SAMPLE-DEPENDENT ESTIMATION FOR THE NATIONAL SURVEY OF FAMILY GROWTH

Dwight K. French and E. Earl Bryant, National Center for Health Statistics

Introduction

The first cycle of the National Survey of Family Growth (NSFG) was conducted by the National Center for Health Statistics under a contractual arrangement with the National Opinion Research Center (NORC) of the University of Chicago between July 1973 and February 1974. The survey was based on a rather complex multi-stage area probability sample of about 10,000 women 15-44 years of age, who had ever been married, or who had never been married but had natural-born children currently living in the household. Personal interviews were conducted to collect information about fertility, family planning and maternal and child health practices. The purpose of this paper is to present the methodology and results of a study to determine the best estimator for the survey based on sample results, a technique known as sample-dependent estimation.

The choice of estimators for the NSFG was between a simple inflation-type estimator based on reciprocals of selection probabilities and a post-stratified estimator using variable bench marks.

Post-stratified estimators are in common use for national surveys which cover the general population, since census-type information needed for bench marks is readily available. For surveys such as the Health Examination Survey and the Current Population Survey, it is known that post-stratification by age, race and sex produces more precise results than are attainable by a simple inflation estimator.[1]

Unfortunately, current census counts for the NSFG target population were not available for the time period that the survey was conducted. It was possible, however, to obtain estimates based on the Bureau of the Census's Current Population Survey (CPS), which collects information on a sample of about 50,000 households each month. Based on our knowledge of the efficiency of post-stratified estimators using census bench marks, it seemed reasonable that a large part of the gain over simple inflation estimators could be achieved using estimates of the bench marks if the estimates were precise enough. Furthermore, the use of large samples reduces the risk of making a wrong decision among alternative sample-dependent estimators.[2]

The criterion for choosing between the two estimators was based on the magnitude of their respective sampling variances. Values of over 200 statistics were calculated using both estimators, and variances were computed using identical balanced half-sample replication procedures. If poststratification decreased the variances of many estimates and did not cause large increases in variance for more than a few, the poststratified estimator would be selected; otherwise, the simple inflation estimator would be used.

Alternative NSFG Estimators

Let $Y = Y_1 + Y_2$ represent an aggregate parameter for ever-married women 15-44 years of age (Y_1) and never-married women in the same age group with natural-born children present in the home at the time of the interview (Y_2). Alternative estimators of Y considered for the NSFG were

$$Y'_{NP} = Y'_1 + Y'_2 \quad \text{and} \qquad (1)$$

$$Y'_P = Y''_1 + Y'_2 \quad \text{where} \qquad (2)$$

Y'_1 and Y'_2 are simple inflation estimators and

$$Y''_1 = \sum_{\alpha=1}^{12} Y'_{\alpha 1} \frac{X^*_{\alpha 1}}{X'_{\alpha 1}} \quad \text{is a poststratified} \qquad (3)$$

estimator.

The $X'_{\alpha 1}$ represent NSFG estimates of ever-married women in each of 12 age-race classes, and the $X^*_{\alpha 1}$ are the corresponding CPS estimates.

Computation of CPS Estimates of Ever-Married Women

Since the NSFG was conducted between July 1973 and February 1974, population values were needed for September 1973, the approximate mid-point of the survey. One possibility was to adjust the March 1973 CPS estimates of ever-married women. Although the total CPS sample is very large, the number of women in some of the desired age-race groups, especially for black women, would be relatively small. Since the CPS sample design allows an accumulation of the sample over time, it is possible to improve the precision of estimates by combining several CPS monthly samples. For the NSFG, two estimates of ever-married women were computed for each of 12 age-color groups using CPS data collected in March of 1970, 1971, 1972, and 1973. One estimator was the average of the four CPS values; the other was a predicted value based on simple linear regression. The predicted value was used if the regression coefficient was significantly different from zero. Otherwise, the average was used. The estimates and their approximate relative standard errors are shown in Table 1.

