Optimum Nonresponse Subsampling Rate for the American Community Survey[1]

Anthony Tersine and Michael Starsinic, U.S. Census Bureau
Anthony Tersine, U.S. Census Bureau, Demographic Statistical Methods Division, Washington, D.C. 20233

1.0 Introduction

The American Community Survey (ACS) is designed as a monthly mail-out survey with follow-up by Computer Assisted Telephone Interviewing (CATI) and Computer Assisted Personal Interviewing (CAPI) operations during a three month interview cycle. The ACS is an annual survey of three million addresses with approximately one-twelfth of the sample mailed out each month. All households with a mailable address are sent a mail questionnaire during the first month of the interviewing cycle. The mailable cases are sent an initial letter and reminder card to return the survey. Also a second questionnaire is delivered if the housing unit does not return the first questionnaire within a few weeks time. If a mail form is incomplete or has more household members than allowed on the form then a telephone failed edit follow-up (TFEFU) operation is conducted to obtain the missing information. During the second month, all households which did not return a mail form and for which we can obtain a telephone number are sent to CATI. During the third month, all households which did not return a mail form or for which we did not obtain a CATI interview are sent to CAPI. Those eligible for CAPI are sub-sampled at two different rates: 2-in-3 for units without a mailable address and 1-in-3 for all other units.

Is the current assumption of CAPI subsampling at 1-in-3, the correct rate? This paper will look at the assumptions behind that rate and see if they are still valid at the current time. The 1-in-3 rate was calculated in Alexander (1993) as 0.38 and changed to 1-in-3 for operational simplicity.

2.0 Costs and Definitions

We will first define variables for each mode and the values for each.

2.1 Mail Costs

n	3,000,000	total annual sample
P_d	0.96	proportion of sample mailable
P_o	0.90	proportion of sample in occupied housing units
C_{m0}	3.92	cost for each mailout case
C_{mr}	14.85	additional cost for each mail return case
C_{mb}	8.88	cost for mailback and processing returns
C_{m2}	2.33	cost for each second mailing
C_{mf}	15.10	cost for each TFEFU
R_{mf}	1/3	proportion of mail returns needing TFEFU

[1] This paper reports the results of research and analysis undertaken by the U.S. Census Bureau staff. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications. This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress.

R_m	0.50	proportion of deliverables returned
R_{m2}	0.40	proportion of mail returns needing second mailing
R_{mo}	0.555556	proportion of occupied deliverables returned (R_m / P_o)

The value of C_{mr} is calculated as follows:

$$C_{mr} = C_{mb} + R_{mf}\ C_{mf} + R_{m2}\ C_{m2} = 8.88 + (1/3) * 15.10 + 0.4 * 2.33 = 14.85$$

2.2 Telephone Costs

C_{ti}	50.94	cost for each telephone interview
C_{tni}	12.73	cost for each telephone noninterview
e_t	0.32	proportion of non-mail returns eligible for CATI (good phone numbers)
f_t	1.00	proportion of non-mail returns selected for CATI (current value)
R_t	0.60	proportion of CATI eligible cases interviewed
R_{to}	0.75	proportion of occupied CATI eligible cases interviewed $[(1- R_m)\ R_t] / [P_o / (1 - R_{mo})]$

2.3 Personal Visit Costs

C_{pi}	145.58	cost for each personal visit interview
C_{pni}	72.79	cost for each personal visit noninterview
f_{pd}	1/3	fraction of mailable noninterviews selected for CAPI (current value)
f_{pu}	2/3	fraction of non-mailables selected for CAPI (current value)
R_p	0.86	proportion of CAPI cases interviewed
R_{po}	0.81998	proportion of occupied CAPI cases interviewed (assume all vacants interviewed) N_{pio} / n_p

The variable, N_{pio}, is defined as the number of occupied interviews in CAPI and is equal to the total number of CAPI interviews minus the number of vacant CAPI interviews. The total number of CAPI interviews is:

$$\text{CAPI Int} = [n\ P_d\ (1 - R_m)\ f_t\ (1 - e_t\ R_t)\ f_{pd} + n\ (1 - P_d)\ f_{pu}]\ R_p = 402{,}342$$

