Weights for Combining Surveys across Time or Space

Adam Chu
J. Michael Brick
Graham Kalton
Westat
1650 Research Blvd.
Rockville, Maryland, U.S.A. 20850
Chua1@westat.com

1. Introduction

Researchers sometimes combine the data from two or more surveys covering the same subject matter in order to increase sample size and hence the precision of estimates. This situation often arises when a survey is repeated across time and when a national survey is supplemented with local surveys in specific areas. In the United States, issues of combining across time arise in the continuous National Health Interview Survey, the National Health and Nutrition Examination Survey, the Medical Expenditure Panel Survey, the Continuing Survey of Food Intakes by Individuals (CSFII), and the proposed American Community Survey. Issues of combining national and local surveys have occurred with the National Assessment of Educational Progress and the National Adult Literacy Survey (NALS). A different type of combining occurred with the Survey of Income and Program Participation when the survey was a repeated panel survey with overlap of panels. Data collected in 2 (sometimes 3) panels at different waves could then be combined to produce estimates for a specified time point or period.

One approach to combining surveys in analysis is to form a weighted average of the estimates from each of the surveys, with weights that minimize the variance of the weighted average. While this composite estimation approach has theoretical attractions, it is time consuming to implement when many estimates are required. For multi-purpose applications, a more attractive approach is to pool the data from the two or more surveys into a single data base that can be analyzed in a standard way. For this latter approach, weights need to be derived for the units in the pooled data base. This paper discusses the development of such weights using the ideas of composite estimation.

2. Composite Weighting

Let $z = \alpha z_1 + (1-\alpha) z_2$ denote a composite estimator for population parameter Z from two independent surveys with separate estimates z_1 and z_2. If z_1 and z_2 are unbiased estimates of Z, then z is unbiased for any choice of α. If z_i is a linear estimate of the form $\Sigma_j w_{ij} z_{ij}$, then z can be computed by pooling the surveys' data and adjusting the weights to αw_{1j} and $(1-\alpha) w_{2j}$ for the two surveys, respectively. This result also holds approximately for nonlinear statistics provided that $\Sigma_j w_{1j} \approx \Sigma_j w_{2j}$.

The choice of α that minimizes the variance $V(z)$ of z is $\alpha_0 = V(z_2)/[V(z_1)+V(z_2)]$, which is applicable when the estimates are unbiased. (Mean square errors may be used in place of variances for biased estimates.) Let $n_i^* = n_i / D^2(z_i)$ be the "effective sample size" of survey i, based on a sample of size n_i and a design effect of $D^2(z_i)$. Then the optimal value of α is $\alpha_0 = n_1^* / (n_1^* + n_2^*)$. If $D^2(z_1) = D^2(z_2)$, then $\alpha_0 = n_1/(n_1+n_2)$, but in practice this condition is often not satisfied. When both samples are epsem samples, then the choice of $\alpha = n_1/(n_1+n_2)$ produces weights that correspond to those obtained from the alternative approach of deriving pooled weights based on the inverses of selection probabilities of units being in either sample (more strictly, on the inverses of the expected number of selections).

Since different estimates have different design effects, and the sample sizes for different subclass estimates for the two surveys may not be in the same proportion as the total sample sizes, no choice of α can be optimal for all analyses. Rather, a compromise α needs to be sought that is satisfactory at least for most purposes. The proportionate increase in variance from using a compromise α rather than the optimal α_0 is $(\alpha-\alpha_0)^2/\alpha_0(1-\alpha_0)$. This increase is generally not large when α is reasonably close to α_0, and α_0 is not very small or large. When two surveys have some strata, or combined strata, in common, it may well be that different compromise α's should be chosen for different strata.

Composite weighting over space was employed in the NALS, which was conducted in 1992 with a national sample of 13,600 adults who completed literacy tasks. In 11 states, supplemental samples of about 1,000 adults were surveyed to provide state level estimates. Both national and state samples were area probability designs, but they had somewhat different designs. Composite weights were developed to improve both national and state estimates. Different compositing factors (α's) were created for demographic subgroups and sample design features (certainty PSUs) based on the general ideas of the approach described above. Burke *et al.* (1994) present a table of the optimal compositing factors for a range of test scores and demographic subgroups for one state. The optimal factors varied from 0.11 to 0.64, indicating that no single compositing factor could be optimal for all estimates.

For the CSFII 1994-96, nationally-representative samples of persons were selected for each of the three years of the study using an area probability sampling design. The study was designed to provide separate annual estimates of mean food and nutrient intakes for specified analytic domains, as well as combined estimates for the three years. Assuming no changes in food consumption patterns over the period, a composite estimate can be formed by pooling the three individual annual estimates as $z = \Sigma\alpha_i z_i$, where z_i is the estimated mean intake for year i, and where $\Sigma\alpha_i = 1$. For most of the analytic domains, the design effects for annual estimates are roughly equal; hence, the optimal α's are approximately proportional to the corresponding sample sizes, n_i. In general, z is not an unbiased estimate of the overall three-year mean if there are appreciable changes over time. When this is the case, the three-year mean may be estimated by $\bar{z} = \Sigma W_i z_i$, where $W_i = N_i/\Sigma N_i$ and N_i is the size of the population in year i (for all practical purposes, $W_i = 1/3$). The combined estimate $\bar{z}$ is unbiased for the three-year mean whether or not the variable of interest changes over time, but it is less precise than the pooled (composite) estimate if there is no change over time. Since the assumption of no change cannot be made for many CSFII items, estimates similar to the combined estimate, $\bar{z}$, were used. This was accomplished by combining the annual samples and poststratifying the existing weights to a common set of control totals using a raking procedure (Chu and Goldman, 1997).

REFERENCES

Chu, A. and Goldman, J. (1997). Weighting procedures for USDA's Continuing Survey of Food Intakes by Individuals 1994-96. Proceedings of the Section on Survey Research Methods, American Statistical Association, 802-807.

Burke, J., Mohadjer, L., Green, J., Waksberg, J., Kirsch, I., and Kolstad, A. (1994). Composite estimation in national and state surveys. Proceedings of the Section on Survey Research Methods, American Statistical Association, 873-878.

RÉSUMÉ

Il arrive que les chercheurs combinent les données provenant de deux ou plusieurs enquêtes en une base de données unique aux fins d'analyse. Cet article décrit comment il est possible d'établir des poids d'échantillon pour une telle base de données en utilisant une approche fondée sur l'estimation composite.