*The Canadian Journal of Statistics*
Vol. 29, No. 2, 2001, Pages 1–yyy
*La revue canadienne de statistique*



# Variance estimation for the finite population distribution function with complete auxiliary information

Changbao WU and Randy R. SITTER

*Key words and phrases:* Design-based; jackknife; model-based; superpopulation model.

*MSC 2000:* Primary 62D05; secondary 62G09.

*Abstract:* The authors develop jackknife and analytical variance estimators for the estimator of Chambers & Dunstan (1986) and Rao, Kovar & Mantel (1990) of the finite population distribution function, using complete auxiliary information. They also describe the associated model and show the design consistency of the variance estimators, whose small-sample performance is examined through a limited simulation study. They highlight the operational advantages of the jackknife in the model-based setting of Chambers & Dunstan (1986) and its better conditional performance in the design-based setting of Rao, Kovar & Mantel (1990).

## Utilisation de l'information auxiliaire complète pour obtenir des estimations de la fonction de répartition d'une population finie

*Résumé :* Les auteurs montrent comment utiliser de l'information auxiliaire complète pour obtenir des estimations de variance analytique et jackknife pour les estimateurs de Chambers & Dunstan (1986) et de Rao, Kovar & Mantel (1990) de la fonction de répartition d'une population finie. Ils décrivent aussi le modèle associé et montrent la cohérence des estimateurs de variance en fonction du plan expérimental. Le comportement à taille finie de ces estimateurs est également étudié par voie de simulation. Ils soulignent les avantages opérationnels du jackknife dans le cadre du modèle de Chambers & Dunstan (1986) et ses meilleures performances conditionnelles du point de vue du plan d'expérimental que privilégient Rao, Kovar & Mantel (1990).

## 1. INTRODUCTION

The use of auxiliary information in estimating the finite population distribution function has attracted increased attention in recent literature. Several estimators which incorporate knowledge of an auxiliary variable known for every unit in the finite population have been proposed and their performances examined and compared. See, for examples, Chambers & Dunstan (1986), Rao, Kovar & Mantel (1990), Chambers, Dorfman & Hall (1992), Kuk (1993) and Wang & Dorfman (1996). For some discussion on situations where auxiliary information is available at the unit level in complex surveys, see Särndal, Swensson & Wretman (1992, Chapter 8). Less attention has been given to the variance estimation problem.

This paper examines variance estimation in this setting for two leading estimators of the finite-population distribution function, the model-based estimator of Chambers & Dunstan (1986) and the design-based estimator of Rao, Kovar & Mantel (1990). For the model-based estimator, analytical variance estimators must be developed for each assumed superpopulation model. In the case of the simple linear regression model, such an estimator is derived, based on an asymptotic result of Chambers, Dorfman & Hall (1992). We do so by using some results in Wang & Dorfman (1996). This variance estimator is less than desirable as it involves kernel density estimators, and must be re-derived for each superpopulation model considered. Rao, Kovar & Mantel (1990) give an analytical variance estimator for the design-based difference estimator. The design consistency of this variance estimator assumes the existence of an expansion in Randles (1982) which is not verifiable for general sampling designs. This point was not considered by Rao, Kovar & Mantel (1990). We establish the existence for some specific and important designs.

One of the goals of this paper is to demonstrate that jackknife variance estimators are an attractive alternative. The jackknife is a commonly used technique to estimate variance which is easy to implement. It is attractive as it merely involves deleting a unit and re-calculating the estimator. In Section 4.1, we investigate the use of the delete-one jackknife for estimating the variance of the model-based estimator and establish consistency results. This method avoids the need for kernel density estimates and remains operationally the same for different superpopulation models. In Section 4.2, we similarly establish the consistency of the jackknife for the design-based difference estimator for some common designs. In the design-based case, the jackknife does not have as great an operational advantage because the analytical variance estimator can quite easily be extended to other models. However, while investigating the small sample performances of these variance estimators through simulation in Section 5, we demonstrate that the jackknife displays better conditional properties in the design-based case. We conclude with a brief discussion in Section 6.

## 2. ESTIMATORS OF THE DISTRIBUTION FUNCTION

Suppose that $y$ is the characteristic of interest and $x$ is the auxiliary variable associated with $y$. The finite population of size $N$ consists of all pairs of $(y_i, x_i)$, $i = 1, \ldots, N$. The finite-population distribution function of $y$ evaluated at $t$ is defined as the proportion of units in the population with $y$ values less than or equal to $t$, viz.

$$F(t) = \frac{1}{N} \sum_{j=1}^{N} I(y_j \leq t),$$

where $I(\,\cdot\,)$ denotes the indicator function. Let $s$ be a sample of $n$ units from the finite population under a general sampling design and let $\bar{s}$ denote the non-sampled units of the finite population. We assume that the auxiliary information $x_i$ is known for all elements in the finite population while $y_i$ is known only for $i \in s$.

The paper of Chambers & Dunstan (1986) motivated much of the later work. In their model-based framework, $x$ and $y$ are assumed to follow a superpopulation model. Though the results can be extended to more complex models, for simplicity of presentation, we will first restrict attention to the simple linear regression model,

$$y_i = \alpha + \beta x_i + \varepsilon_i, \quad i = 1, \ldots, N, \tag{1}$$

where the $\varepsilon_i$'s are independent and identically distributed with $\mathrm{E}\,(\varepsilon_i) = 0$, $\mathrm{var}(\varepsilon_i) = \sigma_\varepsilon^2$ and $\alpha$ and $\beta$ are unknown superpopulation parameters.

Under (1), the model-based estimator

$$\widehat{F}_m(t) = \frac{1}{N}\left[\sum_{i\in s} I_{[y_i \leq t]} + \frac{1}{n}\sum_{j\in\bar{s}}\sum_{i\in s} I\{y_i \leq t - \hat{\beta}(x_j - x_i)\}\right] \tag{2}$$

is asymptotically model-unbiased for $F(t)$, where $\hat{\beta} = \sum_{i\in s}(y_i - \bar{y})(x_i - \bar{x}) / \sum_{i\in s}(x_i - \bar{x})^2$. This is a Royall-type estimator which Chambers & Dunstan (1992) develop by noting that

$$F(t) = N^{-1}\left\{\sum_{i\in s} I(y_i \leq t) + \sum_{i\in\bar{s}} I(y_i \leq t)\right\}$$

and using a model-based predictor of the second term. A crucial point here is that $\widehat{F}_m(t)$ is independent of the sampling design, as $(y_i, x_i)$ for $i = 1, \ldots, N$ are viewed as independent sample values from superpopulation model (1) regardless of whether they belong to the set of sampled units, $s$, or to the set of non-sampled units, $\bar{s}$. The estimator $\widehat{F}_m(t)$ being asymptotically model-unbiased for $F(t)$ is understood to mean $\lim_{n\to\infty} \mathrm{E}\{\widehat{F}_m(t) - F(t)\} = 0$, where $\mathrm{E}$ denotes the model expectation. We should also note that, for asymptotic theory, a sequence of finite populations indexed by $\nu$ is assumed. All limiting processes will be understood to be as $\nu \to \infty$. We also assume $n \to \infty$, $N \to \infty$ as $\nu \to \infty$. However, throughout this paper, we will not make distinctions among $\nu \to \infty$, $n \to \infty$ and $N \to \infty$.

Rao, Kovar & Mantel (1990) proposed a design-based estimator which is asymptotically both design-unbiased under a general sampling design and model-unbiased under a working model such as (1),

$$\widehat{F}_d(t) = \frac{1}{N}\left\{\sum_{i\in s} \pi_i^{-1} I(y_i \le t) + \sum_{j=1}^{N} \widehat{G}_j - \sum_{i\in s} \pi_i^{-1}\widehat{G}_{ic}\right\},$$

where

$$\begin{aligned} \widehat{G}_j &= \sum_{k\in s} \pi_k^{-1} I(\tilde{\varepsilon}_k \le t - \tilde{\alpha} - \tilde{\beta}x_j) \Big/ \sum_{k\in s} \pi_k^{-1}, \\ \widehat{G}_{ic} &= \sum_{k\in s} \frac{\pi_i}{\pi_{ik}} I(\tilde{\varepsilon}_k \le t - \tilde{\alpha} - \tilde{\beta}x_i) \Big/ \sum_{k\in s} \frac{\pi_i}{\pi_{ik}}, \end{aligned} \tag{3}$$

$$\begin{aligned} \tilde{\beta} &= \sum_{i\in s} \pi_i^{-1}(x_i - \tilde{x})(y_i - \tilde{y}) \Big/ \sum_{i\in s} \pi_i^{-1}(x_i - \tilde{x})^2, \\ \tilde{\alpha} &= \tilde{y} - \tilde{\beta}\tilde{x}, \quad \tilde{\varepsilon}_k = y_k - \tilde{\alpha} - \tilde{\beta}x_k, \\ \tilde{x} &= \sum_{i\in s} \pi_i^{-1} x_i \Big/ \sum_{i\in s} \pi_i^{-1}, \quad \tilde{y} = \sum_{i\in s} \pi_i^{-1} y_i \Big/ \sum_{i\in s} \pi_i^{-1} \end{aligned}$$

and $\pi_i$, $\pi_{ij}$ are the first- and second-order inclusion probabilities. Note that the original formulation of $\widehat{F}_d(t)$ given by Rao , Kovar & Mantel (1990) was under a heteroscedastic model which we will mention in Section 4.