Variance Computation for the Two Estimators

Variances for Y'_P and Y'_{NP} were based on the same set of 48 balanced half-sample replicates. Y'_{kNP}, the nonpoststratified estimate for the k^{th} half-sample, was calculated in exactly the same way as Y'_{NP}, except that all records were given an additional weight of 2 to compensate for the selection of the half-sample. Y'_{kP}, the corresponding poststratified half-sample estimate, was calculated in the same way as Y'_{kNP} except that only records for single women were given the additional weight. The poststratification process inflated the estimate for ever-married women to its proper size. Given estimates Y'_{kNP} and Y'_{kP} from each half-sample,

Table 1. Population Estimates Based on CPS Data, and Their Relative Standard Errors

Age	Black Population	Black Rel. Std. Error (%)	Other Population	Other Rel. Std. Error (%)
15-19	115,000	5.2	993,000	1.9
20-24	564,000	2.7	4,967,000	0.8
25-29	622,000	2.4	6,153,000*	1.9
30-34	602,000	2.5	5,371,000*	1.8
35-39	575,000	2.6	4,751,000	0.8
40-44	593,000	2.5	4,940,000*	0.6

*predicted values

the variance Y'_P was estimated by

$$s^2_{Y'_P} = \frac{1}{48} \sum_{k=1}^{48} (Y'_{kP} - Y'_P)^2 \qquad (4)$$

and the variance of Y'_{NP} was estimated by

$$s^2_{Y'_{NP}} = \frac{1}{48} \sum_{k=1}^{48} (Y'_{kNP} - Y'_{NP})^2. \qquad (5)$$

Additional Variance of Y'_P Due to Variance in the $X^*_{\alpha 1}$

The half-sample replication procedure described in the previous section omitted one obvious source of variability in the poststratified estimator. Because each of the 48 half-sample estimates of characteristic Y for ever-married women in the 12 age-color classes was poststratified to the same set of CPS bench marks, the procedure implicitly assumed that the CPS values were known constants. In fact, these values are estimates, subject to sampling variability.

Because the NSFG was designed to interview only one woman from each sample household, estimates for ever-married and never-married women are essentially uncorrelated, as are estimates for women in different age-race classes; i.e.,

$$\text{Cov}\,(Y'_1, Y'_2) = 0$$

and

$$\text{Cov}\,\left(Y'_{\alpha 1} \frac{X^*_{\alpha 1}}{X'_{\alpha 1}},\; Y'_{\beta 1} \frac{X^*_{\beta 1}}{X'_{\beta 1}}\right) = 0, \quad \alpha \neq \beta$$

Therefore, the variance of Y'_P can be expressed as

$$\sigma^2_{Y'_P} = \sum_{\alpha=1}^{12} \text{Var}\,\left(Y'_{\alpha 1} \frac{X^*_{\alpha 1}}{X'_{\alpha 1}}\right) + \text{Var}\,\left(\sum_{\alpha=1}^{12} Y'_{\alpha 2}\right) \qquad (6)$$

The second term of (6) is properly estimated by the half-sample procedure, and will be omitted from the remainder of the discussion.

The approximate sampling variance of

$$Y''_{\alpha 1} = \frac{Y'_{\alpha 1}}{X'_{\alpha 1}} X^*_{\alpha 1},$$

the estimator for an individual α class, is[3]

$$\sigma^2_{Y'_{\alpha 1}} = \left(\frac{Y_{\alpha 1}}{X_{\alpha 1}}\right)^2 (X_{\alpha 1})^2 \left[\frac{\text{Var}\,(Y'_{\alpha 1}/X'_{\alpha 1})}{(Y_{\alpha 1}/X_{\alpha 1})^2} + \frac{\text{Var}\,(X^*_{\alpha 1})}{(X_{\alpha 1})^2} + \frac{2\,\text{Cov}\,\{(Y'_{\alpha 1}/X'_{\alpha 1}),(X^*_{\alpha 1})\}}{(Y_{\alpha 1}/X_{\alpha 1})(X_{\alpha 1})}\right] \qquad (7)$$

The last term inside the brackets is zero since the NSFG and the CPS are independent surveys.

