SORT 46 (1) January-June 2022, 67-94 DOI: 10.2436/20.8080.02.118

Estimation of finite population distribution function with auxiliary information in a complex survey sampling

Mohsin Abbas[†] and Abdul Haq[†], *

Abstract

In this paper, we consider the problem of estimating the finite population cumulative distribution function (CDF) in a complex survey sampling, which includes two-stage and three-stage cluster sampling schemes with and without stratification. We propose two new families of CDF estimators using supplementary information on a single auxiliary variable. Explicit mathematical expressions of the biases and mean squared errors of the proposed CDF estimators are developed under the first order of the approximation. Real datasets are also considered to support the proposed theory.

MSC: *62D05, 62F10.*

Keywords: *Ratio estimator, exponential ratio estimator, auxiliary information, stratification, two-stage and three-stage cluster sampling, relative efficiencies, bias, mean-squared error.*

1. Introduction

An important problem in the inferential statistics is to estimate the cumulative distribution function (CDF) of a finite population. This problem frequently arises when the underlying interest is to determine the proportion of values of a study variable that are less than or equal to a certain value. For instance, for a nutritionist, it is important to know the proportion of a population that consumes 25% or less of the calories from a saturated fat. Likewise, the policy makers, in a developing country, are mostly interested in knowing the proportion of people living below the poverty line. In the context of survey sampling, it is common to develop CDF estimators with different sampling schemes,

* Corresponding author. E-mail address: aaabdulhaq@yahoo.com

† Department of Statistics, Quaid-i-Azam University, Islamabad, Pakistan.

Received: July 2021

Accepted: April 2022

which include simple random sampling (SRS), stratified random sampling, cluster sampling (CS), ranked-set sampling, to name a few. For more details, see Francisco and Fuller (1986), Haq (2017a), Stokes and Sager (1988) and the references cited therein.

A common approach in survey sampling is to increase the precision of an estimator with suitable use of auxiliary information. The ratio, regression and product-type estimators are prime examples as these estimators require supplementary information on one or more auxiliary variables along with the information on a study variable to increase their relative efficiencies. For example, when estimating the total household income, the age and total expenditure may be used as two auxiliary variables. A significant amount of research work has been done in the literature of survey sampling to develop new improved estimators of the population parameters, which include the population mean, total, CDF, median, etc. Here, our focus is on the estimation of the finite population CDF with the auxiliary information. Chambers and Dunstan (1986) considered estimation of the population CDF and quantiles with the model-based approach. On similar lines, Rao, Kovar and Mantel (1990) proposed ratio and difference/regression estimators for estimating the CDF under a general sampling scheme. Singh, Singh and Kozak (2008) considered the problem of estimating the CDF and quantiles with the use of auxiliary information at the estimation stage of a survey. To our knowledge, recent works on the CDF estimation with auxiliary information may be seen in Tarima and Pavlov (2006), Martínez et al. (2010), Berger and Muñoz (2015), Mayor-Gallego, Moreno-Rebollo and Jiménez-Gamero. (2019), Hussain et al. (2020), Yaqub and Shabbir (2020) and Martínez, Rueda and Illescas (2022), to name a few.

In survey sampling, when the available population is in the form of clusters, that is, households in villages and their members, then it is useful to employ CS instead of SRS. In CS, clusters are randomly selected (with a sampling scheme) from a population, and the data pertaining to a study variable are then collected from all of the units of the selected cluster. However, CS is less efficient than SRS when estimating a population parameter and the former restricts the spread of sampling units across the population. One possible solution is to increase the number of clusters in the sample, and then select representative samples via a sampling scheme from the sampled clusters. This sampling scheme has two stages. It is thus called two-stage CS (2SCS), where the first-stage and second-stage units are called primary stage units (PSUs) and secondary stage units (SSUs), respectively. The 2SCS method is an improvement over CS when it may not be possible or difficult to enumerate all the units of the selected clusters, thereby reducing the cost of the survey. A natural extension of a 2SCS is a three-stage CS (3SCS), where third-stage units are called tertiary stage units (TSUs). This scheme is adopted for inpatients' care cost estimation, where hospitals are selected at the first stage, the selection of wards at the second stage, and the patients at the third stage. Moreover, in large-scale health and demographic surveys, where the population is not only heterogeneous but also more graphically spread, both 2SCS and 3SCS schemes may be combined with the stratified random sampling to get more representative samples, where the stratifying variable may be regions, rural and urban, plan and hilly regions, agro-climatic zones,

etc. For more details see, Cochran (1977), Deville and Särndal (1992), Hansen and Hurwitz (1943), Lee, Lee and Shin (2016), Murthy (1967), Nafiu, Oshungade and Adewara (2012), Rustagi (1978) and references cited therein.

In the survey sampling literature, several authors have considered estimation of the population parameters under 2S and 2SCS schemes. Sukhatme et al. (1984) and Sahoo (1987) considered the estimation of the finite population mean using regression-type estimators in 2S sampling. Smith (1969) studied the ratio estimator for estimation of the finite population mean under multi-stage sampling. Särndal, Swensson and Wretman (2003) considered a regression estimator using 2S sampling under a variety of options. In another study, Nematollahi, Salehi and Aliakbari (2008) developed a new estimator of the population mean using 2SCS, where ranked-set sampling (RSS) was considered in the secondary sampling frame. Srivastava and Garg (2009) used multi-auxiliary information for estimating the population mean in 2S sampling, and they proposed separate-type general class of estimators. Following Nematollahi et al. (2008), Haq (2017b) has considered a hybrid RSS scheme in the secondary sampling frame for developing an improved estimator of the population mean in 2SCS. Recently, Haq, Abbas and Khan (2021) have considered estimation of the finite population CDF under a complex survey sampling scheme, which includes 2SCS, 3SCS, stratified 2SCS (S2SCS) and stratified 3SCS (S3SCS). Under these sampling schemes, they have derived unbiased CDF estimators along with their variances, and the unbiased estimators of the variances of these CDF estimators.

In this study, on the lines of Haq et al. (2021), we consider estimation of the finite population CDF with auxiliary information under 2SCS/3SCS and S2SCS/S3SCS schemes. Following the works of Khoshnevisan et al. (2007) and Singh et al. (2009), we propose two families of classical ratio/product and exponential ratio/product-type estimators for estimating the population CDF under the aforementioned sampling schemes. Moreover, on the lines of Sukhatme et al. (1984) and Sahoo (1987), regression/difference estimators CDF are also developed. Explicit mathematical expressions are obtained for the biases and mean squared errors (MSEs) of the proposed estimators. Real datasets are also considered for the application of the proposed estimators.

The rest of the paper is as follows: In Section 2, CDF estimation is reviewed under 2SCS and 3SCS schemes. In Section 3, we develop explicit mathematical expressions for the covariances of the CDF estimators based on 2SCS/3SCS and S2SCS/S3SCS. In addition, the unbiased estimators of the covariances of the CDF estimators are also derived. In Section 4, two families of estimators, say ratio/product and exponential ratio/product, are proposed for estimating the population CDF. An empirical study is conducted in Section 5. Finally, Section 6 summarizes the main findings and concludes the paper.

2. Estimation of the population CDF

In this section, we briefly review the CDF estimators under 2SCS/S2SCS and 3SCS/S3SCS, which will be used in the subsequent sections.

2.1. Two-stage cluster sampling

The 2SCS uses two stages to select a sample. Assume that the target population, denoted by U, comprises N PSUs, where the ith PSU contains M_i SSUs for $i = 1,2,\ldots,N$. Let $Y_{i,j}$ denote the jth SSU that is present in the ith PSU, where $j = 1,2,\ldots,M_i$ with M_i being the total number of SSUs within the ith PSU. Under 2SCS, the population CDF, $F(y)$, may be written as

$$F(y) \;=\; \frac{1}{N\overline{M}}\sum_{i=1}^{N} M_i F_i(y), \tag{1}$$

where

$$\overline{M} \;=\; \frac{1}{N}\sum_{i=1}^{N} M_i \quad \text{and} \quad F_i(y) \;=\; \frac{1}{M_i}\sum_{j=1}^{M_i} I(Y_{i,j} \le y)$$

are the average cluster size and the CDF computed from the ith PSU, respectively.

In order to estimate $F(y)$ under 2SCS, let n denote the number of PSUs selected in the first stage, and let m_i be the number of SSUs selected from the ith PSU. It is to be noted that, with the 2SCS scheme, the samples under both stages are selected using SRS without replacement. An estimator of $F(y)$ under 2SCS, developed by Haq et al. (2021), is given by

$$\hat{F}_{2\mathrm{S}}(y) = \frac{1}{n\overline{M}}\sum_{i=1}^{n} M_i \hat{F}_i(y) \;=\; \frac{1}{n\overline{M}}\sum_{i=1}^{n}\frac{M_i}{m_i}\sum_{j=1}^{m_i} I(Y_{i,j} \le y), \tag{2}$$

where $I(\cdot)$ is an indicator variable. It can be shown that $\hat{F}_{2\mathrm{S}}(y)$ is an unbiased estimator of $F(y)$. The variance of $\hat{F}_{2\mathrm{S}}(y)$ along with its unbiased estimator are given by

$$V(\hat{F}_{2\mathrm{S}}(y)) = \frac{\lambda\sigma^2_{Y,2b}}{n\overline{M}^2} + \frac{1}{nN\overline{M}^2}\sum_{i=1}^{N}\frac{\zeta_i M_i^2\sigma^2_{Y,2i}}{m_i} \quad \text{and} \tag{3}$$

$$\widehat{V}(\hat{F}_{2\mathrm{S}}(y)) = \frac{\lambda\hat{\sigma}^2_{Y,2b}}{n\overline{M}^2} + \frac{1}{nN\overline{M}^2}\sum_{i=1}^{n}\frac{\zeta_i M_i^2\hat{\sigma}^2_{Y,2i}}{m_i}, \tag{4}$$

respectively, where

$$\sigma^2_{Y,2b} \;=\; \frac{1}{N-1}\sum_{i=1}^{N}\left(M_iF_i(y) - \overline{M}F(y)\right)^2,\; \sigma^2_{Y,2i} \;=\; F_i(y)(1-F_i(y)),$$

$$\hat{\sigma}^2_{Y,2b} \;=\; \frac{1}{n-1}\sum_{i=1}^{n}\left(M_i\hat{F}_i(y) - \overline{M}\hat{F}_{2\mathrm{S}}(y)\right)^2, \hat{\sigma}^2_{Y,2i} \;=\; \frac{M_i(m_i-1)}{m_i(M_i-1)}\hat{F}_i(y)(1-\hat{F}_i(y)),$$

$$\lambda \;=\; \left(1-\frac{n}{N}\right), \text{and } \zeta_i \;=\; \frac{(M_i-m_i)}{(M_i-1)}.$$

In an 2SCS scheme, two types of variations may be considered. The first is the variation between the clusters, and the second is the variation within the clusters. In 2SCS, $\sigma^2_{Y,2b}$ denotes the variance between clusters and $\sigma^2_{Y,2i}$ denotes the variance within the ith cluster. Moreover, $\hat{\sigma}^2_{Y,2i}$ is an unbiased estimator of $\sigma^2_{Y,2i}$.

2.2. Three-stage cluster sampling

The 3SCS requires samples to be selected in three different stages. In the first stage, samples are selected from the PSUs; in the second stage, samples are selected from the SSUs of the selected PSUs; and, in the third stage, the tertiary units are selected from the selected SSUs. Similar to 2SCS, the SRS scheme may be used to select samples at three different stages of the 3SCS.

