Estimating the distance distribution of subpopulations for a large-scale complex survey

Jianqiang Wang
Department of Statistics
Iowa State University

Jean Opsomer
Professor, Department of Statistics
Iowa State University

Abstract

Many finite populations targeted by sample surveys consist of homogenous subpopulations with respect to the variables being collected. In on-going survey operations, it is often of interest to be able to assess whether a new observation belongs to one of those subpopulations or should be flagged as not belonging to any of them. For this purpose, we propose a sample-based estimator for the subpopulation distribution functions of the distances between the elements and the subpopulation centers. We explore different ways to define the subpopulation centers and several distance metrics. We describe the theoretical properties of the estimator, and propose several approaches for design-based variance estimation. The practical properties of the procedures are evaluating in a simulation study.

Keywords: jackknife, kernel estimation, estimating equations, generalized median, elliptical distribution, classification.

1. Introduction

A common issue in large-scale complex surveys is the detection of outliers in the data. Such outliers can be caused by frame imperfections, which can lead to ineligible units being selected, or by errors during data collection. If these outliers remain in the survey dataset, they can cause inference based on the survey to be invalid for the population of interest. Most survey operations therefore incorporate data editing and validation as part of the post-data collection steps, where they attempt to identify suspicious observations and either remove or correct them. When outliers exhibit "extreme" values on one or several survey variables, they can be detected relatively easily. However, because surveys often collect large numbers of variables, there is the potential for other outliers which are not extreme on any single variable. Detecting such outliers is more difficult, and identifying unusual or suspicious patterns in the data often requires substantial subject-matter knowledge.

In practice, many finite populations targeted by surveys consist of a number of relatively homogeneous subpopulations, and this structure can be exploited to develop a statistical approach for flagging suspicious observations that does not require detailed knowledge of the relationships between the variables. The type of surveys we are targeting here is one in which recurring surveys are made over time, and the characteristics and composition of the subpopulations remains relatively stable between surveys. The focus of the current article is to propose an estimator for the "outlyingness" of an observation relative to a subpopulation, based on the distance between an observation and the center of the subpopulation, and to derive its statistical properties in an asymptotic design-based context.

The remainder of the paper is as follows. Section 2 defines notation and establishes preliminary results for design-based inference. Section 3 lists our general design assumptions and assumptions on the sequence of finite populations. Section 4 presents our theoretical results showing asymptotic properties of distance distribution functions using either means or medians as subpopulation centers. Simulation results are detailed in Wang (2008) due to space limits.

2. Notation, assumption and preliminary results

2.1 Notation and definitions

Suppose we have an increasing sequence of finite populations $(U_\nu)_{\nu=1}^{\infty}$ of sizes N_ν with $N_\nu < N_{\nu+1}$. Associated with the i-th population element is a p-dimensional vector of observations

$$\boldsymbol{y}_i = (y_{i,1}, ..., y_{i,p}),$$

and let $\mathcal{F}_\nu$ be the power set of ν-th finite population $\{\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_{N_\nu}\}$. The finite population U_ν can be partitioned into G subpopulations, with $U_\nu = \bigcup_{g=1}^{G} U_{\nu g}$. The knowledge of this structure of finite population is usually available for longitudinal surveys of stable populations, and we assume the partition into subpopulations is provided.

We take a sample $\mathcal{S}_\nu = \bigcup_{g=1}^{G} \mathcal{S}_{\nu g}$ of size n_ν from population U_ν, and the sampling design may be a complex design with stratification or multi-stage sampling. Here, $\mathcal{S}_{\nu g}$ are mutually exclusive subsets of $\mathcal{S}_\nu$ which contain elements from subpopulation g only. We assume the inclusion probability of the i-th population element, is known without error,

$$\mathrm{P}(i \in \mathcal{S}_\nu) = \pi_i,$$

and the sample size

$$n_\nu = \sum_{g=1}^{G} n_{\nu g},$$

where $n_{\nu g}$ is the number of elements in sample $\mathcal{S}_\nu$ that come from subpopulation g. We use $n_\nu^* = \mathrm{E}(n_\nu|\mathcal{F}_\nu)$ to denote the expected sample size conditioning on finite population and $n_{\nu g}^*$ to denote the expected sample size for subpopulation g.

We define the population level distribution of distances as

$$D_{\nu g,d}(\boldsymbol{\gamma}_{\nu g}) = \frac{1}{N_{\nu g}} \sum_{U_{\nu g}} \mathrm{I}_{(\|\boldsymbol{y}_i - \boldsymbol{\gamma}_{\nu g}\| \le d)}, \tag{1}$$

where $\boldsymbol{\gamma}_{\nu g}$ is some measure of center of population U_ν. Given a sample $\mathcal{S}_\nu$, $D_{\nu g,d}(\boldsymbol{\gamma}_{\nu g})$ is estimated by,

$$\widehat{D}_{\nu g,d}(\hat{\boldsymbol{\gamma}}_{\nu g}) = \frac{1}{\widehat{N}_{\nu g}} \sum_{\mathcal{S}_{\nu g}} \frac{1}{\pi_i} \mathrm{I}_{(\|\boldsymbol{y}_i - \hat{\boldsymbol{\gamma}}_{\nu g}\| \le d)}, \tag{2}$$

or

$$\widetilde{D}_{\nu g,d}(\hat{\boldsymbol{\gamma}}_{\nu g}) = \frac{1}{N_{\nu g}} \sum_{\mathcal{S}_{\nu g}} \frac{1}{\pi_i} \mathrm{I}_{(\|\boldsymbol{y}_i - \hat{\boldsymbol{\gamma}}_{\nu g}\| \le d)}, \tag{3}$$

where $\hat{\boldsymbol{\gamma}}_{\nu g}$ is an estimator of $\boldsymbol{\gamma}_{\nu g}$, some measure of the center of subpopulation g formed from sample data $\mathcal{S}_{\nu g}$. Here, $\widehat{N}_{\nu g} = \sum_{i \in \mathcal{S}_{\nu g}} \frac{1}{\pi_i}$ is an estimator of size of subpopulation g which is generally not known a priori.

In equation (1), we assume $\boldsymbol{\gamma}_{\nu g}$ is a nonrandom sequence of finite population centers but $\hat{\boldsymbol{\gamma}}_{\nu g}$ is random due to sampling mechanism. Quantities $D_{\nu g,d}(\boldsymbol{\gamma}_{\nu g})$ and $\widehat{D}_{\nu g,d}(\hat{\boldsymbol{\gamma}}_{\nu g})$ are step functions of d and $\boldsymbol{\gamma}$ with jumps of size $O(\frac{1}{N_{\nu g}})$ and $O(\frac{1}{n_{\nu g}^*})$.

