Analyzing Generalized Variances For the American Community Survey 2005 Public Use Microdata Sample[1]

Sirius Fuller, Anthony Tersine, Jr.
U.S. Census Bureau, 4600 Suitland Road, Washington, D.C. 20233

Abstract
The American Community Survey (ACS) releases data every year including a Public Use Microdata Sample (PUMS) File. PUMS allows data users to tabulate their own characteristic estimates by publishing responses from individual households and persons. All identifying information is removed from PUMS to ensure confidentiality. There are two methods available for users to calculate standard errors: generalized variances (design factors) and replicate weights (direct standard errors). Standard errors calculated using replicate weights are more accurate than those using design factors. Using that as a benchmark, this paper investigates whether the design factors for counts, means and medians are an acceptable substitute for their direct standard errors.

Key Words: Generalized Variance Formula, Replicate Weights, Public Use Microdata Sample, PUMS

1. Introduction

The ACS, a rolling monthly survey which collects data historically collected by the census long form, surveys roughly 3 million housing addresses annually. Data is released yearly for areas with a population greater than or equal to 65,000. Three-year estimates are released for areas greater than or equal to 20,000 and five year estimates are released for all areas.

PUMS is a publicly released subsample of ACS data. All sensitive data is removed to ensure confidentiality. Data are released for national, state and Public Use Microdata Areas (PUMA). PUMA are locally defined areas which contain roughly 100,000 people. This data is released to enable data users to calculate their own tabulations.

From 2000 until 2004, the ACS produced Generalized Variance Functions (GVFs) for PUMS to give users the ability to calculate standard errors (SEs). The GVFs for means and medians used a method called the a & b parameter method. The count data used design factors (DF), although there were a few exceptions. The GVFs were based on the Census 2000 Supplementary Survey data.[2]

Beginning in 2005, with the full implementation of the ACS, the a & b parameter method was dropped for means and medians and was replaced by a DF based method. This was to be consistent with Census 2000 and because some estimates in later years which used the a & b parameter method had negative variances. Replicate weights (RW) were also released so that data users could calculate SEs directly. The DFs were calculated based on fitting a simple linear regression (with no intercept) of the SEs calculated for a simple random sampling design to the direct SEs calculated using RWs. An iterative method was then used to remove outliers and refit the DFs. The DFs were calculated using data from the national, state and public use microdata areas (PUMA).

The DF standard errors (SE_{df}) are approximations, and will therefore differ from the RW standard errors (SE_{rw}). Results from previous research has shown that for counts the SE_{df} is a reliable approximation to the direct SEs. However, for means and medians, the SE_{df} were assessed based on a preliminary evaluation designed to see if they represent a good approximation to the SE_{rw}. The results were mixed (see Tersine and Navarro (1998)). The primary research question of this study is: Are the SE_{df} for counts, means and medians acceptable approximations to the SE_{rw}?

[1] This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed on statistical and methodological issues are those of the authors and not necessarily those of the U.S. Census Bureau.
[2] Puerto Rico was not included. Design Factors for Puerto Rico were released in 2005.

2. Methodology

2.1 Variable Selection and Creation of the Dataset

Data from the 2005 ACS PUMS housing and person files were used to address the research question. For this study eighteen variables were chosen, eleven from the housing files and seven from the person files. The variables were selected based on the ability to compute means and medians as well as totals. The list of variables used in the study is given in Table 1.

Table 1: Variables Selected for Study

Housing File	# of Cate-gories	Person File	# of Cate-gories
Family income	16	Age	23
Household income	16	Travel time to work	9
Gross rent	21	Total person's earnings	20
Monthly rent	21	Total person's income	20
Mortgage payment (monthly amount)	15	Wages or salary income	20
Selected monthly owner costs	-	Usual hours worked per week	3
With a mortgage	15	Weeks worked	6
Without a mortgage	11		
Number of bedrooms (categorical)	6		
Number of rooms (categorical)	10		
Categorized yearly real estate taxes	68		
Property value (categorical)	24		
Total Number of Categories:	223		101

Data from the 2005 ACS PUMS

The variables were restricted to match the ACS definitions with three exceptions. 'Total person's earning' was subset so that people who reported a loss for the year were out of scope. In addition, the variables 'number of weeks worked in the past 12 months' and 'usual hours worked per week in the past 12 months' were restricted to people between the ages of 16 and 64.

