Item-Specific Weights in Multipurpose Surveys

Hee-Choon Shin, NORC
55 E. Monroe St., Suite 4800, Chicago, IL 60603

Key Words: Nonresponse, Weighting, Imputation, Multipurpose Surveys, Adjustment Cells

1. Introduction

In most surveys information is collected on more than one item. In fact, interestingly enough, the multiple characteristics or items of interest is regarded as an important dimension of a modern complex sample survey (Wolter, 1985: 2-3). However, sampling methods were and are being developed around the framework of an single item. For example, the classic *Neyman allocation* in stratified sampling is to minimize the variance of an item (Neyman, 1934). In general, the best allocation for one item is not best for another. Even though the problem of allocation with more than one item has been considered by Cochran (1977) and Kish (1961, 1976, 1988), still the Neyman allocation is a theory for an item.

Most surveys are designed to be multipurpose. As a specific example, the primary purpose of the Current Population Survey (CPS) is to estimate the national unemployment rate. However, the unemployment rate is just one piece of a vast amount of information available on the employed, unemployed, and persons not in the labor force. Even if we restrict the CPS only to the estimation of national unemployment rate, we need more than one item to determine the labor force status of an individual. Unemployed persons, for example, are "those who are without work, available for work, and actively seeking work" (Plewes, 1994). To determine the unemployment status of an individual, at a minimum, we need 3 items. In other words, simply it is not possible to design an one-item survey to estimate the national unemployment rate. In addition to the cost and the measurement problem, analytical research objective is another important determinant of multipurpose nature of a modern survey. In many cases, we are not interested in simple description of national unemployment rate. Frequently, we are asked to examine the relationships between the unemployment rate and other relevant factors (or covariates). Due to the non-experimental nature of survey research, many covariates need to be controlled to estimate the independent effect of a factor.

Nowadays most surveys suffer from nonresponse (Bradburn, 1992). About fifty years ago, the problem of nonresponse was limited to the mail questionnaire (Hansen and Hurwitz, 1946). The nonresponse problem has been worsened by the multipurpose nature of major surveys. The problem of nonresponse is not only a matter of *who responds* but also a matter of *on which item* in a survey.

In the presence of non-cooperation or item nonresponse, the most popular technique for improving the estimates based on responding units is to weight the responding units' data to compensate for the nonresponding units' missing data. In other words, the weighting factor for the estimator of a variable or item needs to be adjusted for nonresponse. In practice, the final weights in major surveys does only reflect adjustment for non-cooperation or unit nonresponse, which arises when whole questionnaires are missed because of noncontact, refusal, or some other reasons. In fact, Cochran (1977: 359) uses the term nonresponse to refer to "the failure to measure some of the units in the selected sample." However, our ultimate interest is in the estimation of a specific item or a combination of items. Therefore, the final nonresponse adjustment factor for relevant items should reflect the effect of item nonresponses, where information on particular items in the questionnaire are missing.

In principle, item-specific weights should be provided for all the items in a survey. Each unit should have a different weight for each item (Rubin, 1987: 8). The practical problem is in the fact that the number of items in major surveys approaches several hundreds or thousands. In the following, an alternative approach is considered.

2. Algebra of Item-Specific Weights

Consider a sample survey of I individuals with J items. The Horvitz-Thompson estimator of the population total for jth item (or variable) is

$$\hat{Y}_{HT,j}=\sum_{i=1}^{N}\frac{y_{ij}I_iR_iR_{ij}}{\pi_i}, \qquad (1)$$

where

y_{ij} = measurement for the *i*th individual of the *j*th item,

π_i = probability that the *i*th individual is in the sample,

$I_i = 1$ if the *i*th individual in the sample, and $I_i = 0$ otherwise,

$R_i = 1$ if the *i*th individual cooperates in survey, and $R_i = 0$ otherwise,

$R_{ij} = 1$ if the *i*th individual responds to *j*th item, and $R_{ij} = 0$ otherwise,

N = population size.

The sample size is $\sum_{i}^{N} I_i = n$. R_i is the indicator for cooperation (or unit nonresponse), and R_{ij} is the indicator for item nonresponse.