We can use the usual mean vector as a measure of center in equations (1)-(3),

$$\boldsymbol{\mu}_{\nu g} = \frac{1}{N_{\nu g}} \sum_{U_{\nu g}} \boldsymbol{y}_i, \tag{4}$$

which is estimated by,

$$\hat{\boldsymbol{\mu}}_{\nu g} = \frac{1}{\widehat{N}_{\nu g}} \sum_{\mathcal{S}_{\nu g}} \frac{\boldsymbol{y}_i}{\pi_i}. \tag{5}$$

We can also use a generalized version of median as a measure of center in multivariate case. The definition of spatial median was given in Brown (1983) and Small (1990). We generalize this idea and define the multivariate median of a finite population to be the location with smallest overall distance to all population units with respect to some norm $\|\cdot\|$. The norm is chosen at our choice, and commonly we use Minkoski distance or some shape-respecting quadratic distance.

$$\boldsymbol{q}_{\nu g} = \arg\inf_{\boldsymbol{\gamma}} \sum_{U_{\nu g}} \| \boldsymbol{y}_i - \boldsymbol{\gamma} \| . \tag{6}$$

The sample-based estimator of $\boldsymbol{q}_{\nu g}$ is given as follows,

$$\hat{\boldsymbol{q}}_{\nu g} = \arg\inf_{\boldsymbol{\gamma}} \sum_{i \in \mathcal{S}_{\nu g}} \frac{1}{\pi_i} \| \boldsymbol{y}_i - \boldsymbol{\gamma} \| . \tag{7}$$

We use $\boldsymbol{\gamma}_{\nu g}$ to denote a general measure of center, which can be mean vector $\boldsymbol{\mu}_{\nu g}$ or median $\boldsymbol{q}_{\nu g}$, and $\hat{\boldsymbol{\gamma}}_{\nu g}$ denotes the estimator of $\boldsymbol{\gamma}_{\nu g}$.

For now, we assume that the subpopulation association for each $\boldsymbol{y}_i$ in the sample is known.

3. Assumptions

3.1 General design assumptions

In this paper, we estimate the distribution of subpopulation distances in design-based framework. We assume the sequnce of finite poplations to be fixed and randomness only comes from the sampling mechanism. We do not want to restrict our attention to a specific sampling design but make rather general assumptions to cover various sampling schemes. Assumptions 3.1.1 and 3.1.2 ensure the design consistency and asymptotic normality of our estimator.

Assumption 3.1.1. *The following conditions hold for population size, inclusion probabilities π_i and design variance of Horvitz-Thompson estimator of the mean,*

1. *$N_{\nu g} = O(N_\nu)$, $n_\nu = O_p(N_\nu^\beta)$ with $\beta \in (\frac{1}{2}, 1]$.*
2. *$K_L \le \frac{N_\nu}{n_\nu^*}\pi_i \le K_U$ for all i, where K_L and K_U are positive constances.*
3. *For any vector $\boldsymbol{z}$ with finite $2+\delta$ monents, define $\bar{\boldsymbol{z}}_{\nu,\pi} = \frac{1}{N_\nu}\sum_{\mathcal{S}_\nu} \frac{\boldsymbol{z}_i}{\pi_i}$ as the Horvitz-Thompson estimator of $\bar{\boldsymbol{z}}_\nu = \frac{1}{N_\nu}\sum_{U_\nu} \boldsymbol{z}_i$. We assume*

$$Var(\bar{\boldsymbol{z}}_{\nu,\pi}|\mathcal{F}_\nu) \le K_1 Var_{SRS}(\bar{\boldsymbol{z}}_{\nu,\pi}|\mathcal{F}_\nu),$$

for some constant K_1, where $Var_{SRS}(\bar{\boldsymbol{z}}_{\nu,\pi}|\mathcal{F}_\nu)$ is the design variance-covariance matrix of $\bar{\boldsymbol{z}}_{\nu,\pi}$ under simple random sampling of size n_ν^*.

It can be shown that under Assumption 3.1.1(3), $\frac{n_{\nu g}}{n_{\nu g}^*} \xrightarrow{p} 1$ by bounding its design variance.

We will make the following normality assumption on Horvitz-Thompson estimator for a general vector with moment conditions, similar to Fuller (2007).

Assumption 3.1.2. *For any $\boldsymbol{z}$ with positive variance-covariance matrix and finite fourth moment*

$$n_\nu^{*1/2}(\bar{\boldsymbol{z}}_{\nu,\pi} - \bar{\boldsymbol{z}}_\nu)|\mathcal{F}_\nu \xrightarrow{d} N(\boldsymbol{0}, \Sigma_{\nu,\boldsymbol{z}}), \tag{8}$$

and

$$[V(\bar{\boldsymbol{z}}_{\nu,\pi}|\mathcal{F}_\nu)]^{-1} \hat{V}_{HT}\{\bar{\boldsymbol{z}}_{\nu,\pi}\} - \mathrm{I}_{p\times p} = O_p(n_\nu^{*-1/2}), \tag{9}$$

where $\bar{\boldsymbol{z}}_\nu = \frac{1}{N_\nu}\sum_{i=1}^{N_\nu} \boldsymbol{z}_i, \bar{\boldsymbol{z}}_{\nu,\pi} = \frac{1}{N_\nu}\sum_{\mathcal{S}_\nu} \frac{\boldsymbol{z}_i}{\pi_i}$, and $\hat{V}_{HT}\{\bar{\boldsymbol{z}}_{\nu,\pi}\}$ is the Horvitz-Thompson estimator of the variance of $\bar{\boldsymbol{z}}_{\nu,\pi}|\mathcal{F}_\nu$.

3.2 Assumptions on finite population

To show the design properties of the distance distribution estimator, we need to assume a number of regularity conditions on the sequence of finite populations.

Assumption 3.2.1. *The sequence of population vectors $\boldsymbol{y}_i$'s in subpopulation g has bounded $4+\delta$ moments,*

$$\lim_{N_{\nu g}\to\infty} N_{\nu g}^{-1} \sum_{i\in U_{\nu g}} |\,\boldsymbol{y}_i|^{4+\delta} < \infty,$$

for some $\delta > 0$.

Assumption 3.2.2. *1. The limit of population level distance distribution exists,*

$$\lim_{\nu\to\infty} D_{\nu g,d}(\boldsymbol{\gamma}) = \mathcal{D}_{g,d}(\boldsymbol{\gamma})$$

on $(d,\boldsymbol{\gamma}) \in [0,\infty)\times\Re^p$.

2. The limiting function $\mathcal{D}_{g,d}(\boldsymbol{\gamma})$ is continous in $d \in [0,\infty)$ and $\boldsymbol{\gamma} \in \Re^p$. Additionally, the derivatives $\frac{\partial \mathcal{D}_{g,d}(\boldsymbol{\gamma})}{\partial d}$, $\frac{\partial \mathcal{D}_{g,d}(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}}$ and $\frac{\partial^2 \mathcal{D}_{g,d}(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}^2}$ all exist and are finite in $(d,\boldsymbol{\gamma}) \in [0,+\infty)\times\Re^p$.

Assumption 3.2.3. *The following population quantities converge to zero:*

1.

$$\sqrt{N_{\nu g}}\left\{\frac{1}{N_{\nu g}}\sum_{U_{\nu g}} \mathrm{I}_{(d<\|\boldsymbol{y}_i-\boldsymbol{\gamma}\|\le d+h_{N_{\nu g}})} - \frac{\partial \mathcal{D}_{g,d}(\boldsymbol{\gamma})}{\partial d} h_{N_{\nu g}}\right\}$$

converges to zero, where $h_{N_{\nu g}} = O(N_{\nu g}^{-\alpha})$ and $\alpha \in (\frac{1}{4},1)$.

