the corresponding finite population parameter would be $\Sigma_{i=1}^{N} a_i (y_i - y'_i)$ of H3′ where $a_i = c_i / \Sigma_{i=1}^{N} c_i$. As was true for $\kappa(\delta)$ of (29), there is a combination of elementary estimating functions with expectation $j(\gamma)$, namely

$$J(y_1, \ldots, y_N; \gamma) = \sum_{i=1}^{N} c_i \left\{ (y_i - \theta_i) \frac{z_i}{\alpha_i} - (y'_i - \theta'_i) \frac{(1 - z_i)}{1 - \alpha_i} \right\} \tag{34}$$

It is easy to see that

$$E(J \mid y_i, y'_i, i = 1, \ldots, N; \mathbf{x}) = j(\gamma)$$

but J has the drawback of depending on the θ_i and θ'_i other than through γ. The same is true for its optimal sample estimating function

$$\hat{J} = \sum_{i \in s} c_i \left\{ (y_i - \theta_i) \frac{z_i}{\pi_i \alpha_i} - (y'_i - \theta'_i) \frac{(1 - z_i)}{\pi_i (1 - \alpha_i)} \right\} \tag{35}$$

Conditions under which (35) is useful are fairly narrow, but sometimes met with in practice. As for the estimation of δ in Section 3, suppose we have unbiased allocation of treatments, so that

$$\alpha_i = \frac{1}{2}, \; i = 1, \ldots, N$$

If we suppose further that

$$\sum_{i \in s} \frac{c_i z_i}{\pi_i} = \sum_{i \in s} \frac{c_i (1 - z_i)}{\pi_i} = C(s) \tag{36}$$

then

$$\frac{1}{2}\hat{J} = \left\{ \sum_{i \in s} \frac{c_i y_i z_i}{\pi_i} - \sum_{i \in s} \frac{c_i y'_i (1 - z_i)}{\pi_i} \right\} - \gamma C(s)$$

$$- \sum_{i \in s} \theta'_i \frac{c_i}{\pi_i} (z_i - (1 - z_i)) \tag{37}$$

If (36) is accomplished by pairing units with (approximately) equal values of c_i / π_i and of $\theta'_i c_i / \pi_i$, and by assigning the treatments T, T′ at random within each pair, the last term of (37) is (approximately) 0. (For a design with π_i all equal and c_i all equal, we would be pairing units thought to have approximately equal θ'_i.) Even otherwise, since the last term has expectation 0 conditional on **x**, we might expect with large samples to be able to neglect it. What is left is an estimating function for γ which yields

$$\hat{\gamma} = \left\{ \sum_{i \in s} \frac{c_i y_i z_i}{\pi_i} - \sum_{i \in s} \frac{c_i y'_i (1 - z_i)}{\pi_i} \right\} / C(s) \tag{38}$$

Even when the constant difference assumption itself is not justified, (38) is approximately unbiased for the finite population quantity

$$\sum_{i=1}^{N} c_i (y_i - y'_i) / \sum_{i=1}^{N} c_i$$

Analogously to Section 3, here also one can start with the corresponding 'unconditional model' (that is with **x** varying), implying $E(y_i - y'_i - \gamma) = 0$, and directly obtain optimal estimating functions (in the sense of the joint distribution of all variates) for γ.

6. Estimation of the Propensities

Now return to the stratification case of Section 2, where an estimate of

$$\Delta = \sum_{i=1}^{N} (y_i - y'_i) / N$$

was proposed in a very general setting.

If the propensities α_i are not constant in each stratum $\mathcal{P}_j$, we have $\hat{H}(\hat{\theta}, \hat{\theta}') \neq 0$ in (21). To compute it we need values for the propensities. Now suppose the α_i are not known. Then we might assume α_i to be a function of x_i and t_i where x_i is the covariate in the superpopulation model and t_i is an additional variate outside the superpopulation model. One physical interpretation might be that x_i is the 'size' while t_i is the 'location' of the individual i. For simplicity suppose $\log \{ \alpha_i / (1 - \alpha_i) \} = ax_i + bt_i$. That is,

