AN EVALUATION OF SURVEY DATA ANALYSIS DESIGN EFFECT ADJUSTMENTS

Rick L. Williams, Research Triangle Institute

1. INTRODUCTION

Most large national surveys employ a complex stratified multistage probability sample. Such sample designs allow economical data collection. However, they complicate data analysis since most standard statistical procedures implicitly assume simple random sampling from an infinite population. Classical variance estimates which do not account for the sample design (e.g., using pq/n for a proportion) may seriously under estimate the true variance in the presence of clustering and unequal probability selections. Several authors, such as Shah, Holt and Folsom (1977) and Fellegi (1980), have examined the consequences of ignoring the sample design when analyzing survey data. In general, they have shown that for highly clustered designs this leads to hypothesis tests which reject the null hypothesis too often.

The challenge to researchers is to properly account for the sample design when analyzing survey data. Articles by Koch, Freeman and Freeman (1975), and Freeman, Freeman and Brock (1977) have advocated a weighted least squares approach. This is an extension of the methodology developed by Grizzle, Starmer and Koch (1969) for categorical data analysis. This style of analysis requires two steps. First, for analyzing differences between subpopulations, estimates of the parameter of interest are calculated for the various subpopulations along with the corresponding variance-covariance matrix. Second, weighted least squares is applied to evaluate relevant hypotheses concerning subpopulation effects. During the first stage of the analysis, the parameters and their variance-covariance matrix are estimated in accordance with the sample design. The secondary analysis can then proceed in a more classical vein using weighted least squares or standard multivariate techniques.

Unfortunately, most researchers are ill-equipped to carry out the first step of the proceeding analysis strategy since virtually all traditional software available to them ignores the design structure inherent in sample survey data. Methods to alleviate this problem for tests of goodness of fit and independence in a two way table have been proposed by Rao and Scott (1979) and Fellegi (1980). Under their proposals, classical test statistics for these two problems are scaled by an average sample design effect.

This paper presents an empirical application and assessment of the Rao and Scott methodology using data from the National Assessment of Educational Progress (NAEP). Analyses were conducted first assuming a simple random sampling design and then accounting for the actual clustered NAEP design. Test statistics from these two analyses were compared to determine whether design effect adjustments to the simple random sampling test statistics are effective for NAEP data.

2. THEORETICAL BACKGROUND

To illustrate the design effect adjustment methodology consider the vector $P = (p_1, p_2, \ldots p_r)'$ where p_i is the proportion of students responding correctly to a particular exercise for student subgroup-i. Let $\hat{P}$ be an estimate of P calculated from the sample and assume that $\hat{P}$ is asymptotically multivariate normal with variance-covariance matrix V. Classical analysis methods, which assume simple random sampling, lead to the dispersion matrix $S = \text{diag}\{p_i(1-p_i)/n_i\}$. In this setting, the hypothesis for testing subgroup differences or contrasts may be stated as:

$$H_o: \; CP = \phi \text{ vs. } H_A: \; CP \neq \phi$$

where C is a matrix of d linearly independent contrasts and ϕ is the (d x 1) null vector. The usual test statistic ignoring the sample design is

$$X^2 = (\hat{CP})'[CSC']^{-1}(\hat{CP})$$

On the other hand, the appropriate test of this hypothesis accounting for the sample design is based on the Wald statistic

$$X^2_w = (\hat{CP})'[CVC']^{-1}(\hat{CP})$$

which is asymptotically distributed as a chi-squared with d degrees of freedom under H_o.

A result similar to that shown by Rao and Scott (1981) is that under the null hypothesis $H_o: \; CP = \phi$,

$$X^2 = \sum_{i=1}^{d} \lambda_i Z_i^2$$

where $Z_1, \ldots, Z_d$ are asymptotically independent N(0,1) random variables and $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d$ are the eigenvalues of $D = [CSC']^{-1}[CVC']$. Rao and Scott refer to these eigenvalues as "generalized design effects". Following Rao and Scott's suggestion for tests of independence, X^2 can be scaled by $\bar{\lambda} = (\Sigma\lambda_i)/d$ to bring the test more nearly into line with the Wald statistic X^2_w. Notice that this adjustment factor is dependent on the exact contrast being considered and requires full knowledge of the matrices S and V. Rao and Scott indicate that the average eigenvalue of $S^{-1}V$ may provide an adequate adjustment for a general contrast. In the cases that will be addressed in this paper, S is a diagonal matrix and the average eigenvalue of $S^{-1}V$ is also the average design effect of the subgroups.

