$u_{ij} \sim N(0,10)$ denotes the random cluster effect,
$a_i \sim uniform(2,52)$ is the number of clusters within stratum $i$,
$b_{ij} \sim uniform(10,20)$ is the number of units within cluster $j$ of stratum $i$.

The population for the simulation study has 61,324 subjects. We draw a stratified clustering sampling with unequal probabilities of selection. Specifically, we select two clusters from each stratum with probabilities proportional to cluster size (PPS). Within each selected cluster, we select 1/5 of the population. Thus, the probability that unit $ijk$ is selected is

$$\Pr(\text{cluster ij is selected}) * \Pr(\text{unit ijk is selected} | \text{cluster ij is selected}) \propto b_{ij}.$$

The weights of the sample are calculated by inversing the selection probabilities. Since the number of clusters and units are random, the complex sample size is slightly different across replications, which is approximately 770.

For stratified clustering sample, we generate $L = 100$ synthetic populations following the exact two steps proposed in Section 3.3. Each synthetic population is about 10,000 times ($F = 100, k = 100$) as large as the sample.

The estimands of interest are the marginal means for $x_1$ and $x_2$ and the regression coefficients of $x_1$ on $x_2$. We perform ordinary linear regression analyses to obtain the estimates for the regression coefficients and the standard errors based on the actual data and each synthetic population. The estimates from the synthetic populations are then combined using the synthetic data combining rule developed by Raghunathan *et al.* (2003). We obtain the 95% confidence intervals for the statistics of interest from the synthetic populations and from the actual data.

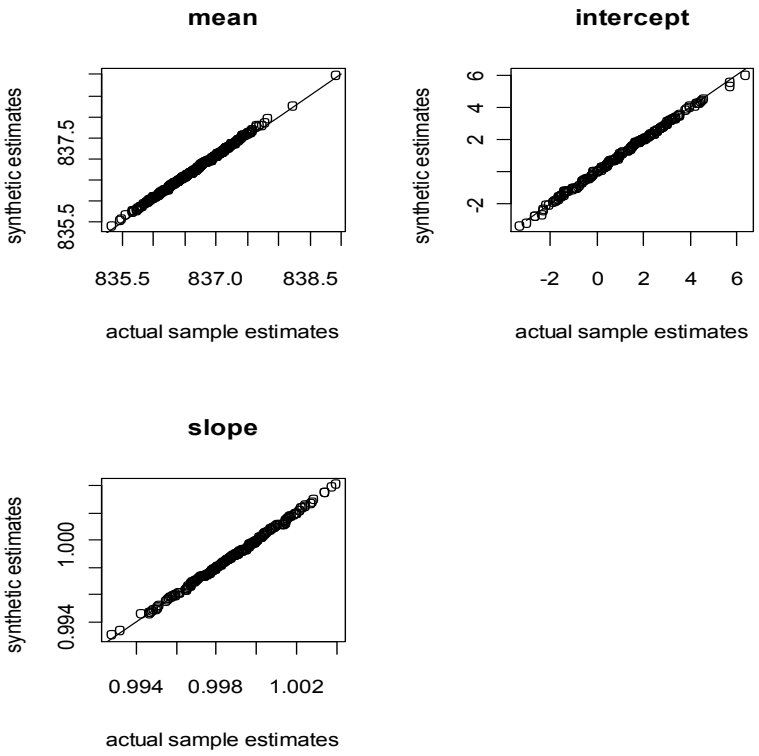
We repeat the process 200 times to evaluate the repeated sampling properties of the nonparametric method. Specifically, we draw 200 complex samples from the population and generate 100 synthetic populations for each sample and analyze them as simple random samples.

Figure 3.3 displays the scatter plot of the 200 pairs of estimated mean, intercept and slope from the actual samples and the corresponding synthetic populations along with a 45-degee line. The sampling distributions of the actual sample and synthetic population estimates are practically the same. Table 3.1 compares the inference of descriptive and analytic statistics from the actual data and the multiple synthetic populations. The point estimates for both types of statistics are identical across the imputed and actual data. The average of standard errors across 200 replications and the standard deviation of the point estimates across 200 replications for mean, intercept and slope are identical. The 95% confidence interval coverage rate for all three statistics is nominal. As we can see, the proposed nonparametric method focuses on the design variables and does not assume models to capture the relationships among the variables. This suggests it is robust for model misspecification, especially in the situations that the relationships among the variables are complicated or the sample size is too small to correctly fit the model.

Table 3.1 Descriptive and Analytic Statistics Estimated from the Actual Data and the Synthetic Populations in the Simulation Evaluation of the Nonparametric Method

| Type | Actual Data | | | | Synthetic Populations | | | | No. of Est. |
|---|---|---|---|---|---|---|---|---|---|
| | Estimate | SE | SD | Coverage (%) | Estimate | SE | SD | Coverage (%) | |
| Mean | 836.701 | 0.461 | 0.491 | 93.25 | 836.793 | 0.476 | 0.493 | 93.75 | 400 |
| Intercept | 1.013 | 1.768 | 1.848 | 93.5 | 1.014 | 1.775 | 1.846 | 92.5 | 200 |
| Slope | 0.999 | 0.002 | 0.002 | 92.0 | 0.999 | 0.002 | 0.002 | 91.5 | 200 |

Figure 3.3 Scatter plot of the descriptive and analytic statistics from the actual and synthetic populations




## 3.6 Applications

In Section 3.6.1 and 3.6.2, we use genuine data (2006 NHIS, MEPS and BRFSS) to evaluate the performance of the nonparametric method under two common sampling designs, stratified clustering sampling and stratified sampling. Then in Section 3.6.3, we produce the combined estimates of people's heath insurance coverage rates by using the information from all three surveys.

### 3.6.1 Estimation of Health Insurance Coverage from the NHIS and MEPS

Most survey data collected under area probability sampling are prepared and released in the format of a stratified clustering sample with weights. For example, both the 2006 NHIS and MEPS data are multistage probability sample that incorporates

stratification, clustering and oversampling of some subpopulations (e.g., Black, Hispanic, and Asian in later years). In this simulation study, we will use the nonparametric method to adjust for the stratified clustering sampling used by the 2006 NHIS and MEPS and generate synthetic populations that can be analyzed as simple random samples. Then we evaluate the method by comparing the estimates of the health insurance coverage rate for the whole population and some subdomains obtained from the synthetic populations to those obtained from the actual data.

Both NHIS and MEPS ask respondents whether they are covered by any health insurance and if so what type health insurance they are using. So there are three health insurance statuses, covered by any private insurance, covered by government insurance and uninsured. We are also interested in estimating the health insurance coverage rates in sub population. So in this study, we choose six demographic variables: gender, race, census region, education level, age (categorical), and income level (categorical) and the subdomains are created by one demographic variable or the combination of 2 or 3 demographic variables.

We delete the cases with item-missing values and focus on our simulation on the complete cases. This results in 20,147 and 20,893 cases in the NHIS and MEPS data respectively. We recode the variables into the same categories. The coding of the variables is shown in 3.2 below.

Table 3.2 Variables and Response Categories for the 2006 NHIS and MEPS

| Variables of Interest | Coding |
|---|---|
| Age | 1: [18,24] 2: [25,34]<br>3: [35,44] 4: [45,54]<br>5: [55,64] 6: >=65 |
| Census Region | 1: Northeast 2: Midwest<br>3: South 4: West |
| Education | 1: Less than high school 2: High school<br>3: Some college 4: College |
| Gender | 1: Male 2: Female |
| Health Insurance Coverage | 1: Any Private Insurance<br>2: Public Insurance<br>3: Uninsured |
| Income | 1: (0,10000) 2: [10000,15000)<br>3: [15000,20000) 4: [20000,25000)<br>5: [25000,35000) 6: [35000,75000)<br>7: >=75000 |
| Race | 1: Hispanic<br>2: Non-Hispanic White<br>3: Non-Hispanic Black<br>4: Non-Hispanic All other race groups |

Using the nonparametric method, we generate 200 synthetic populations for each survey. Specifically, we generate $B = 200$ BB samples and for each BB sample, we generate $F = 10$ FPBB of size 5n ($k = 5$). Thus, each synthetic population is 50 times as big as the actual sample (1,007,350 for NHIS, 1,044,650 for MEPS). Each synthetic population is analyzed as a simple random sample and the estimates are combined using the combining rule for synthetic data (Raghunathan, *et al.* 2003).

The results are summarized in Table 3.3, from which we see the estimates of people's health insurance coverage rates from the synthetic data are almost identical to those obtained from the actual data after complex sampling design features are accounted for. For the NHIS, the variance estimates of the health insurance coverage rates across almost all domains from the synthetic populations are about 30% larger than these from the actual data and less than 10% larger for the MEPS. This implies that this

nonparametric method adjusts for the complex sampling design features with some information loss.

Comparing the results with the ones obtained using the model-based method in Chapter 2, we notice that the nonparametric method does a better job to simulate the actual data for small domains. For example, for the smallest subdomain we consider, the Non-Hispanic white people with Income between 25,000 and 35,000 per year, both the point estimates and the variance estimates from the model-based method are quite different from these from the actual data. For the whole population and other sub domains, the model-based method has less information loss compared to the nonparametric method. This is consistent with our hypothesis. When the model fits the data well, the model-based method is more efficient than the nonparametric method. However, when the assumed model does not fit the data well, the model-based method may produce invalid inference. In such situation, the nonparametric method is robust to model misspecfication.

Table 3.3 Estimates from Actual Data and from the Synthetic Populations for the 2006 NHIS and MEPS

| Domain | Actual Data (Complex Design) | | | Synthetic Populations | |
|---|---|---|---|---|---|
| | Types | NHIS | MEPS | NHIS | MEPS |
| Whole Population | | | Proportion | | |
| | Private | 0.746 | 0.735 | 0.746 | 0.736 |
| | Public | 0.075 | 0.133 | 0.075 | 0.132 |
| | Uninsured | 0.179 | 0.132 | 0.179 | 0.132 |
| | | | Variance | | |
| | Private | 2.46E-05 | 2.78E-05 | 3.15E-05 | 3.31E-05 |
| | Public | 6.29E-06 | 1.44E-05 | 8.06E-06 | 1.59E-05 |
| | Uninsured | 1.84E-05 | 1.41E-05 | 2.29E-05 | 1.71E-05 |
| Male | | | Proportion | | |
| | Private | 0.740 | 0.735 | 0.740 | 0.736 |
| | Public | 0.060 | 0.101 | 0.060 | 0.100 |
| | Uninsured | 0.200 | 0.164 | 0.200 | 0.164 |
| | | | Variance | | |
| | Private | 3.32E-05 | 3.87E-05 | 3.93E-05 | 4.31E-05 |
| | Public | 6.82E-06 | 1.53E-05 | 8.81E-06 | 1.63E-05 |
| | Uninsured | 2.94E-05 | 2.64E-05 | 3.29E-05 | 2.79E-05 |
| Hispanic | | | Proportion | | |
| | Private | 0.494 | 0.506 | 0.495 | 0.508 |
| | Public | 0.096 | 0.161 | 0.097 | 0.158 |
| | Uninsured | 0.410 | 0.334 | 0.409 | 0.334 |
| | | | Variance | | |
| | Private | 1.24E-04 | 1.73E-04 | 1.94E-04 | 1.97E-04 |
| | Public | 2.57E-05 | 8.03E-05 | 3.88E-05 | 8.43E-05 |
| | Uninsured | 1.23E-04 | 1.19E-04 | 1.90E-04 | 1.61E-04 |
| Non-Hispanic White | Proportion | | | | |
| | Private | 0.805 | 0.788 | 0.804 | 0.788 |
| | Public | 0.062 | 0.116 | 0.062 | 0.116 |
| | Uninsured | 0.134 | 0.096 | 0.134 | 0.096 |
| | | | Variance | | |
| | Private | 2.99E-05 | 3.35E-05 | 3.79E-05 | 4.12E-05 |
| | Public | 8.20E-06 | 1.81E-05 | 1.04E-05 | 2.00E-05 |
| | Uninsured | 2.02E-05 | 1.51E-05 | 2.35E-05 | 1.80E-05 |
| Non-Hispanic White & Income [25,000, 35,000) | | | Proportion | | |
| | Private | 0.827 | 0.813 | 0.827 | 0.814 |
| | Public | 0.039 | 0.079 | 0.039 | 0.079 |
| | Uninsured | 0.134 | 0.108 | 0.134 | 0.107 |
| | | | Variance | | |
| | Private | 1.00E-04 | 1.39E-04 | 1.48E-04 | 1.63E-04 |
| | Public | 2.82E-05 | 6.31E-05 | 3.86E-05 | 7.28E-05 |
| | Uninsured | 7.24E-05 | 8.92E-05 | 9.55E-05 | 1.11E-04 |

