Journal of Modern Applied Statistical Methods

Volume 2 | Issue 2 Article 15

11-1-2003

Example Of The Impact Of Weights And Design Effects On Contingency Tables And Chi-Square Analysis

David A. Walker
Northern Illinois University, dawalker@niu.edu

Denise Y. Young
University of Dallas

Example Of The Impact Of Weights And Design Effects On Contingency Tables And Chi-Square Analysis

Journal of Modern Applied Statistical Methods
November, 2003, Vol. 2, No. 2, 425-432

1538 – 9472/03/$30.00

Example Of The Impact Of Weights And Design Effects On Contingency Tables And Chi-Square Analysis

David A. Walker
Educational Research and Assessment
Northern Illinois University

Denise Y. Young
Institutional Research
University of Dallas

Many national data sets used in educational research are not based on simple random sampling schemes, but instead are constructed using complex sampling designs characterized by multi-stage cluster sampling and over-sampling of some groups. Incorrect results are obtained from statistical analysis if adjustments are not made for the sampling design. This study demonstrates how the use of weights and design effects impact the results of contingency tables and chi-square analysis of data from complex sampling designs.

Key words: Design effect, chi-square, weighting

Introduction

Many large-scale data sets used in educational research are constructed using complex designs characterized by multi-stage cluster sampling and over-sampling of some groups. Common statistical software packages such as SAS and SPSS yield incorrect results from such designs unless weights and design effects are used in the analysis (Broene & Rust, 2000; Thomas & Heck, 2001). The objective of this study is to demonstrate how the use of weights and design effects impact the results of contingency tables and chi-square analysis of data from complex sampling designs.

Methodology

In large-scale data collection, survey research applies varied sample design techniques. For example, in a single-stage simple random sample with replacement (SRS), each subject in the study has an equal probability of being selected. Thus, each subject chosen in the sample represents an equivalent total of subjects in the population. More befitting, however, is that data collection via survey analysis often involves the implementation of complex survey design (CSD) sampling, such as disproportional stratified sampling or cluster sampling, where subjects in the sample are selected based on different probabilities. Each subject chosen in the sample represents a different number of subjects in the population (McMillan & Schumacher, 1997).

Complex designs often engender a particular subgroup, due to oversampling or selection with a higher probability, and consequently the sample does not reflect accurate proportional representation in the population of interest. Thus, this may afford more weight to a certain subgroup in the sample than would be existent in the population. As Thomas and Heck (2001) cautioned, “When using data from complex samples, the equal weighting of observations, which is appropriate with data collected through simple random samples, will bias the model’s parameter

This article is based upon work supported by the Association for Institutional Research (AIR), National Center for Education Statistics (NCES), and the National Science Foundation (NSF) through fellowship grants awarded to the authors to participate in the 2001 AIR Summer Data Policy Institute on Databases of NCES and NSF. Correspondence for this article should be sent to: David Walker, Northern Illinois University, ETRA Department, 208 Gabel, DeKalb, IL 60115, (815)-753-7886. E-mail him at: dawalker@niu.edu.

estimates if there are certain subpopulations that have been oversampled” (p. 521).

The National Center for Education Statistics (NCES) conducts various national surveys that apply complex designs to projects such as the Beginning Postsecondary Students study (BPS), the National Educational Longitudinal Study of 1988 (NELS: 88), or the National Study of Postsecondary Faculty (NSOPF). Some statistical software programs, for instance SPSS or SAS, presuppose that data were accumulated through SRS. These statistical programs tend not to use as a default setting sample weights with data amassed through complex designs, but instead use raw expansion weights as a measure of acceptable sample size (Cohen, 1997; Muthen & Satorra, 1995). However, the complex sampling designs utilized in the collection of NCES survey data allocates larger comparative importance to some sampled elements than to others. To illustrate, a complex design identified by the NCES may have a sample selection where 1 subject out of 40 is chosen, which indicates that the selection probability is 1/40. The sample weight of 40, which is inversely proportional to the selection probability, indicates that in this particular case 1 sample subject equals 40 subjects in the population.

