GENERALIZED VARIANCE ESTIMATES IN THE 1996 AMERICAN COMMUNITY SURVEY

Anthony G. Tersine and Alfredo Navarro, U.S. Bureau of the Census[1]
Anthony G. Tersine, U.S. Bureau of the Census, Washington, D.C. 20233

Key Words: Generalized Variances, Design Factor

1.0 Introduction

The 1996 American Community Survey (ACS) data products include population and housing unit estimates based on a sample of housing units and the population residing in the sampled housing units. The ACS Data Products Program (DPP) will tabulate and publish estimates for over 1,000 data items for census block groups, tracts, and places. These data products are equivalent to the decennial census summary tape files (STF-3 and 4.) Although it is possible to both calculate and publish a variance estimate for each published estimate the ACS staff feel obligated to satisfy data needs for most if not all data users. The initial strategy is to provide direct variance estimates for those users who are not limited by computer power and prefer more accurate measures of sampling error. To complement this strategy and serve the needs of data users with computer limitations the ACS is also providing measures of sampling errors based on generalized variance estimates. The decennial census has used the latter approach in every census since 1960. In 1970, a design effect approach was used to generalize the estimated census variances. The 1980 and 1990 censuses used a similar method. The use of a design effect approach by the ACS will provide a smooth transition for decennial data users that want to use and experiment with the ACS data.

The design effect method is very easy to understand and use. The method is based on first computing two variance estimates for each tabulation area (census tract) and data item. The two variance estimates are the direct variance and the variability that would result from a simple random sampling (srs)design and a simple inflated total or mean. The ratio of these two variance estimates is calculated, and the average over all tabulation areas in the demonstration site is calculated. The challenge is to not only provide measures of sampling error for the small areas (i.e., census block groups and tracts), but provide reasonable approximations for larger areas as well. The square roots of the design effects are then published along with tables of simple random sampling standard errors for total and percentages. Users are then instructed to combine the srs standard error estimates in the tables with the design factors to approximate the standard error estimate for any given data item and tabulation area.

A brief overview of the sample design and estimation methods is given in section 2. The procedures for measuring sampling errors associated with ACS estimates are explained in section 3. Section 4 describes the design effect approach and issues related to groupings of data items. Section 5 explains the evaluation methodology and discusses analytical results based on quantitative criteria. Section 6 discusses special cases, such as: small or zero estimates and the estimate of Total number of children ever born.

2.0 Sample Design and Estimation Methods

This section describes the sample design and estimation methods used in the four 1996 ACS demonstration sites: Multnomah County/Portland OR, Rockland County NY, Brevard County FL, and Fulton County PA. These areas had an especially high sampling rate in the first year, 15 percent in most areas and 30 percent in small governmental units, so that detailed estimates could be studied without waiting for multiple years' data. For the 1996 ACS, "small" governmental units are local areas with 1,000 or fewer addresses. The decennial census long form samples small governmental units (defined as less than 2,500 population in 1990) at a higher rate than other areas.

The ACS estimation process (Alexander et al, 1997) consists of:

- editing the responses
- imputing missing responses
- confidentiality edit ("swapping" data for disclosure avoidance)
- weighting sample households and persons

3.0 Method of Variance Estimation

The 1996 ACS used Successive Difference Replication to produce direct variance estimates for all tabulation areas in the demonstration sites. The tabulation areas are census block groups, census tracts, and places. The

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

replicates are formed via VPLX which is a general purpose variance estimation software widely used at the Census Bureau. The ACS variance estimate is based on creating **80** subsamples or replicates of the full sample in a fashion that mimics the original sample design.

The method of successive difference replication pairs neighbors in the order of their selection to take advantage of the systematic nature of the ACS sample selection scheme. This process is repeated for each census tract within a site. Replicate factors are assigned using a Hadamard matrix with dimension 1,600 by 1,600. The use of this matrix maximizes the degrees of freedom and consequently the precision of the variance is improved considerably. Three possible replicate factors were used (1.0,1.7, and 0.3). The replicate factor is computed as $f_{i,r} = 1 + 2^{-3/2} a_{i,r} - 2^{-3/2} a_{i,r}$ where $a_{i,r}$ is either +1 or -1 from the i-th row of the Hadamard matrix. In expectation, about 50 percent will be assigned 1.0, and the other 50 percent will be evenly split between 1.7 and 0.3. Each replicate sample and the original sample are then weighted independently using the 1996 weighting procedures described before. As a result of this process, each housing unit and person is assigned a final replicate weight. The final weights are then integerized using the same controlled rounding routine used for the full sample.

