GENERALIZED VARIANCE ESTIMATES FROM THE NATIONAL SURVEY OF FAMILY GROWTH, CYCLE 5

Jill D. Kavee and Vincent G. Iannacchione, Research Triangle Institute
Jill D. Kavee, Research Triangle Institute, 3040 Cornwallis Rd, PO Box 12194,
Research Triangle Park, NC 27709-2194

Keywords: Generalized variances, Design-based Sampling variance, Design effects

Introduction

Accurate estimates of sampling variances are necessary to assess the reliability of parameter estimates or statistics. Variance estimation procedures for multi-stage sampling designs with clustered data, such as the National Survey of Family Growth (NSFG), must take into account the intricacies of the design. This is called a design-based sampling variance. A variance estimate for complex surveys based on a simple random sample assumption will generally underestimate the actual sampling variance. Design-specific software use complex formulae to compute the actual variances. In the absence of design-based software, generalized variance (GV) estimation is a short-cut method that approximates the correct sampling variance of a population estimate. This paper provides a discussion of the procedures used to create the GV estimates for Cycle 5 of the NSFG (NSFG-5). A comparison of the GV estimates to design-based variances created using Taylor series linearizations is also included.

Background on NSFG-5

The National Survey of Family Growth is administered by the National Center for Health Statistics (NCHS), an agency of the Department of Health and Human Services. The purpose of the survey is to produce national estimates of factors affecting pregnancy—including sexual activity, contraceptive use, infertility, and sources of family planning services—and the health of women and infants. For NSFG-5, interviewing and data processing were conducted by the Research Triangle Institute (RTI), under a contract with NCHS.

A national probability sample of 14,000 civilian noninstitutionalized women ages 15 to 44 was selected for interviewing between mid-January and October 1995. The interviews were conducted in-person by trained female interviewers using laptop computers; this procedure is called computer-assisted personal interviewing (CAPI). The interview, which lasted an average of 105 minutes, collected data on each pregnancy (if any); contraceptive use by her and her partner; her ability to bear children; the use of medical services for contraception, infertility, and prenatal care; her marriage, cohabitation, living situation, and work history; and a variety of demographic and economic characteristics. Additional data were collected in a short self-administered interview in which the respondent heard the questions over headphones and entered her own answers into the notebook computer. This procedure is called audio computer-assisted self-interviewing (A-CASI).

Interviews were obtained from a total of 10,847 women ages 15 to 44 selected from among households that responded to the 1993 National Health Interview Survey (NHIS). The NHIS is a continuous multi-stage household survey conducted by NCHS that covers the U.S. civilian noninstitutionalized population. Women in the NSFG-5 sample were selected from all 198 NHIS primary sampling units, or PSUs. A PSU is a metropolitan statistical area (MSA), a county, or a group of adjacent counties. PSUs were located in nearly every State and included all of the largest metropolitan areas in the United States.

All NHIS households containing Hispanic or non-Hispanic black women were included in the NSFG-5 sample; one woman was selected randomly if more than one woman was eligible for the study. Households of other race or ethnic identification were sampled. The subsample of "other race" households was selected with probability proportional to the number of eligible women in the household.

Design-based variance estimates for the NSFG-5 were calculated after creating analysis weights by adjusting the sampling weights to compensate for nonresponse. The weights were also adjusted so that the sum of the weights across all study respondents would match the population totals by age, race/ethnicity, marital status, and parity generated by the Bureau of the Census from the Current Population Survey (CPS). Due to the complexity of the sampling design, estimation of the sampling variance required the use of SUDAAN, a survey data analysis software package, developed at RTI, designed to accommodate the population parameter being estimated and the sampling design. A shortcut method for estimating the design-based sampling variances was developed. This is known as Generalized Variance Estimation.

Generalized Variance Estimation

Generalized variance (GV) estimates are short-cut methods that approximate the sampling variance of a population estimate when a specific or direct estimate of the variance is not available. GV estimates are based on a modeled relationship between a set of key parameter estimates and their associated direct variances. Because GV estimates reflect a composite of related design-based variance estimates, they usually are smaller or larger than the direct variance estimate of a parameter. Due to this fluctuation, the direct variance estimates obtained from SUDAAN or a similar survey analysis package is strongly preferred to the GV estimates described below.

