International Journal of Statistics and Probability; Vol. 10, No. 6; November 2021
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

# Integration of Nonprobability and Probability Samples via Survey Weights

Balgobin Nandram, Jai Won Choi, Yang Liu

[1] Professor, Worcester Polytechnic Institute, Worcester, 100 Institute Road, Worcester, MA 01609-2280

[2] Statistical Consultant, Meho, Inc, 9504 Mary Knoll Dr., Rockville, MD 20850

[3] Graduate Student, Worcester Polytechnic Institute, Worcester, 100 Institute Road, Worcester, MA 01609-2280

Correspondence: Jai Won Choi, Statistical Consultant, Meho, Inc, 9504 Mary Knoll Dr., Rockville, MD 20850

Received: July 11, 2021 Accepted: August 30, 2021 Online Published: October 19, 2021

doi:10.5539/ijsp.v10n6p5 URL: https://doi.org/10.5539/ijsp.v10n6p5

**Abstract**

Probability sample encounters the problems of increasing cost and nonresponse. The cost has rapidly been increasing in executing a large probability sample survey, and, for some surveys, response rate can be below the 10 percent level. Therefore, statisticians seek some alternative methods. One of them is to use a large nonprobability sample ($\boldsymbol{S_1}$) supplemented by a small probability sample ($S_2$). Both samples are taken from the same population and they include common covariates, and a third sample ($S_3$) is created by combining these two samples; $S_1$ can be biased and $S_2$ may have large sample variance. These two problems are reduced by survey weights and combining the two samples. Although $S_2$ is a small sample, it provides good properties of unbiasedness in estimation and of survey weights. With these known weights, we obtain adjusted sample weights (ASW), and create a sample model from a finite population model. We fit the sample model to obtain its parameters and generate values from the population model. Similarly, we repeat these processes for other two samples, $S_1$ and $S_3$ and for different statistical methods. We show reduced biases of the finite population means and reduced variances.as the combined sample size becomes large. We analyze sample data to show the reduction of these two errors.

**Keywords**: adjusted sample weight, bootstrap, Bayesian method, least squares estimation (LS), doubly robust

## 1. Introduction

Probability sampling has been the main tool for sample surveys since the 1900s. It provides unbiased and consistent estimates. Public and private survey organizations such as Gallop, US Census Bureau, National Center for Health Statistics (**NCHS**), National Institute of Health, and the Bureau of Labor Statistics have been using probability samples to produce official statistics of the US population such as public opinion, personal income, health status, economic indicators, unemployment rates, and other essential official statistics for the United States Government and other researchers.

However, more survey organizations are abandoning probability sampling because of high cost and increasing nonresponses (Beaumont, 2020). Nonprobability sampling is emerging as an attractive to probability sampling for its low cost and convenience. But we must pay for the low cost and convenience because there could be large bias.

A small probability sample is used to reduce such bias of nonprobability sample. Both samples, taken from the same population, have the common covariates. Bias is increased and variance is decreased in the combined sample over the probability sample, and the reverse (bias decreased and variance increased) is true with respect to the nonprobability sample. A small probability sample contributes valuable information of sample weights and unbiased mean. This mean is used to find the bias of nonprobability and combined samples. The known weights are also used to impute unknown weights of $W_1$.(Appendix A). Others often use these covariates in logistic function to obtain sample weights under certain assumptions (Chen et al. 2020). Instead, we use covariates matching that does not require such unreasonable assumptions. These known (or imputed) survey weights or original sample weights (OSW) are then used to obtain ASWs. We use these ASWs to reduce bias and variance in estimation by combining two samples.

For the National Health and Nutrition Examination Survey, NCHS collects a probability sample to investigate the health status of the US population. To present an example, we use a part of this NCHS sample to derive three samples: A rather large sample $S_1$ by dropping the survey weights, a small sample $S_2$, and a third sample $S_3$ is created combining $S_1$ and $S_2$.i.e. $S_3 = (S_1 \cup S_2)$. These three samples are separately used to investigate the bias and variance on these samples in estimating finite population mean of body mass index (**BMI**) of US population. Many government agencies, including the NCHS can benefit enormously from this research to save the cost of large surveys.

We have roughly 1500 observations for $S_1$ and 300 for $S_2$, i.e., $S_1$ is about 5 times for $S_2$. The number of combined sample $S_3$ is 1800. We assume that a probability sample provides unbiased estimate. The reduction of bias is significant when ASWs are applied to probability or nonprobability sample. For example, the unweighted BMI mean of probability sample $S_2$ is 26.94 and its ASW weighted mean is 25.98; reduced bias is about 0.96 (=26.94-25.98), about 3.56%. If only nonprobability sample is used, the unweighted BMI mean is 27.89 and the ASW weighted mean is 27.16, reduced bias is 0.73 (=27.89-27.16) (2.6%). Comparing unweighted mean 27.16 of nonprobability sample to ASW weighted mean 25.98 of probability sample, the reduced bias is 1.18 (=27.16-25.98) (4.34%). This implies that the bias of nonprobability sample is reduced by the help of ASW, it is the same for ASWs with the probability sample. The variance of a small probability sample is reduced by the help of large nonprobability sample by combining the two samples, the variance of nonprobability sample or probability sample is reduced as shown in Tables (3, 4, and 5) in Section 4.

We also use different methods for the estimation, non-Bayesian (least square, and bootstrap) and Bayesian methods to estimate the finite population mean. Within the Bayesian approach, we provide a simple model that has closed form answers if the non-sample covariates are known. Within the non-Bayesian approach, we use least square (LS), and bootstrap method, and bootstrap is used to calculate the variance of LS estimation mainly because the LS variance is too small when OSW is big. The Bayesian and LS methods are separately applied to the three samples, $S_1$, $S_2$ and $S_3$. No matter which method we use, the result shows similar or same results of reducing bias by applying ASW and reducing variance by combining two samples.

Meng (2018) defined the difference between sample mean $\bar{y}_n = \frac{\sum_{i=1}^{n} y_i}{n}$, and unknown true population mean $\bar{Y}_N$, $\bar{Y}_N = \frac{\sum_{i=1}^{N} y_i}{N}$ as the error of the sample: i.e.,

$$\bar{y}_n - \bar{Y}_N =: \rho_{ry} \left(\frac{1-f}{f}\right) \sqrt{\sigma^2}.$$

This error estimate is the product of three terms. The first is data quality, i.e., correlation $\rho_{ry}$ between response indicator $r$=1, 0 and study variable y, data quality is increased by controlling $\rho_{ry}$ at the level of $\frac{1}{\sqrt{N}}$. The second is data quantity, $\sqrt{\frac{(N-n)}{n}} = \sqrt{\frac{(1-f)}{f}}$, where the sample fraction $f = \frac{n}{N}$, N is the population size. When losing the control, the impact of N is no longer cancelled by $\rho_{ry}$, i.e., error estimation increases with $\sqrt{N}$ relative to $\frac{1}{\sqrt{n}}$. Bigger N makes the matter worse for nonprobability sample. He calls this large sample paradox. The bigness of big data for population inference should be measured by the relative size f = n/N, not the absolute value n. The third is problem difficulty, $\sqrt{\sigma^2}$, the standard deviation of study variable y's. When combining data for population inference, those relatively tiny but higher quality ones, i.e., probability samples, should be given far more weights than suggested by their sizes. Rao (2020) reviewed individual terms of the difference between the sample mean and the finite population mean.

Big sample data from the finite population provides small variance. For example, take independently distributed random variables, $y_1, \ldots, y_n \sim (0, \sigma^2)$. Then $\bar{y} \sim (0, \frac{\sigma^2}{n})$ with unspecified distribution. For large n, $\frac{\sigma^2}{n} \approx 0$, this is a common large sample problem for statistical inference. Choi and Nandram (2021) showed how to use the random grouping method to get around this problem.

Any sample, probability or nonprobability, can be tested by $\rho_{ry}$; Here, $\rho_{ry}$= 0 means the sample is biased. We investigate $\rho_{ry}$ for $S_1$. Assume that a finite population of N individuals with finite population mean, $\bar{Y}_N$ of the response variables **y**=$(y_1, \ldots y_N)$. Let **R**=$(r_1, \ldots, r_N)$ denote the sample indicators, and $\bar{R} = \frac{\sum_{i=1}^{N} R_i}{N}$, then the correlation between **y** and **R** is

$$\rho_{ry}=\frac{\sum_{i=1}^{N}(y_i-\bar{Y}_N)\,(r_i-\bar{R})}{\sqrt{\sum_{i=1}^{N}(y_i-\bar{Y}_N)^2}\sqrt{\sum_{i=1}^{N}(r_i-\bar{R})^2}}.$$

Arrange it so that $r_1,\ldots,r_n$ are ones corresponding to the sample, $y_1,\ldots y_n$, for sampled individuals, and $r_{n+1},\ldots,r_N$ are zeros corresponding to non-sampled individuals, $y_{n+1},\ldots,y_N$.

