WHY LARGE DESIGN EFFECTS CAN OCCUR IN COMPLEX SAMPLE DESIGNS: EXAMPLES FROM THE NHANES 1999-2000 SURVEY

David A. Lacher, Lester R. Curtin, Jeffery P. Hughes
Centers for Disease Control and Prevention
National Center for Health Statistics
3311 Toledo Road, Room 4215, Hyattsville, MD 20782

KEY WORDS: Design effect, complex sample, variance estimation, NHANES, laboratory testing

Introduction

For complex sample surveys, the design effect (DEFF) for a point estimate is defined as the ratio of the "true" design-based sampling variance to the hypothetical sampling variance assuming the same point estimate based on a simple random sample of the same size (1,2). Design effects are used for a variety of purposes including comparing alternative survey designs, determining the effective sample size for analysis, and adjusting confidence intervals for estimates based on complex survey designs. Large or small design effects may also indicate data problems including potential sample design problems or potential data quality problems. For example, an outlier, such as an extreme data value, an influential sample weight or cluster, can lead to an unusually large DEFF. For laboratory measures, a large DEFF may occur when there is a change in method detected by a drift in the quality control for the test. In the analysis of the National Health and Nutrition Examination Survey (NHANES) survey data, large design effects have been observed for several laboratory measurements. The paper presents results from one of a series of investigations to explain the large mean design effects seen in NHANES 1999-2000.

The National Health and Nutrition Examination Survey 1999-2000 is the latest in a series of multi-purpose nationally representative health surveys conducted by the National Center for Health Statistics of the Centers for Disease Control and Prevention. NHANES data come from interviews, examinations, and laboratory tests on biological specimens. The sample design is a multistage, area cluster design with differential probabilities of selection for designated demographic groups. Design effects of the means of variables have been examined for the NHANES 1999-2000 survey (3). Although design effects of the estimated means for most variables in NHANES 1999-2000 were less than 3, design effects for means of some laboratory tests were much higher, ranging up to 40.

Previously, the large design effects were examined in terms of the sample design characteristics. Specifically, the within- and between-PSU variation was examined (4). Most of the large design effects occurred when the within-PSU component was relatively small and the between-PSU component was relatively large. Typically, the NHANES geographic heterogeneity is not captured in the hypothetical simple random sample variance assumption, so the DEFF is inflated even though the total sample error is small.

Investigation of large design effects did reveal one particular data quality problem. This was exemplified by serum sodium for NHANES 1999-2000. Serum sodium levels do not vary greatly in a healthy population. After observing a large DEFF for the mean concentration of sodium, the quality control data were re-examined. It was determined that there was a slight drift in the serum sodium method. The difference was not clinically important and no adjustment to the data was needed, but the natural variation in sodium is so small that the estimated standard errors were inflated.

The examination of the sodium data motivated the research summarized in this paper. Specifically, this paper examines the effect of laboratory test method bias (accuracy) and precision (reproducibility) on the design effects of the mean. The effects are evaluated using simulated examples based on the NHANES 1999-2000 sample design. In addition, the relationship of the design effects of the mean of laboratory tests is compared to the relative standard errors (RSE).

Background

The conceptual basis for the DEFF originated with Cornfield (5). The use of DEFF for basic

descriptive survey estimates has been popularized by Kish (1,2). The concept of DEFF has been extended to analytic methods, for example, in regression estimators (6). Generalized DEFF in multivariate models were developed by Rao and Scott (7). Most standard textbooks on sample survey design and analysis now discuss DEFF (8,9). Model based motivation of the typical form of the DEFF, summarizing the effect of differential weighting and a one-stage cluster design, has been provided by Gabler (10) and Spenser (11).

The design effect for a complex sample design is defined as Var_{design}/Var_{srs}, where Var_{design} is the variance of the estimate from the sampling design and Var_{srs} is the variance assuming a simple random sample (srs) at the same point estimate with same number of observation units. The design effect of the mean:

$$deff(\hat{y}_{mean})=V(\hat{y}_{mean})/[(1-(n/N))(s_{srs}^2/n)]$$

where $V(\hat{y}_{mean})$ is the variance of the estimate of the mean, n/N is the sample fraction, and s_{srs}^2 is the variance assuming a simple random sample (8).