The model developed for the generalized variance algorithm assumes that the subpopulations of interest (e.g., the denominator of the proportion) will be combinations of the 108 post-stratification cells used in the computation of the fully adjusted analysis weights. The post-stratification cells are cross-classifications of age, race/ethnicity, marital status, and parity from the CPS. The sampling variance for any combination of the post-stratification cells is assumed to be known without error. For example, if the denominator for a percentage is a combination of the post-stratification cells (e.g., Hispanics ages 25 to 29), then only the numerator contributes to the sampling variance.

The most commonly used model for generalized variances for subpopulation proportions relates the relative sampling variance for an estimate (i.e., the sampling variance divided by the square of the survey estimate) to the inverse of the survey estimate (Wolter, 1985). The model is of the form

$$\begin{aligned} V^2(P) &= S^2 / P^2 \\ &= \alpha + \beta / X \qquad (1) \end{aligned}$$

where,

S^2 = the sampling variance of the estimated proportion P,

α, β = the model coefficients to be estimated, and

X = the survey estimate of the numerator of the proportion P.

Alternatively, the relative variance may be expressed in terms of a design effect (Kish, 1965) which is the ratio of the variance of a survey estimate to the variance that would have been obtained from a simple random sample (SRS) of the same sample size. The design effect (Deff) for an estimated proportion P is defined as

$$\text{Deff} = S^2 / [P(1-P) / n] \qquad (2)$$

where,

S^2 = the sampling variance of the estimated proportion P, and

n = the sample size.

The use of design effects allows the model in (1) to be recast as

$$\begin{aligned} V^2(P) &= (1-P)\text{Deff} / (Pn) \\ &= -\text{Deff} / n + (N\ \text{Deff} / n) / X \qquad (3) \\ &= \alpha + \beta / X \end{aligned}$$

where,

n = the sample size,

X = the survey estimate of the numerator of the proportion P, and

N = the population size or denominator of the proportion P.

Design effects provide a summary measure of the combined effects of stratification, clustering, and unequal weighting on the variance of a survey estimate. The design effects are particularly useful for estimating GVs because they identify the subpopulations that are most affected by the sample design. For example, the design effects for Hispanic and non-Hispanic Black women are generally greater than those obtained for other women because of the oversampling of minorities specified in the NSFG-5 sample design.

Approximation Methods

Two formulas were used for the GV estimation procedures, one for the respondent-level data and one for the pregnancy-level data. However, only respondent-level results will be presented in this paper. Median design effects were used in the formulas instead of average design effects because extreme values can distort measurements based on means. Median design effects for the respondent data, displayed in Table 1, were based on the following nine recoded variables:

1. Indicator for the age of menarche less than 13.
2. Indicator for at least one completed pregnancy.
3. Indicator for at least one live birth.
4. Indicator for the female being fertile.
5. Indicator for current contraceptive being either the Pill or a Male Condom.
6. Indicator for ever using the Pill.
7. Indicator for ever using a Male Condom.
8. Indicator for first method of contraception being either the Pill or a Male Condom.
9. Indicator for intention of additional children.

The direct estimates of the sampling variances were computed using the variance estimator incorporating the single and joint inclusion probabilities (i.e., design=UNEQWOR) in the SUDAAN procedure DESCRIPT. Parameter estimates and the corresponding sampling variances were computed for all woman and for each of the three race/ethnicity domains (i.e., Hispanic, non-Hispanic Black, and Other).

Log-linear Regression Models

Generalized standard errors (GSEs) were obtained from a prediction equation involving the design effect estimates. The model was initially based on the design effect definition for proportion given in equation (2). A log (base 10) linear relationship between the *Deff*, the estimated proportion *p*, and the sample size *n* is the resulting prediction equation:

$$\log[Deff(p)] = \beta_0 + \beta_1 \log(p) + \beta_2 \log(1-p) + \beta_3 \log(n)$$

where,

$\beta_0, \beta_1, \beta_2, \beta_3$ = regression coefficients for the intercept, log(p), log(1-p), and log(n), respectively.

