

## A New Look at Confidence Intervals in Survey Sampling

V.P. GODAMBE<sup>1</sup>

### ABSTRACT

In survey sampling, as in other areas of statistics conventionally, confidence intervals for a parameter are often obtained by inverting the distribution of some approximate pivotal quantity,  $\{(\text{estimate} - \text{parameter}) / (\text{estimated variance})^{1/2}\}$ . Alternatively, estimating function theory suggests a more direct method of constructing pivotal quantity and hence confidence intervals. These alternative confidence intervals perform much better than the conventional ones, in many simulation studies.

KEY WORDS: Confidence Intervals; Estimating Functions; Optimality; Stratification; Survey Sampling.

### RÉSUMÉ

En échantillonnage d'enquête, tout comme c'est habituellement le cas dans d'autres domaines de la statistique, les intervalles de confiance pour un paramètre sont souvent construits en inversant la distribution d'une quantité pivotale approximative,  $\{(\text{valeur estimée-paramètre}) / (\text{variance estimée})^{1/2}\}$ . La théorie des fonctions d'estimation suggère une méthode alternative, plus directe, pour construire la quantité pivotale et, par conséquent, les intervalles de confiance. Ces nouveaux intervalles de confiance se comportent beaucoup mieux que les intervalles traditionnels, dans plusieurs études basées sur des simulations.

MOTS-CLÉS: Intervalles de confiance; fonctions d'estimation; optimalité; stratification; échantillonnage d'enquête.

### 1. HISTORICAL INTRODUCTION

The topic of confidence intervals was first discussed in Neyman's (1934) well-known paper read before the Royal Statistical Society. The paper was on survey sampling. Yet the discussion is nowhere near actual construction of confidence intervals for a survey sampling setup. This possibly could be due to the fact that at the time the distinction between the parameters of a *survey population* on one hand and a *hypothetical population* on the other was far from clearly understood, (Deming, 1950, Godambe, 1976; 1997, Smith, 1997). From hindsight, one can say that Neyman's discussion of confidence intervals referred above relates primarily to the parameters of a hypothetical population. A subsequent publication of Neyman (1937) explicitly demonstrates how confidence intervals could be obtained from a pivotal quantity: a function of observations and the

parameter of interest having a fixed (known) distribution. The availability of such 'pivotal' for some hypothetical populations (characterized completely by a few scalar parameters) can be easily demonstrated. On the other hand to characterize a survey population of size  $N$  one needs a parameter of  $N$ -dimensions, (Basu, 1957, Hájek, 1959). Under the condition, in general no pivotal, that is a function of observations and the parameter of interest, having a fixed distribution can exist, barring trivial cases.

The section VI of Neyman's 1934 paper is entitled Appendix. In addition to other things, the Appendix contains Note I, dealing with confidence intervals followed by Note II. 'The Markoff Method and Markoff Theorem on Least Squares'. The 'Theorem' mentioned here using modern terminology, is the Gauss-Markoff theorem on unbiased minimum variance estimation. Now suppose the unbi-

<sup>1</sup>Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1; vpgodamb@math.uwaterloo.ca

ased minimum variance estimate is approximately normally distributed. Then the inversion of the distribution of the approximate pivotal quantity  $\{(\text{estimate} - \text{parameter}) / (\text{variance})^{1/2}\}$  would provide shortest confidence intervals; that is, assuming the *variance* is known. This however is of no avail, for the ‘variance’ just mentioned is never known in a survey sampling setup. Common practice, as seen from the publications, (cf. Chaudhuri and Vos, 1988), on the subject, is to substitute an ‘estimate’ for the unknown variance. Here the basic question, generally not discussed in the literature is, which of the ‘many estimates’ of the variance would provide a pivot (or approximately so) leading to a set of plausible confidence intervals? This problem for a hypothetical population with an underlying parametric model, is resolved, utilizing the concept of *observed Fisher information* (Efron and Hinkley, 1978). A generalization of the observed Fisher information, for a semiparametric model, provided by the *theory of optimal estimating functions*, (Godambe 1985, Godambe and Thompson 1986) suggests a more *direct approach* to ‘confidence intervals’ within the context of survey sampling.

Following the just suggested ‘direct approach’, confidence intervals are constructed, for different sampling designs and under different superpopulation assumptions in Sections 2, 3 and 5. These confidence intervals are compared with the *conventional* ones, based on approximate pivotal,  $\{(\text{estimate} - \text{parameter}) / (\text{estimated variance})^{1/2}\}$ , with extensive simulation experiments, in Section 6. The results seem to be definitely in favour of the former confidence intervals, (see Section 7).

So far the topic of ‘confidence intervals in survey sampling’ was dealt with within the framework of ‘estimating functions’ only by a couple of authors. Historically, Woodruff (1952), presented an earliest demonstration of confidence intervals for ‘position measures’ of a survey population, utilizing informally estimating functions. This has been commented on in detail by Godambe, (1991). The second author is Binder (1994). The author, in an earlier publication (Binder, 1983) also makes informal use of estimating functions for complex surveys. There, however, the confidence intervals presented are more of conventional type. Now, both the papers, Binder (1994) and the present paper, are based on the theory of estimating functions. Yet

the basic difference between the two is this: Unlike the former, the latter, in an essential manner is tied to the *optimality* criterion of estimating functions. This ‘optimality’ criterion *relates* the present survey populations to a semiparametric superpopulation model, albeit very flexibly. This ‘relationship’, as the present paper demonstrates, provides a guidance, (more specific than one provided by the intuition) as to which estimating function and the implied confidence intervals, are to be used for a given problem. Barring this reference to a superpopulation model, the confidence intervals, presented in this paper are *design-based*. Again, both papers, Binder (1994) and the present one (Section 5) discuss the important case of ‘nuisance parameters’. However the problems treated in the two papers are different and there is no overlap in the results.

## 2. STRATIFIED SIMPLE RANDOM SAMPLING

Here to construct confidence intervals we follow the usual notation. The labelled population of  $N$  individuals (units) is denoted by  $\mathcal{P} = \{i : i = 1, \dots, N\}$ . The population  $\mathcal{P}$  is divided into  $k$  nonoverlapping strata  $\mathcal{P}_j$  of sizes  $N_j, j = 1, \dots, k$ . A variate of study defined for the population  $\mathcal{P}$  is  $y$ , assumed to be scalar for simplicity. For the individual  $i$ ,  $y = y_i, i = 1, \dots, N$ . The population vector  $\mathbf{y} = (y_1, \dots, y_N)$ . To obtain an estimate for the population mean  $\bar{Y} = \sum_1^N y_i / N$  a sample  $s$ , ( $s \subset \mathcal{P}$ ) of size  $n$ ,  $|s| = n$  is drawn from  $\mathcal{P}$ , with a stratified simple random sampling without replacement design. The samples from different strata  $\mathcal{P}_j$  are denoted by  $s_j, |s_j| = n_j, j = 1, \dots, k; \sum n_j = n$ . Further  $\bar{Y}_j$  and  $\bar{y}_j$  denote the stratum  $\mathcal{P}_j$  and sample  $s_j$  means of  $y$  respectively,  $j = 1, \dots, k$ . Thus

$$\bar{Y} = \sum_{j=1}^k N_j \bar{Y}_j / N . \quad (1)$$

Analogously we define

$$\bar{y} = \sum_{j=1}^k N_j \bar{y}_j / N . \quad (2)$$

If the components of the population vector  $\mathbf{y} = (y_1, \dots, y_N)$  are assumed to have been drawn independently from superpopulations with a common

mean  $\mathcal{E}(y_i) = \theta, i = 1, \dots, N$  but possibly different variances  $\mathcal{E}(y_i - \theta)^2, i = 1, \dots, N$  an (approximately) *optimal estimating function* for estimating the population mean  $\bar{Y}$  in (1) is given by

$$g = \sum_{j=1}^k \frac{N_j}{N} \cdot \frac{1}{n_j} \sum_{i \in s_j} (y_i - \bar{Y}) \quad (3)$$