A variance estimate for a given characteristic of interest is computed by the formula:

$$\text{var}(\hat{Y}) = \frac{4}{80}\sum_{r=1}^{80}(\hat{Y}_r - \hat{Y}_0)^2 \text{ where:}$$

$\hat{Y}_r$ - the rth replicate estimate for a characteristic.

$\hat{Y}_0$ - full sample estimate for the characteristic.

4.0 Method of Presenting Sampling Errors

From the beginning a decision was made to make every possible effort to encourage and to promote among data users the use of measures of uncertainty in their data analysis. To achieve this objective we decided that the method of presenting sampling errors must meet two criteria. The method must be **easy to understand and easy to use.** Since the goal of the ACS is to replace the decennial census long form in 2010 it was decided to use a method similar to the one used by decennial. This decision will provide for a smooth transition for most decennial data users.

4.1 Generalized Variance Functions

The sampling errors are presented in the form of factors or design effects named by Kish (1965). The final 1996 ACS design factors are available on the ACS internet site in the Methodology & Documentation section. This method meets the two pre-specified criteria. The design effects approach is easy to understand and most decennial data users are already familiar with it. The design effect is the ratio of the variance estimate of a complex sampling design to the variance estimate resulting from a simple random sampling design. This presentation of sampling errors has been employed in previous censuses (Waksberg et al, 1973) and multi-stage stratified sampling designs (Kalton et al, 1973).

An enormous amount of data based on the ACS sample has been and will be made available to the public through the Census Bureau Data Access and Dissemination System (DADS) on the Internet and in CD-ROMs. It is desirable to reduce the number of design effects or factors provided to the users to a manageable number, say less than 50. To accomplish this goal, data items were arranged into groups in which design effects were thought to be similar. These groups were broad subjects such as: labor force items, occupation items, school enrollment items, etc.

The square root of the design effect is less affected by extreme values and it is therefore preferred when an average effect is required (Kish, pp. 578-579, 1965) Thus, the square root of the design effect rather than the design effect is made available to users and it is referred to as the design effect, hereafter. For each data item (cell estimate) the design effects were averaged over all census tracts. The design effects were then (weighted) averaged over the data items in the group for each demonstration site. This factor minimizes the weighted square error loss function. These group design factors are used in determining the standard error for all data items in the data item group and for all tabulation areas in the demonstration site.

4.2 Data Item Groupings

The initial groupings are given in Table 2, and the final groupings can be gotten from the initial and the changes which follow. For each of the groups there is:

- Subject - subject matter for that group
- group # - number assigned for internal use
- N - tells you what value of N to be used in the denominator of the simple random sampling(SRS) standard error formula
- Table - list of STF tables in this group (can be found on the ACS internet site in the Data Products section)
- # of items - number of cells in all of the STF tables

The groupings define tables that deal with similar types of subject matter. Notice that two tables can be used in more than one group. If a table could fit into more than one subject area, then we recommend using the largest design factor for each of the areas that the table fits. The initial groupings are what we believed were items that would have about the same design factors. After getting the data and computing the design factors, some of the groups needed to be changed for reasons discussed below.

Groups 40 and 41 were changed based on problems with group 41. Group 41 had two items: the number of vacant and the number of occupied housing units. The standard error for the vacant housing units was larger than what would be expected under SRS due to the fact that the majority of the vacant units showed up in the CAPI universe and got large weights.

Old:

Tenure	40	Households	H3,H6, H7, H21	52
Occupancy Status	41	Households	*H4*	2

New:

Occupied by tenure	40	Households	H3,H6, H7, H21	52
Vacant	41	Households	*H5*	*6*

We also had problems with groups where there only a few items in the STF table and most of the respondents fell into just one or two of these items, thus making this item(s) with the majority of the respondents have a smaller design factor then the other item(s) in this STF table. The approach that we used in these situations was to just keep the item or two from the table that was the largest and not use the other table item(s) in the calculation of design factors. The reason for this was that most of these smaller estimates were handled by the small estimate cutoff explained in section 6.1. This happened with initial groups 45-47, 55. We also found that group 55 made up of tables H32 and H33 had different design factors for the two tables and thus this group was also split up. It also was determined that tables H24, H38, and H39 had similar design factors, so these tables were combined into one group.

