A comparsion of methods for combining surveys

Yusa Lin

M.Sc., The University of Auckland

A thesis submitted for the degree of Master of Science at

the University of Auckland in 2022

Department of Statistics

Contents

Copyright notice **v**

Abstract **vii**

Acknowledgements **ix**

1 Introduction **1**

2 Background **3**

2.1 Basic concepts of sampling design 4

2.2 Inclusion probabilities 7

3 Literature Review on combining surveys **9**

3.1 Directed approach to combine multiple surveys 9

3.2 Weight adjustment approach to combine multiple surveys 10

3.3 Calibration weighting methods in surveys 14

4 Methodology **17**

4.1 Simulation: Generating repeated observations between waves 18

4.2 Simulation: Generating variables 21

4.3 Simulation: Selecting samples 24

4.4 Simulation: Combining populations of waves 25

4.5 Simulation: Estimator 27

5 Results **29**

5.1 Estimations for combined waves without repeated observations 30

5.2 Estimations for combined waves with repeated observations included . . . 36

5.3 Estimations for combined waves with calibration (without repeated observations) 39

5.4 Estimations for combined waves with calibration (with repeated observations) 43

5.5 National Health and Nutrition Examination Survey (NHANES) Example . . . 47

6 Conclusion **49**

A Additional results from simulation **51**

A.1 Estimations for combined waves without repeated observations 51

A.2 Estimations for combined waves with repeated observations 57

A.3 Estimations for combined waves with calibration (without repeated observations) . . . 60
A.4 Estimations for combined waves with calibration (with repeated observations) 63

Bibliography **67**

Copyright notice

Abstract

As more and more researchers start to study more complicated survey problems, a single survey might not be sufficient to meet the analytical needs. Therefore, combining multiple surveys to get larger and more diverse samples is useful in some situations. This thesis focus on comparing different methods of combining multiple surveys. A large simulation study where we investigate the properties of different methods for combining surveys is presented in this thesis. In the simulation, we create a series of populations that contain geographic information so the stratified two-stage cluster sampling method can be applied when selecting samples. Then, two methods (weight-adjust method and calibration weighting methods) of combining surveys are applied. The analytical results of simulated data show that the weight-adjust method can reduce the estimation bias (i.e. underestimating or overestimating) if overlapped observations can be identified and need to be removed in order to meet study needs when combining surveys. However, if the variables we are interested in change over time, the estimates produced by the weight-adjust method may contain minor biases. Therefore, using the calibration method when auxiliary variables are available is recommended since it will help reduce the standard error and produce more accurate estimates. Especially when using re-calculated weights to estimate the variables that change over time, using the calibration method can reduce the bias.

Acknowledgements

I would like to thank my supervisor, Dr. Claudia Liliana Rivera Rodriguez, for providing me guidelines and feedback throughout my thesis. I also want to thank my girlfriend, Coco, for staying with me until late night and encouraging me.

Chapter 1

Introduction

Nowadays, companies use surveys to estimate customer satisfaction; biologists use surveys to estimate animal abundance; doctors use surveys to do clinical research, etc. As our society entered the information era, the rapid development of computer technology has given us unprecedented computing power. Survey research started to become much easier than before. More surveys could be conducted in order to investigate more complicated questions that researchers are interested in, for e.g. The National Health Interview Survey (NHIS)(Health Statistics (NCHS), 2020), California Health Interview Survey (CHIS)(Research, 2012), or The National Health and Nutrition Examination Survey (NHANES)(Health Statistics (NCHS), 2017). Then, some interesting questions were raised: what will happen if we combine multiple waves of a period or repeated cross-sectional survey? What are the differences in estimates if repeated observations are known between different waves of period surveys compared to repeated observations that are unknown? In this study, a simulation method will be applied in order to get an insight into these questions. The simulation procedure will be explained in detail in Chapter 4. The simulation results will be discussed in Chapter 5. To be specific, Section 5.1 in Chapter 5 will discuss the simulation results that if repeated observations are known and removed from populations and samples. A comparison will be performed to the estimates of the populations from the samples with re-calculated inclusion probabilities to the samples with original inclusion probabilities. Section 5.2 in Chapter 5 will discuss the simulation results that if repeated observations are unknown and remain in the samples and populations. Section 5.3 and Section 5.4 in Chapter 5 will discuss the simulation results with auxiliary variables included in the dataset to see if the estimates can be improved. Section 5.5 in Chapter 5 will discuss a short example by using NHANES data.

Chapter 2

Background

Surveys are a great and affordable way to help researchers or governments get a general idea of a particular population. For example, governments might want to know the number of people with breast cancer in a country, or social scientists might want to know the unemployment rate for a specific area, etc.

Ornstein (2013) introduced the development of survey research history. Survey research has a very long history in the statistical world. It can be traced to the nineteenth century at the earliest. After taking centuries of census, a *social survey*, which focused on poor people who lived in urban was promoted by English researchers Charles Booth and Joseph Rowntree in the 1880s and 1890s; many other countries started to do similar things in the first third of the twentieth centuries. The real *modern* survey was developed gradually in the mid-1930s by Archibald Crossley, Elmo Roper, and George Gallup, who are American market researchers. They successfully predicted the results of the U.S. presidential election in 1936. The first description of longitudinal survey was established in the second volume of Public Opinion Quarterly in 1937 by Lazarsfeld and Fiske. During World War Two, more than half-million U.S. soldiers were surveyed; they were asked to answer over two hundred questionnaires. The research was led by Samuel Stouffer, who was a sociologist at Harvard University.

2.1 Basic concepts of sampling design

All surveys require a sampling design. Sampling design can be generally divided into two types: Non-probability sampling and Probability sampling.

Non-probability sampling: Non-probability sampling is a group of sampling methods in which samples are selected based on subjective judgments instead of using probabilities. Types of non-probability sampling include purposive sampling, convenience sampling, volunteer sampling, mail-in surveys, tele-voting (or SMS voting), self-selection in web surveys, and network sampling. These designs are explained well by Vehovar, Toepoel, and Steinmetz (2016):

- Purposive sampling: Also called judgmental sampling. Researchers need to select samples based on their experience and judgments. Sometimes samples are selected sequentially until they satisfy researchers.

- Convenience sampling: Convenience sampling is easy to conduct. It simply includes the most accessible samples to the researcher. For example, a researcher may ask his/her colleagues to be the respondents for a survey. However, selecting samples in this way may cause serious bias since there is no way to tell if the samples can represent the population.

- Volunteer sampling: Voluntary response sampling is somewhat *random*. Instead of letting researchers choose respondents, people need to volunteer themselves. Selecting samples in this way may also cause bias since some people may be more likely to volunteer themselves than others.

- Mail-in surveys: Mail-in survey is a type of volunteer sampling. Researchers might hand out leaflets that contain questionnaires and seek people to answer at public locations such as parks, hotels, restaurants, etc.

- Tele-voting (or SMS voting): A type of volunteer sampling. People are invited to participate by calling some toll-free numbers or sending text messages. This sampling method is often used in TV or radio programs.

- Self-selection in web surveys: A type of volunteer sampling. The invitation is posted on the internet and people can participate by clicking the link.
- Network sampling: Also called snowball sampling. It is often used when the population is hard to access. Researchers select a small group of respondents and ask them to recruit more participants in their own social networks.

Probability sampling: Unlike non-probability sampling methods, probability sampling methods require samples to be selected such that each observation has a known and non-zero probability to be selected. Types of probability sampling include simple random sampling, systematic sampling, stratified sampling and cluster sampling. These sampling designs are explained clearly by Fuller (2009):

- Simple random sampling: Every member has the same probability of being selected in the population. Simple random sampling can be replacement or without replacement.
- Systematic sampling: A similar method to simple random sampling, but instead of randomly selecting samples purely from a population, systematic sampling selects samples at regular intervals. For example, a population contains 1000 members in total, and members are labeled as 1,2, . . . , 1000. After randomly selecting a starting point, say 10, every 10th member after the starting point will be selected. Therefore, our sample will be 10, 20, 30, 40, and 50 if we select five members.
- Stratified sampling: In practice, researchers divide the population into different multiple exclusive sub-populations called strata based on specific characteristics such as location, gender, species, etc., and every member of the population should not be included in more than one sub-populations. Then a researcher may apply other sampling methods such as simple random sampling to each sub-population. Stratified sampling is usually used when the population has diverse characteristics and can be represented by different groups clearly.
- Cluster sampling: It is a sampling method that divides a large population into smaller groups named clusters. Researchers may use some pre-existing units such as blocks,

schools, cities as clusters. Then other sampling methods such as simple random sampling can be applied to a cluster population.

In practice, researchers prefer to use more complicated sampling methods when the population is very large and diverse. This generally reduces the cost and increases the efficiency for sample collection. Two popular complex sampling methods are widely used in survey studies: multi-stage sampling and multi-phase sampling. These sampling designs are also explained by Fuller (2009):

- Multi-stage sampling: It can be considered as a more complex form of cluster sampling. In multi-stage sampling, larger clusters can be divided into smaller clusters to meet researchers' needs. It is very useful for reducing costs and selecting more representative samples from the population while conducting large-scale surveys. For example, suppose the New Zealand government wants to investigate the national unemployment rate. In that case, they might choose blocks as the first-stage clusters (or primary sampling units), then households inside blocks as the second stage-units. This method is called *Two-stage cluster sampling*.
- Multi-phase sampling: It is a widely used sampling design for certain survey studies. Researchers collect information from a large sample of certain units, then collect other information in some sub-samples of the whole sample later or at the same time. This is also called *Two-phase Sampling* or *Double Sampling*. For example, suppose we want to investigate the total expenditure of customers in a region, the only available information we have is a list of all households. Then we may choose a large size of the preliminary sample to collect some valuable characteristics information such as occupational status, household size at the first phase. And at the second phase, we may select a small size of the sample to meet the study needs using info collected at the first phase. Multi-phase sampling may be confused with multistage sampling. The main difference is that, when using multistage sampling, we assume the design is independent between primary sampling units (PSUs); the sub-sampling U_k is also independent of sub-sampling U_i. We also have invariance, which means every time we select a PSU, k, the same design $P_k(.)$ will be applied. However, multi-phase

sampling does not have to follow these assumptions. Multi-phase sampling has another advantage: it can handle non-response since respondents are usually viewed at the second phase.

2.2 Inclusion probabilities

This thesis focuses on probability sampling. As Särndal, Swensson, and Wretman (1992) explained in his book, each observation has a known and non-zero probability to be included in a sample for probability sampling. The probability for a sample s can be denoted as $p(s)$, which means the probability of a sample s is selected. Therefore, the probability that a observation i is selected in a sample s can be computed if the sample design is known. This is called *inclusion probability*, denoted as π_i that is $\pi_i = P(i \in s) = \sum_{s \ni i} p(s)$, where $i \in s$ represents the samples that contain i. For example, in simple random sampling without replacement, we have $p(s) = \frac{1}{\begin{pmatrix} N \\ n \end{pmatrix}}$.

Therefore, the inclusion probability will be (Särndal, Swensson, and Wretman, 1992)

$$\pi_i = \sum_{s \ni i} p(s) = \sum_{s \ni i} \frac{1}{\begin{pmatrix} N \\ n \end{pmatrix}} = \frac{\begin{pmatrix} N-1 \\ n-1 \end{pmatrix}}{\begin{pmatrix} N \\ n \end{pmatrix}} = \frac{n}{N}$$

Similarly, the inclusion probability for both i and j are selected in the sample s can be written as $\pi_{ij} = P(i\&j \in S) = \sum_{s \ni i\&j} p(s)$ and this is also called *second order inclusion probability* or *pairwise probabilities*. For example, the second order inclusion probability for simple random sampling without replacement is (Särndal, Swensson, and Wretman, 1992)

$$\pi_{ij} = \sum_{s \ni i\&j} p(s) = \sum_{s \ni i\&j} \frac{1}{\begin{pmatrix} N \\ n \end{pmatrix}} = \frac{\begin{pmatrix} N-2 \\ n-2 \end{pmatrix}}{\begin{pmatrix} N \\ n \end{pmatrix}} = \frac{n(n-1)}{N(N-1)}$$

Chapter 3

Literature Review on combining surveys

Combining multiple surveys to obtain a larger sample is useful and necessary under certain situations, especially when the most recent survey is not sufficient to meet the researchers' analytical needs (Thomas and Wannell, 2009). In this chapter, several articles are reviewed. Although there is a wide range of theories and ideas are covered in the literature, only two main themes will be the focus of this thesis. Specifically the methods for combining multiple surveys (directed approach, weight adjustment approach), and calibration weighting methods in surveys.

3.1 Directed approach to combine multiple surveys

As Thomas and Wannell (2009) mentioned, the target population for a combined survey is no longer the same as the target population of a single survey. Instead, the combined survey is representative of the combined population, and the combined population may overlap. Two methods of combining surveys were introduced in the article: the separate approach and the pooled approach. The separate approach is straightforward, we only need to compute the average of estimates calculated from individual surveys. For example, suppose there are k surveys in total that need to be combined. A simple average can be calculated as $\hat{\theta}_c^{avg} = \frac{\sum_{i=1}^{k}\hat{\theta}_i}{k}$, where $\hat{\theta}_i$ is the estimate of i^{th} survey. The variance can be estimated as

$$\hat{V}\left(\hat{\theta}_c^{avg}\right) = \hat{V}\left(\frac{\hat{\theta}_1 + \hat{\theta}_2 + ... + \hat{\theta}_k}{k}\right) = \frac{1}{k^2}\left[\hat{V}\left(\hat{\theta}_1\right) + \hat{V}\left(\hat{\theta}_2\right) + ... + \hat{V}\left(\hat{\theta}_k\right)\right]$$

if we assume the surveys are independent. The pooled approach is also relatively easy to proceed with. It consists of combining multiple surveys to create a large dataset that can be analyzed as a single sample from a pooled population. However, the pooled approach may not be appropriate for estimating totals since it will overestimate the total if we sum the sample weights from multiple surveys. One way to solve this is to rescale the sampling weights by a factor to represent the population of interest.