(Godambe and Thompson, 1986; Godambe, 1995). Thus the optimal estimate of  $\bar{Y}$  is given by  $\bar{y}$ , the solution of the equation  $g = 0$ . We will call the superpopulation models defined only by a first few moments, as the model just mentioned namely  $\mathcal{E}(y_i) = \theta, i = 1, \dots, N$ , *semiparametric* in contrast to the *fully parametric models* specified by the density functions.

The 'optimum estimating function' for a semi-parametric model has many statistically important properties in common with the 'score function' for a parametric model. Hence in the former situation the 'optimum estimating function' is called a *quasi-score function*. (Godambe, 1985; Godambe and Heyde, 1987; Godambe and Thompson, 1989). For a parametric model, one can construct 'confidence intervals' using Fisher information, (defined as the variance of the score-function) or its natural estimate the observed Fisher information. Similarly in case of a semi-parametric model the confidence intervals can be obtained from the quasi-score function that is the optimum estimating function and its estimated variance. Now though this 'optimality' is tied to the superpopulation model  $\mathcal{E}(y_i) = \theta$ , the 'properties' of the confidence intervals given below are mostly if not entirely 'design-based'. The design-based variance of the optimum estimating function  $g$  in (3) is given by

$$V(g) = \sum_{j=1}^k \frac{N_j^2}{N^2} \left( \frac{1}{n_j} - \frac{1}{N_j} \right) \frac{1}{N_j - 1} \sum_{i \in \mathcal{P}_j} (y_i - \bar{Y}_j)^2. \quad (4)$$

Further since our parameter of interest is  $\bar{Y}$ , in (4), the unobserved  $y_i$ 's and the stratum means  $\bar{Y}_j$  are nuisance parameters. The superpopulation model underlying the estimating function  $g$  in (3) namely  $\mathcal{E}(y_i) = \theta, i = 1, \dots, N$ , suggests for large strata sizes  $N_j$ , ignoring the differences  $\bar{Y}_j - \bar{Y}$ , and replacing  $\bar{Y}$  for  $\bar{Y}_j, j = 1, \dots, k$  in (4).

(Note 1. The models, for which the differences  $\bar{Y}_j - \bar{Y}$  cannot be ignored, are treated in Section 5.)

With this replacement, the estimate of variance  $V(g)$  in (4) is given by,

$$\begin{aligned} \hat{V}(g) &= \sum_{j=1}^k \frac{N_j^2}{N^2} \left( \frac{1}{n_j} - \frac{1}{N_j} \right) \cdot \frac{N_j}{(N_j - 1)} \\ &\quad \cdot \frac{1}{n_j} \sum_{i \in s_j} (y_i - \bar{Y})^2 \\ &= \left\{ \sum_{j=1}^k \frac{N_j^2}{N^2} \left( \frac{1}{n_j} - \frac{1}{N_j} \right) \right. \\ &\quad \left. \cdot \frac{1}{(n_j - 1)} \sum_{i \in s_j} (y_i - \bar{y}_j)^2 \right\} + R \quad (5) \end{aligned}$$

where  $\bar{y}_j$  is the mean of the sample  $s_j, j = 1, \dots, k$  as in (2). In the right hand side of equation (5) the first term is of the  $O(1/n_j)$  while second term  $R$  is of  $O(1/n_j^2)$ . Hence for large samples ignoring the term  $R$ ,  $\hat{V}$  in (5) reduces to the conventional estimate say  $\hat{V}_0$ . This leads to the conventional confidence intervals for  $\bar{Y}$  based on the inversion of the distribution  $\{g/(\hat{V}_0)^{1/2}\}$ . However when sample sizes  $n_j, j = 1, \dots, k$  are not very large the estimating function theory suggests confidence intervals of  $\bar{Y}$ , based on the asymptotic distribution of  $\{g/(\hat{V})^{1/2}\}$ , namely  $N(0, 1)$ :

In case of a simple random sampling from  $N(\mu, \sigma^2)$  population the superiority of the corresponding confidence intervals based on the normal approximation to the distribution of  $\{g/(\hat{V})^{1/2}\}$  to those based on the normal approximation to the distribution of  $\{g/(\hat{V}_0)^{1/2}\}$  is indicated by the analysis of Mach (1988). For a stratified simple random sampling design, let in (3) the estimating function  $g$  be written as

$$g = \sum_{i \in s} g_i, \quad (6)$$

where as before  $s$  denotes the sample (of individuals drawn from all strata). Then, for large  $n_j$  and  $N_j$ , ignoring the finite sample correction  $(1/N_j)$ , and replacing  $(n_j - 1)$  by  $n_j, j = 1, \dots, k$  in (5), we have

$$\hat{V}(g) \simeq \hat{V}_a(g) = \sum_{i \in s} g_i^2. \quad (7)$$

Now it is easy to see that in view of simple random sampling from each stratum (again for reasonably large  $n_j$  and  $N_j, j = 1, \dots, k$ ) the sampling distribution and the superpopulation distribution of the quantity  $\{g/(\hat{V}_a)^{1/2}\}$  would tend to

be the same namely  $N(0, 1)$ . We have already identified the optimum estimating function  $g$  with the quasi-score function. Further just as for a parametric model the inversion of the distribution of  $\{\text{score function}/(\text{observed Fisher information})^{1/2}\}$  provides asymptotically the shortest confidence intervals, for the semi-parametric model  $\{g/(\hat{V}_a)^{1/2}\}$  provides asymptotically shortest confidence intervals, (Wilks, 1938; Godambe and Heyde, 1987).