The number of vacant CAPI interviews is (We assume all vacants are interviewed.):

$$\text{CAPI Vacant Int} = n\ P_d\ (1 - P_o)\ f_t\ f_{pd} + n\ (1 - P_d)\ (1 - P_o)\ f_{pu} = 104{,}000$$

So N_{pio} = CAPI Int - CAPI Vacant Int = 402,342 - 104,000 = 298,342

The variable, n_p, is defined as the number of occupied housing units that were selected in the CAPI subsample:

$$n_p = n\ P_d\ P_o\ (1 - R_{mo})\ (1 - e_t\ R_{to})\ f_t\ f_{pd} + n\ (1 - P_d)\ P_o\ f_{pu} = 363{,}840$$

3.0 Sample Sizes and Proportions by Mode

Let s_m, s_t, and s_p, be the proportions of the occupied housing units represented by the mail, CATI, and CAPI components.

- s_m: proportion of occupied units represented by mail respondents
  $s_m = (n\ P_o\ P_d\ R_{mo}) / (n\ P_o) = P_d\ R_{mo} = 0.533333$
- s_t: proportion of occupied units represented by CATI interviews
  $s_t = (n\ P_o\ P_d\ (1 - R_{mo})\ e_t\ R_{to}) / (n\ P_o) = P_d\ (1 - R_{mo})\ e_t\ R_{to} = 0.102400$
- s_p: proportion of occupied units represented by CAPI universe
  $s_p = 1 - s_m - s_t = 1 - P_d\ R_{mo} - P_d\ (1 - R_{mo})\ e_t\ R_{to} = 0.364267$
  s_p can be split into two components representing mailable and unmailable address.
  - s_{pu}: proportion of occupied units represented by unmailable CAPI cases
    $s_{pu} = 1 - P_d = 0.040000$
  - s_{pd}: proportion of occupied units represented by mailable CAPI cases
    $s_{pd} = s_p - (1 - P_d) = P_d\ [(1 - R_{mo})\ (1 - e_t\ R_{to})] = 0.324267$

We now look at the sample sizes (n_m, n_t, and n_p) for the occupied units in the mail, CATI, and CAPI components.

- n_m: number of sample cases representing occupied unit mail respondents
  $n_m = n\ P_o\ s_m = n\ P_o\ P_d\ R_{mo} = 1{,}440{,}000$
- n_t: number of sample cases representing occupied unit CATI interviews
  $n_t = n\ P_o\ f_t\ s_t = n\ P_o\ f_t\ P_d\ (1 - R_{mo})\ e_t\ R_{to} = 276{,}840$
- n_p: number of sample cases representing occupied unit CAPI universe
  $n_p = n_{pu} + n_{pd} = n\ P_o\ [s_p\ f_t\ f_{pd} + (1 - P_d)\ (f_{pu} - f_t\ f_{pd})] = 363{,}840$
  n_p can be split into two components representing mailable and unmailable address.
  - n_{pu}: number of sample cases representing CAPI universe of unmailable occupied units
    $n_{pu} = n\ P_o\ (1 - P_d)\ f_{pu} = 72{,}000$
  - n_{pd}: number of sample cases representing CAPI universe of mailable occupied units
    $n_{pd} = n\ P_o\ [s_p - (1 - P_d)]\ f_t\ f_{pd} = 291{,}840$

4.0 Cost per Case by Mode and Total Non-Fixed Cost

We determine the overall cost for each interview by mode. The costs of the noninterviews for each mode are apportioned to interview cases as in Alexander, 1993.

Mail: $C_m = C_{m0} / R_m + C_{mr} + [(1 - R_m) / R_m]\ C_{m2} = 3.92 / 0.5 + 14.85 + [(1 - 0.5) / 0.5] * 2.33 = 25.02$

CATI: $C_t = C_{ti} + [(1 - R_t) / R_t]\ C_{tni} = 50.94 + [(1 - 0.6) / 0.6] * 12.73 = 59.43$

CAPI: $C_p = C_{pi} + [(1 - R_p) / R_p]\ C_{pni} = 145.58 + [(1 - 0.86) / 0.86] * 72.79 = 157.43$

