The Annals of Statistics
1993, Vol. 21, No. 3, 1452–1475

ESTIMATORS OF THE FINITE POPULATION DISTRIBUTION FUNCTION USING NONPARAMETRIC REGRESSION

BY ALAN H. DORFMAN AND PETER HALL

Bureau of Labor Statistics and Australian National University

This paper considers estimators of the distribution function of a variable over a finite population, when a sample of units is available and values of a related auxiliary variable are known for the whole population. Theory is offered for several estimators which rely on nonparametric regression, and for calibration estimators which require a parametric model.

1. Introduction. This paper concerns estimation of the distribution function of a variable over the units of a finite population based on a sample of units. For a variable of interest y, the distribution function is given by $F(t) = N^{-1}\{\sum_i I(y_i \le t) + \sum_j I(y_j \le t)\}$; here i indexes units in the sample, j indexes units in the nonsample and $I(\cdot)$ is the standard indicator function. The first term is known, and the task essentially is to estimate the second term. It is assumed that a regression relationship of some sort exists between the variable of interest and an auxiliary variable known for all the units of the population.

In this paper, we consider three distinct regression relationships that can arise in practice:

1. y has a well-defined, typically linear, relation to x, for example, $y_i = a + bx_i + \varepsilon_i$, $i = 1, \dots, N$, with $E\varepsilon_i = 0$, $\operatorname{var}(\varepsilon_i) = \sigma^2$, $\operatorname{cov}(\varepsilon_i, \varepsilon_j) = 0$ and the ε_i have common distribution function $G(u) = P(\varepsilon \le u)$.
2. y has an ill-defined but smooth relationship to x, that is, $y_i = m(x_i) + \varepsilon_i$, $i = 1, \dots, N$, with ε_i as above, and $m(\cdot)$ sufficiently smooth to apply standard nonparametric regression.
3. Instead of a relation between y itself and x, we merely assume one between $h(y_i) \equiv I(y_i \le t)$ and x_i, so that $E(h(y_i)) = H(x_i)$, with $H(\cdot)$ smooth.

Clearly, the above models are not exhaustive. In particular, the heteroscedastic case, with the variance of y dependent on x, is important to survey sampling practice, but is not herein considered, and will require further investigation.

Model-based estimators of $F(t)$ take the form

$$\hat{F}(t) = N^{-1}\left\{\sum_i I(y_i \le t) + \sum_j \hat{H}(x_j)\right\},$$

Received March 1991; revised March 1992.

AMS 1991 *subject classifications*. Primary 62D05; secondary 62G05, 62J99.

Key words and phrases. Auxiliary information, design-based estimation, model-based sampling, survey sample data, bandwidth, kernel estimator.

where $\hat{H}$ is an estimator of H based on one of the three above regression models. In particular, the Chambers and Dunstan (1986) (CD hereafter) estimator relies on (1), using standard linear regression to estimate the parameters, estimating G through the residuals, and relying on the relation $H(x_j) = G(t - a - bx_j)$. One can proceed similarly on the basis of (2), using nonparametric regression to estimate $m(\cdot)$ to get a nonparametric CD estimator. Kuo (1988) suggests applying nonparametric regression directly to $H(\cdot)$ to estimate $F(t)$, in effect using model (3).

In practice, a model such as (1) is unlikely to be strictly correct. This may be an argument for relying on (2) or (3). However, many survey practioners will prefer instead to use a design-based approach. Rao, Kovar and Mantel (1990) (henceforth RKM) suggest a design-based estimator which uses the working model(1) as a calibration device and is of the form

$$\check{F}(t) = N^{-1}\left\{\sum_i \pi_i^{-1} h(y_i) + \sum_{k=1}^{N} \hat{H}(x_k) - \sum_i \pi_i^{-1}\hat{H}(x_i)\right\},$$

where π_i is the probability that unit i is included in the sample; in the case of simple random sampling, $\pi_i = n/N$, where n is the sample size. Estimators of this form based on models (2) and (3) are also of interest and are considered in Section 2.

The use of (1) as a calibration device in the RKM estimator suggests an analogous model-based calibration estimator. For $\pi_i = nN^{-1}$ we can write the RKM estimator as

$$\check{F}(t) = N^{-1}\left\{\sum_i h(y_i) + \sum_j \hat{H}(x_j) + [(N - n)/n]\sum_i r_i\right\},$$

where $r_i = h(y_i) - \hat{H}(x_i)$ are residuals that reflect how well (1) predicts the actual sample values $h(y_i)$. Based on this representation, Chambers, Dorfman and Wehrly (1993) (henceforth CDW) suggest the nonparametric calibration estimator

$$\hat{F}(t) = N^{-1}\left\{\sum_i h(y_i) + \sum_j \hat{H}(x_j) + \sum_j \hat{r}_j\right\},$$

where $\hat{r}_j$ is estimated through nonparametric regression.

This paper provides large sample theory for six estimators of $F(t)$ suggested by these different approaches, taking as a basis for comparison the underlying model (2). Section 2 considers model-based and design-based estimators that arise when the working model is (2) or (3). Section 3 considers the design-calibrated (RKM) estimator, and the nonparametric calibrated (CDW) estimator when the working model is (1).

Results are summarized in Section 4, which contains a convenient tabulation of the asymptotic biases of the estimators. One important result is that the RKM design calibrated estimator is nonrobust against failure of (1), being less generally reliable than estimators relying on nonparametric regression. Several estimators, including the CDW nonparametric calibrated estimator, are left in competition; the way is open for further work to decide on their relative merits.

2. Nonparametric estimators.

2.1. *Nonparametric* CD *and* RKM *estimators.* Suppose $y_i = m(x_i) + \varepsilon_i$, as in schema (2) of the Introduction. Let $G(u) = P(\varepsilon \le u)$ and $g(u) = G'(u)$. Let $K(\cdot)$ be a symmetric density function, and $h > 0$. Consider estimators of $m(x)$ given by

$$\hat{m}(x) = \left[\sum_i y_i K\{(x - x_i)/h\}\right]\left[\sum_i K\{(x - x_i)/h\}\right]^{-1},$$

$$\hat{m}_{i_1}(x) = \left[\sum_{i \ne i_1} y_i K\{(x - x_i)/h\}\right]\left[\sum_{i \ne i_1} K\{(x - x_i)/h\}\right]^{-1},$$

and let

$$\hat{\varepsilon}_i = y_i - \hat{m}_i(x_i), \qquad \hat{G}(u) = n^{-1} \sum_i I(\hat{\varepsilon}_i \le u),$$

$$\tilde{F}(t) = n^{-1} \sum_i I(y_i \le t),$$

$$\hat{F}_1(t) = N^{-1}\left[\sum_i I(y_i \le t) + \sum_j \hat{G}\{t - \hat{m}(x_j)\}\right],$$

$$\breve{F}_1(t) = \tilde{F}(t) + N^{-1} \sum_j \hat{G}\{t - \hat{m}(x_j)\} - (n^{-1} - N^{-1}) \sum_i \hat{G}\{t - \hat{m}_i(x_i)\}.$$

$\hat{F}_1(t)$ and $\breve{F}_1(t)$ are analogues to the original CD and RKM estimators, respectively, in the context of schema (2). The leave-one-out estimator $\hat{m}_i(x_i)$ in $\breve{F}_1(t)$ is used merely to simplify the mathematical argument. $\tilde{F}(t)$ is the naive estimator, which makes no use of the auxiliary information in x. Here, for simplicity, we assume the form of the design-based estimators appropriate for simple random sampling.

NOTATION. Put $k_1 = \int K^2(y)\,dy$, $k_2 = \int y^2 K(y)\,dy$. Assume sample and nonsample values of x are in the interval $[a, b]$, and are generated by design densities d_s and $d_{p \setminus s}$, respectively, both bounded away from zero on $[a, b]$. Here d_s and $dp \setminus s$ are defined by

$$n^{-1} \sum_i I(x_i \le x) \to \int_{-\infty}^{x} d_s(y)\,dy,$$

$$(N - n)^{-1} \sum_j I(x_j \le x) \to \int_{-\infty}^{x} d_{p \setminus s}(y)\,dy,$$

for all x. Under simple random sampling, $d_s = d_{p \setminus s}$. Assume n and N increase together such that $n/N \to \pi$, with $0 < \pi < 1$. Define $\alpha(x) = [(\partial/\partial y)^2\{m(x-y) - m(x)\}d_s(x-y)]_{y=0}$,

$$I_1(T) = \sigma^2 \int g'(t-m) d_s^{-1} d_T + \sigma^2(b-a) \int g'(t-m) d_T,$$

$$I_2(T) = \left(\int \alpha\right) \int g(t-m) d_T - \int \alpha g(t-m) d_s^{-1} d_T,$$

$$I_3(T_1, T_2) = \sigma^2 \int \left\{ g(t-m) d_s^{-1} d_{T_1} - \int g(t-m) d_{T_1} \right\}$$
$$\times \left\{ g(t-m) d_s^{-1} d_{T_2} - \int g(t-m) d_{T_2} \right\} d_s,$$

$$I_4(T_1, T_2) = \int\int G[\{t - m(x)\} \wedge \{t - m(y)\}] d_{T_1}(x) d_{T_2}(y),$$

$$I_5(T) = \int G(t-m) d_T, \qquad I_6(T) = \int \left\{ G(t-m) - G(t-m)^2 \right\} d_T$$

and $I_j(T) = I_j(T,T)$ for $j = 3, 4$.

