

Census 2010 and the American Community Survey

*Counting Everyone Once — and Only Once
— and In the Right Place*

Elana Broch May 28th 2009

Segmentation Clusters for Communications Campaign

Cluster Name	Census Tract Characteristics (Census 2000 data)	Occupied Housing Units	HTC Score	2008 Mail Response
Advantaged Homeowners	- Single family households - Married couples - Homeowners	28%	6	75%
All Around Average I (homeowner skewed)	- Average poverty, education, mobility and unemployment - Highest rural cluster	33%	23	66%
All Around Average II (renter skewed)	- Average poverty, education, mobility and unemployment - Renters in multi-unit buildings	16%	41	66%
Young Mobile Singles	- Non-spousal households - Renters in multi-unit buildings - High income	8%	61	59%
Economically Disadvantaged I (homeowner skewed)	- High poverty, public assistance, unemployment - 1/3 homeowners	6%	65	55%
Economically Disadvantaged II (renter skewed)	- High poverty, low education, high unemployment - Renters in multi-units - Very urban	3%	92	45%
Ethnic Enclave I (homeowner skewed)	- Crowded housing, high poverty - Low education - Mostly spousal households	3%	63	60%
Ethnic Enclave II (renter skewed)	- Urban - Low education - 50% Hispanic, 11% Asian	2%	84	57%

United States
Census 2010

USCENSUSBUREAU

The goal of the 2010 Census is to count all residents living in the United States on April 1, 2010.


- It is used for apportionment of House of Representatives seats
- Redistricting of state legislative districts
- Allocation of hundreds of billions of dollars of federal money

<http://2010.census.gov/2010census/pdf/OverviewRelease.pdf>

2

To keep in mind as we move on to discussing the American Community Survey

- There is no corresponding advertising program for the ACS.
 - Ethnic enclaves are still enclaves
 - Economically disadvantaged are still disadvantaged



5

Besides the enormity of the task there are other factors that complicate the process

- Language barriers
- Undocumented residents' fear of being deported if located
 - The U.S. Census Bureau does not ask about the legal status of respondents in any of its surveys and census programs.
- Distrust of government by some legal residents
 - CB hopes the increase in voter participation in 2008 may foreshadow increased civic participation (census completion)
- Confusion as to where to be counted
- Differential undercount

3

Types of missing data

- Houses that aren't on census rosters
- Missing households=People who don't return forms
Includes but not limited to hard to count groups
- Missing people within households that do return forms
- Missing data from census forms that are returned
- Errors—deliberate or accidental

6

Final Rates, 1970-2000 Censuses (as of end of census year)

	1970	1980	1990	2000
Mail response rate:	78%	75%	65%	67%
Mail return rate:	87%	81%	75%	78%

Hopefully short form only will increase participation ("10 questions, 10 minutes")

7

By early 2011 expect...

Swarms of people asking us for census data help including "when will the long form data become available?"

There will be no long form and no long form data.

10

Techniques for missing data

Send an enumerator to try to get missing information.

If this fails,

- Deleting cases (not a census option)
- Ignoring (report as missing)
- Imputation

What is imputation? Imputation is a group of statistical techniques used to predict values of missing data based on what is known about that person and/or their household.

Imputation on short form (Name, sex, age, date of birth, race, ethnicity, relationship and housing tenure)

e.g., age from date of birth, ethnicity from ethnicity of others in household.

8

**"Long form"**

- Administered to a sample of respondents
- First used in 1940
- Designed to get additional data but without burdening the entire population

Norman Rockwell, Saturday Evening Post cover, April 27, 1940.

11

Summary of 2010 census

- Only short form will be given
- Long form replaced by the American Community Survey
- No plans for statistical adjustment
- Non-response follow-up by enumerators for missing forms not missing responses on returned forms.
- Missing data dealt with via imputation
- There is much missing data

9

Sampling for long form

- Long form given to ON AVERAGE 1/6 households
- Oversampling of smaller geographic areas and areas with lower response rates
- Results are extrapolated to the entire population using statistical techniques (primarily weighting—see next slide)

12

Weighting the long form

Even if there were no missing responses the long form would need to be weighted so that they totals match the short form totals.

Weighting is a complex statistical procedure that adjusts the responses to the long form so that they equal the totals on the short form.

If there are 40 white male Hispanic heads of households containing 3 people in Trenton based on the short form, the long form responses are weighted so it looks like the long form data are based on 40 white Hispanic heads of households containing 3 people. That is, if there are 8 long form responses for white Hispanic heads of household containing 3 people in Trenton, each person's response to the long form questions count five times. So, the 1 of 8 who have a graduate degree gets reported as 5 of 40.

