Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

A comparison between nonparametric estimators for finite population distribution functions

by Leo Pasquazzi and Lucio de Capitani

Release date: June 22, 2016

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

email at STATCAN.infostats-infostats.STATCAN@canada.ca

telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following toll-free numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-877-287-4369

Depository Services Program

- Inquiries line 1-800-635-7943
- Fax line 1-800-565-7757

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Standard table symbols

The following symbols are used in Statistics Canada publications:

. not available for any reference period
.. not available for a specific reference period
... not applicable
0 true zero or a value rounded to zero
0s value rounded to 0 (zero) where there is a meaningful distinction between true zero and the value that was rounded
p preliminary
r revised
x suppressed to meet the confidentiality requirements of the *Statistics Act*
E use with caution
F too unreliable to be published
* significantly different from reference category ($p < 0.05$)

An HTML version is also available.

Cette publication est aussi disponible en français.

Survey Methodology, June 2016
Vol. 42, No. 1, pp. 87-120
Statistics Canada, Catalogue No. 12-001-X

A comparison between nonparametric estimators for finite population distribution functions

Leo Pasquazzi and Lucio de Capitani[1]

Abstract

In this work we compare nonparametric estimators for finite population distribution functions based on two types of fitted values: the fitted values from the well-known Kuo estimator and a modified version of them, which incorporates a nonparametric estimate for the mean regression function. For each type of fitted values we consider the corresponding model-based estimator and, after incorporating design weights, the corresponding generalized difference estimator. We show under fairly general conditions that the leading term in the model mean square error is not affected by the modification of the fitted values, even though it slows down the convergence rate for the model bias. Second order terms of the model mean square errors are difficult to obtain and will not be derived in the present paper. It remains thus an open question whether the modified fitted values bring about some benefit from the model-based perspective. We discuss also design-based properties of the estimators and propose a variance estimator for the generalized difference estimator based on the modified fitted values. Finally, we perform a simulation study. The simulation results suggest that the modified fitted values lead to a considerable reduction of the design mean square error if the sample size is small.

Key Words: Finite population sampling; Distribution function estimator; Fitted values; Kuo estimator.

1 Introduction

Since Chambers and Dunstan's seminal paper Chambers and Dunstan (1986), several estimators for finite population distribution functions have been proposed. Most of them are based either on different types of fitted values or on different ways to combine them into an estimator. The estimator proposed by Chambers and Dunstan (1986), for example, is based on fitted values derived from a superpopulation model where the study variable and an auxiliary variable are linked by a linear regression model with independent error components whose variances are assumed to be known. Substituting the fitted values to the unobserved indicator functions in the definition of the population distribution function of the study variable yields the Chambers and Dunstan estimator. Rao, Kovar and Mantel (1990) incorporate design weights into the fitted values of Chambers and Dunstan and use them in a generalized difference estimator. Kuo (1988) uses nonparametric regression to estimate directly the regression relationship between the indicator functions and the auxiliary variable and obtains fitted values that accommodate virtually any superpopolation model. Like Chambers and Dunstan, she substitutes the unobserved indicator functions with their corresponding fitted values and obtains a model-based estimator. Chambers, Dorfman and Wehrly (1993) combine the fitted values of Chambers and Dunstan (1986) and of Kuo (1988) and propose still another model-based estimator that aims to be more efficient than the Kuo estimator if the linear superpopulation model assumed by Chambers and Dunstan is true, and that does not suffer from model misspecification bias otherwise. Following these early works there has been quite a large number of subsequent proposals with the aim to achieve some gain in efficiency with respect to the Horvitz-Thompson estimator, while preserving robustness and sometimes also one or both of the following desirable properties shared by the Horvitz-Thompson estimator: (i) the fact that it is a linear combination of the sample indicator functions with

1. Leo Pasquazzi and Lucio de Capitani, Università degli Studi di Milano-Bicocca, Milan, Italy. E-mail: leo.pasquazzi@unimib.it, lucio.decapitani1@unimib.it.

coefficients that do not depend on the study variable and (ii) the fact that it gives always rise to nondecreasing estimates for the distribution function.

The present work originates from the idea to improve upon the fitted values proposed by Kuo (1988) through incorporation of an estimate for the mean regression function (see Section 2). This idea has been put forward in a recent textbook of Chambers and Clark (2012) and it is based on the assumption of an underlying superpopulation model with smooth regression relationship between the study variable and an auxiliary variable and with smoothly varying error component distributions. According to this idea, the fitted values are the outcome of a two-step procedure: at the first step the mean regression function is estimated through either parametric or nonparametric regression, and at the second step, using the regression residuals from the first step, the distribution functions of the error components are estimated using nonparametric regression in order to accommodate the possibility of smoothly varying error component distributions. Combining both estimates one may compute fitted values for the indicator functions in the definition of the finite population distribution function of the study variable. Chambers and Clark (2012) analyze the model-based estimator that is obtained by substituting the unobserved indicator functions by their corresponding fitted values and they sketch a proof that leads to an expression for the model variance of the resulting estimator. In that proof they assume that the mean regression function is estimated by a consistent estimator and that the contribution from its estimation error to the model variance of the final distribution function estimator can be neglected. In the present work we consider local linear regression for estimating both the model mean regression function and the error component distributions. We provide asymptotic expansions for the model bias and the model variance of the resulting estimator and compare them with those corresponding to the Kuo estimator based on local linear regression. It turns out that the leading terms in the model variances are the same and that, for appropriately chosen bandwidth sequences, the squared model bias of both estimators goes to zero faster than the model variance. To establish which estimator is asymptotically more efficient from the model-based perspective thus requires knowledge of the second order terms of the model variances. The latter however depend on more specific assumptions than those considered in the present work and, at least for the estimator based on the modified fitted values, it seems no easy task to determine the second order terms of the model variances. Which estimator is more efficient from the model-based perspective remains thus an open question.

In addition to the above model-based estimators, we analyze also the generalized difference estimators based on both types of fitted values in their design weighted versions. The results in Section 3 show that the convergence rates of their model biases and their model variances are the same as those of their model-based counterparts. As for design-based properties, they are discussed to some extent in Section 4 along with the issue of variance estimation. It would of course be of interest to derive and compare asymptotic expansions for the design biases and the design variances. Breidt and Opsomer (2000) derive under mild conditions a general expression for the first order term in the design mean square error of local polynomial regression estimators, of which the generalized difference estimator based on the fitted values of Kuo is a special case. The generalized difference estimator based on the modified fitted values does however not fall into this class. In line with Särndal, Swensson and Wretman (1992), we conjecture that under broad conditions the first order term of its design mean square error is the same as the one of the generalized difference estimator based on the fitted values of Kuo. Formal proofs could perhaps be obtained by adapting and extending some of the results in Wang and Opsomer (2011). To test this conjecture and to compare the performance of the generalized difference and the model-based estimators in various settings, we performed a simulation study whose results are presented in Section 5.

2 Definition of the estimators

Let (y_i, x_i) denote the values taken on by a study variable Y and an auxiliary variable X on unit i of a finite population $U := \{1, 2, \ldots, N\}$. Suppose that

$$y_i = m(x_i) + \varepsilon_i, \qquad i \in U, \tag{2.1}$$

where $m(x)$ is a smooth function and where the ε_i's are independent zero mean random variables whose distribution functions $P(\varepsilon_i \le \varepsilon) = G(\varepsilon \mid x_i)$ depend smoothly on x_i. Let $s \subset U$ be a sample chosen from the population U according to some sample design. As usual in the context of complete auxiliary information we assume that the x_i – values are known for all population units, while the y_i – values are observed only for the population units which belong to the sample s.

To estimate the unknown population distribution function

$$F_N(t) := \frac{1}{N} \sum_{i \in U} I(y_i \le t),$$

Kuo (1988) proposes the estimator given by

$$\hat{F}(t) := \frac{1}{N} \left(\sum_{j \in s} I(y_j \le t) + \sum_{i \notin s} \sum_{j \in s} w_{i,j} I(y_j \le t) \right), \tag{2.2}$$

where in place of $w_{i,j}$ she suggests to use either the local constant regression weights

$$w_{i,j} := \frac{K\left(\frac{x_i - x_j}{\lambda}\right)}{\sum_{k \in s} K\left(\frac{x_i - x_k}{\lambda}\right)}$$

with some (integrable) kernel function in place of $K(u)$ and $\lambda > 0$, or the nearest k neighbor weights

$$w_{i,j} := \begin{cases} 1/k, & \text{if } x_j \text{ is one of the } k \text{ nearest neighbors to } x_i \\ 0, & \text{otherwise.} \end{cases}$$

Note that in the definition $\hat{F}(t)$,

$$\hat{G}_i(t) := \sum_{j \in s} w_{i,j} I(y_j \le t) \tag{2.3}$$

is used as the fitted value in place of the unobserved indicator function $I(y_i \le t)$ for $i \notin s$.

Following an idea put forward in the textbook of Chambers and Clark (2012), we shall analyze an estimator for $F_N(t)$ based on alternative fitted values which incorporate a nonparametric estimate for the mean regression function $m(x)$. The fitted values in question are given by

$$\hat{G}_i^*(t) := \sum_{j \in s} w_{i,j} I(y_j - \hat{m}_j \le t - \hat{m}_i) \tag{2.4}$$

where

$$\hat{m}_i := \sum_{k \in s} w_{i,j} y_j$$

is a nonparametric estimator for $m(x)$ at $x = x_i$, and the resulting estimator for $F_N(t)$ is given by

$$\hat{F}^*(t) := \frac{1}{N}\left(\sum_{j \in s} I(y_j \le t) + \sum_{i \notin s}\sum_{j \in s} w_{i,j} I(y_j - \hat{m}_j \le t - \hat{m}_i)\right). \tag{2.5}$$

The fitted values in (2.3) and (2.4), or appropriately modified versions of them which include sample inclusion probabilities in the regression weights $w_{i,j}$, can obviously be computed also for $i \in s$, and they can be employed for example in generalized difference estimators (Särndal et al. 1992, page 221) or in model calibrated estimators (see for example Wu and Sitter 2001; Chen and Wu 2002; Wu 2003; Montanari and Ranalli 2005; Rueda, Martínez, Martínez and Arcos 2007; Rueda, Sànchez-Borrego, Arcos and Martínez 2010). In addition to the model-based estimators in (2.2) and (2.5), we shall thus consider also the generalized difference estimators given by

$$\tilde{F}(t) := \frac{1}{N}\left(\sum_{i \in U}\sum_{j \in s} \tilde{w}_{i,j} I(y_j \le t)\right) + \sum_{i \in s} \pi_i^{-1}\left(I(y_i \le t) - \sum_{j \in s} \tilde{w}_{i,j} I(y_j \le t)\right)$$

and by

$$\tilde{F}^*(t) := \frac{1}{N}\left(\sum_{i \in U}\sum_{j \in s} \tilde{w}_{i,j} I(y_j - \tilde{m}_j \le t - \tilde{m}_i)\right) + \sum_{i \in s} \pi_i^{-1}\left(I(y_i \le t) - \sum_{j \in s} \tilde{w}_{i,j} I(y_j - \tilde{m}_j \le t - \tilde{m}_i)\right)$$

where π_i denotes the first order sample inclusion probabilities, $\tilde{w}_{i,j}$ denotes design weighted regression weights whose definition is given below, and $\tilde{m}_i := \sum_{k \in s} \tilde{w}_{i,k} y_k$. Note that $\tilde{F}(t)$ and $\tilde{F}^*(t)$ are based on design weighted counterparts of the fitted values $\hat{G}_i(t)$ and $\hat{G}_i^*(t)$ which are given by

$$\tilde{G}_i(t) := \sum_{j \in s} \tilde{w}_{i,j} I(y_j \le t)$$

and

$$\tilde{G}_i^*(t) := \sum_{j \in s} \tilde{w}_{i,j} I(y_j - \tilde{m}_j \le t - \tilde{m}_i),$$

respectively.

As for the regression weights $w_{i,j}$ and $\tilde{w}_{i,j}$, in the present work we consider local linear regression weights in their place. In what follows $w_{i,j}$ and $\tilde{w}_{i,j}$ are thus defined by

$$w_{i,j} := \frac{1}{n\lambda} K\left(\frac{x_i - x_j}{\lambda}\right) \frac{M_{2,s}(x_i) - \left(\frac{x_i - x_j}{\lambda}\right) M_{1,s}(x_i)}{M_{2,s}(x_i) M_{0,s}(x_i) - M_{1,s}^2(x_i)}$$

and

$$\tilde{w}_{i,j} := \frac{1}{\pi_j n\lambda} K\left(\frac{x_i - x_j}{\lambda}\right) \frac{\tilde{M}_{2,s}(x_i) - \left(\frac{x_i - x_j}{\lambda}\right) \tilde{M}_{1,s}(x_i)}{\tilde{M}_{2,s}(x_i) \tilde{M}_{0,s}(x_i) - \tilde{M}_{1,s}^2(x_i)},$$

where n is the number of units in the sample s,

$$M_{r,s}(x) := \sum_{k \in s} \frac{1}{n\lambda} K\left(\frac{x - x_k}{\lambda}\right)\left(\frac{x - x_k}{\lambda}\right)^r, \qquad r = 0,1,2,$$

and

$$\tilde{M}_{r,s}(x) := \sum_{k \in s} \frac{1}{\pi_k n\lambda} K\left(\frac{x - x_k}{\lambda}\right)\left(\frac{x - x_k}{\lambda}\right)^r, \qquad r = 0,1,2.$$

It is worth noting that the nonparametric estimators of this section are not well-defined if the regression weights $w_{i,j}$ and $\tilde{w}_{i,j}$ included in their definitions are not well-defined. This problem occurs for example when the support of the kernel function $K(u)$ is given by the interval $[-1,1]$ (e.g., uniform kernel, Epanechnikov kernel), and when there are not at least two $j \in s$ such that $|x_i - x_j| < \lambda.$ To overcome this problem one can use a kernel function whose support is given by the whole real line (e.g., Gaussian kernel) or choose the bandwidth adaptively. The latter solution may also lead to more efficient estimators (see e.g., Fan and Gijbels 1992). With reference to the estimators $\hat{F}^*(t)$ and $\tilde{F}^*(t)$ based on the modified fitted values, it is moreover worth noting that one could in principle apply different bandwidths and/or regression weights to the y_i – values and to the indicator functions. For the sake of simplicity, in the present work we shall consider neither adaptive bandwidth selection nor the possibility of different regression weights to estimate the mean regression function and the distributions of the error components.

Comparing the definitions of the estimators based on the two types of fitted values, it becomes immediately obvious that $\hat{F}(t)$ and $\tilde{F}(t)$ are easier to compute since they are linear combinations of the observed indicator functions $I(y_j \le t).$ The coefficients of these linear combinations do not depend on the study variable Y and they can therefore be used to estimate averages of other functions than indicator functions, or of functions of several study variables, in particular when there are reasons to believe that the latter are related to the auxiliary variable X. This fact is of particular value to practitioners who want estimates related to several study variables to be consistent with one another. However, there is a strong argument in favor of the estimators $\hat{F}^*(t)$ and $\tilde{F}^*(t)$ based on the modified fitted values too: if $y_i = a + bx_i$ for all $i \in U,$ then it follows that $\hat{F}^*(t) = \tilde{F}^*(t) = F_N(t)$ for every sample s such that the estimators are well-defined. One would therefore expect that $\hat{F}^*(t)$ and $\tilde{F}^*(t)$ be more efficient than $\hat{F}(t)$ and $\tilde{F}(t)$ when there is a strong regression relationship between Y and X.

3 Model-based properties

In this section we provide asymptotic expansions for the model bias and the model variance of the estimators introduced in the previous section. The expansions are based on the following assumptions:

(C1) $N \to \infty$ and the sequence of population x_i – values and of sample designs are such that

$$H_{N,s}(x) := \frac{1}{n} \sum_{i \in s} I(x_i \le x)$$

and

$$H_{N,\bar{s}}(x) := \frac{1}{N-n} \sum_{i \notin s} I(x_i \le x)$$

converge to absolutely continuous distribution functions $H_s(x) := \int_a^x h_s(z)\,dz$ and $H_{\bar{s}}(x) := \int_a^x h_{\bar{s}}(z)\,dz$, respectively. The support of $H_s(x)$ and $H_{\bar{s}}(x)$ is given by a bounded interval $[a,b]$ and the density functions $h_s(x)$ and $h_{\bar{s}}(x)$ have bounded first derivatives for $x \in (a,b)$. $h_s(x)$ is bounded away from zero.

(C2) The kernel function $K(u)$ is symmetric, has support on $[-1,1]$ and has bounded derivative for $u \in (-1,1)$. The bandwidth sequence λ goes to zero slow enough to make sure that

$$\alpha := \max\left\{ \sup_{x\in[a,b]} \left| H_{N,s}(x) - H_s(x) \right|, \sup_{x\in[a,b]} \left| H_{N,\bar{s}}(x) - H_{\bar{s}}(x) \right| \right\}$$

is of order $o(\lambda)$.

