Sampling Variance Estimates for SSA Program Recipients From the 1990 Survey of Income and Program Participation

*by Barry V. Bye and Salvatore J. Gallicchio**

Since 1987 the Social Security Administration (SSA) has published a special set of tabulations on SSA program recipients in the *Annual Statistical Supplement* to the *Social Security Bulletin* using data derived from the Census Bureau's Survey of Income and Program Participation (SIPP). Estimates of sampling errors pertaining to these tabulations were derived from the 1984 SIPP panel. This article provides updated sampling error estimates for the 1990 SIPP panel to be used in conjunction with the SIPP-based tabulations provided in the *Annual Statistical Supplement* for 1992 and 1993. The computational approach is essentially the same as that used in the earlier analysis. Sampling variances are estimated by half-sample replication using the pseudo stratum and half-sample codes available on SIPP public use data files. Generalized tables of standard errors are provided for all SSA program participants. An appendix provides detailed specifications about the calculations. In order that it be self-contained, this article repeats much of the methodological exposition in the previous article that appeared in the October 1988 issue of the *Social Security Bulletin*.

*Barry Bye is with the Division of Statistical Operations and Services, and Salvatore Gallicchio is with the Division of Statistics Analysis, Office of Research and Statistics, Social Security Administration.

The Survey of Income and Program Participation (SIPP) provides data that can be used to study the socioeconomic characteristics of persons participating in programs administered by the Social Security Administration (SSA).[1] The most recent data published by SSA come from the wave 2 of the 1990 panel of the SIPP. The 1990 panel consists of approximately 20,000 households comprising about 54,000 individuals. About 8,500 of these individuals have identified themselves as Old-Age, Survivors, and Disability Insurance (OASDI) or Supplemental Security Income (SSI) program recipients. The latter includes about 900 respondents.

Summary statistics on SSA program participants based on 1990 SIPP data appear in a special set of tables in the *Annual Statistical Supplement* to the *Social Security Bulletin* for 1992 and 1993.[2] The tables pertain to the civilian noninstitutionalized population receiving OASDI or SSI payments. They focus on three major themes: the composition and level of income of persons receiving different types of OASDI benefits, the general characteristics of persons aged 18-64 receiving OASDI or SSI payments based on disability, and similar information about SSI recipients aged 18 or older. The unit of analysis in these tables is the individual recipient.

Many of the distributions and income levels shown in the *Supplement* tables are based on a relatively small number of sample cases. Summary statistics generated from small numbers of cases can be imprecise due to large sampling errors (variances) and often suggest differences between subpopulations when no real differences exist. It is important, therefore, that estimates of sampling errors be provided along with the population estimates.

The Bureau of the Census has provided generalized variance curves for a number of quantities from the 1990 SIPP panel.[3] These curves do not identify OASDI or SSI recipients separately; therefore, the curves do not pertain directly to program participants. Fortunately, provisions were made for the direct calculation of sampling variances of SIPP estimates using special codes available in the SIPP public use data files. The codes allocate the SIPP sample cases to a set of pseudo strata and pseudo primary sampling units. The codes permit direct estimates of sampling variances to be obtained by a number of methods.

The results of direct sampling variance computations for SSA program participants are presented in this article. The approach used to estimate the variances was the method of balanced half-sample replication, the same method that was used previously in connection with the 1984 SIPP Panel.[4,5] The appendix at the end of the article includes the detailed specifications for estimating sampling variances from the SIPP using the same techniques that were used for the computations in this article. The results of the calculations also are provided in sufficient detail to be used as a benchmark.

Sampling variances were computed for 148 population estimates, cross-classifying the recipients by sex, age, and marital status. A curve was fit to the estimated variances using the 126 cells with unweighted counts of 25 or more and was used to produce tables of generalized standard errors. The tables of generalized standard errors can be applied directly to the data presented in the *Supplement* for program participants aged 18 or older and also can be used with other analyses from the 1990 SIPP panel that pertain to SSA program participation of adults. A separate analysis for child beneficiaries under age 18 was not made because the analysis of the 1984 panel data cited above showed that estimated standard errors for this group were strongly associated with family size. As a result, tables of generalized standard errors that would be applicable to a variety of estimates for this subpopulation could not be developed.

The generalized variance curves presented in this article yield variance estimates that are markedly different from those generated by curves provided by the Census Bureau although the functional form of the curves is the same.[6] The differences appear to be due mainly to differences in curve fitting procedures employed by the two agencies and the differences in raw variance items used in the analyses. The SSA estimates are generally smaller than the Census estimates and appear to be more appropriate for OASDI and SSI program participants.

Sampling variances and covariances are also computed for a small set of mean and median income amounts to demonstrate how these calculations can be performed from the SIPP files. The resulting quantities can be used to test differences of means and medians among various subpopulations.

Methodology

Balanced Half-Sample Replications

The method of balanced half-sample replication is an approach to the estimation of sampling variances for complex sample designs that can be implemented easily and has been applied to a wide variety of statistical estimates. For the SIPP, this method presupposes that the primary sampling units for the population have been assigned to one of L strata, and two of the units are selected with replacement from each stratum. Half-sample replicates of this design can be formed by selecting at random one of the two units from each stratum. For a sample design with L strata, there are 2^L such half samples. If an estimate of the statistic of interest is made in each half sample and in the full sample, then the average squared difference between half-sample and full-sample estimates from any subset of half samples provides an estimate of the sampling variance of the statistic. The estimate of the sampling variance is most precise when all 2^L half samples are employed.