Suppose that the target population U consists of N PSUs, where each PSU contains M_i SSUs, and each SSU has T_{ij} TSUs. Let $Y_{ij,k}$ denote the kth TSU with the jth SSU of the ith PSU, where $i = 1,2,\ldots,N$, $j = 1,2,\ldots,M_i$, and $k = 1,2,\ldots,T_{ij}$. Under 3SCS, the population CDF, $F(y)$, may be written as

$$F(y) \;=\; \frac{1}{N\overline{T}} \sum_{i=1}^{N} \sum_{j=1}^{M_i} T_{ij} F_{ij}(y), \tag{5}$$

where

$$\overline{T} \;=\; \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{M_i} T_{ij} \quad \text{and} \quad F_{ij}(y) \;=\; \frac{1}{T_{ij}} \sum_{k=1}^{T_{ij}} I(Y_{ij,k} \le y).$$

Here, $\overline{T}$ denotes the average cluster size and $F_{ij}(y)$ be the CDF computed from the jth SSU of the ith PSU.

In order to estimate $F(y)$ under 3SCS, let n denote the number of PSUs selected in the first-stage, let m_i be the number of SSUs selected from the ith PSU, and let t_{ij} be the number of tertiary units selected from the jth SSU. An estimator of $F(y)$ under 3SCS, developed by Haq et al. (2021), is given by

$$\hat{F}_{3S}(y) = \frac{1}{n\overline{T}} \sum_{i=1}^{n} \frac{M_i}{m_i} \sum_{j=1}^{m_i} \frac{T_{ij}}{t_{ij}} \sum_{k=1}^{t_{ij}} I(Y_{ij,k} \le y) \;=\; \frac{1}{n\overline{T}} \sum_{i=1}^{n} \frac{M_i}{m_i} \sum_{j=1}^{m_i} T_{ij} \hat{F}_{ij}(y) \tag{6}$$

It can be shown that $\hat{F}_{3S}(y)$ is an unbiased estimator of $F(y)$. The variance of $\hat{F}_{3S}(y)$ along with its unbiased estimator are given by

$$V(\hat{F}_{3S}(y)) = \frac{\lambda \sigma^2_{Y,3b}}{n\overline{T}^2} + \frac{1}{nN\overline{T}^2} \sum_{i=1}^{N} \frac{\lambda_i M_i^2 \sigma^2_{Y,3i}}{m_i} + \frac{1}{nN\overline{T}^2} \sum_{i=1}^{N} \frac{M_i}{m_i} \sum_{j=1}^{M_i} \frac{\zeta_{ij} T_{ij}^2 \sigma^2_{Y,3ij}}{t_{ij}} \quad \text{and} \tag{7}$$

$$\widehat{V}(\hat{F}_{3S}(y)) = \frac{\lambda \hat{\sigma}^2_{Y,3b}}{n\overline{T}^2} + \frac{1}{nN\overline{T}^2} \sum_{i=1}^{n} \frac{\lambda_i M_i^2 \hat{\sigma}^2_{Y,3i}}{m_i} + \frac{1}{nN\overline{T}^2} \sum_{i=1}^{n} \frac{M_i}{m_i} \sum_{j=1}^{m_i} \frac{\zeta_{ij} T_{ij}^2 \hat{\sigma}^2_{Y,3ij}}{t_{ij}}, \tag{8}$$

respectively, where

$$\sigma^2_{Y,3b} = \frac{1}{N-1}\sum_{i=1}^{N}\left(M_iF_i(y)-\overline{T}F(y)\right)^2, \hat{\sigma}^2_{Y,3b} = \frac{1}{n-1}\sum_{i=1}^{n}\left(M_i\hat{F}_i(y)-\overline{T}\hat{F}_{3S}(y)\right)^2,$$

$$\sigma^2_{Y,3i} = \frac{1}{M_i-1}\sum_{j=1}^{M_i}\left(T_{ij}F_{ij}(y)-F_i(y)\right)^2, \hat{\sigma}^2_{Y,3i} = \frac{1}{m_i-1}\sum_{j=1}^{m_i}\left(T_{ij}\hat{F}_{ij}(y)-\hat{F}_i(y)\right)^2,$$

$$\sigma^2_{Y,3ij} = F_{ij}(y)(1-F_{ij}(y)), \qquad \hat{\sigma}^2_{Y,3ij} = \frac{t_{ij}(T_{ij}-1)}{T_{ij}(t_ij-1)}\hat{F}_{ij}(y)(1-\hat{F}_{ij}(y)),,$$

$$F_i(y) = \frac{1}{M_i}\sum_{j=1}^{M_i}T_{ij}F_{ij}(y), \qquad \hat{F}_i(y) = \frac{1}{m_i}\sum_{j=1}^{m_i}T_{ij}\hat{F}_{ij}(y),$$

$$\lambda = \left(1-\frac{n}{N}\right), \lambda_i = \left(1-\frac{m_i}{M_i}\right), \quad \zeta_{ij} = \frac{T_{ij}-t_{ij}}{T_{ij}-1},$$

where $\sigma^2_{Y,3b}$, $\sigma^2_{Y,3i}$ and $\sigma^2_{Y,3ij}$ have their usual meanings. Moreover, $\hat{\sigma}^2_{Y,3ij}$ is an unbiased estimator of $\sigma^2_{Y,3ij}$. But, $\hat{\sigma}^2_{Y,3b}$ and $\hat{\sigma}^2_{Y,3i}$ are biased estimators of $\sigma^2_{Y,3b}$ and $\sigma^2_{Y,3i}$, respectively. For more detail, see Haq et al. (2021).

2.3. Stratified two-stage cluster sampling

Suppose that the target population Y may be partitioned into L strata, where the hth stratum contains N_h units for $h = 1,2,\ldots,L$. In addition, there are N_h PSUs within the hth stratum, where the ith PSU contains $M_{i,h}$ SSUs for $i = 1,2,\ldots,N_h$. Let $Y_{i,j,h}$ denote the jth SSU that is present in the ith PSU of the hth stratum, where $j = 1,2,\ldots,M_{i,h}$ with $M_{i,h}$ be the total number of SSUs within the ith PSU. Then the population CDF, $F(y)$, under S2SCS, may be written as

$$F(y) = \sum_{h=1}^{L}W_hF_h(y) = \frac{1}{\sum_{h=1}^{L}N_h\overline{M}_h}\sum_{h=1}^{L}N_h\overline{M}_hF_h(y), \tag{9}$$

where

$$W_h = \frac{N_h\overline{M}_h}{\sum_{h=1}^{L}N_h\overline{M}_h}, \qquad F_h(y) = \frac{1}{N_h\overline{M}_h}\sum_{i=1}^{N_h}M_{i,h}F_{i,h}(y),$$

$$F_{i,h}(y) = \frac{1}{M_{i,h}}\sum_{j=1}^{M_{i,h}}I(Y_{i,j,h}\le y), \quad \overline{M}_h = \frac{1}{N_h}\sum_{i=1}^{N_h}M_{i,h}, \tag{10}$$

are computed for the hth stratum.

In order to estimate $F(y)$ under S2SCS, a two-stage cluster sample of size n_h is selected from the hth stratum, where the sample sizes n_h may be allocated using an allocation scheme, like proportional, equal or Neyman allocation. An estimator of $F(y)$ under S2SCS, developed by Haq et al. (2021), is given by

$$\hat{F}_{S2S}(y) = \sum_{h=1}^{L}W_h\hat{F}_{2S,h}(y), \tag{11}$$

where

$$\hat{F}_{2S,h}(y) = \frac{1}{n_h \overline{M}_h} \sum_{i=1}^{n_h} M_{i,h} \hat{F}_{i,h}(y) \quad \text{and} \tag{12}$$
$$\hat{F}_{i,h}(y) = \frac{1}{m_{i,h}} \sum_{j=1}^{m_{i,h}} I(Y_{i,j,h} \leq y).$$

It can be shown that $\hat{F}_{\text{S2S}}(y)$ is an unbiased estimator of $F(y)$. The variance of $\hat{F}_{\text{S2S}}(y)$ along with its unbiased estimator are given by

$$V(\hat{F}_{\text{S2S}}(y)) = \sum_{h=1}^{L} W_h^2 V(\hat{F}_{2S,h}(y)) \quad \text{and} \tag{13}$$
$$\widehat{V}(\hat{F}_{\text{S2S}}(y)) = \sum_{h=1}^{L} W_h^2 \widehat{V}(\hat{F}_{2S,h}(y)), \tag{14}$$

respectively. Note that the mathematical expressions of $V(\hat{F}_{2S,h}(y))$ and $\widehat{V}(\hat{F}_{2S,h}(y))$ (given in Eqs. (3) and (4)) are similar to $V(\hat{F}_{2S}(y))$ and $\widehat{V}(\hat{F}_{2S}(y))$, respectively, with the exception that the former are computed from the hth stratum for $h = 1,2,\ldots,L$.

2.4. Stratified three-stage cluster sampling

Suppose that the target population U is partitioned into L strata, where the hth stratum contains N_h units for $h = 1,2,\ldots,L$. In addition, there are N_h PSUs in the hth stratum, where the ith PSU contains $M_{i,h}$ SSUs for $i = 1,2,\ldots,N_h$. Moreover, each SSU contain $T_{ij,h}$ TSUs for $j = 1,2,\ldots,M_{i,h}$. Let $Y_{ij,k,h}$ denote the kth TSU that is present in the jth SSU of the ith PSU within the hth stratum, where $k = 1,2,\ldots,T_{ij,h}$, and $T_{ij,h}$ be the total number of TSUs within the jth SSU of the ith PSU. Then the population CDF, $F(y)$, under S3SCS, may be written as

$$F(y) \;=\; \sum_{h=1}^{L} W_h F_h(y) \;=\; \frac{1}{\sum_{h=1}^{L} N_h \overline{T}_h} \sum_{h=1}^{L} N_h \overline{T}_h F_h(y), \tag{15}$$

where

$$W_h \;= \frac{N_h \overline{T}_h}{\sum_{h=1}^{L} N_h \overline{T}_h}, \qquad F_h(y) \;=\; \frac{1}{N_h \overline{T}_h} \sum_{i=1}^{N_h} \sum_{j=1}^{M_{i,h}} T_{ij,h} F_{ij,h}(y),$$
$$F_{ij,h}(y) \;= \frac{1}{T_{ij,h}} \sum_{k=1}^{T_{ij,h}} I(Y_{ij,k,h} \leq y), \quad \overline{T}_h = \frac{1}{N_h} \sum_{i=1}^{N_h} \sum_{j=1}^{M_{i,h}} T_{ij,h}. \tag{16}$$

are computed for the hth stratum.

In order to estimate $F(y)$ with S3SCS, a stratified three-stage cluster sample of size n_h is selected from the hth stratum, where the sample size n_h may be allocated with an

allocation scheme, like equal, proportional or Neyman allocation. An estimator of $F(y)$ under S3SCS, developed by Haq et al. (2021), is given by

$$\hat{F}_{\mathrm{S3S}}(y) \;=\; \sum_{h=1}^{L} W_h \hat{F}_{\mathrm{3S},h}(y), \tag{17}$$

where

$$\begin{aligned}\hat{F}_{\mathrm{3S},h}(y) &= \frac{1}{n_h \overline{T}_h} \sum_{i=1}^{n_h} \frac{M_{i,h}}{m_{i,h}} \sum_{j=1}^{m_{i,h}} \frac{T_{ij,h}}{t_{ij,h}} \sum_{k=1}^{t_{ij,h}} I(Y_{ij,k,h} \le y), \\ &= \frac{1}{n_h \overline{T}_h} \sum_{i=1}^{n_h} \frac{M_{i,h}}{m_{i,h}} \sum_{j=1}^{m_{i,h}} T_{ij,h} \hat{F}_{ij,h}(y) \;=\; \frac{1}{n_h \overline{T}_h} \sum_{i=1}^{n_h} M_{i,h} \hat{F}_{i,h}(y), \end{aligned} \tag{18}$$

and

$$\hat{F}_{ij,h}(y) \;=\; \frac{1}{t_{ij,h}} \sum_{k=1}^{t_{ij,h}} I(Y_{ij,k,h} \le y), \quad \hat{F}_{i,h}(y) \;=\; \frac{1}{m_{i,h}} \sum_{j=1}^{m_{i,h}} T_{ij,h} \hat{F}_{ij,h}(y). \tag{19}$$

It can be shown that $\hat{F}_{\mathrm{S3S}}(y)$ is an unbiased estimator of $F(y)$. The variance of $\hat{F}_{\mathrm{S3S}}(y)$ along with its unbiased estimator are given by

$$V(\hat{F}_{\mathrm{S3S}}(y)) = \sum_{h=1}^{L} W_h^2 V(\hat{F}_{\mathrm{3S},h}(y)) \quad \text{and} \tag{20}$$

$$\widehat{V}(\hat{F}_{\mathrm{S3S}}(y)) = \sum_{h=1}^{L} W_h^2 \widehat{V}(\hat{F}_{\mathrm{3S},h}(y)), \tag{21}$$

respectively. Note that the mathematical expressions of $V(\hat{F}_{\mathrm{3S},h}(y))$ and $\widehat{V}(\hat{F}_{\mathrm{3S},h}(y))$ (given in Eqs. (3) and (4)) are similar to $V(\hat{F}_{\mathrm{3S}}(y))$ and $\widehat{V}(\hat{F}_{\mathrm{3S}}(y))$, respectively, with the exception that the former are computed from the hth stratum for $h = 1,2,\ldots,L$, which can be found in Haq et al. (2021).