Substituting $Y'_{\alpha 1}$, $X'_{\alpha 1}$, and $X^*_{\alpha 1}$ into (7) yields the estimate:

$$\hat{\sigma}^2_{Y''_{\alpha 1}} = \left(\frac{Y'_{\alpha 1}}{X'_{\alpha 1}}\right)^2 (X^*_{1}) \left[\frac{\hat{\text{Var}}\,(Y'_{\alpha 1}/X'_{\alpha 1})}{(Y'_{\alpha 1}/X'_{\alpha 1})^2} + \frac{\text{Var}\,(X^*_{\alpha 1})}{(X^*_{\alpha 1})^2}\right]$$

or

$$\hat{\sigma}^2_{Y''_{\alpha 1}} = (X^*_{\alpha 1})^2\, \hat{\text{Var}}\,(Y'_{\alpha 1}/X'_{\alpha 1}) + \left(\frac{Y'_{\alpha 1}}{X'_{\alpha 1}}\right)^2 \sigma^2_{X^*_{\alpha 1}} \qquad (8)$$

The first term of (8) has been computed by the replication procedure; the second term is the component that must be added on. The computation for estimates of Y for individual α classes is straightforward, since $Y'_{\alpha 1}$, $X'_{\alpha 1}$ and $\sigma^2_{X^*_{\alpha 1}}$ are all known for each α class.

For estimates of Y in any population group G covering more than one α class (such as national estimates, estimates for all eligible black females, or estimates for all eligible females age 20-24) application of (6) and (8) yields

$$\sigma^2_{Y'_P} = \sum_{\alpha \in G} (X^*_{\alpha 1})^2\, \sigma^2_{(Y'_{\alpha 1}/X'_{\alpha 1})} + \sum_{\alpha \in G} \left(\frac{Y'_{\alpha 1}}{X'_{\alpha 1}}\right)^2 \sigma^2_{X^*_{\alpha 1}} \qquad (9)$$

Again, the first summation in (9) is measured by the half-sample variance estimator, while the

Table 2. Poststratified and Nonpoststratified Estimates of the Number of Currently Married U.S. Women, Their Relative Standard Errors, and the Percent Reduction in RSE due to the Poststratified Estimator, by Wife's Religion and Age, National Survey of Family Growth, 1973.

Wife's Religion and Age	Y'_{NP} (1)	RSE of Y'_{NP} (2)	Y'_{P} (3)	RSE of Y'_{P} Excluding CPS Component (4)	RSE of Y'_{P} Including CPS Component (5)	Percent Reduction in RSE: $\frac{(2)-(5)}{(2)}$ (6)
Catholic						
All Ages	7,382,079	.0609	7,661,129	.0575	.0578	5.09
15-19	264,354	.1462	269,556	.1114	.1128	22.85
20-24	1,218,068	.0819	1,319,352	.0763	.0768	6.23
25-29	1,595,694	.0736	1,737,304	.0606	.0636	13.59
30-34	1,599,122	.0928	1,585,113	.0788	.0804	13.36
35-39	1,465,395	.0887	1,393,582	.0814	.0817	7.89
40-44	1,239,446	.0837	1,356,222	.0798	.0801	4.30
Jewish						
All Ages	433,130	.1146	449,013	.1145	.1147	-0.09
15-19	8,391	.7093	8,695	.7173	.7175	-1.16
20-24	30,868	.3721	33,499	.3796	.3797	-2.04
25-29	102,139	.1869	111,566	.1820	.1830	2.09
30-34	74,593	.2434	73,887	.2396	.2401	1.36
35-39	112,247	.2362	106,722	.2193	.2195	7.07
40-44	104,872	.2321	114,644	.2341	.2342	-0.90
Protestant						
All Ages	16,759,541	.0367	17,297,274	.0241	.0248	32.43
15-19	667,533	.0993	662,219	.0607	.0633	36.25
20-24	3,050,430	.0643	3,255,852	.0347	.0357	44.48
25-29	3,586,495	.0467	3,872,671	.0277	.0337	27.84
30-34	3,391,181	.0491	3,349,635	.0345	.0381	22.40
35-39	3,191,781	.0519	3,017,746	.0366	.0373	28.13
40-44	2,872,121	.0492	3,139,151	.0299	.0305	38.01
Other-None						
All Ages	1,188,020	.0822	1,238,770	.0634	.0637	22.51
15-19	85,805	.1996	87,255	.2067	.2075	-3.96
20-24	318,845	.1145	340,689	.1132	.1135	0.87
25-29	314,563	.1292	341,180	.1117	.1133	12.31
30-34	241,614	.1843	239,690	.1685	.1693	8.14
35-39	120,273	.1530	113,626	.1481	.1483	3.07
40-44	106,920	.1739	116,330	.1822	.1823	-4.83