$$\alpha_i = \frac{\exp(ax_i + bt_i)}{1 + \exp(ax_i + bt_i)} \tag{39}$$

$i = 1, \ldots, N$. To estimate a and b on the basis of $z_i, i = 1, \ldots, N$, we have the elementary estimating equations $z_i - \alpha_i = 0$, noting that $E(z_i - \alpha_i) = 0, i = 1, \ldots, N$. Using (39), the population based optimal estimating functions are given by

$$\sum_{i=1}^{N} \frac{(z_i - \alpha_i)(\partial \alpha_i / \partial a)}{E(z_i - \alpha_i)^2} = \sum_{i=1}^{N} (z_i - \alpha_i) x_i \tag{40}$$

$$\sum_{i=1}^{N} \frac{(z_i - \alpha_i)(\partial \alpha_i / \partial b)}{E(z_i - \alpha_i)^2} = \sum_{i=1}^{N} (z_i - \alpha_i) t_i \tag{41}$$

The estimating equations for estimating a and b based on the survey data (4) are obtained from (40) and (41) as

$$\sum_{i \in s} (z_i - \alpha_i) \frac{x_i}{\pi_i} = 0, \quad \sum_{i \in s} (z_i - \alpha_i) \frac{t_i}{\pi_i} = 0 \tag{42}$$

where α_i are as in (39) and π_i are the design inclusion probabilities in (3) (Godambe and Thompson [11]). As noted before, the derivation of equations (42) from (40) and (41) provides a certain kind of protection if the model (39) is not quite appropriate for all individuals i in the population $\mathcal{P}$. If this protection is not considered important, one can obtain possibly more efficient estimating equations by deleting the inclusion probabilities π_i in (42). This of course does not make any difference for a simple random sampling design.

Now if $(\hat{a}, \hat{b})$ solve equations (42), the estimates $\hat{\alpha}_i$ are obtained by replacing (a, b) by $(\hat{a}, \hat{b})$ in (39). As we have seen before, apart from the case when the propensities are constant within strata, $\hat{H}(\hat{\theta}, \hat{\theta}')$ in (21) will generally be a function of the unknown propensities α_i. This unknown function can be

estimated by replacing the α_i in it by their estimates $\hat{\alpha}_i$. Thus we obtain the estimate $\hat{\Delta}$ in (20), very generally.

Note that incorporation of the covariate t, on which y, y′ are assumed not to depend, is important only for the estimation of the $\hat{H}(\hat{\theta}, \hat{\theta}')$ part of $\hat{\Delta}$, and not for the estimation of the θ_j and θ'_j. Thus there is no reason to try to incorporate t in the estimation other than through the $\hat{\alpha}_i$.

7. *Pair Matching*

An illustration of some of the issues in the stratification case is provided by an adaptation of one of the example of Rosenbaum and Rubin [7], namely 'pair matching on balancing scores'. The propensity α_i is a balancing score in their sense if it is a function of covariates explicitly included in the model. We will not assume this, but will nevertheless consider 'pair matching on propensities' as they do.

Since we will be matching on propensities, we imagine a discrete set of propensity values $\alpha_{(r)}$. The stratified population can be considered to be further divided into PSUs (primary sampling units), the rth PSU consisting of individuals with propensity value $\alpha_{(r)}$. Suppose that the rth PSU has M_r individuals M_{r0} of which have $z_i = 0$, and M_{r1} of which have $z_i = 1$. Rosenbaum and Rubin describe a two-step sampling scheme which is close to the following : Within each stratum $\mathcal{P}_j$, select f_j of the PSUs with PSU inclusion probabilities proportional to size; then in each selected PSU, randomly select one individual i_{r1} for which $z_i = 1$, and one individual i_{r0} for which $z_i = 0$. This is an instance where the inclusion probability of an individual depends on its z value :

$$\pi_{i_{r1}} = f_j M_r / N_j M_{r1}$$

and

$$\pi_{i_{r0}} = f_j M_r / N_j M_{r0}$$

If we apply our analysis for the estimation of the parameter Δ of (7), we are led first to the following expression for $\hat{H}$ in (19) :

$$\hat{H}(\theta, \theta') = \sum_j \frac{N_j}{N f_j} \sum_{r \in s_{Ij}} \frac{1}{M_r} \left\{ \frac{M_{r1}(y_{i_{r1}} - \theta_j)}{\alpha_{(r)}} - \frac{M_{r0}(y'_{i_{r0}} - \theta'_j)}{(1 - \alpha_{(r)})} \right\} \quad (43)$$

where s_{Ij} represents the sample PSU labels within the jth stratum.