Because of the use of complex designs, sample weighting for disparate subject representation is employed to bring the sample variance in congruity with the population variance, which supports proper statistical inferences. The NCES incorporates as part of its data sets raw expansion weights to be applied with the data of study to ensure that the issues of sample selection by unequal probability sampling and biased estimates have been addressed. Relative weights can be computed from these raw expansion weights.

Because the NCES accrues an abundance of its data for analysis via CSD, the following formulae present how weights function. The raw expansion weight is the weight that many statistical software programs use as a default setting and should be avoided when working with the majority of NCES data. Instead, the relative weight should be used when conducting statistical analyses with NCES complex designs.

$$\text{Raw Expansion Weight } (W_j) = \sum_{j=1}^{n} w_j = N \qquad (1)$$

$$\text{Weighted Mean } (\bar{x}) = \sum_{j=1}^{n} w_j x_j / \sum w_j \qquad (2)$$

$$\text{Mean Weight } (\bar{w}) = \sum_{j=1}^{n} w_j / n \qquad (3)$$

$$\text{Relative Weight} = w_j / \bar{w} \qquad (4)$$

Notes: n = sample size, j=1 = subject response, w_j = raw weight, x_j = variable value, N = population size

Furthermore, the lack of sample weighting with complex designs causes inaccurate estimates of population parameters. The existence of variance estimates, which underestimate the true variance of the population, induce problems of imprecise confidence intervals, larger than expected degrees of freedom, and an enhancement of Type I errors (Carlson, Johnson, & Cohen, 1993; Lee, Forthofer, & Lorimor, 1989).

Design effect (DEFF) indicates how sampling design influences the computation of the statistics under study and accommodates for the miscalculation of sampling error. As noted previously, since statistical software programs often produce results based on the assumption that SRS was implemented, DEFF is used to adjust for these inaccurate variances. DEFF, as defined by Kish (1965), is the ratio of the variance of a statistic from a CSD to the variance of a statistic from a SRS.

$$\text{DEFF} = \frac{SE^2_{CSD}}{\text{SE}^2_{\text{SRS}}} \qquad (5)$$

The size of DEFF is affined to conditions such as the variables of interest or the attributes of the clusters used in the design (i.e., the extent of in-cluster homogeneity). A DEFF greater than 1.0 connotes that the sampling design decreases precision of estimate compared to SRS, and a DEFF less than 1.0 confirms that the sampling design increases precision of estimate compared to SRS (Kalton, 1983; Muthen & Satorra, 1995). As Thomas and Heck (2001) stated, "If standard errors are underestimated by not taking the complex sample design into account, there exists a greater likelihood of finding erroneously 'significant' parameters in the model that the a priori established alpha value indicates" (p. 529).

Procedures

Three variables were selected from the public-use database of the National Education Longitudinal Study of 1988 to demonstrate the impact of weights and design effects on contingency tables and chi-square analysis. A two-stage cluster sample design was used in NELS: 88, whereby approximately 1,000 eighth-grade schools were sampled from a universe of approximately 40,000 public and private eighth-grade schools (first stage) and 24 eighth-grade students were randomly selected from each of the participating schools (second stage).

An additional 2 to 3 Asian and Hispanic students were selected from each school, which resulted in a total sample of approximately 25,000 eighth-grade students in 1988. Follow-up studies were conducted on subsamples of this cohort in 1990, 1992, 1994, and 2000. Additional details on the sampling methodology for NELS: 88 are contained in a technical report from the U.S. Department of Education (1996).