For the naive estimator $\tilde{F}(t)$, define $t_k = t - m(x_k)$. Then

$$\operatorname{var}\left\{\tilde{F}(t) - F(t)\right\} - N^{-2} \operatorname{var}\left\{ \sum_j I(Y_j \le t_j) \right\}$$
$$= (n^{-1} - N^{-1})^2 \sum_i \left\{ G(t_i) - G(t_i)^2 \right\}$$
$$= n^{-1}(1-\pi)^2 I_6(s) + o(n^{-1}).$$

Under simple random sampling, $E\{\tilde{F}(t) - F(t)\}$ is of size $n^{-1/2}$. For $\hat{F}_1(t)$ and $\breve{F}_1(t)$ we have the following results:

THEOREM 1.

(i)
$$E\left\{\hat{F}_1(t) - F(t)\right\} = \tfrac{1}{2} k_1 (1-\pi)(nh)^{-1} I_1(p \setminus s)$$
$$+ \tfrac{1}{2} k_2 (1-\pi) h^2 I_2(p \setminus s)$$
$$+ o\left\{(nh)^{-1} + h^2\right\},$$

(ii)
$$\operatorname{var}\left\{\hat{F}_1(t) - F(t)\right\} - N^{-2} \operatorname{var}\left\{ \sum_j I(Y_j \le t_j) \right\}$$
$$= n^{-1}(1-\pi)^2 \left\{ I_3(p \setminus s) + I_4(p \setminus s) - I_5(p \setminus s)^2 \right\}$$
$$+ o(n^{-1}) + O\left\{(nh)^{-2} + h^4\right\}.$$

THEOREM 2.

$$\begin{aligned} E\{\breve{F}_1(t) - F(t)\} = {} & \tfrac{1}{2}k_1(1-\pi)(nh)^{-1}\{I_1(p \setminus s) - I_1(s)\} \\ & + \tfrac{1}{2}k_2(1-\pi)h^2\{I_2(p \setminus s) - I_2(s)\} \\ & + o\{(nh)^{-1} + h^2\}, \end{aligned} \tag{i}$$

$$\begin{aligned} \operatorname{var}\{\breve{F}_1(t) - F(t)\} & - N^{-2}\operatorname{var}\Big\{\sum_j I(Y_j \le t_j)\Big\} \\ = n^{-1}(1-\pi)^2\Big[& \{I_3(s) - 2I_3(s, p \setminus s) + I_3(p \setminus s)\} \\ & + \{I_4(p \setminus s) - I_4(s)\} \\ & - \{I_5(p \setminus s)^2 - I_5(s)^2\} + I_6(s)\Big] + o(n^{-1}) \\ + O\{(nh)^{-2} & + h^4\}. \end{aligned} \tag{ii}$$

These results are discussed in subsections 2.2 and Section 4. Proofs are given in the Appendix.

2.2. *The Kuo and design-adjusted Kuo estimators.* Put $H(x) = G\{t - m(x)\}$,

$$v_{i_1 j} = K\{(x_j - x_{i_1})/h\}\Big[\sum_i K\{(x_j - x_i)/h\}\Big]^{-1},$$

$$v^{i_1 i_2} = K\{(x_{i_2} - x_{i_1})/h\}\Big[\sum_{i \ne i_2} K\{(x_{i_2} - x_i)/h\}\Big]^{-1},$$

$$\hat{H}(x_j) = \sum_i v_{ij} I(y_i \le t), \qquad \hat{H}_i(x_i) = \sum_{i_1 \ne i} v^{i,i} I(y_{i_1} \le t),$$

$$\hat{F}_2(t) = N^{-1}\Big\{\sum_i I(y_i \le t) + \sum_j \hat{H}(x_j)\Big\},$$

$$\breve{F}_2(t) = \tilde{F}(t) + N^{-1}\sum_j \hat{H}(x_j) - (n^{-1} - N^{-1})\sum_i \hat{H}_i(x_i).$$

$\hat{F}_2(t)$ is a nonparametric generalization of the naive estimator $\tilde{F}(t)$, originally suggested by Kuo (1988). Unlike the naive estimator, however, $\hat{F}_2(t)$ does rely on the auxiliary information, as incorporated in schema (3). $\breve{F}_2(t)$ is an analogue of the RKM estimator under schema (3). For convenience of comparison, we study these estimators under the assumptions of subsection 2.1, so that $H(x_j) = E(I(y_j \le t) = G(t - m(x_j))$.

NOTATION. $\beta(x) = [(\partial/\partial y)^2\{H(x-y) - H(x)\}d_s(x-y)]_{y=0}$,

$$I_7(T) = \int d_s^{-1} d_T \beta, \qquad I_8(T) = \int \{G(t-m) - G(t-m)^2\} d_s^{-1} d_T^2.$$

THEOREM 3.

$$E\{\hat{F}_2(t) - F(t)\} = \tfrac{1}{2}k_2(1-\pi)h^2 I_7(p \setminus s) + o\{(nh)^{-1} + h^2\},$$

$$\operatorname{var}\{\hat{F}_2(t) - F(t)\} - N^{-2}\operatorname{var}\left\{\sum_j I(y_j \le t_j)\right\}$$

$$= n^{-1}(1-\pi)^2 I_8(p \setminus s) + o(n^{-1}),$$

$$E\{\breve{F}_2(t) - F(t)\} = \tfrac{1}{2}k_2(1-\pi)h^2\{I_7(p \setminus s) - I_7(s)\} + o\{(nh)^{-1} + h^2\},$$

$$\operatorname{var}\{\breve{F}_2(t) - F(t)\} - N^{-2}\operatorname{var}\left\{\sum_j I(y_j \le t_j)\right\}$$

$$= n^{-1}(1-\pi)^2 I_8(p \setminus s) + o(n^{-1}).$$

Proof is given in the Appendix.

NOTES. (i) To ensure that $E\{\hat{F}_1(t) - F(t)\}$, $E\{\breve{F}_1(t) - F(t)\}$ are both of size $o(n^{-1/2})$, the bandwidth h should be chosen to be of smaller order than $n^{-1/4}$ but larger order than $n^{-1/2}$: $n^{1/4}h \to 0$, $n^{1/2}h \to \infty$. For such a choice of h, and for $\bar{F} = \hat{F}_1$ or $\breve{F}_1$,

$$E\{\bar{F}(t) - F(t)\}^2 \text{ is of lower order than } \operatorname{var}\{\bar{F}(t) - F(t)\}.$$

(ii) Choice of h is not quite so critical in the cases of $\hat{F}_2$, $\breve{F}_2$. We need h to be of smaller order than $n^{-1/4}$, but not necessarily of larger order than $n^{-1/2}$.

(iii) Under simple random sampling we have $d_s = d_{p \setminus s}$, whence

$$\operatorname{var}\{\tilde{F}(t) - F(t)\} - \operatorname{var}\{\hat{F}_1(t) - F(t)\}$$

$$= n^{-1}(1-\pi)^2\{I_5(s)^2 + I_6(s) - I_4(s) - I_3(s)\} + o(n^{-1}).$$

It can be shown that

$$I_5^2 + I_6 - I_4 > 0.$$

Furthermore, if $g(x) = \sigma^{-1}l(x/\sigma)$ for a density l, then

$$I_3(s) = \sigma^2\left[\int g(t-m)^2 d_s - \left\{\int g(t-m)d_s\right\}^2\right]$$

$$= \int l\{(t-m)/\sigma\}^2 d_s - \left[\int l\{(t-m)/\sigma\}d_s\right]^2$$

$$\to 0 \quad \text{as } \sigma \to 0 \text{ and as } \sigma \to \infty.$$

This seems to be the only evidence that I_3 might be small; it is not useful, since $I_5^2 + I_6 - I_4 \to 0$ as $\sigma \to 0$ and as $\sigma \to \infty$. Therefore, whether $\hat{F}_1$ is superior to $\tilde{F}$ in terms of variance is unclear.

3. Design and nonparametric calibration estimators.

3.1. *Notation*. We take the working model as $y_k = a + bx_k + \varepsilon_k$, and allow for the possibility it is not the same as the true model $y_k = m(x_k) + \varepsilon_k$. Let

$$\hat{b} = \left(n\sigma_x^2\right)^{-1} \sum (x_i - \bar{x}) y_i,$$

$$\Delta \equiv \left(n\sigma_x^2\right)^{-1} \sum (x_i - \bar{x}) \varepsilon_i,$$

$$\mathring{b} \equiv E(\hat{b}) = \left(n\sigma_x^2\right)^{-1} \sum (x_i - \bar{x}) m(x_i),$$

$$\mathring{\beta} \equiv \sigma_x^{-2} \int (x - \mu) m(x) d_s(x)\, dx,$$

$$t_\nu = t - m(x_\nu),$$

$$\delta_\nu \equiv \delta(x_\nu) \equiv \mathring{b} x_\nu - m(x_\nu)$$

or $\mathring{\beta} x_\nu - m(x_\nu)$, when following integral sign,

$$t_{\nu i} \equiv t - m(x_\nu) - \delta(x_\nu) + \delta(x_i),$$

$$H(x_\nu, x_i) \equiv H_{\nu i} \equiv G(t_{\nu i})$$

and

$$D(x_\nu, x_i) \equiv D_{\nu i} = G(t_{\nu i}) - G(t_\nu).$$

Note that

$$E\Big(I\big(y_i - \hat{a} - \hat{b}x_i \le t - \hat{a} - \hat{b}x_\nu\big)\Big) = E\big(I(\varepsilon_i + \Delta(x_\nu - x_i) \le t_{\nu i})\big)$$
$$= H_{\nu i} + o(1).$$

Let $M = N - n = N(1 - \pi)$ and take $i, i_1, \dots, k, k_1, \dots$ as sample indices and $j, j_1, \dots$ as nonsample indices.