There were only four categorical variables (they are listed at the bottom of the housing variables). For analysis purposes, the rest of the variables were placed in categories based on the data published for the full ACS sample available on the American FactFinder website.

2.2 Data Metrics

Counts, means, medians and their SE_{rw} and SE_{df} were calculated as defined in the PUMS Accuracy of the Data (2005) documentation. The counts, or totals, were calculated using the weighted totals (for example, total number of people between the ages of 5 and 9). The means were the weighted averages for the characteristics. However, the medians were calculated differently. They were originally calculated as would be expected, that is the median represented the midpoint of the data. However, for some variables (especially income-related variables) the responses were rounded, skewing the median values. Therefore, the responses were placed in categories. The median was then calculated using the categorical data.

The SE_{rw} were calculated using the base weight and the 80 replicate weights, denoted as X_0 and X_r in the following equation: $SE_r(X) = \sqrt{\frac{4}{80}\sum_{r=1}^{80}(X_r - X_0)^2}$. The SE_{df} for counts was $SE(X) = DF\sqrt{99\hat{Y}(1-\hat{Y}/N)}$ where $\hat{Y}$ is the estimate and N is the geographic total. The SE_{df} for means was $SE(\bar{Y}) = DF\sqrt{\frac{99}{B}s^2}$ where B is the base, or denominator, of the mean and s^2 is the weighted sample variance. Finally, the median SE is a multi-step process which begins by calculating the 50 percent proportion and using that to calculate the SE using an upper and lower boundary: SE(median)=0.5(Upper

Bound – Lower Bound). The full process is denoted in the 2005 PUMS Accuracy document. There were several metrics calculated to analyze the data. They are discussed below.

2.2.1 The Difference of the Standard Errors, Over- and Under-Estimates

The difference is defined as the SE_{rw} subtracted from the SE_{df} (SE_{df} - SE_{rw}). If the difference is positive then the SE_{df} is an over-estimate; if the difference is negative, then the SE_{df} is an under-estimate. Dividing the number of over-estimates by the total number of differences (including differences of zero) multiplied by 100 gives the percent of differences which are over-estimates. A similar percentage can be computed for the under-estimates.

2.2.2 The Relative Differences and The Absolute Relative Difference

The relative difference (RD) is the difference divided by the SE_{rw} (relative difference = [SE_{df} - SE_{rw}]/SE_{rw}). A RD of 0.25 implies that the SE_{df} is 25 percent larger than the SE_{rw}. While a RD of -0.25 implies that the SE_{df} is 1 - 0.25 = 0.75 or 75 percent of the SE_{rw}. The absolute relative difference (ARD) is the absolute value of the RD. The RD gives a percent difference as well as direction, the ARD gives only the magnitude of the metric.

2.2.3 The Coefficient of Variation

The coefficient of variation (CV) is defined as the SE divided by the estimate. There is a CV for both the DF and RW methods. The CV may be viewed as a percentage (although it can exceed 100 percent). It is used to compare various estimates to one another since it is a unitless measurement. It is also used as a reliability measure. The CV assumes a relationship between the size of the characteristic and the SE. As the sample size increases, the SE decreases and the CV gets smaller. This relationship does not intrinsically exist for the median, therefore only the count and the mean data are included in this metric.

3. Results

3.1 Over- and Under-Estimation of the SE_{rw} by SE_{df}

Examining the data metrics, we will naturally begin by examining the over- and under-estimates between the SEs. The percentage of over-estimates (that is, where SE_{df} is greater than SE_{rw}) at the state level is similar to the percentage of over-estimates at the PUMA level for the count, mean and median data. This can be seen below in Table 2. At the state level, roughly 88 percent of the count data are over-estimates. All of the mean data are over-estimates, whereas only about 51 percent of the median data are over-estimates.

The total number of estimates shown in Table 2 for counts at the state level is found by multiplying the number of states (51, including Washington, D.C.) by the total number of categories for all of the variables (324) which yields a total of 16,524. For the PUMA level, the number of PUMAs (2,071) is substituted for the number of states, yielding 671,004. Not every category had data associated with it. For example, there were several PUMAs which did not have any data for personal income greater than $100,000 (the highest category). Therefore, the actual totals for counts at the state and PUMA levels are lower than these numbers.

