Computing Variances from Data with Complex Sampling Designs:
A Comparison of Stata and SPSS

North American Stata Users Group
March 12-13, 2001

Alicia C. Dowd, Assistant Professor
Univ. Mass. Boston, Graduate College of Education
Wheatley Hall, 100 Morrissey Blvd.
Boston MA 02125
alicia.dowd@umb.edu
617 287-7593 phone
617 287-7664 fax

Michael B. Duggan, Director of Enrollment Research
Suffolk University
8 Ashburton Place
Boston MA 02108
mduggan@admin.suffolk.edu
617 573-8468 phone
617 720-0970 fax

Introduction

        The National Center for Education Statistics (NCES) is responsible for collecting, analyzing, and reporting data related to education in the United States and other countries (U.S. Department of Education, 1996, p. 2). Among the surveys conducted by the NCES, several pertain to postsecondary schooling and outcomes and are of interest to higher education researchers. These include Beginning Postsecondary Students (BPS), Baccalaureate and Beyond (B&B), National Postsecondary Student Aid Study (NPSAS), the National Study of Postsecondary Faculty (NSOPF), and the Integrated Postsecondary Education Data Set (IPEDS). With the exception of IPEDS, these surveys are conducted using complex survey designs, involving stratification, clustering, and unequal probabilities of case selection. Researchers analyzing these data must take the complex sampling designs into account in order to estimate variances accurately.

        Novice researchers and doctoral students, particularly those in colleges of education, will likely encounter issues surrounding the use of complex survey data for the first time if they undertake to analyze NCES data. Doctoral programs in education typically have minimal requirements for the study of statistics, and statistical theories are usually learned based on the assumption of a simple random sample. The NCES, fulfilling their congressional mandate to disseminate data in ways useful to the policymaking and research communities, offers various resources for data analysts, including data in a variety of formats, technical reports, and working papers. For the novice researcher, the NCES presents (along with the National Science Foundation) summer database institutes, designed to introduce the user to the language, methods, and statistical tools associated with complex survey designs.

        In this paper, we compare the steps necessary to conduct analyses of complex data using Stata and SPSS. The primary purpose is to document explicitly these steps so new higher education researchers (and others in related fields) can determine which of these programs (or similar software) they would like to use. In addition, the paper highlights the need to adjust analyses for the complex survey designs and to demonstrate the differences in results obtained if one does not take the complex design into account. Recent published work in higher education journals demonstrates the need for more

information on this topic. Several researchers have analyzed and reported results from NPSAS analyses without accounting for weighting, stratification, or clustering and without discussing the implications of treating the samples as simple random samples.

This paper supplements existing resources. The monographs <u>Analyzing Complex Data</u> (Lee, Forrthofer, & Lorimor, 1989) and <u>Practical Sampling</u> (Henry, 1990), both from Sage Press, provide an introduction to complex sampling designs and the underlying statistical theory. <u>Analysis of Complex Surveys</u> (Skinner, Holt, & Smith, 1989) provides a technically advanced discussion of statistical theory and applications. Broeme and Rust (U.S. Department of Education, 2000) have authored an NCES working paper, <u>Strengths and Limitations of Using SUDAAN, Stata, and WesVarPC for Computing Variances from NCES Data Sets</u>, in which they assess the programming and analysis capabilities of those three software packages. In the recently released second edition of their text <u>Applied Logistic Regression,</u> Hosmer and Lemeshow (2000) include a chapter discussing the steps and statistical tests appropriate for fitting logistic regression models using complex design surveys. Finally, Logistic Regression Analysis in Higher Education: An Applied Perspective by Cabrera (1994), often referenced by educational researchers, is a valuable introduction to logistic analysis, though the author does not address complex sampling issues.