The estimator $\widehat{F}_d(t)$ was motivated as a difference estimator (also called a modified Horvitz–Thompson estimator; see Basu 1971) by assuming the

$$G_j = \frac{1}{N}\sum_{i=1}^{N} I(\varepsilon_i \le t - \alpha - \beta x_j)$$

are known and then replacing them by their estimates in (3). It is design-unbiased and usually has smaller variance than the conventional Horvitz–Thompson estimator. Godambe (1989) derived $\widehat{F}_d(t)$ based on the model- and design-based optimum estimating function theory and showed that $\widehat{F}_d(t)$ is robust against departures from the superpopulation model.

The model-based $\widehat{F}_m(t)$ is model-unbiased but design-inconsistent. Rao, Kovar & Mantel (1990) argue through simulation that the model-based $\widehat{F}_m(t)$ has superior performance in small samples when the superpopulation model is correctly specified but is more vulnerable than $\widehat{F}_d(t)$ to model-misspecification and in such cases can perform poorly in large samples. Chambers, Dorfman & Hall (1992) do a theoretical comparison and conclude that there is no clear winner. One must take care judging whether any model-based or design-based estimator is "better", since they are developed under completely different frameworks. Whether one chooses to work under a model-based framework and use $\widehat{F}_m(t)$ or a design-based framework and use $\widehat{F}_d(t)$, variance estimation will need to be considered.

## 3. ANALYTICAL VARIANCE ESTIMATION

In this section, we propose an analytical variance estimator for the model-based estimator $\widehat{F}_m(t)$ by combining the theoretical results of Chambers, Dorfman & Hall (1992) and some results from Wang & Dorfman (1996). We also introduce the analytical variance estimator for the design-based estimator $\widehat{F}_d(t)$ proposed by Rao, Kovar & Mantel (1990). The establishment of design-consistency of this variance estimator relies on a condition which was not considered in Rao, Kovar and Mantel's original derivation. This condition is given in Section 4.2 and has been verified in Appendix 2.

### *3.1. Model-based estimator.*

Six years after Chambers and Dunstan's original paper, Chambers, Dorfman & Hall (1992) derived the asymptotic model variance of $\widehat{F}_m(t)$ under (1). To summarize their results, assume that $f = n/N \to \pi \in [0,1)$ as $n \to \infty$, and assume the sampled and non-sampled $x_i$s have a common asymptotic density $h(x)$. Let $G$ and $g$ be the error distribution function and the error density function, respectively, let $\sigma_\varepsilon^2$ be the error variance, and

$$\begin{aligned}
\mu_x &= \int x h(x)\,dx, \quad \sigma_x^2 = \int x^2 h(x)\,dx - \mu_x^2,\\
I_1 &= \int (x-\mu_x) g(t-\alpha-\beta x) h(x)\,dx,\\
I_2 &= \int\int G\{(t-\alpha-\beta x)\wedge(t-\alpha-\beta y)\} h(x)h(y)\,dx\,dy,\\
I_3 &= \int G(t-\alpha-\beta x) h(x)\,dx,\\
I_4 &= \int \{G(t-\alpha-\beta x) - G^2(t-\alpha-\beta x)\} h(x)\,dx,
\end{aligned}$$

where $a \wedge b$ denotes the minimum of $a$ and $b$. Then, under model (1),

$$\mathrm{var}\{\widehat{F}_m(t) - F(t)\} = \frac{1}{n}\left\{(1-\pi)^2\left(\frac{\sigma_\varepsilon^2}{\sigma_x^2} I_1^2 + I_2 - I_3^2\right) + \pi(1-\pi) I_4\right\} + o\left(\frac{1}{n}\right). \quad (4)$$

In their construction of a different point estimator of $F(t)$, Wang & Dorfman (1996) suggest estimating $I_1$, $I_2$, $I_3$ and $I_4$ by

$$\begin{aligned}
\hat{I}_1 &= (N-n)^{-1}\sum_{j\in\bar{s}} (x_j - \bar{x})\hat{g}(t-\hat{\alpha}-\hat{\beta}x_j),\\
\hat{I}_2 &= (N-n)^{-2}\sum_{i\in\bar{s}}\sum_{j\in\bar{s}} \widehat{G}_n\big\{(t-\hat{\alpha}-\hat{\beta}x_i)\wedge(t-\hat{\alpha}-\hat{\beta}x_j)\big\},\\
\hat{I}_3 &= (N-n)^{-1}\sum_{j\in\bar{s}} \widehat{G}_n(t-\hat{\alpha}-\hat{\beta}x_j),\\
\hat{I}_4 &= (N-n)^{-1}\sum_{j\in\bar{s}} \widehat{G}_n(t-\hat{\alpha}-\hat{\beta}x_j)\big\{1-\widehat{G}_n(t-\hat{\alpha}-\hat{\beta}x_j)\big\},
\end{aligned} \quad (5)$$

where

$$\widehat{G}_n(u) = n^{-1}\sum_{i\in s} I(\hat{\varepsilon}_i \le u), \quad \hat{\varepsilon}_i = y_i - \hat{\alpha} - \hat{\beta}x_i,$$

and

$$\hat{g}(u) = (nd)^{-1}\sum_{i\in s} K\{(\hat{\varepsilon}_i - u)/d\}$$

is a standard kernel density estimator with bandwidth $d$ and kernel $K(\,\cdot\,)$. Letting $\hat\sigma_x^2 = s_x^2$, $\hat\sigma_\varepsilon^2 = (n-2)^{-1}\sum_{i\in s}\hat\varepsilon_i^2$, we can then estimate $\mathrm{var}\{\widehat F_m(t) - F(t)\}$ by estimating all the unknown components in (4) term-by-term.

We suggest that it is computationally more stable to write $I_{23} = I_2 - I_3^2$ as

$$\iint G\big\{(t-\alpha-\beta x)\wedge(t-\alpha-\beta y)\big\}\big[1 - G\big\{(t-\alpha-\beta x)\vee(t-\alpha-\beta y)\big\}\big]h(x)h(y)\,dx\,dy,$$

where $a \vee b$ denotes the maximum of $a$ and $b$. We can rewrite $\mathrm{var}\{\widehat F_m(t) - F(t)\}$ by replacing $I_2 - I_3^2$ in (4) by $I_{23}$. We therefore propose to estimate $\mathrm{var}\{\widehat F_m(t) - F(t)\}$ by

$$v_m = v\{\widehat F_m(t) - F(t)\} = \frac{1}{n}\left\{(1-f)^2\left(\frac{\hat\sigma_\varepsilon^2}{\hat\sigma_x^2}\hat I_1^2 + \hat I_{23}\right) + f(1-f)\hat I_4\right\},$$

where $\hat I_{23}$ equals

$$(N-n)^{-2}\sum_{i\in\bar s}\sum_{j\in\bar s}\widehat G_n\big\{(t-\hat\alpha-\hat\beta x_i)\wedge(t-\hat\alpha-\hat\beta x_j)\big\}\big[1-\widehat G_n\big\{(t-\hat\alpha-\hat\beta x_i)\vee(t-\hat\alpha-\hat\beta x_j)\big\}\big].$$

It is not difficult to see that $v_m$ is more robust and it is guaranteed that $v_m \ge 0$.

There are still difficulties with this variance estimator $v_m$. One difficulty is that the kernel density estimator $\hat g(\,\cdot\,)$ is computer intensive and involves choosing a kernel $K(\,\cdot\,)$ and a bandwidth $d$ which are always difficult choices to make. In our simulation study, we tried two choices of kernel density estimator: a standard normal kernel with $d = 1.059\hat\sigma_\varepsilon n^{-1/5}$, recommended by Simonoff (1996); and a standard logistic kernel with $d = R/n$, recommended by Kuk (1993), where $R$ is the range of residuals. We chose $R = 4\hat\sigma_\varepsilon$ for each simulation.

Another greater drawback which will be shared by other analytical variance estimators is the fact that the derivation of the analytical variance formula depends heavily on the model. One may, at first glance at the form of (4), intuit that deriving such for a different model would merely amount to a redefinition of the residuals. This is not the case, as can be seen by viewing the development for a linear model with heteroscedastic errors in Chambers & Dunstan (1986). The variance estimator must be completely re-derived for every new model considered.

*3.2. Design-based estimator.*

Rao, Kovar & Mantel (1990) propose estimating the design-variance of $\widehat F_d(t)$ by

$$v_d = v\{\widehat F_d(t)\} = \frac{1}{N^2}\sum_{i\in s}\sum_{j>i}\frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}\left\{\frac{u_i(j)}{\pi_i} - \frac{v_j(i)}{\pi_j}\right\}^2, \tag{6}$$

where $u_i(j) = I(y_i \le t) - \widehat G_{ic}(j)$, $v_j(i) = I(y_j \le t) - \widehat G_{jc}(i)$ and

$$\widehat G_{ic}(j) = \left[\sum_{k\in s}\frac{\pi_{ij}}{\pi_{ijk}} I\left\{y_k \le t - \tilde\beta(x_i - x_k)\right\}\right] \Big/ \left(\sum_{k\in s}\frac{\pi_{ij}}{\pi_{ijk}}\right),$$

which is a conditionally design-unbiased estimator for $G_i$ given the $i$th and the $j$th units are selected in the sample.