There is another way to combine two survey sample frames if we assume that the response variable is measured identically in both sampling frames (Elliott, Raghunathan, and Schenker, 2018). Suppose that the numbers of the elements in frame A only, frame B only are N_a, N_b, and N_{ab} are the overlap of the two frames and identifiable, the estimator of a population total Y is given by

$$\hat{Y}(p) = N_a\bar{y}_a^A + N_b\bar{y}_b^B + N_{ab}\left(p\bar{y}_{ab}^A + (1-p)\bar{y}_{ab}^B\right) = \hat{Y}_a^A + \hat{Y}_b^B + \hat{Y}_{ab}(p)$$

where $\hat{Y}_a^A = N_a\bar{y}_a^A, \hat{Y}_b^B = N_b\bar{y}_b^B$, and $\hat{Y}_{ab}(p) = p\hat{Y}_{ab}^A + (1-p)\hat{Y}_{ab}^B$. Under simple random sampling, the variance of $\hat{Y}(p)$ is minimized when $p = \frac{f_A}{f_A+f_B}$ where $f_A = n_A/N_A$ for $n_A = n_a + n_{ab}$ and $N_A = N_a + N_{ab}$, and similar for f_B if n_A and n_B are fixed.

3.2 Weight adjustment approach to combine multiple surveys

Sometimes we need to adjust the inclusion probabilities (or weights) in order to avoid bias when sample frames are overlapped. It is because that the overlapping part will cause overestimating. Suppose there are two sampling frames, which need to be combined and the unites that appear in both sampling frames (overlapped units) are identifiable. Arcos et al. (2015) proposed to adjust the weights by

$$\tilde{d}_k = \begin{cases} d_k^A & \text{if } k \in a \\ \left(1/d_k^A + 1/d_k^B\right)^{-1} & \text{if } k \in ab \\ d_k^B & \text{if } k \in b \end{cases}$$

Where $\tilde{d}_k$ is the adjusted sampling weight, d_k^A is the original sampling weights in frame A and d_k^B is the original sampling weights in frame B.

Hence, the estimator of combined population total is given as

$$\hat{Y}_{BKA} = \sum_{k \in s_A} \tilde{d}_k^A y_k + \sum_{k \in s_B} \tilde{d}_k^B y_k = \sum_{k \in s} \tilde{d}_k y_k$$

with $s = s_A \cup s_B$.

Lohr (2011) introduced a straightforward way to adjust the weights. If two sample frames are combined and repeated observations are known, the estimator of population total is given by $\hat{Y} = \sum_{i \in \mathcal{S}(A)} \tilde{w}_i^A y_i + \sum_{i \in \mathcal{S}(B)} \tilde{w}_i^B y_i$, where $\tilde{w}_i^A = m_i^A w_i^A$ and $\tilde{w}_i^B = m_i^B w_i^B$ are adjusted weight. Based on the method introduced by Hartley (1962), let $\theta \in [0, 1]$, we have

$$m_{i,\theta}^A = \begin{cases} 1 & \text{if } i \in a \\ \theta & \text{if } i \in ab, \end{cases} \qquad m_{i,\theta}^B = \begin{cases} 1 & \text{if } i \in b \\ 1-\theta & \text{if } i \in ab \end{cases}$$

Thus the previous estimator, $\hat{Y}$, can be expressed as

$$\begin{aligned} \hat{Y}(\theta) &= \sum_{i \in \mathcal{S}(A)} m_{i,\theta}^A w_i^A y_i + \sum_{i \in \mathcal{S}(B)} m_{i,\theta}^B w_i^B y_i \\ &= \hat{Y}_a^A + \theta \hat{Y}_{ab}^A + (1-\theta)\hat{Y}_{ab}^B + \hat{Y}_b^B \end{aligned}$$

where $\theta \in [0, 1]$, $\hat{Y}_a^A = \sum_{i \in \mathcal{S}(A), i \in a} w_i^A y_i$, $\hat{Y}_{ab}^A = \sum_{i \in \mathcal{S}(A), i \in ab} w_i^A y_i$, $\hat{Y}_{ab}^B = \sum_{i \in \mathcal{S}(B), i \in ab} w_i^B y_i$, and $\hat{Y}_b^B = \sum_{i \in \mathcal{S}(B), i \in b} w_i^B y_i$.

A Pseudo-maximum likelihood (PML) estimator can be conducted for probability samples in order to achieve internal consistency. Skinner and Rao (1996) suggested that same weights could be used for all variables in combined sample frames. Let $a = A \cap B^c$,

$b = A^c \cap B$, $ab = A \cap B$ and let N, N_A, N_B, N_a, N_b, N_{ab} denote the size of the sets U, A, B, a, b, ab. Then $N = N_a + N_b + N_{ab}$, $N_A = N_a + N_{ab}$, $N_B = N_b + N_{ab}$, and N_{ab} can be estimated by $\hat{N}_{ab}^{PML}(\theta)$ if it is unknown. That is, $\hat{N}_{ab}^{PML}$ is the smaller root of the equation

$$\left[\frac{\theta}{N_B} + \frac{1-\theta}{N_A}\right] x^2 - \left[1 + \theta\frac{\hat{N}_{ab}^A}{N_B} + (1-\theta)\frac{\hat{N}_{ab}^B}{N_A}\right] x + \theta\hat{N}_{ab}^A + (1-\theta)\hat{N}_{ab}^B = 0.$$

By using the value θ_p for θ, the asymptotic variance of $\hat{N}_{ab}^{PML}(\theta)$ can be minimized, where

$$\theta_P = \frac{N_a N_B V\left(\hat{N}_{ab}^B\right)}{N_a N_B V\left(\hat{N}_{ab}^B\right) + N_b N_A V\left(\hat{N}_{ab}^A\right)}.$$

After substituting θ_p with $\hat{\theta}_p$, the adjusted weight will be

$$m_{i,P}^A = \begin{cases} \frac{N_A - \hat{N}_{ab}^{\text{PML}}\left(\hat{\theta}_P\right)}{\hat{N}_a^A} & \text{if } i \in a \\ \frac{\hat{N}_{ab}^{\text{PML}}\left(\hat{\theta}_P\right)}{\hat{\theta}_P\hat{N}_{ab}^A + \left(1-\hat{\theta}_P\right)\hat{N}_{ab}^B}\hat{\theta}_P & \text{if } i \in ab \end{cases}, \quad m_{i,P}^B = \begin{cases} \frac{N_B - \hat{N}_{ab}^{\text{PML}}\left(\hat{\theta}_P\right)}{\hat{N}_b^B} & \text{if } i \in b \\ \frac{\hat{N}_{ab}^{\text{PM}}\left(\hat{\theta}_P\right)}{\hat{\theta}_P\hat{N}_{ab}^A + \left(1-\hat{\theta}_P\right)\hat{N}_{ab}^B}\left(1-\hat{\theta}_P\right) & \text{if } i \in ab \end{cases}$$

If two sample frames from two different surveys do not have the same quality, i.e. sample frame a has higher response rate or lower measurement error than sample frame b. The adjustment proceeds as follows. In order to reduce bias in b, the adjusted weight for b, as Elliott, Raghunathan, and Schenker (2018) mentioned, can be calculated as

$$w_i^b = d_i^b \frac{\hat{P}\left(S_i = a \mid y_i, \mathbf{x}_i\right) / \hat{P}\left(S_i = a\right)}{\hat{P}\left(S_i = b \mid y_i, \mathbf{x}_i\right) / \hat{P}\left(S_i = b\right)} = d_i^b \frac{\hat{P}\left(S_i = a \mid y_i, \mathbf{x}_i\right)}{\hat{P}\left(S_i = b \mid y_i, \mathbf{x}_i\right)\right)} \frac{\hat{P}\left(S_i = b\right)}{\hat{P}\left(S_i = a\right)}$$

where $S_i = s, s \in \{a, b\}$ is a is an indicator for survey membership, d_i^s is the design weight associated with element i in survey s, y_i is the outcome of interest, and $\mathbf{X}_i$ is a vector of covariates.

In practice, researchers may need to combine more than two sample frames to meet their analytical needs. Under this scenario, a generalized method need to be performed for adjusting weights. As Lohr (2011) mentioned, suppose there are K frames, the probability sample from A_k can be denoted as $S(A_k)$ where $k = 1, 2, ..., K$. The inclusion probability and weight of unit i in sample $S(A_k)$ is $\pi_i^{A_k}$ and $w_i^{A_k}$. Therefore, then generalized multiple frame estimator will be

$$\hat{Y} = \sum_{k=1}^{K} \sum_{i \in S(A_k)} m_i^{A_k} w_i^{A_k} y_i$$

where the weight adjustment is $m_i^{A_k}$ for observation i in $S(A_k)$. If there are D distinct domains in total, the weight adjustment for each frame and domain can be set as $m^{A_k,d}$. Then we have $m_i^{A_k} = m^{(A_k,d)}$ if observation i from $S(A_k)$ is in domain d with assuming that $m^{A_k,d} = 0$ if domain $d \notin A_k$ and $\sum_{k=1}^{K} m^{(A_k,d)} = 1$ for $d = 1, 2, ..., D$. And $m^{(A_k,d)}$ can be calculated as $\frac{1}{Number\ of\ frames\ that\ contain\ domain\ d}$ and Mecatti (2007) called it the multiplicity estimator.

Dong, Elliott, and Raghunathan (2014b) introduced a newly developed method for combining information from multiple surveys with a Bayesian approach involved. The idea is to generate a synthetic population by using Bayesian Bootstrap and finite population Bayesian bootstrap from single survey with complex sampling design (Dong, Elliott, and Raghunathan, 2014a). The first step is adjusting for stratification and clustering by using Bayesian Bootstrap (Rubin, 1981) and select a sample of size m_h from the c_h clusters with each stratum $h = 1, ..., H$ by using simple random sampling with replacement. The bootstrap replicate weights is calculated for each of the n_{hi} observations as $w^{*(l)} = \left\{ w_{hi}^{*(l)}, h = 1, \ldots, H, i = 1, \ldots, c_h, k = 1, \ldots, n_{hi} \right\}$ in each cluster, where $w_{hik}^{*} = w_{hik} \left(\left(1 - \sqrt{(m_h/c_h - 1)} \right) + \sqrt{(m_h/c_h - 1)} \, (c_h/m_h) \, m_{hi}^{*} \right)$ and m_{hi}^{*} denotes the number of times that cluster i is selected. Setting $m_h \leq (c_h - 1)$ to prevent negative weights.

The second step is adjusting for unequal probabilities of selection by using the finite population Bayesian bootstrap (Cohen, 1997; Lo, 1986). Selecting a sample with size of $N_{hi} - n_{hi}$ for each cluster i in stratum h of population size N_{hi}, denoted by $\left(y_1^{*}, \ldots, y_{N_{hi}-n_{hi}}^{*} \right)$. By selecting y_{hik}^{*} from cluster data $(y_1, \ldots, y_{n_{hi}})$ with probability $\frac{w_{hik}^{*} - 1 + l_{hik,j-1} * (N_{hi} - n_{hi})/n_{hi}}{N_{c_H} - n_{c_H} + (j-1) * (N_{hi} - n_{hi})/n_{hi}}$, where the replicate weight of unit k is w_{hik}^{*} in cluster i in stratum h, and the number of bootstrap selections of y_{hik} is $l_{hik,j-1}$ among $y_1^{*}, \ldots, y_{j-1}^{*}$.

The third step is producing finite population Bayesian bootstrap samples for each Bayesian Bootstrap sample, denoted by $S_{l1}, \ldots, S_{lF}, l = 1, \ldots, L$. A synthetic population S_l can be produced by pooling the finite population Bayesian bootstrap samples. Suppose L

synthetic populations, $S_l^{(s)}$ (where $l = 1, ..., L$), are created by using data from a single survey, s. Let $Q_l^{(s)}$ be the estimate of the population quantity Q obtained from synthetic population l, then as Dong, Elliott, and Raghunathan (2014a) showed, under reasonable asymptotic assumptions

$$Q \mid S_1^{(s)}, \ldots, S_L^{(s)} \dot{\sim} t_{L-1} \left(\bar{Q}_L^{(s)}, \left(1 + L^{-1}\right) B_L^{(s)} \right)$$

where Q is the population quantity of interest (such as population total, population mean, etc.) depending the variables Y, which collect from multiple surveys, $\bar{Q}_L^{(s)} = L^{-1} \sum_{l=1}^{L} Q_l^{(s)}$ is the mean of Q for the synthetic populations, and $B_L^{(s)} = (L-1)^{-1} \sum_{l=1}^{L} \left(Q_l^{(s)} - \bar{Q}_L^{(s)} \right)^2$ is the between-imputation variance. When L is large, then

$$Q \mid S_{syn}^{(1)}, \ldots, S_{syn}^{(S)} \dot{\sim} N \left(\bar{Q}_L, B_L \right)$$

where $\bar{Q}_L^{(s)}$ and $B_L^{(s)}$ are combined estimator of the population (such as population total, population mean, etc.) and it's variance, and $\bar{Q}_L = \sum_{s=1}^{S} \left(\bar{Q}_L^{(s)} / B_L^{(s)} \right) / \sum_{s=1}^{S} \left(1/B_L^{(s)} \right)$ and $B_L = 1/ \sum_{s=1}^{S} \left(1/B_L^{(s)} \right)$. When L is not sufficiently large,

$$Q \mid S_{syn}^{(1)}, \ldots, S_{syn}^{(S)} \dot{\sim} t_{v_L} \left(\bar{Q}_L, \left(1 + L^{-1}\right) B_L \right)$$

$\vartheta_L = (L-1)/ \sum_{s=1}^{S} \left(\left(1/b_L^{(s)} \right) / \sum_{s=1}^{S} \left(1/b_L^{(s)} \right) \right)^2$ is the degree of freedom.

3.3 Calibration weighting methods in surveys

Calibration is a method that has been widely applied to estimation on survey samples. The basic idea of the calibration approach is adjust the weights using auxiliary variables, such that calibration constraints are satisfied where the calibration constraint is $\sum_s w_k \mathbf{x}_k = \sum_U \mathbf{x}_k$ (Särndal, 2007). Since generalized regression (GREG) can also take auxiliary information into account, some GREG estimators are also be considered as calibration estimators as long as they can be expressed in terms of a calibrated linear weighting. Calibration estimators do not need to rely on any explicit model. Instead, the key element of the calibration approach is to emphasize the linear weighting of the observed y-values, with weights made to confirm computable aggregates. The calibration approach is robust, it can deal with a wide range of scenarios such as complex sampling design, adjustment for frame error and non-response error.

As Särndal (2007) mentioned, the calibration approach contains many different methods. For example, the minimum distance method is to modify the initial weight $d_k = 1/\pi_k$ into new weight w_k, determined to be close to the d_k. The instrument vector method is an alternative method to minimize the distance. A lot of alternative sets of weights calibrated to the same information can be generated by using this method. The calibration estimation can also be used for more complex parameters such as population quantiles and functions of population totals.