Beginning Postsecondary Students Survey Design

The analysis in subsequent sections presents a topic familiar to higher education researchers. Using logistic regression, the dichotomous outcome of obtaining or not obtaining a bachelor s degree is regressed on several predictor variables. The data analyzed are from the Second Follow-up of the longitudinal BPS study, which has the 1990 NPSAS survey as its base year. The independent variables are theoretically relevant to the outcome and are often included in empirical research on this topic. These particular variables are included in order to use BPS variables for which the NCES has published design effects. Tables VII.1- VII.6 of the BPS Final Technical Report (U.S. Department of Education, 1996) provide the means, standard errors, and design effects for the selected variables, cross tabulated by the dependent variable for this analysis, highest undergraduate degree attained. The analysis is presented strictly for the discussion of

variance estimation under complex sampling designs. No attempt is made to present a fully elaborated or tested model.

The BPS:94 respondents are a subset of NPSAS:90 respondents, selected based on their status as first-time beginning students. As a longitudinal component of NPSAS, the design, weighting, and estimation procedures for BPS 90:94 necessarily follow from those for NPSAS. The NPSAS:90 design involved a multistage probability sample with three stages. In the first stage of sampling, 121 geographic areas, called primary sampling units (PSUs) were selected. In the second stage, 1522 institutions located in the NPSAS PSUs were selected, of which 80 percent were eligible and 1,130 participated (U.S. Department of Education, 1996, p.5). These institutions were identified for selection within 8 strata defined by institutional type and sector (for example, two of these strata are public 4-year and private not-for-profit 2-3 year institutions). The probability of institutional inclusion was proportional to composite measures of institutional enrollment size and student type. The purpose of the strata was to ensure proportional representation of institutional types by region and for small sized institutions.

In the third stage, more than 70,000 students enrolled at these institutions during four time points in 1989-90 were identified for sampling. This student sample effectively represented all students enrolled during the 1989-90 academic year (U.S. Department of Education, 1996, p.5). The BPS sample followed up on a subset of students in the NPSAS sample who were first-time beginning students in 1990. The final sample for BPS:90 included 992 institutions (some in the NPSAS sample were found to be ineligible when none of their sampled students were first-time beginners) and 7,220 students were interviewed.

Each case of students interviewed in BPS is assigned a sampling weight. In a simple random sample, each case is selected with an equal probability and represents an equal number of cases in the population. In BPS and other NCES surveys employing complex survey designs, each case is selected with an unequal probability of selection and represents a different number of cases in the population. The sampling weight assigned to a case indicates the number of cases in the population represented by the respondent. The sampling weight is the inverse of the selection probability. For example,

if a student is selected as 100 out of 1000 students, the selection probability is 1/10, and the sampling weight is 10. (Technical Issues in Using NCES Data, NCES Data Base Institute presentation, June 16, 1999.)

Summary of Statistical Issues

The sampling variance of a survey statistic is affected by the stratification, clustering, and weighting of selected cases. Stratification may increase the precision of the variance estimate, but clustering and weighting decrease precision (U.S. Department of Education, 1996, p.100). Cases within primary sampling units are often correlated, thus violating any assumptions, central to inference using simple random samples (SRS), that the cases are independent and identically distributed (IID). As described in the NCES Final Technical Report for the BPS 90:94 survey: The cumulative effect of the various factors affecting the precision of a survey statistic is often modeled as the survey design effect. The design effect, deff, is defined as the ratio of the sampling variance of the statistic under the actual sampling design divided by the variance that would be expected for a simple random sample of the same size (U.S. Department of Education, 1996, p. 101).

```
Deff= variance(complex design)/variance(SRS)
```

The Deff for linear estimates, such as means, totals, and proportions, is typically much larger than one. The Deff for regression coefficients takes its value from an equation involving the product of the intracluster correlation coefficient of the independent variable and the intracluster correlation coefficient of the dependent variable. By definition the correlation coefficients are each less than one and their product reduces the size of the Deff . The design effect can be less than one for regression coefficients when the product of the intracluster correlation coefficients is negative. (Hosmer & Lemeshow, 2000, p. 220). Recognizing that the size of the design effect for a regression coefficient varies by the dependent and independent variables in the regression highlights that fact that it is accurate to think of a design effect matrix, rather than one design effect pertaining to an entire survey sample.

