Complex Sampling: Implications for Data Analysis

Daniel J. Caplan, DDS, PhD; Gary D. Slade, BDSc, DDPH, PhD; Stuart A. Gansky, MS, DrPH

Abstract

Investigators in dental public health often use strategies other than simple random sampling to identify potential subjects; however, their statistical analyses do not always take into account the complex sampling mechanism. Often it is not clear whether a given strategy requires adjustment for stratification and/or cluster sampling of observations. We propose that the need for such adjustment depends on the primary study objective. As a general rule, we recommend that if the study goal is to estimate the magnitude of either a population value of interest (e.g., prevalence), or an established exposure-outcome association, adjustment of variances to reflect complex sampling is essential because obtaining appropriate variance estimates is a priority. However, if the study goal is to establish the presence of an association, especially in a preliminary investigation of novel conditions or understudied populations, obtaining appropriate variance estimates may not be of primary importance; hence, adjustment of variances for complex sampling is not always required, but often is recommended. This paper describes several types of complex sampling designs, methods of adjusting for complex sampling strategies, examples illustrating the effect of adjustment, and alternative approaches for analysis of complex samples. [J Public Health Dent 1999;59(1):52-59]

Key Words: cluster, epidemiology, exposure, outcome, parameter, risk factors, sampling, statistics, stratification, survey, variance.

Researchers in dental public health, oral epidemiology, and dental health services research often select a sample of individuals from some population, calculate statistics from sample data, and generalize their findings to the population from which the sample was drawn or to other external populations. The purpose of sampling is to estimate a parameter, or population value, from values obtained from a subset of that population. Sampling is undertaken to reduce the resources needed to obtain data about the population while maintaining reasonable confidence that the estimate approximates the population value.

Common sense dictates, and statistical methods require, that subjects comprising the sample be selected at random from the target population (the population about which we would like to make inferences) (1). This objective can be achieved in numerous ways. Because the strategy used to analyze data should mirror the strategy used to collect them, both the sampling mechanism and the primary study goal should influence the analytic methods employed.

The purpose of this paper is to describe several common types of samples, to review methods of accounting for complex sampling strategies, to present an example illustrating the effect of such adjustment, and to recommend alternative analytic approaches for investigators who do not have the resources needed to perform such adjustment. This review does not summarize these issues exhaustively, but may help dental public health practitioners and researchers better conceptualize the manner in which data analyses hinge on study goals and sampling strategies. Readers need not have advanced statistical knowledge to understand this paper; however, some elementary statistical knowledge and vocabulary is assumed.

Classification of Sampling

Nonprobability Sampling. In nonprobability sampling, individual selection is by nonrandom methods, and potential subjects do not have known selection probabilities because the number of individuals in the target population is unknown. One common example is a convenience, haphazard, or fortuitous sample (2), in which subjects are selected based not on chance but on ease of data acquisition. Other types of nonprobability samples include purposive, expert choice, judgment, and quota samples (2).

Results from nonprobability samples may demonstrate a biologic or therapeutic principle by showing, for example, that a given intervention can be efficacious in preventing a disease within a group of volunteers. Yet it would be inappropriate to conclude that the intervention will be equally effective in the entire target population because the volunteers were not sampled randomly and thus cannot be considered representative of the target population. Because inferences drawn from data analyses are based on the observed sample only, nonprobability samples cannot generate estimates that apply to larger populations. Generalizability can be claimed only through nonstatistical arguments, which may not be particularly convincing. For this reason, nonprobability samples will not be discussed further.

Probability Sampling. Probability sampling implies that each individual in the target population has a known and nonzero probability of being selected into the sample (1,2). Two types of probability sampling include random sampling and survey sampling.

In random sampling it is assumed

Send correspondence to Dr. Caplan, Department of Dental Ecology, School of Dentistry, University of North Carolina, CB #7450, Chapel Hill, NC 27599-7450. E-mail: dan_caplan@unc.edu. No reprints will be available. Dr. Slade is with the Department of Dental Ecology, University of North Carolina at Chapel Hill, and Dr. Gansky is with the Department of Dental Public Health and Hygiene, University of California, San Francisco. Web site: http://www.dent.unc.edu/depts/. Manuscript was an invited submission, but underwent peer review. Received: 1/12/99; accepted for publication: 4/12/99.

that the population is infinite and abstract. In contrast, survey sampling assumes that the population is finite and well defined, and though each element in the target population has a known and nonzero probability of selection, selection probabilities may differ among individuals. Survey sampling, rather than random sampling, will be the focus of this paper because so many research efforts in dental public health use survey sampling methods.

In general, survey samples can be classified into one of four types.

Equal Probability Selection Method Samples. In these samples, also called "EPSEM" or "self-weighting" samples, selection probabilities for all elements in the target population are equal. Specifically, the probabilities are equal to the inverse of the number of elements in the sampling frame, i.e., the roster listing all elements in the target population. Depending on the study objectives, population elements can be communities, schools, people, teeth, or some other unit; most often, individuals are the population elements listed in the sampling frame.