Separate models were fit for the data within the race/ethnicity categories: Hispanic, non-Hispanic Black, and Overall. By substituting the fitted model back into the definition of the design effect, a prediction equation for the GSE is

$$se_j(p) = \frac{10^{b_{0j}/2} \; x \; p_j^{(1+b_{1j})/2} \; x \; (1-p_j)^{(1+b_{2j})/2}}{n_j^{(1-b_{3j})/2}}$$

where,

j = race/ethnicity category indicator, and

b_0, b_1, b_2, b_3= estimated regression coefficients for the intercept, log(p), log(1-p). and log(n), respectively.

Initially four Generalized Standard Error tables were produced. The GSE table associated with the "Other" race/ethnicity group was eliminated because the information was basically replicated in the "Overall" race/ethnicity category. GSEs of estimates for Hispanic, non-Hispanic Black, and all women are provided in Table 2. Note that the values provided in these tables are 100 times the standard error since the column headings are percents and not proportions. For example, say a GSE is needed for the percent of Hispanic women who have ever used a specific type of birth control. Table 2 is used to determine the appropriate standard error given the estimate of the percent of Hispanic women in this category (column heading) and the sample count of these women (row heading). The following section compares three standard error estimates for a specific example.

Comparison of GSEs to Direct Estimates

Three methods for calculating standard errors have been discussed to this point, the simple random sample method (SRS), the direct design-based method (SUDAAN), and the generalized method (GSE). A specific example from the NSFG, Cycle 5 survey is presented to compare the standard error results from the three methods. For this example, the proportion of females ever using the Pill is considered. Table 3 displays the results from the SRS, SUDAAN, and GSE methods by overall and by age. Note that the generalized standard errors both over- and under-estimate the design-based standard errors obtained from SUDAAN.

References

Kish L. Survey Sampling. New York: John Wiley & Sons. 1965.

Potter FJ, Iannacchione VG, Mosher WD, Mason RE, Kavee JD. Sample Design, Sampling Weights, Imputation, and Variance Estimation in the 1995 National Survey of Family Growth. National Center for Health Statistics. Vital Health Stat 2(124). 1998.

Shah BV, Barnwell BG, Bieler GS. SUDAAN User's Manual, Release 7.5. Research Triangle Park, NC: Research Triangle Institute. 1997.

Wolter KM. Introduction to Variance Estimation. New York: Springer-Verlag. 1985.

Table 1. Median Design Effects for Nine Respondent-Level Outcomes, by Race/Ethnicity and Demographic Characteristics: 1995 NSFG, Cycle 5.

	RACE/ETHNICITY			
	Hispanic	Black, NH	Other	Total
Total	1.36	1.67	1.30	1.46
Age Category				
15-17	1.21	1.24	1.01	1.10
18-19	1.31	1.43	1.08	1.14
20-24	1.27	1.52	1.23	1.39
25-29	1.44	1.36	1.26	1.47
30-34	1.16	1.37	1.16	1.30
35-39	1.37	1.47	1.09	1.19
40+	1.29	1.51	1.05	1.17
Marital Status				
Married	1.38	1.46	1.13	1.27
Wid/Div/Sep[1]	1.31	1.60	1.16	1.32
Never Married	1.21	1.51	1.19	1.37
Education				
Less Than H.S.	1.37	1.46	1.05	1.24
H.S. Diploma	1.33	1.53	1.19	1.32
Some College	1.33	1.55	1.46	1.57
College Grad+	1.43	1.39	1.15	1.24
Poverty Level				
0-100%	1.45	1.67	1.26	1.53
101-200%	1.35	1.58	1.48	1.54
201-400%	1.23	1.56	1.14	1.22
400%+	1.30	1.44	1.13	1.21
Metro Status				
Metro	1.41	1.61	1.41	1.55
Non-Metro	1.19	1.68	1.11	1.20
Urbanicity				
Urban	1.37	1.60	1.42	1.54
Rural	1.06	1.44	0.98	1.05
Labor Force				
Full-Time Work	1.42	1.48	1.15	1.31
Part-Time Work	1.39	1.69	1.10	1.25
In School	0.97	1.28	1.02	1.10
Other	1.43	1.54	1.21	1.41

[1] Widowed/Divorced/Separated.

