Statistics Canada
Statistique Canada

Methodology Branch

Household Survey
Methods Division

Direction de la méthodologie

Division des méthodes
d'enquêtes des ménages

Canada

WORKING PAPER
METHODOLOGY BRANCH

Skewed Survey Populations: Optimal Construction of "Take-all" and "Sampled" Groups with Application to the Local Government Finance Survey Re-design

SSMD - 95 - 001

Gurupdesh S. Pandher

Methods Development and Analysis Section
Social Survey Methods Division, Statistics Canada

March 1995

The work presented in this paper is the responsibility of the author and does not necessarily represent the views or policies of Statistics Canada

Skewed Survey Populations: Optimal Construction of "Take-all" and "Sampled" Groups
with Application to the Local Government Finance Survey Re-design

Gurupdesh S. Pandher
Methods Development and Analysis Section
Social Survey Methods Division, Methodology Branch
Statistics Canada, Ottawa, K1A 0T6

Abstract

In a survey re-engineering context, the most optimal sample design and estimation strategy holds the promise of offering the largest reduction in the sample size (and survey costs). In this paper, a scheme called the "Transfer Algorithm" is proposed to address the problem of finding an optimal demarcation between the "take-all" and "sampled" groups in surveys of skewed populations (eg. business, agricultural populations). The criterion for constructing these groups is based directly on minimizing the design variance of the regression estimator under a flexible range of sample designs (eg. SRS, pps, stratified). The proposed method also explicitly captures the size-induced heteroscedasticity evident in skewed populations in the determination of the optimal demarcation.

Desirable mathematical properties of this algorithm such as existence and optimality of solution are established. An equivalence result is also obtained allowing the solution to be determined in terms of simple quantities computable directly from the population auxiliary data. These theoretical results are reported in Theorems 1 to 3. The methodology is illustrated using provincial data from the Local Government Finance Survey.

Additionally, an alternative conceptualization of the survey framework for skewed populations is advanced at the onset which places all intermediate and final parameters of interest in a design-based framework. Classical optimal design theory results from the model-assisted framework are then re-cast in terms of finite population quantities in the alternative framework. Although this is not essential to the main results of the paper, this formulation renders a desirable design-based interpretation to all parameters involved and allows the design variance to be used as the criterion for constructing the population partitioning.

Population asymétrique : construction optimale de groupes «à tirage complet» et «échantillons», avec application au remaniement de l'enquête sur les finances des administrations locales

Gurupdesh S. Pandher
Section du développement des enquêtes et des méthodes d'analyse
Division des méthodes d'enquêtes sociales, Direction de la méthodologie
Statistique Canada, Ottawa K1A 0T6

Résumé

Lorsqu'on procède au remaniement d'une enquête, le plan de sondage et la stratégie d'estimation les plus optimaux pourraient se traduire par la réduction de la taille de l'échantillon (et, donc, des coûts de l'enquête) la plus importante. On propose dans cet article un schéma, appelé «algorithme de transfert», pour régler le problème qui est de trouver une ligne de démarcation optimale entre les groupes «à tirage complet» et «échantillons» dans les enquêtes portant sur des populations asymétriques (entreprises, agriculture, etc.). Le critère pour la construction de ces groupes est basé dire[illegible]nt sur la minimisation de la variance du plan de sondage de l'estimateur de régression dans un ensemble souple de plans (SRS, ppt, stratifié, etc.). La méthode proposée saisit également de façon explicite l'hétéroscédasticité observée dans les populations asymétriques lors de la détermination de la démarcation optimale.

On établit les propriétés mathématiques souhaitables de cet algorithme, telles que l'existence et l'optimalité de la solution. On obtient également un résultat d'équivalence, qui permet de déterminer la solution en termes de quantités simples, calculables directement à partir des données auxiliaires de la population. Ces résultats théoriques sont mentionnés dans les théorèmes 1 à 3. La méthodologie est illustrée par des données provinciales tirées de l'enquête sur les finances des administrations locales.

De plus, on présente dès le début une autre conceptualisation du cadre de l'enquête pour les populations asymétriques, dans laquelle tous les paramètres visés, tant ceux de la population que du plan de sondage, sont placés dans un cadre utilisant la base de sondage. Les résultats de la théorie classique du plan de sondage optimal du cadre avec modèle sont ensuite exprimés de nouveau en termes des quantités de la population finie dans le cadre alternatif. Bien que cette formulation ne soit pas essentielle pour les principaux résultats de l'article, elle offre une interprétation souhaitable utilisant le plan de sondage pour tous les paramètres en cause et permet d'employer la variance du plan de sondage comme critère pour la construction des partitions de la population.

Skewed Survey Populations: Optimal Construction of "Take-all" and "Sampled" Groups with Application to the Local Government Finance Survey Re-design

Gurupdesh S. Pandher [*]

1. INTRODUCTION

In many survey situations additional information is available on all population units before the survey is undertaken. This auxiliary information is frequently useful in devising a more efficient sample design and estimation strategy. In a survey re-design context, the most optimal strategy holds the promise of offering the largest reduction in survey costs by requiring the lowest sample size necessary to meet certain quality constraints on the estimates (eg. minimum desired coefficient of variation).

In repeat surveys of skewed populations, maximal reductions in existing sample sizes may be realized by i) taking advantage of the structural relationships present between certain available auxiliary information (eg. population of municipality, employees in company, farm acreage) and the survey variables (eg. expenditures in municipality, value of shipments, farm yield) in estimating the population parameters (eg. totals, ratios, regression coefficients) of interest, and by ii) using this information with design optimality results to identify and implement an efficient sample design.

Furthermore, in many such populations, the characteristics of interest tend to increase in variability as the size of the unit increases. The definite inclusion of these large units then increases the precision of the estimators as the take-all group contributes no sampling variance to the estimates. Therefore, in surveys of skewed populations, it is common to choose the largest units with probability equal to one ("take-all" group) and to employ an efficient sample design (eg. pps, stratified) in the remaining "sampled" group.

In this paper, a new methodology is proposed to address the important issue of finding an

[*] Gurupdesh Pandher is Methodologist with Methods Development and Analysis Section, Social Survey Methods Division, Methodology Branch, Statistics Canada, 16th Floor, Coats Building, Ottawa, Ontario K1A 0T6, Canada. The research reported in the paper was sponsored by Public Institutions Division.

optimal demarcation in the skewed population between the sampled and take-all groups as part of implementing an efficient sampling strategy. This work is motivated by the re-design of the Local Government Finance Survey (LGFS) conducted by Statistics Canada's Public Institutions Division. Financial information (eg. revenues, expenditures, debt, etc.) obtained from local government units is used in the estimation and publication of financial statistics on a provincial and national basis. An important objective of the re-design is to achieve survey cost reductions while maintaining a minimum desired level of precision in the estimates produced.

An overall sample re-design methodology attempting to obtain a maximal reduction in the current sample size needs to address and integrate the solution to three problems:

i) identifying an efficient sample selection scheme,

ii) constructing an efficient demarcation between the take-all and sampled population groups at a given sample size, and

iii) determining the minimal sample size required to meet the precision constraint(s).

The overall methodology devised allows a new optimal sample size and optimal population partitioning to be determined through an iterative scheme while maintaining the desired level of precision in the estimates. The methodology is flexible enough to operate under a variety of sample selection strategies (eg. SRS, pps, generalized pps). Due to space considerations, the main emphasis of this paper is on the first two components of the overall methodology mentioned above.

An iterative scheme - called the "Transfer algorithm" - is developed which finds an optimal allocation of population units between the take-all and sampled population groups in the sense of minimizing the design variance of the generalized regression estimator. Desirable mathematical properties of this algorithm such as existence and optimality of solution are established. An equivalence result is also obtained allowing the solution to be determined in terms of simple quantities computable directly from the population auxiliary data. The methodology is illustrated using provincial data from the Local Government Finance Survey.

Additionally, an alternative conceptualization of the survey framework for skewed finite populations is advanced at the onset which places all parameters involved in a design-based framework. Classical optimal design theory results from the model-assisted framework are then re-cast in terms of finite population quantities in the alternative framework. Although this is not essential to the main results of the paper, the formulation renders a desirable design-based interpretation to all parameters involved and allows the design variance to be used as the criterion for constructing the population partitioning.

Lavallee and Hidiroglou (1988), Hidiroglou and Srinath (1993) (subsequently denoted as L&H and H&S, respectively), and Glaser (1962) have proposed alternative sample re-design methodologies for skewed populations in the context of stratified simple random sampling (SRS) designs. With regard to the construction of the take-all and sampled groups (part ii) above), the proposed approach differs in three respects. Firstly, an optimal population demarcation is obtained for a flexible range of sample designs (eg. SRS, probability proportional to size (pps), stratified) planned for implementation. L&H consider a stratified SRS design on the auxiliary information (x) while H&S consider a stratified SRS design. Secondly, the criterion used by the Transfer Algorithm to find the optimal population allocation is based directly on minimizing the design variance of the regression estimator for the total (for a target survey variable y) under the desired sample design. H&S base their allocation on the total regression sum-of-squares for a survey variable under a regression model with a compulsory intercept assuming SRS while L&H find a stratification minimizing the within-stratum sum-of-squares for an auxiliary variable under SRS. Thirdly, the proposed methodology explicitly captures the size-induced heteroscedasticity evident in skewed survey populations. Other frameworks ignore the heteroscedasticity present in the population and use OLS estimates to determine the partitioning.

2. FINITE POPULATION FRAMEWORK FOR SKEWED POPULATIONS

In this section, after introducing the relevant notation and defining the problem, an alternative survey framework for the skewed population is described which allows all population quantities and important related parameters to be viewed in an entirely design-based perspective. This set-up allows the design variance of the generalized regression estimator to be later used as the criterion for constructing the optimal population partitioning; a superpopulation model is not required. Although this formulation is not critical to the main results of the paper (one could instead base the criterion on the anticipated variance), the design-based connotation may be more appealing to the design-based purist.

2.1 Notation and Problem Definition

Let $C_U=\{(x_1,y_1),\ldots,(x_N,y_N)\}$ represent the characteristics of interest for units in the finite population $U=\{1,\ldots,N\}$. Both x_k and y_k constitute the auxiliary (known before the survey) and the survey variables (unknown before the survey) of interest for the population unit k. They can be vectors but to keep the exposition simple we assume them to be scalars without loss of generality. Further, assume that x is a continuous auxiliary variable and let $x_{(k)}$ represent the kth ordered (ascending) value of the auxiliary set $X_U = \{x_1, \ldots, x_N\}$; ties can be placed in any arbitrary but fixed order. $X_U^* = (x_{(1)}, \ldots, x_{(N)})$ represents the ordered vector of auxiliary population values.

Our purpose is to decompose the finite population U into two disjoint sub-population consisting of the take-all group $U_a = \{1, \ldots, N_a\}$ and the sampled group $U_b = \{1, \ldots, N_b\}$: $U = U_a \bigcup U_b$ and $N = N_a + N_b$. A sample of size $n = n_a + n_b$ is then to be taken from $U = U_a \bigcup U_b$ using the sample design $p(s;\lambda) = (p_a(s_a), p_b(s_b;\lambda))$ such that all units in U_a are selected ($n_a = N_a$) and a subsample of n_b ($< N_b$) units is selected from U_b. The sample design parameter λ determines the type of sample selection implemented in the sampled sub-population U_b. The sample inclusion probabilities due to $p_b(s_b;\lambda)$ are expressible as $\pi_k(\lambda) = n_b \left(x_k^{\lambda/2} / \sum_{U_b} x_j^{\lambda/2}\right)$, $k \epsilon U_b$. Note that the sample design parameter λ defines a broad class of sample designs with SRS ($\lambda = 0$) ana pps ($\lambda = 2$) being just two particular cases.