Now,

$$\sum_{i=1}^{N}(y_i-\bar{Y}_N)(r_i-\bar{R})=\text{n}\,(\bar{y}_n-\bar{Y}_N),$$

$$\sum_{i=1}^{N}(y_i-\bar{Y}_N)^2=N\sigma_y^2 \text{ and } \sum_{i=1}^{N}(r_i-\bar{R})^2=\frac{n}{N}(N-n).$$

Therefore,

$$\rho_{ry}=\frac{\text{n}\,(\bar{y}_n-\bar{Y}_N)}{\sqrt{N\sigma_y^2\,\frac{n}{N}(N-n)}}=\frac{(\bar{y}_n-\bar{Y}_N)}{\sigma_y}\sqrt{\frac{f}{1-f}}.$$

We can now use our probability sample $S_2$: $y_1,\ldots y_{n_2}$, and its known OSWs $W_2$ =:$(W_1,\ldots W_{n_2})$ to estimate unknown population mean $\bar{Y}_N$, variance $\sigma_y^2$, and population size N:

$$\hat{\bar{Y}}_N=\frac{\sum_{i=1}^{n_2}W_i\,y_i}{\sum_{i=1}^{n_2}W_i},\ \hat{\sigma}_y^2=\frac{\sum_{i=1}^{n_2}W_i(y_i-\hat{\bar{y}})^2}{\sum_{i=1}^{n_2}W_i},\ \hat{N}=\sum_{i=1}^{n_2}W_i.$$

Thus, $\hat{\rho}_{ry}=\frac{\bar{y}_n-\hat{\bar{Y}}_N}{\hat{\sigma}_y}\sqrt{\frac{\hat{f}}{1-\hat{f}}}$,= $\frac{\bar{y}_n-\hat{\bar{Y}}_N}{\hat{\sigma}_y}\sqrt{\frac{(\hat{N}-n_2)}{n_2}}$, where $\hat{f}=\frac{n_2}{\hat{N}}$.. Note defect size, $\hat{\rho}_{ry}$ partly depends on $\sqrt{\frac{\hat{N}}{n_2}}$, i.e., the larger $\hat{N}$ is in relation to $n_2$, the bigger the defect becomes. Meng (2018) called it a large data paradox, but it is not true in general. For example, common sense tells us that input of more information (bigger N) in developing self-driving cars, makes safer cars. Medical imagining, which is a pillar in diagnostic health, involves a high volume of data collection i.e. large N, and processing; these data are not biased but they can be unstructured. For our $S_1$ data, $\hat{\rho}_{ry}$=0.006, which means that it has a large selection bias or defect (Meng, 2018) especially if the sample size is large; in our case, it is just about 1500.

Survey organizations are trying to move away from probability sampling to reduce high cost (Sakshaug et al. 2019 and Wisniowski et al. 2020). Instead, they use nonprobability sample (for example, web samples) which is less costly and easily available, but possibly brings in biases into the sample. To reduce this bias of nonprobability sample, they take a small probability sample from the same population. Then propose a Bayesian model to combine these samples in a way that exploits one sample's strengths to overcome the other sample's weakness.

Chen et al. (2020) developed a model for nonprobability sample with a small probability sample under the assumption of ignorable response, i.e., selection probability $\pi_i$ = p($R_i$ =1 |$x_i, y_i$) = p($R_i$=1 | $x_i$). This assumption (Rubin 1976) is also used in nonresponse study (Nandram and Choi 2002a, 2002b, Nandram and Choi 2006, Nandram and Choi 2010).

Potthoff et al. (1992) obtained adjusted survey weights (ASW) corresponding to the original weights. The original weights $W_i$ = 1/$\pi_i$,i=1,…,n, where $\pi_i$ is the known selection probability of the ith unit from a population, and N =$\sum_{i=1}^{n}W_i$, is the finite population size. With these original weights they calculate the adjusted weights that are used to obtain more reasonable variance; as the original sample weights give too small variance. Nandram (2007) used surrogate sampling to sample the entire population after an adjusted sample model is fit to the data.

Our main contribution in this paper is to integrate a non-probability sample and a probability sample to make inference about a finite population mean using a very simple and easy to understand method. Specifically, we have made six important contributions.

(a) Obtain the survey weights for the non-probability sample using record linkage instead of the less robust logistic regression; see Chen et al. (2020).

(b) Use both Bayesian and non-Bayesian methods to predict the finite population mean. Non-Bayesian methods are based on least squares and double robust estimators.

(c) Use bootstrap to obtain the distributions, including expectation and variance, of the least square and double robust estimators. We note that this step is not needed in the Bayesian method.

(d) For the simple situation we discussed, we showed that there is little difference among the different estimators.

However, the Bayesian method will become more useful in more complicated applications, and analytical approximations (e.g., Taylor's series expansion) can be avoided.

(e) We also estimate the correlation between the participation variable and the study variable for the non-probability sample. This is done by supplementing the non-probability sample with the probability sample. In Meng (2018), this correlation is non-identifiable and it cannot be estimated; this leads to a poor representation in Meng (2018) but he did not have a probability sample.

(f) Like Chen et al. (2020), we showed how to use inverse probability weighting when the population size and the nonsample covariates are unknown. This is done even for the Bayesian method, much beyond Chen et al. (2020).

This paper is divided into five sections including the introduction. In Section 2, first we derive the ASWs and estimate the population means with $S_2$,and separately done with $S_1$ and $S_3$ following the same process. In Section 3, we introduce two methods, non-Bayesian and Bayesian. In the non-Bayesian approach, we primarily use the LS method. To obtain the distribution of these estimators, we use the bootstrap method, which provides reasonable variance estimation. Section 4 illustrates our methods by analyzing BMI data of NCHS. Finally, the Section 5 has a brief conclusion.

## 2. ASW and Finite Population Mean

Consider population U with study variable $\mathbf{Y} = (Y_1, \ldots . Y_N)$. Two samples are taken from U. One is a large sample $S_1$ of $\boldsymbol{y_1} = (y_{11}, \ldots . y_{1n_1})'$. The other is a relatively smaller sample $S_2$ of $\boldsymbol{y_2} = (y_{21}, \ldots . y_{2n_2})'$. .A third sample $S_3$ is created by combining the two, $\boldsymbol{y_3} = (y_{11}, \ldots . y_{1n_1}, y_{21}, \ldots . y_{2n_2})'$.

Table 1. Three Types of samples: $\boldsymbol{S_1}$, $\boldsymbol{S_2}$ and $\boldsymbol{S_3} = (\boldsymbol{S_1} \cup \boldsymbol{S_2})$

| $S_1$ | | | $S_2$ | | | $S_3$ | | |
|---|---|---|---|---|---|---|---|---|
| $\boldsymbol{x_1}$ | $\boldsymbol{W_1}$ | $\boldsymbol{y_1}$ | , $\boldsymbol{x_2}$ | $\boldsymbol{W_2}$ | $\boldsymbol{y_2}$ | $\boldsymbol{x_3}$ | $\boldsymbol{W_3}$ | $\boldsymbol{y_3}$ |
| $x_{11}$ | $W_{11}$ | $y_{11}$ | $x_{21}$ | $W_{21}$ | $y_{21}$ | $x_{31}$ | $W_{31}$ | $y_{31}$ |
| . | . | . | . | . | . | | | |
| . | . | . | . | . | . | | | |
| . | . | . | $x_{2n_2}$ | $W_{2n_2}$ | $y_{2n_2}$. | | | |
| $x_{1n_1}$ | $W_{1n_1}$ | $y_{1n_1}$ | | | | | | |
| | | | | | | $x_{3n_3}$ | $W_{3n_3}$ | $y_{3n_3}$ |

In Table 1, $S_1$ does not have the original sample weights (OSWs) $W_1 = (W_{11}, \ldots, W_{1n_1})$, while the probability sample $S_2$ has no missing values with the known OSWs $W_2 = (W_{21}, \ldots, W_{2n_2})$. We change the subscripts $(11,\ldots,1n_1,21,\ldots,2n_2)$ to $(31,\ldots,3n_3)$. From here on, bold face symbolizes a matrix or vector. Let the respective covariates be $x_1 = (x_{11}, x_{12}, \ldots, x_{1n_1})'$, $x_2 = (x_{21}, x_{22}, \ldots . x_{2n_2})'$, and $x_3 = (x_{31}, x_{32}, \ldots . x_{3n_3})'$. Note that each component (e.g., $x_{11}$) of the vectors is a column vector and $x_1$ is an $n_1$x p matrix, where p is the number of common covariates, including an intercept, e.g., age, race, sex, their contents are generally different. For example, age is common in three samples, but the ages of two persons may be different, e.g., one age is 20 years and other 60 years, etc

Missing values may arise from different situations, e.g., cell in a table, sensitive information such as persons' income in a survey, small area missing in probability sample, and nonresponses in survey questions. To estimate missing values, one may use ratio estimation (Cochran, 1977) , covariate matching, nearest neighbor method, regression method (Chen et al. 2020), mean estimation for missing cell in a table or other imputation method. The choice of a method depends on each specific situation and a researcher's preference.

There may be common sample units belonging to both samples $S_1$ and $S_2$ as they are coming from the same population U. The size of nonprobability sample $S_1$ is much bigger than that of probability sample $S_2$, $n_2 \ll n_1$. We assume that these two samples are mutually exclusive and independent.

The purpose of this paper is to estimate the finite population mean $\hat{Y}_N = \sum_{i=1}^{N} \frac{Y_i}{N}$, from these three samples as accurately as possible and reducing bias and variance if possible. In this paper, we created the two samples, $S_1$ and $S_2$ from the NCHS data.

Some researchers use logit function under ignorable assumption (Little and Rubin, 2002, Chen et al. 2020, Rubin,1976). Under this assumption, Chen et al. (2020) obtain propensity scores $\pi_{1i} = \frac{e^{x_{i1}'\beta}}{1+e^{x_{1i}'\beta}}$ for the nonprobility sample; than

unknown weights are $W_{1i} = \frac{1}{\pi_{1i}}$.

Weights are used to reduce the bias introduced by sampling designs (Potthoff et al. 1992) and prediction is done using surrogate sampling (e.g., Nandram 2007) via inverse probability weighting (e.g., Chen et a. 2020). Assume that the original weights are known and fixed; we also assume that the imputed weights are fixed (not random) and known.

