COMBINING DATA SOURCES IN ESTIMATION

S. E. Ahmed and Bashirullah, University of Regina

S. E. Ahmed, Department of Mathematics and Statistics, Regina, Saskatchewan, S4S 0A2 CANADA

KEY WORDS: uncertain prior information, bias and quadratic risk, shrinkage preliminary test estimator and size of the test.

ABSTRACT

The problem of simultaneous estimation of a set of normal means is considered. The properties of the proposed estimator are assessed. It is demonstrated analytically and numerically that the proposed shrinkage preliminary test estimator provides a wider range than the usual preliminary test estimator in which it dominates the classical estimator. An optimal rule for the size of the preliminary test is presented. It is found that the size of the preliminary test for the proposed shrinkage preliminary test estimator is reasonable.

1. INTRODUCTION

In this paper we discuss various estimation techniques to determine whether or not to combine two or more data sources on the basis of preliminary tests of significance and shrinkage principle. It is advantageous to utilize all the data sources in the estimation procedure under some assumptions. But in some experimental situations, it is not certain whether or not these assumptions hold. This uncertain prior information (UPI), in form of the hypothesis, in the estimation procedure can be used in estimation process. It is natural to perform a preliminary test on the validity of the UPI in the form of the parametric constraints, and then choose between the restricted and unrestricted estimation procedure depending upon the outcome of the preliminary test (Bancroft, 1944). The preliminary test estimators are widely used by researchers, as is evident from the extensive bibliographies of Bancroft and Han(1977) and Han *et al.* (1988). For a nice account of the parametric theory of the preliminary test estimation in the finite sample space, we refer to Ahmed (1992a), Ahmed and Saleh (1990) and Judge and Bock (1978), among others. Asymptotic theory of these estimators have been studied by Ahmed (1991, 1992b), Kulperger and Ahmed (1992) and Gupta *et al.* (1989).

In this article, our aim is to focus on the small sample properties (under quadratic loss) of the estimators based on preliminary test estimators and to compare these with usual estimators.

Let $Y_{i1}, Y_{i2}, \cdots, Y_{in_i}$ are independent observations from normal distributions with finite variance σ^2 and mean θ_i, $i = 1, 2, \cdots, k$, and respective sample sizes $n_1, \cdots, n_k$.

We are interested in simultaneous estimation of the mean parameter $\boldsymbol{\theta} = (\theta_1 \cdots, \theta_k)'$. To estimate $\boldsymbol{\theta}$, one need only to consider the sufficient statistic $T_i(y) = \sum_{j=1}^{n_i} Y_{ij}$. The unrestricted estimator (UE) of θ is defined componentwise by $\hat{\theta} = T_i(y)/n_i$, $i = 1, 2, \cdots k$. Moreover, it is suspected that the mean θ_i are presumably equal, but with some degree of uncertainty. We are primarily interested in the estimation of $\boldsymbol{\theta}$ when

$$H_o : \theta_1 = \theta_2 = \cdots = \theta_k = \theta_O \text{ (unknown)}. \quad (1.1)$$

The restricted estimator (RE) of $\boldsymbol{\theta}$ is defined componentwise by $\hat{\boldsymbol{\theta}}^R = \sum_{i=1}^{k} T_i(y)/n$, where $n = n_1 + n_2 + \cdots + n_k$. $\hat{\boldsymbol{\theta}}^R$ of $\boldsymbol{\theta}$ performs better than $\hat{\boldsymbol{\theta}}$ when H_o in (1.1) holds but as the hypothesis error grows, $\hat{\boldsymbol{\theta}}^R$ may be considerably biased, inefficient and inconsistent, while the performance of $\hat{\boldsymbol{\theta}}$ remains constant over such departures. In order to overcome this shortcoming of the $\hat{\boldsymbol{\theta}}^R$, it is often desirable to develop an estimator which is a compromise between $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}^R$ by incorporating a *preliminary test (PT)* on the null hypothesis H_o in (1.1). It is important to remark here that $\hat{\boldsymbol{\theta}}^P$ is a function of α, the size of the preliminary test. It is recommended in the literature to use a level of significance of at least 0.15 for such preliminary testing. Use of such a large significance level helps to maximize the minimum efficiency of $\hat{\boldsymbol{\theta}}^P$. Thus, the use of $\hat{\boldsymbol{\theta}}^P$ may be limited due to the large size of the preliminary test. In this paper, a shrinkage technique will therefore be introduced into the preliminary test estimation to overcome this difficulty. The proposed methodology remarkably improves upon the $\hat{\boldsymbol{\theta}}^P$ with respect to the size of the preliminary test (Ahmed, 1992a). Interestingly, $\hat{\boldsymbol{\theta}}^{SP}$ dominates $\hat{\boldsymbol{\theta}}$ over a wider range than $\hat{\boldsymbol{\theta}}^P$. More importantly, the proposed $\hat{\boldsymbol{\theta}}^{SP}$ provides much more meaningful size for the preliminary test than $\hat{\boldsymbol{\theta}}^P$

2. IMPROVED ESTIMATION

The *preliminary test estimator (PE)*of $\boldsymbol{\theta}$ denoted by $\hat{\boldsymbol{\theta}}^P = (\theta_1^P, \cdots \theta_k^P)'$ is given by

$$\hat{\boldsymbol{\theta}}^P = \hat{\boldsymbol{\theta}}^R I(D_n < c_{(\alpha)}) + \hat{\boldsymbol{\theta}} I(D_n \geq c_{(\alpha)}), \quad (2.1)$$

where $I(A)$ is an indicator function of the set A, and D_n is a test statistic for testing the preliminary hypothesis and $c_{(\alpha)}$ be the upper $100\alpha\%$ $(0 < \alpha < 1)$ point of D_n,

$$D_n = \frac{(\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}^R)' \boldsymbol{\Lambda} (\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}^R)}{k^* S^2}, \quad (2.2)$$

$$\boldsymbol{\Lambda} = diag(n_i), \; S^2 = \frac{1}{m} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \hat{\theta}_i)^2,$$

where $k^* = k - 1$ and $m = n - k$.

Under the null hypothesis the test statistic D_n follows the F-distribution with (k^*, m) degrees of freedom. Thus, for a given level of significance $\alpha (0 < \alpha < 1)$, $c_{(\alpha)} = F_{k^*,m}(\alpha)$, where $F_{v_1,v_2}(\alpha)$ is the upper α level critical value of a central F-distribution with (v_1, v_2) degrees of freedom. The properties of $\hat{\boldsymbol{\theta}}^P$ is reported by Ali (1990). In the present investigation we propose an improved version of $\hat{\boldsymbol{\theta}}^P$ by employing a shrinkage technique in the estimation process.

2.1. Improving on the restricted estimator

When H_o holds in (1.1), then it is reasonable to shrink $\hat{\boldsymbol{\theta}}$ towards $\hat{\boldsymbol{\theta}}^R$ (Thompson, 1968). Thus, a *shrinkage restricted estimator (SRE)* of $\boldsymbol{\theta}$ is defined by

$$\hat{\boldsymbol{\theta}}^{SR} = \pi \hat{\boldsymbol{\theta}}^R + (1 - \pi) \hat{\boldsymbol{\theta}}, \quad \pi \in [0, 1] \quad (2.3)$$

where π is a coefficient reflecting degree of confidence in the prior information. However, the value of π may be completely determined by the experimenter on the basis of his/her belief in the UPI. If the experimenter strongly believes that H_o is true then $\pi = 1$ should be used. On the other hand, if $\pi = 0$ then $\hat{\boldsymbol{\theta}}^{SR} = \hat{\boldsymbol{\theta}}$. Thus, $\hat{\boldsymbol{\theta}}$ is a special case of $\hat{\boldsymbol{\theta}}^{SR}$. However, like $\hat{\boldsymbol{\theta}}^R$, $\hat{\boldsymbol{\theta}}^{SR}$ yields a smaller quadratic risk at and near the null hypothesis at the expense of poorer performance in the rest of the parameter space. It is important to note that the $\hat{\boldsymbol{\theta}}^{SR}$ dominates $\hat{\boldsymbol{\theta}}$ over a large portion of the parameter range which is wider than that of $\hat{\boldsymbol{\theta}}^R$. This motivates replacing $\hat{\boldsymbol{\theta}}^R$ by $\hat{\boldsymbol{\theta}}^{SR}$ in the usual PE given in (2.1). Therefore, we propose the *shrinkage preliminary test estimator (SPE)*, which is a convex combination of $\hat{\boldsymbol{\theta}}^{SR}$ and $\hat{\boldsymbol{\theta}}$ and defined in the following sub-section.

2.2. Improving on preliminary test estimator

The shrinkage preliminary test estimator (SPE) $\hat{\boldsymbol{\theta}}^{SP}$ of $\boldsymbol{\theta}$ is defined below as:

$$\hat{\boldsymbol{\theta}}^{SP} = \hat{\boldsymbol{\theta}}^{SR} I(D_n < c_{(\alpha)}) + \hat{\boldsymbol{\theta}} I(D_n \geq c_{(\alpha)}). \quad (2.4)$$

In this case if the null hypothesis is tenable then $\hat{\boldsymbol{\theta}}^{SR}$ is used, while $\hat{\boldsymbol{\theta}}$ is a sensible choice otherwise. However, both (2.1) and (2.4) involve the test statistic D_n which adjusts the estimator for any empirical departure from the null hypothesis. For large values of D_n both (2.1) and (2.4) yield $\hat{\boldsymbol{\theta}}$, while for small values of D_n their behavior is different. Furthermore, if we substitute $\pi = 1$ in (2.4) then it is the usual preliminary test estimator given in (2.1). Our main objective is to study the finite sample theory of the proposed estimator. Simultaneous analytical and computational comparison of the proposed estimators is presented. The relative performances of the estimators to the unrestricted estimator are discovered. It is important to note that the sampling properties of $\hat{\boldsymbol{\theta}}^{SP}$ depend on among other factors, the size chosen for the preliminary test. A *max-min* rule for the choice of the level of significance for the preliminary test is discussed.

