Methods of Statistical Inference for the Survey of Income and Program Participation that are Suitable for an Online Data Analysis Tool

Reid A. Rottach and David W. Hall, U.S. Bureau of the Census

Key Words: Calibration, Variance estimation, Degrees of freedom, Balanced Repeated Replication (BRR), Residual technique

1. Introduction[*]

Data analysis for the Survey of Income and Program Participation (SIPP) requires special consideration due to the survey's complex design and the methodology used to adjust design weights. Balanced Repeated Replication (BRR) has proven reliable at estimating variances for a broad range of SIPP statistics, and this has been the primary method of design-based estimation since the survey began. But BRR requires access to large replicate weight files, which in some circumstances can be problematic. We present an alternative method following an approach that incorporates linearization by way of regression. The variances of totals, ratios, quantiles, and yearly changes will be considered, along with estimates of their degrees of freedom, which will allow construction of confidence intervals based on the *t* distribution. Data from the 1996 panel of SIPP are used for numerical comparisons with BRR.

Our motivation for pursuing a new method of variance estimation is to improve public access to SIPP data. Currently, the Census Bureau is examining the feasibility of allowing interactive data analysis online through an expanded version of FERRET, the Census Bureau's data extraction tool. Due to computing constraints, reading in a lot of replicate weights, such as the 108 available for the 1996 SIPP panel, is not practical. Significantly reducing the number of replicate weights may be possible, but this has the drawback of reducing the stability of the variance estimator. The linearization approach described in this paper is being considered for the FERRET system. It is non-iterative, requires few additional variables, and results in variance estimates that are comparable to BRR estimates. Furthermore, its added capability of estimating degrees of freedom is especially important for cross-tabulations and statistics related to rare characteristics, where variance stability may be an issue.

2. Background

2.1 The Survey Design and Weighting Procedure

The SIPP uses a two-stage sample design. In the first stage, primary sampling units (PSUs) are selected from strata with probability proportional to size. Some of the larger PSUs form unique (self-representing, or SR) strata, which guarantees they will be selected. In the 1996 panel, two PSUs were selected without replacement from each of the remaining (non-self-representing, or NSR) strata. Within PSUs, clusters of households are selected systematically.

The households are divided into four replicate sample groups, called rotations. Each month, one of the rotations is designated for interview, and the household reference persons are asked questions about each of the previous four months. So for a given reference month, data collection usually spans four consecutive months of interview, one for each of the rotation groups.

When a household is initially selected into SIPP's sample, a design weight is assigned that is equal to the inverse of its probability of selection. A series of adjustments are made to the weights for improved estimation. Some factors are needed to account for the discrepancy between the probability of selection at the design stage and the probability of selecting a survey respondent. Prior to the 1996 panel, a first-stage ratio adjustment was applied to sample in NSR strata to reduce the between-PSU variances. Beginning with the 1996 panel, this adjustment was dropped because there was insufficient evidence that it improved estimation. In all panels, the final adjustment made is the second-stage adjustment, which ensures that certain SIPP estimates of population totals match control totals derived from the Current Population Survey (CPS), and simultaneously forces the final weights of husbands and wives to be equal. The CPS controls can be interpreted as three-way tables, of which only the marginals, formed on combinations of age, marital status, family type, race, sex, and ethnicity, are supplied. So estimates of cell totals remain subject to SIPP sampling variability, while estimates

[*] **Disclaimer:** This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications. This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress.

of marginals do not. These adjustments are done independently among rotations of SIPP, where each rotation is weighted to one-quarter of the CPS controls. The second stage adjustment is really a series of ratio adjustments and husband-wife equalizations applied iteratively until all constraints are met, a process referred to as raking. SIPP statistics closely approximate raking ratio estimators, which are discussed in many sources, including Sarndal, Deville, and Sautory [1993] on the subject of variance estimation. Differences from the raking ratio estimator are due to the nonresponse adjustment and the procedure for achieving Hispanic control totals, whereby these are the first controls to be met, at which point only the weights of non-Hispanics are adjusted to meet the total population controls.

