Published in final edited form as:

Surv Methodol. 2014 December ; 40(2): 347–354.

Combining information from multiple complex surveys

Qi Dong,

Google, Inc., 1R4A, Quad 5, Google Inc, 399 N. Whisman Road, Mountain View, CA 94043. qdong@google.com

Michael R. Elliott, and

Department of Biostatistics, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109 and Survey Methodology Program, Institute for Social Research, University of Michigan, 426 Thompson St., Ann Arbor, MI 48106. mrelliot@umich.edu

Trivellore E. Raghunathan

Department of Biostatistics, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109 and Survey Methodology Program, Institute for Social Research, University of Michigan, 426 Thompson St., Ann Arbor, MI 48106. teraghu@umich.edu

Abstract

This manuscript describes the use of multiple imputation to combine information from multiple surveys of the same underlying population. We use a newly developed method to generate synthetic populations nonparametrically using a finite population Bayesian bootstrap that automatically accounting for complex sample designs. We then analyze each synthetic population with standard complete-data software for simple random samples and obtain valid inference by combining the point and variance estimates using extensions of existing combining rules for synthetic data. We illustrate the approach by combining data from the 2006 National Health Interview Survey (NHIS) and the 2006 Medical Expenditure Panel Survey (MEPS).

Keywords

Synthetic populations; Posterior predictive distribution; Bayesian bootstrap; Inverse sampling

1 Introduction

Survey agencies often repeatedly draw samples from similar populations and collect similar variables, sometimes even using the same frame. For example, the National Health Interview Survey (NHIS) and the National Health and Nutrition Examination Survey (NHANES) are both conducted by the U.S. National Center for Health Statistics. These two surveys target the U.S. non-institutionalized population and have a considerable overlap of questions. By combining information from multiple surveys, we hope to obtain more accurate inference for the population than if we use the data from a single survey.

One of the biggest challenges in such combining is the compatibility of multiple data sources. Surveys may use different sampling designs or modes of data collection, which may result in various sampling and nonsampling error properties. Instead of directly pooling the

data from multiple surveys for a simple analysis, we need to adjust for the discrepancies among the data to make them comparable.

Various methods for combining data collected in two surveys have been proposed in the survey methodology literature (Hartley 1974; Skinner and Rao 1996; Lohr and Rao 2000; Elliott and Davis 2005; Raghunathan, Xie, Schenker, Parsons, Davis, Dodd and Feuer 2007; Schenker, Gentleman, Rose, Hing and Shimizu 2002; Schenker and Raghunathan 2007; Schenker, Raghunathan and Bondarenko 2009). The most recent papers by Raghunathan et al. (2007) and Schenker et al. (2009) applied model-based approaches. The basic idea for the model-based approach is to fit an imputation model to the data of better quality and use the fitted model to impute the values in the other samples of lower quality. As long as the imputation model is correctly specified, this approach can take advantage of the strengths of the multiple data sources and improve the statistical inference. However, as suggested by Reiter, Raghunathan and Kinney (2006), when the sample is collected using complex sampling designs, ignoring those features could result in biased estimates from the design-based perspective. However, fully accounting for the complex sampling design features in practice is very difficult. For example, both Raghunathan et al. (2007) and Schenker et al. (2009) used a simplified method to adjust for stratification and clustering. Raghunathan et al. (2007) used a rudimentary concept of design effect and Schenker et al. (2009) used propensity scores to create adjustment subgroups for modeling.

Here we propose a new method for combining multiple surveys that adjusts for the complex sampling design features in each survey. The unobserved population in each survey will be treated as missing data to be multiply imputed. The imputation model will account for complex design features using a recently developed nonparametric synthetic population generation method (Dong, Elliott and Raghunathan 2014). For each survey, the observed data and the multiply imputed unobserved population produce multiple synthetic populations. Once the whole population is generated, the complex sampling design features such as stratification, clustering and weighting will be of no use in the analysis and the synthetic populations can be treated as equivalent simple random samples. Finally, the estimate for the population quantity of interest will be calculated from each synthetic population and then will be combined first within each individual survey and then across multiple surveys.