2.

$$\frac{{n^*_{\nu g}}^{1/2}}{N_{\nu g}} \sum_{U_{\nu g}} \Big[\mathrm{I}_{(\|\boldsymbol{y}_i-\boldsymbol{\gamma}-n_{\nu g}^{-1/2}\boldsymbol{s}\|\le d)} - \mathrm{I}_{(\|\boldsymbol{y}_i-\boldsymbol{\gamma}\|\le d)} - \mathcal{D}_{g,d}(\boldsymbol{\gamma}+{n^*_{\nu g}}^{-1/2}\boldsymbol{s}) + \mathcal{D}_{g,d}(\boldsymbol{\gamma})\Big]$$

converges to zero uniformly for $\boldsymbol{\gamma}\in\Re^p$ and $\boldsymbol{s}\in C_{\boldsymbol{s}}$, a large enough compact set in $\Re^p$.

Assumptions 3.2.4-3.2.8 are necessary when deriving median-based results, where we need stronger conditions on the spreading of finite population elements and restrictions on the norm. Assumption 3.2.4 guarentees the uniqueness of population median $\boldsymbol{q}_{\nu g}$. Assumption 3.2.6 restricts our concern to an inner product space, and Assumption 3.2.7 will be used in showing asymptotic normality of $\hat{\boldsymbol{q}}_{\nu g}$.

Assumption 3.2.4. *The finite population $U_{\nu g}$ only puts finitely many points $\boldsymbol{y}_i$ on a line in $\Re^p$.*

Assumption 3.2.5. *The norm $\|\cdot\|$ is continous on $\Re^p$, with a continuous gradient vector $\boldsymbol{\psi}(\boldsymbol{\gamma})$, and bounded second derivative matrix $H_s(\boldsymbol{\gamma})$.*

Assumption 3.2.6. *The norm $\|\cdot\|$ has the following inner product representation*

$$\|\boldsymbol{\gamma}\| = \sqrt{<\boldsymbol{\gamma},\boldsymbol{\gamma}>}$$

for $\boldsymbol{\gamma}\in\Re^p$ and some inner product $<\cdot,\cdot>$.

Assumption 3.2.7. *For any $\boldsymbol{\gamma}$ in a neighborhood of $\boldsymbol{q}_{\nu g}$, $\frac{1}{N_{\nu g}}\sum_{U_{\nu g}} H_s(\boldsymbol{y}_i-\boldsymbol{\gamma})$ is a nonsingular matrix. Further, the sequence of $H_s(\boldsymbol{y}_i-\boldsymbol{\gamma}_{\nu g})$ has bounded first two moments at population level.*

Assumption 3.2.8. *Assume that the linearized term $\boldsymbol{\psi}(\boldsymbol{y}_i-\boldsymbol{\gamma}_{\nu g})$ has bounded fourth population moments,*

$$\frac{1}{N_{\nu g}}\sum_{U_{\nu g}} \left|\boldsymbol{\psi}(\boldsymbol{y}_i-\boldsymbol{\gamma}_{\nu g})\right|^4 < \infty.$$

We use a kernel smoothing estimator to estimate the derivative of limiting smooth function as needed in the expression of asymptotic variance. Assumptions 3.2.9 and 3.2.10 give conditions on the choice of kernel function and bandwidth.

Assumption 3.2.9. *The kernel function $K(t)$ is symmetric with $\int_{-\infty}^{\infty} K(t)dt = 1$, and $K(t)$ is an absolutely continuous function with finite derivatives $K'(t)$. Further, let $R(K) = \int_{-\infty}^{\infty} K^2(t)dt < \infty$ and $\sigma_K^2 = \int_{-\infty}^{\infty} t^2K(t)dt < \infty$.*

Assumption 3.2.10. *The smoothing bandwidth $h\to 0$, and $N_{\nu g}h(\log N_{\nu g})^{-1}\to\infty$, as $N_{\nu g}\to\infty$.*

Assumption 3.2.11. *There exists a constant c, such that $\left|\frac{1}{h^2}K'\left(\frac{x}{h}\right)\right| \le c$, for any $x\neq 0$ and h arbitrarily small.*

4. Main results

4.1 Mean-based inference

Lemma 1. *Under Assumptions 3.1.1 and 3.2.2-3.2.3,*

$${n^*_{\nu g}}^{1/2}\left(\widehat{D}_{\nu g,d}(\hat{\boldsymbol{\mu}}_{\nu g}) - \widehat{D}_{\nu g,d}(\boldsymbol{\mu}_{\nu g}) - \mathcal{D}_{g,d}(\hat{\boldsymbol{\mu}}_{\nu g}) + \mathcal{D}_{g,d}(\boldsymbol{\mu}_{\nu g})\right) \tag{10}$$

converges to zero in design.

For a proof of this result, see Wang (2008).

Theorem 1. *Under Assumptions 3.1.1 and 3.2.2-3.2.3, the sample-based quantity $\widehat{D}_{\nu g,d}(\hat{\boldsymbol{\mu}}_{\nu g})$ is $\sqrt{n^*_{\nu g}}-$consistent for the corresponding population quantity $D_{\nu g,d}(\boldsymbol{\mu}_{\nu g})$, namely,*

$${n^*_{\nu g}}^{1/2}\left(\widehat{D}_{\nu g,d}(\hat{\boldsymbol{\mu}}_{\nu g}) - D_{\nu g,d}(\boldsymbol{\mu}_{\nu g})\right) = O_p(1).$$

Proof. We use the following decomposition,

$$\begin{aligned}
& {n^*_{\nu g}}^{1/2}\left(\widehat{D}_{\nu g,d}(\hat{\boldsymbol{\mu}}_{\nu g}) - D_{\nu g,d}(\boldsymbol{\mu}_{\nu g})\right) \\
= \; & {n^*_{\nu g}}^{1/2}\left(\widehat{D}_{\nu g,d}(\hat{\boldsymbol{\mu}}_{\nu g}) - \widehat{D}_{\nu g,d}(\boldsymbol{\mu}_{\nu g}) - \mathcal{D}_{g,d}(\hat{\boldsymbol{\mu}}_{\nu g}) + \mathcal{D}_{g,d}(\boldsymbol{\mu}_{\nu g})\right) \\
& + {n^*_{\nu g}}^{1/2}\left(\widehat{D}_{\nu g,d}(\boldsymbol{\mu}_{\nu g}) - D_{\nu g,d}(\boldsymbol{\mu}_{\nu g})\right) \\
& + {n^*_{\nu g}}^{1/2}\left(\mathcal{D}_{g,d}(\hat{\boldsymbol{\mu}}_{\nu g}) - \mathcal{D}_{g,d}(\boldsymbol{\mu}_{\nu g})\right),
\end{aligned} \tag{11}$$

where the first term is $o_p(1)$ by Lemma 1, and the last two terms are both $O_p(1)$. □

Expression (8) of Assumption 3.1.2 implies the following multivariate normality,

$$\frac{{n^*_{\nu g}}^{1/2}}{N_{\nu g}} \sum_{U_{\nu g}} \underbrace{\begin{bmatrix} \mathrm{I}_{(\|\boldsymbol{y}_i - \boldsymbol{\mu}_{\nu g}\| \le d)} \\ 1 \\ \boldsymbol{y}_i \end{bmatrix}}_{\boldsymbol{b}_{\boldsymbol{\mu},gi}} \left[\frac{\mathrm{I}_{(i \in \mathcal{S}_{\nu g})}}{\pi_i} - 1 \right] \Bigg| \mathcal{F}_\nu$$

$$\xrightarrow{d} N(0, \Sigma_{\boldsymbol{\mu},d}), \tag{12}$$

where

$$\Sigma_{\boldsymbol{\mu},d} = \frac{n^*_{\nu g}}{N^2_{\nu g}} \sum_{U_{\nu g}} \sum_{U_{\nu g}} (\pi_{ij} - \pi_i \pi_j) \frac{\boldsymbol{b}_{\boldsymbol{\mu},gi} \boldsymbol{b}'_{\boldsymbol{\mu},gi}}{\pi_i \pi_{gj}}, \tag{13}$$

as we are estimating a domain quantity and $\boldsymbol{b}_{\boldsymbol{\mu},gi}$ has bounded second moments under Assumption 3.2.1.