Let

$$H(x) = G(t - m(x)),$$

$$\beta(x) = H''(x) d_s(x) + 2H'(x) d_s'(x),$$

$$\gamma(u) = E(\varepsilon I(\varepsilon \le u)),$$

as before, and

$$\beta(x, y) = \frac{\partial^2 H(x, y)}{\partial x^2} d_s(x) + 2 \frac{\partial H(x, y)}{\partial x} d_s'(x).$$

Define

$$J_{00}(T) = \int (y - \mu)\gamma(t - m(y))d_T(y)\,dy,$$

$$J_0(T) = \iint (x - \mu)\gamma(t - m(y) - \delta(y) + \delta(x))d_s(x)d_T(y)\,dx\,dy,$$

$$J_1(T) = \iint (y - x)g(t - m(y) - \delta(y) + \delta(x))d_s(x)d_T(y)\,dx\,dy,$$

$$\begin{aligned} J_2(T_1, T_2) = \iiint \big[& G(t - m(y) - \delta(y) + \delta(x)) \\ & \wedge G(t - m(z) - \delta(z) + \delta(x)) \\ & \times d_s(x)d_{T_1}(y)d_{T_2}(z)\big]\,dx\,dy\,dz, \end{aligned}$$

$$\begin{aligned} J_3(T_1, T_2) = \iiint & G(t - m(y) - \delta(y) + \delta(x))G(t - m(z) - \delta(z) + \delta(x)) \\ & \times d_s(x)d_{T_1}(y)d_{T_2}(z)\,dx\,dy\,dz, \end{aligned}$$

$$\begin{aligned} J_4(T_1, T_2) = \iint & G(t - m(x))G(t - m(y) - \delta(y) + \delta(x)) \\ & \times d_{T_1}(x)d_{T_2}(y)\,dx\,dy, \end{aligned}$$

$$\begin{aligned} J_5(T_1, T_2) = \iint & G(t - m(x)) \\ & \wedge G(t - m(y) - \delta(y) + \delta(x))d_{T_1}(x)d_{T_2}(y)\,dx\,dy, \end{aligned}$$

$$J_6(T) = \int G(t - m(x)) - G(t - m(x))^2 d_T(x)\,dx = I_6(T),$$

$$J_8(T) = \int \{G(t - m(x)) - G(t - m(x))^2\} d_s(x)^{-1} d_T^2(x)\,dx = I_8(T).$$

As always, we assume errors ε_k have distribution function $G(\cdot)$.

3.2. *The design-calibration estimator under misspecification of model.* Let the working model be $y_k = a + bx_k + \varepsilon_k$, $k \in p$, and the underlying model be $y_k = m(x_k) + \varepsilon_k$, $k \in p$, and consider

$$\breve{F}(t) = \tilde{F}(t) + N^{-1}\sum_j \hat{G}\big(t - \hat{a} - \hat{b}x_j\big) - (n^{-1} - N^{-1})\sum_i \hat{G}\big(t - \hat{a} - \hat{b}x_i\big).$$

This is the calibration estimator of RKM, for simple random sampling, for which theory is given in Chambers, Dorfman and Hall (1992) (henceforth CDH) when $E(y) = a + bx$ is the correct model.

3.2.1. *Consistency.*

$$
\begin{aligned}
E\big(\breve{F}(t) - F(t)\big) &= n^{-1}(1-\pi)\sum_i E\{I(y_i \le t) - \hat{G}_i\} \\
&\quad - N^{-1}\sum_j E\big\{I(y_j \le t) - \hat{G}_j\big\} \\
&= n^{-1}(1-\pi)\sum_i \Big\{G\big(t - m(x_i)\big) - n^{-1}\sum_{i'} H_{ii'}\Big\} \\
&\quad - N^{-1}\sum_j \Big\{G\big(t - m(x_j)\big) - n^{-1}\sum_{i'} H_{ji'}\Big\} + O(n^{-1}) \\
&= (1-\pi)\Big[\int G\big(t - m(x)\big) d_s(x)\,dx \\
&\qquad - \iint G\big(t - m(x) - \delta(x) + \delta(y)\big) d_s(y) d_s(x)\,dx\,dy\Big] \\
&\quad -(1-\pi)\Big[\int G\big(t - m(x)\big) d_{p\setminus s}(x)\,dx \\
&\qquad - \iint G\big(t - m(x) - \delta(y) + \delta(y)\big) \\
&\qquad\qquad \times d_{p\setminus s}(x) d_s(y)\,dx\,dy\Big] + o(1).
\end{aligned}
$$

Therefore if sample and nonsample designs are the same we get consistency [it is shown below that $\operatorname{var}(\breve{F}(t) - F(t)) = O(n^{-1})$ even when the model is false]. If the designs are not the same, then $\breve{F}(t) - F(t)$ does not in general converge to zero. This behavior is the same as that of the naive estimator $\tilde{F}(t)$.

3.2.2. *Bias and variance.* Assume $d_{p\setminus s}(x) = d_s(x) = d(x)$. Note that $E(N(\breve{F}(t) - F(t)))$, regarded as a function of x-values, has a bounded mean. Some rather extended calculations yield

$$
\begin{aligned}
\operatorname{var}_{\mathbf{x}}\Big(NE\big(\breve{F}(t) - F(t)\big)\Big) &= n\{\rho\pi^{-1}\mu_{abad} - \rho^2\mu_{abca} \\
&\qquad + \rho^2\mu_{abbd} - (2\rho^2 + \rho)\mu_{ab}^2\} + o(n),
\end{aligned}
$$

where $\rho = (1-\pi)/\pi$, $\mu_{abad} = E_{\mathbf{x}}(D_{\nu i}D_{\nu i'})$, $\mu_{abca} = E_{\mathbf{x}}(D_{\nu i}D_{\nu'\nu})$, $\mu_{abbd} = E_{\mathbf{x}}(D_{\nu i}D_{ii'})$ and $\mu_{ab} = E_{\mathbf{x}}(D_{\nu i})$. For example,

$$
\begin{aligned}
\mu_{abad} = \iiint &\big[G\big(t - m(x) - \delta(x) + \delta(y)\big) - G\big(t - m(x)\big)\big] \\
&\times \big[G\big(t - m(x) - \delta(x) + \delta(z)\big) - G\big(t - m(x)\big)\big] \\
&\times d(x)d(y)d(z)\,dx\,dy\,dz.
\end{aligned}
$$

This variance is not in general zero, so that, we get

$$
E\big(\breve{F}(t) - F(t)\big) = O(n^{-1/2}).
$$

Thus the bias of $\breve{F}$ need be no better than that of the naive estimator $\tilde{F}(t)$, when the model is incorrect. As an example, suppose $m(x) = x^2$ and $d(x)$ is uniform on $[-1, 1]$. Then $\mathring{\beta} = 3\int_{-1}^{1} x^3(1/2)\,dx = 0$, whence $\delta(x) = -m(x)$, and $E(\breve{F}(t) - F(t)) = E(\tilde{F}(t) - F(t)) + O(n^{-1})$.

Under misspecification of the model, we have

$$\begin{aligned}
&\operatorname{var}\bigl(\breve{F}(t) - F(t)\bigr)\\
&\quad = n^{-1}(1-\pi)^2\Bigl\{J_6(s) + J_2(p \setminus s, p \setminus s) + J_2(s,s)\\
&\qquad\qquad -2J_2(s, p \setminus s) - \bigl[J_3(p \setminus s, p \setminus s)\\
&\qquad\qquad\qquad + J_3(s,s) - 2J_3(s, p \setminus s)\bigr]\\
&\qquad\qquad +2\bigl[J_5(s, p(s) - J_5(s,s)\bigr]\\
&\qquad\qquad -2\bigl[J_4(s, p \setminus s) - J_4(s,s)\bigr]\\
&\qquad\qquad +\sigma_x^{-2}\sigma^2\bigl[J_1(p \setminus s) - J_1(s)\bigr]^2\\
&\qquad\qquad -2\sigma_x^{-2}\bigl[J_1(p \setminus s)J_0(p \setminus s) + J_1(s)J_0(s)\\
&\qquad\qquad\qquad -J_1(p \setminus s)J_0(s) - J_1(s)J_0(p \setminus s)\bigr]\\
&\qquad\qquad\qquad -2\sigma_x^{-2}J_{00}(s)\bigl[J_1(p \setminus s) - J_0(s)\bigr]\Bigr\}\\
&\qquad + N^{-1}(1-\pi)J_6(p \setminus s) + o(n^{-1}).
\end{aligned}$$

Proof is given in the Appendix.

NOTE. If $d_{p \setminus s} = d_s$, then this reduces to

$$n^{-1}(1-\pi)^2 I_6(s) + N^{-1}(1-\pi)I_6(s) + o(n^{-1}) \sim \operatorname{var}\bigl(\tilde{F}(t) - F(t)\bigr).$$

Thus, if the model is wrong, $\breve{F}(t)$ has bias of the same order as $\tilde{F}(t)$ (naive estimator) and has the same variance.