Now estimate the finite population parameter Δ as in (20) by $\hat{\Theta} + \hat{H}(\hat{\theta}, \hat{\theta}')$, where $\hat{\Theta}$ is given by (15), and $\hat{H}(\hat{\theta}, \hat{\theta}')$ is $\hat{H}(\theta, \theta')$ with θ_j and θ'_j replaced by their estimates, for example from (12).

It is interesting to note that in this example the arguments can be carried through if we replace the propensity $\alpha_{(r)}$ in H and $\hat{H}$ by its natural estimate $\hat{\alpha}_{(r)} = M_{r1} / M_r$. This estimate would be obtained from the considerations in Section 5 from a model like (39) where the covariates were the PSU indicators. Equation (17) is still valid, conditionally on the values of M_{r1} and M_{r0} for all r. With this modified definition of H, whatever the estimates of θ_j and θ'_j, in (20) the estimate of Δ reduces to

$$\hat{\Delta} = \sum_j \frac{N_j}{N} (\bar{y}_j - \bar{y}'_j) \tag{44}$$

where

$$\bar{y}_j = \frac{1}{\ell_j} \sum_{r \in s_{1j}} y_{i_{r1}}$$

and $\bar{y}'_j$ is defined similarly. This is essentially the estimate suggested by Rosenbaum and Rubin [17]. Thus is the case of matched pairs, matched on propensities, we can incorporate estimation of the propensities very naturally at an early stage.

For an alternative justification of replacing $\alpha_{(r)}$ by $\hat{\alpha}_{(r)}$ in H and $\hat{H}$, note that if we let $g = \Delta - \Theta - \hat{H}$ as before,

$$\frac{\partial g}{\partial \alpha_{(r)}} = K_r \left[\frac{M_{r1} (y_{i_{r1}} - \theta_j)}{\alpha^2_{(r)}} - \frac{M_{r0} (y'_{i_{r0}} - \theta'_j)}{(1 - \alpha_{(r)})^2} \right]$$

where K_r is constant, and

$$E\left(\frac{\partial g}{\partial \alpha_{(r)}} \right) = 0$$

Arguments similar to those given at the end of Section 3 can establish approximate optimality of the estimating function g evaluated at $\hat{\alpha}_{(r)}$ for large population sizes N_j and the sample sizes $\ell_j, j = 1, \ldots, k$.

8. *Interval Estimation*

Tests of the hypotheses of interest can be carried out in the usual manner through confidence intervals for the parameters. For the stratification case with large stratum sample sizes n_j, the following methods of interval construction should be helpful. In this section we will take the variates z_i to be independent of one another.

First, for the model parameter θ_j, j fixed, consider the estimating function

$$\sum_{i \in s_j} \left\{ (y_i - \theta_j) \frac{z_i}{\pi_i} \right\}$$

Its mean under the superpopulation model is 0, and its mean square, conditional on **x** and **z**, is

$$\sum_{i \in s_j} v(x^{(j)}) \frac{z_i}{\pi_i^2} \tag{45}$$

Thus an E-unbiased estimate of the mean square is obtained by replacing $v(x^{(j)})$ in the ith term of (45) by $(y_i - \theta_j)^2$. Assuming approximate normality under the model (or perhaps the model and sampling design combined) suggests constructing an interval by inverting

$$\frac{\sum_{i \in s_j} \{ (y_i - \theta_j) z_i / \pi_i \}}{\sqrt{\sum_{i \in s_j} \frac{z_i}{\pi_i^2} (y_i - \theta_j)^2}} = \pm z_{1-\alpha/2} \tag{46}$$

where $z_{1-\alpha/2}$ is the appropriate N(0,1) percentage point. This leads to a quadratic equation in θ_j to solve for the interval end points. (See Binder and Patak [2]). Alternatively, we could invert (46) with θ_j replaced by $\hat{\theta}_j$ in the denominator.

Intervals for θ'_j, j fixed, can be obtained similarly.

