On Bayesian Inference with Complex Survey Data

Editorial

Volume 3 Issue 5 - 2016

Joseph Kang[1]* and Kyle Bernstein[1]

The Centers for Disease Control and Prevention, USA

***Corresponding author:** Joseph Kang, Statistics Team Lead, 1600 Clifton Rd, MS-02, Atlanta, GA, USA, Email: yma9@cdc.gov

Received: April 18, 2016 | **Published:** May 11, 2016

Editorial

Nationally representative probability sample surveys (e.g., the National Health and Nutrition Examination Survey, or NHANES [1], and the National Survey of Family Growth, NSFG [2]) have immense value in helping develop estimates of the prevalence of disease, morbidity, and risk factors [3,4]. These surveys were designed to utilize survey weights to approximate nationally representative parameters. Any survey data analysis that uses the survey weights is called design-based estimation. Such analyses are implemented in major statistical software programs (e.g. SAS [5] and R [6]).

While commonly used, weighting complex survey data is a foreign concept to Bayesian modelers, according to Gelman [7]. A typical Bayesian analyst does not use weights, but focuses instead on updating assumed prior distributions with observed data likelihood. Much of this disconnect may be a function of differing goals; Bayesian approaches are focused on reliable statistical models [8] rather than on assessing the degree to which their estimates are nationally representative or not. However, Bayesian approaches, which have been successfully applied to multilevel data [8], missing data, and measurement errors [9], may represent a natural partner in complex survey data analysis. Measurement errors, missing data, and multilevel variables in complex survey data sets can be all treated as unobserved random variables in the Bayesian framework and they can be assessed by updating assumed prior distributions of related parameters with observed data sets [8,9].

Rod Little proposed a method called "Calibrated Bayes" [10] which can be used to adopt survey weights within the Bayesian paradigm. Originally the concept of calibration was proposed by non-Bayesian statisticians [11]. The calibration technique modifies survey data sets by changing the survey weights to explain nationally representative features. For example, a complex survey data set's average male age is different from that of the U.S. Census Bureau. By changing survey weights, the calibration method matches the average male age of the complex survey data set to the Census Bureau's. Though theoretically the calibration technique can be used for Bayesian methods, the Calibrated Bayes method has not been used in major health science journals, which is partly due to the fact that the Calibrated Bayes method is relatively new and that its theory has been discussed from a statistical point of view [10,12].

Although Bayesian methods are useful in dealing with complex problems, to our knowledge, none of the popular Bayesian software programs (e.g., BUGS, Bayesian inference Using Gibbs Sampling, Cambridge Institute of Public Health [13]) have code related to adopting survey weights. Despite this and compared to the theoretical Calibrated Bayes method, it is still possible to use Bayesian software and survey weights together via the R software program in a relatively simple way.

Consider, for example, an analyst who wants to use NHANES data to estimate a disease prevalence by using established survey weights. NHANES provides complex survey data with multilevel structures having missing data. The analyst can run the BUGS program in R to build Bayesian models on the basis of well-documented examples (Congdon illustrates extensive examples [9]). Depending on types of target parameters, survey weights can be omitted or treated as a fixed variable in the Bayesian analysis. Again, the analyst can obtain general point estimates of disease conditions for NHANES study subjects by modeling multilevel data structures, missing data, and measurement errors. Design-based variance estimation can be done with the Bayesian point estimates using the jackknife method [14] in the R survey package. That is, for each of the jackknife samples, the Bayesian modeling can be performed to produce point estimates. However, it will be computationally burdensome if the Markov Chain Monte Carlo simulation is performed for each of the jackknife samples to assess the posterior means of the parameters. The computational burden can be relieved if a weighted posterior likelihood is maximized to obtain posterior modes instead.

In terms of a statistical formula, let y denote a binary disease condition, w denote the survey weights, and $\theta = P(y=1)$ denote the probability of having the disease which is the estimand of interest. The usual estimate of θ with w is $\hat{\theta} = \Sigma wy / \Sigma w$, where Σ indicates summation over all sampled units. Because y is subject to multilevel data structures, missing data, and measurement error biases, it can be modeled using a Bayesian probability model $P(y = 1 \mid x)$, where x denotes a vector of auxiliary covariates that are associated with y. $\hat{\theta}$ can be re-paramatized by $E(\Sigma wy / \Sigma w \mid x)$, where the mathematical expectation $E(\cdot)$ is taken with respect to the Bayesian model $P(y = 1 \mid x)$. $\hat{\theta}$ can be estimated within each of the jackknife samples.

To summarize this estimation process, only three steps are needed to conduct the Bayesian analysis with survey weights, as follows:

I. Divide data with the jackknife method.

II. Obtain Bayes' point estimates of target estimands (e.g., a disease outcome) for each jackknife sample.

III. Summarize sample mean and sample variance of jackknifed estimates.

As described previously, Bayesian modeling is generally known to be suitable for handling multilevel data structures, missing data, and measurement errors. However, the Bayesian modeling itself does not provide proper variance estimates in the sense of design-based estimation. Using the jackknife resampling method, the Bayesian point estimates can yield design-based variance estimates. Alternatives to the jackknife method are the bootstrap resampling method, balanced repeated replication, and other resampling methods. In this way, analysts can benefit from the Bayesian methodology for multilevel data, missing data, and measurement errors as well as the calibration technique to report nationally representative estimates.

Acknowledgement

None.

Conflict of Interest

There is no financial interest or conflict of interest.

References

1. http://www.cdc.gov/nchs/nhanes.
2. http://www.cdc.gov/nchs/nsfg.htm.
3. Fanfair RN, Zaidi A, Taylor LD, Xu F, Gottlieb S, et al. (2013) Trends in seroprevalence of herpes simplex virus type 2 among non-Hispanic blacks and non-Hispanic whites aged 14 to 49 years--United States, 1988 to 2010. Sex Transm Dis 40(11): 860-864.
4. Xu F, Sternberg MR, Kottiri BJ, McQuillan GM, Lee FK, et al. (2006) Trends in herpes simplex virus type 1 and type 2 seroprevalence in the United States. JAMA 296(8): 964-973.
5. SAS Institute Inc (2008) Cary, North Carolina 27513, USA.
6. http://www.R-project.org/.
7. Gelman A (2007) Struggles with survey weighting and regression modeling. Statistical Science 22(2): 153-164.
8. Gelman A, John B Carlin, Hal S Stern, Donald B Rubin (2003) Bayesian data analysis, second Edition. Chapman and Hall/CRC 690.
9. Congdon P (2006) Bayesian statistical modeling, 2nd ed. John Wiley & Sons 1-573.
10. Little R (2012) Calibrated Bayes, an alternative inferential paradigm for official statistics. Journal of official statistics 28(3): 309-334.
11. Deville JC, Sarndal CE (1992) Calibration Estimators in Survey Sampling. Journal of the American Statistical Association 87(418): 376-382.
12. Little R (2011) Calibrated Bayes, for statistics in general, and missing data in Particular. Statistical Science 26(2): 162-174.
13. Lunn D, Spiegelhalter D, Thomas A, Best N (2009) The BUGS project: Evolution, critique and future directions. Stat Med 28(25): 3049-3067.
14. Canty AJ, Davison AC (1999) Resampling-based variance estimation for labor force surveys. Journal of the Royal Statistical Society. 48(3): 379-391.