A simple random sample (SRS) is a special type of EPSEM sample in which the population element is the only unit sampled—for example, individuals are sampled, rather than towns, households, schools, or some other unit composed of more than one individual. An SRS was used to assess the prevalence of apical periodontitis among persons aged 30–39 years residing in Porto, Portugal (3). Here, a random drawing of 0.5 percent of the target population was carried out using electorate lists as a sampling frame.

Stratified Samples. Stratified sampling occurs when a target population is stratified, or divided, into two or more mutually exclusive and exhaustive subgroups, and samples are selected within each stratum. Ideally, elements within strata are relatively similar to each other compared with elements from other strata—in other words, stratified samples are most efficient when the strata are internally homogeneous but externally heterogeneous. Stratified samples require both the existence of a sampling frame and a record of each individual's value(s) for the stratification factor(s) so that each potential subject can be placed into a stratum prior to selection of subjects. A stratified sample may use any kind of probability sampling within strata.

Some case-control studies (4) employ stratified sampling. For example, in an investigation of factors related to loss of root canal treated teeth, Caplan and Weintraub (5) used a treatment database to identify a population of HMO patients who underwent endodontic therapy. Next, patients were separated into two mutually exclusive subgroups: those who lost the root canal filled tooth within a specified time interval after treatment (cases) and those who did not (controls). Finally, SRSs were selected from each stratum, and patient- and tooth-level exposures were compared between the two groups.

A stratified sample often is sought when good precision around an estimated parameter is desired but the number of potential subjects in one subgroup is relatively small, or when the study objective is to generate separate estimates for different subgroups. For example, with funding to perform 200 clinical dental examinations and a goal of making separate estimates of the number of decayed, missing, or filled surfaces (DMFS) among Medicaid[1] recipients and nonrecipients in a population of 200 recipient and 800 nonrecipient children, one can randomly select 100 recipient and 100 nonrecipient children. Three aspects of this example should be emphasized:

1. If "Medicaid status" is not already recorded in the sampling frame, one cannot separate the population into the two subgroups for sampling within strata.

2. The total sample is called a stratified SRS because SRSs are selected from each stratum.

3. The proportion of each stratum selected, or sampling fraction, for Medicaid recipients is 100/200=50 percent, while that for nonrecipients is 100/800=12.5 percent. The impact of the sampling fraction on parameter and variance estimates will be discussed later.

Cluster Samples. In cluster samples, groups of population elements are selected, rather than single population elements, with selected groups sharing some common feature(s). Grouping is usually based on region, school, or work place, with the goal of reducing travel and other costs associated with data collection. For example, to examine a cluster sample of schoolchildren, one first can select a sample of schools (clusters) from a list of all schools in the sampling frame, then examine a sample of children within each selected school. Here, children are "clustered" within schools, since schools were sampled and children were examined solely due to their attending selected schools. Three points are worth noting:

1. This sample of schoolchildren satisfies requirements for probability sampling, provided that schools were selected with a known, nonzero probability from the entire population of schools within the region of interest.

2. Cluster samples can include more than one stage, or level, of sampling. In the above example, if 10 percent of each school's students were sampled and examined rather than all children attending the selected schools being examined, the sample would be called a two-stage cluster sample as opposed to a one-stage cluster sample. SRSs are one-stage samples with no clustering.

3. A common example in dental research where the individual represents a cluster involves measurement and analysis of multiple sites within a person. Examples are plentiful, especially in periodontal studies involving assessment of periodontal probing depth or clinical attachment level, where as many as six sites per tooth are analyzed (6).

As a general rule, any sample for which multiple observations per individual are assessed and analyzed should be considered a cluster sample because the assumption of independence among observations is violated. Satisfaction of the independence assumption means that the value of any observation is in no way affected by, or related to, measurements of any other observation. Conventional wisdom in periodontal research holds that clinical attachment level at sites within a mouth are correlated due to person-level factors that affect periodontal disease, such as host resistance, genetics, oral hygiene habits, and diet (7,8). In the school example,

[1]In the US, Medicaid is a program partially funded by the federal government and administered by the states that provides health insurance for certain individuals in need of assistance, primarily those with low incomes.

DMFS among children attending a single school may be correlated due to school-level factors that affect DMFS, such as drinking water fluoride concentration or schoolwide fluoride mouthrinsing programs.

Stratified Cluster Samples. As the name implies, these samples are a combination of startified samples and cluster samples. The North Carolina 1986–87 School Oral Health Survey (9) represents a stratified one-stage cluster sample in which 16 mutually exclusive strata were identified based on cross-classifying four school system geographic regions with two school urbanization categories and two school racial makeup designations (4 x 2 x 2=16). After teachers within each stratum were selected, all students in selected teachers' classrooms were asked to participate. Here, students were clustered within teachers who were stratified by their schools' geographic region, urbanization, and racial makeup.

Many in the dental public health community are familiar with the WHO "pathfinder" survey methodology (10). Briefly, this technique involves sampling a specified number of individuals in certain age groups, geographic locations, or residence types for the purposes of program planning. The pathfinder methodology also has been employed to help procure resources, increase visibility of dental public health programs, and educate the public regarding the benefits of good oral health (11). According to the WHO, the pathfinder methodology is "a practical, economic survey sampling methodology [that uses] a stratified cluster sampling technique, which aims to include the most important population subgroups likely to have differing disease levels" (10).