Table 2: Over-Estimates of SE_{rw} by using SE_{df} for each statistic

Statistic	State		PUMA	
	Over-Estimates	Total	Over-Estimates	Total
Count	14,390 87.9	16,380 100.0	518,762 86.2	601,786 100.0
Mean	918 100.0	918 100.0	37,211 99.8	37,278 100.0
Median	248 51.0	516 100.0	15,502 58.4	27,368 100.0

Data from the 2005 ACS PUMS

For means and medians, there was only one mean and median for each state and one at each PUMA level. Therefore, at the state level there were 51 times 18 or 918 mean and median estimates, and 2,071 times 18 or 37,278 estimates at the PUMA level.

It is worth noting that at the state level there were 299 cases where the SE_{rw} for the median equaled zero. This was because all of the replicate medians calculated using the replicate weights were the same. Since a SE cannot be zero for a sample, these standard errors were not included in Table 2. In addition, the SE_{rw} at the PUMA level was equal to zero in 5,538 cases. For two cases at the state level and 338 cases at the PUMA level the median was in the highest categorical bin, rendering the formula for the SE_{df} to be undefined. In addition there were 101 cases at the state level and 4,034 at the PUMA level where both SE_{rw} equaled zero and SE_{df} was undefined. Therefore 10,312 cases (402 for state and 9,910 for PUMA) were not included in Table 2. This represents roughly 40 percent at the state level and roughly a quarter of the total PUMA level median cases. Before continuing with the analysis, a closer examination of these cases is carried out in the next section.

3.2 Brief Analysis of When the Median SE_{rw} Equals Zero

A list of where the SE_{rw} for the median equaled zero, broken down by variable for state and PUMA data, is given in Table 3. The variables with the highest percentages were: number of bedrooms, number of rooms, property value, travel time to work, usual hours worked per week (past 12 months) and weeks worked (past 12 months). The first three variables are categorical in the PUMS file.

An attempt was made to recalculate the SE_{rw} for categorical variables using an alternate method which is spelled out in Tersine (2000). For each variable the median was calculated categorically. That is, the data are put into categories and interpolation (either linear or Pareto) is used to calculate the median value. After the median was calculated, the medians using replicate weights were calculated and the standard RW formula was used to find SE_{rw}.

Table 3: Number of Median SE_{rw} Equal to Zero by Variable for State and PUMA

Variable	State		PUMA	
	Count	Percent	Count	Percent
Number of Bedrooms	50	12.5	1,804	18.8
Travel Time to work	45	11.3	1,009	10.5
Number of Rooms	43	10.8	1,501	15.7
Categorized yearly real estate taxes	26	6.5	126	1.3
Property Value	43	10.8	1,036	10.8
Usual hours worked per week	51	12.8	2,065	21.6
Weeks worked	50	12.5	1,964	20.5
All other variables	92	23.0	67	0.7
TOTAL:	400	100.0	9,572	100.0

Data from the 2005 ACS PUMS

Using this categorical method, over- and under-estimates for the categorical variables were calculated for states and PUMAs. The results are shown in Table 4. The data from Table 2 and the distribution in Table 4 for the categorical variables are not similar. Thus, the categorical data were not included in the dataset and were excluded from further analysis.

Table 4: Over- and Under-Estimates for Data Where Median SE_{rw} Equaled Zero

Variable	State			PUMA		
	Over-Estimates	Under-Estimates	Total	Over-Estimates	Under-Estimates	Total
Number of Bedrooms	8.3	91.7	100.0	47.4	52.6	100.0
Number of Rooms	23.8	76.2	100.0	72.1	27.9	100.0
Categorized yearly real estate taxes	65.4	34.6	100.0	95.2	4.8	100.0
Property Value	0.0	100.0	100.0	4.9	95.1	100.0

Data from the 2005 ACS PUMS

3.3 Relative and Absolute Relative Differences

Returning to the main analysis, Table 5 shows the ARD for counts, means and medians below various levels. Past Census research, including Tersine and Navarro (1998) and Tersine (1999 a and b), used 0.3 or 0.4 as a benchmark to determine whether the SE_{df} was an acceptable approximation to the SE_{rw}. Most of the count data at the state level and more than half of the PUMA data fall below the 0.4 level. Similarly, more than 50 percent of the state and PUMA data for the median has an ARD below 0.4. However, the mean data for both state and PUMA has the majority of its data above an ARD of 0.4, which indicates a potential problem.