Let $\hat{\boldsymbol{\theta}}^o$ be an estimator of $\boldsymbol{\theta}$ and $\boldsymbol{\Gamma}$ be a positive semi-definite (p.s.d.) matrix, then the quadratic loss function is

$$\mathcal{L}(\hat{\boldsymbol{\theta}}^o, \boldsymbol{\theta}) = (\hat{\boldsymbol{\theta}}^o - \boldsymbol{\theta})' \boldsymbol{\Gamma} (\hat{\boldsymbol{\theta}}^o - \boldsymbol{\theta}). \quad (2.5)$$

3. MAIN RESULTS

In this section, the expressions for the bias, mean squared error matrix and risks of the estimators are provided. First, the joint density of estimators is obtained in the following theorem.

Theorem 3.1: If we define,

$$\mathbf{X}_n = (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}), \quad \mathbf{Z}_n = (\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}^{SR}), \quad (3.1)$$

then

$$\begin{pmatrix} \mathbf{X}_n \\ \mathbf{Z}_n \end{pmatrix} \sim N_{2k} \left\{ \begin{pmatrix} 0 \\ \pi \boldsymbol{\delta} \end{pmatrix}, \; \begin{pmatrix} \boldsymbol{\Sigma} & \pi \boldsymbol{\Sigma} \mathbf{H}' \\ \pi \mathbf{H} \boldsymbol{\Sigma} & \pi^2 \boldsymbol{\Sigma} H' \end{pmatrix} \right\}, \quad (3.2)$$

$$\boldsymbol{\delta} = \mathbf{H} \boldsymbol{\theta}, \; \boldsymbol{\Sigma} = \sigma^2 \boldsymbol{\Lambda}^{-1}, \; \mathbf{H} = \mathbf{I_k} - \frac{\mathbf{J} \boldsymbol{\Lambda}}{n}, \; \mathbf{J} = \mathbf{1}_k \mathbf{1}_k',$$

We note that $\hat{\boldsymbol{\theta}}$ is an unbiased estimator of $\boldsymbol{\theta}$. Further, by virtue of (2.3), (2.4) and Theorem 3.1 we have

$$\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}^{SR} = \pi \mathbf{H} \hat{\boldsymbol{\theta}},$$

$$(\hat{\boldsymbol{\theta}}^{SP} - \boldsymbol{\theta}) = \mathbf{X}_n - \pi \mathbf{Z}_n I(D_n \leq c_{(\alpha)}). \qquad (3.3)$$

The bias of the $\hat{\boldsymbol{\theta}}^{SR}$ and $\hat{\boldsymbol{\theta}}^{SP}$ are derived in the following theorem:

Theorem 3.2: Bias of the $\hat{\boldsymbol{\theta}}^{SR}$ and $\hat{\boldsymbol{\theta}}^{SP}$ are given below in (3.4)-(3.5) respectively.

$$\mathbf{B}_1 = \mathcal{E}(\hat{\boldsymbol{\theta}}^{SR} - \boldsymbol{\theta}) = -\pi \boldsymbol{\delta} \qquad (3.4)$$

$$\mathbf{B}_2 = \mathcal{E}(\hat{\boldsymbol{\theta}}^{SP} - \boldsymbol{\theta}) = -\pi \boldsymbol{\delta} H_{\nu_1,m}\left(c^*_{(\alpha)}; \Delta\right), \quad (3.5)$$

where $\Delta = \boldsymbol{\delta}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}$, $\nu_1 = k+1$ and $H_{v_1,v_2}(\cdot;\Delta)$ is the cumulative distribution function of a noncentral F distribution with (v_1, v_2) degrees of freedom and noncentrality parameter Δ. Furthermore, $c^*_{(\alpha)} = \frac{k^*}{\nu_1} F_{k^*,m}(\alpha)$, and $F_{v_1,v_2}(\alpha)$ is the upper α level critical value of a central F-distribution with (v_1, v_2) degrees of freedom. We conclude that the bias of $\hat{\boldsymbol{\theta}}^{SR}$ is unbounded in $\boldsymbol{\delta}$ which goes to ∞ if $\| \boldsymbol{\delta} \|$ tends to ∞. On the other hand, the bias vector of $\hat{\boldsymbol{\theta}}^{SP}$ is bounded in $\boldsymbol{\delta}$. Hence, the expression for bias of $\hat{\boldsymbol{\theta}}^{P}$ $(\pi = 1)$,

$$\mathbf{B}_3 = \mathcal{E}(\hat{\boldsymbol{\theta}}^{P} - \boldsymbol{\theta}) = -\boldsymbol{\delta} H_{\nu_1,m}\left(c^*_{(\alpha)}; \Delta\right), \quad (3.6)$$

The absolute value of each element in vector $\mathbf{B}_2$ is less than the corresponding element in vector $\mathbf{B}_3$. Thus, the proposed estimator $\hat{\boldsymbol{\theta}}^{SP}$ is superior to $\hat{\boldsymbol{\theta}}^{P}$ from the point of view of the absolute bias depending on the value of π.

Finally, the risk of $\hat{\boldsymbol{\theta}}^{SR}$ and $\hat{\boldsymbol{\theta}}^{SP}$ are presented in the following theorem.

Theorem 3.3: Risk of $\hat{\boldsymbol{\theta}}^{SR}$ and $\hat{\boldsymbol{\theta}}^{SP}$ under the loss function defined in (2.5) are given below by (3.7)-(3.8) respectively:

$$\Re(\hat{\boldsymbol{\theta}}^{SR}; \boldsymbol{\theta}) = \Re(\hat{\boldsymbol{\theta}}; \boldsymbol{\theta}) - \pi^* tr(\boldsymbol{\Gamma}\mathbf{L}) + \pi^2 \Delta_\Gamma, \quad (3.7)$$

$$\Re(\hat{\boldsymbol{\theta}}; \boldsymbol{\theta}) = tr(\boldsymbol{\Gamma}\boldsymbol{\Sigma}), \ \pi^* = \pi(2-\pi), \ \Delta_\Gamma = \boldsymbol{\delta}'\boldsymbol{\Gamma}\boldsymbol{\delta}$$

$$\Re(\hat{\boldsymbol{\theta}}^{SP}; \boldsymbol{\theta}) = \Re(\hat{\boldsymbol{\theta}}; \boldsymbol{\theta}) - tr(\boldsymbol{\Gamma}\mathbf{L})\pi^* H_{\nu_1,m}(c^*_{(\alpha)}; \Delta) +$$

$$\Delta_\Gamma\{2\pi H_{\nu_1,m}(c^*_{(\alpha)}; \Delta) - \pi^* H_{\nu_2,m}(c^o_{(\alpha)}; \Delta)\}, \quad (3.8)$$

where $\nu_2 = \nu_1 + 2$, and $c^o_{(\alpha)} = \frac{k^*}{\nu_2} F_{k^*,m}(\alpha)$. The risk analysis of the estimators is presented in the following section.

4. RISK ANALYSIS

First, we recall that $\hat{\boldsymbol{\theta}}$ is an unbiased estimator of $\boldsymbol{\theta}$ with a constant risk $tr(\boldsymbol{\Gamma}\boldsymbol{\Sigma})$ in the entire parameter space, while the risk for $\hat{\boldsymbol{\theta}}^{SR}$ is a linear line function of Δ_Γ with slope π^2 and intercept $tr(\boldsymbol{\Gamma}\boldsymbol{\Sigma}) - \pi^* tr(\boldsymbol{\Gamma}\mathbf{L})$. The risk function of $\hat{\boldsymbol{\theta}}^{SR}$ and $\hat{\boldsymbol{\theta}}$ intersect at $\Delta\Gamma = tr(\boldsymbol{\Gamma}\mathbf{L})\pi^{-2}\pi^*$. However, it is evident from equation (3.7) that for Δ_Γ close to 0, $\hat{\boldsymbol{\theta}}^{SR}$ performs better than $\hat{\boldsymbol{\theta}}$. But as Δ_Γ moves away from the origin, the risk of $\hat{\boldsymbol{\theta}}^{SR}$ grows and becomes unbounded while the risk of $\hat{\boldsymbol{\theta}}$ remains constant. Hence, the departure from the null hypothesis is exceedingly important to $\hat{\boldsymbol{\theta}}^{SR}$ but is of least concern to $\hat{\boldsymbol{\theta}}$. Further,

$$\frac{\Re(\hat{\boldsymbol{\theta}}^{SR}; \boldsymbol{\theta})}{\Re(\hat{\boldsymbol{\theta}}; \boldsymbol{\theta})} \leq 1 \quad \text{whenever} \quad \Delta_\Gamma \leq \pi^{-2}\pi^* tr(\boldsymbol{\Gamma}\mathbf{L}).$$