3.3. *The nonparametric calibration estimator.*

DEFINITION.

$$\begin{aligned}
\hat{F}_3(t) &\equiv N^{-1}\Bigl\{\sum_i I(y_i \le t) + \sum_j \hat{G}\bigl(t - \hat{a} - \hat{b}x_i\bigr)\\
&\qquad + \sum_j\sum_i v_{ij}\bigl[I(y_i \le t) - \hat{G}\bigl(t - \hat{a} - \hat{b}x_i\bigr)\bigr]\Bigr\}\\
&= N^{-1}\Bigl\{\sum_i I(y_i \le t) + \sum_j \hat{H}(x_j)\\
&\qquad - \sum_j\sum_i v_{ij}\bigl[\hat{G}\bigl(t - \hat{a} - \hat{b}x_i\bigr) - \hat{G}\bigl(t - \hat{a} - \hat{b}x_j\bigr)\bigr]\Bigr\}
\end{aligned}$$

with v_{ij}, $\hat{H}$ as defined in subsection 2.2. We assume the true model is $y_l = m(x_l) + \varepsilon_l$.

3.3.1. *Bias.*

$$E\big(N\big(\hat{F}_3(t)-F(t)\big)\big)=E\Big\{\sum_j\sum_i v_{ij}I(y_i\le t)-I_j(y_j\le t)\Big\}$$
$$-E\Big\{n^{-1}\sum_i\sum_k\sum_j v_{ij}\Big[I\big(y_k-\hat{a}-\hat{b}x_k\le t-\hat{a}-\hat{b}x_i\big)$$
$$-I\big(y_k-\hat{a}-\hat{b}x_k\le t-\hat{a}-\hat{b}x_j\big)\Big]\Big\}.$$

The first term is $(1/2)k_2Mh^2\int d_s^{-1}(x)d_{p\setminus s}(x)\beta(x)\,dx+o(nh^2+h^{-1})$ (compare Appendix A2.2).

For the second term, note that

$$E\Big(I\big(y_k-\hat{\beta}x_k\le t-\hat{a}-\hat{b}x_i\big)-I\big(y_k-\hat{a}-\hat{b}x_k\le t-\hat{a}-\hat{b}x_j\big)\Big)$$
$$=G(t_{ik})-G(t_{jk})+O(n^{-1})$$
$$\equiv H_{ik}-H_{jk}+O(n^{-1}).$$

Therefore the second term is

$$n^{-1}\sum_i\sum_k\sum_j v_{ij}(H_{ik}-H_{jk})+O(1)$$
$$=n^{-1}\sum_k\Big\{\sum_j\hat{d}(x_j)^{-1}(nh)^{-1}\sum_i K\big((x_j-x_k)/h\big)\big(H_{ik}-H_{jk}\big)\Big\}+O(1)$$
$$=\tfrac{1}{2}k_2Mh^2\iint\beta(x,y)d_s^{-1}(x)d_{p\setminus s}(x)d_s(y)\,dx\,dy+o(nh^2+h^{-1}).$$

In sum,

$$NE\big(\hat{F}_3(t)-F(t)\big)=\tfrac{1}{2}k_2Mh^2\Big[\int d_s^{-1}(x)d_{p\setminus s}(x)\beta(x)\,dx$$
$$-\iint d_s^{-1}(x)d_{p\setminus s}(x)d_s(y)\beta(x,y)\,dx\,dy\Big]$$
$$+o(nh^2+h^{-1}).$$

REMARK. If the model is correct, that is, if $m(x)=a+bx$, then $\beta(x,y)=\beta(x)$, and N bias $=o(nh^2+h^{-1})$. However, a stronger result is available:

$$E\big(N\big(\hat{F}_3(t)-F(s)\big)\big)=\sum_j\sum_i v_{ij}G(t_i)-\sum_j G(t_j)$$
$$-\sum_i\sum_j v_{ij}\big(G(t_i)-G(t_j)\big)+O(1)$$
$$=O(1)$$

so that if the model is correct, the bias is actually of order N^{-1}.

TABLE 1
Bias results

Estimator	Name	Parametric model required?	Bias	
			Model correct	Incorrect
$\tilde{F}$	Naive	No		$O(n^{-1/2})$*
$\hat{F}$	Chambers–Dunstan	Yes	$O(n^{-1})$	$O(c)$
$\hat{F}_1$	nonparametric CD	No		$O[(nh)^{-1} + h^2]$**
$\breve{F}$	Rao–Kovar–Mantel	Yes	$O(n^{-1})$	$O(n^{-1/2})$*
$\breve{F}_1$	nonparametric RKM	No		$O[(nh)^{-1} + h^2]$**
$\hat{F}_2$	Kuo	No		$O(h^2) + o[(nh)^{-1}]$**
$\breve{F}_2$	design-calibrated Kuo	No		$O(h^2) + o[(nh)^{-1}]$**
$\hat{F}_3$	nonparametric calibration	Yes	$O(n^{-1})$	$O(h^2) + o[(nh)^{-1}]$**

3.3.2. *Variance.*

$$\operatorname{var}\big(\hat{F}_3(t) - F(t)\big) = n^{-1}(1-\pi)^2 I_8(p \setminus s) + N^{-1}(1-\pi) I_6(p \setminus s) + o(n^{-1}).$$

REMARK. This holds regardless of whether model is correct or not, and is the same expression as for the variance of the simpler Kuo estimator $\hat{F}_2(t)$. The expression reduces to that for var($\tilde{F}(t) - F(t)$) under SRS. Proof is given in the Appendix.

4. Summary. We can tabulate results on bias as in Table 1. Results for $\tilde{F}$, $\hat{F}$ and $\breve{F}$ under the correct model appear in CDH; see also Dorfman (1993).

The entries marked (*) require the sample and nonsample designs to be the same. Under this condition, bias for $\breve{F}_1$ (nonparametric RKM) becomes $o((nh)^{-1} + h^2)$.

The entries marked (**) can be regarded as $o(n^{-1/2})$, for suitable rate of convergence of h to zero ($h = Cn^{-c}$, $1/4 < c < 1/2$ will do it). In these cases and the $O(n^{-1})$ cases, bias is an insignificant part of mean square error. The "bias-vulnerable" cases are then: CD, naive, and RKM, this last being something of a surprise.

When the sample and nonsample designs are the same, as in simple random sampling, then all the estimators except CD and nonparametric CD have the same variance up to $o(n^{-1})$. Thus judgements on the non-CD estimators can be based on bias considerations. Within this group $\hat{F}_3$, the nonparametric calibration (CDW) estimator, stands out: It has bias of same order as the other model-based estimators if the model is correct, but does not share in their vulnerability if the model is wrong. It will do better than $\hat{F}_2$ (Kuo's), if the model is correct, and so presumably if it is nearly correct. Kuo may do better if the model is severely off. The relative worth of $\hat{F}_1$, the nonparametric CD estimator, is as yet unclear.

Comparison of the estimators using simulation studies is a nontrivial step. To compare the estimators as they would do *at their best*, requires development of suitable methods for each of them of choosing correct bandwidths. Standard methods of selecting bandwidth lead to oversmoothing. Promising empirical results appears in CDW for the Kuo and nonparametric calibration estimators, but more study is required. Suitable choice of bandwidth for the nonparametric CD estimator is an open question.

APPENDIX

This Appendix contains proofs of Theorems 1, 2 and 3, and for the expansions of the variances of the calibration estimators, under misspecification of the model. Subsection numbers correspond to sections of the main text.

A2.1.1. Proof of Theorem 1.

(a) *Preliminaries.* Define

$$\Delta_{ij} = \hat{m}(x_j) - \hat{m}_i(x_i) - E\{\hat{m}(x_j) - \hat{m}_i(x_i)\},$$
$$d_{ij} = E\{\hat{m}(x_j) - \hat{m}_i(x_i)\} - \{m(x_j) - m(x_i)\},$$
$$t_j = t - m(x_j), \qquad t_{ij} = t_j - d_{ij}.$$

Then

$$n\sum_j \hat{G}\{t - \hat{m}(x_j)\} = \sum_i \sum_j I\{\hat{\varepsilon}_i \le t - \hat{m}(x_j)\}$$
$$= \sum_i \sum_j I(\varepsilon_i + \Delta_{ij} \le t_{ij}).$$

Put

$$\hat{d}(x) = (nh)^{-1}\sum_i K\{(x - x_i)/h\},$$
$$\hat{d}_{i_1}(x) = \{(n-1)h\}^{-1}\sum_{i \ne i_1} K\{(x - x_i)/h\},$$
$$w_{i_1 i_2 j} = (nh)^{-1}\Big[\hat{d}(x_j)^{-1}K\{(x_j - x_{i_1})/h\}$$
$$-(1 - n^{-1})^{-1}\hat{d}_{i_2}(x_{i_2})^{-1}K\{(x_{i_2} - x_{i_1})/h\}I(i_1 \ne i_2)\Big].$$

In this notation,

$$\Delta_{ij} = \sum_{i_1} w_{i_1 ij}\varepsilon_{i_1}.$$

(b) *Bias*. Observe that

$$
\begin{aligned}
E\left[n\sum_j \hat{G}\{t-\hat{m}(x_j)\}\right]
&= \sum_i\sum_j P(\varepsilon_i+\Delta_{ij}\le t_{ij})\\
&= \sum_i\sum_j E\left[G\left\{\left(t_{ij}-\sum_{i_1\ne i} w_{i_1ij}\varepsilon_{i_1}\right)\Big/(1+w_{iij})\right\}\right]\\
&= \sum_i\sum_j G\left\{t_{ij}(1+w_{iij})^{-1}\right\}+\tfrac{1}{2}\sum_i\sum_j E\left\{\left(\sum_{i_1\ne i} w_{i_1ij}\varepsilon_{i_1}\right)^2(1+w_{iij})^{-2}\right\}\\
&\qquad\times g'\left\{t_{ij}(1+w_{iij})^{-1}\right\}+o(nh^{-1})\\
&= \sum_i\sum_j G(t_{ij})-\sum_i\sum_j w_{iij}t_{ij}g(t_j)\\
&\qquad+\tfrac{1}{2}\sigma^2\sum_i\sum_j g'(t_j)\sum_{i_1} w_{i_1ij}^2+o(nh^{-1}).
\end{aligned}
\tag{A.1}
$$