### 3.6.2 Estimation of Health Insurance Coverage from the BRFSS

Another commonly used data collection method is the random digit dialing (RDD) telephone survey. The RDD samples are usually not clustered. For example, the 2006 BRFSS is collected within each state independently and the telephone numbers are randomly selected within state.

In the second simulation study, we will use the nonparametric method to adjust for the 1-stage stratified sampling used by the 2006 BRFSS and generate synthetic populations that can be analyzed as simple random samples. We are still interested in estimating the health insurance coverage rate for the whole population and some subdomains. However, the BRFSS only asks whether one is insured or not. There is no information about the type of insurance that one uses. So we only calculate the proportion of respondents who are not covered by any insurance. The demographic variables and the coding are the same as in Table 3.2.

We delete the cases with item missing values and focus on our simulation on the complete cases. There are 294,559 complete cases in the 2006 BRFSS data.

Using the proposed method for the 1-stage stratified sampling, we generate 200 synthetic populations for the BRFSS data. Specifically, we generate $B = 200$ BB samples and for each BB sample, we generate $F = 10$ FPBB of size 5n ($k = 5$). Thus, each synthetic population is 50 times as big as the actual sample (around 14,727,950). Each synthetic population is analyzed as a simple random sample and the estimates are combined using the combining rule for synthetic data (Raghunathan, *et al.* 2003).

The results are summarized in Table 3.4, from which we see the estimates of the proportion of uninsured people obtained from the synthetic data are almost identical to those obtained from the actual data after complex sampling design features are adjusted, so are the variance estimates. This implies that this nonparametric method adjusts for the complex sampling design features with little information loss.

Table 3.4 Estimates from Actual Data and from Synthetic Populations for the 2006 BRFSS

| Domain | Actual Data (Complex Design Features) | Synthetic Populations |
|---|---|---|
| Whole population | Proportion | |
| | 0.154 | 0.153 |
| | Variance | |
| | 3.32E-06 | 3.44E-06 |
| Male | Proportion | |
| | 0.167 | 0.167 |
| | Variance | |
| | 8.88E-06 | 8.92E-06 |
| Hispanic | Proportion | |
| | 0.371 | 0.370 |
| | Variance | |
| | 7.18E-05 | 6.72E-05 |
| Non-Hispanic White | Proportion | |
| | 0.106 | 0.106 |
| | Variance | |
| | 2.15E-06 | 2.33E-06 |
| Non-Hispanic White & Income [25,000, 35,000) | Proportion | |
| | 0.173 | 0.173 |
| | Variance | |
| | 2.78E-05 | 3.10E-05 |

### 3.6.3 Combined Estimates of Health Insurance Coverage from the NHIS, MEPS and BRFSS

After we generate the synthetic populations for the three surveys, we produce the combined estimates of people's health insurance coverage rates using the combining survey method in Chapter 2. Since all three surveys have the information about whether people have insurance or not, we can combine the NHIS, BRFSS and MEPS to estimate

the proportion of uninsured people. However, the BRFSS does not ask people what insurance they have, private or public. For these proportions, we can only combine the NHIS and MEPS. The results are summarized in Table 3.5. The variance estimates for the combined estimator are much smaller than the ones obtained from the actual data. Specifically, the precision of the estimates obtained from the NHIS is increased by 43% on average, with the largest increase of 98% obtained by combining the NHIS and MEPS. The gains in precision for the MEPS are even more. The average increase in precision for the MEPS is 101%, with the largest increase being 202%. The precision is further increased when we combine all three surveys. For example, for the proportion of people who have no coverage, on average the precision is increased by 5 times for the NHIS, 0.5 times for the BRFSS and 4.2 times for the MEPS. This implies gains in precision by making use of the information from multiple surveys can be significant, and the more information we combine, the larger the gains are in precision.

Table 3.5 Estimates from Individual Surveys and the Combined Estimates before Missing Information is imputed for the 2006 NHIS, MEPS and BRFSS

| Domain | Actual Data (Complex Design) | | | | Combined Estimates | |
|---|---|---|---|---|---|---|
| | Types | NHIS | BRFSS | MEPS | NHIS and MEPS | NHIS, BRFSS and MEPS |
| Whole Population | Proportion | | | | | |
| | Private | 0.746 | | 0.735 | 0.741 | |
| | Public | 0.075 | | 0.133 | 0.094 | |
| | Uninsured | 0.179 | 0.154 | 0.132 | 0.152 | 0.153 |
| | Variance | | | | | |
| | Private | 2.46E-05 | | 2.78E-05 | 1.61E-05 | |
| | Public | 6.29E-06 | | 1.44E-05 | 5.35E-06 | |
| | Uninsured | 1.84E-05 | 3.32E-06 | 1.41E-05 | 9.80E-06 | 2.55E-06 |
| Male | Proportion | | | | | |
| | Private | 0.740 | | 0.735 | 0.738 | |
| | Public | 0.060 | | 0.101 | 0.074 | |
| | Without | 0.200 | 0.167 | 0.164 | 0.181 | 0.172 |
| | Variance | | | | | |
| | Private | 3.32E-05 | | 3.87E-05 | 2.06E-05 | |
| | Public | 6.82E-06 | | 1.53E-05 | 5.72E-06 | |
| | Uninsured | 2.94E-05 | 8.88E-06 | 2.64E-05 | 1.51E-05 | 5.61E-06 |
| Hispanic | Proportion | | | | | |
| | Private | 0.494 | | 0.506 | 0.5014 | |
| | Public | 0.096 | | 0.161 | 0.1157 | |
| | Without | 0.410 | 0.371 | 0.334 | 0.3684 | 0.3689 |
| | Variance | | | | | |
| | Private | 1.24E-04 | | 1.73E-04 | 9.76E-05 | |
| | Public | 2.57E-05 | | 8.03E-05 | 2.66E-05 | |
| | Uninsured | 1.23E-04 | 7.18E-05 | 1.19E-04 | 8.71E-05 | 3.79E-05 |
| Non-Hispanic White | Proportion | | | | | |
| | Private | 0.805 | | 0.788 | 0.796 | |
| | Public | 0.062 | | 0.116 | 0.081 | |
| | Without | 0.134 | 0.1059 | 0.096 | 0.113 | 0.107 |
| | Variance | | | | | |
| | Private | 2.99E-05 | | 3.35E-05 | 1.97E-05 | |
| | Public | 8.20E-06 | | 1.81E-05 | 6.86E-06 | |
| | Uninsured | 2.02E-05 | 2.15E-06 | 1.51E-05 | 1.02E-05 | 1.90E-06 |
| Non-Hispanic White & Income [25,000, 35,000) | Proportion | | | | | |
| | Private | 0.827 | | 0.813 | 0.821 | |
| | Public | 0.039 | | 0.079 | 0.053 | |
| | Without | 0.134 | 0.173 | 0.108 | 0.122 | 0.154 |
| | Variance | | | | | |
| | Private | 1.0E-04 | | 1.39E-04 | 7.74E-05 | |
| | Public | 2.82E-05 | | 6.31E-05 | 2.52E-05 | |
| | Uninsured | 7.24E-05 | 2.78E-05 | 8.92E-05 | 5.14E-05 | 1.93E-05 |

## 3.7 Discussion

In this chapter, we propose and evaluate a nonparametric method to generate synthetic populations as a counter part of the model-based method in Chapter 2. This method adjusts for the complex sampling design features without assuming any models to the observed data so it is robust to model-misspecification. Also, unlike the model-based method that needs to develop separate imputation models for different variables of interest, the nonparametric method only uses the design variables to generate synthetic populations and thus is not variable-specific.

In the simulation studies where we generate synthetic populations for the 2006 NHIS, BRFSS and MEPS, the estimates of people's health insurance coverage rates and their variance estimates from the synthetic population and from the actual data are very similar. The nonparametric method does not lose much information when imputing the unobserved units and/or clusters. When compared to the model-based method, the nonparametric method outperforms the model-based method for small domains where the assumed model does not fit the data accurately. For the domains where the imputation model is good, both the nonparametric method and the model-based method produce synthetic populations that simulate the actual data well. The model-based method preserves more information from the actual data compared to the nonparametric method.

Beside the fact that the nonparametric method is robust to model misspecification, another advantage is that the nonparametric method only uses the design variables such as stratum, cluster and weight to impute the unobserved part of the population. Unlike the

model-based method, it does not need to model the complicated relationships among the variables of interest, which becomes impossible if there are item missing values in the actual data. The synthetic populations generated by the nonparametric method still preserve the item missing values in the actual data. This potentially fills in a gap in the multiple imputation area that existing imputation methods typically ignore the complex sampling design features in the data and impute the missing values as if they are simple random samples.

A third practical advantage of the nonparametric method is that it is easier to be implemented into the existing statistical software (R, SAS, etc) because it focuses on the design variables and thus need not to develop strategies for various types of variables and data structures.

In the combining survey framework, as we see in the application, combining information from multiple surveys increase the precision of the estimates. Also, when we combine all three surveys to estimate the proportion of people who have no health insurance, the combined estimate is even more precise than the estimates when we only combine the NHIS and MEPS. However, since the BRFSS does not have the information about the types of health insurance, we cannot use the BRFSS data when estimating the proportions of people covered by private or public insurance.

As we mentioned earlier, since we cannot accurately estimate the probabilities of selecting clusters based on the information that survey agencies released to public, we only use an approximated Bayesian bootstrap method to adjust for stratification and clustering. To ensure the replicate weights to be positive, the Bayesian bootstrap method

can only produce fewer clusters within strata than in the actual data. Future research should focus on evaluating the FPBB method in imputing the unobserved clusters.