Hence, $\hat{\boldsymbol{\theta}}^{SR}$ has smaller risk than $\hat{\boldsymbol{\theta}}$ in the interval $[0, \pi^{-2}\pi^* tr(\boldsymbol{\Gamma}\mathbf{L})]$. Clearly, when Δ_Γ moves away from the null hypothesis beyond $\pi^{-2}\pi^* tr(\boldsymbol{\Gamma}\mathbf{L})$, the risk of $\hat{\boldsymbol{\theta}}^{SR}$ grows without a bound. Now, analyzing the risk of $\hat{\boldsymbol{\theta}}^{SP}$ relative to $\hat{\boldsymbol{\theta}}$ and noticing that

$$\frac{\Re(\hat{\boldsymbol{\theta}}^{SP}; \boldsymbol{\theta})}{\Re(\hat{\boldsymbol{\theta}}; \boldsymbol{\theta})} \leq 1 \quad \text{if}$$

$$\Delta_\Gamma \leq \frac{tr(\boldsymbol{\Gamma}\mathbf{L})\pi^* H_{\nu_1,m}(c^*_{(\alpha)}; \Delta)}{\{2\pi H_{\nu_1,m}(c^*_{(\alpha)}; \Delta) - \pi^* H_{\nu_2,m}(c^o_{(\alpha)}; \Delta)\}}.$$

First, we note that $H_{q,m}(c^*_{(\alpha)}; \Delta)$ is decreasing function of q and Δ. For fixed y,

$$lim_{q\to\infty} H_{q,m}(y; \Delta) = lim_{\Delta\to\infty} H_{q,m}(y; \Delta) = 0$$

In particular, we have

$$H_{q+1,m}(c^*_{(\alpha)}; \Delta) < H_{q,m}(c^*_{(\alpha)}; \Delta) \leq H_{q,m}(c^*_{(\alpha)}; 0),$$

for every $q \geq 3$, $\alpha \in (0,1)$ and $\Delta > 0$. Also,

$$H_{q+1,m}(c^*_{(\alpha)}; \Delta) > H_{q+3,m}(c^o_{(\alpha)}; \Delta),$$

for every $q \geq 3$, $\alpha \in (0,1)$ and $\Delta > 0$. Making use of these results in (3.8), we examine the properties of $\hat{\boldsymbol{\theta}}^{SP}$. For $\boldsymbol{\delta} = \mathbf{0}$ the risk of $\hat{\boldsymbol{\theta}}^{SP}$ reduces to $\left\{tr(\boldsymbol{\Gamma}\boldsymbol{\Sigma}) - \pi^* H_{\nu_1,m}(c^*_{(\alpha)}; 0) tr(\boldsymbol{\Gamma}\mathbf{L})\right\}$ which is substantially smaller than the risk of $\hat{\boldsymbol{\theta}}$. The risk reduction depends on the value of α and π. Not only that, for large deviations of $\boldsymbol{\delta}$ from $\mathbf{0}(\Delta \to \infty)$, i.e., departing from the null hypothesis, the risk of $\hat{\boldsymbol{\theta}}^{SP}$ approaches to the risk of $\hat{\boldsymbol{\theta}}$ from the above. Similar results hold when $k \to \infty$. In addition, as $\boldsymbol{\delta}$ moves away from $\mathbf{0}$, the value of the risk of $\hat{\boldsymbol{\theta}}^{SP}$ increases to a maximum after crossing the risk of $\hat{\boldsymbol{\theta}}$, then decreases towards it.

In order to compare $\hat{\boldsymbol{\theta}}^{SR}$ with $\hat{\boldsymbol{\theta}}^{SP}$, from (3.7) -(3.8) we observe that near the null hypothesis $\hat{\boldsymbol{\theta}}^{SR}$ performs better than $\hat{\boldsymbol{\theta}}^{SP}$, further

$$\frac{\Re(\hat{\boldsymbol{\theta}}^{SR};\boldsymbol{\theta})}{\Re(\hat{\boldsymbol{\theta}}^{SP};\boldsymbol{\theta})} \le 1 \quad \text{if}$$

$$\Delta_\Gamma \le \frac{\pi^*\{1-H_{\nu_1,m}(c^*_{(\alpha)};\Delta)\}tr(\boldsymbol{\Gamma}\mathbf{L})}{\pi^2-\{2\pi H_{\nu_1,m}(c^*_{(\alpha)};\Delta)-\pi^* H_{\nu_2,m}(c^o_{(\alpha)};\Delta)\}}.$$

Thus in light of the above discussion, none of the estimators $\hat{\boldsymbol{\theta}}$, $\hat{\boldsymbol{\theta}}^{SR}$ and $\hat{\boldsymbol{\theta}}^{SP}$ is inadmissible with respect to the other.

For some numerical work let us consider a special choice of $\boldsymbol{\Gamma} = \boldsymbol{\Sigma}^{-1}$ in the loss function and then the remaining discussion follows. In this case, $\Delta_\Gamma = \boldsymbol{\delta}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}$ and hence $\Delta_\Gamma = \Delta$. Further, $tr(\boldsymbol{\Gamma}\boldsymbol{\Sigma}) = k$ and $tr(\boldsymbol{\Gamma}\mathbf{L}) = k^*$. Then

$$\Re(\hat{\boldsymbol{\theta}}^{SR};\boldsymbol{\theta}) \le \Re(\hat{\boldsymbol{\theta}}^{SP};\boldsymbol{\theta}) \quad \text{if} \quad \Delta \le \pi^{-2}\pi^* k^*,$$

while

$$\begin{aligned}\Re(\hat{\boldsymbol{\theta}}^{SP};\boldsymbol{\theta}) \le \Re(\hat{\boldsymbol{\theta}};\boldsymbol{\theta}) \quad &\text{if} \\ &\Delta \le k^*\pi^* H_{\nu_1,m}(c^*_{(\alpha)};\Delta) \\ &\quad \{2\pi H_{\nu_1,m}(c^*_{(\alpha)};\Delta) - \\ &\quad \pi^* H_{\nu_2,m}(c^o_{(\alpha)};\Delta)\}^{-1}.\end{aligned}$$

The risk of $\hat{\boldsymbol{\theta}}^{P}$ can be easily deduce from (3.8) as:

$$\begin{aligned}\Re(\hat{\boldsymbol{\theta}}^{P};\boldsymbol{\theta}) =& \Re(\hat{\boldsymbol{\theta}};\boldsymbol{\theta}) - k^* H_{\nu_1,m}(c^*_{(\alpha)};\Delta) + \\ & \Delta\{2H_{\nu_1,m}(c^*_{(\alpha)};\Delta) - H_{\nu_2,m}(c^o_{(\alpha)};\Delta)\}.\end{aligned} \tag{4.1}$$

It is worth pointing out that $\Re(\hat{\boldsymbol{\theta}}^{P};\boldsymbol{\theta}) \le \Re(\hat{\boldsymbol{\theta}};\boldsymbol{\theta})$ if

$$\begin{aligned}\Delta \le & k^* H_{\nu_1,m}(c^*_{(\alpha)};\Delta) \\ & \{2H_{\nu_1,m}(c^*_{(\alpha)};\Delta) - H_{\nu_2,m}(c^o_{(\alpha)};\Delta)\}^{-1}.\end{aligned}$$

Thus $\hat{\boldsymbol{\theta}}^{SP}$ provides data analyst a larger portion in the parameter space in which it dominates $\hat{\boldsymbol{\theta}}$ than that of $\hat{\boldsymbol{\theta}}^{P}$. As an example, if $\alpha \to 0$, then $\hat{\boldsymbol{\theta}}^{SP} \to \hat{\boldsymbol{\theta}}^{SR}$ and hence $\hat{\boldsymbol{\theta}}^{SP}$ dominates $\hat{\boldsymbol{\theta}}$ in the interval $[0, k^*\pi^{-2}\pi^*]$ while $\hat{\boldsymbol{\theta}}^{P}$ performs better than $\hat{\boldsymbol{\theta}}$ in the interval $[0, k^*]$.

Finally, we compare the risk of the $\hat{\boldsymbol{\theta}}^{SP}$ and $\hat{\boldsymbol{\theta}}^{P}$ and examine the conditions under which $\hat{\boldsymbol{\theta}}^{SP}$ dominates $\hat{\boldsymbol{\theta}}^{P}$. Consider the risk difference of $\hat{\boldsymbol{\theta}}^{SP}$ and $\hat{\boldsymbol{\theta}}^{P}$:

$$\Re(\hat{\boldsymbol{\theta}}^{SP};\boldsymbol{\theta}) - \Re(\hat{\boldsymbol{\theta}}^{P};\boldsymbol{\theta}) = k^*(1-\pi)^2 H_{\nu_1,m}(c^*_{(\alpha)};\Delta) - \Delta$$
$$\left\{2(1-\pi)H_{\nu_1,m}(c^*_{(\alpha)};\Delta) - (1-\pi)^2 H_{\nu_2,m}(c^o_{(\alpha)};\Delta)\right\}.$$

It is obvious from the above expression that risk of $\hat{\boldsymbol{\theta}}^{P}$ will be smaller than the risk of $\hat{\boldsymbol{\theta}}^{SP}$ in the neighborhood of the null hypothesis. However, the risk improvement may not be substantial for the larger values of π. More significantly, for large values of Δ, $\hat{\boldsymbol{\theta}}^{SP}$ dominates $\hat{\boldsymbol{\theta}}^{P}$ in the remaining part of the parameter space, and

$$\frac{\Re(\hat{\boldsymbol{\theta}}^{SP};\boldsymbol{\theta})}{\Re(\hat{\boldsymbol{\theta}}^{P};\boldsymbol{\theta})} \le 1 \quad \text{if}$$

$$\Delta \ge \frac{(1-\pi)k^* H_{\nu_1,m}(c^*_{(\alpha)};\Delta)}{2H_{\nu_1,m}(c^*_{(\alpha)};\Delta) - (1-\pi)H_{\nu_2}(c^o_{(\alpha)};\Delta)}.$$

Let Δ_π be the point at which $\Re(\hat{\boldsymbol{\theta}}^{SP};\boldsymbol{\theta})$ and $\Re(\hat{\boldsymbol{\theta}}^{P};\boldsymbol{\theta})$ intersect for each π, then for $\Delta \in [0, \Delta_\pi)$, $\hat{\boldsymbol{\theta}}^{P}$ dominates $\hat{\boldsymbol{\theta}}^{SP}$, while for $\Delta \in [\Delta_\pi, \infty)$, $\hat{\boldsymbol{\theta}}^{SP}$ dominates $\hat{\boldsymbol{\theta}}^{P}$. However, both $\hat{\boldsymbol{\theta}}^{SP}$ and $\hat{\boldsymbol{\theta}}^{P}$ share a common property that as $\Delta \to \infty$ their risk converges to the common limit, i.e., to the risk of $\hat{\boldsymbol{\theta}}$ while the risk of $\hat{\boldsymbol{\theta}}^{SR}$ is unbounded in Δ.