Second, suppose the parameter of interest is

$$\Theta = \sum_{j=1}^{k} N_j (\theta_j - \theta'_j) / N$$

In a sense, all but one of the individual θ_j and θ'_j, $j = 1, \ldots, k$ are now nuisance parameters, and an argument akin to one proposed by Binder and Patak [2] (see also Godambe [9]) suggests the use of approximate normality for the pivot

$$\frac{N(\hat{\Theta} - \Theta)}{\left[\sum_{j=1}^{k}\left\{\left(\frac{N_j}{\hat{N}_{jT}}\right)^2 \sum_{i \in s_j}(y_i - \hat{\theta}_j)^2 \frac{z_i}{\pi_i^2} + \left(\frac{N_j}{\hat{N}_{jT'}}\right)^2 \sum_{i \in s_j}(y'_i - \hat{\theta}'_j)^2 \frac{(1-z_i)}{\pi_i^2}\right\}\right]^{1/2}} \tag{47}$$

where $$\hat{N}_{jT} = \sum_{i \in s_j} z_i / \pi_i, \quad \hat{N}_{jT'} = \sum_{i \in s_j} (1 - z_i)/\pi_i$$

Note that (47) is consistent with the pivot in (46), since

$$N(\hat{\Theta} - \Theta) = \sum_{j=1}^{k}\left[\frac{N_j}{\hat{N}_{jT}} \sum_{i \in s_j}\left\{(y_i - \theta_j)\frac{z_i}{\pi_i}\right\} - \frac{N_j}{\hat{N}_{jT'}} \sum_{i \in s_j}\left\{(y'_i - \theta'_j)\frac{(1-z_i)}{\pi_i}\right\}\right]$$

Finally, consider the finite population parameter

$$\Delta = \sum_{i=1}^{N}(y_i - y'_i)/N$$

If $\hat{\Delta}$ is given by (22), then

$$\hat{\Delta} - \Delta = \left[\sum_{j=1}^{k}\left\{\sum_{i \in s_j} A_{is}(y_i - \theta_j) + \sum_{i \in s_j} B_{is}(y'_i - \theta'_j)\right\} - \sum_{j=1}^{k}\left\{\sum_{i \in P_j}(y_i - \theta_j) - \sum_{i \in P_j}(y'_i - \theta'_j)\right\}\right]/N \tag{48}$$

where

$$A_{is} = \frac{z_i}{\pi_i \alpha_i} - \frac{z_i}{\pi_i}\left(\frac{\sum_{i \in s_j} \frac{z_i}{\pi_i \alpha_i}}{\hat{N}_{jT}} - \frac{N_j}{\hat{N}_{jT}}\right)$$

$$B_{is} = \frac{(1-z_i)}{\pi_i(1-\alpha_i)} - \frac{(1-z_i)}{\pi_i}\left(\frac{\sum_{i \in s_j} \frac{(2-z_i)}{\pi_i(1-\alpha_i)}}{\hat{N}_{jT'}} - \frac{N_j}{\hat{N}_{jT'}}\right)$$

For a large population the first terms in (48) will dominate, and intervals and tests for Δ might be based on approximate normality for

$$\frac{\hat{\Delta} - \Delta}{\sqrt{\frac{1}{N^2}\left\{\sum_{j=1}^{k} \hat{v}_j \sum_{i \in s_j} A_{is}^2 + \sum_{j=1}^{k} \hat{v}'_j \sum_{i \in s_j} B_{is}^2\right\}}} \tag{49}$$

where $\hat{v}_j$ and $\hat{v}'_j$ are sample based estimates of $E\{(y_i - \theta_j)^2 \mid \mathbf{x}\}$ and $E\{(y'_i - \theta'_j)^2 \mid \mathbf{x}\}$, $i \in \mathcal{P}_j$, respectively.

9. *Confounding*

A closer look at the pair matching design discussed in Section 7 suggests some interesting relationships among the basic ideas of randomization, optimal estimating functions, ignorability of treatment assignments (see Section 1) and the related concept of confounding.

Consider a rather extreme case of confounding, namely where the propensities α_i in (1) are functions of y_i, y'_i, so that

$$\alpha_i = \alpha_i(y_i, y'_i), \quad i = 1, \ldots, N \tag{50}$$

Under the dependence of α_i on (y_i, y'_i) as in (50), the assumption of ignorability of treatment assignments is no longer valid. As often the propensities α_i will be unknown, but here we will continue with the assumption made in Section 7, that the α_i take only discrete values, and that individuals i with a common value of the propensity $\alpha_{(r)}$ can be identified.