# CHAPTER 4

# COMBINING INFORMATION FROM MULTIPLE COMPLEX SURVEYS WHEN THERE IS MISSING INFORMATION IN AT LEAST ONE SURVEY

As there are more and more data collected from multiple sources for the same underlying population, there is an increasing demand to make use of all of the information contained in these data sets to produce improved inference. Two statistical techniques have been developed and investigated during the past several decades. The first is data fusion/linkage through statistical matching. The main objective of data fusion is to integrate multiple data collecting different levels of variables into a single complete data set that covers a broader range of variables. The second statistical technique is the combining survey method, which concentrates on reducing the survey errors by making use of the information from multiple surveys. In Chapter 2, we proposed a method that combines information from multiple surveys of different sampling designs. We also provided a simplified simulation study to evaluate the new method in which there is no missing information in both surveys that we combined. However, quite often in practice

multiple surveys using different sampling designs or modes of data collection cover various levels of variables and we need to merge the datasets from these surveys to produce a data file with all the variables of interest. Neither the existing data fusion methods nor the existing combining survey methods can achieve this goal. In this chapter, we extend the new combining survey method proposed in Chapter 2 to this general situation and apply this method to estimate health insurance coverage rates by combining the 2006 National Health Interview Survey (NHIS), the Behavioral Risk Factor Surveillance System (BRFSS) and the Medical Expenditure Panel Survey (MEPS). The combined estimates are shown to be more precise than the estimates from individual surveys.

## 4.1 Introduction

In marketing research and public opinion area, we sometimes gather data from multiple sources for the same group of people. For example, some of the data may be collected by survey agencies. Administrative data that are from external sources such as census data may also be available. Linking these data and analyzing them collectively broaden the scope of our analysis and help us understand the targeting population better and deeper.

One concern of combining data from multiple sources is the discrepancy in data quality. For example, newer data collection methods such as smart phone surveys or web surveys are usually cheaper than the traditional data collection methods such as mail surveys (Cobanoglu, Warae and Morec 2001; Couper 2000; Couper, Traugott and Lamias 2001). However, data obtained from web surveys may have larger nonsampling errors,

such as noncoverage error and nonresponse error. Many methods have been developed to combine data from two surveys to reduce the nonsampling errors in individual surveys. A thorough literature review of the existing combining survey methods can be found in Chapter 2.

Another concern is that the data from multiple sources may collect various levels of variables and thus have different structures. For example, assume a company has one million customers. For each customer, 20 variables are stored in the customer database. Furthermore, a market survey interviews 1000 customers of the company and asks questions corresponding to 50 variables, including 10 variables that overlap those in the customer database. The question is how to make adjustment on the data structure of both the customer database and the market survey data to create a virtual survey with each customer. The most straightforward procedure to perform this kind of data fusion is statistical matching.

Suppose there are two data sets, A and B. Suppose further that A contains variables $X$ and $Y$, whereas B contains variables $X$ and $Z$. $X$, $Y$ and $Z$ could be vector-valued variables. Statistical matching is proposed to combine these two files to obtain at least one file containing $X$, $Y$ and $Z$. In contrast to record linkage or exact matching (Fellegi and Sunter 1969), the two samples to be combined are not assumed to have records for the same entities. Statistical matching assumes the two samples have little or no overlap, and hence records for similar entities are combined instead of records for the same entities.

There are two basic types of statistical matching, constrained and unconstrained (Radner et al. 1980). Constrained statistical matching uses all records in the two samples and preserves the marginal distribution for Y and Z (e.g., see Barr and Turner 1980). Unconstrained matching doesn't have these requirements (Okner 1972).

The inherent assumption in statistical matching is that the random vector Y given X is independent of the random vector Z given X. This is a rather strong assumption that limits the application of statistical matching techniques. For example, statistical matching may not in general be an acceptable procedure for estimating relationships between Y and Z, or for any type of multivariate analysis involving both Y and Z (Rodgers 1984). Only two procedures in the existing literatures can assess the effect of alternative assumptions of the inestimable covariance between Y and Z, the file concatenation method proposed by Rubin (Rubin 1986) and the variance-covariance method proposed by Kadane and supplemented by Moriarity (Kadane 1978; Moriarity and Scheuren 2001).

Another inherent assumption for the statistical matching methods is all samples to be matched are obtained using simple random sampling (SRS) that results in independent and identically distributed (IID) observations. Current methods for stratified matching cannot accommodate complex sampling design features such as stratification, clustering and weighting.

Also, statistical matching has to assume that the bridging variables X in all the samples to be combined do not contain nonsampling errors (Cohen 1991; Ingram et al. 2000). None of the existing methods adjusts for the different survey error properties for either the bridging variables X or the variables of interest Y and Z. If the nonsampling

errors are not taken into account before the samples are combined, the matched data may result in invalid inferences.

In summary, the current combining survey methods concentrate on reducing the survey errors. Statistical matching is proposed to merge datasets from multiple surveys to achieve a complete data file when there is no existing data file containing all variables of interest. Both combining survey methods and statistical matching methods have difficulties when the data to be combined are not comparable in sampling design or nonsampling errors.

In Chapter 2, we propose a new combining survey method that adjusts for the different sampling designs of multiple surveys. We also provided a simplified simulation study in which there is no missing information in both surveys that we combined. This lay out the foundation for the unified combing survey framework that achieves the goals of both the combining survey methods and statistical matching methods. It converts any combining survey situation into a missing data problem that can be handled using the well-established missing data theories. In this chapter, we provide the specific steps of the new combining survey method, develop its theoretical foundation, conduct a simulation study and consider an application to evaluate this new method.

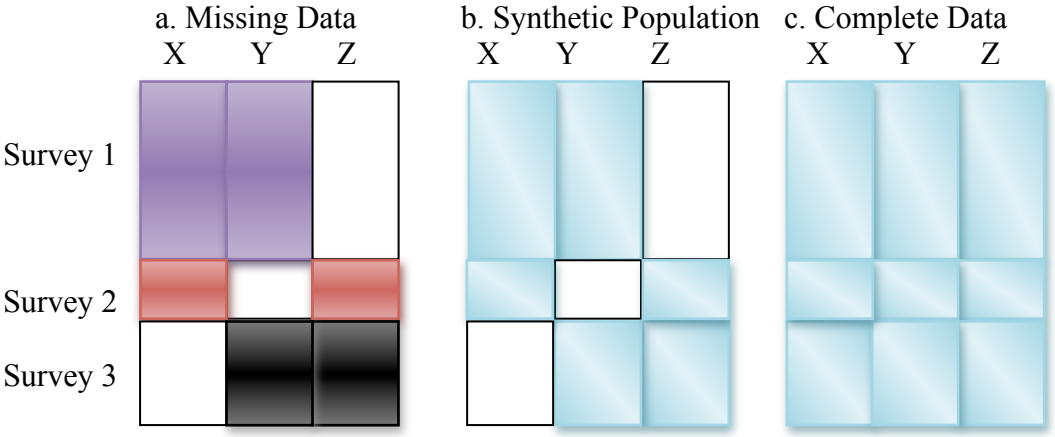
This chapter is organized as follows: Section 4.2 provides a motivating example that covers a broad range of combining survey situations. Section 4.3 provides the proposed steps to produce complete datasets and procedures for combining inference from the complete datasets. In Section 4.4, we derive the combining rule that is appropriate for the two-stage imputation procedure. We also conduct a simulation study

to show that the inferences based on the complete data sets after the missing information is filled and the actual data are very similar. Section 4.5 applies this new method to estimate health insurance coverage rates by combining the data collected from the 2006 BRFSS, NHIS and MEPS. It is shown that the combined estimates are more precise than the estimates from individual surveys. Concluding remarks and discussions are provided in Section 4.6.

### 4.2 An Motivating Example

In this section, we first present a motivating example that covers a large variety of combining survey situations. Figure 4.1.a describes a general combing survey situation, in which the rows represent three surveys that have the same underlying population, i.e., $\mathcal{P} = \{\mathrm{X_i}, \mathrm{Y_i}, \mathrm{Z_i}, i = 1,2, \dots, \mathrm{N}\}$. Suppose that the actual observed data for each survey is $\mathcal{D}^{(1)} = \{\mathrm{X_i}, \mathrm{Y_i}, i = 1,2, \dots, n_1\}$ (Survey 1), $\mathcal{D}^{(2)} = \{\mathrm{X_i}, \mathrm{Z_i}, i = 1,2, \dots, n_2\}$ (Survey 2) and $\mathcal{D}^{(3)} = \{\mathrm{Y_i}, \mathrm{Z_i}, i = 1,2, \dots, n_3\}$ (Survey 3), respectively. The blank cells in the figure denote the missing variables and the shaded cells represent the observed variables. If we are interested in estimating a population quantity that is related to $X, Y$ and $Z$, we have to combine data from all three surveys. We assume that background variables, including design and administrative records, are available in all three surveys and are observed for the whole population. To simplify the notation, the additional background variables are excluded from the figures and the related formulas in the following sections.

Figure 4.1 Data structure in different phases of combining surveys: raw data from surveys with missing variables, synthetic populations after sampling designs are adjusted for, complete data after the missing variables are filled in




## 4.3 Methods

### 4.3.1 Creating the Complete Datasets and Combining Multiple Surveys

For the situation in Figure 4.1, we can combine the three surveys using the following steps illustrated in the flowchart in Figure 4.2:

Step 1: For each survey, generate $L$ synthetic populations that can be analyzed as simple random samples. Then stack the synthetic populations from the 3 surveys so that we have $L$ data sets with a structure as in Figure 4.1.b. We use the nonparametric method (Chapter 3) to generate synthetic populations.

Step 2: Once the sampling design features are adjusted for, the next step is to fill in the missing information. We could multiple impute the $L$ data sets with a structure in Figure 4.1.b so that we have $L * M$ complete data sets as in Figure 4.1.c. For example,