Based on the costs above, we calculate the non-fixed cost to be:

$$n_m\ C_m + n_t\ C_t + [n_{pd}\ R_{po} + n_{pu}\ R_{po} + n\ P_d\ (1 - P_o)\ f_t\ f_{pd} + n\ (1 - P_d)\ (1 - P_o)\ f_{pu}]\ C_p$$

$$= n\ P_o\ s_m\ C_m + n\ P_o\ f_t\ s_t\ C_t + n\ \{P_o\ R_{po}\ [(s_p - (1 - P_d))\ f_t\ f_{pd} + (1 - P_d)\ f_{pu}] + (1 - P_o)[P_d\ f_t\ f_{pd} + (1 - P_d) f_{pu}]\}C_p \quad (4.1)$$

5.0 Average Noninterview Adjustment Factor

First we look at the sum of the weights for occupied interviewed cases.

$(1 / f)\ [n_m + n_t / f_t + (n_{pd} R_{po}) / (f_t f_{pd}) + (n_{pu} R_{po}) / f_{pu}]$ where f is the overall sampling rate

This can be simplified to: $(1 / f)\ n P_o [1 - s_p (1 - R_{po})]$ (5.1)

So the nonresponse adjustment factor that weights this back up to the population total of occupied housing units, $(1 / f)\ n P_o$, is: $1 / [1 - s_p (1 - R_{po})]$

6.0 Variation of an Estimated Proportion

Let the estimated proportions for the three data collection modes be $\hat{P}_m$, $\hat{P}_t$, and $\hat{P}_p$. Then the overall estimator of a weighted proportion is:

$$\hat{P} = \{(1 / f)\ [n_m \hat{P}_m + (n_t / f_t) \hat{P}_t + (n_{pd} R_{po} \hat{P}_p) / (f_t f_{pd}) + (n_{pu} R_{po} \hat{P}_p) / f_{pu}]\} / (5.1)$$

Under the assumption that $\hat{P}_m = \hat{P}_t = \hat{P}_p = P = 1 - Q$, the variance of $\hat{P}$ is:

$$\mathrm{Var}(\hat{P}) = [n_m PQ + (n_t PQ) / f_t^2 + (n_{pd} R_{po} PQ) / (f_t f_{pd})^2 + (n_{pu} R_{po} PQ) / f_{pu}^2] \; / \; \{n P_o [1 - s_p (1 - R_{po})]\}^2$$

Which can be written as:

$$[(PQ) / (n P_o)]\ [1 - s_p (1 - R_{po})]^{-2}\ [s_m + s_t / f_t + \{(s_p - (1 - P_d)) R_{po}\} / (f_t f_{pd}) + \{(1 - P_d) R_{po}\} / f_{pu}] \quad (6.1)$$

7.0 Optimization of Subsampling Rates

7.1 Variance Function

We want to optimize the subsampling rates f_t, f_{pd}, and f_{pu}. Using (4.1) for costs and (6.1) for the variance, we can calculate the optimal subsampling rates.

Choose a reliability, V, for a given P and set V = (6.1). We want to solve this as a function of f_t, f_{pd}, and f_{pu}. V, P, Q, P_o, and $[1 - s_p (1 - R_{po})]^{-2}$ are not functions of the sampling parameters, so we write

$$K^* = [s_m + s_t / f_t + \{(s_p - (1 - P_d)) R_{po}\} / (f_t f_{pd}) + \{(1 - P_d) R_{po}\} / f_{pu}] \mathbf{/} [f N]$$

which does not depend on the sampling parameters.

Letting n = f N and K=1 / K*, we calculate

$$n = K\,[s_m + s_t / f_t + \{(s_p - (1 - P_d))\,R_{po}\} / (f_t\,f_{pd}) + \{(1 - P_d)\,R_{po}\} / f_{pu}] \qquad (7.1)$$

7.2 Cost Function

Substituting (7.1) into (4.1) gives us the objective function to be minimized

$$\{K\,[s_m + s_t / f_t + \{(s_p - (1 - P_d))\,R_{po}\} / (f_t\,f_{pd}) + \{(1 - P_d)\,R_{po}\} / f_{pu}]\}\;\{P_o\,s_m\,C_m + P_o\,f_t\,s_t\,C_t + \{P_o\,R_{po}\,[(s_p - (1 - P_d))\,f_t\,f_{pd} + (1 - P_d)\,f_{pu}] + (1 - P_o)\,[P_d\,f_t\,f_{pd} + (1 - P_d)\,f_{pu}]\}\,C_p\}$$

