American Community Survey: Sample Design Issues and Challenges

Steven P. Hefter, Andre L. Williams
U.S. Census Bureau
Washington, D.C. 20233

Abstract

In 2005, the American Community Survey (ACS) selected and fielded a sample of housing unit addresses in every county and county equivalent in the U.S. and annual samples have been selected each year since. The ACS collects housing and person level data of this sample continuously throughout the year by assigning each address in the annual sample to a particular month of the year. These data have historically been collected by the decennial census long form from a sample of addresses in the census.

The goal of the ACS is to publish data for small geographic areas by cumulating sample over five-year periods. Unlike the census long form, which had an overall sampling rate of approximately 1-in-6 and 100 percent follow-up of non-respondents, the ACS selects a fixed annual sample of approximately 15,000,000 addresses over five years and samples non-respondents with an overall rate of approximately 1-in-3. These two factors, combined with relatively constant growth in the housing unit inventory and a persistent decline in cooperation rates have led to concerns about whether the ACS is meeting its reliability goal for small areas – relative to the Census 2000 Long Form.

This paper looks at the changes in the ACS sample size distribution for counties, places, and census tracts for the years 2005-2007. We also provide a comparison of the distribution of the ACS sampling frame by sampling strata to the same distribution from the Census 2000 Long Form sampling frame and discuss the implications of the fixed sample size on the reliability of the ACS estimates.

Keywords: ACS, Sampling, Census Long Form

1. Differences Between the Census Long Form and the ACS Sample Designs

There are two key differences between the census long form sample design and that of the ACS. These differences, and their impact on ACS estimates are as follows:

Sample Size: The census long form sample was designed with an overall target sampling rate of 1-in-6 (Hefter, 1999) while the ACS design is based on a fixed annual target sample size of three million.
Impact: Over time, due to expected growth in the ACS sampling frame, the percentage of addresses in the ACS sample decreases, presumably at all levels of geography, impacting the reliability of the small area estimates.

Non-response Follow-up: All long form non-responding units were contacted as part of the decennial non-response follow-up operations. Only a sample of non-responding units in the ACS are sent to personal interviewing (Hefter, 2005).
Impact: The ACS, while maintaining weighted response rates of roughly 98%[1], only realizes interviews from approximately 70% of the initial sample. This has a direct negative impact on the variances of the estimates, relative to a full 100% non-response follow-up of cases.

Both of these differences should be carefully considered when discussing the usefulness of the ACS estimates, in terms of reliability, as compared to the census long form sample. This paper focuses on the impact of the sample size difference and the distribution of the initial samples. In analyzing the percent in sample we made no attempt to factor in the ACS Computer Assisted Personal Interview (CAPI) subsampling or the magnitude of Census 2000 Long Form sample cases

1 American Community Survey Quality Measures Webpage: http://www.census.gov/acs/www/UseData/sse/

This report is released to inform interested parties of (ongoing) research and to encourage discussion (of work in progress). Any views expressed on (statistical, methodological, technical, or operational) issues are those of the author(s) and not necessarily those of the U.S. Census Bureau.

where sample data was not collected on the final, realized sample sizes. We also do not address other data quality issues where it has been shown that the ACS out-performed the Census 2000 Long Form, such as item imputation rates[2].

2. ACS Sample Design

2.1 Sampling Frame

The sampling frame for the ACS is made up of addresses in the Master Address File (MAF) maintained by the Census Bureau. We have, over time, developed a specific set of criteria for addresses to be included in the sampling frame from the MAF. The primary source of new addresses in the ACS sampling frame is the Delivery Sequence File that the Census Bureau receives from the U.S. Postal Service at regular intervals.

In 2007 there were 132,841,861 addresses in the U.S. and 1,485,394 addresses in PR eligible for sampling. We have historically seen approximately two percent growth in the number of addresses on the frame. This growth, coupled with the fixed target sample size of three million, has led to decreasing sampling fractions over time at all levels of geography. As this trend continues we have become increasingly concerned that the reliability of the ACS estimates – especially at the lowest levels of geography such as census tract and block group – will be adversely affected.