3. EMPIRICAL INVESTIGATION

Initially, five NAEP exercises per age class were selected for analysis from the Year 09 Mathematics Assessment. Each item was recoded one for correct and zero for incorrect. An additional score was defined for each student as the proportion of the items analyzed on a package that the student answered correctly. This score was analyzed within each age class to form three mean scores for analysis.

Four domain or subgroup defining variables were also selected. These were, with their corresponding levels:

Sex	Race
Male	White
Female	Other

Type of Community (TOC)	Parental Education (PARED)
Extreme Rural	Not High School Graduate
Metro	High School Graduate
Other	Post High School

The ultimate goal of this study was to compare sample design based analyses of NAEP data with those assuming a simple random sample. This approach proceeds by first estimating a vector of domain statistics and its corresponding covariance matrix. Various hypotheses concerning this vector can then be evaluated using weighted least squares and large sample Wald statistics. Two vectors of domain means were formed for each of the 15 item scores and the three mean scores. The first vector contained 12 elements corresponding to the complete cross-classification of Race, Sex and Parents Education (PARED). The second vector was derived from the cross-classification of Sex, Type of Community (TOC) and PARED and was of length 18. For the 15 item scores, these vectors consisted of simple proportion correct p-values. Two covariance matrices were then estimated for each vector. One based upon the actual sample design and the other assuming a simple random sample of students. The covariance matrices were estimated using a Taylor's series linearization approach.

At this point several exercises were excluded from the study because their estimated covariance matrices were singular. For the Race*Sex*PARED cross-classification only one item was excluded. However, for the Sex*TOC*PARED cross-classification it was necessary to exclude five items.

A linear model was then fitted, via weighted least squares, to each of the remaining domain mean vectors. For the Race*Sex*PARED domain cross-classification vectors the model contained the main effects of Race and Sex, a linear effect of PARED and the four possible two- and three-way interactions among these three effects The Sex*TOC*PARED domain classification model had the same form except that TOC was substituted for Race. These models were fitted two ways -- one weighted with the design based covariance matrix and the other weighted with the simple random sampling covariance matrix. The lack of fit of each model and the significance of each effect in the model was then assessed. These tests are labeled one through eight in Tables 1 and 2.

In addition, nine other hypotheses were considered and are labeled nine through 17 in Tables 1 and 2. These hypotheses were tested via direct contrasts of the domain means. The tests labeled "average" (numbers 10, 11, 12 and 13) average the effect over the combined levels of the other two variables. On the other hand, the "nested" tests (numbers 14, 15, 16 and 17) test for all the indicated simple effects being simultaneously null over the combined levels of the other two variables.

Two test statistics were entertained for each hypothesis. The first test was a Wald statistic chi-squared based upon the actual NAEP sample design. A second Wald-like statistic was also calculated assuming a simple random sample of students and will be referred to as the simple random sampling chi-squared. These two test statistics were calculated for each hypothesis for 14 NAEP items and three mean scores for the Race*Sex*PARED cross-classification, as well as for 10 NAEP items plus three mean scores for the Sex*TOC*PARED cross-classification.

The design effects (DEFFs) for each domain p-value and mean score used in the analyses are summarized in Tables 3, 4, and 5. Each table presents the minimum, median, maximum and mean DEFFs for a particular NAEP item or mean score across the levels of the indicated domain defining cross-classification (i.e., Race*Sex*PARED or Sex*TOC*PARED). The design effects reported in these three tables are consistent with previous NAEP experience and tend to average around 1.4. Also, as discussed in section 2, the mean DEFF's given in the last column of each table are the exact quantities proposed by Rao and Scott (1981) and Fellegi (1980) for adjusting simple random sampling (SRS) based Wald statistic chi-squareds to reflect the effects of the sample design. These are the adjustment factors used in the subsequent discussion.