While this method may be useful for other purposes, it should not be used to estimate prevalence or incidence of disease in populations unless the denominator (the total number of people in the population represented by the sample) is known for each cluster and stratum. In the context of the present paper, WHO's use of the term "stratified cluster sampling" is misleading, because it implies that subjects are selected as a probability sample and thus have a known and nonzero probability of selection. This generally is not true in studies that use the pathfinder method. Though subjects are stratified by age, geographic location, or residence status and might be examined in clusters such as schools or factories, selection into the sample generally is a purposive or judgment sample based on convenience, and examined subjects may not be representative of any larger group. Thus, pathfinder samples should be considered nonprobability samples that should not be used to estimate population values such as prevalence or incidence of disease.

Hereafter, the term complex sample will refer to all probability survey sample types mentioned except SRSs.

Variance Estimation for Complex Samples: Issues

Random Variation, Standard Error, and Confidence Intervals. In an EPSEM/SRS, sample statistics should approximate population parameters because sampled individuals were chosen at random with an equal probability of being selected. For example, a 10 percent SRS from the population of 200 Medicaid recipients and 800 nonrecipients mentioned before likely would contain about 20 recipient and 80 nonrecipient children—the same percentage of each subgroup found in the population.

The population value of 20 percent Medicaid recipients does not fluctuate regardless of the population members sampled. In contrast, the estimate obtained from the sample is expected to vary due to random variation, a term reflecting the fact that different samples of identical sizes from a population would include different children and thus could produce different estimates. By chance, one could draw 10 percent SRSs consisting of 100 Medicaid recipients, 100 nonrecipients, or any combination of 100 children; however, "on average" the sample would contain 20 percent Medicaid recipients. The extent of random variation around the 20 percent parameter estimate is expressed mathematically by the variance, which can be viewed as the amount one could expect the estimate to vary if a sample of that size were repeatedly selected at random from the target population. The standard error, a frequently reported measure of random variation, is calculated as the square root of the variance divided by the square root of the number of elements sampled.

Sample data are used to generate a parameter estimate and its associated standard error. These two numbers then can be used to calculate a confidence interval (CI), with greater standard errors leading to wider CIs. Usually, CIs are expressed as 95 percent CIs, where 95 percent implies that if 100 samples of size n were drawn from that population, CIs generated from 95 of those samples would contain the population value. Practically speaking, the 95 percent CI can be viewed as a range in which it is reasonably likely that the true population parameter lies. In the population containing 20 percent Medicaid recipients, 95 of every 100 samples drawn would generate 95 percent CIs containing the value 20 percent.

Weights. In the previous example of a stratified SRS, 100 of 200 Medicaid recipients and 100 of 800 nonrecipients were selected at random, resulting in sampling fractions of 50 percent and 12.5 percent for the two groups, respectively. Sampling fractions affect parameter and variance estimates because the weight of each sampled individual (i.e., the number of individuals in that subgroup represented by that individual) is equal to the inverse of his or her subgroup's probability of selection—here just the sampling fraction. With our stratified SRS example, the statistical likelihood of selection of 50 percent for the Medicaid recipient subgroup results in a weight of 1.0/0.5=2.0 for each sampled recipient, while the statistical likelihood of selection of 12.5 percent for the nonrecipient subgroup results in a weight of 1.0/0.125=8.0 for each sampled nonrecipient.

Weights are important for two reasons. First, if one wants to estimate a parameter from a sample composed of individuals with different weights, one must take into account the relative contribution of each observation to obtain unbiased parameter estimates for the total population. Failure to take weights into account would imply equal importance of all observations, which would lead to biased parameter estimates if, on average, values from observations with smaller weights differed systematically from those obtained from individuals with larger weights.

Second, in stratified SRSs with strata of equal sizes, the greater the sampling fraction within a stratum,

the narrower the CI around that stratum-specific estimate because a greater proportion of individuals are sampled and thus available to "center" the subgroup's estimate near its population value. One easy way to conceptualize this point is to imagine a sample in which all but one individual in the population is sampled—only that person would be available to "deflect" the estimate from the population value. However, we should emphasize that sampling fraction is only one consideration in estimating random variation, and that sample size also is important.

Similarity of Observations Within Clusters. Suppose a study aimed to compare mean DMFS between Medicaid recipients and nonrecipients from the previously mentioned population of 1,000 children, but had only enough resources to examine 100 subjects. Further, assume that the 1,000 children were spread equally across 20 schools in the region of interest, with approximately 50 children attending each school. A 10 percent SRS could be generated; however, examining members of this sample might create logistical problems. One would need to establish contacts with multiple administrators, obtain a list of all 1,000 children, travel to many different schools, and set up and put away portable examination equipment numerous times. Instead, cluster sampling could be employed more conveniently by choosing five schools at random, then selecting at random two classrooms in each school, and examining 10 children selected at random from each classroom. Here, children would be clustered within classrooms and classrooms clustered within schools.

