Survey Methodology, June 1997
Vol. 23, No. 1, pp. 73-75
Statistics Canada

A Modified Random Groups Standard Error Estimator

WILLARD C. LOSINGER[1]

ABSTRACT

The standard error estimation method used for sample data in the U.S. Decennial Census from 1970 through 1990 yielded irregular results. For example, the method gave different standard error estimates for the "yes" and "no" response for the same binomial variable, when both standard error estimates should have been the same. If most respondents answered a binomial variable one way and a few answered the other way, the standard error estimate was much higher for the response with the most respondents. In addition, when 100 percent of respondents answered a question the same way, the standard error of this estimate was not zero, but was still quite high. Reporting average design effects which were weighted by the number of respondents that reported particular characteristics magnified the problem. An alternative to the random groups standard error estimate used in the U.S. census is suggested here.

KEY WORDS: Census; Variance estimation; Random groups; Design effect.

1. INTRODUCTION

During the 1990 Decennial Census, all respondents were asked to provide information on certain data items (called 100-percent data). Most respondents provided this information on the census short form. In addition, a systematic sample (ranging from one-eighth to one-half, but averaging about one-sixth) of respondents provided information for more data items (sample data) on the census long form.

Rather than providing standard error estimates for each published sample data estimate, the Census Bureau published tables of generalized design effects. For any sample data estimate, data users were instructed to create a standard error assuming simple random sampling (either using the standard formula or from a table) and a one-in-six sampling rate. Then, data users were to multiply this standard error by a generalized design effect (provided in another table). The table of generalized design effects listed design effects by data item type and percent of persons or housing units included in the sample (Table 1 provides the design effects published for 1990 U.S. census sample data for Vermont). For example, for all published sample estimates that dealt with occupation, a data user would find four generalized design effects for occupation: one for each of four sampling rate categories for persons in the report. To estimate the standard error for the number of teachers in a published report, a data user would multiply the simple-random-sampling standard error (assuming a one-in-six sampling rate, derived from the formula or table of standard errors) by the design effect for occupation data items for the reported sampling rate. The data user could then use the estimated number of teachers and standard error to construct a confidence interval. More details on the use of the table of design effects are available in the Accuracy of the Data section for any sample data product (U.S. Bureau of the Census 1993, for example).

2. ESTIMATION OF STANDARD ERRORS

A random-groups approach was used to estimate standard errors for the census sample data. The United States was divided into just over 60,000 distinct areas (called weighting areas--areas for which sample weights were derived). For each weighting area, sample units (a sample unit being either a housing unit or a person residing in a group quarter) were assigned systematically among 25 random groups. Thus, it was thought that each random group so formed met the requirement of having approximately the same sampling design as the parent sample (Wolter 1985).

For each of the 25 random groups, a separate estimate of the total for each of 1,804 sample data items was computed by multiplying the weighted count for the sample data item within the random group by 25. For each data item for which the total number of people with a particular characteristic was estimated from the sample data, the random-groups standard error estimate was then computed from the 25 different estimates of the total from the random groups:

$$S_{RG} = \sqrt{(1 - n/N)\sum_{i=1}^{25} \frac{(\hat{Y}_i - \hat{Y})^2}{24}}$$

where n represents the unweighted number of persons in the sample within the weighting area; N represents the census count of persons within the weighting area; $\hat{Y}_i$ represents the estimate of the total for the data item achieved by multiplying the weighted count for the data item within the i-th random group by 25; and $\hat{Y}$ is the weighted count for the data item (*i.e.*, the sample estimate) within the weighting area.

[1] Willard C. Losinger, U.S. Department of Agriculture: APHIS:VS, CEAH, 555 South Howes Street, Suite 200, Fort Collins, CO 80521, U.S.A.