NHANES 1999-2000 was a cross-sectional survey that collected data on the civilian noninstitutionalized U.S. population through questionnaires and medical examinations including laboratory tests. NHANES 1999-2000 used a stratified, multistage probability design to collect a nationally representative sample. Beginning in 1999, NHANES became a continuous survey. The procedures followed to select the sample and conduct the interviews and examinations for NHANES 1999-2000 were similar to those for previous NHANES surveys. NHANES 1999–2000 was designed to over-sample Mexican Americans, non-Hispanic blacks, adolescents, elderly persons 60 years and older, pregnant females, and beginning in the year 2000 low income non-Hispanic whites to improve estimates for these groups. The NHANES 1999-2000 survey consisted of 12160 identified sample persons of all ages, of which 9965 (81.9%) were interviewed and 9282 (76.3%) were interviewed and examined.

The laboratory component of NHANES 1999-2000 consisted of tests performed on blood samples, urine, nasal and vaginal swabs and environmental samples. The laboratory component consisted of general biochemical, nutritional, immunological, environmental, microbiological, and hematological tests.

Methods

NHANES Laboratory Data Collection and Analysis

Laboratory data were collected from NHANES 1999-2000 data files (12). All available laboratory data were used without deletion of outliers. Laboratory methods are described in the laboratory procedure manual for NHANES (13). Statistical analysis was performed using SAS for Windows software (SAS Institute, Cary, NC) and SUDDAN software (RTI, Research Triangle Park, NC). Some tests were non-Gaussian as judged by skewness and kurtosis. These tests were log-transformed to obtain the geometric mean and standard error of the geometric mean. For each test, the sample size, arithmetic or geometric mean, standard error of the mean (SEM), relative standard error of the mean, and the design effect of the mean was calculated. The relative standard error is defined as the standard error of the mean divided by the estimate of the mean. The MEC examined sample weights (WTMEC2YR) were used which takes into account differential probabilities of selection from the sample design, from nonresponse, and from oversampling of subgroups (14). The weights also reflected the complex sample design of NHANES. Standard errors were calculated using the SUDDAN delete 1 jackknife (JK1) method (15). The JK1 method partitioned the sample into 52 sampling units and 52 replicates deleting one unit at a time.

Simulation of Laboratory Test Bias and Precision on Mean DEFF

The effect of bias and precision of laboratory tests on the design effect of the mean was simulated. The NHANES 1999-2000 JK1 weights and sample sizes for 27 stands (survey locations representing 26 PSUs) were used as the sampling design for the simulation. Simulated test data were formed using a Gaussian distribution with a mean of 100 and a standard deviation of 1 for 22 of 27 stands. For 5 other consecutive stands, a method bias of 0, 1, 2, 3, 4, or 5 was added to the mean. For each bias, the method precision was simulated using standard deviations of 0.5, 1, 2, 3, 4, or 5. The design

effect of the mean was calculated for each simulation.

Results

The descriptive statistics and design effect of the mean for 55 laboratory tests for NHANES 1999-2000 are seen in Table 1. The sample size, mean, standard error of the mean, and relative standard error were calculated for each test. The relative standard error of the tests ranged from 0.12% for serum sodium to 8.05% for blood total mercury. The design effects of the mean ranged from 1.56 for lymphocyte percent to 39.52 for the mean corpuscular hemoglobin concentration. The design effect of the mean was compared to the relative standard error for each test (Figure 1). In general, there was an indirect relationship between the relative standard error and the design effect of the mean.

