# Prediction Theory Approach to Multistage Sampling when Cluster Sizes Are Unknown

E. J. Kelly, Los Alamos National Laboratory and W. G. Cumberland, UCLA

A model for two-stage cluster sampling when sample cluster sizes are unknown is used to derive an optimal (model-based) estimator for the population total and to determine robust sampling strategies. In an empirical study using 1970 and 1980 census data for Los Angeles and surrounding counties, comparisons are made between the model based estimator and conventional estimators. The results favor the new estimator over those derived from randomization theory. In addition, the empirical study shows that the robust sampling strategies suggested by the theory can reduce biases, improve efficiency, and decrease the frequencies of large errors.

## 1. Introduction

In previous studies the model-based approach has proved to be a valuable tool for studying finite population sampling. The theoretical results of Royall (1976), and the theoretical and empirical results of Royall and Cumberland (1981a, 1981b, 1982) have brought new understanding of conventional estimators and introduced new, bias robust variance estimators. This paper uses prediction theory to develop criteria for selecting good sampling strategies (sampling plans and estimators) for two-stage sampling when the cluster size is unknown. The traditional approach to this problem, Randomization Theory, assumes that the population variables are fixed constants and that the probability framework is determined by the sampling plan, statistical properties such as bias and MSE are defined in terms of averages over all possible samples. This averaging masks the importance of the sample actually observed in determining bias and MSE. Royall and Cumberland(1981a, 1981b, 1982) show that such masking can be dangerous when making inferences since bias and variance as well as variance estimators can depend on sample characteristics. The prediction approach assumes that the population values are realizations of random variables and the probability framework is described by the joint distributions of the variables in the superpopulation model. Statistical quantities are defined with respect to the model and conditioned on the observed sample. The prediction approach allows us to study bias, variance, and variance estimators as functions of characteristics of the observed sample and thus to determine those samples that produce good estimates. Prediction theory also permits us to study estimators under conditions of model failure and to determine those estimators and sampling plans that are robust to certain types of model failure.

## 2. The Two-Stage Design with Unknown Cluster Sizes

In the two-stage sampling framework the population is divided into N clusters, each contains $M_i$ secondary units. A sample, s, of n clusters is taken and a subsample, $s_i$, of $m_i$ secondary units is drawn from each sampled cluster. All sampling is done without replacement. The situation considered here is the common one where the $M_i$ are known only for the sampled clusters, but there exist related auxiliary variables, $X_i$, that are known for all clusters. After observing $y_{ij}$ for the sampled secondary units one produces $\hat{T}$, an estimate of the population total, $T = \sum_{i=1}^{N}\sum_{j=1}^{M_i} y_{ij}$. The total number of sampled secondary units is $m_s = \sum_{i\in s} m_i$ while the sample means of the $y_{ij}$ for each cluster are given by $\bar{y}_{s_i} = \sum_{j\in s_i}\frac{y_{ij}}{m_i}$ $i = 1,2,...,n$. Letting $r$ denote the clusters that are not in the sample the following examples illustrate the notational convention used for summations of any variable: $X_r = \sum_{i\in r} X_i$, $\bar{X}_s = \frac{1}{n}\sum_{i\in s} X_i$, $\overline{(X^2)}_s = \frac{1}{n}\sum_{i\in s} X_i^2$, $Y_i = \sum_{j=1}^{M_i} y_{ij}$, $M = \sum_{i=1}^{N} M_i$. The quantity $f = n/N$ is the first-stage sampling fraction and $f_i = m_i/M_i$ is the second-stage sampling fraction.

## 3. Conventional Estimators for the Population Total

A sampling design used by many large surveys selects the clusters with probability proportional to a size measure ($\pi$ps sampling), and the secondary units with simple random sampling (SRS). In this case a Horvitz-Thompson estimator is frequently used to estimate the population total: $\hat{T}_P = \frac{N}{n}\sum_{i\in s} M_i\bar{y}_{s_i}\frac{\bar{X}}{X_i}$. For another common sampling design which selects both the clusters and the secondary units by simple random sampling (SRS-SRS), a ratio estimator is often chosen: $\hat{T}_R = \frac{N}{n}\sum_{i\in s} M_i\bar{y}_{s_i}\frac{\bar{X}}{\bar{X}_s}$. Traditionally $\hat{T}_P$ and $\hat{T}_R$ have been the estimators of choice for populations where the $Y_i$ are approximately proportional to the $X_i$. When no such $X_i$ exists then for SRS-SRS designs the "unbiased" estimator, $\hat{T}_U$, can be used (Cochran,1978): $\hat{T}_U = \frac{N}{n}\sum_{i\in s} M_i\bar{y}_{s_i}$. The Horvitz-Thompson estimator, $\hat{T}_P$, is unbiased with respect to a $\pi$ps-SRS plan and $\hat{T}_U$ is unbiased with respect to an SRS-SRS plan. The ratio estimator, $\hat{T}_R$, is biased with respect to SRS-SRS, however the bias is negligible for large n (Cochran,1977). Although these estimators are developed under traditional sampling theory, they will be studied as estimators for the population total under prediction theory.