(C3) The population y_i – values are generated from model (2.1). The function $m(x)$ is such that

$$\left| m(x) - m(x_0) - m'(x_0)(x - x_0) - \frac{1}{2} m''(x_0)(x - x_0)^2 \right| \le C \left| x - x_0 \right|^{2+\delta}$$

for some $\delta > 0$, and the family of error component distribution functions $G(\varepsilon \mid x)$ is such that

$$\begin{aligned} &\Big| G(\varepsilon|x) - G(\varepsilon_0|x_0) - G^{(1,0)}(\varepsilon_0|x_0)(\varepsilon - \varepsilon_0) - G^{(0,1)}(\varepsilon_0|x_0)(x - x_0) \\ &- \frac{1}{2}\Big(G^{(2,0)}(\varepsilon_0|x_0)(\varepsilon - \varepsilon_0)^2 + 2G^{(1,1)}(\varepsilon_0|x_0)(\varepsilon - \varepsilon_0)(x - x_0) + G^{(0,2)}(\varepsilon_0|x_0)(x - x_0)^2 \Big) \Big| \\ &\le C\left(|\varepsilon - \varepsilon_0|^{2+\delta} + |x - x_0|^{2+\delta} \right) \end{aligned}$$

for some $C > 0$ and some $\delta > 0$, where

$$G^{(r,s)}(\varepsilon|x) := \partial^{r+s} G(\varepsilon|x) \big/ \left(\partial\varepsilon^r \partial x^s \right) \quad \text{for} \quad r,s = 0,1,2.$$

Assumption (C1) poses a restriction on how the sample and nonsample x_i – values are generated. Together with assumption (C2) it makes sure that the estimation errors of the kernel density estimators for $h_s(x)$ and $h_{\bar{s}}(x)$ go to zero uniformly for $x \in [a+\lambda, b-\lambda]$ and that they are uniformly bounded for $x \in [a,b]$. Replacing (C1) by more specific assumptions may allow for relaxing (C2) and for improving the uniform convergence rate for the estimation error of the kernel density estimators (see for example the results in Hansen 2008). Assumption (C3) is finally needed to make sure that the model mean square errors of the two estimators converge to zero. It can be relaxed at the cost of slowing down the convergence rates. In addition to assumptions (C1) to (C3) we shall also need the following assumption (C4) to make sure that the model mean square errors of the generalized difference estimators go to zero:

(C4) The first order sample inclusion probabilities are given by

$$\pi_i := n^* \frac{\pi(x_i)}{\sum_{j\in U} \pi(x_j)}, \qquad i \in U,$$

where n^* is the expected sample size and $\pi(x)$ is a function which is bounded away from zero and has bounded first derivative for $x \in (a, b)$.

Proposition 1. *Under assumptions (C1) to (C3) it follows that:*

$$\begin{aligned} E\left(\hat{F}(t) - F_N(t)\right) &= \lambda^2 \frac{N-n}{N} \frac{\mu_2}{2\mu_0} \int_a^b \Big[G^{(2,0)}(t - m(x) \mid x)(m'(x))^2 - G^{(1,0)}(t - m(x) \mid x) m''(x) \\ &\quad - 2G^{(1,1)}(t - m(x) \mid x) m'(x) + G^{(0,2)}(t - m(x) \mid x) \Big] h_{\bar{s}}(x)\, dx + o\left(\lambda^2\right) \end{aligned}$$

and

$$\begin{aligned} \operatorname{var}\left(\hat{F}(t) - F_N(t)\right) &= \frac{1}{n}\left(\frac{N-n}{N}\right)^2 \int_a^b \Big[G(t - m(x) \mid x) - G^2(t - m(x) \mid x) \Big] \Big[h_{\bar{s}}(x) / h_s(x) \Big] h_{\bar{s}}(x)\, dx \\ &\quad + \frac{1}{N-n}\left(\frac{N-n}{N}\right)^2 \int_a^b \Big[G(t - m(x) \mid x) - G^2(t - m(x) \mid x) \Big] h_{\bar{s}}(x)\, dx + o\left(n^{-1}\right), \end{aligned}$$

where $\mu_r := \int_{-1}^{-1} K(u) u^r du$ *for* $r = 0, 1, 2.$

Adding assumption (C4) it can be shown that

$$\begin{aligned} E\left(\tilde{F}(t) - F_N(t)\right) &= \lambda^2 \frac{N-n}{N} \frac{\mu_2}{2\mu_0} \int_a^b \Big[G^{(2,0)}(t - m(x) \mid x)(m'(x))^2 - G^{(1,0)}(t - m(x) \mid x) m''(x) \\ &\quad - 2G^{(1,1)}(t - m(x) \mid x) m'(x) + G^{(0,2)}(t - m(x) \mid x) \Big] h(x)\, dx + o\left(\lambda^2\right), \end{aligned}$$

where

$$h(x) := h_{\bar{s}}(x) + \left(1 - \pi^{-1}(x)\right) h_s(x),$$

and it can be shown that

$$\operatorname{var}\left(\tilde{F}(t) - F_N(t)\right) = \operatorname{var}\left(\hat{F}(t) - F_N(t)\right) + o\left(n^{-1}\right).$$

Proposition 2. *Under assumptions (C1) to (C3) and assuming that*

i) *the function*

$$\sigma^2(x) := \int_{-\infty}^{\infty} \varepsilon^2 dG(\varepsilon \mid x)$$

has bounded first derivative for $x \in (a, b)$

ii)

$$\sup_{x \in [a,b]} \int_{-\infty}^{\infty} \varepsilon^4 dG(\varepsilon \mid x) < \infty,$$

it can be shown that

$$E\left(\hat{F}^*(t)-F_N(t)\right) = \lambda^2 \frac{N-n}{N}\frac{\mu_2}{\mu_0}\int_a^b G^{(0,2)}(t-m(x)|x)h_{\bar{s}}(x)dx + \frac{1}{n\lambda}\frac{N-n}{N}\left[\frac{K(0)-\kappa}{\mu_0}\int_a^b G^{(1,0)}(t-m(x)|x)(t-m(x))h_s^{-1}(x)h_{\bar{s}}(x)dx + \frac{\kappa-\theta}{\mu_0^2}\int_a^b G^{(2,0)}(t-m(x)|x)\sigma^2(x)h_s^{-1}(x)h_{\bar{s}}(x)dx\right] + o\left(\lambda^2+(n\lambda)^{-1}\right),$$

where $\kappa := \int_{-1}^1 K^2(u)du$ *and* $\theta := \int_{-1}^1 K(v)\int_{-1}^1 K(u+v)K(u)dudv,$ *and it can be shown that*

$$\text{var}\left(\hat{F}^*(t)-F_N(t)\right) = \text{var}\left(\hat{F}(t)-F_N(t)\right) + o\left(n^{-1}+\lambda^5\right).$$

Adding assumption (C4) it can also be shown that

$$E\left(\tilde{F}^*(t)-F_N(t)\right) = \lambda^2 \frac{N-n}{N}\frac{\mu_2}{\mu_0}\int_a^b G^{(0,2)}(t-m(x)|x)h(x)dx + \frac{1}{n\lambda}\frac{N-n}{N}\left[\frac{K(0)-\kappa}{\mu_0}\int_a^b G^{(1,0)}(t-m(x)|x)(t-m(x))h_s^{-1}(x)h(x)dx + \frac{\kappa-\theta}{\mu_0^2}\int_a^b G^{(2,0)}(t-m(x)|x)\sigma^2(x)h_s^{-1}(x)h(x)dx\right] + o\left(\lambda^2+(n\lambda)^{-1}\right)$$

and that

$$\text{var}\left(\tilde{F}^*(t)-F_N(t)\right) = \text{var}\left(\hat{F}(t)-F_N(t)\right) + o\left(n^{-1}+\lambda^5\right).$$

The proofs of the Propositions are given in the Appendix. Dorfman and Hall (1993) derived similar expansions for the Kuo estimator with local constant regression weights instead of local linear ones.

Note that in view of the asymptotic expansions it is possible to choose bandwidth sequences λ in such a way as to make sure that the squares of the model biases are of smaller order of magnitude than the corresponding model variances. For the estimators based on the fitted values of Kuo this is achieved whenever $\lambda = o\left(n^{-1/4}\right),$ while for the estimators with the modified fitted values this requires that λ goes to zero faster than $O\left(n^{-1/4}\right)$ and slower than $O\left(n^{-1/2}\right).$ The convergence rates for the model biases of the latter estimators are optimized when $\lambda = O\left(n^{-1/3}\right)$ and in this case the resulting model biases are both of order $O\left(n^{-2/3}\right).$ The model biases for the estimators based on the fitted values of Kuo can be made to converge much faster, depending on the sequences $H_{N,s}(x)$ and $H_{N,\bar{s}}(x)$ and on the bandwidth sequence $\lambda.$

Given the above considerations concerning the model biases and given the fact that the leading terms in the model variances are the same for both types of fitted values, it would be of interest to know the second order terms in the model variances in order to establish which estimator is more efficient from the model-based perspective. The proofs in the Appendix suggest however that the second order terms depend on more specific assumptions than (C1) to (C3) and that, in particular for the estimators based on the modified fitted values, they are difficult to determine.

4 Design-based properties

In the previous section we have shown that the model-based estimators $\hat{F}(t)$ and $\hat{F}^*(t)$ are asymptotically model-unbiased and model mean square error consistent. However, they are not design-unbiased in general and therefore they should not be used when the sample inclusion probabilities are not constant. In these cases the generalized difference estimators $\tilde{F}(t)$ and $\tilde{F}^*(t)$ should be used. In fact, it follows from the results in Breidt and Opsomer (2000) that under fairly general conditions $\tilde{F}(t)$ is asymptotically design-unbiased and that its design mean square error is given by

$$E_d\left(\left|\tilde{F}(t)-F_N(t)\right|^2\right)=\frac{1}{N^2}\sum_{i,j\in U}\frac{\pi_{i,j}-\pi_i\pi_j}{\pi_i\pi_j}\left[I(y_i\le t)-\bar{G}_i(t)\right]\left[I\left(y_j\le t\right)-\bar{G}_j(t)\right]+o\left(n^{-1}\right),$$

where $E_d(\cdot)$ denotes expectation with respect to the sample design, $\pi_{i,j}$ denotes the joint sample inclusion probability for units i and j (it is understood that $\pi_{i,i}=\pi_i$), and where

$$\bar{G}_i(t):=\sum_{j\in U}\bar{w}_{i,j}I\left(y_j\le t\right).$$

The regression weights $\bar{w}_{i,j}$ in the definition of $\bar{G}_i(t)$ refer to the whole finite population U and are given by

$$\bar{w}_{i,j}:=\frac{1}{N\lambda}K\left(\frac{x_i-x_j}{\lambda}\right)\frac{\bar{M}_{2,s}(x_i)-\left(\frac{x_i-x_j}{\lambda}\right)\bar{M}_{1,s}(x_i)}{\bar{M}_{2,s}(x_i)\bar{M}_{0,s}(x_i)-\bar{M}_{1,s}^2(x_i)},$$

where

$$\bar{M}_{r,s}(x):=\sum_{k\in U}\frac{1}{N\lambda}K\left(\frac{x-x_k}{\lambda}\right)\left(\frac{x-x_k}{\lambda}\right)^r,\qquad r=0,1,2.$$

Moreover, according to Breidt and Opsomer (2000),

$$\tilde{V}\left(\tilde{F}(t)\right):=\frac{1}{N^2}\sum_{i,j\in s}\frac{\pi_{i,j}-\pi_i\pi_j}{\pi_{i,j}\pi_i\pi_j}\left[I(y_i\le t)-\tilde{G}_i(t)\right]\left[I\left(y_j\le t\right)-\tilde{G}_j(t)\right]$$

is a consistent estimator for the design mean square error of $\tilde{F}(t)$.

Unfortunately the results in Breidt and Opsomer (2000) cannot be applied to the generalized difference estimator $\tilde{F}^*(t)$ as well, since the latter estimator does not fall into the class of local polynomial regression estimators due to the presence of the regression function estimators $\tilde{m}_i$ and $\tilde{m}_j$ inside the indicator functions in the fitted values $\tilde{G}_i^*(t)$. However, the results for $\tilde{F}(t)$ suggest that in large samples $\tilde{G}_i^*(t)$ and

$$\bar{G}_i^*(t):=\sum_{j\in U}\bar{w}_{i,j}I\left(y_j-\bar{m}_j\le t-\bar{m}_i\right),$$

where $\bar{m}_i := \sum_{j \in U} \bar{w}_{i,j} y_j$, are approximately the same, and that

$$E_d\left(\left|\tilde{F}^*(t) - F_N(t)\right|^2\right) = \frac{1}{N^2} \sum_{i,j \in U} \frac{\pi_{i,j} - \pi_i \pi_j}{\pi_i \pi_j} \left[I(y_i \le t) - \bar{G}_i^*(t)\right]\left[I(y_j \le t) - \bar{G}_j^*(t)\right] + o\left(n^{-1}\right)$$

Based on this conjecture, we tested

$$\tilde{V}\left(\tilde{F}^*(t)\right) := \frac{1}{N^2} \sum_{i,j \in s} \frac{\pi_{i,j} - \pi_i \pi_j}{\pi_{i,j} \pi_i \pi_j} \left[I(y_i \le t) - \tilde{G}_i^*(t)\right]\left[I(y_j \le t) - \tilde{G}_j^*(t)\right].$$

as estimator for the design mean square error of the generalized difference estimator $\tilde{F}^*(t)$ in the simulation study of the following section.

5 Simulation study

In this section we analyze some simulation results. Our goal is to compare efficiency with respect to the sample design of the distribution function estimators introduced in Section 2 and of the variance estimators of Section 4. The simulation results refer to simple random without replacement sampling and to Poisson sampling with unequal inclusion probabilities. As a benchmark, we included also the Horvitz-Thompson distribution function estimator

$$\hat{F}_\pi(t) := \frac{1}{N} \sum_{j \in s} \pi_j^{-1} I\left(y_j \le t\right)$$

and the corresponding variance estimator

$$\tilde{V}\left(\hat{F}_\pi(t)\right) := \frac{1}{N^2} \sum_{i,j \in s} \frac{\pi_{i,j} - \pi_i \pi_j}{\pi_{i,j} \pi_i \pi_j} I(y_i \le t) I\left(y_j \le t\right)$$

in the simulation study.

We considered both artificial and real populations. The former were obtained by generating $N = 1{,}000$ values x_i from i.i.d. uniform random variables with support on the interval $(0,1)$ and by combining them with three types of regression function $m(x)$ and two types of error components ε_i. The regression functions are (i) $m(x) = 0$ (flat), (ii) $m(x) = 10x$ (linear) and (iii) $m(x) = 10x^{1/4}$ (concave), while the error components ε_i are either independent realizations from a unique Student t distribution with $\nu = 5$ d.o.f., or independent realizations from N different shifted noncentral Student t distributions with $\nu = 5$ d.o.f. and with noncentrality parameters given by $\mu = 15x_i$. The shifts applied to the error components in the latter case make sure that the means of the noncentral Student t distributions from which they were generated are zero. The artificial populations are shown in Figure 5.1 to 5.3. As for the real populations, we took the *MU284 Population of Sweden Municipalities* of Särndal et al. (1992) (population size $N = 284$) and considered the natural logarithm of *RMT85 = Revenues from the 1985 municipal taxation (in millions of kronor)* as study variable Y, and the natural logarithm of either *P85 = 1985 population (in thousands)* or *REV 84 = Real estate values according to 1984 assessment (in millions of kronor)* as auxiliary variable X. The real populations are shown in Figure 5.4.

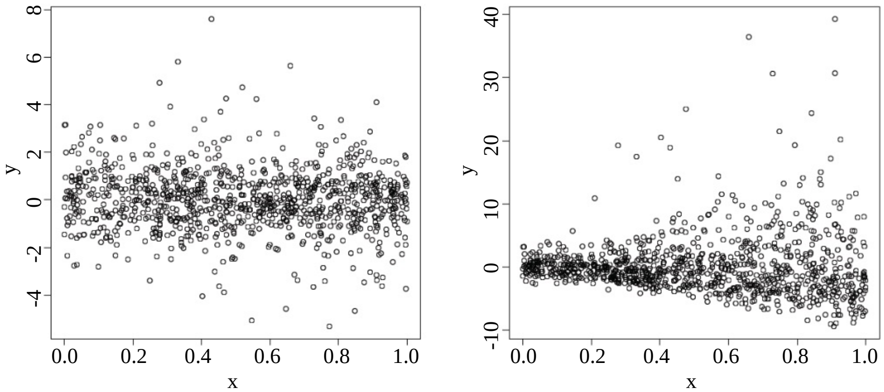

Figure 5.1 Populations generated from $y_i = \varepsilon_i$, where ε_i ~ i.i.d. Student t with $\nu = 5$ (left panel) and ε_i ~ indep. noncentral Student t with $\nu = 5$ and $\mu = 15x_i$ (right panel).