Both the 10 percent SRS and the three-stage cluster sample result in examination of 100 children; however, observations collected from the cluster sample would involve not only variation among individuals, as in the SRS, but also variation within and among clusters. Compared to the variation expected from an SRS, there generally is more variation from a cluster sample of the same size. This may seem counterintuitive, because there probably is less variation within clusters due to cluster members being relatively more homogeneous than members of an SRS. However, this decreased intracluster variation creates an increase in variation across clusters, resulting in cluster samples' generally having more variation than identically sized SRSs.

In the above example, variation in DMFS within a cluster might be less than would be expected from an SRS of the same size because there might be cluster-level factors related to DMFS that affect children in the cluster to a similar degree (2). For example, schools within one school system may share a common water supply and thus a common level of fluoridation. If so, all classrooms within that school system would receive water with similar fluoride concentrations, compared to the greater variation in fluoride concentration across school systems.

Statistical analyses that ignore complex sampling designs treat the data as if they were obtained from an SRS; thus, variation within and across clusters goes unrecognized, generally resulting in underestimation of the true variance. An erroneously small variance estimate, reflected by erroneously small standard errors, would produce narrower CIs and smaller *P*-values than would have been observed had clustering been accounted for, which would result in an increased likelihood of claiming a difference between comparison groups when there is none.

The design effect, or deff (2), is the extent to which the variance generated from a complex sample differs from that which would have been obtained from an SRS of the same size. Greater within-cluster (or within-stratum) homogeneity and more elements per cluster can increase deff substantially (2,12). For example, if deff=3, analysis as an SRS underestimates the true variance by one-third, or the reciprocal of deff. Such underestimation of variance could greatly affect CIs and hypothesis test values. As a rule, if deff>1, as usually happens with complex designs, analyzing data without taking complex sampling into account will underestimate the true variance.

Variance Estimation for Complex Samples: Analytic Strategies

General applications in most statistical analysis programs (e.g., SAS version 6.12, SPSS, BMDP, S-PLUS, StatView, Statistica) assume observations are from an SRS. Although many packages allow for weights, which will provide proper point estimates, including means, proportions, and regression coefficients, they will not give proper variance estimates for complex samples, resulting in inappropriate CIs and hypothesis test values.

To produce appropriate variance estimates from data generated using complex samples, one must either use analytic strategies that account for the correlated (clustered) data structure or modify the data set so that the independence assumption is satisfied. What approaches can be employed to achieve these goals?

- Use a software package that accounts for complex sampling designs. Some packages use Taylor series linearizations to calculate proper variances, while others use simulation methods; both approaches generally are acceptable. Several recent reviews of PC survey sampling analysis software are available (13-15). Some information already is outdated, but most of the content should help in software selection. Software reviewed (13-15) includes: CENVAR, CLUSTERS, EPI-INFO, PC CARP, Stata, SUDAAN, VPLX, and WesVarPC. Though some packages cost up to $12,000, others (CENVAR, CLUSTERS, EPI-INFO's CSAMPLE procedure, VPLX, and WesVarPC version 2.12) are free or inexpensive, so costly software is no longer an excuse for not performing appropriate analyses. SAS version 7.0 has beta test versions of survey data analysis procedures (SURVEYSELECT, SURVEYMEANS, and SURVEYREG) (16,17). Generally, free packages provide fewer features and less customer support than other packages. Programs also can be written in spreadsheets or in software packages, like the SAS macro in Wang (18).
- Regression can be performed to adjust for observation-level covariates and residuals can be aggregated at the cluster level, much like in group randomized trials (19). The main advantage of this approach is that any statistical software can be used. The primary disadvantage is that the analysis is more complex and inefficient.
- Occasionally, deff for similar studies may already have been published. For example, Davies et al. (20) reported deffs ranging from 2–3 in a study of periodontal attachment loss in a stratified cluster sample of elderly North Carolinians. If deff can be approximated from studies with comparable sampling designs, measure-

ments, and populations, one first can analyze the complex sample as though it were an SRS to obtain variances, then estimate the true variance by multiplying the estimated deff by the variance obtained from the SRS analysis. For example, suppose that the mean DMFS and standard error estimates obtained from a complex sample of schoolchildren were 10.0 and 1.5, respectively. Analysis as an SRS would provide an incorrect 95 percent CI of $10.0 \pm (1.96 * 1.5) = [7.06, 12.94]$ (21). If a published study of a comparable population used a similar complex sample and reported that deff=3, the adjusted 95 percent CI would be $10.0 \pm (1.96 * \text{sqrt}(3 * (1.5)^2)) = [4.91, 15.09]$. The main advantage of this approach is that specialized survey sampling software is not required. The main disadvantages are that a range of deffs might be provided in published studies, and any particular deff used would need to be justified. Further, it still would be necessary to compute proper point estimates (e.g., prevalences, means) using correct sampling weights. Methodologic studies addressing estimation of deff are available elsewhere (22-24).