When L is large, one would like to use only a part of the 2^L half samples to estimate the sampling variances without loss of precision. It turns out that special sets of half samples, called balanced, orthogonal sets, are particularly good candidates. Estimates of sampling variances from these special sets are algebraically equivalent to those obtained using all half samples. Also, when the full-sample estimate is a linear function of the observed variables, the average estimate over the balanced, orthogonal set will be equal to the full sample estimate. The minimum number of half samples required for a fully balanced orthogonal set is the smallest multiple of 4 which is greater than the number of strata in the sample design. For designs with many strata, this number will be much smaller than the total number of possible half samples. Descriptions of balanced, orthogonal sets for many designs are provided in the literature.[7]

Once a set of half samples has been identified, estimated sampling variances are particularly easy to compute. Let $\theta_\alpha (\alpha = 1, ..., K)$ denote the estimator of the population parameter of interest computed from the αth half sample, and let θ be the corresponding estimate from the full sample. An estimator of the sampling variance of θ, $V(\theta)$ based on K half samples is given by

$$V(\theta) = \sum_{\alpha=1}^{K} (\theta_\alpha - \theta)^2 / K. \quad (1)$$

When θ is linear and L<K, then

$$\theta = \overline{\theta} = \sum_{\alpha=1}^{K} \theta_\alpha / K,$$

and (1) provides an unbiased estimate of the variance of θ. When θ is not linear (for example, θ is a ratio, a median, a correlation coefficient), then $\theta \neq \overline{\theta}$ and the expected value of $V(\theta)$ differs from the variance of θ by an amount often well approximated by $[E(\overline{\theta}-\theta)]^2$. Thus, if $\overline{\theta}$ is close to θ, equation (1) will provide a good approximation of the sampling variance when θ is not linear.[8]

Variance Curve

A two-parameter curve was fit to the variance estimates obtained by the replication method. The curve specified the relative variance (Rv), the variance divided by the square of the estimate, as a function of the estimate.

$$Rv(x) = a + b/x \qquad (2)$$

where

a and b are coefficients to be estimated,

x is the estimated population total, and

Rv(x) is the estimated relative variance of x--that is,

$$Rv(x) = V(x)/x^2.$$

This functional form has provided a fairly good representation of the relationship between Rv(x) and x in other surveys. Its use is motivated by the following considerations.[9]

The design effect (Deff) for a particular estimate, x, from a complex sample design is defined as the ratio of the sampling variance of x under the design to the sampling variance that would have been obtained from a simple random sample of equal size. For a sample of size n from a population of size N, the simple random sampling variance of an estimated total, x is given by

$$var(x) = var(pN) = N^2PQ/n$$

where

P = X/N, is the true population proportion,

X is the population total estimated by x,

Q = 1-P, and

p is the sample estimate of P.

The variance of x from a complex design of the same size can be expressed as

$$\begin{aligned} var_c(x) &= Deff\ (var(x)) \\ &= Deff\ (N^2PQ/n). \end{aligned}$$

The relative variance of x is given by

$$\begin{aligned} Rv(x) &= var_c(x)/X^2 = Deff\ (Q/Pn) \\ &= -Deff/n + (N/n)Deff/X. \end{aligned} \qquad (3)$$

Equation (3) has the same form as equation (2) where a = -Deff/n and b = (N/n)Deff. If it is reasonable to assume that a constant design effect exists for a particular set of estimates, then the estimated relative variances for those items may be accurately represented by a two-term curve of the form in (2) from which generalized variances can be computed.

The method used to estimate the coefficients in (2) was an iterative procedure that minimized the function

$$\sum_{i=1}^{I} \left[\frac{Rv_i - \hat{R}v_i}{\hat{R}v_i^*} \right]^2$$

where

Rv_i is the computed relative variance for the ith item;

$\hat{R}v_i$ is the estimated relative variance from the curve for the ith item.

Rv_i^* is a weight for the ith item. It is set equal to the computed relative variance, Rv_i, in the first iteration; for all subsequent iterations it is set equal to the estimated relative variance, $\hat{R}v_i$, from the previous iteration.

I is the number of items to be fit.

This estimation approach gives greater weight to items with smaller estimated relative variances (and, thus, generally larger estimated totals) and has been found to work well in other surveys.

Generalized Variances for Counts and Proportions

Having estimated values for the coefficients in equation (2), the relative variance for a specific estimated total, x_o, can be obtained by substituting x_o, into that equation. The variance of the estimated total can be obtained by multiplying the relative variance by the square of the estimate.

$$\begin{aligned} V(x_o) &= Rv(x_o)x_o^2 \\ &= ax_o^2 + bx_o \end{aligned} \qquad (4)$$

Equation (4) can also be used to produce generalized estimates of variances of proportions. A proportion is the ratio of two estimated totals, p = x/y, where the cases counted in the numerator are a subset of the cases counted in the denominator. In large samples, the relative variance of this type of ratio can be approximated by the following formula:

$$Rv(p) = Rv(x/y) = Rv(x) - Rv(y)$$

or

$$V(p) = V(x/y) = (x/y)^2\ [Rv(x) - Rv(y)] \qquad (5)$$

Substitution of estimates from (2) into (5) provides generalized variance estimates for proportions.

$$\begin{aligned} V(p) &= p^2[b(1/x - 1/y)] \\ &= (b/y)\ (p)\ (1 - p) \end{aligned} \qquad (6)$$

Tables of generalized standard errors for estimated totals are often produced from equation (4) by computing and displaying the square root of the estimated variances for a set of predetermined values of x. Similarly, a table of standard errors for estimated proportions can be computed from (6). This table will be two dimensional with the size of the base of the percent on one dimension and the estimated proportion on the other.

