VITAL & HEALTH STATISTICS

A Statistical Methodology for Analyzing Data from a Complex Survey: The first National Health and Nutrition Examination Survey

This report presents an approach to analyzing data from a survey with a complex sample design. The data used to illustrate the approach are from the first National Health and Nutrition Examination Survey, a national probability sample survey that was conducted in 1971–74 with an augmentation survey in 1974–75. Data are examined using regression techniques, analysis of variance, and categorical data analysis.

Data from the National Health Survey
Series 2, No. 92

DHHS Publication No. (PHS) 82–1366

U.S. Department of Health and Human Services
Public Health Service
Office of Health Research, Statistics, and Technology
National Center for Health Statistics
Hyattsville, Md.
September 1982

SUGGESTED CITATION

National Center for Health Statistics, J. Landis, J. Lepkowski, S. Eklund, and S. Stehouwer: A statistical methodology for analyzing data from a complex survey, the first National Health and Nutrition Examination Survey. *Vital and Health Statistics.* Series 2-No. 92. DHHS Pub. No. 82–1366. Public Health Service. Washington. U.S. Government Printing Office, Sept. 1982.

Library of Congress Cataloging in Publication Data

A Statistical methodology for analyzing data from a complex survey.

(Vital and health statistics. Series 2, Data evaluation on methods research; no. 92) (DHHS publication; (PHS) 82–1366)

Authors: J. Richard Landis . . . [et al.].

1. Health surveys—Statistical methods. 2. Nutrition surveys—Statistical methods. I. Landis, J. Richard. II. National Center for Health Statistics (U.S.). III. Series. IV. Series: DHHS publication; (PHS) 82–1366.

RA409.U45 no. 92 [RA408.5] 312'.0723s 82–600240

ISBN 0–8406–0264–2 [614.4'2'0723]

Foreword and acknowledgments

This report is a contribution to the literature on statistical methodology designed to improve the analysis of data from surveys with complex sample designs. Specifically, it is designed to help the users of the data tapes from the National Health and Nutrition Examination Surveys. Many of these users are analyzing data from a survey with a complex sample design for the first time. Some statistical guidance is required to utilize the data available from such studies properly. This report should provide such guidance so that they can proceed with confidence and caution. To make this report possible, the Division of Health Examination Statistics extended the contract (no. 233–79–2092) with the School of Public Health at the University of Michigan.

Dwight Brock of the National Institute on Aging, NIH, Ron Forthofer of the School of Public Health, University of Texas, and Robert Casady of the National Center for Health Statistics, reviewed drafts, discussed statistical methodology, and made suggestions for changes. Their professional involvement helped us all to learn from this project.

Mary Grace Kovar
Special Assistant for Data Policy and Analysis
Office of Interview and Examination Statistics

Contents

A Statistical Methodology for Analyzing Data from a Complex Survey: The first National Health and Nutrition Examination Survey

J. Richard Landis, Ph.D.; James M. Lepkowski, Ph.D.;
Stephen A. Eklund, D.D.S., M.H.S.A., Dr.P.H.; and
Sharon A. Stehouwer, University of Michigan

Introduction

For many large-scale surveys like those conducted by the Bureau of the Census and the National Center for Health Statistics in the United States, and the World Fertility Survey coordinated by the International Statistical Institute in the Netherlands, data are obtained through multi-stage sampling designs involving clustering and stratification, as well as estimation techniques that include post-stratification and non-response adjustments.[1-3] Consequently, the direct application of standard statistical analytic methods may be misleading for such survey data. The inappropriateness of standard methods in this context is due to the complexities in the sample design which induce a non-standard covariance structure among the sample quantities under investigation.

Although the modification of statistical analytic procedures to incorporate the effects of complex sample designs is an important area of research, the methodologies appropriate for such data have not been made readily available to general users of statistical software packages. Exceptions to this are the software packages developed for the analysis of survey data, including OSIRIS IV (University of Michigan)[4] and SUPERCARP (Iowa State University),[5] which are available for purchase and documented for outside users. In addition, the programs SESUDAAN and SURREGR, which are accessed through SAS,[6] can be obtained from Babu Shah of the Research Triangle Institute.

The methods and results presented in this monograph were developed in the process of producing the extensive analyses reported in three NCHS publications[7-9] based on data from the first National Health and Nutrition Examination Survey. Although the general methodology outlined herein is not new or unique, some of these procedures are not well known to users of standard statistical software packages. In particular, several computing stages involving separate algorithms are required to generate the analysis of variance and contingency table analyses.

This document is intended to provide a representative set of analyses illustrated by data from the first National Health and Nutrition Examination Survey. These data are available on public use tapes and can be purchased from the National Technical Information Service. They permit analyses by researchers with varied statistical approaches and available computing software. Even though other users may not have access to the same computing packages used for this report, parallels with other software will be similar. These results have been computed under various assumptions ignoring the weights, ignoring the sample design, or ignoring neither the weights nor the sample design. The importance of the design on estimates of variance, and consequently, of test statistics, is highlighted throughout, both as a research finding of interest for this survey and as an illustration of the critical importance of incorporating these design effects into any analyses of data from the first National Health and Nutrition Examination Survey or from other complex surveys.

Survey design

The sample design for the first National Health and Nutrition Examination Survey (NHANES I) is basically a three-stage, stratified probability sample of clusters of persons in area-based segments. The sample was designed to represent the civilian noninstitutionalized population ages 1–74 years in the coterminous United States, excluding persons residing on lands set aside for the use of American Indians. Successive sampling units used in the sampling were the primary sampling unit (a county or group of counties denoted as a PSU), census enumeration district (ED), segment (a cluster of households), household, eligible person, and sample person.

For the April 1971 through June 1974 period, the design provided for selection of a representative sample of the target population 1–74 years of age. The entire sample was given the nutrition-related interview and examination; a subsample of adults 25–74 years of age received a more detailed examination focused on other aspects of health and health care needs. To increase the size of the subsample of adults and, consequently, the usefulness of the data obtained, the design further provided for selection of an additional national sample of adults 25–74 years of age. This sample was given a detailed examination in July 1974-September 1975. The extension of NHANES I is referred to as the "Augmentation Survey."[10]

The estimated civilian noninstitutionalized U.S. population ages 1–74 years at the time of examination is shown in table 1 by sex, race, and age. Because certain analyses must be done on the basis of age at examination, for the sake of consistency the population estimates also have been based upon age at examination rather than the age at interview.

Sample selection

The first stage of the sample selection began with the 1960 decennial census lists of addresses and nearly 1,900 primary sampling units (PSU's) into which the entire United States had been divided. The 1960 decennial census information was used in the selection of first stage units because the 1970 census information was not available. The 1970 census information was used at subsequent stages of selection as it became available, although it was not used for later stages of selection within all primary selections. Each PSU is either a standard metropolitan statistical area (SMSA), a single county, or a group of two or three contiguous counties. The PSU's were grouped into 357 strata, as for the National Health Interview Survey during 1963–72, and subsequently collapsed into 40 superstrata for use in NHANES I.

For the April 1971-June 1974 period, 15 of the 40 "superstrata" which contained a single large metropolitan area of more than 2 million population were chosen in the sample with certainty. The remaining 25 noncertainty strata were classified into four broad geographic regions of approximately equal population (when the large metropolitan areas selected with certainty were included) and cross-classified into four broad population density groups in each region. A controlled-selection technique[11] was used to select two PSU's from each of the 25 noncertainty superstrata. The probability of selection of a PSU was proportional to its 1960 population. Representation of specified state groups and rate of population change classes in the selections was controlled in the sample selection process. In this manner a total first stage sample of 65 PSU's was selected, 15 large metropolitan certainty areas and $(2)(25) = 50$ paired selections from noncertainty areas. These 65 sample PSU's are the areas within which clusters of sample persons were selected for examination.

Although the 1970 census data were used as the frame for selecting the sample within the PSU's when the data became available, the calendar of operations required that the 1960 census data be used for the first 44 locations in the sample. The 1970 census data were then used for the final 21 locations of the sample and for the Augmentation Survey.

Beginning with the use of the 1970 census data, the segment size was changed from an expected 6 housing

units selected from compact clusters of 18 housing units to an expected 8 housing units. This change was implemented because of operational advantages. Research by the U.S. Bureau of the Census indicated that precision of estimates would not be appreciably affected by such a modification. For large enumeration districts (ED's), the segments were clusters of addresses from the 1960 Census Listing Books (later the corresponding books for 1970). For other ED's, area sampling was employed and consequently some variation in the segment size occurred. To make the sample representative of the then current population of the United States, the address or list segments were supplemented by a sample of housing units that had been constructed since 1960.

Within each selected PSU a systematic sample of clusters of housing units or segments was selected. The ED's selected for the sample were coded into one of two economic classes. The first class, identified as the poverty stratum, was composed of current poverty areas that had been identified by the Bureau of the Census in 1970 based on information obtained prior to the 1970 Census plus other ED's in the PSU with a mean income of less than $3,000 in 1959 (based on 1960 Census). The second economic class, the nonpoverty stratum, included all ED's not designated as belonging to the poverty stratum. All sample segments in ED's classified as being in the poverty stratum were retained in the sample. For those sample segments in nonpoverty stratum ED's, the selected segments were divided into eight random subgroups and one of the subgroups remained in the NHANES I sample. Continuing research during the NHANES I field collection period indicated that efficiency of estimates could be increased by changing the ratio of poverty to nonpoverty segments from 8:1 to 2:1. Therefore, in the later survey locations the selected segments in the nonpoverty stratum ED's were divided into only two random subgroups and one of the subgroups was chosen to remain in the sample. Adequate reliability for separate analyses of those classified as being below the poverty level and those classified as being above the poverty level was achieved through a disproportionate allocation of the sample among poverty and nonpoverty strata within selected PSU's.

After identifying the sample segments, a list was made of all current addresses within the segment boundaries. A household member was interviewed to determine the age and sex of each household member, as well as other demographic and socioeconomic information required for the survey. If no one was at home after repeated calls, or if the household members refused to be interviewed, the interviewer tried to determine the household composition by questioning neighbors.

To select the persons in the sample segments to be examined in NHANES I, all household members ages 1–74 years in each segment were listed on a sample selection worksheet, with each household in the segment listed on the worksheet in the order in which it had been listed by the interviewer. The number of household members in each of six age-sex groups (see table 2) were listed on the worksheet under the appropriate age-sex group column. The sample selection worksheets were then arranged in segment-number order. A systematic sample of persons in each age-sex group was selected to be examined using the sampling rates displayed in table 2.

In general, this procedure resulted in only one person being selected from a household. However, in a few instances, more than one person was selected from a given household. This sampling strategy for the general sample of NHANES I resulted in the selection of 28,043 sample persons 1–74 years of age, a sample that can be regarded as representative of the target population displayed in table 1.

In addition, a subsample of adults 25–74 years of age was designated to receive a detailed health examination in addition to the general health examination given to all selected persons. This detailed sample was chosen systematically after a random start from the general sample of selected persons using sampling rates shown in table 3. For example, adults 45–64 years of age were subsampled for the detailed examination at a somewhat higher rate than 25–44 years of age from among all persons selected within cooperating households.

The mobile examination units were moved from one location to the next during a 39–month period (1971–74) to permit administering single-time examinations to the sample of persons participating in the study. These mobile units were moved throughout the North during the summer months and throughout the Southern areas in the winter months. Consequently, certain measures may reflect seasonal influences.

The sample for the Augmentation Survey, adults 25–74 years of age selected for examination in 35 primary units, also constitutes a national probability sample of the target population. Moreover, when considered jointly with those selected for the NHANES I detailed examinations in the first 65 locations, the entire 100 location sample also represents the adult population at that time.

The sampling frame for selecting the augmentation sample was the 1970 decennial census list of addresses and PSU's. The methods for establishing the sample frame and selecting households were generally similar to those used in selecting the general sample. However, only 5 of the 15 superstrata (composed of only 1 very large metropolitan area of more than 2 million population) were drawn into the augmentation sample with certainty. The remaining 10 of these superstrata were collapsed into 5 groups of 2 each from which only 1 superstrata was selected. Thus, the probability of selection for each of these 10 superstrata is 0.5, even though each of the 5 collapsed pairs is represented in the design with certainty. When these latter 5 locations are con-

sidered a part of the 100 primary sampling unit design, they are selected with certainty.

In this Augmentation Survey there was no economic stratification of enumeration districts and no oversampling among special age-sex groups. One of every two eligible persons within sample households (using a random start among those 25–74 years of age) was selected for participation in the survey.

Nonresponse

In a health examination survey, as well as any survey involving volunteer participation, the survey meets one of its severe problems after the sample is identified and the sample persons are requested to participate in the examination. A sizable number of sample persons who initially are willing to complete the household information, and possibly some of the medical history questionnaires (which are done in the household), usually will not participate in the examination. Full participation by individuals is determined by many factors, some of them uncontrollable by either the sample person or the survey personnel. For example, family health beliefs and practices, employment status, and access to transportation could affect participation in the survey. Because nonresponse is a potential source of bias, intensive efforts were made in NHANES I to develop and implement procedures and inducements to reduce the number of nonrespondents and thereby reduce the potential of bias due to nonresponse. These procedures are discussed in a *Vital and Health Statistics series* report.[1]

Also during the early stages of NHANES I when it became apparent that the response rate for the examinations was lower than in the preceding health examination surveys, a study of the effect of remuneration upon response in NHANES I was undertaken. The findings, published by NCHS,[12] included remuneration as a routine procedure in NHANES I starting with the 21st and 22nd examination locations.

Despite response rates of over 98 percent at the household interview stage and intensive efforts of persuasion, only 20,749 or 74.0 percent of the sample persons from the first 65 survey locations were examined. When adjustments are made for differential sampling for high-risk groups, the weighted response rate becomes 75.2 percent. Consequently, the potential for a sizable bias exists in the estimates from this survey. However, from what is known about the nonrespondents and the nature of the nonresponse, the likelihood of sizable bias is believed to be small.

Using data from NHANES I and from an earlier survey, efforts have been made to examine possible health-related differences between examined and nonexamined persons. An investigation of reasons for participation and nonparticipation in NHANES I was conducted by interviewing a sample of 406 people comprised of 290 examined persons, 35 persons who had made appointments for the examination but who never came to the mobile examination center for the examination, and 81 persons who refused to participate in the survey.[13] The sample persons for this study came from four survey locations: St. Louis, Monterey, New York, and Philadelphia. They were asked to indicate why they did not choose to be examined in NHANES I. The primary reasons given were that they had no need for a physical examination (48 percent), or that the examination times were inconvenient because of work schedules or other demands (15 percent). Only 6 percent of those persons who were not examined indicated that they refused the examination because of sickness, and 3 percent based their refusal on a fear of possible findings.

Data on both examined and nonexamined (but interviewed) persons were analyzed by using information from the first 35 survey locations of NHANES I.[14] For the health characteristics compared, the two groups were quite similar. For example, 20 percent of the examined people reported that a doctor had told them they had arthritis, compared to 17 percent of the unexamined people. Similarly, 18 percent of both the examined and the nonexamined persons had been told by a doctor that they had high blood pressure. Twelve percent of both groups reported that they were on a special diet, and six percent of both groups said that they regularly used medication for nerves.

In another study of factors relating to response in Cycle I of the Health Examination Survey, 36 percent of the nonexamined people viewed themselves as being in excellent health compared with 31 percent of the examined people.[15] A self-appraisal of poor health was made by 5 percent of the nonexamined persons, and by 6 percent of those who were examined. In a different study of Cycle I findings, those who participated in the survey with no persuasion and those who participated only after a great deal of persuasion generally had few differences for numerous selected examination and questionnaire items.[16] This was interpreted as evidence that no large bias exists between these two groups for the items investigated, and was offered as further support for the belief that little bias is introduced to the findings because of differences in health characteristics between examined and nonexamined persons.

Because of the nonparticipation of some sample persons in NHANES I, an adjustment procedure to account for nonresponse (similar to that used in previous National Health Examination Surveys) was used. The reciprocal of the probability of selection of the sample persons is multiplied by a factor that brings estimates based on examined persons up to a level that would have been attained if all sample persons had been examined. This nonresponse adjustment factor was computed separately within relatively homogeneous classes defined by five income groups (under $3,000; $3,000–$6,999; $7,000–$9,999; $10,000–$14,999; and $15,000 or more) within each stand. The factor is the

ratio of the sum of sample weights for all sample persons to the sum of sampling weights for all responding sample persons within the same homogeneous class.

To the degree that groups can be defined which are homogeneous with respect to the characteristics under study, the nonresponse adjustment procedure can be effective in reducing the potential bias from nonresponse. In addition, a poststratified ratio adjustment procedure was employed to force agreement between the final sample estimates of the population and independent controls prepared by the U.S. Bureau of the Census for the noninstitutionalized population of the United States as of November 1, 1972 (the approximate midpoint of the survey) for the cells shown in table 1.

The combined adjustment factor for nonresponse and poststratification among the detailed examinees was 1.45 for the 65 PSU's of the 1971–1974 period and 1.40 for the Augmentation Survey. For the 65-PSU sample of NHANES I, the percent distribution of the adjustment factors used for the 325 cells (determined by the cross-classification of the five income groups by the 65 stands) is shown in table 4.

Missing data and imputation

Examination and other types of surveys in which multiple observations are made on the same person are subject to the loss of information not only through failure to examine all sample persons, but also from the failure to obtain and record all items of information for examined persons. When data for specific items are missing for some of the examinees, values for these items are often imputed to minimize the effect of such item nonresponse on population estimates.

The issues relating to adjustments for missing data in surveys of this magnitude are complex and too numerous to discuss in this report. However, the adjustments for relevant variables used in this research, particularly the dental and blood pressure findings used as examples in the subsequent discussions, are of interest here.

Dental findings were available for 20,218 of the 20,749 examinees in this NHANES I survey. Those 531 (2.6 percent) whose dental records were lost or not obtained through examinations were assigned imputed values. Imputation of dental findings for an examinee was done by randomly selecting a record from among examinees of the same age in years, race, sex, and income group who had dental findings recorded. The values for this matched examinee were then imputed for the missing items for the examinee with missing data. When data for income were not available, the match was limited to age, race, and sex. These imputed values are included in all of the analyses involving the dental variables in this report. The age and sex distribution of the examinees with and without dental data from the survey is shown in table 5.