2.2 Sample Selection

2.2.1 Overview

We select a sample of housing unit addresses from the MAF twice a year (Hefter, 2006a). Main sampling occurs in August and September of the year previous to the sample year and accounts for 99 percent of the sample. In January of the sample year, we select another sample from addresses that have been added to the MAF since main sampling. We refer to this as supplemental sampling. These sample cases account for approximately one percent of the total annual ACS sample. There are two stages of sampling: first-stage and second-stage sampling. The first-stage sample comprises approximately 20 percent of the total number of addresses on the frame. The remaining 80 percent is allocated to four equal groups. Each of these five partitions of the universe is ordered and they are rotated annually. The second-stage sample is selected from the current year's first-stage sample, ensuring no address is eligible for sampling more than once in any five-year period.

2.2.2 Sampling Rate Assignment

Under the differential sampling rate design of the ACS we assign each block to one of five second-stage sampling strata (Hefter, 2006b). These differing rates allow us to sample smaller areas at higher rates thereby selecting enough sample to produce reliable small area estimates. This process uses a measure of size calculated for each design area during main sampling. The set of design areas considered are:

- Counties, County Equivalents, and Municipios in Puerto Rico
- Places – *that are flagged as active*
- School Districts – *elementary, secondary, and unified*
- Minor Civil Divisions in the 12 "strong" MCD states: Connecticut, Maine, Massachusetts, Michigan, Minnesota, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island, Vermont, Wisconsin – *that are flagged as active*
- American Indian Areas
- Alaska Native Village Statistical Areas
- Hawaiian Homelands
- Tribal Subdivisions (starting with the 2007 sample selection) – *that are flagged as active*

With the exception of Tribal Subdivisions, these are the same geographic areas used in the Census 2000 Long Form sample design (Hefter, 1999).

2 See the C2SS/Census 2000 Comparison Studies at: http://www.census.gov/acs/www/AdvMeth/Reports.htm

We calculate a measure of size for each design area by multiplying the number of valid addresses on the frame by the Census 2000 block level occupancy rate. For American Indian and Alaska Native Village Statistical Areas we multiply the occupied housing unit estimate by the proportion of people who responded in Census 2000 as American Indian, alone or in combination. This is done in an effort to ensure that we produce useful (in terms of reliability) estimates of the American Indian and Alaska Native populations in these areas.

Each block is in several design areas, each with its own measure of size. We determine the smallest measure of size for each block and refer to this as the Governmental Unit Measure of Size (GUMOS). We also determine a measure of size for each census tract (TRACTMOS) and assign it to each block as appropriate. We then assign each block to a second-stage sampling stratum using these two measures and the following algorithm:

if (0< GUMOS <200) then second-stage sampling rate = 0.10 (stratum 5)
else if (200≤ GUMOS <800) then second-stage sampling rate = 3 × base rate (stratum 2)
else if (800 ≤ GUMOS ≤ 1200) then second-stage sampling rate = 1.5 × base rate (stratum 3)
else if (TRACTMOS ≥2000) then second-stage sampling rate = 0.735 × base rate (stratum 4)
else second-stage sampling rate = base rate (stratum 1)

The sampling rate for each stratum is determined by first calculating a new base rate (BR) each year. The sampling rate for four of the five strata is then calculated as a function of the BR. We incorporate projected growth between main and supplemental sampling into the base rate calculation to yield an annual target sample size of approximately three million addresses subsequent to supplemental sampling.

The base rate (BR) is rounded to four decimal places and is defined to be the smallest number such that:

$$\sum_{SBSTR=2} 3\times BR + \sum_{SBSTR=3} 1.5\times BR + \sum_{SBSTR=1} BR + \sum_{SBSTR=4} 0.735\times BR + \sum_{SBSTR=5} 0.10 \geq \left(3{,}000{,}000 - \sum_{all\,SBSTR} \text{projected growth} \right)$$

where the index on the summation runs through all valid addresses in the second-stage stratum.