## 4. The Superpopulation Model

The prediction approach to finite population sampling treats the $y_{ij}$ as realizations of random variables $Y_{ij}$. In this application cluster size is unknown except for sampled clusters, therefore, the $M_i$'s are also treated as realizations of random variables. The superpopulation model is a working model that describes the gross structure of many real populations. Deviations from this or any other model are to be expected and an important component of the prediction approach is to study conditions of model failure and determine strategies that are robust to such deviations. The superpopulation model proposed by Royall(1985) describes a population where cluster size is proportional to the previous size measure and cluster totals are increasing linearly conditionally with cluster size. (The $Y_{ij}$ are correlated within clusters but are independent between clusters.) Denoting conditional expectations, variances, and covariances by $E^*$, $Var^*$, and $Cov^*$, the model is as follows:

**MODEL $M_u$:**

$(i)$ $E(M_i) = \beta X_i$ $\quad i=1,2,...,N$

$(ii)$ $Var(M_i) = \tau^2 X_i$ and $Cov(M_i,M_j) = 0$ $\quad i\neq j$

$(iii)$ $Pr(M_i<2) = 0$

$(iv)$ $E^*(Y_{ij}) = \mu$ $\quad j=1,2,...,M_i$

$(v)$ $Var^*(Y_{ij}) = \sigma_i^2$

$(vi)$ $Cov^*(Y_{ij},Y_{kl}) = \begin{cases}\rho_i\sigma_i^2 & i=k, j\neq l\\ 0 & i\neq k\end{cases}$

The parameters $\beta$, $\mu$, and $\tau^2$ are constants. We only consider designs where $m_i\geq 2$ and if $m_i>M_i$ we take $m_i=M_i$. We assume that $\rho_i$ is non negative; this is not a strong restriction since

it can be shown that $\rho_i > -1/(M_i-1)$. In much of the analyses that follow the restrictions $\rho_i = \rho$ and $\sigma_i^2 = \sigma^2$ are made; the model with these restrictions is denoted $M_u'$.

Since the population total can be written as the sum of the observed and unobserved variables: $T = \sum_{i\in s}\sum_{j\in s_i} y_{ij} + \sum_{i\in s}\sum_{j\in r_i} y_{ij} + \sum_{i\in r}\sum_{j=1}^{M_i} y_{ij}$ (where $r_i$ is the set of subunits not included in the sample $s_i$), we note that the problem of estimating $T$ is equivalent to that of predicting the total for the unobserved $y_{ij}$'s: $\sum_{i\in s}\sum_{j\in r_i} y_{ij} + \sum_{i\in r}\sum_{j=1}^{M_i} y_{ij}$. The best linear unbiased estimator for T, $T^*_{BLU}$, is found by adding the observed total to the BLU predictor of the unobserved total and apply a result from prediction theory, which states that given a k+p random vector $X$ with mean $U$ and covariance $\Sigma$

$$X = \begin{bmatrix} X_p \\ X_k \end{bmatrix} \quad U = \begin{bmatrix} U_p \\ U_k \end{bmatrix} \quad \Sigma = \begin{bmatrix} \Sigma_{pp} & \Sigma_{pk} \\ \Sigma_{kp} & \Sigma_{kk} \end{bmatrix}$$

the best linear predictor of $X_p$ given $X_k$ is

$$\hat{X}_p = U_p + \Sigma_{pk}\Sigma_{kk}^{-1}(X_k - U_k) .$$

We exploit this theorem by properly defining $X$. We note that $M_i$ and $r_i$ are random, however, since we condition on the observed sample, $s$, $r$, $m_i$, and $s_i$ are fixed. We define $X$ as the N+2n random vector

$$X = \begin{bmatrix} X_p \\ X_k \end{bmatrix} \quad X_p = \begin{bmatrix} X_I \\ X_{II} \end{bmatrix} \text{ and } \quad X_k = \begin{bmatrix} X_{III} \\ X_{IV} \end{bmatrix}$$

where

$$X_I = ((\sum_{j\in r_i} Y_{ij})) \quad i \in s, \qquad X_{II} = ((\sum_{j=1}^{M_i} Y_{ij})) \quad i \in r,$$

$$X_{III} = ((\sum_{j\in s_i} Y_{ij})) \quad i \in s, \quad \text{and } X_{IV} = ((M_i)) \quad i \in s.$$

$X$ has covariance matrix

$$\Sigma = \begin{bmatrix} V_I & O & V_{I,III} & V_{I,IV} \\ O & V_{II} & O & O \\ V_{I,III} & O & V_{III} & O \\ V_{I,IV} & O & O & V_{IV} \end{bmatrix}$$

where $V_I$ is the n x n covariance matrix of $X_I$, $V_{I,III}$ is the n x n covariance matrix of $X_I$ and $X_{III}$, $V_{II}$ is the N-n x N-n covariance matrix of $X_{II}$, etc. Applying the theorem gives the predictors

$$\hat{X}_i = (M_i - m_i)\mu + (\beta X_i - m_i)\frac{m_i\rho_i\sigma_i^2}{[(1-\rho_i)\sigma_i^2 + m_i\rho_i\sigma_i^2]}(\bar{y}_{s_i} - \mu) \quad i \in s$$

$$\hat{X}_i = \mu\beta X_i \quad i \in r .$$

Therefore, the BLU estimator for T is

$$T^*_{BLU} = \sum_{i\in s}\sum_{j\in s_i} y_{ij} + \sum_{i\in s}(M_i - m_i)\mu +$$

$$\sum_{i\in s}(\beta X_i - m_i)\frac{m_i\rho_i\sigma_i^2}{[(1-\rho_i)\sigma_i^2 + m_i\rho_i\sigma_i^2]}(\bar{y}_{s_i} - \mu) + \beta\mu X_r .$$