The t statistic from a complex sampling design is equal to the t statistic from a simple random sample (SRS) divided by the square root of the Deff (Technical Issues in Using NCES Data, NCES Data Base Institute presentation, June 16, 1999).

```
t statistic(complex)=t statistic(SRS)/square root of Deff
```

This formulation of the relationship between the variance under complex sampling and simple random sampling shows that when the Deff is greater than one, the t statistic (complex) will be smaller than the t statistic (SRS). The opposite is true when the Deff is less than one: the t statistic (complex) is larger than the t statistic (SRS). When researchers treat data from a complex sample as though it were obtained under SRS, they obtain inflated significance levels of regression coefficients when the Deff is greater than one, and reduced significance levels for regression coefficients when the Deff is less than one.

Skinner, Holt, and Smith (1989, p. 29, Table 2.1) show that ignoring the Deff, which is the same as treating the Deff as 1.0, skews significance levels of obtained p. values quite significantly. Reproducing just part of their table, it is clear that researchers will easily misestimate the statistical significance of their results if they do not take account of the design effect. As shown in Table 1, for example, a design effect of 1.5 generates an actual significance level (.11) that is more than twice the nominal alpha level of .05. With a design effect of 3.0, the actual significance level of .26 is far from standard levels of statistical significance.

Table 1 Nominal and Actual Significance Levels with Design Effects

Design Effect	Significance Level	
	Nominal	Actual
.9	.05	.04
1.0	.05	.05
1.5	.05	.11
2.0	.05	.17
2.5	.05	.22
3.O	.05	.26

Steps to Adjusting for Design Effects in Stata and SPSS

Specialized software such as Stata--other packages in this category include WesVarPC and SUDAAN)--compute Taylor series variance estimates that are adjusted for the design effects resulting from complex sampling designs (StataCorp, 2001; U.S. Department of Education, 1996, p 99). The Taylor series computes variance statistics based on information stored in the survey variables for sampling weight, strata, and primary sampling units (PSUs). In the example below using BPS90:94 data, these are, respectively, the variables bps94awt, ofcon2, and psu. In Stata, the researcher uses the svyset command to set the sampling weight equal to bps94awt, the strata equal to ofcon2, and the primary sampling unit equal to psu (StataCorp, 2001, vol. 4, p. 75). The weight variable bps94awt is one of four sampling weights available in BPS 90:94. BPS94awt is the sampling weight for the cross-sectional sample and is appropriate for the analysis that follows. The other weights are appropriate for longitudinal studies or when selecting subsamples of different types of students.

Variance estimates in SPSS (as well as in SAS) are based on the assumption of a simple random sample. Researchers using one of these two popular packages must adjust the estimated variances. Two approaches can be taken. One involves using an average, or generalized, Deff based on the value of the Deff for the dependent variable in a regression analysis. The second involves using the individual Deffs associated with each independent variable and the dependent variable in any given analysis to adjust the t statistics reported under the SRS assumption. The SRS t statistic is divided by the square root of the deff (as indicated above) to provide an adjusted t statistic that will indicate accurate levels of statistical significance.

The first approach is simpler, as the Deff may be set in one step, but it is less accurate. The loss of accuracy depends on the difference between each individual Deff and the selected generalized Deff. The range of values of the Deff in any given data set can be considerable, so the difference between the individual Deff and the generalized Deff can be large, particularly for estimates of variance associated with means and other linear estimates. The second approach is more labor intensive, requiring the selection of a Deff for each variable in the equation, but more accurate.

In this report, generalized Deffs were used and were selected based on published Deffs, as discussed above. The design effects for NCES surveys are available for key variables in the NCES technical report for each survey. The Deff for all variables in user-defined correlation and covariance tables may be obtained through the NCES Data Analysis Systems (DAS), available on CDRom or on line through the world wide web.