*2.1 ASWs With Probability Sample* $\boldsymbol{S_2}$

The study variable $y_{2i}$ is observed and the sample weights are known for sample $S_2$. Note that, although both samples have the common covariates, age, race and sex, their contents are different. Hence, when they match, the matching should be done by content matching.

The OSWs $\boldsymbol{W}_2$ of the sample $S_2$ are already known. We use these weights $\boldsymbol{W}_2$ to reduce possible bias arising from this sampling design. Let $S_2$ be the sample taken from the population U with response variables, $\boldsymbol{y} = (y_1, \ldots . y_{n_2})$. The selection probabilities $\pi=(\pi_1, \ldots . \pi_{n_2})$ are known, hence, OSWs, $\boldsymbol{W_2}=(W_{21}, \ldots . W_{2n_2}), W_{2i}= 1/\pi_{2i}$, i=1,…,$n_2$ are known. We assume that these weights are fixed, not random. We need to rescale the OSWs because they make the variance too small; see Potthoff et al. (1992).

We drop the subscript 2 of $S_2$ for general application. The model $\boldsymbol{y} = \boldsymbol{\beta x} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim (0, \sigma^2 I)$, for any probability sample of size n.

Each $y_i$ is adjusted by its corresponding known OSWs to obtain the sample mean.

$$\bar{y} = \frac{\sum_{i=1}^{n} W_i y_i}{\sum_{i=1}^{n} W_i},$$

Estimator of the finite population mean and

$$\text{var}(\bar{y}) = \frac{\sum_{i=1}^{n} W_i^2 \sigma^2}{\left(\sum_{i=1}^{n} W_i\right)^2} = \frac{\sigma^2}{\frac{\left(\sum_{i=1}^{n} W_i\right)^2}{\sum_{i=1}^{n} W_i^2}} = \frac{\sigma^2}{\hat{n}}, \text{ where } \text{var}(y_i) = \sigma^2.$$

We call $\hat{n} = \frac{\left(\sum_{i=1}^{n} W_i\right)^2}{\sum_{i=1}^{n} W_i^2}$ effective sample size (ESS). $1 \leq \hat{n} \leq n$. Finally, we define the adjusted sample weights (ASWs) as

$$w_i = \hat{n} \frac{W_i}{\sum_{i=1}^{n} W_i}, \text{ i} = 1\ldots,n..$$

Note from here on upper case **W** for OSWs and lower-case **w** for ASWs

For the BMI data in Section 4, the effective sample size $\hat{n}$ = 123 for the example of $S_2$, and the actual sample size $n_2$= 304, out of the population N = 2,360,624.

**Theorem 1**

Suppose, from a population $\mathbf{Y}=(Y_1, \ldots, Y_N)$, we take a probability sample $\mathbf{y}= (y_{1,\ldots\ldots,} y_n)$. $\mathbf{Y}|\boldsymbol{\beta}, \boldsymbol{x} \sim N(\boldsymbol{x\beta}, \sigma^2 I)$. We assume OSWs $\boldsymbol{W}=(W_1, \ldots . W_n)$ are known and fixed, and the ASWs $\boldsymbol{w}=(w_1, w_2, \ldots . w_n)$ are given, then we can have the model,

$$y_i \sim N\left(\boldsymbol{x}_i'\boldsymbol{\beta}, \frac{\sigma^2}{w_i}\right), \text{ for i =1, … , n.}$$

Appendix C shows the proof.

Theorem 1 states that there is actually an increase in variance over W, i.e., the variance with ASW is $\frac{\sigma^2}{w}$, while the variance with OSW is $\frac{\sigma^2}{W}$ which is much smaller than the first. The variance is much too small over the OSW with large W while ASWs provide more reasonable variance with smaller w.

From the sample model, $y_i \sim N\left(\boldsymbol{x}_i'\boldsymbol{\beta}, \frac{\sigma^2}{w_i}\right)$, we can estimate the parameters $\boldsymbol{\beta}$ and $\sigma^2$ by LS method (Appendix B) . First, we can draw the $y_i$, i=1, …, n, sample from this model, this is called surrogate sampling (Nandram 2007).

However, we do not know the non-sampled covariates, especially for big data, and this creates an important practical issue. Therefore, we obtain the estimate of population mean model,

$$\bar{Y}_N|\boldsymbol{\beta},\sigma^2,\boldsymbol{y}\sim N\left(\bar{\boldsymbol{x}}'\boldsymbol{\beta},\frac{\sigma^2}{\mathrm{N}}\right).$$

We can then generate samples from this distribution when $\boldsymbol{\beta},\sigma^2$ become known.

Here average covariate $\bar{\boldsymbol{x}}$ is estimated by inverse probability weighting, $\frac{\sum_{i=1}^{n} W_i x_i}{\sum_{i=1}^{n} W_i}$ and N=$\sum_{i=1}^{n} W_i$ with OSWs and **y** is generated data from the model for $y_i$.

Finally, we present an issue on unbiasedness. We have two samples, one is $S_1$ and the other $S_2$ from the finite population with true unknown mean $\bar{Y}_N$. Note that the $S_1$ is a convenient sample, whose weights are unknown. We have two sample based estimators of unknown population mean $\bar{Y}_N$ are: one is from $S_1$ and the other from $S_2$:

$$\bar{y}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} y_{1i} \text{ and } \bar{y}_2 = \frac{\sum_{i=1}^{n_2} \boldsymbol{W_{2i}y_{2i}}}{\sum_{i=1}^{n_2} W_{2i}},$$

where $W_{2i}$ are the OSWs of the probability sample. Under probability sampling design D, $\bar{y}_2$ is unbiased because it is a Horvitz-Thompson estimator, but $\bar{y}_1$ is biased precisely because it is not probability sample (it does not have the sampling weights). The bias of $\bar{y}_1$ is $\bar{y}_1$- $\bar{Y}_N$ where

$$\bar{y}_1 - \bar{Y}_N = \bar{y}_1 - \bar{y}_2 - (\bar{Y}_N - \bar{y}_2).$$

But $(\bar{Y}_N - \bar{y}_2) \approx 0$ because $E_D(\bar{Y}_N - \bar{y}_2) = 0$, expectation is taken over the probability sampling design D. Then, the bias is $\bar{Y}_N - \bar{y}_2 \approx \bar{y}_1 - \bar{y}_2$,the difference between the two sample estimators.

The ESS and ASWs of $S_1$ and $S_3$ can be obtained in a similar manner as shown in (2.2) and (2.3), respectively.

*2.2 ASW With Nonprobability Sample* $\boldsymbol{S_1}$

For probability sample $S_2$, the selection probabilities $\boldsymbol{\pi_2}$ (hence the weights) are known. When the contents of covariates (i.e., age, sex, race) of units in probability sample $S_2$ match with those of sample $S_1$.The missing OSWs $\boldsymbol{W_1}$ for $S_1$ are filled with those of the known OSWs $\boldsymbol{W_2}$ of $S_2$ by covariate matching via Mahalanobis distances (Appendix A).

Let the response variable $R_i = I(i \in S_1)$, the indicator variable**.** The response variable $R_i$ = 1 if the unit i $\in S_1$ and 0 otherwise. Assume $P_D(R_i = 1\,|x_i, y_i) = P_D(R_i = 1\,|x_i)$ of missing at random (Rubin 1976, Little and Rubin 2002, Chen et al 2020) regardless the design D of the sampling of $S_1$. Under this assumption of missing at random, the propensity score can be obtained by the logistic function $\pi_i=\frac{e^{x_i'\boldsymbol{\beta}}}{1-e^{x_i'\boldsymbol{\beta}}}$, i=1, …,$n_1$. Hence, the weights of nonprobability sample $S_1$ is $W_i=\frac{1}{\pi_i}$, i=1, …,$n_1$. However, this propensity score makes too much unreasonable assumptions on logistic function, and hardly reflects the true situation. Thus, in this paper, we do not use logistic function to obtain its unknown OSWs of $S_1$. The main reason that Mahalanobis distances does not require such unnecessary assumptions.

Assume the imputed OSWs $W_1$ are also fixed, not random, as done with the probability sample $S_2$. In actual sample in Section 4, size $n_1$= 1,563 for $S_1$, its ESS $\hat{n}_1 = \frac{(\sum_{i=1}^{n_1} W_i)^2}{\sum_{i=1}^{n_1} W_i^2}$ = 722, and the population size N=2,370,624. ASWs are $w_{1i} = \hat{n}_1 \frac{W_i}{\sum_{i=1}^{n_1} W_i}$, i = 1, …,$n_1$, $1 \le \hat{n}_1 \le n_1$.

For $S_1$, the weight-adjusted model takes the same form for $y_{1i}\sim N\left(x_i'\beta, \frac{\sigma^2}{w_{1i}}\right)$ and $\bar{Y}_N \mid \beta,\sigma^2,y\sim N\left(\bar{x}_1'\beta,\frac{\sigma^2}{\mathrm{N}}\right)$, as those of $S_2$, where adjusted covariates $\bar{x}_1$ is estimated by inverse probability weighting $\frac{\sum_{i=1}^{n_1} W_i x_i}{\sum_{i=1}^{n_1} W_i}$ and N=$\sum_{i=1}^{n_1} W_{1i}$ and $W_{1i}$

are imputed; ASWs $w_{1i}$ are used only in estimation, not in prediction.