13

<http://www.amstat.org/sections/srms/Proceedings/y2002/Files/JSM2002-000555.pdf>

IMPUTATION on long form (continued)

- But how do you determine the variables to base the imputation on?
- How do you handle the additional complexities of allowing people to report more than one race?
- What do you do if some of the values you want to base your imputation on are themselves missing?
- What level of geography should be used?
- How do you get your data users to understand the estimation error involved in these estimates?

16

Weighting the long form (continued)

The process of matching the long form counts to the short form counts has to be repeated for each level of geography (down to block groups)

You can see the problems that arise for the weights for American Indian Hispanics in households with 7 people.

Imagine how complicated this gets when respondents can pick more than one race category.

14

Population estimates program annually produces population estimates and projections down to the county level. This includes estimates of age, sex, race, and Hispanic origin. It is based on extrapolating the most recent decennial census count, taking into account births, deaths, and migrations.

<http://www.census.gov/popest/overview.html>

17

IMPUTATION for missing data on long form

!!!! On the long form there is the extrapolation from the sample to the population to begin with.

In addition, 29.7% of long form records in 2000 had at least some income imputed, compared to 13.4% in 1990 (<http://www.census.gov/srd/papers/pdf/rrs2008-13.pdf>).

Take the case of someone whose salary is missing.
The idea behind multiple imputation is that you could guess at salary by
Assigning the mean salary OR
Assigning the mean salary for that person's gender, race, age, and occupation (or some combination of these variables) OR
Assigning your neighbor's salary.

The resulting estimates are just that—estimates. Proper use of them requires standard errors of the estimate.

15

THE American Community Survey

U.S. DEPARTMENT OF COMMERCE
Economic and Research Administration
U.S. CENSUS BUREAU

La Encuesta sobre la Comunidad Estadounidense

DEPARTAMENTO DE COMERCIO DE LOS EE.UU.
Administración de Economía y Estadísticas
OFICINA DEL CENSO DE LOS EE.UU.

18

The American Community Survey

- Major new continuous survey designed to provide small-area data
- Replaces long form
- Monthly survey of 250,000 households
- Participation is required by law
- Available in English and Spanish
- Very detailed (10 pages of questions per person)
- Census bureau estimates it takes 38 minutes per household to complete

19

ACS creates a new statistical landscape (but with many similarities to the long form)

- As was the case with the long form
 - Sampling rates vary based on size of geographic unit and anticipated response rate
 - Results not available at certain geographies
 - Weighting to population controls
 - Imputation of missing data

22

ACS very similar in content to long form

- ACS items not on the 2000 long form:
- (1) whether the household received food stamps in the previous 12 months and their value;
- (2) the length of time and main reason for staying at the address (for example, permanent home, vacation home, to attend school or college);
- (3) for women ages 15– 50, whether they gave birth to any children in the past 12 months.

20

ACS Sampling rates

- Sampling rate is based on size of geographic unit and anticipated response rate in census tract.
- We always think of long form sampling as 1/6 but really different place sizes are sampled differently to make sure there is adequate sample size for smaller places. The same logic applies to the ACS.
- Likewise there has always been oversampling of groups and areas that have lower response rates.

23

ACS follow-up using trained enumerators

- The used of trained enumerators is supposed to greatly increase the completeness of the data (of course, at a huge expense)
- About 33 percent of mail questionnaires in 2005 required telephone follow-up because key items were missing or because households reported more members that there was room to provide information.

21

TABLE 2-3a. Housing Unit Addresses, 2005 ACS and 2000 Census Long-Form Sample: Approximate Initial Block-Level Sampling Rates

Type and Size of Smallest Area Containing a Block	2005 American Community Survey		2000 Long-Form
	Annual Initial Sampling Rate	Cumulative 5-Year Initial Sampling Rate	Sample Census Day Sampling Rate
Governmental unit (county, place, township in 12 states, school district, American Indian or Alaska Native area)			
With < 200 occupied housing units (fewer than about 800 people)	10.0% (1 in 10)	10.0% (1 in 2)	10.0% (1 in 2)
With 200-800 occupied housing units (about 500-2,000 people)	6.9% (1 in 14)	34.5% (1 in 3)	50.0% (1 in 2)
With 800-1,200 occupied housing units (about 2,000-3,000 people)	3.5% (1 in 28)	17.5% (1 in 6)	25.0% (1 in 4)
Census tract with > 2,000 occupied housing units (more than about 5,000 people) ^a	1.7% (1 in 59) or 1.6% (1 in 63)	8.5% (1 in 12) or 8.0% (1 in 13)	12.5% (1 in 8)
Other area ^b	2.3% (1 in 44) or 2.1% (1 in 48)	11.5% (1 in 9) or 10.5% (1 in 10)	16.7% (1 in 6)
Overall	2.3% (1 in 44)	11.5% (1 in 9)	16.7% (1 in 6)