The sampling rates in strata 1 and 4 are then reduced by eight percent for blocks in tracts with high expected mail and Computer Assisted Telephone Interview (CAPI) cooperation rates. This is to offset the cost of the differential CAPI sampling in areas with low cooperation that are sampled at higher rates (Asiala, 2005)

3. Analysis of the Percent in Sample

3.1 State Level Sampling Fractions – 2005 to 2007

Table 1 shows the sampling fraction distribution by state. The percent in sample clearly shows the effect of the fixed sample size as the address frame has grown. Every state except New Mexico has seen a decrease in the percent in sample in the first three years of full implementation. Arkansas shows the largest decrease in the percent in sample with a drop of 0.26 percent.

3.2 Distribution of Counties by State by Two Percent Threshold

In Table 2, we show the distribution of states by ranges of the percentage of counties with initial sample sizes below two percent. The number of states where all counties in the state have a sample size of more than two percent has decreased from 10 in 2005 to four in 2007. The percentage of counties by state with a sample size of less than two percent has not only grown from 2005 to 2007, but the number of counties with a sample size less than two percent has also increased over this time period. Rhode Island, has a sample size of less than two percent in all counties in the state. In 2007, only Hawaii, North Dakota, Vermont, and Puerto Rico had a sample size of greater than two percent in every county in the state.

3.3 Distribution of Places by State by Two Percent Threshold

In Table 3, we show the distribution of states by ranges of the percentage of places with initial sample sizes below two percent. No state has an ACS sampling fraction greater than two percent in every place in 2007. The percentage of places by state with a sample size of less than two percent has remained relatively stable from 2005-2007. Only the District of

Columbia (which is one place) was above two percent in sample in 2005. The District of Columbia has sample size of less than two percent for 2006 and 2007. This may be due to a higher level of urbanization in incorporated places and therefore there may be less opportunity for growth in the housing unit inventory.

3.4 Distribution of Tracts by State by Two Percent Threshold

Table 4 provides the distribution of states by ranges of the percentage of tracts with initial sample sizes below two percent. No states have a two percent or greater ACS sample size in every tract. The percentage of tracts by state with a sample size of less than two percent has changed from 2005-2007. There are twice as many states in the 40%-60% range from 2005 to 2006, and twice as many in 2007 as were present in 2006. No states have more than 60 percent of it's tracts with a sample size of less than two percent.

3.5 Comparison of Census 2000 Long Form and the ACS by Sampling Stratum

In Table 5, we show a comparison of the percentage of eligible addresses and selected sample for the Census 2000 Long Form and the 2005-2007 initially selected ACS sample by each sampling stratum. Note that the counts (eligible addresses and selected sample) shown from the long form sampling are only those addresses that appeared on the list of addresses included in Census operations (Decennial Master Address File) at the time of sample selection. Subsequent to the initial long form sampling, several field sampling operations occurred which sampled additions to the universe discovered during the update/leave, list/enumerate, and the non-response follow-up operations. The counts reflected in Table 5 only reflect the address frame based sampling that occurred.

The base rates used to define the ACS sampling rates for 2005-2007 were, 2.3%, 2.26%, and 2.23% respectively. This alone highlights the decrease in the percent in sample for the ACS over just the first three years of full implementation. Under the current design the base rate will continue to decrease over time.

Table 5 shows a higher percentage of the Census 2000 Long Form sample in the higher sampling rate strata as compared with the ACS. Significant change to the housing unit inventory over the five to seven years since the census long form universe was created leads to smaller proportions of the ACS sample being selected at the highest rates relative to the long form.

The differences seen in the distribution of the frame in the two lowest sampling rate strata as compared to the long form can most likely be attributed the fact that the ACS samples ungeocoded units at the base rate which is comparable to the long from rate of 1-in-6. Only addresses geocoded to a census collection block were eligible to be included Census 2000 operations, and therefore the long form sampling frame. Also, the reduction of the ACS sample in the two lowest sampling strata where there is overlap with tracts having the highest expected cooperation rates leads to a smaller percentage of the ACS sample in the $0.735 \times$ BR stratum.

4. Current Research

In response to the diminishing sampling fractions, in particular at the smaller levels of geography, we have begun to explore several sample design alternatives. In order to gauge how well the ACS is performing, we have completed a preliminary assessment of the reliability of tract level ACS estimates relative to the Census 2000 Long Form.