2.2 Inferential Statistics

The variance codes available to public users are pseudo-stratum and half-sample (cluster). Direct variance estimation is based on the two-cluster-per-stratum estimator when sampling occurs with replacement. For confidentiality reasons, these codes do not provide complete information on the sample design, such as the information needed to use the "without replacement" estimator, but they can be used for reliable estimates of variance. Generally, the estimator will have a small positive bias.

Given a statistic of the form

$$\hat{Y} = \sum_{h,i,k} a_{hik} y_{hik} \qquad (1)$$

where $h = 1, 2, ..., L$ represents pseudo-stratum; $i = 1, 2$ represents half-sample within pseudo-stratum h; k represents a sample person within pseudo-stratum h, cluster i; y_{hik} is a numeric variable assigned to every person in sample; a_{hik} is the design weight;

the variance estimator is

$$\hat{V}(\hat{Y}) = \sum_{h=1}^{L} \{\hat{Y}_{h1} - \hat{Y}_{h2}\}^2 \qquad (2)$$

where $\hat{Y}_{hi} = \sum_{k} a_{hik} y_{hik}$.

This formula is only appropriate for linear statistics; that is, statistics that are linear functions of the design weights. To be of any real use, the formula must be generalized to nonlinear statistics, which includes any statistic computed using final weights. Typically, variance estimation for nonlinear statistics follows either a replication approach or a linearization approach. The SIPP has supported replication estimates of variance through the use of replicate weight files provided to public users. These replicate weights extend equation (2), via a method developed by Robert Fay [1984, 1989], to statistics computed using SIPP final weights. To calculate variances, the user must first calculate the statistic independently for each weight in the replicate array, and then input the set of statistics to a quadratic function. Refer to the SIPP Users' Guide [U.S. Census Bureau, 2001] for details. One of the appeals of this estimator is the uniform treatment of any type of statistic that is computed with the final weights, regardless of the statistic's complexity.

Linearization methods modify the statistic, and leave the form of the variance estimator mostly unchanged. To estimate the variance of a nonlinear statistic, create an artificial linear statistic that has approximately the same variance as that of the nonlinear statistic and use equation (2) directly.

The primary difficulty in applying linearization methods is finding an expression for the artificial statistic. Linearizations of common statistics, such as ratios, means, and medians, are easy to find in statistical literature, but in most cases they are not intended for use with poststratified weights. The procedure for creating SIPP final weights is very complex, and methods of linearization based strictly on Taylor Series approximations do not easily lend themselves to statistics computed with these weights. An alternative method has been developed in recent years that is much simpler than the pure Taylor Series approach. It is the residual technique outlined by Deville [1999]. We have extended this method of variance estimation to estimation of degrees of freedom using the Satterthwaite approximation, to allow statistical tests based on the *t* distribution.

Due to its suitability to our survey design, we recommend the estimator described by Eltinge and Jang [1996] that was derived using method of moment arguments. With a statistic in the form of equation (1) and variance estimator in the form of equation (2), the degrees of freedom approximation is

$$\hat{d}\{\hat{V},\hat{Y}\} = \frac{3\{\hat{V}(\hat{Y})\}^2}{\sum_{h=1}^{L}\{\hat{V}_h(\hat{Y})\}^2} - 2 \qquad (3)$$

where $\hat{V}_h(\hat{Y}) = \{\hat{Y}_{h1} - \hat{Y}_{h2}\}^2$ is the component of variance from stratum *h*.

The methodology of BRR does not readily lend itself to Satterthwaite-like estimates of degrees of freedom.