The unknown parameters $\beta$ and $\mu$ are estimated by the BLU estimators

$$\hat{\beta} = \frac{\bar{M}_s}{\bar{X}_s}, \qquad \hat{\mu} = \sum_{i\in s} u_i \bar{y}_{s_i}$$

where $u_i = (m_i/[(1-\rho_i)\sigma_i^2 + m_i\rho_i\sigma_i^2])/\sum_{i\in s} m_i/[(1-\rho_i)\sigma_i^2 + m_i\rho_i\sigma_i^2]$ . The BLU estimator with $\hat{\beta}$ and $\hat{\mu}$ substituted for $\beta$ and $\mu$ is denoted $\hat{\hat{T}}_{BLU}$. Substituting $M_i$ for $\beta X_i$ in the third term of $\hat{\hat{T}}_{BLU}$ produces a non linear estimator, $\hat{\hat{T}}_{NL}$,

$$\hat{\hat{T}}_{NL} = \sum_{i\in s}\sum_{j\in s_i} y_{ij} + \sum_{i\in s}(M_i - m_i)[w_i\bar{y}_{s_i} + (1-w_i)\hat{\mu}] + \hat{\beta}\hat{\mu}X_r ,$$

$$w_i = m_i\rho_i\sigma_i^2/[(1-\rho_i)\sigma_i^2 + m_i\rho_i\sigma_i^2] .$$

$\hat{\hat{T}}_{NL}$ is unbiased with respect to $M_\mu$ and comparing MSE's we find

$$E(\hat{\hat{T}}_{BLU} - T)^2 - E(\hat{\hat{T}}_{NL} - T)^2 =$$

$$\sum_{i\in s}[Var(M_i) - Var(\beta X_i)][w_i^2(Var(\bar{y}_{s_i}) - Var(\hat{\mu}))].$$

This sum is non-negative therefore this estimator has smaller MSE than the BLU estimator. If $\rho_i = 0$, $n=1$, or $r^2 = 0$ then $\hat{\hat{T}}_{NL} = \hat{\hat{T}}_{BLU}$.

The estimator $\hat{\hat{T}}_{NL}$ depends on $u_i$ and $w_i$ which depend on $\rho_i\sigma_i^2$ and $(1-\rho_i)\sigma_i^2$ and are generally unknown. If $\rho_i = \rho$ and $\sigma_i^2 = \sigma^2$ then Rustagi (1978) notes that $M_u'$ is equivalent to a one way random effects model: $y_{ij} = \mu + \alpha_i + \epsilon_{ij}$, $i=1,2,...,n$, $j=1,2,...,m_i$, where $\rho\sigma^2 = Var(\alpha_i) = \sigma_\alpha^2$ and $(1-\rho)\sigma^2 = Var(\epsilon_{ij}) = \sigma_\epsilon^2$. In the following analysis the unweighted sum of squares estimators (USS) from random effects analysis are used to estimate $\rho\sigma^2$ and $(1-\rho)\sigma^2$ and the resulting estimator is denoted $\hat{T}_{NL1}$.

## 5. Model-Based Analysis

### 5.1 Bias

When expectation is taken with respect to $M_u$, $\hat{T}_P$ and $\hat{T}_R$ are unbiased and the model-based estimator $\hat{T}_{NL1}$ is asymptotically unbiased. The "unbiased" estimator $\hat{T}_U$ is biased; its bias is:

$$E(\hat{T}_U - T) = N\beta\mu(\bar{X}_s - \bar{X}) . \qquad (5.1.1)$$

The traditional theory says that $\hat{T}_U$ is unbiased with respect to an SRS-SRS sampling plan and the ratio estimator is biased (this bias is negligible for large n). In practice when the conditions of the model are reasonable $\hat{T}_U$ will generally have a larger variance than $\hat{T}_R$ and $\hat{T}_P$ (Cochran, 1977). The model-based analysis gives insight into what is causing this large variance.

### 5.2 Failure of the Model - Misspecified Expectation

The model $M_u$ assumes that the regressions of cluster size $M_i$ on the known size variable $X_i$ and cluster totals $Y_i$ conditioned on the $M_i$'s are straight lines through the origin. Royall and Cumberland show in their empirical studies that, for the six real populations they studied, biases can be explained by assuming that the model fails and that departures from the simple linear model can be described by a model that has an intercept and quadratic term (Royall and Cumberland,1981a,b), (Cumberland and Royall,1982). The model failure considered here is denoted as TYPE I failure and describes the situation where cluster size is not proportional to the previous size measure, but can be described by a polynomial with an intercept and quadratic term:

$$TYPE\ I \quad E(M_i) = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 . \qquad (5.2.1)$$