*2.3 ASW With Combined Sample* $\boldsymbol{S_3}$

We assumed that $S_1$ and $S_2$ are independent, and $\boldsymbol{y_1}\sim N(\boldsymbol{x_1'\beta}, \sigma^2 W_1^{-1})$ and $\boldsymbol{y_2}\sim N(\boldsymbol{x_2'\beta}, \sigma^2 W_2^{-1})$. Combining these two samples $n_3=n_1+n_2$. $\boldsymbol{y_3}$ is $n_3$ x1 vector, $\boldsymbol{y_3} = (\boldsymbol{y_1}', \boldsymbol{y_2}')'$, and $\boldsymbol{x_3}= (\boldsymbol{x_1}, \boldsymbol{x_2})$..

$$\boldsymbol{y_3} = \boldsymbol{x_3'\beta} + \boldsymbol{\varepsilon}, \ \boldsymbol{\varepsilon}\sim \boldsymbol{N(0, \sigma^2 W_3^{-1})}.$$

$W_3= (W_1, W_2), W_2 = (W_{21}, \dots, W_{2n_2})$ is known OSW for $\boldsymbol{S_2}$ and $W_1 = (W_{11} \dots, W_{1n_1})$ imputed OSW for $\boldsymbol{S_1}$ and the combined OSW $W_3= (W_{31}, \dots, W_{3n_3})$ for $\boldsymbol{S_3}$. Following the same process for $S_2$. the sample variance for $\boldsymbol{y_3}$ is $\frac{\sigma^2}{\hat{n}_3}$

The ESS $\hat{n}_3 = \frac{(\sum_{i=1}^{n_3} W_{3i})^2}{\sum_{i=1}^{n_3} W_{3i}^2}$, and ASWs $w_{3i} = \hat{n}_3 \frac{W_{3i}}{\sum_{i=1}^{n_3} W_{3i}}$, i = 1,…,$n_3$, for the combined sample. Bias adjusted sample model for the combined sample $S_3$ is

$y_{3i}\sim N\left(\boldsymbol{x_i'\beta}, \frac{\sigma^2}{w_{3i}}\right)$, i = 1,…,$n_3$, by Theorem 1.

We estimate the parameters $\boldsymbol{\beta}$ and $\sigma^2$ by LS method from this model. The finite population mean with the combined sample $S_3$ is

$$\bar{Y}_N \mid \boldsymbol{\beta}, \sigma^2, \boldsymbol{y_3}\sim N\left(\bar{\boldsymbol{x}}_3'\boldsymbol{\beta}, \frac{\sigma^2}{N}\right)$$

where adjusted covariates $\bar{\boldsymbol{x}}_3$ is estimated by $\frac{\sum_{i=1}^{n_3} W_i \boldsymbol{x_i}}{\sum_{i=1}^{n_3} W_i}$.

## 3. Estimation of Finite population Mean Using Different Methods

We use two different approaches for the estimation of the finite population mean, Bayesian and non-Bayesian. The Bayesian approach is a completely closed form 95% highest posterior density interval (HPDI) for the finite population mean. We also present a sampling-based method, which is slightly more convenient. The non-Bayesian approach is basically LS. It also includes doubly robust method (DR), (e.g., Chen et al. 2020). When the variance from LS is too small for large N, we use the bootstrap to calculate reasonable variance and to obtain distributions. Each method uses data sets $S_1$, $S_2$, and $S_3$, separately for the estimation. DR method uses different estimation by combining $S_1$ and $S_2$, not $S_3$.The Bayesian and non-Bayesian methods are presented in Section 3.1 and 3.2, respectively.

*3.1 Bayesian Approach*

To fit a general model, we drop the subscriptions for sample numbers, suppose a random sample $y_{1,\dots,}y_n$ is taken from the population $y_1, \dots, y_N|\beta, \sigma^2\sim$f(y|$\beta, \sigma^2, x$), and $\beta$ p x 1 vector with the adjusted weights. We assume f(y|$\beta, \sigma^2, x$) is normal function, then

$$y_i|\beta, \sigma^2\sim N(x_i'\beta, \frac{\sigma^2}{w_i}),\ \text{i=1,…,n,}$$

and a priori,

$$\pi(\boldsymbol{\beta}, \sigma^2) \propto \frac{\mathbf{1}}{\boldsymbol{\sigma^2}}.$$

This is Jeffreys' improper prior, and it is well-known that the joint posterior density is in closed form and proper. The finite population mean is

$$\bar{Y}_N \ |\boldsymbol{\beta}, \sigma^2\sim N(\bar{\boldsymbol{X}}'\boldsymbol{\beta}, \frac{\sigma^2}{N}),\ \bar{X} = \frac{\sum_{i=1}^{N} x_i}{N}.$$

Let **x** is n x p matrix of covariates and **W**= diag($w_1$,…,$w_n$) is n x n diagonal matrix, and **y** is the n x 1 vector of response variables. Then, letting

$$\hat{\beta} = (x'\,W\,x)^{-1}(x'\,W\,y),\ \text{and}\ S^2=\frac{\sum_{i=1}^{n} w_i(y_i - x_i'\,\hat{\beta})}{n-p},$$

It is easy to show that

$$\frac{\bar{Y}-\bar{X}'\hat{\beta}}{\bar{S}^*} \mid \mathbf{y} \sim t_{n-p},$$

has a Student's t density on n-p degrees of freedom, where $\bar{S}^* = S\sqrt{\frac{1}{N} + \bar{X}'(\sum_{i=1}^{n} w_i x_i x_i')^{-1}\bar{X}}$.

It follows that $E(\bar{Y}_N|y)=\bar{X}'\hat{\beta}$,

$$\text{Var}(\bar{Y}_N|y)=\frac{n-p}{n-p-2}S^{*2},\ n > p+2.$$

Two-sided 95% highest posterior density interval for $\bar{Y}_N$ is $\bar{X}'\hat{\beta} \pm SD(\bar{Y}_N|y)\ t_{(n-p),0.025}$, where $t_{(n-p),0.025}$, is the 97.5 percentage point of t-distribution. This is nice, but the non-sampled covariates are all known, and it is convenient to use a sampling based method to make posterior inference about $\bar{Y}_N$. We can use inverse probability weighting, where

$$\bar{X} = \frac{\sum_{i=1}^{N} W_i x_i}{N} \text{ and } N=\sum_{i=1}^{n} W_i$$

Note that $w_1$,$w_2$,and $w_3$ are the ASWs for the sample $S_1$, $S_2$, and $S_3$, respectively. To fit the model for the individual data $S_1, S_2,$ or $S_3$, we follow the above steps by adding sample subscript 1, 2 or 3. The model is the same for the individual data $S_1, S_2,$ or $S_3$; so we present only $S_2$.

Specifically, we discuss how to fit the model for $S_2$,

$$y_{2i}|\ \beta, \sigma^2 \sim N(x'_{2i}\beta, \frac{\sigma^2}{w_{2i}}),\ i=1,,,,n_2,$$

$$\pi(\beta, \sigma^2) \propto \frac{1}{\sigma^2}.$$

Let $\hat{\beta} = (x_2'W_2x_2)^{-1}(x_2'W_2y_2)$, and $\Delta=(x_2'W_2x_2)^{-1}$.
Then,

$$\sigma^2|\ y_2 \sim \text{Gam}(\frac{n_2-p}{2},\ \frac{\sum_{i=1}^{n_2} W_{2i}\,(y_{2i}-x'_{2i}\widehat{\beta})^2\,(y_{2i}-x'_{2i}\widehat{\beta})}{2}),$$

$$\beta|\sigma^2,\ y_2\ \sim N(\hat{\beta}, \sigma^2\Delta),$$

$$\bar{Y}_N|\ \beta, \sigma^2, y_2\ \sim N(\bar{x}'_2\beta, \frac{\sigma^2}{N}).$$

We sample $\sigma^2|y_2\ and,\ \beta|\sigma^{2,}, y_2$, and finally $\bar{Y}_N|\beta, \sigma^2, y_2$.

The finite population mean $\bar{Y}_N|\beta, \sigma^2 \sim N(\bar{x}'_2\beta, \frac{\sigma^2}{N})$, $\bar{x}_2$ is calculated by inverse probability weighting $\frac{\sum_{i=1}^{n_2} W_i x_i}{\sum_{i=1}^{n_2} W_i}$.

For the simple model we consider here, there is virtually no practical difference between the non-Bayesian method (via the bootstrap) and the Bayesian method; the interpretations are different. However, for a more complicated model, a Bayesian method will be more attractive as it avoids analytical approximations.

*3.2 Non-Bayesian Approach*

Two methods will be introduced to estimate finite population mean: LS and DR estimator. When LS produces variance which is too small for inference, we use Bootstrap to get larger variance. Distributions are obtained by using the bootstrap.

**Least Square (LS) estimation**

LS uses linear model and minimizes variance for estimation. The linear model for all three data sets is the same. Below we show the model for combined data. Set a linear relationship of study variable $y_k$: to the predictor variable $x_k$ with the common parameter $\beta_k = (\beta_1, \beta_2, \ldots \beta_p)'$ for k = 1, 2, 3. So we are using one data of k=3 to show the LSE. Other data follow the same steps. Here,

$y_3 = (y_1, y_2)$, $x_3 = (x_1, x_2)$, original weights $W_3 = (W_1, W_2)$, adjusted weights $w_3 = (w_1, w_2)$, and $n_3 = n_1 + n_2$,

$$\boldsymbol{y_3} = \boldsymbol{x_3'\beta} + \boldsymbol{\varepsilon}$$

$\boldsymbol{\beta}$ and $\varepsilon \sim (0, \sigma^2 W_3^{-1})$ estimated by LSE:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{x_3' W_3 x_3})^{-1}(\boldsymbol{x_3' W_3 y_3}), \text{ and } \widehat{\boldsymbol{\sigma}}^{\boldsymbol{2}} = \frac{(\boldsymbol{y_3} - \boldsymbol{x_3}\widehat{\boldsymbol{\beta}})' \boldsymbol{W_3} (\boldsymbol{y_3} - \boldsymbol{x_3}\widehat{\boldsymbol{\beta}})}{n_3 - p}.$$

Appendix B has a brief and clear presentation.