Generally, when Δ is close to 0, $\hat{\boldsymbol{\theta}}^{SR}$ performs better than both $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}^{SP}$. However, when Δ moves away from the null hypothesis, the risk of $\hat{\boldsymbol{\theta}}^{SR}$ increases and become unbounded while the risk of $\hat{\boldsymbol{\theta}}$ is independent of Δ and risk of $\hat{\boldsymbol{\theta}}^{SP}$ is bounded in Δ. Hence, departure from the null hypothesis is fatal to $\hat{\boldsymbol{\theta}}^{SR}$. Thus, $\hat{\boldsymbol{\theta}}^{SP}$ has an edge over $\hat{\boldsymbol{\theta}}^{SR}$ with respect to the risk. We conclude that none of the four estimators $\hat{\boldsymbol{\theta}}$, $\hat{\boldsymbol{\theta}}^{SR}$, $\hat{\boldsymbol{\theta}}^{P}$ and $\hat{\boldsymbol{\theta}}^{SP}$ of $\boldsymbol{\theta}$ is inadmissible with respect to any one of the other three.

Under the null hypothesis in (1.1), the risks of $\hat{\boldsymbol{\theta}}^{SR}$, $\hat{\boldsymbol{\theta}}$, $\hat{\boldsymbol{\theta}}^{P}$ and $\hat{\boldsymbol{\theta}}^{SP}$ may be ordered according to the magnitude of the risks in the following theorem.

Theorem 4.1: Under the null hypothesis dominance picture of the estimators is:

$$\hat{\boldsymbol{\theta}}^{SR} \succ \hat{\boldsymbol{\theta}}^{P} \succ \hat{\boldsymbol{\theta}}^{SP} \succ \hat{\boldsymbol{\theta}}, \qquad \text{for a range of } \pi,$$

where $\succ$ denotes domination.

Proof: Consider risk difference under the null hypothesis

$$\Re(\hat{\boldsymbol{\theta}};\boldsymbol{\theta}) - \Re(\hat{\boldsymbol{\theta}}^{SP};\boldsymbol{\theta}) = k^*\pi^* H_{\nu_1,m}(c^*_{(\alpha)};0)$$
$$\Re(\hat{\boldsymbol{\theta}}^{SP};\boldsymbol{\theta}) - \Re(\hat{\boldsymbol{\theta}}^{P};\boldsymbol{\theta}) = k^*(1-\pi)^2 H_{\nu_1,m}(c^*_{(\alpha)};0)$$
$$\Re(\hat{\boldsymbol{\theta}}^{P};\boldsymbol{\theta}) - \Re(\hat{\boldsymbol{\theta}}^{SR};\boldsymbol{\theta}) = k^*\left\{\pi^* - H_{\nu_1,m}(c^*_{(\alpha)};0)\right\},$$

the risk difference is positive in all above cases and the last relation holds whenever $\pi \ge 1 - \sqrt{1 - H_{\nu_1,m}(c^*_{(\alpha)};0)}$.

As we observed that $\hat{\theta}^{SP}$ is function of α and π. One method to determine α and π is to employ an *optimal* rule given by Ahmed (1992a) among others which is discussed in the following section.

5. RISK EFFICIENCY ANALYSIS

The *relative efficiency* of the two estimators of θ defined by

$$E(\hat{\theta}^{*} : \hat{\theta}^{o}) = \Re(\hat{\theta}^{o}; \theta)/\Re(\hat{\theta}^{*}; \theta).$$

Bear in mind that the value of $E(\hat{\theta}^{*} : \hat{\theta}^{o})$ greater than 1 signifies improvement of $\hat{\theta}^{*}$ over $\hat{\theta}^{o}$. The purpose of calculating the relative efficiency is two fold; one is to see the performance of estimators and the other is to determine the significance levels of the preliminary tests. Usually the investigator wishes to use an estimator with the efficiency larger than unity. The relative efficiency of $\hat{\theta}^{SP}$ compared to $\hat{\theta}$ is given by

$$E(\alpha, \Delta, \pi) = \Re(\hat{\theta}; \theta)/\Re(\hat{\theta}^{SP}; \theta) = \frac{1}{1 + h(\Delta)},$$

where

$$\begin{aligned} h(\Delta) =& \frac{\Delta}{k}\{2\pi H_{\nu_1,m}(c^{*}_{(\alpha)}; \Delta) - \pi^{*} H_{\nu_2,m}(c^{o}_{(\alpha)}, \Delta)\} \\ &- k^{-1} k^{*} \pi^{*} H_{\nu_1,m}(c^{*}_{(\alpha)}; \Delta). \end{aligned} \tag{5.1}$$

For a given n and k, efficiency is a function of α, Δ, π and the maximum of the function occurs at $\Delta = 0$ with value

$$E_{max} = \{1 - k^{-1} k^{*} \pi^{*} H_{\nu_1,m}(c^{*}_{(\alpha)}; 0)\}^{-1} (> 1). \tag{5.2}$$

Noting that for fixed values of α and π, $E(\alpha, \Delta, , \pi)$ decreases as Δ increases from 0, crossing the line $E(\alpha, \Delta, , \pi) = 1$, attains a minimum value at a point Δ_o and then increases asymptotically to 1. However, for fixed π, E_{max} is a decreasing function of α while the minimum efficiency (E_{min}) is an increasing function of α. Alternatively, for any fixed α, the maximum value of efficiency is a decreasing function of π and the minimum efficiency is an increasing function of π. The shrinkage factor π may also be viewed as a variation controlling factor among the maximum and minimum efficiencies.

In an effort to help the user in choosing an estimator with maximum relative efficiency, we adopt the following procedure. If $\Delta \leq k^{*}\pi^{-2}\pi^{*}$, then $\hat{\theta}^{SR}$ may be used because the behavior of $\hat{\theta}^{SR}$ is superior in this range as compared to the other estimators discussed here. Generally, Δ and γ are unknown, there is no way of choosing a uniformly best estimator.

In order to determine the size of the preliminary test we need to pre-determine a value of the *minimum efficiency* (E_{min}) that we are going to accept. Consider the set

$$W = \{\alpha, \pi | E(\alpha, \pi, \Delta) \geq E_{min}, \ \forall \Delta\}.$$

Thus, the estimator is chosen which maximizes $E(\alpha, \pi, \Delta)$ over all $\alpha, \pi \in W$ and Δ. Thus, we solve for α^{*} and π^{o} such that

$$\sup_{\alpha,\pi \in W} \left\{ \inf_{\Delta} E(\alpha, \pi, \Delta) \right\} = E(\alpha^{*}, \pi^{o}, \Delta) = E_{min}. \tag{5.3}$$

For given π we determine the value of α such that

$$\sup_{\alpha \in W} \left\{ \inf_{\Delta} E(\alpha, \pi, \Delta) \right\} = E(\alpha^{*}, \pi, \Delta) = E_{min}.$$

Tables have been prepared for the values of E_{max}, E_{min} along with Δ_{min}, the value of Δ at which minimum occurs, for various values of k and $\pi = 0.2(0.2)1.0$. However, the tables are not appended here due to page limit and may be obtained from the first authour.

It is observed that when π increases E_{max} increases and E_{min} decreases. Hence, there does not exist a π^{o} satisfying (5.3). The value of π can be determined by the researcher according to his prior belief in the null hypothesis. However, we recommend the following two steps for selecting the size of the preliminary test:

1. Suppose the experimenter does not know the size of the test but believes that $\pi = \pi_o$ and wishes to accept an estimator with at least efficiency E_{min}. Then the max-min principle determines $\alpha = \alpha^{*}$ such that

$$E(\alpha^{*}, \pi_o, \Delta) = E_{min}.$$

Therefore, a user who wishes to find a reasonable alternative to the $\hat{\theta}$ or $\hat{\theta}^{SR}$ then should be able to specify E_{min}.