Table 5: Percent of Counts, Means and Medians Below Various ARD Ranges

ARD Range	Count		Mean		Median	
	State	PUMA	State	PUMA	State	PUMA
Below 0.1	20.4	14.0	0.0	0.3	15.5	19.5
Below 0.2	42.1	28.5	0.0	0.8	32.2	37.1
Below 0.3	63.5	43.3	0.0	1.5	45.5	51.8
Below 0.4	79.7	57.2	0.2	2.5	57.4	64.4

Data from the 2005 ACS PUMS

To see the data in greater detail, the RDs for counts and means by state and PUMA are displayed in Table 6. The distribution of RDs are displayed in each category. The categories are asymmetrical. The categories below an RD of -0.4 have been combined into one category. The categories with an RD between -0.4 and 0.4 (i.e. an ARD of between 0.0 and 0.4) were combined into a second category.

The count data at the state level has some over-estimation. Outside the -0.4 to 0.4 range the data falls mostly between 0.4 and 1.0. For PUMA there is clearer evidence of over-estimation. The means demonstrate clear over-estimation with the majority of the data for both state and PUMA being above 1.0.

Table 6: Ranges for Relative Differences by State and PUMA for Counts and Means

Relative Difference	State		PUMA		Type of Estimate
	Counts	Means	Counts	Means	
Less than or equal to -0.4	0.3	0.0	1.0	0.0	Under-estimate
Greater than -0.4 to 0.4	79.7	0.2	57.2	2.5	Acceptable
Greater than 0.4 to 1.0	19.9	22.7	38.3	17.7	Over-estimate
Greater than 1.0 to 1.5	0.1	23.3	2.9	20.0	
Greater than 1.5 to 2.0	0.0	19.5	0.4	13.7	
Greater than 2.0 to high	0.0	34.3	0.3	46.1	
TOTAL	100.0	100.0	100.0	100.0	-

Data from the 2005 ACS PUMS

The RD for medians are in Table 7. The ranges are slightly different than the ones displayed in Table 6. This is because the median data shows different behavior with a roughly equal distribution both below an RD of -0.4 and above 0.4. However, there are more under-estimates at the state level and more over-estimates at the PUMA level.

Table 7: Ranges for Relative Differences by State and PUMA for Medians

Relative Difference	Medians		Type of Estimate
	State	PUMA	
Less than or equal to -0.8	5.2	1.9	Under-estimate
Greater than -0.8 to -0.6	8.9	3.7	
Greater than -0.6 to -0.4	11.6	6.8	
Greater than -0.4 to 0.4	57.4	64.4	Acceptable
Greater than 0.4 to 0.6	4.8	9.0	Over-estimate
Greater than 0.6 to 0.8	4.7	5.3	
Greater than 0.8 to high	7.5	8.9	
Total	100.0	100.0	-

Data from the 2005 ACS PUMS

3.4 Adjustments

To correct for the over-estimations in the count data, two possibilities were examined. The first was the finite population correction factor, while the second was the creation of DFs. Park and Lee (2004) found that means and count data have different DFs. The count DFs tend to be much larger than the mean DFs. The problem posed by the mean data is therefore not wholly unexpected.

With regards to the finite population correction factor, the sampling fraction used in the standard error calculation was the designated sampling rate for the full ACS of 2.5 percent. However, the actual sampling rate was closer to 2.4. To correct for this, the SE_{df} was multiplied by a correction factor of approximately 0.98 or the square root of [(97.5/2.5)*(2.4/97.6)]. The 97.5/2.5 portion undoes the original finite population correction factor, while the 2.4/97.6 applies the new correction factor. Applying the correction factor to the data did not have an effect on the results and the percentages within tables 6 and 7 did not change.

Concerning the second possibility, recall that the DFs were calculated based on fitting a simple linear regression of the SEs calculated for a simple random sampling design to the actual (direct) SEs for each variable. An iterative method was then used to remove outliers and refit the DFs. There was a maximum of five iterations.

For the 2005 data, when the DF was calculated, there were five linear regression models to choose from:

1. DF using national level data
2. DF using state level data
3. DF using state and PUMA data calculated for each state
4. DF using national, state and PUMA data
5. DF using the original 2000 DF.

The DF with the lowest average ARD fitted to the full ACS 2005 data was chosen as the final publicly released DF. The DF chosen could vary from variable to variable.