The three variables used in this example are F2RHMA_C (total Carnegie Units in mathematics taken in high school), RMATH (flag for whether one or more courses in remedial math were taken since leaving high school), and F3TRSCWT (1994 weight to be used with 1992 transcript data). Five categories for the number of Carnegie Units of math taken in high school were created (up through 1.99, 2.00 through 2.99, 3.00 through 3.99, 4.00 through 4.99, 5.00 or more). The other variable of interest was whether a student had taken a postsecondary remedial math course by the time of the 1994 follow-up study. Four chi-square contingency tables were developed for these two variables using SPSS. Differences in the four tables are due to use of weights and DEFF.

Only those observations where RMATH > 0 and F3TRSCWT > 0 were selected for this analysis, which resulted in 6,948 students. Although there were 14,915 students in the 1994 follow-up of NELS: 88, only 12,509 had high school transcript data (F3TRSCWT > 0) from which F2RHMA_C was obtained. Of these, 6,948 participated in post-secondary education by the time of the third follow-up in 1994.

Missing values were not a problem with RMATH. Of the 14,915 students in the 1994 follow-up of NELS: 88, 6,943 had a legitimate missing value because they had not participated in postsecondary education (i.e., not of interest for this paper), 16 had missing values, and 7,956 had a value (yes or no) for postsecondary remedial math.

There were some missing values for high school transcript data, but the transcript weight (F3TRSCWT) provided in NELS: 88 takes into account missing transcript data. The Carnegie units of high school math (F2RHMA_C) came from high school transcript data. There were 14,915 students in the 1994 follow-up of NELS: 88; however, only 12,509 had high school transcript data. That is why NCES provides a separate weight (F3TRSCWT) that is to be used specifically with variables from high school transcript data.

This weight has already been adjusted by NCES for missing high school transcript observations. Of the 7,956 students with a value for RMATH, 1,008 did not have high school transcript data. These 1,008 students were not included in the analysis presented here (7,956-1,008 = 6948 students for analysis in this paper). After selecting the 7,956 students with a value for RMATH, only those observations with F3TRSCWT>0 were selected. No further adjustment was necessary for missing values since F3TRSCWT had already been adjusted by NCES for missing values.

Effect sizes are reported for each chi-square statistic addressed in the research. For the chi-square statistic, a regularly used effect size is based on the coefficient of contingency (C), which is not a true correlation but a "scaled" chi-squared (Sprinthall, 2000). As a caveat with the use of C, it has been noted that its highest value cannot attain 1.00, as is common with other effect sizes, which makes concordance with akin effect sizes arduous.

In fact, C has a maximum approaching 1.0 only for large tables. In tables smaller than 5 x 5, C may underestimate the level of association (Cohen, 1988; Ferguson, 1966). As an alternative to C, Sakoda's Adjusted C (C*) may be used, which varies from 0 to 1 regardless of table size. For chi-square related effect sizes, Cohen (1988) recommended that .10, .30, .50 represent small, medium, and large effects.

$$C = SQRT\ [\chi^2 / (\chi^2 + n)] \qquad (6)$$

$$C^* = C / SQRT\ [(k-1)/k] \qquad (7)$$

k = number of rows or columns, whichever is smaller.

Results

A total of 6,948 observations met the selection criteria (i.e., availability of high school transcripts and participation in post-secondary education by the time of the third follow-up in 1994). The first contingency table (Table 1), without any weights or design effects, has a total count of 6,948 and a chi-square value of 130.92. This table is useful for determining minimum cell sizes, but the percentages in each of the cells and the overall chi-square (130.92) are incorrect because the sample observations were not weighted to represent the population.

Table 1. Contingency Table Without Weights: Carnegie Units of High School Math by Postsecondary Education (PSE) Remedial Math. $\chi^2(4) = 130.92$, C = .136, 95% CI (.112, .160), C* = .192, 95% CI (.168, .216).