3. The Residual Technique

3.1 Overview and Notation

In this discussion, $\hat{T}$ is a statistic computed using final weights. The residual technique is a method of linearization, so the first objective is to create a synthetic variable such that its weighted sum has approximately the same variance as $\hat{T}$. This occurs in two stages, where first an intermediate variable, s_{hik}, is derived from $\hat{T}$ following a linearization approach but replacing design weights with final weights. That is, the variable s_{hik} has the theoretical property

$$V(\hat{S}) \approx V(\hat{T})$$

where $\hat{S} = \sum_{h,i,k|R} w_{hik} s_{hik}$; *R* represents a restriction to survey respondents; w_{hik} is the final weight.

Secondly, we create a new statistic that is a linear function of the non-interview adjusted weights and has approximately the same variance as $\hat{S}$. If the response probabilities were known, this would complete the linearization procedure since replacing design weights with the non-interview adjusted weights (the ratio of the design weight and the response probability) in equations (1) and (2) would provide similarly valid estimators. If the response probabilities are not known, but estimated adequately, it is often acceptable to treat them as known quantities. Note that the BRR estimator does not rely on this assumption, although our linearization estimator does. For simplicity, the language used when describing the linearization approach will be consistent with defining a linear statistic to be linear in the non-interview adjusted weights.

The artificial statistic $\hat{S}$ is treated as a generalized regression estimator of a population total. As such, the following relationship is established in the theory of generalized regression estimators:

$$V(\hat{S}) \approx V\left(\sum_{h,i,k|R} b_{hik} e_{hik}\right)$$

where b_{hik} is the non-interview adjusted weight; e_{hik} is a residual in the weighted least squares regression of *s* onto indicators of control groups.

The regression procedure completes the linearization, and a variance estimator derived from equation (2) is constructed. This is accomplished following the arguments made by Stukel, Hidiroglou, and Sarndal [1996, Section 3]. The estimator is

$$\hat{V}(\hat{T}) = \sum_{h=1}^{L}\{\hat{E}_{h1} - \hat{E}_{h2}\}^2 \qquad (4)$$

where $\hat{E}_{hi} = \sum_{k|R} w_{hik} e_{hik}$.

Extending equation (3) to these circumstances gives the implied degrees of freedom approximation:

$$\hat{d}\{\hat{V},\hat{T}\} = \frac{3\{\hat{V}(\hat{T})\}^2}{\sum_{h=1}^{L}\{\hat{E}_{h1} - \hat{E}_{h2}\}^4} - 2 \qquad (5)$$

3.2 The First Transformation

Expressions for s_{hik} depend on the form of the statistic $\hat{T}$ and are often easy to find in statistical literature. Some examples are given in the following table. These expressions were used to generate the numerical results in this paper.

Table 1

Examples of Common Transformations[a]

Description	$\hat{T}$	s
Total	$\sum wy$	y
Ratio or Mean	$\frac{\sum wy}{\sum wz}$	$\frac{1}{\sum wz}(y-\hat{T}z)$
Quantile	Y_p^w	$\frac{Y_{p_2}^w - Y_{p_1}^w}{(p_2-p_1)\sum w}\{I(y \le Y_p^w)-p\}$

[a]All summations are performed over survey respondents; *w* represents the final weight; *y* and *z* are numeric variables; Y_p^w is the p^{th} weighted sample quantile of the variable *y*; p_1 and p_2 have the property $0<p_1<p<p_2<1$; $I(\cdot)$ is an indicator function.

3.3 The Second Transformation

A weighted least squares regression (weighted by final weights) is required to create the variable e_{hik} from the variable s_{hik}. For the 1996 and 2001 panels, the control groups we need for the regression are marginals in a three dimensional table. One dimension consists of sex by age categories of Hispanics, another dimension is categories of sex by race by age of the total population, and the last is categories of sex by race by household and family member type of the total population.