Weight adjusted sample model is $y_{3i} \sim N\left(x_{3i}'\boldsymbol{\beta}, \frac{\sigma^2}{w_{3i}}\right)$. The population mean $\bar{Y}_N$ is distributed by

$$\bar{Y}_N | \boldsymbol{\beta}, \sigma^2, \boldsymbol{y} \sim N\left(\bar{\boldsymbol{x}}_3'\boldsymbol{\beta}, \frac{\sigma^2}{N}\right),$$

$$\text{Var}(\bar{Y}_N) = \boldsymbol{a}_3' \text{var}(\hat{\beta}) a_3, \ \boldsymbol{a}_3 = \sum_{i=1}^{n} W_{3i}\, \boldsymbol{x}_{3i},$$

The variance is too small when N is large, as an alternative, we use bootstrap method to calculate variance as shown before in the bootstrap method.

**Doubly robust method (DR)**

DR estimator of population mean (Chen et al 2020) is given by

$$\bar{Y}_{DR1} = \frac{1}{\hat{N}^{S_1}} \sum_{i=1}^{n_1} \frac{(y_i - \boldsymbol{x_i'}\widehat{\boldsymbol{\beta}})}{\pi_{S_1 i}} + \frac{1}{\hat{N}^{S_2}} \sum_{i=1}^{n_2} \frac{\boldsymbol{x_i'}\widehat{\boldsymbol{\beta}})}{\pi_{S_2 i}},$$

$\hat{N}^{S_1} = \sum_{i \in S_1} \frac{1}{\pi_{S_1 i}}$ and $\hat{N}^{S_2} = \sum_{i \in S_2} \frac{1}{\pi_{S_2 i}}$ are the two estimates of finite population totals, using two samples, $S_1$.and, $S_2$,

respectively. If $\hat{N}^{S_1} = \hat{N}^{S_2} = \sum_{i \in S_2} \frac{1}{\pi_{S_2 i}} = \sum_{i=1}^{n_2} W_{2i}$, DR estimator of the finite population mean is given by

$$\bar{Y}_{DR2} = \frac{\sum_{i=1}^{n_1} W_{1i}(y_{1i} - \boldsymbol{x_{1i}'}\widehat{\boldsymbol{\beta}}) + \sum_{i=1}^{n_2} W_{2i} \boldsymbol{x_{2i}'}\widehat{\boldsymbol{\beta}}}{\sum_{i=1}^{n_2} W_{2i}},$$

where $\widehat{\boldsymbol{\beta}}$ is the LS estimator obtained from the probability sample $S_2$. From the numerical example on BMI, the estimated finite population mean is $\bar{Y}_{DR2}$=27.075. Chen et al (2020) claim that the first equation $\bar{Y}_{DR1}$is doubly robust estimator; it is not clear why it is doubly robust. But since they use the logistic function for propensity scores and LS estimation for $\boldsymbol{\beta}$, they made several assumptions which may not be robust. Therefore, we used the matching method to estimate the propensity scores. Although we do not have a participation model, it is still useful to make comparisons with the DR estimators.

Because we have the **y** values in $S_2$ data, we replace $\boldsymbol{x_{2i}'\beta}$ with **y**, and we have a third DR estimator,

$$\bar{Y}_{DR3} = \frac{\sum_{i=1}^{n_1} W_{1i}(y_{1i} - \boldsymbol{x_{1i}'}\widehat{\boldsymbol{\beta}}) + \sum_{i=1}^{n_2} W_{2i}\, y_{2i}}{\sum_{i=1}^{n_2} W_{2i}}.$$

Note that $\bar{Y}_{DR3}$ does require LS estimation of $\beta$ from $S_1$, not $S_2$ .

The inverse probability weighting estimators are sensitive to misspecified models for the propensity scores especially when they are very small; see Chen et al. (2020). If the participation model or the study variable model is correctly specified, this provides the double robustness property that is widely used in recent literature on missing data problems. We do not have a model for the participation variable because we use record linkage to fill in the propensity scores for the non-probability sample with the probability sample being the donor. So the double robust estimator is not really needed in our case, but we use it for comparison.

**Bootstrap used to calculate variance for LS and DR estimates**

We have seen that LS procedure gives very small, overly optimistic, estimates. This is mostly due to the large population size N. So, we have used the bootstrap method. We draw B=1,000 bootstrap samples from each of $S_1$, $S_2$, and $S_3$, get $x_{s_k b}$, $\boldsymbol{y}_{s_k b}$, $W_{s_k b}$, $w_{S_k b}$, b = 1,…,B, k = 1, 2, 3.

Here we drop the subscript k of $S_k$, and $n_k$ for sample indicator as they are applicable in general.

$a_b = \sum_{i=1}^{n} W_{ib}\, x_{ib}$, b=1,…,B. Let $\hat{\beta}_b$ denote the LS estimators.

Then we compute the finite population mean as

$$\bar{Y}_{Nb} = \boldsymbol{a}_b' \widehat{\boldsymbol{\beta}}_b, \text{ b = 1,…,B,.}$$

and bootstrap variance obtained. The bootstrap variances are shown in the tables for the LS and DR methods in Section 4. Note that the formula for LS estimation gives too small variance as shown Table 3 while bootstrap estimators in Table 4 give better representation for the variance of $\bar{Y}_{Nb}$.

We bootstrap the data, both $S_1$ and $S_2$, 1000 times in the conventional way. That is, we sample the $S_1$ data and $S_2$ data respectively with replacement to constitute a single sample. For each bootstrap sample, we compute the least square estimator and the double robust estimator. This gives us a sample of 1000 values of the estimators, and a kernel density estimator is obtained. The mean and variance of the estimator are obtained as summaries from these 1000 values and the 95\% confidence intervals are obtained by using the percentile method; further refinement can be done, of course.

## 4. Data Analysis

We now present detailed results from the BMI data using our estimators. We are assuming, when the weights of $S_2$ are used, it provides unbiased estimate of the finite population mean. Therefore, we use the estimate from the $S_2$ as a 'gold' standard for inference. However, because the $S_2$ is relatively small, we expect it to provide an estimate with a relatively large standard error. There are two different methods, Bayesian and non-Bayesian. In the Bayesian approach, we provide posterior summaries and distributions (e.g., posterior mean, posterior standard deviation, posterior coefficient of variation, and 95% highest posterior density interval – HPDI). In the non-Bayesian approach, we use LS and DR estimates. The bootstrap is used to get distributions of the corresponding estimators and its variances. Analogous to the Bayesian approach, we have used LS estimate, its standard error, coefficient of variation and 95% confidence interval for the finite population mean.

The relative standard error (CV) of an estimate is obtained by dividing the standard error of an estimate by the estimate itself, well known as coefficient of variation, and it is usually expressed as a percentage, CV100%= $\frac{\sigma_{\bar{x}}}{\bar{x}}$ x 100. (Choi, 1977, Appendix). NCHS put an asterisk on the number when CV is greater than 30% to let the data users know the number is not reliable. Note that CV is often not accurate and one needs to be careful to use it for statistical inference.

For example, for x ~ Binomial(n, p), an unbiased estimator of p is $\hat{p} = x/n$; so var($\hat{p}$) = $\frac{p(1-p)}{n}$, E($\hat{p}$) = p and CV($\hat{p}$) = $\frac{\sqrt{p(1-p)}}{\sqrt{n}\ p}$ is very small as n becomes large (Choi and Nandram, 2021). The large sample causes problems not only here but other statistical tests (e.g., normal test and t-test) for hypothesis. In general, sample variance is also a function of n.

We use the body mass index (BMI) data from the National Health and Nutrition Examination Survey III (NHANES III), conducted from 1988 to 1994 in two phases by National Center for Health Statistics. There were 30,818 people examined from the US population of about 300 million and overall response rate was 78%. BMI is person's weight in kilograms divided by the square of height in meters. BMI is an inexpensive and easy screening method for health status. Table 2 below shows the weight status of a person.

Table 2. Weight (wt) Category and Range of BMI Classes

| Category | Under wt | Healthy wt | Over wt | obese I | obese II |
|---|---|---|---|---|---|
| Range | BMI<18.5 | 18.5<BMI<25 | 25<BMI<30 | 30<BMI<35 | BMI>35 |

We only use a small portion of sampled people from the California counties. We have used 6 counties for the $S_1$ and 2 counties for the $S_2$. The sample size of $S_1$ is about 1500 and that of $S_2$ is about 300; so $S_1$ is five times the size of the $S_2$ (i.e., the $S_2$ is relatively small although it is expected to give unbiased estimate) . Our $S_1$data, presented here, defect correlation $\hat{\rho}_{ry}$=0.006, which means that it has a large selection bias or defect (Meng, 2018).

We have non-Bayesian methods, LS and DR estimators, and the Bayesian method. Each of these uses the three data sets, separately, $S_1$, $S_2$, and $S_3$ for estimation. The bootstrap is used to get distributions for LS and DR estimators. As stated earlier, we obtain $S_1$, $S_2$ and $S_3$ from NCHS BMI data. For bootstrapping or the Bayesian method, we generate 1,000 population means of $\bar{Y}_N$. The distributions of these finite population means for the BMI data from $S_1$, $S_2$, and $S_3$, are shown in in Tables 3, 4, 5 and Figure 1.

