〈Research Paper〉

Journal of the Korean
Statistical Society
Vol. 12, No. 1, 1983.

Estimation and Variance Estimation for the U.S. Consumer Expenditures Surveys Redesign Research

Jong-Ik Kim*

ABSTRACT

After every decennial census in the U.S., national surveys such as the Consumer Expenditures surveys are redesigned. The redesigned samples will be multi-stage systematic samples. Many sampling schemes have been proposed for comparison which requires the estimation and variance estimation formula. This paper deals with the estimation and variance estimation procedure only for the Consumer Expenditures(CE) surveys redesign research which concerns the sample design within the Primary Sampling Unit (PSU). In constructing the estimators it deals with the problem of which first stage inflation factor to use. The expected value of the proposed estimators is also derived.

I. Introduction

After every decennial census in the U.S., national surveys such as the Consumer Expenditures(CE) surveys conducted by the Bureau of the Census have been redesigned. The purpose of the redesign is to reflect in the newly designed surveys the changes in the demographic and housing characteristics of the nation and subnational areas occurred during the ten year span between two censi. The redesign research involves re-examination of various aspects of the surveys, i.e., stratification and sampling of Primary Sampling Unit's(PSU's), identification of stratifiers, determination of strata and sorting units within the selected PSU's, and rotation and "phase-in" schemes, etc. In the case of rotation schemes, currently 4-8-4 system is used, that is, a sample housing unit is in the sample for four consecutive months, out of the sample for the following eight

* Visiting Associate Professor of Applied Statistics, Yonsei University and Mathematical Statistician, U.S. Bureau of the Census

months, returns to the sample the next four months and then retires from the sample for good. In place of 4-8-4 system, 3-9-3 system has been proposed and examined in detail at the Bureau.

Once new sample is designed and selected, a question arises as to how to introduce the new sample, i.e., whether to replace the old sample units by the new ones all at once or replace them portion by portion is the "phase-in" problem, which has been studied carefully at the Bureau. These are two examples of the redesign research the Bureau has conducted.

The redesign research concerning the sampling schemes is carried out in roughly three steps:

Step 1. Propose viable sampling schemes at the PSU and within PSU level;

Step 2. Come up with appropriate estimators and variance estimators applicable to each sampling scheme;

Step 3. Apply the schemes on the most recent decennial census data, calculate the variance for each scheme and choose the best scheme.

This paper deals with estimation and variance estimation mentioned in step 2 for the U.S. Consumer Expenditures surveys redesign sampling schemes within PSU only. It provides only the theoretical background for the empirical comparison described in step 3 which is being carried out at the Bureau.

It should be mentioned that due to impossibility of calculating bias or conditional bias, mean squared error or conditional mean squared error can not be used as a criterion of selecting one scheme over others.

Ⅱ. Preliminaries and Sampling Situations

During the census, some items were asked at every household in the nation and others only at the sampled households. The sample part of the census, i.e., census sample provided more detailed information on the nation and subnational areas, and thus ample ground for the experimentation of the redesign research.

The C.E. surveys redesign research involved selection of a subsample from the census sample using a variety of sampling schemes which are based on the systematic sampling with probability proportional to a measure of size (PPS). The schemes can be classified into two broad categories: one is the unit sampling and the other the non-compact cluster

sampling of size 2.[1]

Parapharizing, the experimental survey is a two-stage sampling: i) in the first stage, census sample was selected using systematic sampling and ii) in the second stage, the experimental CE sample was selected using PPS systematic sampling.

For exposition, define

M : number of units[2]/clusters of the census based on the complete count;

m : number of units/clusters selected in the census sample;

n : number of units/clusters selected in the CE sample;

w_i : census weight assigned to the i-th unit/cluster;

y_i : value of the estimation variable y for the i-th unit/cluster;

w_{ij} : census weight assigned to the j-th unit in the i-th cluster, where $\sum_{j}^{J_i} w_{ij}=w_i$, $J_i=1$ or 2 depending on the size of cluster i;

y_{ij} : value of the estimation variable y for the j-th unit in the i-th cluster, where $\sum_{j}^{J_i} y_{ij}=y_i$;

$\Pi_i=\frac{nw_i}{M}$: relative frequency that the i-th unit/cluster is selected in the CE sample;

R_i : Expected sampling rate of the Enumeration District(ED) from which the i-th unit/cluster was selected for the CE sample; and

SR_i : observed sampling rate corresponding to R_i.