In the absence of both non-cooperation and item nonresponse, the Horvitz-Thompson estimator of the population total for the *j*th item is

$$\hat{Y}_{HT,j}=\sum_{i=1}^{n} w_i y_{ij}, \qquad (2)$$

where $w_i=1/\pi_i$. By design, w_i is not a random variable but a known quantity. However, in large-scale multipurpose surveys, it is extremely difficult to collect complete information from all the sampled individuals and on every survey item. Now, suppose that first m ($\leq$ n) individuals cooperate in the survey and respond to all the survey items, i.e.,

$$\sum_{i}^{N} R_i = \sum_{i}^{N} R_{ij} = m. \qquad (3)$$

The Horvitz-Thompson estimator of the population total for the *j*th item can be decomposed into two components, i.e.,

$$\hat{Y}_{HT,j}=\sum_{i=1}^{m} w_i y_{ij} + \sum_{i=m+1}^{n} w_i y_{ij}. \qquad (4)$$

The first term of the right-hand side in (4) is the weighted sum of measurements on *j*th item for m cooperating individuals, and the second term of the right-hand side in (4) is the weighted sum of measurements for (n-m) non-cooperating individuals. Weighting (instead of imputation) is usually used to handle this kind of non-cooperation or unit nonresponse. Unobserved units are omitted from the sample, and the sampling weights (w_i) for cooperating units are adjusted for the non-cooperation. Using the sampling or background variables, define an adjustment cell variable C that takes value c for all individuals in cell c. The population cooperation rate in cell c is

$$\phi_c = \frac{M_c}{N_c}, \qquad (5)$$

where M_c is the number of individuals that cooperate if sampled in cell c and N_c is the number of individuals. In practice the population cooperation rate is not known. The ϕ_c is estimated from the sample. The estimated cooperation rate, $\hat{\phi}_c$, in cell c is

$$\hat{\phi}_c = \frac{\sum_{k=1}^{m_c} w_k}{\sum_{k=1}^{n_c} w_k}, \qquad (6)$$

where m_c is the number of cooperating individuals in cell c and n_c is the number of sampled individuals in cell c. If all the w_i's are equal within cell c, the $\hat{\phi}_c = m_c / n_c$. Now, the Horvitz-Thompson estimator of the population total for the *j*th item is

$$\hat{Y}_{HT,j}=\sum_{i=1}^{m} w_i' y_{ij}, \qquad (7)$$

where

$$w_i' = w_i\hat{\phi}_c^{-1}. \qquad (8)$$

Now, suppose that first m ($\leq$ n) individuals cooperate in the survey and further that first m_j ($\leq$ m) responds to the *j*th item among m cooperating individuals. That is

$$\sum_{i}^{N} R_i = m, \text{ and } \sum_{i}^{N} R_{ij} = m_j. \qquad (9)$$

In most surveys, there are complex skip patterns. However, we assume no skip patterns for the sake of argument in this paper. We assume that every sampled individual is supposed to respond to every item. Now, the Horvitz-Thompson estimator of the population total for the *j*th item can be decomposed into three components, i.e.,

$$\hat{Y}_{HT,j} = \sum_{i=1}^{m_j} w_i y_{ij} + \sum_{i=m_j+1}^{m} w_i y_{ij} + \sum_{i=m+1}^{n} w_i y_{ij}. \qquad (10)$$

The second term of the right-hand side in (10) is the weighted sum of measurements for ($m-m_j$) cooperating-but-nonresponding individuals. Now the item-specific response rate should reflects the effect of ($n-m_j$) nonresponses instead of that of (n-m) noncooperations. The estimated response rate, $\hat{\phi}_{c,j}$, for item j in cell c is

$$\hat{\phi}_{c,j} = \frac{\sum_{k=1}^{m_{c,j}} w_k}{\sum_{k=1}^{n_c} w_k}, \qquad (11)$$

where $m_{c,j}$ is the number of responding individuals on item j in cell c and n_c is the number of sampled individuals in cell c. The Horvitz-Thompson estimator of the population total for the *j*th item is

$$\hat{Y}_{HT,j} = \sum_{i=1}^{m_j} w'_{i,j} y_{ij}, \qquad (12)$$

where

$$w'_{i,j} = w_i \hat{\phi}_{c,j}^{-1}. \qquad (13)$$

If J is small, the item-specific weights could be included in the data file. However, there are several hundreds or thousands of items in a large-scale survey. Practically, it is not desirable to provide several hundreds or thousands of weights in a data file.

Another argument against providing all the item-specific weights in a data file is in the fact that the determination of nonresponse status of each item is not that straightforward. In many social surveys, one of the response category is *Don't Know* (DK). A common practice is to treat those DKs as nonresponse. However, DKs in attitudinal items are very different from DKs in demographic or behavioral items. For example, DKs in respondents' age can be safely regarded as nonresponse or missing value. Meanwhile, DKs in attitudinal items (e.g., abortion, capital punishment, residential segregation) might be regarded as a valid response category (Clogg, 1982, 1984). In other words, ultimately nonresponse status of an item should be determined by each substantive user or analyst.