3. Covariance computation and estimation under a complex survey sampling

In this section, we develop explicit mathematical expressions for the covariances of the CDF estimators based on aforementioned complex survey sampling schemes. In addition, the unbiased estimators of these covariances of the CDF estimators are also derived, which may be used to develop regression-type estimators of the population CDF.

3.1. Two-stage and stratified two-stage cluster sampling

Let Y be the study variable and let X be an auxiliary variable in a finite population U. In order to estimate $(F(y),F(x))$ under 2SCS and S2SCS, let $(\hat{F}_{\mathrm{2S}}(y),\hat{F}_{\mathrm{2S}}(x))$ and $(\hat{F}_{\mathrm{S2S}}(y),\hat{F}_{\mathrm{S2S}}(x))$ be the respective CDF estimators that are based on (Y,X), respectively.

Lemma 1. *Under 2SCS scheme, the covariance between $\hat{F}_{2S}(y)$ and $\hat{F}_{2S}(x)$, along with its unbiased estimator are given by*

$$C(\hat{F}_{2S}(y),\hat{F}_{2S}(x)) = \frac{\lambda\sigma_{XY,2b}}{n\overline{M}^2} + \frac{1}{nN\overline{M}^2}\sum_{i=1}^{N}\frac{\lambda_i M_i^2\sigma_{XY,2i}}{m_i} \quad \textit{and} \tag{22}$$

$$\widehat{C}(\hat{F}_{2S}(y),\hat{F}_{2S}(x)) = \frac{\lambda\hat{\sigma}_{XY,2b}}{n\overline{M}^2} + \frac{1}{nN\overline{M}^2}\sum_{i=1}^{n}\frac{\lambda_i M_i^2\hat{\sigma}_{XY,2i}}{m_i}, \tag{23}$$

respectively, where

$$\sigma_{XY,2b} = \frac{1}{N-1}\sum_{i=1}^{N}\Big((M_iF_i(y)-\overline{M}F(y))(M_iF_i(x)-\overline{M}F(x))\Big), \tag{24}$$

$$\hat{\sigma}_{XY,2b} = \frac{1}{n-1}\sum_{i=1}^{n}\left((M_i\hat{F}_i(y)-\overline{M}\hat{F}_{2S}(y))(M_i\hat{F}_i(x)-\overline{M}\hat{F}_{2S}(x))\right), \tag{25}$$

$$\sigma_{XY,2i} = \frac{1}{M_i-1}\sum_{j=1}^{M_i}\Big((I(Y_{i,j}\le y)-F_i(y))(I(X_{i,j}\le x)-F_i(x))\Big), \tag{26}$$

$$\hat{\sigma}_{XY,2i} = \frac{1}{m_i-1}\sum_{j=1}^{m_i}\left((I(Y_{i,j}\le y)-\hat{F}_i(y))(I(X_{i,j}\le x)-\hat{F}_i(x))\right). \tag{27}$$

Proof. Here, $\sigma_{XY,2b}$ and $\sigma_{XY,2i}$ have their usual meanings. The proof of this Lemma may be seen in the Appendix.

Lemma 2. *Under S2SCS scheme, the covariance between $\hat{F}_{S2S}(y)$ and $\hat{F}_{S2S}(x)$, along with its unbiased estimator are given by*

$$C(\hat{F}_{S2S}(y),\hat{F}_{S2S}(x)) = \sum_{h=1}^{L}W_h^2\,C(\hat{F}_{2S,h}(y),\hat{F}_{2S,h}(x)) \quad \textit{and} \tag{28}$$

$$\widehat{C}(\hat{F}_{S2S}(y),\hat{F}_{S2S}(x)) = \sum_{h=1}^{L}W_h^2\,\widehat{C}(\hat{F}_{2S,h}(y),\hat{F}_{2S,h}(x)), \tag{29}$$

respectively, where W_h is given in Eq. (10).

Proof. The proof of Lemma 2 is similar to that of Lemma 1. Note that the mathematical expressions of $C(\hat{F}_{2S,h}(y),\hat{F}_{2S,h}(x))$ and $\widehat{C}(\hat{F}_{2S,h}(y),\hat{F}_{2S,h}(x))$ are similar to those of $C(\hat{F}_{2S}(y),\hat{F}_{2S}(x))$ and $\widehat{C}(\hat{F}_{2S}(y),\hat{F}_{2S}(x))$, respectively, with the exception that the former are computed from the hth stratum for $h=1,2,\ldots,L$.

3.2. Three-stage and stratified three-stage cluster sampling

In order to estimate $(F(y),F(x))$ under 3SCS and S3SCS, let $(\hat{F}_{3S}(y),\hat{F}_{3S}(x))$ and $(\hat{F}_{S3S}(y),$ $\hat{F}_{S3S}(x))$ be the respective CDF estimators that are based on (Y,X), respectively.

Lemma 3. *Under 3SCS scheme, the covariance between $\hat{F}_{3S}(y)$ and $\hat{F}_{3S}(x)$, along with its unbiased estimators are given by*

$$C\big(\hat{F}_{3S}(y),\hat{F}_{3S}(x)\big) = \frac{\lambda\,\sigma_{XY,3b}}{n\overline{T}^2}+\frac{1}{nN\overline{T}^2}\sum_{i=1}^{N}\frac{\lambda_i M_i^2\sigma_{XY,3i}}{m_i}+\frac{1}{nN\overline{T}^2}\sum_{i=1}^{N}\frac{M_i}{m_i}\sum_{j=1}^{M_i}\frac{\lambda_{ij}T_{ij}^2\sigma_{XY,3ij}}{t_{ij}}, \tag{30}$$

and

$$\widehat{C}\big(\hat{F}_{3S}(y),\hat{F}_{3S}(x)\big) = \frac{\lambda\,\hat{\sigma}_{XY,3b}}{n\overline{T}^2}+\frac{1}{nN\overline{T}^2}\sum_{i=1}^{n}\frac{\lambda_i M_i^2\hat{\sigma}_{XY,3i}}{m_i}+\frac{1}{nN\overline{T}^2}\sum_{i=1}^{n}\frac{M_i}{m_i}\sum_{j=1}^{m_i}\frac{\lambda_{ij}T_{ij}^2\hat{\sigma}_{XY,3ij}}{t_{ij}}, \tag{31}$$

respectively, where

$$\sigma_{XY,3b}=\frac{1}{N-1}\sum_{i=1}^{N}\Big((M_iF_i(y)-\overline{T}F(y))(M_iF_i(x)-\overline{T}F(x))\Big), \tag{32}$$

$$\hat{\sigma}_{XY,3b}=\frac{1}{n-1}\sum_{i=1}^{n}\big((M_i\hat{F}_i(y)-\overline{T}\hat{F}_{3S}(y))(M_i\hat{F}_i(x)-\overline{T}\hat{F}_{3S}(x))\big), \tag{33}$$

$$\sigma_{XY,3i}=\frac{1}{M_i-1}\sum_{j=1}^{M_i}\Big((T_{ij}F_{ij}(y)-F_i(y))(T_{ij}F_{ij}(x)-F_i(x))\Big), \tag{34}$$

$$\hat{\sigma}_{XY,3i}=\frac{1}{m_i-1}\sum_{j=1}^{m_i}\big((T_{ij}\hat{F}_{ij}(y)-\hat{F}_i(y))(T_{ij}\hat{F}_{ij}(x)-\hat{F}_i(x))\big), \tag{35}$$

$$\sigma_{XY,3ij}=\frac{1}{T_{ij}-1}\sum_{k=1}^{T_{ij}}\Big((I(Y_{ij,k}\le y)-F_{ij}(y))(I(X_{ij,k}\le x)-F_{ij}(x))\Big), \tag{36}$$

$$\hat{\sigma}_{XY,3ij}=\frac{1}{t_{ij}-1}\sum_{k=1}^{t_{ij}}\big((I(Y_{ij,k}\le y)-\hat{F}_{ij}(y))(I(X_{ij,k}\le x)-\hat{F}_{ij}(x))\big), \tag{37}$$

and $\lambda_{ij}=(1-t_{ij}/T_{ij})$.

Proof. Here, $\sigma_{XY,3b}$ and $\sigma_{XY,3i}$ have their usual meanings. The proof of this Lemma may be seen in the Appendix.

Lemma 4. *Under S3SCS scheme, the covariance between $\hat{F}_{S3S}(y)$ and $\hat{F}_{S3S}(x)$, along with its unbiased estimator are given by*

$$C(\hat{F}_{S3S}(y),\hat{F}_{S3S}(x))=\sum_{h=1}^{L}W_h^2\,C(\hat{F}_{3S,h}(y),\hat{F}_{3S,h}(x)) \quad \textit{and} \tag{38}$$

$$\widehat{C}(\hat{F}_{S3S}(y),\hat{F}_{S3S}(x))=\sum_{h=1}^{L}W_h^2\,\widehat{C}(\hat{F}_{3S,h}(y),\hat{F}_{3S,h}(x)), \tag{39}$$

respectively, where W_h is given in Eq. (16).

Proof. The proof of Lemma 4 is similar to that of Lemma 3. Note that the mathematical expressions of $C(\hat{F}_{3S,h}(y),\hat{F}_{3S,h}(x))$ and $\widehat{C}(\hat{F}_{3S,h}(y),\hat{F}_{3S,h}(x))$ are similar to those of $C(\hat{F}_{3S}(y),\hat{F}_{3S}(x))$ and $\widehat{C}(\hat{F}_{3S}(y),\hat{F}_{3S}(x))$, respectively, with the exception that the former are computed from the hth stratum for $h = 1,2,\ldots,L$.

4. The CDF estimation with auxiliary information

In this section, we develop two auxiliary-information-based families of estimators, say ratio/product and exponential ratio/product, for estimating the population CDF $F(y)$ under the aforementioned complex survey sampling schemes.