Among the 13,671 examinees ages 18–74 years in the total, or nutrition, sample for 1971–1974, there were 76 (0.6 percent) examinees missing either the single measurement of systolic or diastolic blood pressure or both. Out of the 6,913 examinees ages 25–74 years in the detailed and augmentation sample, only 28 (0.4 percent) were missing measurements of either systolic or diastolic blood pressure or both in the first sitting position. For the recumbent position, 59 (0.9 percent) were missing measurements of either systolic or diastolic blood pressure or both, while for the second sitting position, 64 (0.9 percent) were missing measurements for either or both blood pressures. In no case was a diastolic measurement present without an accompanying systolic measurement.

For the statistical analysis of the blood pressure variables reported in *Vital and Health Statistics,* Series 11-No. 203,[17] imputed values for missing systolic and diastolic blood pressures were assigned from the records of matched examinees with the same age, sex, and race, with similar arm girth, weight, and height. However, these missing value imputations are not recorded on the public use tapes; the imputation process would need to be repeated prior to statistical analysis of the data if identical analyses to those reported in the Series 11-No. 203 report were desired. Because there are so few of them, persons with missing blood pressures can be excluded from investigations of hypotheses involving these variables without seriously altering population inferences. Thus, to simplify the analyses in this report, records with missing data for blood pressure variables were excluded for estimates or hypothesis tests in which that required variable was missing.

In general, missing data cannot be ignored in the analysis. For these analyses values were imputed for the missing dental variables and persons with missing blood pressures were excluded when the necessary value was missing. However, for variables with excessive rates of missing data (for example, greater than 10 percent), the data analyst must exercise caution in making estimates and drawing inferences from the survey findings.

Design considerations for examined persons

Although the sample design for this survey is described in extensive detail in the previous sections and in another document,[1] aspects of the design pertaining to data analysis considerations will be discussed further in this section. All 20,749 examined persons in the first 65 survey locations received a specifically designed nutrition-related examination. In addition, approximately 20 percent of those ages 25–74 years (3,854 persons) received a more detailed examination concerning other aspects of health and health care needs. An additional 3,059 persons ages 25–74 years were examined in the 35-location Augmentation Survey to increase the size of the detailed sample, and hence, the

reliability of the estimates. The data collection forms for the entire sample, together with the additional forms for the detailed sample, are published elsewhere.[18]

Although the sample design for this survey was complex, the essential feature is the selection of primary sampling units (PSU's) consisting of counties or groups of counties from each of the defined strata. In particular, the NHANES I design involved the selection with certainty of the 15 large standard metropolitan statistical areas with more than 2 million population.

For data analysis purposes, several of the 15 certainty strata were combined by NCHS to form only 10 strata. The data tapes from NCHS reflect this revised indexing of the certainty strata, although this recombination of strata is not documented completely in previous NCHS publications. Each of these "certainty PSU's" consists of a large number of enumeration districts which were treated as PSU's. Each of the remaining 25 strata can be considered as being composed of exactly two PSU's.

The Augmentation Survey discussed in *Vital and Health Statistics,* Series 1-No. 14[9] poses additional complications for analysis. The 3,059 examined persons selected for this Augmentation Survey represent a national probability sample of the target population when used as a separate 35-location sample. The Augmentation Survey can also be combined with the 65 location detailed sample to form a 100-PSU national probability sample, in which the combined number of persons is 6,913. Of the PSU's, 10 were included on both the Augmentation Survey and the initial survey. There was oversampling of the elderly in the initial detailed sample group (tables 2 and 3), but not in the Augmentation Survey.

The number of PSU's and the corresponding number of examined persons in each of these strata for each of these survey components are summarized in table 6. Thus, for analytical purposes, this design can be characterized as having the following:

1) 10 strata with selection of segments as PSU's and with multiple PSU's for all survey components (survey locations 1–65);
2) 25 strata with
   a) paired selections of PSU's for the general and detailed sample (survey locations 1–100);
   b) selection of a single PSU for survey locations 1–35 and for the augmentation sample (locations 66–100).

Throughout the remainder of this report, these paired or multiple selections will be referred to as sampling error computing units (SECU's) indicating their role as basic units in variance calculations. For example, if all strata have exactly two SECU's, a paired selection model involving squared differences of SECU totals within each stratum can be used to obtain Taylor series approximations to variances and covariances of sample estimators. Thus, if a particular design has exactly 2 PSU's per stratum, these PSU's play the role of SECU's without further recombination. On the other hand, the NHANES I design summarized in table 6 requires that the multiple PSU's in strata 1–10 be combined into two SECU's each in order to have a paired design. Although the analyses of this report do not deal with 35-location design where only one PSU was selected for the noncertainty strata, it should be noted that NCHS recommends that the 25-noncertainty strata be collapsed into 13-SECU's for variance computational purposes in the documentation available with the microdata tapes.

Even though the overall number of examined persons in this survey is quite large, subclass analyses still can lead to estimators with unstable properties, particularly estimators of their variances based on Taylor series approximations for which the SECU sample sizes are small. For example, in the general sample the number of examined persons for the "other race" category is extremely sparse in some of the strata as shown in table 7. Moreover, as shown in table 8, the number in some strata is quite sparse both for black people and those of other races in the detailed survey. Consequently, analyses by racial subclasses requires particular attention to the coefficient of variation of the denominator for the estimators involving ratio means; for the detailed sample, certain analyses such as multiple regressions by racial subgroups may lead to serious computational difficulties or analyses of questionable reliability. This issue will be addressed further in subsequent sections.

Another important aspect of the NHANES I design is the oversampling of the following subgroups thought to be at high risk of malnutrition:

1) Persons with low income;
2) Preschool children;
3) Women of childbearing age; and
4) Elderly persons.

Adjusted sampling weights that reflect these unequal selection probabilities, in addition to the basic probability of selection and the adjustments for nonresponse and poststratification, were computed and are on the public-use data tapes.

An additional design complication arises because there was no oversampling of the subset of the sample persons ages 25–74 years who received the more detailed health examination. Women of childbearing age were not oversampled as they were for the major nutrition component of NHANES I. However, some oversampling remained among the elderly and poor people. There are separate adjusted sampling weights on the data tapes for the 3,854 persons given this detailed examination.

Consequently, when computing estimates of ana-

lytic statistics and their estimated variance-covariance structure, the appropriate sampling weights need to be utilized in the weighted analyses. Thus, in this report hypotheses involving variables from the initial detailed sample of persons ages 25–74 (survey locations 1–65) were investigated using the adjusted sampling weights associated with those sample persons. Analyses involving the augmentation detailed sample (survey locations 66–100) used the adjusted sampling weights for this group. When hypotheses were investigated across the combined detailed sample groups (survey locations 1–100), a third adjusted sampling weight was used for the combined groups. Hypotheses involving variables from the entire nutrition-related initial sample (survey locations 1–65) utilized the adjusted sampling weights for that sample.

Analytical strategies

Because of the complexities in the sample design, an analysis could be performed in any one of at least three different ways depending on whether the sampling weights were used or whether the sample design features were incorporated in the estimation procedure. For simplicity, the following three options will be discussed:

Option	Use of sampling	
	Weights	Design features
1	No	No
2	Yes	No
3	Yes	Yes

Although the analyses could be performed under any of these options, it will be demonstrated that option 3 is more appropriate for making final inferences from these NHANES I data. However, as a practical matter, most hypotheses initially were investigated under option 1, since the implementation of each option in successive order from 1–3 involved considerably more preparation and computing costs. Relationships found to be statistically significant under option 1 were subjected to more definitive analyses under option 3 utilizing the sample weights and the survey design effects. Consequently, the estimated covariance structure for the sample estimators, based on the complexities of the survey design, was utilized in all final models and inferential conclusions.

There is a certain risk associated with this sequential strategy. Relationships found to be nonsignificant under option 1, the "screening stage," may in fact be significant if the complex sample effects on the variances of the estimators actually reduce the estimated variances. Although this situation is rare in highly clustered data such as those obtained in the NHANES I, substantive relationships thought to be important should be investigated more rigorously under option 3, even if the statistical tests indicate the lack of significance under option 1.

In survey research, the design effect is commonly defined as the ratio of the variance for a statistic from a complex sample to the corresponding variance from a simple random sample of the same size. These effects are used by survey designers and analysts for a variety of purposes. Frequently the design effect has been used to summarize conveniently the effects of a complex sample design on the precision of estimates from the survey data and to specify design features for new surveys. Increasingly, design effects are being used to adjust estimates and statistics computed under simple random sampling assumptions for the effects of the complexities in the sample design on measures of precision. Given the importance of these effects to those designing and analyzing surveys, simple useful models have been sought for design effects. Such models are useful for deriving estimates of design effects for statistics for which they are not available and for suggesting methods to adjust estimates computed under the assumption of independent selections for complexities in the sample design. A review of these design effect considerations and analytical strategies for survey data from complex sample designs was presented by Lepkowski.[19] Throughout this publication, the estimated design effects will be shown to illustrate the importance of these effects in definitive hypothesis tests or model fitting calculations.

All analyses under option 1 can be performed quite simply and relatively inexpensively using standard statistical software packages. In this option sampling weights and design effects are totally ignored. Thus, the data are regarded as coming from a simple random sample with equal probability of selection for every element in the population. Analyses under option 2 incorporate the sampling weights in estimating the analytic statistics, but simple random sampling computations are still utilized as under option 1 for the variance estimation. Analyses under option 3 utilize both the sampling weights and the complex sampling design in calculating the estimates and the estimated variance-covariance structure of analytic statistics. The calculations for options 2 and 3

were performed with the OSIRIS IV software package developed by the Computer Support Group within the Survey Research Center of the Institute for Social Research at the University of Michigan.[4] Alternatively, other statistical software packages could be used if they can incorporate the sampling weights and the design structure into the analysis.

In particular, for this report the computer program &PSALMS was used for estimating ratio means and the program &REPERR was utilized to fit regression models. For relatively simple statistics such as ratio means, differences of ratios, and totals, the &PSALMS routine approximates the complex sample variance of these estimators using a linearized Taylor series expansion. For more complex statistics, such as regression coefficients, either a balanced half sample (BHS) or a Jack-knife replicated variance estimation procedure is available. The BHS option within the &REPERR routine was utilized to fit simple and multiple regression models to the NHANES I data. Both of these routines are available within the OSIRIS IV library, and are described in more detail by Vinter.[20]

Because of the multiple SECU's within the certainty strata 1–10, the estimation procedure to implement option 3 can be extremely time consuming and expensive, particularly if replication procedures are used to fit regression models. On the other hand, if each stratum has exactly 2 SECU's, the BHS approach to fitting regression models is straightforward and economical.

To alleviate these cost and computing time difficulties, the multiple SECU identification codes in each of the certainty strata (i.e., 1–10) were randomly allocated into 2 pseudo-replicates within the stratum. Consequently, the paired selection computation procedures could be utilized across all 35 strata for all statistical analyses, not just those involving multiple regression. The effects of randomly assigning the multiple SECU's to two paired pseudo-replicates was investigated by comparing standard errors and design effects for estimates of proportions and means within the age groups shown in table 10 for variables such as decayed, missing, and filled (DMF) teeth, systolic blood pressure (SBP), and calories. The means and standard errors were computed under the multiple SECU classification scheme and under the paired SECU groupings. For these variables, it is apparent that the random allocation of SECU's in the certainty strata to form a complete paired design has not substantially altered the estimates of variances or the corresponding design effects for overall means and subclass means.

As a result of this pairing for the 10 certainty strata, all variance-covariance computations can be obtained directly as appropriate sums of squares and cross-products of differences between SECU or replicate totals across the 35 strata in the initial sample utilizing 70 paired SECU's. Consequently, all the analyses under option 3 for the data from the 65 survey locations were performed assuming this paired selection design. On the other hand, analyses under option 3 for the combined data from the detailed and augmentation surveys (the 100–location survey) require a multiple selection model for variance computations because the design cannot be paired for the 25 noncertainty strata; each of the strata 11–35 have 3 SECU's in this combined design. Consequently, when combining the data from the detailed and augmentation survey, the user needs to utilize a variance-covariance estimation procedure that permits multiple SECU's per stratum. For example, either the multiple selection model in the &PSALMS program of OSIRIS IV at the University of Michigan or the replication methodology discussed by Gurney[21] can be used for these calculations.

Continuous variables: Means

Means and standard errors were estimated for several variables to investigate the relative effects of the sampling weights and the sampling design on the estimates. These results are displayed in table 11 for four variables—number of decayed, missing, and filled teeth, systolic blood pressure, calories consumed daily, and age.

For the total sample, the unweighted and weighted analyses (options 1 and 2) for these variables are similar for the means and variances. However, the complex sample design introduces a considerable increase in the estimated variance of the mean (option 3). The ratio of the standard error of the mean under option 3 to that obtained under option 2 (shown in the last column in table 11) ranges from 1.71 to 2.73. Consequently, the design effects range from 2.92 to 7.43.

One might expect the design effects to be smaller when stratifying into subclasses such as age groups. This expected reduction is due to the clustering effect which is both a function of subclass size as well as the homogeneity coefficient. The latter is the extent to which persons in the same subclass tend to have similar responses within clusters. Thus, unless the homogeneity coefficient increases for smaller subclasses, the design effect will be smaller for the age subclasses than for the overall sample. To investigate this possibility, means, standard deviations, and standard errors of the means of these variables were computed within age groups shown in table 12. Although the design effects are somewhat reduced, they are not negligible, ranging from 1.48 to 5.07.

Subgroup comparisons: Means

Many hypotheses involve the comparison of two subgroup means. Because of the clustered design and the sampling weights, the difference between the mean response for each subgroup was computed as the difference between two weighted ratio means within the context of the &PSALMS routine described by Vinter.[20]

To assess the effects of the sampling weights and

the complex sample design on the magnitude of the t-statistics associated with the tests for these differences, a representative analysis was investigated under options 1–3. In particular, the mean systolic blood pressure was compared for two subclasses determined by the lowest 15th percentile and highest 15th percentile of skinfold thickness in selected age by race subgroups. These results are shown in table 13 under each of the three analysis options. In all subgroups, the simple random sample estimates for the unweighted and weighted analyses are similar, both for the means and variances. However, the complex sample design introduces a considerable increase in the estimated variance of the difference in the means between the two subclasses. Specifically, the ratio of the standard error of the difference of the mean under option 3 to that obtained under option 2 in the last column in table 13 ranges from 1.3 to 2.0. Thus, the design effects for estimated means range from 1.7 to 4.0. In other words, the t-statistic computed under option 2 is from 1.3 to 2.0 times larger than that computed under option 3 because the variances under option 3 are larger.

Continuous variables: Multiple regression models

The basic model used for assessing the joint effects of several predictor variables on the variation of a continuous response variable is the multiple regression model. The general super-population model is

$$Y_i = B_1 + B_2X_{2i} + B_3X_{3i} + \ldots + B_kX_{ki} + E_i \quad (1)$$

where Y_i denotes the ith observation of the dependent variable, X_{ki} denotes the ith observation of the kth independent or explanatory variable, and E_i is the random variation of the ith observation of Y. The subscripts $2, 3, \ldots, k$ identify the specific explanatory variables. B_1 is the intercept term, and B_k is the change in the expected value of Y_i corresponding to a unit change in the kth explanatory variable, holding all other explanatory variables constant. $B_2, B_3, \ldots, B_k$ are often referred to as the regression slopes or (partial) regression coefficients.

Alternatively, multiple regression models can be developed in terms of standardized independent variables. This approach leads to standardized estimators usually referred to as beta coefficients. The beta coefficients are the result of a linear regression in which each variable is "normalized" by subtracting its mean and dividing by its estimated standard deviation or sum of squares about the mean. In other words, the beta coefficient adjusts the estimated slope parameter by the ratio of the standard deviation of the independent variable to the standard deviation of the dependent variable. In this formulation, the model does not have a constant or intercept term. A beta coefficient of 0.3 may be interpreted to mean that a standard deviation change of 1.0 in the independent variable will lead to a 0.3 standard deviation change in the dependent variable. Beta coefficients are also used to make statements about the relative importance of the X variables in the model.

Assumptions of the multiple regression model

The classical assumptions associated with the regression model are

1. The model specification is correct.
2. The X's are nonstochastic. In addition, no exact linear relationship exists among two or more of the independent variables.
3. The E_i are independent, identically distributed as $N(0,\sigma^2)$.

Any set of real data is unlikely to meet all of these assumptions, particularly one utilizing a complex survey design such as the NHANES I. However, certain violations of these assumptions may not seriously affect statistical inferences. For example, under simple random sampling arguments, it is straightforward to show that the least squares estimators of the regression coefficients retain their desirable asymptotic properties (unbiased, consistent and efficient) when the X's are stochastic (i.e., a violation of the second assumption) provided that the explanatory variables are each distributed independently of the true errors in the model (see, for example, Kmenta[22]). More detailed discussions of the properties of regression model estimates from complex sample surveys can be found in Holt, Smith, and Winter.[23]

Specification error

If any variables are omitted from the regression equation that are correlated with both the dependent variable and the independent variable(s) included in the model, the estimates of those regression coefficients will be biased. This particular problem is the reason a multivariable (rather than a series of bivariate) estimation procedure may be required when investigating a phenomenon that has multiple, interrelated causes. For example, in the relationship between dietary intake patterns and dental caries experience, if a variable such as age is omitted, biased estimates of that relationship emerge because there is a correlation between age and dental caries experience. In spite of the effort to include all of the theoretically important variables in the model, if some have been omitted, either because they were not part of the data collected or theory has not yet advanced sufficiently to implicate them, the estimators given by the model could be biased.

Another concern with specification error is the actual mathematical relationship between the response variable and the joint distribution of the independent variables in the model. If the true relationship is, for

example, logarithmic, the specification of the model as linear may lead to biased and inconsistent parameter estimates. Therefore, careful attention to all available theoretical knowledge concerning the relationships involved is essential.

Some of the relationships studied might be better represented by a series of simultaneous interdependent equations. For example, the symptoms of periodontal disease could influence the frequency of dental visits, dental visits could influence toothbrushing behavior, and toothbrushing could affect periodontal disease. In such a circumstance, ordinary least squares estimation of individual equations can lead to biased and inconsistent parameter estimates. While these forms of possible misspecification probably do not pose a serious threat to the conclusions reached by Burt,[7] they do warrant future exploration to more precisely assess the underlying form of these relationships.