For this paper, the five possible DFs were rounded to one decimal place to mimic the publicly released DFs. Five SE_{df} were then recalculated using the five possible DFs. The ARD for each was recalculated and the minimum was chosen as 'best'. The DF considered 'best' often did not change form the publicly released DF. These alternate DFs were calculated for each geographic level and for each estimate within each characteristic. This is an artificial method for calculating alternate DFs. A more natural method would be to recalculate the DFs for each geographic level and characteristic. The former method was chosen to maximize the benefit derived from using alternate DFs. The results are discussed in the following sections

3.4.1 Count Adjustment

The results for counts can be seen in Table 8. The use of alternate DFs moved roughly three percent of the data within the acceptable range of a RD between –0.4 and 0.4.

Table 8: Count Data for States and PUMAs, Original Relative Differences and Relative Differences Using Alternate Design Factors

Relative Differences	Original Counts		Alternate DF Counts		Type of Estimate
	State	PUMA	State	PUMA	
Less than or equal to -0.4	0.3	1.0	0.2	0.5	Under-estimate
Greater than -0.4 to 0.4	79.7	57.2	83.2	60.5	Acceptable
Greater than 0.4 to 1.0	19.9	38.3	16.6	36.0	Over-estimate
Greater than 1.0 to 1.5	0.1	2.9	0.1	2.5	
Greater than 1.5 to 2.0	0.0	0.4	0.0	0.4	
Greater than 2.0 to high	0.0	0.3	0.0	0.2	
Total	100.0	100.0	100.0	100.0	-

Data from the 2005 ACS PUMS

3.4.2 Mean Adjustment

As discussed in section 3.3, the means clearly demonstrated over-estimation for both the state and PUMA data. The results of applying alternate DFs are displayed in Table 9 below. This approach improved the data slightly, shifting roughly two percent of the state data and less than one percent of the PUMA data into the acceptable range.

Table 9: Mean Data for States and PUMAs, Original Relative Differences and Relative Differences Using Alternate Design Factors

Relative Differences	Original Means		Alternate DF Means		Type of Estimate
	State	PUMA	State	PUMA	
Less than or equal to -0.4	0.0	0.0	0.0	0.0	Under-estimate
Greater than -0.4 to 0.4	0.2	2.5	2.1	3.3	Acceptable
Greater than 0.4 to 1.0	22.7	17.7	22.7	17.8	Over-estimate
Greater than 1.0 to 1.5	23.3	20.0	23.5	19.6	
Greater than 1.5 to 2.0	19.5	13.7	19.8	14.2	
Greater than 2.0 to high	34.3	46.1	31.9	45.1	
Total	100.0	100.0	100.0	100.0	-

Data from the 2005 ACS PUMS

3.4.3 Median Adjustment

The medians have both over- and under-estimation. Even after alternate DFs are employed, there are still more under-estimates for the state level and more over-estimates for the PUMA level.

Table 10: Median Data for States and PUMAs, Original Relative Differences and Relative Differences Using Alternate Design Factors

Relative Differences	Original Medians		Alternate DF Medians		Type of Estimate
	State	PUMA	State	PUMA	
Less than or equal to -0.8	5.2	1.9	4.1	1.5	Under-estimate
Greater than -0.8 to -0.6	8.9	3.7	7.4	3.2	
Greater than -0.6 to -0.4	11.6	6.8	9.9	4.4	
Greater than -0.4 to 0.4	57.4	64.4	62.6	68.3	Acceptable
Greater than 0.4 to 0.6	4.8	9.0	4.7	8.8	Over-estimate
Greater than 0.6 to 0.8	4.7	5.3	4.5	5.1	
Greater than 0.8 to high	7.4	8.9	7.0	8.6	
Total	100.0	100.0	100.0	100.0	-

Data from the 2005 ACS PUMS

3.5 Coefficient of Variation

Finally, to conclude our analysis, we examine the CV. As previously mentioned, the CV assumes a relationship between the size of the estimate and the SE. Therefore, data for the median are excluded. The CVs can be seen as percentage. The acceptable limit for a CV varies depending on the context.

The RW and DF CVs for counts were roughly consistent for the state but not the PUMA levels. There were higher CVs at the PUMA level than at the state level. Roughly 15 percent of the CVs were above the 0.60 level for the PUMA data using RW, while roughly a quarter was above 0.60 for the CVs created using the DFs.