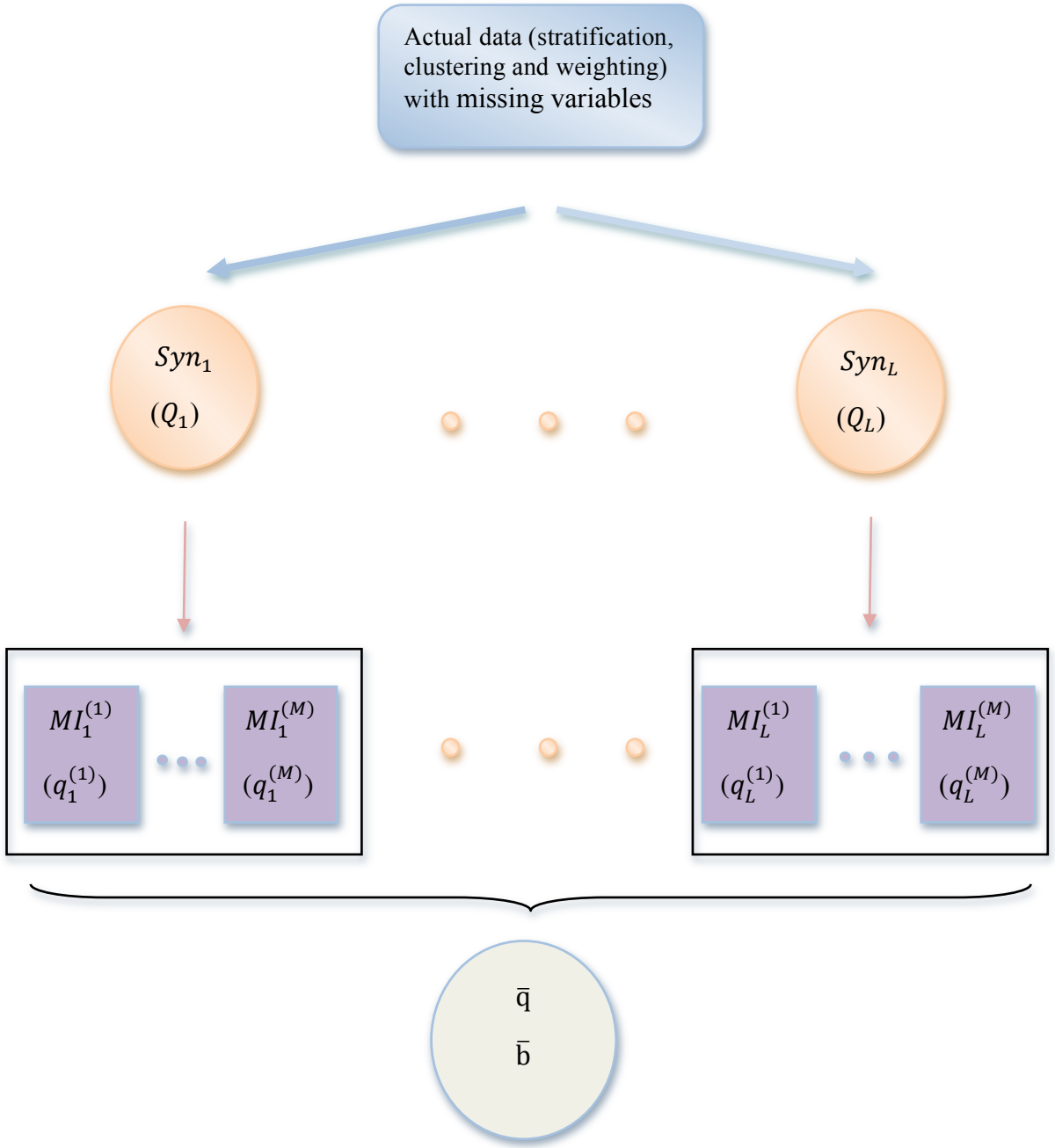
we could use sequential regression (Raghunathan *et al* 2001) to multiple impute the missing variables.

Step 3: Analyze each of the $L * M$ complete data sets using the statistical methods for simple random samples and denote the estimates for $Q(X, Y, Z)$ as $q_l^{(m)}, l = 1, \ldots., L, m = 1, \ldots, M.$

Step 4: For each survey, combine $q_l^{(m)}, l = 1, \ldots., L, m = 1, \ldots, M,$ using the two-stage combining rule developed later in this chapter and denote the combined point and variance estimate as $\overline{q_\infty^{(s)}}$ and $\overline{b_\infty^{(s)}}$, $\mathrm{s} = 1,2,3$ respectively.

Step 5: Combine the estimate from each survey using the combining rule for multiple surveys developed in Chapter 2 to produce the final combined point and variance estimate.

Figure 4.2 Flowchart for combining surveys with missing variables




### 4.3.2 Analyzing the Complete Datasets

In presence of missing information in some survey, we have to impute $M$ times for the item-missing data after we generate the synthetic populations. This procedure

involves two-stage imputation and produces $L * M$ complete data. However, neither the standard multiple imputation inference (Rubin 1987) nor the multiple synthetic data inference (Raghunathan *et al* 2003) will result in valid inference because both of them ignore the fact that it is a two-stage imputation, imputing the nonsampled units to produce synthetic populations and imputing the item-missing values to produce complete data, which brings in two separate sources of variability. Therefore, a new combining rule is needed. We propose to estimate $Q$ by

$$\bar{q} = \frac{1}{ML}\sum_{l=1}^{L}\sum_{m=1}^{M} q_l^{(m)} = \frac{1}{L}\sum_{l=1}^{L} \bar{q}_l$$

with variance

$$\bar{b} = \left(1+\frac{1}{L}\right)\frac{1}{L-1}\sum_{l=1}^{L}\left(\frac{1}{M}\sum_{m=1}^{M} q_l^{(m)} - \frac{1}{ML}\sum_{l=1}^{L}\sum_{m=1}^{M} q_l^{(m)}\right)^2 +$$

$$\frac{1}{L}\sum_{l=1}^{L}\left(1+\frac{1}{M}\right)\frac{1}{M-1}\sum_{m=1}^{M}\left(q_l^{(m)} - \frac{1}{M}\sum_{m=1}^{M} q_l^{(m)}\right)^2$$

$$= \left(1+\frac{1}{L}\right)\widetilde{B_L} + \left(1+\frac{1}{M}\right)\frac{1}{L}\sum_{l=1}^{L} b_l,$$

where $q_l^{(m)}$ is the computed value of the population quantity of interest $Q$ from the complete dataset, $\bar{q}_l = \frac{1}{M}\sum_{m=1}^{M} q_l^{(m)}$ is the combined point estimate for synthetic population $l$, $\widetilde{B_L} = \frac{1}{L-1}\sum_{l=1}^{L}(\bar{q}_l - \bar{q})^2$ is the between-variance for the estimates from the $L$ synthetic populations and $b_l = \frac{1}{M-1}\sum_{m=1}^{M}\left(q_l^{(m)} - \frac{1}{M}\sum_{m=1}^{M} q_l^{(m)}\right)^2$ is the between-variance for the estimates from the $M$ multiple imputed datasets for synthetic population $l$.

For small or moderate number of synthetic populations, the inference about $Q$ is made by approximating its posterior distribution by a $t$ distribution with degree of freedom $\gamma_L = (L-1)((1+L^{-1})\widetilde{B_L}/T)^{-2}$.

### 4.4 Evaluating the Two-stage Combining Rule

In this section, we derive the combining rule for the two-stage imputation in a single survey setting. Then we conduct a simulation study to evaluate the inference from the complete datasets obtained using the two-stage imputation procedure.

#### 4.4.1 Theoretical Justification for the Two-stage Combining Rule

Assume we generate $L$ synthetic populations, denoted as $Syn = \{Syn_1, \dots, Syn_L\}$ from the observed data. The next step is to multiply impute the missing information in each synthetic population, i.e., for $Syn_l, l = 1, \dots L$, we generate $M$ complete data $MI = \{MI_l^{(m)}, m = 1, \dots, M\}$. The goal is to derive the posterior predictive distribution of population quantity of interest $Q$ based on the complete datasets, i.e., $\pi\left(Q \middle| MI = \{MI_l^{(m)}, m = 1, \dots, M\}\right)$.

The conceptual framework for creating the complete datasets $MI$ outlined in Section 4.3.1 suggests the following natural decomposition,

$$\pi(Q|MI) = \int [\int \pi(Q|\, Syn, MI, Obs)\, \pi(Obs|\, Syn, MI)\, dObs]\, \pi(Syn|\, MI) dSyn,$$

where $Obs$ denotes the observed data. Obviously, $Syn$ and $MI$ are irrelevant after conditioning on $Obs$ because both are random functions of $Obs$. Similarly, $MI$ is

irrelevant after conditioning on $Syn$. Thus, $\pi(Q|\, Syn, MI, Obs) = \pi(Q|\, Obs)$ and $\pi(Obs|\, Syn, MI) = \pi(Obs|\, Syn)$. Thus, the expression for $\pi(Q|MI)$ simplifies to

$$\pi(Q|MI) = \int [\int \pi(Q|\, Obs)\, \pi(Obs|\, Syn)\, dObs]\, \pi(Syn|\, MI) dSyn$$

$$= \int \pi(Q|\, Syn)\ \pi(Syn|\, MI) dSyn.$$

Throughout this chapter, we assume that $L$ and $M$ are large enough to permit normal approximations for these posterior distributions. Thus, we only require the first two moments for each distribution. To derive these conditional moments, we use standard large sample Bayesian arguments. For example, to derive $\pi(Q|\, Syn)$, we treat the first two moments of $Q$ given $Syn$ as unknown and use $Syn$ as the data. Similarly, for the first two moments of $\pi(Syn|\, MI)$, we treat the first two moments based on $Syn$ as unknown and use $MI$ as the data. Diffuse priors are assumed for all parameters.

**First-stage inference: $\boldsymbol{\pi(Q|Syn_1, \dots, Syn_L)}$**

Let $\{Q_1, \dots, Q_L\}$ denote the estimator of the population quantity $Q$ from the synthetic populations. The nonsampled units can be treated as missing data and the standard multiple imputation framework (Rubin 1987) can be applied. Since each synthetic data is an entire population, the within-imputation variance can be ignored. Then based on Equations (3.1.5) and (3.1.6) from Rubin (1987, pp. 76-77), we have

$$Q|Syn \sim t_{L-1}(\bar{Q}, (1 + 1/L)B_L),$$

where $\bar{Q} = \frac{1}{L}\sum_{l=1}^{L} Q_l$ and $B_L = \frac{1}{L-1}\sum_{l=1}^{L}(Q_l - \bar{Q})(Q_l - \bar{Q})'$. When $L$ is large, it becomes a normal distribution.

**Second-stage inference: $\boldsymbol{\pi\left(Q_l \middle| MI_l^{(m)}, m = 1, \dots, M\right), l = 1, \dots, L}$**

Assume the computed values of the population quantity $Q$ from the $M$ complete data $\{MI_l^{(m)}, m = 1, \dots, M\}$ are denoted by $\{q_l^{(m)}, m = 1, \dots, M\}$. Since each $MI_l^{(m)}, m = 1, \dots, M$, is an entire complete population, we can ignore the within-imputation variance. Based on the combining rule for multiple imputation inference (Rubin 1987), we have

$$Q_l \mid \{MI_l^{(m)}, m = 1, \dots, M\} \sim t_{M-1}(\bar{q}_l, (1 + 1/M)b_l),$$

where $\bar{q}_l = \frac{1}{M}\sum_{m=1}^{M} q_l^{(m)}$ and $b_l = \frac{1}{M-1}\sum_{m=1}^{M}\left(q_l^{(m)} - \bar{q}_l\right)\left(q_l^{(m)} - \bar{q}_l\right)'$. When $M$ is large, it becomes a normal distribution. And the posterior distribution of $\bar{Q}$ becomes $N(\bar{q}, (1 + 1/M)\frac{b_l}{L})$, where $\bar{q} = \frac{1}{L}\sum_{l=1}^{L}\bar{q}_l = \frac{1}{ML}\sum_{l=1}^{L}\sum_{m=1}^{M} q_l^{(m)}$ .

**Derivation of $\boldsymbol{\pi(Q|MI_l^{(m)}, l = 1, \dots, L, m = 1, \dots, M)}$**

Since both $\pi(Q|\, Syn)$ and $\pi(Q|\, MI)$ are approximated by a normal distribution under the assumption that we generate a large number of synthetic populations and multiple imputed datasets, the posterior distribution $\boldsymbol{\pi}(Q|MI)$ can be approximated by a normal distribution with mean $E(Q|MI)$ and variance $Var(Q|MI)$.