The above analysis can be easily extended to include a covariate. Suppose for the population  $\mathcal{P} = \{i : i = 1, \dots, N\}$  in addition to the variate  $y$  under study is defined a covariate  $x$ , again for simplicity assumed to be a scalar like  $y$ . For the individual  $i$ ,  $x = x_i$  is known,  $i = 1, \dots, N$ . Now the superpopulation model  $\mathcal{E}(y_i - \theta) = 0$ , underlying the forgoing discussion is extended to  $\mathcal{E}(y_i - \theta x_i) = 0, i = 1, \dots, N$ . Along the lines of (1) and (2) we define

$$\bar{X} = \sum_{j=1}^k N_j \bar{X}_j / N \quad (8)$$

and

$$\bar{x} = \sum_{j=1}^k N_j \bar{x}_j / N \quad (9)$$

where  $\bar{X}_j$  and  $\bar{x}_j$  are stratum  $\mathcal{P}_j$  and sample  $s_j$  means of  $x$  respectively. The optimal estimating function  $g$  in (3) is now replaced by

$$g = \sum_{j=1}^k \frac{N_j}{N} \cdot \frac{1}{n_j} \sum_{i \in s_j} \left( y_i - \frac{\bar{Y}}{\bar{X}} x_i \right). \quad (10)$$

As before the solution of the equation  $g = 0$ , provides the optimal (or approximately so), estimate for  $\bar{Y}$ . Again the superpopulation model  $\mathcal{E}(y_i - \theta x_i) = 0, i = 1, \dots, N$ , for large strata sizes  $N_j$  suggests, taking  $\frac{\bar{Y}_j}{\bar{X}_j} = \frac{\bar{Y}}{\bar{X}}$  and ignoring the differences  $\bar{Y}_j - \frac{\bar{Y}}{\bar{X}} \bar{X}_j, j = 1, \dots, k$ . This leads to the following estimate of variance of  $g$  in (10):

$$\hat{V}(g) = \sum_{j=1}^k \frac{N_j^2}{N^2} \left( \frac{1}{n_j} - \frac{1}{N_j} \right) \cdot \frac{N_j}{(N_j - 1) n_j} \sum_{i \in s_j} \left( y_i - \frac{\bar{Y}}{\bar{X}} x_i \right)^2. \quad (11)$$

(Note 2. The models, for which the differences  $\bar{Y}_j - (\bar{Y}/\bar{X})\bar{X}_j$  cannot be ignored are treated in Section 5.)

The equation (11) reduces to (5) if  $x_i = \text{constant}, i = 1, \dots, N$ . Again the confidence intervals for  $\bar{Y}$ , according to the estimating function theory, can be obtained by inversion of the sampling distribution of the (approximate) pivot  $g/\{\hat{V}(g)\}^{1/2}$ ; the distribution asymptotically is  $N(0, 1)$ .

### 3. STRATIFIED CLUSTER SAMPLING

In this section we assume the whole population of individuals (units) is divided as before into a number of nonoverlapping strata. But now in addition, each stratum is divided into a number of nonoverlapping clusters of individuals. The first stage sampling consists of drawing from each stratum a small number of clusters with *simple random sampling*. Next, from each selected cluster a sample of (ultimate) individuals is drawn possibly with a multistage 'sampling design'. This sampling design is 'specific' to the 'cluster' and does *not* depend on what other clusters have been selected at the first stage of selection.

To accommodate the above situation in our framework we use the following extension of the previous notation. As before  $i$  denotes the 'individual'. A 'cluster' is denoted by  $c$ ; that is  $i \in c$ . The elements of strata  $\mathcal{P}_j, j = 1, \dots, k$  are now clusters  $c$ ; the stratum  $\mathcal{P}_j$  consists of  $N_j$  clusters,  $j = 1, \dots, k$ . A sample of individuals from the cluster  $c$  is denoted by  $s^c$  and the set of clusters selected from the stratum  $\mathcal{P}_j$  is denoted by  $s_j, |s_j| = n_j$  and  $|\mathcal{P}_j| = N_j, j = 1, \dots, k$ . Otherwise we use the same notation as before. Again the superpopulation model as before is  $\mathcal{E}(y_i - \theta x_i) = 0$  for all individuals  $i$ , in the population.

Now suppose the sampling design for the cluster  $c$  is such that, (once the cluster is selected) the probability of including an individual  $i$  in the sample, that is  $Prob(i \in s^c | c) = \pi'_i, i \in c$ . Hence if the cluster  $i \in \mathcal{P}_j$ , the unconditional (inclusion)  $Prob(i) = \pi'_i (n_j/N_j), j = 1, \dots, k$ . Thus if the population totals of  $y$  and  $x$  are denoted by  $Y$  and  $X$  respectively, the optimal estimating function for  $Y$  or  $(Y/X)$ , with respect to the superpopulation model just mentioned is given by replacing in (10)  $g$  by

$$g = \sum_{j=1}^k \frac{N_j}{n_j} \sum_{c \in s_j} \sum_{i \in s^c} \left\{ \frac{y_i - \left( \frac{Y}{X} \right) x_i}{\pi'_i} \right\}. \quad (12)$$

Further if  $Y_c$  and  $X_c$  denote the cluster totals of  $y$  and  $x$  respectively the optimal estimating function (Godambe, 1995) for estimating  $Y_c$  or  $(Y_c/X_c)$  is obtained from (12) as,

$$g_c = \sum_{i \in s^c} \left\{ \frac{(y_i - \frac{Y_c}{X_c} x_i)}{\pi_i'} \right\}. \quad (13)$$

Now we assume that for each cluster  $c$ , the sampling design is *calibrated*; that is for each sample  $s^c$  of non zero selection probability,

$$\sum_{i \in s^c} \frac{x_i}{\pi_i'} = X_c.$$

For such calibrated sampling designs, if  $\hat{Y}_c$  denotes the estimate of  $Y_c$  obtained from the equation  $g_c = 0$ , where  $g_c$  is as in (13), then from (12) we have

$$g = \frac{1}{N} \sum_{j=1}^k \frac{N_j}{n_j} \sum_{c \in s_j} \left\{ \hat{Y}_c - \left( \frac{Y}{X} \right) X_c \right\}. \quad (14)$$

With a fairly straightforward algebra it can be shown that the variance of  $g$  in (14),

$$V(g) = E \left\{ \frac{1}{N^2} \sum_{j=1}^k \frac{N_j^2}{n_j^2} \sum_{c \in s_j} \left( \hat{Y}_c - \frac{Y}{X} X_c \right)^2 \right\} + 0 \left( \frac{1}{N} \right).$$

Hence if all strata sizes  $N_j$  are large enough we have

$$V(g) \simeq E \left\{ \frac{1}{N^2} \sum_{j=1}^k \frac{N_j^2}{n_j^2} \sum_{c \in s_j} \left( \hat{Y}_c - \frac{Y}{X} X_c \right)^2 \right\}. \quad (15)$$

A natural estimate of the variance  $V(g)$  in (15) is given by

$$\hat{V}(g) \simeq \frac{1}{N^2} \sum_{j=1}^k \frac{N_j^2}{n_j^2} \sum_{c \in s_j} \left( \hat{Y}_c - \frac{Y}{X} X_c \right)^2; \quad (16)$$

the confidence intervals for  $(Y/X)$ ,  $(\bar{Y}/\bar{X})$  or  $\bar{Y}$  can be obtained as before, by inverting the sampling distribution of the approximate pivot  $g/\sqrt{\{\hat{V}(g)\}}$ ; asymptotically the distribution is  $N(0,1)$ .

The confidence intervals discussed above do not require the knowledge of sampling design for any cluster, provided at the cluster level the estimates  $\hat{Y}_c$  of  $Y_c$  are available for  $c \in s_j, j = 1, \dots, k$ . These

confidence intervals, though valid, cannot be expected to be as efficient as the ones based on the entire data, if and when available.