Table 11: Coefficient of Variation for Counts

Range of CV	State		PUMA	
	CV using SE_{rw}	CV using SE_{df}	CV using SE_{rw}	CV using SE_{df}
0.00 to 0.01	0.9	0.8	0.0	0.0
Greater than 0.01 to 0.02	7.0	4.3	0.0	0.0
Greater than 0.02 to 0.05	33.9	27.9	0.2	0.7
Greater than 0.05 to 0.10	29.4	30.7	2.7	1.4
Greater than 0.10 to 0.15	11.7	14.2	10.5	5.9
Greater than 0.15 to 0.30	11.5	13.8	43.8	35.2
Greater than 0.30 to 0.60	4.3	5.8	28.0	30.9
Greater than 0.60	1.3	2.5	14.8	25.9
TOTAL	100.0	100.0	100.0	100.0

Data from the 2005 ACS PUMS

Previously, when examining the relative differences, there was evidence of over-estimation for both the count and mean data. Although not a solution to the problem, CVs can offer perspective on this issue. Small CVs can mitigate the effect of over-estimation. For example, if the CV using the SE_{rw} is approximately one percent and the corresponding CV using the SE_{df} is three percent, the SE_{df} may still be a useful approximation to the SE_{rw} even though the RD is 2.0.

The data for the means are given in Table 12. The majority of the CVs created using the SE_{rw} at the state level are below two percent, while their corresponding CVs using the SE_{df} were below five percent. The CVs for PUMAs were higher, with the CVs using SE_{rw} mostly between two and ten percent and the CVs for the SE_{df} between five and 15 percent.

Table 12: Coefficient of Variation for Means

Range of CV	State		PUMA	
	CV using SE_{rw}	CV using SE_{df}	CV using SE_{rw}	CV using SE_{df}
0.00 to 0.01	73.9	19.5	2.3	0.0
Greater than 0.01 to 0.02	20.5	40.4	24.1	0.0
Greater than 0.02 to 0.05	5.7	36.4	56.3	8.8
Greater than 0.05 to 0.10	0.0	3.7	16.6	57.1
Greater than 0.10 to 0.15	0.0	0.0	0.6	24.0
Greater than 0.15	0.0	0.0	0.2	10.1
TOTAL	100.0	100.0	100.0	100.0

Data from the 2005 ACS PUMS

4. Conclusion

The goal of this paper was to examine whether the SE_{df} was an acceptable approximation for the SE_{rw}. At the state level, count SE_{df} are an acceptable approximation to the count SE_{rw}, with roughly 80 percent of the data lying within the acceptable range of a relative difference of –0.4 to 0.4. At the PUMA level there is some evidence of over-estimation, however about 60 percent of the data is acceptable. The mean SE_{df} over-estimates the SE_{rw} at both the state and PUMA level with less than one percent and three percent, respectively, falling in the acceptable range. However, while the CVs using the SE_{df} are larger than those using the SE_{rw}, both are relatively small. An over-estimate with a small CV is less catastrophic than if the CVs were larger. For medians, roughly 60 percent of both the state and PUMA level data are in the acceptable range. However, the median SE_{df} both over- and under-estimates the SE_{rw}. Under-estimates (especially large under-estimates) and large over-estimates make the SE_{df} an inadequate approximation to the SE_{rw}. Using alternate DFs to correct for over- and under-estimation shifts a small amount of the data into the acceptable range for all three statistics but fails to make a relevant impact. Thus, future research may explore new methods for producing DFs for the means and medians.

References:

1. U.S. Census Bureau, American FactFinder. http://factfinder.census.gov/home/saff/main.html?_lang=en

2. U.S. Census Bureau, 2005. "PUMS Accuracy Document (2005)." Washington, D.C., 2005 http://www.census.gov/acs/www/Downloads/2005/AccuracyPUMS.pdf.

3. Tersine, A. (2000), "Specification for Calculating Medians/Quartile and their Variances Based on Categorical Data", From Tupek A. to Bowie C., Internal Memorandum, U.S. Census Bureau

4. Tersine, A. and Navarro, A., Generalized Variance Estimates in the 1996 American Community Survey, *1998 Proceedings of the Section on Survey Research Methods*, American Statistical Association, St. Louis, MO., pp. 184-189

5. Tersine, A., (1999a), Generalized Design Factors for the 1996 ACS Data Products, Internal Census Documentation

6. Tersine, A., (1999b), Generalized Variance Research, Internal Census Documentation

7. Park, I., and Lee, H (2004), Design Effects for the Weighted Mean and Total Estimators Under Complex Survey Sampling. Survey Methodology 30, 183 - 193.