This paper proceeds as follows: Section 2 summarizes generating synthetic population while accounting for complex sampling design features using the nonparametric approach. Section 3 describes methodology to produce combined estimates from these multiple synthetic populations. In Section 4, we apply the proposed method to combine the 2006 NHIS and the Medical Expenditure Panel Survey (MEPS) to estimate the health insurance coverage rates of the US population. Section 5 concludes with discussion and directions for future research

2 Generating synthetic populations from single survey data that accounts for complex sampling designs

Dong et al. (2014) extended work in the finite population Bayesian bootstrap to develop a non-parametric approach to the generation of posterior predictive distributions. A summary

of the algorithm to draw the l-th of l=1,…,L synthetic populations for stratified, clustered sample designs with unequal probabilities of selection is as follows:

1. Use the Bayesian Bootstrap (BB) (Rubin 1981) to adjust for stratification and clustering. Draw a simple random sample with replacement (SRSWR) of size m_h from the c_h clusters within each stratum h = 1, …, H and calculate bootstrap replicate weights for each of the n_{hi} observations in each cluster as $w^{*(l)}=\left\{w_{hi}^{*(l)}, h=1,\ldots,H, i=1,\ldots,c_h, k=1,\ldots,n_{hi}\right\}$, 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, i = 1, …,c_h is selected. To ensure all the replicate weights are non-negative, $m_h \le (c_h - 1)$; here and below we take $m_h = (c_h - 1)$.

2. Use the finite population Bayesian bootstrap (FPBB) (Lo 1986; Cohen 1997) for unequal probabilities of selection to adjust for unequal probabilities of selection. For each cluster i in stratum h of population size N_{hi}, draw a sample of size $N_{hi} - n_{hi}$, denoted by $\left(y_1^{*},\ldots,y_{N_{hi}-n_{hi}}^{*}\right)$, by drawing y_{hik}^{*} from cluster data (y_1, …, $y_{n_{hi}}$) with probability $\frac{w_{hik}^{*}-1+I_{hik,j-1}*(N_{hi}-n_{hi})/n_{hi}}{N_{c_H}-n_{c_H}+(j-1)*(N_{hi}-n_{hi})/n_{hi}}$, where w_{hik}^{*} is the replicate weight of unit k in cluster i in stratum h, and $I_{hik,j-1}$ is the number of bootstrap selections of y_{hik} among $y_1^{*},\ldots,y_{j-1}^{*}$. Form the FPBB population y_1, …, $y_{n_{hi}}$, $y_1^{*},\ldots,y_{N_{hi}-n_{hi}}^{*}$.

3. Produce F FPBB samples for each BB sample, denoted by S_{l1}, …, S_{lF}, l = 1, …, L. Pool the F FPBB samples to produce one synthetic population, S_l. (Because $N = \Sigma_h \Sigma_i N_{hi}$ may be unrealistically large, generating a sample of size $k * n$ for large k is sufficient.)

3 Combining rule for the synthetic populations from multiple surveys

Assume that $Q = Q(Y)$ is the population quantity of interest depending upon the set of variables Y that are collected in multiple surveys: for example, a population mean, proportion or total, a vector of regression coefficients, etc. For simplicity of exposition we assume Q to be scalar. Assume that, using data from a single survey s, we create L synthetic populations, $S_l^{(s)}$, l = 1, …, L, using the methods summarized in Section 2. Denote $Q_l^{(s)}$ as the corresponding estimate of the population quantity Q obtained from synthetic population l generated using data from survey s (note this estimate can be obtained under a simple random sampling assumption). Dong et al. (2014) shows that, under reasonable asymptotic assumptions (sufficient sample size for the sample quantity of interest to be normally distributed, synthetic populations generated consistent with the survey design),

$$Q|S_1^{(s)},\ldots,S_L^{(s)} \overset{\bullet}{\sim} t_{L-1}\left(\overline{Q}_L^{(s)},\left(1+L^{-1}\right)B_L^{(s)}\right) \quad (3.1)$$

where $\overline{Q}_L^{(s)}=L^{-1}\sum_{l=1}^{L}Q_l^{(s)}$ is the mean of Q across the L synthetic populations and $B_L^{(s)}=(L-1)^{-1}\sum_{l=1}^{L}\left(Q_l^{(s)}-\overline{Q}_L^{(s)}\right)^2$ is the between-imputation variance. The result follows immediately from Section 4.1 of Raghunathan, Reiter and Rubin (2003), and is based on the standard Rubin (1987) multiple imputation combining rules. The average "within" imputation variance is zero, since the entire population is being synthesized; hence the posterior variance of Q is entirely a function of the between-imputation variance.

The combining rule obtained in (3.1) may not yield valid inference for the parameters of interest for multiple surveys, since the models to generate synthetic populations for the multiple surveys may be different. Thus, a new rule for combining estimates across multiple surveys needs to be developed.