Old (Changes appear in *italic.*):

Kitchen Facilities	45	Households	*H39*	*2*
Source of Water and Plumbing Facilities	46	Households	*H24, H38*	*6*
Sewage Disposal	47	Households	H25	*3*
Air Conditioning and Heating System	55	Households	*H32, H33*	*5*

New:

Kitchen Facilities, Source of Water, and Plumbing Facilities	45	Households	*H39, H24, H38* *(First items in each table)*	*3*
Heating system	46	Households	*H33* *(First item in the table)*	*1*
Sewage Disposal	47	Households	H25 *(First item in the table)*	*1*
Air Conditioning	55	Households	*H32* *(First two items of the table)*	2

5.0 Design Factors Evaluation

How good are the predicted standard errors? Using the design factor approach the user will not get the replicate standard error value. Rather, the user gets an approximation of the standard error. Once the group design factors are determined we need to assess how good of a job they do at predicting the standard errors. To do this, we will use the group design factors to approximate the standard errors. Compare the predicted standard errors to the direct standard errors via the relative absolute difference function (RAD) (1). To get the relative absolute difference, take the absolute difference between the predicted and the actual standard error divided by the actual standard error. This will give us a statistic for each site, tract, group and item combination, but we would like a summary statistic for each site and group combination only. So we are going to get two weighted averages based on that item's estimate to get what we call the WRAD(g) statistic. First we will take a weighted average of the RADs across all tracts for each combination of site, group and item, and we will denote this as WRAD (2). This makes sure that if an item is particularly large in a few tracts that it contributes a lot to the WRAD(g), but that a tract with a small estimate (compared to the other tracts) and a large RAD will not contribute too much to the WRAD(g). So once we compute the WRAD we will then take the same approach as computing it, but this time we will take the weighted average across all items within each site and group, to produce our final comparative statistic WRAD(g) (3).

We set a cutoff for the WRAD(g) of .40, and any values that fell above this cutoff point were checked out to make sure that the predicted standard errors were not too far off from the actual standard errors. If there were any values where this was the case, then all the items were looked at to see if the group definition could be changed. In all cases that we saw either the group was changed or the predicted standard errors were not off that much from the actual standard errors. In those that the group was not changed, it was just one or two tracts or items that was

making the WRAD(g) so high, and in most cases the predicted standard error was overestimating the actual standard error, and we feel that being a little conservative is a good approach. Over half of the WRAD(g)'s were less than .15, and less than 7% were greater than .3.

We also used scatter plots of the SRS standard error vs. the replicate standard error for all combinations of site and group. These helped in determining outliers and changes to the groupings. Examples for Brevard for initial and final group 41 are given in Plots1 & 2.

6.0 Special Cases

There are three cases for which the generalized variance estimates are not appropriate or not available because it is not feasible to produce a generalized variance estimate.

- The estimate of the number or proportion of people, households, housing units, or families in a geographic area with a specific characteristic is zero or very small.
- The estimate for "Total number of children ever born."
- Estimates of median for a given tabulation area or population group, for example median family income for black householders.

6.1 Small or Zero Estimates

The srs standard errors of zero estimates or very small estimates of totals and percentages approach zero as the size of the estimate gets smaller. The same is true for very large estimates of percentages and totals that are close to the size of the tabulation area. In other words when the estimated proportion is close to 1, the srs standard error of the estimate is close to zero. Obviously, a sample estimate of zero is subject to both sampling and nonsampling errors. Users should not conclude that there is complete or very high certainty the population value is zero, very small or very large. An observed zero sample estimate is possible even though there are units with the characteristic of interest in the population, so that the actual variance is not zero. We used a Poisson approximation to model the variance and compared that to the simple random sampling variance (Tersine, 1998). Table 1 provides the variance under the model and the srs variance for the estimate $\hat{Y}$. It also gives the estimate. The N used in the formulas was based on the average tract in 1990 having about 4,000 population. Assuming a 1-in-7 sampling fraction $\hat{p} = x / 571$.