In the last factor we combine the terms with $f_t\,f_{pd}$ and with f_{pu}.

$$\{K\,[s_m + s_t / f_t + \{(s_p - (1 - P_d))\,R_{po}\} / (f_t\,f_{pd}) + \{(1 - P_d)\,R_{po}\} / f_{pu}]\}\;\{P_o\,s_m\,C_m + P_o\,f_t\,s_t\,C_t + \{[P_o\,R_{po}\,(s_p - (1 - P_d)) + (1 - P_o)\,P_d]\,f_t\,f_{pd} + (1 - P_d)\,[P_o\,R_{po} + (1 - P_o)]f_{pu}\}\,C_p\} \qquad (7.2)$$

7.3 Minimization

Define

$a_m = s_m / s_m^{½}$ $\qquad$ $b_m = (P_o\,s_m\,C_m)^{½}$

$a_t = s_t / (s_t\,f_t)^{½}$ $\qquad$ $b_t = (P_o\,f_t\,s_t\,C_t)^{½}$

$a_{pd} = [(s_p - (1 - P_d))\,R_{po}] / [(s_p - (1 - P_d))\,R_{po}\,f_t\,f_{pd}]^{½}$

$b_{pd} = \{[P_o\,R_{po}\,(s_p - (1 - P_d)) + (1 - P_o)\,P_d]\,f_t\,f_{pd}\,C_p\}^{½}$

$a_{pu} = [(1 - P_d)\,R_{po}] / [(1 - P_d)\,R_{po}\,f_{pu}]^{½}$

$b_{pu} = \{(1 - P_d)\,[P_o\,R_{po} + (1 - P_o)]f_{pu}\,C_p\}^{½}$

So minimizing (7.2) is the same as minimizing

$$(a_m^2 + a_t^2 + a_{pd}^2 + a_{pu}^2)\,(b_m^2 + b_t^2 + b_{pd}^2 + b_{pu}^2)$$

By the Cauchy-Schwartz inequality this is minimized if and only if

$$a_m / b_m = a_t / b_t = a_{pd} / b_{pd} = a_{pu} / b_{pu}$$

Calculating and simplifying the four individual ratios we get

$$a_m / b_m = 1 / (P_o\,C_m)^{½} \qquad (7.3)$$

$$a_t / b_t = 1 / [(P_o\,C_t)^{½}\,f_t] \qquad (7.4)$$

$$a_{pd} / b_{pd} = [(s_p - (1 - P_d))R_{po}]^{½} / \{[P_o\,R_{po}\,(s_p - (1 - P_d)) + (1 - P_o)P_d]\,C_p\}^{½}\,f_t\,f_{pd} \qquad (7.5)$$

$$a_{pu} / b_{pu} = R_{po}^{1/2} \;/\; \{[P_o R_{po} + (1 - P_o)] C_p\}^{1/2} f_{pu} \qquad (7.6)$$

Equating (7.3) and (7.4) we get

$$f_t = (C_m / C_t)^{1/2} \qquad (7.7)$$

Equating (7.4) and (7.5) we get

$$f_{pd} = [C_t P_o R_{po} (s_p - (1 - P_d))]^{1/2} / \{C_p [P_o R_{po} (s_p - (1 - P_d)) + (1 - P_o)P_d]\}^{1/2} \qquad (7.8)$$

Equating (7.3) and (7.6) we get

$$f_{pu} = (C_m P_o R_{po})^{1/2} / \{C_p [P_o R_{po} + (1 - P_o)]\}^{1/2} \qquad (7.9)$$

7.4 Minimization - No CATI Subsampling

Suppose we have no CATI subsampling (i.e. $f_t = 1$), how does this affect the optimization?

Define

$a_{mt} = (s_m + s_t) / (s_m + s_t)^{1/2}$ $\qquad$ $b_{mt} = [P_o (C_m s_m + C_t s_t)]^{1/2}$

a_{pd} and b_{pd} are as in section 7.3, but with $f_t = 1$. a_{pu} and b_{pu} are exactly the same as section 7.3.