In SPSS, the researcher may choose the weight cases option and indicate the weighting variable. Again, the correct weighting variable for the following example is bps94awt. However, SPSS computes the population size by summing the weights associated with each case. Therefore, it is necessary to calculate the relative weight, or normalized weight for each case and to use the relative weight in the analyses. The relative weight is defined as the weight for each case divided by the mean weight of all the cases in the sample. This value is multiplied by 1/Deff, where Deff is the generalized Deff selected for the analysis, to obtain a relative weight adjusted for the design effect (Technical Issues in Using NCES Data, NCES Data Base Institute presentation, June 16, 1999). For this analysis, the relative weight was multiplied by _, which is a Deff of 2, selected as a generalized Deff based on the published design effects associated with the variable for highest degree attained (U.S. Department of Education, 1996, p.112).

Example Using BPS 90:94

A binary logistic regression predicted the dichotomous outcome of attaining or not attaining a bachelor s degree as the highest degree. The independent variables and their descriptions are shown in Table 2. The indicator variables for race were created using Stata s automatic indicator coding command, xi, where the largest group was established as the comparison group (natural coding). The automatically generated names for the i.bpsrace categories have been renames in Tables 3 and 4 for ease of reference. The results for the regression in Stata are displayed in Table 3 as logit coefficients in the column labeled Coef. These results are converted into odds ratios, Exp(B), and reported again in Table 4, where they are presented in the form of confidence intervals side-by-side with the results from the regression in SPSS. The Deffs generated by Stata are also reported in Table 4, column 2.

Table 2 BPS 90:94 Variables included in the Analysis

Bach	Dependent Variable 1= bachelor s degree highest degree; else=0
Level2	1=First institution was a 2yr. college; else=0 Compared to first institution was a 4yr. college
Level1	1=First institution was <2yr. college; else=0 Compared to first institution was a 4yr. college
Gender	1=Male; Female=0
Sesperc	Socioeconomic percentile
Noaid	1= no financial aid received; else=0
Nontrad	1=older than 21yrs of age 1990; else=0
i.bpsrace	Indicator variables created for race with white as omitted group and Black, Hispanic, Asian, and American Indian as comparison groups

Table 3 Command Syntax and Results from Stata

```
.
. xi: svylogit bach level2 level1 gender sesperc noaid nontrad i.bpsrace,
p ci deff
i.bpsrace              Ibpsra_1-5   (naturally coded; Ibpsra_1 omitted)

Survey logistic regression

pweight:  bps94awt                    Number of obs= 5887
Strata:   ofcon2                      Number of strata = 8
PSU:      psu                         Number of PSUs   = 1144
                                      Population size  = 2226144.6
                                      F( 10,   1127) = 62.71
                                      Prob > F       = 0.0000


    ------------------------------------------------------------------------
       bach |     Coef.   Std. Err.       t    P>|t|    [95% Conf. Interval]
    ---------+--------------------------------------------------------------
      level2 | -2.402613   .1709494   -14.055   0.000    -2.738025   -2.067201
      level1 | -4.099195   .4051118   -10.119   0.000    -4.894047   -3.304344
      gender | -.3647276   .0962272    -3.790   0.000    -.5535307   -.1759245
     sesperc |  .0219936   .0023353     9.418   0.000     .0174116    .0265756
       noaid | -.7131809   .0894584    -7.972   0.000     -.888703   -.5376587
     nontrad | -1.212565   .2546131    -4.762   0.000     -1.71213   -.7130007
    Black    | -.3306523   .1587093    -2.083   0.037    -.6420485   -.0192561
    Hispanic |   .074014   .2367039     0.313   0.755    -.3904119    .5384398
    Asian    |   .350513   .2357684     1.487   0.137    -.1120775    .8131035
    AmIndian | -.1546564   .4104237    -0.377   0.706      -.95993    .6506171
       _cons | -.8940138   .1934085    -4.622   0.000    -1.273492   -.5145357
    ------------------------------------------------------------------------
```

The results reported by Stata (Table 3) include an adjusted Wald test of the overall significance of the model (Prob>F = .0000). The —2LogLikelihood is not generated in Stata, though it is in SPSS, because under a complex sampling design the estimates are not true maximum likelihood estimates, but an approximation (Hosmer &Lemeshow, 2000). For the same reason, the pseudo Rsquared, often reported in educational research using logistic regression, is not indicated in the output.