In the analysis and discussion that follow TYPE I failure implies the conditions of model $M_u$ except for the misspecification of the expectation. It should be noted that even though these models are crude approximations to the underlying population, they are useful for explaining the behavior of the estimators. The biases under TYPE I failure are :

$$E(\hat{T}_P - T) = N\mu\beta_0\left[\overline{(X^{-1})}_s\,\bar{X} - 1\right] + N\mu\beta_2\left[\bar{X}\,\bar{X}_s - \overline{(X^2)}\right] \qquad (5.2.2)$$

$$E(\hat{T}_R\text{-}T) = N\mu\beta_0\left[\frac{\overline{X}\text{-}\overline{X}_s}{\overline{\overline{X}}_s}\right] + N\mu\beta_2\left[\frac{\overline{X(X^2)}_s\text{-}\overline{X}_s\overline{(X^2)}}{\overline{\overline{X}}_s}\right]$$

$$E(\hat{T}_U\text{-}T) = N\mu\beta_1\left[\overline{X}_s\text{-}\overline{X}\right] + N\mu\beta_2\left[\overline{(X^2)}_s - \overline{(X^2)}\right]$$

$$E(\hat{T}_{NL1}\text{-}T) \approx N\mu\beta_0\left[\frac{\overline{X}\text{-}\overline{X}_s}{\overline{\overline{X}}_s}\right] + N\mu\beta_2\left[\frac{\overline{X(X^2)}_s\text{-}\overline{X}_s\overline{(X^2)}}{\overline{\overline{X}}_s}\right]$$

(Throughout this failure analysis $\rho_i = \rho$, $\sigma_i^2 = \sigma^2$, and $E(\hat{T}_{NL1}\text{-}T)$ is an asymptotic result unless $\rho_i$ and $\sigma_i^2$ are known.) The estimators can be protected from bias caused by TYPE I failure if restrictions are placed on the sampled X's. The restrictions for $\hat{T}_P$ are: $\overline{(X^{-1})}_s = 1/\overline{X}$ and $\overline{X}\,\overline{X}_s = \overline{(X^2)}$. The condition, $\overline{X(X^{j-1})}_s = \overline{X^j}$, is called $\pi$-balance and is expected under the probability sampling distribution when a $\pi$ps sampling plan is used (Cumberland and Royall,1982). $\hat{T}_R$, $\hat{T}_U$, and $\hat{T}_{NL1}$ have zero biases when $\overline{X} = \overline{X}_s$ and $\overline{(X^2)}_s = \overline{(X^2)}$. This condition is called balance on the first and second moments. In general, $\overline{(X^j)}_s = \overline{X^j}$ is called balance on the $j^{th}$ moment and is expected under the probability sampling distribution when an SRS sampling plan is used. Although balanced samples have long been recognized as useful, the traditional theory offered little support for their use nor could it explain why poorly balanced samples were bad (Royall and Cumberland, 1981a).

The analysis of the bias of the estimators when expectation with respect to the model is misspecified dramatizes the importance of the sample in determining the reliability of the estimators. The analysis shows that under certain types of model failure, samples that are properly balanced can protect estimators from serious errors. On the average, probability-proportional-to-size sampling plans will be $\pi$-balanced and simple random sampling plans will be balanced. The failure analysis gives new insight into the success of these conventional sampling procedures but introduces the question as to whether an expected balanced sample is good enough. Cumberland and Royall (1982,1981a) have shown that SRS and $\pi$ps sampling alone do not provide adequate protection against bias. In the six populations they studied, departures from the appropriate balance conditions appeared in a significant percentage of the samples. Those samples which deviated the farthest from balance produced biases which were large compared to the $(MSE)^{1/2}$. These analyses suggest that sampling techniques that force balance will protect against such errors, thus introducing the notion of robust sampling strategies.

## 6. Empirical Study

The empirical study uses real data to test the theoretical results. The real data set allows us to investigate the robustness of the model-based theoretical results when the model failure is more complex than that described in section 5.2.

The empirical study is a survey, utilizing previous data, to predict the total population for a rapidly growing area, the outlying areas of Los Angeles county, and all of Ventura and Orange County. The data are block statistics from the 1970 and 1980 California census tapes. The number of blocks in a census tract in 1970 is the previous size measure, and the number in 1980 is the cluster size. The variable of interest, $y_{ij}$, is the block population in 1980. If a census tract had fewer than twenty blocks in the 1970 census then it was combined with its nearest neighbor. This process continued until the resulting tract had twenty or more blocks. After this adjustment, 420 census tracts (clusters) remained. There were a total of 23,001 blocks in 1970, and by 1980 the number had grown to 29,102. The total population of these regions for 1980 was 4,045,074.

Figures 6.1, 6.2, and 6.3 contain plots of the census data showing that the superpopulation model is a reasonable description for this data. The model assumes that cluster size is increasing linearly with the previous cluster size. Figure 6.1 shows that this assumption is reasonable, and that there may be an intercept term and a slight convex curvature (TYPE I failure), however, the model failure is more complex than that described in 5.2.

### 6.1 Sampling Strategies

The model failure analysis of 5.2 supports previous studies by Royall and Cumberland (1981a) and Herson (1976), which show that under certain types of model failure appropriately balanced sampling can protect against biases and reduce MSE. We call such balanced first-stage sampling plans with the appropriate estimators robust strategies and compare these strategies to conventional sampling.

Exact satisfaction of the balance conditions is not possible, however, for SRS plans one can exclude samples that are badly unbalanced by imposing the following conditions:

$$\left|\frac{\sqrt{n}(\overline{X}_s\text{-}\overline{X})}{t_1}\right| \le \epsilon_1 \qquad (6.1)$$

$$\left|\frac{\sqrt{n}(\overline{(X^2)}_s\text{-}\overline{(X^2)})}{t_2}\right| \le \epsilon_2 \qquad (6.2)$$

where $t_1$ and $t_2$ are the SRS finite population standard deviations of $X_i$ and $X_i^2$ respectively. The sampling procedure $r_1$-SRS consists of taking a simple random sample and rejecting those samples that fail condition (6.1). The $r_2$-SRS rejects those samples that do not satisfy conditions (6.1) and (6.2).