It is important to distinguish the estimates  $\hat{V}(g)$  in (5), (11) and (16), (of the variances  $V(g)$  of the estimating function  $g$ ), from the conventional estimates (of the variances of estimates). The former generally in an essential way contain the parameter of interest  $Y$  or  $\bar{Y}$ . The latter by definition must be free of the parameter. The distribution of  $g/\sqrt{\{\hat{V}(g)\}}$  would generally tend to its limit faster than the corresponding distribution of

$$(\hat{Y} - \bar{Y})/\{\text{estimate of the variance } \hat{Y}\}^{1/2}. \quad (17)$$

For unlike the  $\{\text{estimate of the variance } \hat{Y}\}$  in (17),  $\hat{V}(g)$  would be a sum of independently distributed random variates, and would be stabler. An unpublished result due to M.E. Thompson (1997) suggests that the distribution of  $\hat{V}$  in (16) would be closer to normal than that of the quantity in (17). This is also supported by many simulations reported in Section 6.

The estimate  $\hat{V}$  in (16) depends on the sample variates only through the estimates of the cluster totals or means; a property also shared by the *traditional* 'estimate of the variance  $\hat{Y}$ ' in (17). In connection with the latter, early references could be traced back to Mahalanobis' interpenetrating samples, in the thirties while some of the recent ones are Sarndal, Swensson and Wretman (1992); Yung and Rao (1996).

#### 4. BOOTSTRAP

In this section we present bootstrap versions of the estimates of the variance  $\hat{V}(g)$  given in (5), (11) and (16). We illustrate the method in case of (11) in some detail; the estimates (5) and (16) could be obtained as special cases.

Our bootstrap method is different than usual in the sense that we obtain the bootstrap variance of the estimating function  $g$  in (10), holding the parameter value  $(\bar{Y}/\bar{X})$  in it *fixed*. As before our data consists of  $(y_i, x_i) : i \in s_j, j = 1, \dots, k$ . The stratified resampling is done as follows. The  $n_j$  number of draws are made with replacement from  $(y_i, x_i) : i \in s_j, j = 1, \dots, k$ . If  $q$  denotes a generic draw, a generic bootstrap value of the estimating

function of  $g$  in (10),  $g_b$  say, is given by

$$g_b = \sum_{j=1}^k \frac{N_j}{N} \cdot \frac{1}{n_j} \sum_{q=1}^{n_j} \left( y_q - \frac{\bar{Y}}{\bar{X}} \cdot x_q \right). \quad (18)$$

Denoting by  $E_B$  and  $V_B$  the bootstrap expectation and variance respectively, we have

$$E_B(g_b) = g.$$

And

$$V_B(g_b) = E_B(g_b^2) - \{E_B(g_b)\}^2 = E_B(g_b^2) - g^2. \quad (19)$$

In (19),

$$E_B(g_b^2) = A + B + C \quad (20)$$

where

$$\begin{aligned} A &= \sum_{j=1}^k \frac{N_j^2}{N^2} \frac{1}{n_j^2} \sum_{q=1}^{n_j} E_B \left( y_q - \frac{\bar{Y}}{\bar{X}} x_q \right)^2 \\ &= \sum_{j=1}^k \frac{N_j^2}{N^2} \cdot \frac{1}{n_j^2} \sum_{i \in s_j} \left( y_i - \frac{\bar{Y}}{\bar{X}} \cdot x_i \right)^2. \\ B &= \sum_{j=1}^k \frac{N_j^2}{N^2} \cdot \frac{1}{n_j^2} \sum_{\substack{q, q' = 1 \\ \text{stratum } j}}^{n_j} E_B \left( y_q - \frac{\bar{Y}}{\bar{X}} x_q \right) \\ &\quad \cdot \left( y_{q'} - \frac{\bar{Y}}{\bar{X}} x_{q'} \right), \\ &= \sum_{j=1}^k \frac{N_j^2}{N^2} \cdot \frac{1}{n_j^2} n_j (n_j - 1) \left( \bar{y}_j - \frac{\bar{Y}}{\bar{X}} \cdot \bar{x}_j \right)^2. \\ C &= \sum_{\substack{j, j' = 1 \\ j \neq j'}}^k \frac{N_j N_{j'}}{N^2} \cdot \frac{1}{n_j n_{j'}} \\ &\quad \sum_{\substack{q=1 \\ \text{stratum } j}}^{n_j} \sum_{\substack{q'=1 \\ \text{stratum } j'}}^{n_{j'}} \cdot E_B \left( y_q - \frac{\bar{Y}}{\bar{X}} x_q \right) \left( y_{q'} - \frac{\bar{Y}}{\bar{X}} x_{q'} \right) \\ &= g^2 - \sum_{j=1}^k \frac{N_j^2}{N^2} \left( \bar{y}_j - \frac{\bar{Y}}{\bar{X}} \bar{x}_j \right)^2. \end{aligned}$$

We have, from the above equalities,

$$(B + C) = g^2 - \sum_{j=1}^k \frac{N_j^2}{N^2} \cdot \frac{1}{n_j} \left( \bar{y}_j - \frac{\bar{Y}}{\bar{X}} \bar{x}_j \right)^2.$$

Now because of the assumption that the variates  $y_i$  are drawn from a superpopulation satisfying  $\mathcal{E}(y_i - \theta x_i) = 0, i = 1, \dots, N$ , in the above expression for  $(B + C)$ ,  $\bar{y}_j - (\bar{Y}/\bar{X})\bar{x}_j \simeq 0(1/\sqrt{n_j}), j = 1, \dots, k$ . Therefore in (20), for large  $n_j, j = 1, \dots, k$ ,

$$E_B(g_b^2) \simeq A + g^2.$$

That is in (19)

$$V_B(g_b) \simeq A = \sum_{j=1}^k \frac{N_j^2}{N^2} \frac{1}{n_j^2} \sum_{i \in s_j} \left( y_i - \frac{\bar{Y}}{\bar{X}} x_i \right)^2 \simeq \hat{V}(g) \quad (21)$$

in (11).

The variance estimate  $\hat{V}(g)$  in (5) is obtained as a special case of (21) when  $x_i = 1, i = 1, \dots, N$ . Similarly the variance estimate  $\hat{V}(g)$  in (16) is obtained by replacing in (21) individual 'i' by a cluster 'c' and correspondingly replacing  $y_i$  and  $x_i$  by  $\hat{Y}_c$  and  $X_c$  respectively.

## 5. STRATA WITH DIFFERING MEANS

The optimality of the estimating functions  $g$  in (3), (10) and (14) discussed in the previous sections depends in an *essential* manner on the superpopulation model  $\mathcal{E}(y_i - \theta x_i) = 0, i = 1, \dots, N$ . The optimality however is not much affected by the superpopulation variances  $\mathcal{E}(y_i - \theta x_i)^2, i = 1, \dots, N$  (Godambe, 1995). This is also supported by the simulation studies reported in the next section.

It is interesting to note that the optimality of the estimating function  $g$  in (3) continues to hold even when the superpopulation model  $\mathcal{E}(y_i - \theta) = 0, i \in \mathcal{P}$  is replaced by the extended model  $\mathcal{E}(y_i - \theta_j) = 0, i \in \mathcal{P}_j, j = 1, \dots, k$ . That is now  $\theta$  is allowed to vary from stratum to stratum (Godambe, 1995). However now, the variance of  $g$  cannot be approximated by replacing the stratum mean  $\bar{Y}_j$  by the population mean  $\bar{Y}$  in (4). The earlier approximation and the subsequent estimate  $\hat{V}(g)$  in (5) were based on the assumption that (for large strata) the differences  $\bar{Y}_j - \theta$  or  $\bar{Y}_j - \bar{Y}, j = 1, \dots, k$  can be ignored. With  $\theta$  replaced in the stratum  $\mathcal{P}_j$  by  $\theta_j$ , the terms  $\bar{Y}_j - \bar{Y}$  are no more ignorable,  $j = 1, \dots, k$ . Here we note that the practice of stratifying the population so as to make each stratum 'internally' homogeneous tends to enlarge the differences  $\bar{Y}_j - \bar{Y}, j = 1, \dots, k$ . The title of this section is intended to reflect this situation.