### *4.1 LS Estimation*

The finite population mean of BMI is 27.175 for the nonprobability sample $S_1$, 25.985 for probability sample $S_2$, and

26.982 for the combined data $S_3$. It also shows the bias of the means of $S_1$ and $S_3$ are 1.190 (=27.175 -25.985) and 0.997(=26.982-25.985), respectively. The sample size of $S_2$ is the smallest $n_2$=304 and that of $S_1$ is $n_1$ =1,563 while the combined sample size is $n_3$=1,867. But the SDs remain the same, very small because all the three sample sizes are rather big, i.e., 304, 1,563, and 1,867, respectively. We used bootstrap, which is distribution free, to find the realistic variances in Table 4.

Table 3. LS estimation of the finite population mean for $S_1$, $S_2$, and $S_3$ (Mark * on CV greater than 30%)

| $S_1$ | Estimate | S.D. | CV100% | 95% C.I. |
|---|---|---|---|---|
| $\beta_0$ mean | 25.144 | 0.385 | 1.533 | 24.388 25.899 |
| $\beta_1$ age | 0.036 | 0.008 | 21.585 | 0.021 0.052 |
| $\beta_2$ sex | 0.710 | 0.386 | 54.319* | . - 0.046 1.466 |
| $\beta_3$ race | 0.558 | 0.261 | 46.793* | 0.046 1.070 |
| Finite Pop Mean | 27.175 | 0.002 | 0.008 | 27.171 27.180 |
| $S_2$ | | | | |
| $\beta_0$ mean | 23.277 | 0.835 | 3.587 | 21.608 24.947 |
| $\beta_1$ age | 0.056 | 0.016 | 28.581 | 0.025 0.087 |
| $\beta_2$ sex | 2.264 | 1.645 | 72.683* | -0'961 5.488 |
| $\beta_3$ race | 0.052 | 0.609 | 1163.536* | -1.141 1.246 |
| Finite Pop Mean | 25.985 | 0.002 | 0.008 | 25.981 25.990 |
| $S_3$ | | | | |
| $\beta_0$ mean | 24.960 | 0.350 | 1.400 | 24.275 25.645 |
| $\beta_1$ age | 0.037 | 0.007 | 18.888 | 0.023 0.051 |
| $\beta_2$ sex | 0.930 | 0.372 | 40.019* | 0.200 1.659 |
| $\beta_3$ race | 0.452 | 0.240 | 53.038* | - 0.018 0.922 |
| Finite Pop Mean | 26.982 | 0.002 | 0.008 | 26.977 26.986 |

Note that some of CVs* of the $\beta'$s are greater than 30%, for sex and race. But the contribution of sex and race to BMI are bigger, 0.710 and 0.558 than age contribution, 0.036 for $S_1$. This trend remains the same for other two tables 4 and 5. Note CV of race in the table of $S_2$, is very large, 1154%, mainly because of the small $\beta_3$ ( 0.052) and large SD (0.609). Finally, we note that the SD of the finite population mean is extremely small and this is due to the large sample size that appears in the variance of the least square formula.

*4.2 Bootstrap and DR Estimation*

Table 4 gives us the bootstrap estimations of the finite population means and variances of $\bar{Y}_B$, B=1,000 for $S_1$, $S_2$, and $S_3$. For $S_1$, $S_2$, and $S_3$ in the last row of Table 4 we add the bootstrap variance of LS estimation as the LS variances were too small. The SD of $S_1$ and $S_3$ are 0.141 and 0.127, respectively, while the SD of $S_2$ is 0.298 which is almost twice bigger than those of $S_1$ and $S_3$.This reduction is largely due to the large sample sizes. As a DR estimator of the finite population mean and SD for our data, $S_3$, we got BMI mean $\bar{Y}_{DR2}$= 26.161 (SD =0.524) which is smaller than the $\bar{Y}_{DR3}$= 27.093 (SD =0.267) in Table 4 with much smaller variance. Note that this is different combination from the one used in the LS or Bayesian, and that $\bar{Y}_{DR2}$, and $\bar{Y}_{DR3}$ are defined in DR method in Section 3.2

Table 4. Bootstrap estimation of the finite population mean for $S_1$, $S_2$, and $S_3$ (Mark * on CV greater than 30%)

| $S_1$ | Estimate | SD | CV100% | 95% C.I. |
|---|---|---|---|---|
| $\beta_0$ mean | 25.172 | 0.663 | 2.634 | 23.972 26.579 |
| $\beta_1$ age | 0.037 | 0.013 | 34.850* | 0.014 0.065 |
| $\beta_2$ sex | 0.788 | 0,867 | 110.036* | -0.794 2.602 |
| $\beta_3$ race | 0.585 | 0.493 | 84.370* | -0.456 1.492 |
| $\sigma^2$ variance | 9.219 | 2.469 | 26.783 | 4.660 13.129 |
| Finite Pop Mean | 27.264 | 0.284 | 1.041 | 26.752 27.852 |
| LS | 27.175 | 0.141 | 0.518 | 26.899 27.451 |
| $S_2$ | | | | |
| $\beta_0$ mean | 23.308 | 0.823 | 3.532 | 21.769 24.959 |
| $\beta_1$ age | 0.055 | 0.018 | 32.169* | 0.020 0.090 |
| $\beta_2$ sex | 2.233 | 1.559 | 69.807* | -0.658 5.305 |
| $\beta_3$ race | 0.086 | 0.727 | 841.517* | -1.359 1.412 |
| $\sigma^2$ variance | 11.274 | 2.707 | 24.015 | 6.668 16.613 |
| Finite Pop Mean | 25.995 | 0.390 | 1.501 | 25.334 26.865 |

| | | | | |
|---|---|---|---|---|
| LS | 25.985 | 0.298 | 1.148 | 25.401 26.570 |
| $S_3$ | | | | |
| $\beta_0$ mean | 24.900 | 0.568 | 2.281 | 23.84-0 25.957 |
| $\beta_1$ age | 0.039 | 0.011 | 28.713 | 0.019 0.062 |
| $\beta_2$ sex | 1.096 | 0.831 | 75.829* | -0.393 2.844 |
| $\beta_3$ race | 0.431 | 0.420 | 97.549* | - 0.317 1.287 |
| $\sigma^2$ variance | 9.832 | 2.229 | 22.668 | 5.937 13.675 |
| Finite Pop Mean | 26.976 | 0.249 | 0.924 | 26.478 27.454 |
| LS | 26.982 | 0.127 | 0.470 | 26.733 27.230 |
| $S_1$ & $S_2$ | | | | |
| DR estimation | 27.093 | 0.267 | 0.010 | 26.575 27.629 |

We assume the probability sample provides unbiased estimates of mean. The mean is 25.995 from the probability sample $S_2$ compared to the means 27.264 of nonprobability sample $S_1$, and to 27.093 of the combined sample $S_3$. Note that the bias of nonprobability sample is 1.269 (=27.264 – 25.995) and the bias of combined sample is 0.981 (=26.976 – 25.995), about three standard deviations.

On the other hand, the SD of $S_2$ is 0.390 while those of $S_1$ and $S_3$ are much smaller at 0.284 and 0.249, respectively. Here SD reflects sample size. i.e., SD of $S_2$ is 0.390, bigger than those, 0.284 and 0.249 of respective for $S_1$ and $S_3$. Some of the CV's of the $\beta$'s are greater than 30%, especially race and sex although the race and sex contribute more to the BMI than age for all three samples. But CV is not a good estimator as described above. SD is better for statistical inference than CV.

*4.3 Bayesian Estimation*

Table 5 shows posterior summaries. The finite population means of BMI are all close: 26.974, 26.074, and 26.887 for $S_1$, $S_2$ and $S_3$ respectively with standard deviations of. 0.129, 0.305, and 0,130.

Table 5. Posterior summaries of finite population mean for $S_1$, $S_2$, and $S_3$ (Mark * on CV greater than 30%)

| **$S_1$** | **Estimate** | **S.D.** | **CV100%** | **95% C.I.** |
|---|---|---|---|---|
| $\beta_0$ mean | 25.062 | 0.374 | 1.494 | 24.323 25.796 |
| $\beta_1$ age | 0.033 | 0.007 | 21.163 | 0.019 0.047 |
| $\beta_2$ sex | 0.832 | 0.337 | 40.527* | 0.163 1.489 |
| $\beta_3$ race | 0.774 | 0.259 | 33.452* | 0.287 1.287 |
| $\sigma^2$ variance | 11.593 | 0.616 | 5.311 | 10.423 12.819 |
| Finit Pop Mean | 26.974 | 0.129 | 0.480 | 26.725 27.234 |
| **$S_2$** | | **.** | | **. .** |
| $\beta_0$ mean | 23.574 | 0.821 | 3.482 | 21.924 25.117 |
| $\beta_1$ age | 0.045 | 0.015 | 33.402 | 0.015 0.073 |
| $\beta_2$ sex | 1.927 | 0.975 | 50.617* | -0.004 3.814 |
| $\beta_3$ race | 0.635 | 0.608 | .95.742* | -0.562 1.843 |
| $\sigma^2$ variance | 10.359 | 1.348 | 13.012 | 7.853 13.050 |
| Finit Pop Mean | 26.074 | 0.305 | 1.171 | 25.482 26.676 |
| **$S_3$** | | | | |
| $\beta_0$ mean | 25.311 | 0.348 | 1.374 | 24.652 26.006 |
| $\beta_1$ age | 0.029 | 0.007 | 22.673 | 0.015 0.041 |
| $\beta_2$ sex | 1.266 | 0.326 | 25.754 | 0.624 1.900 |
| $\beta_3$ race | 0.333 | 0.240 | 71.993* | -0.147 0.803 |
| $\sigma^2$ variance | 12.266 | 0.591* | 4.820 | 11.156 13.449 |
| Finit Pop.Mean | 26.887 | 0.130 | 0.483 | 26.630 27.140 |

The SD of $S_2$ is much bigger than those of $S_1$ or $S_3$ while the bias of the mean of $S_1$ is 0.900 (=26.974-26.074) (3.45%) and that of $S_3$ is 0.813 (=26.887-26.074) (3.12%) when compared to the mean BMI, 26.074, of $S_2$.The CVs of $\beta's$ for sex and race are 40.5* %, and 33.5%, respectively, for $S_1$, the CVs are 50.61* %, and 95.74* % for $S_2$, and CVs are 25% and 71.99* % for $S_3$. This means they are unreliable estimates except one 25%. As seen in these tables,

the pattern is consistent regardless of the method used for the estimation of finite population mean. i.e., larger variance with $S_2$, compared to other data sets, $S_1$ or $S_3$, and larger bias with $S_1$ or $S_3$ compared to that of $S_2$.