Tables 6 and 7 present results of this initial research designed to determine various levels of reliability for the ACS estimates. This approach determines the necessary annual sampling rate and sample size for ACS 5-year estimates to achieve various levels of reliability. The ACS levels of reliability are described as a function of the Census 2000 Long Form (LF) reliability, measured by the coefficient of variation (CV) of a fixed, generic 10 percent characteristic estimate. The CVs have been calculated for the proposed sampling rate and sample size changes with the current overall sampling rate provided for comparison.

4.1 Methodology and Definitions

4.1.1 Formulas

The following CV formula was used: $CV = \sqrt{((1-f)\times DE \times q))/(f \times p \times N)}$

Where: f =LF sampling rate = 0.17, DE = design effect for LF (see section 4.1.2), N = population size of an average tract = 4,200 (as of 2000), p = percentage of interest = 0.10, and $q = 1 - p = 0.90$.

The resulting LF CV was then used in the following formula: $f_{ACS} = (DE \times q)/((R * CV)^2 \times p \times N + (DE \times q))$

where: CV=the LF CV=0.153, R=inflation factors for the LF CV=1.25, 1.33, 1.47, 1.63, or 1.75, DE=design effect for ACS, N=population size of an average tract=4,500 (as of 2006), p=percentage of interest=0.10, and q=1–p=0.90.

The five-year ACS sample size needed for each level of reliability was calculated as n = f_{ACS} × the total number of address, where: f_{ACS} = the ACS sampling rate, and total number of addresses = 130,683,466 (as of 2006).

The margin of error (MOE) is calculated as: MOE=SE × 1.645, where SE=the standard error for an estimate. Note that 1.645 is used for to generate a 90 percent confidence interval.

4.1.2 Design Factor

The DE used for the CV_{LF} was 2.25, which is the square of the published LF DF of 1.5 for estimates of people in poverty. The *DE* used for the CV_{ACS} was 4.41, which is the square of the average DF for three ACS poverty statistics:

- Poverty Rate of Children 5 - 17
- Poverty Rate of Families
- Poverty Rate of the Population

4.1.3 Assumptions

The following assumptions were necessary to generalize our research to the entire U.S. population:

- The CV calculations assume average values of the characteristic throughout the population.
- The calculation of the CVs is based on the assumption that the proportion (P) is fixed through the estimation period.
- The population growth rate is assumed to be uniform across all geographic areas and across years.
- All calculations are based on the number of addresses and not occupied housing units.

4.1.4 Limitations

- ACS sample sizes needed to match the reliability of the LF is based on an overall sampling rate.
- ACS sampling takes place two times each year, with an adjustment for growth made between phases. No adjustment for growth has been included in this preliminary analysis; all calculations are made assuming all sample is selected in one phase.

4.2 Results

Table 6 shows the summary for five different designs. The sampling fraction of 2.20 percent represents the state of the ACS as of 2008. The other four designs show larger sample sizes, increased reliability, and improved margins of error at 90 percent. The most reliable design has an annual sampling rate of 3.9 percent with an annual sample size over 5.2 million addresses. Another way we can look at this information is the impact on the confidence level for a fixed MOE of 0.0371. This MOE gives a 90 percent confidence level for a generic 10 percent estimate using a sampling rate of 3.0 percent. Table 7 shows how the confidence level changes for this fixed MOE as the sampling rate changes.

So, in our most reliable design we have a one in twenty chance of the true value being outside the confidence interval (formed by the MOE = 0.0371) while under our current design we have a one in six chance.

5. Conclusions

It is clear that over the first three years of full implementation of the ACS, the effect of the requirement that the annual target housing unit address sample be fixed at three million is reflected at many levels of geography. The downward trend

of the sampling fractions at the county and tract level could lead to concerns about the standard errors of the ACS estimates for each estimate period (1-, 3-, and 5-year). We do note that the sampling fractions for places appears to be relatively stable.

The distribution of the address frame by sampling stratum for the ACS compared to the Census 2000 Long Form has changed as well. A smaller percentage of the ACS universe is being sampled at the higher rates relative to the Census 2000 Long Form universe.