Theorem 2. *Under Assumptions 3.1.1-3.1.2 and 3.2.1-3.2.3,,*

$${n_{\nu g}}^{1/2} \left[V\left(\widehat{D}_{\nu g,d}(\hat{\boldsymbol{\mu}}_{\nu g}) \right) \right]^{-1/2} \left(\widehat{D}_{\nu g,d}(\hat{\boldsymbol{\mu}}_{\nu g}) - D_{\nu g,d}(\boldsymbol{\mu}_{\nu g}) \right)$$

$$\xrightarrow{d} N(0,1),$$

where

$$V\left(\widehat{D}_{\nu g,d}(\hat{\boldsymbol{\mu}}_{\nu g}) \right) = \boldsymbol{a}'_{\boldsymbol{\mu}} \Sigma_{\boldsymbol{\mu},d} \boldsymbol{a}_{\boldsymbol{\mu}}, \tag{14}$$

$$\boldsymbol{a}_{\boldsymbol{\mu}} = \left(1, -D_{\nu g,d}(\boldsymbol{\mu}_{\nu g}) - \left(\frac{\partial \mathcal{D}_{g,d}(\boldsymbol{\mu}_{\nu g})}{\partial \boldsymbol{\mu}_{\nu g}} \right)' \boldsymbol{\mu}_{\nu g}, \left(\frac{\partial \mathcal{D}_{g,d}(\boldsymbol{\mu}_{\nu g})}{\partial \boldsymbol{\mu}_{\nu g}} \right)' \right) \tag{15}$$

and $\Sigma_{\boldsymbol{\mu},d}$ is defined in (13).

Proof. We still use decomposition (11), then it follows from Lemma 1, expression (4.1), Taylor Linearization and Slutsky's Theorem. □

The asymptotic variance of $\widehat{D}_{\nu g,d}(\hat{\boldsymbol{\mu}}_{\nu g})$ consists of two pieces, the piece from estimating the distribution function with known center and the piece due to the uncertainty of estimating population center. The first piece of variance can be easily estimated using plug-in estimator or replication procedure, but the second piece involves an unknown derivative of limiting smooth function. The derivative can be estimated by kernel smoothing and incorporated into either plug-in or replication estimator.

Remarks: In Lemma 1, we use the limiting function $\mathcal{D}_{g,d}(\cdot)$ instead of population quantity $D_{\nu g,d}(\cdot)$, because we want the third term in (11) to be a smooth function so we can use linearization to quantify the randomness due to estimating subpopulation center.

4.2 Median-based inference

Now let us look at the case when we use median as a measure of subpopulation center. We introduce the following estimating equations at population and sample level, respectively,

$$\sum_{U_{\nu g}} \boldsymbol{\psi}(\boldsymbol{y}_i - \boldsymbol{\gamma}) = \mathbf{0}, \tag{16}$$

and

$$\sum_{i \in \mathcal{S}_{\nu g}} \frac{\boldsymbol{\psi}(\boldsymbol{y}_i - \boldsymbol{\gamma})}{\pi_i} = \mathbf{0}. \tag{17}$$

The medians defined by (6)-(7) are related to the roots of the estimating equations (16)-(17). More details are held until Appendix B.

Lemma 2. *Under Assumptions 3.2.4 and 3.2.6, for a large enough population, $\sum_{U_{\nu g}} \|\boldsymbol{y}_i - \boldsymbol{\gamma}\|$ has only one local minimum, which is also its global minimum.*

Remark 1: Lemma 2 states a stronger result than $\boldsymbol{q}_{\nu g}$ being unique, and it also says there are no other local minimums for $\sum_{U_{\nu g}} \|\boldsymbol{y}_i - \boldsymbol{\gamma}\|$.

Remark 2: If we make a similar assumption on the sample $\mathcal{S}_{\nu g}$, then it is obvious that $\sum_{\mathcal{S}_{\nu g}} \frac{\|\boldsymbol{y}_i - \boldsymbol{\gamma}\|}{\pi_i}$ has a unique global minimizer and no other local minimizers. But under a complex sampling design, we may have nonzero probability of selecting a sample where all points are on the same line. But considering the increasing sequence of finite populations and sequence of sampling designs, the probability of selecting a finite-sized sample will go to zero.

In establishing the weak convergence of $\hat{\boldsymbol{q}}_{\nu g}$ for $\boldsymbol{q}_{\nu g}$, we adopt the definition of weak convergence from P.24 of Billingsley (1968) and use general norm $\|\cdot\|$ as a discrepancy measure.

Theorem 3. *Under Assumptions 3.1.1, 3.2.4 and 3.2.6, any sequence $\hat{\boldsymbol{q}}_{\nu g}$ that satisfies (7) is design consistent for $\boldsymbol{q}_{\nu g}$.*

Proof. Negation. Suppose $\hat{\boldsymbol{q}}_{\nu g}$ does not converge to $\boldsymbol{q}_{\nu g}$ in probability, then there exists $\epsilon_1, \delta > 0$, such that

$$\mathrm{P}\left(\| \hat{\boldsymbol{q}}_{\nu g} - \boldsymbol{q}_{\nu g} \| > \epsilon_1 \right) > \delta, \tag{18}$$

for all ν. Equation (18) together with Assumption 3.2.4 would imply $\exists\ \epsilon_2 > 0$, such that

$$\mathrm{P}\left(\frac{1}{N_{\nu g}} \sum_{U_{\nu g}} \|\boldsymbol{y}_i - \hat{\boldsymbol{q}}_{\nu g}\| - \frac{1}{N_{\nu g}} \sum_{U_{\nu g}} \|\boldsymbol{y}_i - \boldsymbol{q}_{\nu g}\| > \epsilon_2 \right) > \delta, \tag{19}$$

for all ν.