Variances of Means and Medians

Balanced half-sample replication can also be used to estimate sampling variances for means and medians. The sampling variance is obtained by estimating the mean (median) is each half sample and then applying equation (1). This approach is demonstrated below with OASDI benefit payments. The mean benefit payments are computed in the usual way: the sum of the weighted benefit amount divided by the sum of the weights. The medians are estimated from distributions of benefit amounts using the following formula:

$$M = L_j + \left[\frac{S_{50} - S_j}{N_j} \right] W_j$$

where

- j indexes the interval containing the 50th percentile;
- L_j is the lower limit of the jth interval;
- S_{50} is the estimated population at the 50th percentile;
- S_j is the estimated population with values below the jth interval;
- N_j is the estimated population in the jth interval; and
- W_j is the width of the jth interval.

A distribution of equal intervals with width of $100 was used for the OASDI income distribution.

Covariance Matrix

One advantage of the half-sample replication approach to sampling variance estimation is that the computation of the full covariance matrix for a set of estimates is straightforward. Having a full covariance matrix permits the testing of simple and complex hypotheses among the members of the set. Generally, statistical tests require that the estimates have a multivariate normal distribution and that a consistent estimate of the covariance matrix is available.[10] Although in suitably large samples, the normality assumption is reasonable for the kinds of estimates described here, the consistency of the estimates of the covariance matrix based on pseudo strata and primary sampling units is problematic. Still, it is believed that test statistics based on these matrices provide some useful information about the relative sizes of the population estimates even if the significance levels are not known precisely.

The sampling covariance matrix is obtained through the balanced half-sample method by a computation similar to that of equation (1). If population estimates have been computed for some set of classifications, then the (i,j)th element of the covariance matrix for the set of estimates is given by

$$\sum_{\alpha=1}^{K} [M_\alpha^{(i)} - M^{(i)}][M_\alpha^{(j)} - M^{(j)}]/K$$

where

$M^{(r)}$ is the estimate of the statistic (for example, mean or median) for the rth population category,

$M_a^{(r)}$ is the estimate of the statistic for the rth category the αth half sample,

K is the number of half samples.

Results

Counts and Proportions

Appendix table I presents the population estimates, standard errors, and relative variances for 148 items cross-classifying the SSA recipient population by age, sex, and marital status. Of these estimates, 126 had unweighted cell counts of 25 or more, and were used to derive the parameters of the generalized variance curve. The estimated parameters are:

a = .00047

b = 5931.5.

Note that the estimated constant, a, is positive. Although the rationale for the two-parameter curve indicates that a should be negative, the algorithm used to estimate the parameters does not impose this constraint.

Table 1 provides standard errors for estimated population totals from the curve. Table 2 provides standard errors for estimated proportions from equation (6). Generalized curves were fit separately to OASDI and SSI subpopulations. Although there was some variation in a and b parameters--generally a small tradeoff between a and b, for example slightly larger a for slightly smaller b--the resulting lookup tables were very similar.[11]

Means

To demonstrate variance estimates for means, table 3 presents estimated standard errors for mean Social Security benefit amounts for persons receiving only OASDI benefits. The first four columns of the table give the unweighted sample count, the estimated population total, the estimated mean benefit amount and its standard deviation, based on weighted data. The next column gives an estimate of the standard error of the mean based on the half-sample replication method. The coefficients of variation (that is, the estimated mean divided by the standard deviation) range from a low of 0.6 percent for the overall estimate (one standard error of 3.3 on an estimated mean of $537) to almost 4 percent for the never married female estimate.

The next column of table 3 provides estimates of standard errors that would have been obtained from simple random samples of the same size as indicated by the unweighted sample counts and using the weighted standard deviations as estimates of the population standard deviation. The formula for the estimated standard error of a mean, M, in a simple random sample is

$$\text{StdErr}(M) = \hat{\sigma}/\sqrt{n},$$

where $\hat{\sigma}$ is the estimated population standard deviation, n is the unweighted sample size.

Estimated design effects (the square of the ratio of the replication standard error to the simple random standard error) range from a low of 0.88 for the single males to a high of 1.74 for the never married females. Most of the values are in the neighborhood of 1.6 .

The last column of the table provides estimates of standard errors of the mean

Table 1.--Standard errors for estimated population totals

Estimate	Standard error
75,000	21,154
100,000	24,451
250,000	38,887
500,000	55,527
750,000	68,650
1,000,000	80,008
2,500,000	133,284
5,000,000	203,473
7,500,000	266,289
10,000,000	326,023
25,000,000	664,744
40,000,000	994,419

derived from a formula suggested by the Census Bureau.

$$\mathrm{StdErr}(M) = \sqrt{(b/Y)}\ \hat{\sigma},$$

where Y is the estimated base of the mean, b is the parameter of the generalized variance curve and the weighted standard deviation is again used as an estimate of the population standard deviation.[12] The advantage of using this formula is that half-sample calculations are not required; however, one must assume that the design effect derived from the estimated b parameter is accurate and appropriate for means. As indicated in table 3, the estimated standard errors from this formula have the same order of magnitude as the replication estimates and there is no apparent pattern to the differences.

Medians

To demonstrate estimated variances for medians, table 4 presents standard errors for estimated medians for the same cells as were used for the estimated means in the previous section. The third column