Measurement error

When variables are measured with error, they can affect the results of statistical procedures applied to them. In general, considerable effort was expended to ensure a minimum of observer error in the gathering of NHANES I data. Potential problems concerning some variables and the procedures employed to minimize some of these are described in *Vital and Health Statistics, Series 11-No. 225*.[7] Consider, for example, the group of nutritional and dietary variables from the 24-hour recall record. There are short-term and long-term variations in what people eat. Therefore, the 24-hour recall record is an imperfect measure of long-term dietary patterns. This kind of random error in an independent variable in a regression equation will bias the estimate of the regression coefficient of that variable toward zero. Under simple random sampling arguments, it is possible to demonstrate that the form of the bias is

$$B' = B/(1 + \lambda)$$

where

- B' is the biased estimate of the regression parameter as computed by ordinary least squares,
- B is the unbiased estimator, and
- λ is the ratio of the true variance to the additional variance attributable to the measurement error.

(See for example, Snedecor and Cochran.[24]) The extent to which the bias in estimates of regression coefficients can be expressed in this formulation under the complexities of option 3, utilizing both the sampling weights and the survey design effects, requires further investigation.

Because some empirical work has provided estimates of the ratio of interindividual (true between subject variation in individual intake) to intraindividual (day-to-day variation in individual intake) variation, rough estimates of this bias are possible.[25] These data suggest that values of λ of at least one or two are not unreasonable. Based on this information, the relationships between dental caries experience and diet in one 24-hour period are, as estimates of the relationship between dental caries experience and lifetime dietary patterns, underestimates by a factor of 1/2 to 1/3. Stated another way, estimates based on lifetime data are likely to be two or three times larger than those provided by the 24-hour data.

When a variable with this type of error is used as a dependent variable, as in the investigation of the effect of dentulous state on dietary patterns, the problem encountered is less severe. Standard errors will be overestimated, but estimators will be unbiased. Therefore the only real hazard is the failure to reject the null hypothesis when it should be rejected.

Heteroscedasticity

When assumption 3 is violated, standard errors estimated by ordinary least squares tend to be inefficient. Because the variance of variables such as DMF and PI measures tends to increase with age, the possibility of this phenomenon influencing the results presented should be investigated. Weighted least squares procedures may be required when heteroscedasticity is a problem. The extent to which this correction procedure is sufficient under the complexities of option 3 requires further investigation.

Nonnormality of random variation term

Dependent variables such as DMF teeth and PI have distributions that are skewed toward zero in the younger age groups. The random variation term is, therefore, not normally distributed. In some instances, transformations may be employed to provide reasonable approximations of normality. In others, where transformations are of little value, it still may be possible to employ multiple regression models as though the disturbances are normally distributed, because the procedure is considered to be relatively robust when sample sizes as large as occur in analysis of the NHANES I are used.

Empirical results for regression models

To investigate predictive relationships among continuous variables, multiple regression models also can be fitted under either option 1, 2 or 3. Specifically, the effects of the sampling weights and complex design on the precision of regression coefficients were investigated

under options 1–3 for the number of decayed, missing, and filled (DMF) teeth, systolic blood pressure (SBP) and calories consumed daily regressed on age within race-sex subclasses as summarized in table 14. First, it can be observed in the corresponding entries under options 1 and 2 that the results are similar, particularly for DMF teeth on age and SPB on age. These both have a strong linear relationship in all the race-sex subclasses. However, for calories on age, which has extremely small R^2 values for all subgroups, the estimate of the slope is quite different for some subclasses. For the "other males" there is a 12-fold increase in the slope under option 2 compared to option 1 and for the "other females" it differs by a factor of nearly three. Of course, in both of these subclasses the sample size is relatively small.

The square root of the design effects for the slope parameter estimates in these simple linear regression models for the white race data are displayed in the last column of table 14. In particular, these quantities range from 0.98 to 1.85 for each of these three variables; the design effect is smaller for men than for women.

In table 14, the results under option 3 are reported only for the white subgroups even though the number of black persons examined appears to be reasonably large. This omission is due to the failure of the balanced half sample routine in the weighted regression program when entire strata had no data in an early version of the &REPERR program in OSIRIS IV (this routine now has been modified to allow for empty strata). The problem of missing data for black persons in some SECU's as shown in table 7 is even more pronounced within the more restrictive detailed examination, as displayed in table 8, and for persons of other races. Consequently, due to the sparse design across strata, only the data for the white and black races were used in many of the analyses.

In addition to simple linear regression models, multiple regression models also can be fitted within this same framework. As discussed previously, the paired SECU's for each of the 70 strata were utilized in the balanced half sample routine &REPERR to generate estimated variances for the estimated slope parameters. Table 15 summarizes the results for 6,349 persons ages 11–30 years of DMF regressed jointly on age (in single years), race (1 = white, 2 = black), sex (1 = male, 2 = female), and sweets, which is the sum of the reported frequencies for the ingestion of food from the three categories of desserts and sweets, candy, and beverages (sweetened, carbonated, and noncarbonated). In this model, the design effects for the regression coefficients range from 1.25 to 4.49. Similarly, the results of a multiple regression model for 13,573 people ages 18–74 years of systolic blood pressure regressed jointly on age, race, sex and Quetelet's Index of body mass expressed in kg/cm^2 units are displayed in table 16. Here again, the design effects for the regression coefficients range from 2.69 to 3.61.

These empirical results, expressed in terms of estimated design effects, demonstrate the effects of incorporating the sampling weights and the survey design adjustments into multiple regression models. The decision of when to incorporate sampling weights and design features into the analysis depends on more than a recognition of the potential errors in inference that can arise because of such effects. Some analysts argue that when making a model-based inference from survey data about a super-population model, one may ignore the sampling design features, even in a design as complex as NHANES I. However, many survey practitioners argue that a design-based inference, as illustrated here, is more appropriate for survey data, especially when examining exploratory models for which the specification of the model is likely to be in error. Accounting for unequal probabilities of selection and other design features in the design-based approach recognizes that the model may be misspecified and that somewhat conservative inferences are desired. Further, the model of interest in the design-based approach is appropriately one in which the model refers directly to the finite population from which the sample was selected. In this and subsequent sections, the analytic perspective for survey data is the design-based view of inference for complex sample survey data.

Continuous variables: Analysis of variance

The familiar analysis of variance (ANOVA) situation involves a set of factors $X_1, X_2, \ldots, X_p$ each of which may have several levels. These factors are used to explain the variability in a response variable Y. In general, an appropriate measure of total variation for Y, such as the total corrected sum of squares, is partitioned into individual components each attributable to a factor or group of factors. The usual hypothesis tests require the assumption of equality of variances and zero covariances among subgroups and the assumption of simple random sampling. ANOVA for data from complex surveys such as NHANES I requires alternative considerations. Because of unequal probabilities of selection, the clustered design, and the adjustment weights for nonresponse and poststratification, the mean response for each subclass (or domain), determined by the cross-classification of the relevant factors, is computed as a weighted ratio mean. Consequently, the variance-covariance structure of these weighted ratio means must be incorporated into the ANOVA tests when attempting to identify the statistically important sources of variation.

One approach is the large sample methodology utilizing weighted least squares algorithms for the computation of Wald statistics originally described by Grizzle, Starmer and Koch[26] for the analysis of multivariate categorical data. This general methodology was modified and applied to data from complex sample

surveys in a series of papers[26–30] using data from another NCHS Survey, the National Health Interview Survey. A brief outline of the application of this methodology to data from the NHANES I is presented in appendix I of the *Vital and Health Statistics Series 11-No. 209*,[31] Koch and Stokes[32] and Koch, Stokes, and Brock.[33] In essence, this strategy involves a vector $\underset{\sim}{F}$ of subclass or domain ratio means, together with an appropriate, valid, and consistent estimate of the covariance matrix V_F of these means, and the framework of a general linear model. Consequently, the usual ANOVA hypotheses about which factors or combinations of factors make statistically significant contributions to the variation among these domain means can be investigated by fitting linear models to the vector of means by the method of weighted least squares relative to the estimated covariance matrix.

Quite a few different approaches can be used to estimate the covariance structure of the ratio means. One method is to use the balanced repeated replication (BRR) strategy described by McCarthy[34] and Kish and Frankel.[35] Several different variations of this replication approach were investigated empirically within the context of National Health Interview Survey (NHIS) data as reported in Freeman, Freeman, Brock and Koch.[29] On the other hand, for paired designs in which there are exactly 2 SECU's within each stratum such as the NHANES I design, as well as for other multistage sample designs, direct methods involving sums of squared differences and cross-products can be utilized to obtain estimates of the variances of the numerators and denominators of the ratio means. These variances and covariances can be incorporated directly into a linear Taylor series expansion for the estimated covariance structure of the ratio means. This particular direct approach to the estimation of the covariance matrix is described for contingency table proportions expressed as ratio means in Lepkowski and Landis[36] and Lepkowski.[37] These calculations are directly analogous to those for ratio means in general.

The ANOVA results presented here were obtained in two stages of computing and data analysis. First, the vector of ratio means of the dependent variable, together with their estimated variances and covariances, were computed directly within the OSIRIS.IV package using the &PSALMS routine. Any computing algorithm designed to generate a vector of ratio means and a consistent estimate of its covariance structure under the complex sampling design can be utilized to obtain these estimates. At the second stage of computing, the vector of sample means and its covariance matrix were entered directly into the weighted least squares program, GENCAT (see Landis, Stanish, Freeman, and Koch[38]) to perform the various ANOVA hypothesis tests and final model-fitting computations. The specific command files used to generate the results for the ANOVA example in the subsequent section are listed in appendix II.

ANOVA methodology

Consider a linear model for the vector of g subclass or domain ratio means $\underset{\sim}{F}$ as

$$E_A\{\underset{\sim}{F}\} = X\underset{\sim}{B}, \tag{2}$$

where X is a $(g \times u)$ matrix of known constants with $u \leq g$ and rank $(X) = u$, $\underset{\sim}{B}$ is a $(u \times 1)$ vector of unknown parameters, and $E_A\{.\}$ denotes the asymptotic expected value of the argument $\{.\}$. The weighted least squares (WLS) estimator for the parameter vector $\underset{\sim}{B}$ can be obtained from the survey data as

$$b = (X'V_{\underset{\sim}{F}}^{-1}X)^{-1}X'V_{\underset{\sim}{F}}^{-1}\underset{\sim}{F}. \tag{3}$$

A useful feature of the WLS procedure is the ability to characterize the variation among the functions $\underset{\sim}{F}$ in terms of the factors specified in the design matrix X. The goodness of fit of such models to the data can be evaluated and more parsimonious models (i.e., models with fewer parameters) can be developed that may lend insight to the substantive phenomena underlying the data. To test the goodness of fit of the model to the data, a Wald statistic

$$Q = (\underset{\sim}{F} - X\underset{\sim}{b})'V_{\underset{\sim}{F}}^{-1}(\underset{\sim}{F} - X\underset{\sim}{b}) \tag{4}$$

is computed. When the model X in (2) holds, and the data come from the usual multinomial or product multinomial distribution with simple random sampling, Q asymptotically follows the chi-square distribution with $(g - u)$ degrees of freedom. Moreover, even when the covariance matrix is estimated from a complex sample design, authors such as Shuster and Downing[39] assert that the same result holds for large samples.

If the goodness of fit of the model to the data is adequate by the Wald statistic criteria in (4), it usually is desirable to identify more parsimonious models by examining individual parameters in the model. The exploration of reduced models is conducted through tests of hypotheses of the form

$$H_0: C\underset{\sim}{B} = \underset{\sim}{0} \tag{5}$$

for some $(d \times u)$ hypothesis matrix of known constants C for which $d \leq u$. The test statistic

$$Q_C = (C\underset{\sim}{b})'[C(X'V_{\underset{\sim}{F}}^{-1}X)^{-1}C']^{-1}C\underset{\sim}{b} \tag{6}$$

is asymptotically distributed as a chi-square random variable with d degrees of freedom when H_0 is true. Repeated application of hypothesis tests such as in (5) and (6) leads to the development of a reduced model for which predicted values then can be obtained as

$$\hat{\underset{\sim}{F}} = X_R\underset{\sim}{b}_R, \tag{7}$$

where X_R is the "reduced model design matrix" and $\underset{\sim}{b}_R$ is the WLS vector of estimated reduced model parameters. Within the context of X_R, the estimated variance-covariance matrix of $\underset{\sim}{\hat{F}}$ is then

$$V_{\underset{\sim}{\hat{F}}} = X_R(X_R' V_{\underset{\sim}{F}}^{-1} X_R)^{-1} X_R'. \quad (8)$$

Such predicted values will adequately characterize the statistically important sources of variation in the data and will have smaller sampling errors than the original estimates. The estimated variances of the functions in $\underset{\sim}{F}$ in (2) are based on the sample sizes in each subclass, whereas the estimated variances of the predicted functions in (7) and (8) are based on the entire sample. Thus, the sampling errors of parameters in the reduced model are smaller than for the original estimates because of the "smoothing" across the entire sample relative to the reduced model.

ANOVA example

Consider the results shown in table 17 for which the dependent variable under investigation is periodontal index (PI), a continuous variable ranging in value between 0.0 and 8.0. The individual and joint effects of current drinking and smoking habits on the variation of PI are to be modeled in the ANOVA framework outlined in the previous section. Drinking—classified as none, little, moderate, and heavy, and smoking—classified as never, past, and now, are to be the subclass or independent variables (see appendix I for detailed descriptions).

The mean PI score under option 1, together with the corresponding standard error and total number of persons examined, is displayed in table 17 for the 12 subclasses determined by the cross-classification of drinking and smoking. Smoking history is available only for those individuals included in the detailed survey; thus, the usable sample size is limited to a maximum of 3,854 adults ages 25–74 years. For this detailed sample, 2,943 adults had complete data both for the drinking and smoking variables. The loss in available data was due primarily to missing data on the smoking variable. If the examined persons with missing smoking information differ from the 2,943 people in the analysis with smoking data, the ANOVA results may be biased.

This example is not necessarily the most informative from either a methodological or substantive point of view. The data ultimately are not characterized by a simple model with strong substantive implications, primarily because they do not adjust for other covariates known to be important in these relationships (see *Vital and Health Statistics* Series 11-No. 225[6]). However, the example is representative of many data screening and model-fitting investigations in large, complex data sets where hypotheses are considered for preliminary inquiry and where significant relationships among variables do no necessarily lead to a simple final model.

Since these data are available only for the detailed sample, the weighted analyses under both options 2 and 3 require the sampling weights associated with the detailed sample (tape location 170–175 on all public-use tapes). If the sampling weight for the ith examined person, w_i, is used to compute a standardized weight $v_i = 3{,}854[w_i/\Sigma_i w_i]$, then the sum of the standardized weights, $\Sigma_i v_i$, is precisely the total number of examined persons, *viz.*, 3,854. The relative size of this weight v_i for each person, when compared to weights for the others in the sample, remains unchanged. All analyses using the $\{v_i\}$ will lead to identical results as the same analysis using the original $\{w_i\}$.

An examination of the sum of the standardized weights $\{v_i\}$ for a given subclass, labeled "Weighted Number Examined" in table 17, can reveal the extent to which a particular subclass is over represented in the sample. For example, for all three smoking categories within nondrinkers, the number examined is larger than the weighted number examined. This indicates that nondrinkers are overly represented in the sample relative to the reference population. On the other hand, all other subclasses have larger weighted sample sizes than the actual number examined (except for heavy drinkers who never smoked for which the weighted and unweighted totals are the same, 30 vs. 29.42). This indicates that this group is under represented in the sample. Moreover, it is interesting to note that the sum of the standardized weights is 3,126.23, indicating that even though the actual number examined with complete data is only 2,943, they actually represent 3,126 of the available 3,854 weighted people in the detailed sample. That is, those with missing data on smoking or drinking had relatively less weight per examined person than those contributing data to the analysis.

Although this discussion and presentation of the weighted number examined by subclass is not required to complete the ANOVA analysis, these weighted totals are critical in the contingency table analyses reported in the next section, and illustrated with some of these same variables. Furthermore, even though the variation in the weighted ratio means can be assessed without these weighted sample sizes being stated explicitly, they are incorporated into both the estimates of the means and the variances through the use of sampling weights.

To incorporate the sampling weights and the complex sample design in analyzing these mean PI scores, the subclass ratio mean PI scores and their corresponding variances and covariances were computed. They utilized simple random sampling calculations under option 2 and Taylor series based calculations under option 3. As discussed in appendix II, these calculations were obtained within the OSIRIS IV software package using the &USTATS and &PSALMS routines. As noted in the last column of table 17, the square root of

the design effect ranges from 0.85 to 1.48; the design effects range from 0.72 to 2.19. Since some of these design effects are less than 1.0, the chi-square statistics will be influenced by the complex sample design in unexpected ways. In particular, if all the design effects were greater than 1.0 for the subclass ratio means, the chi-square statistics for the complex design-based estimates can be expected to be smaller than for estimates computed under options 1 or 2. On the other hand, with some design effects less than 1.0, a chi-square statistic under option 3 may actually be greater than those obtained under options 1 or 2.

For purposes of model fitting and hypothesis testing, the vector $\underset{\sim}{F}$ of the 12 cross-class ratio mean PI scores, together with a diagonal matrix $V_{\underset{\sim}{F}}$ with the corresponding variances under either option 1 or 2, were used to generate the ANOVA results in the formulation outlined in the previous section. The full covariance matrix obtained from the Taylor series based estimated covariances was utilized to perform the ANOVA tests under option 3. Note that this covariance matrix is not a diagonal matrix; subclass means are not independent because subclass elements are not selected independently between subclasses in the sample design.

Initially, the variation among these mean PI scores was investigated using the saturated, reference cell design matrix X_1 in the linear model formulation $E_A(\underset{\sim}{F}) = X_1\underset{\sim}{B}_1$, where

$$X_1 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}, \text{ and}$$

$$\underset{\sim}{B}_1 = \begin{bmatrix} \text{Reference value for non-drinkers who never smoked} \\ \text{Increment for past or current smokers} \\ \text{Increment for current smokers} \\ \text{Increment for drinkers} \\ \text{Increment for moderate or heavy drinkers} \\ \text{Increment for heavy drinkers} \\ \text{(6 additional parameters reflecting drinking by smoking interaction)} \end{bmatrix}.$$

Hypotheses of the form given in (5) can be used to generate chi-square statistics to test the significance of each of the parameters in the model. Moreover, the factorial effects in this design can be obtained by letting C_D be the contrast matrix for the drinking effects, C_S be the contrast matrix for the smoking effects, and C_{DS} be the contrast matrix for the drinking and smoking interaction effects. Specifically, for this design matrix X_1, the corresponding contrast matrices are given by

$$C_D = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

$$C_S = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \text{ and}$$

$$C_{DS} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

The corresponding test statistics associated with these three sources of variation in the mean PI scores under each of the options 1–3 are shown in table 18.