Now,

$$\sum_i\sum_j w_{iij}t_jg(t_j)=\sum_j t_jg(t_j)=O(n)=o(nh^{-1}),$$

$$
\begin{aligned}
&\sum_i\sum_j g'(t_j)\sum_{i_1} w_{i_1ij}^2\\
&\quad=(N-n)h^{-1}\left[(N-n)^{-1}\sum_j g'(t_j)\hat{d}(x_j)^{-2}(nh)^{-1}\sum_{i_1}K\{(x_j-x_{i_1})/h\}^2\right.\\
&\qquad\qquad-2(N-n)^{-1}(1-n^{-1})^{-1}\\
&\qquad\qquad\times\sum_j g'(t_j)\hat{d}(x_j)^{-1}n^{-1}\sum_i \hat{d}_i(x_i)^{-1}(nh)^{-1}\\
&\qquad\qquad\times\sum_{i_1\ne i}K\{(x_j-x_{i_1})/h\}K\{(x_i-x_{i_1})/h\}\\
&\qquad\qquad+\left\{(N-n)^{-1}\sum_j g'(t_j)\right\}\\
&\qquad\qquad\left.\times(1-n^{-1})^{-2}n^{-1}\sum_i \hat{d}_i(x_i)^{-2}(nh)^{-1}\sum_{i_1\ne i}K\{(x_i-x_{i_1})/h\}^2\right]\\
&\quad=N(1-\pi)h^{-1}\left(\int K^2\right)\left[\int g'\{t-m(x)\}d_s(x)^{-1}d_{p\setminus s}(x)\,dx\right.\\
&\qquad\qquad\left.+(b-a)\int g'\{t-m(x)\}d_{p\setminus s}(x)\,dx\right]+o(nh^{-1}),
\end{aligned}
$$

$$\begin{aligned}
&\sum_i \sum_j \{G(t_{ij}) - G(t_j)\} \\
&\quad = -\sum_i \sum_j g(t_j) d_{ij} + O\Big(\sum_i \sum_j d_{ij}^2\Big) \\
&\quad = -\sum_i \sum_j g\{t - m(x_j)\}\tfrac{1}{2}k_2 \\
&\qquad \times \Big\{d_s(x_j)^{-1}\alpha(x_j) - d_s(x_i)^{-1}\alpha(x_i)\Big\}h^2 + o(n^2h^2 + nh^{-1}) \\
&\quad = \tfrac{1}{2}k_2 n(N-n)h^2\Big[\Big\{\int g(t-m)d_{p\setminus s}\Big\}\int\alpha \\
&\qquad\qquad - \int g(t-m)d_s^{-1}d_{p\setminus s}\alpha\Big] \\
&\qquad + o(n^2h^2 + nh^{-1}).
\end{aligned}$$

Combining the results from (A.1) down we see that

$$\begin{aligned}
E\big\{\hat{F}_1(t) - F(t)\big\} &= N^{-1}\sum_j \Big(E\big[\hat{G}\{t - \hat{m}(x_j)\}\big] - G\{t - m(x_j)\}\Big) \\
&= \tfrac{1}{2}k_1(1-\pi)(nh)^{-1}I_1(p\setminus s) + \tfrac{1}{2}k_2h^2(1-\pi)I_2(p\setminus s) \\
&\quad + o\big\{(nh)^{-1} + h^2\big\}.
\end{aligned}$$

(c) *Variance*. Observe that

$$\begin{aligned}
\operatorname{var}\big\{\hat{F}_1(t) - F(t)\big\} &= (nN)^{-2}\operatorname{var}\Big\{\sum_i\sum_j I(\varepsilon_i + \Delta_{ij} \le t_{ij})\Big\} \\
&\quad + N^{-2}\operatorname{var}\Big\{\sum_j I(Y_j \le t_j)\Big\}.
\end{aligned} \tag{A.2}$$

Now,

$$\operatorname{var}\Big\{\sum_i\sum_j I(\varepsilon_i + \Delta_{ij} \le t_{ij})\Big\} = \sum_i\sum_{i'}\sum_j\sum_{j'}\eta(i,i',j,j'), \tag{A.3}$$

where

$$\begin{aligned}
\eta(i,i',j,j') &= P(\varepsilon_i + \Delta_{ij} \le t_{ij}, \varepsilon_{i'} + \Delta_{i'j'} \le t_{i'j'}) \\
&\quad - P(\varepsilon_i + \Delta_{ij} \le t_{ij})P(\varepsilon_{i'} + \Delta_{i'j'} \le t_{i'j'}).
\end{aligned}$$

Since

$$\eta(i,i,j,j') = G(t_j \wedge t_{j'}) - G(t_j)G(t_{j'}) + o(1),$$

then

$$\begin{aligned}
\sum_{i=i'}\sum_j\sum_{j'}\eta(i,i',j,j') &= n\sum_j\sum_{j'}\big\{G(t_j \wedge t_{j'}) - G(t_j)G(t_{j'})\big\} \\
&\quad + o(n^3).
\end{aligned} \tag{A.4}$$

Next we treat the sum over $i \neq i'$.

It may be shown that if c_1, c_2 are constants then, defining $\gamma(u) = E\{\varepsilon I(\varepsilon \le u)\}$,

$$P(\varepsilon_1 + c_1\varepsilon_2 \le u_1, \varepsilon_2 + c_2\varepsilon_1 \le u_2) = G(u_1)G(u_2) - c_1 g(u_1)\gamma(u_2) - c_2 g(u_2)\gamma(u_1) + O(c_1^2 + c_2^2)$$

as $|c_1| + |c_2| \to 0$. Therefore, defining

$$U_{ii'j} = \sum_{i_1 \ne i, i'} w_{i_1 ij}\varepsilon_{i_1},$$

we have

$$\begin{aligned} &P(\varepsilon_i + \Delta_{ij} \le t_{ij}, \varepsilon_{i'} + \Delta_{i'j'} \le t_{i'j'}) \\ &\quad = P\{(1 + w_{iij})\varepsilon_i + w_{i'ij}\varepsilon_{i'} \\ &\qquad \le t_{ij} - U_{ii'j}, (1 + w_{i'i'j'})\varepsilon_{i'} + w_{ii'j'}\varepsilon_i \le t_{i'j'} - U_{i'ij'}\} \\ &\quad = E\big[G\{(t_{ij} - U_{ii'j})/(1 + w_{iij})\}G\{(t_{i'j'} - U_{i'ij'})/(1 + w_{i'i'j'})\}\big] \\ &\qquad - w_{i'ij}g(t_{ij})\gamma(t_{i'j'}) - w_{ii'j'}g(t_{i'j'})\gamma(t_{ij}) + O\{(nh)^{-2}\}. \end{aligned} \tag{A.5}$$

Similarly,

$$P(\varepsilon_i + \Delta_{ij} \le t_{ij}) = E\big[G\{(t_{ij} - U_{ii'j})/(1 + w_{iij})\}\big] + O(w_{i'ij}^2). \tag{A.6}$$

Observe that $t_{ij} = t_j - d_{ij} = t_j + O(h^2)$,

$$\sum_{i \ne i'}\sum w_{i'ij} = \sum_i \sum_{i'} w_{i'ij} - \sum_i w_{iij} = (n - n) - 1 = -1, \tag{A.7}$$

$$\begin{aligned} \sum_i \sum_{i'} w_{i'ij}^2 &\le 2(nh)^{-2} \sum_i \sum_{i'} \Big[\hat{d}(x_j)^{-2} K\{(x_j - x_{i'})/h\}^2 \\ &\qquad + (1 - n^{-1})^{-2} \hat{d}_{i'}(x_i)^{-2} K\{(x_i - x_{i'})/h\}^2 I(i' \ne i)\Big] \\ &= O(h^{-1}). \end{aligned} \tag{A.8}$$

From (A.7) we see that

$$\begin{aligned} \sum_{i \ne i'}\sum w_{i'ij}g(t_{ij})\gamma(t_{i'j'}) &= \sum_{i \ne i'}\sum w_{i'ij}g(t_j)\gamma(t_{j'}) + O\Big(h^2 \sum_{i \ne i'}\sum |w_{i'ij}|\Big) \\ &= O(1 + nh^2), \end{aligned}$$

whence by (A.5),

$$\begin{aligned} &\sum_{i \ne i'}\sum \sum_j \sum_{j'} P(\varepsilon_i + \Delta_{ij} \le t_{ij}, \varepsilon_{i'} + \Delta_{i'j'} \le t_{i'j'}) \\ &\quad = \sum_{i \ne i'}\sum \sum_j \sum_{j'} E\big[G\{(t_{ij} - U_{ii'j})/(1 + w_{iij})\} \\ &\qquad\qquad \times G\{(t_{i'j'} - U_{i'ij'})/(1 + w_{i'i'j'})\}\big] \\ &\quad + O(n^3h^2 + n^2h^{-2}). \end{aligned} \tag{A.9}$$

By (A.6) and (A.8),

$$\begin{aligned}
&\sum_{i\neq i'}\sum_{j}\sum_{j'}\sum P(\varepsilon_i+\Delta_{ij}\le t_{ij})P(\varepsilon_{i'}+\Delta_{i'j'}\le t_{i'j'})\\
&\quad=\sum_{i\neq i'}\sum_{j}\sum_{j'}\sum E\big[G\{(t_{ij}-U_{ii'j})/(1+w_{iij})\}\big]\\
&\quad\quad\times E\big[G\{(t_{i'j'}-U_{i'ij'})/(1+w_{i'i'j'})\}\big]+O(n^2h^{-1}).
\end{aligned}$$