Using the results in the first and second stage approximation,

$$E(Q|MI) = E[E(Q|Syn)|MI] = E(\bar{Q}|MI) = \bar{q}.$$

Since $\frac{\widetilde{B_L}}{B_L + (1+1/M)b_l}|MI \sim \chi^2_{L-1}/(L-1)$ (Raghunathan, Reiter and Rubin 2003) and when $L$ is large, $\chi^2_{L-1}/(L-1) \approx 1$, we can approximate $B_L$ by $\widetilde{B_L}$. Here we omit

$(1 + 1/M)b_l$ because the missing information due to the nonresponse is trivial compared to the missing information from the nonsampled part of the population. Thus, we have

$$Var(Q|MI) = E[Var(Q|Syn)|MI] + Var[E(Q|Syn)|MI] = \left(1 + \frac{1}{L}\right)\widetilde{B_L} + \left(1 + \frac{1}{M}\right)\frac{b_l}{L}.$$

For small or moderate number of synthetic populations, the inference about $Q$ is made by approximating its posterior distribution by a *t* distribution with degree of freedom $\gamma_L = (L - 1)((1 + L^{-1})\widetilde{B_L}/T)^{-2}$.

### 4.4.2 Simulation Validation for the Two-stage Combining Rule

This section presents a simulation study for the two-stage combining rule. We first generate a population of size $N = 1{,}000$ from a 5-variate normal distribution with means equal to 0, variances equal to 1 and a common covariance of 0.5. We denote the 5 variables as $x_1, x_2, x_3, x_4, x_5$, respectively. Then we draw 500 independent actual samples of size $n = 100$ using simple random sampling. For each sample, we create item-missing data on $x_4$ and $x_5$ and the missing data pattern simulates the situation in Figure 4.1, i.e., we simple random sample 80 subjects from the sample and make $x_4$ missing for the first 60 subjects and make $x_5$ missing for the last 20 subjects. Then for each sample with item missing data, we generate $L = 5$ synthetic populations of size 1000. For each synthetic population, we multiple impute the item missing data $M = 5$ times using Markov chain Monte Carlo (MCMC) method (Schafer, 1997) which is realized by the proc mi function in SAS. Thus, we obtain 500*5*5=12500 data sets, which will be combined using the two-stage combining rule.

**Generating synthetic data**

We use the same synthetic data generation model as described in the Simulation Study 1 in Raghunathan et al (2003). The imputation model assumes multivariate normal distribution with unknown mean μ and covariance matrix Σ. A noninformative Jeffrey prior, $\pi(\mu, \Sigma) \propto |\Sigma|^{-1/2}$, is applied (Jeffreys, 1961). Suppose $\bar{y}$ is the sample mean and $S$ is the sample covariance matrix for a particular sample. Standard Bayesian calculations lead to the following procedure for creating synthetic data sets:

- Generate a random variate, $W$, from a Wishart distribution with $n-1$ degrees of freedom and the associated matrix $\frac{S^{-1}}{n-1}$. Define $\Sigma^* = W^{-1}$.

- Generate $\mu^*$ from a multivariate normal distribution with mean $\bar{y}$ and covariance matrix $\Sigma^*/n$.

- Generate $N = 1000$ independent multivariate normal random vectors with mean $\mu^*$ and covariance matrix $\Sigma^*$.

- Repeat this process $L = 5$ times to create five synthetic populations of size 1000 each.

For the part of the sample with missing $x_4$, we impute the nonsampled part of the population for $(x_1, x_2, x_3\ x_5)$ from a 4-vairate normal distribution using the same steps and set $x_4$ as missing in the synthetic populations. We use the same approach for generating synthetic populations for the part of the sample with missing $x_5$. Since the imputation model matches the underlying true model, the synthetic data are created under the best scenario. This setup allows for the evaluation of our inference method without unnecessary implications from other factors.

**Simulation results**

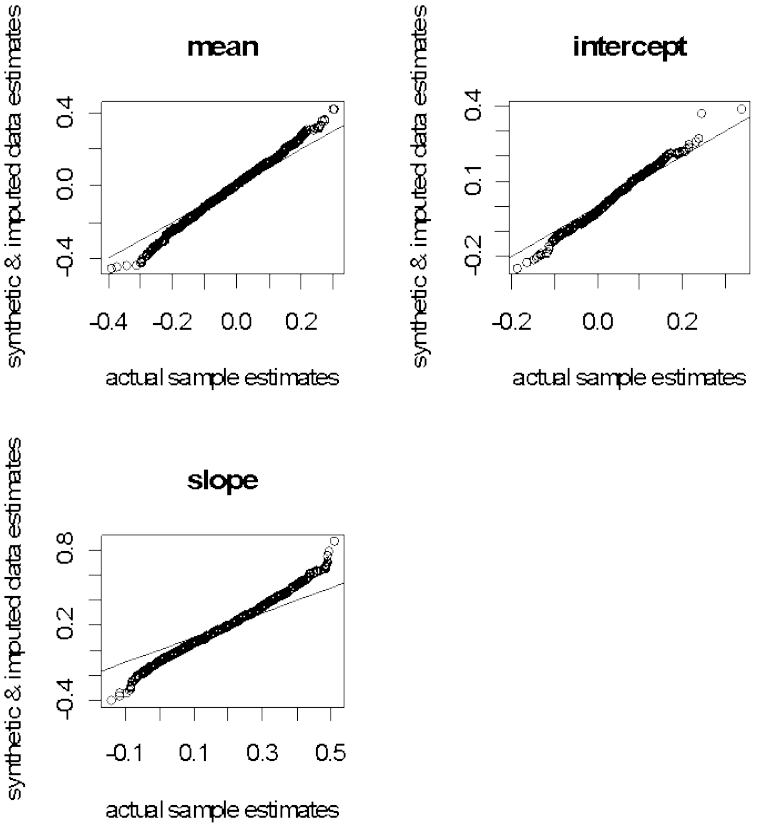
The estimands of interest are the marginal means for all five variables and the regression coefficients of $x_1$ on $(x_2, x_3, x_4\ x_5)$. We perform ordinary linear regression analyses to obtain the estimates for the regression coefficients and the standard errors based on the actual data and each imputed data set. We obtain the 95% confidence intervals for the statistics of interest from the imputed data sets using the combining rule and from the actual data using the standard t-distribution.

Figure 4.3 displays the scatter plot of the 500 pairs of estimated mean, intercept and slope from the actual samples and the corresponding imputed data sets along with a 45-degee line. The sampling distributions of the actual sample and imputed sample estimates are practically the same. Table 4.1 compares the inference of descriptive and analytic statistics from the actual data and the multiple imputed data. The point estimates for both types of statistics are very similar across the imputed and actual data. The standard errors from the imputed data are a little bigger than the standard errors from the actual data suggesting some loss of efficiency. Also provided in Table 4.1 are the proportions of 95% confidence intervals that contain the true value. Consistent results are found with the coverage rates as the imputed data provides an over nominal coverage than the actual sample for both descriptive and analytic statistics. The 95% confidence intervals from the imputed data are wider than those from the actual samples as the average interval lengths are 0.39 for the actual data and 0.79 for the imputed data. Based on this new combining rule, the repeated sampling properties of the inference from the actual and imputed data are almost identical as predicted. In conclusion, the combining rule yields valid inference.

Table 4.1 Descriptive and analytic statistics estimated from the actual data and the imputed data in the simulation evaluation of combining rule

| Type | Actual | | | Imputed | | | No. of estimates |
|---|---|---|---|---|---|---|---|
| | Estimate | SE | Coverage (%) | Estimate | SE | Coverage (%) | |
| Mean | -0.025 | 0.101 | 95.8 | -0.026 | 0.132 | 96.56 | 2500 |
| Intercept | 0.0399 | 0.081 | 95.6 | 0.0391 | 0.106 | 97.0 | 500 |
| Slope | 0.197 | 0.101 | 96.25 | 0.198 | 0.171 | 96.1 | 2000 |

Figure 4.3 Scatter plot of the descriptive and analytic statistics from the actual and imputed data sets




## 4.5 Application

In this section, we apply the new method to combine three surveys to make inference on health insurance coverage. The surveys we consider are the 2006 Behavioral Risk Factor Surveillance System (BRFSS), the 2006 National Health Interview Survey

(NHIS) and the 2006 Medical Expenditure Panel Survey (MEPS). The three surveys use different modes of data collection and have different sampling designs, but they have similar target populations and share a considerable amount of questions.

### 4.5.1 Data Sources

The Behavioral Risk Factor Surveillance System is an ongoing telephone survey of the health behaviors of U.S. adults and was established in 1984 by Centers for Disease Control and Prevention (CDC). The BRFSS was designed to provide state-specific estimates of the prevalence of risk behaviors. By 1994, all states and the District of Columbia (DC) participated. The BRFSS sample households are obtained using a list-assisted random digit dial (RDD) telephone sampling and thus are not clustered. The BRFSS allows the states to implement their own protocols, though some features have been standardized. It interviews more than 350,000 adults each year and thus is able to produce reliable estimates on both state level and national level.

The National Health Interview Survey is a principle source of national health information for the U.S., non-institutionalized, civilian population. It has been conducted annually since 1957 by National Center for Health Statistics (NCHS) and Centers for Disease Control and Prevention (CDC). The NHIS utilizes a stratified, multi-stage probability sampling design. The sample is drawn from a geographic frame designated using the most recent decennial Census. Names and addresses are derived in a separate listing activities conducted specifically for NHIS. From 1995 through 2005, African American and Hispanic households were oversampled in order to facilitate better estimates for these populations. Beginning in 2006, households with at least one Asian member are also oversampled. The NHIS is currently conducted via a personal household

interview with a knowledgeable adult household representative using computer-assisted personal interviewing (CAPI) technology.

A third data source is the Medical Expenditure Panel Survey. The MEPS has been conducted annually since 1996 by Agency for Healthcare Research and Quality (AHRQ). It currently has two major components: the Household Component and the Insurance Component. The Household Component collects data from a sample of families and individuals in selected communities across the United States. The sample is drawn from a nationally representative subsample of households that participated in the prior year's National Health Interview Survey. The MEPS household survey is also conducted via a personal household interview. The Insurance Component collects data from a sample of private and public sector employers on the health insurance plans they offer their employees. Thus, MEPS provides a very rich database that includes medical care utilization data.

Both the NHIS and MEPS ask people whether they are covered by insurance and if so what type of insurance (government vs private) they use, while the BRFSS does not have information about the type of insurance people use. However, the sample size of the BRFSS is about 10 times as big as the NHIS and MEPS and it has small area identifier available to public. For example, the sample sizes for the 2006 BRFSS, the 2006 NHIS and the 2006 MEPS are 355,710, 75,716 and 34,145 respectively. Thus, the NHIS and MEPS may not be able to produce precise estimates for health insurance coverage especially for small domains of the population. If we impute the missing information about the types of insurance that people use in the BRFSS data, we could use the complete data to produce precise estimates on a small area or small domain level.

### 4.5.2 Combining the NHIS, BRFSS and MEPS

Following the proposed steps in Section 4.3.1, we first use the nonparametric approach to generate synthetic populations to adjust for the different sampling designs that the three surveys used. Then we stack the synthetic populations from the three surveys and create a missing data problem as in Figure 4.1.b. Next we fill in the missing information in the BRFSS by using the sequential regression method (Raghunathan *et al.* 2001) implemented by IVEware software (Raghunathan, Solenberger and Van Hoewyk 2002). Finally, we compute the health insurance coverage rates for the whole population and some sub-domains from each complete dataset and combine the estimates using the two-stage combining rules developed in this chapter. The results are summarized in Table 4.2. Since the BRFSS has a much larger sample than the NHIS and MEPS, after the missing information in the BRFSS is imputed, we have more precise estimates compared to the ones from the NHIS and MEPS.