		PSE Remedial Math		
Units HS Math		Yes	No	Row Total
0 – 1.99	Count	84	231	315
	% of Grand Total	1.2%	3.3%	4.5%
2 – 2.99	Count	215	661	876
	% of Grand Total	3.1%	9.5%	12.6%
3 – 3.99	Count	495	1,875	2,370
	% of Grand Total	7.1%	27.0%	34.1%
4 – 4.99	Count	382	2,504	2,886
	% of Grand Total	5.5%	36.0%	41.5%
>= 5	Count	43	458	501
	% of Grand Total	0.6%	6.6%	7.2%
Column Total	Count	1,219	5,729	6,948
	% of Grand Total	17.5%	82.5%	100.0%

Asian and Hispanic students were over-sampled in NELS: 88, so the sample contained higher proportions of these ethnic groups than did the reference population. Sampling weights must be applied to the observations to adjust for the over-sampling. In contrast, a chi-square table without weights or design effects is appropriate for a simple random sample because each observation represents the same number of cases in the population.

The variable F3TRSCWT, a raw expansion weight, is used as the weight in Table 2. This is one of several raw expansion weights provided by NCES, and it is the weight that is to be used when analyzing variables from the 1994 follow-up (e.g., RMATH) in conjunction with high school transcript variables such as F2RHMA_C. The raw expansion weight is the number of cases in the population that the observation represents. Unlike simple random sampling, the weights are not the same for each subject. The weights for these 6,948 observations range from 7 to 12,940 with a mean of 228.50. The total count of 1,587,646 in this table represents the number of students from the 1988 eighth-grade cohort that met the selection criteria. This table contains correct population counts and percentages in the cells; however, the overall chi-square (27,500.88) is too high because the cell sizes are overstated. The cell sizes represent counts of the population rather than the sample.

Table 2. Contingency Table With Raw Expansion Weight F3TRSCWT: Carnegie Units of High School Math by Postsecondary Education (PSE) Remedial Math. $\chi^2(4) = 27{,}500.88$, C = .130, 95% CI (.128, .132), C* = .184, 95% CI (.182, .186).

		PSE Remedial Math		
Units HS Math		Yes	No	Row Total
0 – 1.99	Count	24,353	63,532	87,885
	% of Grand Total	1.5%	4.0%	5.5%
2 – 2.99	Count	53,767	167,485	221,252
	% of Grand Total	3.4%	10.5%	13.9%
3 – 3.99	Count	118,230	427,763	545,993
	% of Grand Total	7.4%	26.9%	34.4%
4 – 4.99	Count	81,325	537,884	619,209
	% of Grand Total	5.1%	33.9%	39.0%
>= 5	Count	14,951	98,356	113,307
	% of Grand Total	0.9%	6.2%	7.1%
Column Total	Count	292,626	1,295,020	1,587,646
	% of Grand Total	18.4%	81.6%	100.0%

The relative weight of F3TRSCWT is used in Table 3 to bring the cell counts in Table 2 back into congruence with the sample counts. For each of the 6,948 observations, the relative weight of F3TRSCWT is computed by dividing F3TRSCWT by 228.50, which is the mean of F3TRSCWT for the 6,948 observations. The total count in Table 3 is 6,947, which differs from Table 1 only because of rounding (note: although displayed in whole numbers by SPSS, Table 3 actually contains fractional numbers of observations in each cell). Table 3 contains correct cell percentages, but the cell sizes and chi-square (120.62) are overstated due to the two-stage clustered sample design of NELS: 88.

Table 3. Contingency Table With Relative Weight = F3TRSCWT / 228.5. Carnegie Units of High School Math by Postsecondary Education (PSE) Remedial Math. $\chi^2(4) = 120.62$, C= .131, 95% CI (.107, .155), C* = .185, 95% CI (.161, .209).