Now suppose (50) holds, and therefore that the assumption of 'no confounding' is not true, yet $0 < \alpha_i < 1$, $i = 1, \ldots, N$. Then some of the previous results still hold. In (43) the design expectation $E(\hat{H}) = H$; further, for large N, $H \cong \Delta - \Theta$ as in (18). Obviously, also, the substitution of the unknown $\alpha_{(r)}$ by its estimate M_{r1} / M_r is still natural. To this extent the estimation of $\hat{\Delta}$ given by (44) is justified even under a possible confounding such as (50). However, in what follows one will see that the estimating function $\hat{H}$ in (43), unlike the corresponding one in (19), is not optimal for H.

In fact, this is partly because the underlying sampling design depends on the treatment allocation vector **z**. Before demonstrating this, however, we briefly discuss the significance of 'optimality' in the present context.

For a general survey sampling setup, a definition of optimal estimating function is given by Godambe and Thompson [11]. Suppose that according to

this definition, in the class of all design unbiased estimating functions for H a candidate for the 'optimal' one is given by $\hat{H}$. A sufficient condition for this optimality is that if any estimating function $\delta = \delta\{(i,y_i),\ i \in s, \theta, \theta'\}$ has design expectation zero, then δ is *uncorrelated* with $\hat{H}$, with respect to the sampling design and the underlying superpopulation semiparametric model. That is, the estimating function $\hat{H}$ mimics for a semiparametric model an important property of a complete sufficient statistic for a parametric model, namely that the statistic is independent of every ancillary statistic, (Godambe [8], Lehman [15]). In this sense the estimation based on the optimal estimating function may be said to be utilizing all the information in the sample.

To study the conditions under which the correlation $E\,E(\hat{H}\,\delta) = 0$, we note from (19) that

$$\hat{H} = \sum_{i \in s} \frac{\psi_i}{\pi_i} \text{ where } \psi_i = \frac{y_i - \theta_i}{\alpha_i} z_i - \frac{y'_i - \theta'_i}{1 - \alpha_i}(1 - z_i)$$

Hence, using $E(\delta) = 0$, we have

$$E\,E(\hat{H}\,\delta) \mid \mathbf{y}, \mathbf{y}') = -\sum_{i=1}^{N} E\left\{ \frac{\psi_i}{\pi_i} \sum_{s: i \notin s} (\delta_p) \mid \mathbf{y}, \mathbf{y}' \right\} \qquad (51)$$

where p is the probability of sample s as in (2). Now if in (51) the z_i are independent and the sampling design p is independent (i) of the treatment allocation vector z and (ii) of the response vectors $(\mathbf{y}, \mathbf{y}')$, $E\,E(H\,\delta) = 0$. Otherwise in general $E\,E(\hat{H}\,\delta) \neq 0$. Apart from the 'pair matching' design discussed in Section 7, and the paired unbiased allocation designs of Sections 4 and 5, all the conditions above are satisfied for the examples considered in this paper. For the pair matching design, however, neither of the conditions (i) or (ii) is satisfied.

It is interesting to note that in the present case, estimation for Δ similar to one given by (44) can be obtained by a sampling design satisfying conditions (i) and (ii) above. This estimation is 'optimal', however, only for large samples, or in other words *asymptotically*.

Consider a stratified simple random sampling design with strata as in Section 7 and stratum sample sizes $|s_j| = n_j,\ j = 1, \ldots, k,\ s = \cup s_j$ as before being the total sample. Now let

$$s_{j(r)} = \{ i : i \in s_j \text{ and } \alpha_i = \alpha_{(r)} \},\ |s_{j(r)}| = n_{jr}$$

where $\alpha_{(r)}$ are the same as in Section 7. Then

$$\hat{H} = \frac{1}{N} \sum_{j} \sum_{(r)} \sum_{i \in s_{j(r)}} \left\{ \frac{y_i - \theta_j}{\alpha_{(r)}} z_i - \frac{y'_i - \theta'_j}{1 - \alpha_{(r)}} (1 - z_i) \right\} / \left(\frac{n_j}{N_j} \right) \qquad (52)$$