Several observations can be cited at this stage of analysis. First, both drinking and smoking have statistically significant ($P < 0.05$) effects on the variation of the mean PI score under all three analysis options. Based on the non-significant interaction between drinking and smoking, there is little evidence that these individual effects are not the same at each level of the other variable. As expected from the small design effects (with some even less than 1.0) shown in table 17, the test criteria are not influenced strongly by the weights or the complex sample design for this particular variable. The chi-square for the drinking effect is reduced from 29.54 under option 1 to only 26.91 under option 3; the smoking chi-square is reduced from 13.58 to 5.14. On the other hand, the interaction of chi-square is actually increased slightly (2.52 to 3.78), although neither test statistic is statistically significant even at $P = 0.25$.

A reduced variational model reflecting no interaction between drinking and smoking can be fit to the function vector $\underset{\sim}{F}$ by deleting the last 6 columns of X_1, those associated with the interaction parameters. As shown in table 19, the chi-square criteria for both the drinking and smoking effects are increased substantially from their counterparts in table 18 under the saturated model. Note that the test for interaction in table 18 is equivalent to the lack of fit chi-square in the reduced model results presented in table 19, since the interaction

effects are precisely the ones left out of the saturated model to obtain the reduced model.

Although alternative models could be considered for these data, these results illustrate the effects of the weights and the design effects on ANOVA test criteria. For these particular hypotheses, the results were not influenced greatly by either the weights or the complex design effects. However, for other variables within the NHANES I survey this similarity of results is not likely to hold.

Categorical variables

The multifactor, multiresponse framework for categorical data described by Bhapkar and Koch[40] provides a general approach to the analysis of multidimensional contingency tables from sample surveys. For many types of categorical data analyses, independent variables or factors and dependent variables or responses are specified. The observations are then cross-classified by these variables to create a multiway contingency table of frequencies. Within this framework, the independent variables are cross-classified into a set of mutually exclusive subpopulations or subclasses. Similarly, the dependent variables are cross-classified to determine the levels of the potentially multivariate response profiles. Thus, the entire distribution of these response profiles (or selected functions of this distribution) are estimated within each subpopulation or subclass. This terminology—factors and responses—has been borrowed from the classical experimental design setting to provide an analogous context for the analytical strategies for these observational categorical data, even though they were not obtained from a similarly designed experimental research investigation.

Contingency table notation and methodology

Consider a finite population of N elements from which a probability sample is to be selected. Suppose that this population is divided into s distinct subclasses and subclass elements can be classified into one and only one category of an r dimensional response profile. Specifically, let N_{ij} denote the number of population elements in the ith subclass classified into the jth category of the response profile. Let $N_{i.}$ denote the total number of elements in the ith subclass. Analysis of categorical data from this finite population is concerned with the proportions $P_{ij} = N_{ij}/N_{i.}$, the proportion of elements in the ith subclass that are in category j of the response profile, where $i = 1, \ldots, s$ and $j = 1, \ldots, r$.

Suppose a sample of size n is selected from the finite population and $n_{i.}$ sample elements are in the ith subclass. Let n_{ij} denote the number of sample elements in the ith subclass in category j of the response profile. The sample proportion of elements in subclass i and response profile category j can be denoted as $p_{ij} = n_{ij}/n_{i.}$. Thus, the resulting sample contingency table can be displayed as indicated in table 20.

Suppose now that a complex multistage sample design has been used to select the sample of size n. For simplicity, suppose the N population elements have been divided into H subgroups or strata and that the population elements within the strata are grouped into A clusters. Let A_h denote the number of clusters in stratum h, where $h = 1, 2, \ldots, H$. From each stratum a probability selection of a_h clusters is selected; within each selected cluster, a probability selection of n_{ha} sample elements is selected. Thus, the sample size $n = \sum_h \sum_a n_{ha}$. Assume that the stratum sizes are large so that the finite population corrections can be ignored, or that the design can be closely approximated by a with replacement sampling strategy.

Given the contingency table shown in table 20, the analytic task is to make inferences about the characteristics comprising the factors and the responses given that the data were generated from a sample with a complex design. Before discussing an analytic strategy for such inferences, some computational features of estimating the sample proportions and associated variances and covariances from such a survey design should be reviewed.

Suppose that each sample element has been assigned some weighting factor w_{hak}, where $h = 1, \ldots, H$, $a = 1, \ldots, a_h$, and $k = 1, \ldots, n_{ha}$. These weights may reflect several of the following features of the design simultaneously: unequal probabilities of selection, poststratification, and nonresponse adjustment. A set of indicator functions can be created that allows the estimation of sample proportions and other estimates from the survey observation.

For each sample element, define the $(s \times r)$ indicator variables

$$y_{ijhak} = \begin{cases} w_{hak}, & \text{if the } (hak)\text{-th sample element is in the } i\text{th subclass and } j\text{th category of the response profile,} \\ 0, & \text{otherwise,} \end{cases} \quad (9)$$

where $i = 1, \ldots, s$ and $j = 1, \ldots, r$. For a single sample element, the indicator variables are all zero except for one corresponding to the cell of the table into which the element is classified.

These variables can be used to compute quantities useful for the estimation of the proportions and other parameters. In particular, the weighted total sample elements in the (i,j)-th cell of the table for the (h,a)-th cluster can be expressed as

$$n_{ijha} = \sum_k y_{ijhak}, \quad (10)$$

with the ith subclass total as

$$n_{iha} = \sum_j \sum_k y_{ijhak}. \quad (11)$$

Summing these cluster totals across clusters (and strata), the overall sample weighted cell and subclass totals obtained are

$$n_{ij} = \sum_h \sum_a n_{ijha} \tag{12}$$

and

$$n_{i.} = \sum_j n_{ij}. \tag{13}$$

From these sample totals, the proportion of elements in subclass i and response category j can be estimated as

$$p_{ij} = n_{ij}/n_{i.}. \tag{14}$$

The denominator of (14) is not fixed by the sample design and is therefore a random variable. Hence, p_{ij} is a ratio mean and subject to some theoretical difficulties. For one, the variances and covariances of the p_{ij} generally are not known exactly. In practice, however, a first order Taylor series expansion approximation can easily be computed for the variance of ratio means, provided by the &PSALMS routine in OSIRIS IV. For example, the estimated variance of (14) can be approximated by

$$\text{var}\,(p_{ij}) = (p_{ij})^2[(n_{ij})^{-2}\,\text{var}\,(n_{ij}) + (n_{i.})^{-2}\,\text{var}\,(n_{i.}) - 2(n_{ij}n_{i.})^{-1}\,\text{cov}\,(n_{ij},n_{i.})], \tag{15}$$

where the variances and covariances var (n_{ij}), var $(n_{i.})$, and cov $(n_{ij},n_{i.})$ are estimated in a manner consistent with the sample design. If, for instance, there are exactly two primary selections in each stratum (i.e., $a_h = 2$ for $h = 1, \ldots, H$), the variances and covariances of the cell and subclass totals can be computed as

$$\text{var}\,(n_{ij}) = \sum_h (n_{ijh1} - n_{ijh2})^2, \tag{16}$$

$$\text{var}\,(n_{i.}) = \sum_h (n_{i.h1} - n_{i.h2})^2, \tag{17}$$

and

$$\text{cov}\,(n_{ij},n_{i.}) = \sum_h (n_{ijh1} - n_{ijh2})(n_{i.h1} - n_{i.h2}). \tag{18}$$

An approximation similar to (15) can be given for the covariance of two proportions. In particular,

$$\text{cov}\,(p_{ij},p_{i'j'}) = (p_{ij}p_{i'j'})\,[(n_{ij}n_{i'j'})^{-1}\,\text{cov}\,(n_{ij},n_{i'j'}) + (n_{i.}n_{i'.})^{-1}\,\text{cov}\,(n_{i.},n_{i'.}) - (n_{i'j'}n_{i.})^{-1}\,\text{cov}\,(n_{i'j'},n_{i.}) - (n_{i'.}n_{ij})^{-1}\,\text{cov}\,(n_{i'.},n_{ij})].$$

Replication procedures also can be used to compute estimates of these variances and covariances.

The estimation procedures for the proportions p_{ij} and the corresponding variances and covariances can be summarized compactly in vector notation. For each sample element, let the $(sr \times 1)$ vector of observed values y_{ijhak} be denoted as

$$\underset{\sim}{y}'_{hak} = [y_{11hak}, y_{12hak}, \ldots, y_{srhak}]. \tag{19}$$

Summing these vectors across elements in each primary selection, the weighted cluster totals are obtained as

$$\underset{\sim}{y}_{ha} = \sum_k \underset{\sim}{y}_{hak}. \tag{20}$$

Let K denote an $[s(r+1) \times sr]$ linear operator matrix that can be used to generate a vector of cell and subclass totals for each cluster from the $\underset{\sim}{y}_{ha}$. Specifically, let

$$K = \left[\frac{I_{sr}}{\underset{\sim}{1}'_r \otimes I_s}\right], \tag{21}$$

where I_{sr} and I_s denote sr and s dimension identity matrices, respectively, $\underset{\sim}{1}_r$ denotes an r dimension vector of one's, and $\otimes$ denotes the Kronecker or direct product (see section 8.8 of Graybill).[41] Then the $[s(r+1) \times 1]$ vector of cell and subclass totals denoted as $\underset{\sim}{n}_{ha}$ can be expressed as

$$\underset{\sim}{n}_{ha} = K\underset{\sim}{y}_{ha}. \tag{22}$$

Summing the cluster vectors $\underset{\sim}{n}_{ha}$ across clusters, the $[s(r+1) \times 1]$ vector of weighted cell and subclass totals is

$$\underset{\sim}{n} = \sum_h \sum_a \underset{\sim}{n}_{ha}. \tag{23}$$

or,

$$\underset{\sim}{n}' = [\underset{\sim}{n}_{11}, \ldots, \underset{\sim}{n}_{sr}, \underset{\sim}{n}_1, \ldots, \underset{\sim}{n}_s].$$

As noted previously, the variances and covariances of the elements of $\underset{\sim}{n}$ depend on the nature of the sample design. If, as before, a paired selection type of design is used, then the $[s(r+1) \times s(r+1)]$ variance-covariance matrix of $\underset{\sim}{n}$ can be estimated as

$$V_{\underset{\sim}{n}} = \sum_h (\underset{\sim}{n}_{h1} - \underset{\sim}{n}_{h2})(\underset{\sim}{n}_{h1} - \underset{\sim}{n}_{h2})'. \tag{24}$$

A more complete discussion of the conditions under which estimators such as (24) are appropriate is given in Kish and Hess.[42]

A series of transformations can now be applied to $\underset{\sim}{n}$ and $V_{\underset{\sim}{n}}$ to obtain an $(sr \times 1)$ vector of the sample proportions p_{ij}, and an $(sr \times sr)$ matrix of the variances and covariances of the p_{ij}. Let $\exp(\cdot)$ and $\log(\cdot)$ denote matrix operators which take the natural exponent and logarithm, respectively, of every element of the matrix

argument $(\cdot)$. Also let A denote the $[s(r+1) \times sr]$ linear operator matrix

$$A = [I_{sr} \mid -\underset{\sim}{1}_r \otimes I_s]. \tag{25}$$

Then the $(sr \times 1)$ vector of sample proportions, denoted $\underset{\sim}{p}$, can be obtained as

$$\underset{\sim}{p} = \exp[A(\log(\underset{\sim}{n}))]. \tag{26}$$

The Taylor series approximation to the variance-covariance matrix of $\underset{\sim}{p}$, denoted $V_{\underset{\sim}{p}}$, is the transformation of $V_{\underset{\sim}{n}}$

$$V_{\underset{\sim}{p}} = D_{\underset{\sim}{p}} A D_{\underset{\sim}{n}}^{-1} V_{\underset{\sim}{n}} D_{\underset{\sim}{n}}^{-1} A' D_{\underset{\sim}{p}}, \tag{27}$$

where $D_{\underset{\sim}{p}}$ and $D_{\underset{\sim}{n}}$ denote $(sr \times sr)$ diagonal matrices with the elements of the vectors $\underset{\sim}{p}$ and $\underset{\sim}{n}$, respectively, along the diagonal.

The matrix operations presented in expressions (19) through (27) are straightforward generalizations of the results in expressions (10) through (18). In addition, compounded sets of transformations, as in (26), can be developed for the vector $\underset{\sim}{p}$ to obtain functions of the sample proportions, such as $f_1(\underset{\sim}{p}), \ldots, f_g(\underset{\sim}{p})$, with their corresponding Taylor series approximations to the variances and covariances as discussed in Forthofer and Koch[43] and appendix I of the report by Koch, Landis, Freeman, and others.[44] Thus, a vector of functions of sample proportions, denoted as $\underset{\sim}{F}(\underset{\sim}{p}) = [f_1(\underset{\sim}{p}), \ldots, f_g(\underset{\sim}{p})]$, and its variance-covariance matrix approximation $V_{\underset{\sim}{F}}$, can be computed directly from the "raw" survey data.

The utility of this method is enhanced by the ability to obtain estimates of $V_{\underset{\sim}{n}}$ in (24) for other survey sample designs besides paired selections. Multistage designs with multiple clusters in each stratum, stratified random samples, or systematic selections of clusters, are a few of the other designs that can be handled by this method.

The integration of this computational procedure for categorical data from sample surveys with the flexible WLS procedure for analyzing categorical data has been implemented by Lepkowski[37] and, with a slightly different computational procedure, by Freeman.[45] Provided a consistent estimate of the variance-covariance matrix of the vector of functions $\underset{\sim}{F}(\underset{\sim}{p})$ is available, the WLS procedure outlined in ANOVA methodology can be applied directly to the function vector and its covariance matrix associated with the contingency table.

Contingency table example

The data in table 21 were obtained from the cross-classification of current cigarette smoking status, race, and Periodontal Index. Since smoking history is available only for those individuals included in the detailed survey, the usable sample size is limited to a maximum of 3,854 adults ages 25–74. For this detailed sample, smoking data was available for 2,948 persons. After eliminating persons whose race was not white or black, there were 2,919 examined persons with complete data on these variables.

The primary hypothesis addressed in this analysis is the relationships between the factors, cigarette smoking and race, and the response variable, periodontal index (PI). However, as discussed in considerable detail in *Vital and Health Statistics,* Series 11-No. 225,[6] the relationship between PI and such factors as smoking and race are highly influenced by other covariates such as frequency of tooth brushing. Consequently, the analyses of these frequency data will illustrate the contingency table methodology. They do not suggest substantive conclusions.

The weighted frequency distribution of PI score and the proportion classified PI (Some) for each of these subclasses is shown in table 22. They are based on the weights for the detailed survey (tape location 170–175 on all public-use tapes) after standardization of the weights to sum to the total number examined in the detailed sample. In this context, it is critical that the weights be standardized to the number of examined persons in the detailed sample. Otherwise, the weighted frequencies will be population estimates and the analysis will not be conducted relative to the sample sizes actually utilized in the survey. In particular, note that the weighted frequencies for black people are all smaller than those actually examined. This reflects the oversampling in the design. Overall, the proportion of people estimated to have some periodontal disease is approximately 4 percent lower using the weights as compared to the unweighted estimates displayed in table 21 (51.8 percent vs. 55.7 percent).

To incorporate the effects of the complex sample design in the analysis of these contingency table data, the proportions in table 22 were also computed under option 3 as a vector of ratio means using the methodology discussed in the previous section. Consequently, the Taylor series-based variance estimates for these proportions were generated directly from the &PSALMS routine for this analysis. These results, together with the estimates obtained under options 1 and 2, are shown in table 23. As noted in the last column, for these proportions the square root of the design effects ranges from 1.22 to 1.88; the design effects range from 1.48 to 3.53. Note that the design effects for the proportions with PI (Some) are smaller for the current cigarette smokers than for those not smoking, regardless of race.

For purposes of model fitting and hypothesis testing, the vector $\underset{\sim}{F}$ of the subclass proportions with PI (Some) and its corresponding covariance matrix $V_{\underset{\sim}{F}}$ are shown in table 24.

Initially, the variation among these proportions was investigated using the usual 2^2 factorial design matrix X_1 in the linear model formulation $E_A(\underset{\sim}{F}) = X_1 \underset{\sim}{B}_1$,

where

$$X_1 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}$$

and

$$\underset{\sim}{B}_1 = \begin{bmatrix} \text{Overall mean} \\ \text{Differential effect for whites} \\ \text{Differential effect for smokers} \\ \text{Interaction effect} \end{bmatrix}$$

The estimated parameter vector for this model is obtained from the weighted least squares routine as $\underset{\sim}{b}_1 = (0.606, -0.095, 0.029, 0.033)$. Thus, hypotheses of the form $H_0: C\underset{\sim}{B} = \underset{\sim}{0}$ can be used to investigate the relative importance of these factors and their interaction in contributing to the variation among the proportion with PI (Some). Specifically, the three hypotheses for this model, together with their corresponding contrast matrices and resulting test statistics, are shown in table 25.

For comparative purposes, these hypotheses were also tested under options 1 and 2, using the frequency data from tables 21 and 22, respectively. As shown in table 26, the importance of incorporating the design effects into the test criteria is quite pronounced. In particular, under option 1, each of the 3 hypotheses would have been rejected at the usual 5 percent level of significance.

Although the race effect is highly significant, there is evidence among the estimates in $\underset{\sim}{F}$ and from the marginally significant test ($p = 0.09$) for H_3 in table 25 that smoking has a differential effect across race subclasses. To investigate this possibility formally, the variation among the estimates was characterized by the linear model $E_A(\underset{\sim}{F}) = X_2\underset{\sim}{B}_2$, where

$$X_2 = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & -1 \\ 1 & -1 & 0 \\ 1 & -1 & 0 \end{bmatrix}$$

$$\underset{\sim}{B}_2 = \begin{bmatrix} \text{Overall mean} \\ \text{Differential effect for race} \\ \text{Smoking effect for whites} \end{bmatrix}$$

As shown in table 27, the goodness of fit statistic (i.e., $Q = 0.01$) indicates that X_2 provides an adequate characterization for the variation among these proportions. Moreover, the smoking effect for the white population is highly significant ($Q = 38.63$) compared to the nonsignificant average smoking effects indicated in table 26 under X_1. Moreover, the square root of the design effect is less than 1.05 for this smoking test statistic, whereas it was 1.45 under X_1.