Table 2

First Marginal Control Categories

Age	Hispanic Male	Hispanic Female	Non-Hispanic
0-14	1	5	9
15-24	2	6	9
25-44	3	7	9
45+	4	8	9

Table 3

Second Marginal Control Categories

Age	Black Male	Black Female	Non-black Male	Non-black Female
<1	1	24	48	80
1	2	25	49	81
2	2	25	50	82
3	2	25	51	83
4	3	26	52	84
5	3	26	53	85
6	4	27	54	86
7	4	27	55	87
8	5	28	56	88
9	5	28	57	89
10-11	6	29	58	90
12-13	7	30	59	91
14	8	31	60	92
15	9	32	61	93
16-17	10	33	62	94
18-19	11	34	63	95
20-21	12	35	64	96
22-24	13	36	65	97
25-29	14	37	66	98
30-34	15	38	67	99
35-39	16	39	68	100
40-44	17	40	69	101
45-49	18	41	70	102
50-54	19	42	71	103
55-59	20	43	72	104
60-61	21	44	73	105
62-64	21	44	74	106
65-69	22	45	75	107
70-74	23	46	76	108
75-79	23	47	77	109
80-84	23	47	78	110
85+	23	47	79	111

Table 4

Third Marginal Control Categories[b]

Household & Family Member Type	Black Male		Black Female		Non-black Male		Non-black Female	
	C	A	C	A	C	A	C	A
HH1	-	3	-	11	-	19	-	27
HH2	-	4	-	12	-	20	-	28
HH3	-	5	-	13	-	21	-	29
HH4	1	6	9	14	17	22	25	30
HH5	-	7	-	15	-	23	-	31
HH6	2	8	10	16	18	24	26	32

[b] C = Child (Age 0-14); A = Adult (Age 15+); HH1 = Household with family, spouse in primary family; HH2 = Household with family, householder, no spouse present; HH3 = Household with family, spouse in subfamily; HH4 = Household with family, other household member; HH5 = Household without family, householder; HH6 = Household without family, other household member.

Since each rotation of SIPP is raked to match one-quarter of the population controls within the marginal categories, the marginals should be viewed as being nested within rotation group.

Note that in some instances, such as when the sample representing a control group is considered too small, a control group will be collapsed with another. In effect, there will be fewer control groups than are listed in the tables. The rules for collapsing are complicated and are not discussed here. To obtain the numerical results for the linearization method presented in this paper, we have ignored the effect of collapsing control groups.

Summary of the regression model: The residuals e_{hik}, analogous to those expressed in equation (9.1) of Deville, Sarndal, and Sautory [1993], are computed from the w_{hik} -weighted least squares regression of s_{hik} onto the three categorical marginals, all nested within rotation group. Every respondent in the cross-section with a positive weight should be included in the regression.

4. Cross-Sectional Differences

Due to the longitudinal design of SIPP, different cross-sections of the survey within the same panel are likely to contain many of the same persons in sample, resulting in a high correlation between estimates calculated at the two points in time. This will have a substantial effect on variance estimates of cross-sectional differences, so treating the statistic as the difference between two independent samples is unreasonable. A more thorough discussion of this issue may be found in Roberts, et al [2001]. Let $\hat{\Delta} = \hat{T}_2 - \hat{T}_1$ be the estimated difference between time=2 and time=1 of a measurement. Then we have

$$V(\hat{\Delta}) = V(\hat{T}_1) + V(\hat{T}_2) - 2Cov(\hat{T}_1, \hat{T}_2)$$

which leads to the estimator

$$\hat{V}(\hat{\Delta}) = \hat{V}(\hat{T}_1) + \hat{V}(\hat{T}_2) - 2C\hat{o}v(\hat{T}_1, \hat{T}_2)\,.$$

If equation (4) is generalized to estimates of covariance by replacing the squared term with a cross-product, the variance estimator given above reduces to

$$\hat{V}(\hat{\Delta}) = \sum_{h=1}^{L} \{\tilde{\Delta}_{h1} - \tilde{\Delta}_{h2}\}^2 \qquad (6)$$

where $\tilde{\Delta}_{hi} = \sum_{k|R_2} w_{2,hik} e_{2,hik} - \sum_{k|R_1} w_{1,hik} e_{1,hik}$; R_1 and R_2 are survey respondents at times 1 and 2, respectively; $w_{1,hik}$ and $w_{2,hik}$ are final weights at times 1 and 2, respectively; $\hat{T}_1$ has a linearized variable $e_{1,hik}$ and $\hat{T}_2$ has a linearized variable $e_{2,hik}$.

