

VARIANCE ESTIMATION FOR THE FINITE POPULATION DISTRIBUTION FUNCTION WITH COMPLETE AUXILIARY INFORMATION

C. Wu and R.R. Sitter¹

ABSTRACT

Jackknife variance estimators for the model-based estimator (Chambers & Dunstan, 1986) and the design-based estimator (Rao, Kovar & Mantel, 1990) of the finite population distribution function, using complete auxiliary information, have been implemented and their model and design consistency established, respectively. Operational advantages of the jackknife in the model-based setting and better conditional performance of the jackknife in the design-based setting are highlighted.

KEY WORDS: Design-based; Jackknife; Model-based; Superpopulation model.

RÉSUMÉ

Les estimateurs analytiques et de type Jackknife de la variance pour l'estimateur basé sur le modèle (Chambers et Dunstan, 1986) et l'estimateur basé sur le plan de sondage (Rao, Kovar et Mantel, 1990) de fonction de distribution d'une population finie, en utilisant l'information auxiliaire complète, ont été mis en application et leur convergence selon le modèle et le plan de sondage ont été respectivement établies. Les avantages opérationnels du Jackknife dans le cadre de travaux basé sur le modèle et une meilleure performance conditionnelle du Jackknife dans le cadre de travaux basé sur le design sont mis en évidence.

MOTS CLÉS : Basé sur le plan de sondage; Jackknife; basé sur le modèle; modèle de super population.

1. INTRODUCTION

This paper examines variance estimation for two leading estimators of the finite population distribution function using complete auxiliary information, 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 are difficult to derive and must be developed one-at-a-time for each assumed superpopulation model. In the case of the simple linear regression model, such an estimator can be derived based on an asymptotic result of Chambers, Dorfman & Hall (1992) and some results in Wang & Dorfman (1996). See Wu (1999) for detailed discussions of this estimator. 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) gives an

analytical variance estimator for the design-based difference estimator.

We demonstrate that jackknife variance estimators are an attractive alternative. In Section 3.1, we investigate the use of the delete-1 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 3.2 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 studies, it has been shown that the jackknife

¹ Changbao Wu and Randy R. Sitter, *Department of Mathematics and Statistics, Simon Fraser University Burnaby, BC, Canada V5A 1S6* cwua, sitter}@cs.sfu.ca

displays better conditional properties in the design-based case (Wu, 1999). We conclude with a brief discussion in Section 4. All proofs can be found in Wu (1999).

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) , $i=1, \dots, N$. 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 , $F(t) = N^{-1} \sum_{j=1}^N I_{[y_j \leq t]}$, 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. We will restrict attention to the simple linear regression model,

$$y_i = \alpha + \beta x_i + \varepsilon_i, \quad i=1, \dots, N, \quad (2.1)$$

where ε_i 's are independent and identically distributed with $E(\varepsilon_i) = 0$, $Var(\varepsilon_i) = \sigma_\varepsilon^2$ and α and β are unknown superpopulation parameters.

Under (2.1), the model-based estimator of $F(t)$,

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

is asymptotically model-unbiased, where $\hat{\beta} = \sum_{i \in s} (y_i - \bar{y})(x_i - \bar{x}) / \sum_{i \in s} (x_i - \bar{x})^2$.

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 (2.1),

$$\hat{F}_d(t) = \frac{1}{N} \left\{ \sum_{i \in s} \pi_i^{-1} I_{[y_i \leq t]} + \sum_{j=1}^N \hat{G}_j - \sum_{i \in s} \pi_i^{-1} \hat{G}_{ic} \right\},$$

where $\hat{G}_j = \sum_{k \in s} \pi_k^{-1} I_{[\tilde{\varepsilon}_k \leq t - \tilde{\alpha} - \tilde{\beta}x_j]} / \sum_{k \in s} \pi_k^{-1}$,

$$\hat{G}_{ic} = \sum_{k \in s} \frac{\pi_i}{\pi_{ik}} I_{[\tilde{\varepsilon}_k \leq t - \tilde{\alpha} - \tilde{\beta}x_i]} / \sum_{k \in s} \frac{\pi_i}{\pi_{ik}},$$

$$\tilde{\beta} = \sum_{i \in s} \pi_i^{-1} (x_i - \bar{x})(y_i - \bar{y}) / \sum_{i \in s} \pi_i^{-1} (x_i - \bar{x})^2,$$

$$\tilde{\alpha} = \bar{y} - \tilde{\beta} \bar{x}, \quad \tilde{\varepsilon} = y_k - \tilde{\alpha} - \tilde{\beta} x_k,$$

$\tilde{x} = \sum_{i \in s} \pi_i^{-1} x_i / \sum_{i \in s} \pi_i^{-1}$, $\tilde{y} = \sum_{i \in s} \pi_i^{-1} y_i / \sum_{i \in s} \pi_i^{-1}$, and π_i , π_{ij} are the first- and second-order inclusion probabilities.