Ⅲ. Estimation of Population Total

Ⅲ.1 Estimation of Population Total

Estimation formula for population total is found for the cluster sample case which can be easily extended to the unit sample case. Horvitz-Thompson estimator($\hat{Y}$) of population total Y is

$$\hat{Y}=\sum_{i=1}^{n}\frac{y_i}{\rho_i}$$

where ρ_i is the selection probability of the i-th cluster, i.e.,

$\rho_i=\Pr(i\text{-th cluster units in census sample})\ \Pr(i\text{-th cluster units in CE sample}$

1) Sometimes the cluster was of size 1, which was very rare.

2) Unit means the housing unit or household.

i-th cluster units in the census sample)

Assuming that the cluster is of size 2 and denoting each unit of the cluster by i_1 and i_2, respectively, ρ_i can be re-expressed as

Pr(i_1 and i_2 are in the census sample) · Pr(i_1 and i_2 are in CE sample | i_1 and i_2 are in the census sample). ……(1)

The second factor in the above equation is the second stage sampling rate. Note that before the cluster sampling was performed, two units within an ED were combined to form a cluster. Clusters within the PSU[3] were sorted by some criteria and then selected in the CE sample using systematic sampling with probability proportional to the weight size of the cluster. Thus, the second factor is

$$\Pi_i = \frac{nw_i}{M}.$$

The first factor of equation (1) can be further rewritten as

Pr(i_1 and i_2 are in the census sample) = Pr(i_1 is in the census sample) ×

Pr(i_2 is in the census sample | i_1 is in the census sample) ……(2)

In systematic sampling, given a specific order of units on the sampling frame, once the first unit, i.e., random start, and the sampling interval are determined, next units are automatically determined. Hence, the second factor in equation (2) is 1. The first factor of (2) is R_i as defined in section II. Note that R_i was calculated using all units on the sampling frame which were actually either observed or unobserved during the census. However, the CE sample is selected from the observed units only. Thus, the use of R_i is inappropriate in this situation. One approach which can overcome this defficiency of R_i is use the observed sampling rate SR_i. Another approach is based on the census sample weight, w_i or w_{ij}, assigned to the census sample units. w_i's or w_{ij}'s were calculated by i) inflating each sample unit by the inverse of the observed sampling rate for the ED to which the unit belongs and ii) putting it in a matrix, i.e., weighting matrix, and then adjusting the cell counts of the matrix to make them conform to certain control counts (1). Thus, in general, w_i or w_{ij} is not the same as SR_i. The value of w_i or w_{ij} takes into account the differential sampling rate for each subgroup within the ED. Two approaches are available for using w_{ij}. One is use w_{i1} and w_{i2} separately to inflate the first and second units of the cluster, respectively, and the other is use the mean weight, $\bar{w}_i = (w_{i1} + w_{i2})/2$, to weight up the whole cluster.

3) For the CE, a county is defined as a PSU.

In short, four forms of ρ_i or ρ_{ij} can be derived:

i) $\rho_i=R_i\pi_i$,

ii) $\rho_i=SR_i\pi_i$,

iii) $\rho_i=\pi_i/\bar{w}_i$,

iv) $\rho_{ij}=\pi_i/w_{ij}$.

Thus, four forms of $\hat{Y}$ corresponding to the above ρ_i's or ρ_{ij} can be obtained as follows:

i) $\hat{Y}_1=\sum_{i=1}^{n}\frac{y_i}{R_i\pi_i}$,

ii) $\hat{Y}_2=\sum_{i=1}^{n}\frac{y_i}{SR_i\pi_i}$,

iii) $\hat{Y}_3=\sum_{i=1}^{n}\frac{y_i\bar{w}_i}{\pi_i}$,

iv) $\hat{Y}_4=\sum_{i=1}^{n}\frac{y_{ij}w_{ij}}{\pi_i}$.