Once the sample has been selected, our ultimate objective is to estimate the population total $t = t_a + t_b$ for the survey variable y where t_a and t_b represent the sub-totals for the take-all and sampled groups, respectively. Since a census is taken in U_a ($\pi_{a,k} = 1$, $k \epsilon U_a$), the estimator of t_a is t_a itself with design variance zero: $\hat{t}_a = t_a = \sum_{U_a} y_k$. To estimate t_b, the class of estimators defined by the generalized regression estimator will be considered. These estimators take account of the correlation existing between the survey variable y_k and available auxiliary covariates x_k (note that x_k can in general be a vector) to efficiently estimate the population total. For example, a combined ratio-based regression estimator for the total is given by

$$\hat{t}_{Rb} = \sum_{U_b} x_k \hat{B} + \sum_{s_b} \frac{(y_k - x_k \hat{B})}{\pi_k} \tag{2.1}$$

where $\hat{B} = \dfrac{\sum_U y_k / \pi_k}{\sum_U x_k / \pi_k}$ is the sample-based probability weighted estimate of the regression parameter B.

In terms of the notation introduced above, this paper addresses the problem of optimally constructing the population demarcation $U = U_a \bigcup U_b$ under the mixed regression-based estimator $\hat{t}_R = t_a + \hat{t}_{Rb}$ for any desired sample design $p(s;\lambda)$. The optimality criterion under which the population demarcation is obtained depends on whether a model-assisted approach or an extended design-based formulation for the skewed population (developed below) is adopted. Since there is particular interest in identifying an optimal sample design $p^*(s;\lambda^*)$ in order to minimize the sample size, an analogue of the design optimality result from the model-assisted framework is also obtained for the alternative finite population framework.

2.1 Design Optimality in the Model-Assisted Survey Sampling Framework

In the model-assisted approach, underlying the class of generalized regression estimators for the population total are regression models (Sarndal, p.255) assumed to link the survey variables y with their auxiliary covariates x. For example, a ratio-form linear heteroscedastic model for the estimator (2.1) in which variability in the survey response increases with the size of the unit is represented by the model

$$y_k = \beta x_k + \epsilon_k \tag{2.2}$$

where $\epsilon_k \sim (0, \sigma_k^2)$ is the random error component.

Under model (2.2), optimal design theory (Godambe and Joshi, 1965) suggests that choosing the probabilities of inclusion $\pi_k \alpha \sigma_k$, $k \epsilon U$, would minimize the anticipated variance of the estimator $\hat{t}_{Rb}$ taken with respect to both the sample design $p(s)$ and the model ξ given by

$$\begin{aligned} \varepsilon_\xi V_p(\hat{t}_{Rb}) &\doteq \Sigma_{U_b}(\frac{1}{\pi_k} - 1)\, \varepsilon_\xi(y_k^2) \\ &= \Sigma_{U_b}(\frac{1}{\pi_k} - 1)\, \sigma_k^2. \end{aligned} \tag{2.3}$$

Furthermore, if σ_k^2 depends on the auxiliary measure x_k according to the formulation $\sigma_k^2 = c\, x_k^\gamma$ as proposed by Sarndal et. al. (1992, p. 462), then the design optimality condition leads to the result that $\pi_k \propto x_k^{\gamma/2}$, $k \epsilon U_b$.

The approach described above falls within the realm of model assisted survey sampling and estimation. The estimator $\hat{t}_{Rb}$ uses the relationship between the survey data y_k and the auxiliary data x_k

(based on a linear model) to improve its efficiency. If the model fit is good, then the efficiency is high and confidence intervals for point estimates are narrow; if the model is bad, then the efficiency declines and confidence intervals for the point estimators are large. In either event, the probability weights π_k^{-1} ensure that the estimator $\hat{t}_{Rb}$ is (asymptotically) unbiased and consistent with respect to the sample design.

With regard to the optimal sample design, model assisted sampling requires that first order probabilities be chosen so that $\pi_k \propto x_k^{\gamma/2}$. Putting the model at the forefront leads to the conclusion that the heteroscedasticity parameter γ is a superpopulation parameter.

2.2 Design Optimality in the Extended Framework for Skewed Finite Populations

The survey framework for skewed populations described below allows all the intermediate parameters such as B and γ to be viewed entirely as design-based constructs. The heteroscedasticity parameter γ is introduced into a purely design-based view of the population data witho[illegible]quiring an explicit specification of a model for the survey variable y_k in terms of a deterministic superpopulation component $x_k \beta$ and a stochastic error component ϵ_k. The essential concepts are laid out below.

The skewed survey population is seen as composed of a three tuple:

$$\{(x_k, y_k; E_k)\}_{k=1}^{N} \tag{2.4}$$

where $E_k = y_k - Bx_k$ is a residual parameter to the data combination x_k and y_k defined through the finite population regression coefficient B.

It is helpful to visualize the population data given by (2.4) as a scatterplot of N points. Conceptually, the finite population parameter B is the slope of a population line (Bx) running through the middle of the population scatter. Mathematically, it is defined as a value which minimizes the sum of squared population residuals $E_k = y_k - B x_k$ around the population line $B x$:

$$B_0^{*} = \left\{ B : \min_{B} \sum_{k=1}^{N} (y_k - B x_k)^2 = SS_N(B) \right\}. \tag{2.5}$$

One also observes in skewed survey populations that variability in the survey response y

tends to increase with the size of the population unit x. In the population scatter (see Figure 1) this behaviour shows up as a "fanning out" pattern of points (x_k, y_k) along the population line $B\,x$ as the value of x increases. Hence, the population residuals are an increasing function of x_k. This relationship may be specified as $E_k^2 \propto h(x_k)$. A parameterized form of $h(x_k)$ which is general, yet simple enough to be practically useful is $h(x_k) \alpha x_k^\gamma$, where $\gamma \geq 0$ is the finite population heteroscedasticity parameter. Furthermore, the relationship between E_k^2 and $h(x_k)$ is not perfect, meaning that all E_k^2 do not lie on the curve $h(x_k)$. Keeping this in view, the relationship between the magnitude of the residual squared E_k^2 and the size x_k of unit k can be posited as

$$E_k^2 \propto x_k^\gamma\, \eta_k \qquad (2.6)$$

where η_k is the multiplicative deviation explaining the fact that all E_k^2 do not lie on the curve $c\,x_k^\gamma$. The deviations $\eta_1, \ldots, \eta_N$ are not assumed to be realizations from

Fig.1 Example of Population Scatterplot (X, Y)

X = Population Size, Y = Revenues (Ontario)

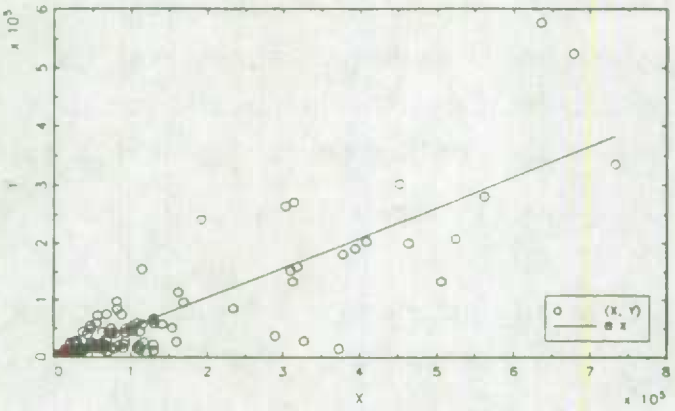
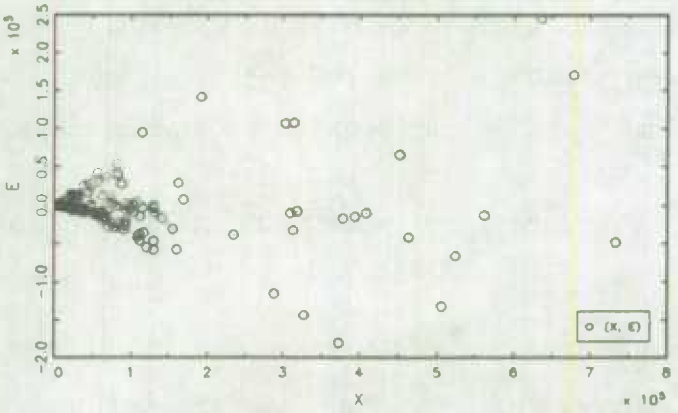


Fig.2 Example of Population Residual Scatterplot (X, E)

X = Population Size, Y = Revenues, E = Y − B*X (Ontario)



some probability distribution, but instead are unknown and unobservable constants in the population. In the case where γ is unknown (most plausible situation), they are also inestimable.

The relationship between the squared population residual E_k^2, the transformed size value $x_k^{\gamma/2}$, and the deviation η_k spelled out in (2.6) is also a finite population model. We can think of it as a "second order" model on the population residuals. It is directly analogous to the (1st order) finite population regression model in which the errors are additive; here the disturbances are multiplicative. Also note that the second order model defined by (2.6) is weaker than the probability model on the residuals $\epsilon_k \sim (0, \sigma_k^2)$ assumed in the model-assisted set-up.

Using this additional knowledge about the behaviour of population residuals, we define our finite population regression parameter B in the size-proportional variance setting as follows

$$B^*(\gamma) = \left\{ B : \min_{B} \sum_{k=1}^{N} \frac{(y_k - B x_k)^2}{x^\gamma} = SS_N(B;\gamma) \right\} \tag{2.5a}$$

Given γ, $SS_N(B;\gamma)$ can be seen as a weighted sum of the population squared residuals $E_k = y_k - B x_k$ around the population line Bx.

In the alternative framework described above, the parameter γ is a population parameter quantifying the spread of the residuals E_k around the finite population line Bx. Of course, γ is unknown but in repeat surveys may be estimated using sample data available from a previous survey. The finite population heteroscedasticity parameter γ is now defined.

Taking the natural logarithm on both sides of (2.6) yields the linearized equation

$$\ln(E_k^2) = \gamma \ln x_k + \ln(\eta_k c) \tag{2.7}$$

where c is the constant of proportionality. Further, defining $\eta_k^* = \ln(\eta_k c)$ as the new additive error term, suggests the following finite population least squares definition of γ:

$$\gamma^*(B) = \left\{ \gamma : \min_{\gamma} \sum_{k=1}^{N} \left(\ln(E_k)^2 - \gamma \ln x_k \right) = SS_N(\gamma;B) \right\} \tag{2.8a}$$

or, alternatively, using $E_k = y_k - B x_k$, as

$$\gamma^*(B) = \left\{ \gamma : \min_{\gamma} \sum_{k=1}^{N} \left(\ln(y_k - Bx_k)^2 - \gamma \ln x_k \right) = SS_N(\gamma\,; B) \right\}. \tag{2.8}$$

Next, let B^* ($\equiv B^*(\gamma^*)$) and γ^* ($\equiv \gamma^*(B^*)$) represent the finite population parameter values which jointly minimize $SS_N(B;\gamma)$ and $SS_N(\gamma\,;B)$ in (2.5a) and (2.8), respectively. In this formulation, the survey problem is to first estimate B^* and γ^* from the sample data and then employ the regression estimator for the total $\hat{t}_{Rb}$ given in (2.1). Moreover, in a repeat survey setting, knowledge of γ^* derived from a previous sample may be used to identify a new efficient sample design $p^*(s;\lambda=\hat{\gamma}^*)$ (generalized pps) defining the first order inclusion probabilities $\pi_k(\lambda) = n \left(x_k^{\lambda/2} / \sum_U x_j^{\lambda/2} \right)$, $k \epsilon U_b$.