Now, let us show that

$$\frac{1}{N_{\nu g}} \sum_{i \in \mathcal{S}_{\nu g}} \frac{\|\boldsymbol{y}_i - \hat{\boldsymbol{q}}_{\nu g}\|}{\pi_i} - \frac{1}{N_{\nu g}} \sum_{U_{\nu g}} \|\boldsymbol{y}_i - \hat{\boldsymbol{q}}_{\nu g}\| \xrightarrow{p} 0. \tag{20}$$

Define

$$Q_n(\boldsymbol{\gamma}) = \frac{1}{N_{\nu g}} \sum_{i \in \mathcal{S}_{\nu g}} \frac{\|\boldsymbol{y}_i - \boldsymbol{\gamma}\|}{\pi_i} - \frac{1}{N_{\nu g}} \sum_{U_{\nu g}} \|\boldsymbol{y}_i - \boldsymbol{\gamma}\|,$$

note that,

$$\mathrm{P}(|Q_n(\hat{\boldsymbol{q}}_{\nu g})| > \epsilon) \leq \mathrm{P}[\sup_{\boldsymbol{\gamma} \in C} |Q_n(\boldsymbol{\gamma})| > \epsilon'] + \mathrm{P}(\hat{\boldsymbol{q}}_{\nu g} \notin C),$$

where C is a large enough compact set. Now it is left to show that

$$\sup_{\boldsymbol{\gamma} \in C} |Q_n(\boldsymbol{\gamma})| \xrightarrow{p} 0. \tag{21}$$

The equation above can be shown by covering technique. Equations (19) and (20) imply that, there exists $\epsilon_3 > 0$ and $\delta_1 > 0$, such that

$$\mathrm{P}\left(\frac{1}{N_{\nu g}} \sum_{i \in \mathcal{S}_{\nu g}} \frac{\|\boldsymbol{y}_i - \hat{\boldsymbol{q}}_{\nu g}\|}{\pi_i} - \frac{1}{N_{\nu g}} \sum_{U_{\nu g}} \|\boldsymbol{y}_i - \boldsymbol{q}_{\nu g}\| > \epsilon_3 \right) > \delta_1. \tag{22}$$

As $\frac{1}{N_{\nu g}} \sum_{i \in \mathcal{S}_{\nu g}} \frac{\|\boldsymbol{y}_i - \boldsymbol{q}_{\nu g}\|}{\pi_i} \geq \frac{1}{N_{\nu g}} \sum_{i \in \mathcal{S}_{\nu g}} \frac{\|\boldsymbol{y}_i - \hat{\boldsymbol{q}}_{\nu g}\|}{\pi_i}$, for the same ϵ_3 and δ_1,

$$\mathrm{P}\left(\frac{1}{N_{\nu g}} \sum_{i \in \mathcal{S}_{\nu g}} \frac{\|\boldsymbol{y}_i - \boldsymbol{q}_{\nu g}\|}{\pi_i} - \frac{1}{N_{\nu g}} \sum_{U_{\nu g}} \|\boldsymbol{y}_i - \boldsymbol{q}_{\nu g}\| > \epsilon_3 \right) > \delta_1$$

as $\nu \to \infty$. Contradicting the fact that $\frac{1}{N_{\nu g}} \sum_{i \in \mathcal{S}_{\nu g}} \frac{\|\boldsymbol{y}_i - \boldsymbol{q}_{\nu g}\|}{\pi_i}$ is design consistent for $\frac{1}{N_{\nu g}} \sum_{U_{\nu g}} \|\boldsymbol{y}_i - \boldsymbol{q}_{\nu g}\|$. □

Theorem 4. *Under Assumptions 3.1.1-3.1.2, 3.2.5 and 3.2.7, and let $\hat{\boldsymbol{q}}_{\nu g}$ be a design consistent sequence, then we have the following asymptotic normality for sample median,*

$$n_{\nu g}^{*\,1/2} (\hat{\boldsymbol{q}}_{\nu g} - \boldsymbol{q}_{\nu g}) \xrightarrow{d} N_p(\boldsymbol{0}, \Sigma_{\nu g, \boldsymbol{q}}), \tag{23}$$

where

$$\Sigma_{\nu g, \boldsymbol{q}} = A \Sigma_{\nu g, \boldsymbol{\psi}} A', A = \left[\frac{1}{N_{\nu g}} \sum_{i=1}^{N_{\nu g}} H_s(\boldsymbol{y}_i - \boldsymbol{q}_{\nu g}) \right]^{-1}$$

and

$$\Sigma_{\nu g, \boldsymbol{\psi}} = \frac{n_{\nu g}}{N_{\nu g}^2} \sum_{U_{\nu g}} \sum_{U_{\nu g}} (\pi_{ij} - \pi_i \pi_j) \frac{\boldsymbol{\psi}(\boldsymbol{y}_i - \boldsymbol{q}_{\nu g})}{\pi_i} \frac{\boldsymbol{\psi}(\boldsymbol{y}_{gj} - \boldsymbol{q}_{\nu g})}{\pi_j}.$$

And further,

$$n_{\nu g}^{1/2} (\hat{\boldsymbol{q}}_{\nu g} - \boldsymbol{q}_{\nu g}) \xrightarrow{d} N_p(\boldsymbol{0}, \Sigma_{\nu g, \boldsymbol{q}}). \tag{24}$$

Proof. The consistency of $\hat{\boldsymbol{q}}_{\nu g}$ for $\boldsymbol{q}_{\nu g}$ gives,

$$\begin{aligned}
&\sum_{i \in \mathcal{S}_{\nu g}} \frac{1}{\pi_i} \boldsymbol{\psi}(\boldsymbol{y}_i - \hat{\boldsymbol{q}}_{\nu g}) = \boldsymbol{0} \\
\Leftrightarrow \quad &\sum_{i \in \mathcal{S}_{\nu g}} \frac{1}{\pi_i} \boldsymbol{\psi}(\boldsymbol{y}_i - \boldsymbol{q}_{\nu g} - (\hat{\boldsymbol{q}}_{\nu g} - \boldsymbol{q}_{\nu g})) = \boldsymbol{0} \\
\Leftrightarrow \quad &\sum_{i \in \mathcal{S}_{\nu g}} \frac{1}{\pi_i} \left\{ \boldsymbol{\psi}(\boldsymbol{y}_i - \boldsymbol{q}_{\nu g}) - H_s(\boldsymbol{y}_i - \boldsymbol{q}_{\nu g})(\hat{\boldsymbol{q}}_{\nu g} - \boldsymbol{q}_{\nu g}) \right. \\
&\left. + o_p(\hat{\boldsymbol{q}}_{\nu g} - \boldsymbol{q}_{\nu g}) \right\} = \boldsymbol{0},
\end{aligned}$$

which implies

$$\begin{aligned}
\hat{\boldsymbol{q}}_{\nu g} \quad = \quad & \boldsymbol{q}_{\nu g} \\
& + \left[\frac{1}{N_{\nu g}} \sum_{i \in \mathcal{S}_{\nu g}} \frac{H_s(\boldsymbol{y}_i - \boldsymbol{q}_{\nu g})}{\pi_i} \right]^{-1} \frac{1}{N_{\nu g}} \sum_{i \in \mathcal{S}_{\nu g}} \frac{\boldsymbol{\psi}(\boldsymbol{y}_i - \boldsymbol{q}_{\nu g})}{\pi_i} \\
& + o_p(\hat{\boldsymbol{q}}_{\nu g} - \boldsymbol{q}_{\nu g})
\end{aligned} \tag{25}$$

after using non-singularity condition in Assumption 3.2.7 to take the inverse of $\frac{1}{N_{\nu g}} \sum_{i=1}^{N_{\nu g}} H_s(\boldsymbol{y}_i - \boldsymbol{q}_{\nu g})$.