To obtain an estimate  $\hat{V}(g)$ , under the extended model  $\mathcal{E}(y_i - \theta_j) = 0$ , we set out to estimate the nuisance parameters  $\theta_j$  or  $\bar{Y}_j, j = 1, \dots, k$  holding  $\bar{Y}$  fixed. Note that this problem of estimation is entirely different, conceptually and also mathematically, from that of the estimation of the variance of  $\bar{y}$  in (4).

Now assuming the strata sizes  $N_j, j = 1, \dots, k$  are large, we would approximate the superpopulation parameters  $\theta_j$ , with the survey population stratum means  $\bar{Y}_j, j = 1, \dots, k$ . Further, the problem of estimating  $\bar{Y}_j, j = 1, \dots, k$ , subject to holding the population mean  $\bar{Y}$  fixed, can be solved following the usual Lagrangian technique: For variations of  $\bar{Y}_j, j = 1, \dots, k$  minimize the function

$$\phi = \sum_{j=1}^k \sum_{i \in \mathcal{P}_j} (y_i - \bar{Y}_j)^2 - \lambda \left\{ \left( \sum_{j=1}^k \frac{N_j \bar{Y}_j}{N} \right) - \bar{Y} \right\}, \quad (22)$$

where  $\lambda$  is the Lagrangian multiplier. This technique of estimation has intuitive appeal even without reference to the superpopulation model just mentioned. It is easy to check that (22) is minimized for  $\hat{\bar{Y}}_j = \bar{Y}_j$ , (that is for the estimate  $\hat{\bar{Y}}_j$  of  $\bar{Y}_j$ ) where,

$$\hat{\bar{Y}}_j = \bar{y}_j - \frac{N_j / (2n_j N)}{\sum_{j=1}^k [N_j^2 / (2n_j N^2)]} (\bar{y} - \bar{Y}), \quad j = 1, \dots, k, \quad (23)$$

$n_j$  as before being the sample sizes from stratum  $j, j = 1, \dots, k$ . Note, when strata sizes  $N_j$  and sample sizes  $n_j$  are 'proportional' that is  $(n_j/N_j) = (n/N), j = 1, \dots, k$  the equations, (23) reduce to

$$\hat{\bar{Y}}_j = \bar{y}_j - (\bar{y} - \bar{Y}), \quad j = 1, \dots, k. \quad (24)$$

This simple relationship can also be used when strata sizes and sample sizes are not exactly proportional but are only approximately so. Note with reference to (24), that the estimating function  $(\bar{Y}_j - \bar{y}_j) - (\bar{Y} - \bar{y})$  is design-unbiased.

The above discussion also suggests estimation of the stratum means  $\bar{Y}_j$  when there is a covariate  $x$ . The superpopulation model underlying the estimating function  $g$  in (10), as noted before was  $\mathcal{E}(y_i - \theta_j) = 0$ , for all individuals  $i \in \mathcal{P}$ , with a common parameter  $\theta$ . Suppose this model is to be replaced by a more flexible and realistic model where the parameter  $\theta$  is allowed to vary from stratum to

stratum. That is now  $\mathcal{E}(y_i - \theta_j x_i) = 0, i \in \mathcal{P}_j, \mathcal{P}_j$  as before denoting the stratum  $j, j = 1, \dots, k$ . As in (4), the stratum means  $\bar{Y}_j$  enter the variance  $V(g)$  of the estimating function  $g$  in (10);

$$V(g) = \sum_{j=1}^k \frac{N_j^2}{N^2} \left( \frac{1}{n_j} - \frac{1}{N_j} \right) \frac{1}{N_j - 1} \sum_{i \in \mathcal{P}_j} \left\{ (y_i - \bar{Y}_j) - \frac{\bar{Y}}{\bar{X}} (x_i - \bar{X}_j) \right\}^2, \quad (25)$$

$\bar{X}_j, j = 1, \dots, k$  as before denoting the stratum means of  $x$ 's. The estimate  $\hat{V}(g)$  in (11) was obtained by ignoring the terms  $\bar{Y}_j - \frac{\bar{Y}}{\bar{X}} \bar{X}_j$  assuming large stratum sizes  $N_j$  and the superpopulation model,  $\mathcal{E}(y_i - \theta x_i) = 0$  for all individuals  $i \in \mathcal{P}$ , with a 'common' parameter  $\theta$ . With the new, more flexible model  $\mathcal{E}(y_i - \theta_j x_i) = 0, i \in \mathcal{P}_j, j = 1, \dots, k$ , the terms  $\bar{Y}_j - \frac{\bar{Y}}{\bar{X}} \bar{X}_j$  cannot be ignored any more. Now the appropriate estimate namely  $\hat{V}(g)$  of the variance  $V(g)$  in (25) is given by

$$\hat{V}(g) = \sum_{j=1}^k \frac{N_j^2}{N^2} \left( \frac{1}{n_j} - \frac{1}{N_j} \right) \frac{N_j}{N_j - 1} \cdot \frac{1}{n_j} \sum_{i \in \mathcal{P}_j} \left\{ (y_i - \bar{Y}_j) - \frac{\bar{Y}}{\bar{X}} (x_i - \bar{X}_j) \right\}^2. \quad (26)$$

However now the usual confidence intervals for  $(\bar{Y}/\bar{X})$  obtained by inverting the distribution of the approximate pivot  $\{g/\sqrt{\hat{V}(g)}\}$  contain nuisance parameters  $\bar{Y}_j, j = 1, \dots, k$ , assuming the covariate stratum means  $\bar{X}_j, j = 1, \dots, k$  are known.

As before now we have to estimate the nuisance parameters, namely the stratum means  $\bar{Y}_j, j = 1, \dots, k$ , holding the population mean  $\bar{Y}$  fixed. Note that now the underlying superpopulation model is  $\mathcal{E}(y_i - \theta_j x_i) = 0, i \in \mathcal{P}_j, j = 1, \dots, k$ . Further denoting the superpopulation variances  $\mathcal{E}(y_i - \theta_j x_i)^2 = \sigma_i^2, i \in \mathcal{P}_j$  and assuming as before, the strata sizes  $|\mathcal{P}_j| = N_j, j = 1, \dots, k$  to be large, we replace the function  $\phi$  in (22) by

$$\psi = \sum_{j=1}^k \sum_{i \in \mathcal{P}_j} (\sigma_i^2)^{-1} \left( y_i - \frac{\bar{Y}_j}{\bar{X}_j} \cdot x_i \right)^2 - \lambda \left\{ \left( \sum_{j=1}^k \frac{\bar{Y}_j}{\bar{X}_j} N_j \bar{X}_j \right) - N \bar{Y} \right\}; \quad (27)$$