Table 1
Design Effects Published for 1990 U.S. Census
Sample Data for Vermont

Characteristic	Percent of persons or housing units in sample			
	< 15%	15 - 30%	30 - 45%	≥ 45%
Age	1.2	1.0	0.6	0.5
Sex	1.2	1.0	0.6	0.5
Race	1.2	1.0	0.6	0.5
Hispanic origin (of any race)	1.2	1.0	0.6	0.5
Marital status	1.1	0.9	0.6	0.5
Household type and relationship	1.2	1.0	0.6	0.5
Children ever born	2.5	2.2	1.3	1.2
Work disability and mobility limitation status	1.2	1.0	0.6	0.5
Ancestry	1.8	1.5	1.0	0.8
Place of birth	1.9	1.6	1.0	0.9
Citizenship	1.7	1.4	1.0	0.8
Residence in 1985	1.9	1.7	1.0	0.9
Year of entry	1.3	1.0	0.6	0.5
Language spoken at home and ability to speak English	1.6	1.3	0.9	0.7
Educational attainment	1.3	1.1	0.6	0.5
School enrollment	1.6	1.4	1.0	0.8
Type of residence (urban/rural)	1.7	1.7	1.4	1.4
Household type	1.2	1.0	0.6	0.5
Family type	1.1	1.0	0.6	0.5
Group quarters	1.0	1.1	0.9	0.8
Subfamily type and presence of children	1.1	0.9	0.5	0.5
Employment status	1.2	1.0	0.6	0.5
Industry	1.2	1.0	0.6	0.5
Occupation	1.2	1.0	0.6	0.5
Class of worker	1.2	1.0	0.6	0.5
Hours per week and weeks worked in 1989	1.4	1.2	0.7	0.6
Number of workers in family	1.3	1.1	0.7	0.6
Place of work	1.4	1.2	0.8	0.6
Means of transportation to work	1.4	1.2	0.7	0.6
Travel time to work	1.3	1.1	0.6	0.5
Private vehicle occupancy	1.4	1.2	0.7	0.6
Time leaving to go to work	1.2	1.0	0.6	0.5
Type of income in 1989	1.3	1.1	0.6	0.5
Household income in 1989	1.1	1.0	0.6	0.5
Family income in 1989	1.1	1.0	0.6	0.5
Poverty status in 1989 (persons)	1.5	1.2	0.7	0.7
Poverty status in 1989 (families)	1.1	0.9	0.5	0.5
Armed forces and veteran status	1.4	1.1	0.7	0.6

Source: U.S. Bureau of the Census (1993). 1990 Census of Population: Social and Economic Characteristics: Vermont. Report Number 1990 CP-2-47. Page C-11.

A standard error based upon simple random sampling and a one-in-six sampling rate was computed thus:

$$S_{SRS} = \sqrt{5\,\hat{Y}\,(1 - \hat{Y}/N)}$$

developed from standard formulas displayed in Cochran (1977).

For each data item within the weighting area, a design effect was computed as the ratio of the S_{RG} to S_{SRS}:

$$F = \frac{S_{RG}}{S_{SRS}}.$$

For a state report of sample data, the design effects for each data item were averaged across the weighting areas in the state. Then, a generalized design effect for each data item type (for example, all data items that dealt with occupation) was computed. The generalized design effect was weighted in favor of data items that had higher population estimates. Details on most of the procedures followed are available in a Census Bureau document (U.S. Bureau of the Census 1991). The same basic method was also used for sample data products in both the 1970 and 1980 census.

3. A HYPOTHETICAL EXAMPLE OF RANDOM GROUPS

Table 2 presents a hypothetical example of data that might have arisen from the random-groups method. For a weighting area in Vermont, weighted counts of whites and blacks are listed for the 25 random groups. In this hypothetical weighting area, there are no persons of other race. The standard errors assuming simple random sampling are the same for whites and blacks (as one would expect for a binomial variable). However, S_{RG} is much higher for the estimate of whites than the estimate of blacks. And, the design effect is nearly five times higher for the estimate of whites than the estimate of blacks. Since the generalized design effect computed for groups of data items was weighted in favor of data items that had higher population estimates, the generalized design effect computed for race for the state of Vermont was quite high.