The estimator $\hat{F}_d(t)$ was motivated as a difference estimator. It is design-unbiased and usually has smaller variance than the conventional Horvitz-Thompson estimator. The model-based $\hat{F}_m(t)$ is model-unbiased but design-inconsistent. Rao, Kovar & Mantel (1990) demonstrates through simulation that the model-based $\hat{F}_m(t)$ has superior performance in small samples when the superpopulation model is correctly specified but is much more vulnerable than $\hat{F}_d(t)$ to model-misspecification and can perform poorly in large samples. Chambers, Dorfman & Hall (1992) do a theoretical comparison under simple random sampling and conclude that there is no clear winner. Whether one chooses to work under a model-based framework and use $\hat{F}_m(t)$ or a design-based framework and use $\hat{F}_d(t)$, variance estimation will need to be considered.

3. JACKKNIFE VARIANCE ESTIMATION

3.1 Using the Jackknife to Estimate the Variance of $\hat{F}_m(t) - F(t)$

Let us consider $var[\hat{F}_m(t) - F(t)]$ under model (2.1). First, note that we cannot ignore the variability induced by estimation of α and β . Also note that

$$\hat{F}_m(t) - F(t) = \frac{1}{nN} \sum_{j \in \bar{s}} \sum_{i \in s} I_{[y_i \leq t - \hat{\beta}(x_j - x_i)]} - \frac{1}{N} \sum_{j \in \bar{s}} I_{[y_j \leq t]},$$

and thus $var[\hat{F}_m(t) - F(t)] = V_1 + V_2$, where

$$V_2 = var \left\{ \frac{1}{N} \sum_{j \in \bar{s}} I_{[y_j \leq t]} \right\} = \frac{f(1-f)}{n(N-n)} \sum_{j \in \bar{s}} G(t - \alpha - \beta x_j) [1 - G(t - \alpha - \beta x_j)]$$

and

$$V_1 = var \left\{ \frac{1}{n} \sum_{i \in s} \left[\frac{1}{N} \sum_{j \in \bar{s}} I_{[y_i \leq t - \hat{\beta}(x_j - x_i)]} \right] \right\}.$$

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 $f \rightarrow \pi = 0$. We will then consider the case where $\pi > 0$.

There are two things to observe if $f \rightarrow 0$. First, that $V_2 = o(1/n)$ and second that

$$\hat{F}_m(t) = (nN)^{-1} \sum_{j \in \bar{s}} \sum_{i \in s} I_{[y_i \leq t - \hat{\beta}(x_j - x_i)]} + o_p(n^{-1/2}).$$

Thus, the leading term in both $\text{var}[\hat{F}_m(t) - F(t)]$ and $\text{var}[\hat{F}_m(t)]$ is V_1 . We are now ready to establish the consistency of the jackknife variance estimator when $f \rightarrow 0$, Theorem 1 below.

Theorem 1. (i) Under certain regularity conditions and model (2.1), and assuming $f \rightarrow 0$,

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

is a consistent estimator of $\text{var}[\hat{F}_m(t) - F(t)]$, where

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

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

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

Thus, in the case where f is negligible, one can use the usual delete-1 jackknife variance estimator. The regularity conditions used for the theorem were described in Wu (1999).

Now, let us consider the case where $f \rightarrow \pi \in (0,1)$. In this case, we cannot merely apply the jackknife, as the last term in (4.1) 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, which arise due to the first term of (4.1) 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 and consistent. We summarize this in Theorem 2.

Theorem 2. (i) Under certain regularity conditions and model (2.1), let $v_2 = f(1-f)\hat{I}_4/n$ and, where

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

\hat{I}_4 is a simple substitution estimator given in Wu (1999), $\hat{\beta}_i$ is calculated from s_i and $\bar{F}^{**} = \sum_{i=1}^n F_i^{**} / n$. Then $v_{Jm2} = v_{J1} + v_2$ is a consistent estimator of $\text{var}[\hat{F}_m(t; \hat{\alpha}, \hat{\beta}) - F(t)]$. (ii) With the same conditions as in (i), $(v_{Jm2})^{-1/2} [\hat{F}_m(t) - F(t)]$ converges to $N(0,1)$ in distribution.

Thus, we jackknife term 1 of (4.1) 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 (2.1) to other more complex models can be made easily under revised regularity conditions.