The simulation of the effects of bias and precision of laboratory tests on the design effect of the mean is seen in Figure 2. A mean DEFF of 1.80 was seen with no bias present and a precision of 1%. The mean DEFF increased dramatically with the introduction of bias for each level of precision. For a RSE of 1%, the mean DEFF increased from 1.80 when no bias was present to 73.6 for a bias of 5%. The effect of bias on the mean DEFF moderated when more imprecision was introduced. For a RSE of 5%, the mean DEFF increased from 1.70 with no bias to 46.4 for a bias of 5%. The mean DEFF decreased as more imprecision was introduced for a given bias. For a bias of 1%, the mean DEFF decreased from 11.9 for a RSE of 0.5% to mean DEFF of 6.2 for a RSE of 5%. For a bias of 5%, the mean DEFF decreased from 75.0 for a RSE of 0.5% to a mean DEFF of 46.4 for a RSE of 5%.

Discussion

The design effect is a variance correction factor when using complex survey designs instead of the simple random sample design. The design effect is used in complex surveys to adjust confidence intervals around estimates, determine effective sample size of analysis and compare alternate designs for fixed costs and for different variables. Also, averaging design effects is used in general variance functions. NHANES 1999-2000 was designed to have design effects below 3.0 for most estimates. Design effects for subgroups are higher than DEFFs for the total sample in NHANES because they are over-sampled or under-sampled which leads to inflated variances. Also, design effects are increased in NHANES compared to simple random samples due to positive correlations among sample persons in the same area cluster. In addition, large design effects can occur with large between-PSU and small within-PSU components of sampling error which is exacerbated by the relatively small number of PSUs (26) found in NHANES 1999-2000.

Laboratory tests are subject to systematic error (bias) and random error (precision) (16). In addition, biological variation between persons and within a person are components of variation for laboratory tests. Furthermore, laboratory tests are subject to pre-analytical variation (improper collection of specimens) or post-analytical errors (reporting of results). In NHANES 1999-2000, large mean design effects for laboratory tests ranging up to 40 were seen (Table 1). Laboratory tests with small RSE tended to have the largest mean DEFF (Figure 1). Tests such as serum electrolytes (sodium, potassium, bicarbonate, and chloride) and serum calcium are known to be tightly regulated in the body to maintain physiologic homeostasis. For example, serum sodium for NHANES 1999-2000 had a sample mean of 139.4 mEq/L with a standard deviation of 2.5 mEq/L with a between-person coefficient of variation (CV) of 1.8%.

Many laboratory test methods seen in Table 1 have excellent precision with analytical coefficients of variation below 5%. For example, serum sodium had an analytical CV of 1.3% for NHANES 1999-2000. If analytical bias is introduced in tests that are very precise, there is a potential for elevated mean DEFF due to increased between-PSU variation for tests with small RSE. For NHANES 1999-2000, the mean serum sodium for 27 stands ranged from 136.9-141.5 mEq/L. The mean serum sodium for 4 consecutive stands showed a downward systematic shift of -2.0. This was caused by a method shift of -2.0 in the serum sodium as detected by quality controls during these 4 stands. Although a negative bias of -2.0 had no clinical significance, it had a significant effect on the mean design effect for serum sodium. When 2.0 was added to every sample person in the 4 stands with analytical bias, the mean design effect for serum sodium decreased from 25.9 to 16.6.

A systematic approach examining the effect of laboratory test bias and precision on the design effect of the mean was done through simulation based on the NHANES 1999-2000 sample design (Figure 2). The effect of test bias had a more profound effect than test imprecision. An increase in bias increases the design effect of the mean for a given level of precision. An increase in imprecision led to a smaller mean DEFF. Also, a change in bias has the most influence on the design effect when the test is more precise. Changes in laboratory test method performance should be investigated as a cause of extreme mean design effects.

References

1. Kish L. Survey sampling. New York: John Wiley, 1965.

2. Kish L. Methods for design effects. J Official Statistics 1995; 11: 53-77.

3. National Center for Health Statistics. 1999-Current National Health and Nutrition Examination Survey (NHANES). http://www.cdc.gov/nchs/about/major/nhanes/currentnhanes.htm (Accessed September 2004).