Ranalli et al. (2016) introduced a calibration estimation method for dual-frame designs. Assuming the population total $\boldsymbol{t}_x = \sum_{k=1}^{N} \boldsymbol{x}_k$ is known, where $x_k = (x_{1k}, \ldots, x_{pk})$ is the value taken on unit k by a vector of auxiliary variables x.The vector of totals may pertain to A only, B only, U (entire population) or to a combination of those three sets. The basic weights d_k° is modified to obtain new weights w_k° by using the calibration paradigm. Then, a new calibration estimator of the total Y can be defined as

$$\hat{Y}_{\text{CAL}} = \sum_{k \in s} w_k^\circ y_k$$

where $k \in s$ is to account for auxiliary information, and w_k° is such that

$$\min \sum_{k \in s} q_k G\left(w_k^\circ, d_k^\circ\right) \quad s.t. \quad \sum_{k \in s} w_k^\circ \boldsymbol{x}_k = \boldsymbol{t}_x$$

and $G(w, d)$ is a distance measure satisfying the calibration paradigm conditions, and the arbitrary constant q_k may or may not be a function of x_k. An appropriate choice of q_k may help us reproduce ratio estimator or pseudo-optimal calibration estimator. Different distance measures can be used to obtain different calibration estimators given the set of constraints such as Euclidean distance.

Chapter 4

Methodology

This chapter presents a large simulation study where we investigate the properties of different methods for combining waves of repeated surveys. The main goal of the simulation is to create a series of populations that contain geographic information so a stratified two-stage cluster sampling method can be applied when selecting samples. There are four populations (or waves) created with 100,000 observations each, and all populations are created under the following assumptions or rules.

1. The populations (waves) contain all the people between 20 and 30 years old in a fictitious country for specific years.
2. Each population represents a different year. For example, if population 1 (or wave 1) covers all the people between 20 and 30 years old in the year of 2000, then population 2 (or wave 2) can only cover all the people between 20 and 30 years old in a given year after 2000. The same rule applies to other two waves. The populations of waves will have overlapping if the gap between two waves is less then or equal to ten years.
3. Nobody moves out of the country or accidentally dies between 20 and 30 years old.
4. Nobody moves into the country between 20 and 30 years old.
5. The fertility and mortality of the country remain stable over time. That means *baby boom* will never happen in this country. Therefore, the total population of the country also remain stable over time.

4.1 Simulation: Generating repeated observations between waves

In practice, repeated observations might occur when we combine repeated surveys. For example, a household might be a part of the target population for multiple waves of a repeated survey, say National Health Interview Survey (NHIS). That is, a household might present more than once after multiple surveys are combined. Therefore, we need to consider this situation in our simulating process. Repeated observations between different waves will be generated under the assumptions mentioned previously. Suppose we want wave 1 and wave 2 to have repeated observations in population between each other, the year difference between wave 1 and wave 2 should not be longer than 10 years. For example, suppose wave 1 covers all the people between 20 and 30 years old in the year of 2010 and wave 2 covers all the people between 20 and 30 years old in the year of 2020; the year difference is 10 years between wave 1 and wave 2. That means all the people who are 20 years old in wave 1 will be those who are 30 years old in wave 2. That is, we have about 9.1% overlapping in the population between wave 1 and wave 2 if each age group has the same number of people as Figure 4.1 shows.

Similarly, if we want three waves to have repeated observations, or if we want some observations to show up in all three different waves, the same idea can be applied. For example, suppose wave 1 covers all the people between 20 and 30 years old in the year of 2010, wave 2 covers all the people between 20 and 30 years old in the year of 2015, and wave 3 covers all the people between 20 and 30 years old in the year of 2020. Then, the year difference is five years between every two consecutive waves. That means all the people who are 20 to 25 years old in wave 1 will be those who are 25 to 30 years old in wave 2. Likewise, all the people who are 20 to 25 years old in wave 2 will be those who are 25 to 30 years old in wave 3. Therefore, people who are 20 years old in wave 1 will be 25 years old in wave 2 and then become 30 years old in wave 3 as Figure 4.2 shows. That is, we have about 54.5% overlapping in the population between every two consecutive waves and 9.1% overlapping among all three waves.

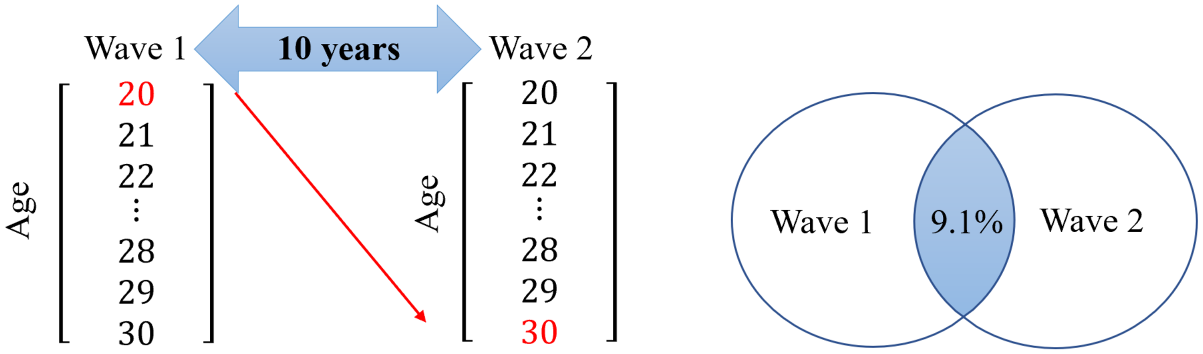

Figure 4.1: *Two waves with 9.1% overlapping.*

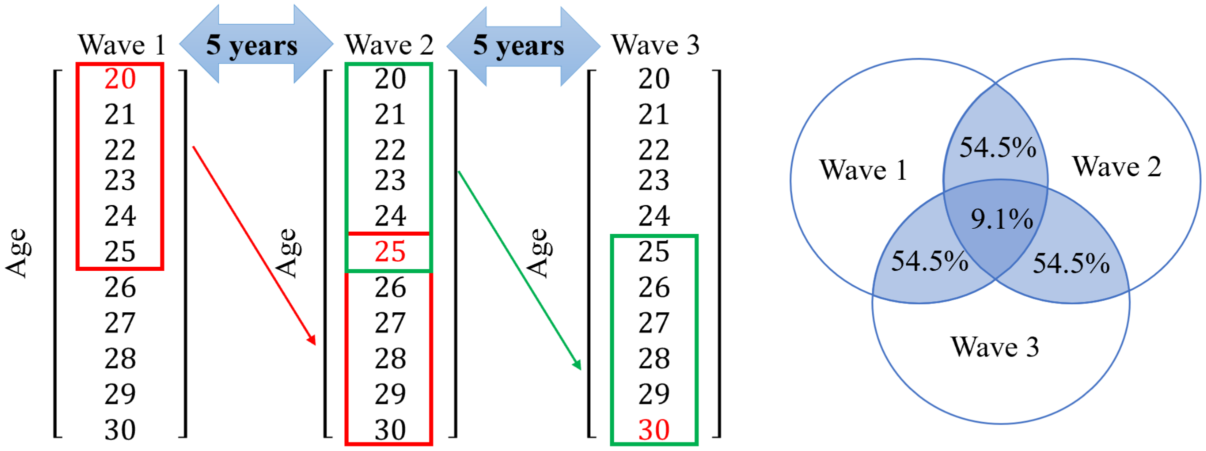

Figure 4.2: *9.1% overlapping among three waves.*

Figure 4.3: *72.7% overlapping among four waves.*

We can also apply the same idea to four waves. Suppose we want to generate repeated observations between four waves, and the year difference is 1 year between every two consecutive waves. That means the people between 20 and 29 years old in wave 1, 2 and 3 will also show up in the next wave. Similarly, the people between 20 and 28 years in wave 1 and 2 will also show up in the subsequent two waves, and the people between 20 and 27 years old in wave 1 will show up in all four waves. Therefore, there will be about 91% overlapping between every two consecutive waves, 81.8% overlapping between every three successive waves, and 72.7% overlapping in the population among all four waves, as Figure 4.3 shows.

We can change the year difference between every two consecutive waves to adjust the overlapping percentage in order to meet the study needs. The same idea can be applied to five or more waves, but we only focus on at most four waves in this study.

4.2 Simulation: Generating variables

There are 16 variables are generated initially in each wave in order to meet the study needs such as creating a larger and more diverse dataset to conduct a comprehensive analysis. Variable details can be found below:

- **Regions** (Character): The purpose of generating the Regions variable is for applying stratified sampling when selecting samples. There are eight different regions generated in each wave, and each region has a unique size. The sizes are Region 1 (Reg1): 12%, Region 2 (Reg2): 15%, Region 3 (Reg3): 21%, Region 4 (Reg4): 28%, Region 5 (Reg5): 5%, Region 6 (Reg6): 10%, Region 7 (Reg7): 8%, and Region 8 (Reg8): 1%. That is, there are 12% of people live in Region 1 in each wave, 15% of people live in Region 2, and so on.
- **Cities** (Character): There are 100 cities in total for each wave. The number of cities in each region is proportional to the region size. For example, the size for Region 1 is 12%. Therefore there are 12 cities located in Region 1.
- **City ID** (Character): The unique ID for cities in each region in each wave. For example, ID for City 1 in Region 1 will be Reg1City1, and for City 1 in Region 2 will be Reg2City1.
- **Resident Number** (Integer): The variable of Resident_num is generated on a random basis. There are 100000 residents are randomly generated in the first wave. Each city in each region contains about 1000 residents. Therefore, Region 1 contains about 12000 residents. The number of residents that be generated in the rest waves depend on the year gaps between waves. For example, if two waves are generated and the year gap is ten years, all 100000 residents are randomly generated in the first wave. Then, 90900 residents are randomly generated in the second wave for residents are between 20 and 29 years old. The remaining 9100 residents with age of 30 in wave 2 come from the residents with age of 20 wave 1. Therefore, the total population in wave 2 is also 100000 and there are 9100 residents are contained in both wave 1 and wave 2. In other words, we have 9.1% overlapping between wave 1 and wave 2.

- **Resident ID** (Character): The unique ID for residents in each city in each region. For example, Resident 1 in City 1 Region 1 is Reg1City1_1, and Resident 1 in City 1 Region 2 is Reg2City1_1.

- **Unique ID** (Character): The unique ID for each resident for all waves. This variable is used to identify repeated observations between different waves. For example, Unique ID: Wave_1Reg1city12_1 means Resident 1 in City 12 Region 1 in Wave 1.

- **Age** (Integer): The age of residents. All residents are between 20 and 30 years old in all waves. The variable Age is generated from a uniform distribution. All age group contains the same number of residents. That is, we have roughly 9091 residents in each age group in a wave (the total population is 100000 in each wave).

- **Gender** (Binary): The gender of residents. We assume that 50% of residents are males and 50% residents are females.

- **Type II Diabetes** (Binary): This variable is an indicator to indicate if a resident has Type II diabetes. According to the CDC National Diabetes Statistics Report 2020, about 13% (95% confidence interval is (12%, 14.1%)) of U.S. adults have diabetes in 2018, and 90% - 95% of them have Type II diabetes approximately (Disease Control and (CDC), 2020). Therefore, we set 12.35% as the Type II diabetes rate in our simulated waves. That is, there are approximately 12350 residents who have Type II diabetes in each wave. We assume Type II Diabetes cannot be cured (even it can be in real life). That is, the residents who have Type II Diabetes in the preceding waves will also have Type II Diabetes if they show up in later waves.

- **Handedness** (Binary): This variable is an indicator to indicate if a resident prefers to use his/her left hand in daily life. According to Statista.com, there are approximately 13.1% of people prefer to use their left hand in their daily life in the U.S. (McCarthy, 2020). Therefore, we use this number as the percentage of the residents who prefer to use their left hand in our simulated wave. Therefore, there are approximately 13100 residents prefer to use their left hand in each wave. We assume people will not change their preference of handedness over time.

- **Health Insurance Cost** (Numeric): The health insurance cost for residents. According to william-russell.com, the average health insurance cost for an individual is 7470 U.S. dollars per year in the U.S. in 2020 (Talabani, 2021). The health insurance cost variable is assumed to relate to the inflation rate, and the inflation rate is assumed to be 3% per year. The values of health insurance cost in our dataset are generated from a normal distribution with mean equals 7470 and standard deviation equals 1500. Therefore, if the health insurance cost for a resident is 8000 U.S. dollars in 2010, it will increase to $8240 in 2011 due to the inflation rate. We assume that all residents do not change nor cancel their health insurance plan overtime in the simulated waves.

- **Body mass index** (Numeric): This variable represents the BMI values for the people in each wave. The BMI value is assumed to relate to the age and other health conditions of a person. For example, we expect to see a person with Type II diabetes also have a higher BMI value than a healthy person. We also expect to see an older person have a slightly higher BMI value than a younger person. The BMI variable is generated from a normal distribution with mean = 18.5 + Age * 0.2 + Type II diabetes * 10 the standard deviation = 1.7, where 18.5 is assumed as a standard BMI value for a healthy adult. That means, the mean BMI value for a person with Type II Diabetes is 10 units higher than a healthy person at the same age; the mean BMI value for an older person is 0.2 units higher than a person who is 1 year younger under the same health condition.

- **Parents Left-Handedness** (Binary): This is an auxiliary variable for doing calibration. We assume that if an individual is left-handedness, there will be a 95% of chance that at least one of the individual's parents is left-handedness. Similarly, if an individual is not left-handedness, then there will be a 5% of chance that at least one of the individual's parents is left-handedness.

- **Overweight** (Binary): This is an auxiliary variable for doing calibration. An individual will be considered to be overweight if the individual's BMI value is higher than 27. We assume people do not lose weight. That is, the residents who are overweight in the preceding waves will also be overweight if they show up in later waves.