Several remarks of caution about the model building approach used here are appropriate. First, the reduced model could be criticized from the vantage point of "overfitting" models to data. Clearly, the lack of fit statistic is extremely small. On the other hand, the model is consistent with the data in that the proportions are nearly identical for the black population regardless of smoking status. However, the objective of the model building is to find a linear model that adequately describes the variation in observed proportions and offers substantively appealing explanations for the relationships among the variables of interest. Thus, the reduced model is in a certain sense "overfitted." However, it offers the substantive expert insight to complex relationships through a relatively straightforward linear model framework. Second, the model was not obtained by successively fitting models until the best one was discovered. From a classical hypothesis testing point of view, one should always investigate H_3, the interaction hypothesis, prior to testing H_1 and H_2, the main effects hypothesis. However, in this case, a more informative model was proposed by noting the "nested" effect of smoking. Thus, the issues of hierarchical testing should be considered carefully when proceeding with model reduction. Finally, the appropriate significance level to be applied for the individual hypothesis tests in the model building is not clearly specified. Generally a "level" of 0.05 has been used as an acceptable criteria in choosing among models. But the particular significance level is arbitrary in a model building framework. The reduced model is the result of careful examination of the observed proportions in each subclass and the goodness of fit and hypothesis test statistics. The application of formal hypothesis testing is inappropriate in such an approach, but the methods and terminology are utilized nonetheless.

Summary

The analysis of data from large complex sample surveys is not a straightforward task. The analyst must consider many issues to develop an appropriate and efficient strategy for conducting the analysis. Such considerations should include not only the technical issues of the sample design, weights, or underlying assumptions in the analytic procedures to be applied, but the fundamental inference issues concerning the nature of models to be developed from the data.

Perhaps the first consideration should be inferential: are inferences to be made to the finite population from which the sample was selected or to some superpopulation or theoretical model of which the finite population may be a single realization? The two approaches to inference for survey data—model-based and design-based—each offer the analyst difficult choices. The choice of a model-based inference leads to analytic strategies that ignore the complexities of the sample design and allow analysts to use routine statistical software for calculations. On the other hand, the model-based inference often requires stronger assumptions for the particular problem than does design-based inference. Further, the model is assumed to be perfectly specified.

The approach presented here has been a design-based type of inference. Computationally, the design-based approach to inference is more difficult to develop and more costly to apply than is the model-based approach. But many survey practitioners feel that it offers advantages when developing exploratory models from survey data. For example, consideration of weights, which account for unequal probabilities of selection adjustment for nonresponse, and adjustment for coverage errors in the analysis, can protect the analyst from some types of misspecification error. Further, with use of variance and covariance estimates, which account for the complexity of the design, design-based inferences tend to be somewhat conservative compared with the model-based approach.

Some analysts argue that an hypothesis testing framework for survey data from a finite population concerning finite population parameters is inappropriate. The perspective taken here has been that if the complexity of the survey design is taken into account, substantively useful inferences are possible by applying existing model building methodologies to survey data.

If a design-based inference approach is chosen, the technical details of the sample design must be considered. The discussion in previous sections has indicated that the following design features ought to be carefully considered:

- What is the nature of the sample design? Was a stratified multistage sample design used? Were unequal probabilities of selection applied?
- Were there adjustments for nonresponse or coverage errors? Is there a weight variable or are there several weight variables that must be applied when different parts of the sample are analyzed?
- Are there important measurement issues that could affect survey analyses? Is item nonresponse an important problem for some variables? Do inter- and intra-observer variability contribute to errors in the data?
- Given the sample design and various sources of error present in the data collection operations, how can estimates be formulated? How can such estimation procedures be incorporated into existing analytic procedures? How can the results be interpreted, and what kind of inferences are appropriate in view of the complex survey design?

These are not all of the issues that can or should be raised in this context. In many instances the analyst probably should consult with a survey practitioner or sampling specialist to resolve the technical issues.

The effects of a complex sample design on inference can be quite dramatic, as illustrated in previous sections for several types of analytic procedures (e.g., regression analysis, ANOVA). In most analyses, design effects are not negligible, even for means within subclasses, regression coefficients, or chi-square criteria computed from contingency table analyses. On the

other hand, the costs of computing can be large when the complexity of the sample design is used in the estimation process. Based on findings in this publication, the following recommendations about analytic strategy are suggested for users of NHANES I data from public use tape:

- A design-based inference, although difficult and costly to apply, is appropriate for such data. Estimation procedures, which account for the complexity of the survey design, should be used for the final analysis.
- Investigate all preliminary hypotheses without regard to the design effects. Since estimated means and other statistics may change greatly when sampling weights are considered, sampling weights and weighted estimates should be used.
- Based either on significant results at the previous step or on relationships thought to be important from previous substantive considerations regardless of whether they were significant at the previous step, proceed with a more rigorous analysis using both the appropriate weights and sample design effects.

Such a two stage design-based inference approach will be both less costly and more appropriate than other strategies that could be applied to the NHANES I data.

References

[1]National Center for Health Statistics, H. W. Miller: Plan and operation of the Health and Nutrition Examination Survey, United States, 1971–1973. *Vital and Health Statistics.* Series 1-No. 10a. DHEW Pub. No. (PHS) 79–1310. Public Health Service, Washington. U.S. Government Printing Office, Dec. 1978.

[2]Kish, L., Groves, R. M. and Krotki, K.: Sampling Errors for Fertility Surveys, WFS Occasional Paper No. 17. The Hague, International Statistical Institute. 1976.

[3]Verma, V., Scott, C. and O'Muircheartaigh, C.: Sample designs and sampling errors for the World Fertility Survey. *J R Stat Soc A.* 143(4):431-473, 1980.

[4]Survey Research Center Computer Support Group, OSIRIS IV User's Manual, Institute for Social Research, 1979.

[5]Hidiroglou, M. A., Fuller, W. A., and Hickman, R. D.: SUPERCARP. Survey Section of the Statistical Laboratory, Iowa State University, 6th ed., 1980.

[6]SAS Institute, Inc.: *SAS User's Guide, 1979 edition,* J. T. Helvig and K. A. Council, eds., Raleigh, N.C. SAS Institute, Inc.

[7]National Center for Health Statistics, B. A. Burt, S. A. Eklund, J. R. Landis, and others: Diet and dental health, a study of relationships, United States, 1971–74. *Vital and Health Statistics.* Series 11-No. 225. DHHS Pub. No. (PHS) 82–1675. Public Health Service. Washington. U.S. Government Printing Office, Jan. 1981.

[8]National Center for Health Statistics, W. R. Harlan, A. L. Hull, R. P. Schmonder, and others: Dietary intake and cardiovascular risk factors, United States, 1971–75, Part I, blood pressure. *Vital and Health Statistics.* Series 11-No. 226. DHHS Pub. No. (PHS) 82–1676. Public Health Service. Washington. U.S. Government Printing Office. In preparation.

[9]National Center for Health Statistics, W. R. Harlan, S. A. Eklund, J. R. Landis, and others: Dietary intake and cardiovascular risk factors, United States, Part II, serum urate, serum cholesterol, and correlates. *Vital and Health Statistics.* Series 11-No. 227. DHHS Pub. No. (PHS) 82–1677. Public Health Service. Washington. U.S. Government Printing Office. In preparation.

[10]National Center for Health Statistics, A. Engel, R. Murphy, K. Maurer, and E. Collins: Plan and operation of the HANES I Augmentation Survey of adults 25–74 years, United States, 1974–75. *Vital and Health Statistics.* Series 1-No. 14. DHEW Pub. No. (PHS) 78–1314. Public Health Service. Washington. U.S. Government Printing Office, June 1978.

[11]Goodman, R. and Kish, L.: Controlled selection—a technique in probability sampling. *J. Am. Stat. Assoc.* 45, 350–373. 1950.

[12]National Center for Health Statistics, E. E. Bryant, M. G. Kovar, and H. Miller: A study of the effect of renumeration upon response in the Health and Nutrition Examination Survey, United States. *Vital and Health Statistics.* Series 2-No. 67. DHEW Pub. No. (HRA) 76–1341, Health Resources Administration. Washington. U.S. Government Printing Office, Oct. 1975.

[13]National Center for Health Statistics: *The HANES Study, Final Report.* Prepared by the Institute for Survey Research. HSM–110–73–376. Health Services and Mental Health Administration. Philadelphia. Temple University.

[14]U.S. Department of Health, Education, and Welfare: *A Comparison and Analysis of Examined and Unexamined Persons on Medical History Characteristics for the First Round of the Health and Nutrition Examination Study HSM–110–73–371.* Prepared by Westat Inc. Rockville, MD. Jan. 24, 1974.

[15]National Center for Health Statistics: Factors related to response in a Health Examination Survey: United States, 1960–1962. *Vital and Health Statistics.* Series 2-No. 36. DHEW Pub. No. 36 (HSM) 73–1263. Health Services and Mental Health Administration. Washington. U.S. Government Printing Office, Aug. 1969.

[16]Wesley L. Schaible, Ph.D., Acting Chief, Methodological Research Branch, National Center for Health Statistics: Memorandum to Arthur J. McDowell, Director. Division of Health Examination Statistics. June 21, 1974.

[17]National Center for Health Statistics: Blood Pressure Levels of Persons 6–74 Years, United States, 1971–1974. *Vital and Health Statistics,* Series 11-No. 203. U.S. Department of Health, Education, and Welfare. DHEW Publication No. (PHS) 78–1648. 1977.

[18]National Center for Health Statistics, H. W. Miller: Plan and operation of the Health and Nutrition Examination Survey, United States, 1971–1973. *Vital and Health Statistics.* Series 1-No. 10b. DHEW Pub. No. (PHS) 73-1310. Public Health Service. Washington. U.S. Government Printing Office. 1978.

[19]Lepkowski, J. M.: Design effects for multivariate categorical interactions, doctoral thesis, University of Michigan, 1980.

[20]Vinter, S. T.: Survey sampling errors with OSIRIS IV. Paper presented at the COMPSTAT conference, Aug. 1980.

[21]Gurney, M. and Jewett, R. S.: Constructing Orthogonal Replications for Variance Estimation. *J Am Stat Assoc.* 70:819–821. 1975.

[22]Kmenta, J.: *Elements of Econometrics.* Macmillan Publishing Co., Inc. New York. 1971.

[23]Holt, D., Smith, T. M. F. and Winter P. D.: Regression Analysis of Data from Complex Surveys. *J R Stat Soc A,* 143(4):1980. 474–487.

[24]Snedecor, G. W. and Cochran, W. G.: *Statistical Methods.* The Iowa State University Press, Ames, Iowa. Sixth Ed., 1967.

[25]Beaton, G. H., Milner, J., Corey, P., and others: Sources of variance in 24-hour dietary recall data, implications for nutrition study design and interpretation. *Am J Clin Nutr* 32(6), Dec. 1979.

[26]Grizzle, J. E., Starmer, C. F. and Koch, G. G.: Analysis of categorical data by linear models. *Biometrics* 25:489–503, 1969.

[27]Koch, G. G. and Lemeshow, S.: An application of multivariate analysis to complex sample survey data. *J Am Stat Assoc* 67:780–782, 1972.

[28]Koch, G. G., Freeman, D. H., Jr. and Freeman, J. L.: Strategies in the multivariate analysis of data from complex surveys. *Int Stat Rev* 43:59–78, 1975.

[29]Freeman, D. H., Jr. et al.: Strategies in the multivariate analysis of data from complex surveys II, an application to the United States National Health Interview Survey. *Int Stat Rev.* 44:317–330, 1976.

[30]Freeman, D. H., Jr. and Brock, D. B.: The role of covariance matrix estimation in the analysis of complex sample survey data, in N. K. Namboudri, ed. *Survey Sampling and Measurement.* New York. Academic Press, 121–140, 1978.

[31]National Center for Health Statistics, S. Abraham, M. Carroll, C. Johnson, and C. Dresser: Caloric and selected nutrient values for persons 1–74 years of age, first Health and Nutrition Examination Survey, United States, 1971–74. *Vital and Health Statistics.* Series 11-No. 209. DHEW Pub. No. (PHS) 79–1657. Public Health Service. Washington. U.S. Government Printing Office, June 1979.

[32]Koch, G. G. and Stokes, M. E.: Annotated computer applications of weighted least squares methods for illustrative analyses of examples involving health survey data. Technical report prepared for the National Center for Health Statistics. 1980.

[33]Koch, G. G., Stokes, M. E. and Brock, D.: Applications of weighted least squares methods for fitting variational models to health survey data. *Proceedings of the American Statistical Association Section on Survey Research Methods,* 218–223. 1980.

[34]McCarthy, P.J.: Pseudoreplication, Half samples. *Rev Int Stat Inst* 37, 239–264, 1969.

[35]Kish, L. and Frankel, M. R.: Balanced repeated replications for standard errors. *J Am Stat Assoc,* 65:1071–1094.

[36]Lepkowski, J. M. and Landis, J. R.: Strategies in the analysis of the dental data from the HES and HANES. *Contributed papers to the Third Data Use Conference, November 14–16, 1978, Part II,* 99–115. 1979.

[37]Lepkowski, J. M., Design Effects for Multivariate Categorical Interactions. Unpublished doctoral dissertation, University of Michigan, 1980.

[38]Landis, J. R. et al.: A computer program for the generalized chi-square analysis of categorical data using weighted least squares (GENCAT). *Comput Programs Biomed,* 6:196–231, 1976.

[39]Shuster, J. J. and Downing, D. J.: Two-way contingency tables for complex sampling schemes. *Biometrika,* 63:271–276, 1976.

[40]Bhapkar, V. P. and Koch, G. G.: Hypotheses of 'no interaction' in multidimensional contingency tables. *Technometrics,* 10:107–123, 1968.

[41]Graybill, F. A.: Introduction to Matrices with Applications in Statistics. Belmont, California: Wadsworth Publishing Company, Inc., 1969.

[42]Kish, L. and Hess, I. On variances of ratios and their differences in multistage samples. *J Am Stat Assoc,* 54:416–446. 1959.

[43]Forthofer, R. N. and Koch, G. G. An analysis for compounded functions of categorical data. *Biometrics,* 29:143–157. 1973.

[44]Koch, G. G., Landis, J. R., Freeman, J. L., Freeman, D. H., Jr. and Lehnen, R. G. A general methodology for the analysis of experiments with repeated measurement of categorical data. *Biometrics,* 33:133–158. 1977.

[45]Freeman, D. H., Jr., The regression analysis of data from complex sample surveys: an empirical investigation of covariance matrix estimation. Unpublished doctoral dissertation, University of North Carolina, 1975.

List of detailed tables

Table 1. NHANES I population estimates for examination locations 1–65, by sex, race, and age at examination: United States, 1971–74

Age at examination	Estimated population						
	Total	Male			Female		
		All races	White	Black	All races	White	Black
Total	193,976,381	94,239,866	82,740,899	10,413,986	99,736,515	86,867,546	11,999,935
1 year	3,313,458	1,693,074	1,401,508	280,212	1,620,384	1,327,657	257,289
2–3 years	6,963,162	3,553,765	2,997,107	479,362	3,409,397	2,872,581	505,442
4–5 years	6,672,346	3,378,503	2,866,374	485,872	3,293,843	2,755,016	511,134
6–7 years	7,193,663	3,652,322	3,060,888	573,867	3,541,341	2,951,927	576,578
8–9 years	7,696,597	3,880,396	3,279,649	586,419	3,816,201	3,257,936	539,855
10–11 years	8,465,793	4,381,730	3,732,593	563,823	4,084,063	3,424,070	617,793
12–14 years	12,335,321	6,312,519	5,397,061	879,377	6,022,802	5,122,189	836,252
15–17 years	12,318,434	6,207,169	5,311,596	812,321	6,111,265	5,233,091	853,294
18–19 years	7,352,200	3,673,321	3,206,467	404,045	3,678,879	3,158,930	504,417
20–24 years	17,325,038	8,109,775	7,094,036	866,201	9,215,263	7,972,486	1,073,358
25–34 years	26,936,001	13,002,514	11,594,115	1,231,793	13,933,487	12,160,578	1,646,337
35–44 years	22,268,477	10,675,731	9,515,530	1,004,953	11,592,746	10,111,458	1,318,050
45–54 years	23,313,316	11,150,110	10,039,124	1,056,837	12,163,206	10,879,167	1,237,459
55–64 years	19,049,001	9,072,586	8,274,948	702,647	9,976,415	9,037,157	871,098
65–74 years	12,773,574	5,496,351	4,969,903	486,257	7,277,223	6,603,303	651,579

Table 2. Sampling rates by age-sex groups for general sample of the NHANES I: United States, 1971–74

Age and sex	Sampling rate
1–5 years	1/2
6–19 years	1/4
20–44 years (men)	1/4
20–44 years (women)	1/2
45–64 years	1/4
65–74 years	1

Table 3. Subsampling rates by age-sex groups for detailed sample of the NHANES I: United States, 1971–74

Age and sex	Subsampling rate
25–44 years (men)	2/5
25–44 years (women)	1/5
45–64 years	3/5
65–74 years	1/4

Table 4. Percent distribution of adjustment factors for the NHANES I: United States, 1971–74 National Health and Nutrition Examination Survey, survey locations 1–65, United States, 1971–74

Size of adjustment factor	Number of cells	Percent distribution
Total	325	100.0
1.00–1.24	106	32.6
1.25–1.49	125	38.4
1.50–1.74	59	18.2
1.75–1.99	24	7.4
2.00–2.49	9	2.8
2.50–2.99	1	0.3
3.00–3.03	1	0.3

Table 5. Total number of examinees and those without dental examination records, by sex and age: National Health and Nutrition Examination Survey, 1971–74

Age	Both sexes	Male	Female	Both sexes	Male	Female
	Total number examined			Number without dental examination records		
All ages, 1–74 years	20,749	8,820	11,929	531	207	324
1–5 years	2,895	1,469	1,425	78	36	42
6–11 years	2,057	1,026	1,031	63	30	33
12–17 years	2,126	1,064	1,062	48	18	30
18–24 years	2,296	773	1,523	60	14	46
25–34 years	2,700	804	1,896	80	26	54
35–44 years	2,328	664	1,664	55	14	41
45–54 years	1,601	765	836	43	22	21
55–64 years	1,267	598	669	33	10	23
65–74 years	3,479	1,657	1,822	71	37	34

Table 6. Number of primary sampling units (PSU's) and number of examined persons for the general, detailed, and augmentation survey by stratum number for the NHANES I design: United States, 1971–75

Stratum number	Number of PSU's in sample survey design		Number of examined persons		
	General and detailed	Augmentation	General and detailed	Detailed only	Augmentation
Total	1,263	236	20,749	3,854	3,059
1–10	1,213	211	4,511	853	701
1	169	21	621	112	55
2	106	17	367	80	63
3	125	18	482	87	59
4	156	21	737	129	60
5	197	24	741	143	97
6	83	22	250	48	82
7	108	23	395	71	72
8	61	21	188	42	80
9	89	21	304	57	64
10	119	23	429	84	69
11–35	50	25	16,235	3,001	2,358

NOTE: In the certainty strata 1–10, PSU's are enumeration districts. In the noncertainty strata 11–35, PSU's are counties or groups of contiguous counties.