Table 1. Estimated and Actual Variances

Value of		Weighted ACS Est.	Variance Estimate	Variance of x	Variance of $\hat{Y}$
x	$\hat{p}$	$\hat{Y}$	$6\hat{Y}(1-\frac{\hat{Y}}{N})$	λ_1	$49\lambda_1$
0	0	0	0	0.82	40.30
1	.002	7	41.93	1.90	93.10
2	.004	14	83.71	2.81	137.56
3	.005	21	125.34	3.65	179.02
4	.007	28	166.82	4.46	218.64
5	.009	35	208.16	5.25	257.36
6	.011	42	249.35	6.02	295.04
7	.012	49	290.24	6.78	332.28
8	.014	56	331.30	7.53	369.06
9	.016	63	372.05	8.27	405.11
10	.018	70	412.65	9.00	441.15
11	.019	77	453.11	9.73	476.74
12	.021	84	493.42	10.45	512.04
13	.023	91	533.57	11.17	547.19
14	.025	98	573.59	11.88	582.19

Inspecting the values in the 4th and 6th columns of Table 1, we concluded that for x=11 or a sample estimate of 77, the actual and estimated variances are getting very close. Basically the standard errors are identical. This lead to the recommendation of 75 as the point at which the srs standard error and the variance under the model would be approximately the same. A standard error of 21 for smaller estimates is recommended.

An alternative procedure would be to suggest standard error bounds for several levels of small or large estimated totals and percentages. However the benefits of this option will be out-weighted by the additional undue complexity for data users.

6.2 Total Number of Children Ever Born

The usual standard error formulae for estimates of proportions or totals can not be used to get standard error estimates for the aggregate number of children ever born. This item is an aggregate and therefore the use of standard formulas for Bernoulli characteristics are not appropriate. The elements or units being tabulated, number of children ever born, are not necessarily members of the tabulation area's universe. A description of the methodology used is included in the technical documentation prepared for the ACS data products.

6.3 Reliability of Sample Medians

To obtain reliability measures for estimates of medians an approximate method is suggested. The first step is to determine the size of the universe on which the median is based. Let N be the size of the universe and compute N/2. The second step is to treat N/2 as an ordinary estimate and use the srs standard error formula for a total to approximate the standard error. The next step is to compute a given confidence interval (i.e., 90 or 95 percent confidence interval) around N/2. The final step is to use linear interpolation to approximate the lower and upper limits of the confidence interval.

7.0 References

ACS Internet site, http://www.census.gov/CMS/www/.

Alexander, C., Dahl, S. and Weidman, L. (1997), "Making Estimates From the American Community Survey," *Proceedings of the Section on Government Statistics and Section on Social Statistics*, American Statistical Association, Washington, DC, pp. 88-97.

Kalton, G. and Blunden, R. M. (1973) "Sampling Errors in the British General Household Survey" *Proceedings of the 39th session International Statistical Institute*, Vienna, August 1973, pp. 83-87.

Kim, J. and Thompson, J. (1980), "Standard Error for Small or Zero Sample Estimates," Internal Census Bureau Report, September 25, 1980.

Kish, L. (1965), *Survey Sampling*, New York: John Wiley & Sons, Inc.

Tersine, A. (1998), "Standard Errors for Small or Zero Estimates for 1996 ACS Data Products," Internal Census Bureau Report, August 3, 1998.

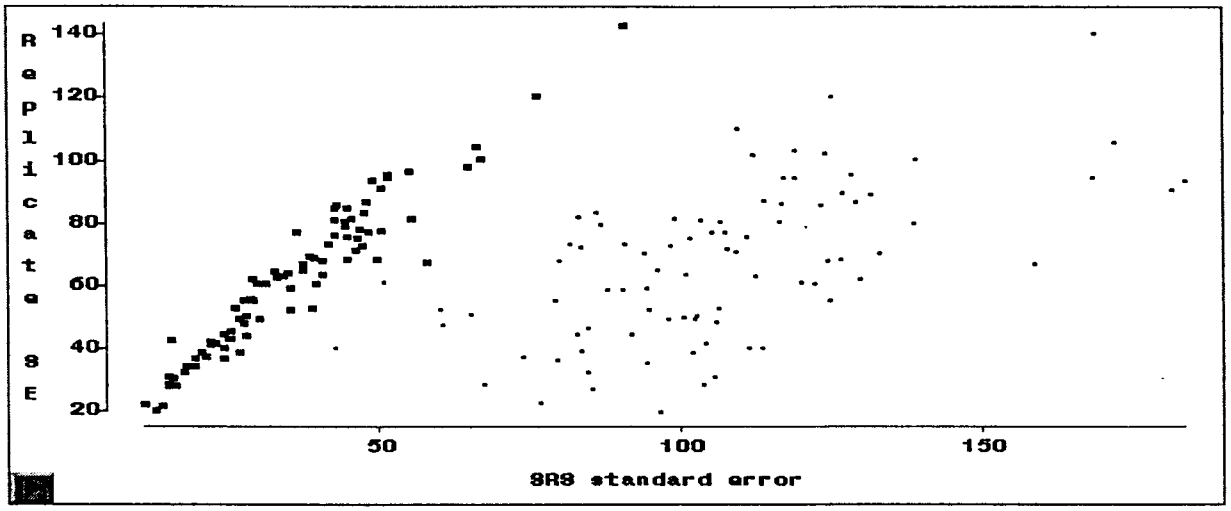
Waksberg, J., Hanson, R., and Bounpane, P. (1973), "Estimation and Presentation of Sampling Errors for Sample Data from the 1970 U.S. Census", *Proceedings of the 39th session International Statistical Institute*, Vienna, August 1973, pp. 67-82.