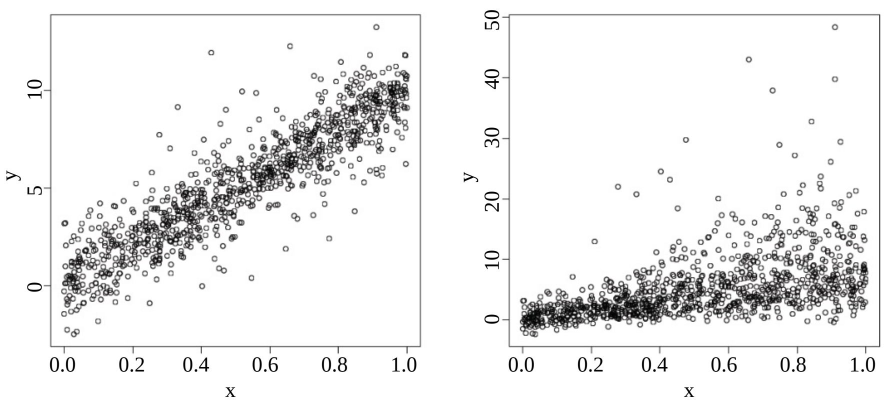

Figure 5.2 Populations generated from $y_i = 10x_i + \varepsilon_i$, where ε_i ~ i.i.d. Student t with $\nu = 5$ (left panel) and ε_i ~ indep. noncentral Student t with $\nu = 5$ and $\mu = 15x_i$ (right panel).

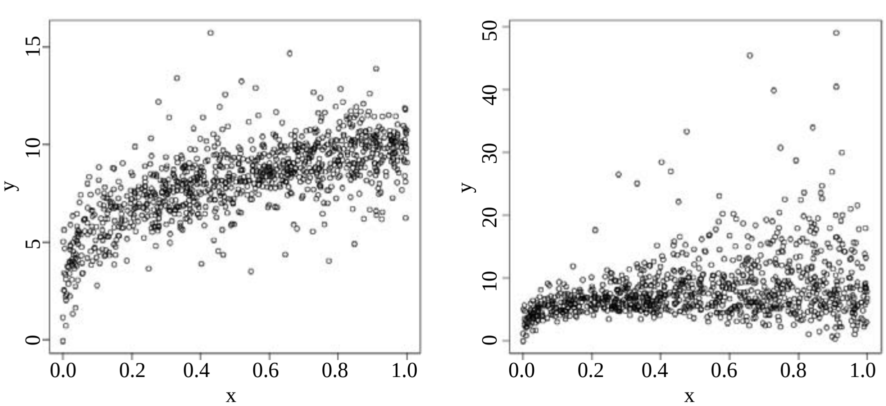

Figure 5.3 Populations generated from $y_i = 10x_i^{1/4} + \varepsilon_i$, where ε_i ~ i.i.d. Student t with $\nu = 5$ (left panel) and ε_i ~ indep. noncentral Student t with $\nu = 5$ and $\mu = 15x_i$ (right panel).

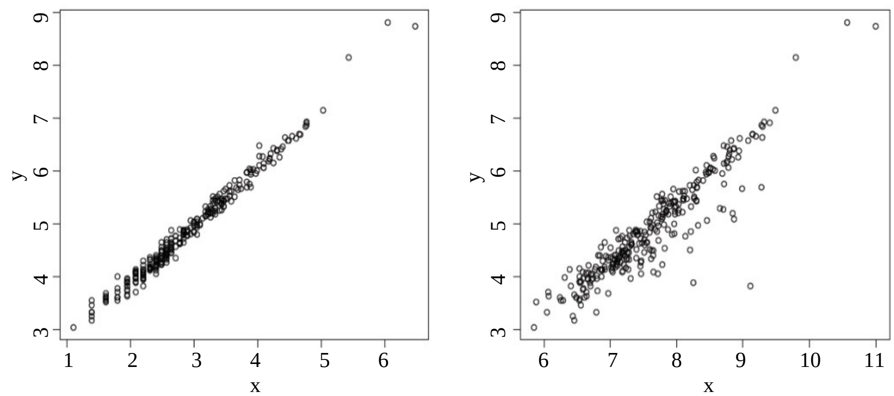

Figure 5.4 ***MU284 Population of Sweden Municipalities*** **of Särndal et al (1992).** $y_i = \ln RMT85_i$ **for the** i^{th} **municipality, and** $x_i = \ln P85_i$ **(left panel) or** $x_i = \ln REV84_i$ **(right panel).**

From each population we selected independently $B = 1{,}000$ samples. When sampling from the artificial populations we set the sample size equal to $n = 100$ in case of simple random without replacement sampling and, in case of Poisson sampling, we set the expected sample size equal to $n^* = 100$ and made the sample inclusion probabilities proportional to the standard deviations of the shifted noncentral Student t distributions of above. When sampling from the real populations, we set the sample size equal to $n = 30$ in case of simple random without replacement sampling. In case of Poisson sampling, we set the expected sample size equal to $n^* = 30$ and made the sample inclusion probabilities proportional to the absolute values of the residuals from the linear least squares regressions of the population y_i values on the population x_i values.

As for the definition of the nonparametric estimators, we used the Epanechnikov kernel function $K(u) := 0.75(1 - u^2)$ with $\lambda = 0.15$ or $\lambda = 0.3$ for the samples taken from the artificial populations, and the Gaussian kernel function $K(u) := 1/\sqrt{2\pi}\, e^{-(1/2)u^2}$ with $\lambda = 1$ or $\lambda = 2$ for the samples taken from the real populations. In the tables with the simulation results the nonparametric estimators corresponding to the small and large bandwidth values are identified with an s (small) or an l (large) in the subscript. We resorted to the Gaussian kernel function for the samples taken from the real populations to avoid singularity problems that occur in case of holes in the sampled set of x_i – values. Such holes are much more likely to occur with the real populations than with the artificial ones, because the distributions of the auxiliary variables are asymmetric in the former. In fact, in the artificial populations the nonparametric estimators were well-defined for all the $B = 1{,}000$ samples selected according to the simple random without replacement sampling design. For the Poisson sampling design, on the other hand, 47 among the $B = 1{,}000$ simulated samples were such that the nonparametric estimators with the small bandwidth value could not be computed and just one of these samples was such that the nonparametric estimators with the large bandwidth value were undefined. The simulation results referring to the nonparametric estimators in Tables 5.2 and 5.5 account only for the samples where they were well-defined and thus they are based on a little less than $B = 1{,}000$ realizations.

Tables 5.1 to 5.4 report the simulated bias (BIAS) and the simulated root mean square error (RMSE) for each distribution function estimator at different levels of t at which $F_N(t)$ has been estimated: based, for example, on the values $\tilde{F}_b(t)$, $b = 1, 2, \ldots, B$, taken on by the estimator $\tilde{F}(t)$,

$$\text{BIAS} := \frac{1}{B}\sum_{b=1}^{B}\left(\tilde{F}_b(t) - F_N(t)\right) \times 10{,}000$$

and

$$\text{RMSE} := \sqrt{\frac{1}{B}\sum_{b=1}^{B}\left(\tilde{F}_b(t) - F_N(t)\right)^2} \times 10{,}000.$$

The RMSE's show that the estimators based on the modified fitted values are usually more efficient. In sampling from the real populations the gain in RMSE is sometimes quite large. As expected, the model-based estimators tend to be more efficient than the generalized difference estimators in case of simple random without replacement sampling when both types of estimator are approximately unbiased. Under the Poisson sampling scheme the BIAS of the model-based estimators increases, but nonetheless they remain competitive. More variability in the sample inclusion probabilities would certainly change this outcome, because it would increase the BIAS of the model-based estimators. The simulation results should therefore not be seen to be in contrast with Johnson, Breidt and Opsomer (2008) who argue in favor of generalized difference estimators (called model-assisted estimators in their paper) as "a good overall choice for distribution function estimators".

Table 5.1
Artificial populations (population size $N = 1{,}000$). BIAS and RMSE of distribution function estimators under simple random without replacement sampling. Sample size $n = 100$

	$t = F_N^{-1}(0.05)$		$t = F_N^{-1}(0.25)$		$t = F_N^{-1}(0.50)$		$t = F_N^{-1}(0.75)$		$t = F_N^{-1}(0.95)$	
	BIAS	**RMSE**	**BIAS**	**RMSE**	**BIAS**	**RMSE**	**BIAS**	**RMSE**	**BIAS**	**RMSE**
$y_i = \varepsilon_i$, with $\varepsilon_i \sim$ i.i.d. central Student t with $\nu = 5$										
$\hat{F}_s(t)$	6	216	-3	433	31	512	23	434	12	207
$\hat{F}_l(t)$	15	219	10	430	0	502	-10	429	3	213
$\hat{F}_s^*(t)$	6	209	-30	411	22	484	22	414	3	200
$\hat{F}_l^*(t)$	15	214	-9	409	10	477	1	407	-10	207
$\tilde{F}_s(t)$	6	213	8	425	24	504	-4	430	8	207
$\tilde{F}_l(t)$	6	210	10	417	22	494	-8	422	6	206
$\tilde{F}_s^*(t)$	8	213	9	426	25	503	-5	432	5	206
$\tilde{F}_l^*(t)$	7	210	10	417	23	494	-6	424	4	206
$\tilde{F}_\pi(t)$	7	208	11	411	19	489	-5	417	6	200
$y_i = \varepsilon_i$, with $\varepsilon_i \sim$ indep. noncentral Student t with $\nu = 5$ and $\mu = 15x_i$										
$\hat{F}_s(t)$	26	225	33	376	8	477	26	419	33	209
$\hat{F}_l(t)$	52	236	23	374	-5	475	38	421	29	213
$\hat{F}_s^*(t)$	20	195	-29	351	-89	471	11	407	30	202
$\hat{F}_l^*(t)$	36	201	-11	357	-94	473	28	410	21	204
$\tilde{F}_s(t)$	8	211	11	370	-7	473	4	415	16	211
$\tilde{F}_l(t)$	5	208	8	367	-5	468	5	411	16	212
$\tilde{F}_s^*(t)$	11	210	11	372	-11	475	4	416	15	210
$\tilde{F}_l^*(t)$	7	208	11	368	-7	468	8	412	15	211
$\tilde{F}_\pi(t)$	1	211	1	391	-6	477	8	399	18	210

Table 5.1 (continued)
Artificial populations (population size $N = 1{,}000$). BIAS and RMSE of distribution function estimators under simple random without replacement sampling. Sample size $n = 100$

	$t = F_N^{-1}(0.05)$		$t = F_N^{-1}(0.25)$		$t = F_N^{-1}(0.50)$		$t = F_N^{-1}(0.75)$		$t = F_N^{-1}(0.95)$	
	BIAS	**RMSE**	**BIAS**	**RMSE**	**BIAS**	**RMSE**	**BIAS**	**RMSE**	**BIAS**	**RMSE**
$y_i = 10x_i + \varepsilon_i$, with $\varepsilon_i \sim$ i.i.d. Student t with $\nu = 5$										
$\hat{F}_s(t)$	32	201	25	275	13	250	-14	264	-36	217
$\hat{F}_l(t)$	114	250	152	304	12	236	-180	312	-86	242
$\hat{F}_s^*(t)$	-50	165	12	226	51	216	26	230	13	172
$\hat{F}_l^*(t)$	-46	155	-14	199	69	195	23	211	17	156
$\tilde{F}_s(t)$	-5	186	4	275	15	248	11	269	-2	201
$\tilde{F}_l(t)$	-5	184	7	274	17	250	5	269	-2	196
$\tilde{F}_s^*(t)$	-10	180	5	275	16	245	14	266	-1	200
$\tilde{F}_l^*(t)$	-9	176	3	272	15	242	13	262	-1	194
$\tilde{F}_\pi(t)$	-7	203	14	413	37	472	17	405	1	206
$y_i = 10x_i + \varepsilon_i$, with $\varepsilon_i \sim$ indep. noncentral Student t with $\nu = 5$ and $\mu = 15x_i$										
$\hat{F}_s(t)$	24	204	23	351	27	403	26	382	29	208
$\hat{F}_l(t)$	94	242	135	372	51	392	13	380	15	212
$\hat{F}_s^*(t)$	55	182	-9	301	-18	368	-23	359	37	202
$\hat{F}_l^*(t)$	124	210	-31	278	-63	363	-8	356	48	200
$\tilde{F}_s(t)$	-2	194	-4	349	11	401	18	377	13	208
$\tilde{F}_l(t)$	-2	190	-5	345	12	398	17	374	11	209
$\tilde{F}_s^*(t)$	0	191	-5	352	14	401	20	376	13	207
$\tilde{F}_l^*(t)$	-1	189	-6	344	13	397	18	375	12	209
$\tilde{F}_\pi(t)$	-4	205	-5	401	21	470	24	401	14	207
$y_i = 10x_i^{1/4} + \varepsilon_i$, with $\varepsilon_i \sim$ i.i.d. Student t with $\nu = 5$										
$\hat{F}_s(t)$	81	207	44	316	17	384	-2	376	23	203
$\hat{F}_l(t)$	138	258	183	356	35	367	-50	374	8	208
$\hat{F}_s^*(t)$	7	146	-14	274	16	352	-8	358	15	197
$\hat{F}_l^*(t)$	9	144	10	246	-2	323	-18	339	24	186
$\tilde{F}_s(t)$	3	175	3	319	10	383	17	374	10	203
$\tilde{F}_l(t)$	0	178	5	316	11	380	17	370	8	202
$\tilde{F}_s^*(t)$	1	167	5	320	12	383	17	374	9	203
$\tilde{F}_l^*(t)$	-1	164	6	316	13	379	20	368	8	201
$\tilde{F}_\pi(t)$	4	209	11	412	25	477	27	422	10	200
$y_i = 10x_i^{1/4} + \varepsilon_i$, with $\varepsilon_i \sim$ indep. noncentral Student t with $\nu = 5$ and $\mu = 15x_i$										
$\hat{F}_s(t)$	59	234	95	402	66	455	51	395	26	208
$\hat{F}_l(t)$	94	259	190	441	147	467	98	400	16	212
$\hat{F}_s^*(t)$	30	184	33	343	-123	435	-34	385	40	203
$\hat{F}_l^*(t)$	57	201	58	331	-148	437	2	382	34	203
$\tilde{F}_s(t)$	1	205	7	386	12	449	17	392	13	208
$\tilde{F}_l(t)$	-1	204	0	385	9	445	20	389	11	209
$\tilde{F}_s^*(t)$	3	201	8	389	7	449	13	392	14	207
$\tilde{F}_l^*(t)$	0	198	6	383	9	446	19	390	13	208
$\tilde{F}_\pi(t)$	0	205	-2	399	9	463	25	398	14	208

Table 5.2
Artificial populations (population size $N = 1{,}000$). BIAS and RMSE of distribution function estimators under Poisson sampling with sample inclusion probabilities π_i proportional to the standard deviations of the noncentral Student t distributions with $\nu = 5$ d.o.f. and with noncentrality parameters $\mu = 15x_i$. Expected sample size $n^* = 100$