This form of the variance estimator allows for the following degrees of freedom approximation, which is analogous to equation (5):

$$\hat{d}\{\hat{V}, \hat{\Delta}\} = \frac{3\{\hat{V}(\hat{\Delta})\}^2}{\sum_{h=1}^{L} \{\tilde{\Delta}_{h1} - \tilde{\Delta}_{h2}\}^4} - 2 \qquad (7)$$

5. Numerical Comparisons with BRR

We used data from the 1996 panel public use files of SIPP for numerical comparisons. These files have 105 pseudo-strata. Our BRR estimator used the replicate weights currently provided to public users, which contains an array of 108 replicate weights for each person within each cross-section. Since the number of replicate weights is greater than the number of pseudo-strata, there is no expected loss in degrees of freedom between the linearization and replication variance estimators.

The comparisons were generally favorable, showing the linearization and replicate standard errors to be quite close in most situations. For example, Table 5 displays a comparison for a typical statistic of interest. We estimated the average Social Security payment per recipient for the months of November 1996 and November 1997. We were interested not only in this ratio of two population totals (total payments over number of recipients), but also their change between the two years. Social Security recipiency was chosen because of its high correlation with age, which is controlled for in post-stratification. The variances of Social Security statistics are substantially reduced by post-stratification.

As Table 5 indicates, the linearization standard errors are in accord with the BRR standard errors. The largest departure is for the number of Social Security recipients in November 1996, which is 4.24% larger than the BRR standard error. Most of the linearization standard errors were within two percent of the corresponding BRR standard errors.

Another useful comparison that can be made is the ratio of the length of a confidence interval using linearization versus BRR. A 90% confidence interval using the linearization standard error is found from the *t* distribution using our degrees of freedom approximation, whereas the BRR confidence interval is found from the normal curve since degrees of freedom have not been computed with BRR. The comparisons demonstrate that the use of the linearization standard error and calculated degrees of freedom would have little impact on the length of a 90% confidence interval for these statistics.

Alternatively, Table 6 demonstrates that not all statistics will necessarily have large degrees of freedom. We calculated participation rates for five programs in the three west coast states. The proportion of the civilian non-institutional population of Oregon with health insurance in November 1996, for example, has only an estimated two degrees of freedom.

Finally, Table 7 displays estimated quartiles for the distribution of total household incomes for the month of November 1996. The linearization estimate of standard error was less than that of the BRR standard error for each quartile.

References

Deville, J. (1999), “Variance Estimation for Complex Statistics and Estimators: Linearization and the Residual Techniques,” Survey Methodology, Vol. 25, No. 2, pp 193-203.

Deville, J., Sarndal, C.E., and Sautory, O. (1993), “Generalized Raking Procedures in Survey Sampling,” Journal of the American Statistical Association, Vol 88, Issue 423, pp 1013-1020.

Eltinge, J.L., and Jang, D.S. (1996) “Stability Measures for Variance Component Estimators Under a Stratified Multistage Design,” Survey Methodology, 22, 157-165.

Fay, R.E. (1984), “Some Properties of Estimators of Variance Based on Replication Methods,” Proceedings of the Section on Survey Research Methods, American Statistical Association, Washington, DC, pp. 495-500.