- **Wave** (Factor): This is an auxiliary variable for doing calibration. It is for indicating the wave of an individual belongs to.
- **Stage 1 Probability** (Numeric): We will use stratified two-stage cluster sampling method to select samples from each wave and region as the strata. The Stage 1 probability is the inclusion probabilities for strata.
- **Stage 2 Probability** (Numeric): We will use stratified two-stage cluster sampling method to select samples from each wave, and cities in each region are the first stage clusters. Approximately 30% of the cities in each region will be selected randomly. Stage 2 Probability indicates the inclusion probabilities of first stage clusters.
- **Stage 3 Probability** (Numeric): We will use stratified two-stage cluster sampling method to select samples from each wave and residents in each is the second stage clusters. 100 residents will be randomly selected from each city. Stage 3 Probability indicates the inclusion probabilities of second-stage clusters.
- **Final inclusion probability** (Numeric): The final inclusion probability for each resident. It is the product of inclusion probabilities of Stage 1 Probability, Stage 2 Probability, and Stage 3 Probability.

4.3 Simulation: Selecting samples

A stratified two-stage cluster sampling method is applied for sample selection. The variable Regions is treated as strata. All eight regions are selected from each wave. Therefore, the inclusion probability for all eight regions is

$$\pi_{i,strata} = \frac{\text{number of selected regions}}{\text{total number of regions}} = \frac{8}{8} = 1$$

The variable Cities is the first-stage cluster, approximately 30% of cities in each region are randomly selected without replacement. Since the number of cities in each region is proportional to the region size, therefore, the first-stage inclusion probabilities can be calculated as

$$\pi_{i,stage1} = \frac{\text{number of selected cities in Region } i}{\text{total number of cities in Region } i}$$

The variable Resident ID is the second-stage cluster, 100 residents in each city are randomly selected without replacement. Therefore, the inclusion probability for the second-stage can be calculated as

$$\pi_{i,stage2} = \frac{\text{number of selected residents in City } i}{\text{total number of residents in City } i}$$

Therefore, the final inclusion probability can be calculated as

$$\begin{aligned}\pi_i &= \pi_{\text{i,Strata}} * \pi_{\text{i,Stage1}} * \pi_{\text{i,Stage 2}} \\ &= 1 * \frac{\text{Number of selected cities in Region } i}{\text{Total number of cities in Region } i} * \frac{\text{Number of selected residents in city } i}{\text{Total number of residents in City } i}\end{aligned}$$

4.4 Simulation: Combining populations of waves

In this study, we are trying to compare the differences in estimations when repeated observations are known in different waves and when repeated observations are unknown. When we combine different waves and the repeated observations are known, the repeated observations need to be removed. For example, suppose there are two waves need to be combined. Wave 1 covers all the residents between 20 and 30 years old in year 2010, and Wave 2 covers all the residents between 20 and 30 years old in year 2020. That means all residents who are 20 years old in Wave 1 will become 30 years old in Wave 2. Then all residents who are 20 years old will be removed from the population of Wave 1 since they are also present in Wave 2 as Figure 4.4 shows.

Similarly, if we want to combine three waves where overlapped observations are known and year difference is 5 years between every two consecutive waves, as Figure 4.5 shows, residents between 20 to 25 years old in Wave 1 will become 25 to 30 years old in Wave 2, and residents between 20 to 25 years old in Wave 2 will become 25 to 30 years old in Wave 3. Therefore, residents who are 20 years old in Wave 1 will be 25 years old in Wave 2 and 30 years old in Wave 3. Thus, residents between 20 to 25 years will be removed from Wave 1 and Wave 2.

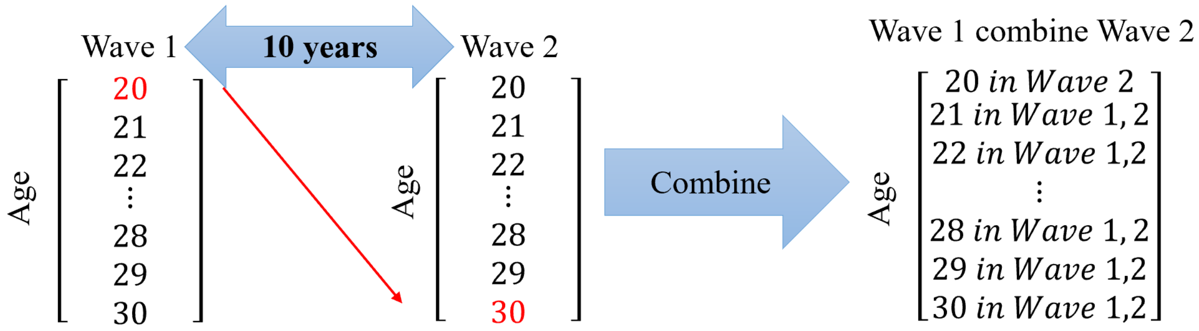

Figure 4.4: *Combine Wave 1 and Wave 2 with 10 years difference if overlapped observations are known*

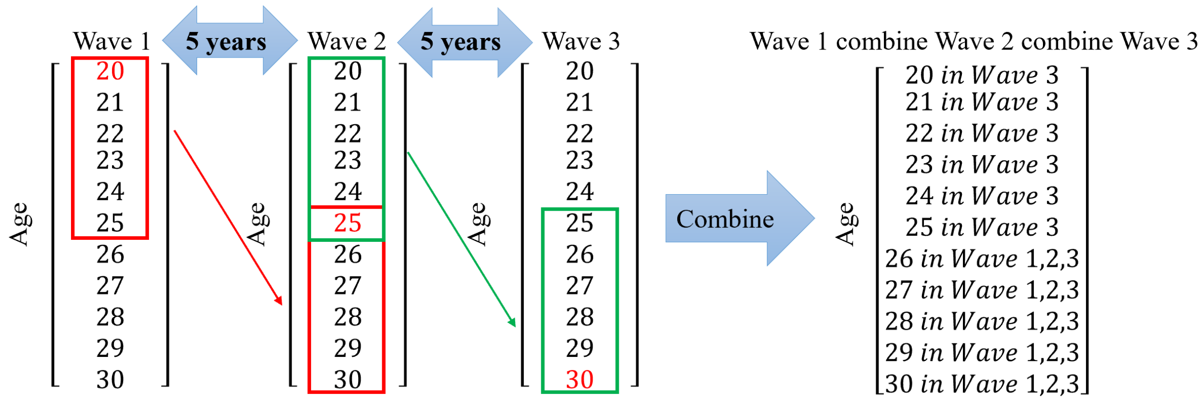

Figure 4.5: *Combine Wave 1, Wave 2, and Wave 3 with 5 years difference if overlapped observations are known*

We need to combine samples and remove overlapped observations to estimate the population of the combined waves. However, the original inclusion probabilities of the samples can no longer be used, otherwise, the estimates may be biased. Therefore, the inclusion probabilities need to be re-calculated. The general formula for re-calculating the inclusion probability for the combined sample is defined by

$$\pi_{i,Combined} = 1 - (1 - \pi_{i,\text{Wave 1}})(1 - \pi_{i,\text{Wave 2}}) \cdots (1 - \pi_{i,\text{Wave k}})$$

For example, suppose we want to combine sample (S_1) from Wave 1 and sample (S_2) from Wave 2 to estimate the combination of Wave 1 and Wave 2. If a resident presents in both S_1 and S_2 with inclusion probabilities 0.03 and 0.025 respectively, the new inclusion probability ($\pi_{i,Combined}$) after S_1 and S_2 are combined is $1 - (1 - 0.03)(1 - 0.025) = 0.05425$.

Here is the simple proof for the formula of re-calculating inclusion probability.

$$\begin{aligned} P(\text{ include }) &= P\left(i \in S_1 \text{ or } i \in S_2 \text{ or } \cdots \text{ or } i \in S_k\right) \\ &= 1 - P\left(i \notin S_1 \text{ and } i \notin S_2 \text{ and } \cdots \text{ and } i \notin S_k\right) \\ &= 1 - \left(1 - P\left(i \in S_1\right)\right)\left(1 - P\left(i \in S_2\right)\right) \cdots \left(1 - P\left(i \in S_k\right)\right) \\ &= 1 - \left(1 - \pi_{i,1}\right)\left(1 - \pi_{i,2}\right) \cdots \left(1 - \pi_{i,k}\right) \\ &= \pi_{i,Combined} \end{aligned}$$

If the overlapped observations of the population of the combined waves are assumed to be unknown, the overlapped observations of the population will not be removed as well as the overlapped observations in samples. That is, the original inclusion probabilities from the stratified two-stage cluster sampling will be used to do the estimations.

4.5 Simulation: Estimator

After removing repeated observations from combined samples and re-calculating inclusion probabilities, the totals or means of combined waves can be estimated without bias from combined samples. The following estimators are described in Cochran (1977). The estimator of total $\hat{Y}^C$ for stratified two-stage cluster sampling can be written as

$$\hat{Y}^C = \sum_{h=1}^{L} \frac{N_h}{n_h} \sum_{i=1}^{n_h} \frac{M_{hi}}{m_{hi}} \sum_{j=1}^{m_{hi}} y_{hij}$$

where L is the number of strata in the wave, N_h is the number of clusters in stratum h, n_h is the number of sample cluster in stratum h, M_{hi} is the number of population units in cluster i stratum s, m_{hi} is the size of sample observations in sampled cluster i stratum h, and y_{hij} is the j^{th} observation in i^{th} cluster h^{th} stratum.

The variance of estimator $\hat{Y}^{Total}$ is

$$\hat{V}\left(\hat{Y}^C\right) = \sum_{h=1}^{L} \left[\frac{N_h^2}{n_h} \left(1 - f_{1h}\right) \frac{\sum_{i=1}^{n_h} \left(\hat{Y}_{hi} - \overline{\hat{Y}}_h^C \right)^2}{n_h - 1} + \frac{N_h}{n_h} \sum_{i=1}^{n_h} M_{hi}^2 \left(1 - f_{2hi}\right) \frac{s_{2hi}^2}{m_{hi}} \right]$$

where $\hat{Y}_{hi} = M_{hi}\bar{y}_{hi}$, $\bar{y}_{hi} = m_{hi}^{-1} \sum_{j=1}^{m_{hi}} y_{hij}$, $\overline{\hat{Y}}_h^C = n_h^{-1} \sum_{i=1}^{n_h} M_{hi}\overline{y_{hi}}$, $f_{1h} = n_h / N_h$ (the finite population correction for first-stage), $f_{2hi} = m_{hi} / M_{hi}$ (the finite population correction for second-stage), and $s_{2hi}^2 = \sum_{j=1}^{m_{hi}} \left(y_{hij} - \bar{y}_{hi}\right)^2 / \left(m_{hi} - 1\right)$

The ratio estimator of stratified two-stage cluster sampling $\hat{Y}^{Ratio}$ can be written as

$$\hat{Y}^{Ratio} = \sum_{s=1}^{L} \frac{M_s}{\sum_{i=1}^{n_s} M_{si}} \sum_{i=1}^{n_s} \frac{M_{si}}{m_{si}} \sum_{j=1}^{m_{si}} y_{sij}$$

if the the number of units in each population stratum is known.

And the variance of ratio estimator $\hat{Y}^{Ratio}$ can be written as

$$\hat{V}\left(\hat{Y}^{Ratio}\right) = \sum_{s=1}^{L} \left[\frac{N_s^2}{n_s} \frac{\sum_{i=1}^{n_s} M_{si}^2 \left(\overline{y_{si}} - \hat{y}_s\right)^2}{n_{s-1}} \left(1 - f_{1s}\right) + \frac{N_s}{n_s} \sum_{i=1}^{n_s} M_{si}^2 \left(1 - f_{2si}\right) \frac{s_{2si}^2}{m_{si}} \right]$$

where $\hat{\bar{y}}_s = n_s^{-1} \sum_{i=1}^{n_s} \overline{y_{si}}$

Chapter 5

Results

Simulation results will be discussed in this chapter. Section 5.1 will discuss the estimations for combined waves with repeated observations removed. Section 5.2 will discuss the estimations for combined waves with repeated observations remain in the data. Section 5.3 and Section 5.4 will discuss the estimations for combined waves with auxiliary variables provided. There are four waves generated during the simulation, and we will compare the estimations for two waves combined, three waves combined, and four waves combined.

A sample of 3400 observations is selected from each wave populations by using stratified two-stage cluster sampling method. The variable `Regions` are strata; the variable `Cities` and `Residents` are receptively the first- and second-stage clusters. The samples from each wave will be combined, and repeated observations will be removed to estimate the population of combined waves without repeated observations if the repeated observations in the combined population of waves are assumed to be known. The inclusion probabilities of the combined samples are re-calculated by applying the method mentioned in Section 4.4. These processes are repeated 1000 times, so 1000 independent samples are selected from each wave, and 1000 estimations will be conducted. The final results are the average of these 1000 estimates.

5.1 Estimations for combined waves without repeated observations

This part discusses the estimates for the population of combined waves without repeated observations such as total number of people who have Type II Diabetes. That is, all repeated observations are removed from the populations and samples after waves are combined. For example, if a person presents in the populations and samples of both Wave 1 and Wave 2, after two waves are combined, the observation of that person in Wave 1 will be removed but the observation of that person in Wave 2 will remain in the data. Please refer to Chapter 4 for more details about simulation methodology. The standard errors will be calculated over 1000 estimates.

5.1.1 Estimations for combination of two waves

Year Diff.	Diabetes	True BMI in Combined Pop.	Re-calculated Weight
10 years	False	23.54	23.49(0.02)
10 years	True	33.54	33.49(0.06)
5 years	False	23.69	23.51(0.03)
5 years	True	33.69	33.50(0.07)
1 years	False	23.59	23.51(0.02)
1 years	True	33.57	33.49(0.06)

Table 5.1: *Average BMI of residents who have Type II Diabetes for two waves combined.*

Year Diff.	True Cost in Combined Pop.	Re-calculated Weight
10 years	8812.49	8753.16(92.12)
5 years	8290.85	8069.18(46.48)
1 years	7672.74	7580.16(20.33)

Table 5.2: *Average health insurance cost for two waves combined.*

Year Diff.	True Proportion in Combined Pop.	Re-calculated Weight
10 years	0.1236	0.1237(0.0040)
5 years	0.1227	0.1228(0.0041)
1 years	0.1231	0.1230(0.0040)

Table 5.3: *The proportion of people who have Type II Diabetes for two waves combined.*

Year Diff.	Diabetes	True BMI in Combined Pop.	Re-calculated Weight
10 years	False	23.50	23.50(0.02)
10 years	True	33.49	33.49(0.06)
5 years	False	23.50	23.50(0.04)
5 years	True	33.50	33.49(0.07)
1 years	False	23.50	23.50(0.02)
1 years	True	33.49	33.49(0.06)

Table 5.4: *Average BMI of residents who have Type II Diabetes for two waves combined (taking average of repeated observations).*

Year Diff.	True Cost in Combined Pop.	Re-calculated Weight
10 years	8752.32	8751.54(87.96)
5 years	8065.48	8065.78(32.92)
1 years	7579.24	7580.27(18.87)

Table 5.5: *Average health insurance cost for two waves combined (taking average of repeated observations).*

Table 5.1 and Table 5.2 show that the estimates with using re-calculated weights contain some minor biases. The standard error also indicates that the estimates of average cost of health insurance contains larger bias. This may be due to the characteristics of BMI and health insurance cost since the values of these two variables change over time. Therefore, we changed our method a little bit to see if better estimates could be produced. That is, we do not directly delete repeated observations from populations and samples this time. Instead, we calculate the average values of the repeated observations for the same person, then do the estimations. For example, if the BMI value for a person is 28 in Wave 1 and 29 in Wave 2, then the final BMI value for that person after two waves combined will be 28.5.