As an example, if the investigator suspects that $\pi = 0.6$ and is looking for an estimator with a minimum efficiency of at least 0.73, then from Table 3, α^{*} is 0.10. Such a choice of α^{*} would yield an estimator with a maximum efficiency of 1.89 at $\Delta = 0$ with a minimum guaranteed efficiency of 0.73 at $\Delta_{min} = 7.60$. Alternatively, if the user wishes to rely on data completely and uses $\hat{\theta}^{P}$, then from one of the Tables the size of the preliminary test will be approximately 0.20 and the maximum efficiency drops from 1.89 to 1.75. Thus, the use of $\hat{\theta}^{P}$

may be limited by the large size of α, the level of significance, as compared to $\hat{\boldsymbol{\theta}}^{SP}$. Hence, $\hat{\boldsymbol{\theta}}^{SP}$ has a remarkable edge over $\hat{\boldsymbol{\theta}}^{P}$ with respect to the size of the preliminary test.

2. Suppose the experimenter does not know the value of α and π and wants to use $\hat{\boldsymbol{\theta}}^{SP}$ which has efficiency at least E_{min}. Then, among the set of estimators with $\alpha \in \mathcal{W}$, where $\mathcal{W} = \{\alpha : E(\alpha, \gamma, \Delta) \geq E_{min} \; \forall \; \Delta \text{ and } \pi\}$, the estimator is chosen to maximize the efficiency over all $\alpha \in \mathcal{W}$ and all Δ, π. In short, we use the following equation for α:

$$\sup_{\alpha \in \mathcal{W}} \left\{ \inf_{\Delta} E(\alpha, \pi, \Delta) \right\} = E(\alpha^*, \pi, \Delta) = E_{min}, \forall \; \pi.$$

The solution α^* is the optimum level of significance. Tables may be used again for finding α^*. For example, suppose $\pi \in \prod$, where

$$\prod = \{\pi : \pi = 0.2(0.2)1.0\},$$

and the experimenter does not know the value of π but is willing to use an estimator with efficiency at least E_{min}. Then $\alpha = \alpha^*$ can be located from the tables using the above method. As a result, the user can attain efficiency larger than 1.

4. CONCLUSIONS AND OUTLOOK

We conclude this article with the following remarks. The proposed estimator $\hat{\boldsymbol{\theta}}^{SP}$ dominates the usual preliminary test estimator $\hat{\boldsymbol{\theta}}^{P}$ in a larger portion of the parameter space. Besides it provides more appropriate size for the preliminary test. More significantly, $\hat{\boldsymbol{\theta}}^{SP}$ gives a wider range in which it performs better than $\hat{\boldsymbol{\theta}}$ as compared to $\hat{\boldsymbol{\theta}}^{P}$.

The distribution theory of $\hat{\boldsymbol{\theta}}^{SP}$ and risk of all the estimators studied in this paper rests on the multinormality of the unrestricted and the restricted maximum likelihood estimators as well as on the noncentral F-distribution of test statistic.

ACKNOWLEDGEMENTS

The research of first author has been supported by the NSERC grant and President's NSERC grant, University of Regina. The second author, who is on leave from the Peshawar University, Pakistan, is grateful to the Canadian Common-Wealth scholorship plan for providing scholorship for postgraduate studies at the University of Regina.

REFERENCES

Ahmed, S.E. (1991). To pool or not to pool: The discrete data Probability and Statistics Letters **11**, 233-237.

Ahmed, S. E.(1992a). Shrinkage preliminary test estimation in multivariate normal distributions. Journal of Statistical Computation and Simulation **43**, 177-195.

Ahmed, S. E.(1992b). Large-sample pooling procedure for correlation. The Statistician **41**, 425-438.

Ahmed, S. E. and A.K.M.E. Saleh (1990). Estimation strategies for the intercept vector in a simple linear multivariate regression model. Journal of Computational Statistics and Data Analysis **10**, 193-206.

Ali, M. A. (1990). Interface of preliminary test approach and empirical Bayes approach to shrinkage estimation. Unpublished Ph.D. thesis, Carleton University, Canada.

Bancroft, T.A. (1944). On biases in estimation due to use of preliminary test of significance. Annals of Statistics, 15 190-204.

Bancroft, T.A. and C.P. Han (1977). Inference based on conditional specification: A note and a bibliography. International Statistical Review, 45, 117-127.

Gupta A. K., A.K.M.E. Saleh and P.K. Sen (1989). Improved estimation in a contingency table: independence structure. Journal of American Statistical Association **84**, 525-532.

Han, C.P., Rao, C.V. and J. Ravichandran (1988). Inference based on conditional specification: A second bibliography. Communications in Statistics - Theory and Methods, 17, 1945-1964.

Judge, G. G. and M. E. Bock (1978). The statistical implication of pre-test and Stein-rule estimators in econometrics. Amsterdam: North-Holland.

Kulperger, R. J. and S.E. Ahmed. (1992). A bootstrap theorem for a preliminary test estimator. Communications in Statistics–Theory and Methods **21**, 2071-2082.

Thompson, J.C. (1968). "Some shrinkage techniques for estimating the mean. Journal of American Statistical Association **63**, 113-122.

UNBIASED ESTIMATION IN THE PRESENCE OF FRAME DUPLICATION

Orrin Musser, National Agricultural Statistics Service, USDA
Research Division, Room 305, 3251 Old Lee Hwy, Fairfax, VA 22030

INTRODUCTION

Survey organizations which conduct surveys on an ongoing basis devote much effort and expense to the maintenance of their sampling frame. Estimation of population parameters may suffer from two main types of frame deficiency: incomplete population coverage and duplication. In this paper we will focus on the problem of duplication, with emphasis on the computation of correct inclusion probabilities as a means to achieve unbiased estimation.

Duplication in the sampling frame is a serious problem which undermines the assumption of known inclusion probabilities for each population element. For a large sampling frame, while it may be too costly to determine all duplication in the frame, it may be reasonable to assume that for a given population element it may be possible to determine all duplicates in the frame. If so, then for many sampling designs, unbiased estimation is possible.

A sampling frame is a device which associates a collection or list of sampling units with a finite population of elements. It is helpful to formally describe the relationship between the sampling frame and the population. Suppose we have a population U = $\{E_1, E_2,\ldots,E_{k,}\ldots,E_N\}$, a collection of N elements E_k and a sampling frame F = $\{F_1, F_2,\ldots,F_{i,} \ldots,F_M\}$, a collection of M sampling units F_i. For each unit F_i and each element E_k, let the indicator variable δ_{ik} be defined:

$$\delta_{ik} = \begin{cases} 1 \textit{ if } F_i \textit{ represents } E_k \\ 0 \textit{ otherwise.} \end{cases}$$

And let $M_k = \sum_i \delta_{ik}$ be the count of frame units which represent or "link to" population element k. We will call the collection or set of frame units linked to population element k, link-group k. Duplication exists in the frame when there are some population elements which are linked to more than one frame unit, that is $M_k > 1$ for some k. We assume that while M_k's are unknown (difficult and/or expensive to determine for a large entire population) M_k can be determined exactly for a particular unit k and thus for a sample. Assume a simple random sample of size m without replacement. We will denote this sample of frame units by s. For each sampled unit, we obtain a listing of all frame units that link to it. This set of frame units represents a single unique population element k. Since other members of this linkage group could have been sampled, it is possible to sample a population element k more than once. If we think of our sampling as sampling without replacement of population elements, we also obtain a sample s_P which contains $n(\leq m)$ distinct population elements. While each frame unit in our original sample s had equal probability of selection, each population element in our sample s_P did not have equal probability of selection, and thus estimators which assume we are sampling population elements with equal probability will be biased if there is duplication in the frame.

2. CORRECT INCLUSION PROBABILITIES

We know that the Horwitz-Thompson estimator:

$$\hat{Y}=\sum_{i=1}^{n}\frac{y_k}{\pi_k}$$

is an unbiased estimator of the population total, Y. Thus if we can compute π_k for each sampled element k, we can get unbiased estimates for population size and in general for any variable of interest, even in the presence of duplication. The inclusion probabilities, π_k, are straightforward to calculate if we know M_k for each sampled unit. If we are interested in the probability of selection for population unit k, or equivalently linkage group k, of size M_k, we use the fact that this is equivalent to 1 minus the probability of selecting a sample of size m from the frame such that no units of linkage group k were selected. This is just the ratio of the number of possible samples of size m chosen from the (M - M_k) frame units which do not include any member of linkage group k over the number of possible samples:

$$\pi_k=1-\frac{\binom{M-M_k}{m}}{\binom{M}{m}}$$

Example: If we take a sample of m=5 frame unit from a frame of size M=100 in which there is duplication, what is the probability that our sample s_P of distinct population elements contains a particular population unit k for which $M_k = 1, 2$?

For $M_k = 1$(Population element k is represented only once on the frame):

$$\pi_k=1-\frac{\binom{99}{5}}{\binom{100}{5}}$$
$$=1-\frac{99!}{94!5!}\,\frac{95!5!}{100!}$$
$$=1-\frac{95}{100}$$
$$=\frac{5}{100}=\frac{m}{M}.$$

This is true in general, for each population unit for which there is no duplication, i.e. $M_k=1$, the probability of selection is just m/M, or f, to be denoted as π^*. (Recall that since we may sample a given population element more than once, n, the number of distinct population elements sampled, is a random variable and $\pi_k \neq n/N$.)