The $\pi$ps sampling plan used in this study is the one investigated by Hartley and Rao (1962) and used by Cumberland and Royall (1982). Cumberland and Royall give a concise description of this procedure:

> "the procedure consists of a random permutation of the first-stage units followed by a random start systematic sample of step size X/n over the interval (0,X). Unit $i$ is selected if $\sum_1^{i-1} X_j \le U < \sum_1^{i} X_j$ for one of the systematic sample points $U$."

The following restrictions are used to force $\pi$-balance: $\left|\frac{\sqrt{n}(\overline{(X^{-1})}_s\text{-}1/\overline{X})}{t_1}\right| \le \epsilon_1$ and $\left|\frac{\sqrt{n}(\overline{X}_s\text{-}\overline{(X^2)}/\overline{X})}{t_2}\right| \le \epsilon_2$, where $t_1$ is the standard deviation of $1/X$ and $t_2$ is the standard deviation of $X$. The values of $t_1$ and $t_2$ are found by using the finite population variance formulas. For $r_1$ sampling $\epsilon_1 = .15$ gave a 10% acceptance rate. Using $\epsilon_1$ and $\epsilon_2 = .15$ in the $r_2$ sampling procedure, yields an acceptance percentage of 5% for SRS sampling and 3%for $\pi$ps sampling.

Basket sampling was introduced by Wallenius (1973) as a technique for getting extremely well balanced samples while retaining some randomness. This technique has the advantage of being much less expensive to implement on a computer than restricted sampling, and, at least for this study population, it produces samples that are almost perfectly balanced on the first and second moments and extremely well balanced on the third and fourth moments.

All second-stage samples are selected by simple random sampling. In a preliminary analysis constant, self-weighting, and model-based optimal second-stage allocation procedures were used. The second-stage allocation procedures did not yield significantly different results for any estimator or any first-stage sampling plan (Cohen,1984). Therefore, constant second-stage allocation was used in the final analysis. For each sampling scheme $n = 42$ and $m_i = 20$ producing a 3% sample.

### 6.2 Performance Evaluation Criteria

To study bias and MSE as functions of sample characteristics, the samples are arranged in order of increasing value of the characteristic. The samples are then grouped in equal sets, so

that the first set contains samples with the smallest values of the characteristic, the second set contains the next smallest, etc. For each group, the average values of $(\hat{T}-T)$ and MSE are calculated and used to plot $(\hat{T}-T)$ and $(MSE)^{1/2}$ against the average of the sample characteristic. Figure 6.4 shows the plots of the averages of $(\hat{T}_P-T)$ (B), and $(MSE)^{1/2}$ (V) versus the average $\overline{(X^{-1})}_s$. The 450 samples are arranged in 10 groups of 45 samples each. Plots of the cumulative distributions of the errors are used to determine how different sampling plans and different estimators affect the error distributions. Figure 6.7 contains the cumulative error distributions for $\hat{T}_{NL1}$ illustrating the differences between errors generated by SRS and Basket first-stage sampling plans.

**6.3 Results**

Traditional sampling theory assures us that $\hat{T}_P$ and $\hat{T}_U$ are unbiased, and $\hat{T}_R$ is approximately unbiased under their appropriate sampling procedures. The errors, averaged over all 450 replications (Table 6.1) support this theory. However, the error curves show that this net effect is deceiving, a result of negative biases on one side of a population balance point and positive biases on the other side. The error curve for $\hat{T}_P$ (Figure 6.4) shows that negative biases when $\overline{(X^{-1})}_s$ is less than $\frac{1}{\overline{X}}$ are balanced by positive biases when $\overline{(X^{-1})}_s$ is greater than $\frac{1}{\overline{X}}$. This result is what the model-based theory predicts when the underlying model has a positive intercept term as in TYPE I failure (5.2.2). The error curve for $\hat{T}_U$ (Figure 6.5) demonstrates the extreme biases of this estimator, an estimator that traditional theory calls the *unbiased* (SRS-SRS) estimator. Prediction theory maintains that $\hat{T}_U$ will be biased under $M_u$ unless the sample is balanced and equation (5.1.1) indicates that this bias will show a linear dependency on $\overline{X}_s$ with negative biases when $\overline{X}_s$ is less than $\overline{X}$ and positive biases when $\overline{X}_s$ is greater than $\overline{X}$. This is exactly what the error curve reveals. The error curves for $\hat{T}_{NL1}$ and $\hat{T}_R$ show the same agreement with the theoretical results.

The model-based theory predicts that restricted sampling ($r_1$-$\pi$ps and $r_1$-SRS) will eliminate much of the remaining bias due to the intercept term and that the biases will be functions of $\overline{X}_s$ ($\hat{T}_P$;$r_1$-$\pi$ps) and $\overline{(X^2)}_s$ ($\hat{T}_R$,$\hat{T}_U$,$\hat{T}_{NL1}$;$r_1$-SRS). The error curves for all estimators show the appropriate trends, giving further support to the TYPE I failure model for the census data. The plot of $(\hat{T}_P-T)$, versus $\overline{X}_s$ for $r_1$-$\pi$ps sampling, Figure 6.6, illustrates this agreement. As the theory predicts for TYPE I failure with $\beta_2$ positive, the biases are negative for $\overline{X}_s$ less than $\overline{(X^2)}/\overline{X}$ and positive for $\overline{X}_s$ greater than $\overline{(X^2)}/\overline{X}$.