Ⅲ.2 Expected Value of the Estimators

Let

$$t_i=\begin{cases}1, & \text{if the } i\text{-th unit is in the } CE \text{ sample,}\\ 0, & \text{otherwise,}\end{cases}$$

and

$$d_i=\begin{cases}1, & \text{if the } i\text{-th unit is in the census sample,}\\ 0, & \text{otherwise.}\end{cases}$$

Then t_i and d_i follow the binomial distribution for a sample of size 1 with probability π_i and R_i, respectively.

$$E(\hat{Y}_1\,|\,\text{census})=E\left(\sum_{i=1}^{n}\frac{y_i}{R_i\pi_i}\;\Big|\text{census}\right)$$

$$=E\left(\sum_{i=1}^{m}\frac{y_i}{R_it_i}\;\Big|\text{census}\right)$$

$$=\sum_{i=1}^{m}\frac{y_i}{R_i\pi_i}\;E(t_i|\text{census})=\sum_{i=1}^{m}\frac{y_i}{R_i}\ ,$$

which follows since $E(t_i|\text{census})=\pi_i$.

$$E(\hat{Y}_1)=E\left[\;E(\hat{Y}_1|\text{census})\right]=E\left(\sum_{i=1}^{m}\frac{y_i}{R_i}\right)$$

$$=E\left(\sum_{i=1}^{M}\frac{y_i}{R_i}d_i\right)=\sum_{i=1}^{M}\frac{y_i}{R_i}E(d_i)$$

$$=\sum_{i=1}^{M}y_i=Y.$$

Thus, $\hat{Y}_1$ is an unbiased estimator of Y, as Horvitz and Thompson established.

The other estimators are all biased because of the first stage inflation factors which are different from $1/R_i$ in $\hat{Y}_1$. That is,

$$E(\hat{Y}_2)=\sum_{i=1}^{M}\frac{y_iR_i}{SR_i},$$

$$E(\hat{Y}_3)=\sum_{i=1}^{M}y_iw_iR_i,$$

$$E(\hat{Y}_4|\text{census})=E\left(\sum_{i=1}^{m}\sum_{j=1}^{J_i}\frac{y_{ij}w_{ij}}{\pi_i}t_i\Big|\text{census}\right)$$

$$=\sum_{i=1}^{n}\frac{E(t_i|\text{census})}{\pi_i}\sum_{j=1}^{J_i}y_{ij}w_{ij}=\sum_{i=1}^{n}\sum_{j=1}^{J_i}y_{ij}w_{ij},$$

where $J_i=1$ or 2 depending on the size of the i-th cluster.

$$E(\hat{Y}_4)=E\left(\sum_{i=1}^{m}\sum_{j=1}^{J_i}y_{ij}w_{ij}\right)=\sum_{i=1}^{M}\sum_{j=1}^{J_i}y_{ij}w_{ij}R_i$$

In the case of unit sampling, $\hat{Y}_3$ and $\hat{Y}_4$ reduce to the same estimator, i.e.,

$$\hat{Y}_3=\sum_{i=1}^{n}y_i\frac{M}{n}.$$

It is interesting to note that $\frac{M}{n}$ is the sampling interval (SI) in the second stage sampling and $\hat{Y}_3$ looks like a usual unbiased estimator of Y. However, in comparing $\hat{Y}_3$ with $\hat{Y}_1$, it is evident that $\hat{Y}_3$ is not unbiased.

Ⅳ. Variance Estimation

It is well-known that the variance of an estimator obtained from a two-stage sampling has the following form:

$$V(\hat{Y})=V[E(\hat{Y}|\text{census})]+E[V(\hat{Y}|\text{census})] \quad \cdots\cdots(3)$$

where the first term is the "between the census sample component" and the second term is the "within the census sample component." In this paper, we will obtain the variance

formula for the "within" component.