- If investigators produce statistically significant results using variances calculated as if from an SRS, they could propose a range of deff values that, if applied to the SRS variance, still would produce statistically significant results. Again, the main advantage of this approach is that specialized survey sampling software is not required, and the main disadvantages relate to justification of the specified range of deffs. While this approach provides a crude method to adjust variances, it should be reemphasized that sampling weights must be used to obtain correct point estimates from complex samples, regardless of software used.

Variance Estimation for Complex Samples: Example

In this section we demonstrate the potential influence of analytic strategy on variance estimates. Data sets used here were selected for illustrative purposes only; readers should not infer that differences of this magnitude would be seen in similar analyses of other data sets. Such differences would depend on the deff for those studies.

The example uses data from both phases of the third National Health and Nutrition Examination Survey (NHANES III), a cross-sectional study of the United States population aged 2 months and older conducted between 1988–94. Full documentation of the survey has been provided elsewhere (25) and descriptive findings on oral health status from the first phase of the study have been reported (26).

Briefly, NHANES III used a multistage, stratified cluster sampling design to assess health characteristics of the US civilian, noninstitutionalized population (the target population for this survey). The design oversampled young children, persons at least 65 years old, African-Americans, and Mexican-Americans to provide sufficient numbers of subjects for analysis of these relatively small population subgroups. Hence, unit record weights are provided that adjust for different probabilities of subject selection and rates of nonresponse. The public-release data set contains variables for the sample design, including the stratum, primary sampling unit, and sampling weight for each subject.

Interviews were conducted in respondents' homes and standardized oral examinations were conducted at mobile examination centers. Dental caries experience was recorded for all tooth surfaces except third molars, and a separate assessment was made for the presence of fissure sealants (27).

For this example, we computed parameter estimates for two outcomes among children aged 5–17 years: mean DMFS and percent of children with one or more fissure sealants. Estimates were compared among socioeconomic subgroups as categorized by the income:poverty ratio, which represents the midpoint of the family income category recorded in the household interview divided by the poverty threshold for the subject's family in the year of the interview.

Documentation accompanying the public-release data sets cautions against conducting analyses that do not account for the complex sampling design, and methods are described for conducting appropriate analyses using SUDAAN software. However, in order to demonstrate some of the principles described above, we examined the two outcomes using three different analytic methods:

SAS (Version 6.12 for Windows 95) Unweighted Calculation. This method ignored the sampling design and sampling weights, thus treating the NHANES III sample as an SRS. This method has the effect of assigning each participant a weight of one, and thus is expected to produce parameter and variance estimates that cannot be generalized to the target population.

SAS (Version 6.12 for Windows 95) Weighted Calculation. This method used "normalized weights" for each observation that were obtained by dividing the weight for each observation by the mean weight of all observations. This has the effect of maintaining the sample size (hence, degrees of freedom) represented by the number of people examined, while providing proportions within comparison subgroups that mirror those in the target population. This method is expected to produce parameter estimates that can be generalized to the target population. However, because most SAS procedures do not account for stratification and clustering—including the MEANS, GLM, FREQ, and LOGISTIC procedures used in this analysis—variance estimates cannot be generalized to the target population.

SUDAAN (Version 7.50 for Windows 95) Weighted Calculation. This method adjusts both for sampling design and sampling weights, so it provides weighted sample sizes that are representative of the target population. Both parameter and variance estimates can be generalized to the target population. This method is consistent with the analysis guidelines provided with the public-release NHANES III data set.

Findings from the three analytic methods are presented in Table 1. Although the analyses revealed similar trends, including lower mean DMFS values and higher proportions of people with sealants in groups with higher income:poverty ratios, several important distinctions exist. For the SAS unweighted calculation, the values in the weighted sample size column are equivalent to the number of subjects examined, incorrectly implying that within the US population there are more than twice as many children in the lowest income:poverty category as in each of the two highest categories of socioeconomic status. In contrast, weighted sample sizes from the two weighted methods indicate a population that is fairly evenly distributed among the four income:poverty

TABLE 1
Unweighted and Weighted Calculations of Dental Caries Experience (DMFS) and Frequency of Fissure Sealants in NHANES III Among Persons Aged 5–17 Years with One or More Permanent Teeth