4. Curtin L, Carroll M, Dohrmann S, Winters F, Lacher D. Extreme design effects, why they occur and what to do about them. Presented at the Joint Statistical Meetings, New York, NY, 2002.

5. Cornfield J. Statistical relationships and proof in medicine. .American Statistician 1954; 8: 19-21.

6. Kish L, Frankel MR. Inference from complex samples. J. Royal Statistical Society, Series B 1974; 36: 1-37.

7. Rao JNK, Scott AJ. On simple adjustments to chi-square tests with sample survey data. Annals of Statistics 1987;15: 385-97.

8. Lohr SL. Sampling: design and analysis. Pacific Grove, CA: Brooks/Cole Publishing Co., 1999.

9. Korn E, Graubard B. Analysis of health surveys. New York: John Wiley and Sons, 1999.

10. Gabler S, Haeder S, Lahiri P. A model based justification of Kish's formula for design effects for weighting and clustering. Survey Methodology 1999; 25: 105-6.

11. Spenser B. An approximate design effect for unequal weighting when measurements may correlate with selection probabilities. Survey Methodology 2000; 26: 137-8.

12. National Center for Health Statistics. NHANES 1999-2000 Data Files, Data Docs, Codebooks, SAS Code. http://www.cdc.gov/nchs/about/major/nhanes/NHANES99_00.htm (Accessed September 2004).

13. National Center for Health Statistics. Laboratory Procedures Manual. http://www.cdc.gov/nchs/about/major/nhanes/lab_methods.htm (Accessed September 2004).

14. National Center for Health Statistics. NHANES 1999-2000 Addendum to the NHANES III Analytic Guidelines. http://www.cdc.gov/nchs/data/nhanes/guidelines1.pdf (Accessed September 2004).

15. Wolter KM. Introduction to variance estimation. New York: Springer-Verlag, 1990.

16. Burtis CA, Ashwood ER, eds., Tietz textbook of clinical chemistry, 2nd Ed., Philadelphia: W.B. Saunders Co., 1994.

Table 1. Descriptive statistics and design effects of the mean for laboratory tests for NHANES 1999-2000 sorted by design effect

Test[1]	N	Mean	SEM	RSE	Mean DEFF
Lymphocyte Percent	7949	31.4	0.133	0.42	1.56
Aspartate Aminotransferase*	6231	22.6	0.13	0.56	1.69
Neutrophil Percent	7949	56.9	0.173	0.30	2.12
Glucose, serum*	6231	91.0	0.34	0.37	2.24
Monocyte Percent*	7949	7.86	0.037	0.47	2.33
Uric Acid	6231	5.27	0.030	0.57	2.51
Microalbumin, urine*	7606	8.8	0.21	2.40	2.59
Creatinine, serum*	6231	0.68	0.005	0.67	2.77
Alanine Aminotransferase*	6231	21.7	0.24	1.09	2.88
Gamma Glutamyl Transferase*	6231	21.6	0.31	1.41	2.95
Total Bilirubin*	6231	0.52	0.005	0.96	3.04
Iron, serum	7877	89.4	0.74	0.83	3.06
Eosinophil Percent	7949	2.98	0.050	1.68	3.22
Total Cholesterol	7420	193.9	0.89	0.46	3.42
Homocysteine*	7599	6.76	0.058	0.86	3.51
Ferritin*	7860	61.1	1.41	2.31	3.53
Blood Urea Nitrogen*	6231	13.0	0.10	0.80	3.73
Vitamin B12*	7524	487.3	5.0	1.02	3.76
Phosphorus	6231	3.52	0.014	0.40	3.85
Mean Corpuscular Volume	7982	89.3	0.12	0.13	3.98
Fibrinogen*	2680	352.4	2.89	0.82	4.19
Platelet Count	7982	272.7	1.57	0.58	4.35
C -reactive protein*	7493	0.14	0.005	3.58	4.49
Hematocrit	7982	42.0	0.10	0.24	4.74
Bone Alkaline Phosphatase*	6365	19.0	0.33	1.76	4.85
Red Blood Cell Count	7982	4.71	0.012	0.25	4.90
Alkaline Phoshatase*	6231	82.1	0.97	1.18	4.97
Triglycerides*	6233	110.8	1.85	1.67	5.06
Methylmalonic Acid*	7598	0.13	0.0016	1.20	5.39
White Blood Cell Count	7981	7.34	0.056	0.76	5.46
N-Telopeptides, urine*	6721	426.1	14.4	3.38	5.47
Creatinine, urine	7606	136.4	2.31	1.69	5.54
Mean Corpuscular Hemoglobin	7982	30.2	0.055	0.18	5.55
Erythrocyte Protoporphyrin*	7985	47.6	0.43	0.90	5.93
Hemoglobin	7982	14.2	0.040	0.28	6.18
Lactate Dehydrogenase*	6231	147.9	0.96	0.65	6.79
Total Iron Binding Capacity	7847	370.5	1.89	0.51	7.75
Varicella Antibody	5113	13.5	0.26	1.93	7.78
Measles Antibody	5113	10.4	0.28	2.69	8.01
HDL Cholesterol	7415	47.8	0.48	1.00	8.89
Basophil Percent	7949	0.65	0.015	2.31	9.34
Blood Lead*	7970	1.66	0.038	2.28	9.50
Folate, serum*	7526	14.04	0.27	1.95	9.76
Mean Platelet Volume	7982	8.21	0.031	0.38	9.96
Total Mercury*	2414	0.85	0.07	8.05	10.59
Rubella Antibody	5113	4.98	0.169	3.39	10.95