$\lambda$  as before being the Lagrangian multiplier. Underlying the minimization of  $\phi$  in (22), there was a tacit *assumption* that the corresponding superpopulation model variances namely  $\mathcal{E}(y_i - \theta_j)^2 = \sigma_i^2$  were constant ( $\sigma^2$ ) for all  $i \in \mathcal{P}$ . Similarly now in minimizing  $\psi$  in (27) we make the 'assumption' that for the superpopulation model  $\mathcal{E}(y_i - \theta_j x_i) = 0$ , the variance functions  $\mathcal{E}(\theta_i - \theta_j x_i)^2 = \sigma_i^2 = \sigma^2 x_i$ ,  $i \in \mathcal{P}$ . That is  $\sigma_i^2$  is proportional to the covariate value  $x_i$ ,  $i \in \mathcal{P}$ . As stated in the beginning of this section, both the assumptions just mentioned, are primarily for 'simplicity' and are of no important statistical consequence (Godambe, 1995). It is easy to check that the values (estimates)  $\hat{Y}_j$  of  $\bar{Y}_j$  which minimize (27) are given by

$$\frac{\bar{y}_j}{\bar{x}_j} - \frac{\hat{Y}_j}{\bar{X}_j} = \left[ \left\{ \left( \sum_{j=1}^k N_j \bar{X}_j \frac{\bar{y}_j}{\bar{x}_j} \right) - N\bar{Y} \right\} / \left\{ \sum_{j=1}^k \frac{(N_j \bar{X}_j)^2}{2n_j \bar{x}_j} \right\} \right] \cdot \frac{N_j \bar{X}_j}{2n_j \bar{x}_j}, \quad (28)$$

$j = 1, \dots, k$ . Again as in case of (23), the equations (28) are considerably simplified for large samples, in case the sample sizes  $n_j$  are proportional to the strata sizes  $N_j$ , that is  $(n_j/N_j) = (n/N)$ ,  $j = 1, \dots, k$ :

$$\frac{\bar{y}_j}{\bar{x}_j} - \frac{\hat{Y}_j}{\bar{X}_j} \simeq \frac{\bar{y} - \bar{Y}}{\bar{X}}, \quad j = 1, \dots, k. \quad (29)$$

Since in (29), the right hand side is  $O(1/\sqrt{n})$  we have

$$\hat{Y}_j \simeq \frac{\bar{y}_j}{\bar{x}_j} \bar{X}_j \quad j = 1, \dots, k. \quad (30)$$

The simple estimates of  $\bar{Y}_j$  namely  $\hat{Y}_j$  given by (30) are quite intuitive and can also be used even in case the sample sizes  $n_j$  are not 'exactly' proportional to  $N_j$  but are only 'approximately' so. Note that analogous to (24), underlying (30) is the design-unbiased estimating function  $(\bar{x}_j \bar{Y}_j - \bar{y}_j \bar{X}_j)$ .

Now for the estimating function  $g$  in (3), the variance  $V(g)$  is given by (4). Further if  $\hat{V}_1(g)$  is the estimate of  $V(g)$ , based on the estimates  $\hat{Y}_j$  of  $\bar{Y}_j$ ,  $j = 1, \dots, k$  given by (24), then

$$\hat{V}_1(g) = \sum_{j=1}^k \frac{N_j^2}{N^2} \left( \frac{1}{n_j} - \frac{1}{N_j} \right) \frac{1}{(n_j - 1)} \cdot \sum_{i \in s_j} \left\{ y_i - (\bar{y}_j - \bar{y} + \bar{Y}) \right\}^2. \quad (31)$$

The confidence intervals for  $\bar{Y}$  are obtained by inverting the distribution of the approximate pivot  $[g / \{\hat{V}_1(g)\}^{1/2}]$ ; asymptotically,

$$g / \{\hat{V}_1(g)\}^{1/2} \sim N(0, 1). \quad (32)$$

Similarly in case of a covariate, for the estimating function  $g$  in (10), if  $\hat{V}_2(g)$  denotes the estimate of the variance  $V(g)$  in (25), based on the estimates  $\hat{Y}_j$  given by (30), then

$$\hat{V}_2(g) = \sum_{j=1}^k \frac{N_j^2}{N^2} \left( \frac{1}{n_j} - \frac{1}{N_j} \right) \frac{1}{(n_j - 1)} \sum_{i \in s_j} \left\{ (y_i - \frac{\bar{y}_j}{\bar{x}_j} \bar{X}_j) - \frac{\bar{Y}}{\bar{X}} (x_i - \bar{X}_j) \right\}^2. \quad (33)$$

Again the confidence intervals for  $\bar{Y}$  are based on the inversion of the distribution of  $[g / \{\hat{V}_2(g)\}^{1/2}]$ ; asymptotically,

$$g / \{\hat{V}_2(g)\}^{1/2} \sim N(0, 1). \quad (34)$$

*Note 3.* Even if in (22) and (27), the strata weights  $N_j$  are introduced to define the functions  $\phi$  and  $\psi$  respectively, the final estimates  $\hat{Y}_j$  would remain unaltered; they would as in (24) and (30).

## 6. APPLICATIONS

In the preceding section we have provided construction of confidence intervals when the superpopulation model  $\mathcal{E}(y_i - \theta) = 0$  or  $\mathcal{E}(y_i - \theta x_i) = 0$ , with a 'common' value of  $\theta$  for all individuals  $i \in \mathcal{P}$ , is replaced by the model  $\mathcal{E}(y_i - \theta_j) = 0$  or  $\mathcal{E}(y_i - \theta_j x_i) = 0$ ,  $i \in \mathcal{P}_j$ ,  $j = 1, \dots, k$ . That is now  $\theta$  can vary from stratum to stratum. Generally in practice, one cannot be sure if for the survey population at hand, the parameter  $\theta$  has a 'common' value for all individuals  $i \in \mathcal{P}$ . Theoretical as well as numerical investigations clearly indicate that the performance of the confidence intervals computed on the assumption of a common value of  $\theta$  (i.e. the once based on the pivots  $[g / \{\hat{V}(g)\}^{1/2}]$  of Section 2), is very susceptible even to 'small deviations' of  $\theta$ , from stratum to stratum. In contrast, as said before, the just mentioned confidence intervals are not much affected by

the differential model variances or even the distributions of  $y$ , from stratum to stratum. Again, as said before, in stratifying a population, a prior assessment of the mean values  $\theta$  for different individuals can lead to construction of strata  $\mathcal{P}_j$ , with differing mean values  $\theta_j$ ,  $j = 1, \dots, k$ .