In contrast, under the model-assisted set-up, the parameter γ entered through the superpopulation model $y_k = B x_k + \epsilon_k$ with $Var(\epsilon_k) = \sigma_k^2 = c x_k^{\gamma}$. Design optimality $(\pi_k^* \propto \sigma_k)$, then implied that choosing a sample with $\pi_k^* \propto x_k^{\gamma/2}$ would minimize the anticipated design variance of the regression estimator of the total.

The alternative formulation enables design optimality to be defined through the design variance of $\hat{t}_R$ as shown below. First note that the variance of regression estimator $\hat{t}_{R,b}$ under the sampling distribution $p(s)$ (with replacement sampling) is given by

$$V_p(\hat{t}_{Rb}) = \sum_{U_b} (\frac{1}{\pi_k} - 1)\, E_k^2 \tag{2.9}$$

where $E_k = (y_k - B x_k)$ is the population residual. Minimizing (2.9) subject to the constraint $\sum_{U_b} \pi_k = n_b$ leads to the optimal specification of inclusion probabilities given by

$$\pi_k \propto |E_k|\,,\ k \epsilon U_b. \tag{2.10}$$

Using the relationship described between E_k^2 and x_k given by (2.6) in (2.10) yields

$$\pi_k^* \propto x_k^{\gamma/2}\, |\eta_k|\,,\ k \epsilon U_b. \tag{2.11a}$$

The problem with (2.11a) is that although γ can be estimated from past sample data, the population deviations $\eta_1, \ldots, \eta_N$ are unknown and inestimable. If, however, the relationship between the squared residual E_k^2 and the size value x_k defined in $E_k^2 \propto x_k^{\gamma}\, \eta_k$ (2.6) holds well, then the influence of the disturbance η_k on x_k^γ will be small and selecting population units according to the inclusion probabilities

$$\hat{\pi}_k^* \propto x_k^{\gamma/2},\ k\in U_b, \tag{2.11}$$

will be close to optimal. The loss in efficiency due to this approximation is given by

$$V_p(\hat{t}_{R,b};\pi_1^*,\dots,\pi_N^*) - V_p(\hat{t}_{R,b};\hat{\pi}_1^*,\dots,\hat{\pi}_N^*) = \sum\nolimits_{U_b}\left[\frac{\sum_{U_b} x_j^{\gamma/2}\eta_j}{\eta_k} - \sum\nolimits_{U_b} x_j^{\gamma/2}\right]\frac{E_k^2}{n\,x_k^{\gamma/2}}. \tag{2.12}$$

The conditions under which this loss is minimum are apparent from (2.12). Firstly, the smaller the variation among the deviations $\eta_1,\dots,\eta_N$, the smaller will be the departure from the optimal variance. In the extreme case when all the deviations are of constant size ($\eta_k=\eta$, k=1,...,N,) the efficiency loss is zero. Secondly, a good population fit of the relation (2.6) implies that the impact of the disturbances $\eta_1,\dots,\eta_N$ on the corresponding x_k^γ will be small so that the differences $\left[\frac{\sum_{U_b} x_j^{\gamma/2}\eta_j}{\eta_k} - \sum_{U_b} x_j^{\gamma/2}\right]$, k=1,...,N_b, will also small, leading to a smaller loss in efficiency.

The design optimality condition given by (2.11a) - and estimated by (2.11) - is close to the result obtained under the model-assisted framework where instead of minimizing the design variance $V_p(\hat{t}_{Rb})$, the "anticipated design variance" $\varepsilon_\xi\, V_p(\hat{t}_{Rb})$ (defined as the superpopulation expectation of the design variance) is minimized. Therefore, when (2.6) holds, essentially the same design optimality criteria is obtained as under the model-assisted approach without requiring the superpopulation model and, furthermore, γ is a finite population quantity parameterizing the relationship between the residual E_k^2 and the size variable x_k observed in the population scatterplot.

The advantage of the proposed formulation is that the design-optimality result ($\pi_k \propto x^{\gamma/2}$) can be derived under an augmented design-based set-up in which all parameters such as B and γ are finite population parameters. Moreover, this survey framework for skewed populations enables the design-based variance of the estimator $\hat{t}_{Rb}$ to be used as the criterion for optimally constructing the take-all and sampled groups; the superpopulation assumptions are not required.

Three methods for estimating the finite population heteroscedasticity parameter γ from past survey data called the "Least Squares Approach", the "Maximum Likelihood Approach", and the

"Graphical Approach" are described in Appendix A. Results from applying these methods to Local Government Finance Survey data are also reported.

3. OPTIMAL DETERMINATION OF TAKE-ALL AND SAMPLED SUB-POPULATIONS

In surveys of skewed populations, units with the largest sizes are included in the sample with certainty and are not subject to randomized selection. Taken together, these few large-sized units may constitute a considerable segment of the total quantity being estimated and their definite inclusion in the sample lowers the overall sample size needed to accurately measure the population parameters of interest. Moreover, in these populations, the population characteristics of interest tend to increase in variability as the size (eg. number of employees, farm acreage) of the unit increases. The definite inclusion of these large units then increases the precision of the estimators as the take-all group contributes no sampling variance to the estimates.

In this section, a methodology is proposed to determine the optimal demarcation between the take-all and sampled sub-populations. Recall that the skewed population U is partitioned into two groups $U = U_a \bigcup U_b$ where U_a contains the take-all units, all sampled with probability 1 ($\pi_i = 1,\ i \in U_a$), and U_b forms the sampled group of units to be selected with a probability mechanism ($0 < \pi_i < 1,\ i \in U_b$). The total population and sample sizes are represented by $N = N_a + N_b$ and $n = n_a + n_b$, respectively.

3.1 The Transfer Algorithm

The proposed methodology used to define the population allocation iteratively between the take-all and sampled sub-populations, U_a and U_b, respectively, is based on the following idea. Initially, place all population units in the sampled group, labelling it $U_b^{(0)}$ (the superscript represents the iteration cycle). Hence, the take-all group is an empty set $U_a^{(0)}=\{\varnothing\}$. The resulting population and sample sizes at $l=0$ are given by $N_a^{(0)}=0$, $n_a^{(0)}=0$, $N_b^{(0)}=N$, and $n_b^{(0)}=n$ where n is the current sample size of the survey to be re-designed. Furthermore, the design variance of $\hat{t}_R = t_a + \hat{t}_{Rb}$ (using the fact that $V(\hat{t}_R)=V(\hat{t}_{Rb})$) is given by

$$V^{(l)}(\hat{t}_{Rb};\lambda,N_b^{(l)},n_b^{(l)})=\sum_{U_b^{(l)}}\left[\frac{1}{\pi_k(\lambda)}-1\right]E_k^2 \tag{3.1}$$

where

$$\pi_k(\lambda)=n_b^{(l)}\left[x_k^{\lambda/2}\ /\ \sum_{k=1}^{N_b^{(l)}}x_k^{\lambda/2}\right] \tag{3.2}$$

is the probability with which unit k is chosen. Note that λ is used here to parameterize the sample design to allow greater generality when $\lambda \neq \gamma$. The component t_a of the estimator $\hat{t}_R = t_a + \hat{t}_{Rb}$ is a census total of the take-all sub-population in U_a and, therefore, has design variance zero.

If the relationship between the squared residual E_k^2 and the size value x_k defined in $E_k^2 = c\ x_k^{\gamma}\ \eta_k$ (2.6) holds well in the population, then the impact of the disturbance η_k on x_k^{γ} will be small. The values E_k^2 in (3.1) can then be empirically modelled as $E_k^2 = c\,x_k^{\gamma}$ (2.6a) where γ and c are estimated from current sample data by employing the methods of Appendix A. Using the estimated version of (2.6a) in (3.1) yields the following estimator for $V^{(l)}(\hat{t}_{Rb};\cdot)$:

$$\hat{V}^{(l)}(\hat{t}_R;\lambda,N_b^{(l)},n_b^{(l)})=\sum_{U_b^{(l)}}\left[\frac{1}{\pi_k(\lambda)}-1\right]\hat{c}\,x_k^{\hat{\gamma}}. \tag{3.3}$$

In the iterative algorithm, we start initially with all population units placed in $U_a^{(0)}$. Using this population configuration, $V^{(0)}(\hat{t}_{Rb};\cdot)$ is computed. At the first iteration ($l=1$), the largest x-valued unit in $U_b^{(0)}$ is transferred to $U_a^{(0)}=\{\varnothing\}$. As a result, $U_a^{(0)}$ now has one population unit and one sampled unit since all units in the take-all group are placed in the sample: $N_a^{(1)}=N_a^{(0)}+1$ and $n_a^{(1)}=n_a^{(0)}+1$. On the other hand, $U_b^{(0)}$ has lost both a population and sample unit, yielding population and sample sizes $N_b^{(1)}=N_b^{(0)}-1$ and $n_b^{(1)}=n_b^{(0)}-1$, respectively. In general, for any iteration l, $1 \leq l < n$, the relationship between population and sample sizes is described by the following relations: $N_b^{(l)}=N-l$, $n_b^{(l)}=n-l$, and $N_a^{(l)}=n_a^{(l)}=l$. These relations hold because the overall population and sample sizes must remain constant ($N=N_a^{(l)}+N_b^{(l)}$ and $n=n_a^{(l)}+n_b^{(l)}$) for all iterations $1 \leq l < n$.

At iteration l, the gain (loss) in efficiency due to the transfer of the lth largest x-valued unit, denoted by $x_{(N-l)}$, is given by the difference

$$\Delta(l) = V^{(l+1)}(\hat{t}_{Rb}\,;\,\lambda, N-l-1\,, n-l-1) - V^{(l)}(\hat{t}_{Rb}\,;\,\lambda, N-l, n-l)\,. \tag{3.4}$$

Negative values of $\Delta(l)$ mean that the transfer of the unit corresponding to the ordered value$x_{(N-l)}$ from $U_b^{(l-1)}$ to $U_a^{(l-1)}$ lead to a decrease in the design variance. Moreover, such transfers will continue to result in a reduction in the variance of $\hat{t}_{Rb}$ as long as $\Delta(l) < 0\,, 1 \le l < n$. An estimator of $\Delta(l)$ may be readily obtained by using the estimator $\hat{V}^{(l)}(\hat{t}_{Rb}; \cdot)$ for $V^{(l)}(\hat{t}_{Rb}; \cdot)$ as defined by (3.3) in (3.4) to yield $\hat{\Delta}(l)$.