Test[1]	N	Mean	SEM	RSE	Mean DEFF
Total Protein	6231	7.53	0.019	0.25	11.49
Potassium	6231	4.14	0.014	0.34	11.79
Folate, Red Blood Cell*	7614	280.8	4.54	1.62	13.78
Albumin, serum	6231	4.51	0.016	0.35	14.89
Calcium	6231	9.46	0.027	0.29	25.63
Sodium	6231	139.4	0.17	0.12	25.94
Bicarbonate	6231	23.8	0.16	0.67	30.02
Chloride	6231	102.4	0.23	0.22	34.10
Mean Corpuscular Hemoglobin Conc.	7982	33.8	0.055	0.16	39.52

[1] *indicates tests were non-Gaussian and were log-transformed to obtain geometric mean and stand error of the geometric mean

Figure 1. Laboratory tests mean design effects vs. relative standard error

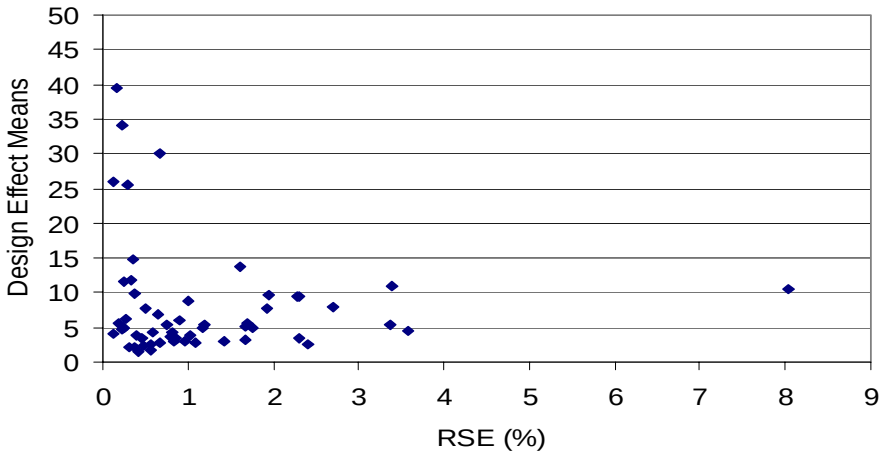

Figure 2. Effect of laboratory test bias and precision on the design effect of the mean