3.1 Normal Approximation when L is large

Let $\overline{Q}_L^{(s)}$ and $B_L^{(s)}$ be the combined estimator of the population quantity of interest and its variance for survey s obtained using the combining formulas for synthetic populations $S_{syn}^{(s)}=\left\{S_l^{(s)},l=1,\ldots,L\right\}$, s = 1, …, S in a single survey setting. When L is large, we have

$$Q|S_{syn}^{(1)},\ldots,S_{syn}^{(s)}\dot{\sim}N\left(\overline{Q}_L,B_L\right) \quad (3.2)$$

where $\overline{Q}_L=\sum_{s=1}^{S}(\overline{Q}_L^{(s)}/B_L^{(s)})/\sum_{s=1}^{S}\left(1/B_L^{(s)}\right)$ and $B_L=1/\sum_{s=1}^{S}(1/B_L^{(s)})$. Equation (3.2) follows immediately from standard Bayesian results, assuming that 1) the true variance of $\overline{Q}_L^{(s)}$, B_s, can be approximated by $B_L^{(s)}$ obtained from the synthetic populations as in Section 3, i.e., $(\overline{Q}_L^{(s)}|Q,B_s)=(\overline{Q}_L^{(s)}|Q,B_L^{(s)})\sim N(Q,B_L^{(s)})$, 2) each survey is independent, and 3) Q has a non-informative prior $\pi\left(Q|B_L^{(s)}\right)\propto 1$.

3.2 T-corrected Distribution for Small/Moderate *L*

For small to moderate L, the posterior distribution of Q is better approximated by

$$Q|S_{syn}^{(1)},\ldots,S_{syn}^{(s)}\dot{\sim}\mathrm{t}_{\nu_L}\left(\overline{Q}_L,\left(1+L^{-1}\right)B_L\right) \quad (3.3)$$

where $\bar{Q}_L$ and B_L are defined as in 3.1, and degrees of freedom $\vartheta_L=(L-1)/\sum_{s=1}^{S}\left(\left(1/b_L^{(s)}\right)/\sum_{s=1}^{S}\left(1/b_L^{(s)}\right)\right)^2$. Details are available in Dong (2012), and follow the extensions of Raghuanthan et al. (2003) that were used to derive the large L results.

4 Combined estimates of health insurance coverage from the NHIS, MEPS and BRFSS

The 2006 NHIS and MEPS data are multistage probability samples that incorporate stratification, clustering and oversampling of some subpopulations (e.g., Black, Hispanic, and Asian in later years). For confidentiality reasons the true strata and PSUs are suppressed. The NHIS is released with 300 pseudo-strata and two pseudo-PSUs per stratum; MEPS, which is a subsample of the households which participate in the NHIS, is released with 203 pseudo-strata and up to three pseudo-PSUs per stratum (Ezzati-Rice, Rohde and Greenblatt 2008; National Center for Health Statistics 2007). The NHIS and MEPS ask one randomly-sampled adult in each household whether they are covered by any health insurance and, if so, whether they are covered by private or government insurance. We consider this trinomial distribution of insurance status in the overall adult population, as well as in subpopulations consisting of males, Hispanics, non-Hispanic whites, and non-Hispanic whites earning between $25,000 and $35,000 per year. We delete the cases with item-missing values and focus on our study on the complete cases. This results in 20,147 and 20,893 cases in the NHIS and MEPS data respectively.

The 2006 BRFSS is obtained via random digit dialing (RDD) using list-assisted sampling, stratified by state. While such designs avoid clustering, unequal probability of selection is introduced because the sample size is roughly equal in each state; in addition only one adult is sampled per household. In contrast to the NHIS and MEPS, the BRFSS only asks whether one is insured or not, so we only calculate the proportion of respondents who are not covered by any insurance. 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.

We generate the synthetic populations for the three surveys from 200 BB samples, each consisting of 10 FPBB samples of size $5n$ ($B = 200$, $F = 10$, $k = 5$). We then produce the combined estimates of people's health insurance coverage rates using the combining survey method described above. 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 type insurance they have (private vs. public). For these proportions, we can only combine the NHIS and MEPS. The results are summarized in Table 4.1. 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, 1.5 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.

5 Discussion

In this paper, we propose a new method to combine information from multiple complex surveys. We apply the new method to combine information about health insurance status from the 2006 NHIS, MEPS, and BRFSS. Results show that the combined estimate is more precise compared to the estimates from individual surveys. As previous work has shown (Dong et al. 2014), we have little information loss in the sense that the sampling properties of inferences from the synthetic population and the actual sample are very similar. Thus when we combine the estimates from three samples, the combined estimate is substantially more efficient that the estimates from individual surveys. (We note that this application is primarily for illustrative purposes; similar inferences could be made by computing the design-based estimates and variances for each of the surveys, then applying the combining rule in (3.2) on the design-based estimates.)