Further if the sample sizes n_j are large compared to the distinct number of propensities and if the number of individuals i in $s_{j(r)}$ with $z_i = 1$ is $n'_{j(r)}$, the natural estimate of $\alpha_{(r)}$ is given by $\hat{\alpha}_{(r)} = n'_{j(r)} / n_{j(r)}$. Substituting $\hat{\alpha}_{(r)}$ for $\alpha_{(r)}$ in (52) we have

$$\tilde{H} = \frac{1}{N} \sum_{j} \sum_{(r)} n_{j(r)} \left\{ (\bar{y}_{j(r)} - \theta_j) - (\bar{y}'_{j(r)} - \theta'_j) \right\} / \left(\frac{n_j}{N_j} \right) \qquad (53)$$

where $\bar{y}_{j(r)}$ is the mean of all y_i for $i \in s_{j(r)}$ having $z_i = 1$, and $\bar{y}'_{j(r)}$ is defined similarly. From (53) we have

$$\tilde{H} = \frac{1}{N} \sum_{j} \frac{N_j}{n_j} \sum_{(r)} n_{j(r)} \{ \bar{y}_{j(r)} - \bar{y}'_{j(r)} \} - \frac{1}{N} \sum_{j} N_j (\theta_j - \theta'_j)$$

providing the estimate of Δ as

$$\tilde{\Delta} = \frac{1}{N} \sum_{j} \frac{N_j}{n_j} \sum_{(r)} n_{j(r)} \{ \bar{y}_{j(r)} - \bar{y}'_{j(r)} \} \qquad (54)$$

It is interesting to compare the two estimates of Δ, one gives by $\hat{\Delta}$ in (44) and the other $\tilde{\Delta}$ in (54). For large samples the estimate $\tilde{\Delta}$ is approximately optimal *regardless* of the assumption of 'no confounding'. Though no such optimality is available for the estimate $\tilde{\Delta}$, it seems natural, granting no confounding, to pool together estimates $\bar{y}_{j(r)}$ corresponding to different values of the propensities $\alpha_{(r)}$ to get what possibly is a more efficient estimate of Δ, namely $\hat{\Delta}$. Perhaps the difference $\hat{\Delta} - \tilde{\Delta}$ can provide a test statistic for testing the hypothesis of 'no confounding'. Yet a more direct test of 'confounding' is provided as follows.

Under the assumption of no confounding we have $E(y_i) = \theta_j$, $E(y'_i) = \theta'_j$ for the individual $i \in \mathcal{P}_j$. However, if there is confounding as given by (50), then for $\alpha_i = \alpha_{(r)}$, $E(y_i) = \theta_{j(r)}$ and $E(y'_i) = \theta'_{j(r)}$ where $\theta_{j(r)}$ and $\theta'_{j(r)}$ are determined by the underlying model and $\alpha_{(r)}$. Here no confounding would be expressed by the null hypothesis,

$$H_0 : \theta_{j(r)} = \theta_j \; : \; \theta'_{j(r)} = \theta'_j \text{ for all j and r}$$

Further, if under the null hypothesis and the model with $\alpha_i = \alpha_{(r)}$ fixed the variates (y_i, y'_i) are independent and normally distributed with known variances σ^2, the likelihood ratio test for H_0 is given by

$$\sum_j \frac{1}{\sigma^2} \sum_r \{ n'_{j(r)} (\bar{y}_{j(r)} - \bar{y}_j)^2 + n^0_{j(r)} (\bar{y}'_{(r)} - \bar{y}'_j)^2 \} \cong \chi^2 \{ 2 \sum_j (f_j - 1) \} \tag{55}$$

$n^0_{j(r)}$ being $n_{j(r)} - n'_{j(r)}$. If in (55) the variance σ^2 is unknown and to be estimated, we can construct an F test in the usual manner.