Rao, Kovar & Mantel (1990) argue that $v_d$ is design-consistent by referring to a result from Randles (1982). Though this claim may be true for many commonly used designs, it relies on the existence of an asymptotic expansion in Randles's paper which may or may not exist for a particular design and is difficult to verify generally. This point is discussed further in Section 4.2 since similar problems must be overcome to establish the consistency of the delete-one jackknife. In the process, we verify the design-consistency of $v_d$ for some common situations.

As noted by Rao, Kovar & Mantel (1990), $v_d$ "involves the computation of third order inclusion probabilities, $\pi_{ijk}$, which could be cumbersome under unequal probability sampling". However, it is often the case that the sampling fraction (in the first stage for multi-stage sampling) is small. If so, it is common practice to assume with-replacement sampling (at the first stage for multi-stage sampling) for the purpose of variance estimation. This usually yields a conservative variance estimate and also simplifies (6) by eliminating the need for second- and third-order inclusion probabilities. This is further discussed in Section 4.2.

## 4. JACKKNIFE VARIANCE ESTIMATION

### *4.1. Using the jackknife to estimate the variance of* $\widehat{F}_m(t) - F(t)$.

Let us consider $\mathrm{var}\{\widehat{F}_m(t) - F(t)\}$ under model (1). First, notice that we cannot ignore the variability induced by estimation of $\alpha$ and $\beta$. To see this, denote $\widehat{F}_m(t)$ by $\widehat{F}_m(t;\hat\alpha,\hat\beta)$ and compare

$$\mathrm{var}\{\widehat{F}_m(t;\alpha,\beta) - F(t)\} = \frac{1}{n}\big\{(1-\pi)^2 I_{23} + \pi(1-\pi)I_4\big\} + o\left(\frac{1}{n}\right) \tag{7}$$

to $\mathrm{var}\{\widehat{F}_m(t;\hat\alpha,\hat\beta) - F(t)\}$ given in (4). We see that, under (1), the asymptotic model variance of $\widehat{F}_m(t;\hat\alpha,\hat\beta) - F(t)$ is the same as that of $\widehat{F}_m(t;\alpha,\beta) - F(t)$ if and only if

$$I_1 = \int (x-\mu_x) g(t-\alpha-\beta x) h(x)\, dx = 0. \tag{8}$$

For a particular population quantile $t$ of interest, (8) is usually not true and the impact of replacing $\alpha$, $\beta$ by $\hat\alpha$, $\hat\beta$ on asymptotic variance depends on the value of $|I_1|$ and the ratio of $\sigma_\varepsilon^2/\sigma_x^2$. This suggests that any jackknife variance estimator will have to recalculate $\hat\alpha$ and $\hat\beta$ with each deleted unit. We mention this because, as we will see in Section 4.2, this is not the case for the design-based estimator. It is also worth noting that this remains true even for $\pi = 0$. It is interesting that Chambers, Dorfman & Hall (1992) identified the size of $I_1$ as an indicator of cases where $\widehat{F}_d(t)$ outperformed $\widehat{F}_m(t)$ even when the model was correct.

Next, note that

$$\widehat{F}_m(t) - F(t) = \frac{1}{nN}\sum_{j\in\bar s}\sum_{i\in s} I\{y_i \le t - \hat\beta(x_j - x_i)\} - \frac{1}{N}\sum_{j\in\bar s} I(y_j \le t), \tag{9}$$

and thus $\mathrm{var}\{\widehat{F}_m(t) - F(t)\} = V_1 + V_2$, where

$$\begin{aligned} V_1 &= \mathrm{var}\left[\frac{1}{n}\sum_{i\in s}\frac{1}{N}\sum_{j\in\bar s} I\{y_i \le t - \hat\beta(x_j - x_i)\}\right] \quad \text{and} \\ V_2 &= \mathrm{var}\left\{\frac{1}{N}\sum_{j\in\bar s} I(y_j \le t)\right\} = \frac{f(1-f)}{n(N-n)}\sum_{j\in\bar s} G(t-\alpha-\beta x_j)\{1 - G(t-\alpha-\beta x_j)\}. \end{aligned}$$

The jackknife cannot be applied directly to $\widehat{F}_m(t) - F(t)$ since it involves the unobserved values of the $y$ variable. Typically, the jackknife is applicable in surveys when $f = n/N$ is small, which is often the case. By ignoring $f$, we may induce a positive bias which will hopefully be small. We will first consider this case, by assuming $f \to \pi = 0$. We will then consider the case where $\pi > 0$.

There are two things to observe if $f \to 0$. First, $V_2 = o(1/n)$, and second, $\mathrm{var}\{\widehat{F}_m(t)\} = V_1 + o(1/n)$. This last follows from viewing $\widehat{F}_m(t)$ as the sum of two terms in (2) and noting that the variance of the first term is $V_1$,

$$\mathrm{var}\left\{N^{-1}\sum_{i\in s} I(y_i \le t)\right\} = f^2 \cdot \mathrm{var}\left\{n^{-1}\sum_{i\in s} I(y_i \le t)\right\} = o\left(\frac{1}{n}\right)$$

and that the covariance term can also be shown to be $o(1/n)$.

Thus, the leading term in both $\mathrm{var}\{\widehat{F}_m(t) - F(t)\}$ and $\mathrm{var}\{\widehat{F}_m(t)\}$ is $V_1$. We are now ready to state conditions needed for establishing the consistency of the jackknife variance estimator when $f \to 0$, Theorem 1 below. Recall that $G(\cdot)$ and $g(\cdot)$ are the error cumulative distribution function and density function under model (1). Let $G_n(\cdot)$ be the corresponding empirical distribution function based on a sample of size $n$.

A1. $N^{-1}\sum_{j=1}^N x_j^2 = O(1)$.

A2. For a random sample $s$ of size $n$, $\max_{i\in s}|x_i| = O_p(n^q)$ for some $q \in (0, 1/4)$.

A3. $G(\cdot)$ has bounded first and second derivative on $\Theta$, where $\Theta$ is the closure of all possible values of $t - \alpha - \beta x_j$, $j = 1, \ldots, N$ in the limiting process.

A4. For fixed $t$, $N^{-1}\sum_{j=1}^N g(t-\alpha-\beta x_j) = O(1)$ and $N^{-1}\sum_{j=1}^N x_j g(t-\alpha-\beta x_j) = O(1)$.

Condition A1 is not restrictive. Condition A2 is needed to furnish our proofs and a sufficient condition for this is $\mathrm{E}(X^{1/q}) = O(1)$, where $X$ is the auxiliary variable. For example, for $q = 1/5$, finite fifth moment is enough.

THEOREM 1. (i) *Under conditions A1–A4, model* (1), *and assuming* $f \to 0$,

$$v_{Jm1} = \frac{(n-1)}{n}\sum_{i=1}^n (F_i^* - \overline{F}^*)^2$$

*is a consistent estimator of* $\mathrm{var}\{\widehat{F}_m(t) - F(t)\}$, *where*

$$F_i^* = \frac{n}{N}\frac{1}{n-1}\sum_{k\in s_i} I(y_k \le t) + \frac{1}{n-1}\sum_{k\in s_i}\left[\frac{1}{N}\sum_{j\in\bar{s}} I\{y_k \le t - \hat{\beta}_i(x_j - x_k)\}\right],$$

$\hat{\beta}_i$ *is calculated based on* $s_i$, *the sample data with the* $i$*th observation excluded, and* $\overline{F}^* = \sum_{i=1}^n F_i^*/n$.

(ii) *With the same conditions as in* (i), $(v_{Jm1})^{-1/2}\{\widehat{F}_m(t) - F(t)\}$ *converges to* $\mathrm{N}(0,1)$ *in distribution.*

*Proof of Theorem 1.* See Appendix 1.

Thus, in the case where $f$ is negligible, one can use the usual delete-one jackknife variance estimator.

Now let us consider the case where $f \to \pi \in (0, 1)$. In this case, we cannot merely apply the jackknife, as the last term in (9) involves unobserved $y$'s and its variance, $V_2$, is not negligible. On the other hand, we would prefer to avoid the problems with the analytical variance estimators of Section 3, which arise due to the first term of (9) and estimation of its variance $V_1$. It turns out that we can combine the jackknife and analytical approaches to get a variance estimator which is easy to implement (much easier than $v_m$) and consistent. We summarize this in Theorem 2.