It is easy to argue that

$$\left[\frac{1}{N_{\nu g}} \sum_{i \in \mathcal{S}_{\nu g}} \frac{H_s(\boldsymbol{y}_i - \boldsymbol{q}_{\nu g})}{\pi_i} \right]^{-1} \xrightarrow{p} \left[\frac{1}{N_{\nu g}} \sum_{U_{\nu g}} H_s(\boldsymbol{y}_i - \boldsymbol{q}_{\nu g}) \right]^{-1}, \tag{26}$$

and

$$\frac{n_{\nu g}^{*\,1/2}}{N_{\nu g}} \sum_{i \in \mathcal{S}_{\nu g}} \frac{\boldsymbol{\psi}(\boldsymbol{y}_i - \boldsymbol{q}_{\nu g})}{\pi_i} \xrightarrow{d} N_p(\boldsymbol{0}, \Sigma_{\nu g, \boldsymbol{\psi}}), \tag{27}$$

where

$$\Sigma_{\nu g, \boldsymbol{\psi}} = \frac{n_{\nu g}^*}{N_{\nu g}^2} \sum_{U_{\nu g}} \sum_{U_{\nu g}} (\pi_{ij} - \pi_i \pi_j) \frac{\boldsymbol{\psi}(\boldsymbol{y}_i - \boldsymbol{q}_{\nu g})}{\pi_i} \frac{\boldsymbol{\psi}(\boldsymbol{y}_{gj} - \boldsymbol{q}_{\nu g})}{\pi_j}$$

.

The proof is then completed by applying Slutsky's Theorem to obtain (23). Then we can use Slutsky's Theorem again and the fact that $\frac{n_{\nu g}}{n_{\nu g}^*} \xrightarrow{p} 1$ to obtain (24). □

Theorem 5. *Suppose Assumptions 3.1.1 and 3.2.4-3.2.7 are satisfied, then for any sequence $\hat{\boldsymbol{q}}_{\nu g}$ that satisfies (7), the estimated distance distribution $\widehat{D}_{\nu g,d}(\hat{\boldsymbol{q}}_{\nu g})$ is $\sqrt{n_{\nu g}^*}$−consistent for the corresponding population quantity $D_{\nu g,d}(\boldsymbol{q}_{\nu g})$, namely,*

$$n_{\nu g}^{*\,1/2} \left(\widehat{D}_{\nu g,d}(\hat{\boldsymbol{q}}_{\nu g}) - D_{\nu g,d}(\boldsymbol{q}_{\nu g}) \right) = O_p(1).$$

Proof. Similar to the proof of Theorem 2, we use the following decomposition,

$$\begin{aligned}
& n_{\nu g}^{*\,1/2} \left(\widehat{D}_{\nu g,d}(\hat{\boldsymbol{q}}_{\nu g}) - D_{\nu g,d}(\boldsymbol{q}_{\nu g}) \right) \\
= \quad & n_{\nu g}^{*\,1/2} \left(\widehat{D}_{\nu g,d}(\hat{\boldsymbol{q}}_{\nu g}) - \widehat{D}_{\nu g,d}(\boldsymbol{q}_{\nu g}) - \mathcal{D}_{g,d}(\hat{\boldsymbol{q}}_{\nu g}) + \mathcal{D}_{g,d}(\boldsymbol{q}_{\nu g}) \right) \\
& + n_{\nu g}^{*\,1/2} \left(\widehat{D}_{\nu g,d}(\boldsymbol{q}_{\nu g}) - D_{\nu g,d}(\boldsymbol{q}_{\nu g}) \right) \\
& + n_{\nu g}^{*\,1/2} \left(\mathcal{D}_{g,d}(\hat{\boldsymbol{q}}_{\nu g}) - \mathcal{D}_{g,d}(\boldsymbol{q}_{\nu g}) \right).
\end{aligned} \tag{28}$$

We can show the first term converges to zero in design in a similar fashion to the proof of Lemma 1. The remainder of the proof follows as $\sqrt{n_{\nu g}^*}(\hat{\boldsymbol{q}}_{\nu g} - \boldsymbol{q}_{\nu g}) = O_p(1)$. □

Assumption 3.2.8 together with Assumption 3.1.2 gives

$$\frac{n_{\nu g}^{*\,1/2}}{N_{\nu g}} \sum_{U_{\nu g}} \underbrace{\begin{bmatrix} \mathrm{I}_{(\|\boldsymbol{y}_i - \boldsymbol{q}_{\nu g}\| \leq d)} \\ 1 \\ \boldsymbol{\psi}(\boldsymbol{y}_i - \boldsymbol{q}_{\nu g}) \end{bmatrix}}_{\boldsymbol{b}_{\boldsymbol{q},gi}} \left[\frac{\mathrm{I}_{(i \in \mathcal{S}_{\nu g})}}{\pi_i} - 1 \right] \Bigg| \mathcal{F}_\nu \xrightarrow{d} N(0, \Sigma_{\boldsymbol{q},d}), \tag{29}$$

where

$$\Sigma_{\boldsymbol{q},d} = \frac{n^*_{\nu g}}{N^2_{\nu g}} \sum_i \sum_j (\pi_{ij} - \pi_i \pi_j) \frac{\boldsymbol{b}_{\boldsymbol{q},gi} \boldsymbol{b}'_{\boldsymbol{q},gi}}{\pi_i \pi_j}. \tag{30}$$

Theorem 6. *Under Assumptions 3.1.1-3.1.2 and 3.2.4-3.2.8, for any sequence $\hat{\boldsymbol{q}}_{\nu g}$ satisfying (7),*

$$n_{\nu g}{}^{1/2} \left[V\left(\widehat{D}_{\nu g,d}(\hat{\boldsymbol{q}}_{\nu g}) \right) \right]^{-1/2} \left(\widehat{D}_{\nu g,d}(\hat{\boldsymbol{q}}_{\nu g}) - D_{\nu g,d}(\boldsymbol{q}_{\nu g}) \right) \Big| \mathcal{F}_\nu$$

$$\xrightarrow{d} N(0,1),$$

where

$$V\left(\widehat{D}_{\nu g,d}(\hat{\boldsymbol{q}}_{\nu g}) \right) = \boldsymbol{a}'_{\boldsymbol{q}} \Sigma_{\boldsymbol{q},d} \boldsymbol{a}_{\boldsymbol{q}}, \tag{31}$$

and

$$\boldsymbol{a}_{\boldsymbol{q}} = \begin{pmatrix} 1 \\ -D_{\nu g,d}(\boldsymbol{q}_{\nu g}) - \left(\frac{\partial \mathcal{D}_{g,d}(\boldsymbol{q}_{\nu g})}{\partial \boldsymbol{q}_{\nu g}} \right)' H^{-1}_{s,\,N_{\nu g}} \boldsymbol{q}_{\nu g} \\ H^{-1}_{s,\,N_{\nu g}} \left(\frac{\partial \mathcal{D}_{g,d}(\boldsymbol{q}_{\nu g})}{\partial \boldsymbol{q}_{\nu g}} \right) \end{pmatrix}, \tag{32}$$

where $H_{s,N_{\nu g}} = \frac{1}{N_{\nu g}} \sum_{i=1}^{N_{\nu g}} H_s(\boldsymbol{y}_i - \boldsymbol{q}_{\nu g})$ and $\Sigma_{\boldsymbol{q},d}$ is defined in (30).