Table 4 Confidence Intervals for Exp(B) from Stata and SPSS

```
--------------------------------------------------------------------------------
  bach  |      DEFF   [exp(B) 95% Conf. Interval]    [exp(B)95% Conf. Interval]
        |      Stata              Stata                    SPSS, DEFF=2
--------+-----------------------------------------------------------------------
 level2 |   3.462604   .065        .127                 .072         .117
 level1 |   .7571926   .007        .037                 .055         .060
h_gendr |   1.665473   .575        .838                 .572         .855
 sesperc|   1.325557  1.017       1.027                1.016        1.027
  noaid |   1.298902   .411        .584                 .402         .614
 nontrad|   1.522288   .181        .490                 .182         .484
Black   |   1.138383   .526        .981                 .482        1.058
Hispanic|   1.813159   .677       1.713                 .687        1.655
Asian   |   1.654303   .894       2.255                 .867        2.330
AmIndian|   .6061156   .383       1.917                 .205        3.842
  _cons |   1.481074
--------------------------------------------------------------------------------
```

While the estimation of the variance statistic is affected by the design effect, the estimation of the coefficient is not. Therefore, the estimated coefficients were the same in Stata and in SPSS. When the Deff is actually greater than the generalized Deff of 2 specified in the SPSS analysis (see Table 4, column 2), the SPSS confidence intervals are expected to be smaller than the Stata confidence intervals. This result is observed in Table 4 for the level2 variable, where the actual Deff is 3.5. When the Deff is actually smaller than 2, the SPSS confidence intervals are expected to be larger. This pattern is generally observed in this output. However, when the actual Deff is close to two, the results deviate from that expectation, but the differences in the size of the intervals are small. When the difference between the generalized Deff and the actual Deff is very large and the coefficient is large, as it is on the variable for American Indian, the differences between the confidence intervals is considerable.

Stata also provides the capacity to run an adjusted Wald test (through the command svytest) to test the significance of individual and joint coefficients in a manner appropriate to the complex survey design. Table 5 reports a test on the significance of the group of indicator variables for race (In this table, the automatically named indicator variables are not renamed). The significance of this group of variables is rejected (prob>F=.1051). In contrast, entering race as a block of categorical variables in SPSS yields an F test in which the joint significance of these variables is not rejected (p<.05).

Table 5 Adjusted Wald Test for Joint Coefficients

```
. svytest Ibpsra_2 Ibpsra_3 Ibpsra_4 Ibpsra_5

Adjusted Wald test

 ( 1)   Ibpsra_2 = 0.0
 ( 2)   Ibpsra_3 = 0.0
 ( 3)   Ibpsra_4 = 0.0
 ( 4)   Ibpsra_5 = 0.0

       F(  4,  1133) =    1.92
             Prob > F =    0.1051
```

Discussion

In the results above, the use of the generalized Deff in SPSS typically yields confidence intervals that are similar to those generated in Stata using the survey (svy) commands. There are no substantive differences in the estimation of the effect on the attainment of a bachelor s degree of the type of first institution attended, gender, SES percentile, aid receipt, and nontraditional student age. This is true also for the categories of race, with the exception of American Indian, a category that has a small number of cases. These reasonably accurate results in SPSS depend on the choice of a generalized Deff that falls within the mid-range of the actual Deffs. In this example, the chosen value of 2 falls in the mid-range of the actual Deff minimum of .61 and the maximum of 3.4. In this respect, then, the approach provides reasonable approximations and would not generate inaccurate interpretations. However, the analyst must be vigilant in obtaining the

actual Deffs from the NCES Data Analysis System (or from published Deff tables) in order to be confident that the generalized Deff is appropriate.