Combining this result with (A.9), and Taylor expanding the functions G, we obtain

$$\begin{aligned}
&\sum_{i\neq i'}\sum_{j}\sum_{j'}\sum \eta(i,i',j,j')+O(n^3h^2+n^2h^{-2})\\
&\quad=\sum_{i\neq i'}\sum_{j}\sum_{j'}\sum E(U_{ii'j}U_{i'ij'})(1+w_{iij})^{-1}(1+w_{i'i'j'})^{-1}\\
&\quad\quad\times g\Big\{t_{ij}(1+w_{iij})^{-1}\Big\}g\Big\{t_{i'j'}(1+w_{i'i'j'})^{-1}\Big\}\\
&\quad\quad+O\Bigg[\sum_{i\neq i'}\sum_{j}\sum_{j'}\sum\Big\{\Big|E\big(U^2_{ii'j}U_{i'ij'}\big)\Big|+E\big(U^4_{ii'j}\big)\Big\}\Bigg]\\
&\quad=\sigma^2\sum_{i}\sum_{i'}\sum_{j}\sum_{j'}\Bigg(\sum_{i_1}w_{i_1ij}w_{i_1i'j'}\Bigg)g(t_j)g(t_{j'})\\
&\quad\quad+O\Bigg[\{(nh)^{-1}+h^2\}\sum_{i}\sum_{i'}\sum_{j}\sum_{j'}\big|E(U_{ii'j}U_{i'ij'})\big|\\
&\quad\quad\quad+\sum_{i}\sum_{j}\sum_{j'}\big|E(U_{iij}U_{iij'})\big| \qquad\qquad \text{(A.10)}\\
&\quad\quad\quad+\sum_{i}\sum_{i'}\sum_{j}\sum_{j'}\Bigg\{\sum_{i_1}\big|w^2_{i_1ij}w_{i_1i'j'}\big|+\Bigg(\sum_{i_1}w^2_{i_1ij}\Bigg)^2+\sum_{i_1}w^4_{i_1ij}\Bigg\}\Bigg]\\
&\quad=\sigma^2\sum_{i_1}\Bigg\{\sum_{i}\sum_{j}w_{i_1ij}g(t_j)\Bigg\}^2\\
&\quad\quad+O\Bigg[\{(nh)^{-1}+h^2\}\sum_{i}\sum_{i'}\sum_{j}\sum_{j'}(nh)^{-1}\\
&\quad\quad\quad+\sum_{i}\sum_{j}\sum_{j'}(nh)^{-1}+\sum_{i}\sum_{i'}\sum_{j}\sum_{j'}(nh)^{-2}\Bigg]\\
&\quad=nN^2\sigma^2(1-\pi)^2\int\Bigg\{d_s^{-1}d_{p\setminus s}g(t-m)-\int g(t-m)d_{p\setminus s}\Bigg\}^2d_s\\
&\quad\quad+o(n^3)+O\big[n^4\{(nh)^{-2}+h^4\}\big].
\end{aligned}$$

Equations (A.2), (A.3), (A.4) and (A.10) together yield the claimed result. □

A2.1.2. Proof of Theorem 2.

(a) *Preliminaries*. Adopt notation from the proof of Theorem 1, and in addition define

$$\begin{aligned} \Delta^{i_1 i_2} &= \hat{m}_{1_2}(x_{i_2}) - \hat{m}_{i_1}(x_{i_1}) - E\{\hat{m}_{1_2}(x_{i_2}) - \hat{m}_{i_1}(x_{i_1})\}, \\ d^{i_1 i_2} &= E\{\hat{m}_{i_2}(x_{i_2}) - \hat{m}_{i_1}(x_{i_1})\} - \{m(x_{i_2}) - m(x_{i_1})\}, \\ t_i &= t - m(x_i), \qquad t^{i_1 i_2} = t_{i_2} - d^{i_1 i_2}, \\ w^{i_1 i_2 i_3} &= \{(n-1)h\}^{-1}\big[\hat{d}_{i_3}(x_{i_3})^{-1} K\{(x_{i_3} - x_{i_1})/h\} I(i_1 \neq i_3) \\ &\qquad - \hat{d}_{i_2}(x_{i_2})^{-1} K\{(x_{i_2} - x_{i_1})/h\} I(i_1 \neq i_2)\big]. \end{aligned}$$

In this notation,

$$\begin{aligned} \Delta^{i_2 i_3} &= \sum_{i_1} w^{i_1 i_2 i_3} \varepsilon_{i_1}, \\ n \sum_i \hat{G}\{t - \hat{m}_i(x_i)\} &= \sum_{i_1} \sum_{i_2} I\big(\varepsilon_{i_1} + \Delta^{i_1 i_2} \le t^{i_1 i_2}\big), \end{aligned}$$

$$\begin{aligned} \check{F}_1(t) - F(t) &= (1-\pi) n^{-1} \sum_i I(\varepsilon_i \le t_i) \\ &\quad + (nN)^{-1} \sum_i \sum_j I(\varepsilon_i + \Delta_{ij} \le t_{ij}) \\ &\quad - (1-\pi) n^{-2} \sum_{i_1} \sum_{i_2} I\big(\varepsilon_{i_1} + \Delta^{i_1, i_2} \le t^{i_1, i_2}\big) \\ &\quad - N^{-1} \sum_j I(\varepsilon_j \le t_j). \end{aligned} \tag{A.11}$$

(b) *Bias*. We know from part (b) of the proof of Theorem 1 that

$$\begin{aligned} &E\left[\sum_j \hat{G}\{t - \hat{m}(x_j)\}\right] - \sum_j G(t_j) \\ &\quad = \tfrac{1}{2} k_1 (1-\pi) N (nh)^{-1} I_1(p \setminus s) + \tfrac{1}{2} k_2 (1-\pi) N h^2 I_2(p \setminus s) \\ &\qquad + o\Big\{n\big[(nh)^{-1} + h^2\big]\Big\}. \end{aligned}$$

Similarly it may be shown that

$$\begin{aligned} &E\left[\sum_i \hat{G}\{t - \hat{m}_i(x_i)\}\right] - \sum_i G(t_i) \\ &\qquad = \tfrac{1}{2} k_1 h^{-1} I_1(s) + \tfrac{1}{2} k_2 n h^2 I_2(s) + o\Big\{n\big[(nh)^{-1} + h^2\big]\Big\}. \end{aligned}$$

The desired result follows on combining these two formulae.

(c) *Variance*. By (A.11)

$$\begin{aligned} \operatorname{var}\{\check{F}_1(t) - F(t)\} &= (1-\pi)^2 n^{-2} v_1 + (nN)^{-2} v_2 + (1-\pi)^2 n^{-4} v_3 \\ &\quad + 2(1-\pi)(n^2 N)^{-1} v_4 - 2(1-\pi)^2 n^{-3} v_5 \\ &\quad - 2(1-\pi)(n^3 N)^{-1} v_6 + N^{-2} v_7, \end{aligned} \tag{A.12}$$

where, writing $\alpha_n = n^4\{(nh)^{-2} + h^4\}$,

$$v_1 = \operatorname{var}\Big\{\sum_i I(\varepsilon_i \le t_i)\Big\} = nI_6(s) + o(n),$$

$$v_2 = \operatorname{var}\Big\{\sum_i \sum_j I(\varepsilon_i + \Delta_{ij} \le t_{ij})\Big\}$$

$$= nN^2(1-\pi)^2\big\{I_3(p \setminus s) + I_4(p \setminus s) - I_5(p \setminus s)^2\big\} + o(n^3) + O(\alpha_n),$$

$$v_3 = \operatorname{var}\Big\{\sum_{i_1} \sum_{i_2} I\big(\varepsilon_{i_1} + \Delta^{i_1 i_2} \le t^{i_1 i_2}\big)\Big\}$$

$$= n^3\big\{I_3(s) + I_4(s) - I_5(s)^2\big\} + o(n^3) + O(\alpha_n),$$

$$v_4 = \operatorname{cov}\Big\{\sum_i I(\varepsilon_i \le t_i), \sum_i \sum_j I(\varepsilon_i + \Delta_{ij} \le t_{ij})\Big\}$$

$$= \sum_i \operatorname{cov}\Big\{I(\varepsilon_i \le t_i), \sum_j I(\varepsilon_i + \Delta_{ij} \le t_{ij})\Big\} + o(n^2) + O\big(n^{-1}\alpha_n\big)$$

$$= nN(1-\pi)\{I_4(s, p \setminus s) - I_5(s)I_5(p \setminus s)\} + o(n^2) + O\big(n^{-1}\alpha_n\big),$$

$$v_5 = \operatorname{cov}\Big\{\sum_i I(\varepsilon_i \le t_i), \sum_{i_1} \sum_{i_2} I\big(\varepsilon_{i_1} + \Delta^{i_1 i_2} \le t^{i_1 i_2}\big)\Big\}$$

$$= n^2\big\{I_4(s) - I_5(s)^2\big\} + o(n^2) + O\big(n^{-1}\alpha_n\big),$$

$$v_6 = \operatorname{cov}\Big\{\sum_i \sum_j I(\varepsilon_i + \Delta_{ij} \le t_{ij}), \sum_{i_1} \sum_{i_2} I\big(\varepsilon_{i_1} + \Delta^{i_1 i_2} \le t^{i_1 i_2}\big)\Big\}$$

$$= n^2N(1-\pi)\{I_3(s, p \setminus s) + I_4(s, p \setminus s) - I_5(s)I_5(p \setminus s)\}$$
$$+o(n^3) + O(\alpha_n),$$

$$v_7 = \operatorname{var}\Big\{\sum_j I(\varepsilon_j \le t_j)\Big\}.$$

(The formula for v_2 was derived during the proof of Theorem 1, and the formulae for v_1 and v_6 may be obtained similarly.) The desired result follows on combining the estimates from (A.12) down. □

A2.2. Proof of Theorem 3.