	$t = F_N^{-1}(0.05)$		$t = F_N^{-1}(0.25)$		$t = F_N^{-1}(0.50)$		$t = F_N^{-1}(0.75)$		$t = F_N^{-1}(0.95)$	
	BIAS	**RMSE**	**BIAS**	**RMSE**	**BIAS**	**RMSE**	**BIAS**	**RMSE**	**BIAS**	**RMSE**
$y_i = \varepsilon_i$, with $\varepsilon_i \sim$ i.i.d. central Student t with $\nu = 5$										
$\hat{F}_s(t)$	-10	252	-11	593	-22	738	-20	743	6	357
$\hat{F}_l(t)$	-1	237	9	543	-15	621	-5	590	11	302
$\hat{F}_s^*(t)$	22	244	-29	485	-3	555	9	515	-17	297
$\hat{F}_l^*(t)$	14	238	-10	492	-5	564	14	524	-1	283
$\tilde{F}_s(t)$	-6	247	0	579	-27	724	-40	736	3	349
$\tilde{F}_l(t)$	-2	231	11	526	-1	598	-10	566	7	285
$\tilde{F}_s^*(t)$	23	248	23	505	-4	562	-27	531	-20	304
$\tilde{F}_l^*(t)$	12	240	20	504	1	573	-13	538	-6	287
$\tilde{F}_\pi(t)$	-6	220	-7	543	-37	741	-44	929	-48	1,058
$y_i = \varepsilon_i$, with $\varepsilon_i \sim$ indep. noncentral Student t with $\nu = 5$ and $\mu = 15x_i$										
$\hat{F}_s(t)$	17	164	30	411	4	749	14	590	15	190
$\hat{F}_l(t)$	47	173	19	383	-1	602	57	498	15	187
$\hat{F}_s^*(t)$	21	175	-7	378	-89	554	-11	473	3	192
$\hat{F}_l^*(t)$	29	152	-3	367	-99	555	27	481	3	184
$\tilde{F}_s(t)$	1	159	10	406	-11	737	-5	579	-2	194
$\tilde{F}_l(t)$	1	158	9	388	-5	586	14	482	-1	192
$\tilde{F}_s^*(t)$	14	186	27	409	-3	562	-17	487	-10	200
$\tilde{F}_l^*(t)$	3	160	22	399	-11	566	-5	482	-2	193
$\tilde{F}_\pi(t)$	-3	162	-7	451	-31	738	-29	980	-55	1,067
$y_i = 10x_i + \varepsilon_i$, with $\varepsilon_i \sim$ i.i.d. Student t with $\nu = 5$										
$\hat{F}_s(t)$	8	461	21	561	-12	259	-18	218	-30	164
$\hat{F}_l(t)$	78	429	183	451	2	248	-161	261	-79	189
$\hat{F}_s^*(t)$	-69	306	12	340	10	267	15	199	6	143
$\hat{F}_l^*(t)$	-59	294	4	302	56	205	15	172	17	124
$\tilde{F}_s(t)$	-25	441	4	560	-10	257	9	219	5	153
$\tilde{F}_l(t)$	-14	372	35	410	-10	262	4	219	5	151
$\tilde{F}_s^*(t)$	-31	333	-2	386	-29	294	4	227	-1	161
$\tilde{F}_l^*(t)$	-20	339	15	372	-10	259	11	215	4	151
$\tilde{F}_\pi(t)$	-15	385	3	746	-37	917	-35	1,004	-48	1,070
$y_i = 10x_i + \varepsilon_i$, with $\varepsilon_i \sim$ indep. noncentral Student t with $\nu = 5$ and $\mu = 15x_i$										
$\hat{F}_s(t)$	-4	516	30	671	7	453	11	344	6	182
$\hat{F}_l(t)$	63	409	129	539	61	421	9	341	1	180
$\hat{F}_s^*(t)$	44	300	-29	433	-45	422	-47	345	12	180
$\hat{F}_l^*(t)$	107	314	-41	420	-60	397	-22	323	31	171
$\tilde{F}_s(t)$	-27	502	8	667	-8	450	0	344	-8	185
$\tilde{F}_l(t)$	-10	364	16	510	11	425	-2	345	-7	182
$\tilde{F}_s^*(t)$	-6	325	-9	479	-25	447	-14	356	-10	187
$\tilde{F}_l^*(t)$	-7	332	-9	489	-5	426	-3	344	-6	182
$\tilde{F}_\pi(t)$	-16	349	-2	705	-21	886	-42	1,013	-61	1,069

Table 5.2 (continued)
Artificial populations (population size $N = 1{,}000$). BIAS and RMSE of distribution function estimators under Poisson sampling with sample inclusion probabilities π_i proportional to the standard deviations of the noncentral Student t distributions with $\nu = 5$ d.o.f. and with noncentrality parameters $\mu = 15x_i$. Expected sample size $n^* = 100$

	$t = F_N^{-1}(0.05)$		$t = F_N^{-1}(0.25)$		$t = F_N^{-1}(0.50)$		$t = F_N^{-1}(0.75)$		$t = F_N^{-1}(0.95)$	
	BIAS	**RMSE**	**BIAS**	**RMSE**	**BIAS**	**RMSE**	**BIAS**	**RMSE**	**BIAS**	**RMSE**
$y_i = 10x_i^{1/4} + \varepsilon_i$, with $\varepsilon_i \sim$ i.i.d. Student t with $\nu = 5$										
$\hat{F}_s(t)$	36	497	47	629	9	418	-11	320	15	191
$\hat{F}_l(t)$	56	393	186	490	43	383	-48	308	13	184
$\hat{F}_s^*(t)$	-29	276	-19	383	-18	380	-43	335	-1	204
$\hat{F}_l^*(t)$	-29	274	10	355	7	336	-29	290	23	179
$\tilde{F}_s(t)$	-30	475	12	630	4	421	7	317	6	191
$\tilde{F}_l(t)$	-42	336	31	452	11	390	8	312	8	186
$\tilde{F}_s^*(t)$	-31	306	5	429	-18	406	-14	344	-8	210
$\tilde{F}_l^*(t)$	-28	308	14	424	7	387	5	315	7	191
$\tilde{F}_\pi(t)$	-15	380	10	739	-23	891	-37	993	-47	1,064
$y_i = 10x_i^{1/4} + \varepsilon_i$, with $\varepsilon_i \sim$ indep. noncentral Student t with $\nu = 5$ and $\mu = 15x_i$										
$\hat{F}_s(t)$	24	308	69	687	53	690	38	406	2	188
$\hat{F}_l(t)$	47	301	131	553	139	561	91	393	-2	186
$\hat{F}_s^*(t)$	15	237	2	435	-135	513	-59	411	12	186
$\hat{F}_l^*(t)$	27	235	18	435	-149	506	-5	374	13	179
$\tilde{F}_s(t)$	-28	274	-8	673	4	688	3	403	-10	191
$\tilde{F}_l(t)$	-29	251	-12	512	17	541	7	395	-9	188
$\tilde{F}_s^*(t)$	-3	255	-12	481	-7	536	-20	422	-12	196
$\tilde{F}_l^*(t)$	-12	251	-16	489	2	538	-4	399	-9	189
$\tilde{F}_\pi(t)$	-10	267	-8	608	-4	860	-38	1,009	-63	1,066

Table 5.3
Real populations (population size $N = 284$). BIAS and RMSE of distribution function estimators under simple random without replacement sampling. Sample size $n = 30$

	$t = F_N^{-1}(0.05)$		$t = F_N^{-1}(0.25)$		$t = F_N^{-1}(0.50)$		$t = F_N^{-1}(0.75)$		$t = F_N^{-1}(0.95)$	
	BIAS	**RMSE**	**BIAS**	**RMSE**	**RBIAS**	**RMSE**	**BIAS**	**RMSE**	**BIAS**	**RMSE**
MU284 population with $Y = \ln RMT85$ and $X = \ln P85$										
$\hat{F}_s(t)$	133	421	339	625	180	529	-265	490	-187	439
$\hat{F}_l(t)$	52	380	67	588	45	555	-63	469	-87	370
$\hat{F}_s^*(t)$	8	81	-154	203	90	130	62	123	6	54
$\hat{F}_l^*(t)$	28	66	-170	212	69	112	57	109	2	50
$\tilde{F}_s(t)$	-28	300	-24	497	8	483	-48	421	-38	319
$\tilde{F}_l(t)$	-28	326	-96	569	-52	544	3	466	1	319
$\tilde{F}_s^*(t)$	26	177	-11	302	0	244	1	308	-18	102
$\tilde{F}_l^*(t)$	29	179	-10	302	-2	243	-1	308	-21	104
$\tilde{F}_\pi(t)$	22	388	-10	771	9	864	5	731	-43	394
MU284 population with $Y = \ln RMT85$ and $X = \ln REV84$										
$\hat{F}_s(t)$	143	449	303	643	138	554	-217	543	-166	446
$\hat{F}_l(t)$	62	395	62	611	36	582	-49	519	-71	376
$\hat{F}_s^*(t)$	-11	204	-32	300	-101	328	42	285	31	155
$\hat{F}_l^*(t)$	36	183	-40	288	-149	345	6	261	34	122
$\tilde{F}_s(t)$	5	340	-22	548	4	557	-30	498	-23	332
$\tilde{F}_l(t)$	-2	349	-78	599	-36	588	10	522	8	331
$\tilde{F}_s^*(t)$	24	303	7	446	-6	494	2	439	-13	209
$\tilde{F}_l^*(t)$	29	304	4	443	-6	495	-1	432	-18	192
$\tilde{F}_\pi(t)$	34	395	1	766	16	880	9	744	-37	398

Table 5.4
Real populations (population size $N = 284$). BIAS and RMSE of distribution function estimators under Poisson sampling with inclusion probabilities proportional to the absolute value of the residuals of the linear regression of the population y_i – values on the population x_i – values. Expected size $n^* = 30$

	$t = F_N^{-1}(0.05)$		$t = F_N^{-1}(0.25)$		$t = F_N^{-1}(0.50)$		$t = F_N^{-1}(0.75)$		$t = F_N^{-1}(0.95)$	
	BIAS	**RMSE**	**BIAS**	**RMSE**	**RBIAS**	**RMSE**	**BIAS**	**RMSE**	**BIAS**	**RMSE**
MU284 population with $Y = \ln RMT85$ and $X = \ln P85$										
$\hat{F}_s(t)$	204	420	485	668	239	519	-412	626	-90	317
$\hat{F}_l(t)$	180	424	417	684	319	614	-239	548	-148	348
$\hat{F}_s^*(t)$	-41	97	-118	199	132	178	40	140	-71	104
$\hat{F}_l^*(t)$	11	70	-147	211	63	128	-25	122	-85	106
$\tilde{F}_s(t)$	24	360	30	649	0	675	-68	614	58	368
$\tilde{F}_l(t)$	9	390	-63	737	-64	774	-7	682	75	414
$\tilde{F}_s^*(t)$	16	184	-14	307	36	283	16	323	-11	103
$\tilde{F}_l^*(t)$	25	187	-15	312	30	286	14	328	-11	112
$\tilde{F}_\pi(t)$	40	445	73	1,983	12	2,498	-43	3,094	-49	3,341
MU284 population with $Y = \ln RMT85$ and $X = \ln REV84$										
$\hat{F}_s(t)$	349	660	1,185	1,373	890	1,059	458	654	-32	270
$\hat{F}_l(t)$	287	601	1,003	1,236	771	989	484	695	42	263
$\hat{F}_s^*(t)$	317	453	739	866	761	879	624	701	159	207
$\hat{F}_l^*(t)$	364	471	720	842	718	824	572	647	96	158
$\tilde{F}_s(t)$	35	488	82	818	-31	772	7	634	-8	326
$\tilde{F}_l(t)$	22	500	3	878	-98	852	40	704	27	354
$\tilde{F}_s^*(t)$	37	317	32	498	-13	513	32	412	7	157
$\tilde{F}_l^*(t)$	51	313	30	498	-30	518	12	411	-10	149
$\tilde{F}_\pi(t)$	32	671	19	1,658	-172	2,354	-173	2,787	-191	2,935

Consider finally the simulation results referring to the variance estimators of Section 4. Tables 5.5 to 5.8 report the relative bias (RBIAS) and the relative root mean square error (RRMSE) for each of them. For example, based on the variance estimates $\tilde{V}_b(\tilde{F}(t)),\ b = 1, 2, \ldots, B,$ obtained from the estimator $\tilde{V}(\tilde{F}(t)),$

$$\text{RBIAS} := \frac{1}{B}\sum_{b=1}^{B} \frac{\tilde{V}_b(\tilde{F}(t)) - V_B(\tilde{F}(t))}{V_B(\tilde{F}(t))} \times 10{,}000$$

and

$$\text{RRMSE} := \frac{\sqrt{\frac{1}{B}\sum_{b=1}^{B}\left(\tilde{V}_b(\tilde{F}(t)) - V_B(\tilde{F}(t))\right)^2}}{V_B(\tilde{F}(t))} \times 10{,}000$$

where

$$V_B(\tilde{F}(t)) := \frac{1}{B}\sum_{b=1}^{B}\left(\tilde{F}_b(t) - F_N(t)\right)^2.$$

As a benchmark, we report also the RBIAS and RRMSE of the estimator

$$\tilde{V}\left(\tilde{F}_{\pi}(t)\right) := \frac{1}{N^2} \sum_{i,j \in s} \frac{\pi_{i,j} - \pi_i \pi_j}{\pi_{i,j} \pi_i \pi_j} I(y_i \le t) I\left(y_j \le t\right).$$

for the variance of the Horvitz-Thompson estimator.

Table 5.5
Artificial populations (population size $N = 1{,}000$). RBIAS and RRMSE of variance estimators under simple random without replacement sampling. Sample size $n = 100$

	$t = F_N^{-1}(0.05)$		$t = F_N^{-1}(0.25)$		$t = F_N^{-1}(0.50)$		$t = F_N^{-1}(0.75)$		$t = F_N^{-1}(0.95)$	
	RBIAS	**RRMSE**	**RBIAS**	**RRMSE**	**RBIAS**	**RRMSE**	**RBIAS**	**RRMSE**	**RBIAS**	**RRMSE**
$y_i = \varepsilon_i$, with $\varepsilon_i \sim$ i.i.d. central Student t with $\nu = 5$										
$\tilde{V}(\tilde{F}_s(t))$	-1,092	32,442	-1,249	3,895	-1,714	3,077	-1,536	3,828	-824	34,601
$\tilde{V}(\tilde{F}_l(t))$	-576	31,726	-603	3,838	-1,122	3,374	-951	3,758	-441	33,055
$\tilde{V}(\tilde{F}_s^*(t))$	-1,091	32,579	-1,292	3,914	-1,708	3,085	-1,640	3,828	-802	34,809
$\tilde{V}(\tilde{F}_l^*(t))$	-556	31,881	-622	3,857	-1,148	3,361	-1,025	3,749	-425	33,184
$\tilde{V}(\tilde{F}_\pi(t))$	42	30,952	57	3,928	-592	3,776	-287	3,825	551	33,462
$y_i = \varepsilon_i$, with $\varepsilon_i \sim$ indep. noncentral Student t with $\nu = 5$ and $\mu = 15x_i$										
$\tilde{V}(\tilde{F}_s(t))$	-1,900	29,622	50	4,707	-917	3,557	-998	3,695	-1,480	29,417
$\tilde{V}(\tilde{F}_l(t))$	-1,359	29,623	535	4,572	-395	3,881	-527	3,736	-1,277	28,267
$\tilde{V}(\tilde{F}_s^*(t))$	-1,832	30,119	-101	4,710	-991	3,530	-1,077	3,704	-1,398	29,927
$\tilde{V}(\tilde{F}_l^*(t))$	-1,362	29,713	465	4,559	-420	3,865	-591	3,718	-1,236	28,489
$\tilde{V}(\tilde{F}_\pi(t))$	-351	29,132	1,096	4,215	-78	4,074	574	4,067	-638	29,507
$y_i = 10x_i + \varepsilon_i$, with $\varepsilon_i \sim$ i.i.d. Student t with $\nu = 5$										
$\tilde{V}(\tilde{F}_s(t))$	-2,170	11,624	-1,027	2,480	-816	3,274	-1,424	2,583	-1,946	8,681
$\tilde{V}(\tilde{F}_l(t))$	-1,534	11,605	-529	2,632	-148	2,975	-859	2,590	-1,151	9,015
$\tilde{V}(\tilde{F}_s^*(t))$	-1,765	12,107	-1,108	2,529	-714	3,366	-1,318	2,660	-1,905	8,658
$\tilde{V}(\tilde{F}_l^*(t))$	-1,062	11,948	-671	2,735	-212	3,291	-762	2,785	-1,048	8,590
$\tilde{V}(\tilde{F}_\pi(t))$	254	31,545	-52	3,726	136	4,152	267	3,992	35	30,264
$y_i = 10x_i + \varepsilon_i$, with $\varepsilon_i \sim$ indep. noncentral Student t with $\nu = 5$ and $\mu = 15x_i$										
$\tilde{V}(\tilde{F}_s(t))$	-1,642	25,809	-855	3,541	-1,076	3,038	-1,081	3,030	-1,361	21,157
$\tilde{V}(\tilde{F}_l(t))$	-950	25,692	-323	3,509	-597	3,312	-617	3,164	-1,124	20,231
$\tilde{V}(\tilde{F}_s^*(t))$	-1,385	26,406	-997	3,505	-1,089	3,045	-1,096	3,033	-1,310	21,393
$\tilde{V}(\tilde{F}_l^*(t))$	-832	26,212	-292	3,556	-614	3,317	-716	3,154	-1,135	20,286
$\tilde{V}(\tilde{F}_\pi(t))$	105	29,621	507	3,857	209	4,244	425	3,910	-337	29,082
$y_i = 10x_i^{1/4} + \varepsilon_i$, with $\varepsilon_i \sim$ i.i.d. Student t with $\nu = 5$										
$\tilde{V}(\tilde{F}_s(t))$	-2,465	30,612	-1,121	4,594	-1,512	3,183	-1,958	3,076	-863	19,720
$\tilde{V}(\tilde{F}_l(t))$	-1,780	28,103	-663	4,420	-1,092	3,319	-1,491	3,140	-439	18,985
$\tilde{V}(\tilde{F}_s^*(t))$	-2,052	33,980	-1,150	4,619	-1,537	3,217	-1,948	3,127	-954	19,637
$\tilde{V}(\tilde{F}_l^*(t))$	-1,194	33,573	-691	4,472	-1,124	3,368	-1,438	3,228	-357	19,245
$\tilde{V}(\tilde{F}_\pi(t))$	-81	30,001	9	3,756	-110	3,996	-598	3,661	440	32,455
$y_i = 10x_i^{1/4} + \varepsilon_i$, with $\varepsilon_i \sim$ indep. noncentral Student t with $\nu = 5$ and $\mu = 15x_i$										
$\tilde{V}(\tilde{F}_s(t))$	-1,873	29,437	-758	3,759	-621	3,476	-709	3,599	-1,298	27,679
$\tilde{V}(\tilde{F}_l(t))$	-1,267	28,511	-284	3,661	-131	3,758	-321	3,552	-1,075	26,790
$\tilde{V}(\tilde{F}_s^*(t))$	-1,710	30,670	-928	3,741	-628	3,510	-777	3,603	-1,245	27,972
$\tilde{V}(\tilde{F}_l^*(t))$	-939	30,486	-270	3,764	-171	3,803	-375	3,581	-1,014	26,926
$\tilde{V}(\tilde{F}_\pi(t))$	178	29,640	599	3,816	533	4,324	590	3,874	-404	28,917