For the reasons given above we propose a general use of the confidence intervals based on the pivot (32), when there is no covariate; and the confidence intervals based on the pivot (34) for a covariate case. In the following illustrations the above confidence intervals are compared with the *conventional* confidence intervals: They are obtained, in case of no covariate, from the approximate  $N(0, 1)$  pivot,

$$\left\{ \sum_{j=1}^k \frac{N_j}{N} \cdot \frac{1}{n_j} \sum_{i \in s_j} (y_i - \bar{Y}) \right\} / \left\{ \sum_{j=1}^k \frac{N_j^2}{N^2} \left( \frac{1}{n_j} - \frac{1}{N_j} \right) \frac{1}{(n_j - 1)} \sum_{i \in s_j} (y_i - \bar{y}_j)^2 \right\}^{1/2} ; \quad (35)$$

in case of a covariate the approximate  $N(0, 1)$  pivot is

$$\sum_{j=1}^k \frac{N_j}{N} \cdot \frac{1}{n_j} \sum_{i \in s_j} \left( y_i - \frac{\bar{Y}}{\bar{X}} x_i \right) / \left[ \sum_{j=1}^k \frac{N_j^2}{N^2} \cdot \left( \frac{1}{n_j} - \frac{1}{N_j} \right) \frac{1}{n_j - 1} \sum_{i \in s_j} \left\{ (y_i - \bar{y}_j) - \frac{\bar{y}}{\bar{x}} (x_i - \bar{x}_j) \right\}^2 \right]^{1/2} . \quad (36)$$

(Cochran, 1977). In general we will refer to (32) and (34) as the *new* pivots and (35) and (36) as the *conventional* pivots. Extensive simulation experiments were conducted to compare confidence intervals based on the new and the conventional pivots. However the results reported below are primarily for small samples. Here we have sixteen survey populations, each of which is divided into four strata; samples of sizes 2, 3, 4, 2 are drawn from the respective strata. Such samples of (total) sizes as small as 11, can bring out best, as in Tables 1 and 2 to follow, the superior performance of the new pivots over the conventional ones. Though to a less degree than for the small size samples, just mentioned, the superiority of the new pivots over the conventional ones continues to hold, for moderate size (25) samples,

as in Table 3 and 4. Our unreported simulation studies included populations divided into 16 strata each, with total sample size of about 50. Even for such large samples the new pivots appear to perform better than the conventional ones. Eventually, of course for very large sample sizes, the distinctive performances between the two pivots, the new and the conventional tends to disappear.

The sixteen survey populations (1) - (16) in Tables 1 and 2 below, barring populations (7), (8) are of sizes 1000 each; populations (7), (8) are of sizes 2000 each. Each one of the sixteen populations, as said before, is divided in 4 strata. Tables 1 and 2, each have six columns (i), (ii), ... (vi). Column (i) gives the population number ( $\cdot$ ). Column (ii) provides, corresponding to the four strata, of the population ( $\cdot$ ), the superpopulation distributions from which the strata have been drawn. The distribution can be Chi-square (C), Normal (N), or Uniform (U).

When there is no covariate as in Table 1, column (ii) refers to just the distribution, of the variate  $y$ ; on the other hand in Table 2, it refers to the distributions of both, the variate  $y$  and the covariate  $x$ . Column (iii) gives the sample sizes from different strata. Column (iv) shows the nominal coverage probability. Columns (v) and (vi) provide the actual coverage probabilities attained and the average length of the confidence intervals, under 4000 simulations. Thus a typical horizontal line in Table 1, starting with (6) say, is to be read as follows. The four strata of the population (6) are drawn from the superpopulation distributions Normal, Chi-square, Normal, Chi-square respectively; the sample sizes from different strata being (2, 3, 4, 2) respectively. The interpretation of the columns (iv), (v), (vi) is straightforward. The Table 2 needs no extra explanation, excepting that as said above, here column (ii) in addition to giving the distribution of the variate  $y$ , also gives the distribution of the covariate  $x$ . Unlike the populations (1) - (16) above, the populations (17) and (18) in Tables 3 and 4 are divided into 8 strata, each, the population (17) being without a covariate and (18) with a covariate.

TABLE 3  
 In the population numbered (17) below, the mean value  $\theta$ , varies  
 from stratum to stratum between  $\theta = 100$  to  $\theta = 800$ .

(i) Population	(ii) Superpopulation distribution $y$	(iii) Sample sizes	(iv) Nominal coverage	(v) Actual coverage probability pivot (32) pivot (35)	(vi) Average length pivot (32) pivot (35)
(17)	$\{N, U, C, U, C, N, U, U\}$	(2, 3, 4, 2, 3, 4, 3, 4)	.95	.93 .889	12.76 10.94

TABLE 4  
 In the population numbered (18) below, the regression coefficient  $\theta$  varies  
 from stratum between  $\theta = 3$  to  $\theta = 6$

(i) Population	(ii) Superpopulation distribution	(iii) Sample sizes	(iv) Nominal coverage	(v) Actual coverage probability pivot (34) pivot (36)	(vi) Average length pivot (34) pivot (36)
(18)	x: $\{C, C, C, C, C, C, C, C\}$ y: $\{N, U, N, C, C, C, C, N\}$	(2, 3, 4, 2, 3, 4, 3, 4)	.95	.937 .90	22.80 20.89

TABLE 1

In the populations numbered (1) - (4) below the mean value  $\theta$  is held fixed from stratum to stratum,  $\theta = 100$ . For the remaining populations, (5) to (8), the mean value  $\theta$  varies from stratum to stratum, between  $\theta = 100$  to  $\theta = 400$

(i) Population	(ii) Superpopulation distribution $y$	(iii) Sample sizes	(iv) Nominal coverage probability	(v) Actual coverage probability pivot (32) pivot (35)	(vi) Average length pivot (32) pivot (35)
(1)	{N, C, U, C}	(2, 3, 4, 2)	.95	.967	19.83
(2)	{N, C, U, C}	(2, 3, 4, 2)	.90	.90	13.11
(3)	{N, N, N, N}	(2, 3, 4, 2)	.90	.90	4.85
(4)	{U, U, U, U}	(2, 3, 4, 2)	.90	.90	4.90
(5)	{N, C, N, C}	(2, 3, 4, 2)	.95	.946	34.34
(6)	{N, C, N, C}	(2, 3, 4, 2)	.90	.866	22.71
(7)	{N, U, C, N}	(2, 3, 4, 2)	.95	.97	20.76
(8)	{N, U, C, N}	(2, 3, 4, 2)	.90	.908	13.69

TABLE 2

In the populations numbered (9) - (12) below the regression coefficient  $\theta$  is held fixed for all strata,  $\theta = 3$ . For the remaining populations (13) - (15), the regression coefficients  $\theta$ , varies from stratum to stratum, between  $\theta = 2$  to  $\theta = 4$ .

(i) Population	(ii) Superpopulation distribution $x$	(iii) Sample sizes	(iv) Nominal coverage probability	(v) Actual coverage probability pivot (34) pivot (36)	(vi) Average length pivot (34) pivot (36)
(9)	{U, U, U, U}	(2, 3, 4, 2)	.90	.879	91.12
(10)	{U, U, U, U}	(2, 3, 4, 2)	.95	.926	113.49
(11)	{U, U, U, U}	(2, 3, 4, 2)	.90	.88	12.21
(12)	{C, C, C, C}	(2, 3, 4, 2)	.90	.83	7.07
(13)	{C, C, C, C}	(2, 3, 4, 2)	.95	.926	113.53
(14)	{C, C, C, C}	(2, 3, 4, 2)	.90	.869	35.89
(15)	{C, C, C, C}	(2, 3, 4, 2)	.95	.92	42.88
(16)	{U, C, C, U}	(2, 3, 4, 2)	.95	.959	31.17

## 7. CONCLUSIONS

The following conclusions are based on the theoretical investigations of the preceding sections and the simulation results reported in Section 6 and many other simulation results, as mentioned earlier, not reported in this paper.