In order to obtain the biases and MSEs of the proposed families of estimators of $F(y)$, we may consider the following relative error terms: Let

$$\xi_0 \;=\; \frac{\hat{F}_S(y)-F(y)}{F(y)} \quad \text{and} \quad \xi_1 \;=\; \frac{\hat{F}_S(x)-F(x)}{F(x)},$$

such that $E(\xi_0) = E(\xi_1) = 0$. Let us denote

$$V_{rs} = E\left(\xi_0^r \xi_1^s\right) \;=\; E\left[\left(\frac{\hat{F}_S(y)-F(y)}{F(y)}\right)^r \left(\frac{\hat{F}_S(x)-F(x)}{F(x)}\right)^s\right], \tag{40}$$

which gives

$$\begin{aligned}
V_{20} &= E(\xi_0)^2 \;=\; E\left(\frac{\hat{F}_S(y)-F(y)}{F(y)}\right)^2 \;=\; \frac{V(\hat{F}_S(y))}{(F(y))^2},\\
V_{02} &= E(\xi_1)^2 \;=\; E\left(\frac{\hat{F}_S(x)-F(x)}{F(x)}\right)^2 \;=\; \frac{V(\hat{F}_S(x))}{(F(x))^2},\\
V_{11} &= E(\xi_0\xi_1) = E\left[\left(\frac{\hat{F}_S(y)-F(y)}{F(y)}\right)\left(\frac{\hat{F}_S(x)-F(x)}{F(x)}\right)\right] \;=\; \frac{C(\hat{F}_S(y),\hat{F}_S(x))}{F(y)F(x)},
\end{aligned}$$

where $\hat{F}_S$ denotes an CDF estimator based on an S sampling scheme, where S = 2S, S2S, 3S and S3S.

4.1. First proposed family of CDF estimators

On the lines of Khoshnevisan et al. (2007), we propose a family of ratio/product-type estimators for estimating the population CDF $F(y)$, given by

$$\hat{F}_R(y) \;=\; \hat{F}_S(y)\left(\frac{aF(x)+b}{\alpha(a\hat{F}_S(x)+b)+(1-\alpha)(aF(x)+b)}\right)^g, \tag{41}$$

where $a \neq 0$ and b are either real numbers or functions of the known parameters of the auxiliary variable X such as coefficient of variation (C_X), correlation coefficient (ρ_{XY}), coefficient of skewness $(\beta_{1,X})$ and coefficient of kurtosis $(\beta_{2,X})$ etc. Here, $g \in \{-1,1\}$

and $\alpha\,(0 \leq \alpha \leq 1)$ are suitably chosen scalars which make the MSE of $\hat{F}_R(y)$ minimum. It is possible to develop different estimators of $\hat{F}_R(y)$ with suitable choices of a, b, g and α. In Table 1, some members of $\hat{F}_R(y)$ are given for different values of a, b, α, and g.

In order to derive approximate mathematical expressions for the bias and MSE of $\hat{F}_R(y)$, we can write $\hat{F}_{\mathrm{S}}(y) = F(y)(1+\xi_0)$ and $\hat{F}_{\mathrm{S}}(x) = F(x)(1+\xi_1)$. Express the right-hand side (RHS) of (41) in terms of ξs to get:

$$\hat{F}_R(y) = F(y)(1+\xi_0)(1+\alpha v \xi_1)^{-g}, \tag{42}$$

where $v = aF(x)/(aF(x)+b)$. Expand the RHS of Eq. (42) and retain terms up to 2nd power of ξs, we have

$$\hat{F}_R(y) \;\approx\; F(y)\left(1+\xi_0-\alpha v g\xi_1+\frac{g(g+1)}{2}\alpha^2 v^2\xi_1^2-\alpha v g\xi_0\xi_1\right) \tag{43}$$

Take expectation on both sides of Eq. (43) after subtracting $F(y)$ on both sides to get the bias of $\hat{F}_R(y)$ up to the first order of approximation, which is given by

$$\mathrm{Bias}(\hat{F}_R(y)) \;\approx\; F(y)\left(\frac{g(g+1)}{2}\,\alpha^2 v^2 V_{02}-\alpha v g V_{11}\right). \tag{44}$$

From Eq. (43), we can write

$$\hat{F}_R(y)-F(y) \approx F(y)(\xi_0-\alpha v g\xi_1) \tag{45}$$

Take square on both sides of Eq. (45) and then taking its expectation to get the MSE of $\hat{F}_R(y)$ under first order of approximation, which is given by

$$\mathrm{MSE}(\hat{F}_R(y)) \;\approx\; F^2(y)\left(V_{20}+\alpha^2 v^2 g^2 V_{02}-2\alpha v g V_{11}\right), \tag{46}$$

The minimum MSE at the optimum value of $(\alpha v g)$, say $(\alpha v g)_{\mathrm{opt}} = V_{11}/V_{02}$, is given by

$$\mathrm{MSE}_{\min}(\hat{F}_R(y)) \approx F^2(y)\left(V_{20}-\frac{V_{11}^2}{V_{02}}\right) \tag{47}$$

$$\approx F^2(y)V_{20}\,(1-\rho^2), \tag{48}$$

where $\rho = V_{11}/\sqrt{V_{20}V_{02}}$ is the correlation coefficient between $\hat{F}_{\mathrm{S}}(y)$ and $\hat{F}_{\mathrm{S}}(x)$ with an S sampling scheme.

4.2. Second proposed family of CDF estimators

On the lines of Singh et al. (2009), we propose another family of exponential ratio/product-type estimators for estimating the population CDF $F(y)$, given by

$$\hat{F}_E(y) \;=\; \hat{F}_{\mathrm{S}}(y)\exp\left(\frac{(agF(x)+b)-(ag\hat{F}_{\mathrm{S}}(x)+b)}{(aF(x)+b)+(a\hat{F}_{\mathrm{S}}(x)+b)}\right), \tag{49}$$

where $a = 0$ and b are either real numbers or functions of the known parameters of the auxiliary variable X, but $g \in \{-1, 1\}$. In Table 1, some members of $\hat{F}_E(y)$ are given for different values of a, b, α, and g.

In order to obtain the bias and MSE of $\hat{F}_E(y)$, express $\hat{F}_E(y)$ in terms of ξs to get

$$\begin{aligned}\hat{F}_E(y) &= F(y)(1+\xi_0)\exp\left(\frac{agF(x)-agF(x)(1+\xi_1)}{aF(x)+2b+aF(x)(1+\xi_1)}\right)\\ &= F(y)(1+\xi_0)\exp\left(-\theta g\xi_1(1+\theta\xi_1)^{-1}\right),\end{aligned} \tag{50}$$

where $\theta = aF(x)/(2aF(x)+2b)$. After expanding the RHS of Eq. (50) up to 2nd power of ξs, we have

$$\hat{F}_E(y) \;\approx\; F(y)\left(1+\xi_0-\theta g\xi_1+\frac{g(g+1)}{2}\,\theta^2\xi_1^2-\theta g\xi_0\xi_1\right). \tag{51}$$

Take expectation after subtracting $F(y)$ on both sides of Eq. (51) to get the bias of $\hat{F}_E(y)$, which under the first order of approximation is given by

$$\text{Bias}(\hat{F}_E(y)) \;\approx\; F(y)\left(\frac{g(g+1)}{2}\,\theta^2V_{02}-\theta gV_{11}\right). \tag{52}$$

From Eq. (51), we can write

$$\hat{F}_E(y)-F(y) \approx F(y)(\xi_0-\theta g\xi_1). \tag{53}$$

Take square on both sides of Eq. (53), and then take it expectation to get the MSE of $\hat{F}_E(y)$ under the first order of approximation, which is given by

$$\text{MSE}(\hat{F}_E(y)) \approx F(y)^2\left(V_{20}+\theta^2g^2V_{02}-2\theta gV_{11}\right). \tag{54}$$

The minimum MSE at the optimum value of (θg), say $(\theta g)_{\text{opt}} = V_{11}/V_{02}$, is given by

$$\text{MSE}_{\min}(\hat{F}_E(y)) \approx F^2(y)V_{20}\,(1-\rho^2), \tag{55}$$

which is equivalent to that of $\hat{F}_R(y)$.

In addition to these estimators a large number of estimators can also be generated from the proposed families of estimators $\hat{F}_R(y)$ and $\hat{F}_E(y)$ given in Eq. (41) and Eq. (49) respectively, just by putting values of a, b, α, and g.

It is observed that the expression of the first order approximation of bias and MSE/Variance of the given member of the families $\hat{F}_R(y)$ and $\hat{F}_E(y)$ can be obtained by mere substituting the values of α, g, a and b in (Eq. (44) and Eq. (46)) and (Eq. (52) and Eq. (54)), respectively. It is to be noted that, based on S scheme, the proposed families of estimators, $\hat{F}_R(y)$ and $\hat{F}_E(y)$, are more precise than $\hat{F}_R(y)$ when the following conditions hold in practice:

$$\begin{aligned}&\text{MSE}(\hat{F}_R(y)) < V(\hat{F}_S(y)) \implies \nu < \frac{2V_{11}}{(\alpha gV_{02})},\\ &\text{MSE}(\hat{F}_E(y)) < V(\hat{F}_S(y)) \implies \theta < \frac{2V_{11}}{(gV_{02})}.\end{aligned} \tag{56}$$

Table 1. *Some members of proposed families of CDF estimators.*

$\hat{F}_R(y)$	$\hat{F}_E(y)$	g	α	a	b
$\hat{F}_R^{(1)}(y)=\hat{F}_S(y)\left(\frac{F(x)}{\hat{F}_S(x)}\right)$	$\hat{F}_E^{(1)}(y)=\hat{F}_S(y)\exp\left(\frac{F(x)-\hat{F}_S(x)}{F(x)+\hat{F}_S(x)}\right)$	1	1	1	0
$\hat{F}_R^{(2)}(y)=\hat{F}_S(y)\left(\frac{F(x)+\rho_{XY}}{\hat{F}_S(x)+\rho_{XY}}\right)$	$\hat{F}_E^{(2)}(y)=\hat{F}_S(y)\exp\left(\frac{F(x)-\hat{F}_S(x)}{F(x)+\hat{F}_S(x)+2\rho_{XY}}\right)$	1	1	1	ρ_{XY}
$\hat{F}_R^{(3)}(y)=\hat{F}_S(y)\left(\frac{F(x)+C_X}{\hat{F}_S(x)+C_X}\right)$	$\hat{F}_E^{(3)}(y)=\hat{F}_S(y)\exp\left(\frac{F(x)-\hat{F}_S(x)}{F(x)+\hat{F}_S(x)+2C_X}\right)$	1	1	1	C_X
$\hat{F}_R^{(4)}(y)=\hat{F}_S(y)\left(\frac{F(x)+\beta_{2,X}}{\hat{F}_S(x)+\beta_{2,X}}\right)$	$\hat{F}_E^{(4)}(y)=\hat{F}_S(y)\exp\left(\frac{F(x)-\hat{F}_S(x)}{F(x)+\hat{F}_S(x)+2\beta_{2,X}}\right)$	1	1	1	$\beta_{2,X}$
$\hat{F}_R^{(5)}(y)=\hat{F}_S(y)\left(\frac{C_XF(x)+\beta_{2,X}}{C_X\hat{F}_S(x)+\beta_{2,X}}\right)$	$\hat{F}_E^{(5)}(y)=\hat{F}_S(y)\exp\left(\frac{C_X(F(x)-\hat{F}_S(x))}{C_X(F(x)+\hat{F}_S(x))+2\beta_{2,X}}\right)$	1	1	C_X	$\beta_{2,X}$
$\hat{F}_R^{(6)}(y)=\hat{F}_S(y)\left(\frac{\beta_{2,X}F(x)+C_X}{\beta_{2,X}\hat{F}_S(x)+C_X}\right)$	$\hat{F}_E^{(6)}(y)=\hat{F}_S(y)\exp\left(\frac{\beta_{2,X}(F(x)-\hat{F}_S(x))}{\beta_{2,X}(F(x)+\hat{F}_S(x))+2C_X}\right)$	1	1	$\beta_{2,X}$	C_X
$\hat{F}_R^{(7)}(y)=\hat{F}_S(y)\left(\frac{\rho_{XY}F(x)+C_X}{\rho_{XY}\hat{F}_S(x)+C_X}\right)$	$\hat{F}_E^{(7)}(y)=\hat{F}_S(y)\exp\left(\frac{\rho_{XY}(F(x)-\hat{F}_S(x))}{\rho_{XY}(F(x)+\hat{F}_S(x))+2C_X}\right)$	1	1	ρ_{XY}	C_X
$\hat{F}_R^{(8)}(y)=\hat{F}_S(y)\left(\frac{C_XF(x)+\rho_{XY}}{C_X\hat{F}_S(x)+\rho_{XY}}\right)$	$\hat{F}_E^{(8)}(y)=\hat{F}_S(y)\exp\left(\frac{C_X(F(x)-\hat{F}_S(x))}{C_X(F(x)+\hat{F}_S(x))+2\rho_{XY}}\right)$	1	1	C_X	ρ_{XY}
$\hat{F}_R^{(9)}(y)=\hat{F}_S(y)\left(\frac{F(x)+\beta_{1,X}}{\hat{F}_S(x)+\beta_{1,X}}\right)$	$\hat{F}_E^{(9)}(y)=\hat{F}_S(y)\exp\left(\frac{F(x)-\hat{F}_S(x)}{F(x)+\hat{F}_S(x)+2\beta_{1,X}}\right)$	1	1	1	$\beta_{1,X}$

ρ_{XY} is correlation coefficient between X and Y, C_X is coefficient of variation of X
$\beta_{1,X}$ is coefficient of skewness of X, $\beta_{2,X}$ is coefficient of kurtosis of X

4.3. Difference and regression CDF estimators

It is possible to further enhance the precision of the aforementioned families of estimators $(\hat{F}_S(y),\hat{F}_S(x))$ when the supplementary information in terms of the covariance between the CDF estimators based on Y and X, and on the variance of the CDF estimator of X are utilized.