Table 5.6
Artificial populations (population size $N = 1{,}000$). RBIAS and RRMSE of variance estimators under Poisson sampling with sample inclusion probabilities π_i proportional to standard deviation of noncentral Student t distribution with $\nu = 5$ d.f. and with noncentrality parameter $\mu = 15x_i$. Expected sample size $n^* = 100$

	$t = F_N^{-1}(0.05)$		$t = F_N^{-1}(0.25)$		$t = F_N^{-1}(0.50)$		$t = F_N^{-1}(0.75)$		$t = F_N^{-1}(0.95)$	
	RBIAS	**RRMSE**	**RBIAS**	**RRMSE**	**RBIAS**	**RRMSE**	**RBIAS**	**RRMSE**	**RBIAS**	**RRMSE**
$y_i = \varepsilon_i$, with $\varepsilon_i \sim$ i.i.d. central Student t with $\nu = 5$										
$\tilde V(\tilde F_s(t))$	-3,306	65,777	-4,248	8,032	-5,093	4,242	-6,258	4,844	-5,652	32,037
$\tilde V(\tilde F_l(t))$	-2,048	47,035	-2,656	4,705	-2,434	3,116	-3,310	3,939	-3,092	29,380
$\tilde V(\tilde F_s^*(t))$	-3,362	36,855	-2,488	4,409	-1,910	3,147	-2,869	3,910	-4,329	23,247
$\tilde V(\tilde F_l^*(t))$	-2,696	39,509	-2,076	4,450	-1,768	3,163	-2,648	3,811	-3,244	26,343
$\tilde V(\tilde F_\pi(t))$	113	129,637	259	15,120	618	6,327	193	5,429	273	6,097
$y_i = \varepsilon_i$, with $\varepsilon_i \sim$ indep. noncentral Student t with $\nu = 5$ and $\mu = 15x_i$										
$\tilde V(\tilde F_s(t))$	-740	125,975	-2,522	14,864	-5,466	3,658	-4,896	6,691	-1,551	83,262
$\tilde V(\tilde F_l(t))$	-391	83,047	-1,503	8,946	-2,428	4,099	-2,228	5,526	-1,154	54,680
$\tilde V(\tilde F_s^*(t))$	-3,260	58,072	-2,649	7,661	-2,260	3,936	-2,795	5,011	-2,116	48,739
$\tilde V(\tilde F_l^*(t))$	-716	77,935	-2,000	7,979	-1,934	4,235	-2,279	5,243	-1,243	52,531
$\tilde V(\tilde F_\pi(t))$	666	251,134	-564	26,553	-87	7,344	-2	6,029	407	6,610
$y_i = 10x_i + \varepsilon_i$, with $\varepsilon_i \sim$ i.i.d. Student t with $\nu = 5$										
$\tilde V(\tilde F_s(t))$	-6,801	7,898	-6,470	4,281	-1,059	22,596	-398	32,401	-1,650	72,632
$\tilde V(\tilde F_l(t))$	-4,978	5,826	-2,898	4,473	-603	9,530	206	15,226	-1,157	40,466
$\tilde V(\tilde F_s^*(t))$	-4,520	6,691	-2,710	4,213	-3,245	6,723	-1,156	12,681	-2,458	32,907
$\tilde V(\tilde F_l^*(t))$	-4,226	6,206	-1,674	5,062	-978	7,874	55	12,781	-1,283	33,737
$\tilde V(\tilde F_\pi(t))$	-707	47,550	118	7,214	609	4,409	743	4,628	435	4,800
$y_i = 10x_i + \varepsilon_i$, with $\varepsilon_i \sim$ indep. noncentral Student t with $\nu = 5$ and $\mu = 15x_i$										
$\tilde V(\tilde F_s(t))$	-7,398	8,847	-6,235	3,667	-2,493	8,171	-1,051	16,299	-1,440	71,943
$\tilde V(\tilde F_l(t))$	-4,548	9,463	-3,136	3,282	-1,187	4,246	-832	7,638	-982	45,182
$\tilde V(\tilde F_s^*(t))$	-3,902	11,727	-2,808	3,409	-2,411	3,501	-1,721	6,737	-1,671	41,389
$\tilde V(\tilde F_l^*(t))$	-3,598	10,771	-2,610	3,462	-1,284	3,988	-852	7,008	-972	43,017
$\tilde V(\tilde F_\pi(t))$	146	57,044	-42	8,708	520	4,784	214	4,686	390	5,085
$y_i = 10x_i^{1/4} + \varepsilon_i$, with $\varepsilon_i \sim$ i.i.d. Student t with $\nu = 5$										
$\tilde V(\tilde F_s(t))$	-7,731	8,568	-6,597	3,484	-2,442	7,775	-903	16,067	-1,967	56,480
$\tilde V(\tilde F_l(t))$	-4,611	9,378	-2,990	3,252	-874	4,119	-347	7,420	-1,310	35,051
$\tilde V(\tilde F_s^*(t))$	-4,747	11,909	-2,679	3,298	-1,896	3,272	-2,248	5,747	-3,382	27,222
$\tilde V(\tilde F_l^*(t))$	-4,223	10,380	-2,100	3,494	-788	3,731	-550	5,975	-1,795	29,856
$\tilde V(\tilde F_\pi(t))$	-428	47,038	-206	7,350	641	4,504	738	4,708	487	4,943
$y_i = 10x_i^{1/4} + \varepsilon_i$, with $\varepsilon_i \sim$ indep. noncentral Student t with $\nu = 5$ and $\mu = 15x_i$										
$\tilde V(\tilde F_s(t))$	-4,936	40,696	-6,111	4,579	-5,549	4,035	-1,864	14,381	-1,509	84,892
$\tilde V(\tilde F_l(t))$	-3,004	29,404	-2,764	3,962	-2,436	3,606	-1,234	7,357	-1,103	53,875
$\tilde V(\tilde F_s^*(t))$	-4,328	27,704	-2,516	4,235	-2,671	3,332	-2,586	5,955	-1,939	47,601
$\tilde V(\tilde F_l^*(t))$	-3,454	28,267	-2,263	4,160	-2,329	3,574	-1,433	6,682	-1,171	50,985
$\tilde V(\tilde F_\pi(t))$	152	98,607	663	12,879	15	5,376	20	5,080	429	5,619

Table 5.7
Real populations (population size $N = 284$). RBIAS and RRMSE of variance estimators under simple random without replacement sampling. Sample size $n = 30$

	$t = F_N^{-1}(0.05)$		$t = F_N^{-1}(0.25)$		$t = F_N^{-1}(0.50)$		$t = F_N^{-1}(0.75)$		$t = F_N^{-1}(0.95)$	
	RBIAS	**RRMSE**	**RBIAS**	**RRMSE**	**RBIAS**	**RRMSE**	**RBIAS**	**RRMSE**	**RBIAS**	**RRMSE**
MU284 population with $Y = \ln RMT85$ and $X = \ln P85$										
$\tilde V(\tilde F_s(t))$	-2,853	16,809	-1,700	3,037	-1,554	2,984	-1,100	4,633	-5,503	16,257
$\tilde V(\tilde F_l(t))$	-1,110	16,374	-1,827	2,760	-1,683	2,847	-927	4,387	-3,016	18,685
$\tilde V(\tilde F_s^*(t))$	-1,043	19,081	-91	7,728	-448	9,120	-484	7,715	-1,877	65,298
$\tilde V(\tilde F_l^*(t))$	-424	18,971	104	7,819	-382	9,110	-301	7,799	-1,058	62,968
$\tilde V(\tilde F_\pi(t))$	-186	29,720	-603	3,901	31	3,971	500	4,383	-74	28,418
MU284 population with $Y = \ln RMT85$ and $X = \ln REV84$										
$\tilde V(\tilde F_s(t))$	-2,283	16,303	-1,450	3,538	-945	3,526	-1,071	4,300	-4,832	19,401
$\tilde V(\tilde F_l(t))$	-1,095	16,755	-1,427	3,181	-938	3,390	-780	4,051	-2,753	20,551
$\tilde V(\tilde F_s^*(t))$	-1,737	14,642	-298	5,648	-546	5,282	-736	5,679	-3,564	38,344
$\tilde V(\tilde F_l^*(t))$	-1,174	14,111	-27	5,856	-422	5,452	-228	5,974	-1,433	43,923
$\tilde V(\tilde F_\pi(t))$	-307	28,421	-460	3,963	-344	3,850	112	4,235	-401	27,987

Table 5.8
Real populations (population size $N = 284$). RBIAS and RRMSE of variance estimators under Poisson sampling with inclusion probabilities proportional to the absolute value of the residuals of the linear regression of the population y_i – values on the population x_i – values. Expected size $n^* = 30$

	$t = F_N^{-1}(0.05)$		$t = F_N^{-1}(0.25)$		$t = F_N^{-1}(0.50)$		$t = F_N^{-1}(0.75)$		$t = F_N^{-1}(0.95)$	
	RBIAS	**RRMSE**	**RBIAS**	**RRMSE**	**RBIAS**	**RRMSE**	**RBIAS**	**RRMSE**	**RBIAS**	**RRMSE**
MU284 population with $Y = \ln RMT85$ and $X = \ln P85$										
$\tilde V(\tilde F_s(t))$	-3,502	26,342	-1,841	14,037	-2,691	12,087	-3,415	9,674	-5,932	26,823
$\tilde V(\tilde F_l(t))$	-2,159	27,610	-1,782	14,010	-2,840	12,002	-3,186	10,177	-4,455	26,802
$\tilde V(\tilde F_s^*(t))$	-434	22,455	515	15,503	-506	31,296	-1,460	23,496	-2,649	78,527
$\tilde V(\tilde F_l^*(t))$	-80	22,921	677	15,575	-280	33,294	-1,283	26,612	-1,597	72,166
$\tilde V(\tilde F_\pi(t))$	-294	361,991	522	75,891	43	48,764	-241	36,354	90	32,354
MU284 population with $Y = \ln RMT85$ and $X = \ln REV84$										
$\tilde V(\tilde F_s(t))$	-5,220	18,699	-3,667	8,749	-3,222	7,537	-3,018	9,279	-4,955	44,597
$\tilde V(\tilde F_l(t))$	-4,254	20,765	-3,100	9,180	-3,435	7,231	-3,196	8,540	-3,461	43,206
$\tilde V(\tilde F_s^*(t))$	-2,938	18,922	-1,110	11,828	-1,265	8,726	-1,040	10,963	-3,682	89,262
$\tilde V(\tilde F_l^*(t))$	-1,938	19,997	-699	12,641	-1,003	9,305	-599	11,545	-1,558	98,798
$\tilde V(\tilde F_\pi(t))$	-143	128,401	493	33,934	-255	18,473	-91	17,904	327	16,463

As can be seen from the simulation results, the variance estimators suffer from large variability. This problem is shared by the variance estimator for the Horvitz-Thompson estimator, which occasionally exhibits extremely large RRMSE's. It is further interesting to note that while the RBIAS of the variance estimators for the generalized difference estimators is almost always negative and at times rather large in absolute value, the RBIAS of the variance estimator for the Horvitz-Thompson estimator is in most of the considered cases positive.

Acknowledgements

This research was partially supported by the FAR 2014-ATE-0200 grant from University of Milano-Bicocca.

Appendix

Let β denote a sequence of real numbers. Throughout this appendix we shall indicate by $O_{i_1,i_2,\ldots,i_k}(\beta)$ rest terms that may depend on $x_{i_1}, x_{i_2}, \ldots, x_{i_k}$ and that are of the same order as the sequence β uniformly for $i_1, i_2, \ldots, i_k \in U$. Formally, $R\left(x_{i_1}, x_{i_2}, \ldots, x_{i_k}\right) = O_{i_1,i_2,\ldots,i_k}(\beta)$ if

$$\sup_{i_1,i_2,\ldots,i_k \in U} \left| R\left(x_{i_1}, x_{i_2}, \ldots, x_{i_k}\right) \right| = O(\beta).$$

Moreover, to simplify the notation, we shall write m_i in place of $m(x_i)$ and σ_i^2 in place of $\sigma^2(x_i)$.

Bias of the model-based Kuo estimator

$$\begin{aligned}
E\left(\hat{F}(t) - F_N(t)\right) &= E\left(\frac{1}{N}\sum_{i\notin s}\sum_{j\in s} w_{i,j}\left[I\left(\varepsilon_j \le t - m_j\right) - I\left(\varepsilon_i \le t - m_i\right)\right]\right) \\
&= \frac{1}{N}\sum_{i\notin s}\sum_{j\in s} w_{i,j}\left[G\left(t - m_j \middle| x_j\right) - G\left(t - m_i \middle| x_i\right)\right] \\
&= \frac{1}{2N}\sum_{i\notin s}\Big[G^{(2,0)}\left(t - m_i \middle| x_i\right)\left(m_i'\right)^2 - G^{(1,0)}\left(t - m_i \middle| x_i\right) m_i'' \\
&\qquad - 2G^{(1,1)}\left(t - m_i \middle| x_i\right) m_i' + G^{(0,2)}\left(t - m_i \middle| x_i\right)\Big]\sum_{j\in s} w_{i,j}\left(x_j - x_i\right)^2 + o\left(\lambda^2\right) \\
&= \lambda^2 \frac{N-n}{N}\frac{\mu_2}{2\mu_0}\int_a^b \Big[G^{(2,0)}\left(t - m(x) \middle| x\right)\left(m'(x)\right)^2 - G^{(1,0)}\left(t - m(x) \middle| x\right) m''(x) \\
&\qquad -2G^{(1,1)}\left(t - m(x) \middle| x\right) m'(x) + G^{(0,2)}\left(t - m(x) \middle| x\right)\Big] h_{\bar{s}}(x)\, dx + o\left(\lambda^2\right).
\end{aligned}$$

Bias of the generalized difference Kuo estimator

Write

$$\begin{aligned}
\tilde{F}(t) - F_N(t) &= \frac{1}{N}\Bigg\{\sum_{i\notin s}\sum_{j\in s} \tilde{w}_{i,j}\left[I\left(\varepsilon_j \le t - m_j\right) - I\left(\varepsilon_i \le t - m_i\right)\right] \\
&\qquad + \sum_{i\in s}\left(1 - \frac{1}{\pi_i}\right)\sum_{j\in s} \tilde{w}_{i,j}\left[I\left(\varepsilon_j \le t - m_j\right) - I\left(\varepsilon_i \le t - m_i\right)\right]\Bigg\}.
\end{aligned}$$

Similar steps as those seen for $\hat{F}(t)$ show that

$$E\left(\tilde{F}(t)-F_N(t)\right) = \lambda^2 \frac{N-n}{N}\frac{\mu_2}{2\mu_0}\int_a^b\Big[G^{(2,0)}(t-m(x)|x)\left(m'(x)\right)^2 - G^{(1,0)}(t-m(x)|x)m''(x)$$

$$-2G^{(1,1)}(t-m(x)|x)m'(x)+G^{(0,2)}(t-m(x)|x)\Big]h(x)\,dx+o\left(\lambda^2\right),$$

where

$$h(x) := h_{\bar{s}}(x)+\left(1-\pi^{-1}(x)\right)h_s(x).$$

Variance of the model-based Kuo estimator

$$\operatorname{var}\left(\hat{F}(t)-F_N(t)\right) = \operatorname{var}\left(\frac{1}{N}\sum_{i\notin s}\sum_{j\in s}w_{i,j}I\left(\varepsilon_j \le t-m_j\right)-\frac{1}{N}\sum_{i\notin s}I\left(y_i\le t\right)\right)$$

$$= \frac{1}{N^2}\sum_{i_1\notin s}\sum_{i_2\notin s}\sum_{j\in s}w_{i_1,j}w_{i_2,j}\left[G\left(t-m_j|x_j\right)-G^2\left(t-m_j|x_j\right)\right]$$

$$+\frac{1}{N^2}\sum_{i\notin s}\left[G(t-m_i|x_i)-G^2(t-m_i|x_i)\right]$$

$$= A_1 + A_2,$$

where

$$A_1 := \frac{1}{N^2}\sum_{i_1\notin s}\sum_{i_2\notin s}\sum_{j\in s}w_{i_1,j}w_{i_2,j}\left[G\left(t-m_j|x_j\right)-G^2\left(t-m_j|x_j\right)\right]$$

$$= \frac{1}{N^2}\sum_{j\in s}\left[G\left(t-m_j|x_j\right)-G^2\left(t-m_j|x_j\right)\right]\left(\sum_{i\notin s}w_{i,j}\right)^2$$

$$= \frac{1}{n}\left(\frac{N-n}{N}\right)^2\int_a^b\left[G(t-m(x)|x)-G^2(t-m(x)|x)\right]\left[h_{\bar{s}}(x)/h_s(x)\right]h_{\bar{s}}(x)\,dx$$

$$+O\left((n\lambda)^{-1}\alpha\right)$$

and

$$A_2 := \frac{1}{N^2}\sum_{i\notin s}\left[G(t-m_i|x_i)-G^2(t-m_i|x_i)\right]$$

$$= \frac{1}{N-n}\left(\frac{N-n}{N}\right)^2\int_a^b\left[G(t-m(x)|x)-G^2(t-m(x)|x)\right]h_{\bar{s}}(x)\,dx+O\left(n^{-1}\alpha\right).$$