As mentioned before, $\hat{Y}_1$ is not proper for this empirical study, hence only for $\hat{Y}_2$, $\hat{Y}_3$ and $\hat{Y}_4$ variance formula will be derived.

Ⅳ.1 Variance of $\hat{Y}_2$

For finding the variance of the estimators, Hartley's(2) approach will be employed.[4]

$$V(\hat{Y}_2|\text{census})=E\left[\left\{\sum_{i=1}^{n}\frac{y_i}{SR_i\pi_i}-E\left(\sum_{i=1}^{n}\frac{y_i}{SR_i\pi_i}\Big|\text{census}\right)\right\}\Big|\text{census}\right]^2 \quad \cdots\cdots(4)$$

However,

$$E\left(\sum_{i=1}^{n}\frac{y_i}{SR_i\pi_i}\Big|\text{census}\right)=E\left(\sum_{i=1}^{m}\frac{y_i}{SR_i\pi_i}t_i\Big|\text{census}\right)$$

$$=\sum_{i=1}^{m}\frac{y_i}{SR_i}\ ,$$

which follows from $E(t_i)=\pi_i$.

Hence, (4) can be re-expressed as

$$E\left[\left\{\sum_{i=1}^{n}\frac{y_i}{SR_i\pi_i}-\sum_{i=1}^{m}\frac{y_i}{SR_i}\right\}\Big|\text{census}\right]^2. \quad \cdots\cdots(5)$$

To evaluate the above expectation, define

s_k : the probability that the k-th CE sample is selected,

L : total number of all possible CE samples.

Then (5) can be further re-expressed as

$$\sum_{k=1}^{L}s_k\left[\left(\sum_{i=1}^{n}\frac{y_i}{SR_i\pi_i}-\sum_{i=1}^{m}\frac{y_i}{SR_i}\right)\Big|\text{census}\right]^2$$

$$=\sum_{k=1}^{L}s_k\left(\sum_{i=1}^{n}\frac{y_i}{SR_i\pi_i}\Big|\text{census}\right)^2-\left(\sum_{i=1}^{m}\frac{y_i}{SR_i}\right)^2. \quad \cdots\cdots(6)$$

Define

$$t_i^{(k)}=\begin{cases}1, & \text{if the } i\text{-th unit is in the } k\text{-th } CE \text{ sample,}\\ 0, & \text{otherwise.}\end{cases}$$

The first term of equation (6) can be re-expressed as

4) For deriving the variance formula, Horvitz-Thompson approach can also be taken. For the equivalence between those two, see my "A Note on Equivalency between Hartley's Variance Formula and Horvitz-Thompson's in the Case of Systematic Sampling with Probability Proportional to A Measure of Size and without Replacement", Yonsei Business Review, vol. 20, No. 2, 1983.

$$\sum_{k=1}^{L} s_k\left[\left(\sum_{i=1}^{n}\frac{y_i}{SR_i\pi_i}\Big|\text{census}\right)\right]^2=\sum_{k=1}^{L} s_k\left(\sum_{i=1}^{m}\frac{y_i}{SR_i\pi_i}t_i^{(k)}\Big|\text{census}\right)^2$$

$$=\sum_{k=1}^{L} s_k\left(\sum_{i=1}^{m}\frac{y_i^2}{SR_i^2\pi_i^2}t_i^{(k)}\Big|\text{census}\right)$$

$$+\sum_{k=1}^{L} s_k\left(\sum_{i\neq j}^{m}\frac{y_iy_j}{SR_iSR_j\pi_i\pi_j}t_i^{(k)}t_j^{(k)}=\sum_{i=1}^{m}\frac{y_i^2}{SR_i^2\pi_i^2}\sum_{k=1}^{L}s_kt_i^{(k)}\right.$$

$$+\sum_{i\neq j}^{m}\frac{y_iy_j}{SR_iSR_j\pi_i\pi_j}\sum_{k=1}^{L}s_k\ t_i^{(k)}t_j^{(k)}=\sum_{i=1}^{m}\frac{y_i^2}{SR_i^2\pi_i}+\sum_{i\neq j}^{m}\frac{y_iy_j\pi_{ij}}{SR_iSR_j\pi_i\pi_j},$$

which follows since $\sum_{k=1}^{L}s_k\ t_i^{(k)}=\pi_i$ and $\sum_{k=1}^{L}s_kt_i^{(k)}t_j^{(k)}=\pi_{ij}$,

where π_{ij} is the joint probability that both i-th and j-th units are selected in the CE sample.