As was noted earlier two different methods of analyses or hypothesis testing often used by researchers was considered. The first fitted a linear model to the estimated domain statistics. Relevant hypotheses were then tested via contrasts of the estimated linear model parameters. The parameters were estimated weighting inversely proportional to the SRS covariance matrix of the domain statistics to obtain the SRS test statistics. Another set of parameter estimates was obtained by weighting by the inverse of the design based covariance matrix and the asymptotically correct test statistics were calculated. The second method of analysis evaluated hypotheses via direct contrasts of the domain statistics. Again this was first accomplished using the SRS covariance matrix to obtain the SRS test statistics, and was then repeated using the design based covariance matrix to obtain the asympotically correct tests. Results in the rest of this section will be presented separately for these two modes of analysis (i.e., contrasts of linear model coefficients and contrasts of cell means).

For each hypothesis test entertained in this portion of the investigation, the ratio of the SRS based test statistic to the asymptotically correct sample design based Wald statistic chi-squared was calculated. These ratios are another measure of the effect of the sample design and are referred to in the remaining tables as hypothesis test design effects. Two issues will be addressed by way of these test DEFFs. First, an indication of the ordinal relationship between the two test statistics will be sought. That is, does the SRS statistic tend to be generally smaller or larger than the design based chi-squared? Second, are the test DEFFs fairly constant, at least within an item or mean score? This second point is important

if a simple multiplicative adjustment to the SRS test statistics is to be successful. Tables 6, 7, 8, present a summary of the test DEFFs for each mean or item score for the indicated cross-classification. The minimum, median, maximum and mean test design effects are shown separately for linear model coefficient contrasts (test numbers 1 through 8 in Tables 1 and 2) and cell mean contrasts (test numbers 9 through 17 in Tables 1 and 2).

The most striking feature of these three tables is the extreme instability of the test DEFFs for linear model coefficients. In virtually every case the mean is far greater than the median, indicating a skewed distribution with a long right hand tail. It appears that adjusting the SRS test statistic for the linear model coefficient contrasts will not prove fruitful because of the extreme range they cover. This may result from using the SRS covariance matrix to estimate the linear model parameters for the SRS test statistic. This process does not properly account for the correlated nature of the domain statistics and leads to less precise estimates of the model coefficients. The theory presented in section 2 does not strictly apply in this situation. These results are included to illustrate the problems that arise when SRS is assumed. Conversely, Tables 6, 7, and 8 indicate that the cell mean contrast hypothesis test design effects tend to be more symmetrically distributed over a narrower range than their linear model counterparts. However, they still exhibit enough variation on both sides of unity to make a simple multipicative adjustment questionable.

As indicated earlier, theoretical considerations suggest that the mean design effects presented in Tables 3, 4 and 5 may provide serviceable adjustments to the SRS test statistics. This conclusion is drawn into question by comparing the standard mean DEFFs in these three tables with the average test DEFFs for cell mean contrasts in Tables 6, 7, and 8. Almost without exception the mean test DEFFs are less than their corresponding p-value DEFF average. In addition, the mean hypothesis test DEFFs are generally near unity or less while the standard mean DEFFs are generally much greater than unity. This implies that dividing the SRS test statistic by the mean design effect will produce a test that is generally much too conservative. In fact, the adjustment suggested by Rao and Scott (1981) or Fellegi (1980) is in the wrong direction for the examples presented here.

REFERENCES

Fellegi, I. P. (1980), "Approximate Tests of Independence and Goodness of Fit Based on Stratified Multistage Samples," Journal of the American Statistical Association, Volume 75, Number 370.

Grizzle, J. E., Starmer, C. F., Koch, G. G. (1969), "Analysis of Categorical Data by Linear Models," Biometrics, Volume 25.