Thus,

$$\operatorname{var}\left(\hat{F}(t)-F_N(t)\right) = \frac{1}{n}\left(\frac{N-n}{N}\right)^2 \int_a^b \left[G(t-m(x)\mid x)-G^2(t-m(x)\mid x)\right]\left[h_{\bar{s}}(x)/h_s(x)\right]h_{\bar{s}}(x)\,dx$$
$$+\frac{1}{N-n}\left(\frac{N-n}{N}\right)^2 \int_a^b \left[G(t-m(x)\mid x)-G^2(t-m(x)\mid x)\right]h_{\bar{s}}(x)\,dx + O\left((n\lambda)^{-1}\alpha\right).$$

Variance of the generalized difference Kuo estimator

Note that

$$\tilde{F}(t)-F_N(t) = \frac{1}{N}\left\{\sum_{j\in s} I\left(y_j \le t\right)\left[\sum_{i\notin s}\tilde{w}_{i,j} - \sum_{i\in s}\tilde{w}_{i,j}\left(\pi_i^{-1}-1\right)+\left(\pi_j^{-1}-1\right)\right]-\sum_{i\notin s} I\left(y_i \le t\right)\right\}$$

so that

$$\begin{aligned}\operatorname{var}\left(\tilde{F}(t)-F_N(t)\right) &= \operatorname{var}\left(\frac{1}{N}\sum_{j\in s} I\left(y_j\le t\right)\left[\sum_{i\notin s}\tilde{w}_{i,j}+\left(\pi_j^{-1}-1\right)-\sum_{i\in s}\tilde{w}_{i,j}\left(\pi_i^{-1}-1\right)\right]\right)\\ &\quad+\operatorname{var}\left(\frac{1}{N}\sum_{i\notin s} I\left(y_i\le t\right)\right)\\ &= B_1 + A_2,\end{aligned}$$

where A_2 is the same as in the variance of $\hat{F}(t)$, and where

$$\begin{aligned}B_1 &:= \operatorname{var}\left(\frac{1}{N}\sum_{j\in s} I\left(y_j\le t\right)\left[\sum_{i\notin s}\tilde{w}_{i,j}+\left(\pi_j^{-1}-1\right)-\sum_{i\in s}\tilde{w}_{i,j}\left(\pi_i^{-1}-1\right)\right]\right)\\ &= \frac{1}{N^2}\sum_{j\in s}\left[G\left(t-m_j\mid x_j\right)-G^2\left(t-m_j\mid x_j\right)\right]\left[\sum_{i\notin s}\tilde{w}_{i,j}+\left(\pi_j^{-1}-1\right)-\sum_{i\in s}\tilde{w}_{i,j}\left(\pi_i^{-1}-1\right)\right]^2\\ &= \frac{1}{N^2}\sum_{j\in s}\left[G\left(t-m_j\mid x_j\right)-G^2\left(t-m_j\mid x_j\right)\right]\left[\sum_{i\notin s}\tilde{w}_{i,j}+\left(\pi_j^{-1}-1\right)\left(1-\sum_{i\in s}\tilde{w}_{i,j}\right)\right]^2+O\left(\lambda n^{-1}\right)\\ &= \frac{1}{n}\left(\frac{N-n}{N}\right)^2\int_a^b\left[G(t-m(x)\mid x)-G^2(t-m(x)\mid x)\right]\left[h_{\bar{s}}(x)/h_s(x)\right]h_{\bar{s}}(x)\,dx\\ &\quad+O\left((n\lambda)^{-1}\alpha+\lambda n^{-1}\right)\\ &= A_1+O\left((n\lambda)^{-1}\alpha+\lambda n^{-1}\right).\end{aligned}$$

Thus,

$$\operatorname{var}\left(\tilde{F}(t)-F_N(t)\right)=\operatorname{var}\left(\hat{F}(t)-F_N(t)\right)+O\left((n\lambda)^{-1}\alpha+\lambda n^{-1}\right).$$

Bias of the model-based estimator with modified fitted values

Let $\hat{\hat{m}}_i := \sum_{k\in s} w_{i,k} m_k$, $c_{i,j} := 1 - w_{j,j} + w_{i,j}$ and

$$d_{i,j} := \frac{1}{c_{i,j}}\left[\left(1-c_{i,j}\right)(t-m_i)+\left(\hat{\hat{m}}_j-m_j\right)-\left(\hat{\hat{m}}_i-m_i\right)+\sum_{k\in s,k\neq j}\left(w_{j,k}-w_{i,k}\right)\varepsilon_k\right].$$

Observe that $w_{i,j}=O_{i,j}\left((n\lambda)^{-1}\right)$ so that

$$y_j-\hat{m}_j\leq t-\hat{m}_i$$

is (asymptotically, as soon as $c_{i,j}>0$) equivalent to

$$\varepsilon_j\leq t-m_i+d_{i,j}.$$

Since $d_{i,j}$ does not depend on ε_j, it follows that

$$\begin{aligned}E\left(I\left(y_j-\hat{m}_j\leq t-\hat{m}_i\right)\right) &= E\left(I\left(\varepsilon_j\leq t-m_i+d_{i,j}\right)\right)\\ &= E\left(E\left(I\left(\varepsilon_j\leq t-m_i+d_{i,j}\right)\middle|\varepsilon_k,k\neq j\right)\right)\\ &= E\left(G\left(t-m_i+d_{i,j}\middle|x_j\right)\right).\end{aligned} \tag{A.1}$$

Now, using the fact that

$$d_{i,j}=\left(1-c_{i,j}\right)(t-m_i)+\left(\hat{\hat{m}}_j-m_j\right)-\left(\hat{\hat{m}}_i-m_i\right)+\sum_{k\in s,k\neq j}\left(w_{j,k}-w_{i,k}\right)\varepsilon_k+R\left(d_{i,j}\right), \tag{A.2}$$

where

$$E^{1/4}\left(\left|R\left(d_{i,j}\right)\right|^4\right)=O_{i,j}\left(\lambda n^{-1}+(n\lambda)^{-3/2}\right), \tag{A.3}$$

it is seen from (A.1) that

$$\begin{aligned}E\left(I\left(y_j-\hat{m}_j\leq t-\hat{m}_i\right)\right) &= E\left(G\left(t-m_i+d_{i,j}\right)\middle|x_j\right)\\ &= G\left(t-m_i\middle|x_j\right)+G^{(1,0)}\left(t-m_i\middle|x_j\right)E\left(d_{i,j}\right)\\ &\quad+\frac{1}{2}G^{(2,0)}\left(t-m_i\middle|x_j\right)E\left(d_{i,j}^2\right)+o_{i,j}\left(\lambda^4+(n\lambda)^{-1}\right).\end{aligned} \tag{A.4}$$

Thus,

$$\begin{aligned}E\left(\hat{F}^*(t)-F_N(t)\right) &= E\left(\frac{1}{N}\sum_{i\notin s}\sum_{j\in s}w_{i,j}\left(I\left(y_j-\hat{m}_j\leq t-\hat{m}_i\right)-I\left(y_i\leq t\right)\right)\right)\\ &= \frac{1}{N}\sum_{i\notin s}\sum_{j\in s}w_{i,j}\left[G\left(t-m_i\middle|x_j\right)-G\left(t-m_i\middle|x_i\right)\right]\\ &\quad+\frac{1}{N}\sum_{i\notin s}\sum_{j\in s}w_{i,j}G^{(1,0)}\left(t-m_i\middle|x_j\right)E\left(d_{i,j}\right)\\ &\quad+\frac{1}{2N}\sum_{i\notin s}\sum_{j\in s}w_{i,j}G^{(2,0)}\left(t-m_i\middle|x_j\right)E\left(d_{i,j}^2\right)+o\left(\lambda^4+(n\lambda)^{-1}\right)\\ &:= C_1+C_2+C_3+o\left(\lambda^4+(n\lambda)^{-1}\right).\end{aligned} \tag{A.5}$$

Consider first C_1 and note that

$$\begin{aligned} C_1 &:= \frac{1}{N}\sum_{i\notin s}\sum_{j\in s} w_{i,j}\left[G\left(t-m_i \mid x_j\right)-G\left(t-m_i \mid x_i\right)\right] \\ &= \frac{1}{2N}\sum_{i\notin s} G^{(0,2)}\left(t-m_i \mid x_i\right)\sum_{j\in s} w_{i,j}\left(x_j-x_i\right)^2+o\left(\lambda^2\right) \\ &= \lambda^2\,\frac{N-n}{N}\frac{\mu_2}{\mu_0}\int_a^b G^{(0,2)}\left(t-m(x)\mid x\right)h_{\bar{s}}(x)\,dx+o\left(\lambda^2\right). \end{aligned}$$

Consider next C_2. (A.2) and (A.3) imply that

$$\begin{aligned} E\left(d_{i,j}\right) &= \left(1-c_{i,j}\right)\left(t-m_i\right)+\left(\hat{\hat{m}}_j-m_j\right)-\left(\hat{\hat{m}}_i-m_i\right)+O_{i,j}\left(\lambda n^{-1}+(n\lambda)^{-3/2}\right) \\ &= \left(w_{j,j}-w_{i,j}\right)\left(t-m_i\right)+m_j''\sum_{k\in s} w_{j,k}\left(x_k-x_j\right)^2-m_i''\sum_{k\in s} w_{i,k}\left(x_k-x_i\right)^2 \\ &\quad +o_{i,j}\left(\lambda^2\right)+O_{i,j}\left(\lambda n^{-1}+(n\lambda)^{-3/2}\right) \\ &= \left(w_{j,j}-w_{i,j}\right)\left(t-m_i\right)+\left(m_j''-m_i''\right)\sum_{k\in s} w_{j,k}\left(x_k-x_j\right)^2 \\ &\quad +m_i''\left(\sum_{k\in s} w_{j,k}\left(x_k-x_j\right)^2-\sum_{k\in s} w_{i,k}\left(x_k-x_i\right)^2\right) \\ &\quad +o_{i,j}\left(\lambda^2\right)+O_{i,j}\left(\lambda n^{-1}+(n\lambda)^{-3/2}\right) \end{aligned}$$

so that

$$C_2 = C_{2,a}+C_{2,b}+C_{2,c}+o\left(\lambda^2\right)+O\left(\lambda n^{-1}+(n\lambda)^{-3/2}\right),$$

where

$$\begin{aligned} C_{2,a} &:= \frac{1}{N}\sum_{i\notin s}\sum_{j\in s} w_{i,j}G^{(1,0)}\left(t-m_i \mid x_j\right)\left(w_{j,j}-w_{i,j}\right)\left(t-m_i\right) \\ &= \frac{1}{N}\sum_{i\notin s} G^{(1,0)}\left(t-m_i \mid x_i\right)\left(t-m_i\right)\sum_{j\in s} w_{i,j}\left(w_{j,j}-w_{i,j}\right)+O\left(n^{-1}\right) \\ &= \frac{1}{n\lambda}\frac{N-n}{N}\frac{K(0)-\kappa}{\mu_0}\int_a^b G^{(1,0)}\left(t-m(x)\mid x\right)\left(t-m(x)\right)\left[h_{\bar{s}}(x)/h_s(x)\right]dx \\ &\quad +O\left((n\lambda)^{-1}\lambda^{-1}\alpha+n^{-1}\right) \end{aligned}$$

with $\kappa := \int_{-1}^{1} K^2(u)\,du$,

$$\begin{aligned} C_{2,b} &:= \frac{1}{N}\sum_{i\notin s}\sum_{j\in s} w_{i,j}G^{(1,0)}\left(t-m_i \mid x_j\right)\left(m_j''-m_i''\right)\sum_{k\in s} w_{j,k}\left(x_k-x_j\right)^2 \\ &= o\left(\lambda^2\right) \end{aligned}$$

and

$$\begin{aligned} C_{2,c} &:= \frac{1}{N}\sum_{i\notin s}\sum_{j\in s} w_{i,j} G^{(1,0)}\left(t-m_i \mid x_j\right) m_i''\left(\sum_{k\in s} w_{j,k}\left(x_k-x_j\right)^2-\sum_{k\in s} w_{i,k}\left(x_k-x_i\right)^2\right) \\ &= \frac{1}{N}\sum_{i\notin s} G^{(1,0)}\left(t-m_i \mid x_i\right) m_i''\left(\sum_{j\in s} w_{i,j}\sum_{k\in s} w_{j,k}\left(x_k-x_j\right)^2-\sum_{k\in s} w_{i,k}\left(x_k-x_i\right)^2\right)+o\left(\lambda^2\right) \\ &= o\left(\lambda^2\right). \end{aligned}$$

Consider finally C_3. Note that from (A.2) and (A.3)

$$E\left(d_{i,j}^2\right)=\sum_{k\in s}\left(w_{j,k}-w_{i,k}\right)^2\sigma_k^2+O_{i,j}\left(\lambda^4+(n\lambda)^{-2}\right) \tag{A.6}$$

so that

$$\begin{aligned} C_3 &= \frac{1}{2N}\sum_{i\notin s}\sum_{j\in s} w_{i,j} G^{(2,0)}\left(t-m_i \mid x_j\right)\sum_{k\in s}\left(w_{j,k}-w_{i,k}\right)^2\sigma_k^2+O\left(\lambda^4+(n\lambda)^{-2}\right) \\ &= \frac{1}{2N}\sum_{i\notin s} G^{(2,0)}\left(t-m_i \mid x_i\right)\sigma_i^2\sum_{j\in s} w_{i,j}\sum_{k\in s}\left(w_{j,k}-w_{i,k}\right)^2+o\left((n\lambda)^{-1}\right)+O\left(\lambda^4\right) \\ &= \frac{1}{n\lambda}\frac{N-n}{N}\frac{\kappa-\theta}{\mu_0^2}\int_a^b G^{(2,0)}\left(t-m(x)\mid x\right)\sigma^2(x)\left[h_{\bar{s}}(x)/h_s(x)\right]dx+o\left((n\lambda)^{-1}\right)+O\left(\lambda^4\right) \end{aligned}$$

with $\theta:=\int_{-1}^{1}K(v)\int_{-1}^{1}K(u+v)K(u)\,du\,dv$.