Table 2.--Standard errors for estimated percents

Base of percents	Percent											
	1 or 99	2 or 98	5 or 95	8 or 92	10 or 90	15 or 85	20 or 80	25 or 75	30 or 70	35 or 65	40 or 60	50
75,000............	2.80	3.94	6.13	7.63	8.44	10.04	11.25	12.18	12.89	13.41	13.78	14.06
100,000...........	2.42	3.41	5.31	6.61	7.31	8.70	9.74	10.55	11.16	11.62	11.93	12.18
250,000...........	1.53	2.16	3.36	4.18	4.62	5.50	6.16	6.67	7.06	7.35	7.55	7.70
500,000...........	1.08	1.52	2.37	2.95	3.27	3.89	4.36	4.72	4.99	5.20	5.34	5.45
750,000...........	.88	1.25	1.94	2.41	2.67	3.18	3.56	3.85	4.08	4.24	4.36	4.45
1,000,000........	.77	1.08	1.68	2.09	2.31	2.75	3.08	3.33	3.53	3.67	3.77	3.85
2,500,000........	.48	.68	1.06	1.32	1.46	1.74	1.95	2.10	2.23	2.32	2.39	2.44
5,000,000........	.34	.48	.75	.93	1.03	1.23	1.38	1.49	1.58	1.64	1.69	1.72
7,500,000........	.28	.39	.61	.76	.84	1.00	1.12	1.22	1.29	1.34	1.38	1.41
10,000,000......	.24	.34	.53	.66	.73	.87	.97	1.05	1.12	1.16	1.19	1.22
25,000,000......	.15	.22	.34	.42	.46	.55	.62	.67	.71	.73	.75	.77
40,000,000......	.12	.17	.27	.33	.37	.43	.49	.53	.56	.58	.6	.61

Table 3.--Estimated standard errors for mean Social Security benefits

Sex and marital status	Count	Population	Mean	Standard deviation	Standard error		
					Replication	Simple random	Census
Total..	7,116	33,067,110	$537	227.2	3.3	2.7	3.0
Married......................................	3,980	19,100,452	528	240.8	4.6	3.8	4.2
Widowed..................................	2,220	9,911,401	558	202.9	4.3	4.3	5.0
Single..	492	2,134,975	532	210.3	9.5	9.5	11.1
Never married...........................	424	1,921,182	512	216.9	13.5	10.5	12.1
Male...	2,958	14,107,315	637	216.7	4.9	4.0	4.4
Married......................................	2,198	10,557,505	656	214.2	5.9	4.6	5.1
Widowed..................................	383	1,730,759	622	218.3	14.1	11.2	12.8
Single..	194	959,900	563	211.8	14.2	15.2	16.6
Never married...........................	183	8,542,947	371	167.6	5.1	4.0	4.4
Female..	4,158	18,959,715	463	205.4	4.1	3.2	3.6
Married......................................	1,782	859,150	511	185.8	14.4	13.7	15.4
Widowed..................................	1,837	8,180,641	545	196.8	4.8	4.6	5.3
Single..	298	1,174,175	507	205.7	13.4	11.9	14.6
Never married...........................	241	1,062,032	514	239.1	20.3	15.4	17.9

shows the estimated medians and the fourth column, the replication standard errors. In general, the coefficients of variation for the medians are slightly larger than for the estimated means but of the same general order of magnitude.

The last column of table 4 provides estimates of standard errors for the medians, again suggested by the Census Bureau, that do not require repeated calculations of the median.[13] The standard errors for each cell are obtained by forming a 68-percent confidence interval about an estimate of 50 percent with a population size equal to the base of the distribution used to calculate the median. The upper and lower bounds of this interval can be obtained from the generalized curve. Then one standard error on the median can be estimated by halving the corresponding 68-percent confidence interval about the median. This interval is obtained by computing the percentile scores corresponding to the upper and lower points of the confidence interval on 50 percent using a distribution of the median variable, in this case the OASDI benefit amount.

In calculating the last column of table 4, the same distributions were used to obtain the upper and lower bounds of the 68-percent confidence interval about the median as those used to compute the medians themselves. Also, the same formula was used with the 50th percentile replaced successively by the upper and lower limits about 50 percent. As shown in table 4, the estimated standard errors under this procedure appear to be generally larger than those obtained by replication.

Covariances

Tables 5 and 6 provide full estimated covariance matrices for the detail cells (the last eight estimates) in tables 3 and 4, respectively. Sampling covariances can be important when calculating standard errors of the differences between estimates because, in general, the variance of the difference between two estimates is equal to the sum of the variances minus twice the covariance.[14] The sum of the variances, which is often used for the variance of the difference when estimates of covariances are not available, may over or understate the variance of the difference depending on the sign and size of the covariance.

Substantial covariances between population estimates arise when the estimates have an underlying structural relationship that is preserved by the clustering in the sample design.[15] As an example, consider the estimates of average Social Security benefits for married men and married women. In the OASDI program, there is a strong connection between husbands' and wives' benefits. Generally, though not in all cases, a wife who is entitled to benefits on her husband's account will receive half of her husband's benefits. In some cases, a wife may not be immediately entitled when her husband is (for example, the younger wife of a retired-worker beneficiary). In other cases, a wife may be entitled to benefits on her own account that are larger than half her husbands. Still, there remains a strong positive association between spousal benefit amounts.

If men and women were sampled independently, survey estimates of the OASDI benefits for married men and women would not be correlated. However, in household surveys such as the SIPP in which both husbands and wives are interviewed (if they are both noninstitutionalized and residing together), one might expect that the positive association between spousal benefit levels to result in a positive correlation between the estimated benefit levels of married men and women.

As shown in table 5, the estimated covariance for mean benefits is fairly large relative to the variances. The variances for married men and women

Table 4.--Estimated standard errors for median Social Security benefits

Sex and marital status	Count	Population	Median	Standard error	
				Replication	Census
Total	7,116	33,067,110	$ 533	3.9	4.1
Married	3,980	19,100,452	517	6.9	7.4
Widowed	2,220	9,911,401	554	4.1	5.3
Single	492	2,134,975	525	12.2	14.5
Never Married	424	1,921,182	505	12.6	14.1
Male	2,958	14,107,315	646	4.1	4.5
Married	2,198	10,557,505	666	5.0	5.1
Widowed	383	1,730,759	611	12.7	12.8
Single	194	959,900	580	19.9	22.7
Never Married	183	859,150	503	20.1	22.9
Female	4,158	18,959,715	440	5.0	5.0
Married	1,782	8,542,947	351	3.7	4
Widowed	1,837	8,180,641	542	4.9	5.8
Single	298	1,174,175	487	13.9	15.9
Never Married	241	1,062,032	505	16.4	17.9

are 34.7 and 25.5, respectively; and the covariance is 12.2. One standard error on the difference, assuming a zero covariance, is 7.8. Subtracting twice the estimated covariance from the sum of the variances, one standard error on the difference is 6.0, about 23 percent smaller than the estimate assuming zero covariance. Although this difference may not be particularly important here because of the large spread in mean benefit amounts ($656 for married men, compared with $371 for married women), such differences could be important in other contexts. A similar reduction in the estimated standard error for the difference in medians is obtained from the figures in table 6.