Let $l^{*}(\lambda)\,, 0 \le l^{*} < n$, represent the solution to the Transfer Algorithm under the sample design $p(s\,;\lambda)$. The optimal population allocation to the take-all group $U_a^{*}(l^{*})$ is then given by the population units coinciding with the l^{*} ordered units transfered to the auxiliary vector $X_a^{*} = \left(x_{(N-l^{*})}, x_{(N-.\ -1)}, \ldots, x_{(N)}\right)$. Correspondingly, the optimal population allocation to the sampled group $U_b^{*}(l^{*})$ coinciding with the units in $X_b^{*} = \left(x_{(1)}, x_{(2)}, \ldots, x_{(N-l^{*}-1)}\right)$. The solution $l^{*}(\lambda), 0 \le l^{*} < n$, is also constrained by the condition $\pi_k(\lambda) < 1\,, k \epsilon\, U_b^{*}(l^{*})$. Note that if $\pi_{(N-l^{*})} < 1$, then $\pi_{(N-k)} < 1\,, l^{*} \le k \le n$, since $x_{(N-k)}^{\lambda/2} \le x_{(N-l^{*})}^{\lambda/2}\,, l^{*} \le k \le n$. These observations suggest that the solution for the optimal construction of the take-all and sampled sub-population, represented by $U_a^{*}(l^{*})$ and $U_b^{*}(l^{*})$, respectively, under the Transfer algorithm is given by

$$l^{*}(\lambda) = \min_{l} \left\{ l\!:\ \left[\pi_{(N-l)}(\lambda) < 1\right] \mathit{and}\ \hat{\Delta}(l) = \left[\hat{V}^{(l+1)}(\hat{t}_{Rb}\,;\,\lambda, N-l-1\,, n-l-1) - \hat{V}^{(l)}(\hat{t}_{Rb}\,;\,\lambda, N-l, n-l)\right] \ge 0\,, \atop 0 \le l < n \right\} \tag{3.5}$$

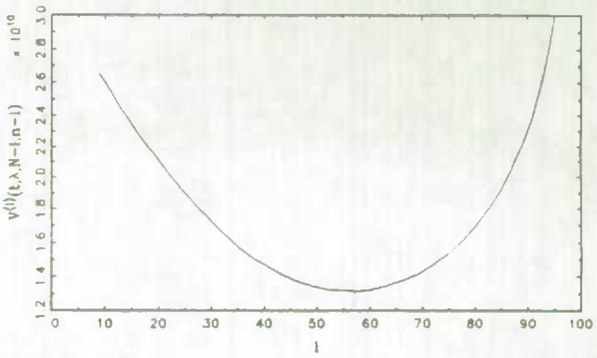
It is also clear from (3.5) that the solution to the Transfer Algorithm $l^{*}(\lambda)$ also depends on the sample design $p(s\,;\lambda)$ - indexed by λ - in effect.

Mathematical properties of the solution (3.5) of the Transfer algorithm such as existence and optimality are studied in detail in Section 4.2, however, the following general observations can be made here. Transferring a unit from $U_b^{(l-1)}$ to $U_a^{(l-1)}$ causes two opposite effects on the variance

$V^{(l)}(\hat{t}_{Rb};\cdot)$. The reduction in the population size $(N_b^{(l)} = N_b^{(l-1)} - 1)$ has the impact of decreasing the variance, while the equivalent reduction in the sample size $(n_b^{(l)} = n_b^{(l-1)} - 1)$ has the reverse effect of increasing $V^{(l)}(\hat{t}_{Rb}^{(l)};\cdot)$. Somewhere in this process, a critical value l^*, $0 \le l^* < n$, exists which gives the optimal breakdown $\{U_a^*(l^*),\, U_b^*(l^*)\}$.

Fig 3 Variance of Regression Estimator at l (λ=1)

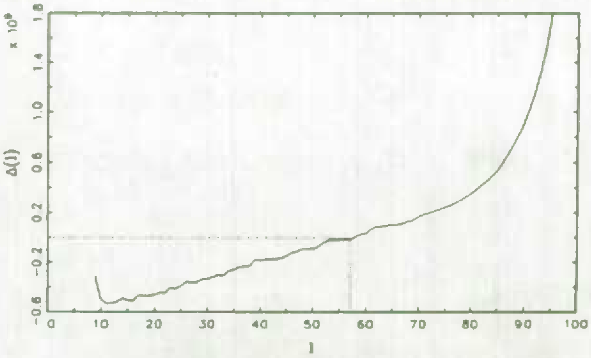
N=Population Size, n=Current Sample Size, l=Units Transferred



The behaviour of this system is also affected by the initial sample size $n^{(0)} = n$ and the distribution of the values $x_k^{\gamma/2}$, $k \in U$, in the population. It is also possible that for certain configurations, $\Delta(l) > 0$ holds for all $1 \le l < n$. This means that no efficiency gains can be realized from transferring units as described in the proposed methodology; the optimal construction of take-all and sampled groups is then given by $U_a^* = \{\varnothing\}$ and $U_b^* = U$, with $l^* = 0$.

Fig.4 Changes in Variance of Regression Estimator (λ=1) :

$\Delta(l) = V^{(l+1)}(t;1,N-l-1,n-l-1) - V^{(l)}(t;1,N-l,n-l)$



An example of the application of the Transfer Algorithm to the LGF survey population of local municipalities in Ontario (with N=793, n=108, γ=2, and λ=1) is given in Figures 3 and 4. The curves are plotted for $l > 8$ because in the interval $0 < l \le 8$, the first condition of (3.5), namely $[\pi_{(N-l)}(\lambda) < 1]$, is not satisfied. Note that in Figure 3, the minimum value of $\hat{V}^{(l)}(\hat{t}_{Rb})$ is achieved at $l^* = 57$ and in Figure 4 this point coincides with

$\Delta(l^*) = \hat{V}^{(l^*+1)} - \hat{V}^{(l^*)} \geq 0$. In Table 3.0, the solution $l^*(\lambda)$ to the Transfer Algorithm as defined in (3.5) and (3.18) are reported for $0 \leq \lambda \leq 2\gamma$ $(\gamma=2)$.

Table 3.0 Solution to Transfer Algorithm $l^*(\lambda)$ for $0 \leq \lambda \leq 2\gamma$ $(\gamma=2)$

λ	$l^*(\lambda)$ **Definition (3.5)**	$l^*(\lambda)$ **Definition (3.18)**
0	64	64
.5	60	60
1.0	57	57
1.5	50	50
2.0	39	39
2.5	50	50
3.0	57	57
3.5	60	60
4.0	64	64

3.2 Analysis of the Transfer Algorithm

In this Section, the Transfer algorithm described in the previous section is analyzed. This is done with two questions in mind: i) does the algorithm converge to a solution? and ii) is the solution optimal? Furthermore, the analysis below reveals that the solution defined by (3.5) can be expressed equivalently in terms of quantities which are much simpler to compute. This equivalence result is established first before investigating the properties of the solution.

3.2.1 Equivalence Result

From the expression for the variance of $V^{(l)}(\hat{t}_{Rb};\cdot)$ given in (3.1), we have after substituting for $\pi_k(\lambda)$

$$V^{(l)}(\hat{t}_{Rb};\lambda, N^{(l)}, n^{(l)}) = \sum_{k=1}^{N^{(l)}} \left[\frac{\sum_{j=1}^{N^{(l)}} x_{(j)}^{\lambda/2}}{n^{(l)}\, x_{(k)}^{\lambda/2}} - 1 \right] E_{(k)}^2 \qquad (3.6a)$$

or, equivalently,

$$V^{(l)}(\hat{t}_{Rb};\lambda, N-l, n-l) = \sum_{k=1}^{N-l-1} \left[\frac{\sum_{j=1}^{N-l-1} x_{(j)}^{\lambda/2} + x_{(N-l)}^{\lambda/2}}{(n-l)\, x_{(k)}^{\lambda/2}} - 1 \right] E_{(k)}^2 + \left[\frac{\sum_{j=1}^{N-l} x_{(j)}^{\lambda/2}}{(n-l)\, x_{(N-l)}^{\gamma/2}} - 1 \right] E_{(N-l)}^2 . \qquad (3.6)$$

The subscript b in $N_b^{(l)}$ and $n_b^{(l)}$ may be dropped since by definition $N^{(l)} = N_b^{(l)} = N - l$ is the population size of $U_b^{(l)}$ and $n^{(l)} = n_b^{(l)} = n - l$ is the resulting sample size. Moreover, at iteration $l+1$, the variance expression $V^{(l-1)}(\hat{t}_{Rb};\cdot)$ may be written as

$$V^{(l+1)}(\hat{t}_{Rb};\lambda, N-l-1, n-l-1) = \sum_{k=1}^{N-l-1} \left[\frac{\sum_{j=1}^{N-l-1} x_{(j)}^{\lambda/2} - (n-l-1)\, x_{(k)}^{\lambda/2}}{n-l-1} \right] \frac{E_{(k)}^2}{x_{(k)}^{\lambda/2}} \qquad (3.7$$

After matching common terms and some further reductions, the difference of the variances $\Delta(l) = V^{(l+1)} - V^{(l)}$ may be written as

$$V^{(l+1)} - V^{(l)} = \sum_{k=1}^{N-l-l} \left[\frac{\sum_{j=1}^{N-l-1} x_{(j)}^{\lambda/2} - (n-l-1)\, x_{(N-l)}^{\lambda/2}}{(n-l)(n-l-1)} \right] \frac{E_{(k)}^2}{x_{(k)}^{\lambda/2}} - \left[\frac{\sum_{j=1}^{N-l} x_{(j)}^{\lambda/2} - (n-l)\, x_{(N-l)}^{\lambda/2}}{(n-l)} \right] \frac{E_{(N-l)}^2}{x_{N-l}^{\lambda/2}} . \qquad (3.8)$$

An estimator of the above expression based on (3.3) is given by

$$\hat{V}^{(l+1)} - \hat{V}^{(l)} = c \sum_{k=1}^{N-l-l} \left[\frac{\sum_{j=1}^{N-l-1} x_{(j)}^{\lambda/2} - (n-l-1)\, x_{(N-l)}^{\lambda/2}}{(n-l)(n-l-1)} \right] x_{(k)}^{\gamma-\lambda/2} - c \left[\frac{\sum_{j=1}^{N-l} x_{(j)}^{\lambda/2} - (n-l)\, x_{(N-l)}^{\lambda/2}}{(n-l)} \right] x_{N-l}^{\gamma-\lambda/2} . \qquad (3.9a)$$

where $c > 0$ and $\gamma \geq 0$ are estimated from modelling the relation $E_k^2 = c\, x_k^{\gamma}$ (2.6a) using the methods of in Appendix A.

Expression (3.9a) further reduces to

$$\hat{V}^{(l+1)} - \hat{V}^{(l)} = c \, \frac{A(l)\; B(l)}{(n-l)\;(n-l-1)} . \qquad (3.9)$$

where $A(l) = \sum_{j=1}^{N-l} x_{(j)}^{\lambda/2} - (n-l)\, x_{(N-l)}^{\lambda/2}$ and $B(l) = \sum_{k=1}^{N-l} x_{(k)}^{\gamma-\lambda/2} - (n-l)\, x_{(N-l)}^{\gamma-\lambda/2}$. For notational convenience, in the remainder of the paper $V^{(l)}$ will be re-defined to represent the estimator of the design variance defined in (3.3) and (3.9).