		PSE Remedial Math		
Units HS Math		Yes	No	Row Total
0 – 1.99	Count	107	278	385
	% of Grand Total	1.5%	4.0%	5.5%
2 – 2.99	Count	235	733	968
	% of Grand Total	3.4%	10.6%	13.9%
3 – 3.99	Count	517	1,872	2,389
	% of Grand Total	7.4%	26.9%	34.4%
4 – 4.99	Count	356	2,354	2,710
	% of Grand Total	5.1%	33.9%	39.0%
>= 5	Count	65	430	495
	% of Grand Total	0.9%	6.2%	7.1%
Column Total	Count	1,280	5,667	6,947
	% of Grand Total	18.4%	81.6%	100.0%

Table 4 was obtained by dividing the relative weight for F3TRSCWT by the NELS: 88 average DEFF (2.94), extrapolated via Taylor series methods, which resulted in effective cell sizes with correctly weighted cell counts and proportions and the appropriate overall chi-square (40.81) for this clustered design. The counts in Table 4 are the effective sample size after accounting for the clustered sample design (i.e., a sample of 6,948 from this clustered design is equivalent to a sample size of 2,363 randomly selected students). Essentially, a mean DEFF of 2.94 tells us that if a SRS design had been conducted, only 33% as many subjects when compared against a CSD, would have been necessary to observe the statistic of study.

DEFFs that range between 1.0 and 3.0 tend to be indicative of a well-designed study. The current study's DEFF of 2.94 indicated that the variance of the NELS: 88 estimates was increased by 194% due to variations in the weights. The square root of DEFF, the DEFT, yields the degree by which the standard error has been increased by the CSD. The DEFT (1.71) implied that the standard error was 1.71 times as large as it would have been had the present results been realized through a SRS design, or the standard error was increased by 71%. An intra-class correlation coefficient (ICC) of .20 or less is desirable for indicating the level of association between the responses of the members in the cluster. Since an ICC was not used in the computation of the Taylor series-derived average DEFF for NELS: 88, an estimated, average ICC was calculated from the following formula for determining

$$\text{DEFF: } 1 + \delta\,(n - 1), \qquad (8)$$

where δ is the ICC and n is the typical size of a cluster (Flores-Cervantes, Brick, & DiGaetano, 1999). The low ICC (.0844) indicated that the members in the same cluster were only about 8%, on average, more probable of having corresponding characteristics than if compared to another member selected randomly from the population.

Table 4. Contingency Table With Weight = (F3TRSCWT / 228.5) / 2.94: Carnegie Units of High School Math by Postsecondary Education (PSE) Remedial Math. $\chi^2(4) = 40.81$, C = .130, 95% CI (.090, .170), C* = .184, 95% CI (.144, .224).

		PSE Remedial Math		
Units HS Math		Yes	No	Row Total
0 – 1.99	Count	36	95	131
	% of Grand Total	1.5%	4.0%	5.5%
2 – 2.99	Count	80	249	329
	% of Grand Total	3.4%	10.5%	13.9%
3 – 3.99	Count	176	637	813
	% of Grand Total	7.4%	27.0%	34.4%
4 – 4.99	Count	121	801	922
	% of Grand Total	5.1%	33.9%	39.0%
>= 5	Count	22	146	168
	% of Grand Total	0.9%	6.2%	7.1%
Column Total	Count	435	1,928	2,363
	% of Grand Total	18.4%	81.6%	100.0%

NELS: 88 used a clustered sample design, in which schools were randomly selected, and then students within those schools were randomly selected. Students selected from such a sampling design would be expected to be more homogeneous than students selected from a simple random design across all schools. The chi-square values from SPSS cross-tabulations and SAS Proc Freq tables presume simple random samples. One method for estimating the proper chi-square for the two variables under investigation from NELS: 88 is to divide the relative weight for F3TRSCWT by the average DEFF (2.94), and use the result as the weight in SPSS cross-tabulations or SAS Proc Freq. The results in Table 4 were obtained by such a computation, which yields effective cell sizes and correctly weighted proportions. Furthermore, the chi-square (40.81) is an appropriate approximation of the true chi-square for this clustered design. These are the values that should be used in a chi-square analysis of Carnegie Units of high school math by whether or not a student took a postsecondary education remedial math course. Notice that the cell counts and the total count in Table 4 are equal to those in Table 3 divided by 2.94. The counts in Table 4 are the effective sample size after accounting for the clustered sample design.