In order to be responsive to the ever changing population of the United Sates, and to the increased demands being placed on the ACS to serve as a key decision making tool for numerous stakeholders and data users, the following options should be considered:

- Consider making the shift – sooner rather than later – from a fixed target sample size to a constant, target sampling rate. This would entail an annual sample size increase of approximately 1.6 percent each year to account for growth in the frame.
- Increasing the sample size to roughly 3.9 million per year would only provide estimates with reliability comparable to the Census 2000 Long Form under the assumptions given. A more detailed investigation is in process, which accounts for growth in the frame and estimates the five-year ACS sample size needed post-2010.
- Research into the optimal number of sampling strata is needed. This research could lead to an improvement in sampling efficiency, which could be implemented with little or no increase in cost, while providing an even more comparable distribution of standard errors across all levels of geography. This is important goal of the ACS, specifically of the multi-year estimates.

Acknowledgements

We wish to thank the following people who provided significant help by generating and assisting us in analyzing the data contained in this paper, or aided by providing many clear and useful comments and suggestions: Edward C. Castro Jr., Karen E. King, Alfredo Navarro, Robyn Sirkis.

References

Asiala, M. 2005. "American Community Survey Research Report: Differential Sub-Sampling in the Computer Assisted Personal Interview Sample Selection in Areas of Low Cooperation Rates". Draft - Internal U.S. Census Bureau Memorandum to R. Singh from D. Hubble, February 15, 2005.

Hefter, S. 1999. "Long Form Sampling Specifications for Census 2000." Internal U.S. Census Bureau Memorandum to M. Longini from H. Hogan, Washington, DC, November 17, 1999.

Hefter, S. 2005. "American Community Survey: Specifications for Selecting the Computer Assisted Personal Interview Samples." Draft - Internal U.S. Census Bureau Memorandum to L. McGinn from R. Singh, Washington, DC, July 27, 2005.

Hefter, S. 2006a. "Specifications for Selecting the Main and Supplemental Housing Unit Address Samples for the American Community Survey. Draft - Internal U.S. Census Bureau Memorandum to S. Schechter from D. Whitford August 23, 2006.

Hefter, S. 2006b. "Creating the Governmental Unit Measure of Size (GUMOS) Datasets for the American Community Survey and the Puerto Rico Community Survey". Draft - Internal U.S. Census Bureau Memorandum to S. Schechter from D. Whitford, June 6, 2006.

Table 1. State Level Sampling Fractions – 2005 to 2007

State	2005	2006	2007	State	2005	2006	2007
Alabama	2.36	2.29	2.25	Montana	3.22	3.12	3.05
Alaska	3.45	3.40	3.33	Nebraska	3.22	3.15	3.05
Arizona	1.99	1.95	1.97	Nevada	2.00	1.96	1.94
Arkansas	2.55	2.46	2.29	New Hampshire	2.54	2.46	2.40
California	2.03	1.99	1.96	New Jersey	2.05	2.01	1.98
Colorado	2.16	2.10	2.07	New Mexico	2.32	2.27	2.35
Connecticut	1.97	1.94	1.91	New York	2.26	2.22	2.18
Delaware	2.46	2.42	2.39	North Carolina	2.08	2.03	1.95
District of Columbia	2.05	2.00	1.97	North Dakota	3.77	3.68	3.61
Florida	1.87	1.83	1.80	Ohio	2.14	2.09	2.06
Georgia	2.03	1.98	1.96	Oklahoma	2.85	2.77	2.73
Hawaii	2.48	2.39	2.35	Oregon	2.12	2.08	2.04
Idaho	2.49	2.40	2.35	Pennsylvania	2.59	2.54	2.50
Illinois	2.23	2.18	2.14	Rhode Island	1.90	1.86	1.84
Indiana	2.16	2.11	2.08	South Carolina	2.04	1.99	1.95
Iowa	2.91	2.85	2.80	South Dakota	3.33	3.25	3.18
Kansas	2.66	2.60	2.56	Tennessee	2.01	1.97	1.94
Kentucky	2.18	2.12	2.09	Texas	2.16	2.10	2.07
Louisiana	2.34	2.28	2.25	Utah	2.31	2.26	2.24
Maine	3.41	3.32	3.26	Vermont	3.91	3.79	3.73
Maryland	1.98	1.94	1.91	Virginia	1.94	1.90	1.87
Massachusetts	1.93	1.89	1.87	Washington	2.14	2.10	2.07
Michigan	2.71	2.65	2.61	West Virginia	2.39	2.33	2.29
Minnesota	3.41	3.33	3.28	Wisconsin	3.27	3.20	3.14
Mississippi	2.18	2.11	2.07	Wyoming	2.52	2.46	2.45
Missouri	2.42	2.36	2.32	Puerto Rico	2.43	2.41	2.43