$M_k = 2$: (Population element k is represented twice in the frame)

$$\pi_k=1-\frac{\binom{98}{5}}{\binom{100}{5}}$$
$$=1-\frac{98!}{93!5!}\,\frac{95!5!}{100!}$$
$$=1-\frac{95}{100}\,\frac{94}{99}$$
$$=.0979797$$

Note that this probability is **not** double the selection probability for a population unit without duplication. It is interesting to look at this ratio of selection probabilities in general. Looking at the general formula for the selection probability when $M_k=2$, we may express π_k approximately as a function of the sampling fraction, f = m/M:

$$\pi_k=1-\frac{\binom{M-M_k}{m}}{\binom{M}{m}}$$
$$=1-\frac{(M-M_k)!}{(M-M_k-m)!m!}\,\frac{(M-m)!m!}{M!}$$
$$=1-\frac{M-m}{M}\,\frac{M-m-1}{M-1}$$
$$\approx 1-(1-f)^2=2f(1-\frac{f}{2}).$$

Thus the ratio $r_k = \pi_k/\pi^*$ where $M_k = 2$, and π^* is the probability of selection for any population element for which there is no duplication, may be expressed:

$$r_k=\frac{\pi_k}{\pi^*}\approx\frac{2f(1-\frac{f}{2})}{f}=2(1-\frac{f}{2}).$$

Thus, as the sampling fraction approaches 0, r_k approaches two. As f gets large and approaches 1, r_k approaches 1. Thus as the likelihood of selection gets smaller, the bias due to incorrect assumptions of equal probabilities is increased. This ratio r_k is interesting because it expresses the degree to which the data y_k is "over expanded" due to the assumption of known equal selection probability. In our example where f = .05, the "pi estimator" would over expand y_k by a factor of .09797/.05 = 1.96. If N were 10(f= 1/2), then r_k will be approximately 1.5 and thus estimates which ignore duplication will over-expand data for elements with one duplicate by a factor of 1.5. If N were 10,000 then r_k would be essentially 2.

Since we assume that we may determine M_k for any population element k, then clearly we may compute π_k for each sampled element and thus use the Horwitz-Thompson estimator to obtain unbiased estimates for population totals and means. If we define $y_k=1$ for each population element k, then we could obtain an unbiased estimate of N, the true population size. This is just the sum of the reciprocals of the inclusion probabilities for the n distinct population elements in s_P. This estimator is unbiased for N:

$$\hat{N}=\sum_{i=1}^{n}\frac{1}{\pi_k}$$

$$E[\hat{N}]=E[\sum_{i=1}^{n}\frac{1}{\pi_k}]$$
$$=\sum_{i=1}^{N}E[\frac{1}{\pi_k}I_k]$$
$$where\ I_k=1\ if\ k\in s_p$$
$$=\sum_{i=1}^{N}\frac{1}{\pi_k}E[I_k]$$
$$=\sum_{i=1}^{N}\frac{1}{\pi_k}\pi_k=\sum_{i=1}^{N}1=N\ .$$

The variance of the Horvitz-Thompson estimator of a population total Y is given by

$$V(\hat{Y}_\pi)=\sum\sum\nolimits_U(\pi_{kl}-\pi_k\pi_l)\frac{y_ky_l}{\pi_k\pi_l}$$

and an unbiased estimator of the variance is given by

$$\hat{V}(\hat{Y}_\pi)=\sum\sum\nolimits_s\frac{(\pi_{kl}-\pi_k\pi_l)}{\pi_{kl}}\frac{y_ky_l}{\pi_k\pi_l}$$

(Sarndal, Swensson, and Wretman, section 2.8). These general formulae are very useful for this situation where duplication results in unequal selection probabilities.

The second order inclusion probabilities needed for these formulae may be determined for the sample by the following formula which uses analogous reasoning to that for the first order inclusion probabilities.

$$\pi_{kl}=P((k\in s)\cap(l\in s))$$
$$=1-P((k\notin s)\cup(l\notin s))$$
$$=1-\frac{\binom{M-M_k}{m}}{\binom{M}{m}}-\frac{\binom{M-M_l}{m}}{\binom{M}{m}}+\frac{\binom{M-(M_k+M_l)}{m}}{\binom{M}{m}}$$

3. ALTERNATIVE STRATEGIES

One alternative "adjustment" for list duplication, and one that is currently used by NASS surveys, is the common survey practice of using a weight or data adjustment factor to account for the effect of duplication. If a population element k appears on the sampling frame M_k times, then when sampled the data is multiplied by $1/M_k$. Even if the same population element appears multiple times in the sample, every sampled unit reports. Cox(1993) describes this procedure as an adjustment of the weight "associated with sampled frame units to reflect the multiple selection opportunities for the desired population unit." This adjustment obtained by multiplying the sampling weight M/m, and the adjustment $1/M_{k,}$ results in an overall weight of $M/(m^*M_k)$. This new weight is not, in general, equal to the reciprocal of the probability of selection. As shown earlier the ratio r_k depends on the sampling fraction. Nonetheless, this procedure does result in unbiased estimation.

Suppose x is a data item with x_k being the data for each true population element k. Then for each frame unit l which is linked to element k, we define $y_{kl} = x_k/M_k$, $l = 1$ .. M_k. Thus we are letting each frame unit account for the proportion, $1/M_k$, of the data for population element k. Clearly the total of the y's is equal to the total of the x's:

$$\sum_{i=1}^{M}y_i=\sum_{k=1}^{N}\sum_{l=1}^{M_k}y_{kl}$$
$$=\sum_{k=1}^{N}\sum_{l=1}^{M_k}\frac{x_k}{M_k}$$
$$=\sum_{k=1}^{N}M_k\frac{x_k}{M_k}$$
$$=\sum_{k=1}^{N}x_k$$

Thus a reasonable estimate for X would be

$$\hat{Y}=\sum_{i=1}^{m}\frac{M}{m}y_i\ .$$

Note again that this sum is over the entire sample of frame units. This is clearly unbiased for X, since this approach is equivalent to a simple random sample with the frame being the population. Thus $\hat{Y}$ is unbiased for $Y = X$.

This technique really obtains unbiased estimation by redefining the relationship of the frame to the population. If a population element appears M_k times on the frame, then each of those M_k records accounts only for the proportion $1/M_k$ of the data for population element k. This eliminates the duplication of the data.

Another approach used to obtain unbiased estimation in the presence of frame duplication is to define a "unique counting rule" which links each population element to a single frame unit. An example would be to link each population element k to the frame unit in M_k with the largest frame id, etc. In this case, population element k is sampled only if this particular frame unit is selected. Thus, if duplication were detected after data collection, there could be loss of data.

REFERENCES

Cox, B. (1993), "Weighting Class Adjustments for Nonresponse in Integrated Surveys: Framework for Hog Estimation," NASS Research Report SRB-93-03.

Lessler, J., Kalsbeek, W. (1992) *Nonsampling Error in Surveys*, New York: John Wiley.

Sarndal, C., Swensson, B., Wretman, J. (1992), *Model Assisted Survey Sampling*, New York: Springer-Verlag.

AN EVALUATION OF ROUNDING TECHNIQUES USED TO OBTAIN ESTIMATED TOTALS

Melinda Kraus and Franklin Winters, U.S. Bureau of the Census
Melinda Kraus, U.S. Bureau of the Census, Agriculture Division, Research and Methods, Washington, DC 20233

KEY WORDS: Relative efficiency, random variable weights, unbiased estimates, mean square error.

ABSTRACT. Estimators in a complex survey usually have noninteger weights which must be rounded to integer weights to help facilitate data review. Using integer weights also eliminates rounding discrepancies between tabulated sample data for a given area. This evaluation focuses on the effects of systematically rounding noninteger sample weights to integer weights to produce estimates for the U.S. Census of Agriculture. The relative efficiency of the rounding technique is measured to determine if there is any potential increase in the variance of the estimated totals incurred by utilizing integer weights. This evaluation pointed to some factors which effect the performance of the systematic rounding process, and it indicated that the over and under estimating occurring at the strata level balances out at the published level. The systematic rounding technique increases the variances, but has a limited effect on the estimated totals at the published levels.

1. INTRODUCTION

Several investigators at the Bureau of the Census have studied the process of systematically rounding noninteger weights to integer weights and its effect on the variance of estimated totals. This procedure has also, in previous research, been referred to as "Weighting Random Subsamples of an Original Sample". However, to date only one study has focused on agriculture data. Hanson (1969) noted that, assuming simple random sampling without replacement, a potential increase in variances of estimated totals resulted from assigning integer weights. Thompson (1978) presented analogous results to those found in the previous study, but included the finite population correction factor that arises in the variance expression for simple random sampling. He noted that in areas where a 1-in-2 sampling rate is used, the effect of the finite population correction could not be ignored. Hanson and Thompson studied this problem in connection with the 1970 and 1980 Census of Population and Housing. Griffin (1987) found that controlled rounding for disclosure avoidance in the 1990 Decennial Census produced significant increases in variance for both 1-in-2 and 1-in-6 sampling rates. Chapman (1981) investigated this problem in connection with the 1978 Census of Agriculture. He noted increases in the variances of between 1.2 % and 3.0 % due to integer weighting. However, the rounding procedure has changed since this study was conducted. This paper reinvestigates the relative loss in precision due to integer weighting for estimated totals from the 1987 Census of Agriculture.

Data for several characteristics across two states from the 1987 Census of Agriculture is used to investigate the effects of the rounding technique. For analysis, the ratio of the variances of the integer and noninteger weights is used as a measure of loss. Before presenting the results of the investigation we first describe the sampling design and weighting procedures used to produce estimates for the census of agriculture, followed by the methodology used to evaluate the rounding technique.