Within an adjustment cell c, the estimated item-specific response rate is equivalent to the product of the estimated cooperation rate (or unit response rate) and the estimated conditional response rate given cooperation, i.e.,

$$\hat{\phi}_{c,j} = \hat{\phi}_c(\hat{\phi}_{c,j}|R_i=1), \qquad (14)$$

where

$$(\hat{\phi}_{c,j}|R_i=1) = \frac{\sum_{k=1}^{m_{c,j}} w_k}{\sum_{k=1}^{m_c} w_k}. \qquad (15)$$

Now the item-specific weight, $w'_{i,j}$, is equivalent to

$$w'_{i,j} = w_i \hat{\phi}_c^{-1}(\hat{\phi}_{c,j}|R_i=1)^{-1} \qquad (16)$$

$$= w'_i(\hat{\phi}_{c,j}|R_i=1)^{-1} \qquad (17)$$

Now consider a typical situation where $w'_i = w_i \hat{\phi}_c^{-1}$, weights adjusted for non-cooperation is provided in the data file. Additionally let us assume that the adjustment cell C is provided. *Within an adjustment cell c*, we have the following equality:

$$\frac{\sum_{k=1}^{m_{c,j}} w'_k}{\sum_{k=1}^{m_c} w'_k} = \frac{\hat{\phi}_c^{-1} \sum_{k=1}^{m_{c,j}} w_k}{\hat{\phi}_c^{-1} \sum_{k=1}^{m_c} w_k}. \qquad (18)$$

Accordingly, we can calculate $(\hat{\phi}_{c,j}|R_i=1)$ using w_i' instead of w_i.

3. Conclusion

In a modern multipurpose survey with several hundreds or thousands of items, nonresponse (unit and item) is the norm rather than an exception. In particular, the nonresponse problem has been worsened by the multipurpose nature of major surveys.

Traditionally, the weighting method is utilized to deal with the problem of non-cooperation or unit nonresponse. Imputation is the popular way to deal with the item nonresponse. The quality of imputation completely depends on the goodness of a chosen model. As long as the model is good, the imputation is our magical medicine to cure the disease of nonresponse. However, can we find a reasonable model for each of the items in a major survey? In fact, if the answer to this question were affirmative, we would not have to worry about nonresponse problem. Further, we would not need to implement costly surveys; we just need to collect information from a small number of individuals. The model would take care of the rest.

Item-specific weights are an alternative approach to imputation with *questionable* models. It is not desirable to provide several hundreds or thousands of item-specific weights in a data file, especially because of the subjective nature of the nonresponse status or missingness of attitudinal items.

The item-specific weights can be easily calculated without massive efforts, if we are provided with the following two things: 1) weights adjusted for the non-cooperation, and 2) adjustment cells used for non-cooperation adjustment. Usually, the weights adjusted for non-cooperation are provided in the data files for major surveys. Let us add an additional item, the adjustment cells, to the data file!

REFERENCES

Bradburn, Norman M. 1992. "Presidential address: A response to the non-response problem." *Public Opinion Quarterly* **56**: 391-398.

Clogg, Clifford C. 1982. "Using association models in sociological research: Some examples." *American Journal of Sociology* **88**: 114-134.

Clogg, Clifford C. 1984. "Some statistical models for analyzing why surveys disagree." In C. F. Turner and E. Martin (eds.). *Surveying Subjective Phenomena, Vol. 2.* New York: Sage.

Cochran, William G. 1977. *Sampling Techniques, 3rd ed.* New York: Wiley.

Hansen, Morris H. and William N. Hurwitz. 1946. "The problem of non-response in sample surveys." *Journal of American Statistical Association* **41**: 517-529.

Kish, Leslie. 1961. "Efficient allocation of a multipurpose sample." *Econometrica* **29**: 363-385.

Kish, Leslie. 1976. "Optima and Proxima in linear sample designs." *Journal of Royal Statistical Society, Ser. A*, **139**: 80-95.

Kish, Leslie. 1988. "Multipurpose sample designs." *Survey Methodology* **14**: 19-32.

Neyman, Jerzy. 1934. "On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection (with discussion)." *Journal of Royal Statistical Society* **97**: 558-625.

Plewes, Thomas J. 1994. "Federal agencies introduce redesigned Current Population Survey." *Chance* **7**: 35-41.

Rubin, Donald B. 1987. *Multiple Imputation for Nonresponse in Surveys.* New York: Wiley.

Wolter, Kirk M. 1985. *Introduction to Variance Estimation*. New York: Springer-Verlag.

ACKNOWLEDGEMENT

I am grateful to the late Clifford C. Clogg for his guidance and comments.