Bias.

$$NE\big\{\hat{F}_2(t) - F(t)\big\} = \sum_i \sum_j v_{ij}\{H(x_i) - H(x_j)\}$$

$$= \sum_j \hat{d}(x_j)^{-1}(nh)^{-1} \sum_i K\{(x_j - x_i)/h\}\{H(x_i) - H(x_j)\}$$

$$= \sum_j d_s(x_j)^{-1} \tfrac{1}{2}k_2\beta(x_j)h^2 + o(nh^2 + h^{-1})$$

$$= \tfrac{1}{2}k_2(N - n)h^2 \int d_s^{-1} d_{p \setminus s}\beta + o(nh^2 + h^{-1}).$$

Variance.

$$
\begin{aligned}
\operatorname{var}\{\hat{F}_2(t) - F(t)\} &- N^{-2} \operatorname{var}\Big\{ \sum_j I(Y_j \le t)\Big\} \\
&= N^{-2} \sum_i \{G(t_i) - G(t_i)^2\}\Big(\sum_j v_{ij}\Big)^2 \\
&= (nN)^{-2}(N-n)^2 \sum_i \{G(t_i) - G(t_i)^2\}\{d_s(x_i)^{-1} d_{p\setminus s}(x_i)\}^2 + o(n^{-1}) \\
&= n^{-1}(1-\pi)^2 \int \{G(t-m) - G(t-m)^2\} d_s^{-1} d_{p\setminus s}^2 + o(n^{-1}).
\end{aligned}
$$

Bias.

$$
\begin{aligned}
E\{\breve{F}_2(t) - F(t)\} & \\
&= N^{-1} \sum_i \sum_j v_{ij}\{H(x_i) - H(x_j)\} \\
&\quad - (n^{-1} - N^{-1}) \sum_{i_1 \ne i_2} \sum v^{i_1 i_2}\{H(x_{i_1}) - H(x_{i_2})\} \\
&= N^{-1} \sum_j d_s(x_j)^{-1} \tfrac{1}{2} k_2 \beta(x_j) h^2 - (n^{-1} - N^{-1}) \sum_{i_2} d_s(x_{i_2})^{-1} \tfrac{1}{2} k_2 \beta(x_{i_2}) h^2 \\
&\quad + o\{(nh)^{-1} + h^2\} \\
&= \tfrac{1}{2} k_2 (1-\pi) h^2 \Big(\int d_s^{-1} d_{p\setminus s} \beta - \int \beta \Big) + o\{(nh)^{-1} + h^2\}.
\end{aligned}
$$

Variance.

$$
\begin{aligned}
\operatorname{var}\{\breve{F}_2(t) - F(t)\} &- N^{-2} \operatorname{var}\Big\{ \sum_j I(Y_j \le t)\Big\} \\
&= n^{-2} \sum_i \{G(t_i) - G(t_i)^2\}\Big\{(1-\pi) + \pi \sum_j v_{ij} - (1-\pi) \sum_{i_2 \ne i} v_{ii_2}\Big\}^2 \\
&= n^{-2}(1-\pi)^2 \sum_i \{G(t_i) - G(t_i)^2\}\{d_s(x_i)^{-1} d_{p\setminus s}(x_i)\}^2 + o(n^{-1}) \\
&= n^{-1}(1-\pi)^2 \int \{G(t-m) - G(t-m)^2\} d_s^{-1} d_{p\setminus s}^2 + o(n^{-1}). \ \square
\end{aligned}
$$

A3.2.2. Proof of expression for $\operatorname{var}(\breve{F}(t) - F(t))$, under model misspecification.

We have

$$
\begin{aligned}
\breve{F}(t) - F(t) = {} & (n^{-1} - N^{-1}) \sum_i I(\varepsilon_i \le t_i) \\
& + (nN)^{-1} \sum_i \sum_j I\{\varepsilon_i + \Delta(x_j - x_i) \le t_{ji}\} \\
& - (n^2 N)^{-1} (N - n) \sum_{i_1} \sum_{i_2} I\{\varepsilon_{i_1} + \Delta(x_{i_2} - x_{i_1}) \le t_{i_2 i_1}\} \\
& - N^{-1} \sum_j I(\varepsilon_j \le t_{ji}),
\end{aligned}
$$

whence it follows that

$$
\begin{aligned}
\operatorname{var}\{\breve{F}(t) - F(t)\} = {} & (n^{-1} - N^{-1})^2 v_1 + (nN)^{-2} v_2 + (n^2 N)^{-2} (N - n)^2 v_3 \\
& + 2(n^{-1} - N^{-1})(nN)^{-1} v_4 - 2(n^3 N^2)^{-1} (N - n)^2 v_5 \\
& - 2(n^3 N^2)^{-1} (N - n) v_6 + N^{-2} v_7,
\end{aligned}
$$

where

$$
v_1 = \operatorname{var}\Big\{ \sum_i I(\varepsilon_i \le t_i) \Big\} = \sum_i \{ G(t_i) - G(t_i)^2 \} = n J_6(S) + o(n),
$$

$$
\begin{aligned}
v_2 &= \operatorname{var}\Big\{ \sum_i \sum_j I\big(\varepsilon_i + \Delta(x_j - x_i) \le t_{ji}\big) \Big\} \\
&= \sum_i \sum_{j_1} \sum_{j_2} \{ G(t_{j,i}) \wedge G(t_{j_2 i}) - G(t_{j_1 i}) G(t_{j_2 i}) \} \\
&\quad + n^{-1} \sigma_x^{-2} \sigma^2 \Big\{ \sum_i \sum_j (x_j - x_i) g(t_{ji}) \Big\}^2 \\
&\quad - n^{-1} \sigma_x^{-2} \sum_{i_1 \ne i_2} \sum_{j_1} \sum_{j_2} \{ (x_{j_1} - x_{i_1})(x_{i_2} - \bar{x}) g(t_{j_1 i_1}) \gamma(t_{j_2 i_2}) \\
&\qquad\qquad + (x_{j_2} - x_{i_2})(x_{i_1} - \bar{x}) g(t_{j_2 i_2}) \gamma(t_{j_1 i_1}) \} + o(n^3) \\
&= n M^2 \big[J_2(p \setminus s, p \setminus s) - J_3(p \setminus s, p \setminus s) \\
&\qquad + \sigma_x^{-2} \sigma^2 J_1(p \setminus s)^2 - 2\sigma_x^{-2} J_1(p \setminus s) J_0(p \setminus s) \big] + o(n^3).
\end{aligned}
$$

Likewise,

$$
\begin{aligned}
v_3 &= \operatorname{var}\Big\{ \sum_{i_1} \sum_{i_2} J\big(\varepsilon_{i_1} + \Delta(x_{i_2} - x_{i_1}) \le t_{i_2 i_1}\big) \Big\} \\
&= n^3 \big[J_2(s, s) - J_3(s, s) + \sigma_x^{-2} \sigma^2 J_1(s)^2 \\
&\qquad - 2\sigma_x^{-2} J_1(s) J_0(s) \big] + o(n^3),
\end{aligned}
$$

$$
\begin{aligned}
v_4 &= \operatorname{cov}\left\{ \sum_i J(\varepsilon_i \le t_i),\ \sum_i \sum_j J\big(\varepsilon_i + \Delta(x_j - x_i) \le t_{ji}\big) \right\} \\
&= \sum_i \sum_j \big[G(t_i) \wedge G(t_{ji}) - G(t_i) G(t_{ji}) \big] \\
&\qquad - n^{-1} \sigma_2^{-2} \sum_{i \ne k} \sum \sum_j (x_j - x_k)(x_i - \bar{x}) g(t_{jk}) \gamma(t_i) + o(n^2) \\
&= Mn\big[J_5(s, p \setminus s) - J_4(s, p \setminus s) - \sigma_x^{-2} J_1(p \setminus s) J_{00}(s) \big] + o(n^2), \\
v_5 &= \operatorname{cov}\left\{ \sum_i J(\varepsilon_i \le t_i),\ \sum_{i_1} \sum_{i_2} J\big(\varepsilon_{i_1} + \Delta(x_{i_2} - x_{i_1}) \le t_{i_2 i_1}\big) \right\} \\
&= n^2 \big[J_5(s, s) - J_4(s, s) - \sigma_x^{-2} J_1(s) J_{00}(s) \big] + o(n^2), \\
v_6 &= \operatorname{cov}\left\{ \sum_i \sum_j J\big(\varepsilon_i + \Delta(x_j - x_i) \le t_{ji}\big),\ \sum_{i_1} \sum_{i_2} J\big(\varepsilon_{i_1} + \Delta(x_{i_2} - x_{i_1}) \le t_{i_2 i_1}\big) \right\} \\
&= n^2 M \big\{ J_2(p \setminus s, s) - J_3(p \setminus s, s) + \sigma_x^{-2} \sigma^2 J_1(p \setminus s) J_1(s) \\
&\qquad\qquad - \sigma_x^{-2} \big[J_1(p \setminus s) J_0(s) + J_1(s) J_0(p \setminus s) \big] \big\} + o(n^3), \\
v_7 &= \operatorname{var}\left\{ \sum_j I(\varepsilon_j \le t_j) \right\} = N(1 - \pi) J_6(p \setminus s) + o(n).
\end{aligned}
$$

These combined give the result.