Under a sampling scheme S, let β denote the ratio of the covariance of $\hat{F}_S(y)$ and $\hat{F}_S(x)$ to the variance of $\hat{F}_S(x)$, i.e.

$$\beta_S \;=\; \frac{C(\hat{F}_S(y),\hat{F}_S(x))}{V(\hat{F}_S(x))}. \tag{57}$$

In addition, it is also possible to have information available on the population CDF of X, say $F(x)$. The difference estimator of the population CDF $F(y)$, say $\hat{F}_D(y)$, that requires information on β_S, $\hat{F}_S(y)$ and $\hat{F}_S(x)$ is given by

$$\hat{F}_D(y) \;=\; \hat{F}_S(y)+\beta_S\left(F(x)-\hat{F}_S(x)\right), \tag{58}$$

where $\hat{F}_D(y)$ is a linear combination of $\hat{F}_S(y)$ and $\hat{F}_S(x)$. It can easily be shown that the $\hat{F}_D(y)$ is an unbiased estimator of $F(y)$.

In order to obtain the variance of $\hat{F}_D(y)$, we express $\hat{F}_D(y)$ in terms of ξs, i.e.

$$\begin{aligned} \hat{F}_D(y) &= F(y)(1+\xi_0) - \beta_S F(x)\xi_1 \\ \hat{F}_D(y) - F(y) &= F(y)\xi_0 - \beta_S F(x)\xi_1. \end{aligned} \tag{59}$$

Take square on both sides of Eq. (59) and then apply expectation to get the variance of $\hat{F}_D(y)$, which is given by

$$V(\hat{F}_D(y)) \;=\; F^2(y)V_{20} + \beta_S^2 F^2(x)V_{02} - 2\beta_S F(x)F(y)V_{11}. \tag{60}$$

The simplified expression for the variance of $\hat{F}_D(y)$, after replacing the value of β_S into $V(\hat{F}_D(y))$, is given by

$$V(\hat{F}_D(y)) \;=\; F^2(y)V_{20}(1-\rho^2), \tag{61}$$

which is equivalent to the minimum MSE of $\hat{F}_R(y)$ and $\hat{F}_E(y)$.

It is to be noted that the value of β_S may be taken from previous studies, surveys or census. In case the value of β_S is not known, then it is possible to estimate it with a large sample size. The estimated value of β_S may be obtained by replacing the covariance of $(\hat{F}_S(y), \hat{F}_S(x))$ and the variance of $\hat{F}_S(x)$ by their respective unbiased estimators, which gives

$$\hat{\beta}_S \;=\; \frac{\widehat{C}(\hat{F}_S(y), \hat{F}_S(x))}{\widehat{V}(\hat{F}_S(x))}. \tag{62}$$

It is a well-known fact under SRS that the sample covariance $\widehat{C}(\hat{F}_S(y), \hat{F}_S(x))$ and sample variance $\widehat{V}(\hat{F}_S(x))$ are weakly-consistent estimators of $C(\hat{F}_S(y), \hat{F}_S(x))$ and $V(\hat{F}_S(x))$, respectively. Thus, for a large sample size, $\hat{\beta}_S$ is also a weakly-consistent estimator of β_S.

In the survey sampling literature, the difference estimator $\hat{F}_D(y)$ with estimated value of β_S is called a regression estimator, given by

$$\hat{F}_{Reg}(y) \;=\; \hat{F}_S(y) + \hat{\beta}_S\left(F(x) - \hat{F}_S(x)\right). \tag{63}$$

It can be shown that $\hat{F}_{Reg}(y)$ is a biased estimator of $F(y)$. Moreover, for a large sample size, we have

$$\mathrm{MSE}(\hat{F}_{Reg}(y)) \;\approx\; V(\hat{F}_D(y)) \;=\; F^2(y)V_{20}(1-\rho^2). \tag{64}$$

5. Empirical Study

In this section, real datasets are considered and the relative efficiencies (REs) of the proposed CDF estimators of $F(y)$ are computed with respect to $\hat{F}_S(y)$ based on sampling scheme S.

5.1. Population I

This dataset is taken from Social & Household Integrated Economic Survey (HIES), conducted in Pakistan during the years 2011-12, which comprises 14722 households (after removing the missing observations). The entire dataset is partitioned into two strata, where Stratum-I and Stratum-II correspond to Urban and Rural (U-R) areas. These areas are further partitioned into four provinces of Pakistan, namely Punjab, Khyber Pakhtunkhwa (KPK), Sindh and Balochistan, where (Punjab - KPK) and (Sindh - Balochistan) belong to Stratum-I and Stratum-II, respectively. Moreover, where each province is further partitioned into different enumeration blocks (EBs). This dataset may be downloaded from the Pakistan Bureau of Statistics web-page via the link: https://www.pbs.gov.pk/content/microdata. The study variable Y and the auxiliary variable X are total income and total expenditure of a household (HH), respectively. Here, our objective is to estimate the proportion of HH whose yearly total income is less than or equal to $y = \$1.9 \times 365$, which is considered as the poverty line for Pakistan according to the World bank's website: https://data.worldbank.org/indicator/SI.POV.NAHC?locations=PK. The yearly total income is converted from USD to PKR by multiplying $1.9 \times 365 \times 86.3198$ PKR. For example, if the total income of a HH is less than or equal to 59862.7813 PKR, it is then considered on or below the poverty line using auxiliary variable X while $x = 226386.0582$ (yearly average expenditure of a HH). Note that (province and yearly total income of a HH) and (province, EB and yearly total income of a HH) are taken as (PSU and SSU) and (PSU, SSU and TSU) for 2SCS/S2SCS and 3SCS/S3SCS, respectively. The values of the population parameters are given below:

$$F(y) = 0.0474, F(x) = 0.6587, C_X = 0.8161,$$
$$\rho_{XY} = 0.7662, \beta_{1,X} = 4.5387 \text{ and } \beta_{2,X} = 43.4005.$$

The values of V_{rs} based on an S sampling scheme are computed and then reported in Table 2, where

$$V_{rs} = E\left(\xi_0^r \xi_1^s\right) = E\left[\left(\frac{\hat{F}_{\mathrm{S}}(y) - F(y)}{F(y)}\right)^r \left(\frac{\hat{F}_{\mathrm{S}}(x) - F(x)}{F(x)}\right)^s\right],$$

where $r, s = 0, 1, 2$.

Table 2. *The V_{rs} values based on scheme S using Population-I.*

S	$\mathrm{S_{tr}}-$Variable	PSU	SSU	TSU	n	m_i	t_{ij}	V_{20}	V_{02}	V_{11}
2SCS	$--$	Province	HH	$--$	3	40	$--$	0.31145	0.03899	0.04976
S2SCS	U/R	Province	HH	$--$	1	40	$--$	0.44729	0.12185	0.10904
3SCS	$--$	Province	EB	HH	3	15	4	0.27805	0.04275	0.05609
S3SCS	U/R	Province	EB	HH	1	15	4	0.39719	0.12749	0.11854

Note: Stratifying is abbreviated as $\mathrm{S_{tr}}$.

5.2. Population II

Another dataset is taken from Center of Disease Control (CDC), which is related to the Second National Health and Nutrition Examination Survey (NHANES-II). The NHANES sample (comprising 10351 units) represents the total non-institutionalized civilian (NIC) US population that resides in 50 states and the district of Columbia. This dataset is divided into four regions (REGs), namely southern, western, mid-western and north-eastern, where each REG is further divided into different locations (LOCs). The entire dataset is stratified into two strata, which are formed by generating random numbers from Bernoulli distribution with 0.50 as the probability of success, where 0 and 1 correspond to Stratum-I and Stratum-II, respectively. This dataset is available at the https://www.stata-press.com/data/r15/svy.html. Here, the body mass index (BMI) is taken as the study variable Y and weight is taken as the auxiliary variable X. Our objective is to estimate the proportion of people (in the NIC US population) that are under-weight, i.e., an individual is classified as under-weight if the BMI values are less than or equal to $y = 18.50$ using auxiliary variable X while $x = 71.8975$ (average weight of NIC US population) under sampling scheme S. Note that the (REG and BMI) and (REG, LOC and BMI) are taken as (PSU and SSU) and (PSU, SSU and TSU) for the 2SCS/S2SCS and 3SCS/S3SCS, respectively. The values of the population parameters are given below:

$$F(y) = 0.0318, F(x) = 0.5401, C_X = 0.2136,$$
$$\rho_{XY} = 0.8338, \beta_{1,X} = 0.7364 \text{ and } \beta_{2,X} = 4.0614.$$

The values of V_{rs} based on an sampling scheme S are computed and then reported in Table 3, where

$$V_{rs} = E\left(\xi_0^r \xi_1^s\right) = E\left[\left(\frac{\hat{F}_{\mathrm{S}}(y) - F(y)}{F(y)}\right)^r \left(\frac{\hat{F}_{\mathrm{S}}(x) - F(x)}{F(x)}\right)^s\right],$$

where $r, s = 0, 1, 2$.

Table 3. *The V_{rs} values based on scheme S using Population-II.*

S	$\mathrm{S_{tr}}-$Variable	PSU	SSU	TSU	n	m_i	t_{ij}	V_{20}	V_{02}	V_{11}
2SCS	$--$	REG	BMI	$--$	3	50	$--$	0.20634	0.00721	0.00766
S2SCS	0/1	REG	BMI	$--$	3	50	$--$	0.10207	0.00359	0.00386
3SCS	$--$	REG	LOC	BMI	3	3	50	0.07641	0.00892	0.00937
S3SCS	0/1	REG	LOC	BMI	3	3	50	0.03714	0.00476	0.00478

Using the aforementioned datasets, the REs of the CDF estimators based on a sampling scheme S are computed with different values of n, m_i and t_{ij}. The REs of the

proposed CDF estimators of $F(y)$ with auxiliary information with respect to usual unbiased CDF estimator of $\hat{F}_{\mathrm{S}}(y)$ without auxiliary information are given by

$$\mathrm{RE}_R = \frac{V(\hat{F}_{\mathrm{S}}(y))}{\mathrm{MSE}(\hat{F}_R^{(t)}(y))}, \quad \mathrm{RE}_E = \frac{V(\hat{F}_{\mathrm{S}}(y))}{\mathrm{MSE}(\hat{F}_E^{(t)}(y))}, \quad \mathrm{RE}_D = \frac{V(\hat{F}_{\mathrm{S}}(y))}{V(\hat{F}_D(y))}, \tag{65}$$

where $t = 1, 2, \ldots, 9$. The REs of these CDF estimators are reported in Tables 4 and 5.

