ACHIEVING AGREEMENT BETWEEN THE AMERICAN COMMUNITY SURVEY AND THE CURRENT POPULATION SURVEY

Nanak Chand, Charles H. Alexander, U.S. Bureau of the Census
Nanak Chand, U.S. Bureau of the Census, Washington, D.C. 20233

Key Words: Micro data, Bernoulli trials

I. Introduction

The Current Population Survey (CPS) is designed to provide unbiased estimates of labor force characteristics at the national and state levels. While the American Community Survey (ACS) has a larger sample size, the resulting estimates may be biased compared to the CPS estimates due to differences in interviewer training, questionnaire design and similar other factors.

In this paper, we develop procedures to attain agreement between two surveys by imputing modified responses to the more biased survey. The objective is to arrive at micro data capable of providing unbiased estimates rather than to match the corresponding estimates from the two surveys. The process is based on the estimates of conditional probabilities of classification. The problem is complicated since the two surveys have unequal weights and the microdata from the surveys are not matchable.

Section II describes the procedure to modify one of the surveys to attain agreement when the population elements fall into one of the two classes on the basis of a single labor force characteristic. Sections III adapts this procedure to multiple and nested classifications. The resulting data provide estimates with equal expected values for the two surveys for various domains of the underlying population.

Section IV contains examples of differences in proportions in the CPS and the census long form sample. These differences illustrate the bases of imputing the microdata in achieving the desired agreement. Section V illustrates the application of the underlying theory.

II. Achieving Agreement Between Two Surveys - Two Classes

Let D be a domain of known size N within the population. $S^{(1)}$ and $S^{(2)}$ are two surveys which take samples of sizes $n^{(1)}$ and $n^{(2)}$ respectively from D. While $S^{(1)}$ provides an unbiased estimate of number of units in D with characteristic C, the corresponding estimate resulting from $S^{(2)}$ may contain some bias.

Let $p^{(1)}$ and $p^{(2)}$ be the estimated proportions of units with characteristic C in domain D according to the two surveys, and let $d = p^{(1)} - p^{(2)} > 0$. The following analysis also holds for $d < 0$, when we consider the complement of class C.

Let $C^{\sim}$ denote the set of k sampled units in $S^{(2)}$ which are classified as not having characteristic C. We sequentially number these units as 1, ..., k, with their final survey weights $w_1, ..., w_k$. Let

$$w = \sum_{i=1}^{k} w_i, \text{ and } R = w/N.$$

For $S^{(1)}$ and $S^{(2)}$ to provide estimates of the population proportion with equal expected values, some of these k units, would be imputed as having characteristic C. The weighted proportion of the units to be imputed is

$$p = d/R$$

The following algorithm generates a random sample (Cochran (1977)) from $C^{\sim}$. One of the results of Theorem 1. is that the expected weighted proportion of the resulting sampled units is p.

Algorithm A:

Selecting a Random Sample from $C^{\sim}$:

1. Let u be a number generated from the uniform distribution on the interval (0,1), and let $r = .5 + ku$.

2. We select the sample unit i with the weight w_i if $i - .5 <= r < i + .5$, $i = 1, ..., k$.

3. We repeat steps 1. and 2. and stop sampling when the sum of the weights of the selected units first becomes greater than or equal to pw. Let units $i_1, ..., i_s$, be selected. Then

$$g < pw <= h, \text{ where}$$

$$g = \sum_{j=1}^{s-1} w_{i_j}, \quad h = \sum_{j=1}^{s} w_{i_j}.$$

4. If pw = h, then the selected sample is $\{i_1, ..., i_s\}$.

5. If pw < h, then let π = (h - pw)/(h - g). We generate a Bernoulli trial with probability of success π. If the outcome is 0, we select the sample units $i_1, ..., i_s$. If the outcome is 1, we select the sample units $i_1, ..., i_{s-1}$.