As was found with the chi-square statistics, weighting, or lack thereof, also influenced effect size values. For example, the coefficient of contingency and Sakoda's Adjusted C in Table 1, where the default of no weighting occurred, had higher values than any of the reported C or C* estimations where a form of weighting transpired. It should be noted that the C values ranged from .130 to .136, or in the case of adjusted C from .184 to .192, which means that regardless of weighting scheme, or none at all, the practical implication of the chi-square statistics of study was that they had a small effect. Thus, although the chi-square statistics were all statistically significant, they had a small effect, which indicates that the results derived from the chi-square statistics would not be deemed very important practically and also in terms of accounting for much of the total variance of the outcome.

Conclusion

Some sampling designs over-sample certain groups (i.e., their proportion in the sample is greater than their proportion in the population) in order to obtain sufficiently large numbers of observations in these categories so that statistical analyses can be conducted separately on these groups. When analyzing the entire sample, relative weights should be used to bring the sample proportions back in congruence with the population proportions. When clustered sampled designs are used, then relative weights should be divided by the DEFF to adjust for the fact that a sample from a clustered design is more homogeneous than if a simple random sampling scheme had been employed. The chi-square values from SPSS cross-tabulations and SAS Proc Freq tables presume simple random samples. Design effects must be used with such software in order to obtain an appropriate approximation for the true chi-square, and its accurate effect size, of a clustered design.

References

Broene, P., & Rust, K. (2000). *Strengths and limitations of using SUDAAN, Stata, and WesVarPC for computing variances from NCES data sets*. (U.S. Department of Education, National Center for Education Statistics Working Paper No. 2000-03). Washington, DC: U.S. Department of Education.

Carlson, B. L., Johnson, A. L., & Cohen, S. B. (1993). An evaluation of the use of personal computers for variance estimation with complex survey data. *Journal of Official Statistics, 9*, 795-814.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, S. B. (1997). An evaluation of alternative PC-based software packages developed for the analysis of complex survey data. *The American Statistician, 51*, 285-292.

Ferguson, G. A. (1966). *Statistical analysis in psychology and education* (2nd ed.). New York: McGraw-Hill.

Flores-Cervantes, I., Brick, J. M., & DiGaetano, R. (1999). *Report no. 4: 1997 NSAF variance estimation*. http://newfederalism.urban.org/nsaf/methodology1997.html

Kalton, G. (1983). Introduction to survey sampling. In J. L. Sullivan & R. G. Niemi (Series Eds.), *Quantitative applications in the social sciences*. Beverly Hills, CA: Sage Publications.

Kish, L. (1965). *Survey sampling*. New York: John Wiley & Sons.

Lee, E. S., Forthofer, R. N., & Lorimor, R. J. (1989). Analyzing complex survey data. In M. S. Lewis-Beck (Series Ed.), *Quantitative applications in the social sciences*. Newbury Park, CA: Sage Publications.

McMillan, J. H., & Schumacher, S. (1997). *Research in education: A conceptual introduction* (4th ed.). New York: Longman.

Muthen, B. O., & Satorra, A. (1995). Complex sample data in structural equation modeling. In P. Marsden (Ed.), *Sociological methodology* (pp. 267-316). Washington, DC: American Sociological Association.

Sprinthall, R. C. (2000). *Basic statistical analysis* (6th ed.). Boston, MA: Allyn and Bacon.

Thomas, S. L., & Heck, R. H. (2001). Analysis of large-scale secondary data in higher education research: Potential perils associated with complex sampling designs. *Research in Higher Education, 42*, 517-540.

U.S. Department of Education. (1996). *Methodology report, National Education longitudinal study: 1988-1994* (NCES 96-174). Washington, DC: Author.