The error curves demonstrate that inferences about the reliability of these estimators based on probability sampling theory, which says that the bias is zero no matter what sample is observed, can have serious errors. The analysis also demonstrates the robustness of the model-based theory. Even on a real population where the underlying structure is difficult (if not impossible) to model mathematically, the model-based theory remains valid and useful.

We have seen the advantage of using the restricted sampling plans in eliminating bias. One might also ask if these restricted sampling plans reduce MSE and decrease the probability of observing large errors. Figure 6.7 compares the cumulative distributions of $\hat{T}_{NL1}$-$T$ for SRS, restricted SRS and Basket sampling. The comparison shows a slight reduction in the probability of errors greater than 450,000 and less than -450,000 for $r_2$-SRS sampling (not shown since this curve is almost indistinguishable from the SRS curve ) and a larger reduction for Basket sampling. With SRS sampling, 17.4% of the errors will be greater than 450,000 or less than -450,000. The comparable percentage for Basket sampling is 7%. Comparable curves for $\hat{T}_R$, $\hat{T}_P$ and $\hat{T}_U$ show similar improvements in the error distributions for appropriately balanced samples. Comparing MSE's (Table 6.1) tells the same story. The relative efficiencies of Basket to SRS sampling (the ratio of SRS MSE to Basket MSE) range from 1.2 to 1.9, for $r_2$-SRS to SRS the ranges are 1.0 to 1.3, and for $r_2$-$\pi$ps to $\pi$ps the relative efficiency is 1.2. The analysis demonstrates that balanced samples can reduce bias, improve efficiency, and decrease the probability of observing large errors.

Figure 6.8 contains the error distributions for $\hat{T}_R$, $\hat{T}_{NL1}$, and $\hat{T}_U$ for SRS sampling and $\hat{T}_P$ for $\pi$ps sampling. These curves show that $\hat{T}_{NL1}$ performs best. The frequency of errors greater than 550,000 and less than -550,000 is 24% for $\hat{T}_U$, 25% for $\hat{T}_P$, 17% for $\hat{T}_R$ and 9% for $\hat{T}_{NL1}$. $\hat{T}_{NL1}$ also has the smallest MSE and the relative efficiencies of the other estimators to it are 0.51 for $\hat{T}_U$, 0.56 for $\hat{T}_P$, and 0.74 for $\hat{T}_R$. The relative efficiency of $\hat{T}_P$ to $\hat{T}_R$ is 0.69. In Figure 6.9, we compare the error distributions for these estimators under their optimal sampling procedures. $\hat{T}_R$ and $\hat{T}_U$ are equivalent when the samples are balanced and their error distributions are indistinguishable. $\hat{T}_{NL1}$ has no errors greater than 750,000 or less than -450,000, while 11% of $\hat{T}_R$'s and 22% of $\hat{T}_P$'s errors fall outside these bounds. The relative efficiencies of $\hat{T}_R$ and $\hat{T}_P$ to $\hat{T}_{NL1}$ are 0.51 and 0.36 respectively. The relative efficiency of $\hat{T}_P$ to $\hat{T}_R$ is 0.70. The superior performance of $\hat{T}_R$ to $\hat{T}_P$ on the census data may be due to the erratic growth of the small clusters. First-stage $\pi$ps designs in conjunction with the estimator $\hat{T}_P$ are commonly used in large surveys. This analysis cautions that $\hat{T}_P$ can have serious biases and suggests that $\pi$-balance will help protect $\hat{T}_P$ against such model failures. Although $\hat{T}_{NL1}$ is the optimal estimator under the superpopulation model, it is difficult to see why it out performs $\hat{T}_R$ on the census data. The census data fails many of the model assumptions, including the assumption that all $Y_{ij}$'s have the same mean. Just why $\hat{T}_{NL1}$ remains robust to the census population's deviations from the model and how it will perform on other real populations are areas for further investigation.

**7. Summary**

This analysis shows that the behavior of the estimators depends on characteristics of the underlying population and the prediction model approach is an appropriate way to study this dependency. The prediction model was used to derive a best estimator for the population total under a basic model for the two-stage design problem with unknown cluster sizes. In the empirical study this estimator performed better than the traditional estimators and deserves further study as an alternative estimator for the population total.

The model-based theory indicated that estimators, unbiased under the probability sampling distribution, could have important biases that depended on observable characteristics of the sample. The theoretical analysis showed that the traditional "unbiased" estimator can have severe biases for samples that are badly unbalanced and is not appropriate for populations described by the basic model. These results support traditional practices, but give additional insight into why this estimator performs poorly. The other traditional estimators considered, the ratio estimator with a two-stage simple random sampling plan and the Horvitz-Thompson estimator with a probability-proportional-to-size first-stage plan and a simple random sampling second-stage plan, are unbiased under the basic model. However, the theory indicates that they can have biases under certain types of model failure. The empirical study showed that these estimators were biased for samples that were badly balanced. For the traditional random sampling plans, twenty percent or more of the samples produced estimates with severe biases.

The empirical study confirmed the theoretical findings for a real population that was only very roughly described by the model indicating that, rather than looking for increasingly complex models to try to describe the population exactly, samples and estimators should be chosen with robustness in mind. Robust strategies will yield good estimates for a variety of

different populations. Robust strategies include plans that restrict the sample so that it has the necessary characteristics to produce reliable estimators. The theoretical results suggested and the empirical study confirmed that techniques for restricting the first-stage samples so that they were well balanced reduced biases, improved efficiency, and decreased the frequencies of large errors.