Table 7. Number of examined persons by race, sex, and stratum number in the NHANES I design: United States, 1971–74

Stratum number	Total	Number of examined persons ages 1–74 years by race and sex					
		White males	Black males	Other males	White females	Black females	Other females
Total	20,749	7,004	1,707	109	9,347	2,456	126
1	621	169	88	2	220	138	4
2	367	146	24	0	157	38	4
3	482	123	85	1	171	102	0
4	737	198	102	11	255	162	9
5	741	232	65	13	328	88	15
6	250	67	35	2	85	57	4
7	395	85	90	0	93	127	0
8	188	67	16	0	79	26	0
9	304	109	13	1	149	32	0
10	429	138	32	13	190	37	19
11	481	205	4	0	267	3	2
12	517	198	14	0	286	17	2
13	531	232	2	2	290	4	1
14	701	273	15	2	396	14	1
15	486	185	20	4	226	43	8
16	563	178	68	5	211	98	3
17	594	235	6	0	346	6	1
18	505	176	39	2	224	62	2
19	585	237	12	4	317	14	1
20	446	171	13	1	246	14	1
21	790	344	0	0	446	0	0
22	551	114	107	3	141	185	1
23	619	167	85	0	249	116	2
24	499	131	73	0	170	122	3
25	728	225	73	0	311	119	0
26	887	232	156	0	305	194	0
27	684	262	23	1	379	17	2
28	1,001	259	174	0	327	241	0
29	634	222	51	1	292	68	0
30	868	284	84	1	371	124	4
31	651	221	34	5	334	52	5
32	691	250	22	8	367	32	12
33	619	222	3	21	345	10	18
34	545	236	5	5	295	1	3
35	1,059	411	74	1	479	93	1

Table 8. Number of examined persons by race, sex, and stratum number in the NHANES I design for the detailed sample: United States, 1971–74

Stratum number	Total	Number of examined persons ages 25–74 years by race and sex					
		White males	Black males	Other males	White females	Black females	Other females
Total	3,854	1,541	277	21	1,667	335	13
1	112	37	13	1	34	27	0
2	80	38	4	0	27	11	0
3	87	23	18	0	29	17	0
4	129	46	15	1	43	23	1
5	143	60	11	4	55	12	1
6	48	17	7	1	12	11	0
7	71	16	18	0	17	20	0
8	42	19	0	0	18	5	0
9	57	25	1	0	27	4	0
10	84	34	8	4	30	5	3
11	100	45	0	0	53	1	1
12	93	40	3	0	49	0	1
13	92	45	1	0	46	0	0
14	129	54	1	0	70	4	0
15	78	43	2	1	27	5	0
16	101	29	13	0	41	18	0
17	107	52	1	0	54	0	0
18	81	41	4	1	28	7	0
19	109	45	2	1	59	2	0
20	81	34	2	0	44	1	0
21	162	72	0	0	90	0	0
22	89	28	17	1	23	20	0
23	112	33	16	0	48	15	0
24	81	28	8	0	30	15	0
25	156	67	8	0	67	14	0
26	150	45	22	0	65	18	0
27	141	65	6	0	68	1	1
28	182	57	26	0	64	35	0
29	126	50	10	0	58	8	0
30	152	63	14	0	64	11	0
31	113	49	3	1	51	8	1
32	123	51	2	2	61	6	1
33	119	45	0	2	69	0	3
34	100	46	2	0	52	0	0
35	224	99	19	1	94	11	0

Table 9. Number of survey locations, type of examination, years of data collection, age of target population, number of examined persons, and location of appropriate weights on public use tapes for NHANES I data

Survey locations and examination in sample design	Year	Age in years of target population	Number of examined persons	Tape locations of weights
1–65 detail	1971–74	25–74	3,854	170–175
1–65 nutrition	1971–74	1–74	20,749	176–181
66–100[1] detail	1974–75	25–74	3,059	182–187
1–100[2] detail	1971–75	25–74	6,913	188–193

[1]Augmentation sample

[2]Includes augmentation sample

Table 10. Comparative analyses of standard errors and design effects for multiple and paired sampling error computational units (SECU's) within certainty strata for the number of decayed, missing, and filled (DMF) teeth, systolic blood pressure (SBP), and calories consumed daily by age for NHANES I data: United States, 1971–74

Age	Number of examined persons	Mean	Multiple SECU's		Paired SECU's	
			Standard error of mean	Square root of design effect	Standard error of mean	Square root of design effect
			DMF teeth			
1–74 years	20,749	14.723	0.166	2.094	0.161	2.034
1–17 years	7,104	3.965	0.071	1.545	0.070	1.538
18–24 years	2,297	11.924	0.237	1.766	0.237	1.768
25–34 years	2,694	16.918	0.261	1.823	0.262	1.826
35–44 years	2,327	21.436	0.249	1.560	0.248	1.555
45–54 years	1,599	22.826	0.216	1.085	0.232	1.164
55–64 years	1,262	25.744	0.291	1.278	0.279	1.224
65–74 years	3,466	27.727	0.154	1.283	0.154	1.278
			SBP			
6–74 years	17,658	123.95	0.424	2.292	0.409	2.211
6–17 years	4,085	108.24	0.492	2.207	0.498	2.234
18–24 years	2,290	118.89	0.466	1.573	0.441	1.489
25–34 years	2,675	120.93	0.445	1.534	0.440	1.515
35–44 years	2,317	125.64	0.580	1.479	0.603	1.536
45–54 years	1,589	134.14	1.015	1.746	1.037	1.783
55–64 years	1,255	142.11	0.826	1.214	0.804	1.181
65–74 years	3,447	150.01	0.793	1.820	0.784	1.799
			Calories			
1–74 years	20,749	2000.0	17.80	2.923	17.88	2.937
1–17 years	7,104	2011.0	20.75	2.106	20.03	2.033
18–24 years	2,297	2294.8	37.02	1.660	35.32	1.584
25–34 years	2,694	2177.5	27.66	1.479	29.44	1.573
35–44 years	2,327	2042.9	28.33	1.545	28.94	1.578
45–54 years	1,599	1897.3	31.76	1.515	30.41	1.451
55–64 years	1,262	1723.2	33.06	1.418	33.45	1.435
65–74 years	3,466	1518.9	20.68	1.870	19.99	1.808

Table 11. Number of examined persons, estimated means, standard deviations, standard errors of the mean, and design effects for the number of decayed, missing, and filled (DMF) teeth, systolic blood pressure (SBP), calories consumed daily, and age under analysis options 1–3 for NHANES I data: United States, 1971–74

Option number	Inclusion of sampling		Sample size	Mean	Standard error of mean	Square root of design effect
	Weights	Design				
				DMF teeth		
1	No	No	20,749	14.93	0.079	[1]...
2	Yes	No	20,749	14.72	0.075	[1]...
3	Yes	Yes	20,749	14.72	0.161	2.156
				SBP		
1	No	No	17,658	126.91	0.185	[1]...
2	Yes	No	17,658	123.95	0.168	[1]...
3	Yes	Yes	17,658	123.95	0.409	2.442
				Calories		
1	No	No	20,749	1827.5	6.088	[1]...
2	Yes	No	20,749	2000.0	6.560	[1]...
3	Yes	Yes	20,749	2000.0	17.883	2.726
				Age		
1	No	No	20,749	32.23	0.159	[1]...
2	Yes	No	20,749	30.61	0.140	[1]...
3	Yes	Yes	20,749	30.61	0.239	1.707

[1]Category not applicable.

Table 12. Number of examined persons, estimated means, standard deviations, standard errors of the mean, and design effects for the number of decayed, missing, and filled (DMF) teeth, systolic blood pressure (SBP), and calories consumed daily within age groups under analysis options 1–3 for NHANES I data: United States, 1971–74

Age	Number of examined persons	Option 1			Option 2			Option 3		
		Mean	Standard deviation	Standard error of mean	Mean	Standard deviation	Standard error of mean	Mean	Standard error of mean	Square root of design effect
					DMF teeth					
1–74 years	20,749	14.935	11.4180	0.0793	14.723	10.7760	0.0748	14.723	0.1613	2.156
1–17 years	7,104	3.338	3.8493	0.0457	3.965	4.0810	0.0484	3.965	0.0703	1.452
18–24 years	2,297	12.050	6.4173	0.1339	11.924	6.2566	0.1305	11.924	0.2367	1.813
25–34 years	2,694	16.872	7.4408	0.1434	16.918	7.2497	0.1397	16.918	0.2618	1.874
35–44 years	2,327	21.271	7.6962	0.1595	21.436	7.3482	0.1523	21.436	0.2481	1.629
45–54 years	1,599	22.515	7.9700	0.1993	22.826	7.5709	0.1893	22.826	0.2320	1.226
55–64 years	1,262	25.234	8.0990	0.2280	25.744	7.6022	0.2140	25.744	0.2790	1.304
65–74 years	3,466	27.608	7.0741	0.1202	27.727	6.7742	0.1151	27.727	0.1536	1.334
					SBP					
6–74 years	17,658	126.91	24.585	0.1850	123.95	22.262	0.1675	123.95	0.4090	2.442
6–17 years	4,085	108.67	14.245	0.2229	108.24	14.132	0.2211	108.24	0.4980	2.252
18–24 years	2,290	117.96	14.166	0.2960	118.89	13.794	0.2883	118.89	0.4407	1.529
25–34 years	2,675	119.90	15.006	0.2901	120.93	14.710	0.2844	120.93	0.4397	1.546
35–44 years	2,317	125.76	18.885	0.3923	125.64	17.665	0.3670	125.64	0.6026	1.642
45–54 years	1,589	135.10	23.176	0.5814	134.14	22.782	0.5715	134.14	1.0365	1.814
55–64 years	1,255	143.13	24.126	0.6810	142.11	23.453	0.6620	142.11	0.8040	1.215
65–74 years	3,447	151.02	25.580	0.4357	150.01	25.056	0.4268	150.01	0.7840	1.836
					Calories					
1–74 years	20,749	1827.5	877.00	6.088	2000.0	944.91	6.560	2000.0	17.883	2.726
1–17 years	7,104	1880.4	830.42	9.853	2011.0	874.24	10.372	2011.0	20.033	1.931
18–24 years	2,297	2084.6	1068.70	22.298	2294.8	1136.60	23.715	2294.8	35.317	1.489
25–34 years	2,694	1954.5	971.00	18.708	2177.5	1050.10	20.232	2177.5	29.435	1.455
35–44 years	2,327	1829.0	884.65	18.339	2042.9	966.51	20.036	2042.9	28.935	1.444
45–54 years	1,599	1840.4	838.33	20.965	1897.3	816.17	20.411	1897.3	30.410	1.490
55–64 years	1,262	1679.2	828.08	23.310	1723.2	814.02	22.914	1723.2	33.454	1.460
65–74 years	3,466	1497.2	651.06	11.059	1518.9	649.50	11.032	1518.9	19.991	1.812

Table 13. Number of examined persons in subclasses determined by lowest 15 percentile and highest percentile of skinfold thickness, means, standard errors, test statistics, and design effects for serum cholesterol: United States, 1971–74

Option	Low skinfold			High skinfold			t statistic	Square root of design effect
	Number of examinees	Mean	Standard error	Number of examinees	Mean	Standard error		
				All males				
1	1,030	198.7	1.41	1,015	223.1	1.55	11.6	[1]...
2	1,030	191.6	1.34	1,015	221.8	1.60	14.3	[1]...
3	1,030	191.6	1.59	1,015	221.8	2.39	9.5	1.5
				Black males				
1	282	200.1	2.73	155	222.2	4.03	4.7	[1]...
2	282	193.4	2.81	155	226.6	4.89	6.4	[1]...
3	282	193.4	4.00	155	226.6	8.80	3.4	2.1
				White males				
1	748	198.2	1.65	860	223.3	1.68	10.6	[1]...
2	748	191.2	1.54	860	221.2	1.70	12.8	[1]...
3	748	191.2	1.90	860	221.2	2.26	9.7	1.3
				All females				
1	1,652	197.4	1.19	1,637	224.7	1.24	15.9	[1]...
2	1,652	196.1	1.19	1,637	225.9	1.25	17.3	[1]...
3	1,652	196.1	2.00	1,637	225.9	1.83	13.4	1.3
				Black females				
1	288	191.8	2.40	488	221.0	2.25	8.5	[1]...
2	288	193.0	2.39	488	224.2	2.11	9.6	[1]...
3	288	193.0	3.00	488	224.2	3.33	6.6	1.4
				White females				
1	1,364	198.5	1.34	1,149	226.2	1.49	13.8	[1]...
2	1,364	196.5	2.93	1,149	226.3	1.52	14.8	[1]...
3	1,364	196.5	2.21	1,149	226.3	2.21	11.6	1.3

[1]Category not applicable.

Table 14. Summary of simple regression models of the number of decayed, missing, and filled (DMF) teeth, systolic blood pressure (SBP), and calories consumed daily on age under analysis options 1–3 by race and sex for NHANES I data: United States, 1971–74

Race and sex	Number of examined persons	Unweighted design (option 1)				Weighted design						
								(Option 2)		(Option 3)		Square root of design effect
		R^2	Slope	Standard error	t-statistic	R^2	Slope	Standard error	t-statistic	Standard error	t-statistic	
DMF on age												
Total	20,749	0.67	0.408	0.0020	206.89	0.65	0.432	0.0022	196.52	0.0032	135.09	1.45
White males	7,004	0.73	0.416	0.0030	138.91	0.67	0.440	0.0037	118.93	0.0042	105.49	1.13
Black males	1,707	0.63	0.335	0.0062	54.44	0.47	0.308	0.0080	38.52	[1]	[1]	[1]
Other males	109	0.53	0.317	0.0287	11.04	0.45	0.294	0.0316	9.28	[1]	[1]	[1]
White females	9,347	0.67	0.414	0.0030	136.49	0.68	0.439	0.0031	139.50	0.0053	82.76	1.69
Black females	2,456	0.59	0.391	0.0065	59.91	0.54	0.385	0.0072	53.29	[1]	[1]	[1]
Other females	126	0.40	0.337	0.0372	9.07	0.25	0.244	0.0376	6.50	[1]	[1]	[1]
SBP on age												
Total	17,658	0.40	0.730	0.0068	107.45	0.35	0.696	0.0071	98.11	0.0131	53.14	1.85
White males	5,854	0.36	0.605	0.0106	57.24	0.33	0.610	0.0115	53.14	0.0113	54.06	0.98
Black males	1,326	0.46	0.815	0.0240	33.91	0.43	0.848	0.0269	31.53	[1]	[1]	[1]
Other males	89	0.35	0.762	0.1118	6.81	0.14	0.401	0.1064	3.77	[1]	[1]	[1]
White females	8,243	0.41	0.767	0.0102	75.57	0.38	0.734	0.0104	70.39	0.0188	39.03	1.80
Black females	2,037	0.47	0.979	0.0230	42.55	0.44	1.008	0.0252	40.05	[1]	[1]	[1]
Other females	109	0.40	0.920	0.1086	8.47	0.37	0.818	0.1040	7.87	[1]	[1]	[1]
Calories on age												
Total	20,749	0.02	−4.90	0.2629	−18.64	0.01	−5.50	0.3238	−16.99	.3171	−17.35	0.98
White males	7,004	0.01	−3.39	0.4873	−6.95	0.00	−3.52	0.6102	−5.78	.6314	−5.58	1.04
Black males	1,707	0.01	−3.74	0.9217	−4.05	0.00	−1.08	1.212	−0.89	[1]	[1]	[1]
Other males	109	0.00	1.00	3.598	0.28	0.05	12.50	5.101	2.45	[1]	[1]	[1]
White females	9,347	0.04	−5.89	0.3034	−19.41	0.04	−6.44	0.3315	−19.43	.4339	−14.85	1.31
Black females	2,456	0.06	−8.39	0.6578	−12.75	0.06	−9.45	0.7420	−12.74	[1]	[1]	[1]
Other females	126	0.00	−1.23	3.474	−0.35	0.01	−3.35	3.899	−0.86	[1]	[1]	[1]

[1] Category not applicable.

Table 15. Summary of multiple regression models for the number of decayed, missing, and filled (DMF) teeth on age, race, sex, and sweets for 6,349 examined persons ages 11–30 under analysis options 1–3: United States, 1971–74

Variable	Regression coefficient	Standard error of coefficient	t-statistic	Square root of design effect
	Unweighted SRS design (option 1)			
Age	0.685	0.0130	52.42	[1]
Race	0.875	0.0899	9.73	[1]
Sex	−0.491	0.0752	−6.52	[1]
Sweets	0.057	0.0070	8.21	[1]
	Weighted SRS design (option 2)			
Age	0.705	0.0125	56.29	[1]
Race	0.795	0.1072	7.42	[1]
Sex	−0.465	0.0698	−6.65	[1]
Sweets	0.049	0.0068	7.17	[1]
	Weighted complex sampling design (option 3)			
Age	0.705	0.0209	33.67	1.67
Race	0.795	0.2277	3.50	2.12
Sex	−0.465	0.0928	−5.01	1.33
Sweets	0.049	0.0077	6.43	1.12

[1] Category not applicable.

Table 16. Summary of multiple regression models for systolic blood pressure (SBP) on age, race, sex, and Quetelet's Index for 13,573 examined persons ages 18–74 under analysis options 1–3: United States, 1971–74

Variable	Regression coefficient	Standard error of coefficient	t-statistic	Square root of design effect
		Unweighted SRS design (option 1)		
Age	0.667	0.0096	69.44	[1]...
Race	3.896	0.3938	9.89	[1]...
Sex	−1.885	0.3495	−5.39	[1]...
Quetelet's Index	1.135	0.0335	33.88	[1]...
		Weighted SRS design (option 2)		
Age	0.584	0.0102	57.49	[1]...
Race	2.908	0.4422	6.58	[1]...
Sex	−2.871	0.3162	−9.08	[1]...
Quetelet's Index	1.177	0.0331	35.56	[1]...
		Weighted complex sampling design (option 3)		
Age	0.584	0.0177	32.92	1.75
Race	2.908	0.8266	3.52	1.87
Sex	−2.871	0.5206	−5.52	1.64
Quetelet's Index	1.177	0.0630	18.69	1.90

[1]Category not applicable.