The situation when there is no covariate seems to be fairly clear from Tables 1 and 3 of Section 6. For small samples the conventional confidence intervals, that is the ones based on the pivot (35), can be very misleading: The 'asserted' probability of coverage can be very different than the 'actual' one. Further, this gap between 'asserted' and 'actual' coverage probabilities, for the conventional confidence intervals seem to increase as the 'variation in the strata means' increases. Interestingly, as noted in Section 5, this increased variation in the strata means, can often be a result of stratifying a population into (internally) homogeneous strata for efficient point estimation. The confidence intervals based on the new pivot (32), as it can be seen from the Tables 1 and 3 of Section 6, perform much better than the ones based on the conventional pivot (35). From our simulations, based on the three distributions namely Normal, Chi-square and Uniform, it seems that the comparison between performance of the new pivot (32) and the conventional one (35) depends on the distributions mostly through their variations of the mean values from stratum to stratum. Particularly, the comparison is not much affected by the variances or the forms of the distributions. This is to be expected from our underlying semi-parametric model,  $\mathcal{E}(y_i - \theta_j) = 0$ ,  $i \in \mathcal{P}_j$ ,  $j = 1, \dots, k$ . This thus extends the conclusion previously drawn in the beginning of Section 6. We emphasize here that the optimality of the estimating function  $g$  in (3), continues to hold even when  $\theta$  varies from stratum to stratum.

For large samples, according to our simulation results mentioned earlier (unreported here) the difference between the two sets of confidence intervals, one based on the pivot (32) and the other on (35) tend to diminish. This, also is in line with the theory.

Tables 2 and 4 of Section 6, provide results concerning confidence intervals for populations admitting a covariate. Here a comparison of the performances of the new pivot (34) and the conventional

one (36) is rather subtle. We consider two situations: *One*, when the regression coefficient  $\theta$ , is the same for all strata. *Two*, when  $\theta$  varies (though not very much) from stratum to stratum. Only in the former situation the estimating function  $g$  in (10) is optimal. For this situation, (i.e. same  $\theta$  for all strata), which is mostly of academic interest, the pivot given at the end of Section 2, as our simulation studies (unreported here) show, performs very well. The situation two, above, is more realistic. Hence it is practically very important to study the performance of the estimating function  $g$ , that is the performance of the confidence intervals based on the new pivot (34), when  $\theta$  varies from stratum to stratum. Under this situation, it is clear from Tables 2 and 4, that the confidence intervals based on the new pivot (34) provide 'actual' coverage probabilities *closer* to the 'asserted' ones than the confidence intervals based on the conventional pivot (36). Also under situation one, i.e. the same  $\theta$  for all strata, as the Table 2 indicates, the performance of the new pivot (34) is at least as good as the conventional pivot (36). The phenomenon seems to be more striking as more variation in the covariate values is introduced. Actually as the covariate values within each stratum tend to be uniform the difference between the performances of the two pivots, the new (34) and the conventional (36), tends to diminish. This is also true as the sample sizes go on increasing.

It is interesting to note that the confidence intervals based on the new pivots are generally longer than the ones based on the conventional pivots. All the same, the comparison between the two pivots, new and conventional, of the 'coverage probabilities', holding the respective 'lengths' fixed was found to be distinctly in favour of the former in the cases investigated; albeit very few. Of course what is operationally most important is the fact that the usual confidence statements based on the new pivots are far more accurate than the ones based on the conventional pivots.

## ACKNOWLEDGEMENTS

I am grateful to D.R. Cox, A.C. Singh and M.E. Thompson for their valuable comments on the earlier draft of this paper. My thanks are also due to Jiahua Chen and Lianxiang Wang for their advice and assistance in computations.

## REFERENCES

- Basu, D. (1958), "On Sampling with Replacement", *Sankhya*, **20**, 287-294.
- Binder, D.A. (1983), "On the Variances of Asymptotically Normal Estimators from Complex Surveys", *International Statistical Review*, **51**, 279-292.
- Binder, D.A. and Patak, Z. (1994), "Use of Estimating Functions for Estimation from Complex Surveys", *Journal of the American Statistical Association*, **39**, 1035-1043.
- Chandhuri, A. and Vos, J.W.E. (1988), *Unified Theory and Strategies of Survey Sampling*, Amsterdam: North Holland.
- Cochran, W.G. (1977), *Sampling Techniques* (3rd ed.), New York: John Wiley.
- Deming, W.E. (1950), *Some Theory of Sampling*, John Wiley & Sons Inc., New York: Chapman & Hall, London.
- Efron, B. and Hinkley, D.V. (1978), "Assessing the Accuracy of the Maximum Likelihood Estimator: Observed versus Expected Fisher Information", (with discussion), *Biometrika*, **65**, 457-487.
- Godambe, V.P. (1976), "A Historical Perspective of Recent Developments in the Theory of Sampling from Actual Populations", *Journal of the Indian Society of Agricultural Statistics*, **28**, 1-12.
- \_\_\_\_\_ (1985), "The Foundations of Finite Sample Estimation in Stochastic Processes", *Biometrika*, **72**, 419-428.
- \_\_\_\_\_ (1991), "Orthogonality of Estimating Functions and Nuisance Parameters", *Biometrika*, **78**, 143-151.
- \_\_\_\_\_ (1995), "Estimation of Parameters in Survey Sampling: Optimality", *Canadian Journal of Statistics*, **23**, 227-243.
- \_\_\_\_\_ (1997), "Estimation of Parameters in Survey Sampling", *Proceedings of the Survey Methods Section SSC Annual Meetings, June 1996*, 1-8.
- Godambe, V.P. and Heyde, C.C. (1987), "Quasi-likelihood and Optimal Estimation", *International Statistical Review*, **55**, 231-244.
- Godambe, V.P. and Thompson, M.E. (1986), "Parameters of Superpopulation and Survey Population: Their Relationship and Estimation", *International Statistical Review*, **54**, 127-138.
- \_\_\_\_\_ (1989), "An Extension of Quasi-likelihood Estimation" (with discussion), *Journal of Statistical Planning and Inference*, **22**, 137-172.
- Hajek, J. (1959), "Optimum Strategy and other Problems in Probability Sampling", *Casopis Pro Pěstování Matematiky*, **84**, 387-423.
- Mach, L. (1988), "The Use of Estimating Functions for Confidence Interval Construction: The Case of the Population Mean. Working Paper No. BSMD-88-028 E Methodology Branch, Statistics Canada.
- Neyman, J. (1934). "On Two Different Aspects of Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection", (with discussion), *Journal of the Royal Statistical Society*, **97**, 558-652.
- \_\_\_\_\_ (1937), "Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability", *Philosophical Transactions of Royal Society Series A*, **236**, 333-380.
- Sarndal, C.E., Swensson, B. and Wretman, J. (1992), *Model Assisted Survey Sampling*, Springer-Verlag, New York.
- Smith, T.M.F. (1997), "Social Surveys and Social Science", (with discussion), *Canadian Jour. Statist.* **25**, 23-44.
- Wilks, S.S. (1938), "Shortest Average Confidence Intervals from Large Samples" *Annals of Mathematical Statistics*, **9**, 166-175.

Woodruff, R.S. (1952), "Confidence Intervals for Medians and other Position Measures. *Journal of the American Statistical Association*, 47, 635-646.

Young, W. and Rao, J.N.K. (1996), "Jackknife Linearization Variance Estimators Under Stratified Multi-Stage Sampling", *Survey Methodology*, 22, 23-31.