Table 2. Generalized standard error for estimated percentages and corresponding sample sizes from the respondent file: 1995 NSFG, Cycle 5.

Hispanic Women

	Percentage									
Sample Size	50%	45 or 55%	40 or 60%	35 or 65%	30 or 70%	25 or 75%	20 or 80%	15 or 85%	10 or 90%	5 or 95%
100	5.42	5.51	5.50	5.39	5.17	4.83	4.35	3.73	2.91	1.82
200	3.90	3.96	3.95	3.87	3.72	3.47	3.13	2.68	2.09	1.31
300	3.21	3.27	3.26	3.19	3.06	2.86	2.58	2.21	1.72	1.08
400	2.80	2.85	2.84	2.79	2.67	2.50	2.25	1.93	1.50	0.94
500	2.52	2.56	2.56	2.50	2.40	2.24	2.02	1.73	1.35	0.85
600	2.31	2.35	2.34	2.30	2.20	2.06	1.86	1.59	1.24	0.78
700	2.15	2.18	2.18	2.13	2.05	1.91	1.72	1.48	1.15	0.72
800	2.01	2.05	2.04	2.00	1.92	1.79	1.62	1.38	1.08	0.68
900	1.90	1.94	1.93	1.89	1.82	1.70	1.53	1.31	1.02	0.64
1,000	1.81	1.84	1.84	1.80	1.73	1.61	1.46	1.25	0.97	0.61
1,100	1.73	1.76	1.76	1.72	1.65	1.54	1.39	1.19	0.93	0.58
1,200	1.66	1.69	1.69	1.65	1.58	1.48	1.33	1.14	0.89	0.56
1,300	1.60	1.62	1.62	1.59	1.52	1.42	1.28	1.10	0.86	0.54
1,400	1.54	1.57	1.57	1.53	1.47	1.37	1.24	1.06	0.83	0.52
1,500	1.49	1.52	1.52	1.48	1.42	1.33	1.20	1.03	0.80	0.50
1,553	1.47	1.49	1.49	1.46	1.40	1.31	1.18	1.01	0.79	0.49

For the above table, the coefficients in equation (4) are: b_0=0.4279; b_1=1.0017; b_2=0.5047; b_3=0.0479

non-Hispanic Black Women

	Percentage									
Sample Size	50%	45 or 55%	40 or 60%	35 or 65%	30 or 70%	25 or 75%	20 or 80%	15 or 85%	10 or 90%	5 or 95%
100	5.75	5.76	5.71	5.60	5.40	5.13	4.75	4.25	3.57	2.59
200	4.13	4.15	4.11	4.03	3.89	3.69	3.42	3.06	2.57	1.86
300	3.41	3.42	3.39	3.32	3.21	3.04	2.82	2.52	2.12	1.53
400	2.97	2.98	2.96	2.90	2.80	2.65	2.46	2.20	1.85	1.34
500	2.68	2.68	2.66	2.60	2.52	2.39	2.21	1.98	1.66	1.20
600	2.45	2.46	2.44	2.39	2.31	2.19	2.03	1.81	1.52	1.10
700	2.28	2.29	2.27	2.22	2.14	2.03	1.89	1.69	1.42	1.03
800	2.14	2.15	2.13	2.08	2.01	1.91	1.77	1.58	1.33	0.96
900	2.02	2.03	2.01	1.97	1.90	1.81	1.67	1.50	1.26	0.91
1,000	1.92	1.93	1.91	1.87	1.81	1.72	1.59	1.42	1.20	0.87
1,100	1.84	1.84	1.83	1.79	1.73	1.64	1.52	1.36	1.14	0.83
1,200	1.76	1.77	1.75	1.72	1.66	1.57	1.46	1.31	1.10	0.79
1,300	1.70	1.70	1.69	1.65	1.60	1.52	1.40	1.26	1.06	0.76
1,400	1.64	1.64	1.63	1.60	1.54	1.46	1.36	1.21	1.02	0.74
1,600	1.54	1.54	1.53	1.50	1.45	1.37	1.27	1.14	0.96	0.69
1,800	1.46	1.46	1.45	1.42	1.37	1.30	1.20	1.08	0.90	0.65
2,000	1.38	1.39	1.38	1.35	1.30	1.24	1.14	1.02	0.86	0.62
2,200	1.32	1.33	1.32	1.29	1.24	1.18	1.09	0.98	0.82	0.59
2,400	1.27	1.27	1.26	1.24	1.19	1.13	1.05	0.94	0.79	0.57
2,446	1.26	1.26	1.25	1.22	1.18	1.12	1.04	0.93	0.78	0.57