Table 5.4 and Table 5.5 show that there are no notable biases anymore after taking the average of BMI and health insurance cost for repeated observations instead of deleting repeated observations directly from populations and samples.

Table 5.3 shows the estimates of the proportion of people who have Type II Diabetes for two waves combined. Unlike the estimates in Table 5.1 and Table 5.2, estimates in Table 5.3 does not contain notable biases. This may be because Type II Diabetes variable does

Year Diff.	True Total in Combined Pop.	Re-calculated Weight
10 years	23605	23633.23(1045.69)
5 years	17803	17822.14(803.85)
1 years	13419	13402.95(582.52)

Table 5.6: *Total number of people who have Type II Diabetes for two waves combined.*

not change over time (we assume that Type II Diabetes cannot be cured when generating the data).

Table 5.6 shows the estimates of the total number of people who have Type II Diabetes. The result shows that the estimates using re-calculated weights do not contain notable bias.

5.1.2 Estimations for combination of three waves

Year Diff.	Diabetes	True BMI in Combined Pop.	Re-calculated Weight
10 years	False	23.56	23.50(0.02)
10 years	True	33.56	33.49(0.05)
5 years	False	23.78	23.50(0.02)
5 years	True	33.79	33.51(0.05)
1 years	False	23.66	23.51(0.02)
1 years	True	33.65	33.50(0.06)

Table 5.7: *Average BMI of residents who have Type II Diabetes for three waves combined.*

Year Diff.	True Cost in Combined Pop.	Re-calculated Weight
10 years	10321.26	10282.05(113.87)
5 years	9102.93	8742.56(57.24)
1 years	7868.74	7697.17(18.26)

Table 5.8: *Average health insurance cost for three waves combined.*

Year Diff.	True Proportion in Combined Pop.	Re-calculated Weight
10 years	0.1227	0.1228(0.0033)
5 years	0.1233	0.1236(0.0034)
1 years	0.1231	0.1232(0.0034)

Table 5.9: *The proportion of people who have Type II Diabetes for three waves combined.*

Year Diff.	Diabetes	True BMI in Combined Pop.	Re-calculated Weight
10 years	False	23.50	23.50(0.02)
10 years	True	33.49	33.49(0.05)
5 years	False	23.50	23.50(0.03)
5 years	True	33.50	33.50(0.06)
1 years	False	23.50	23.51(0.02)
1 years	True	33.49	33.49(0.06)

Table 5.10: *Average BMI of residents who have Type II Diabetes for three waves combined (Taking average of repeated observations).*

Year Diff.	True Cost in Combined Pop.	Re-calculated Weight
10 years	10280.490	10280.28(113.18)
5 years	8735.304	8735.91(47.04)
1 years	7692.870	7692.25(16.10)

Table 5.11: *Average health insurance cost for three waves combined (Taking average of repeated observations).*

Similar to the results of two waves combined, the estimates of average BMI for diabetes and average of health insurance cost (Table 5.7 and Table 5.8) contain some minor biases. After using the average values of BMI and health insurance cost for the repeated observations, the biases are not notable anymore (Table 5.10 and Table 5.11).

Estimates in Table 5.9 does not contain notable biases. This result is consistent with the results of two waves combined.

Table 5.12 shows the estimates of the total number of people who have Type II Diabetes. Similar to the two waves combined results, the estimates of using re-calculated weights do not contain notable biases.

5.1.3 Estimations for combination of four waves

Similar to the results of two waves combined and three waves combined, the estimates of average BMI for diabetes and average of health insurance cost (Table 5.13 and Table 5.14) contain some minor biases. After using the average values of BMI and health insurance cost for the repeated observations, the biases are not notable anymore (Table 5.16 and Table 5.17).

Year Diff.	True Total in Combined Pop.	Re-calculated Weight
10 years	34604	34622.9(1343.09)
5 years	23560	23605.36(939.634)
1 years	14540	14558.48(554.697)

Table 5.12: *Total number of people who have Type II Diabetes for three waves combined.*

Year Diff.	Diabetes	True BMI in Combined Pop.	Re-calculated Weight
10 years	False	23.54	23.48(0.02)
10 years	True	33.54	33.48(0.04)
5 years	False	23.85	23.51(0.02)
5 years	True	33.85	33.51(0.05)
1 years	False	23.72	23.51(0.02)
1 years	True	33.71	33.50(0.05)

Table 5.13: *Average BMI of residents who have Type II Diabetes for four waves combined.*

Estimates in Table 5.15 does not contain notable biases. This result is consistent with the results of two waves combined.

Table 5.18 shows the estimates of the total number of people who have Type II Diabetes. Similar to the two waves combined results, the estimates of using re-calculated weights do not contain notable biases.

Year Difference	True Cost in Combined Pop.	Re-calculated Weight
10 years	12319.55	12242.73(132.64)
5 years	9659.19	9338.50(55.72)
1 years	8059.92	7816.59(17.08)

Table 5.14: *Average health insurance cost for four waves combined.*

Year Difference	True Proportion in Combined Pop.	Re-calculated Weight
10 years	0.1231	0.1233(0.0028)
5 years	0.1231	0.1232(0.0030)
1 years	0.1229	0.1230(0.0030)

Table 5.15: *The proportion of people who have Type II Diabetes for four waves combined.*

Year Diff.	Diabetes	True BMI in Combined Pop.	Re-calculated Weight
10 years	False	23.48	23.48(0.02)
10 years	True	33.48	33.48(0.04)
5 years	False	23.50	23.50(0.02)
5 years	True	33.50	33.51(0.05)
1 years	False	23.50	23.50(0.02)
1 years	True	33.50	33.49(0.05)

Table 5.16: *Average BMI of residents who have Type II Diabetes for four waves combined (Taking average of repeated observations).*

Year Diff.	True Cost in Combined Pop.	Re-calculated Weight
10 years	12241.73	12241.18(131.53)
5 years	9333.47	9333.99(51.03)
1 years	7809.99	7810.16(14.84)

Table 5.17: *Average health insurance cost for four waves combined (Taking average of repeated observations).*

Year Diff.	True Total in Combined Pop.	Re-calculated Weight
10 years	45896	45967(1505.08)
5 years	29072	29081.53(984.67)
1 years	15617	15620.22(519.79)

Table 5.18: *Total number of people who have Type II Diabetes for four waves combined.*

5.2 Estimations for combined waves with repeated observations included

This part discusses the estimates for combined waves repeated observations. That is, all repeated observations will remain in the populations (waves) and samples after waves are combined. For example, if a person presents in both Wave 1 and Wave 2, after two waves are combined, the observation of that person in Wave 1 and the observation of that person in Wave 2 will remain in the data. Please refer to Chapter 4 for more details about simulation methodology.

5.2.1 Estimations for combination of two waves

Year Diff.	Diabetes	True BMI in Combined Pop.	Original Weight
10 years	False	23.50	23.50(0.02)
10 years	True	33.49	33.49(0.06)
5 years	False	23.50	23.50(0.02)
5 years	True	33.50	33.50(0.06)
1 years	False	23.51	23.50(0.02)
1 years	True	33.49	33.49(0.06)

Table 5.19: *Average BMI of residents who have Type II Diabetes for two waves combined.*

Year Diff.	True Cost in Combined Pop.	Original Weight
10 years	8752.62	8752.49(92.03)
5 years	8065.38	8065.27(45.95)
1 years	7579.51	7578.90(20.16)

Table 5.20: *Average health insurance cost for two waves combined.*

Year Diff.	True Proportion in Combined Pop.	Original Weight
10 years	0.1237	0.1239(0.0039)
5 years	0.1235	0.1236(0.0040)
1 years	0.1232	0.1231(0.0040)

Table 5.21: *The proportion of people who have Type II Diabetes for two waves combined.*

Table 5.19 and Table 5.20 show the estimates of the average BMI of residents who have Type II Diabetes and average health insurance cost. Unlike the results in Section 5.1, we

Year Diff.	True Total in Combined Pop.	Original Weight
10 years	24741	24770.94(1092.53)
5 years	24701	24712.82(1103.25)
1 years	24643	24624.11(1099.81)

Table 5.22: *Total number of people who have Type II Diabetes for two waves combined.*

do not see notable biases this time. This makes sense because we did not delete any time dependent variables or observations from populations nor samples; therefore, we do not lose any information.

Table 5.21 contains the estimates of the proportion of people who have Type II Diabetes; Table 5.22 contains the estimates of the total number of people who have Type II Diabetes. The estimates in those tables do not contain notable biases.

5.2.2 Estimations for combination of three waves

Year Diff.	Diabetes	True BMI in Combined Pop.	Original Weight
10 years	False	23.50	23.50(0.02)
10 years	True	33.49	33.50(0.05)
5 years	False	23.50	23.50(0.02)
5 years	True	33.50	33.50(0.05)
1 years	False	23.51	23.51(0.02)
1 years	True	33.50	33.50(0.05)

Table 5.23: *Average BMI of residents who have Type II Diabetes for three waves combined.*

Year Diff.	True Cost in Combined Pop.	Original Weight
10 years	10227.36	10228.15(109.22)
5 years	8725.21	8725.46(51.41)
1 years	7693.47	7693.88(17.74)

Table 5.24: *Average health insurance cost for three waves combined.*

Similar to the previous part, there is no notable bias for estimates of average BMI of residents who have Type II Diabetes and average health insurance cost if we use original weight to do the estimation for three waves combined with repeated observations remain in the data (Table 5.23 and Table 5.24). The estimates of the proportion of people who

Year Diff.	True Proportion in Combined Pop.	Original Weight
10 years	0.1228	0.1229(0.0032)
5 years	0.1236	0.1239(0.0033)
1 years	0.1234	0.1235(0.0033)

Table 5.25: *The proportion of people who have Type II Diabetes for three waves combined.*

Year Diff.	True Total in Combined Pop.	Original Weight
10 years	36846	36876.33(1422.98)
5 years	37092	37157.02(1431.74)
1 years	37015	37049.66(1432.76)

Table 5.26: *Total number of people who have Type II Diabetes for three waves combined.*

have Type II Diabetes and the estimates of the total number of people who have Type II Diabetes do not contain notable biases either (Table 5.25 and Table 5.26).

5.2.3 Estimations for combination of four waves

Year Diff.	Diabetes	True BMI in Combined Pop.	Original Weight
10 years	False	23.47	23.47(0.02)
10 years	True	33.47	33.47(0.04)
5 years	False	23.50	23.50(0.02)
5 years	True	33.51	33.51(0.04)
1 years	False	23.51	23.51(0.02)
1 years	True	33.50	33.50(0.05)

Table 5.27: *Average BMI of residents who have Type II Diabetes for four waves combined.*

Year Diff.	True Cost in Combined Pop.	Original Weight
10 years	12185.36	12185.07(127.91)
5 years	9277.26	9276.61(45.87)
1 years	7810.05	7810.03(16.19)

Table 5.28: *Average health insurance cost for four waves combined.*

Similar to the previous part, there is no notable bias for estimates of average BMI of residents who have Type II Diabetes and average health insurance cost if we use original weight to do the estimation for three waves combined with repeated observations remain

Year Diff.	True Proportion in Combined Pop.	Original Weight
10 years	0.1231	0.1232(0.0028)
5 years	0.1235	0.1235(0.0028)
1 years	0.1235	0.1234(0.0029)

Table 5.29: *The proportion of people who have Type II Diabetes for four waves combined.*

Year Diff.	True Total in Combined Pop.	Original Weight
10 years	49231	49296.37(1603.34)
5 years	49401	49413.83(1618.97)
1 years	49381	49379.02(1641.45)

Table 5.30: *Total number of people who have Type II Diabetes for four waves combined.*

in the data (Table 5.27 and Table 5.28). The estimates of the proportion of people who have Type II Diabetes and the estimates of the total number of people who have Type II Diabetes do not contain notable biases either (Table 5.29 and Table 5.30).

The results in this section suggest that the standard error of the estimates with overlapped observations remain in the combined population are smaller compared to the standard error of the estimates without overlapped observations remain in the combined population shown in Section 5.1.

5.3 Estimations for combined waves with calibration (without repeated observations)

This section will discuss the results of calibration by using auxiliary variables. The auxiliary variables are `Wave`, `Overweight`, `Parents Left-Handedness` and `Age`. In the simulation process, we assume that there is positive correlations between `overweight` and `Type II Diabetes`. That is, people who are overweight have higher chance to have Type II Diabetes. We also assume `Parents Left-Handedness` have positive correlation with `Handedness`. That is, if one of the parents are left-handedness, their children will have higher chance to be left-handedness. The variables `Wave` and `Age` provide the time effects. Please refer to Chapter 4 for more details about simulation methodology. For calibration without repeated observations, the repeated observations will be removed from populations and samples after waves are combined. Then the inclusion probabilities will be re-calculated by using the formula mentioned in Section 4.4 and calibration method will be applied to the survey design.