THEOREM 2. (i) *Under conditions A1–A4 and model* (1), *let* $v_2 = f(1-f)\hat{I}_4/n$ *and*

$$v_{J1} = \frac{n-1}{n}\sum_{i=1}^n (F_i^{**} - \overline{F}^{**})^2,$$

*where*

$$F_i^{**} = \frac{1}{n-1}\sum_{k\in s_i}\left[\frac{1}{N}\sum_{j\in\bar{s}} I\{y_k \le t - \hat{\beta}_i(x_j - x_k)\}\right],$$

$\hat{I}_4$ *is given in* (5), $\hat{\beta}_i$ *is calculated from* $s_i$ *and* $\overline{F}^{**} = \sum_{i=1}^{n} F_i^{**}/n$. *Then* $v_{Jm2} = v_{J1} + v_2$ *is a consistent estimator of* $\mathrm{var}\{\widehat{F}_m(t;\hat{\alpha},\hat{\beta}) - F(t)\}$.

(ii) *With the same conditions as in* (i), $(v_{Jm2})^{-1/2}\{\widehat{F}_m(t) - F(t)\}$ *converges to* $\mathrm{N}(0,1)$ *in distribution.*

*Proof of Theorem 2.* The result follows easily from the fact that $\mathrm{var}\{\widehat{F}_m(t) - F(t)\} = V_1 + V_2$, $v_2 \to V_2$ and $v_{J1} \to V_1$ by the proof of Theorem 1.

Thus, we jackknife term 1 of (9) and analytically estimate the variance of term 2 with a substitution estimator which does not require kernel density estimation.

Generalizations of the simple linear regression model in (1) to other models are important in practice. For example, in the original papers of both Rao, Kovar & Mantel (1990) and Chambers & Dunstan (1986), the following heteroscedastic linear model was assumed:

$$y_i = \beta x_i + v(x_i)\varepsilon_i, \quad i = 1, \ldots, N, \tag{10}$$

where $v(x)$ is a strictly positive function of $x$ only and the $\varepsilon_i$'s are independent and identically distributed. Under model (10), we need modified versions of A1–A4 to validate the proposed variance estimation strategy. If we redefine $\varepsilon_j = (y_j - \beta x_j)/v(x_j)$, $\hat{\varepsilon}_j = (y_j - \hat{\beta} x_j)/v(x_j)$ and

$$\hat{\beta} = \Big\{\sum_{i\in s} x_i y_i / v^2(x_i)\Big\} \Big/ \Big\{\sum_{i \in s} x_i^2/v^2(x_i)\Big\},$$

then we can replace A1–A4 by A1′–A4′ obtained by replacing $x_i$ by $x_i/v(x_i)$ in A1 and A2, replacing $x_j$ by $x_j/v(x_j)$ for $x_j$ not in the argument of $g$ in A4, and replacing $t - \alpha - \beta x_j$ by $(t - \beta_j)/v(x_j)$ in A3 and A4.

Noting that

$$V_2 = N^{-2} \sum_{j \in \bar{s}} G\{(t - \beta x_j)/v(x_j)\}\big[1 - G\{(t - \beta x_j)/v(x_j)\}\big]$$

can be estimated by

$$v_2 = N^{-2} \sum_{j \in \bar{s}} \widehat{G}_n\{(t - \hat{\beta} x_j)/v(x_j)\}\big[1 - \widehat{G}_n\{(t - \hat{\beta} x_j)/v(x_j)\}\big],$$

Theorems 1 and 2 can then both be extended to model (10) if we assume A1′–A4′. The method extends as easily to more complex models.

### *4.2. Jackknife variance estimation for* $\widehat{F}_d(t)$.

In this section, we will consider the delete-one jackknife variance estimator for $\widehat{F}_d(t)$. By Randles (1982), if $X_1, \ldots, X_n$ is a random sample and if $T_n(\hat{\lambda}) = T_n(X_1, \ldots, X_n; \hat{\lambda})$, where $\hat{\lambda}$ is a consistent estimator for parameter $\lambda$, then assuming that the following expansion holds,

$$T_n(\hat{\lambda}) - \mu(\lambda) = T_n(\lambda) - \mu(\lambda) + (\hat{\lambda} - \lambda)' \left\{ \frac{\partial}{\partial \nu}\mu(\nu) \Big|_{\nu=\lambda} \right\} + o_p\left(\frac{1}{\sqrt{n}}\right), \tag{11}$$

$T_n(\lambda)$ and $T_n(\hat{\lambda})$ have the same limiting distribution provided $\partial\mu(\nu)/\partial\nu|_{\nu=\lambda} = 0$, where $\mu(\nu) = \lim_{n\to\infty} \mathrm{E}_{\lambda}\{T_n(\nu)\}$ and the expectation is taken when the true parameter is $\lambda$.

Denote $\widehat{F}_d(t)$ by $\widehat{F}_d(t;\tilde{\alpha},\tilde{\beta})$. Let $\widehat{F}_d(t;\alpha,\beta)$ be the same estimator but with $\alpha$, $\beta$ not estimated from the sample. Rao, Kovar & Mantel (1990) show that $\partial\mu(\nu_1,\nu_2)/\partial\nu_i|_{\nu=\lambda} = 0$ for $i = 1, 2$, where $\mu(\nu_1,\nu_2) = \lim_{n\to\infty} \mathrm{E}_{\alpha,\beta}\{\widehat{F}_d(t;\nu_1,\nu_2)\}$, $\widehat{F}_d(t;\nu_1,\nu_2)$ denotes $\widehat{F}_d(t;\alpha,\beta)$ but replacing $\lambda = (\alpha, \beta)$ by mathematical symbol $\nu = (\nu_1, \nu_2)$. Using the above argument, they conclude that

$$\frac{\mathrm{var}\{\widehat{F}_d(t; \tilde{\alpha}, \tilde{\beta})\}}{\mathrm{var}\{\widehat{F}_d(t; \alpha, \beta)\}} \to 1, \tag{12}$$

as $n \to \infty$, by quoting Randles (1982). They did not mention the crucial condition, that expansion (11) exists. Although (11) holds for many statistics $T_n(\nu)$ which are not smooth functions of $\nu$, it is mathematically unverifiable for $\widehat{F}_d(t; \nu_1, \nu_2)$ for general sampling designs.

We will look at a couple of important sampling designs. First consider unequal probability sampling with replacement. In this case, it is possible to verify (11) and (12) under assumptions similar to those made by Shao & Rao (1993) when they consider estimators of the low income proportion in the context of stratified multi-stage sampling. In our case, we must derive the result uniformly over $t_j = t - \alpha - \beta x_j$. This is done in Appendix 2. In practice the units are sampled without replacement, but if $f$ is small, assuming with-replacement sampling for the purpose of variance estimation is a common practice. This simplifies the calculations and is likely to create only a small positive bias in the resulting variance estimates. The same arguments hold for stratified multi-stage sampling with first-stage clusters sampled with replacement.

Once (11) and (12) have been established in this setting, consistency of the jackknife variance estimator follows quite easily. To see this, let

$$F_d^*(t) = \frac{1}{N}\left\{\sum_{i \in s} w_i I(y_i \le t) + \sum_{j=1}^{N} G_j - \sum_{i \in s} w_i G_i\right\},$$

where $w_i = 1/\pi_i$,

$$G_i = \frac{1}{N}\sum_{j=1}^{N} I(\varepsilon_j \le t - \alpha - \beta x_i)$$

is a population characteristic, and $\varepsilon_j = y_j - \alpha - \beta x_j$. It is straightforward to show that $F_d^*(t) - \widehat{F}_d(t; \alpha, \beta) = o_p(n^{-1/2})$, i.e., $\widehat{F}_d(t; \alpha, \beta)$ and $F_d^*(t)$ have the same asymptotic design variance. One can then take note that

$$\mathrm{var}\{F_d^*(t)\} = \mathrm{var}\left[\frac{1}{N}\sum_{i \in s} w_i\{I(y_i \le t) - G_i\}\right] \tag{13}$$

is the design variance of a weighted average. Assuming certain regularity conditions, the conventional delete-one jackknife variance estimator, denoted by $v_{Jd}^*(G_i)$, will thus be a consistent estimator of $\mathrm{var}\{F_d^*(t)\}$ (Shao & Tu, 1995, p. 261, Theorem 6.1). The jackknife variance estimator, $v_{Jd}^*(G_i)$, can further be approximated by replacing $G_i$ by

$$\widehat{G}_i = \sum_{k \in s} w_k I(\tilde{\varepsilon}_k \le t - \tilde{\alpha} - \tilde{\beta} x_i) \Big/ \sum_{k \in s} w_k.$$

More formally,

THEOREM 3. *Let* $s_i = s - \{i\}$ *and*

$$F_{di}^{(1)} = N^{-1}\left\{\sum_{k \in s} w_k^{(i)} I(y_k \le t) + \sum_{j=1}^{N} \widehat{G}_j - \sum_{k \in s} w_k^{(i)} \widehat{G}_k\right\}$$

*with* $w_k^{(i)} = n w_k/(n-1)$ *for* $k \in s_i$ *and* $w_i^{(i)} = 0$. *For single stage sampling satisfying* $\max_{i \in s} n w_i/N = O(1)$,

$$v_{Jd1} = \frac{n-1}{n}\sum_{i \in s}\left\{F_{di}^{(1)} - \overline{F}_d^{(1)}\right\}^2 \tag{14}$$