We do the same minimization as above. Calculating and simplifying the three individual ratios we get

$$a_{mt} / b_{mt} = \{(s_m + s_t) / [P_o (C_m s_m + C_t s_t)] \}^{1/2} \qquad (7.10)$$

$$a_{pd} / b_{pd} = [(s_p - (1 - P_d))R_{po}]^{1/2} / \{[P_o R_{po} (s_p - (1 - P_d)) + (1 - P_o)P_d] C_p\}^{1/2} f_{pd} \qquad (7.11)$$

$$a_{pu} / b_{pu} = R_{po}^{1/2} \;/\; \{[P_o R_{po} + (1 - P_o)] C_p\}^{1/2} f_{pu} \quad \text{(Note: same as (7.6))} \qquad (7.12)$$

Equating (7.10) and (7.11) we get

$$f_{pd} = \{[(C_m s_m + C_t s_t) / (s_m + s_t)] P_o R_{po} (s_p - (1 - P_d))\}^{1/2} / \{C_p [P_o R_{po} (s_p - (1 - P_d)) + (1 - P_o)P_d]\}^{1/2} \qquad (7.13)$$

Equating (7.10) and (7.12) we get

$$f_{pu} = \{[(C_m s_m + C_t s_t) / (s_m + s_t)] P_o R_{po}\}^{1/2} / \{C_p [P_o R_{po} + (1 - P_o)]\}^{1/2} \qquad (7.14)$$

8.0 Results

What are the optimal subsampling rates? Using (7.7), (7.8), and (7.9), we calculate the optimal subsampling rates as:

- f_t = 0.648863
- f_{pd} = 0.519043
- f_{pu} = 0.374116

If we assume that f_t = 1, and use (7.13) and (7.14) to calculate the optimal subsampling rates we get:

- f_{pd} = 0.372223
- f_{pu} = 0.413479

What is the affect on the variance and total cost for these optimal rates as compared to the current rates. We assume an annual sampling rate of 2.5 percent and an estimate of 10 percent. The standard error will be calculated for an average tract which has 4000 people, which means an annual initial sample of 100 people and a sample over five years of 500 people. For the calculation of the standard error, we use a design factor of 1.6.

Table 1. Estimated Variances and Total Cost for Different Subsampling Rates

Variable	Current Rates	Optimum with f_t not equal to 1		Optimum with f_t = 1		
		Actual Rates	Rounded Rates	Actual Rates	Rounded Rates 1	Rounded Rates 2
f_t	1.000000	0.648863	0.666667	1.000000	1.000000	1.000000
f_{pd}	0.333333	0.519043	0.500000	0.372223	0.400000	0.333333
f_{pu}	0.666667	0.374116	0.400000	0.413479	0.400000	0.400000
Standard Error	0.027972	0.028769	.028754	0.027466	0.027011	0.028280
Relative Change in Standard Error from Current		2.85%	2.79%	-1.81%	-3.44%	1.10%
Coefficient of Variation	27.97%	28.77%	28.75%	27.47%	27.01%	28.28%
90% Confidence Interval	5.40%, 14.60%	5.27%, 14.73%	5.27%, 14.73%	5.48%, 14.52%	5.56%, 14.44%	5.35%, 14.65%
Total Cost	115,800,000	105,950,000	106,100,000	117,950,000	122,140,000	111,580,000

9.0 Conclusions

The results suggest that the efficiency of the ACS could be improved by starting the subsampling in the CATI phase. With this change the standard error would be about 3 percent larger, but the cost would be reduced by about $10 million. The other option which would save money is the last column in Table 1. Under this scenario the only subsampling rate that changes is for the unmailables from 2-in-3 to 2-in-5, and this would save about $4 million.

10.0 References

Alexander, C. H. (1993), "Determination of Sample Size for the Intercensal Long Form Survey Prototype," Internal Census Bureau Report, August 10, 1993.

Cochran, W. G. (1977), *Sampling Techniques*, New York: Wiley.

Elliot, M. R., Little, R. J. A., and Lewitzky, S. (2000), “Subsampling Callbacks to Improve Survey Efficiency,” *Journal of the American Statistical Association*, 95, 730-738.