For the above table, the coefficients in equation (4) are: b_0=0.0876; b_1=0.1915; b_2=0.0262; b_3=0.0495.

Table 2. Generalized standard error for estimated percentages and corresponding sample sizes from the respondent file: 1995 NSFG, Cycle 5 (continued).

All Women and White Women

Sample Size	Percentage									
	50%	45 or 55%	40 or 60%	35 or 65%	30 or 70%	25 or 75%	20 or 80%	15 or 85%	10 or 90%	5 or 95%
100	4.95	4.93	4.86	4.73	4.54	4.28	3.94	3.51	2.92	2.09
200	3.61	3.60	3.54	3.45	3.31	3.13	2.88	2.56	2.13	1.53
300	3.00	2.99	2.95	2.87	2.76	2.60	2.39	2.13	1.77	1.27
400	2.64	2.63	2.59	2.52	2.42	2.28	2.10	1.87	1.56	1.11
500	2.38	2.37	2.34	2.28	2.18	2.06	1.90	1.69	1.41	1.01
600	2.19	2.18	2.15	2.09	2.01	1.90	1.75	1.55	1.29	0.93
700	2.04	2.04	2.01	1.95	1.87	1.77	1.63	1.45	1.21	0.86
800	1.92	1.92	1.89	1.84	1.76	1.66	1.53	1.36	1.14	0.81
900	1.82	1.82	1.79	1.74	1.67	1.58	1.45	1.29	1.08	0.77
1,000	1.74	1.73	1.71	1.66	1.59	1.50	1.38	1.23	1.03	0.73
1,200	1.60	1.59	1.57	1.53	1.47	1.38	1.27	1.13	0.94	0.68
1,600	1.40	1.40	1.38	1.34	1.29	1.21	1.12	0.99	0.83	0.59
2,000	1.27	1.26	1.24	1.21	1.16	1.10	1.01	0.90	0.75	0.54
3,000	1.05	1.05	1.04	1.01	0.97	0.91	0.84	0.75	0.62	0.45
4,000	0.93	0.92	0.91	0.88	0.85	0.80	0.74	0.66	0.55	0.39
5,000	0.84	0.83	0.82	0.80	0.77	0.72	0.67	0.59	0.49	0.35
6,000	0.77	0.77	0.76	0.74	0.71	0.67	0.61	0.54	0.45	0.33
8,000	0.68	0.67	0.66	0.65	0.62	0.58	0.54	0.48	0.40	0.29
10,000	0.61	0.61	0.60	0.58	0.56	0.53	0.49	0.43	0.36	0.26
10,847	0.59	0.59	0.58	0.56	0.54	0.51	0.47	0.42	0.35	0.25

For the above table, the coefficients in equation (4) are: b_0=-0.1513; b_1=0.0810; b_2=0.0493; b_3=0.0908.

Table 3. Comparison of Simple Random Sample (SRS), Design-Based (SUDAAN), and Generalized (GSE) Standard Errors for Percent of Pill Use Estimates by Age Category and White Protestants and Catholics: 1995 NSFG, Cycle 5.

Reporting Domain	Sample Size	Percent Using The Pill	Standard Errors		
			SRS	Design-Based[1]	GSE
Total	10,847	74.3%	0.17	0.49	0.54
15-19	1,416	29.4%	1.47	1.25	1.38
20-24	1,519	73.6%	1.28	1.30	1.21
25-29	1,739	83.6%	0.79	1.10	0.99
30-34	2,148	85.3%	0.58	0.88	0.90
35-39	2,144	84.0%	0.63	0.87	0.90
40-44	1,881	83.8%	0.72	0.93	0.90

[1] Design-based variances were produced using SUDAAN.