Koch, G. G., Freeman, D. H., Freeman, J. L. (1975), "Strategies in the Multivariate Analysis of Data from Complex Surveys," International Statistical Review, Volume 43, Number 1.

Rao, J. N. K., Scott, A. J. (1981) "The Analysis of Categorical Data from Complex Sample Surveys: Chi-Squared Test for Goodness of Fit and Independence in Two-Way Tables," Journal of the American Statistical Association, Volume 76, Number 374.

"The work upon which this publication is based was performed pursuant to Grant NIE-G-80-0003 of the National Institute of Education. It does not, however, necessarily reflect the view of that agency."

Table 1. Hypothesis Tests for the Race*Sex*PARED Cross-Classification

Test Number	d.f.	Description
Linear Model Tests		
1	4	Lack of fit
2	1	Race
3	1	Sex
4	1	PARED linear
5	1	Race*Sex
6	1	Race*PARED linear
7	1	Sex*PARED linear
8	1	Race*Sex*PARED linear
Contrast Tests		
9	11	All cells equal
10	1	Average Race effect
11	1	Average Sex effect
12	2	Average PARED effect
13	1	Average PARED linear effect
14	6	Nested Race effect
15	6	Nested Sex effect
16	8	Nested PARED effect
17	4	Nested PARED linear effect

Table 2. Hypothesis Tests for the Sex*TOC*PARED Cross-Classification

Test Number	d.f.	Description
Linear Model Tests		
1	6	Lack of fit
2	1	Race
3	2	TOC
4	1	PARED linear
5	2	Sex*TOC
6	1	Sex*PARED linear
7	2	TOC*PARED linear
8	2	Sex*TOC*PARED linear
Contrast Tests		
9	17	All cells equal
10	1	Average Sex effect
11	2	Average TOC effect
12	2	Average PARED effect
13	1	Average PARED linear effect
14	9	Nested Sex effect
15	12	Nested TOC effect
16	12	Nested PARED effect
17	6	Nested PARED linear effect

Table 3. NAEP Item Design Effects for the Race * Sex * PARED Cross-Classification

NAEP Item	Minimum DEFF	Median DEFF	Maximum DEFF	Mean DEFF
NO222A	.79	1.23	3.08	1.48
NO227A	.80	1.36	1.94	1.40
NO305C	.62	1.39	1.93	1.35
NO323A	.59	1.27	1.67	1.14
TO105A	.91	1.50	2.84	1.63
TO110A	.56	1.26	2.38	1.43
TO203A	.99	1.72	2.29	1.66
TO223A	.69	1.13	2.32	1.28
TO224A	1.00	1.31	2.82	1.47
SO108A	.63	.94	1.99	1.11
SO117A	.61	1.17	2.44	1.23
SO121A	.39	1.09	3.71	1.37
SO206A	.72	1.25	3.44	1.40
SO225A	.59	.84	1.83	.99
Average	.71	1.25	2.48	1.35

Table 4. NAEP Item Design Effects for the Sex * TOC * PARED Cross-Classification

NAEP Item	Minimum DEFF	Median DEFF	Maximum DEFF	Mean DEFF
NO222A	.21	1.17	2.49	1.25
NO305C	.37	1.53	2.21	1.35
TO105A	.49	1.40	4.32	1.61
TO110A	.64	1.28	3.02	1.31
TO203A	.27	1.36	4.46	1.62
TO223A	.68	1.14	2.10	1.25
SO108A	.44	1.03	2.01	1.14
SO117A	.35	1.11	2.14	1.14
SO206A	.48	1.53	4.17	1.66
SO225A	.47	.93	2.37	1.04
Average	.44	1.25	2.93	1.34