Figure 1 below includes seven density curves, A,B,C, and D,E,F,G. They are informative about the distributions of finite population mean by three data sets, $S_1$, $S_2$, $S_3$, and three methods: Bayesian (ABC), bootstrap (DEF), and DR (G) that uses both data sets, not $S_3$, and bootstrap. Vertical axis is for the heights of the distribution of population means and horizonal line is scale of BMI. The seven curves depend on both methods as well as sample size. A,D use $S_1$, A with Bayesian and D with Bootstrap. B,E use $S_2$, B with Bayesian and E with Bootstrap, and C,F use combined sample of $S_3 = (S_1 \cup S_2)$, C with Bayesian and F with Bootstrap. The G curve gives the bootstrap distribution of DR estimator.

They show the bias of the means, i.e., B,E curves of probability sample $S_2$, are assumed to have unbiased mean, which is used to find the size of bias of mean estimates from other two samples $S_1$ and $S_3$. The curves of B,E are shifted to the left side of other curves. This implies that the curves A,D of nonprobability sample, those of C,F of combined samples, and the double robust curve of G with combined sample, are biased to the right side of unbiased curves B,E. On the other hand, the bases of B,E of $S_2$ are wider than those of others, $S_1$ and $S_3$, implying that variance of B,E are bigger than those of others due to smaller sample size. The two highest ones among the seven curves are C of $S_3$ and A of $S_1$. The shortest ones are B,E of $S_2$ because of its smaller sample size.

## 5. Conclusion

Original sample weights are available in probability sample, but they are missing for nonprobability sample. We used these known weights to impute the unknown weights of nonprobability sample. Others use logistic model to find the propensity score to get unknown weights. But we use covariate matching via Mahalanobis distances to avoid unnecessary parametric assumptions.

We show a way to reduce possible bias arising from not using a sample design via ASWs obtained from the OSWs (known or imputed). The combined data help to reduce the variance compared to that of $S_2$ or $S_1$ alone. Therefore, it is another benefit of our methods, a potentially useful for data integration.

The finite population mean is estimated by two different methods using both the non-Bayesian and Bayesian approaches. In the non-Bayesian approach, we use least squares estimators and the DR estimator. Bootstrap is used to calculate variance for the LS and DR estimators, especially when variance is too small for large N. We also obtain the distribution of the finite population mean using the bootstrap in the non-Bayesian approach. With the two methods, non-Bayesian and Bayesian, each method separately uses three different data sets, $S_1$, $S_2$, and $S_3$ to find the finite population mean. The DR estimator combines $S_1$ and $S_2$ in a different way, not the combined sample $S_3$ and Bootstrap method, as the result shown in the last row of Table 2.

Each data set gives an estimate of the finite population mean. The estimates are close within normal range between 26 to 27 regardless of the data types and methods used. The tables and Figure 1 show the significant bias of the means from $S_1$ and $S_3$ compared to that of $S_2$. The variances of finite population mean from $S_1$ and $S_3$ are smaller than that of $S_2$. These results show consistent pattern, reduced bias and variance regardless the method used.

The estimates of regression coefficients $\beta$ of sex and race have unreliable CVs except age although they contributed more to the BMI estimation. Since CV is not a good measurement of reliability, one needs to be careful to use CV for statistical inference.

We can extend our method to include a second probability sample which does not have the study variable. We can now use record linkage to fill in the missing weights in the nonprobability sample and the missing study variable in the probability sample. Therefore, there can be four data sets to do data integration. Our methods can be applied much the same way. Chen et al. (2020) used two data sets, non-probability sample without survey weights and probability sample without study variable. To obtain survey weights in the nonprobability sample, they used logit function. It is also possible to have a nonprobability sample without both the survey weights and the study variable. The study variable from the probability sample, if available, can be used to impute missing variables using record linkage.

It is more sensible to use the probability sample, supplemented by the nonprobability sample, to obtain the finite population mean. However, because the nonprobability sample is expected to be biased, together with its relatively large size, it will shrink the probability sample away from its expected unbiased position. Therefore, a method is much needed to partially penalize the nonprobability sample. This is one of areas in nonprobability sampling and data integration with a probability sample that is of enormous current practical interest. We have been working on this topic.

Truth is simple, but hard to find, and this is also true in statistics. This paper is trying to find a true finite population mean with nonprobability sample supplemented by a small probability sample. We hope that our efforts are a small step forward to find the truth.

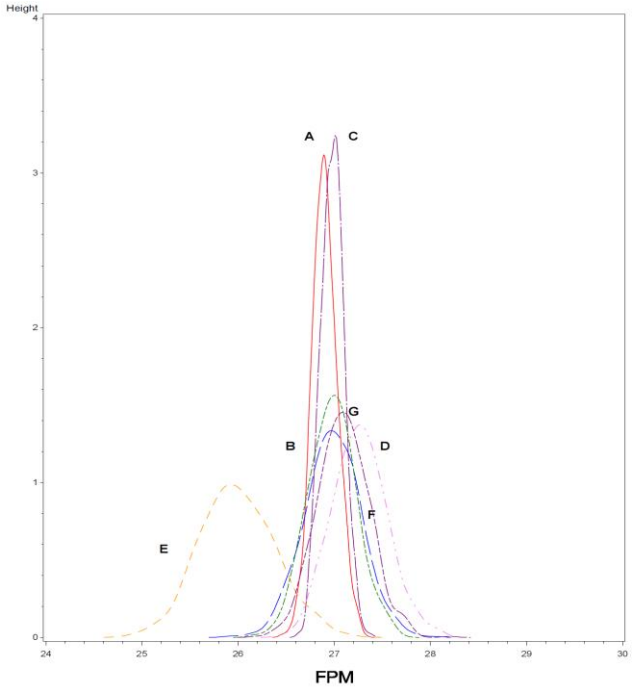


Figure 1. Comparison of bootstrap distributions (DEF), the posterior distributions (ABC) and DR (G) of the finite population mean

**Appendix A.** Imputation of OSWs for $S_1$

Data fusion (Castanedo, 2013) is widely used in practice to fill the missing data. There are many ways to perform statistical data fusion; see Kedem, Oliveira and Sverchkov (2017). Data fusion, combination or integration is very important in today's world. For example, to build self-driving cars, the more information (large N), the safer cars can be. Here one can use all sources of information for safer self-driving cars.

Mahalanobis distance (Stuart 2010) can be used to impute unknown weights of nonprobability sample with known weights of a probability sample.

We show how to impute the unknown survey weights in nonprobability sample using covariates matching or record linkage via Mahalanobis distance. We simply consider the two samples, $S_1$ with covariates, $x_{1i}$, i = 1, . . . , $n_1$ only, and $S_2$ with covariates $\boldsymbol{x_{2i}}$, i = 1, . . . , $n_2$ and survey weights, $W_{2i}$, i = 1, . . . , $n_2$, where $\sum_{i=1}^{n_2} W_{2i} = N$. Define the p ×p matrix,

$$\mathrm{S} = \frac{\sum_{i=1}^{n_2}(x_{2i}-\bar{x}_2)(\boldsymbol{x_{2i}}-\bar{\boldsymbol{x}}_2)'}{(n_2-1)},$$

where $\bar{\boldsymbol{x}}_2 = \sum_{i=1}^{n_2} \frac{x_{2i}}{n_2}$. We use the following steps.

a. Compute the Mahalanobis distances (e.g., Stuart 2010). Define $D_{ij}$,

$$D_{ij} = (\boldsymbol{x_{ji}} - x_{2j})'S^{-1}(\boldsymbol{x_{ji}} - \boldsymbol{x_{2j}}),\ \mathrm{i} = 1, \ldots, n_1,\ \mathrm{j} = 1, \ldots, n_2.$$

1. For each i, find the smallest value $D_{ij}$, j = 1,. . . ,$n_2$. Because of discreteness, it is possible that there are more than one unit with the smallest distance.
2. Randomly sample one of the units in 2. Suppose this value is k. Then, the weight assigned to unit i is $W_{2k}$, which we denote by $W_{1i}^* = W_{2k}$. Repeat this step for all units in the $S_1$.
3. Now, calibrate the $W_{1i}^*$ using a raking procedure,

$$W_{1i} = N\frac{W_{1i}^*}{\sum_{i=1}^{n_1} W_{1i}^*}, \mathrm{i}=1,\ldots,n_1.$$

These are surrogates for the original weights.