2. BACKGROUND

To reduce respondent burden, a representative sample of the 1987 Census of Agriculture mail list universe is selected to respond to questions deemed to be of a sensitive nature, such as farm production expenditures, value of land and buildings, and other farm-related income. A discussion of the sampling design and weighting procedure follows. Sample weighting is the terminology used by the census of agriculture to describe our method of inflating the sample to represent the census universe. Certainty farms, cases which were expected to have large total value of agricultural products sold or large acreage or other "special" characteristics, are included in the sample with probability equal to one, but are excluded from these weighting processes and assumed to have no sampling variability. All farms in counties containing less than 100 farms were also classified as certainty farms. The other farms in counties containing 100 to 199 farms in 1987 were systematically sampled at a rate of 1 in 2, and farms in counties containing 200 farms or more in 1987 were systematically sampled at a rate of 1 in 6. This differential sampling scheme was used to provide reliable data for all counties.

The weighting procedure produces final sample weights which account for nonresponse and inflate the sample data to represent the census universe. As mentioned in the paragraph above we start out with integer weights (i.e., 1, 2, or 6), for inflating the sample to represent the census universe, but in order to improve the precision of our estimates we use post-stratification and an iterative raking ratio adjustment procedure which in turn produces noninteger weights. The focus here is on the technique we use to round these noninteger weights back to integer.

The post-stratification procedure assigns sample respondents to one of 32 initial sample post-strata (ISPS) based on: total value of products sold (TVP), acres of land owned, and standard industrial classification. Final sample post-strata (FSPS) are assigned after collapsing initial strata which do not meet a specified criteria. Since the sample records are subject to both sampling and nonresponse variability, the final sample weight is a combination of a nonresponse weight and an adjusted sample weight.

After the final sample post-strata have been determined, a base sample weight is computed for each stratum. This is the total number of noncertainty farms divided by the number of noncertainty sample farms.

This base sample weight under goes an iterative raking ratio adjustment which smooths the weights across strata within county and produces an adjusted sample weight. This in turn is multiplied by a nonresponse weight, which accounts for nonresponse to the data collection, to produce the final sample weight.

All records in the census universe, respondent and nonrespondent, are assigned to one of five nonresponse strata, which are defined based on TVP, previous census status and whether the record was identified by a discriminant model as having a low probability of being a farm and therefore would receive the census screener form (i.e., form type is assigned based on the probability of a particular census record being a farm. There are three different types of census report forms, the nonsample census form, the screener form, and the sample form. The sections on the sample form are identical to sections on the nonsample census form. The sample form contains additional questions such as farm production expenditures, value of land and buildings, and farm-related income. The screener form is identical to the nonsample census form with questions added to allow quick identification of nonfarm addresses). Estimates of the proportion of census nonrespondents that operate farms are computed for each stratum in a state using results from the census of agriculture nonresponse survey and applied to the total number of census nonrespondents in that stratum. The number of census nonrespondents that operate farms for each county by stratum is then derived. Within each stratum in a county, a noninteger nonresponse weight is calculated and assigned to each eligible respondent farm record. The noninteger nonresponse weight is the ratio of the sum of the estimated number of nonrespondent farms from the nonresponse survey and the number of eligible census respondent farms to the number of eligible census respondent farms. Stratum controls are established to ensure that this weight is never greater that 2.0. This noninteger nonresponse weight is used in the calculation of the final sample weight for sample items.

The integer weighting process takes place within nonresponse strata (NRS) within final sample post-strata (FSPS) within a county. The resulting value within a stratum is usually a noninteger number of the form $(W + p)$, where W is an integer and $0 \leq p < 1$. All records within a nonresponse strata within a final sample post-strata have the same noninteger weight. We assume, as did the authors of the previous studies, that the sample weighting process involves selecting a simple random sample without replacement of n units from a population of N units and that the integer weighting process involves selecting a random subsample of size n_1, which is equal to $\lceil np \rceil$ with probability $P = np - \lfloor np \rfloor$ or equal to $\lfloor np \rfloor$ with probability $Q = 1 - P$. The symbol $\lceil x \rceil$ denotes the smallest integer greater than or equal to x and $\lfloor x \rfloor$ denotes the largest integer less than or equal to x. Each of these n_1 units is given a weight of $(W + 1)$, where W is equal to integer part of the noninteger value. The weight W is assigned to the remaining $n_2 = n - n_1$ sample units in the stratum. Actually systematic sampling is used, but for estimating the variances due to rounding we are assuming simple random sampling.

Noncertainty nonsample records are assigned an integer sample weight of zero. Noncertainty sample records proceed through the systematic rounding process as shown in the example below.

Consider a FSPS/NRS combination within a county with 25 total records and 7 sample records. Each of the 7 sample records have a noninteger sample weight (NISW) of 6.20 which needs to be converted to integer sample weights. To accomplish this, all 7 records first have their sample weight truncated and a systematic sample of these records is then selected to increase their weights by one to account for the remainder.

The noninteger sample weight is separated into an integer portion, (W), and a fractional portion, the remainder. The sampling interval is the reciprocal of the remainder. Here, the sampling interval is (1/0.20) = 5.00 Multiplying the number of sample cases (7) by the remainder indicates that 1.40 records should be rounded up, which implies that 1 or 2 records will be rounded up. Therefore, the probability is 40 % that a second case will be rounded up to 7 and 60 % that it will remain 6. The integer sample weight is equal to the initial integer weight for five or six of the sample cases. One or two of the seven cases becomes part of the systematic sample which is increased to the initial integer weight plus one (W + 1), to account for the remainder. This implies that n_1 is a random variable. The sampling scheme in Table 1 illustrates the rounding process.

Table 1. Example of systematic rounding

SAMPLE CASE NUMBER	NON-INTEGER SAMPLE WEIGHT (NISW)	INITIAL INTEGER WEIGHT (W)	INTEGER SAMPLE WEIGHT AFTER SYSTEMATIC ROUNDING	
			ONE CASE ROUNDED UP	TWO CASES ROUNDED UP
1	6.20	6	6	6
2	6.20	6	6	7
3	6.20	6	6	6
4	6.20	6	6	6
5	6.20	6	7	6
6	6.20	6	6	6
7	6.20	6	6	7
Total	43.40	42	43	44

3. METHODOLOGY

This study uses the data from the 1987 Census of Agriculture to examine the rounding technique. Explanation of the methodology follows.

Assume sample variable X, has the following stratum total (1), population mean (2) and population variance of the X_i's (3).

$$X = \sum_{i=1}^{N} X_i \tag{1}$$

$$\overline{X_N} = \frac{\sum_{i=1}^{N} X_i}{N} \tag{2}$$

$$S^2 = \frac{\sum_{i=1}^{N} (X_i - \overline{X}_N)^2}{N-1} \tag{3}$$

Here our stratum refers to the level at which the integer weighting take place (i.e., the nonresponse strata within final sample post-strata, FSPS/NRS).

3.1 Estimate and Variance of the Estimated Total Using Integer Weights

We restate the assumptions from the previous section (i.e., that a random subsample of size n_1 is selected from the n original sample units to be given a weight (W + 1) and n_2 is selected to be given a weight W). Applying systematically rounded integer weights to the variable X produces the following estimate of total at the strata level.

$$\hat{X}_I = n_1 \ (W + 1) \ \overline{X}_{n_1} + n_2 \ W \overline{X}_{n_2}$$

where $\overline{X}_{n_1} = \frac{\sum_{i=1}^{n_1} X_i}{n_1}$ and $\overline{X}_{n_2} = \frac{\sum_{i=1}^{n_2} X_i}{n_2}$ are the sample means of n_1 and n_2 respectively.

The covariance of the sample means $\overline{X}_{n_1}$ and $\overline{X}_{n_2}$ is

$$COV(\overline{X_{n_1}}, \overline{X_{n_2}}) = -\frac{S^2}{N}$$

because the samples of size n_1 and n_2 are simple random samples from the population of size N.

The finite population correction factors for the samples n_1 and n_2 are

$$f_{11} = \frac{\lfloor np \rfloor}{N}, \ f_{21} = \frac{n - \lfloor np \rfloor}{N}$$

if the number of sample cases receiving the weight (W+1) is $\lfloor$ np $\rfloor$, and

$$f_{12} = \frac{\lceil np \rceil}{N}, \ f_{22} = \frac{n - \lceil np \rceil}{N}$$

if the number of sample cases receiving the weight (W+1) is $\lceil$ np $\rceil$.

The variance of the integer weighted estimator, $VAR(\hat{X}_I)$, is made up of two components. The first is the variability arising from the fact that the number of cases, n_1, selected to receive a weight (W+1) is a random variable.

$$P\,\overline{X}_N^2\,[\lceil np \rceil (W+1) + (n - \lceil np \rceil) W - n\,(W+p)]^2$$
$$+\,Q\,\overline{X}_N^2\,[\lfloor np \rfloor (W+1) + (n - \lfloor np \rfloor) W - n\,(W+p)]^2$$

The second is variability among the data items which receive integer weights (W) or (W+1), (i.e., the variance arising from the sample cases).

$$P\,[\,(\lceil np \rceil (W+1)^2\,(1 - f_{12})\,S^2)$$
$$+\,((n - \lceil np \rceil)\,W^2\,(1 - f_{22})\,S^2)$$
$$-\,(2\lceil np \rceil (n - \lceil np \rceil)\,W\,(W+1)\,\frac{S^2}{N})]$$
$$+\,Q\,[\,(\lfloor np \rfloor (W+1)^2\,(1 - f_{11})\,S^2)$$
$$+\,((n - \lfloor np \rfloor)\,W^2\,(1 - f_{21})\,S^2)$$
$$-\,(2\lfloor np \rfloor (n - \lfloor np \rfloor)\,W\,(W+1)\,\frac{S^2}{N})]$$

To simplify the work, n and W are considered fixed even though they are random variables.