Next, note that $V^{(l+1)} - V^{(l)} < 0$ in the cases i) $[A(l) > 0 \text{ and } B(l) < 0]$ and ii) $[A(l) < 0 \text{ and } B(l) > 0]$ and $V^{(l+1)} - V^{(l)} \geq 0$ when iii) $[A(l) \geq 0 \text{ and } B(l) \geq 0]$ and iv) $[A(l) \leq 0 \text{ and } B(l) \leq 0]$. In case i), the condition on $n^{(l)} = n - l$ under which $B(l) < 0$ is determined to be

$$n^{(l)} > \left[\frac{\sum_{k=1}^{N-l} x_{(k)}^{\gamma-\lambda/2}}{x_{(N-l)}^{\gamma-\lambda/2}} \right]. \tag{3.10}$$

Moreover, defining

$$R(l;\gamma-\lambda/2) = \left[\frac{\sum_{k=1}^{N-l} x_{(k)}^{\gamma-\lambda/2}}{x_{(N-l)}^{\gamma-\lambda/2}} \right] \tag{3.11}$$

allows (3.9) to be written compactly as

$$n^{(l)} > R(l;\gamma-\lambda/2). \tag{3.12}$$

Similarly, the condition on $n^{(l)} = n - l$ required for $A(l) > 0$ is obtained by solving $A(l) = \sum_{j=1}^{N-l} x_{(j)}^{\lambda/2} - (n-l)\, x_{(N-l)}^{\lambda/2} > 0$. This yields

$$n^{(l)} < \left[\frac{\sum_{k=1}^{N-l} x_{(k)}^{\lambda/2}}{x_{(N-l)}^{\lambda/2}} \right] \tag{3.13}$$

which using an analogous definition to (3.12) may be re-expressed as

$$n^{(l)} < R(l;\, \lambda/2). \tag{3.14}$$

Similar conditions for $n^{(l)} = n - l$ can be derived for the remaining cases ii), iii), and iv) mentioned above. The results are summarized in Table 3.1.

Table 3.1 Outcomes for $V^{(l+1)}-V^{(l)}<0$ and $V^{(l+1)}-V^{(l)}\geq 0$ in Terms of $n^{(l)}=n-l$.

	$V^{(l+1)}-V^{(l)}<0$		$V^{(l+1)}-V^{(l)}\geq 0$
Behaviour of A and B.	Condition on $n^{(l)}=n-l$	Behaviour of A and B.	Condition on $n^{(l)}=n-l$
$A(l)>0$ $B(l)<0$	$R(l;\gamma-\lambda/2)<n-l<R(l;\lambda/2)$ (T.1)	$A(l)>0$ $B(l)\geq 0$	$n-l\leq \min\{R(l;\lambda/2),\ R(l;\gamma-\lambda/2)\}$ (T.2)
$A(l)<0$ $B(l)>0$	$R(l;\lambda/2)<n-l<R(l;\gamma-\lambda/2)$ (T.3)	$A(l)\leq 0$ $B(l)\leq 0$	$n-l\geq \max\{R(l;\lambda/2),\ R(l;\gamma-\lambda/2\}$ (T.4)

The first and second columns of Table 3.1 describes the behaviour of $A(l)$ and $B(l)$ leading to the outcomes $V^{(l+1)}-V^{(l)}<0$ and $V^{(l+1)}-V^{(l)}\geq 0$. The second and fourth columns describe the equivalent condition in terms of $n^{(l)}=n-l$ which yield $V^{(l)}-V^{(l-1)}<0$ and $V^{(l)}-V^{(l-1)}\geq 0$. Since all ranges for $n^{(l)}=n-l$ depend on $R(l;\lambda/2)$ and $R(l;\gamma-\lambda/2)$, the solution to the Transfer algorithm $l^*(\lambda)$ given in (3.5), along with the mathematical properties of the solution, will depend on the distribution of the size measure x and the values of γ and λ.

The behaviour of the system described in Table 3.1 depends on the sample design$p(s;\lambda)$ (indexed by the heteroscedasticity parameter λ) employed. Three cases are distinguished and discussed below: a) $\lambda<\gamma \Rightarrow [R(l;\gamma-\lambda/2)<R(l;\lambda/2)]$, b) $\lambda=\gamma \Rightarrow [R(l;\gamma-\lambda/2)=R(l;\lambda/2)]$, and c) $\gamma<\lambda\leq 2\gamma \Rightarrow [R(l;\gamma-\lambda/2)>R(l;\lambda/2)]$. Although $\lambda>2\gamma$ is also possible and mathematically defined, this situation is arbitrarily ruled out because it leads to $\gamma-\lambda/2<0$ (this term appears as the exponent in $R(l;\gamma-\lambda/2)$). Using the implications of each case for λ on the relative ordering of $R(l;\gamma-\lambda/2)$ and $R(l;\lambda/2)$ in Table 3.1, leads to the systems described in Tables 3.2 to 3.4.

Table 3.2 Case a) $\lambda < \gamma$: Outcomes for $V^{(l+1)}-V^{(l)} < 0$ and $V^{(l+1)}-V^{(l)} \geq 0$ in Terms of $n^{(l)} = n-l$.

	$V^{(l+1)}-V^{(l)} < 0$		$V^{(l+1)}-V^{(l)} \geq 0$
Behaviour of A and B.	Condition on $n^{(l)} = n-l$	Behaviour of A and B.	Condition on $n^{(l)} = n-l$
$A(l) > 0$ $B(l) < 0$	$R(l;\gamma-\lambda/2) < n-l < R(l;\lambda/2)$ (T.1)	$A(l) > 0$ $B(l) \geq 0$	$n-l \leq R(l;\gamma-\lambda/2)$ (T.2)
$A(l) < 0$ $B(l) > 0$	*NOT POSSIBLE* (T.3)	$A(l) \leq 0$ $B(l) \leq 0$	$n-l \geq R(l;\lambda/2)$ (T.4)

Table 3.3 Case b) $\lambda = \gamma$: Outcomes for $V^{(l+1)}-V^{(l)} < 0$ and $V^{(l+1)}-V^{(l)} \geq 0$ in Terms of $n^{(l)} = n-l$.

	$V^{(l+1)}-V^{(l)} < 0$		$V^{(l+1)}-V^{(l)} \geq 0$
Behaviour of A and B.	Condition on $n^{(l)} = n-l$	Behaviour of A and B.	Condition on $n^{(l)} = n-l$
$A(l) > 0$ $B(l) < 0$	*NOT POSSIBLE* (T.1)	$A(l) > 0$ $B(l) \geq 0$	$n-l \leq R(l;\gamma/2)$ (T.2)
$A(l) < 0$ $B(l) > 0$	*NOT POSSIBLE* (T.3)	$A(l) \leq 0$ $B(l) \leq 0$	$n-l \geq R(l;\gamma/2)$ (T.4)

Table 3.4 Case c) $\gamma < \lambda \leq 2\gamma$: Outcomes for $V^{(l+1)} - V^{(l)} < 0$ and $V^{(l+1)} - V^{(l)} \geq 0$ in Terms of $n^{(l)} = n - l$.

	$V^{(l+1)} - V^{(l)} < 0$		$V^{(l+1)} - V^{(l)} \geq 0$
Behaviour of A and B.	Condition on $n^{(l)} = n - l$	Behaviour of A and B.	Condition on $n^{(l)} = n - l$
$A(l) > 0$ $B(l) < 0$	*NOT POSSIBLE* (T.1)	$A(l) > 0$ $B(l) \geq 0$	$n-l \leq R(l;\lambda/2)$ (T.2)
$A(l) < 0$ $B(l) > 0$	$R(l;\lambda/2) < n-l < R(l;\gamma-\lambda/2)$ (T.3)	$A(l) \leq 0$ $B(l) \leq 0$	$n-l \geq R(l;\gamma-\lambda/2)$ (T.4)

An important condition required for the solution to the Transfer Algorithm

$$l^*(\lambda) = \min_{l} \left\{ l: \; \left[\pi_{(N-l)}(\lambda) < 1\right] \; \textit{and} \; \hat{\Delta}(l) = \left[\hat{V}^{(l+1)}(\hat{t}_{Rb}\,;\, \lambda, N-l-1, n-l-1) - \hat{V}^{(l)}(\hat{t}_{Rb}\,;\, \lambda, N-l, n-l)\right] \geq 0, \; 0 \leq l < n \right\} \tag{3.5}$$

is that $\pi_{(N-l)}(\lambda) < 1$. It is easily to show that $\pi_{(N-l)}(\lambda) < 1 \Leftrightarrow A(l) > 0$. In terms of the descriptions given for the Transfer Algorithm in Tables 3.1 to 3.4, this condition means that the solution can occur only when both $A(l) > 0$ and $B(l) \geq 0$ or, equivalently, when $n-l$ satisfies condition (T.2) in each table. The important results from the above analysis are summarized in the following equivalence theorem.

Theorem 1. Equivalence Theorem

Let $p(s;\lambda)$ represent the sample design in effect, defining the inclusion probabilities $\pi_k^{(l)}(\lambda) = (n-l)\, x_{(k)}^{\lambda/2} \,/\, \sum_{k=1}^{N-l} x_{(k)}^{\lambda/2}, \; k \in U_b$. Further, with $R(l;\lambda/2) = \sum_{k=1}^{N-l} x_{(k)}^{\lambda/2} \,/\, x_{(N-l)}^{\lambda/2}$ and $R(l;\gamma-\lambda/2) = \sum_{k=1}^{N-l} x_{(k)}^{\gamma-\lambda/2} \,/\, x_{(N-l)}^{\gamma-\lambda/2}$ defining the critical values for $n-l$, the following equivalences hold for the Transfer Algorithm.

a) $\lambda < \gamma$:

$$\left[\left[\pi_{(N-l)}(\lambda) < 1\right] \textit{ and } \left[V^{(l+1)}(\hat{t}_{Rb}\,;\,\lambda, N-l-1, n-l-1) - V^{(l)}(\hat{t}_{Rb}\,;\,\lambda, N-l, n-l)\right] \geq 0\right] \tag{3.15a}$$

$$\Leftrightarrow \quad \left[n-l \leq R(l;\gamma-\lambda/2)\right]$$

$$\left[V^{(l+1)}(\hat{t}_{Rb}\,;\,\lambda, N-l-1, n-l-1) - V^{(l)}(\hat{t}_{Rb}\,;\,\lambda, N-l, n-l) < 0\right] \tag{3.16a}$$

$$\Leftrightarrow \quad \left[R(l;\gamma-\lambda/2) < n-l < R(l;\lambda/2)\right]$$

$$\left[\left[\pi_{(N-l)}(\lambda) \geq 1\right] \textit{ and } \left[V^{(l+1)}(\hat{t}_{Rb}\,;\,\lambda, N-l-1, n-l-1) - V^{(l)}(\hat{t}_{Rb}\,;\,\lambda, N-l, n-l)\right] \geq 0\right] \tag{3.17a}$$

$$\Leftrightarrow \quad \left[n-l \geq R(l;\lambda/2)\right]$$

b) $\lambda = \gamma$:

$$\left[\left[\pi_{(N-l)}(\gamma) < 1\right] \textit{ and } \left[V^{(l+1)}(\hat{t}_{Rb}\,;\,\gamma, N-l-1, n-l-1) - V^{(l)}(\hat{t}_{Rb}\,;\,\gamma, N-l, n-l)\right] \geq 0\right] \tag{3.15b}$$