Income:Poverty Ratio	Subjects Examined	Weighted Sample Size	Caries Experience (DMFS)		Presence of 1+ Sealants		
			Mean (SE)	95% CI	% of Persons (SE)	Odds Ratio	95% CI
SAS unweighted calculations							
<1.0 (low SES)	2,203	2,203	2.32 (0.10)	2.12–2.52	7.8 (0.6)	ref	
1.0–<2.0	1,507	1,507	2.40 (0.13)	2.15–2.65	9.0 (0.7)	1.2	0.9–1.5
2.0–<3.0	955	955	2.26 (0.15)	1.97–2.55	17.5 (1.2)	2.5	2.0–3.1
3.0 or more (high SES)	862	862	1.96 (0.14)	1.69–2.23	28.3 (1.5)	4.7	3.8–5.8
P-value			.15		<.01		
All persons	5,527	5,527	2.28 (0.06)	2.16–2.40	13.0 (0.5)		
SAS normalized weighted calculations							
<1.0 (low SES)	2,203	1,326	2.66 (0.13)	2.41–2.91	11.5 (0.9)	ref	0.9–1.4
1.0–<2.0	1,507	1,351	2.78 (0.14)	2.51–3.05	12.7 (0.9)	1.1	1.9–2.9
2.0–<3.0	955	1,286	2.68 (0.15)	2.39–2.97	23.5 (1.2)	2.4	3.3–4.9
3.0 or more (high SES)	862	1,564	2.06 (0.10)	1.86–2.26	34.0 (1.2)	4.0	
P-value			<.01		<.01		
All persons	5,527	5,527	2.52 (0.06)	2.40–2.64	21.0 (0.5)		
SUDAAN weighted calculations							
<1.0 (low SES)	2,203	10,113,839	2.66 (0.23)	2.21–3.11	11.5 (2.3)	ref	
1.0–<2.0	1,507	10,306,702	2.78 (0.33)	2.13–3.43	12.7 (2.2)	1.1	0.6–2.0
2.0–<3.0	955	9,812,149	2.68 (0.36)	1.97–3.39	23.5 (3.5)	2.4	1.4–3.9
3.0 or more (high SES)	862	11,935,149	2.06 (0.20)	1.67–2.45	34.0 (2.4)	4.0	2.7–5.9
P-value			.09		<.01		
All persons	5,527	42,167,841	2.52 (0.18)	2.17–2.87	21.0 (2.0)		

categories. Finally, the SAS unweighted computation underestimates both mean DMFS and percent of children with fissure sealants, the latter by approximately one-third (13 percent of all persons aged 5–17 years have one or more sealants according to the unweighted analysis, compared with 21 percent as calculated in the weighted analyses).

As expected, SAS and SUDAAN weighted analyses provided similar point estimates but different standard errors, CIs, and *P*-values. SAS weighted standard errors were similar in magnitude to SAS unweighted standard errors, but were substantially smaller than SUDAAN weighted standard errors. Thus, for mean DMFS, SAS weighted standard errors suggested a statistically significant difference among income:poverty groups, whereas the SUDAAN weighted calculation showed a trend that was not significant (*P*=.09).

Though not shown in Table 1, the larger standard errors for the SUDAAN weighted analysis reflect moderately large deffs for these estimates. For example, deff for mean DMFS among all persons was 8.1, implying that the SUDAAN variance for mean DMFS was more than eight times that of the SAS unweighted variance. For the statistics reported in Table 1, deffs ranged from 2.3 to 12.8.

Finally, additional results obtained using survey data analysis procedures in SAS 7.0 and a macro available from the authors (28) were almost identical to those obtained in SUDAAN, except that CIs around parameter estimates were slightly wider, and global hypothesis test values for differences in odds ratios were not obtained.

Impact of Study Objective on Analytic Method: Recommendations

It is not clear whether all studies that employ complex sampling necessarily require adjustment of variances. Some investigators may have philosophical reasons for always adjusting for complex sampling, while others base their recommendation on the amount of "inefficiency" produced by accounting for complex sampling when it may not be useful. Specifically, Korn and Graubard (29) report that such inefficiency is related to the number of primary sampling units minus the number of strata, and that depending on the degree of inefficiency observed, the data analyst either should use clustering and weighting, or control for the covariates used to create the weights. Models should contain all of the important adjustment variables and the form of those variables must be correctly specified, e.g., linear and quadratic components of age should be included if that combination describes the true relationship (30).

We propose that the decision to adjust for complex sampling should hinge primarily on the study's main objective, as follows:

Adjustment of variances to reflect complex sampling is essential when:

- The study goal is to estimate the magnitude of a population value such as prevalence or incidence of a given condition. As stated previously, inferences to populations are based on point estimates and CIs generated from sample data. In public health practice, parameter estimates are used as a basis for comparison of characteristics among populations, and also for policy making, budget justification, and resource allocation. CIs around parameter estimates take on added importance in this setting because they describe a range in which one would expect the true population value to fall. If CIs are erroneously small, as would likely occur with complex samples analyzed with no correction for clustering and/or stratification, expectations for budgeting and resources would be based on too narrow a range of values, which ultimately could have adverse impact on the delivery of public health care. Thus, when one wishes to generalize estimates obtained from a sample to a larger population, it is imperative that variance estimates, and thus CIs, be as accurate as possible.
- The research objective is to estimate the magnitude of an exposure-outcome association in a hypothesis-testing study. These studies usually are conducted when there is previous evidence that a relationship exists, but the strength of the association has not been agreed upon. When such studies employ multivariable regression to obtain explanatory models, decisions to include or exclude variables from models often are based on the extent to which factors confound the association of interest or affect the precision of that estimate. In generating prediction models for a given outcome, decisions to include variables often hinge on *P*-values and/or estimates of sensitivity and specificity, which also are affected by estimates of variance. Thus, if investigators want to assess the magnitude of already established associations through the development of either explanatory or predictive regression models, appropriate estimates of variance are essential because they impact the decision-making process regarding factors to be included in and excluded from models. [Note: the above sentiment is that of the present authors and does not necessarily represent a consensus opinion. Discussion of this issue and presentation of comparative views can be found elsewhere (31,32).]
- Analysis is undertaken to estimate the presence of an exposure-outcome association in a sample that had been selected using stratification on some other variable (i.e., neither the exposure nor the outcome). For example, in Table 1, the association between socioeconomic status and presence of fissure sealants requires adjustment for complex sampling because subjects were not sampled on the basis of their income:poverty ratio or their fissure sealant status. Without accounting for complex sampling, biased estimates could result. This type of bias is discussed in detail elsewhere (4).
- Multiple observations per population element are analyzed, such as in studies of periodontal disease that analyze multiple sites within individuals. It should be noted that adjustment is required only if analysis is conducted for several observations within a cluster, as often is done at the surface-, site-, or tooth-level in periodontal or caries studies. If multiple observations are recorded, but the analysis is performed at the cluster level (e.g., if site-level data are aggregated into a "mean clinical attachment level" per randomly sampled individual), adjustment of variances is not needed.