Table 4. *REs of proposed CDF estimators with respect to $\hat{F}_S(y)$ using Population-I.*

$\hat{F}_R(y)$	2SCS	S2SCS	3SCS	S3SCS	$\hat{F}_E(y)$	2SCS	S2SCS	3SCS	S3SCS
$\hat{F}_R^{(1)}(y)$	1.2412	1.2741	1.3328	1.3810	$\hat{F}_E^{(1)}(y)$	1.1474	1.2131	1.1952	1.2791
$\hat{F}_R^{(2)}(y)$	1.1376	1.2007	1.1815	1.2616	$\hat{F}_E^{(2)}(y)$	1.0720	1.1088	1.0929	1.1374
$\hat{F}_R^{(3)}(y)$	1.1335	1.1953	1.1758	1.2540	$\hat{F}_E^{(3)}(y)$	1.0697	1.1053	1.0898	1.1329
$\hat{F}_R^{(4)}(y)$	1.0048	1.0073	1.0060	1.0089	$\hat{F}_E^{(4)}(y)$	1.0024	1.0036	1.0030	1.0045
$\hat{F}_R^{(5)}(y)$	1.0039	1.0060	1.0049	1.0073	$\hat{F}_E^{(5)}(y)$	1.0020	1.0030	1.0025	1.0037
$\hat{F}_R^{(6)}(y)$	1.2381	1.2764	1.3279	1.3829	$\hat{F}_E^{(6)}(y)$	1.1438	1.2087	1.1902	1.2728
$\hat{F}_R^{(7)}(y)$	1.1158	1.1717	1.1517	1.2213	$\hat{F}_E^{(7)}(y)$	1.0599	1.0908	1.0770	1.1140
$\hat{F}_R^{(8)}(y)$	1.1242	1.1830	1.1631	1.2369	$\hat{F}_E^{(8)}(y)$	1.0645	1.0976	1.0830	1.1228
$\hat{F}_R^{(9)}(y)$	1.0400	1.0609	1.0512	1.0758	$\hat{F}_E^{(9)}(y)$	1.0201	1.0307	1.0256	1.0379
$\hat{F}_D(y)$	1.2562	1.2790	1.3600	1.3841					

Table 5. *REs of proposed CDF estimators with respect to $\hat{F}_S(y)$ using Population-II.*

$\hat{F}_R(y)$	2SCS	S2SCS	3SCS	S3SCS	$\hat{F}_E(y)$	2SCS	S2SCS	3SCS	S3SCS
$\hat{F}_R^{(1)}(y)$	1.0409	1.0421	1.1473	1.1488	$\hat{F}_E^{(1)}(y)$	1.0292	1.0299	1.1030	1.1072
$\hat{F}_R^{(2)}(y)$	1.0244	1.0249	1.0850	1.0887	$\hat{F}_E^{(2)}(y)$	1.0134	1.0137	1.0457	1.0479
$\hat{F}_R^{(3)}(y)$	1.0365	1.0375	1.1309	1.1349	$\hat{F}_E^{(3)}(y)$	1.0226	1.0231	1.0786	1.0821
$\hat{F}_R^{(4)}(y)$	1.0083	1.0085	1.0279	1.0293	$\hat{F}_E^{(4)}(y)$	1.0043	1.0043	1.0142	1.0149
$\hat{F}_R^{(5)}(y)$	1.0020	1.0021	1.0067	1.0071	$\hat{F}_E^{(5)}(y)$	1.0010	1.0010	1.0034	1.0035
$\hat{F}_R^{(6)}(y)$	1.0402	1.0413	1.1448	1.1473	$\hat{F}_E^{(6)}(y)$	1.0273	1.0279	1.0958	1.0999
$\hat{F}_R^{(7)}(y)$	1.0355	1.0364	1.1269	1.1310	$\hat{F}_E^{(7)}(y)$	1.0216	1.0221	1.0749	1.0783
$\hat{F}_R^{(8)}(y)$	1.0086	1.0088	1.0289	1.0303	$\hat{F}_E^{(8)}(y)$	1.0044	1.0045	1.0147	1.0154
$\hat{F}_R^{(9)}(y)$	1.0258	1.0264	1.0903	1.0942	$\hat{F}_E^{(9)}(y)$	1.0143	1.0146	1.0489	1.0513
$\hat{F}_D(y)$	1.0410	1.0424	1.1477	1.1488					

It can be seen that the proposed CDF estimators under complex survey sampling with auxiliary information are slightly more efficient than those that are without the auxiliary information, that is, all values of the REs are greater than one. It can also be seen that the proposed CDF estimators under a sampling scheme S with stratification are slightly more efficient than those without stratification and the REs tend to increase with increasing the sampling stages. Generally, with an increase in the sample size at the primary, secondary or tertiary stage of sampling, the REs may tend to increase and vice versa. Among all estimators, as expected, the REs of $\hat{F}_D(y)$ are higher than those of other considered CDF estimators.

It is to be noted that the proposed families of estimators, $\hat{F}_R(y)$ and $\hat{F}_E(y)$, are conditionally better than $\hat{F}_S(y)$, i.e. when the conditions given in Eq. (56) hold. However, the difference and regression estimators, $\hat{F}_D(y)$ and $\hat{F}_{Reg}(y)$, respectively, are always more precise than $\hat{F}_S(y)$, $\hat{F}_R(y)$ and $\hat{F}_E(y)$. In usual practice, if no information is available to check these conditions, it is preferable to use $\hat{F}_{Reg}(y)$ when estimating the population CDF under scheme S.

6. Conclusion

In this paper, we have considered the problem of estimating the finite population CDF in 2SCS and 3SCS schemes with and without stratification. Two families of classical ratio/product-type and exponential ratio/product-type CDF estimators have been proposed that require supplementary information on a single auxiliary variable. In addition, difference and regression estimators of the CDF have also been proposed. Explicit mathematical expressions of the biases and MSEs of the proposed CDF estimators have been developed under first order of the approximation. Real datasets were also considered to support the proposed theory.

Along the lines of Nematollahi et al. (2008) and Haq et al. (2021), it is also possible to increase the precision of proposed families of the CDF estimators by employing RSS and double RSS schemes in the secondary and tertiary sampling frames. Moreover, the current work may be extended to develop new CDF estimators that require supplementary information on two or more auxiliary variables. In addition, it may be possible to develop the CDF estimators when using probability proportional to size sampling to select units at the first stage of sampling under the 2SCS/3SCS and S2SCS/S3SCS schemes.

Acknowledgments

The authors are thankful to the editor and two anonymous reviewers for providing many useful comments that led to an improved version of the article.

Appendix

In this Appendix, we present the proofs of the Lemmas in Section 3.

1: Proof of Lemma 1

Here, the indices 1 and 2 are used for the first-stage and second-stage of sampling under 2SCS, respectively.

1. The covariance between $\hat{F}_{2S}(y)$ and $\hat{F}_{2S}(x)$ can be written as:

$$C(\hat{F}_{2S}(y),\hat{F}_{2S}(x)) \;=\; C_1[E_2(\hat{F}_{2S}(y),\hat{F}_{2S}(x))]+E_1[C_2(\hat{F}_{2S}(y),\hat{F}_{2S}(x))]. \tag{66}$$

It can be shown that $E_2(\hat{F}_{2S}(y))=\sum_{i=1}^{n}M_iF_i(y)/(n\overline{M})$. Based on this result, we have

$$C_1\left[E_2(\hat{F}_{2S}(y),\hat{F}_{2S}(x))\right]=C_1\left(\frac{1}{n\overline{M}}\sum_{i=1}^{n}M_iF_i(y),\frac{1}{n\overline{M}}\sum_{i=1}^{n}M_iF_i(x)\right) \;=\; \frac{\lambda\sigma_{XY,2b}}{n\overline{M}^2} \tag{67}$$

$$\begin{aligned}E_1\left[C_2(\hat{F}_{2S}(y),\hat{F}_{2S}(x))\right] &= E_1\left[\frac{1}{n^2\overline{M}^2}\sum_{i=1}^{n}M_i^2C_2(\hat{F}_i(y),\hat{F}_i(x))\right],\\ &= \frac{1}{nN\overline{M}^2}\sum_{i=1}^{N}\frac{\lambda_iM_i^2\sigma_{XY,2i}}{m_i},\end{aligned} \tag{68}$$

which completes the proof.

2. An unbiased estimator of $C\left(\hat{F}_{2S}(y),\hat{F}_{2S}(x)\right)$ is given by

$$\widehat{C}\left(\hat{F}_{2S}(y),\hat{F}_{2S}(x)\right)=\frac{\lambda\hat{\sigma}_{XY,2b}}{n\overline{M}^2}+\frac{1}{nN\overline{M}^2}\sum_{i=1}^{n}\frac{\lambda_iM_i^2\hat{\sigma}_{XY,2i}}{m_i}, \tag{69}$$

From Eq. (25), we can write

$$\hat{\sigma}_{XY,2b} \;=\; \frac{n}{n-1}\left[\frac{1}{n}\sum_{i=1}^{n}(M_i\hat{F}_i(y)M_i\hat{F}_i(x))-\overline{M}\hat{F}_{2S}(y)\overline{M}\hat{F}_{2S}(x)\right]. \tag{70}$$

Consider the mathematical expectation on the RHS of Eq. (70) to get:

$$\begin{aligned}E\left[\frac{1}{n}\sum_{i=1}^{n}\left(M_i\hat{F}_i(y)M_i\hat{F}_i(x)\right)\right] &= E_1\left[\frac{1}{n}\sum_{i=1}^{n}E_2\left(M_i\hat{F}_i(y)M_i\hat{F}_i(x)\right)\right]\\ &= E_1\left[\frac{1}{n}\sum_{i=1}^{n}\left(C_2(M_i\hat{F}_i(y),M_i\hat{F}_i(x))\right)\right.\\ &\qquad \left.+\frac{1}{n}\sum_{i=1}^{n}\left(E_2(M_i\hat{F}_i(y))E_2(M_i\hat{F}_i(x))\right)\right]\end{aligned}$$

$$= E_1\left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\lambda_i M_i^2\sigma_{XY,2i}}{m_i} + M_iF_i(y)M_iF_i(x)\right)\right]$$
$$= \frac{1}{N}\sum_{i=1}^{N}\frac{\lambda_i M_i^2\sigma_{XY,2i}}{m_i} + \frac{1}{N}\sum_{i=1}^{N}M_iF_i(y)M_iF_i(x), \tag{71}$$

and

$$E\left[\left(\overline{M}\hat{F}_{2S}(y)\overline{M}\hat{F}_{2S}(x)\right)\right] = C\left(\overline{M}\hat{F}_{2S}(y),\overline{M}\hat{F}_{2S}(x)\right) + E(\overline{M}\hat{F}_{2S}(y))E(\overline{M}\hat{F}_{2S}(x))$$
$$= \frac{\lambda\sigma_{XY,2b}}{n} + \frac{1}{nN}\sum_{i=1}^{N}\frac{\lambda_i M_i^2\sigma_{XY,2i}}{m_i} + \overline{M}F(y)\overline{M}F(x). \tag{72}$$

Using Eqs. (71) and (72) and then take the mathematical expectation of Eq. (25) to get:

$$E(\hat{\sigma}_{XY,2b}) \;=\; \sigma_{XY,2b} + \frac{1}{N}\sum_{i=1}^{N}\frac{\lambda_i M_i^2\sigma_{XY,2i}}{m_i}, \tag{73}$$

which shows that $\hat{\sigma}_{XY,2b}$ is a biased estimator of $\sigma_{XY,2b}$.