In short, the conditional "within-census sample component" of the variance of $\hat{Y}_2$ is,

$$\sum_{i=1}^{m}\frac{y_i^2}{SR_i^2\pi_i}+\sum_{i\neq j}^{m}\frac{y_iy_j\pi_{ij}}{SR_iSR_j\pi_i\pi_j}-\left(\sum_{i=1}^{m}\frac{y_i}{SR_i}\right)^2.$$

Ⅳ.2 Variance of $\hat{Y}_3$

$$V(\hat{Y}_3|\text{census})=\left[\left\{\sum_{i=1}^{n}\frac{y_i\bar{w}_i}{\pi_i}-E\left(\sum_{i=1}^{n}\frac{y_i\bar{w}_i}{\pi_i}\Big|\text{census}\right)\right\}\Big|\text{census}\right]^2, \qquad \cdots\cdots(7)$$

where $\bar{w}_i$ is the average weight for the i-th cluster. However,

$$E\left(\sum_{i=1}^{n}\frac{y_i\bar{w}_i}{\pi_i}\Big|\text{census}\right)=\sum_{i=1}^{n}y_i\bar{w}_i.$$

Hence, equation (7) can be expressed as

$$E\left[\left\{\sum_{i=1}^{n}\frac{y_i\bar{w}_i}{\pi_i}-\sum_{i=1}^{m}y_i\bar{w}_i\right\}\Big|\text{census}\right]^2$$

$$=\sum_{k=1}^{L}s_k\left[\left\{\sum_{i=1}^{n}\frac{y_i\bar{w}_i}{\pi_i}-\sum_{i=1}^{m}y_i\bar{w}_i\right\}\Big|\text{census}\right]^2$$

$$=\sum_{k=1}^{L}s_k\left\{\sum_{i=1}^{n}\frac{y_i\bar{w}_i}{\pi_i}\Big|\text{census}\right\}^2-\left(\sum_{i=1}^{m}y_i\bar{w}_i\right)^2$$

$$=\sum_{i=1}^{m}\frac{y_i^2w_i^2}{\pi_i}+\sum_{i\neq j}^{m}\frac{y_iy_j\bar{w}_i\bar{w}_j\pi_{ij}}{\pi_i\pi_j}-\left(\sum_{i=1}^{m}y_i\bar{w}_i\right)^2.$$

Ⅳ.2 Variance of $\hat{Y}_4$

$$V(\hat{Y}_4|\text{census})=E\left[\left\{\sum_{i=1}^{n}\sum_{j=1}^{J_i}\frac{y_{ij}w_{ij}}{\pi_i}-E\left(\sum_{i=1}^{n}\sum_{j=1}^{J_i}\frac{y_{ij}w_{ij}}{\pi_i}\Big|\text{census}\right)\right\}\Big|\text{census}\right]^2\cdots(8)$$

However, $E\left(\sum_{i=1}^{n}\sum_{j=1}^{J_i}\frac{y_{ij}w_{ij}}{\pi_i}\middle|\text{census}\right)=\sum_{i=1}^{m}\sum_{j=1}^{J_i}y_{ij}w_{ij}$.