The results of the joint test of significance of the group of race indicator variables are more problematic, in that the SPSS results encourage the inclusion of race as a significant group of variables and interpretation of the individual t statistics as significant predictors. The Stata results using the adjusted Wald test lead to the opposite conclusion. Therefore, it appears that the model-building process is compromised under assumptions of SRS inherent in the SPSS calculations. The example above includes a large sample size. Contradictory results regarding tests of significance would likely have been observed more often in this simulation if the sample size had been smaller.

Conclusion

While NCES disseminates information about the use of generalized design effects in SPSS or SAS, it recommends that researchers use programs like Stata (or SUDAAN or WestVar) that compute variance estimates corrected for design effects. Use of generalized design effects is recommended only for very limited analyses (see, for example, U.S. Department of Education, 1997). Learning about complex sampling designs presents a challenge for novice educational researchers. If they use SPSS or SAS when working with data from a complex design and create a generalized Deff, they must consider the loss of accuracy in estimating confidence intervals. If they use individual Deffs, they must be attentive at each step of model building and testing to make appropriate adjustments. The process of adjusting for the survey design effects is considerably simpler in Stata, where setting the variables for weighting (using a pweight, or probability weight), strata, and PSU can be accomplished at the outset using the svyset command. In addition, Stata does not generate results that depend on the estimation of a true likelihood through maximum likelihood estimation, since the likelihoods are only approximations under complex designs. This is a benefit to the novice researcher, who may otherwise assume that any generated output can appropriately be applied to interpret results. The absence of typically reported statistics forces the researcher to find alternate approaches to determine goodness of fit. (These approaches are discussed in Hosmer and Lemeshow, 2000).

If educational researchers already own a package more typically used in educational research, the purchase of Stata or one of the specialized software packages for complex survey data creates an additional financial cost. In addition, a time cost is incurred to learn a new statistical package and to move from a menu-driven to a syntax-based computing environment. The novice educational researcher will also find that Stata reference manuals assume a higher level of statistical knowledge than do SPSS manuals. However, once familiar with the survey (svy) commands, Stata provides a simpler and more accurate estimation and model-building process.

Bibliography

Cabrera, A. (1994). Logistic Regression Analysis in Higher Education: An Applied Perspective. In J.C. Smart, ed. <u>Higher Education: Handbook of Theory and Research</u>, vol.10. Edison, N.J.: Agathon Press.

Henry, G. T. (1990). <u>Practical Sampling</u>. Newbury Park, CA: Sage Press.

Hosmer, D. W. & Lemeshow, S. (2000). <u>Applied Logistic Regression</u>. New York: John Wiley and Sons.

Lee, E.S.; Forthofer, R.N.; & Lorimor, R. J. (1989). <u>Analyzing Complex Survey Data</u>. Newbury Park, CA: Sage Press.

Skinner, C. J.; Holt, D.; Smith, T.M.F. (1989). <u>Analysis of Complex Surveys.</u> New York: John Wiley and Sons.

U.S. Department of Education (2000). National Center for Education Statistics. Strengths and Limitations of Using SUDAAN, Stata, and WesVarPC for Computing Variances from NCES Data Sets, Working Paper No. 2000-03, by Pam Broene and Keith Rust. Project Officer, Susan Ahmed. Washington, D.C.

U.S. Department of Education (1999). National Center for Education Statistics. Technical Issues in Using NCES Data, NCES Data Base Institute, June 16.

U.S. Department of Education (1997), National Center for Education Statistics. <u>National Postsecondary Student Aid Study, 1995-96 (NPSAS:96), Methodology Report,</u> NCES 98-073, by John A. Riccobono, Roy W. Whitmore, Timothy J. Gabel, Mark A. Traccarella, and Daniel J. Pratt, Research Triangle Institute; Lutz K. Berkner, MPR Associates, Inc. Andrew G. Malizio, project officer. Washington, DC.

U.S. Department of Education (1996). National Center for Education Statistics. <u>Beginning Postsecondary Students Longitudinal Study Second Follow-up (BPS:90/94) Final Technical Report</u>, NCES 96-153. Washington, D.C.