NOTES: Number of occupied housing units is estimated from the MAE. Because the initial ACS sample size will be kept at approximately 1 million residential addresses per year, the initial sampling rates shown will be slightly reduced as the number of occupied housing units grows. Township and other minor civil divisions are recognized for sampling purposes in 12 states where they are functioning governments: Connecticut, Maine, Massachusetts, Michigan, Minnesota, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island, Vermont, and Wisconsin.

^aThe smaller of the two ACS sampling rates shown applies for blocks in census tracts with predicted mail/CAIT response rates greater than 60% (see Table 2-3b).

SOURCE: Adapted from U.S. Census Bureau (2006/Tables 4.3, 4.2) for the ACS.

24

TABLE 2-3b Housing Unit Addresses, 2005 ACS and 2000 Census Long-Form Sample: Census Tract-Level CAPI Sampling Rates in the 2005 ACS for Mail/CAPI Nonrespondents^a

Sampling Rate Category	CAPI Subsampling Rate	Illustrative Completed Sample Cases as Percentage of Initial Sample
Tracts with predicted mail/CAPI use rate less than or equal to 35%	50% (1 in 2)	If, say, 20% mail/CAPI response, then completed sample will be 60% of initial sample—20% mail/CAPI plus 40% CAPI (1/2 of 80%)
Tracts with predicted mail/CAPI use rate between 36 and 50%	40% (2 in 5)	If, say, 40% mail/CAPI response, then completed sample will be 64% of initial sample—40% mail/CAPI plus 24% CAPI (2/5 of 60%)
Tracts with predicted mail/CAPI use rate between 51 and 60%	33% (1 in 3)	If, say, 55% mail/CAPI response, then completed sample will be 70% of initial sample—55% mail/CAPI plus 15% CAPI (1/3 of 45%)
Tracts with predicted mail/CAPI use rate greater than 60% ^b	33% (1 in 3)	If, say, 80% mail/CAPI response, then completed sample will be 87% of initial sample—80% mail/CAPI plus 7% CAPI (1/3 of 20%)

^aUnmailable addresses are followed up in the CAPI data collection phase at a rate of 67% (2 in 3).
^bCensus tracts outside oversampled governmental units with high predicted mail/CAPI response rates, the initial sample is reduced by a factor of 2 (see Table 2-3a). This reduction is implemented to satisfy a budget constraint for personal interviewing.

NOTE: Adapted from U.S. Census Bureau (2006:Tables 4.1, 4.2) for the ACS.

25

Another aspect of the statistical landscape that is used with the census but one you may not have been aware of.

- In much statistical work the standard error is computed by dividing the observed standard deviation by the (sq. rt of) the sample size of the data with which you are working.
- Because of the division, the larger the sample size the smaller the standard error.

28

Weighting to Population Controls

- The long form was matched to the short form census counts.
- Estimates of housing units and people are controlled to the population estimates derived from the Population Estimates Program
- It is also weighted "...for persons ...to agree with independent estimates of people in terms of age, sex, race, and ethnic groups in the area as of July 1" (p. 47-48)

26

Dirty little secret on standard errors

- In the ACS, the standard errors of reported statistics values are not computed in the manner just described but estimated using complex statistical techniques.
- This procedure (or something similar) has been used with the census and the CPS for decades.
- This procedure is commonly used with complex sampling designs, particularly when there is missing data.
- Like "real" standard errors it is greatly influenced by the size of the sample you are working with.
- These standard errors are then used to compute the Margins of Error that are reported in the data tables (we'll see this later)

Weighting to controls, pg. 2

- "The population controls used in the ACS are midyear population estimates (PEP) based on different residence rules and different sources than the yearly accumulation of ACS monthly samples....It cannot be assumed that they necessarily improve the quality of the area level and therefore are subject to appreciable error, particularly since they are applied at the estimation area level and therefore are subject to appreciable error." (p. 207)
- Because of the way multiyear estimates are weighted, users should not expect the ACS demographic estimates to match any individual year's population estimate within the time period.

from Citro and Kalton (2007)

27

How are these standard errors estimated?