3.2 Estimate and Variance of the Estimated Total Using Noninteger Weights

Applying noninteger weights to variable X produces the following weighted estimate at the strata level.

$$\hat{X}_{NI} = \frac{N}{n}\sum_{i=1}^{n} X_i$$

The noninteger weighted estimator has a variance of

$$VAR(\hat{X}_{NI}) = (W+p)^2 n(1-f)\,S^2$$

where $f = \frac{n}{N}$ is the finite population correction factor.

As a measure of accuracy we calculated the relative efficiency of the two variances as follows,

$$Relative\ Efficiency = \frac{VAR(\hat{X}_I)}{VAR(\hat{X}_{NI})}.$$

4. RESULTS

Data from the states of Delaware and Massachusetts were used to investigate the effects of the rounding technique. Due to space constraints, only the results from Delaware are listed. The integer weight estimator and the noninteger (constant) weight estimator were compared for three major sample characteristics -- value of land and buildings, interest paid, and total farm production expenses. The differences in estimates of the characteristics and a comparison of their sampling variances were examined. The relative efficiency of the two estimators was measured to determine accuracy.

4.1 Differences in Estimates of Farm Characteristics

In summarizing the results, the focus is first on the effect of systematic rounding on the estimates of several variables at the county level, which is the lowest level at which the census of agriculture publishes data. The results indicate that minor discrepancies at the strata level due to systematic rounding accumulate to sizable differences at the published level for some farm characteristics. Differences between the weighted estimates for several characteristic are shown below at the published level. The percent difference is the ratio of the differences in the estimates. Generally, the difference due to rounding is between ± 3 % at the county level for Delaware and Massachusetts. The tables below show how the discrepancies due to rounding vary by state, county, and variable.

Table 2. Weighted Estimates and Differences of the Total Farm Production Expenses Variable for All Counties in Delaware.

Delaware -- Total Farm Production Expenses ($)				
	Weighted Estimate		Difference	Percent Difference
County	Integer (I)	Noninteger (NI)	DIFF =I-NI	100*DIFF/NI
New Castle	26,843,931	26,852,456.00	- 8,525.00	-0.03
Kent	75,839,137	75,714,197.63	124,939.37	0.17
Sussex	274,418,465	274,434,736.98	- 16,217.98	-0.01
State	377,101,533	377,001,390.61	100,142.39	0.03

The integer weighted estimate for total farm production expenses is over estimated as compared to the noninteger estimates for the state of Delaware. A large part of this is due to Kent county. New Castle and Sussex counties integer weighted estimates are relatively small under estimates.

Table 3. Weighted Estimates and Differences of the Interest Paid Variable for All Counties in Delaware.

Delaware -- Interest Paid ($)				
	Weighted Estimate		Difference	Percent Difference
County	Integer	Noninteger		
New Castle	1,512,831	1,509,379.48	3,451.52	0.23
Kent	4,994,486	4,999,221.60	- 4,735.60	-0.09
Sussex	8,364,715	8,354,747.05	9,967.95	0.12
State	14,872,032	14,863,348.14	8,683.86	0.06

The integer weighted estimate for the interest paid is over estimated as compared to the noninteger estimates for the State of Delaware by 0.06 percent. Kent county is under estimated while New Castle and Sussex counties are over estimated.

Table 4. Weighted Estimates and Differences of the Value of Land and Buildings Variable for All Counties in Delaware.

Delaware -- Value of Land and Buildings ($)				
	Weighted Estimate		Difference	Percent Difference
County	Integer	Noninteger		
New Castle	254,777,140	254,205,301.88	571,838.12	0.22
Kent	344,669,483	344,155,516.99	513,966.01	0.15
Sussex	496,821,534	497,426,164.00	-604,630.00	-0.12
State	1,096,268,157	1,095,786,982.87	481,174.13	0.04

Again, the integer weighted estimate for the value of land and buildings is over estimated as compared to the noninteger estimates for the State of Delaware by 0.04 percent. Sussex county is under estimated while New Castle and Kent counties are over estimated.

The above tables show that all three sample variables are over estimated at the state level while the county level is both over and under estimated. Even though the integer and noninteger weighted estimates have very large absolute difference, the percent difference is less than 1% for both states. Large absolute differences are not significant because the expected values of the integer and noninteger estimates are equal.

Over half of the counties in Massachusetts are over estimated for all three sample variables, total farm production expenses, interest paid, and value of land and buildings. These counties percent differences range from 0.02 to 0.96 percent. Estimates for about 15 % of the counties are identical (i.e., integer vs. noninteger estimates) and about 20 % of the counties are underestimated with percent differences ranging from -3.99 to -0.06 percent. Overall, the difference in weighted estimates for Massachusetts is over estimated at 11 % for total farm production expenses and 31 % for interest paid and value of land and buildings.

4.2 Comparison of Sampling Variances

This section focuses on the effect systematic rounding has on the variance of the estimates. To help understand the results, we refer you back to our sample weighting process which was discussed in the background section. To summarize, the sampling design and weighting procedure used was to assume a simple random sample without replacement of n units from a population of N units. The noninteger weights were assigned independently within FSPS/NRS strata within a county and were of the form (W + p), where W is an integer and $0 \leq p < 1$. A proportion p of the n original sample units (n_1 units) were selected and given the weight (W + 1). The remaining units within the FSPS/NRS combination (n_2 units) were given the weight W.

The tables below shows variances and relative efficiency scores for strata and counties in Delaware. Variances at the county level are assumed independent among strata. These estimates are used for illustrative purposes only. In general, the relative efficiencies ranged from 1.0000 to 1.9705 at the strata level and 1.0023 to 1.0171 at the county level.

Table 5. Relative Efficiency of Sample Variables in Delaware at the County Level.

Delaware			
	Relative Efficiency		
County	Total Farm Production Expenses	Interest Paid	Value of Land and Buildings
New Castle	1.0023	1.0044	1.0043
Kent	1.0171	1.0055	1.0091
Sussex	1.0059	1.0047	1.0074

The counties and variables in Delaware have up to a 2 % increase in variance due to integer rounding. The majority of the strata in Massachusetts have up to a 2 % increase in variance.

Table 6. Variance and Relative Efficiency of Selected Sample Variables in Delaware at the Strata Level.

Variable	Stratum	Variance		Relative Efficiency
		Integer Weight	Noninteger Weight	
Total Farm Production Expenses	New Castle (17, 4)	282,318,282,458	281,859,153,259	1.0016
	Kent (17, 0)	801,339,046	621,919,519	1.2885
Interest Paid	Kent (5, 3)	45,959,184	45,930,944	1.0006
	Sussex (1, 5)	994,416,441	976,977,655	1.0179
	Kent (17, 0)	3,836,062	396,700	9.6699
Value of Land and Buildings	Kent (13, 0)	4,454,982,478	3,780,038,494	1.1786
	Sussex (1, 3)	915,339,717,410	914,451,825,739	1.0010

Each stratum is identified by the county and (FSPS, NRS).

About 53 % of the strata in Delaware and 45 % of the strata in Massachusetts produce a relative efficiency between 1.0000 and 1.0199. About 44 % of the relative efficiencies scores produced in Delaware and Massachusetts are undefined because the sampling variance is zero or the finite population correction factor is 1.

The investigation pointed to some factors which effect the performance of the systematic rounding process. The data showed that because our integer weighting process takes place within NRS within FSPS the sample sizes within strata are usually small, generally between one and ten. As we can see in the interest paid variable (Delaware, Kent county, FSPS 17, NRS 0), the relative efficiency created is very high because the S^2 is small and the variance due to integerization is very large. Even if $S^2 = 0$, the integer variance, $VAR(\hat{X}_I)$, can be large because the integerization variance is the primary term.

We have seen that systematic rounding produces large absolute differences in the weighted estimates but small percent differences. The same is true for the variances. The relative efficiency between variances is small even though large differences occur between the estimated variances. The problems found in this study occur at the strata level and tend to even out at the published level. The above factors (problem areas) point to the level at which the systematic rounding takes place. Rounding within the FSPS/NRS combination limits the performance of the rounding technique, due to the small number of sample cases available. This was not a problem when Chapman (1981) investigated the integer weighting process, because the noninteger nonresponse weights and the base sample weights, were rounded at different stages and then combined to produce the final sample weights. The above data indicate that although there is some over and under estimating occurring at the strata level, it seems to balance out at the publishing level and implies that the systematic rounding technique does increase the variances, but has a limited effect on the estimated totals at the published level.

ACKNOWLEDGEMENTS

The authors would like to thank Larry Ernst for his contributions to this paper. This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the authors and do not necessarily reflect those of the Census Bureau.

REFERENCES

Chapman, David (1981), "Statistical Properties of Linear Survey Estimators with Integer Valued Random Variable Weights", American Statistical Association, 1981 Proceedings of the Section on Survey Research, pp. 607-612.

Cochran, William (1977), Sampling Techniques, 3rd Edition, John Wiley & Sons, New York.

Griffin, Richard (1987), Unpublished Memorandum, Bureau of the Census, Washington, D.C.

Hanson, Robert (1969), Unpublished Memorandum, Bureau of the Census, Washington, D.C.

Thompson, John (1978), Unpublished Memorandum, Bureau of the Census, Washington, D.C.