Conclusion

This article provided sampling variance estimates for Social Security program participants from the SIPP. The methodology employed, balanced half-sample replication, was the same as that reported in a 1988 *Bulletin* article in connection with the 1984 SIPP panel. Formulas for computing sampling variances and covariances have been presented and demonstrated for count data, means and medians. Because replication variance estimation is not difficult to implement for the SIPP and facilitates a wide range of hypothesis testing techniques, it was recommended that direct variance calculations be used.

For those who cannot compute variances directly, a generalized curve and standard error tables for counts and proportions have been provided for Social Security program participants aged 18 or older. The standard error tables pertain directly to the SIPP tables in the *Annual Statistical Supplement* to the *Social Security Bulletin* for 1992 and 1993, and can be used for other analyses as well. These generalized variances appear to be more appropriate for estimates pertaining to Social Security program participants than curves provided by the Census Bureau.

This article has also provided some indication of the usefulness of the estimators of sampling variances for

Table 5.--Covariance matrix for estimated means

Sex and marital status	Count	Mean	Male				Female			
			Married	Widowed	Single	Never married	Married	Widowed	Single	Never married
Male:										
Married	2,198	$656	34.7	....	...	...	...	...	...	...
Widowed	383	622	-7.4	198.9	...	...	...	...	...	...
Single	194	563	7.5	-21.8	202.6	...	...	...	...	...
Never married	183	511	-6.1	3.7	47.5	207.4	...	...	...	...
Female:										
Married	1,782	371	12.2	-5.3	7.5	-2.3	25.5	...	...	...
Widowed	1,837	545	3.3	-7.1	-14.8	1.5	0.0	23.0	...	...
Single	298	507	22.9	27.7	-7.2	9.5	16.5	-2.4	179.6	...
Never married	241	514	-7.7	-9.9	35.8	29.7	-9.2	23.4	-44.5	413.0

Table 6.--Covariance matrix for estimated medians

Sex and marital status	Count	Median	Male				Female			
			Married	Widowed	Single	Never married	Married	Widowed	Single	Never married
Male:										
Married	2,198	$666	25.2	...	...	...	...	...	...	...
Widowed	383	611	-7.0	160.7	...	...	...	...	...	...
Single	194	580	12.0	-4.8	395.9	...	...	...	...	...
Never married	183	503	-1.9	33.2	64.0	405.5	...	...	...	...
Female:										
Married	1,782	351	6.5	-3.7	4.2	-6.4	13.8	...	...	...
Widowed	1,837	542	3.3	-4.2	-9.7	0.1	1.2	23.6	...	...
Single	298	487	5.1	-1.5	-47.4	33.6	4.2	1.6	192.2	...
Never married	241	505	-17.1	-18.7	-34.4	-16.2	-17.2	15.7	-6.7	268.8

means and medians suggested by the Census Bureau that are based on the generalized variance curve parameters. These estimates of standard errors have roughly the same order of magnitude as those computed directly, and the curve-based estimates for medians appear to be more conservative than the direct computations.

One issue concerning the appropriateness of the methodology raised in the previous report on the 1984 SIPP panel has been addressed. Variance calculations for estimated population totals using the pseudo sample design indicators provided in the 1984 public use file have been compared with internal Census Bureau calculations using the actual sample design.[16] The results from the public use file were quite similar to the sampling variance calculations performed internally at the Census Bureau, giving much support to the approach recommended here. Although such comparisons were not repeated for the 1990 panel, there is no reason to believe that similar results would not be obtained.

An issue that still requires investigation concerns the raw sample sizes that are required before the assumption of normality in the sampling distributions of the various statistics is appropriate. If sampling distributions from estimates derived from small numbers of cases differ markedly from the normal, then it might be quite misleading to form confidence intervals and perform statistical tests assuming a normal distribution (for example, assuming that symmetric intervals of one standard error about an estimate yields a 68-percent confidence interval or two standard errors provides a 95-percent confidence interval). The true confidence intervals and significance levels may be larger or smaller than those calculated assuming normality, and symmetric confidence intervals may not be appropriate. More information is needed on the shape of the sampling distributions of the survey estimates.

Notes

[1] General information on the SIPP can be found in Dawn Nelson, David McMillen, and Daniel Kasprzyk, *An Overview of the Survey of Income and Program Participation* (SIPP Working Paper Series, No. 8401, update 1), Bureau of the Census, Department of Commerce, 1985.

[2] *Annual Statistical Supplement* to the *Social Security Bulletin, 1992 (1993)*, Office of Research and Statistics, Social Security Administration, 1992 (1993), tables 3.C9-C11, 3.D1, 5.A11-A13, and 7.A6-A7.

[3] *Source and Accuracy Statement for 1990 Public Use Files From the Survey of Income and Program Participation*, Bureau of the Census, Department of Commerce, May 1992.

[4] Kirk Wolter, *Introduction to Variance Estimation*, Springer-Verlag, New York, 1985.

[5] Barry V. Bye and Salvatore J. Gallicchio, "A Note on Sampling Variance Estimates for Social Security Program Participants From the Survey of Income and Program Participation," *Social Security Bulletin*, Vol. 51, No. 10 (1988), pp. 4-21.