5.3.1 Estimations for combination of two waves

Year Diff.	Diabetes	True BMI in Combined Pop.	Original Weight	Calibrated Weight
10 years	False	23.54	23.50(0.02)	23.54(0.02)
10 years	True	33.54	33.49(0.06)	33.54(0.06)
5 years	False	23.69	23.50(0.02)	23.69(0.02)
5 years	True	33.69	33.50(0.06)	33.69(0.07)
1 years	False	23.59	23.50(0.02)	23.59(0.03)
1 years	True	33.57	33.49(0.06)	33.58(0.08)

Table 5.31: *[Calibration] Average BMI of residents who have Type II Diabetes for two waves combined.*

Year Diff.	True Cost in Combined Pop.	Original Weight	Calibrated Weight
10 years	8812.49	8752.49(92.02)	8812.32(18.32)
5 years	8290.85	8065.27(45.90)	8290.81(21.57)
1 years	7672.74	7578.90(19.89)	7672.57(24.17)

Table 5.32: *[Calibration] Average health insurance cost for two waves combined.*

Year Diff.	True Proportion in Combined Pop.	Original Weight	Calibrated Weight
10 years	0.1236	0.1238(0.0039)	0.1237(0.0017)
5 years	0.1227	0.1236(0.0040)	0.1227(0.0021)
1 years	0.1231	0.1231(0.0039)	0.1231(0.0023)

Table 5.33: *[Calibration] The proportion of people who have Type II Diabetes for two waves combined.*

Table 5.31 and Table 5.32 show the estimates of average BMI of residents who have Type II Diabetes and the estimates of average health insurance cost from calibration. Unlike the results in Section 5.1 (Table 5.1 and Table 5.2), the estimates from calibration do not contain notable biases even through the BMI value and average health insurance cost change over time because `Wave` and `Age` are included in the calibration and these essentially produce the time effects. Table 5.33 and Table 5.34 show the estimates of the proportion of people

Year Diff.	True Total in Combined Pop.	Original Weight	Calibrated Weight
10 years	23605	24770.94(1092.01	23628.54(331.85)
5 years	17803	24712.82(1099.10)	17793.79(307.17)
1 years	13419	24624.11(1090.56)	13421.5(258.64)

Table 5.34: *[Calibration] Total number of people who have Type II Diabetes for two waves combined.*

who have Type II Diabetes and the total number of people who have Type II Diabetes. The results are fairly good with no notable biases and the standard error is much smaller than the results in Section 5.1 (Table 5.6)

5.3.2 Estimations for combination of three waves

Year Diff.	Diabetes	True BMI in Combined Pop.	Original Weight	Calibrated Weight
10 years	False	23.56	23.50(0.02)	23.56(0.02)
10 years	True	33.56	33.50(0.05)	33.56(0.05)
5 years	False	23.78	23.50(0.02)	23.78(0.02)
5 years	True	33.79	33.50(0.05)	33.79(0.06)
1 years	False	23.66	23.50(0.02)	23.67(0.03)
1 years	True	33.65	33.50(0.05)	33.65(0.07)

Table 5.35: *[Calibration] Average BMI of residents who have Type II Diabetes for three waves combined.*

Year Diff.	True Cost in Combined Pop.	Original Weight	Calibrated Weight
10 years	10321.26	10228.15(109.22)	10361.85(17.20)
5 years	9102.93	8725.46(51.34)	9103.55(19.14)
1 years	7868.74	7693.88(17.33)	7867.51(22.29)

Table 5.36: *[Calibration] Average health insurance cost for three waves combined.*

Similar to the previous part, the estimates from calibration show very good performance. Table 5.35 and Table 5.36 show no biases for the estimates of BMI and average health insurance cost. The estimates of Type II Diabetes and the total number of people who have Diabetes in Table 5.37 and Table 5.38 also show good performance. The standard error of the estimates is much smaller compared to the results in Table 5.12.

Year Diff.	True Proportion in Combined Pop.	Original Weight	Calibrated Weight
10 years	0.1227	0.1229(0.0032)	0.1228(0.0014)
5 years	0.1233	0.1239(0.0032)	0.1232(0.0019)
1 years	0.1231	0.1235(0.0032)	0.1231(0.0023)

Table 5.37: *[Calibration] The proportion of people who have Type II Diabetes for three waves combined.*

Year Diff.	True Total in Combined Pop.	Original Weight	Calibrated Weight
10 years	34604	36876.33(1422.13)	34616.80(408.32)
5 years	23560	37157.02(1424.08)	23543.49(367.36)
1 years	14540	37049.66(1411.57)	14534(269.54)

Table 5.38: *[Calibration] Total number of people who have Type II Diabetes for three waves combined.*

5.3.3 Estimations for combination of four waves

Year Diff.	Diabetes	True BMI in Combined Pop.	Original Weight	Calibrated Weight
10 years	False	23.54	23.47(0.01)	23.55(0.02)
10 years	True	33.54	33.47(0.04)	33.54(0.04)
5 years	False	23.85	23.50(0.02)	23.84(0.02)
5 years	True	33.85	33.51(0.04)	33.86(0.05)
1 years	False	23.72	23.51(0.02)	23.72(0.02)
1 years	True	33.70	33.50(0.04)	33.71(0.07)

Table 5.39: *[Calibration] Average BMI of residents who have Type II Diabetes for four waves combined.*

Year Diff.	True Cost in Combined Pop.	Original Weight	Calibrated Weight
10 years	12319.55	12185.07(127.90)	12366.2(20.47)
5 years	9659.19	9276.61(45.79)	9608.03(19.03)
1 years	8059.92	7810.03(15.68)	8058.91(20.63)

Table 5.40: *[Calibration] Average health insurance cost for four waves combined.*

Similar to the previous part, the estimates from calibration show good performance. All the estimates (Table 5.39, Table 5.40, Table 5.41, Table 5.42) show no notable biases. The standard error of the estimates is much smaller compared to the results in Section 5.1.

Year Diff.	True Proportion in Combined Pop.	Original Weight	Calibrated Weight
10 years	0.1231	0.1232(0.0028)	0.1230(0.0012)
5 years	0.1231	0.1235(0.0028)	0.1228(0.0018)
1 years	0.1229	0.1234(0.0028)	0.1231(0.0022)

Table 5.41: *[Calibration] The proportion of people who have Type II Diabetes for four waves combined.*

Year Diff.	True Total in Combined Pop.	Original Weight	Calibrated Weight
10 years	45896	49296.37(1602.16)	45860.47(464.64)
5 years	29072	49413.83(1607.68)	28999.46(419.34)
1 years	15617	49379.02(1605.49)	15647.13(274.58)

Table 5.42: *[Calibration] Total number of people who have Type II Diabetes for four waves combined.*

The results in this section suggest that the calibration method show good performance in estimating ratios and totals when repeated observations are not included in the population of combined waves. The estimates are more accurate. The standard errors are smaller especially for estimating the totals compared to the results in Section 5.1 (Table 5.6, Table 5.12, Table 5.18). This is because the auxiliary variables provide the time effects (`Wave`, `Age`) and additional information (`Parents Left-Handedness`, `Overweight`).

5.4 Estimations for combined waves with calibration (with repeated observations)

For calibration with repeated observations, the repeated observations will remain in populations and samples after waves are combined. Then the inclusion probabilities will not be re-calculated but calibration method will be applied to the survey design.

5.4.1 Estimations for combination of two waves

For repeated observation remain in the data, the calibration method also shows very good performance. The estimates (Table 5.43, Table 5.44, Table 5.45, Table 5.46) are unbiased and standard error of the estimates is smaller than the results in Section 5.2.

Year Diff.	Diabetes	True BMI in Combined Pop.	Original Weight	Calibrated Weight
10 years	False	23.50	23.50(0.02)	23.50(0.02)
10 years	True	33.49	33.49(0.06)	33.49(0.06)
5 years	False	23.50	23.50(0.02)	23.50(0.02)
5 years	True	33.50	33.50(0.06)	33.50(0.06)
1 years	False	23.51	23.50(0.02)	23.50(0.02)
1 years	True	33.49	33.49(0.06)	33.49(0.06)

Table 5.43: *[Calibration] Average BMI of residents who have Type II Diabetes for two waves combined.*

Year Diff.	True Cost in Combined Pop.	Original Weight	Calibrated Weight
10 years	8752.62	8752.49(92.02)	8752.70(18.28)
5 years	8065.38	8065.27(45.95)	8066.15(18.88)
1 years	7579.51	7578.90(20.16)	7580.21(18.56)

Table 5.44: *[Calibration] Average health insurance cost for two waves combined.*

5.4.2 Estimations for combination of three waves

Similar to the previous part, the calibration method shows very good performance. The estimates (Table 5.47, Table 5.48, Table 5.49, Table 5.50) are un-biased and standard error of the estimates is smaller than the results in Section 5.2.

5.4.3 Estimations for combination of four waves

Similar to the previous part, the calibration method shows very good performance. The estimates (Table 5.51, Table 5.52, Table 5.53, Table 5.54) are un-biased and standard error of the estimates is smaller than the results in Section 5.2.

The results in this section suggest that the calibration method show good performance in estimating ratios and totals when repeated observations are included in the population of combined waves. The estimates are more accurate. The standard errors are smaller especially for estimating the totals compared to the results in Section 5.2 (Table 5.22, Table 5.26, Table 5.30). This is because the auxiliary variables provide the time effects (`Wave`, `Age`) and additional information (`Parents Left-Handedness`, `Overweight`).

Year Diff.	True Proportion in Combined Pop.	Original Weight	Calibrated Weight
10 years	0.1237	0.1238(0.0040)	0.1237(0.0017)
5 years	0.1235	0.1236(0.0040)	0.1235(0.0017)
1 years	0.1232	0.1231(0.0040)	0.1232(0.0017)

Table 5.45: *[Calibration] The proportion of people who have Type II Diabetes for two waves combined.*

Year Diff.	True Total in Combined Pop.	Original Weight	Calibrated Weight
10 years	24741	24770.94(1092.52)	24746.02(332.17)
5 years	24701	24712.82(1103.25)	24705.35(333.35)
1 years	24643	24624.11(1099.81)	24632.01(337.42)

Table 5.46: *[Calibration] Total number of people who have Type II Diabetes for two waves combined.*

Year Diff.	Diabetes	True BMI in Combined Pop.	Original Weight	Calibrated Weight
10 years	False	23.50	23.50(0.02)	23.50(0.02)
10 years	True	33.50	33.50(0.05)	33.49(0.05)
5 years	False	23.50	23.50(0.02)	23.50(0.02)
5 years	True	33.50	33.50(0.05)	33.50(0.05)
1 years	False	23.51	23.51(0.02)	23.51(0.02)
1 years	True	33.50	33.50(0.05)	33.50(0.06)

Table 5.47: *[Calibration] Average BMI of residents who have Type II Diabetes for three waves combined.*

Year Diff.	True Cost in Combined Pop.	Original Weight	Calibrated Weight
10 years	10227.36	10228.15(109.22)	10227.96(17.00)
5 years	8725.22	8725.46(51.41)	8725.97(15.93)
1 years	7693.47	7693.88(17.74)	7693.27(15.51)

Table 5.48: *[Calibration] Average health insurance cost for three waves combined.*

Year Diff.	True Proportion in Combined Pop.	Original Weight	Calibrated Weight
10 years	0.1228	0.1229(0.0032)	0.1228(0.0014)
5 years	0.1236	0.1239(0.0033)	0.1237(0.0014)
1 years	0.1234	0.1235(0.0033)	0.1233(0.0014)

Table 5.49: *[Calibration] The proportion of people who have Type II Diabetes for three waves combined.*

Year Diff.	True Total in Combined Pop.	Original Weight	Calibrated Weight
10 years	36846	36876.33(1422.98)	36839.54(408.23)
5 years	37092	37157.02(1431.75)	37105.51(410.08)
1 years	37015	37049.66(1432.76)	36982.59(418.02)

Table 5.50: *[Calibration] Total number of people who have Type II Diabetes for three waves combined.*

Year Diff.	Diabetes	True BMI in Combined Pop.	Original Weight	Calibrated Weight
10 years	False	23.47	23.47(0.02)	23.47(0.02)
10 years	True	33.47	33.47(0.04)	33.47(0.04)
5 years	False	23.50	23.50(0.02)	23.50(0.02)
5 years	True	33.50	33.51(0.04)	33.51(0.04)
1 years	False	23.51	23.51(0.02)	23.51(0.02)
1 years	True	33.50	33.50(0.05)	33.50(0.05)

Table 5.51: *[Calibration] Average BMI of residents who have Type II Diabetes for four waves combined.*

Year Diff.	True Cost in Combined Pop.	Original Weight	Calibrated Weight
10 years	12185.36	12185.07(127.91)	12185.78(19.69)
5 years	9277.26	9276.61(45.87)	9276.78(14.25)
1 years	7810.05	7810.03(16.19)	7809.74(13.78)

Table 5.52: *[Calibration] Average health insurance cost for four waves combined.*

Year Diff.	True Proportion in Combined Pop.	Original Weight	Calibrated Weight
10 years	0.1231	0.1232(0.0028)	0.1230(0.0012)
5 years	0.1235	0.1235(0.0028)	0.1235(0.0012)
1 years	0.1235	0.1234(0.0029)	0.1235(0.0012)

Table 5.53: *[Calibration] The proportion of people who have Type II Diabetes for four waves combined.*

Year Diff.	True Total in Combined Pop.	Original Weight	Calibrated Weight
10 years	49231	49296.37(1603.33)	49208.8(464.09)
5 years	49401	49413.83(1618.97)	49389.56(474.36)
1 years	49381	49379.02(1641.45)	49379.89(488.54)

Table 5.54: *[Calibration] Total number of people who have Type II Diabetes for four waves combined.*

5.5 National Health and Nutrition Examination Survey (NHANES) Example

Some survey research programs provide their way to combine waves for researchers to create larger samples and it is useful for analyzing rare events and small groups (Chen et al., 2020). For example, the National Health and Nutrition Examination Survey (NHANES). The NHANES is a survey conducted by the National Center for Health Statistics (NCHS) that aims to study the health and nutritional status of people in the United States (Health Statistics (NCHS), 2017). The NHANES uses a four stages sample design. The first stage is PSUs (i.e. counties), the second stage is segments within PSUs such as census blocks, the third stage is dwelling units such as households, and the forth stage is individuals within households (Chen et al., 2020). The NHANES data for public use contains 2-year weights and survey designers suggest dividing the 2-year weights by the number of waves we want to combine to obtain the new weights (Chen et al., 2020). In this short example, the first two, first three, and all four waves (2011-2012, 2013-2014, 2015-2016, and 2017-2018) will be combined and the alcohol use for adults in the USA will be estimated. The original data can be found at NHANES website: `http://www.cdc.gov/nchs/nhanes/index.htm`

Wave	12 or More Drink	Total Number	Standard Error
2011-2012	Yes	149940424	9913768
2011-2012	No	34322360	2219761
2013-2014	Yes	150488689	10096123
2013-2014	No	43546287	5221265
2015-2016	Yes	146126610	9445203
2015-2016	No	45418498	2361064
2017-2018	Yes	106723249	5344214
2017-2018	No	80889209	4535038