$$\Leftrightarrow \quad \left[n-l \leq R(l;\gamma/2)\right]$$

$$\left[\left[\pi_{(N-l)}(\gamma) \geq 1\right] \textit{ anu } \left[V^{(l+1)}(\hat{t}_{Rb}\,;\,\gamma, N-l-1, n-l-1) - V^{(l)}(\hat{t}_{Rb}\,;\,\gamma, N-l, n-l)\right] \geq 0\right] \tag{3.16b}$$

$$\Leftrightarrow \quad \left[n-l \geq R(l;\gamma/2)\right]$$

c) $\gamma < \lambda \leq 2\gamma$:

$$\left[\left[\pi_{(N-l)}(\lambda) < 1\right] \textit{ and } \left[V^{(l+1)}(\hat{t}_{Rb}\,;\,\lambda, N-l-1, n-l-1) - V^{(l)}(\hat{t}_{Rb}\,;\,\lambda, N-l, n-l)\right] \geq 0\right] \tag{3.15c}$$

$$\Leftrightarrow \quad \left[n-l \geq R(l;\lambda/2)\right]$$

$$\left[V^{(l+1)}(\hat{t}_{Rb}\,;\,\lambda, N-l-1, n-l-1) - V^{(l)}(\hat{t}_{Rb}\,;\,\lambda, N-l, n-l) < 0\right] \tag{3.16c}$$

$$\Leftrightarrow \quad \left[R(l;\lambda/2) < n-l < R(l;\gamma-\lambda/2)\right]$$

$$\left[\left[\pi_{(N-l)}(\lambda) \geq 1\right] \textit{ and } \left[V^{(l+1)}(\hat{t}_{Rb}\,;\,\lambda, N-l-1, n-l-1) - V^{(l)}(\hat{t}_{Rb}\,;\,\lambda, N-l, n-l)\right] \geq 0\right] \tag{3.17c}$$

$$\Leftrightarrow \quad \left[n-l \geq R(l;\gamma-\lambda/2)\right]$$

3.2.2 Simpler Method of Solution

The methodology of finding the optimal allocation of the population to the take-all and sampled groups was originally developed in terms of the behaviour of the difference

$$\Delta(l) = V^{(l+1)}(\hat{t}_{Rb}\,;\,\gamma, N-l-1, n-l-1) - V^{(l)}(\hat{t}_{Rb}\,;\,\gamma, N-l, n-l). \tag{3.4}$$

The analysis performed above of the Transfer Algorithm, and particularly the equivalence established between the behaviour of $V^{(l)}-V^{(l-1)}$ and $\pi^{(l)}_{(N-l)}(\lambda)$ and the terms $n^{(l)}=n-l$, $R(l;\lambda/2)$, and $R(l;\gamma-\lambda/2)$ in Theorem 1, allows the solution $l^*(\lambda)$ to be stated in a greatly simplified - yet equivalent - form. This result is stated in Theorem 2. It is directly obtained from the three components of Theorem 1 keeping in view that the solution $l^*(\lambda)$ must satisfy the two conditions: i) $V^{(l^*+1)}-V^{(l^*)}\geq 0$ and ii) $\left[\pi^{(l^*)}_{(N-l^*)}<1\right]\Leftrightarrow\left[A(l^*)>0\right]$.

Theorem 2. Alternative Solution to Transfer Algorithm

The solution $l^*(\lambda)$ to the Transfer Algorithm stated in (3.5) in terms of $V^{(l)}-V^{(l-1)}$ and $\pi^{(l)}_{(N-l)}(\lambda)$ may also be equivalently expressed as

$$l^*(\lambda)=\left\{\begin{array}{l}\min\limits_{l}\{l:\ n-l\ \leq\ R(l;\gamma-\lambda/2)\ ,\ 0\leq l<n\}\ ,\ \lambda<\gamma\\ \min\limits_{l}\ \{l:\ n-l\ \leq\ R(l;\gamma/2)\ ,\ 0\leq l<n\}\ ,\ \lambda=\gamma\\ \min\limits_{l}\ \{l:\ n-l\ \leq\ R(l;\lambda/2)\ ,\ 0\leq l<n\}\ ,\ \gamma<\lambda<2\gamma\end{array}\right\}. \tag{3.18}$$

An example of how (3.18) can be used to find the optimal population allocation is illustrated in Figure 5 (the same Ontario data for the population of local municipalities is used as in Figures 3 and 4) with γ=2 and λ=1. In this case $\lambda<\gamma$, and the solution is determined by the behaviour of functions $R(l;\gamma-\lambda/2)$ and $n-l$ (see Theorem 2). The same solution $l^*=57$ is obtained as before. Moreover, near l=8, the functions $R(l;\lambda/2)$ and $n-l$ cross. From Table 3.2 it is clear that $\left[\pi^{(l)}_{(N-l)}\geq 1\right]\Leftrightarrow\left[A(l)\leq 0\right]$ for $l\leq 8$ and $\left[\pi^{(l)}_{(N-l)}<1\right]\Leftrightarrow\left[A(l)>0\right]$ for $l>8$.

3.2.3 Existence and Optimality

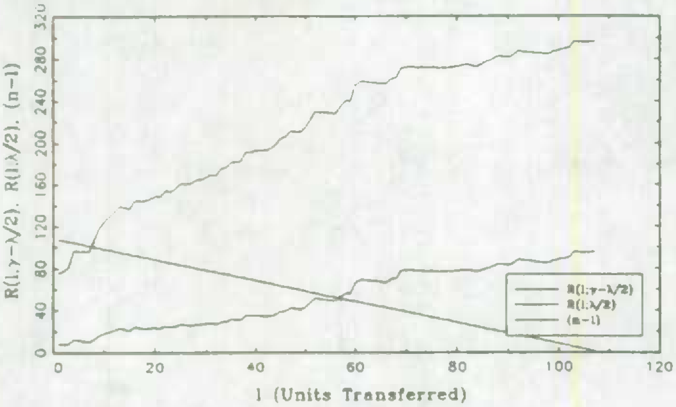


Fig 5 Use of R(l;γ−λ/2), R(l;λ/2), and (n−l) to Construct Optimal Take-all/Sampled Groups (Ontario)

The issue of existence and optimality is concerned with the question of whether 1) the Transfer Algorithm always converge to a solution and 2) is the solution reached globally optimal ? Note that by construction, the solution $l^*(\lambda)$ guarantees local optimality in the region $[0, l^*(\lambda)]$. The second question is concerned with the conditions required for globally optimality. The solution $l^*(\lambda)$ will be optimal if the conditions leading to $l^*(\lambda)$ remain unchanged (stable) in the system defined over the remaining region $(l^*(\lambda), n-1]$. Stability (of conditions) in this region ensures the (global) optimality of $l^*(\lambda)$. In terms of the original formulation, these concepts may be defined as follows:

1) Existence: $\exists\, l^*, 0 \le l^* < n$, such that $V^{(l^*+1)} - V^{(l^*)} \ge 0$ and $\pi^{(l^*)}_{(N-l^*)} < 1$.

2) Stability: If $V^{(l^*+1)} - V^{(l^*)} \ge 0$, then $V^{(l+1)} - V^{(l)} \ge 0$ and $\pi^{(l)}_{(N-l)} < 1$ for $0 \le l^* < l < n$3.

In Section 3.2.1, $V^{(l+1)} - V^{(l)}$ was expressed as

$$V^{(l+1)} - V^{(l)} = \frac{A(l)\ B(l)}{(n-l)\ (n-l-1)}. \tag{3.9}$$

where $A(l) = \sum_{j=1}^{N-l} x_{(j)}^{\lambda/2} - (n-l)\, x_{(N-l)}^{\lambda/2}$ and $B(l) = \sum_{k=1}^{N-l} x_{(k)}^{\gamma-\lambda/2} - (n-l)\ x_{(N-l)}^{\gamma-\lambda/2}$. Additionally, denote l_A^* to be the smallest value of $0 \le l < n$ satisfying $A(l) > 0$; similarly, let l_B^* be the smallest value of $0 \le l < n$ satisfying $B(l) > 0$. At the solution $l^*(\lambda)$, the following two conditions are required: i) $V^{(l^*+1)} - V^{(l^*)} \ge 0$ and ii) $\left[\pi^{(l^*)}_{(N-l^*)} < 1\right] \Leftrightarrow \left[A(l^*) > 0\right]$. Keeping (3.9) in view, this implies that at the solution, the condition $B(l^*) > 0$ must also hold. Therefore, the solution to the Transfer

Algorithm can also be stated as

$$l^* = \max\ \{l_A^*\ , l_B^*\}\ \ ,\ 0 \leq l^* < n. \qquad (3.19)$$

Further, note that because l_A^* and l_B^* are solutions to two independent systems defined over $0 \leq l < n$, we can re-define existence and stability as follows:

1) Existence: $\exists\ l_A^*\ , 0\ \leq l_A^*\ < n$, such that $A(l_A^*) > 0$, and

$\exists\ l_B^*\ , 0\ \leq l_B^*\ < n$, such that $B(l_B^*) > 0$.

2) Stability: If $A(l_A^*) > 0$, then $A(l) > 0$ for $0 \leq l_A^* < l < n$, and

If $B(l_B^*) > 0$, then $B(l) > 0$ for $0 \leq l_B^* < l < n$.

These properties and the conditions under which they hold are now established for the three cases a) $\lambda < \gamma \Rightarrow [R(l;\gamma\text{-}\lambda/2)\ < R(l;\lambda/2)]$, b) $\lambda = \gamma \Rightarrow [R(l;\gamma\text{-}\lambda/2)\ = R(l;\lambda/2)]$, and c) $\gamma < \lambda \leq 2\gamma$ $\Rightarrow [R(l;\gamma\text{-}\lambda/2)\ > R(l;\lambda/2)]$

3.A Existence and Optimality of l^* $(\lambda < \gamma)$

Existence of l^* $(\lambda < \gamma)$

It follows from (3.19) above that l^* exists if both l_A^* and l_B^* exist. Recall that $\lambda < \gamma$ $\Rightarrow [R(l;\gamma\text{-}\lambda/2)\ < R(l;\lambda/2)]$. The fact that $n-l$ is a decreasing function over $0 \leq l < n$ means that the event $A(l_A^*) > 0 \Leftrightarrow \left[n-l_A^*\ \ < R(l_A^*\ ;\lambda/2)\right]$ will occur before $B(l_B^*) > 0 \Leftrightarrow \left[n-l_B^*\ \ < R(l_B^*\ ;\gamma\text{-}\lambda/2)\right]$: $l_A^*\ < l_B^*$.

Existence of l_A^* $(\lambda < \gamma)$

Initially $(l=0)$, two outcomes are possible: i) either $[A(l)\ \leq 0] \Leftrightarrow [n\text{-}l \geq R(l;\lambda/2)]$ or ii) $[A(l) > 0] \Leftrightarrow [n\text{-}l < R(l;\lambda/2)]$. The outcome for l^* will depend on which of these cases occurs.

i) $[A(0)\ \leq 0] \Leftrightarrow [n\ \geq R(0;\lambda/2)]$

Note that $n^{(l)} = n-l$ is a strictly decreasing linear function of l with $\lim\limits_{l \to n} n^{(l)} = 0$. On the other hand, $\lim\limits_{l \to n} R(l;\lambda/2) \geq \lim\limits_{l \to N} R(l;\lambda/2) = 0$. Therefore, given the fact that

$R(0;\lambda/2) \le n$, there exists a $l_A^*, 0 \le l_A^* < n$, such that $R(l_A^*;\gamma/2) > n - l_A^*$ (with $A(l_A^*) > 0$).

ii) $[A(0) > 0] \Leftrightarrow [n < R(0;\lambda/2)]$

Here, initially the function value $R(0;\lambda/2)$ is above $n-0$ so that $A(0) > 0$. Therefore, existence is satisfied at $l_A^* = 0$.