Adjustment of variances to reflect complex sampling is recommended, but not required, when:

- The study goal is to determine the presence of an exposure-outcome association by examining differences in exposure across subgroups stratified by outcome (or differences in outcome across subgroups stratified by exposure). For example, Selwitz et al. (33) used a cross-sectional design to describe caries and fluorosis levels in three towns with different fluoride concentrations in their drinking water. The purpose of this type of study is not to estimate the magnitude of an exposure-disease association, but instead to demonstrate "in principle" that an association exists (i.e., to determine the presence of an association).

Studies of this type likely do not affect the practice of dental public health to a great extent. Most often their objective is to identify potential risk factors, predictors, or causative agents for disease, which in turn helps to improve our understanding of biologic processes and generate new hypotheses for future testing rather than build a foundation on which to make policy decisions or population comparisons. The impact of erroneous CIs on the practice of dental public health likely is minor, so adjustment of variances for complex sampling is not considered a requirement; however, it still is recommended on a philosophical basis. Further, valid point estimates for any association can be obtained only by using weights that reflect the selection probability for each sampled individual.

Adjustment of variances to reflect complex sampling is not required when:

- The primary goal is to report new information about unique population subgroups or novel conditions. Reports of this type are valuable because they elicit information about previously unstudied populations (34) or new diseases (35). Sophisticated analyses are not required because such reports, rather than being considered epidemiologic studies, can instead be viewed as "case studies" in which the "case" is a population subgroup rather than a person. In these situations, investigators can give the range of deff values that still provide significance.
- The sample is a nonprobability sample, such as a group of volunteers or dental clinic patients (36). In these situations, or when sampling design is not known, generalizations from obtained estimates should not be made because there is no sampling frame or denominator for the target population.
- The sample is an SRS. Because SRSs are not complex samples, this recommendation applies regardless of the study objective.

We strongly encourage researchers proposing investigations in dental public health, oral epidemiology, and dental health services research to consider what resources will be available at the time statistical analyses are conducted (because unavailable software or personnel at the time of analysis could influence the proposed sampling strategy), and to include an adequate budget for appropriate analyses. In addition, we recommend that they employ simple random sampling with one observation per person if they cannot analyze the data accounting for complex sampling because this method offers the greatest flexibility

with respect to analytic software and future secondary analyses. However, we recognize that real-life logistics and funding restrictions often preclude the use of this strategy. If complex sampling is to be used, investigators should use a predicted deff, or range of deffs, as a multiplicative factor for sample size calculations using SRS formulas.

In preparing manuscripts for submission, investigators should describe their sampling strategy fully so that reviewers and other readers can evaluate whether the statistical analyses used were appropriate. If clustering and/or stratification were employed, but adjustment for complex sampling was not carried out, investigators' reasons for not doing so should be stated. In addition, they should comment on the change in precision around estimates that might be expected if such adjustment were undertaken, drawing on findings from comparable studies that have considered sampling design effects. Finally, authors should describe the populations to which they feel their study findings can be generalized.