Substituting the above expansions for C_1, C_2 and C_3 into (A.5) yields finally

$$\begin{aligned} E\left(\hat{F}^*(t)-F_N(t)\right) &= \lambda^2\frac{N-n}{N}\frac{\mu_2}{\mu_0}\int_a^b G^{(0,2)}\left(t-m(x)\mid x\right)h_{\bar{s}}(x)\,dx \\ &\quad+\frac{1}{n\lambda}\frac{N-n}{N}\left[\frac{K(0)-\kappa}{\mu_0}\int_a^b G^{(1,0)}\left(t-m(x)\mid x\right)\left(t-m(x)\right)h_s^{-1}(x)h_{\bar{s}}(x)\,dx\right. \\ &\quad\left.+\frac{\kappa-\theta}{\mu_0^2}\int_a^b G^{(2,0)}\left(t-m(x)\mid x\right)\sigma^2(x)h_s^{-1}(x)h_{\bar{s}}(x)\,dx\right] \\ &\quad+o\left(\lambda^2+(n\lambda)^{-1}\right). \end{aligned}$$

Bias of the generalized difference estimator with modified fitted values

Let $\tilde{d}_{i,j}$ be the design-weighted counterpart of $d_{i,j}$ and observe that

$$\begin{aligned} \tilde{F}^*(t)-F_N(t) &= \frac{1}{N}\left[\sum_{i\notin s}\sum_{j\in s}\tilde{w}_{i,j}\left(I\left(\varepsilon_j\le t-m_i+\tilde{d}_{i,j}\right)-I\left(y_i\le t\right)\right)\right. \\ &\quad\left.+\sum_{i\in s}\left(1-\pi_i^{-1}\right)\sum_{j\in s}\tilde{w}_{i,j}\left(I\left(\varepsilon_j\le t-m_i+\tilde{d}_{i,j}\right)-I\left(y_i\le t\right)\right)\right]. \end{aligned} \tag{A.7}$$

Adapting the proof that leads to (A.4), it is seen that the asymptotic expansion in (A.4) holds also with $\tilde{d}_{i,j}$ in place of $d_{i,j}$. Adapting the remaining part of the proof finally leads to

$$\begin{aligned} E\left(\tilde{F}^*(t)-F_N(t)\right) &= \lambda^2 \frac{N-n}{N}\frac{\mu_2}{\mu_0}\int_a^b G^{(0,2)}(t-m(x)\mid x)h(x)dx \\ &+\frac{1}{n\lambda}\frac{N-n}{N}\left[\frac{K(0)-\kappa}{\mu_0}\int_a^b G^{(1,0)}(t-m(x)\mid x)(t-m(x))h_s^{-1}(x)h(x)dx\right. \\ &\left.+\frac{\kappa-\theta}{\mu_0^2}\int_a^b G^{(2,0)}(t-m(x)\mid x)\sigma^2(x)h_s^{-1}(x)h(x)dx\right] \\ &+o\left(\lambda^2+(n\lambda)^{-1}\right), \end{aligned}$$

where

$$h(x):=h_{\bar{s}}(x)+\left(1-\pi^{-1}(x)\right)h_s(x).$$

Variance of the model-based estimator with modified fitted values

Write

$$\hat{F}^*(t)-F_N(t)=\frac{1}{N}\left(\sum_{i\notin s}\sum_{j\in s}w_{i,j}I\left(\varepsilon_j\le t-m_i+d_{i,j}\right)-\sum_{i\notin s}I\left(\varepsilon_i\le t-m_i\right)\right)$$

and observe that

$$\operatorname{var}\left(\hat{F}^*(t)-F_N(t)\right)=D_1+D_2+D_3,$$

where

$$D_1:=\frac{1}{N^2}\sum_{i_1\notin s}\sum_{i_2\notin s}\sum_{j\in s}w_{i_1,j}w_{i_2,j}\operatorname{cov}\left(I\left(\varepsilon_j\le t-m_{i_1}+d_{i_1,j}\right),I\left(\varepsilon_j\le t-m_{i_2}+d_{i_2,j}\right)\right),$$

$$D_2:=\frac{1}{N^2}\sum_{i_1\notin s}\sum_{i_2\notin s}\sum_{j_1\in s}\sum_{j_2\in s,\, j_2\ne j_1}w_{i_1,j_1}w_{i_2,j_2}\times\operatorname{cov}\left(I\left(\varepsilon_{j_1}\le t-m_{i_1}+d_{i_1,j_1}\right),I\left(\varepsilon_{j_2}\le t-m_{i_2}+d_{i_2,j_2}\right)\right)$$

and where $D_3:=A_2$ from the variance of the model-based Kuo estimator.

Consider D_1. Observe that

$$\begin{aligned} \operatorname{cov}\left(I\left(\varepsilon_j\le t-m_{i_1}+d_{i_1,j}\right),I\left(\varepsilon_j\le t-m_{i_2}+d_{i_2,j}\right)\right) &= E\left(G\left(t-m_{i_1}+d_{i_1,j}\wedge t-m_{i_2}+d_{i_2,j}\middle| x_j\right)\right) \\ &-E\left(G\left(t-m_{i_1}+d_{i_1,j}\middle| x_j\right)\right)E\left(G\left(t-m_{i_2}+d_{i_2,j}\middle| x_j\right)\right). \end{aligned} \quad (A.8)$$

Since

$$\left|\left(t-m_{i_1}+d_{i_1,j}\wedge t-m_{i_2}+d_{i_2,j}\right)-\left(t-m_{i_1}\wedge t-m_{i_2}\right)\right|\le\left|d_{i_1,j}\right|+\left|d_{i_2,j}\right|,$$

it follows from (A.6) that

$$E\left(G\left(t-m_{i_1}+d_{i_1,j}\wedge t-m_{i_2}+d_{i_2,j}\mid x_j\right)\right)=G\left(t-m_{i_1}\wedge t-m_{i_2}\mid x_j\right)\quad+O_{i_1,i_2,j}\left(\lambda^2+(n\lambda)^{-1/2}\right).\tag{A.9}$$

Moreover, from (A.1), (A.4) and (A.6) it follows that

$$E\left(G\left(t-m_i+d_{i,j}\mid x_j\right)\right)=G\left(t-m_i\mid x_j\right)+O_{i,j}\left(\lambda^2+(n\lambda)^{-1/2}\right).\tag{A.10}$$

Using (A.9) and (A.10) to get an asymptotic expansion for the covariance in (A.8), and substituting the outcome into the definition of D_1 yields

$$\begin{aligned}
D_1 &:= \frac{1}{N^2}\sum_{i_1\notin s}\sum_{i_2\notin s}\sum_{j\in s} w_{i_1,j}w_{i_2,j}\operatorname{cov}\left(I\left(\varepsilon_j\le t-m_{i_1}+d_{i_1,j}\right),I\left(\varepsilon_j\le t-m_{i_2}+d_{i_2,j}\right)\right)\\
&=\frac{1}{N^2}\sum_{i_1\notin s}\sum_{i_2\notin s}\sum_{j\in s} w_{i_1,j}w_{i_2,j}\Big[E\left(G\left(t-m_{i_1}+d_{i_1,j}\wedge t-m_{i_2}+d_{i_2,j}\mid x_j\right)\right)\\
&\qquad\qquad -E\left(G\left(t-m_{i_1}+d_{i_1,j}\mid x_j\right)\right)E\left(G\left(t-m_{i_2}+d_{i_2,j}\mid x_j\right)\right)\Big]\\
&=\frac{1}{N^2}\sum_{i_1\notin s}\sum_{i_2\notin s}\sum_{j\in s} w_{i_1,j}w_{i_2,j}\Big[G\left(t-m_{i_1}\wedge t-m_{i_2}\mid x_j\right)-G\left(t-m_{i_1}\mid x_j\right)G\left(t-m_{i_2}\mid x_j\right)\Big]\\
&\quad +O\left(\lambda^2 n^{-1}+(n\lambda)^{-1/2}n^{-1}\right)\\
&=\frac{1}{N^2}\sum_{j\in s}\Big[G\left(t-m_j\mid x_j\right)-G^2\left(t-m_j\mid x_j\right)\Big]\left(\sum_{i\notin s}w_{i,j}\right)^2+O\left(\lambda n^{-1}+(n\lambda)^{-1/2}n^{-1}\right)\\
&=\frac{1}{n}\left(\frac{N-n}{N}\right)^2\int_a^b\Big[G\left(t-m(x)\mid x\right)-G^2\left(t-m(x)\mid x\right)\Big]\Big[h_{\bar{s}}(x)/h_s(x)\Big]h_{\bar{s}}(x)\,dx\\
&\quad +O\left((n\lambda)^{-1}\alpha+n^{-1}\lambda+n^{-1}(n\lambda)^{-1/2}\right).
\end{aligned}\tag{A.11}$$

Consider next

$$D_2:=\frac{1}{N^2}\sum_{i_1\notin s}\sum_{i_2\notin s}\sum_{j_1\in s}\sum_{j_2\in s,\,j_2\ne j_1} w_{i_1,j_1}w_{i_2,j_2}\times\operatorname{cov}\left(I\left(\varepsilon_{j_1}\le t-m_{i_1}+d_{i_1,j_1}\right),I\left(\varepsilon_{j_2}\le t-m_{i_2}+d_{i_2,j_2}\right)\right).$$

Since

$$\operatorname{cov}\left(I\left(\varepsilon_{j_1}\le t-m_{i_1}+d_{i_1,j_1}\right),I\left(\varepsilon_{j_2}\le t-m_{i_2}+d_{i_2,j_2}\right)\right)=0$$

if $\left|x_{i_1}-x_{i_2}\right|>2\lambda,$ it follows that rest terms R_{i_1,j_1,i_2,j_2}, whose contribution to the above covariance is of order $O_{i_1,j_1,i_2,j_2}(\beta)$ for some sequence β that goes to zero, contribute to D_2 a term of order $O(\lambda\beta)$. Now, let

$$b_{i,j_1,j_2}:=c_{i,j_1}^{-1}\left(w_{j_1,j_2}-w_{i,j_2}\right),$$

$$a_{i,j_1,j_2}:=t-m_i+d_{i,j_1}-b_{i,j_1,j_2}\varepsilon_{j_2}$$

and note that

$$t - m_i + d_{i,j_1} = a_{i,j_1,j_2} + b_{i,j_1,j_2}\varepsilon_{j_2}.$$

Since a_{i,j_1,j_2} does not depend on ε_{j_1} and ε_{j_2}, it follows that

$$\begin{aligned} & E\left(I\left(\varepsilon_{j_1} \le t - m_{i_1} + d_{i_1,j_1}\right) I\left(\varepsilon_{j_2} \le t - m_{i_2} + d_{i_2,j_2}\right)\right) \\ & \quad = E\left(E\left(I\left(\varepsilon_{j_1} \le a_{i_1,j_1,j_2} + b_{i_1,j_1,j_2}\varepsilon_{j_2}\right) I\left(\varepsilon_{j_2} \le a_{i_2,j_2,j_1} + b_{i_2,j_2,j_1}\varepsilon_{j_1}\right) \mid \varepsilon_k, k \ne j_1, j_2\right)\right) \\ & \quad = E\left(\int_{-\infty}^{\varepsilon^*_{i_1,i_2,j_1,j_2}} G\left(a_{i_2,j_2,j_1} + b_{i_2,j_2,j_1}\varepsilon \mid x_{j_2}\right) dG\left(\varepsilon \mid x_{j_1}\right)\right) \\ & \quad + E\left(\int_{-\infty}^{\varepsilon^*_{i_2,i_1,j_2,j_1}} G\left(a_{i_1,j_1,j_2} + b_{i_1,j_1,j_2}\varepsilon \mid x_{j_1}\right) dG\left(\varepsilon \mid x_{j_2}\right)\right) \\ & \quad - E\left(G\left(\varepsilon^*_{i_1,i_2,j_1,j_2} \mid x_{j_1}\right) G\left(\varepsilon^*_{i_2,i_1,j_2,j_1} \mid x_{j_2}\right)\right), \end{aligned} \tag{A.12}$$

where

$$\varepsilon^*_{i_1,i_2,j_1,j_2} := \frac{a_{i_1,j_1,j_2} + a_{i_2,j_2,j_1} b_{i_1,j_1,j_2}}{1 - b_{i_1,j_1,j_2} b_{i_2,j_2,j_1}}.$$

Note that the two expectations in the third and fourth lines in (A.12) are the same if i_1 and j_1 are interchanged with i_2 and j_2, respectively. Thus it suffices to analyze the first expectation. Using the fact that

$$\varepsilon^*_{i_1,i_2,j_1,j_2} = t - m_{i_1} + d_{i_1,j_1} + b_{i_1,j_1,j_2}\left(t - m_{i_2} - \varepsilon_{j_2}\right) + R\left(\varepsilon^*_{i_1,i_2,j_1,j_2}\right),$$

where

$$E^{1/4}\left(\left|R\left(\varepsilon^*_{i_1,i_2,j_1,j_2}\right)\right|^4\right) = O_{i_1,i_2,j_1,j_2}\left(\lambda n^{-1} + (n\lambda)^{-3/2}\right),$$

it is seen that

$$\begin{aligned} & E\left(\int_{-\infty}^{\varepsilon^*_{i_1,i_2,j_1,j_2}} G\left(a_{i_2,j_2,j_1} + b_{i_2,j_2,j_1}\varepsilon \mid x_{j_2}\right) dG\left(\varepsilon \mid x_{j_1}\right)\right) \\ & \quad = G\left(t - m_{i_1} \mid x_{j_1}\right) G\left(t - m_{i_2} \mid x_{j_2}\right) \\ & \quad + G^{(1,0)}\left(t - m_{i_1} \mid x_{j_1}\right) G\left(t - m_{i_2} \mid x_{j_2}\right)\left[E\left(d_{i_1,j_1}\right) + b_{i_1,j_1,j_2}\left(t - m_{i_2}\right)\right] \\ & \quad + G^{(1,0)}\left(t - m_{i_2} \mid x_{j_2}\right) G\left(t - m_{i_1} \mid x_{j_1}\right) E\left(d_{i_2,j_2}\right) + G^{(1,0)}\left(t - m_{i_2} \mid x_{j_2}\right) b_{i_2,j_2,j_1} \int_{-\infty}^{t - m_{i_1}} \varepsilon dG\left(\varepsilon \mid x_{j_1}\right) \\ & \quad + \frac{1}{2} G^{(2,0)}\left(t - m_{i_1} \mid x_{j_1}\right) G\left(t - m_{i_2} \mid x_{j_2}\right) E\left(d^2_{i_1,j_1}\right) + \frac{1}{2} G^{(2,0)}\left(t - m_{i_2} \mid x_{j_2}\right) G\left(t - m_{i_1} \mid x_{j_1}\right) E\left(d^2_{i_2,j_2}\right) \\ & \quad + G^{(1,0)}\left(t - m_{i_1} \mid x_{j_1}\right) G^{(1,0)}\left(t - m_{i_2} \mid x_{j_2}\right) E\left(d_{i_1,j_1} d_{i_2,j_2}\right) \\ & \quad + o_{i_1,i_2,j_1,j_2}\left(\lambda^4 + (n\lambda)^{-1}\right), \end{aligned} \tag{A.13}$$

and that

$$
\begin{aligned}
& E\left(G\left(\varepsilon^{*}_{i_1,i_2,j_1,j_2} \,\middle|\, x_{j_1}\right) G\left(\varepsilon^{*}_{i_2,i_1,j_2,j_1} \,\middle|\, x_{j_2}\right)\right) \\
&\quad = G\left(t - m_{i_1} \,\middle|\, x_{j_1}\right) G\left(t - m_{i_2} \,\middle|\, x_{j_2}\right) \\
&\quad\quad + G^{(1,0)}\left(t - m_{i_1} \,\middle|\, x_{j_1}\right) G\left(t - m_{i_2} \,\middle|\, x_{j_2}\right)\left[E\left(d_{i_1,j_1}\right) + b_{i_1,j_1,j_2}\left(t - m_{i_2}\right)\right] \\
&\quad\quad + G^{(1,0)}\left(t - m_{i_2} \,\middle|\, x_{j_2}\right) G\left(t - m_{i_1} \,\middle|\, x_{j_1}\right)\left[E\left(d_{i_2,j_2}\right) + b_{i_2,j_2,j_1}\left(t - m_{i_1}\right)\right] \\
&\quad\quad + \frac{1}{2} G^{(2,0)}\left(t - m_{i_1} \,\middle|\, x_{j_1}\right) G\left(t - m_{i_2} \,\middle|\, x_{j_2}\right) E\left(d^2_{i_1,j_1}\right) \\
&\quad\quad + \frac{1}{2} G^{(2,0)}\left(t - m_{i_2} \,\middle|\, x_{j_2}\right) G\left(t - m_{i_1} \,\middle|\, x_{j_1}\right) E\left(d^2_{i_2,j_2}\right) \\
&\quad\quad + G^{(1,0)}\left(t - m_{i_1} \,\middle|\, x_{j_1}\right) G^{(1,0)}\left(t - m_{i_2} \,\middle|\, x_{j_2}\right) E\left(d_{i_1,j_1} d_{i_2,j_2}\right) \\
&\quad\quad + o_{i_1,i_2,j_1,j_2}\left(\lambda^4 + (n\lambda)^{-1}\right).
\end{aligned}
\tag{A.14}
$$

Using the asymptotic expansions in (A.4), (A.13) and (A.14) yields

$$
\begin{aligned}
& \operatorname{cov}\left(I\left(\varepsilon_{j_1} \le t - m_{i_1} + d_{i_1,j_1}\right), I\left(\varepsilon_{j_2} \le t - m_{i_2} + d_{i_2,j_2}\right)\right) \\
&\quad = G^{(1,0)}\left(t - m_{i_2} \,\middle|\, x_{j_2}\right) b_{i_2,j_2,j_1} \gamma_{i_1,j_1} + G^{(1,0)}\left(t - m_{i_1} \,\middle|\, x_{j_1}\right) b_{i_1,j_1,j_2} \gamma_{i_2,j_2} \\
&\quad\quad + G^{(1,0)}\left(t - m_{i_1} \,\middle|\, x_{j_1}\right) G^{(1,0)}\left(t - m_{i_2} \,\middle|\, x_{j_2}\right) \operatorname{cov}\left(d_{i_1,j_1}, d_{i_2,j_2}\right) \\
&\quad\quad + o_{i_1,i_2,j_1,j_2}\left(\lambda^4 + (n\lambda)^{-1}\right),
\end{aligned}
\tag{A.15}
$$

where

$$
\gamma_{i,j} := \int_{-\infty}^{t - m_i} \varepsilon \, dG\left(\varepsilon \,\middle|\, x_j\right).
$$