Existence of l_B^* ($\lambda < \gamma$)

The proof for the existence of l_B^* is analogous to that for l_A^* with $A(l)$ and $R(l;\lambda/2)$ replaced by $B(l)$ and $R(l;\gamma-\lambda/2)$, respectively.

Optimality of l^* ($\lambda < \gamma$)

For optimality, the conditions which lead to the solution l^* must hold stable in the region $(l^*(\lambda), n-1]$. By $l^* = \max\{l_A^*, l_B^*\}$, $0 \le l^* < n$, (3.19), stability prevails if the conditions leading to l_A^* and l_B^* in the two independent sub-systems of the Transform Algorithm (defined by $A(l)$ and $B(l)$, respectively) continue to hold in $(l_A^*, n-1]$ and $(l_B^*, n-1]$, respectively.

Stability in $(l_A^*(\lambda), n-1]$ ($\lambda < \gamma$)

Again consider the two possible cases initially ($l=0$) possible.

i) $[A(0) \le 0] \Leftrightarrow [n \ge R(0;\lambda/2)]$

We are assured of at least one solution by existence, however, the system may be unstable if the function $R(l;\lambda/2)$ crosses $n^{(l)} = n-l$ more than once in the range $[0, n)$.

The behaviour of the function $R(l;\lambda/2)$ depends on the value of λ and the distribution of the auxiliary characteristic x in the population. For example, if $\lambda = 0$, then $R(l;\lambda/2) = N-l$ always lies above $n-l$ so that $A(l) > 0, 0 \le l < n$ with $l_A^* = 0$. Similarly, the distribution of the x-values has an effect on the shape of $R(l;\lambda/2)$. For example, if all $x_{(j)} = c, j \in U$, are constant, then again $R(l;\lambda/2) = N-l$ and $l_A^* = 0$. In these situations, no crossings of $R(l;\lambda/2)$ and $n-l$ are realized over $[0,n)$. The more interesting non-trivial cases occur for $\lambda > 0$ and non-homogeneous values of x.

The step change in the function $R(l;\lambda/2)$ over consecutive values of l roughly parallels the idea of a slope for continuous functions. This jump may be expressed as

$$R(l+1\,;\lambda/2) - R(l\,;\lambda/2) = \sum_{k=1}^{N-l-1} x_{(k)}^{\lambda/2} \left[\frac{1}{x_{(N-l-1)}^{\lambda/2}} - \frac{1}{x_{(N-l)}^{\lambda/2}} \right] - 1\,. \tag{3.20}$$

Although the term in ( ) of (3.20) is non-negative since $\left(x_{(N-l)}^{\lambda/2} - x_{(N-l-1)}^{\lambda/2}\right) \geq 0$, the jump can be either positive or negative. However, if all negative jumps of $R(l;\lambda/2)$ are greater than or equal to all jumps of $n^{(l)} = n-l$ given by $n^{(l+1)} - n^{(l)} = -1$, $0 \leq l < n$, then stability is guaranteed: $R(l;\lambda/2)$ crosses $n^{(l)} = n-l$ in only one period $[l_A^*, l_A^* + 1]$. Formally, this condition may be expressed as

$$R(l+1\,;\lambda/2) - R(l;\lambda/2) \;\geq -1\,,\; 0 \leq l < n\,. \tag{3.21}$$

Solving (3.21) yields

$$\left(x_{(N-l)}^{\lambda/2} - x_{(N-l-1)}^{\lambda/2}\right) \geq 0\,,\; 0 \leq l < n, \tag{3.22a}$$

or, with $\lambda > 0$,

$$\left(x_{(N-l)} - x_{(N-l-1)}\right) \geq 0\,,\; 0 \leq l < n. \tag{3.22}$$

Barring the trivial cases ($\lambda = 0$ or $x_{(j)} = c\,, j \in U,$ are constant), the strict inequality of (3.22) holds; thereby, ensuring stability for $A(l)$ in $(l_A^*(\lambda), n-1]$. Note that condition (3.22) allows for ties among consecutive auxiliary values.

ii) $[A(0) > 0] \Leftrightarrow [n < R(0;\lambda/2)]$

This is the second outcome initially possible when $l = 0$. It was shown earlier that in this case $l_A^* = 0$. Stability in this case requires that $R(l;\lambda/2)$ remain above $n^{(l)} = n-l$ for all $0 \leq l \leq n$, with the two curves never crossing. Again conformity of non-trivial cases with condition (3.22) assures that this does not happen.

The strict inequality of condition (3.22) is not met in the cases when $\lambda = 0$ or $x_{(j)} = c\,, j \in U,$ are constant. These cases lead to $R(l;\lambda/2) = N-l \;>\; n^{(l)} = n-l$, $0 \leq l < n$, and all jumps $R(l+1;\lambda/2) - R(l;\lambda/2) = -1$, $0 \le l < n$, of the function $R(l;\lambda/2)$ are negative and equal to the jumps of $n^{(l)} = n-l$. The two curves never cross (with $A(l) > 0$, $0 \le l < n$), yielding stability in $(l_A^*(\lambda), n-1]$. Hence, stability exists even in the special cases under which condition (3.22) is not met.

Stability in $(l_B^*(\lambda), n-1]$ $(\lambda < \gamma)$

The proof for stability in $(l_B^*(\lambda), n-1]$ is analogous to that for l_B^* , with $A(l)$ and $R(l;\lambda/2)$ replaced by $B(l)$ and $R(l;\gamma-\lambda/2)$, respectively. The condition required for stability turns out to be the same as (3.22).

3.2.3.B Existence and Optimality of l^* $(\lambda = \gamma$ and $\gamma < \lambda \le 2\gamma)$

The proof for the existence and optimality of $l^*(\lambda)$ in the remaining two cases b) $\lambda = \gamma$ and c) $\gamma < \lambda \le 2$ [illegible] analogous to that for case a) $\lambda < \gamma$ given above and lead to the same results. These results follow easily from using the relevant choice of the functions $R(l;\lambda/2)$ and $R(l;\gamma-\lambda/2)$ and are not repeated again in the interest of brevity.

The results regarding the existence and optimality of the solution delivered by the Transfer Algorithm proved above are summarized in the theorem below.

Theorem 3 Existence and Optimality of Solution to Transfer Algorithm

The Transfer Algorithm always converges to a solution $0 \le l^*(\lambda) < n$ defined in Theorem 2 as

$$l^*(\lambda) = \left\{ \begin{array}{l} \min\limits_{l} \{l : n-l \le R(l;\gamma-\lambda/2) \text{ , } 0 \le l < n\} \text{ , } \lambda < \gamma \\ \min\limits_{l} \{l : n-l \le R(l;\gamma/2) \text{ , } 0 \le l < n\} \text{ , } \lambda = \gamma \\ \min\limits_{l} \{l : n-l \le R(l;\lambda/2) \text{ , } 0 \le l < n\} \text{ , } \gamma < \lambda < 2\gamma \end{array} \right\} . \qquad (3.17)$$

where λ indexes the sample design $p(s;\lambda)$ defining the inclusion probabilities $\pi_k^{(l)}(\lambda) = (n-l)\ x_{(k)}^{\lambda/2} \ / \ \sum_{k=1}^{N-l} x_{(k)}^{\lambda/2}$, $k \in U_b$ and $R(l;\lambda/2) = \sum_{k=1}^{N-l} x_{(k)}^{\lambda/2} \ / \ x_{(N-l)}^{\lambda/2}$ and $R(l;\gamma-\lambda/2) = \sum_{k=1}^{N-l} x_{(k)}^{\gamma-\lambda/2} \ / \ x_{(N-l)}^{\gamma-\lambda/2}$ define the critical values for $n-l$.

For $0 \le \lambda \le 2\gamma$, $\gamma \ge 0$, the solution $l^*(\lambda)$ reached is optimal under the conditions stated below:

a) $\lambda = 0$ and/or $\gamma = 0$ and/or $x_{(j)} = c$, $j \in U$, are constant: the solution $l^*(\lambda) = 0$ is optimal.

b) $\gamma > 0$ and $0 < \lambda \le 2\gamma$: the solution $0 \le l^*(\lambda) < n$ is optimal if

$$\left(x_{(N-l)} - x_{(N-l-1)}\right) \ge 0,\ 0 \le l < n.$$

Note that due to the ordering imposed on the population auxiliary values $\{x_1, x_2, \ldots, x_N\}$, condition $\left(x_{(N-l)} - x_{(N-l-1)}\right) \ge 0$ holds for all $0 \le l < n$. Graphically, this ensures that i) the$R(l;\lambda/2)$ and $R(l;\gamma-\lambda/2)$ curves do not cross $n^{(l)} = n-l$ from above and ii) the $R(l;\lambda/2)$ and$R(l;\gamma-\lambda/2)$ curves cross $n^{(l)} = n-l$ from below only once.

4. CONCLUSIONS

In on-going surveys of skewed populations, the most optimal sample design and estimation strategy holds the promise of offering the largest reduction in the sample size (and hence survey resources) while maintaining a given minimum desired level of precision in the estimates produces. It was mentioned earlier that an overall sample re-design methodology aimed at this purpose needs to address and integrate the solution to three problems:

i) identifying an efficient sample selection scheme,

ii) constructing an efficient demarcation between the take-all and sampled population groups at a given sample size, and

iii) determining the minimal sample size required to meet the precision constraint(s).

This paper dealt with problem of constructing the take-all and sampled groups in surveys of skewed populations through the "Transfer Algorithm" as well as identifying an optimal sample selection scheme. The Transfer Algorithm allows the survey designer to find an optimal allocation of population units between the take-all and sampled population groups in the sense of minimizing directly the design variance of the regression estimator under the desired sample design. Desirable mathematical properties of this algorithm such as existence and optimality of solution were established and an equivalence result was obtained allowing the solution to be determined in terms of simple quantities computable directly from the population auxiliary data.

Due to space, the integration of the ii) and iii) components of the overall sample design methodology was not discussed. This involves a) application of the Transfer Algorithm at the current overall sample size to create the take-all and sampled sub-populations, b) using the precision constraint (in terms of the coefficient of correlation) to find the required sample size in the resulting sampled group, and c) repeating the above two steps as long as further reductions are observed in the overall sample size. Iteration is critical to arrive at the globally minimal sample size and allocation ($n^* = n_a^* + n_b^* = l^* + n_b^*$) because the solution to the Transfer Algorithm $l^*(\lambda, n)$ ($N_a^* = l^*$ and $N_b^* = N - l^*$) also depends on the current total prevailing sample size.

At the final stage, the desired sample design $p(s;\lambda)$ (indexed by λ) may be implemented in the sampled group U_b^*. Note that the solution $l^*(\lambda)$ given by the proposed Transfer Algorithm applies generally to any sample design defined in the range $0 \leq \lambda \leq 2\gamma$; it is not merely defined at the optimal value of the design parameter $\lambda = \gamma$. The presence of λ in the specification of first order inclusion probabilities gives rise to a wide class of generalized pps designs which yield the SRS ($\lambda=0$) and the standard pps design ($\lambda=2$) as special cases. One good approximation to the optimal design ($\lambda = \gamma$) is a stratified design based on the transformed size values $x_k^{\gamma/2}, k \epsilon U_b^*$.