Table 2. Number of States by Percentage Range of Counties With A Sample Size Less than Two Percent

Year	Number of States with ...					
	all counties > 2% in Sample	less than 20% of counties with < 2% in Sample	20% - 40% of counties with < 2% in Sample	41% - 60% of counties with < 2% in Sample	61% - 80% of counties with < 2% in Sample	81% - 100% of counties with < 2% in Sample
2005	10	25	11	2	3	1
2006	6	25	10	6	3	2
2007	4	24	11	8	3	2

Table 3. Number of States by Percentage Range of Places With A Sample Size Less than Two Percent

Year	Number of States with ...					
	all places > 2%	less than 20% of places with < 2% in Sample	20% - 40% of places with < 2% in Sample	41% - 60% of places with < 2% in Sample	61% - 80% of places with < 2% in Sample	81% - 100% of places with less than 2% in Sample
2005	1	18	21	9	3	0
2006	0	15	23	9	4	1
2007	0	14	23	9	5	1

Table 4. Number of States by Percentage Range of Tracts With A Sample Size Less than Two Percent

Year	Number of States with ...					
	all tracts > 2% in Sample	less than 20% of tracts with < 2% in Sample	21% - 40% of tracts with < 2% in Sample	41% - 60% of tracts with < 2% in Sample	61% - 80% of tracts with < 2% in Sample	81% - 100% of tracts with < 2% in Sample
2005	0	12	37	3	0	0
2006	0	8	38	6	0	0
2007	0	5	34	13	0	0

Table 5. 2005-2007 ACS Universe and Sample, and the Census 2000 Long Form Universe and Sample by Stratum

Survey	Count	Sampling Stratum (Census 2000 Long Form; ACS)			
		1-in-2; 10%, 3 × BR	1-in-4; 1.5 × BR	1-in-6; BR	1-in-8; 0.735 × BR
Census 2000 Long Form	Addresses	6.9	2.8	39.6	50.7
	Sample	20.3	4.1	38.5	37.0
2005 ACS	Addresses	5.6	2.5	48.0	43.8
	Sample	18.4	4.0	46.7	31.0
2006 ACS	Addresses	5.5	2.5	47.4	44.7
	Sample	18.1	3.8	46.3	31.7
2007 ACS	Addresses	5.3	2.5	47.1	45.1
	Sample	17.8	3.8	46.2	32.1

Table 6. ACS Sampling Rates and Sample Sizes For Various Levels of Reliability

ACS Annual Sampling Rate (f)	ACS Annual Address Sample Size (*n*) in millions	CV_{ACS} Level of Reliability	CV_{ACS} as a function of the CV_{LF}	MOE at the 90 percent Confidence Level
3.90%	5.2	19.10%	1.25 CV_{LF}	0.0315
3.50%	4.6	20.40%	1.33 CV_{LF}	0.0335
3.00%	3.9	22.50%	1.47 CV_{LF}	0.0371
2.50%	3.3	24.90%	1.63 CV_{LF}	0.0411
2.20%	2.9	26.80%	1.75 CV_{LF}	0.0442

Table 7. Impact on Reliability of a Fixed Margin of Error by Sampling Rate

ACS Annual Sampling Rate (f)	ACS Annual Address Sample Size (*n*) (in millions)	CV_{ACS} Level of Reliability	CV_{ACS} as a function of the CV_{LF}	Confidence Level of the MOE=0.0371
3.90%	5.2	19.10%	1.25 CV_{LF}	95%
3.50%	4.6	20.40%	1.33 CV_{LF}	93%
3.00%	3.9	22.50%	1.50 CV_{LF}	90%
2.50%	3.3	24.90%	1.63 CV_{LF}	86%
2.20%	2.9	26.80%	1.75 CV_{LF}	83%