- ACS standard errors are estimated using a computer simulation.
- This process is called replicate-based variance estimation
- Calculate the statistic of interest for each subsample, and then use these subsamples to estimate the variance of the full-sample statistic.
- The variation between the replicate estimates and the full-sample estimate is then used to estimate the variance for the full sample.
- Like "real" standard errors it is greatly influenced by the size of the sample you are working with.

<http://www.wesstat.com/wesvar/about/WV4.2%20Manual.pdf> pg. A-2

30

New ACS statistical concepts

- Reference years
- 1-, 3-, 5-years worth of data depending of size of geographic unit
- Annual updates
- Acknowledgement of margin of error
- Acknowledgement of imputation rates

32

ACS data availability will depend on the size of the population you are studying

Data Product	Population Size of Area	CY 2006	CY 2007	CY 2008	CY 2009	CY 2010	CY 2011	CY 2012	CY 2013
1-Year Estimates for Data Collected in:	65,000+	2005	2006	2007	2008	2009	2010	2011	2012
3-Year Estimates for Data Collected in:	70,000+			2005-2007	2006-2008	2007-2009	2008-2010	2009-2011	2010-2012
5-Year Estimates for Data Collected in:	All Areas*					2005-2009	2006-2010	2007-2011	2008-2012

* Five-year estimates will be available for areas as small as census tracts and block groups.
Source: US Census Bureau

35

New Statistical Concept—Reference Periods

- Questions about the previous year (e.g., about income earned) will refer to different 12-month periods--known as reference years--and this window will move as the data collection progresses through the year.

33

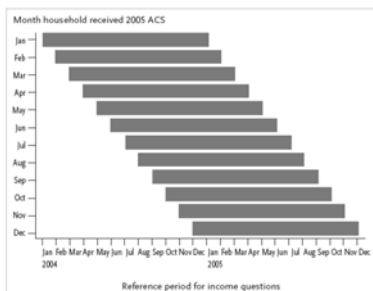
ACS provides annual updates

- By 2010, data should be available for the smallest areas around the same time we would have had long form data for these places. Once the data are available for a particular place, they will be updated annually.
- This is one of the strengths of the ACS, but also one of its limitations because it is nearly impossible to interpret consecutive multiyear estimates.

36

Reference Periods for ACS

from Mather 2005



34

An embarrassment of riches for large places (65,000 or more people)

- There will be 1-, 3-, and 5-year estimates available. The user needs to decide which would be most appropriate to use.
- The larger samples (those combined across years) will have smaller standard errors (more precise results).
- A 65,000+ place will have much smaller subsamples (e.g., female heads of household) and if you are interested in subgroups you may need to use the 3- or 5-year.
- To preserve confidentiality the year on which the data are based is not reported in the multi-year results.
- Disadvantages of multi-year
 - Harder to interpret
 - Not as recent

37

Interpretation of multi-year data

<http://www.census.gov/acs/www/Downloads/ACSResearch.pdf>

“Comparisons should be based on nonoverlapping periods (e.g., comparing estimates from 2006–2008 with estimates from 2009–2011). The comparison of two estimates for different, but overlapping periods, is challenging since the difference is driven by the nonoverlapping years, but diluted across the number of years in common. While the interpretation of this difference is difficult, these comparisons can be made with caution.” (p. A-20)

Because of the differences in the sampling errors associated with the different numbers of years you cannot compare a 1-year with a 3-year or with a 5-year.

When making comparisons across geographic areas do not cross-compare one-year, three-year, and five-year estimates.

38

Why all the fuss about MoE?

The reason MoE gets mentioned more in the context of the ACS is that it is more pertinent. This is because sample sizes are much smaller in the ACS than they were in the long form. For national level data the sample sizes are adequate but for smaller geographies or subgroups of people (e.g., female heads of households) the sampling error may be so large that it renders the results meaningless or of limited value.

Remember we talked about how the standard error is estimated using the replicate-based variance estimation.

Once the standard error is estimated it is used to compute the Margin of Error.

41

Understanding Multiyear Estimates from the American Community Survey

http://www.census.gov/acs/www/Downloads/ACS_Understnd_Multiyr_Est.ppt

39

Sample ACS results showing Margin of Error

PRINCETON REGIONAL, New Jersey
 S1901, Facility
 Data for: 2005-2007 American Community Survey 3-Year Estimates
 Survey: American Community Survey

NOTE: For information on confidentiality protection, sampling error, nonresponse error, and definitions, see Current Methodology.