Then we apply the combining rule for multiple surveys developed in Chapter 2. The combined estimates are summarized in Table 4.3. By making use of the large sample of the BRFSS, we produce the combined estimates with much smaller variance estimates than the ones from the NHIS and MEPS.

Table 4.2 Estimates from Actual Data and from Synthetic Populations after Missing Information is Imputed for the 2006 NHIS, MEPS and BRFSS

| Domain | Actual Data (Complex Design) | | | | Synthetic Populations | | |
|---|---|---|---|---|---|---|---|
| | Types | NHIS | BRFSS | MEPS | NHIS | BRFSS | MEPS |
| Whole Population | Proportion | | | | | | |
| | Private | 0.746 | | 0.735 | 0.746 | 0.769 | 0.736 |
| | Public | 0.075 | | 0.133 | 0.075 | 0.078 | 0.132 |
| | Uninsured | 0.179 | 0.154 | 0.132 | 0.179 | 0.153 | 0.132 |
| | Variance | | | | | | |
| | Private | 2.46E-05 | | 2.78E-05 | 3.15E-05 | 7.52E-06 | 3.31E-05 |
| | Public | 6.29E-06 | | 1.44E-05 | 8.06E-06 | 5.39E-06 | 1.59E-05 |
| | Uninsured | 1.84E-05 | 3.32E-06 | 1.41E-05 | 2.29E-05 | 3.52E-06 | 1.71E-05 |
| Male | Proportion | | | | | | |
| | Private | 0.740 | | 0.735 | 0.740 | 0.770 | 0.736 |
| | Public | 0.060 | | 0.101 | 0.060 | 0.063 | 0.100 |
| | Uninsured | 0.200 | 0.167 | 0.164 | 0.200 | 0.167 | 0.164 |
| | Variance | | | | | | |
| | Private | 3.32E-05 | | 3.87E-05 | 3.93E-05 | 1.31E-05 | 4.31E-05 |
| | Public | 6.82E-06 | | 1.53E-05 | 8.81E-06 | 5.66E-06 | 1.63E-05 |
| | Uninsured | 2.94E-05 | 8.88E-06 | 2.64E-05 | 3.29E-05 | 9.17E-06 | 2.79E-05 |
| Hispanic | Proportion | | | | | | |
| | Private | 0.494 | | 0.506 | 0.495 | 0.519 | 0.508 |
| | Public | 0.096 | | 0.161 | 0.097 | 0.112 | 0.158 |
| | Uninsured | 0.410 | 0.371 | 0.334 | 0.409 | 0.369 | 0.334 |
| | Variance | | | | | | |
| | Private | 1.24E-04 | | 1.73E-04 | 1.94E-04 | 7.24E-05 | 1.97E-04 |
| | Public | 2.57E-05 | | 8.03E-05 | 3.88E-05 | 3.39E-05 | 8.43E-05 |
| | Uninsured | 1.23E-04 | 7.18E-05 | 1.19E-04 | 1.90E-04 | 6.84E-05 | 1.61E-04 |
| Non-Hispanic White | Proportion | | | | | | |
| | Private | 0.805 | | 0.788 | 0.804 | 0.831 | 0.788 |
| | Public | 0.062 | | 0.116 | 0.062 | 0.063 | 0.116 |
| | Uninsured | 0.134 | 0.106 | 0.096 | 0.134 | 0.106 | 0.096 |
| | Variance | | | | | | |
| | Private | 2.99E-05 | | 3.35E-05 | 3.79E-05 | 7.55E-06 | 4.12E-05 |
| | Public | 8.20E-06 | | 1.81E-05 | 1.04E-05 | 6.26E-06 | 2.00E-05 |
| | Uninsured | 2.02E-05 | 2.15E-06 | 1.51E-05 | 2.35E-05 | 2.44E-06 | 1.8E-05 |
| Non-Hispanic White & Income [25,000, 35,000) | Proportion | | | | | | |
| | Private | 0.827 | | 0.813 | 0.827 | 0.755 | 0.814 |
| | Public | 0.039 | | 0.079 | 0.039 | 0.072 | 0.079 |
| | Uninsured | 0.134 | 0.173 | 0.108 | 0.134 | 0.173 | 0.107 |
| | Variance | | | | | | |
| | Private | 1.00E-04 | | 1.39E-04 | 1.48E-04 | 6.53E-05 | 1.63E-04 |
| | Public | 2.82E-05 | | 6.31E-05 | 3.86E-05 | 3.15E-05 | 7.28E-05 |
| | Uninsured | 7.24E-05 | 2.78E-05 | 8.92E-05 | 9.55E-05 | 3.25E-05 | 1.11E-04 |

Table 4.3 Estimates from Individual Surveys and the Combined Estimates after the Missing Information is Imputed for the 2006 NHIS, MEPS and BRFSS

| Domain | Actual Data (Complex Design) | | | | Combined Estimates |
|---|---|---|---|---|---|
| | Types | NHIS | BRFSS | MEPS | |
| Whole Population | Proportion | | | | |
| | Private | 0.746 | | 0.735 | 0.760 |
| | Public | 0.075 | | 0.133 | 0.086 |
| | Uninsured | 0.179 | 0.154 | 0.132 | 0.153 |
| | Variance | | | | |
| | Private | 2.46E-05 | | 2.78E-05 | 5.13E-06 |
| | Public | 6.29E-06 | | 1.44E-05 | 2.68E-06 |
| | Uninsured | 1.84E-05 | 3.32E-06 | 1.41E-05 | 2.59E-06 |
| Male | Proportion | | | | |
| | Private | 0.740 | | 0.735 | 0.758 |
| | Public | 0.060 | | 0.101 | 0.069 |
| | Uninsured | 0.200 | 0.167 | 0.164 | 0.172 |
| | Variance | | | | |
| | Private | 3.32E-05 | | 3.87E-05 | 8.01E-06 |
| | Public | 6.82E-06 | | 1.53E-05 | 2.84E-06 |
| | Uninsured | 2.94E-05 | 8.88E-06 | 2.64E-05 | 5.71E-06 |
| Hispanic | Proportion | | | | |
| | Private | 0.494 | | 0.506 | 0.512 |
| | Public | 0.096 | | 0.161 | 0.114 |
| | Uninsured | 0.410 | 0.371 | 0.334 | 0.369 |
| | Variance | | | | |
| | Private | 1.24E-04 | | 1.73E-04 | 4.16E-05 |
| | Public | 2.57E-05 | | 8.03E-05 | 1.49E-05 |
| | Uninsured | 1.23E-04 | 7.18E-05 | 1.19E-04 | 3.83E-05 |
| Non-Hispanic White | Proportion | | | | |
| | Private | 0.805 | | 0.788 | 0.822 |
| | Public | 0.062 | | 0.116 | 0.071 |
| | Uninsured | 0.134 | 0.106 | 0.096 | 0.107 |
| | Variance | | | | |
| | Private | 2.99E-05 | | 3.35E-05 | 5.46E-06 |
| | Public | 8.2E-06 | | 1.81E-05 | 3.27E-06 |
| | Uninsured | 2.02E-05 | 2.15E-06 | 1.51E-05 | 1.97E-06 |
| Non-Hispanic White & Income [25,000, 35,000) | Proportion | | | | |
| | Private | 0.827 | | 0.813 | 0.785 |
| | Public | 0.039 | | 0.079 | 0.062 |
| | Uninsured | 0.134 | 0.173 | 0.108 | 0.153 |
| | Variance | | | | |
| | Private | 1.00E-04 | | 1.39E-04 | 3.54E-05 |
| | Public | 2.82E-05 | | 6.31E-05 | 1.40E-05 |
| | Uninsured | 7.24E-05 | 2.78E-05 | 8.92E-05 | 1.99E-05 |

## 4.6 Discussion

In this chapter, we present the steps to combine any number of surveys of different sampling designs and survey error properties. The biggest advantage of this new method is that it could combine surveys that use different sampling designs and share disjointed subsets of information. This actually fulfills the objects of both combining surveys and statistical matching.

To make valid inference from the two-stage imputation procedure, we develop a new combining rule from a Bayesian perspective and verify it via a simulation study. Then we apply this approach to combine the NHIS, BRFSS and MEPS after we fill in the missing information in the BRFSS. The combined variance estimates are reduced dramatically due to the use of the large sample size in the BRFSS and the estimates for small domains are more precise than the ones from the NHIS and MEPS. Since the BRFSS has county-level indicator in its data, we can produce the county-level estimates for the three types of health insurance coverage rates using the complete BRFSS data, something impossible to estimate from any of the three individual surveys.

Even though this approach generates a large number of synthetic populations and multiply imputes the missing information, the computational burden is manageable. Our stacked data contains about 17 millions observations and it takes less than a day to produce the final results.

One limitation is that in the application where we combine the 2006 BRFSS, NHIS and MEPS to estimate health insurance coverage rates, we only focus on filling in

the missing information about the type of the health insurance in the BRFSS data. We ignore the fact that the BRFSS is a telephone survey and has noncoverage error because it excludes people without telephone from the sample and may have larger nonresponse error because of the lower response rates compared to the NHIS and MEPS. A more comprehensive application of this new combining survey method that adjusts for all the discrepancies in the data from multiple surveys will be the focus of our future research.

# CHAPTER 5

# DISCUSSIONS AND FUTURE WORK

## 5.1 Summary of this Dissertation

This dissertation develops a new method for combining information from multiple complex surveys from a missing data perspective. This method could be applied to combine multiple surveys that are conducted independently but cover the same underlying population. The new method first imputes the unobserved population and generates synthetic populations to adjust for the complex sampling design features of the multiple surveys. Then the synthetic populations from multiple surveys can be treated as simple random samples from the same population and thus can be stacked to impute the missing variables. Also, we could adjust for nonsampling errors of individual surveys by borrowing information from the surveys with smaller or no error. Once we have the complete data, we could estimate the population quantity of interest from each of them and combine the estimates using the appropriate combining rules to produce the combined estimates.

Since the imputation models for multiple surveys could be different, the current combining rule for the synthetic populations that are generated from one imputation model is not appropriate. This dissertation derives the posterior predictive distribution of the population quantity of interest given the data from multiple surveys, which is approximated by a normal distribution when the number of synthetic populations is infinite and by a t distribution when we generate a limited number of synthetic

populations. The combined estimate is a weighted average of the estimates from individual surveys. This suggests that after we adjust for the nonsampling error to eliminate the biases from individual surveys, the combined estimate is unbiased. The combined estimate is also more precise than the ones from individual surveys. The combining rule is then evaluated via a simulation study, which shows that the combined estimate is unbiased for the population true value and has nominal coverage rate even though it has smaller variance estimate and empirical mean square error (eMSE) compared to the estimates from individual surveys.

In situations where there are missing variables in one or multiple surveys, we first impute the unobserved population and then fill in the missing variables by borrowing information from other surveys. This dissertation derives a two-stage combining rule to adjust for the extra uncertainty due to simultaneously generating synthetic populations and imputing the missing variables. We prove the randomization validity of the combing rule and evaluate it using a simulation study. This simulation study shows the two-stage combining rule produces identical point estimates for both descriptive statistic and analytical statistics. The variance estimates are also well maintained despite there is a little inflation over the ones from the actual data that results in slightly wider confidence interval coverage.