The proposed approach differs from existing methods for constructing the take-all and sampled groups in the literature in three respects. Firstly, an optimal population demarcation can be obtained for a flexible range of sample designs (eg. SRS, pps, generalized pps, stratified) instead of simply stratified SRS designs. Secondly, the criterion used to find the optimal population allocation is based directly on minimizing the design-based variance of the regression estimator under the desired sample design. Thirdly, the proposed methodology explicitly captures the size-induced heteroscedasticity evident in skewed survey populations.

Acknowledgements

I would like to acknowledge the support of Public Institutions Division in sponsoring this work and thank M.S. Kovacevic, S. Wu, P. Lavallee, and H. Mantel for their valuable comments in the development of this paper.

References

Glasser, G.J. (1962). On the Complete Coverage of large Units in a Statistical Study. International Review of the International Statistical Institute, 30, 28-32.

Godambe, V.P., Joshi, V.M. (1965). Admissibility and Bayes Estimation in Sampling Finite Populations. Annals of Mathematical Statistics, 36, 1702-1722.

Hidiroglou, M.A., Srinath, K.P. (1993). Problems Associated With Designing Subannual Business Surveys. Journal of Business and Economic Statistics, Vol. 11, 4, 397-405.

Lavallee, P., Hidiroglou, M.A. (1988). On the Stratification of Skewed Populations. Survey Methodology, Vol. 14.1, 33-43.

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

APPENDIX A. ESTIMATION OF FINITE POPULATION HETEROSCEDASTICITY PARAMETER γ

This Appendix is concerned with the estimation of the heteroscedasticity parameter γ (a finite population definition of γ is given in expression (2.8)). In Section 2, the relationship between the magnitude of the residual squared E_k^2 and the size x_k of unit k was posited as

$$E_k^2 \propto x_k^\gamma \, \eta_k \tag{2.6}$$

where η_k is the multiplicative error explaining the fact that all E_k^2 do not lie on the curve $c\,x_k^\gamma$. The finite population heteroscedasticity parameter γ captures the relationship existing between the magnitude of the disturbance E_k and the size measure x_k for the population unit k. The problem is that γ is not known for the population to be sampled. In repeat surveys like the LGF survey, however, data from previous sampling permits estimation of γ.

Three methods are discussed below for estimating γ. No one method was used uniformly to determine the value of γ in each province. Estimates from the different methods were compared and a value of γ was identified from these comparisons. In order to ascertain the stability of γ over the observation set, these methods were applied to variants of the same data set created by excluding a different number of the largest observations - those with the largest x-values. The approach gave a profile of the behaviour of γ over different size ranges. The values of the heteroscedasticity parameter finally chosen in each province also took this analysis into account. Estimates of γ values obtained after excluded the largest observations - these went into the take-all group, sampled with $\pi_k = 1$ - were favoured across the different γ estimation methodologies. The estimation methods are now described separately below.

A.1 Least Squares (LS) Approach

This LS approach involves linearizing the relationship between residuals E_k and x_k given in equation (2.6) and using the sample estimates of E_k to then fit the linearized equation. First, estimates of residuals $E_k = y_k - x_k B$ are obtained by fitting the regression

$$y_k = x_k B + E_k \tag{A.1}$$

where the estimated residuals are given by $\hat{E}_k = y_k - x_k \hat{B}$. Secondly, the linearized version of (2.6) is fitted using $\hat{E}_k$, $k \epsilon s$.

Upon taking the natural logarithm of both sides of (2.6) we have

$$\ln(E_k^2) = \gamma \ln x_k + \eta_k^* \qquad \text{(A.3)}$$

where $\eta_k^* = \ln(\eta_k c)$ is the additive error component in the linearized form of (2.6). Using $\hat{E}_k^2$ for E_k^2 in (A.3) gives the least squares estimate of the heteroscedasticity parameter γ.

A.2 Maximum Likelihood Approach

This method uses a totally model-based approach to the estimation of γ. Empirical investigations into the relationship between the survey variables y based on past sample data reveal that the model given below captures quiet well the scatter-plot phenomena - an increasing linear trend and increasing heteroscedasticity with x_k - between expenditures (and revenues) y and the population size x:

$$y_k = \beta x_k + \epsilon_k \qquad k = 1, \ldots, n \text{ (A.4)}$$

where

$$\epsilon_k \sim N(0, \sigma^2 x_k^\gamma) \qquad \text{(A.5)}$$

and $\sigma^2 x_k^\gamma$ is the variance function completely specified upon determining γ. The MLE. for γ based on the assumption of normal errors is formed upon solving

$$g(\gamma, \hat{\beta}_\gamma) = \sum_{k=1}^{n} \frac{(y_k - x_k \hat{\beta}_\gamma)}{x_k^\gamma} \left[\ln x_k - \sum_{k=1}^{n} \frac{\ln x_k}{n} \right] = 0 \qquad \text{(A.6)}$$

where

$$\hat{\beta}_\gamma = \frac{\sum_{i=1}^{n} y_i x_i^{(1-\gamma)}}{\sum_{i=1}^{n} x_i^{(2-\gamma)}}. \qquad \text{(A.7)}$$

Expression (A.6) is obtained by first solving the score functions for σ^2 and substituting it into the log-likelihood. Finally, the equation for $g(\gamma, \beta_\gamma)$ above is then obtained by differentiating the resulting concentrated-log likelihood function, with respect to γ and setting it to zero. The estimator for $\hat{\beta}_\gamma$ given in (A.7) follows from solving the score function for β and using this expression in conjunction with the score $g(\gamma, \beta_\gamma)$ enables both parameters γ and β to be solved iteratively. A Newton-Raphson algorithm was programmed in GAUSS to obtain estimates for γ and β_γ.

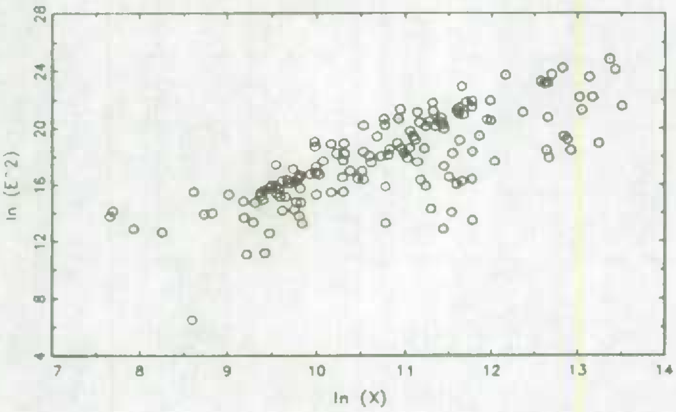
The assumption of normality implicit in the maximum likelihood approach is a drawback since the distribution of local government financial information (e.g., revenues and expenditures) strongly departs from normality. However, the method does yield an alternative estimation methodology which may be used to check and compare the results obtained under other approaches.

A.3 Graphical Approach

In some provinces, due to small sample sizes, the estimation methods discussed above yielded suspicious and unstable estimates for γ when the larger observations are sequentially dropped. This problem was addressed by obtaining some graphical insights into the value for γ. Plots of $\ln(\hat{E}_k^2)$ and $\ln x_k$ (see Figure 3) were examined to ascertain visually the slope of a line through the sample cluster. This rough estimate of the slope should be close to the least squares estimate of γ if a sufficiently large number of points had been available.

Fig 3 Graphical Method to Find Gamma

Plot of ln(E^2) vs. ln(X) (Ontario)



Information about plausible values of γ using this approach was used in addition to the numerical methods discussed above in provinces with small sample sizes and in cases where estimates of γ showed a large degree of instability over reduced observations.

A.4 Application to Local Government Finance Survey Data

The estimates of the heteroscedasticity parameter γ under the least squares (LS) and MLE methods (denoted $\hat{\gamma}_{LS}$ and $\hat{\gamma}_{MLE}$, respectively), after excluding the m largest x-valued observations (effective sample size $n-m$), are reported in Table 2.1. The dependent (survey) variable y was defined as the revenues reported by local government units in the 1989 actual estimates; the independent variable x is the 1991 census count for the municipality. The actual estimates are prepared 30 months after the end of the survey year from municipal financial statements submitted by the local government units.

Table A.1 Least Squares and Maximum Likelihood Estimates of γ

Largest Units Removed (m)	Effective Sample Size (n-m)	$\hat{\gamma}_{LS}$	$\hat{\gamma}_{MLE}$
0	108	1.97	2.05
1	106	1.72	2.07
8	100	1.90	2.14
18	90	1.94	2.10
28	80	2.15	2.14
38	70	2.18	2.07

The graph of $\ln(\hat{E}_k^2)$ vs. $\ln(x_k)$ is exhibited in Figure 3. The slope of this cluster of points is a rough estimate of the value of γ. Based on the insights given by the three methods for possible estimates of γ, the value of the heteroscedasticity parameter was set to $\hat{\gamma}=2$. A similar methodology - which synthesized and checked the information about γ obtained under the three estimation approaches - is used to determine the most plausible estimate for the heteroscedasticity parameter in the other provinces.

A.5 Estimation of Proportionality Constant *c*

For the purpose of estimating the design variance of the regression estimator given in (2.8), modelling of the squared residuals E_k^2 , $k \epsilon U$, is required. The difficulty posed by $E_k^2 \propto x_k^\gamma \eta_k$ (2.6) is that although γ can be estimated from past sample data, the unknown population deviations $\eta_1, \ldots, \eta_N$ are unknown. If, however, the relationship between the squared residual E_k^2 and the size value x_k defined in (2.6) holds well in the population, then the disturbance η_k will have a small influence on x_k^γ and modelling the relation

$$E_k^2 = c\ x_k^\gamma \tag{2.6a}$$

gives a justifiable empirical approximation to (2.6).

After an estimate for γ has been identified, the proportionality constant appearing in (2.6a) is estimated. This value is needed to facilitate the estimation of the design variance $V_p(\hat{t}_R)$ given in (3.3). Expression (2.6a) can also be written as $c = E_k^2 / x_k^\gamma$, thereby suggesting the following estimator for c:

$$\hat{c} = \frac{1}{n} \sum_{k=1}^{n} \hat{E}_k^2 / x_k^{\hat{\gamma}} \tag{A.8}$$

where $\hat{E}_k = y_k - \hat{B} x_k$ is the estimated residual and $\hat{\gamma}$ is the value estimated earlier.

Estimates of c using the estimator (A.8) over different subsets of the Ontario data (y = revenues for 1989, x = 1991 census population counts) excluding the m largest x-valued observations are given in Table 2.2. These estimates over the reduced datasets give some indication as to the sensitivity and stability of the estimation procedure and the behaviour of the data.

Table A.2 Estimates of the Proportionality Constant *c*.

Largest Units Removed (*m*)	Effective Sample Size (*n*-*m*)	$\hat{c}$
0	108	.0825
2	106	.0803
8	100	.0853
18	90	.0817
28	80	.0857
38	70	.0737

The estimates $\hat{c}$ for $m=1,\ldots,38$ are relatively stable. The value of $\hat{c}$ at $m=0$ was chosen for later work.

DATE DUE