3.2 Jackknife Variance Estimation for $\hat{F}_d(t)$

One crucial result for establishing the consistency of jackknife variance estimator for $\hat{F}_d(t)$ is to show that the estimated model parameters do not change the asymptotical design variance. That is,

$$\text{var}[\hat{F}_d(t; \tilde{\alpha}, \tilde{\beta})] / \text{var}[\hat{F}_d(t; \alpha, \beta)] \rightarrow 1, \quad \text{as } n \rightarrow \infty.$$

Rao, Kovar & Mantel (1990) conclude (4.2) by quoting a result from Randles (1982) which depends on an asymptotical expansion that is unverifiable here for a general sampling design. Wu (1999) has justified (4.2) for some commonly used designs.

Once (4.2) have been established in this setting, consistency of the jackknife variance estimator follows quite easily. To see this, note that the asymptotic design variance of $\hat{F}_d(t; \alpha, \beta)$ is the same as the asymptotic design variance of

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

where $w_i = 1/\pi_i$, $G_i = N^{-1} \sum_{j=1}^N I_{[\varepsilon_j \leq t - \alpha - \beta x_i]}$ is a population characteristic and $\varepsilon_j = y_j - \alpha - \beta x_j$. It follows that $\text{var}[F_d^*(t)] = \text{var}[N^{-1} \sum_{i \in s} w_i (I_{[y_i \leq t]} - G_i)]$ is the design variance of a weighted average. Assuming certain regularity conditions, the conventional delete-1 jackknife variance estimator, denoted by $v_{Jd}^*(G_i)$, will thus be a consistent estimator of $\text{var}[F_d^*(t)]$ (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 $\hat{G}_i = \sum_{k \in s} w_k I_{[\bar{y}_k \leq t - \bar{\alpha} - \bar{\beta} x_i]} / \sum_{k \in s} w_k$.

More formally,

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

$$F_{di}^{(1)} = N^{-1} \left[\sum_{k \in s_i} w_k^{(i)} I_{[y_k \leq t]} + \sum_{j=1}^N \hat{G}_j - \sum_{k \in s_i} w_k^{(i)} \hat{G}_k \right]$$

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

$$v_{Jd1} = \frac{n-1}{n} \sum_{i \in s} (F_{di}^{(1)} - \bar{F}_d^{(1)})^2$$

is a design-consistent estimator of $\text{var}[\hat{F}_d(t)]$, where

$$\bar{F}_d^{(1)} = \frac{1}{n} \sum_{i \in S} F_{di}^{(1)} = \frac{1}{N} \left[\sum_{i \in S} w_i I_{[y_i \leq t]} + \sum_{j=1}^N \hat{G}_j - \sum_{i \in S} w_i \hat{G}_i \right]$$

One could also use the true delete-1 jackknife variance estimator, v_{Jd2} . That is, to merely delete a unit and recalculate the estimator. In Theorem 3, we do not quite do this as \hat{G}_k is not recalculated with each unit deleted. We are able to do this because of (4.2). If the sampling is in fact without replacement, one may choose to multiply (4.3) by $1-f$.

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

4. CONCLUDING REMARKS

Based on the theoretical development and our limited simulation study, we suggest that, for the model-based estimator $\hat{F}_m(t)$, the true delete-1 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. For the design-based case, in terms of conditional performance, v_{Jd2} , the true delete-1 jackknife variance estimator, performs the best. However, it tends to have large positive unconditional bias when $F(t)$ is close to 0 or 1 and n is not large.

ACKNOWLEDGMENTS

This research was supported by a grant from the Natural Sciences and Engineering Council of Canada. The first author was partially supported by the Edward C. Bryant Scholarship of the American Statistical Association. Our thanks are due to Dr. Jiahua Chen for helpful comments and suggestions.

REFERENCES

- CHAMBERS, R. L. & DUNSTAN, R. (1986). "Estimating Distribution Function from Survey Data", *Biometrika*, 73, 597-604.
- CHAMBERS, R. L., DORFMAN, A. H. & HALL, P. (1992). "Properties of Estimators of the Finite Distribution Function", *Biometrika*, 79, 577-582.
- RANDLES, R. H. (1982). "On the Asymptotic Normality of Statistics with Estimated Parameters", *Annals of Statistics*, Vol. 10, No. 2, 462-474.
- RAO, J. N. K., KOVAR, J. G. & MANTEL, H. J. (1990). "On Estimating Distribution Functions and Quantiles from Survey Data Using Auxiliary Information", *Biometrika*, 77, 365-375.
- SHAO, J. & TU, D. (1995). *The Jackknife and Bootstrap*, New York: Springer-Verlag.
- WANG, S. & DORFMAN, A. H. (1996). "A New Estimator for the Finite Population Distribution Function", *Biometrika*, 83, 639-652.
- WU, C. (1999). "The Effective Use of Complete Auxiliary Information From Survey Data", Unpublished Ph.D. dissertation, Simon Fraser University, Burnaby, BC.