Data on race were frequently included in 1990 U.S. census sample data products. Because race was asked of every census respondent (*i.e.*, it was a census 100-percent data item), and because the weighting process used by the Census Bureau effectively forced the sample estimates by race to match the 100-percent Census counts by race, the standard errors for estimates of race probably should have been considered to be zero. However, generalized design effects were still published by race, although set to arbitrary constants for all reports (rather than as computed by this method).

4. A MODIFIED APPROACH TO THE RANDOM GROUPS METHOD

A slight modification of the random groups method (essentially applying a ratio-estimation technique) can achieve much more satisfactory results in the estimation of standard errors. Rather than using $\hat{Y}_i$ as defined above for the estimate of the total for the i-th random group, one could instead use

$$\hat{Y}_i = N X_i / W_i$$

Table 2
Hypothetical example of data that could have resulted from the Random Groups method used to estimate standard errors for census sample data.
For a weighting area in Vermont, people are asked their race.
A few (110) are black; most (2,518) are white.
A sampling rate of one-in-six is assumed ($N = 2{,}628$, $n = 438$).

Random Group	Weighted count of blacks*	Weighted count of whites*	Total weighted population count #
1	10	90	100
2	0	100	100
3	0	110	110
4	0	140	140
5	5	70	75
6	8	50	58
7	12	103	115
8	20	60	80
9	0	65	65
10	0	100	100
11	0	125	125
12	0	130	130
13	10	90	100
14	0	100	100
15	0	110	110
16	0	140	140
17	5	70	75
18	8	52	60
19	12	103	115
20	20	160	180
21	0	65	65
22	0	100	100
23	0	125	125
24	0	130	130
25	0	130	130
Sum of weighted counts ($\hat{Y}$)	110	2,518	2,628
S_{RG}	145.98	687.96	
S_{SRS}	22.96	22.96	
F	6.36	29.96	

* The first 25 figures in this column represent X_i for the i-th random group under the modified random groups method. Multiplying the figure by 25 yields $\hat{Y}_i$ for the random groups method employed by the U.S. Bureau of the Census.

\# The first 25 figures in this column represent W_i under the modified random groups method.

where X_i represents the weighted count for the data item within the i-th random group, W_i is the weighted count of all persons in the i-th random group, and N represents the census count of persons in the weighting area. The modified random groups standard error estimate is then

$$S_L = \sqrt{(1 - n/N) \sum_{i=1}^{25} \frac{(\hat{L}_i - \hat{Y})^2}{24}}.$$

Using this method, S_L is 160.78 for both blacks and whites in the hypothetical weighting area of Table 1 (close to the value of S_{RG} for blacks). In this case, the requirement for standard error estimates for both responses for a binomial variable to be identical is met. Moreover, if all sample units have the same response for some variable, S_L becomes zero, whereas S_{RG} only becomes zero when each random group has the same weighted count.

This modified standard error estimation procedure could be useful for researchers who do not have access to any of the many computer programs now available for computing estimates from sample data (such as SUDAAN, STATA, PC-CARP, VPLX, *etc.*). In addition, the U.S. Bureau of the Census ought to consider modifying its approach for estimating standard errors for sample data from the 2000 census. Moreover, with the U.S. Bureau of the Census' current emphasis on quality management, the U.S. Bureau of the Census may wish to poll users of sample data products to determine how useful the presentation of standard errors (through design effects) was to them, and involve a number of the data users in improving the presentation of standard errors for the next census.

REFERENCES

COCHRAN, W.G. (1977). *Sampling Techniques* (third edition). New York: John Wiley & Sons.

WOLTER, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.

U.S. BUREAU OF THE CENSUS (1991). Computer Specifications for the 1990 Decennial Census Variance Estimation Operation. STSD Decennial Census Memorandum Series #Z-65.

U.S. BUREAU OF THE CENSUS (1993). Appendix C. Accuracy of the Data. Pp. C-1 to C-11 in 1990 Census of Population: Social and Economic Characteristics: Vermont. Bureau of the Census Document 1990 CP-2-47.