Subject	Total		Rate per 1,000 women		Percent of women who had a birth in the past 12 months who were unemployed			
	Margin of Error	Number	Margin of Error	Margin of Error	Margin of Error	Margin of Error		
Women 15 to 54 years	1.566	2,768	288	<1.88	29	+1.14	45.5%	+25.7
15 to 19 years	1.512	<1,000	0	<1.62	0	+1.27	-	-
20 to 24 years	2.522	<1,000	117	<1.61	48	+1.33	47.0%	+37.2
25 to 54 years	1.548	<1,000	92	<1.69	29	+1.22	42.4%	+41.1

Subject	Total		Rate per 1,000 women		Percent of women who had a birth in the past 12 months who were unemployed			
	Margin of Error	Number	Margin of Error	Margin of Error	Margin of Error	Margin of Error		
Facility	3.07%	(X)	(X)	(X)	(X)	(X)	(X)	(X)
Facility	2.4%	(X)	(X)	(X)	(X)	(X)	(X)	(X)

42

Summary of statistical changes in ACS

- Reference Periods
- Multi-year data and the issues around interpreting the results.
- Greater acknowledgement of imputation rates and margins of error
- Differences in the way data are collected (no ad campaign), less foreign languages, not tied to the short form census that may result in nonresponses, particularly differential non-responses.
- Our last topic is the way the sampling error is quantified in the ACS and how that can be used in interpreting data.

40

Using the margin of error

- The MoE is 1.65 times the Standard Error
- The margin of error can be used to establish a confidence interval
- To address this the coefficient of variation is introduced to provide guidance in interpreting results. The Census Bureau uses 10-12% as a tolerable range for CV.

43

BOX 2-5
Brief Descriptions of Statistical Terms Used in This Report

- **Standard error of an estimate:** A commonly used statistic that expresses the imprecision in an estimate that is due to sampling. This imprecision is known as sampling error. It is to be distinguished from nonsampling errors from such sources as misreporting and nonresponse, which are often systematic in nature and result in biased survey estimates (see Box 2-3).
- **Coefficient of variation (CV) or relative standard error:** The standard error expressed as a percentage of the estimate. CVs of 10-12 percent or less are often accepted as a reasonable standard of precision for an estimate.
- **90 percent margin of error (MOE):** Plus or minus 1.65 times the standard error of an estimate.
- **90 percent confidence interval (CI):** The 90 percent MOE expressed as a range around the estimate.

Example Calculations

Consider the example of MEDIUM CITY, 5-Year Period ACS Estimate (see Tables 2-7a, 2-7b, and 2-7c). Assume that MEDIUM CITY has a population of 100,000 with an estimated 20,000 school-age children, of whom 3,000 (15 percent) are estimated to be poor. For a 15.0 percent poverty rate for school-age children with a 1.20 percentage point standard error:

- CV = 0.5 percent (1.20/15.0)

44

These concepts can be extended to making comparisons

Statistical differences – Users may conduct a statistical test to see if the difference between an ACS estimate and any other census estimates is statistically significant at a given confidence level. “Statistically significant” means that the difference is not likely due to random chance alone. With the two estimates (E_{1t} and E_{2t}) and their respective standard errors (SE_1 and SE_2), calculate

$$Z = \frac{E_{1t} - E_{2t}}{\sqrt{(SE_1)^2 + (SE_2)^2}}$$

If $Z > 1.645$ or $Z < -1.645$, then the difference can be said to be statistically significant at the 90 percent confidence level. [Note that we are now recommending that -1.645 be used to determine significance. Previous ACS Accuracy of the Data documents suggested using -1.65 .] Any estimate can be compared to an ACS estimate using this method, including other ACS estimates from the current year, the ACS estimate for the same characteristic and geographic area but from a previous year, Census 2000 100 percent counts and long form estimates, estimates from other Census Bureau surveys, and estimates from other sources. Not all estimates have sampling error – Census 2000 100 percent counts do not, for example, although Census 2000 long form estimates do – but they should be used if they exist to give the most accurate result of the test.

Users are also cautioned to not rely on looking at whether confidence intervals for two estimates overlap to determine statistical significance, because there are circumstances where that method will not give the correct test result. The Z calculation above is recommended in all cases.

All statistical testing in ACS data products is based on the 90 percent confidence level. Users should understand that all testing was done using unrounded estimates and standard errors, and it may not be possible to replicate test results using the rounded estimates and margins of error as published.

45

<http://www.hulu.com/watch/4165/saturday-night-live-census-taker>

46