Theorem 1:

Let $\hat{p}$ and $\hat{n}$ be respectively the weighted proportion of sample units and size of the sample from the set $C^{\sim}$ as given by algorithm A. With g given by the algorithm, define

$$\pi^{(2)} = p^{(2)} + \hat{p}.R,$$

$$\Omega = \{w_i\text{: } 1 \leq i \leq k,\ w_i \geq (pw-g)\},$$

$$m = Min_{\Omega}\ (w_i), \text{ and}$$

$$M = Max_{\Omega}\ (w_i).$$

Then conditional on the sample outcome of the vector $(s,\ g,\ ,h,\ \pi)$, as defined in the algorithm, the following results hold:

1. $E\ [\pi^{(2)}] = p^{(1)}$,

2. The actual difference a between $p^{(1)}$ and $\pi^{(2)}$ is at most $\bar{a} = M/N$,

3. The lower and upper bounds on the probability that unit i_s is retained in the final sample are respectively

   $L = (pw-g)/M$, and $U = (pw-g)/m$,

4. $E[\hat{n}] = s - \pi$, and

5. $v = var[\hat{p}] = (pw-g)(h-pw)/w^2$.

Proof:

1. If pw=h, then the sampled units are $\{i_1, ..., i_s\}$, and $\hat{p} = h/w = p$. If pw < h, then $\hat{p}$ takes values g/w and h/w respectively with probabilities π and $1 - \pi$, resulting in

$$E[\hat{p}] = \pi(g-h)/w + h/w = p, \text{ giving}$$

$$E[\pi^{(2)}] = p^{(2)} + p.R = p^{(1)}.$$

2.

$$a = p^{(1)} - \pi^{(2)} = p^{(1)} - p^{(2)} - \hat{p}.R = R(p - \hat{p}).$$

Since $p \leq h/w$ and $\hat{p} > g/w$, we have,

$$a \leq \frac{h-g}{N} = \frac{w_{i_s}}{N} \leq M/N = \bar{a}.$$

3. The probability that the s_{th} unit is retained in the final sample is

$$1 - \pi = \frac{pw - g}{w_{i_s}},\ \textit{giving}$$

$$\frac{pw - g}{M} \leq 1 - \pi \leq \frac{pw - g}{m}$$

4.

$\hat{n}$ attains values s - 1 and s respectively with probabilities π *and* $1 - \pi$, giving

$$E[\hat{n}] = s - \pi$$

5.

$$E[\hat{p}^2] = \frac{\pi g^2 + (1 - \pi)h^2}{w^2}, \text{ giving}$$

$$v = E[\hat{p}^2] - p^2 = \frac{\pi(g^2 - h^2) + (h^2 - p^2w^2)}{w^2},$$

This simplifies to

$$v = \frac{(pw - g)(h - pw)}{w^2}$$

Algorithm B:

Multiple Samples from $C^{\sim}$

To make the imputed data appropriate for application of complete-data methods, we repeat algorithm A, m times resulting in m independent samples, the j_{th} sample being of size $\hat{n}_j$ with the corresponding vector

(s_j, g_j, h_j, π_j), and estimated weighted

proportion $\hat{p}_j$, with variance v_j given by Theorem 1 as

$$v_j = (pw-g_j)\ (h_j-pw)/w^2,\ j=1,\ ...,\ m.$$

The m complete - data sets consisting of the modified sample outcome result in m complete - data estimates of the population proportion. These are given by

$$\pi_j^{(2)} = p^{(2)} + \hat{p}_j.R,\ j=1,\ ...,\ m$$

with the average estimate given by

$$\bar{\pi}^{(2)} = p^{(2)} + \hat{\bar{p}}\ .R,\ \textit{with}$$

$$\hat{\bar{p}} = m^{-1}\sum_{j=1}^{m} \hat{p}_j$$

Theorem 2:

Let $\bar{\pi}^{(2)}$ be the average of the complete - data estimates given by Algorithm B. Then conditional on the sample outcome, we have

1. $E\ [\ \bar{\pi}^{(2)}\ ] = p^{(1)}$, and

2. Increase in variance due to misclassification in $S^{(2)}$ is given by

$$i_m = \frac{m+1}{m(m-1)} R^2 \sum_{j=1}^{m} (\hat{p}_j - \hat{\bar{p}})^2$$

Proof:

1. The result follows from definition of $\bar{\pi}^{(2)}$ since by Theorem 1,

$$E\ [\hat{p}_j] = p,\ j = 1,\ ...,\ m$$

2. Since increase in variance is equal to the finite imputation factor ($1+m^{-1}$) times the between imputation variance B_m, we have

$$i_m = \frac{m+1}{m} B_m,\ \textit{where}$$

$$B_m = \frac{1}{m-1} \sum_{j=1}^{m} (\pi_j^{(2)} - \bar{\pi}^{(2)})^2,$$

and the result follows since

$$\pi_j^{(2)} - \bar{\pi}^{(2)} = R(\hat{p}_j - \hat{\bar{p}}),\ j = 1,\ ...,\ m$$

III. Achieving Agreement - Three or More Classes

If the surveys classify the population units into three or more classes and the classes are mutually exclusive, we achieve agreement between the surveys with respect to these classes, by applying the above procedure to each of the classes treating the remaining units as belonging to the complementary class.

However, if a class is a subset of another class, the larger classes precede their subsets in applying the procedure.

The imputation procedure will also change the proportions in different categories for the classifications not yet imputed.

This can be demonstrated by observing classification tables for two classes C_{i-1} and C_i, C_i is a subset of C_{i-1} and is classified within C_{i-1}. An example of this situation is where C_{i-1} represents the labor force, and C_i represents the persons classified as employed within the labor force. O_{i-1} and O_i represent the complements of C_{i-1} and C_i respectively.

Proportions for Class C_{i-1}

	$S^{(1)}$	$S^{(2)}$
C_{i-1}	$p_{i-1}^{(1)}$	$p_{i-1}^{(2)}$
O_{i-1}	$1 - p_{i-1}^{(1)}$	$1 - p_{i-1}^{(2)}$

Proportions for Class C_i

	$S^{(1)}$	$S^{(2)}$
C_i	$p_i^{(1)}$	$p_i^{(2)}$
O_i	$1 - p_i^{(1)}$	$1 - p_i^{(2)}$

After $S^{(2)}$ has been modified for C_{i-1}, the second table becomes:

Proportions for Class C_i After Imputation of C_{i-1}

	$S^{(1)}$	$S^{(2)}$
C_i	$p_i^{(1)}$	$p_i^{(2)}.d_{i-1}$
O_i	$1 - p_i^{(1)}$	$1 - p_i^{(2)}.d_{i-1}$, where

$$d_{i-1} = p_{i-1}^{(2)}/\pi_{i-1}^{(2)}.$$

IV. Examples of Differences in Proportions

Example 1:

The CPS ($S^{(1)}$) (U.S. Department of Labor (1993)) and the census long form sample ($S^{(2)}$) (U.S. Department of Commerce (1993)) classify the 1990 non-institutionalized civilian U.S. population, into labor force (C_{i-1}) and other than labor force (O_{i-1}) categories, resulting in the following proportions:

	$S^{(1)}$	$S^{(2)}$
C_{i-1}	.664	.661
O_{i-1}	.336	.339

The imputation procedure is applied to sample units classified as not being in labor force in the census long form.

Example 2:

In the surveys of example 1, let categories C_i and O_i respectively represent the employed and unemployed segments of the labor force. The two surveys provide the following proportions:

	$S^{(1)}$	$S^{(2)}$
C_i	.945	.937
O_i	.055	.063

When a proportion of the long form sample units in category O_{i-1} has been imputed as being in labor force by the above procedure, then assuming $\pi_{i-1}^{(2)} = .664$, the classification table changes to

	$S^{(1)}$	$S^{(2)}$
C_i	.945	.933
O_i	.055	.067

The imputation procedure is applied to sample units classified as unemployed according to the long form sample.

V. Illustration of the Imputation Procedure

We illustrate the above procedure by generating an ACS sample assuming that the census long form and the CPS results of the above examples are the same as given respectively by an ACS ($S^{(2)}$) and a CPS ($S^{(1)}$) sample taken from a given domain, and we want to impute the ACS results to match with the CPS results.