Similarly, from Eq. (27), we have

$$\hat{\sigma}_{XY,2i} \;=\; \frac{m_i}{m_i-1}\left[\frac{1}{m_i}\sum_{j=1}^{m_i}\Big(I(Y_{i,j}\le y)I(X_{i,j}\le x)\Big) - \hat{F}_i(y)\hat{F}_i(x)\right]. \tag{74}$$

Consider the mathematical expectation on the RHS of Eq. (74) to get:

$$E_2\left[\frac{1}{m_i}\sum_{j=1}^{m_i}\left(I(Y_{ij}\le y)I(X_{ij}\le x)\right)\right] = \frac{1}{M_i}\sum_{j=1}^{M_i}\left(I(Y_{ij}\le y)I(X_{ij}\le x)\right) \tag{75}$$
$$E_2\left[\hat{F}_i(y)\hat{F}_i(x)\right] = C_2\left(\hat{F}_i(y),\hat{F}_i(x)\right) - E_2(\hat{F}_i(y))E_2(\hat{F}_i(x))$$
$$= \frac{\lambda_i\sigma_{XY,2i}}{m_i} - F_i(y)F_i(x). \tag{76}$$

Using Eqs. (75)-(76) and then take the expectation of Eq. (27) to get

$$E_2(\hat{\sigma}_{XY,i}) \;=\; \sigma_{XY,2i}, \tag{77}$$

which shows that $\hat{\sigma}_{XY,2i}$ is an unbiased estimator of $\sigma_{XY,2i}$.

Now take the mathematical expectation of Eq. (69), and use the results given in Eqs. (73) and (77) to show that

$$E\left[\widehat{C}\left(\hat{F}_{2S}(y),\hat{F}_{2S}(x)\right)\right] \;=\; C\left(\hat{F}_{2S}(y),\hat{F}_{2S}(x)\right), \tag{78}$$

which completes the proof.

2: Proof of Lemma 3

Here, the indices 1, 2 and 3 are used for the first stage, second stage and third stage of sampling under 3SCS, respectively.

1. The covariance between $\hat{F}_{3S}(y)$ and $\hat{F}_{3S}(x)$ can be written as:

$$C\left(\hat{F}_{3S}(y),\hat{F}_{3S}(y)\right) = C_1E_2E_3\left[\hat{F}_{3S}(y),\hat{F}_{3S}(x)\right] + E_1C_2E_3\left[\hat{F}_{3S}(y),\hat{F}_{3S}(x)\right] + E_1E_2C_3\left[\hat{F}_{3S}(y),\hat{F}_{3S}(x)\right]. \quad (79)$$

It can be shown that $E_3(\hat{F}_{3S}(y)) = \sum_{i=1}^{n}(M_i/m_i)\sum_{j=1}^{m_i} T_{ij}F_{ij}(y)/n\overline{T}$. Based on this result, we have

$$\begin{aligned}
C_1E_2E_3\left[\hat{F}_{3S}(y),\hat{F}_{3S}(x)\right] &= C_1E_2\left[\frac{1}{n\overline{T}}\sum_{i=1}^{n}\frac{M_i}{m_i}\sum_{j=1}^{m_i}T_{ij}F_{ij}(y),\frac{1}{n\overline{T}}\sum_{i=1}^{n}\frac{M_i}{m_i}\sum_{j=1}^{m_i}T_{ij}F_{ij}(x)\right] \\
&= C_1\left[\frac{1}{n\overline{T}}\sum_{i=1}^{n}M_iF_i(y),\frac{1}{n\overline{T}}\sum_{i=1}^{n}M_iF_i(x)\right] = \frac{\lambda\sigma_{XY,3b}}{n\overline{T}^2}. \quad (80) \\
E_1C_2E_3\left[\hat{F}_{3S}(y),\hat{F}_{3S}(x)\right] &= E_1C_2\left[\frac{1}{n\overline{T}}\sum_{i=1}^{n}\frac{M_i}{m_i}\sum_{j=1}^{m_i}T_{ij}F_{ij}(y),\frac{1}{n\overline{T}}\sum_{i=1}^{n}\frac{M_i}{m_i}\sum_{j=1}^{m_i}T_{ij}F_{ij}(x)\right] \\
&= E_1\left[\frac{1}{n^2\overline{T}^2}\sum_{i=1}^{n}M_i^2\,C_2\Big(\frac{1}{m_i}\sum_{j=1}^{m_i}T_{ij}F_{ij}(y),\frac{1}{m_i}\sum_{j=1}^{m_i}T_{ij}F_{ij}(x)\Big)\right] \\
&= E_1\left[\frac{1}{n^2\overline{T}^2}\sum_{i=1}^{n}\frac{\lambda_iM_i^2\sigma_{XY,3i}}{m_i}\right] \\
&= \frac{1}{nN\overline{T}^2}\sum_{i=1}^{N}\frac{\lambda_iM_i^2\sigma_{XY,3i}}{m_i}. \quad (81) \\
E_1E_2C_3\left[\hat{F}_{3S}(y),\hat{F}_{3S}(x)\right] &= E_1E_2\left[\frac{1}{n^2\overline{T}^2}\sum_{i=1}^{n}\frac{M_i^2}{m_i^2}\sum_{j=1}^{m_i}T_{ij}^2\,C_3\big(\hat{F}_{ij}(y),\hat{F}_{ij}(x)\big)\right] \\
&= E_1E_2\left[\frac{1}{n^2\overline{T}^2}\sum_{i=1}^{n}\frac{M_i^2}{m_i^2}\sum_{j=1}^{m_i}\frac{\lambda_{ij}T_{ij}^2\sigma_{XY,3ij}}{t_{ij}}\right] \\
&= E_1\left[\frac{1}{n^2\overline{T}^2}\sum_{i=1}^{n}\frac{M_i}{m_i}\sum_{j=1}^{M_i}\frac{\lambda_{ij}T_{ij}^2\sigma_{XY,3ij}}{t_{ij}}\right] \\
&= \frac{1}{nN\overline{T}^2}\sum_{i=1}^{N}\frac{M_i}{m_i}\sum_{j=1}^{M_i}\frac{\lambda_{ij}T_{ij}^2\sigma_{XY,3ij}}{t_{ij}}. \quad (82)
\end{aligned}$$

Add Eqs. (80)–(82), which completes the proof.

2. An unbiased estimator of $C(\hat{F}_{3S}(y),\hat{F}_{3S}(x))$ is given by

$$\widehat{C}(\hat{F}_{3S}(y),\hat{F}_{3S}(x)) = \frac{\lambda\,\hat{\sigma}_{XY,3b}}{n\overline{T}^2} + \frac{1}{nN\overline{T}^2}\sum_{i=1}^{n}\frac{\lambda_i M_i^2\hat{\sigma}_{XY,3i}}{m_i} + \frac{1}{nN\overline{T}^2}\sum_{i=1}^{n}\frac{M_i}{m_i}\sum_{j=1}^{m_i}\frac{\lambda_{ij}T_{ij}^2\hat{\sigma}_{XY,3ij}}{t_{ij}}, \tag{83}$$

From Eq. (33), we can write

$$\hat{\sigma}_{\mathrm{XY},3b} = \frac{n}{n-1}\left[\frac{1}{n}\sum_{i=1}^{n}\left(M_i\hat{F}_i(y)M_i\hat{F}_i(x)\right) - \left(\overline{T}\hat{F}_{3S}(y)\overline{T}\hat{F}_{3S}(x)\right)\right]. \tag{84}$$

Consider the mathematical expectation of the RHS of the above equation to get:

$$\begin{aligned}
E\left[\frac{1}{n}\sum_{i=1}^{n}(M_i\hat{F}_i(y)M_i\hat{F}_i(x))\right] &= E_1E_2\left[\frac{1}{n}\sum_{i=1}^{n}E_3\left(M_i\hat{F}_i(y)M_i\hat{F}_i(x)\right)\right]\\
&= E_1E_2\left[\frac{1}{n}\sum_{i=1}^{n}\left(C_3\left(M_i\hat{F}_i(y),M_i\hat{F}_i(x)\right)\right)\right.\\
&\quad\left.+\frac{1}{n}\sum_{i=1}^{n}\left\{E_3(M_i\hat{F}_i(y))E_3(M_i\hat{F}_i(x))\right\}\right]\\
&= E_1E_2\left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{M_i^2}{m_i^2}\sum_{j=1}^{m_i}C_3\left(\hat{F}_{ij}(y),\hat{F}_{ij}(x)\right)\right)\right.\\
&\quad\left.+\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{M_i}{m_i}\sum_{j=1}^{m_i}E_3(T_{ij}\hat{F}_{ij}(y))\frac{M_i}{m_i}\sum_{j=1}^{m_i}E_3(T_{ij}\hat{F}_{ij}(x))\right\}\right]\\
&= E_1E_2\left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{M_i^2}{m_i^2}\sum_{j=1}^{m_i}\frac{\lambda_{ij}T_{ij}^2\sigma_{XY,3ij}}{t_{ij}}\right)\right.\\
&\quad\left.+\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{M_i}{m_i}\sum_{j=1}^{m_i}T_{ij}F_{ij}(y)\frac{M_i}{m_i}\sum_{j=1}^{m_i}T_{ij}F_{ij}(x)\right\}\right]\\
&= E_1\left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{M_i}{m_i}\sum_{j=1}^{M_i}\frac{\lambda_{ij}T_{ij}^2\sigma_{XY,3ij}}{t_{ij}}\right)\right.\\
&\quad\left.+\frac{1}{n}\sum_{i=1}^{n}\left\{E_2\left(\frac{M_i}{m_i}\sum_{j=1}^{m_i}T_{ij}F_{ij}(y)\frac{M_i}{m_i}\sum_{j=1}^{m_i}T_{ij}F_{ij}(x)\right)\right\}\right]
\end{aligned}$$

$$= E_1\left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{M_i}{m_i}\sum_{j=1}^{M_i}\frac{\lambda_{ij}T_{ij}^2\sigma_{XY,3ij}}{t_{ij}}+\frac{\lambda_i M_i^2\sigma_{XY,3i}}{m_i}\right)\right.$$
$$\left.+\frac{1}{n}\sum_{i=1}^{n}\left(M_iF_i(y)M_iF_i(x)\right)\right]$$
$$= \frac{1}{N}\sum_{i=1}^{N}\frac{M_i}{m_i}\sum_{j=1}^{M_i}\frac{\lambda_{ij}T_{ij}^2\sigma_{XY,3ij}}{t_{ij}}+\frac{1}{N}\sum_{i=1}^{N}\frac{\lambda_i M_i^2\sigma_{XY,3i}}{m_i}$$
$$+\frac{1}{N}\sum_{i=1}^{N}M_iF_i(y)M_iF_i(x), \quad \text{and} \tag{85}$$

$$E\left[\left(\overline{T}\hat{F}_{3S}(y)\overline{T}\hat{F}_{3S}(x)\right)\right] = C\left(\overline{T}\hat{F}_{3S}(y),\overline{T}\hat{F}_{3S}(x)\right)+E(\overline{T}\hat{F}_{3S}(y))E(\overline{T}\hat{F}_{3S}(x)),$$
$$= \frac{\lambda\sigma_{XY,3b}}{n}+\frac{1}{nN}\sum_{i=1}^{N}\frac{\lambda_i M_i^2\sigma_{XY,3i}}{m_i}+\frac{1}{nN}\sum_{i=1}^{N}\frac{M_i}{m_i}\sum_{j=1}^{M_i}\frac{\lambda_{ij}T_{ij}^2\sigma_{XY,3ij}}{t_{ij}}$$
$$+\overline{T}F(y)\overline{T}F(x). \tag{86}$$

Using Eqs. (85) and (86) in Eq. (33), and then take expectation to show that

$$E(\hat{\sigma}_{XY,3b}) \;=\; \sigma_{XY,3b}+\frac{1}{N}\sum_{i=1}^{N}\frac{\lambda_i M_i^2\sigma_{XY,3i}}{m_i}+\frac{1}{N}\sum_{i=1}^{N}\frac{M_i}{m_i}\sum_{j=1}^{M_i}\frac{\lambda_{ij}T_{ij}^2\sigma_{XY,3ij}}{t_{ij}}, \tag{87}$$

which shows that $\hat{\sigma}_{XY,3b}$ is a biased estimator of $\sigma_{XY,3b}$.