This new combining survey method has two major advantages over the existing methods. First, the approach used here to generate synthetic populations, discussed in detail in Dong et al. (2014), accounts for the complex sample design nonparametrically using extensions of finite population Bayesian bootstrap methods. Since the resulting synthetic populations can be analyzed as simple random samples, information from other surveys can be used to adjust for the nonsampling errors and/or filling in the missing variables. Another advantage of this method is it has no limitation on the number of surveys to be combined as long as the surveys have the same underlying population. The proposed method that adjusts for the complex sampling design features can be applied to each survey independently. After the missing information is imputed, regardless the number of surveys to be combined, we only need to combine the estimates from each survey using the combing rule developed in this manuscript. A final advantage of the proposed approach is the ability of the synthetic populations generated by the nonparametric method to 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. We consider this application in future work.

References

Cohen, MP. Proceedings of the Survey Research Methods Section. American Statistical Association; 1997. The Bayesian bootstrap and multiple imputation for unequal probability sample designs; p. 635-638.

Dong, Q. Unpublished PhD thesis. University of Michigan; 2012.

Dong Q, Elliott MR, Raghunathan TE. A nonparametric method to generate synthetic populations to adjust for complex sampling design features. Survey Methodology. 2014; 40(1):29–46.

Elliott MR, Davis WW. Obtaining cancer risk factor prevalence estimates in small areas: combining data from two surveys. Journal of the Royal Statistical Society C: Applied Statistics. 2005; 54:595–609.

Ezzati-Rice, TM., Rohde, F., Greenblatt, J. Methodology Report No 22. Agency for Healthcare Research and Quality; Rockville, MD: 2008. Sample design of the medical expenditure panel survey household component, 1998–2007. Accessed at: http://www.meps.ahrq.gov/mepsweb/data_files/publications/mr22/mr22.pdf [February 2014]

Hartley HO. Multiple frame methodology and selected applications. The Indian Journal of Statistics, C. 1974; 38:99–118.

Lo AY. Bayesian statistical inference for sampling a finite population. Annals of Statistics. 1986; 14:1226–1233.

Lohr SL, Rao JNK. Inference from dual frame surveys. Journal of the American Statistical Association. 2000; 95:271–280.

National Center for Health Statistics. National Center for Health Statistics. Centers for Disease Control and Prevention; Hyattsville, Maryland: 2007. Data file documentation, National Health Interview Survey, 2006 (machine readable data file and documentation). Accessed at: ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Dataset_Documentation/NHIS/2006/srvydesc.pdf [February 2014]

Raghunathan TE, Reiter JP, Rubin DB. Multiple imputation for statistical disclosure limitation. Journal of Official Statistics. 2003; 19:1–16.

Raghunathan TE, Xie DW, Schenker N, Parsons VL, Davis WW, Dodd KW, Feuer DJ. Combining information from two surveys to estimate county-level prevalence rates of cancer risk factors and screening. Journal of the American Statistical Association. 2007; 102:474–486.

Reiter JP, Raghunathan TE, Kinney SK. The importance of modeling the sampling design in multiple imputation for missing data. Survey Methodology. 2006; 32:143–149.

Rubin DB. The Bayesian bootstrap. The Annals of Statistics. 1981; 9:131–134.

Rubin, DB. Multiple Imputation for Nonresponse in Surveys. New York: Wiley; 1987.

Schenker N, Gentleman JF, Rose D, Hing E, Shimizu IM. Combining estimates from complementary surveys: A case study using prevalence estimates from national health surveys of households and nursing homes. Public Health Reports. 2002; 117:393–407. [PubMed: 12477922]

Schenker N, Raghunathan TE. Combining information from multiple surveys to enhance estimation of measures of health. Statistics in Medicine. 2007; 26:1802–1811. [PubMed: 17278184]

Schenker N, Raghunathan TE, Bondarenko I. Improving on analyses of self-reported data in a large-scale health survey by using information from an examination-based survey. Statistics in Medicine. 2009; 29:533–545.

Skinner CJ, Rao JNK. Estimation in dual frame surveys with complex designs. Journal of the American Statistical Association. 1996; 91:349–356.

Table 4.1

Individual and combined estimates 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	

Domain		Actual Data (Complex Design)			Combined Estimates	
	Types	**NHIS**	**BRFSS**	**MEPS**	**NHIS and MEPS**	**NHIS, BRFSS and MEPS**
	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

Surv Methodol. Author manuscript; available in PMC 2017 November 30.