References

1. Last JM. A dictionary of epidemiology. 2nd ed. New York: Oxford University Press, 1988.
2. Kish L. Survey sampling. New York: J. Wiley, 1965.
3. Marques MD, Moreira B, Eriksen HM. Prevalence of apical periodontitis and results of endodontic treatment in an adult, Portuguese population. Int Endodont J 1998;31:161-5.
4. Schlesselman JJ. Case-control studies. Design, conduct, analysis. New York: Oxford University Press, 1982.
5. Caplan DJ, Weintraub JA. Factors related to loss of root canal filled teeth. J Public Health Dent 1997;57:31-9.
6. Beck JD, Sharp T, Koch GG, Offenbacher S. A study of attachment loss patterns in survivor teeth at 18 months, 36 months, and 5 years in community-dwelling older adults. J Periodont Res 1997;32:497-505.
7. Koch GG, Paquette DW. Design principles and statistical considerations in periodontal clinical trials. Ann Periodontol 1997;2:42-63.
8. McDonald BW, Pack ARC. Concepts determining statistical analysis of dental data. J Clin Periodontol 1990;17:153-8.
9. Rozier RG, Dudley GG, Spratt CJ. The 1986-87 North Carolina school oral health survey. Raleigh, NC: North Carolina Department of Environment, Health, and Natural Resources, 1991.
10. World Health Organization. Oral health surveys. Basic methods. 4th ed. Geneva: World Health Organization, 1997.
11. Siegal MD, Martin B, Kuthy RA. Usefulness of a local oral health survey in program development. J Public Health Dent 1988;48:121-4.
12. Cochran WG. Sampling techniques. 3rd ed. New York: John Wiley & Sons, 1977.
13. Lepkowski J, Bowles J. Sampling error software for personal computers. Survey Statistician 1996;35:10-17. In: University of Michigan Internet web site, Ann Arbor, MI, Apr 1999. [Available from: http://www.fas.harvard.edu/~stats/survey-soft/ iass.html.]
14. American Statistical Association, Section on Survey Research Methods. Summary of Survey Analysis Software. In: American Statistical Association Internet web site, Apr 1999. [Available from: http://www.fas.harvard.edu/~stats/survey-soft/survey-soft.html.]
15. Cohen S. An evaluation of alternative PC-based software packages developed for the analysis of complex survey data. Am Statistician 1997;51:285-92.
16. An A, Watts D. New SAS* procedures for analysis of sample survey data. In: SAS Institute, Inc., Internet web site, Cary, NC, Apr 1999. [Available from: http://www.sas.com/rnd/app/papers/survey.pdf.]
17. SAS Institute Statistics and Operations Research Group. Sample survey design and analysis. In: SAS Institute, Inc., Internet web site, Cary, NC, Apr 1999. [Available from: http://www.sas. com/rnd/app/da/new/dasurvey.html.]
18. Wang ST. Analysis of survey data—a solution using SAS (letter). Public Health 1998;112:273-5. In: Univeristy of California at San Francisco Internet web site, Apr 1999. [Available from: http://itsa.ucsf.edu/~sgansky/wangsurv.sas.]
19. Murray DM. Design and analysis of group-randomized trials. New York: Oxford University Press, 1998.
20. Davies GM, Koch GG, Beck J. Statistical strategies for event rate comparisons in dental studies. J Biopharm Stat 1997;7:625-34.
21. Kleinbaum DG, Kupper LL, Morgenstern H. Epidemiologic research. Principles and quantitative methods. New York: Van Nostrand Reinhold, 1982.
22. Katz J, Zeger SL. Estimation of design effects in cluster surveys. Ann Epidemiol 1994;4:295-301.
23. Mickey RM, Goodwin GD. The magnitude and variability of design effects for community intervention studies. Am J Epidemiol 1993;137:9-18.
24. Ferguson JA, Corey PN. Adjusting for clustering in survey research. DICP 1990;24:310-13.
25. US Department of Health and Human Services. National Center for Health Statistics. Third National Health and Nutrition Examination Survey, 1988-1994, NHANES III Examination Data File (CD-ROM). Public Use Data File Documentation #76200. Hyattsville, MD: Centers for Disease Control and Prevention, 1996.
26. Kleinman DV, Drury TF. Oral health in the United States, 1988-1991: the first three years of the third National Health and Nutrition Examination Survey. J Dent Res 1996;75(Spec Iss):617-19.
27. Selwitz RH, Winn DM, Kingman A, Zion GR. The prevalence of dental sealants in the US population: findings from NHANES III, 1988-1991. J Dent Res 1996;75(Spec Iss):652-60.
28. Gansky SA. A SAS 7.0 program and macro for "Complex sampling: implications for data analysis." In: University of California at San Francisco Internet web site, Jun 1999. [Available from: http://itsa.ucsf.edu/~sgansky/jphdsurv. sas.]
29. Korn EL, Graubard BI. Epidemiologic studies utilizing surveys: accounting for the sampling design. Am J Public Health 1991;81:1166-73.
30. Pfeffermann D, Lavange LM. Regression models for stratified multi-stage cluster samples. In: Skinner CJ, Hoilt D, Smith TMF, eds. Analysis of complex surveys. New York: John Wiley & Sons, 1989.
31. Sarndal C-E, Swensson B, Wretman J. Model assisted survey sampling. New York: Springer-Verlag, 1992.
32. Hansen MH, Madow WG, Tepping BJ. An evaluation of model-dependent and probability-sampling inferences in sample surveys. J Am Stat Assoc 1983;78:776-93.
33. Selwitz RH, Nowjack-Raymer RE, Kingman A, et al. Dental caries and dental fluorosis among schoolchildren who were lifelong residents of communities having either low or optimal levels of fluoride in drinking water. J Public Health Dent 1998;58:28-35.
34. Kaste LM, Bolden AJ. Dental caries in homeless adults in Boston. J Public Health Dent 1995;55:34-6.
35. Greenspan D, Greenspan JS, Conant M, et al. Oral "hairy" leukoplakia in male homosexuals: evidence of association with both papillomavirus and a herpesgroup virus. Lancet 1984;2:831-4.
36. Marcenes W, Pankhurst CL, Lewis DA. Oral health behavior and the prevalence of oral manifestations of HIV infection in a group of HIV-positive adults. Int Dent J 1998;48:557-62.