*is a design-consistent estimator of* $\mathrm{var}\{\widehat{F}_d(t)\}$*, where*

$$\overline{F}_d^{(1)} = \frac{1}{n}\sum_{i\in s} F_{di}^{(1)} = N^{-1}\Big\{\sum_{i\in s} w_i I(y_i \le t) + \sum_{j=1}^{N}\widehat{G}_j - \sum_{i\in s} w_i \widehat{G}_i\Big\}.$$

One could redefine

$$F_{di}^{(1)} = N^{-1}\sum_{k\in s} w_k^{(i)}\{I(y_k \le t) - \widehat{G}_k\} \quad \text{and} \quad \overline{F}_d^{(1)} = N^{-1}\sum_{i\in s} w_i\{I(y_i \le t) - \widehat{G}_i\}$$

in Theorem 3. We chose to write it more like a jackknife might be applied. That is, to merely delete a unit and recalculate the estimator. In Theorem 3, we do not quite do this as $\widehat{G}_k$ is not recalculated with each unit deleted. We are able to do this because of (12). For small or moderate sample size $n$, it will be of interest to consider a modified version of the jackknife variance estimator which recalculates $\tilde{\alpha}$, $\tilde{\beta}$ and $\widehat{G}_k$ for each deletion. That is, defining $v_{Jd2}$ as in (14) but using

$$F_{di}^{(2)} = \frac{1}{N}\Big\{\sum_{k\in s} w_k^{(i)} I(y_k \le t) + \sum_{j=1}^{N}\widehat{G}_{ji} - \sum_{k\in s} w_k^{(i)}\widehat{G}_{ki}\Big\},$$

where

$$\widehat{G}_{ji} = \Big\{\sum_{k\in s} w_k^{(i)} I(\hat{\varepsilon}_k \le t - \tilde{\alpha}_i - \tilde{\beta}_i x_j)\Big\}\Big/\Big\{\sum_{k\in s} w_k^{(i)}\Big\},$$

with $\tilde{\alpha}_i$ and $\tilde{\beta}_i$ calculated from the sample data using the jackknife weights, $w_k^{(i)}$ for $k = 1,\ldots,n$. This is a true delete-one jackknife as it completely recalculates the estimator with each unit deleted. If the sampling is in fact without replacement, one may choose to multiply (14) by $1-f$. This is not strictly correct for unequal probability sampling but may be used to attempt to reduce the positive bias induced by using a variance estimator based on with-replacement sampling (see Durbin 1953, and Särndal, Swensson & Wretman 1992, p. 99). This was done in the simulations of Section 5.

If the design is stratified multistage sampling, the formulation of $v_{Jd1}$ and the required regularity conditions need to be modified, (see Shao & Tu 1995, Chapter 6; Wu (1999). However, the basic ideas behind the verification of (11) and (12) remain similar provided the sampling of first-stage clusters is assumed to be with-replacement for the purposes of variance estimation. The usual jackknife variance estimator, which deletes each cluster one at a time is then design consistent as the number of strata gets large (Wu 1999).

The formulation of $\widehat{F}_d(t)$ can be easily extended to other superpopulation models and the corresponding jackknife variance estimator is still design consistent.

A simplified version of the analytical variance estimator can also be obtained from (13). In the variance formula (6) proposed by Rao, Kovar & Mantel (1990), we could simply replace $u_i(j)$ by $u_i = I(y_i \le t) - \widehat{G}_i$, with no third order inclusion probabilities involved (the difference between $u_i$ and $u_i(j)$ is that $u_i$ is a simple substitution estimator while $u_i(j)$ is a conditionally design-unbiased estimator for $I(y_i \le t) - G_i$ given the $i$th and $j$th units are selected in the sample). Further, if the sampling fraction $f$ is small, we could adapt the sampling-with-replacement variance estimator which involves $\pi_i$s only. Thus, in view of the theoretical developments in this section, we can say that the only real advantage, if any, of the jackknife variance estimator in the design-based case is operational convenience. We will demonstrate through simulation, however, that the jackknife does seem to have better conditional properties than the analytical variance estimator. This has been noted in other contexts (Rao & Sitter 1995; Sitter 1997).

## 5. A SIMULATION

In this section, we present the results of a limited simulation study on the small sample performance of variance estimators proposed in Sections 3 and 4. The finite populations used in the simulation were generated from the simple linear regression model (1) with $\alpha = \beta = 1$. The covariates $x_i$ were generated as an independent and identically distributed sample from a lognormal$(2.0, 0.25)$ and $\varepsilon_i$'s were independent and identically distributed from a $\mathrm{N}(0, 0.6)$. The population size was chosen as $N = 2000$.

For the model-based estimators, a new finite population was created for each simulation and then a simple random sample of size $n = 50$ was drawn from the population. The variance estimators, $v_m$, $v_{Jm1}$, $v_{Jm2}$ and $v_{J1}$ were computed from each sample. This process was repeated $B = 1000$ times. We then reconducted the above simulation with $n = 200$. The same procedure was repeated for the design-based estimators except that only one finite population was generated and used for all simulations. The variance estimators $v_d$, $v_{Jd1}$ and $v_{Jd2}$ were computed from each simulated sample from this finite population. The sample sizes $n = 50$ and $n = 200$ used here were trying to mimic two situations: sampling fraction negligible ($f = 50/2000 = 0.025$) and sampling fraction non-negligible ($f = 200/2000 = 0.1$).

The performance of variance estimators was measured and compared in terms of relative percentage bias ($\mathrm{RB\%}$) and instability ($\mathrm{INST}$). The simulated values of $\mathrm{RB\%}$ and $\mathrm{INST}$ for a particular variance estimator $v$ were computed as

$$\mathrm{RB\%}(v) = 100 \times \frac{\bar{v} - \mathrm{MSE}}{\mathrm{MSE}} \quad \text{and} \quad \mathrm{INST}(v) = \frac{s_v}{\mathrm{MSE}},$$

where $\bar{v} = B^{-1} \sum_{b=1}^{B} v_b$, $s_v^2 = B^{-1} \sum_{b=1}^{B} (v_b - \mathrm{MSE})^2$, $\mathrm{MSE} = B^{-1} \sum_{b=1}^{B} \{\widehat{F}_b(t) - F(t)\}^2$ is the estimated mean square error of $\widehat{F}(t)$ from another independent $B$ simulations, and $\widehat{F}_b(t)$ and $v_b$ are the values of $\widehat{F}(t)$ and $v$ from the $b$th simulation, respectively. $\mathrm{RB\%}$ and $\mathrm{INST}$ were computed for $t = \xi_p$ at $p = 0.10$, $p = 0.25$, $p = 0.50$, $p = 0.75$ and $p = 0.90$, where $\xi_p$ is the $p$th population quantile.

Table 1 reports the values of $\mathrm{RB\%}$ and $\mathrm{INST}$ of variance estimators $v_m$, $v_{Jm1}$, $v_{Jm2}$ and $v_{J1}$ for the model-based estimator $\widehat{F}_m(t)$. We observe that: (a) For $n = 50$ ($f = 0.025$): (i) the jackknife variance estimators $v_{Jm1}$ and $v_{Jm2}$ perform well; (ii) $v_{J1}$, the leading term in both $v_{Jm1}$ and $v_{Jm2}$, also provides valid estimated variance, but has negative bias in all cases; (iii) the analytical variance estimator $v_m$ has the smallest value of $\mathrm{INST}$ in all cases, but it has the largest negative bias among $v_m$, $v_{Jm1}$ and $v_{Jm2}$; (b) For $n = 200$ ($f = 0.1$): (i) $v_m$, $v_{Jm1}$ and $v_{Jm2}$ perform quantitatively similar in terms of both $\mathrm{RB\%}$ and $\mathrm{INST}$; (ii) they all have larger and positive bias for $F(t)$ close to $0.50$ and smaller or negative bias when $F(t)$ is close to $0$ or $1$; (iii) by ignoring the sampling fraction, $v_{J1}$ seriously underestimates the true variance; (iv) it is interesting to notice the good performance of $v_{Jm1}$ for $n = 200$, since it is not clear whether $v_{Jm1}$ is consistent or not when $f$ is not negligible; and (v) we should note that $v_m$ took in the order of 30 times longer to calculate than the jackknife variance estimators because of the density estimation.