Proof. We can use decomposition (28), then the leading term in decomposition is asymptotically normally distributed follows by using Slutsky's Theorem. □

4.3 Variance Estimation

This section deals with estimating the variances of $\widehat{D}_{\nu g,d}(\hat{\boldsymbol{\mu}}_{\nu g})$ and $\widehat{D}_{\nu g,d}(\hat{\boldsymbol{q}}_{\nu g})$. We will introduce a naive estimator, a kernel estimator estimating the effect of error in $\hat{\boldsymbol{\gamma}}$ and a jackknife estimator. The naive estimator ignores the error in estimating population center, and the extra piece of variance can be estimated by kernel smoothing and included in analytic variance estimator. An alternative is to incorporate a smoothing term in jackknife to get a consistent variance estimator. The smoothing term is only estimated once using the whole sample, so the benefit of replication procedure is not much affected.

4.3.1 Naive estimator

To estimate $V\left(\widehat{D}_{\nu g,d}(\hat{\boldsymbol{\mu}}_{\nu g}) \right)$ and $V\left(\widehat{D}_{\nu g,d}(\hat{\boldsymbol{q}}_{\nu g}) \right)$ as defined in (14) and (31), we need to estimate gradient vectors $\frac{\partial \mathcal{D}_{g,d}(\boldsymbol{\mu}_{\nu g})}{\partial \boldsymbol{\mu}_{\nu g}}$ and $\frac{\partial \mathcal{D}_{g,d}(\boldsymbol{q}_{\nu g})}{\partial \boldsymbol{q}_{\nu g}}$. We can either use some nonparametric smoothing method or propose a parametric model for the subpopulation distribution. First, we will show that if the subpopulation is a realization of an elliptical distribution, the variance due to estimating subpopulation center can be ignored.

Lemma 3. *Assume random variable $\boldsymbol{Y}_g \sim EC_p(\boldsymbol{\mu}_g, \Lambda_g, \phi)$ with mean vector $\boldsymbol{\mu}_g$, and Λ_g is a non-negative definite matrix. We define the norm as,*

$$\| \boldsymbol{u} \| = \sqrt{\boldsymbol{u}' B \boldsymbol{u}}, \tag{33}$$

where B is a non-negative definite matrix. Then

1. *the partial derivative evaluated at superpopulation mean is $\mathbf{0}$, $\frac{\partial \mathcal{D}_{g,d}(\boldsymbol{\mu}_g)}{\partial \boldsymbol{\mu}_g} = \mathbf{0}$.*
2. *mean $\boldsymbol{\mu}_g$ coincide with median $\boldsymbol{q}_g$, $\boldsymbol{\mu}_g = \boldsymbol{q}_g$.*

Lemma 4. *Assume 3.2.2, $\frac{\partial \mathcal{D}_{g,d}(\boldsymbol{\gamma}_g)}{\partial \boldsymbol{\gamma}_g} = \mathbf{0}$ for some constant vector $\boldsymbol{\gamma}_g$, and the sequence of subpopulation centers $\boldsymbol{\gamma}_{\nu g}$ converges to $\boldsymbol{\gamma}_g$, $\lim_{\nu \to \infty} \boldsymbol{\gamma}_{\nu g} = \boldsymbol{\gamma}_g$. Then $\frac{\partial \mathcal{D}_{g,d}(\boldsymbol{\gamma}_{\nu g})}{\partial \boldsymbol{\gamma}_{\nu g}} = o(1)$.*

Proof. The proof follows from Taylor expansion,

$$\frac{\partial \mathcal{D}_{g,d}(\boldsymbol{\gamma}_{\nu g})}{\partial \boldsymbol{\gamma}_{\nu g}} = \frac{\partial \mathcal{D}_{g,d}(\boldsymbol{\gamma}_g)}{\partial \boldsymbol{\gamma}_g} + \frac{\partial^2 \mathcal{D}_{g,d}(\boldsymbol{\gamma}^*_{\nu g})}{\partial \boldsymbol{\gamma}^*_{\nu g}{}^2} (\boldsymbol{\gamma}_{\nu g} - \boldsymbol{\gamma}_g) = o(1).$$

□

Lemma 3 and 4 imply that the extra variance due to estimating subpopulation centers can be ignored in elliptical distributions with a norm specified by (33). This special case is similar to case A of Randles (1982), where we can pretend that we are using true population center $\boldsymbol{\gamma}_{\nu g}$ without affecting the leading variance. So we can propose the following naive plug-in variance estimator for the leading term,

$$\widehat{V}_{NV}\left(\widehat{D}_{\nu g,d}(\hat{\boldsymbol{\gamma}}_{\nu g}) \right) = \left(1, -\widehat{D}_{\nu g,d}(\hat{\boldsymbol{\gamma}}_{\nu g}) \right) \widehat{\Sigma}_{\boldsymbol{\gamma},d} \left(1, -\widehat{D}_{\nu g,d}(\hat{\boldsymbol{\gamma}}_{\nu g}) \right)', \tag{34}$$

with

$$\widehat{\Sigma}_{\boldsymbol{\gamma},d,NV} = \frac{n_{\nu g}}{\hat{N}^2_{\nu g}} \sum_i \sum_j \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \frac{\begin{bmatrix} \mathrm{I}_{(\|\boldsymbol{y}_i - \hat{\boldsymbol{\gamma}}_{\nu g}\| \le d)} \\ 1 \end{bmatrix} \left[\mathrm{I}_{(\|\boldsymbol{y}_i - \hat{\boldsymbol{\gamma}}_{\nu g}\| \le d)}, 1 \right]}{\pi_i \pi_j}.$$

4.3.2 Estimating the effect of error in $\hat{\boldsymbol{\gamma}}$

The naive estimator (34) ignores the piece of variance due to estimating population center and tends to underestimate the true variance for a general suppopulation. So we need to estimate $\frac{\partial \mathcal{D}_{g,d}(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}}$ and incorporate the extra piece of variance. Let $\boldsymbol{\zeta}_{g,d}(\boldsymbol{\gamma}) = \frac{\partial \mathcal{D}_{g,d}(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}}$. We have proposed an estimator of $\boldsymbol{\zeta}_{g,d}(\boldsymbol{\gamma})$ using kernel smoothing,

$$\hat{\boldsymbol{\zeta}}_{\nu g,d}(\boldsymbol{\gamma}) = \frac{1}{\hat{N}_{\nu g} h} \sum_{\mathcal{S}_{\nu g}} K\left(\tfrac{d - \|\boldsymbol{y}_i - \boldsymbol{\gamma}\|}{h} \right) \boldsymbol{\psi}(\boldsymbol{y}_i - \boldsymbol{\gamma}) \frac{1}{\pi_i}. \tag{35}$$

The idea of estimator $\hat{\boldsymbol{\zeta}}_{\nu g,d}(\boldsymbol{\gamma})$ is to estimate $\mathcal{D}_{g,d}(\boldsymbol{\gamma})$ using primitive function of kernel $K(\cdot)$, and then take derivative with respect to $\boldsymbol{\gamma}$.