Similarly, we can write from Eq. (35):

$$\hat{\sigma}_{XY,3i} \;=\; \frac{m_i}{m_i-1}\left[\frac{1}{m_i}\sum_{j=1}^{m_i}(T_{ij}\hat{F}_{ij}(y)T_{ij}\hat{F}_{ij}(x))-(\hat{F}_i(y)\hat{F}_i(x))\right]. \tag{88}$$

Consider the mathematical expectation on the RHS of the above equation to get:

$$E_2\left[\frac{1}{m_i}\sum_{j=1}^{m_i}(T_{ij}\hat{F}_{ij}(y)T_{ij}\hat{F}_{ij}(x))\right] = E_2\left[\frac{1}{m_i}\sum_{j=1}^{m_i}E_3\left(T_{ij}\hat{F}_{ij}(y)T_{ij}\hat{F}_{ij}(x)\right)\right]$$
$$= E_2\left[\frac{1}{m_i}\sum_{j=1}^{m_i}\left(T_{ij}^2\,C_3\left(\hat{F}_{ij}(y),\hat{F}_{ij}(x)\right)\right)\right.$$
$$\left.+\frac{1}{m_i}\sum_{j=1}^{m_i}\left\{E_3\left(T_{ij}\hat{F}_{ij}(y)\right)E_3\left(T_{ij}\hat{F}_{ij}(x)\right)\right\}\right]$$
$$= E_2\left[\frac{1}{m_i}\sum_{j=1}^{m_i}\left(\frac{\lambda_{ij}T_{ij}^2\sigma_{XY,3ij}}{t_{ij}}+T_{ij}F_{ij}(y)T_{ij}F_{ij}(x)\right)\right]$$
$$= \frac{1}{M_i}\sum_{j=1}^{M_i}\frac{\lambda_{ij}T_{ij}^2\sigma_{XY,3ij}}{t_{ij}}+\frac{1}{M_i}\sum_{j=1}^{M_i}T_{ij}F_{ij}(y)T_{ij}F_{ij}(x) \tag{89}$$

and

$$\begin{aligned}
E_2\left[(\hat{F}_i(y)\hat{F}_i(x))\right] &= E_2\left[E_3\big(\hat{F}_i(y)\hat{F}_i(x)\big)\right] \\
&= E_2\left[C_3\big(\hat{F}_i(y),\hat{F}_i(x)\big)+E_3\big(\hat{F}_i(y)\big)E_3\big(\hat{F}_i(x)\big)\right] \\
&= E_2\left[\frac{T_{ij}^2}{m_i^2}\sum_{j=1}^{m_i}C_3\big(\hat{F}_{ij}(y),\hat{F}_{ij}(x)\big)+\frac{1}{m_i}\sum_{j=1}^{m_i}T_{ij}F_{ij}(y)\frac{1}{m_i}\sum_{j=1}^{m_i}T_{ij}F_{ij}(x)\right] \\
&= \frac{1}{m_iM_i}\sum_{j=1}^{M_i}\frac{\lambda_{ij}T_{ij}^2\sigma_{XY,3ij}}{t_{ij}}+E_2\Big(\frac{1}{m_i}\sum_{j=1}^{m_i}T_{ij}F_{ij}(y)\frac{1}{m_i}\sum_{j=1}^{m_i}T_{ij}F_{ij}(x)\Big) \\
&= \left[\frac{1}{m_iM_i}\sum_{j=1}^{M_i}\frac{\lambda_{ij}T_{ij}^2\sigma_{XY,3ij}}{t_{ij}}+C_2\Big(\frac{1}{m_i}\sum_{j=1}^{m_i}T_{ij}F_{ij}(y),\frac{1}{m_i}\sum_{j=1}^{m_i}T_{ij}F_{ij}(x)\Big)\right. \\
&\quad \left.+\left\{E_2\big(\frac{1}{m_i}\sum_{j=1}^{m_i}T_{ij}F_{ij}(y)\big)E_2\big(\frac{1}{m_i}\sum_{j=1}^{m_i}T_{ij}F_{ij}(x)\big)\right\}\right] \\
&= \frac{1}{m_iM_i}\sum_{j=1}^{M_i}\frac{\lambda_{ij}T_{ij}^2\sigma_{XY,3ij}}{t_{ij}}+\frac{\lambda_i\sigma_{XY,3i}}{m_i}+F_i(y)F_i(x) \qquad (90)
\end{aligned}$$

Using Eqs. (89) and (90) in Eq. (35), and then take expectation to show that

$$E_2(\hat{\sigma}_{XY,3i}) \;=\; \sigma_{XY,3i}+\frac{1}{M_i}\sum_{j=1}^{M_i}\frac{\lambda_{ij}T_{ij}^2\sigma_{XY,3ij}}{t_{ij}}, \qquad (91)$$

which shows that $\hat{\sigma}_{XY,3i}$ is also a biased estimator of $\sigma_{XY,3i}$.

Similarly, we can write from Eq. (37):

$$\hat{\sigma}_{XY,3ij}=\frac{t_{ij}}{t_{ij}-1}\left[\frac{1}{t_{ij}}\sum_{k=1}^{t_{ij}}(I(Y_{ij,k}\le y)I(X_{ij,k}\le x))-\hat{F}_{ij}(x)\hat{F}_{ij}(y)\right]. \qquad (92)$$

Consider the RHS of Eq. (92):

$$E_3\left[\frac{1}{t_{ij}}\sum_{k=1}^{t_{ij}}\Big(I(Y_{ij,k}\le y)I(X_{ij,k}\le x)\Big)\right] \;=\; \frac{1}{T_{ij}}\sum_{k=1}^{T_{ij}}\big(I(Y_{ij,k}\le y)I(X_{ij,k}\le x)\big), \qquad (93)$$

and

$$\begin{aligned}
E_3\left[(\hat{F}_{ij}(x)\hat{F}_{ij}(y))\right] &= C_3\big(\hat{F}_{ij}(x),\hat{F}_{ij}(y)\big)+E_3\big(\hat{F}_{ij}(x)\big)E_3\big(\hat{F}_{ij}(y)\big) \\
&= \frac{\lambda_{ij}\sigma_{XY,3ij}}{t_{ij}}+F_{ij}(y)F_{ij}(x). \qquad (94)
\end{aligned}$$

Use Eqs. (93) and (94) in Eq. (37), and then take expectation to show that

$$E_3(\hat{\sigma}_{XY,3ij}) \;=\; \sigma_{XY,3ij} \qquad (95)$$

which show that $\hat{\sigma}_{XY,3ij}$ is an unbiased estimator of $\sigma_{XY,3ij}$. Now Eq. (83) follows from the results given in Eqs. (87), (91) and (95), which completes the proof.

References

Berger, Y. G. and Muñoz, J. F. (2015). On estimating quantiles using auxiliary information. *Journal of Official Statistics*, 31(1):101–119.

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

Cochran, W. G. (1977). *Sampling Techniques, 3rd Edition.* John Wiley.

Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418):376–382.

Francisco, C. A. and Fuller, W. A. (1986). Estimation of the distribution function with a complex survey. In *JSM Proceedings, Survey Research Methods Section, American Statistical Association*, pages 37–45.

Hansen, M. H. and Hurwitz, W. N. (1943). On the theory of sampling from finite populations. *The Annals of Mathematical Statistics*, 14(4):333–362.

Haq, A. (2017a). Estimation of the distribution function under hybrid ranked set sampling. *Journal of Statistical Computation and Simulation*, 87(2):313–327.

Haq, A. (2017b). Two-stage cluster sampling with hybrid ranked set sampling in the secondary sampling frame. *Communications in Statistics-Theory and Methods*, 46(17):8450–8467.

Haq, A., Abbas, M., and Khan, M. (2021). Estimation of finite population distribution function in a complex survey sampling. *Communications in Statistics - Theory and Methods*, 0(0):1–23.

Hussain, S., Ahmad, S., Saleem, M., and Akhtar, S. (2020). Finite population distribution function estimation with dual use of auxiliary information under simple and stratified random sampling. *Plos one*, 15(9):e0239098.

Khoshnevisan, M., Singh, R., Chauhan, P., and Sawan, N. (2007). A general family of estimators for estimating population mean using known value of some population parameter (s). *Far East Journal of Theoretical Statistics*, 22(2):181–191.

Lee, S. E., Lee, P. R., and Shin, K.-I. (2016). A composite estimator for stratified two stage cluster sampling. *Communications for Statistical Applications and Methods*, 23(1):47–55.

Martínez, S., Rueda, M., Arcos, A., and Martínez, H. (2010). Optimum calibration points estimating distribution functions. *Journal of Computational and Applied Mathematics*, 233(9):2265–2277.

Martínez, S., Rueda, M., and Illescas, M. (2022). The optimization problem of quantile and poverty measures estimation based on calibration. *Journal of Computational and Applied Mathematics*, 405(15):113054.

Mayor-Gallego, J. A., Moreno-Rebollo, J. L., and Jiménez-Gamero, M. D. (2019). Estimation of the finite population distribution function using a global penalized calibration method. *AStA Advances in Statistical Analysis*, 103(1):1–35.

Murthy, M. N. (1967). *Sampling Theory and Methods*. Calcutta-35: Statistical Publishing Society.

Nafiu, L., Oshungade, I., and Adewara, A. (2012). Alternative estimation method for a three-stage cluster sampling in finite population. *American Journal of Mathematics and Statistics*, 2(6):199–205.

Nematollahi, N., M, M. S., and Saba, R. A. (2008). Two-stage cluster sampling with ranked set sampling in the secondary sampling frame. *Communications in Statistics - Theory and Methods*, 37(15):2404–2415.

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

Rustagi, R. K. (1978). *Some theory of the prediction approach to two stage and stratified two stage cluster sampling*. PhD thesis, The Ohio State University.

Sahoo, L. (1987). A regression-type estimator in two-stage sampling. *Calcutta Statistical Association Bulletin*, 36(1-2):97–100.

Särndal, C.-E., Swensson, B., and Wretman, J. (2003). *Model Assisted Survey Sampling*. Springer Science & Business Media.

Singh, H. P., Singh, S., and Kozak, M. (2008). A family of estimators of finite-population distribution function using auxiliary information. *Acta Applicandae Mathematicae*, 104(2):115–130.

Singh, R., Chauhan, P., Sawan, N., and Smarandache, F. (2009). Improvement in estimating the population mean using exponential estimator in simple random sampling. *International Journal of Statistics & Economics*, 3(A09):13–18.

Smith, T. (1969). A note on ratio estimates in multistage sampling. *Journal of the Royal Statistical Society: Series A (General)*, 132(3):426–430.

Srivastava, M. and Garg, N. (2009). A general class of estimators of a finite population mean using multi-auxiliary information under two stage sampling scheme. *Journal of Reliability and Statistical Studies*, 2(1):103–118.

Stokes, S. L. and Sager, T. W. (1988). Characterization of a ranked-set sample with application to estimating distribution functions. *Journal of the American Statistical Association*, 83(402):374–381.

Sukhatme, P. V., Sukhatme, B., Sukhatme, S., and Asok, C. (1984). Sampling theory with applications. *Indian Society of Agricultural Statistics, New Delhi & IOWA State University Press, Ames, USA*.

Tarima, S. and Pavlov, D. (2006). Using auxiliary information in statistical function estimation. *ESAIM: Probability and Statistics*, 10:11–23.

Yaqub, M. and Shabbir, J. (2020). Estimation of population distribution function involving measurement error in the presence of non response. *Communications in Statistics - Theory and Methods*, 49(10):2540–2559.