TABLE 1: Relative percentage bias and instability of variance estimators for the model-based estimator $\widehat{F}_m(t)$ at $t = \xi_p$.

| | | $p=0.10$ | $p=0.25$ | $p=0.50$ | $p=0.75$ | $p=0.90$ |
|---|---|---|---|---|---|---|
| | | $N=2000$, $n=50$ | | | | |
| $v_m$ | RB% | −9.47 | −6.56 | −1.73 | −3.01 | −5.83 |
| | INST | 0.254 | 0.208 | 0.210 | 0.251 | 0.307 |
| $v_{Jm1}$ | RB% | −1.72 | 1.05 | 1.33 | 3.24 | 0.99 |
| | INST | 0.346 | 0.293 | 0.243 | 0.382 | 0.499 |
| $v_{Jm2}$ | RB% | −0.25 | 2.21 | 3.95 | 6.48 | 4.51 |
| | INST | 0.338 | 0.287 | 0.244 | 0.380 | 0.493 |
| $v_{J1}$ | RB% | −6.61 | −4.04 | −4.14 | −2.39 | −4.09 |
| | INST | 0.338 | 0.285 | 0.236 | 0.368 | 0.487 |
| | | $N=2000$, $n=200$ | | | | |
| $v_m$ | RB% | −5.94 | 5.27 | 11.19 | 4.06 | 0.57 |
| | INST | 0.145 | 0.126 | 0.160 | 0.127 | 0.152 |
| $v_{Jm1}$ | RB% | −5.48 | 7.84 | 8.91 | 2.04 | −1.71 |
| | INST | 0.186 | 0.171 | 0.151 | 0.163 | 0.226 |
| $v_{Jm2}$ | RB% | −3.12 | 8.12 | 12.55 | 6.48 | 0.030 |
| | INST | 0.166 | 0.163 | 0.173 | 0.167 | 0.210 |
| $v_{J1}$ | RB% | −26.34 | −16.86 | −18.84 | −25.46 | −28.14 |
| | INST | 0.302 | 0.213 | 0.214 | 0.290 | 0.341 |

Table 2 reports the results for the design-based estimator $\widehat{F}_d(t)$: (a) For simple random sampling, the jackknife variance estimator $v_{Jd1}$ is identical to the analytical variance estimator $v_d$; (b) (i) $v_{Jd2}$ is more stable than $v_d$ in all cases; (ii) for $n = 50$, $v_{Jd2}$ has larger positive bias than $v_d$ at $F(t) = 0.10$ and $0.90$, but this difference disappears when $n = 200$.

Turning to conditional properties, we chose two values of $p$, $0.1$ and $0.5$, ordered the simulated samples on the values of $\bar{x}$ and then grouped them into twenty successive groups. We ran 10,000 simulations to reduce Monte Carlo error so that each group is of size $500$. For each group, the conditional mean of each variance estimator $v$ was calculated as

$$\mathrm{E}_c(v) = \frac{1}{500}\sum_{b=1}^{500} v_b.$$

Independently, we generated $100{,}000$ simulated samples and similarly grouped them on $\bar{x}$ into $20$ groups and calculated the simulated conditional MSE of each of $\widehat{F}_m(t)$ and $\widehat{F}_d(t)$ as

$$\mathrm{MSE}_c\{\widehat{F}(t)\} = \frac{1}{5000}\sum_{b=1}^{5000}\{\widehat{F}_b(t) - F(t)\}^2,$$

in obvious notation.

TABLE 2: Relative percentage bias and instability of variance estimators for the design-based estimator $\widehat{F}_d(t)$ at $t = \xi_p$.

| | | $p = 0.10$ | $p = 0.25$ | $p = 0.50$ | $p = 0.75$ | $p = 0.90$ |
|---|---|---|---|---|---|---|
| | | | | $N = 2000$, $n = 50$ | | |
| $v_d$ | RB% | 7.87 | 1.58 | −3.74 | −3.70 | 3.87 |
| | INST | 0.484 | 0.3134 | 0.274 | 0.369 | 0.581 |
| $v_{Jd1}$ | RB% | 7.87 | 1.58 | −3.74 | −3.70 | 3.87 |
| | INST | 0.484 | 0.313 | 0.274 | 0.369 | 0.581 |
| $v_{Jd2}$ | RB% | 14.84 | 6.24 | −0.24 | 0.88 | 13.45 |
| | INST | 0.452 | 0.301 | 0.247 | 0.343 | 0.558 |
| | | | | $N = 2000$, $n = 200$ | | |
| $v_d$ | RB% | −2.89 | −5.03 | 1.14 | 3.57 | 5.36 |
| | INST | 0.201 | 0.146 | 0.134 | 0.192 | 0.282 |
| $v_{Jd1}$ | RB% | −2.89 | −5.03 | 1.14 | 3.57 | 5.36 |
| | INST | 0.201 | 0.146 | 0.134 | 0.192 | 0.282 |
| $v_{Jd2}$ | RB% | −0.98 | −4.00 | 1.87 | 4.50 | 7.56 |
| | INST | 0.176 | 0.129 | 0.120 | 0.176 | 0.258 |

For the design-based case, the values of $\mathrm{E}_c(v)$ ($\times 10^4$) for $v = v_d (= v_{Jd1})$ and $v_{Jd2}$ and $MSE_c\{\widehat{F}_d(t)\}$ ($\times 10^4$) were plotted against the group averages of $\bar{x}$. Figure 1 gives this plot for $n = 50$, and $t = \xi_p$ for $p = 0.1$ and $0.5$. As we can see, all of the variance estimators for $\widehat{F}_d(t)$ perform well in tracking the conditional MSE for the case $p = 0.5$, but $v_{Jd2}$, the full jackknife, significantly outperforms the other variance estimators in this respect for $p = 0.1$. In other simulations which are not presented here, it can be seen that this property is a function of $p$ and becomes more pronounced as one moves away from $p = 0.5$ in either direction. This suggests that despite the asymptotic (and unconditional) arguments of Section 4 and the previously presented unconditional simulations, which suggest these variance estimators are essentially equivalent, there is an advantage to completely recalculating $\widehat{F}_d(t)$ for each deleted unit when applying the jackknife. This also suggests that the jackknife when so applied has better conditional properties than the analytical variance estimator of Section 3.

For the model-based case, similar plots were examined but are not presented as all of the variance estimators for $\widehat{F}_m(t)$ performed well in tracking the conditional MSE for all values of $p$.

## 6. CONCLUDING REMARKS

Based on the theoretical development and our limited simulation study, we suggest that, for the model-based estimator $\widehat{F}_m(t)$, the true delete-one jackknife variance estimator $v_{Jm1}$ is recommended if $f$ is small; in cases where $f$ is not negligible, it is safe to use $v_{Jm2}$, although simulation results suggest that $v_{Jm1}$ can also be used in these cases. $v_{J1}$ was included in the simulation to serve the purpose of illustrating the asymptotic results only, it should not be used in practice even though it is consistent when $f$ is small. It should be emphasized that our solution (Theorem 2, $v_{Jm2}$) for this latter case is a hybrid of the jackknife and the analytical approach. Compared to the pure analytical approach, where the derivation of analytical variance for more complex models is very difficult (if not impossible) and the estimation of the variance requires a kernel smoother, the analytical component of this hybrid estimator involves a simple empirical distribution estimator of the residual distribution. Thus, this hybrid estimator is simple, stable and easily applicable, and its extension to more complex models is obvious and easy.

FIGURE 1: Plot of conditional performance of variance estimators: Conditional means $\mathrm{E}_c(v_d)$, $\mathrm{E}_c(v_{Jd2})$ ($\times 10^4$) and conditioned mean squared error, $\mathrm{MSE}_c\{\widehat{F}_d(t)\}$ ($\times 10^4$) for (a) $t = \xi_{0.1}$ and (b) $t = \xi_{0.5}$, versus group average of $\bar{x}$.

For the design-based case, in terms of conditional performance, $v_{Jd2}$, the true delete-one jackknife variance estimator, performs the best. However, it tends to have larger positive unconditional bias when $F(t)$ is close to $0$ or $1$ and $n$ is not large. Jackknife variance estimator is usually less stable. In our settings here, both the analytical and the jackknife variance estimators are approximately unbiased, the jackknife may be preferred due to its operational simplicity.

## APPENDIX 1: PROOF OF THEOREM 1

The following lemma is used in our proof.

LEMMA. *Let $a_n = c_0 n^{-q}$, $q \in (0, 1/2)$ be a constant. If $G$ has bounded first derivative over $\Theta$, then*

$$\sup_{|x| \le a_n} |\{G_n(u+x) - G_n(u)\} - \{G(u+x) - G(u)\}| \le R_n,$$

*where $R_n = o(n^{-1/2})$, independent of $u \in \Theta$.*

*Proof.* The proof follows along the lines of Bahadur (1966) (see also Serfling 1980, pp. 97–99). Using his notation, let $b_n = n^{\frac{1}{2}(1-q)}$ so that $b_n^2 = n^{1-q}$ and $a_n b_n^{-1} = o(n^{-1/2})$. Next, letting $\gamma_n = c_1 n^{-\frac{1}{2}(1+q)} (\log n)^{1/2} = o(n^{-1/2})$, we can show that

$$\delta_n = \frac{n\gamma_n^2}{2(c_2 a_n + \gamma_n)} > 2 \log n$$

for sufficiently large $n$. The fact that $|G'(u)| = |g(u)| \le M$, $u \in \Theta$ for some constant $M$ indicates that the choices of $c_1$ and $c_2$ can be independent of $u$, which also implies the uniformity of $R_n$ over $\Theta$.