Lemma 5. *Under Assumptions 3.1.1, 3.2.5 and 3.2.8-3.2.10, the estimator $\hat{\boldsymbol{\zeta}}_{\nu g,d}(\boldsymbol{\gamma})$ is design consistent for $\boldsymbol{\zeta}_{\nu g,d}(\boldsymbol{\gamma})$.*

Lemma 6. *Under Assumptions 3.1.1, 3.2.5 and 3.2.7-3.2.11, and assume the sequence of populations is such that*

$$\sup_{\boldsymbol{\gamma}\in\Re^p} \left|\boldsymbol{\zeta}_{\nu g,d}(\boldsymbol{\gamma}) - \boldsymbol{\zeta}_{g,d}(\boldsymbol{\gamma})\right| \to \mathbf{0}, \tag{36}$$

the kernel estimator $\hat{\boldsymbol{\zeta}}_{\nu g,d}(\hat{\boldsymbol{\gamma}}_{\nu g})$ is design consistent for $\boldsymbol{\zeta}_{g,d}(\boldsymbol{\gamma}_{\nu g})$ for every d.

Remarks: We have established uniform strong consistency of $\boldsymbol{\zeta}_{g,d}(\boldsymbol{\gamma})$ for $\frac{\partial \mathcal{D}_{g,d}(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}}$, under appropriate superpopulation assumptions. Assumption (36) assumes that we are not working with the populations where $\boldsymbol{\zeta}_{\nu g,d}(\boldsymbol{\gamma})$ does not converge to $\boldsymbol{\zeta}_{g,d}(\boldsymbol{\gamma})$, which is on a zero-probability set.

We can directly plug the estimator (35) into estimators (14) and (31) to get a design-consistent variance estimator, and it performance will be evaluated in simulation study as shown in Wang (2008).

4.3.3 Jackknife variance estimator

This section only applies formally to mean-based estimator, but can be modified to include median-based estimator. To introduce the jackknife variance estimator for our application, we borrowed some strength from Da Silva and Opsomer (2006) and start by assuming there already exists a design consistent jackknife variance estimator for linear estimators. Then we define jackknife replicates in our case and show its consistency.

Theorem 7. *Let $\hat{\theta}$ be a linear estimator for subpopulation g with*

$$\hat{\theta}_g = \sum_{\mathcal{S}_{\nu g}} w_i z_i,$$

where w_i is the survey weight and z_i has bounded $4+\delta$ moments. Assume there is a jackknife replication procedure that generates L replicated estimates

$$\hat{\theta}_g^{(l)} = \sum_{\mathcal{S}_{\nu g}} w_i^{(l)} z_i,$$

with $l = 1, 2, \cdots, L$ and $w_i^{(l)}$ is replication weight for unite i in the $l-th$ replicate. The replication variance estimator is defined as

$$\hat{V}_{JK}(\hat{\theta}_g) = \sum_{l=1}^{L} c_l \left(\hat{\theta}_g^{(l)} - \hat{\theta}_g\right)^2, \tag{37}$$

where c_l is a set of constants for the $l-th$ replicate. Assumptions similar to (D1)-(D4) and (D6) in Da Silva and Opsomer (2006) are assumed.

We define the l-th jackknife replicate as

$$\begin{aligned}\widehat{D}^{(l)}(\hat{\boldsymbol{\mu}}_{\nu g}) &= \widehat{D}^{(l)}_{\nu g,d}(\hat{\boldsymbol{\mu}}_{\nu g}) \\ &\quad + \frac{1}{\hat{N}_{\nu g} h} \sum_{\mathcal{S}_{\nu g}} \frac{1}{\pi_i} K\left(\tfrac{d-\|\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_{\nu g}\|}{h}\right) \boldsymbol{\psi}^T(\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_{\nu g})(\hat{\boldsymbol{\mu}}^{(l)}_{\nu g} - \hat{\boldsymbol{\mu}}_{\nu g}),\end{aligned} \tag{38}$$

where $\widehat{D}^{(l)}_{\nu g,d}(\hat{\boldsymbol{\mu}}_{\nu g}) = \frac{1}{\hat{N}^{(l)}_{\nu g}} \sum_{i\in\mathcal{S}_{\nu g}} w_i^{(l)} \mathrm{I}_{(\|\boldsymbol{y}_i - \hat{\mu}_{\nu g}\| \le d)}, \hat{N}^{(l)}_{\nu g} = \sum_{i\in\mathcal{S}_{\nu g}} w_i^{(l)}$ and $\hat{\boldsymbol{\mu}}^{(l)}_{\nu g} = \frac{1}{\hat{N}^{(l)}_{\nu g}} \sum_{i\in\mathcal{S}_{\nu g}} w_i^{(l)} \boldsymbol{y}_i$.

Then the jackknife variance estimator

$$\widehat{V}_{JK}\left(\widehat{D}_{\nu g,d}(\hat{\boldsymbol{\mu}}_{\nu g})\right) = \sum_{l=1}^{L} c_l \left(\widehat{D}^{(l)}(\hat{\boldsymbol{\mu}}_{\nu g}) - \widehat{D}(\hat{\boldsymbol{\mu}}_{\nu g})\right)^2 \tag{39}$$

is design consistent for $V\left(\widehat{D}_{\nu g,d}(\hat{\boldsymbol{\mu}}_{\nu g})\right)$.

Proof. The use of this jackknife variance estimator is suggested by expression (11). The first term in (38) is used to approximate the variance of $\widehat{D}_{\nu g}(\boldsymbol{\mu}_{\nu g})$, and the second term approximates the variance due to estimating the center. □

To examine the connection between the jackknife estimator and kernel estimator in section 4.3.2, we first ignore the second piece in (38) and compare it with the naive estimator (34). The naive estimator only approximates the linearized variance for the ratio, but the jackknife estimator usually overestimates this linearized variance. If we compare the whole jackknife replicate (38) with kernel estimator, the difference exists because of the nonlinearity of $\widehat{D}_{\nu g,d}(\boldsymbol{\mu}_{\nu g})$ and $\hat{\boldsymbol{\mu}}_{\nu g}$.

A great advantage of jackknife variance estimator over the plug-in estimator is that we do not need to estimate the covariance matrix (13), which can be complicated in a large-scale complex survey. In jackknife variance estimation, we estimate the gradient vector based on the whole sample only once, but $\widehat{D}^{(l)}_{\nu g,d}(\hat{\boldsymbol{\mu}}_{\nu g})$ and $\hat{\boldsymbol{\mu}}^{(l)}_{\nu g}$ will change with replicate. Variations of delete-1 jackknife like delete-d or delete-a-group jackknife can be used in complex surveys or in case of using median as measures of center.

Acknowledgement

This research was supported in part by the USDA Natural Resources Conservation Service cooperative agreement NRCS-68-3A75-4-122.

References

Billingsley, P. (1968). *Convergence of Probability Measures.* John Wiley & Sons.

Brown, B.M. (2007). Statistical uses of the spatial median. *Journal of the Royal Statistical Society, Series B: Methodology 45.* 25-30.

Da Silva, D. and Opsomer, J(2001). A kernel smoothing method to adjust for unit nonresponse in sample surveys. *the Canadian Journal of Statistics 34*. ??? - ???.

Fuller, W. (2007). *Sampling statistics*.

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

Small, C.G. (2007). A survey of multidimensional medians *International Statistical Review 58*. 263-277.

Wang, J. (2008). *Estimating the distance distribution of subpopulations for a large-scale complex survey*. Ph.D. thesis, Iowa State University.