It is interesting to compare the estimates of Δ given by (23), (44) and (54), in conjunction with their corresponding sampling designs. Underlying (23), in the context of simple random sampling, is the assumption that the propensities depend exclusively on the covariates, and as a result, in (21) $\hat{H} = 0$. Thus from (20), $\hat{\Delta} = \hat{\Theta}$. That is, the estimate (23) of Δ is obtained *indirectly* from the estimate $\hat{\Theta}$. On the other hand the estimates in (44) and (54), in their respective contexts of pair matching and a kind of post stratification on propensity *classes*, do not require estimation of Θ at all : in (19),

$$\hat{H}(\theta, \theta') = [\text{estimate (44) or (54)}] - \Theta$$

which with (18) provides directly the corresponding estimate of Δ. Thus estimation in (44) and (54) is less dependent on the superpopulation model of Section 1 then is the estimation in (23). Particularly, the 'variance' assumption of the superpopulation model, which plays a crucial role in the estimation of Θ by $\hat{\Theta}$, is not required for derivation of estimates (44) and (54). Also, unlike the derivation of (23), the derivations of (44) and (54) do not require the assumption that the propensities are uniquely determined by the covariates. In this sense the properties of the estimates (44), (54) are more robust than those of the estimate in (23). Of course when the above mentioned 'assumptions' are satisfied the estimate (23) will be more efficient than that given by (54); it will actually be approximately optimal. However, the asymptotic optimality of the estimate (54), in a smaller class than that with which (23) is compared, is insensitive to some departures from the assumption of 'no confounding', as discussed in the preceding paragraphs. This in some situations could be the crucial point in support of the strategy leading to the estimate (54).

ACKNOWLEDGEMENT

The authors would like to thank D.R. Cox, J.N.K. Rao, B. Sutradhar, and K.R. Shah for helpful comments.

REFERENCES

[1] Bell, J.S., 1964. On Einstein-Podolsky-Rosen paradox. *Physica*, **1**, 195-200.

[2] Binder, D.A. and Patak, Z., 1994. Use of estimating functions for estimation from complex surveys. *J. Amer. Statist. Assoc.*, **89**, 1035-1043.

[3] Cox, D.R., 1958. *The Planning of Experiments*. John Wiley and Sons, New York.

[4] Cox, D.R., 1992. Causality : some statistical aspects. *J. Roy. Statist. Soc.*, **A155**, 281-301.

[5] Fisher, R.A., 1926. The arrangement of field experiments. *J. Ministry Agri.*, **33**, 503-513.

[6] Fisher, R.A., 1935. *The Design of Experiments*. Oliver and Boyd, Edinbugh.

[7] Fisher, R.A., 1958. Cigarettes, cancer and statistics. *The Centennial Review*, **2**, 151-166.

[8] Godambe, V.P., 1966. A new approach to sampling from finite populations I : sufficiency and linear estimation. *J. Roy. Statist. Soc.*, **B28**, 310-319.

[9] Godambe, V.P., 1991. Orthogonality of estimating functions and nuisance parameters. *Biometrika*, **78**, 143-151.

[10] Godambe, V.P. and Kunte, S., 1993. Optimum estimation under biased allocation of treatments. *Biometrika*, **80**, 797-806.

[11] Godambe, V.P. and Thompson, M.E., 1986. Parameters of superpopulation and survey population : their relationships and estimation. *Internat Statist. Rev.*, **54**, 127-138.

[12] Godambe, V.P. and Thompson, M.E., 1989. An extension of quasi-likelihood estimation (with discussion). *J. Statist. Plann. Inf.*, **22**, 137-172.

[13] Holland, P.W., 1986. Statistics and causal inference (with discussion). *J. Amer. Statist. Assoc.*, **81**, 945-970.

[14] Kempthorne, O., 1952. *The Design and Analysis of Experiments*. John Wiley and Sons, New York.

[15] Lehman, E.L., 1981. An interpretation of completeness and Basu's theorem. *J. Amer. Statist. Assoc.*, **76**, 335-340.

[16] Robbins, H. and Zhang, C.H., 1991. Estimating a multiplicative treatment effect under biased allocations. *Biometrika*, **78**, 349-354.

[17] Rosenbaum, P.R. and Rubin, D.B., 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika*, **70**, 41-55.

[18] Rubin, D.B., 1980. Discussion of "Randomization analysis of experimental data : The Fisher randomization test," by D. Basu. *J. Amer. Statist. Assoc.*, **75**, 591-593.

[19] Sprott, D.A. and Farewell, V.T., 1993. Randomization in experimental science. *Statistics Papers*, **34**, 89-94.

[20] Stone, Richard, 1993. Assumptions on which causal inference rests. *J. Roy. Statist. Soc.*, **B55**, 455-466.