*Proof of Theorem 1.* (i) We need only to show that $V_1 = \mathrm{var}\{\widehat{H}(t)\}$ can be consistently estimated by a jackknife estimator, where

$$\widehat{H}(t) = \frac{1}{nN} \sum_{i \in s} \sum_{j \in \bar{s}} I\{y_i \le t - \hat{\beta}(x_j - x_i)\}.$$

To do this, let $A(u; t, \alpha, \beta) = \int_{t-\alpha-\beta x \ge u} h(x)\, dx$, $B(t, \alpha, \beta) = \int g(t - \alpha - \beta x) h(x)\, dx$ and $C(t, \alpha, \beta) = \int x g(t - \alpha - \beta x) h(x)\, dx$ and note that

$$\begin{aligned} \frac{1}{N-n} \sum_{j \in \bar{s}} g(t - \alpha - \beta x_j) &= B(t, \alpha, \beta) + o(1), \\ \frac{1}{N-n} \sum_{j \in \bar{s}} x_j g(t - \alpha - \beta x_j) &= C(t, \alpha, \beta) + o(1), \qquad (15) \\ \frac{1}{N-n} \sum_{j \in \bar{s}} I(u \le t - \alpha - \beta x_j) &= A(u; t, \alpha, \beta) + R_{N,n}(u; t, \alpha, \beta), \end{aligned}$$

where $R_{N,n}(u; t, \alpha, \beta) = o(1)$. Under model (1), $\varepsilon_i = y_i - \alpha - \beta x_i$ and $G_n(x) = n^{-1} \sum_{i \in s} I(\varepsilon_i \le x)$. For a given sample, $s$, let $a_n = |\hat{\beta} - \beta| \max_{i \in s} |x_i|$ and observe that

$$G_n(x + \hat{\alpha} - \alpha - a_n) \le \frac{1}{n} \sum_{i \in s} I(\hat{\varepsilon}_i \le x) \le G_n(x + \hat{\alpha} - \alpha + a_n).$$

This implies that $n^{-1}\sum_{i\in s} I(\hat{\varepsilon}_i \le x) = G_n(x+d_n)$, for some $d_n = (\hat{\alpha}-\alpha)+(\hat{\beta}-\beta)c_n$ and $|c_n| \le \max_{i\in s}|x_i|$. Now, let

$$\widetilde{H}(t) = \frac{1}{nN}\sum_{i\in s}\sum_{j\in\bar{s}} I\{y_i \le t-\beta(x_j-x_i)\} = \frac{1}{N}\sum_{j\in\bar{s}} G_n(t_j), \tag{16}$$

where $t_j = t-\alpha-\beta x_j$. Noting that $\mathrm{var}\{R_{N,n}(\varepsilon_i;t,\alpha,\beta)\} = o(1)$, we get

$$\begin{aligned} \widetilde{H}(t) &= \left(1-\frac{n}{N}\right)\frac{1}{n}\sum_{i\in s}\frac{1}{N-n}\sum_{j\in\bar{s}} I(\varepsilon_i \le t-\alpha-\beta x_j) \\ &= \left(1-\frac{n}{N}\right)\frac{1}{n}\sum_{i\in s}\{A(\varepsilon_i;t,\alpha,\beta)+R_{N,n}(\varepsilon_i;t,\alpha,\beta)\} \\ &= \left(1-\frac{n}{N}\right)\frac{1}{n}\sum_{i\in s} A(\varepsilon_i;t,\alpha,\beta) + o_p\left(\frac{1}{\sqrt{n}}\right). \end{aligned} \tag{17}$$

Also note that

$$\begin{aligned} \widehat{H}(t) &= \frac{1}{nN}\sum_{i\in s}\sum_{j\in\bar{s}} I\{y_i \le t-\hat{\beta}(x_j-x_i)\} \\ &= \frac{1}{N}\sum_{j\in\bar{s}} G_n(t-\hat{\alpha}-\hat{\beta}x_j+d_n) \\ &= \frac{1}{N}\sum_{j\in\bar{s}} G_n\{t_j+(\hat{\beta}-\beta)(c_n-x_j)\}, \end{aligned} \tag{18}$$

where the second equality holds since $I\{y_i \le t-\hat{\beta}(x_j-x_i)\} = I(\hat{\varepsilon}_i \le t-\hat{\alpha}-\hat{\beta}x_j)$.

By Condition A1 which implies $\hat{\beta}-\beta = O_p(n^{-1/2})$, Conditions A1 and A2 which together imply $(\hat{\beta}-\beta)(c_n-x_j) = o_p(n^{-1/4})$, and the lemma,

$$\left|G_n\{t_j+(\hat{\beta}-\beta)(c_n-x_j)\} - G_n(t_j) - G\{t_j+(\hat{\beta}-\beta)(c_n-x_j)\} + G(t_j)\right| \le R_n,$$

where $R_n = o_p(n^{-1/2})$. Since $G$ has bounded second derivative and $N^{-1}\sum_1^N x_j^2 = O(1)$, by applying (16), (18) and a Taylor expansion of $G\{t_j+(\hat{\beta}-\beta)(c_n-x_j)\}$ at $t_j$ to the second order, we get

$$\begin{aligned} \widehat{H}(t) &= \{\widehat{H}(t)-\widetilde{H}(t)\} + \widetilde{H}(t) \\ &= \frac{1}{N}\sum_{j\in\bar{s}}\big[G_n\{t_j+(\hat{\beta}-\beta)(c_n-x_j)\} \\ &\qquad - G_n(t_j) - G\{t_j+(\hat{\beta}-\beta)(c_n-x_j)\} + G(t_j)\big] \\ &\quad + \frac{1}{N}\sum_{j\in\bar{s}}\big[G\{t_j+(\hat{\beta}-\beta)(c_n-x_j)\} - G(t_j)\big] + \widetilde{H}(t) \\ &= (\hat{\beta}-\beta)\left\{c_n\frac{1}{N}\sum_{j\in\bar{s}} g(t_j) - \frac{1}{N}\sum_{j\in\bar{s}} x_j g(t_j)\right\} + \widetilde{H}(t) + o_p\left(\frac{1}{\sqrt{n}}\right). \end{aligned}$$

We know from (4) and (7) that $\mathrm{var}\{\widehat{H}(t)\} = O(1/n)$ and $\mathrm{var}\{\widetilde{H}(t)\} = O(1/n)$. From the fact that $\mathrm{var}(\hat{\beta}-\beta) = O(1/n)$, $\mathrm{var}\{o_p(n^{-1/2})\} = o(1/n)$, and for any $X, Y$, $\mathrm{cov}(X,Y) \le \{\mathrm{var}(X)\,\mathrm{var}(Y)\}^{1/2}$, we conclude that $c_n = O(1)$ or $o(1)$. This together with (15) and (17) implies the asymptotic variance of $\widehat{H}(t)$ is the same as that of

$$\left(1-\frac{n}{N}\right)\left[(\hat{\beta}-\beta)\{c_n B(t,\alpha,\beta) - C(t,\alpha,\beta)\} + \frac{1}{n}\sum_{i\in s} A(\varepsilon_i;t,\alpha,\beta)\right],$$

which is a smooth function of sample means of properly defined population characteristics. The conventional delete-one jackknife variance estimator is therefore consistent and construction of this estimator is straightforward.

(ii) The asymptotic normality of $\widehat{F}_m(t)$ follows from the fact that $\widehat{F}_m(t)$ is asymptotically equivalent to a smooth function of sample means.

## APPENDIX 2: ILLUSTRATION OF (11) AND (12) UNDER SAMPLING WITH REPLACEMENT

We first consider unequal probability sampling with replacement with inclusion probabilities $\pi_i$. Let $\varepsilon_i = y_i - \alpha - \beta x_i$ for $i = 1, \ldots, N$, $G_N(u) = N^{-1} \sum_{j=1}^N I(\varepsilon_j \le u)$ and

$$\widehat{G}_n(u) = \sum_{i \in s} \pi_i^{-1} I(\varepsilon_i \le u) \Big/ \sum_{i \in s} \pi_i^{-1}.$$

We assume

C1. $\max_{i \in s} n(N\pi_i)^{-1} = O(1)$.

C2. There exists $g_N(x) \ge 0$, for $\delta_N = O(N^{-q})$, $q \in (0, 1/2)$, such that

$$\{G_N(x + \delta_N) - G_N(x)\}/\delta_N - g_N(x) = r_{Nx}, \quad x \in \Theta,$$

where $|r_{Nx}| \le R_N$ and $R_N = o(1)$, independent of $x$, and $\Theta$ is the closure of all possible values of $t - \alpha - \beta x_j$ for $j = 1, \ldots, N$ in the limiting process.

C3. $|g_N(x)| \le M$, for any $x \in \Theta$, where $M$ is a constant. C4. $\tilde{\alpha} - \alpha = O_p(n^{-1/2})$, $\tilde{\beta} - \beta = O_p(n^{-1/2})$, $N^{-1} \sum_{j=1}^N x_j^2 = O(1)$ and $\max_{i \in s} |x_i| = O_p(n^{1/2-q})$ for some $q \in (0, 1/2)$.

Conditions C2 and C3 used here are stronger than (C2) and (C3) in Shao & Rao (1993) in order to achieve certain uniformity (C2 and C3 together is equivalent to A3 used in Theorem 1); C4 holds for most common situations.