[6] Bureau of the Census (1992) *Op. cit.*, Generalized Variance Parameters, Program Participation and Benefits, Poverty.

[7] R. L. Plackett and J. P. Burman, "The Design of Optimum Multifactor Experiments," *Biometrika*, 33 (1946), pp. 305 and 325.

[8] Wolter (1985), *op. cit.*, references a number of empirical investigations supporting the use of equation (1).

[9] See, for example, *The Current Population Survey: Design and Methodology* (Tech Paper 40), Bureau of the Census, Department of Commerce, January 1978.

[10] J. R. Grizzle, C. F. Starmer, and G. C. Koch, "Analysis of Categorical Data by Linear Models," *Biometrics*, September 1969, pp. 489-504. The test procedures suggested by Grizzle *et al.* are implemented in the SAS CATMOD procedure (*SAS Procedure Guide*, Version 6, Third Edition, SAS Institute Inc., 1990).

[11] The variance estimates for the 1990 panel are similar to those of the 1984 panel. This result would be expected because the sample sizes and first stage designs are similar. The generalized curve for 1990 has a slightly different orientation than the 1984 curve, giving slightly larger estimates of standard errors for population estimates below 5 million and slightly smaller estimates over that number.

[12] Bureau of the Census (1992), *op. cit.*, This formula is apparently motivated by the following. If the design effect is constant for estimated means, then

$$\mathrm{StdErr}(M) = \sqrt{\mathrm{Deff}}\,(\hat{\sigma}/\sqrt{n}).$$

Assuming that an estimate of the design effect can be obtained from equation (3),

$$\sqrt{\mathrm{Deff}} = \sqrt{b(n/Y)}$$

where Y is the base of the mean. Substitution of this equation into the previous equation yields the Census Bureau formula.

[13] Bureau of the Census (1992) *op. cit.*

[14] Another way to obtained standard errors for the differences of means or medians is to compute the difference in each half sample and then estimate the standard error of the difference directly using equation (1).

[15] Correlations between sample estimates can be introduced by interviewer error, even when respondents are sampled independently.

[16] Barry Bye and Salvatore Gallicchio, *Two Notes on Sampling Variance Estimates from the 1984 SIPP Public-Use Files* (SIPP Working Paper Series 8902), Bureau of the Census, Department of Commerce, April 1989.

Appendix: Detailed Sampling Variance Specifications

Assignment of Half-Sample Codes

Respondents in the 1990 SIPP file have been assigned a pseudo-stratum code and a pseudo primary sampling unit (PSU) code within each pseudo stratum.[1]

Generally, a self-representing (SR) PSU from the original design was associated with two non-self-representing (NSR) PSUs to form a pseudo stratum. Segments of the SR PSU were assigned to one of the two PSUs at random; each of the NSR PSUs was assigned, in its entirety, to one or the other of the pseudo units. In some cases, two SR PSUs or four NSR PSUs were grouped to form a pseudo stratum. The assignment result in the formation of 72 pseudo strata with two pseudo PSUs per stratum.[2]

The set of orthogonal half samples used in the variance computations is shown in chart I. The array represents a string of 72 1s and 0s that indicate half-sample membership for cases in each of the 144 combinations of stratum and PSUs. Each row of the array identifies the half samples to which the case with the corresponding stratum and PSU code is to be assigned. A "1" in the kth position in the string indicates that the case is to be included in the kth half sample; a "0" means the case is not to be included in the corresponding half sample. For example, a case in Stratum 1, PSU 1 is in the first seven half samples, not in the 8th, in the 9th, 10th and 11th, not in the 12th, and so on.

Item Specification for Generalized Variances

Replication variances were obtained for estimated totals of Social Security program recipients, aged 18 or older. Recipiency status was determined by the responses for May, 1990. Estimated population totals were obtained in each half sample by multiplying the sum of the weights by 2.[3] The recipients were cross-classified by age, sex, and marital status. The cross-classification yielded 148 distinct detailed and subtotal cells, of which 126 had 25 or more cases.

The May, 1990, recipient universe consists of those persons, 18 or older in the sample who meet the following test:[4]

[IO1AMT-*>0 or IO3AMT-*>0]
and
[AGE-* >17] and [FNLWGT-*>0]

where

IO1AMT-*	refers to the OASDI benefit amount-,
IO3AMT-*	refers to the SSI amount-,
AGE-*	is age in May, 1990, and
FNLWGT-*	is the case weight.

Each variable group is selected for May based on the rotation group of the sample case shown below.

Rotation group	Month
1	1
2	4
3	3
4	2

The cross-classications were constructed as follows:

Age (AGE-*):

Under 18	55–64
18–24	65–69
25–34	70–74
35–44	75–84
45–54	85 or older

Sex:
Male, Female

Marital status (MS-*)	Code
Married	Under 3
Widowed	3
Separated	4, 5
Never married	6 or over

Table I presents the estimated sampling variances for the 148 items described above.