This dissertation develops both a model-based method and a nonparametric method to generate synthetic populations from the observed data of each survey. The model-based method uses the asymptotic normal distribution of the model parameters to approximate the posterior distribution of the model parameters given the observed data. We use a jackknife repeated replication method to adjust for stratification, clustering and

unequal probability of selection when estimating the point estimate and covariance matrix of the model parameters. We evaluate the model-based method under two situations: when the underlying model is linear and when the underlying model is log-linear. The simulation study shows that when the imputation model that generates the synthetic populations is the exact model that generates the population, the model-based method adjusts for the sampling design features without losing any information. In applications, as long as we correctly specify the imputation model, i.e., selecting the correct type of model, including the important variables and taking into consideration the sampling design features, the model-based method does a good job in adjusting for the complex sampling design features when generating synthetic populations. However, when the imputation model fails to capture the variability in the data for small domains, the model-based method could have problems in adjusting for the complex sampling design features. This suggests the model-based method may fail when the relationship among the variables of interest is too complicated to be specified by models or when we have samples too small to fit the imputation models.

A nonparametric method is developed to overcome the potential model misspecification for the model-based method. The nonparametric method only focuses on the design variables such as stratum, cluster and weight in the observed data and does not specify any model. Once the unobserved units are drawn from the observed data, all the variables of the selected units will be drawn. When the synthetic populations are analyzed as simple random samples, we prove the point estimate obtained from the combining rule for synthetic data is unbiased and the variance estimate is unbiased for the actual sampling variance. We also show the sampling properties of inferences on population

mean and regression coefficients from the synthetic populations are very similar to the actual sample. We then apply this method to generate synthetic populations for a stratified clustering sample and a clustering sample. For both sampling designs, the method adjusts for the complex sampling design features even for small domains. For the whole population and large domains where the imputation model performs well, the nonparametric is a little less efficient than the model-based method.

Then we apply the new combining survey method to estimate the percentage of the population that is covered by private insurance, is covered by public insurance and is uninsured by combining the 2006 NHIS, MEPS and BRFSS. When we combine the NHIS and the MEPS, we see the increase in precision for the combined estimates could be as high as 191% over the NHIS and 266% over the MEPS. The BRFSS has a sample about 15 times as large as the ones from the NHIS and MEPS. However, it only asks whether one is insured or not. When we combine the BRFSS with the NHIS and MEPS, we first adjust for the different sampling designs of the three surveys and then impute the missing information in the BRFSS by borrowing from the NHIS and MEPS. The combined estimates are more precise than the ones from any individual surveys and more precise than the combined estimates from only the NHIS and MEPS.

## 5.2 Future Work

### 5.2.1 Hierarchical Bayesian Model-based Method to Fully Adjust for Complex Sampling Design Features

In Chapter 2, we use the asymptotic normal distribution to approximate the posterior distribution of the imputation model parameters. This method captures the

relationship of the variables on national level. However, it does not consider the relationships of the clusters within strata and the intracluster correlation. To fully consider these relationships and adjust for stratification, clustering and unequal probability of selection, a hierarchical Bayesian approach will be developed. Under the Bayesian approach, the prior distribution can be of a hierarchical form, the Bayesian formulation provides a natural setting for hierarchical modeling, which allows for sharing information across clusters and provides a convenient means of incorporating clustering and stratification. This hierarchical Bayesian model will adjust for the complex sampling design features on both national level and subdomain level.

### 5.2.2 Applying the Method to Adjust for the Nonsampling Errors

This dissertation mainly focuses on adjusting for the different sampling designs of multiple surveys. It ignores the fact that data collected by different modes (face-to-face, telephone, web, etc.) could have different survey error properties. The generation of equivalent synthetic populations for multiple surveys makes it possible to convert many combining survey situations into missing data problems to adjust for the nonsampling errors by using the information from the data of higher quality. In future research, the new method will be extended to adjust for nonsampling errors. A comprehensive simulation study will be conducted, from which we will evaluate the new method in a variety combining survey scenarios that aim to fix different types of nonsampling errors. Both accuracy and precision of the combined estimates will be evaluated in the simulation study. The method will also be applied to combine real survey data to correct deficiencies in each data source and produce more precise estimates.

Next, we describe how to convert common problems of combining surveys into a unified missing data problem that can be combined using the method in this dissertation. We provide two scenarios where combining information from multiple surveys could adjust for noncoverage error, nonresponse error and measurement error.

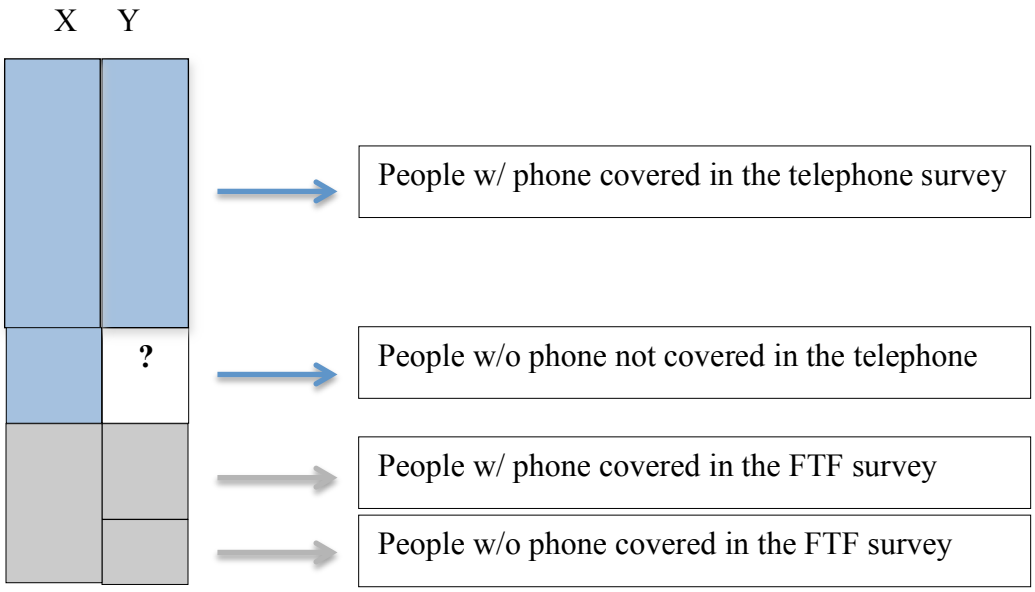
#### 5.2.2.1 Reducing Noncoverage Error/Nonresponse Error

The last 50 years have seen a gradual replacement of face-to-face (FTF) surveys with telephone surveys as the dominant mode of data collection in the United States. The main reason is that telephone surveys are much cheaper (Hochstim 1967; Groves and Kahn 1979) because they don't require interviewers to travel to the respondent's location to conduct a personal interview. Thus, they are able to produce a large sample that is widely distributed geographically and is suitable for small area estimation. However, telephone surveys cannot contact people without a phone. When people with telephones and people without telephones have different mean values for the quantity of interest, telephone surveys could introduce noncoverage error. Also, the response rate of a telephone survey is usually lower that that of a face-to-face survey (de Leeuw and Edith Desiree 1992), which means telephone surveys could potentially introduce larger nonresponse error. One reason that researchers combine telephone surveys with face-to-face surveys is that the information obtained by face-to-face surveys can be used to reduce the noncoverage error and nonresponse error in telephone data. We can then use the adjusted telephone data to produce statistics of interest on a small area level (Elliott and Davis 2005; Raghunathan 2007).

Assume the telephone coverage rate is 95%, the sample size of the telephone survey is $n$. To adjust for the noncoverage error in a telephone survey, we could add

$\frac{n}{95\%} - n$ rows below the telephone sample with all variables missing to represent the sample without telephones. Figure 5.1 shows the missing data structure for this combining survey situation in which we also exclude the background variables that are observed for the whole population. Then we can borrow the face-to-face survey data to impute the missing data using the new method proposed in this chapter.

Figure 5.1 Converting a combining surveys problem into a missing data problem to adjust for noncoverage error



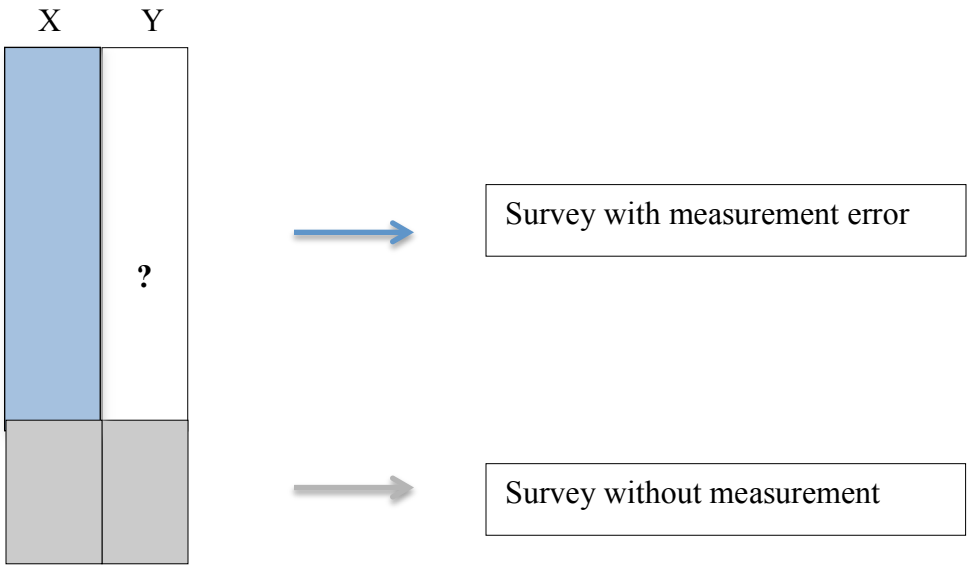

### 5.2.2.2 Reducing Measurement Error

Measurement errors are deviations of respondents' answers from their true values on the measure. Respondents tend to misreport when they are asked sensitive questions. For example, respondents tend to over-report socially desirable behaviors and under-report socially undesirable behaviors. This effect is more likely to happen in interviewer-administered surveys than self-administered surveys (Aquilino and Losciuto 1990;

Hochstim 1967; Locander, Sudman and Bradburn 1976; Turner, Lessler and Devore 1992).

One way to reduce measurement error is to combine the survey with larger error with the one with smaller or no error (Schenker *et al.* 2009). For example, assume there are two variables of interest, *X* and *Y* and there are two surveys collecting the variables. Further assume both surveys measure *X* accurately. However, since *Y* collects sensitive information, one survey measures *Y* with error and the other one measuring *Y* without error. We could treat the variable of interest measured with error as missing so that we have a missing data problem as in Figure 4.1.

Figure 5.2 Converting a combining survey problem into a missing data problem to adjust for measurement error




### 5.2.3 Developing Relevant Statistical Packages

One advantage of this new combining survey method is that it is possible to implement it as a standard package for the existing statistical software – especially for the nonparametric method that only focuses on the design variables and thus has the same general approach for all types of data structure. An R package (tentatively called "CombineSurvey") and a new module for the Imputation and Variance Estimation Software (tentatively called "COMBINE") will be developed to improve the usability and practicality, which will have two functions: 1) to generate synthetic populations to adjust for the sampling designs; and 2) to produce combined estimates using the appropriate combining rule.