## BIBLIOGRAPHY


Cochran, W.G. (1977), *Finite Population Sampling*, John Wiley and Sons.

Cohen, B.J. (1984), Prediction Approach to Multistage Sampling when Cluster Size is Unknown", Dissertation, University of California at Los Angeles.

Cumberland W.G., and Royall, R.M. (1982), "Prediction Models and Unequal Probability Sampling," *JRSS B*, 43, No. 3, 353-367.

Hartley, H.O. and Rao, J.N.K (1962), "Sampling with Unequal Probabilities and Without Replacement," *Ann. Math. Statist.*, 33, 350-374.

Herson,J. (1976), "An Investigation of Relative Efficiencey of Least Squares Prediction to Conventional Probability Sampling Plans," *JASA*, 71, 700-703

Royall, R.M. and Cumberland, W.G. (1981a), "An Empirical Study of the Ratio Estimator and Estimators of Its Variance," *JASA*, 76, 66-88.

Royall, R.M. and Cumberland, W.G. (1981b), "The Finite Population Linear Regression Estimator and Estimators of Its Variance," *JASA*, 76, 924-930.

Royall, R.M. (1976) "The Least Squares Linear Approach to Two-Stage Sampling," *JASA*, 71, 657-664.

Royall, R.M. (1985), "The Prediction Approach to Robust Variance Estimation in Two-Stage Cluster Sampling," *John Hopkins University #559.*

Rustagi, R.K. (1978), "Some Theory of the Prediction Approach to Two Stage and Stratified Two-Stage Sampling," Dissertation, Ohio State University.

Wallenius, K.J. (1973), "On Statistical Methods in Contract Negotiations - Part III," Report N45, Department of Mathematical Sciences, Clemson University.


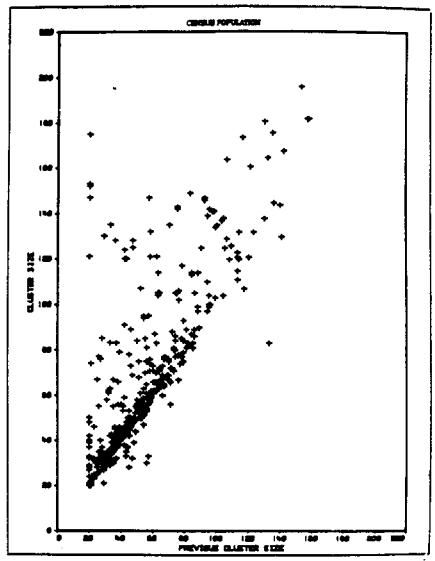


Figure 6.1 Plots of cluster size ( number of blocks in 1980 census tract ) versus previous cluster size ( number of blocks in 1970 census ).

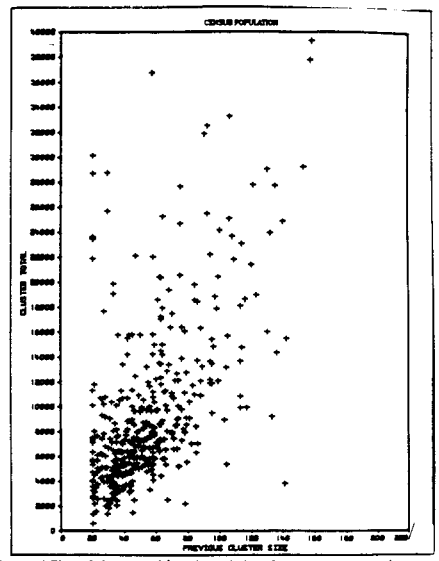


Figure 6.2 Plots of cluster total ( total population of census tract in 1980 ) versus previous cluster size ( number of blocks in 1970 census).

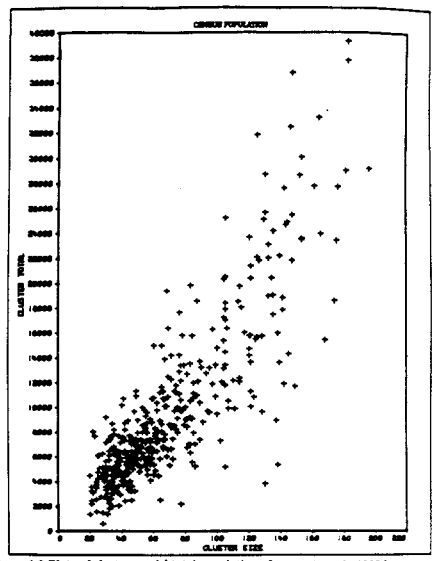


Figure 6.3 Plots of cluster total ( total population of census tract in 1980 ) versus cluster size ( number of blocks in 1980 census).

| Census Averages ( thousands ) | | | |
|---|---|---|---|
| Estimator | First-Stage Sampling Plan | Error | $(MSE)^{\frac{1}{2}}$ |
| $\hat{T}_P$ | πps | -22.2 | 480 |
| | $\pi_1$-πps | -25.2 | 430 |
| | $\pi_2$-πps | 3.5 | 441 |
| $\hat{T}_R$ | SRS | 12.6 | 398 |
| | $\pi_1$-SRS | -27.1 | 370 |
| | $\pi_2$-SRS | 24.4 | 397 |
| | Basket | 19.3 | 362 |
| $\hat{T}_U$ | SRS | 2.0 | 457 |
| | $\pi_1$-SRS | -29.1 | 371 |
| | $\pi_2$-SRS | 27.1 | 397 |
| | Basket | 19.3 | 362 |
| $\hat{T}_{WLS}$ | SRS | 0.2 | 343 |
| | $\pi_1$-SRS | -31.9 | 331 |
| | $\pi_2$-SRS | 17.9 | 347 |
| | Basket | 3.6 | 259 |

Table .6.1 The results averaged over 450 replications.