Chart I.--Half sample assignments by stratum and primary sampling unit

Stratum	PSU	Half sample
01	1	111111101110100110111000110101101000111010010100111000100110100010000000
01	2	000000010001011001000111001010010111000101101011000111011001011101111111
02	1	111111011101001101110001101011010001110100101001110001001101000100000010
02	2	000000100010110010001110010100101110001011010110001110110010111011111101
03	1	111110111010011011100011010110100011101001010011100010011010001000000110
03	2	000001000101100100011100101001011100010110101100011101100101110111111001
⋮		
71	1	011111110111010011011100011010110100011101001010011100010011010001000000
71	2	100000001000101100100011100101001011100010110101100011101100101110111111
72	1	000000000000000000000000000000000000000000000000000000000000000000000000
72	2	111111111111111111111111111111111111111111111111111111111111111111111111

Table I.--Variance estimates for SSA recipients

Age	Sex	Marital status [1]	Unweighted count	Estimate	Standard error	Relative variance
Total	Total	Total	8,024	36,944,301	679,343	.0003381
Total	Total	M	4,156	19,853,513	509,472	.0006585
Total	Total	W	2,470	10,934,407	323,317	.0008743
Total	Total	S	721	3,066,786	151,859	.0024520
Total	Total	NM	677	3,089,595	148,592	.0023131
Total	Male	Total	3,236	15,438,107	364,529	.0005575
Total	Female	Total	4,788	21,506,194	437,547	.0004139
Total	Male	M	2,280	10,917,384	270,021	.0006117
Total	Male	W	406	1,820,380	124,721	.0046942
Total	Male	S	239	1,163,723	95,863	.0067858
Total	Male	NM	311	1,536,620	105,324	.0046981
Total	Female	M	1,876	8,936,129	259,930	.0008461
Total	Female	W	2,064	9,114,026	293,846	.0010395
Total	Female	S	482	1,903,063	109,573	.0033151
Total	Female	NM	366	1,552,975	94,486	.0037017
18-24	Total	Total	95	439,041	61,861	.0198528
25-34	Total	Total	181	909,662	86,800	.0091049
35-44	Total	Total	214	899,070	83,866	.0087014
45-54	Total	Total	263	1,120,021	82,068	.0053690
55-64	Total	Total	1,239	5,719,844	206,451	.0013028
65-69	Total	Total	1,956	8,990,481	256,781	.0008158
70-75	Total	Total	1,900	8,855,979	280,586	.0010038
75-79	Total	Total	948	4,331,678	194,538	.0020170
80+	Total	Total	1,228	5,678,525	238,377	.0017622
18-24	Total	M	7	38,580	17,284	.2007143
18-24	Total	W	2	8,246	5,893	.5107441
18-24	Total	S	4	11,586	6,919	.3565935
18-24	Total	NM	82	380,629	59,732	.0246266
25-34	Total	M	37	168,397	34,344	.0415952
25-34	Total	W	8	28,881	10,328	.1278757
25-34	Total	S	29	141,983	38,428	.0732516
25-34	Total	NM	107	570,401	75,138	.0173525
35-44	Total	M	72	302,923	37,144	.0150353
35-44	Total	W	31	132,612	31,174	.0552606
35-44	Total	S	48	186,254	33,360	.0320803
35-44	Total	NM	63	277,280	41,243	.0221239
45-54	Total	M	107	489,658	50,835	.0107780
45-54	Total	W	36	129,598	25,500	.0387155
45-54	Total	S	68	270,681	38,394	.0201192
45-54	Total	NM	52	230,084	30,717	.0178233
55-64	Total	M	740	3,512,365	174,236	.0024608
55-64	Total	W	249	1,083,619	80,521	.0055216
55-64	Total	S	176	811,112	78,589	.0093877
55-64	Total	NM	74	312,748	38,593	.0152278
65-69	Total	M	1,284	6,138,688	221,690	.0013042
65-69	Total	W	411	1,756,779	102,164	.0033819
65-69	Total	S	165	676,301	56,012	.0068593
65-69	Total	NM	96	418,713	48,644	.0134966
70-75	Total	M	1,085	5,295,145	234,868	.0019674
70-75	Total	W	610	2,672,453	138,370	.0026808
70-75	Total	S	124	511,233	50,773	.0098635
70-75	Total	NM	81	377,148	51,312	.0185101

See footnote at end of table.

Table I.--Variance estimates for SSA recipients -- *Continued*

Age	Sex	Marital status [1]	Unweighted count	Estimate	Standard error	Relative variance
75-79	Total	M	439	2,080,293	125,132	.0036182
75-79	Total	W	397	1,793,735	111,402	.0038572
75-79	Total	S	57	231,636	37,785	.0266092
75-79	Total	NM	55	226,014	34,922	.0238743
80+	Total	M	385	1,827,463	154,284	.0071276
80+	Total	W	726	3,328,483	151,663	.0020762
80+	Total	S	50	226,000	40,596	.0322656
80+	Total	NM	67	296,579	37,742	.0161946
18-24	Male	Total	51	251,690	49,500	.0386785
18-24	Female	Total	44	187,351	29,864	.0254085
25-34	Male	Total	90	485,964	58,599	.0145404
25-34	Female	Total	91	423,699	58,341	.0189599
35-44	Male	Total	78	359,007	51,523	.0205964
35-44	Female	Total	136	540,063	58,220	.0116214
45-54	Male	Total	105	474,332	55,253	.0135690
45-54	Female	Total	158	645,690	59,138	.0083884
55-64	Male	Total	499	2,465,795	105,392	.0018269
55-64	Female	Total	740	3,254,049	161,892	.0024752
65-69	Male	Total	833	3,919,132	154,492	.0015539
65-69	Female	Total	1,123	5,071,348	167,215	.0010872
70-75	Male	Total	775	3,693,878	148,711	.0016208
70-75	Female	Total	1,125	5,162,102	200,263	.0015050
75-79	Male	Total	385	1,811,737	110,884	.0037458
75-79	Female	Total	563	2,519,941	128,496	.0026001
80+	Male	Total	420	1,976,573	125,365	.0040228
80+	Female	Total	808	3,701,952	179,298	.0023458
18-24	Male	M	3	14,778	8,574	.3366688
18-24	Male	NM	48	236,913	48,751	.0423442
18-24	Female	M	4	23,803	13,735	.3329805
18-24	Female	W	2	8,246	5,893	.5107441
18-24	Female	S	4	11,586	6,919	.3565935
18-24	Female	NM	34	143,716	27,353	.0362232
25-34	Male	M	12	59,129	19,969	.1140531
25-34	Male	W	1	3,414	3,414	1.0000063
25-34	Male	S	7	39,736	18,981	.2281646
25-34	Male	NM	70	383,684	54,661	.0202959
25-34	Female	M	25	109,268	26,922	.0607043
25-34	Female	W	7	25,467	9,747	.1464879
25-34	Female	S	22	102,246	27,662	.0731926
25-34	Female	NM	37	186,717	42,157	.0509759
35-44	Male	M	29	125,238	23,566	.0354082
35-44	Male	W	5	27,132	12,619	.2163004
35-44	Male	S	14	72,098	24,288	.1134872
35-44	Male	NM	30	134,538	22,732	.0285483
35-44	Female	M	43	177,685	26,726	.0226238
35-44	Female	W	26	105,480	28,013	.0705327
35-44	Female	S	34	114,157	20,491	.0322199
35-44	Female	NM	33	142,741	31,137	.0475842