## Bibliography


Agresti, A. (2002). Categorical Data Analysis, New York, Wiley & Sons.

Aquilino, W.S. and Losciuto, L.A. (1990). Effects of Interview on Self-Reported Drug Use, Public Opinion Quarterly, Vol. 54(3), 362-391.

Barr, R.S. and Turner, J.S. (1980). Merging the 1977 Statistics of Income and the March 1978 Current Population Survey, Prepared for the Office of Tax Analysis, Washington, D.C., U.S. Dept. of the Treasury.

Bickel, PJ. And Freedman, D.A. (1984). Asymptotic normality and the bootstrap in stratified sampling, The Annals of Statistics, 12, 470-482.

Chao, M.T. and Lo, S.H. (1985). A bootstrap method for finite population, Sankhyā Ser, A 47, 399-405.

Cobanoglu, C., Warde, B. and Moreo, P.J. (2001). A Comparison of Mail, Fax, and Web Survey Methods, International Journal of Market Research, 43:441–52.

Cochran, W.G. (2007). Sampling Techniques, Chapter 12, Wiley: New York.

Cohen, M.P. (1991). Statistical matching and microsimulation models, in improving information for social policy decisions: the uses of microsimulation modeling, Vol. II: Technical Papers, eds. C.F. Citro and E.A. Hanushek, Washington, D.C.: National Academy Press, 62-88.

Cohen, M.P. (1997). The bayesian bootstrap and multiple imputation for unequal probability smaple designs, Proceedings of the Survey Research Methods Section, Journal of the American Statistical Association.

Couper, M.P. (2000). Web Surveys: A Review of Issues and Approaches, Public Opinion Quarterly, 64:464–94.

Couper, M.P., Traugott, M.W., and Lamias, M.J. (2001). Web Survey Design and Administration, Public Opinion Quarterly, 65:230–53.

Davis, W.W., Parsons, V.L., Xie, DW, Schenker, N., Town, M., Raghunathan, T.E. and Feuer, E.J. (2010). State-Based Estimates of Mammography Screening Rates Based on Information from Two Health Sruveys, Public Health Report, Vol 125, 567-578.

de Leeuw and Edith Desiree. (1992). Data Quality in Mail, Telephone and Face to Face Surveys, T. T. Publikaties, Plantage Daklaan 40, 1018CN Amsterdam.

Efron, B. (1979). Bootstrap methods: another look at the jackknife, The Annals of Statistics, 7 (1): 1–26

Elliott, M.R. and Davis, W.W. (2005). Obtaining Cancer Risk Factor Prevalence Estimates in Small Areas: Combining Data from Two Surveys, Applied Statistics, Part 3, 595-609.

Ericson, W.A. (1969). Subjective Bayesian modeling in sampling finite populations, Journal of the Royal Statistical Society, Series B, 31, 195-234.

Fellegi, I.P. and Sunter, A.B. (1969). A theory for record linkage, Journal of the American Statistical Association, 64, 1183-1210.

Ferguson, T.S. (1973). A Bayesian analysis of some nonparametric problems, The Annals of Statistics, 7, 1-26.

Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (2004). Bayesian Data Analysis, Chapman & Hall/CRC

Gross, S. (1980). Median estimation in sample surveys, presented at the 1980 Joint Statistical Mettings.

Groves, R.M. and Kahn, R.L. (1979). Surveys by telephone: A national comparison with personal interviews, Academic Press, New York.

Hansen, M. (1987). Some history an reminiscences on survey sampling. Statistical Science, 2, 162-179.

Hansen, M.H. and Hurwitz, W.N. (1943). On the Theory of Sampling from Finite Populations. The Annals of Mathematical Statistics, Vol. 14, No. 4, 333-362.

Hansen, M.H., Madow, W.G. and Tepping, B.J. (1983). An Evaluation of Model-Dependent and Probability-Sampling Inferences in Sample Surveys, Journal of the American Statistical Association, Vol. 78, 384, 776-793

Hartley, H. O. (1974). Multiple Frame Methodology and Selected Applications, The Indian Journal of Statistics, Vol. 38, Series C. Pt, 3, 99-118.

He, Y. and Raghunathan, T.E. (2006). Tukey's gh Distribution for Multiple Imputation, Journal of the American Statistical Association, Vol. 60, no3, 251-256.

Hinkins, S., Oh, H.L. and Scheuren, F. (1997). Inverse sampling design algorithms, Survey Methodology, 23, No. 1, 11-21.

Hochstim, J.R. (1967). A Critical Comparison of Three Strategies of Collecting Data from Households, Journal of the American Statistical Association, Vol. 62, No. 319.

Holt, D. and Smith, T.M.F. (1979). Poststratification, Journal of the Royal Statistical Society, Series A, 142, 33-46.

Huber, P.J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions, Proc 5th Berkeley Symp. In Math. Stat. and Prob, 1, 221-233.

Ingram, D.D., O'Hare, J., Scheuren, F., and Turek, J. (2000). Statistical matching: a new validation case study, in Proceedings of the Survey Research Methods Section, American Statistical Association, 746-751.

Jebe, D.H. (1952). Estimation for Sub-Sampling Designs Employing the County as a Primary Sampling Unit. Journal of the American Statistical Association, Vol. 47, No. 257, 49-70.

Kadane, J.B. (1978). Some statistical problems in merging data files, Journal of Official Statistics, 17, 3, 423-433.

Kish, L. (1995). The Hundred Years Wars of Survey Sampling, Centennial representative Sampling Conference, Rome, May 31, 1995.

Knoke, D. and Burke, P.J. (1980). Log-linear Models, Sage Publications, Inc, Newberry Park, California, USA.

Lacander, W., Sudman, S. and Bradburn., N. (1976). An Investigation of Interview Method, Threat and Response Distortion, Journal of the American Statistical Association Vol 71, 269-275.

Little, R.J.A. (1983). Poststratification: A modeler's perspective, Journal of the American Statistical Association, 88, 1001-1012.

Little, R.J.A. (2004). To model or not to model? Competing modes of inference for finite population sampling, Journal of the American Statistical Association, 99, No. 466, 546-556

Little, R.J.A and Rubin D.B. (2002). Statistical Analysis with Missing Data, Wiley series in probability and statistics, New York: Wiley.

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

Lo, A.Y. (1987). A large sample study of the Bayesian bootstrap, Annals of Statistics, 15, 360-375.

Lo, A.Y. (1988). A Bayesian bootstrap for a finite population, Annals of Statistics, 16, 1684-1695.

Madow, W.G., Nisselson, H., Olkin, I. and Rubin, D.B. (1983). Incomplete Data in Sample Surveys. 1, 2, and 3, New York: Academic Press.

Moriarity, C. and Scheuren, F. (2001). Statistical matching: a paradigm for assessing the uncertainty in the procedure, Journal of Official Statistics, 17, 407-422.

Okner, B.A. (1972). Constructing a new data base from existing microdata sets: the 1966 merge file, Annals of Economic and Social Measurement, 1, 325-342.

Radner, D.B., Allen, R., Gonzalez, M.E., Jabine, T.B. and Muller, H.J. (1980). Report an exact and statistical matching techniques, Statistical Policy Working Paper 5, U.S. Dept. of Commerce, Washington, D.C.. U.S. Government Printing Office.

Raghunathan, T.E. (2006). Combining Information from Multiple Surveys for Assessing Health Disarities, Allgemeines Statistches Archiv, 90, 515-526.

Raghunathan, T.E., Lepkowski, J.M., Van Hoewyk, J. and Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models, Survey Methodology, Vol. 27, No.1, 85-95.

Raghunathan, T.E., Reiter, J.P. and Rubin, D.B. (2003). Multiple Imputation for Statistical Disclosure Limitation, Journal of Official Statistics, Vol.19, No.1, 1-16.

Raghunathan, T.E., Xie, DW, Schenker, N., Parsons, V.L., Davis, W.W., Dodd, K.W. and Feuer, D.J. (2007). Combining Information from Two Surveys to Estimate County-level Prevalence Rates of Cancer Risk Factors and Screening, Journal of the American Statistical Association, Vol. 102, No. 478.

Rao, J.N.K. (2003). Small Area Estimation, New York: Wiley.

Rao, J. N. K. and Wu, C. F. J. (1988). Resampling inference with complex survey data, Journal of the American Statistical Association, 83(401), 231–241.

Reiter, J.P., Raghunathan, T.E. and Kinney, S.K. (2006). The Importance of Modeling the Sampling Design in Multiple Imputation for Missing Data, Survey Methodologh, Vol. 32, 143-149.

Rodgers, W.L. (1984). An evaluation of statistical matching, Journal of Business and Economic Statistics, 2, 91-102.

Rubin, D.B. (1981). The Bayesian bootstrap, The Annals of Statistics, 9, No. 1, 131-134.

Rubin, D.B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputation, Journal of Business and Economic Statistics, 4, 87-94.

Rubin, D.B. (1987). Multiple Imputation for nonresponse in Surveys, New York, Wiley & Sons.

Rubin, D.B. (1996). Multiple Imputation after 18+ Years. *Journal of the American Statistical Association*, Vol. 91, 473--489.

Rubin, D.B. and Schafer, J.L. (1990). Efficiently creating multiple imputations for incomplete multivariate normal data, Proceeding of the Statistical Coputing Section of the American Statistical Association, 83-88.

Schafer, J.L. (1997). Analysis of Incomplete Multivariate Data by Simulation, New York: Chapman and Hall.

Schenker, N., Gentleman, J.F., Rose, D, Hing, E and Shimizu, I.M. (2002). Combining Estimates from Complementary Surveys: A Case Study Using Prevalence Estimates from National Health Surveys of Households and Nursing Homes, Public Health Reports, Vol 117, 393-407.

Schenker, N., Raghunathan, T.E. and Bondarenko, I. (2009). Improving on Analyses of Self-reported Data in a Large-scale Health Survey by Using Information from an Examination-based Survey, Statistics in Medicine, Vol 29, Issue 5, 533 – 545.

Schenker, N. and Raghunathan, T.E. (2007). Combining Information from Multiple Surveys to Enhance Estimation of Measures of Health, Statistics in Medicine, Vol. 26, 1802-1811.

Scott, A.J. (1977). Large Sample Posterior Distributions in Finite Populations, The Annals of Mathematical Statistics, 42, 1113-1117.

Skinner, D.J., Holt, D. and Smith, T.M.F. (1979). Analysis of Complex Surveys, Journal of the Royal Statistical Society, Series A, 142, 33-46.

Skinner, C., Holt, D., and Smith, T. (1989). Analysis of Complex Surveys, New York: Wiley.

Skinner, N., Raghunathan, T.E. and Bondarenko, I. (2006). A Case Study of Combining Information from An Examination-based Health Survey and An Interview-based Health Survey to Improve on Analyses of Self-reported Data, Working Paper Series, Department of Biostatistics, University of Michigan, Ann Arbor, MI, U.S.A.

Skinner, C.J. and Rao, J.N.K. (1996). Estimation in Dual Frame Surveys with Complex Designs, Journal of the American Statistical Association, Vol. 91, No. 433.

Turner, C.F., Lessler, J.T. and Devore, J. (1992). Effect of Mode of Administration and Wording on Reporting of Drug Use, In Survey Measurement of Drug Use: Methodological Studies, ed, 177-220.