Example 3:

$S^{(2)}$ classifies the basic domain population into labor force and other than labor force categories. Taking $n^{(2)}$ equal to 100,000 and two possible survey weights of magnitude 2 and 5 each with probability .5, we performed the simulation in three steps by:

1. generating the domain population with size approximately given by the sum of the survey weights as

$$N \fallingdotseq 350{,}408,$$

2. simulating $S^{(2)}$ results with expected probability .661 of a population unit being in the labor force, resulting in

$$p^{(2)} = .660967,\ k = 33{,}903,$$
$$w = 118{,}800,\ pw = 1{,}051.38, \text{ and}$$

3. applying algorithms A and B, with m=3, to $C^{\sim} = \{1, ..., 33{,}903\}$ with the following results:

Selected Labor Force Statistics - Example 3 $(\text{Sample})_j$

j	1	2	3
g_j	1,050	1,050	1,048
h_j	1,052	1,055	1,053
$(pw-g)_j$	1.38	1.38	3.38
π_j	.310	.724	.324
m_j	2	2	5
M_j	5	5	5
$\bar{a}_j$	.000014	.000014	.000014
L_j	.276	.276	.676
U_j	.690	.690	.676
s_j	304	304	300
$\hat{p}_j$	.008855	.008838	.008864
$\hat{d}_j$	.003002	.002996	.003005
$\hat{n}_j$	304	303	300
$\pi_j^{(2)}$	.664002	.663996	.664005
v_j	6.062 $x10^{-11}$	3.54 $x10^{-10}$	3.88 $x10^{-10}$

Combining the results from the three imputations, we get

$$\hat{p} = .008852,\ R = .339033,\ \bar{\pi}^{(2)} = .664001$$

$$B_3 = 2^{-1}\,(.339033)^2 x10^{-10}\{.3^2+1.4^2+1.2^2\}$$

$$= .200576x10^{-10}$$

$$i_3 = \frac{4}{3}\,B_3 = .267435x10^{-10}$$

Example 4:

The different labor force sizes as imputed by each of the samples in example 3 are approximately given by

$$N_1 \fallingdotseq 232{,}660,\ N_2 \fallingdotseq 232{,}658,\ \textit{and}\ N_3 \fallingdotseq 232{,}661$$

We simulated the employment component of the labor force sample outcome of $S^{(2)}$ by generating number of employed persons in each of these three populations, with expected probability of a person being employed given the person is in labor force, being equal to .933.

The parameters of the sets $C^{\sim}$ and the results of applying algorithms A and B are as follows:

Selected Employment Statistics - Example 4 (Sample)$_j$

j	1	2	3
k_j	4,421	4,421	4,419
w_j	15,529	15,544	15,519
$(pw)_j$	2,791.9589	2,791.7024	2,791.868
g_j	2,789	2,791	2,787
h_j	2,794	2,796	2,792
$(pw\text{-}g)_j$	2.9589	.7024	4.868
π_j	.4082	.8595	.0264
m_j	5	2	5
M_j	5	5	5
$\bar{a}_j$	.000022	.000022	.000022
L_j	.591782	.14048	.97362
U_j	.591782	.3512	.97362
s_j	779	810	793
$\hat{p}_j$	.179599	.179555	.179908
$\hat{d}_j$	.012033	.012030	.012054
$\hat{n}_j$	778	809	793
$\pi_j^{(2)}$	.945033	.945030	.945054
v_j	2.5044 $x10^{-8}$	1.2494 $x10^{-8}$	2.6661 $x10^{-8}$

The combined results from the three imputations are:

$$\hat{p} = .179687 \quad \bar{\pi}^{(2)} = .945039$$

$$R_1 = .066745, \quad R_2 = .066811, \quad R_3 = .066702$$

$$B_3 = 2^{-1}x10^{-10}\{(.587356)^2 + (.881905)^2 + (1.474114)^2\}$$

$$= 1.647878x10^{-10} \text{ and}$$

$$i_3 = \frac{4}{3}B_3 = 2.197170x10^{-10}.$$

REFERENCES

[1] Cochran, W.G. (1977) Sampling Techniques, John Wiley, New York.

[2] U.S. Department of Commerce (1993) 1990 Census of Population, U. S. Department of Commerce, Bureau of the Census.

[3] U.S. Department of Labor (1993) Employment and Earnings, U. S. Department of Labor, Bureau of Labor Statistics.

(This paper reports the general results of research undertaken by the Census Bureau staff. The views expressed are attributable to the authors and do not necessarily reflect those of the Census Bureau.)