See footnote at end of table.

Table I.--Variance estimates for SSA recipients -- *Continued*

Age	Sex	Marital status [1]	Unweighted count	Estimate	Standard error	Relative variance
45-54	Male	M	52	227,559	38,044	.0279500
45-54	Male	W	8	26,463	10,779	.1658976
45-54	Male	S	20	91,698	22,932	.0625401
45-54	Male	NM	25	128,612	23,574	.0335977
45-54	Female	M	55	262,099	36,128	.0190002
45-54	Female	W	28	103,135	23,379	.0513868
45-54	Female	S	48	178,984	30,814	.0296400
45-54	Female	NM	27	101,473	20,482	.0407437
55-64	Male	M	373	1,788,406	91,400	.0026119
55-64	Male	W	20	89,325	20,911	.0548021
55-64	Male	S	70	408,739	56,971	.0194272
55-64	Male	NM	36	179,326	28,582	.0254046
55-64	Female	M	367	1,723,959	127,483	.0054683
55-64	Female	W	229	994,295	79,263	.0063549
55-64	Female	S	106	402,372	48,680	.0146370
55-64	Female	NM	38	133,422	24,209	.0329240
65-69	Male	M	677	3,239,904	137,264	.0017949
65-69	Male	W	58	224,101	35,622	.0252663
65-69	Male	S	57	255,538	36,766	.0207003
65-69	Male	NM	41	199,589	39,641	.0394479
65-69	Female	M	607	2,898,784	132,134	.0020778
65-69	Female	W	353	1,532,678	95,338	.0038693
65-69	Female	S	108	420,763	44,026	.0109483
65-69	Female	NM	55	219,124	27,194	.0154014
70-75	Male	M	610	2,959,629	145,967	.0024324
70-75	Male	W	96	407,808	48,343	.0140526
70-75	Male	S	40	173,553	32,989	.0361308
70-75	Male	NM	29	152,887	31,087	.0413453
70-75	Female	M	475	2,335,517	129,585	.0030785
70-75	Female	W	514	2,264,645	132,267	.0034112
70-75	Female	S	84	337,680	46,407	.0188871
70-75	Female	NM	52	224,260	33,270	.0220085
75-79	Male	M	263	1,266,748	86,378	.0046497
75-79	Male	W	87	405,117	54,211	.0179063
75-79	Male	S	17	73,157	20,547	.0788817
75-79	Male	NM	18	66,715	15,731	.0556012
75-79	Female	M	176	813,545	66,383	.0066582
75-79	Female	W	310	1,388,618	88,826	.0040918
75-79	Female	S	40	158,480	30,641	.0373826
75-79	Female	NM	37	159,299	30,762	.0372907
80+	Male	M	261	1,235,993	95,708	.0059960
80+	Male	W	131	637,020	71,296	.0125263
80+	Male	S	14	49,204	15,309	.0968034
80+	Male	NM	14	54,356	17,398	.1024504
80+	Female	M	124	591,470	79,481	.0180576
80+	Female	W	595	2,691,463	135,382	.0025301
80+	Female	S	36	176,796	36,399	.0423876
80+	Female	NM	53	242,223	36,518	.0227294

[1] M = married, W = widowed, S = single, NW = never married.

Notes

[1] The fields are identified as H*-STRAT and H*-HSC in the public use file data dictionary. The version of the wave 2, 1990, file used for these calculations was not the public use version and did not have pseudo stratum and half-sample codes assigned to new entrants to the panel at wave 2. (The public use version has codes for all cases.) Of the 8,024 adult SSA program recipients, 41 had no codes assigned. These cases contributed to overall population estimates but not to the half-sample estimates. Because the number of cases was small, the impact on variance calculations is not important.

[2] The 72 order design in Plackett and Burman (1946), *op. cit.*, was used. The half-sample indicators for Strata 2-71 for cases with PSU = 1 can be generated by from the row for Stratum 1 by shifting the first 71 digits one digit to the left, successively for each subsequent Stratum. The half sample indicators for stratum 72 and PSU = 1 are all "0"s. The indicators for cases with PSU = 2 are the complements of the indicators ("1"s are replaced by "0"s and vice versa) for PSU = 1, within each stratum.

Note that for the 1990 panel, the number of pseudo strata, 72, is equal to the number of half samples used for variance estimation. The 1984 panel had 71 pseudo strata. Also note that chart I of the 1988 *Bulletin* article (which showed rows only for cases with PSU = 1) was incorrect. The rows of the array contain only 71 items; the last item of each row should have been a "0" but was inadvertently omitted.

[3] This half-sample estimator does not fully replicate original SIPP esitmates in each half sample because the noninterview and post-stratification adjustments in the construction of case weights were not repeated in each half sample. The overall effect on the estimated variance is not known.

[4] All variables are referred to by their public use file names.