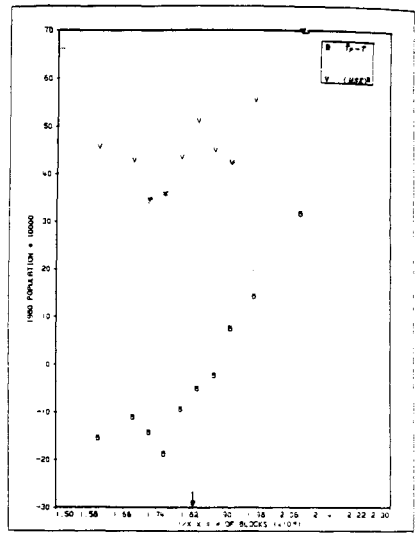

Figure 6.4 Census population, πps first-stage sampling, constant second-stage sampling, estimator $\hat{T}_P$. The symbol | marks the balance point.

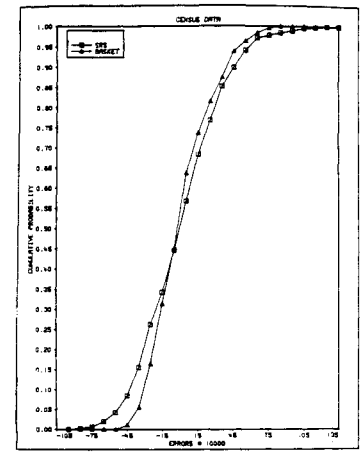

Figure 6.7 Error distribution for $\hat{T}_{NLI}$.

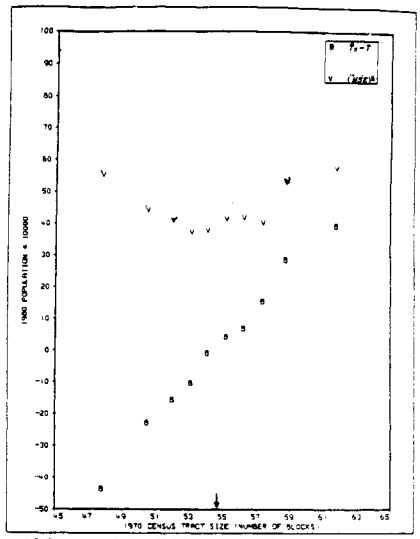

Figure 6.5 Census population, SRS first-stage sampling, constant second-stage sampling, estimator $\hat{T}_E$. The symbol | marks the balance point.

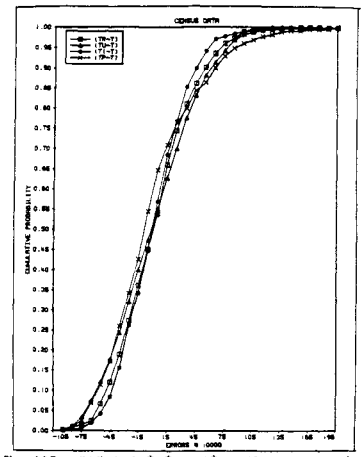

Figure 6.8 Error distribution for $\hat{T}_E$, $\hat{T}_{NLI}$, and $\hat{T}_U$ with SRS sampling and for $\hat{T}_P$ with πps sampling.

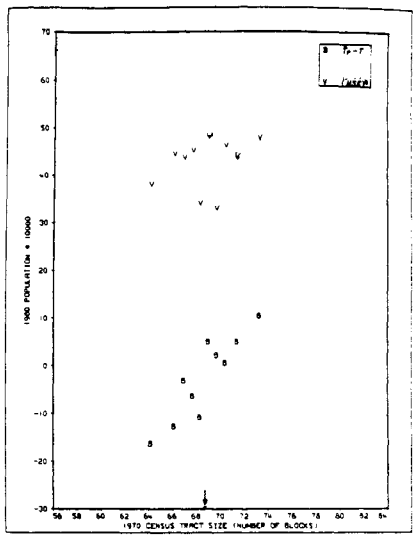

Figure 6.6 Census population, $r_1$-πps first-stage sampling, constant second-stage sampling, estimator $\hat{T}_P$. The symbol | marks the balance point.

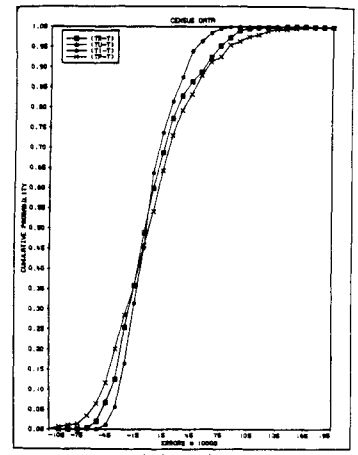

Figure 6.9 Error distribution for $\hat{T}_E$, $\hat{T}_{NLI}$, and $\hat{T}_U$ with Basket sampling and for $\hat{T}_P$ with $r_1$-πps sampling.