Table 5. Mean Scores Design Effects

Model/Age	Minimum DEFF	Median DEFF	Maximum DEFF	Mean DEFF
RACE*SEX*PARED				
9-year-olds	.57	1.45	3.32	1.50
13-year-olds	.78	1.31	2.33	1.46
17-year-olds	.49	1.09	2.57	1.16
Average	.61	1.28	2.74	1.37
SEX*TOC*PARED				
9-year-olds	.80	1.52	3.47	1.66
13-year-olds	.59	1.50	3.57	1.66
17-year-olds	.75	1.30	2.61	1.45
Average	.71	1.44	3.32	1.59

Table 6. Hypothesis Test Design Effects by NAEP Item for the Race * Sex * PARED Cross-Classification

NAEP Item	Contrast of Linear Model Coefficients				Contrast of Cell Means			
	Minimum	Median	Maximum	Mean	Minimum	Median	Maximum	Mean
N0222A	.04	.82	5.42	1.41	.19	.74	1.81	.88
N0227A	.00	.57	900.26	112.96	.23	1.02	1.60	.88
N0305C	.09	.57	18.69	3.88	.62	1.33	2.40	1.38
N0323A	.00	.48	1.08	.57	.51	1.08	2.01	1.12
T0105A	.32	.99	15.73	4.02	.44	1.16	1.98	1.27
T0110A	.16	.63	1.72	.81	.56	1.18	2.18	1.19
T0203A	.10	.86	2.29	1.03	.53	1.51	2.21	1.50
T0223A	.49	5.10	284.87	45.21	.72	1.11	1.63	1.10
T0224A	.80	1.68	34.09	9.05	.65	1.10	2.41	1.27
S0108A	.03	.71	47.13	6.47	.55	.84	1.50	.93
S0117A	.19	.59	3.62	.97	.53	.75	1.75	1.00
S0121A	.00	.47	26.19	3.91	.60	.95	2.23	1.19
S0206A	.59	1.51	2.67	1.58	.59	1.10	2.09	1.12
S0225A	.34	.65	2.33	.87	.43	.92	1.09	.84
Average	.23	1.12	96.15	13.77	.51	1.06	1.92	1.12

Table 7. Hypothesis Test Design Effects by NAEP Item for the Sex * TOC * PARED Cross-Classification

NAEP Item	Contrast of Linear Model Coefficients				Contrast of Cell Means			
	Minimum	Median	Maximum	Mean	Minimum	Median	Maximum	Mean
N0222A	.48	4.28	55.14	11.51	.11	.48	2.82	.75
N0305C	.10	1.08	190.09	29.98	.19	.97	1.81	.89
T0105A	.04	.60	6.97	1.70	.13	.39	3.23	.98
T0110A	.37	.76	1.57	.80	.19	.55	3.41	.84
T0203A	.14	.44	3.93	.91	.27	1.08	1.84	.93
T0223A	.22	1.23	10.30	2.40	.45	.86	1.13	.77
S0108A	.02	.14	.64	.22	.10	.36	2.62	.73
S0117A	.46	.97	2.80	1.22	.03	.36	2.46	.70
S0206A	.11	.47	1.27	.54	.10	.64	1.27	.59
S0225A	.05	.75	2.98	.98	.23	.45	1.43	.60
Average	.20	1.07	27.57	5.03	.18	.61	2.20	.78

Table 8. Hypothesis Test Design Effects for Mean Scores

Model/Age	Contrast of Linear Model Coefficients				Contrast of Cell Means			
	Minimum	Median	Maximum	Mean	Minimum	Median	Maximum	Mean
Race*Sex*PARED								
9-year-olds	.11	.22	3.74	.85	.29	.91	1.67	1.00
13-year-olds	.09	1.86	7064.23	885.11	.59	1.19	2.23	1.26
17-year-olds	.00	.43	1.16	.56	.40	1.08	1.32	.89
Average	.07	.84	2356.38	295.51	.43	1.06	1.74	1.05
Sex*TOC*PARED								
9-year-olds	.23	.39	1.39	.55	.19	.62	2.53	.91
13-year-olds	.05	.50	1.96	.74	.17	.72	2.87	1.09
17-year-olds	.02	.65	223.55	28.54	.03	.50	1.27	.53
Average	.10	.51	75.63	9.94	.13	.61	2.22	.84