Table 5.55: *Estimated total number of people who drink at least 12 alcohol drinks per year in USA.*

The estimated total numbers of people who drink at least 12 alcohol drinks per year in USA (as Table 5.55 and Figure 5.1 shows) are 149,940,424 (2011-2012), 150,488,689 (2013-2014), 146,126,610 (2015-2016), 106,723,249 (2017-2018). Table 5.56 shows the estimated total number of people who drink at least 12 alcohol drinks per year in the USA for combined waves, and the estimated number is just the average of the estimated number of individual

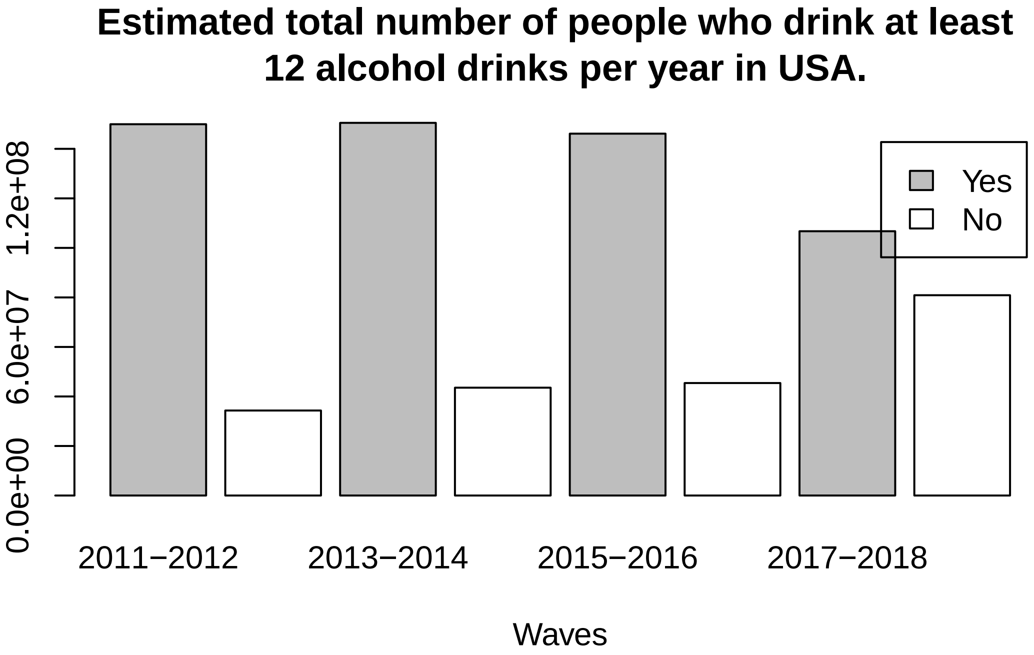

Figure 5.1: *Estimated total number of people who drink at least 12 alcohol drinks per year in USA.*

Number of Waves	12 or More Drink	Total Number	Standard Error
2 Waves	Yes	150214557	7074858
2 Waves	No	38934324	2836765
3 Waves	Yes	148851908	5670845
3 Waves	No	41095715	2048402
4 Waves	Yes	138319743	4458047
4 Waves	No	51044089	1909354

Table 5.56: *Estimated total number of people who drink at least 12 alcohol drinks per year in USA (combined waves).*

waves. For example, the estimated total number of people who drink at least 12 alcoholic drinks per year for the first two waves combined is 150,214,557, which is the average value of 2011-2012 and 2013-2014. The results suggest that the total number of people who drink at least 12 alcohol drinks per year decreases over years in the US.

Chapter 6

Conclusion

In practice, it is not always possible to identify the repeated observations when combining surveys (waves). However, if repeated observations can be identified and need to be removed in order to meet study needs, re-calculating the inclusion probabilities (or weights) is necessary. Otherwise, the estimates will contain significant biases as more surveys (waves) are combined if using original weights. If the variables we are interested in change over time, the estimates produced by re-calculated weights may contain minor biases. One possible solution is to use the average values of the interested variables for repeated observations instead of directly removing them from the populations and samples. However, this solution is not always recommended since sometimes the average values of certain variables do not have any scientific significance.

If the repeated observations cannot be identified when combining surveys (waves), using original weights to estimate will not produce biased estimates. Therefore, re-calculating weights is not necessary. In fact, using re-calculated weights to do estimations when repeated observations remain in the data will yield biases.

Results show that the calibration method has good performance no matter the repeated observations are removed or not. Therefore, using the calibration method when auxiliary variables are available is recommended since it will help reduce the standard error and produce more accurate estimates. And the calibration method can particularly reduce bias for time varying variables when re-calculated weights are used.

Appendix A

Additional results from simulation

This part contains some additional results from simulation. All results are consistent with the findings we concluded in Chapter 6.

A.1 Estimations for combined waves without repeated observations

Year Diff.	Gender	True BMI in Combined Pop.	Re-Calculated Weight
10 years	Female	24.77	24.73(0.06)
10 years	Male	24.78	24.74(0.06)
5 years	Female	24.91	24.73(0.07)
5 years	Male	24.93	24.74(0.07)
1 years	Female	24.80	24.72(0.07)
1 years	Male	24.83	24.75(0.07)

Table A.1: *Average BMI of different genders for two waves combined.*

Year Diff.	Gender	True BMI in Combined Pop.	Re-Calculated Weight
10 years	Female	24.73	24.73(0.06)
10 years	Male	24.73	24.73(0.06)
5 years	Female	24.72	24.73(0.07)
5 years	Male	24.74	24.74(0.07)
1 years	Female	24.72	24.72(0.07)
1 years	Male	24.75	24.75(0.07)

Table A.2: *Average BMI of different genders for two waves combined (taking average of repeated observations).*

Year Diff.	Overlap Perc.	Left Hand	True BMI in Combined Pop.	Re-Calculated Weight
10 years	9.1	False	24.77	24.73(0.05)
10 years	9.1	True	24.78	24.75(0.13)
5 years	54.5	False	24.92	24.73(0.05)
5 years	54.5	True	24.93	24.74(0.13)
1 years	91.0	False	24.81	24.73(0.05)
1 years	91.0	True	24.83	24.74(0.13)

Table A.3: *Average BMI of different Handedness for two waves combined.*

Year Diff.	Left Hand	True BMI in Combined Pop.	Re-Calculated Weight
10 years	False	24.73	24.73(0.05)
10 years	True	24.74	24.74(0.12)
5 years	False	24.73	24.73(0.06)
5 years	True	24.74	24.75(0.13)
1 years	False	24.73	24.73(0.05)
1 years	True	24.75	24.75(0.13)

Table A.4: *Average BMI of different Handedness for two waves combined (taking average of repeated observations).*

Year Diff.	True Total in Combined Pop.	Re-Calculated Weight
10 years	1683670369	1672198619(53989097)
5 years	1202588286	1170660997(35810800)
1 years	836435980	826201438(23902405)

Table A.5: *Total health insurance cost of the population for two waves combined.*

Year Diff.	True Total in Combined Pop.	Re-Calculated Weight
10 years	1672174403	1672041113(53859868)
5 years	1169898347	1169793518(35272097)
1 years	826243999	826279195(23815561)

Table A.6: *Total health insurance cost of the population for two waves combined (taking average of repeated observations).*

Year Diff.	Gender	True BMI in Combined Pop.	Re-Calculated Weight
10 years	Female	24.78	24.72(0.05)
10 years	Male	24.79	24.73(0.05)
5 years	Female	25.01	24.73(0.05)
5 years	Male	25.03	24.75(0.06)
1 years	Female	24.87	24.72(0.05)
1 years	Male	24.90	24.75(0.06)

Table A.7: *Average BMI of different genders for three waves combined.*

Year Diff	Gender	True BMI in Combined Pop.	Re-Calculated Weight
10 years	Female	24.72	24.72(0.05)
10 years	Male	24.73	24.73(0.05)
5 years	Female	24.72	24.72(0.06)
5 years	Male	24.74	24.74(0.06)
1 years	Female	24.72	24.72(0.06)
1 years	Male	24.75	24.75(0.06)

Table A.8: *Average BMI of different genders for three waves combined (taking average of repeated observations).*

Year Diff.	Left Hand	True BMI in Combined Pop.	Re-Calculated Weight
10 years	False	24.79	24.72(0.04)
10 years	True	24.80	24.74(0.10)
5 years	False	25.02	24.74(0.04)
5 years	True	25.00	24.73(0.10)
1 years	False	24.87	24.74(0.04)
1 years	True	24.90	24.76(0.11)

Table A.9: *Average BMI of different Handedness for three waves combined.*

Year Diff.	Left Hand	True BMI in Combined Pop.	Re-Calculated Weight
10 years	False	24.72	24.72(0.04)
10 years	True	24.74	24.74(0.10)
5 years	False	24.74	24.74(0.04)
5 years	True	24.71	24.72(0.10)
1 years	False	24.73	24.73(0.04)
1 years	True	24.76	24.76(0.11)

Table A.10: *Average BMI of different Handedness for three waves combined (taking average of repeated observations).*

Year Diff.	True Total in Combined Pop.	Re-Calculated Weight
10 years	2910255044	2899184057(87853760)
5 years	1739159691	1670110384(49184580)
1 years	929361229	909326782(24005421)

Table A.11: *Total health insurance cost of the population for three waves combined.*

Year Diff.	True Total in Combined Pop.	Re-Calculated Weight
10 years	2898759078	2899238740(87274044)
5 years	1668923450	1669099983(48337940)
1 years	908589289	908734196(23991361)

Table A.12: *Total health insurance cost of the population for three waves combined (taking average of repeated observations).*

Year Diff.	Gender	True BMI in Combined Pop.	Re-Calculated Weight
10 years	Female	24.77	24.71(0.05)
10 years	Male	24.78	24.72(0.05)
5 years	Female	25.07	24.73(0.05)
5 years	Male	25.08	24.74(0.06)
1 years	Female	24.93	24.72(0.05)
1 years	Male	24.96	24.75(0.05)

Table A.13: *Average BMI of different genders for four waves combined.*

Year Diff.	Gender	True BMI in Combined Pop.	Re-Calculated Weight
10 years	Female	24.71	24.71(0.05)
10 years	Male	24.72	24.72(0.05)
5 years	Female	24.72	24.72(0.05)
5 years	Male	24.74	24.74(0.05)
1 years	Female	24.72	24.72(0.05)
1 years	Male	24.75	24.74(0.05)

Table A.14: *Average BMI of different genders for four waves combined (taking average of repeated observations).*

Year Diff.	Left Hand	True BMI in Combined Pop.	Re-Calculated Weight
10 years	False	24.77	24.72(0.03)
10 years	True	24.79	24.73(0.09)
5 years	False	25.08	24.74(0.04)
5 years	True	25.07	24.73(0.09)
1 years	False	24.94	24.73(0.04)
1 years	True	24.82	24.77(0.10)

Table A.15: *Average BMI of different Handedness for four waves combined.*

Year Diff.	Left Hand	True BMI in Combined Pop.	Re-Calculated Weight
10 years	False	24.71	24.71(0.03)
10 years	True	24.72	24.72(0.09)
5 years	False	24.73	24.73(0.04)
5 years	True	24.72	24.72(0.09)
1 years	False	24.73	24.72(0.04)
1 years	True	24.77	24.77(0.10)

Table A.16: *Average BMI of different Handedness for four waves combined (taking average of repeated observations).*

Year Diff.	True Total in Combined Pop.	Re-Calculated Weight
10 years	4592888404	4564070374(117329171)
5 years	2280707765	2205172576(54365346)
1 years	1024366856	993026944(22166049)

Table A.17: *Total health insurance cost of the population for four waves combined.*

Year Diff.	True Total in Combined Pop.	Re-Calculated Weight
10 years	4563877658	4563193950(117358206)
5 years	2203800999	2204514732(53891527)
1 years	992602888	992445644(22161090)

Table A.18: *Total health insurance cost of the population for four waves combined (taking average of repeated observations).*

A.2 Estimations for combined waves with repeated observations

Year Diff.	Gender	True BMI in Combined Pop.	Original Weight
10 years	Female	24.73	24.73(0.06)
10 years	Male	24.73	24.74(0.06)
5 years	Female	24.73	24.73(0.06)
5 years	Male	24.75	24.75(0.06)
1 years	Female	24.72	24.73(0.06)
1 years	Male	24.75	24.75(0.06)

Table A.19: *Average BMI of different genders for two waves combined.*

Year Diff.	Left Hand	True BMI in Combined Pop.	Original Weight
10 years	False	24.73	24.73(0.05)
10 years	True	24.73	24.74(0.12)
5 years	False	24.74	24.74(0.05)
5 years	True	24.75	24.75(0.12)
1 years	False	24.73	24.73(0.05)
1 years	True	24.75	24.74(0.13)

Table A.20: *Average BMI of different Handedness for two waves combined.*

Year Diff.	True Total in Combined Pop.	Original Weight
10 years	1750523631	1750374923(56597867)
5 years	1613074920	1612775642(50401440)
1 years	1515901444	1515758026(46840777)

Table A.21: *Total health insurance cost of the population for two waves combined.*

Year Diff.	Gender	True BMI in Combined Pop.	Original Weight
10 years	Female	24.72	24.72(0.05)
10 years	Male	24.73	24.73(0.05)
5 years	Female	24.73	24.73(0.05)
5 years	Male	24.75	24.75(0.05)
1 years	Female	24.73	24.73(0.05)
1 years	Male	24.75	24.75(0.05)

Table A.22: *Average BMI of different genders for three waves combined.*

Year Diff.	Left Hand	True BMI in Combined Pop.	Original Weight
10 years	False	24.72	24.73(0.04)
10 years	True	24.73	24.73(0.10)
5 years	False	24.74	24.74(0.04)
5 years	True	24.72	24.73(0.10)
1 years	False	24.74	24.74(0.04)
1 years	True	24.75	24.75(0.10)

Table A.23: *Average BMI of different Handedness for three waves combined.*

Year Diff.	True Total in Combined Pop.	Original Weight
10 years	3068207196	3068596194(92438795)
5 years	2617564838	2617542506(74756050)
1 years	2308041163	2308379127(64473249)

Table A.24: *Total health insurance cost of the population for three waves combined.*