Table 2. Initial Data Item Groupings for the 1996 ACS Generalized Variance Estimates

Subject	Group #	N	Table	# of Items
POPULATION				
Age	1	Persons	P9	31
Sex	2	Persons	P5, P10(collapsed)	12
Race	3	Persons	P6, P8	15
Hispanic Origin	4	Persons	P7, P8	15
Marital Status	5	Persons	P20, P27	20
Ancestry	6	Persons	P24, P25	72
Household Size	7	Households	P12	7
Household Type and Relationship	8	Persons	P14, P19	20
Children Ever Born	9	Female 15 years old and older	P28	8
Work Disability and Functional Limitation	10	Persons	P46, P47	32
Place of Birth	11	Persons	P29, P30	20
Residence in 1985	12	Persons	P31, P32	22
Year of Entry	13	Persons	P26	5
Language Spoken at Home and English Ability	14	Persons	P21	10
Educational Attainment	15	Persons	P42, P43	20
School Enrollment	16	Persons	P41 P43(1-6)	11
Family Type	17	Families	P86 (collapsed)	12
Employment Status	18	Persons	P50, P51	20
Industry	19	Persons	P54	13
Occupation	20	Persons	P55	13
Class of Worker	21	Persons	P56	7
Hours Per Week and Weeks Worked in 1989	22	Persons	P53	10
Number of Workers in Family	23	Families	P77	4
Place of Work	24	Persons	P33, P34, P35	14
Means of Transportation to Work	25	Persons	P36	9
Travel Time to Work	26	Persons	P37	13
Private Vehicle Occupancy	27	Persons	P40	8
Time Leaving Home to Go to Work	28	Persons	P39	15
Type of Income in the Past 12 Months	29	Households	P61 - P67	14
Household Income in the Past 12 Months	30	Households	P57	25
Family Income in the Past 12 Months	31	Families	P79	25
Poverty Status in the Past 12 Months (persons)	32	Persons	P83, P85	66
Poverty Status in the Past 12 Months (families)	33	Families	P86	24
Armed Forces and Veteran Status	34	Persons	P44, P45	18
HOUSING				
Age of Householder	35	Households	H8	14

Race of Householder	36	Households	H7	20
Hispanic Origin of Householder	37	Households	H7	10
			(6-10, 16-20)	
Condominium Status	38	Households	H23	2
Units in Structure	39	Households	H20, H21	30
Tenure	40	Households	H3, H6, H7, H21	52
Occupancy Status	41	Households	H4	2
Gross Rent	42	Households	H42	18
Year Structure Built	43	Households	H26, H27, H29	19
Rooms, Bedrooms	44	Households	H9, H34, H35	27
Kitchen Facilities	45	Households	H39	2
Source of Water and Plumbing Facilities	46	Households	H24, H38	6
Sewage Disposal	47	Households	H25	3
House Heating Fuel	48	Households	H31	9
Telephone	49	Households	H36	4
Vehicles Available	50	Households	H37	12
Length at Residence	51	Households	H30	14
Mortgage Status and Selected Monthly Owner Costs(SMOC)	52	Households	H47, H49	41
Gross Rent as a Percent of Household Income	53	Households	H45	10
SMOC as a Percent of Household Income	54	Households	H49	20
Air Conditioning and Heating System	55	Households	H32, H33	5
Value	56	Households	H16	20

Plot 1: Initial Group 41 (Occupancy Status) for Brevard County, FL (Vacants are the darker boxes on the left)



Plot 2: Final Group 41 (Vacants) for Brevard County, FL