Table 17. Summary of mean periodontal index (PI) score and estimated standard errors and design effects by drinking and smoking classification for NHANES I detailed sample: United States, 1971–74

Subclass		Unweighted design SRS (option 1)			Weighted design				
					Weighted number examined	SRS (option 2)		Complex (option 3)	
Drinking	Smoking	Number examined	Mean PI score	Standard error		Mean PI score	Standard error	Standard error	Square root of design effect
None	Never	417	1.618	0.1074	354.73	1.424	0.1018	0.1371	1.35
None	Past	101	2.038	0.2366	74.74	1.836	0.2171	0.3044	1.40
None	Now	195	2.349	0.1733	162.63	1.904	0.1578	0.1339	0.85
Little	Never	479	0.961	0.0733	547.18	0.800	0.0663	0.0583	0.88
Little	Past	214	1.280	0.1282	251.74	0.966	0.1137	0.1325	1.17
Little	Now	483	1.738	0.0968	571.00	1.516	0.0896	0.1267	1.41
Moderate	Never	178	1.003	0.1303	196.80	0.853	0.1271	0.1887	1.48
Moderate	Past	166	1.148	0.1341	198.95	0.930	0.1195	0.1120	0.94
Moderate	Now	483	1.731	0.0984	540.25	1.463	0.0902	0.1256	1.39
Heavy	Never	30	1.774	0.3815	29.42	1.420	0.3592	0.4219	1.17
Heavy	Past	32	1.769	0.3391	36.06	1.754	0.3290	0.3646	1.11
Heavy	Now	165	2.029	0.1801	198.79	1.690	0.1676	0.2211	1.32

Table 18. Hypothesis tests for variation in mean periodontal index (PI) score by cross-classification of drinking and smoking variables for NHANES I detailed sample: United States, 1971–74

Source of variation	Degree of freedom	Chi-square test criteria and significance levels					
		Unweighted SRS design		Weighted SRS design		Weighted complex design	
		Q	P-value	Q	P-value	Q	P-value
Drinking (D)	3	29.54	0.00	28.83	0.00	26.91	0.00
Smoking (S)	2	13.58	0.00	7.89	0.02	9.14	0.01
$D \times S$	6	2.52	0.87	5.54	0.48	3.78	0.71

Table 19. Hypothesis tests for variation in mean periodontal index (PI) score by cross-classification of drinking and smoking variables (model with no interaction) for NHANES I detailed sample: United States, 1971–74

Source of variation	Degree of freedom	Chi-square test criteria and significance levels					
		Unweighted SRS design		Weighted SRS design		Weighted complex design	
		Q	p-value	Q	p-value	Q	p-value
Drinking (D)	3	51.58	0.00	48.89	0.00	40.92	0.00
Smoking (S)	2	76.85	0.00	65.45	0.00	42.70	0.00
Lack of Fit	6	2.52	0.87	5.54	0.48	3.78	0.71

Table 20. Distribution of sample elements according to the r levels of the response profile by the s subclasses

Subclass	Response profile				Total
	1	2	...	r	
1	n_{11}	n_{12}	...	n_{1r}	$n_{1.}$
2	n_{21}	n_{22}	...	n_{2r}	$n_{2.}$
.	.	.	...	.	.
.	.	.	...	.	.
.	.	.	...	.	.
s	n_{s1}	n_{s2}	...	n_{sr}	$n_{s.}$

Table 21 Number of examined persons ages 25–74 with periodontal index (PI) scores of zero (none) and greater than zero (some) by race and current smoking status for NHANES I detailed sample: United States, 1971–74

Race	Current cigarette smoker	Number examined	PI score		Proportion PI (some)
			None	Some	
All subjects		2,919	1,294	1,625	0.557
White	Yes	851	351	500	0.588
White	No	1,574	821	753	0.478
Black	Yes	230	58	172	0.748
Black	No	264	64	200	0.758

Table 22 Weighted number of examined persons ages 25–74 with periodontal index (PI) scores of zero (none) and greater than zero (some) by race and current smoking status for NHANES I detailed sample: United States, 1971–74

Race	Current cigarette smoker	Weighted number examined	PI score		Proportion PI (some)
			None	Some	
All subjects		3,137.5	1,511.3	1,626.2	0.518
White	Yes	1,076.3	459.0	617.3	0.574
White	No	1,727.2	952.5	774.7	0.449
Black	Yes	171.2	51.8	119.4	0.697
Black	No	162.8	48.0	114.8	0.705

Table 23. Distribution of proportion of some periodontal index (PI > 0.0) and estimated standard errors and design effects by race and current cigarette smoking classification for NHANES I detailed sample: United States, 1971–74

Subclass		Unweighted design SRS (option 1)			Weighted design				
					Weighted number examined	SRS (option 2)		Complex (option 3)	
Race	Current cigarette smoking	Number examined	Proportion PI (some)	Standard error		Proportion PI (some)	Standard error	Standard error	Square root of design effect
White	Yes	851	0.588	0.0169	1,076.3	0.574	0.0151	0.0250	1.66
White	No	1,574	0.478	0.0126	1,727.2	0.449	0.0120	0.0225	1.88
Black	Yes	230	0.748	0.0286	171.2	0.697	0.0351	0.0427	1.22
Black	No	264	0.758	0.0264	162.8	0.705	0.0357	0.0547	1.53

Table 24. Vector of subclass proportions of some periodontal index (PI > 0.0) and estimated covariance matrix by race and current cigarette smoking classification for NHANES I detailed sample: United States, 1971–74

Subclass		Proportion PI (some)	Estimated covariance matrix × 10^3			
Race	Current cigarette smoker					
White	Yes	0.574	0.626	0.364	0.059	0.000
White	No	0.449		0.506	0.070	0.020
Black	Yes	0.697			1.825	−0.411
Black	No	0.705				2.995

Table 25. Hypotheses, hypothesis matrices, and test statistics for the model X_1 relating the variation in the proportion of some periodontal index (PI > 0.0) to race and current cigarette smoking classification using sample weights and design effects for NHANES I detailed sample: United States, 1971–74

Hypothesis	Hypothesis matrix	Chi-square statistic	Degree of freedom	P-value
H_1: There is no variation due to the effect of race	[0 1 0 0]	26.02	1	<0.01
H_2: There is no variation due to the effect of smoking	[0 0 1 0]	2.27	1	0.13
H_3: There is no variation due to the interaction between race and smoking	[0 0 0 1]	2.92	1	0.09

Table 26. Hypothesis tests for variation in the proportion of some periodontal index (PI > 0.0) by cross-classification of race and smoking cross-classification for NHANES I detailed sample: United States, 1971–74

Source of variation	Degree of freedom	Chi-square test criteria and significance levels						Square root of design effect
		Unweighted SRS design		Weighted SRS design		Weighted complex design		
		Q	P-value	Q	P-value	Q	P-value	
Race (R)	1	98.59	0.00	50.20	0.00	26.02	<0.01	1.39
Smoking (S)	1	5.04	0.02	4.78	0.03	2.27	0.13	1.45
$R \times S$	1	7.22	0.01	6.11	0.01	2.92	0.09	1.45

Table 27. Hypothesis tests for variation in the proportion of some periodontal index (PI > 0.0) by cross-classification of race and smoking (reduced model) for NHANES I detailed sample: United States, 1971–74

Source of variation	*Degree of freedom*	*Chi-square test criteria and significance levels*						*Square root of design effect*
		Unweighted SRS design		*Weighted SRS design*		*Weighted complex design*		
		Q	P-*value*	Q	P-*value*	Q	P-*value*	
White versus black mean.....	1	99.47	0.00	50.17	0.00	27.20	<0.01	1.36
Smoking with whites.....	1	26.87	0.00	42.19	0.00	38.63	<0.01	1.05
Lack of fit.....	1	0.06	0.80	0.02	0.88	0.01	0.92	1.41

Appendixes

Contents

Appendix I. Definitions of terms and variables

Dietary variables

Calories—The total energy intake determined from the 24-hour dietary recall measured in kilocalories.

Sweets—The sum of the reported frequencies for the ingestion of food from the three categories of desserts and sweets, candy, and beverages (sweetened, carbonated, and non-carbonated).

Dental and medical variables

Periodontal index—Periodontal index score for entire mouth as given in the data provided.

DMF—Sum of decayed (D), missing (M), and filled (F) permanent teeth.

Quetelet's index—Body mass index which standardizes weight for height and permits indirect prediction of adiposity. Defined as weight÷height^2 using weight in kilograms and height in centimeters.

Behavioral variables

Drinking—Categorical variable concerning alcohol consumption derived from three other variables. The four categories are

1) Those who claimed not to have had a drink in the past (called "none" in the tables),
2) Those who claimed to drink no more than once a week *and* when they did drink had three or fewer drinks (called "little" in the tables),
3) Those who stated they drank more often than once a week but have three or fewer drinks at a time, or those who drink no more often than once a week but have four or more drinks when they do drink (called "moderate" in the tables), and
4) Those who claimed to drink more often than once a week *and* have four or more drinks at a time (called "heavy" in tables).

Smoking—Categorical variable derived from several other variables. The three categories are

1) Never have used tobacco in quantities up to or equal to the amounts stated in the medical history questionnaire, that is, at least 100 cigarettes, 50 cigars, or three packages of pipe tobacco during the subjects' lifetime.
2) Have used tobacco at least up to the amounts stated for at least one of the categories stated in the questionnaire, but do not use tobacco now, and
3) Used tobacco at the time of the interview, in amounts at least as large as those stated in the questionnaire.

Current cigarette smoker—Categorical variable for current cigarette smoking status. The categories are

1) Have smoked more than 100 cigarettes and smoke cigarettes now, and
2) Have never smoked more than 100 cigarettes or do not smoke cigarettes now.

Appendix II. Computing control card files

The following subsections contain representative control card files that were used for the run illustrated in the corresponding previous section. All the program statements are intended for the OSIRIS IV system available at the University of Michigan, except for the ANOVA and contingency table analyses which were processed sequentially through OSIRIS IV and then through the GENCAT weighted least squares program.

Means and variances

Example 1 shows the OSIRIS IV commands were used to generate the multiple SECU results for DMF teeth and calories displayed in table 10. &USTATS computes means, standard deviations, and standard errors using the weighted data, whereas &PSALMS computes estimates and sampling errors for ratio means from stratified clustered sample designs.

Regression models

Example 2 shows the OSIRIS IV commands were used to generate the simple regression model results for DMF teeth on age and calories on age shown in table 14. &REGRESSN computes standard regressions using the weighted data ignoring the sample design, whereas &REPERR computes regression statistics and their sampling errors for data from clustered sample designs using balanced half sample replications.

ANOVA

Example 3 shows the five command files were used to generate the results under options 1–3 displayed in tables 17–19. Step 1 uses the OSIRIS IV &USTATS command to generate means and their standard errors (unweighted and weighted) for the 12 drinking and smoking classifications. The vector of means and the corresponding covariance matrix then were read under the direct input option of GENCAT to generate the analyses for options 1 and 2, according to whether the sampling weights were included in the analysis in Steps 2 and 3, respectively. Finally, Steps 4 and 5 were used to generate the ratio means and the covariance matrix under the cluster sample design and to produce the chi-square statistics under option 3.

Contingency tables

Example 4 shows the five command files were used to generate the results in tables 21–27. Step 1 uses the OSIRIS IV &TABLES command to obtain the 4×2 table with race-current cigarette smoking categories forming the rows and periodontal index forming the columns (none, some). These unweighted and weighted tables then were used as input to GENCAT in Steps 2 and 3. Step 4 utilizes the &PSALMS routine to obtain the variance-covariance matrix for the weighted proportions taking into account the complex sample design. Finally, step 5 contains the control cards needed to run GENCAT on the weighted data.

Example 1

```
$RUN ISR:OSIRIS.IV SPRINT=*PRINT*
&RECODE
 R1=1
 NAME R1'COUNTER'
&END
&USTATS DICTIN=-DICT DATAIN=-DATA
WEIGHTED STATS FOR DMF AND CALORIES BY AGE GROUP
VARS=V6042,V203 WTVAR=V90-
REP=(V49=1-17/18-24/25-34/35-44/45-54/55-64/65-74/1-74)
&END
&PSALMS DICTIN=-DICT DATAIN=-DATA
SAMPLING ERROR ANALYSIS OF DMF AND CALORIES BY AGE GROUP
R=1 WTVAR=V90 SECU=V96 ST=V91-
REP=(V49=1-17/18-24/25-34/35-44/45-54/55-64/65-74/1-74)
MOD=MULT ST=1-10 SECU=169,106,125,156,197,83,108,61,89,119
MOD=PAIR ST=11-35
PAR=V6042/R1 SUB=1,1
PAR=V6042/R1 SUB=2,2
PAR=V6042/R1 SUB=3,3
PAR=V6042/R1 SUB=4,4
PAR=V6042/R1 SUB=5,5
PAR=V6042/R1 SUB=6,6
PAR=V6042/R1 SUB=7,7
PAR=V6042/R1 SUB=8,8
PAR=V203/R1 SUB=1,1
PAR=V203/R1 SUB=2,2
PAR=V203/R1 SUB=3,3
PAR=V203/R1 SUB=4,4
PAR=V203/R1 SUB=5,5
PAR=V203/R1 SUB=6,6
PAR=V203/R1 SUB=7,7
PAR=V203/R1 SUB=8,8
&END
```

Example 2

```
$RUN ISR:OSIRIS.IV SPRINT=*PRINT*
&REGRESSN DICTIN=-DICT DATAIN=-DATA
REGRESSION OF DMF AND CALORIES ON AGE
WTVAR=V90
V=V49 DEPV=V6042
V=V49 DEPV=V203
&END
&REGRESSN
INCLUDE V50=1 AND V51=1
REGRESSION OF DMF AND CALORIES ON AGE BY RACE-SEX CATEGORIES
WTVAR=V90
V=V49 DEPV=V6042
V=V49 DEPV=V203
&END
&REGRESSN
INCLUDE V50=2 AND V51=1
REGRESSION OF DMF AND CALORIES ON AGE BY RACE-SEX CATEGORIES
WTVAR=V90
V=V49 DEPV=V6042
V=V49 DEPV=V203
&END
&REGRESSN
INCLUDE V50=3 AND V51=1
REGRESSION OF DMF AND CALORIES ON AGE BY RACE-SEX CATEGORIES
WTVAR=V90
V=V49 DEPV=V6042
V=V49 DEPV=V203
&END
&REGRESSN
INCLUDE V50=1 AND V51=2
REGRESSION OF DMF AND CALORIES ON AGE BY RACE-SEX CATEGORIES
WTVAR=V90
V=V49 DEPV=V6042
V=V49 DEPV=V203
&END
&REGRESSN
INCLUDE V50=2 AND V51=2
REGRESSION OF DMF AND CALORIES ON AGE BY RACE-SEX CATEGORIES
WTVAR=V90
V=V49 DEPV=V6042
V=V49 DEPV=V203
&END
&REGRESSN
INCLUDE V50=3 AND V51=2
REGRESSION OF DMF AND CALORIES ON AGE BY RACE-SEX CATEGORIES
WTVAR=V90
V=V49 DEPV=V6042
V=V49 DEPV=V203
&END
&REPERR DICTIN=-DICT DATAIN=-DATA
BHS MODEL FOR DMF AND CALORIES
SECU=V97 ST=V91 WTVAR=V94 VAR=49,6042,203-
STATS=(MEANS,RCOEFF,MULTR) REGR=TOT
ST=1-35 MOD=BHS
V=49 DEPV=6042,203
&END
```

```
&REPERR
INCLUDE V50=1 AND V51=1
BHS MODEL FOR DMF AND CALORIES BY RACE-SEX CATEGORIES
SECU=V97 ST=V91 WTVAR=V94 VAR=49,6042,203-
STATS=(MEANS,RCOEFF,MULTR) REGR=TOT
ST=1-35 MOD=BHS
V=49 DEPV=6042,203
&END
&REPERR
INCLUDE V50=2 AND V51=1
BHS MODEL FOR DMF AND CALORIES BY RACE-SEX CATEGORIES
SECU=V97 ST=V91 WTVAR=V94 VAR=49,6042,203-
STATS=(MEANS,RCOEFF,MULTR) REGR=TOT
ST=1-35 MOD=BHS
V=49 DEPV=6042,203
&END
&REPERR
INCLUDE V50=3 AND V51=1
BHS MODEL FOR DMF AND CALORIES BY RACE-SEX CATEGORIES
SECU=V97 ST=V91 WTVAR=V94 VAR=49,6042,203-
STATS=(MEANS,RCOEFF,MULTR) REGR=TOT
ST=1-35 MOD=BHS
V=49 DEPV=6042,203
&END
&REPERR
INCLUDE V50=1 AND V51=2
BHS MODEL FOR DMF AND CALORIES BY RACE-SEX CATEGORIES
SECU=V97 ST=V91 WTVAR=V94 VAR=49,6042,203-
STATS=(MEANS,RCOEFF,MULTR) REGR=TOT
ST=1-35 MOD=BHS
V=49 DEPV=6042,203
&END
&REPERR
INCLUDE V50=2 AND V51=2
BHS MODEL FOR DMF AND CALORIES BY RACE-SEX CATEGORIES
SECU=V97 ST=V91 WTVAR=V94 VAR=49,6042,203-
STATS=(MEANS,RCOEFF,MULTR) REGR=TOT
ST=1-35 MOD=BHS
V=49 DEPV=6042,203
&END
&REPERR
INCLUDE V50=3 AND V51=2
BHS MODEL FOR DMF AND CALORIES BY RACE-SEX CATEGORIES
SECU=V97 ST=V91 WTVAR=V94 VAR=49,6042,203-
STATS=(MEANS,RCOEFF,MULTR) REGR=TOT
ST=1-35 MOD=BHS
V=49 DEPV=6042,203
&END
```

```
STEP 1 ** MEANS AND VARIANCES (WEIGHTED AND UNWEIGHTED) **
      $RUN ISR:OSIRIS.IV SPRINT=*PRINT*
      &RECODE=1
       R1=BRAC(V4501,1=0,2=1,3=2,4=3)
       R2=BRAC(V7501,1=0,2=1,3=2)
       R4=COMBINE R2(3),R1(4)
       R6=V1089
       MDATA R4(99)
      &END
      &USTATS DICTIN=DICTREP4 DATAIN=DATAREP4
      UNIVARIATE STATISTICS
      RECODE=1 VARS=6005,6008 REP=(R4=0/1/2/3/4/5/6/7/8/9/10/11)
      &END
      &USTATS
      WEIGHTED UNIVARIATE STATISTICS
      RECODE=1 VARS=6005,6008 REP=(R4=0/1/2/3/4/5/6/7/8/9/10/11)
      WT=R6
      &END

STEP 2 ** GENCAT ANALYSIS USING UNWEIGHTED DATA **

      $RUN SJS6:GENCAT 1=*SOURCE* 3=*PRINT* 8=-TEMP
          5    3    1                1  UNWEIGHTED ANALYSIS OF P.I.
         12    1                       (6F9.4)
       1.617914 2.037624 2.348564 .961378 1.280140 1.738261
       1.003090 1.148133 1.731346 1.773667 1.768750 2.028848
          3                            (6F9.3)
        0.0115326 0.0560125  0.0300307  0.0053.778  0.0164335
        0.0093619
        0.0169859  0.0179936  0.0096814 1.455348 1.149650
        0.0324495
          7    1   12          1       (12F1.0)          FULL MODEL
      111111111111
      011011011011
      001001001001
      000111111111
      000000111111
      000000000111
      000011011011
      000000011011
      000000000011
      000001001001
      000000001001
      000000000001
          8    1    2                  (12F1.0)          SMOKE
          EFFECT / FULL MODEL
      01
      001
          8    1    3                  (12F1.0)          DRINK
          EFFECT / FULL MODEL
      0001
      00001
      000001
          7    1    6          1       (12F1.0)          MODEL WITH
          NO INTERACTION
      111111111111
```

```
011011011011
001001001001
000111111111
000000111111
000000000111
    8    1    2                    (6F1.0)          SMOKE
    EFFECT / NO INTERACTION
01
001
    8    1    3                    (6F1.0)          DRINK
    EFFECT / NO INTERACTION
0001
00001
000001
    7    1    4          1         (12F1.0)
111000000111
000111111000
012000000012
000001001000
    8    1    1                    (4F2.0)
 1-1 0 0
    8    1    1                    (4F1.0)
0010
    8    1    1                    (4F1.0)
0001

STEP 3 ** GENCAT ANALYSIS USING WEIGHTED DATA **

    5    3    1                1  WEIGHTED ANALYSIS OF P.I.
   12    1                        (6F9.4)
 1.423589 1.835681 1.903959  0.800040  0.965758 1.516229
 0.852681  0.929804 1.462482 1.419985 1.753583 1.689494
    3                             (6F9.3)
 0.0103659  0.0471113  0.0249002   0.0044020  0.0129284
 0.0080.248
  0.0161660  0.0142776   0.0081406 1.290554 1.082204
  0.0280853
    7    1   12          1         (12F1.0)          FULL MODEL
111111111111
011011011011
001001001001
000111111111
000000111111
000000000111
000011011011
000000011011
000000000011
000001001001
000000001001
000000000001
    8    1    2                    (12F1.0)          SMOKE
    EFFECT / FULL MODEL
01
001
    8    1    3                    (12F1.0)          DRINK
    EFFECT / FULL MODEL
0001
00001
```

```
000001
    7    1    6          1       (12F1.0)          MODEL WITH
    NO INTERACTION
111111111111
011011011011
001001001001
000111111111
000000111111
000000000111
    8    1    2                  (6F1.0)          SMOKE
    EFFECT / NO INTERACTION
01
001
    8    1    3                  (6F1.0)          DRINK
    EFFECT / NO INTERACTION
0001
00001
000001
    7    1    4          1       (12F1.0)
111000000111
000111111000
012000000012
000001001000
    8    1    1                  (4F2.0)
 1-1 0 0
    8    1    1                  (4F1.0)
0010
    8    1    1                  (4F1.0)
0001
```