Now observe that

$$
b_{i,j_1,j_2} = w_{j_1,j_2} - w_{i,j_2} + O_{i,j_1,j_2}\left((n\lambda)^{-2}\right)
$$

and that

$$
\begin{aligned}
\operatorname{cov}\left(d_{i_1,j_1}, d_{i_2,j_2}\right) &= \frac{1}{c_{i_1,j_1} c_{i_2,j_2}} \sum_{k \in s; k \ne j_1, j_2} \left(w_{j_1,k} - w_{i_1,k}\right)\left(w_{j_2,k} - w_{i_2,k}\right) \sigma^2_k \\
&= \sum_{k \in s} \left(w_{j_1,k} - w_{i_1,k}\right)\left(w_{j_2,k} - w_{i_2,k}\right) \sigma^2_k + O_{i_1,i_2,j_1,j_2}\left((n\lambda)^{-2}\right)
\end{aligned}
$$

so that

$$
D_2 = 2D_{2a} + D_{2b} + o\left(\lambda^5 + n^{-1}\right),
\tag{A.16}
$$

where

$$\begin{aligned} D_{2a} &:= \frac{1}{N^2} \sum_{i_1 \notin s} \sum_{i_2 \notin s} \sum_{j_1 \in s} \sum_{j_2 \in s, j_2 \neq j_1} w_{i_1, j_1} w_{i_2, j_2} G^{(1,0)}\left(t - m_{i_1} \left| x_{j_1}\right.\right)\left(w_{j_1, j_2} - w_{i_1, j_2}\right) \gamma_{i_2, j_2} \\ &= \frac{1}{N^2} \sum_{i_1 \notin s} \sum_{i_2 \notin s} \sum_{j_1 \in s} \sum_{j_2 \in s} w_{i_1, j_1} w_{i_2, j_2} G^{(1,0)}\left(t - m_{i_1} \left| x_{j_1}\right.\right)\left(w_{j_1, j_2} - w_{i_1, j_2}\right) \gamma_{i_2, j_2} + O\left(n^{-1} (n\lambda)^{-1}\right) \\ &= \frac{1}{N^2} \sum_{j_2 \in s} G^{(1,0)}\left(t - m_{j_2} \left| x_{j_2}\right.\right) \gamma_{j_2, j_2} \left[\sum_{j_1 \in s} w_{j_1, j_2} \sum_{i_1 \notin s} w_{i_1, j_1} \sum_{i_2 \notin s} w_{i_2, j_2} - \left(\sum_{i \notin s} w_{i, j_2}\right)^2\right] \\ &\quad + O\left(n^{-1}\lambda + n^{-1} (n\lambda)^{-1}\right) \\ &= O\left((n\lambda)^{-1} \alpha + n^{-1}\lambda + n^{-1} (n\lambda)^{-1}\right) \end{aligned} \tag{A.17}$$

and

$$\begin{aligned} D_{2b} &:= \frac{1}{N^2} \sum_{i_1 \notin s} \sum_{i_2 \notin s} \sum_{j_1 \in s} \sum_{j_2 \in s, j_2 \neq j_1} w_{i_1, j_1} w_{i_2, j_2} G^{(1,0)}\left(t - m_{i_1} \left| x_{j_1}\right.\right) G^{(1,0)}\left(t - m_{i_2} \left| x_{j_2}\right.\right) \\ &\quad \times \sum_{k \in s} \left(w_{j_1,k} - w_{i_1,k}\right)\left(w_{j_2,k} - w_{i_2,k}\right) \sigma_k^2 \\ &= \frac{1}{N^2} \sum_{i_1 \notin s} \sum_{i_2 \notin s} \sum_{j_1 \in s} \sum_{j_2 \in s} w_{i_1, j_1} w_{i_2, j_2} G^{(1,0)}\left(t - m_{i_1} \left| x_{j_1}\right.\right) G^{(1,0)}\left(t - m_{i_2} \left| x_{j_2}\right.\right) \\ &\quad \times \sum_{k \in s} \left(w_{j_1,k} - w_{i_1,k}\right)\left(w_{j_2,k} - w_{i_2,k}\right) \sigma_k^2 + O\left(n^{-1} (n\lambda)^{-1}\right) \\ &= \frac{1}{N^2} \sum_{k \in s} \sigma_k^2 \left[G^{(1,0)}\left(t - m_k \left| x_k\right.\right)\right]^2 \left(\sum_{i \notin s} \sum_{j \in s} w_{i,j}\left(w_{j,k} - w_{i,k}\right)\right)^2 + O\left(n^{-1}\lambda + n^{-1}(n\lambda)^{-1}\right) \\ &= \frac{1}{N^2} \sum_{k \in s} \sigma_k^2 \left[G^{(1,0)}\left(t - m_k \left| x_k\right.\right)\right]^2 \left(\sum_{j \in s} w_{j,k} \sum_{i \notin s} w_{i,j} - \sum_{i \notin s} w_{i,k}\right)^2 + O\left(n^{-1}\lambda + n^{-1}(n\lambda)^{-1}\right) \\ &= O\left((n\lambda)^{-1} \alpha + n^{-1}\lambda\right). \end{aligned} \tag{A.18}$$

Putting everything together finally yields

$$\begin{aligned} \operatorname{var}\left(\hat{F}^*(t) - F_N(t)\right) &= \frac{1}{n}\left(\frac{N-n}{N}\right)^2 \int_a^b \left[G\left(t - m(x) | x\right) - G^2\left(t - m(x) | x\right)\right]\left[h_{\bar{s}}(x) / h_s(x)\right] h_{\bar{s}}(x)\, dx \\ &\quad + \frac{1}{N-n}\left(\frac{N-n}{N}\right)^2 \int_a^b \left[G\left(t - m(x) | x\right) - G^2\left(t - m(x) | x\right)\right] h_{\bar{s}}(x)\, dx + o\left(\lambda^5 + n^{-1}\right). \end{aligned}$$

Variance of the generalized difference estimator with modified fitted values

In view of (A.7), we shall show that

$$\operatorname{var}\left(\tilde{F}^*(t) - F_N(t)\right) = \operatorname{var}\left(\hat{F}^*(t) - F_N(t)\right) + o\left(n^{-1}\right) \tag{A.19}$$

by showing that

$$\operatorname{var}\left(\frac{1}{N}\sum_{i\in s}\left(1-\pi_i^{-1}\right)\sum_{j\in s}\tilde{w}_{i,j}\left(I\left(\varepsilon_j \le t-m_i+\tilde{d}_{i,j}\right)-I\left(y_i \le t\right)\right)\right)=o\left(n^{-1}\right). \quad \text{(A.20)}$$

To prove (A.20) observe that the variance on the left hand side may be written as

$$E_1+E_2+E_3-2E_4-2E_5,$$

where

$$E_1 := \frac{1}{N^2}\sum_{i_1\in s}\sum_{i_2\in s}\sum_{j\in s}\tilde{w}_{i_1,j}\tilde{w}_{i_2,j}\left(1-\pi_{i_1}^{-1}\right)\left(1-\pi_{i_2}^{-1}\right)\times\operatorname{cov}\left(I\left(\varepsilon_j \le t-m_{i_1}+\tilde{d}_{i_1,j}\right), I\left(\varepsilon_j \le t-m_{i_2}+\tilde{d}_{i_2,j}\right)\right),$$

$$E_2 := \frac{1}{N^2}\sum_{i_1\in s}\sum_{i_2\in s}\sum_{j_1\in s}\sum_{j_2\in s,\, j_2\ne j_1}\tilde{w}_{i_1,j}\tilde{w}_{i_2,j_2}\left(1-\pi_{i_1}^{-1}\right)\left(1-\pi_{i_2}^{-1}\right)\times\operatorname{cov}\left(I\left(\varepsilon_{j_1} \le t-m_{i_1}+\tilde{d}_{i_1,j_1}\right), I\left(\varepsilon_{j_2} \le t-m_{i_2}+\tilde{d}_{i_2,j_2}\right)\right),$$

$$E_3 := \frac{1}{N^2}\sum_{i\in s}\left(1-\pi_i^{-1}\right)^2\operatorname{var}\left(I\left(\varepsilon_i \le t-m_i\right)\right),$$

$$E_4 := \frac{1}{N^2}\sum_{i\in s}\sum_{j\in s}\tilde{w}_{i_1,j}\left(1-\pi_i^{-1}\right)\left(1-\pi_j^{-1}\right)\operatorname{cov}\left(I\left(\varepsilon_j \le t-m_i+\tilde{d}_{i,j}\right), I\left(\varepsilon_j \le t-m_j\right)\right),$$

and finally

$$E_5 := \frac{1}{N^2}\sum_{i_1\in s}\sum_{i_2\in s}\sum_{j\in s,\, j\ne i_2}\tilde{w}_{i_1,j}\left(1-\pi_{i_1}^{-1}\right)\left(1-\pi_{i_2}^{-1}\right)\times\operatorname{cov}\left(I\left(\varepsilon_j \le t-m_{i_1}+\tilde{d}_{i_1,j}\right), I\left(\varepsilon_{i_2} \le t-m_{i_2}\right)\right).$$

To begin with, consider E_1 and E_2. Observe that except for (i) the fact that the summation indexes i_1 and i_2 range over s instead of the complement of s in U, (ii) the presence of the factors $\left(1-\pi_i^{-1}\right)$ and (iii) the fact that the $w_{i,j}$'s and the $d_{i,j}$'s are substituted by their design-weighted counterparts $\tilde{w}_{i,j}$ and $\tilde{d}_{i,j}$, E_1 and E_2 are the same as D_1 and D_2 from $\operatorname{var}\left(\hat{F}^*(t)-F_N(t)\right)$, respectively. Adapting the proofs that lead to the asymptotic expansions for D_1 and D_2 shows thus that

$$E_1 = \frac{1}{n}\left(\frac{N-n}{N}\right)^2\int_a^b\left[G\left(t-m(x)\,|\,x\right)-G^2\left(t-m(x)\,|\,x\right)\right]\left[1-\pi^{-1}(x)\right]^2 h_s(x)\,dx+o\left(n^{-1}\right)$$

and that

$$E_2 = o\left(\lambda^5+n^{-1}\right).$$

As for E_3 it is immediately seen that

$$E_3 = E_1+o\left(n^{-1}\right),$$

while in order to deal with E_4 and E_5 we shall need asymptotic expansions for

$$\operatorname{cov}\left(I\left(\varepsilon_j \le t - m_{i_1} + \tilde{d}_{i_1,j}\right), I\left(\varepsilon_{i_2} \le t - m_{i_2}\right)\right) \tag{A.21}$$

for the case when $j = i_2$ and the case when $j \ne i_2$. In the former case we may employ arguments similar to those for proving (A.9) and (A.10), which lead to

$$\begin{aligned}&\operatorname{cov}\left(I\left(\varepsilon_j \le t - m_{i_1} + \tilde{d}_{i_1,j}\right), I\left(\varepsilon_j \le t - m_j\right)\right)\\ &\quad = G\left(t - m_{i_1} \wedge t - m_j \,|\, x_j\right) - G\left(t - m_{i_1} \,|\, x_j\right) G\left(t - m_j \,|\, x_j\right) + O\left(\lambda^2 + (n\lambda)^{-1/2}\right).\end{aligned}$$

When $j \ne i_2$, on the other hand, the covariance in (A.21) is different from zero only if $\left|x_j - x_{i_2}\right| \le \lambda$ or $\left|x_{i_1} - x_{i_2}\right| \le \lambda$, and adapting (A.12) it can be shown that

$$\begin{aligned}&E\left(I\left(\varepsilon_j \le t - m_{i_1} + \tilde{d}_{i_1,j}\right) I\left(\varepsilon_{i_2} \le t - m_{i_2}\right)\right)\\ &\quad = E\left(E\left(I\left(\varepsilon_j \le \tilde{a}_{i_1,j,i_2} + \tilde{b}_{i_1,j,i_2}\varepsilon_{i_2}\right) I\left(\varepsilon_{i_2} \le t - m_{i_2}\right) \middle| \varepsilon_k, k \ne i, j\right)\right)\\ &\quad = E\left(\int_{-\infty}^{t-m_{i_2}} G\left(\tilde{a}_{i_1,j,i_2} + \tilde{b}_{i_1,j,i_2}\varepsilon \,|\, x_j\right) dG\left(\varepsilon \,|\, x_{i_2}\right)\right)\\ &\quad = G\left(t - m_{i_1} \,|\, x_j\right) G\left(t - m_{i_2} \,|\, x_{i_2}\right) + G\left(t - m_{i_2} \,|\, x_{i_2}\right) G^{(1,0)}\left(t - m_{i_1} \,|\, x_j\right) E\left(d_{i_1,j}\right)\\ &\qquad + G^{(1,0)}\left(t - m_{i_1} \,|\, x_j\right) \tilde{b}_{i_1,j,i_2}\gamma_{i_2,i_2} + \frac{1}{2} G\left(t - m_{i_2} \,|\, x_{i_2}\right) G^{(2,0)}\left(t - m_{i_1} \,|\, x_j\right) E\left(d_{i_1,j}^2\right)\\ &\qquad + o_{i_1,i_2,j}\left(\lambda^4 + (n\lambda)^{-1}\right),\end{aligned}$$

where $\tilde{a}_{i,j,k}$ and $\tilde{b}_{i,j,k}$ are the design-weighted counterparts of $a_{i,j,k}$ and $b_{i,j,k}$, respectively. Adapting also (A.4) to account for the design-weights, it is seen that

$$\begin{aligned}\operatorname{cov}\left(I\left(\varepsilon_j \le t - m_{i_1} + \tilde{d}_{i_1,j}\right), I\left(\varepsilon_{i_2} \le t - m_{i_2}\right)\right) &= G^{(1,0)}\left(t - m_{i_1} \,|\, x_j\right) \tilde{b}_{i_1,j,i_2}\gamma_{i_2,i_2} + o_{i_1,i_2,j}\left(\lambda^4 + (n\lambda)^{-1}\right)\\ &= G^{(1,0)}\left(t - m_i \,|\, x_j\right)\left(\tilde{w}_{j,i_2} - \tilde{w}_{i_1,i_2}\right)\gamma_{i_2,i_2} + o_{i_1,i_2,j}\left(\lambda^4 + (n\lambda)^{-1}\right)\end{aligned}$$

so that (cfr. the steps that lead to the asymptotic expansions of the terms D_1 and D_2 in the variance of the model-based two-step estimator)

$$E_4 = E_1 + o\left(n^{-1}\right)$$

and

$$E_5 = o\left(\lambda^5 + n^{-1}\right).$$

This completes the proof of (A.20) and thus (A.19) follows.

References

Breidt, F.J., and Opsomer, J.D. (2000). Local polynomial regression estimators in survey sampling. *The Annals Statistics*, 28(4), 1026-1053.

Chambers, R.L., and Clark, R. (2012). *An Introduction to Model-Based Survey Sampling with Applications*, Oxford Statistical Science Series 37.

Chambers, R.L., and Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika*, 73(3), 597-604.

Chambers, R.L., Dorfman, A.H. and Wehrly, T.E. (1993). Bias robust estimation in finite populations using non-parametric calibration. *Journal of the American Statistical Association*, 88(421), 268-277.

Chen, J., and Wu, C. (2002). Estimation of distribution function and quantiles using the model-calibrated pseudo empirical likelihood method. *Statistica Sinica*, 12, 1223-1239.

Dorfman, A.H., and Hall, P. (1993). Estimators of the finite population distribution function using nonparametric regression. *The Annals of Statistics*, 21(3), 1452-1475.

Fan, J., and Gijbels, I. (1992). Variable bandwidth and local linear regression smoothers. *The Annals of Statistics*, 20(4), 2008-2036.

Hansen, B.E. (2008). Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory*, 24, 726-748.

Johnson, A.A., Breidt, F.J. and Opsomer, J.D. (2008). Estimating distribution functions from survey data using nonparametric regression. *Journal of Statistical Theory and Practice*, 2(3), 419-431.

Kuo, L. (1988). Classical and prediction approaches to estimating distribution functions from survey data. In *Proceedings of the Survey Research Methods Section,* American Statistical Association, Alexandria, VA, 280-285.

Montanari, G.E., and Ranalli, M.G. (2005). Nonparametric model calibration estimation in survey sampling. *Journal of the American Statistical Association*, 100(472), 1429-1442.

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(2), 365-375.

Rueda, M., Martínez, S., Martínez, H. and Arcos, A. (2007). Estimation of the distribution function with calibration methods. *Journal of Statistical Planning and Inference*, 137(2), 435-448.

Rueda, M., Sànchez-Borrego, I., Arcos, A. and Martínez, S. (2010). Model-calibration estimation of the distribution function using nonparametric regression. *Metrika*, 71(1), 33-44.

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

Wang, J.C., and Opsomer, J.D. (2011). On asymptotic normality and variance estimation for nondifferentiable survey estimators. *Biometrika*, 98(1), 91-106.

Wu, C. (2003). Optimal calibration estimators in survey sampling. *Biometrika*, 90(4), 937-951.

Wu, C., and Sitter, R.R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96(453), 185-193.