Fay, R.E. (1989), “Theory and Application of Replicate Weighting for Variance Calculations,” Proceedings of the Section on Survey Research Methods, American Statistical Association, Washington, DC, pp. 212-217

Roberts, G., Kovacevic, M., Mantel, H., Phillips, O. (2001), “Cross-sectional Inference Based on Longitudinal Surveys: Some Experiences with Statistics Canada Surveys,” Federal Committee for Statistical Methods Conference, Washington, DC, October 2001, Working Paper 34, Part 4, pp. 25-34.

Stukel, D. M., Hidiroglou, M. A., and Sarndal, C. E. (1996) “Variance Estimation for Calibration Estimators: A Comparison of Jackknifing Versus Taylor Linearization,” Survey Methodology, 22, 117-125.

U. S. Census Bureau (2001), *Survey of Income and Program Participation Users’ Guide, 3rd Ed.* Washington, DC: U. S. Census Bureau.

Table 5
Average Social Security Payments

	Estimate	DF	SE (Lin)	SE (BRR)	% Diff	CI Ratio	CV (Lin)
Nov. 1996							
Social Security Payments ($1000)	23,856,013	118	177,365	173,659	2.13	1.03	0.007
Number of Recipients	37,478,979	78	179,784	172,480	4.24	1.05	0.005
SS Payments/Recipient ($)	636.52	91	3.209	3.156	1.68	1.03	0.005
Nov. 1997							
Social Security Payments ($1000)	24,655,365	88	164,588	163,394	0.73	1.02	0.007
Number of Recipients	38,054,239	70	188,863	181,563	4.02	1.05	0.005
SS Payments/Recipient ($)	647.90	102	2.959	3.010	-1.69	0.99	0.005
Change: Nov. 1996 to Nov. 1997							
Social Security Payments ($1000)	799,352	92	134,962	135,698	-0.54	1.00	0.169
Number of Recipients	575,261	108	163,035	161,601	0.89	1.02	0.283
SS Payments/Recipient ($)	11.39	48	2.821	2.858	-1.30	1.01	0.248

Table 6
November 1996 Program Participation Rates in Three States

	Estimate	DF	SE (Lin)	SE (BRR)	% Diff	CI Ratio	CV (Lin)
California							
Medicaid	0.151	40	0.0050	0.0053	-5.44	0.97	0.033
Health Insurance	0.612	26	0.0081	0.0082	-0.73	1.03	0.013
Social Security	0.117	31	0.0041	0.0041	-1.09	1.02	0.035
WIC	0.027	21	0.0019	0.0019	-1.28	1.03	0.070
Food Stamps	0.088	30	0.0041	0.0042	-3.34	1.00	0.046
Oregon							
Medicaid	0.096	4	0.0210	0.0211	-0.47	1.29	0.220
Health Insurance	0.743	2	0.0215	0.0226	-4.87	1.69	0.030
Social Security	0.134	6	0.0124	0.0133	-6.77	1.10	0.099
WIC	0.018	4	0.0071	0.0074	-4.05	1.24	0.411
Food Stamps	0.069	7	0.0177	0.0176	0.57	1.16	0.255
Washington							
Medicaid	0.126	4	0.0117	0.0121	-3.31	1.25	0.096
Health Insurance	0.732	7	0.0120	0.0134	-10.45	1.03	0.018
Social Security	0.137	16	0.0090	0.0133	-32.33	0.72	0.097
WIC	0.014	6	0.0040	0.0038	5.26	1.24	0.271
Food Stamps	0.079	7	0.0111	0.0118	-5.93	1.08	0.149

Table 7
November 1996 Total Household Income Quartiles

	Estimate	DF	SE (Lin)	SE (BRR)	% Diff	CI Ratio	CV (Lin)
Quartile							
0.25	1,727	117	14.20	16.20	-12.35	0.88	0.008
0.50	3,234	117	19.80	22.77	-13.04	0.88	0.006
0.75	5,248	126	32.01	34.12	-6.18	0.95	0.006