Hence, equation (8) can be re-expressed as

$$E\left[\left\{\sum_{i=1}^{n}\sum_{j=1}^{J_i}\frac{y_{ij}w_{ij}}{\pi_i}-\sum_{i=1}^{m}\sum_{j=1}^{J_i}y_{ij}w_{ij}\right\}\middle|\text{census}\right]^2$$

$$=\sum_{k=1}^{L}s_k\left(\sum_{i=1}^{n}\sum_{j=1}^{J_i}\frac{y_{iij}w_{ij}}{\pi_i}-\sum_{i=1}^{m}\sum_{j=1}^{J_i}y_{ij}w_{ij}\right)\middle|\text{census}\Bigg]^2$$

$$=\sum_{k=1}^{L}s_k\left[\left(\sum_{i=1}^{n}\sum_{j=1}^{J_i}\frac{y_{ij}w_{ij}}{\pi_i}\middle|\text{census}\right)^2-\left(\sum_{i=1}^{m}\sum_{j=1}^{J_i}y_{ij}w_{ij}\right)^2\right]$$

$$=\sum_{k=1}^{L}s_k\left[\left(\sum_{i=1}^{m}\sum_{j=1}^{J_i}\frac{y_{ij}w_{ij}}{\pi_i}t_i\middle|\text{census}\right)^2-\left(\sum_{i=1}^{m}\sum_{j=1}^{J_i}y_{ij}w_j\right)^2\right]$$

$$=\sum_{i=1}^{m}\frac{\left(\sum_{j=1}^{J_i}y_{ij}w_{ij}\right)^2}{\pi_i}+\sum_{i\neq l}^{m}\left[\left(\sum_{j=1}^{J_i}\frac{y_{ij}w_{ij}}{\pi_i}\right)\left(\sum_{j=1}^{J_l}\frac{y_lw_{lj}}{\pi_l}\right)\pi_{il}\right]$$

$$-\left(\sum_{i=1}^{m}\sum_{j=1}^{J_i}y_{ij}w_{ij}\right)^2.$$

Thus far, the conditional variance formulae given the census sample have been derived. A few comments are required for the formulae.

i) The complete formula for the "within the census sample" component of the variance of $\hat{Y}$ involves taking the expected value of the conditional variances so far obtained which can be done easily. However, the conditional variance formulae obtained thus far are sufficient for our comparison since the quantities involved in the formulae are all available on the file for the census sample units;

ii) Since only one census sample was taken during the census, the "between the census sample" component of the variance of $\hat{Y}$ must be estimated using some special technique, i.e., random group, balanced half-sample replication, Jack-knife or general variance function. Estimation of this component by means of any of the above methods will incur vast amount of expenses, hence no attempt is being made to estimate this component at the Bureau.

iii) The variance formula involves the joint probability (π_{ij}) of selecting two distinct units in the CE sample which can be calculated without difficulty using an algorithm (3) developed by this author.

V. Concluding Remarks

Thus far, alternative forms of estimators and corresponding variance estimators have been shown which were developed for comparing competing within-PSU sampling schemes for the U.S. Consumer Expenditures surveys. There are several limitations to this research:

i) Due to unavailability of population total Y, the conditional bias given the census sample can not be obtained. Hence, even if the conditional mean saquared error is undoubtedly the best criterion for comparing those schemes, the conditional variance is the only criterion available for the comparison;

ii) The variances could not be mathematically compared, and hence numerical comparison approach has been adopted;

iii) Lastly, calcuation of the "between the census sample" component of the variance of $\hat{Y}$ would be expensive, hence no attempt is being made to compute the component. The "within the census sample" component is the only component which is being calculated at the Bureau. It should be added that the results of the empirical comparison have not been out, hence there is no telling which estimator is most efficient at this moment.

REFERENCES

1. Woltman, H., et. al. (1981). 1980 Census Estimation and Variance Estimation Studies, Design and Methodology, *Amer. Statis. Assoc. Surv. Res. Meth. Sec.*, 144—149.
2. Hartley, H.(1966). Systematic Sampling with Unequal Probability and without Replacement, *J. Amer. Statis. Assoc.* 61, 739—748.
3. Kim, J. (1983). Algorithm for Calculating Joint Probability (π_{ij}) When Systematic Sampling is Performed without Replacement, *technical paper, U.S. Bureau of the Census.*