Year Diff.	Gender	True BMI in Combined Pop.	Original Weight
10 years	Female	24.70	24.70(0.05)
10 years	Male	24.71	24.71(0.05)
5 years	Female	24.73	24.73(0.05)
5 years	Male	24.74	24.74(0.05)
1 years	Female	24.73	24.73(0.05)
1 years	Male	24.76	24.76(0.05)

Table A.25: *Average BMI of different genders for four waves combined.*

Year Diff.	Left Hand	True BMI in Combined Pop.	Original Weight
10 years	False	24.70	24.70(0.03)
10 years	True	24.71	24.71(0.09)
5 years	False	24.74	24.74(0.03)
5 years	True	24.72	24.71(0.09)
1 years	False	24.74	24.74(0.04)
1 years	True	24.76	24.76(0.09)

Table A.26: *Average BMI of different Handedness for four waves combined.*

Year Diff.	True Total in Combined Pop.	Original Weight
10 years	4874145865	4873882986(124351821)
5 years	3710904861	3710433232(88336065)
1 years	3124020563	3124085827(72979161)

Table A.27: *Total health insurance cost of the population for four waves combined.*

A.3 Estimations for combined waves with calibration (without repeated observations)

Year Diff.	Gender	True BMI in Combined Pop.	Original Weight	Calibrated Weight
10 years	Female	24.77	24.73(0.06)	24.77(0.06)
10 years	Male	24.78	24.74(0.06)	24.78(0.06)
5 years	Female	24.91	24.73(0.06)	24.90(0.06)
5 years	Male	24.93	24.74(0.06)	24.92(0.06)
1 years	Female	24.80	24.72(0.06)	24.81(0.06)
1 years	Male	24.83	24.75(0.06)	24.83(0.06)

Table A.28: *[Calibration]Average BMI of different genders for two waves combined.*

Year Diff.	Left Hand	True BMI in Combined Pop.	Original Weight	Calibrated Weight
10 years	False	24.77	24.73(0.05)	24.78(0.03)
10 years	True	24.79	24.74(0.12)	24.79(0.12)
5 years	False	24.92	24.74(0.05)	24.91(0.03)
5 years	True	24.93	24.75(0.12)	24.93(0.13)
1 years	False	24.82	24.73(0.05)	24.82(0.04)
1 years	True	24.83	24.74(0.12)	24.83(0.15)

Table A.29: *[Calibration]Average BMI of different Handedness for two waves combined.*

Year Diff.	True Total in Combined Pop.	Original Weight	Calibrated Weight
10 years	1683670369	1750374923(56597553)	1683637570(3500523)
5 years	1202588286	1612775642(50398494)	1202582592(3129304)
1 years	836435980	1515758026(46833912)	836418014(2634775)

Table A.30: *[Calibration]Total health insurance cost of the population for two waves combined.*

Year Diff.	Gender	True BMI in Combined Pop.	Original Weight	Calibrated Weight
10 years	Female	24.78	24.72(0.05)	24.78(0.04)
10 years	Male	24.79	24.73(0.05)	24.79(0.04)
5 years	Female	25.00	24.73(0.05)	25.00(0.05)
5 years	Male	25.03	24.75(0.05)	25.02(0.05)
1 years	Female	24.87	24.73(0.05)	24.88(0.06)
1 years	Male	24.90	24.76(0.05)	24.91(0.06)

Table A.31: *[Calibration]Average BMI of different genders for three waves combined.*

Year Diff.	Left Hand	True BMI in Combined Pop.	Original Weight	Calibrated Weight
10 years	False	24.79	24.73(0.04)	24.79(0.02)
10 years	True	24.80	24.73(0.10)	24.80(0.10)
5 years	False	25.02	24.74(0.04)	25.02(0.03)
5 years	True	25.00	24.73(0.10)	24.99(0.11)
1 years	False	24.89	24.74(0.04)	24.89(0.03)
1 years	True	24.91	24.75(0.10)	24.92(0.14)

Table A.32: *[Calibration]Average BMI of different Handedness for three waves combined.*

Year Diff.	True Total in Combined Pop.	Original Weight	Calibrated Weight
10 years	2910255044	3068596194(92438006)	2921698386(4850348)
5 years	1739159691	2617542506(74749521)	1739277928(3657696)
1 years	929361229	2308379127(64453897)	929216181(2633067)

Table A.33: *[Calibration]Total health insurance cost of the population for three waves combined.*

Year Diff.	Gender	True BMI in Combined Pop.	Original Weight	Calibrated Weight
10 years	Female	24.77	24.70(0.04)	24.77(0.04)
10 years	Male	24.78	24.71(0.05)	24.78(0.04)
5 years	Female	25.07	24.73(0.04)	25.06(0.04)
5 years	Male	25.08	24.74(0.05)	25.08(0.04)
1 years	Female	24.93	24.73(0.04)	24.94(0.05)
1 years	Male	24.96	24.75(0.05)	24.97(0.05)

Table A.34: *[Calibration]Average BMI of different genders for four waves combined.*

Year Diff.	Left Hand	True BMI in Combined Pop.	Original Weight	Calibrated Weight
10 years	False	24.77	24.70(0.03)	24.77(0.02)
10 years	True	24.78	24.71(0.09)	24.78(0.08)
5 years	False	25.08	24.74(0.03)	25.07(0.02)
5 years	True	25.07	24.71(0.09)	25.06(0.10)
1 years	False	24.94	24.74(0.03)	24.95(0.03)
1 years	True	24.98	24.76(0.09)	25.00(0.13)

Table A.35: *[Calibration]Average BMI of different Handedness for four waves combined.*

Year Diff.	True Total in Combined Pop.	Original Weight	Calibrated Weight
10 years	4592888404	3068596194(92438006)	4610279090(7633913)
5 years	2280707765	2617542506(74749521)	2268628229(4493368)
1 years	1024366856	2308379127(64453897)	1024239077(2621887)

Table A.36: *[Calibration]Total health insurance cost of the population for four waves combined.*

A.4 Estimations for combined waves with calibration (with repeated observations)

Year Diff.	Gender	True BMI in Combined Pop.	Original Weight	Calibrated Weight
10 years	Female	24.73	24.73(0.06)	24.73(0.05)
10 years	Male	24.73	24.74(0.06)	24.75(0.05)
5 years	Female	24.73	24.73(0.06)	24.73(0.05)
5 years	Male	24.75	24.75(0.06)	24.75(0.05)
1 years	Female	24.72	24.72(0.06)	24.72(0.05)
1 years	Male	24.75	24.75(0.06)	24.75(0.06)

Table A.37: *[Calibration]Average BMI of different genders for two waves combined.*

Year Diff.	Left Hand	True BMI in Combined Pop.	Original Weight	Calibrated Weight
10 years	False	24.73	24.73(0.05)	24.73(0.03)
10 years	True	24.73	24.74(0.12)	24.74(0.12)
5 years	False	24.74	24.74(0.05)	24.74(0.03)
5 years	True	24.75	24.75(0.12)	24.75(0.12)
1 years	False	24.73	24.73(0.05)	24.73(0.03)
1 years	True	24.75	24.74(0.13)	24.74(0.12)

Table A.38: *[Calibration]Average BMI of different Handedness for two waves combined.*

Year Diff.	True Total in Combined Pop.	Original Weight	Calibrated Weight
10 years	1750523631	1750374923(56597867)	1750539652(3656215)
5 years	1613074920	1612775642(50401440)	1613229773(3775496)
1 years	1515901444	1515758026(46840777)	1516042594(3711189)

Table A.39: *[Calibration]Total health insurance cost of the population for two waves combined.*

Year Diff.	Gender	True BMI in Combined Pop.	Original Weight	Calibrated Weight
10 years	Female	24.72	24.72(0.05)	24.72(0.04)
10 years	Male	24.73	24.73(0.05)	24.73(0.04)
5 years	Female	24.73	24.73(0.05)	24.73(0.04)
5 years	Male	24.75	24.75(0.05)	24.75(0.04)
1 years	Female	24.73	24.73(0.05)	24.73(0.04)
1 years	Male	24.75	24.76(0.05)	24.75(0.04)

Table A.40: *[Calibration]Average BMI of different genders for three waves combined.*

Year Diff.	Left Hand	True BMI in Combined Pop.	Original Weight	Calibrated Weight
10 years	False	24.72	24.73(0.04)	24.72(0.02)
10 years	True	24.73	24.73(0.10)	24.73(0.10)
5 years	False	24.74	24.74(0.04)	24.74(0.02)
5 years	True	24.72	24.73(0.10)	24.72(0.10)
1 years	False	24.74	24.74(0.04)	24.74(0.02)
1 years	True	24.75	24.75(0.10)	24.76(0.10)

Table A.41: *[Calibration]Average BMI of different Handedness for three waves combined.*

Year Diff.	True Total in Combined Pop.	Original Weight	Calibrated Weight
10 years	3068207196	3068596194(92438006)	3068388002(5098809)
5 years	2617564838	2617542506(74749521)	2617790860(4779160)
1 years	2308041163	2308379127(64453897)	2307982205(4654487)

Table A.42: *[Calibration]Total health insurance cost of the population for three waves combined.*

Year Diff.	Gender	True BMI in Combined Pop.	Original Weight	Calibrated Weight
10 years	Female	24.70	24.70(0.05)	24.70(0.04)
10 years	Male	24.71	24.71(0.05)	24.71(0.04)
5 years	Female	24.73	24.73(0.05)	24.73(0.04)
5 years	Male	24.74	24.74(0.05)	24.74(0.04)
1 years	Female	24.73	24.73(0.05)	24.73(0.04)
1 years	Male	24.76	24.76(0.05)	24.76(0.04)

Table A.43: *[Calibration]Average BMI of different genders for four waves combined.*

Year Diff.	Left Hand	True BMI in Combined Pop.	Original Weight	Calibrated Weight
10 years	False	24.70	24.70(0.03)	24.70(0.02)
10 years	True	24.71	24.71(0.09)	24.71(0.08)
5 years	False	24.74	24.74(0.03)	24.74(0.02)
5 years	True	24.72	24.71(0.09)	24.72(0.08)
1 years	False	24.74	24.74(0.04)	24.74(0.02)
1 years	True	24.76	24.76(0.09)	24.76(0.09)

Table A.44: *[Calibration]Average BMI of different Handedness for four waves combined.*

Year Diff.	True Total in Combined Pop.	Original Weight	Calibrated Weight
10 years	4874145865	4873882986(124351821)	4874313634(7874745)
5 years	3710904861	3710433232(88336065)	3710713602(5701913)
1 years	3124020563	3124085827(72979161)	3123896637(5511294)

Table A.45: *[Calibration]Total health insurance cost of the population for four waves combined.*

Bibliography

A Arcos, D Molina, MG Ranalli, and M del Mar Rueda (2015). Frames2: A Package for Estimation in Dual Frame Surveys. *The R Journal* **7**(1), 54.

T Chen, J Clark, M Riddles, L Mohadjer, and T Fakhouri (2020). National Health and Nutrition Examination Survey, 2015-2018: Sample design and estimation procedures. *National Center for Health Statistics* **2**(184), 8.

WG Cochran, (1977). *Sampling Techniques*. 3rd ed. Hohn Wiley and Sons.

M Cohen, (1997). The Bayesian bootstrap and multiple imputation for unequal probability sample designs. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 635–638.

Centers for Disease Control and Prevention (2020). National Diabetes Statistics Report, 2020. *Centers for Disease Control and Prevention*, 2.

Q Dong, M Elliott, and T Raghunathan (2014a). A nonparametric method to generate synthetic populations to adjust for complex sampling design features. *Survey Methodology* **40**(1), 29–46.

Dong, Q, MR Elliott, and TE Raghunathan (2014b). Combining information from multiple complex surveys. *Statistics Canada* **12**(1), 347–354.

Elliott, MR, TE Raghunathan, and N Schenker (2018). Combining Estimates from Multiple Surveys. *Wiley StatsRef: Statistics Reference Online*, 1–10.

Fuller, WA (2009). *Sampling Statistics*. John Wiley and Sons.

Hartley, HO (1962). Multiple frame surveys. *Proceedings of the American Statistical Association, Social Statistics Sections*, 203–206.

Health Statistics (NCHS), NC for (2017). About the National Health and Nutrition Examination Survey. **Published only on the Internet: cite as**, https://www.cdc.gov/nchs/nhanes/about–nhanes.htm.

Health Statistics (NCHS), NC for (2020). National Health Interview Survey. **Published only on the Internet: cite as**, https://www.cdc.gov/nchs/nhis/participants/aboutnhis.htm.

Lo, A (1986). Bayesian statistical inference for sampling a finite population. *Annals of Statistics* **14**(3), 1226–1233.

Lohr, SL (2011). Alternative survey sample designs: Sampling with multiple overlapping frames. *Survey Methodology* **37**(2), 197–213.

McCarthy, N (2020). The Countries With The Most Left-Handed People. **Published only on the Internet: cite as**, https://www.statista.com/.

Mecatti, F (2007). A single frame multiplicity estimator for multiple frame surveys. *Survey Methodology* **33**(2), 151–157.

Ornstein, M (2013). *A Companion to Survey Research*. SAGE Publications Ltd.

Ranalli, MG, A Arcos, M del Mar Rueda, and A Teodoro (2016). Calibration estimation in dual-frame surveys. *Statistical Methods and Applications* **25**(12), 321–349.

Research, UCFHP (2012). California Health Interview Survey. **Published only on the Internet: cite as**, https://healthpolicy.ucla.edu/chis/Pages/default.aspx.

Rubin, D (1981). The Bayesian bootstrap. *The Annals of Statistics* **9**(1), 130–134.

Särndal, CE (2007). The calibration approach in survey theory and practice. *Survey Methodology* **33**(2), 99–119.

Särndal, CE, B Swensson, and J Wretman (1992). *Model Assisted Survey Sampling*. 1st ed. Springer Science+Business Media.

Skinner, CJ and JNK Rao (1996). Estimation in dual frame surveys with complex designs. *Journal of the American Statistical Association* **91**(433), 349–356.

Talabani, A (2021). Health Insurance in the USA: What's the Cost? **Published only on the Internet: cite as**, https://www.william-russell.com/blog/health-insurance-usa–cost/.

Thomas, S and B Wannell (2009). Combining cycles of the Canadian Community Health Survey. *Statistics Canada: Health Reports* **20**(1), 53–58.

Vehovar, V, V Toepoel, and S Steinmetz (2016). *The SAGE Handbook of survey Methodology*. SAGE Publications Ltd.