The following development is similar to the proof of Theorem 1. Note that C1 also implies $N^{-1} \sum_{i \in s} \pi_i^{-1} - 1 = O_p(n^{-1/2})$. Following the lines of Shao & Rao (1993, p. 400), we can show that, under conditions C1–C3 for $\hat{\theta} - \theta = O_p(n^{-q})$, $q \in (0, 1/2)$, $\theta \in \Theta$,

$$\widehat{G}_n(\hat{\theta}) = \widehat{G}_n(\theta) - G_N(\theta) + G_N(\hat{\theta}) + u_n(\theta), \quad \theta \in \Theta, \tag{19}$$

where $|u_n(\theta)| \le u_N$, $u_N = o_p(n^{-1/2})$, independent of $\theta$. Recall that $\varepsilon_i = y_i - \alpha - \beta x_i$, $\tilde{\varepsilon}_i = y_i - \tilde{\alpha} - \tilde{\beta} x_i$ and

$$\sum_{i \in s} \pi_i^{-1} I(\varepsilon_i \le t + \tilde{\alpha} - \alpha - a_n) \le \sum_{i \in s} \pi_i^{-1} I(\tilde{\varepsilon}_i \le t) \le \sum_{i \in s} \pi_i^{-1} I(\varepsilon_i \le t + \tilde{\alpha} - \alpha + a_n),$$

where $a_n = |\tilde{\beta} - \beta| \max_{i \in s} |x_i|$. Hence,

$$\sum_{i \in s} \pi_i^{-1} I(\tilde{\varepsilon}_i \le t) \Big/ \sum_{i \in s} \pi_i^{-1} = \widehat{G}_n(t + d_n),$$

where $d_n = (\tilde{\alpha} - \alpha) + (\tilde{\beta} - \beta) c_n$ and $|c_n| \le \max_{i \in s} |x_i|$. Let $\hat{\theta}_j = t - \tilde{\alpha} - \tilde{\beta} x_j$ and $\theta_j = t - \alpha - \beta x_j$. It is easy to see that $\widehat{G}_j = \widehat{G}_n(\hat{\theta}_j + d_n)$ in the formulation of $\widehat{F}_d(t; \tilde{\alpha}, \tilde{\beta})$. Also note that $\widehat{G}_{ic} = \widehat{G}_i$ under sampling with replacement. Therefore,

$$\widehat{F}_d(t; \alpha, \beta) = \frac{1}{N} \left\{ \sum_{i \in s} \pi_i^{-1} I(y_i \le t) + \sum_{j=1}^N \widehat{G}_n(\theta_j) - \sum_{i \in s} \pi_i^{-1} \widehat{G}_n(\theta_i) \right\}.$$

With condition C4, $\hat{\theta}_j - \theta_j = -(\tilde{\alpha} - \alpha) - (\tilde{\beta} - \beta)x_j = O_p(n^{-1/2})$ and $d_n = O_p(n^{-q})$. Assuming $\theta_j + d_n \in \Theta$, from (19),

$$\widehat{G}_n(\hat{\theta}_j + d_n) = \widehat{G}_n(\theta_j + d_n) - G_N(\theta_j + d_n) + G_N(\hat{\theta}_j + d_n) + u_n(\theta_j + d_n)$$

and

$$\widehat{G}_n(\theta_j + d_n) = \widehat{G}_n(\theta_j) - G_N(\theta_j) + G_N(\theta_j + d_n) + u_n(\theta_j).$$

Now it follows that

$$\begin{aligned}
\widehat{F}_d(t; \tilde{\alpha}, \tilde{\beta}) &= \frac{1}{N}\Bigg[\sum_{i \in s} \pi_i^{-1} I(y_i \le t) \\
&\quad + \sum_{j=1}^{N} \big\{\widehat{G}_n(\theta_j) - G_N(\theta_j) + G_N(\hat{\theta}_j + d_n) + u_n(\theta_j + d_n) + u_n(\theta_j)\big\} \\
&\quad - \sum_{i \in s} \pi_i^{-1} \big\{\widehat{G}_n(\theta_i) - G_N(\theta_i) + G_N(\hat{\theta}_i + d_n) + u_n(\theta_i) + u_n(\theta_i + d_n)\big\}\Bigg] \\
&= \widehat{F}_d(t; \alpha, \beta) + \frac{1}{N}\sum_{j=1}^{N}\{G_N(\hat{\theta}_j + d_n) - G_N(\theta_j)\} \\
&\quad - \frac{1}{N}\sum_{i \in s} \pi_i^{-1}\{G_N(\hat{\theta}_i + d_n) - G_N(\theta_i)\} + o_p\left(\frac{1}{\sqrt{n}}\right).
\end{aligned}$$

Under condition C4, $\hat{\theta}_j - \theta_j + d_n = (\tilde{\beta} - \beta)(c_n - x_j) = O_p(n^{-q})$. From condition C2,

$$G_N(\hat{\theta}_j + d_n) - G_N(\theta_j) = (\tilde{\beta} - \beta)(c_n - x_j)\{g_N(\theta_j) + r_{N\theta_j}\},$$

where $|r_{N\theta_j}| < O_p(n^{-q})$, uniformly over $\theta_j$. Therefore, we have the following expansion,

$$\widehat{F}_d(t; \tilde{\alpha}, \tilde{\beta}) = \widehat{F}_d(t; \alpha, \beta) + (\tilde{\beta} - \beta)\mu_N(\alpha, \beta) + o_p\left(\frac{1}{\sqrt{n}}\right), \tag{20}$$

where

$$\mu_N(\alpha, \beta) = \frac{1}{N}\sum_{j=1}^{N}(c_n - x_j)g_N(\theta_j) - \frac{1}{N}\sum_{i \in s}\pi_i^{-1}(c_n - x_i)g_N(\theta_i).$$

It is interesting to notice that $\tilde{\alpha}$ does not appear in the right-hand side of (20) explicitly. In fact, $\widehat{F}_d(t; \tilde{\alpha}, \tilde{\beta})$ contains no $\tilde{\alpha}$ at all. If all the finite population moments involved in $\mu_N(\alpha, \beta)$ are bounded (similar to those assumed in A4), we immediately conclude that $\mu_N(\alpha, \beta) = o_p(1)$, that is, $\widehat{F}_d(t; \tilde{\alpha}, \tilde{\beta}) = \widehat{F}_d(t; \alpha, \beta) + o_p(n^{-1/2})$, which implies (12).

For stratified multi-stage sampling with first-stage clusters sampled with replacement, the above development is similar, except that condition C1 needs to be re-formulated (cf. Shao & Rao 1993, for the exact formulation) and the triple index $(hij)$ should be used in place of $i$ throughout.

## ACKNOWLEDGEMENTS

The authors thank Dr. Jiahua Chen for helpful comments and suggestions, as well as the referees and Associate Editor.

## REFERENCES


R. R. Bahadur (1966). A note on quantiles in large samples. *The Annals of Mathematical Statistics*, 37, 577–580.

D. Basu (1971). *Foundations of Statistical Inference, A Symposium.* (V. P. Godambe and D. A. Sprott, eds.), Holt, Rinehart and Winston, Toronto, Ontario, Canada.

R. L. Chambers, A. H. Dorfman & P. Hall (1992). Properties of estimators of the finite distribution function. *Biometrika*, 79, 577–582.

R. L. Chambers & R. Dunstan (1986). Estimating distribution function from survey data. *Biometrika*, 73, 597–604.

J. Durbin (1953). Some results in sampling theory when the units are selected with unequal probabilities. *Journal of the Royal Statistical Society Series B*, 15, 262–269.

V. P. Godambe (1989). Estimation of cumulative distribution of a survey population. Technical Report No. STAT-89-17. Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada.

A. Y. C. Kuk (1993). A kernel method for estimating finite population distribution functions using auxiliary information. *Biometrika*, 80, 385–392.

R. H. Randles (1982). On the asymptotic normality of statistics with estimated parameters. *The Annals of Statistics*, 10, 462–474.

J. N. K. Rao (1966). Alternative estimators in pps sampling for multiple characteristics. *Sankhyā, Series A*, 28, 47–60.

J. N. K. Rao, J. G. Kovar & H. J. Mantel (1990). On estimating distribution functions and quantiles from survey data using auxiliary information. *Biometrika*, 77, 365–375.

J. N. K. Rao & R. R. Sitter (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika*, 82, 453–460.

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

R. J. Serfling (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York.

J. Shao & J. N. K. Rao (1993). Standard errors for low income proportions estimated from stratified multi-stage samples. *Sankhyā, Series B*, 55, 393–414.

J. Shao & D. Tu (1995). *The Jackknife and Bootstrap*. Springer-Verlag, New York.

J. S. Simonoff (1996). *Smoothing Methods in Statistics*. Springer-Verlag, New York.

R. R. Sitter (1997). Variance estimation for the regression estimator in two-phase sampling. *Journal of the American Statistical Association*, 92, 780–787.

S. Wang & A. H. Dorfman (1996). A new estimator for the finite population distribution function. *Biometrika*, 83, 639–652.

C. Wu (1999). Complete title of Thesis. Ph. D. Dissertation, Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, BC.


---


*Received 19 October 1999*
*Accepted 26 September 2000*



Changbao WU: cbwu@math.uwaterloo.ca
*Department of Statistics and Actuarial Science*
*University of Waterloo, Waterloo*
*Ontario, Canada N2L 3G1*

Randy R. SITTER: sitter@stat.sfu.ca
*Department of Statistics and Actuarial Science*
*Simon Fraser University, Burnaby*
*British Columbia, Canada V5A 1S6*