4. Finally, compute the ASWs.

**Appendix B. Least square estimation of the parameters**

Although this is not completely necessary, for convenience, we present the least square estimators. Let $\boldsymbol{x_1'} = (\boldsymbol{x_{11}}, \boldsymbol{x_{12}}, \ldots, \boldsymbol{x_{1n_1}})'$, $\boldsymbol{x_2'} = (\boldsymbol{x_{21}}, \boldsymbol{x_{22}} \ldots, \boldsymbol{x_{2n_2}})'$, each $\boldsymbol{x_k}$ is $n_k$ x p matrix, k =1, 2 and $\boldsymbol{y_1}$ = $(y_{11}, y_{12}, \ldots, y_{1n_1})'$, $\boldsymbol{y_2}$ = $(y_{21}, y_{22}, \ldots, y_{2n_2})'$, each $\boldsymbol{y_k}$ is $n_k$x1 vector, k =1, 2. Also let $\boldsymbol{\beta_p} = (\beta_1, \beta_2, \ldots, \beta_p)'$ p x1 vector of regression coefficients, and

$$\boldsymbol{w_1} = \begin{pmatrix} w_{11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & w_{1\boldsymbol{n_1}} \end{pmatrix}, \ \boldsymbol{w_2} = \begin{pmatrix} w_{21} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & w_{2\boldsymbol{n_2}} \end{pmatrix},$$

two diagonal matrices of dimensions $n_1$ and $n_2$.

Assume $y_k \sim x_k \beta_p + \varepsilon_k$, $\varepsilon_k \sim \sigma^2 W_k^{-1}$, k=1, 2 and the $y_k$ are independent. Consider the $S_1$ sample; the answers are similar for $S_2$, and combined sample $S_3$.

We have

$$y_1 = x_1\beta_p + \varepsilon, \ \ \varepsilon \sim (0, \sigma^2 W_1^{-1}).$$

Then, the least squares estimators are

$$\hat{\beta} = (x_1' w_1 x_1)^{-1}(x_1' w_1 y_1)$$

and an estimator of $\sigma^2$ is

$$\hat{\sigma}^2 = \frac{(\boldsymbol{y_1} - x_1\boldsymbol{\beta})' \boldsymbol{w_1}(\boldsymbol{y_1} - x_1\boldsymbol{\beta})}{(\boldsymbol{n_1} - \boldsymbol{p})}.$$

Similar formulas are used for the ps and the combined sample.

**Appendix C. Proof of theorem 1**

Now we drop the subscript of sample number since our discussion is applicable in general. Here lower case $w_i$ are the adjusted sample weights and upper case $W_i$ are the original weights.

Suppose the population $y_{1,} \ldots, y_N|\boldsymbol{\theta} \sim$ f($\mathbf{y}|\boldsymbol{\theta}$). Take an independent sample.$y_{1,\ldots\ldots,} y_n|\boldsymbol{\theta} \sim$ f($\mathbf{y}|\boldsymbol{\theta}$), $\boldsymbol{\theta}$ is parameter and f(.) is general function linking $\boldsymbol{\theta}$ **to y**. Then, the log-likelihood for the entire population is $\sum_{i=1}^{N} \log f(y_i|\boldsymbol{\theta})$, and the Horvitz-Thompson estimator of the log-likelihood is given by

$$\sum_{i=1}^{n} W_i \log f(y_i|\theta).$$

Exponentiating, we have pseudo-likelihood, $\prod_{i=1}^{n} f(y_i|\boldsymbol{\theta})^{W_i}$.

We make the two adjustments to this pseudo-likelihood. First, to reflect the correct variability we replace the original weights $W_i$ by the adjusted sample weights $w_i$, so we get $\prod_{i=1}^{n} f(y_i|\boldsymbol{\theta})^{w_i}$. This first step can be presented more rigorously, but this is not necessary. Second, following a full Bayesian approach, we need to normalize this density to get

$$\prod_{i=1}^{n} \frac{(f(y_i|\boldsymbol{\theta}))^{w_i}}{\int (f(y_i|\boldsymbol{\theta}))^{w_i} dy_i}.$$

It is worth noting that this pseudo density is a new formulation as it includes the normalization constant, thereby providing a proper density. This is different from what is presented in the literature, and it should make a difference when normality does not hold.

For example, suppose we have an independent sample $y_{1,\ldots,} y_n|\mu, \sigma^2$ from Normal$(\mu, \sigma^2)$ taken with unequal selection probabilities. Then the joint probability density becomes

$$\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\frac{\sigma^2}{w_i}}} e^{-\frac{w_i}{2\sigma^2}(y_i - \mu)^2}.$$

That is, the sample model with above adjusted weights $w_i$ is given by

$$y_i \sim N\left(\boldsymbol{x_i'}\boldsymbol{\beta}, \frac{\sigma^2}{w_i}\right).$$

Note that if we have used the original weights, $W_i$, the variance would have been much too small. The original survey weight for the i-th unit is how many units it represents in the entire population, including itself, but the data do not exist for all these units.

Note that normalization constant does not make a significant difference in this example with normality from the standard survey literature, but it will be important in other examples (e.g., Binomial data), and this is under investigation.

**References**

Beaumont, J. F. (2020). Are Probability Surveys Bound to Disappear for the Production of Official Statistics? *Survey Methodology, 46*(1), 1-28.

Beaumont, J. F., & Rao, J. N. K. (2020). Pitfalls of Making Inferences from Non-probability Samples: Can Data Integration through Probability Samples Provide Remedies? *New and Emerging Methods*.

Castanedo, F. (2013). A Review of Data Fusion Technologies. *The Scientific World Journal*, 704504. https://doi.org/10.1155/2013/704504

Chen, Y., L., P. &, Wu, C. (2020). Doubly Robust Inference with Nonprobability Survey Samples. *Journal of the American Statistical Association, 115*(532), 2011-2021. https://doi.org/10.1080/01621459.2019.1677241

Choi, J. W. (1977). Out of Pocket Cost and Acquisition of Prescribed Medicine. *Publication Series No.10. No. 108*, U.S. Department of Health Education and Welfare, Public Health Service. National Center for Health Statistics, Appendix: Reliability of Estimates, 33-34.

Choi, J. W., & Nandram, B. (2021). Large Sample Problems. *International Journal of Statistics and Probability, 10*(2), 81-89. https://doi.org/10.5539/ijsp.v10n2p81

Choi, J., & Nandram, B. (2000). A Measure of Concordance When There Are Many Traits. *ASA 1999 Proceedings of the Section on the Survey Research Methods*, 837-842.

Cochran, W. G. (1977). *Sampling Techniques, Third Edition*. John Wiley and Sons, New York.

Kedem, B., Oliveira, V. D., & Sverchkov, M. (2017). *Statistical Data Fusion.* World Scientific, New Jersey. https://doi.org/10.1142/10282

Little, R. J., & Rubin, D. B. (2002). *Statistical Analysis with Missing Data* (2nd ed.). New York: Wiley. https://doi.org/10.1002/9781119013563

Meng, X. (2018). Statistical Paradises and Paradoxes in Big Data (1): Law of Large Populations, Big Data Paradox, and the 2016 US Presidential Election. *The Annals of Applied Statistics, 12*(2), 685-726. https://doi.org/10.1214/18-AOAS1161SF

Nandram, B. (2007). Bayesian Predictive Inference Under Informative Sampling vi Surrogate Samples. *Bayesian Statistics and its Application* edited by S.K. Upadhyay, U. Singh, and D. K. Dey. Anamaya Publishing, New Delhi, India.

Nandram, B., & Choi, J. W. (2002a). A Hierarchical Bayesian Nonresponse Models for Binary Data from Small Areas with Uncertainty about Ignorability, *Journal of American Statistical Association, 97*(457), 381-388. https://doi.org/10.1198/016214502760046934

Nandram, B., & Choi, J. W. (2002b). A Bayesian Analysis of a Proportion under Nonignorable Nonresponse. *Statistics in Medicine, 21,* 1189-1212. https://doi.org/10.1002/sim.1100

Nandram, B., & Choi, J. W. (2006). Hierarchical Bayesian Nonignorable Nonresponse Regression Models for Small Areas: An Application to the NHANES III Data. *Survey Methodology, II* (1), 73-84.

Nandram, B., & Choi, J. W. (2010). A Bayesian Analysis of Body Mass Index Data from Small Domains under nonignorable Nonresponse and Selection. *Journal of American Statistical Association, 105*(489), 120-133. https://doi.org/10.1198/jasa.2009.ap08443

Potthoff, R. F., Max A. W., & Kenneth, G. M. (1992). Equivalent Sample Size and Equivalent Degrees of Freedom Refinements for Inference Using Survey weights Under Superpopulation Models, *Journal of American Statistical Association, 87*(418), 383-396.

Rao, J. N. K. (2020). On Making Valid Inferences by Integrating Data from Surveys and other Sources, Sankhya B: *The Indian Journal of Statistics.* Published on Line: 03 April 2020. 1-31. https://doi.org/10.1007/s13571-020-00227-w

Rubin, D. B. (1976). Inference and Missing Data. *Biometrika, 63*, 581-592. https://doi.org/10.1093/biomet/63.3.581

Sakshaug, J. W., Wisniowski, A. R., Ruiz, D. A. P., & Blom, A. G. (2019) Supplementing Small Probability Samples with Nonprobability Samples: A Bayesian Approach. *Journal of Official Statistics, 35*(3), 653-681. https://doi.org/10.2478/jos-2019-0027

Stuart, E. A. (2010). Matching Methods for Causal Inference: A Review and a Look Forward, *Statistical Science, 25*(1), 1–21. https://doi.org/10.1214/09-STS313

Wisniowski, A., Sakshaug, J. W., Ruiz, D. A. P., & Blom, A. G. (2020). Integrating Probability and Nonprobability Samples for Survey Inference, Supplementing Small Probability Samples, *Journal of Survey Statistics and Methodology, 8,* 120-147. https://doi.org/10.1093/jssam/smz051