second summation is not. A problem for these estimates is that the $Y'_{\alpha 1}$ are not readily available. One method of approximating the second sum is to assume that $Y_{\alpha 1} = C \cdot X_{\alpha 1}$ for all α. Under this assumption, a natural estimate of $Y'_{\alpha 1}$ is

$$Y'_{\alpha 1} = \frac{Y'_P \cdot X'_{\alpha 1}}{\sum_{\alpha \varepsilon G} X'_{\alpha 1}} \qquad (10)$$

and the extra component of variance is given by

$$\sum_{\alpha \varepsilon G} \left(\frac{Y'_{\alpha 1}}{X_{\alpha 1}}\right)^2 \sigma^2_{X^*_{\alpha 1}} = \sum_{\alpha \varepsilon G} \left[\frac{(Y'_P)^2 \; (X'_{\alpha 1})^2}{(\sum_\alpha X'_{\alpha 1})^2}\right] \cdot \left[\frac{\sigma^2_{X^*_{\alpha 1}}}{(X'_{\alpha 1})^2}\right]$$

$$= \sum_{\alpha \varepsilon G} \left[\frac{(Y'_P)^2}{(\sum_\alpha X'_{\alpha 1})^2}\right] \cdot \sigma^2_{X^*_{\alpha 1}}$$

$$= \frac{(Y'_P)^2}{(\sum_\alpha X'_{\alpha 1})^2} \sum_{\alpha \varepsilon G} \sigma^2_{X^*_{\alpha 1}} \qquad (11)$$

which can be computed.

Results

Simple inflation and poststratified estimates of the number of currently married U.S. women and the relative standard errors (RSE) of both estimates were computed for more than 200 subdomains of the population. Tabulations were made by race, wife's education, husband's education, family income, wife's religion, ethnic origin, labor force status and working status, each cross-classified by the wife's age. Table 2 shows the statistics produced for one cross-classification, namely wife's religion and wife's age. The poststratified estimator was better than the inflation estimator for more than 80 percent of the domains. The improvement in precision was 20 percent or better for almost two-fifths of the domains (Table 3).

The percent reductions in RSE for the tabulations by wife's education, income, religion and working status were each tested by the Wilcoxon Signed-Rank Test. The null hypothesis for each table was that the two estimators were equally precise; i.e., the average percent reduction in RSE due to the poststratified estimator was 0. The null hypothesis was rejected at the two-tailed .01 level of significance for all four tables. The remaining four tables generally showed similar results, which was expected because wife's education and husband's education are correlated, labor force status and working status are highly correlated, and race and ethnic origin age groups closely correspond to the poststratification classes. The one exception to the pattern occurred for persons of Spanish origin. The RSE of the poststratified estimator was higher for five of the six age groups, but never more than 6 percent higher.

As a result of these findings, the poststratified estimator was adopted for the NSFG.

References

1. Simmons, Walt R. and Bean, Judy A., Impact of Design and Estimation Components on Inference, *New Developments in Survey Sampling,* Wiley--InterScience, New York, 1969.

2. Dalenius, Tore and Ziegler, Walter Joh. (Ed.), Sample Dependent Estimation in Survey Sampling. *Contributions to Applied Statistics dedicated to Arthur Linder,* Birkhauser Verlag, 1976.

3. Hansen, Morris H., Hurwitz, William N. and Madow, William G., *Sample Survey Methods and Theory,* Vol. 1, John Wiley & Sons, N.Y., 1953, pg. 513.

Table 3. Distribution of the percent reduction in relative standard error of the simple inflation estimator by use of poststratified estimator. National Survey of Family Growth, 1973.

		Percent Reduction*						
	Total	-10 to 0	0-9	10-19	20-29	30-39	40-49	50+
Number of estimates	215	35	55	43	32	18	13	19
Percent of estimates	100.0	16.3	25.6	20.0	14.9	8.4	6.0	8.8

$$\text{*Percent Reduction} = \frac{RSE\ (Y'_{NP}) - RSE\ (Y'_P)}{RSE\ (Y'_{NP})}$$