STEP 4 ** &PSALMS RUN TO GENERATE RATIO MEANS & THEIR COVARIANCE STRUCTURE UNDER THE CLUSTERED DESIGN **

```
$RUN ISR:OSIRIS.IV  SPRINT=-PR
&RECODE
 R1=V4501
 R2=V7501
 TABLE A,COLS 1-3,ROWS 1(1-3),2(4-6),3(7-9),4(10-12) ENDTAB
 R3=TABLE(R1,R2,TAB=A)
 R100=1
 R101=V1089
 MDATA R1(99),R2(99)
&END
&PSALMS DICTIN=DICTREP4 DATAIN=DATAREP4 OUTPUT=-SE4
HANES MEAN P.I. BY DRINK-SMOKE CATEGORIES (DETAILED WEIGHTS)
R=1 SORT=4000 PSU=V97 ST=V91 W=R101
REP=(R3=1/2/3/4/5/6/7/8/9/10/11/12) OUT
ST=1-35 MOD=PAIR
PAR=V6008/R100-V6008/R100 SUB=1,1,2,2 P=FULL
SUB=1,1,3,3 P=FULL
SUB=1,1,4,4 P=FULL
SUB=1,1,5,5  P=FULL
SUB=1,1,6,6  P=FULL
SUB=1,1,7,7  P=FULL
SUB=1,1,8,8  P=FULL
SUB=1,1,9,9  P=FULL
SUB=1,1,10,10  P=FULL
SUB=1,1,11,11  P=FULL
```

```
SUB=1,1,12,12  P=FULL
SUB=2,2,3,3  P=FULL
SUB=2,2,4,4  P=FULL
SUB=2,2,5,5  P=FULL
SUB=2,2,6,6  P=FULL
SUB=2,2,7,7  P=FULL
SUB=2,2,8,8  P=FULL
SUB=2,2,9,9  P=FULL
SUB=2,2,10,10  P=FULL
SUB=2,2,11,11  P=FULL
SUB=2,2,12,12  P=FULL
SUB=3,3,4,4  P=FULL
SUB=3,3,5,5  P=FULL
SUB=3,3,6,6  P=FULL
SUBB=3,3,7,7  P=FULL
SUB=3,3,8,8  P=FULL
SUB=3,3,9,9  P=FULL
SUB=3,3,10,10  P=FULL
SUB=3,3,11,11  P=FULL
SUB=3,3,12,12  P=FULL
SUB=4,4,5,5  P=FULL
SUB=4,4,6,6  P=FULL
SUB=4,4,7,7  P=FULL
SUB=4,4,8,8  P=FULL
SUB=4,4,9,9  P=FULL
SUB=4,4,10,10  P=FULL
SUB=4,4,11,11  P=FULL
SUB=4,4,12,12  P=FULL
SUB=5,5,6,6  P=FULL
SUB=5,5,7,7  P=FULL
SUB=5,5,8,8  P=FULL
SUB=5,5,9,9  P=FULL
SUB=5,5,10,10  P=FULL
SUB=5,5,11,11  P=FULL
SUB=5,5,12,12  P=FULL
SUB=6,6,7,7  P=FULL
SUB=6,6,8,8  P=FULL
SUB=6,6,9,9  P=FULL
SUB=6,6,10,10  P=FULL
SUB=6,6,11,11  P=FULL
SUB=6,6,12,12  P=FULL
SUB=7,7,8,8  P=FULL
SUB=7,7,9,9  P=FULL
SUB=7,7,10,10  P=FULL
SUB=7,7,11,11  P=FULL
SUB=7,7,12,12  P=FULL
SUB=8,8,9,9  P=FULL
SUB=8,8,10,10  P=FULL
SUB=8,8,11,11  P=FULL
SUB=8,8,12,12  P=FULL
SUB=9,9,10,10  P=FULL
SUB=9,9,11,11  P=FULL
SUB=9,9,12,12  P=FULL
SUB=10,10,11,11  P=FULL
SUB=10,10,12,12  P=FULL
SUB=11,11,12,12  P=FULL
&END
&SM15:MATGEN INPUT=-SE4 3=-T
```

```
   12    1   70
$ENDFILE

STEP 5 ** GENCAT ANALYSIS USING CLUSTER DESIGN COVARIANCE MATRIX

$RUN SJS6:GENCAT 1=*SOURCE* 3=*PRINT* 4=-T 8=-U
    5    3    4                1  DETAILED SAMPLE WEIGHT
    ANALYSIS OF P.I.
   12   12    1
    2
    7    1   12          1       (12F1.0)          FULL MODEL
111111111111
011011011011
001001001001
000111111111
000000111111
000000000111
000011011011
000000011011
000000000011
000001001001
000000001001
000000000001
    8    1    2                  (6F1.0)          SMOKE
    EFFECT / FULL MODEL
01
001
    8    1    3                  (6F1.0)          DRINK
    EFFECT / FULL MODEL
0001
00001
000001
    7    1    6          1       (12F1.0)          MODEL WITH
    NO INTERACTION
111111111111
011011011011
001001001001
000111111111
000000111111
000000000111
    8    1    2                  (6F1.0)          SMOKE
    EFFECT / NO INTERACTION
01
001
    8    1    3                  (6F1.0)          DRINK
    EFFECT / NO INTERACTION
0001
00001
000001
    7    1    4          1       (12F1.0)          4
```

```
    PARAMETER MODEL
111000000111
000111111000
012000000012
000001001000
     8     1     1                    (4F2.0)
 1-1  0  0
     8     1     1                    (4F1.0)
001
     8     1     1                    (4F1.0)
0001
```

Example 4

```
STEP 1 ** UNWEIGHTED AND WEIGHTED FREQUENCIES **

      DIS INT FI=NEW.REP3
      R @NEW INT FI=NEW.REP3 V=ALL
      DES V=89,90,91
      ONEWAY V=50,100,7339,6008 OP=*
      TWOWAY V=50,7339 OP=*
      TWOWAY V=* OP=* C=V92:1
      TWOWAY V=* OP=* C=V92:1*V100:NONE
      TWOWAY V=* OP=* C=V92:1*V100:SOME
      $R ISR:OSIRIS.IV SPRINT=*PRINT*
      &MIDASFILE INPUT=NEW.REP3
      &RECODE
       RECODE=1
       IF MDATA(V91) THEN REJECT
       R1=BRAC(V50,1=0,2=1)
       R2=BRAC(V7339,1=0,2=1)
       R3=BRAC(V100,1=0,2=1)
       R4=COMBINE R2(2),R1(2)
       R5=BRAC(R4,0=0,1=1,2=2,3=3)
       R6=3854.*V91
       MDATA R3(99),R5(99)
      &END
      &TABLES
      BIVARIATE FREQUENCIES:  UNWEIGHTED
      RECODE=1
      VAR=R3 ST=R5
      &END
      &TABLES
      BIVARIATE FREQUENCIES:  WEIGHTED
      RECODE=1 WTVAR=R6
      VAR=R3 ST=R5
      &END

STEP 2 ** GENCAT ANALYSIS OF UNWEIGHTED FREQUENCIES **

      $RUN SJS6:GENCAT 1=*SOURCE* 3=-PRINT 8=-V
          5     1     1                    UNWEIGHTED ANALYSIS
          4     2                          (2F4.0)
       351 500
       821 753
        58 172
        64 200
          1     2     4     1     1     1  (2F1.0)
      01
          7     1     4           1        (4F2.0)          FULL MODEL
       1 1 1 1
       1 1-1-1
       1-1 1-1
       1-1-1 1
          8     1     1                    (4F1.0)          TEST FOR R1
          EFFECT (RACE)
      0100
          8     1     1                    (4F1.0)          TEST FOR R2
          EFFECT (SMOKE)
      0010
```

```
      8     1     1                    (4F1.0)          TEST FOR R1
     BY R2 INTERACTION
0001
      7     1     3          1         (4F2.0)          MODEL WITH
     NO INTERACTION
 1 1 1 1
 1 1-1-1
 1-1 1-1
      8     1     1                    (3F1.0)          R1 EFFECT /
     NO INTERACTION
010
      8     1     1                    (3F1.0)          R2 EFFECT /
     NO INTERACTION
001
```

STEP 3 ** GENCAT ANALYSIS OF WEIGHTED FREQUENCIES **

```
      5     1     1                    WEIGHTED ANALYSIS
      4     2                          (2F7.2)
 458.98 617.31
 952.49 774.69
  51.80 119.36
  48.04 114.80
      1     2     4     1     1     1  (2F1.0)
01
      7     1     4          1         (4F2.0)          FULL MODEL
 1 1 1 1
 1 1-1-1
 1-1 1-1
 1-1-1 1
      8     1     1                    (4F1.0)          TEST FOR R1
     EFFECT (RACE)
0100
      8     1     1                    (4F1.0)          TEST FOR R2
     EFFECT (SMOKE)
0010
      8     1     1                    (4F1.0)          TEST FOR R1
     BY R2 INTERACTION
0001
      7     1     3          1         (4F2.0)          MODEL WITH
     NO INTERACTION
 1 1 1 1
 1 1-1-1
 1-1 1-1
      8     1     1                    (3F1.0)          R1 EFFECT /
     NO INTERACTION
010
      8     1     1                    (3F1.0)          R2 EFFECT /
     NO INTERACTION
001
```

STEP 4 ** &PSALMS RUN TO GENERATE COVARIANCE MATRIX OF WEIGHTED FREQUENCIES UNDER CLUSTER SAMPLE DESIGN **

```
$COPY -PRINT *MSINK*
$RUN ISR:OSIRIS.IV SPRINT=*PRINT*
&MIDASFILE INPUT=NEW.REP3
&RECODE
```

```
 R1=V50
 R2=V7339
 TABLE A,COLS 1-2,ROWS 1(1-2),2(3-4) ENDTAB
 R3=TABLE (R1,R2,TAB=A)
 R4=BRAC(V100,1=0,2=1)
 R100=1
 R101=3854.*V91
 MDATA R1(99),R2(99),R4(99)
&END
&PSALMS OUTPUT=-SE
HANES 4 X 2 TABLES (SMOKING VS RACE)
R=1 PSU=V97 ST=V191 W=R101 REP=(R3=1/2/3/4) OUT SORT=4000
ST=1-35 MOD=PAIR
NUM=12 PAR=R4/R100-R4/R100 SUB=1,1,2,2 P=FULL
NUM=13 SUB=1,1,3,3 P=FULL
NUM=14 SUB=1,1,4,4 P=FULL
NUM=23 SUB=2,2,3,3 P=FULL
NUM=24 SUB=2,2,4,4 P=FULL
NUM=34 SUB=3,3,4,4 P=FULL
&END
&SM15:MATGEN INPUT=-SE 3=-T
    4    1   70
$ENDFILE

STEP 5 ** GENCAT ANALYSIS OF WEIGHTED FREQUENCIES UNDER OPTION 3
**

$RUN SJS6:GENCAT 1=*SOURCE* 3=-OUT 4=-T 8=-O
    5    3    4              1  WEIGHTED ANALYSIS (VIA
    PSALMS)
    4    4    1
    2
    7    1    4         1        (4F2.0)          FULL MODEL
 1 1 1 1
 1 1-1-1
 1-1 1-1
 1-1-1 1
    8    1    1                  (4F1.0)          TEST FOR R1
    EFFECT (RACE)
0100
    8    1    1                  (4F1.0)          TEST FOR R2
    EFFECT (SMOKE)
0010
    8    1    1                  (4F1.0)          TEST FOR R1
    BY R2 INTERACTION
0001
    7    1    3         1        (4F2.0)          MODEL WITH
    NO INTERACTION
 1 1 1 1
 1 1-1-1
 1-1 1-1
    8    1    1                  (3F1.0)          R1 EFFECT /
    NO INTERACTION
010
    8    1    1                  (3F1.0)          R2 EFFECT /
    NO INTERACTION
001
$COPY -OUT *PRINT*
```


☆U.S. GOVERNMENT PRINTING OFFICE: 1985 461 357 20493

Vital and Health Statistics series descriptions

SERIES 1. **Programs and Collection Procedures.**—Reports describing the general programs of the National Center for Health Statistics and its offices and divisions and the data collection methods used. They also include definitions and other material necessary for understanding the data.

SERIES 2. **Data Evaluation and Methods Research.**—Studies of new statistical methodology including experimental tests of new survey methods, studies of vital statistics collection methods, new analytical techniques, objective evaluations of reliability of collected data, and contributions to statistical theory.

SERIES 3. **Analytical and Epidemiological Studies.**—Reports presenting analytical or interpretive studies based on vital and health statistics, carrying the analysis further than the expository types of reports in the other series.

SERIES 4. **Documents and Committee Reports.**—Final reports of major committees concerned with vital and health statistics and documents such as recommended model vital registration laws and revised birth and death certificates.

SERIES 10. **Data From the National Health Interview Survey.**—Statistics on illness, accidental injuries, disability, use of hospital, medical, dental, and other services, and other health-related topics, all based on data collected in the continuing national household interview survey.

SERIES 11. **Data From the National Health Examination Survey and the National Health and Nutrition Examination Survey.**—Data from direct examination, testing, and measurement of national samples of the civilian noninstitutionalized population provide the basis for (1) estimates of the medically defined prevalence of specific diseases in the United States and the distributions of the population with respect to physical, physiological, and psychological characteristics and (2) analysis of relationships among the various measurements without reference to an explicit finite universe of persons.

SERIES 12. **Data From the Institutionalized Population Surveys.**—Discontinued in 1975. Reports from these surveys are included in Series 13.

SERIES 13. **Data on Health Resources Utilization.**—Statistics on the utilization of health manpower and facilities providing long-term care, ambulatory care, hospital care, and family planning services.

SERIES 14. **Data on Health Resources: Manpower and Facilities.**—Statistics on the numbers, geographic distribution, and characteristics of health resources including physicians, dentists, nurses, other health occupations, hospitals, nursing homes, and outpatient facilities.

SERIES 15. **Data From Special Surveys.**—Statistics on health and health-related topics collected in special surveys that are not a part of the continuing data systems of the National Center for Health Statistics.

SERIES 20. **Data on Mortality.**—Various statistics on mortality other than as included in regular annual or monthly reports. Special analyses by cause of death, age, and other demographic variables; geographic and time series analyses; and statistics on characteristics of deaths not available from the vital records based on sample surveys of those records.

SERIES 21. **Data on Natality, Marriage, and Divorce.**—Various statistics on natality, marriage, and divorce other than as included in regular annual or monthly reports. Special analyses by demographic variables; geographic and time series analyses; studies of fertility; and statistics on characteristics of births not available from the vital records based on sample surveys of those records.

SERIES 22. **Data From the National Monthly and Natality Surveys.**—Discontinued in 1975. Reports from these sample surveys based on vital records are included in Series 20 and 21, respectively.

SERIES 23. **Data From the National Survey of Family Growth.**—Statistics on fertility, family formation and dissolution, family planning, and related maternal and infant health topics derived from a periodic survey of a nationwide probability sample of ever-married women 15–44 years of age.

For a list of titles of reports published in these series, write to:

Scientific and Technical Information Branch
National Center for Health Statistics
Public Health Service
Hyattsville, Md. 20782