A3.3.2. Proof of expression for var($\hat{F}_3(t) - F(t)$).

$$
N^2 \operatorname{var}\big(\hat{F}_3(t) - F(t)\big) = v_1 + v_2 + v_3 - 2v_4 + 2v_5 - 2v_6 + v_7,
$$

where, letting $\phi_i = \sum_j v_{ij}$,

$$
\begin{aligned}
v_1 &= \operatorname{var}\left(\sum_j \hat{H}(x_j) \right) = \operatorname{var}\left(\sum_i \phi_i I(y_i \le t) \right) \\
&= n^{-1} (1 - \pi)^2 J_8(p \setminus s) + o(n^{-1}) \quad \text{as in the proof of Theorem 3}, \\
v_2 &= \operatorname{var}\left(-n^{-1} \sum_j \sum_i v_{ij} \sum_k I\big(\varepsilon_k + \Delta(x_i - x_k) \le t_{ik}\big) \right) \\
&= n^{-2} \sum_{i_1} \sum_{i_2} \sum_{k_1} \sum_{k_2} \varphi_{i_1} \varphi_{i_2} \big[P\big(\varepsilon_{k_1} + \Delta(x_{i_1} - x_{k_1}) \le t_{i_1 k_1}, \\
&\qquad\qquad \varepsilon_{k_2} + \Delta(x_{i_2} - x_{k_2}) \le t_{i_2 k_2}\big) \\
&\qquad\qquad - P\big(\varepsilon_{k_1} + \Delta(x_{i_1} - x_{k_1}) \le t_{i_1 k_1}\big) \\
&\qquad\qquad \times P\big(\varepsilon_{k_2} + \Delta(x_{i_2} - x_{k_2}) \le t_{i_2 k_2}\big) \big]
\end{aligned}
$$

$$
\begin{aligned}
&= n^{-2} \sum_k \sum_{i_1} \sum_{i_2} \varphi_{i_1}\varphi_{i_2}\big[G(t_{i,k}) \wedge G(t_{i_2k}) - G(t_{i_1k})G(t_{i_2k})\big] \\
&\quad + n^{-3}\sigma_x^{-2}\sigma^2\bigg\{\sum_i \sum_k \varphi_i(x_i - x_k)g(t_{ik})\bigg\}^2 \\
&\quad - n^{-3}\sigma_x^{-2} \sum_{i_1}\sum_{i_2}\sum_{k_1}\sum_{k_2} \varphi_{i_1}\varphi_{i_2}\big\{(x_{i_1} - x_{k_1})(x_{k_2} - \bar{x})g(t_{i_1k_1})\gamma(t_{i_2k_2}) \\
&\qquad\qquad + (x_{i_2} - x_{k_2})(x_{k_1} - \bar{x})g(t_{i_2k_2})\gamma(t_{i_1k_1})\big\} + o(n) \\
&= M^2 n^{-4} \sum_k \sum_{i_1} \sum_{i_2} d_{p\setminus s}(x_{i_1})d_s(x_{i_1})^{-1}d_{p\setminus s}(x_{i_2})d_s(x_{i_2})^{-1} \\
&\qquad \times\big[G(t_{i_1k} \wedge Gt_{i_2k}) - G(t_{i_1}k)G(t_{i_2k})\big] \\
&\qquad + n^{-5}M^2\sigma_x^{-2}\sigma^2\bigg\{\sum_i\sum_k d_{p\setminus s}(x_i)d_s(x_i)^{-1}(x_i - x_k)g(t_{ik})\bigg\}^2 \\
&\quad - M^2 n^{-5}\sigma_x^{-2}\sum_{i_1}\sum_{i_2}\sum_{k_1}\sum_{k_2}\Big\{d_{p\setminus s}(x_{i_1})d_s(x_{i_1})^{-1}d_{p\setminus s}(x_{i_2})d_s(x_{i_2})^{-1} \\
&\qquad \times\big[(x_{i_1} - x_{k_1})(x_{k_2} - \bar{x})g(t_{i_1i_1})\gamma(t_{i_1k_2}) \\
&\qquad + (x_{i_2} - x_{k_2})(x_{k_1} - \bar{x})g(t_{i_2k_2})\gamma(t_{i_1k_1})\big]\Big\} + o(n) \\
&= M^2 n^{-1}\big\{J_2(p \setminus s, p \setminus s) - J_3(p \setminus s, p \setminus s) + \sigma_x^{-2}\sigma^2 J_1^2(p \setminus s) \\
&\qquad - 2\sigma_x^{-2}J_1(p \setminus s)J_0(p \setminus s)\big\} + o(n),
\end{aligned}
$$

$$
\begin{aligned}
v_3 &= \operatorname{var}\bigg(n^{-1}\sum_j\sum_k I\big(\varepsilon_k + \Delta(x_j - x_k) \le t_{jk}\big)\bigg) \\
&= J_2(p \setminus s, p \setminus s) - J_3(p \setminus s, p \setminus s) + \sigma_x^{-2}\sigma^2 J_1^2(p \setminus s) \\
&\quad - 2\sigma_x^{-2}J_1(p \setminus s)J_0(p \setminus s) + o(n),
\end{aligned}
$$

$$
\begin{aligned}
v_4 &= n^{-1}\operatorname{cov}\bigg\{\sum_{k_1}\varphi_{k_1}I(y_{k_1} \le t),\ \sum_i\sum_k \varphi_i I\big(\varepsilon_k + \Delta(x_i - x_k) \le t_{ik}\big)\bigg\} \\
&= n^{-1}\sum_k\sum_i \operatorname{cov}\big\{\varphi_k I(y_k \le t),\ \varphi_i I\big(\varepsilon_k + \Delta(x_i - x_k) \le t_{ik}\big)\big\} \\
&\quad + n^{-1}\sum_{k_1 \ne k}\sum\sum_i \operatorname{cov}\big\{\varphi_{k_1}I(y_{k_1} \le t),\ \varphi_i I\big(\varepsilon_k + \Delta(x_i - x_k) \le t_{ik}\big)\big\} \\
&= n^{-3}M^2\sum_k\sum_i\Big\{d_{p\setminus s}(x_k)d_s(x_k)^{-1}d_{p\setminus s}(x_i)d_s(x_i)^{-1} \\
&\qquad \times\big[G(t_k) \wedge G(t_{ik}) - G(t_k)G(t_{ik})\big]\Big\} \\
&\quad - n^{-4}M^2\sigma_x^{-2}\sum_{k_1}\sum_k\sum_i\Big\{d_{p\setminus s}(x_{k_1})d_s(x_{k_1})^{-1}d_{p\setminus s}(x_i)d_s^{-1}(x_i) \\
&\qquad \times(x_i - x_k)(x_{k_1} - \bar{x})g(t_{ik})\gamma(t_{k_1})\Big\} + o(n)
\end{aligned}
$$

$$= M^2 n^{-1}\{J_5(p \setminus s, p \setminus s) - J_4(p \setminus s, p \setminus s) - \sigma_x^2 J_{00}(p \setminus s) J_1(p \setminus s)\} + o(n),$$

$$v_5 = n^{-1} \operatorname{cov}\left\{ \sum_i \varphi_i I(y_i \le t), \sum_j \sum_k I\left(\varepsilon_k + \Delta(x_j - x_k) \le t_{jk}\right)\right\}$$

$$= M^2 n^{-1}\{J_5(p \setminus s, p \setminus s) - J_4(p \setminus s, p \setminus s) - \sigma_x^{-2} J_{00}(p \setminus s) J_1(p \setminus s)\} + o(n),$$

$$v_6 = n^2 \operatorname{cov}\left\{ \sum_i \sum_k \varphi_i I(\varepsilon_k + \Delta(x_i - x_k) \le t_{ik}), \sum_j \sum_{k_1} I\left(\varepsilon_{k_1} + \Delta(x_j - x_{k_1}) \le t_{jk_1}\right)\right\}$$

$$= M^2 n^{-1}\{J_2(p \setminus s, p \setminus s) - J_3(p \setminus s, p \setminus s) + \sigma_x^{-2}\sigma^2 J_1^2(p \setminus s) - 2\sigma_x^{-2} J_1(p \setminus s) J_0(p \setminus s)\} + o(n)$$

and

$$v_7 = \operatorname{var}\left(\sum_j I(y_j \le t_j)\right) = N(1 - \pi) J_6(p \setminus s) + o(n).$$

Combining terms gives the result. □

Acknowledgment. The authors thank R. L. Chambers for helpful comments.

REFERENCES

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

Chambers, R. L., Dorfman, A. H. and Wehrly, T. (1993). Bias robust estimation in finite populations using nonparametric calibration. *J. Amer. Statist. Assoc.* **88** 268–277.

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

Dorfman, A. H. (1993). A comparison of design-based and model-based estimators of the finite population distribution function. *Austral. J. Statist.* To appear.

Kuo, L. (1988). Classical and prediction approaches to estimating distribution functions from survey data. In *Proc. Survey Research Methods Section* 280–285. Amer. Statist. Assoc., Alexandria, VA.

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

Office of Survey Methods Research
Bureau of Labor Statistics
Washington, D.C. 20212

Department of Statistics
Australian National University
GPO Box 4, Canberra
ACT 2601
Australia