ESTIMATION IN THE SOUTHWEST COMPONENT OF THE HISPANIC HEALTH AND NUTRITION EXAMINATION SURVEY

Joe Fred Gonzalez, Jr. and Trena Ezzati, National Center for Health Statistics;
Josefina Lago and Joseph Waksberg, Westat, Inc.

INTRODUCTION

The Hispanic Health and Nutrition Examination Survey (HHANES), sponsored by the National Center for Health Statistics (NCHS), was the first large scale multistage probability sample survey to assess the health and nutritional status of Hispanics in the United States. The HHANES was a multi-purpose survey consisting of personal household interviews, dietary interviews, and a physical examination consisting of an examination by a physician, a dental examination, various physiological measurements and laboratory tests. The HHANES was carried out in the period July 1982 to December 1984. The HHANES was a subnational survey and consisted of three separate target populations: persons 6 months to 74 years of age and of Mexican origin residing in the Southwest (Arizona, California, Colorado, New Mexico, and Texas); persons in the same age group who were of Cuban origin residing in Dade County (Miami, Florida); and, persons in the same age group who were of Puerto Rican origin residing in the New York City area. Separate estimates will be produced for each of the three populations. This paper will focus on the estimation procedures used for the Southwest component.

SAMPLE DESIGN OF THE SOUTHWEST HHANES

It is useful to start off with a brief description of the sample design of the Southwest HHANES . A more detailed description of the HHANES sample design can be found in two previous papers [1,2]. Although the general structure of the HHANES sample design and operation was similar to both of NCHS' first National Health and Nutrition Examination Survey (NHANES I) [3] and the second National Health and Nutrition Examination Survey (NHANES II) [4], there was a major difference between the HHANES and these previous NCHS surveys. The HHANES was a subnational survey of a special subgroup of the U.S. population.

For the Southwest HHANES, a complex, multistage, stratified, probability cluster design was used to survey persons of Mexican origin. The four stages of selection were primary sampling units or PSUs (counties or small groups of contiguous counties), segments (clusters of households), households, and persons. The sampling units at the PSU and segment stage were stratified prior to selection.

Three population subgroups were defined as being of primary analytic interest - persons who were 6 months to 19 years, 20 to 44 years, and 45 to 74 years. In order to assure that the sample size for each of the three groups would be sufficient to support the analyses expected to be made, the sampling plan selected persons in these three groups at somewhat different sampling rates: persons 6 months to 19 years were selected at a rate 3/4 that of persons 45-74, and the rate for persons 20-44 was 1/2 that of persons 45-74. Within each of the three age groups, the sample was designed to be approximately (though not exactly) self-weighting.

Definition of the Southwest HHANES Universe

Although the target population for the Southwest HHANES was conceived to be all households with at least one member of Mexican origin, sampling and data collection were restricted to counties that had a sufficient number and/or percentage of Hispanics to render it economically feasible to establish and operate a medical examination center (MEC) over a four to seven week time period [5]. For purposes of sampling and data collection the Southwest HHANES Universe consisted of 193 PSUs which included about 84 percent of the 1980 Mexican origin population in the United States and about 97 percent of the 1980 Mexican origin population in the five southwestern states.

In addition, in order to reduce screening costs even further, a small percentage (usually less than 10 percent and averaging about 7 percent) of the Mexican origin population within each sample PSU was not covered because block groups (BGs) or enumeration districts (EDs) that the 1980 Census reported as having less than a minimum number (between 50 and 100) of "eligible" Hispanics were excluded.

The count of "eligible" Hispanics within a given BG/ED was defined as the number of Mexican origin persons plus a certain (PSU-specific) percentage of persons of "other Spanish" origin who were assumed to be of Mexican origin. The net coverage rate of the 1980 Mexican origin population in the Southwest was approximately 90 percent (.97 x .93). As will be seen later, one of the goals of the estimation procedure was to adjust the data to compensate as much as possible for the undercovered population.

Stratification of Primary Sampling Units [6]

Information for Hispanics from the 1980 Census was used to stratify the Southwest PSUs. The five PSU characteristics that were used as stratification variables were:

- number of Hispanics
- percent Hispanic
- ratio of the 1980 to the 1970 Hispanic population
- median income
- percent urban

These variables were believed to be correlated with the survey variables of interest.

A critical sample design feature for the Southwest HHANES was that the strata be of equal Hispanic population size. Equal-size strata generally come close to minimizing sampling variances, and at the same time provide efficient work loads since they permit approximately the same number of sample interviews and examinations at each survey location. This requirement was satisfied by forming equal size strata and then applying the same sampling fraction to each stratum.

The Statistical Analysis System (SAS) PROC CLUSTER [7] was the technique that was chosen to stratify the PSUs. The SAS routine PROC CLUSTER (outlined by Johnson [8]) is a multivariate proce-

dure which uses a hierarchical algorithm for grouping similar vector observations. A major drawback of the algorithm is its inability to impose constraints on the cluster sizes. However, by iteratively applying this SAS procedure, the clustering process was controlled to yield strata of approximately equal size. Fourteen strata (2 certainty and 12 non-certainty) were formed for the Southwest.

Selection of Primary Sampling Units

One PSU was selected from each stratum with probability proportionate to size. It was desirable to maximize the probability that counties in the five Southwest States in the universe would be included in the final sample. Therefore, during PSU selection a slightly modified version of a procedure by Goodman and Kish [9,10] was used to obtain a balanced sample with respect to State while retaining a true probability sample design. A detailed description of this controlled selection process and its application to other health examination surveys is given in other NCHS reports [4, 11].

Within PSU Design

Within the PSUs selected, the in-scope population consisted of all households and residents of group quarters (noninstitutional) containing one or more "eligible" Hispanics. An "eligible" Hispanic was anyone whose self-reported national origin was Mexican origin. Because Hispanics constitute a minority of the population in most PSUs, considerable screening of households was required to locate a sample of "eligible" Hispanic households within a PSU. As a means of reducing screening costs, BGs/EDs with very low "eligible" Hispanic density were excluded within each sample PSU and were considered out-of-scope. The overall goal was to attain a minimum of 90 percent coverage of the eligible Hispanic population within each sample PSU. The percentage was well above 90% in most PSUs and averaged about 93% in the entire Southwest sample. In addition, certain types of living quarters were considered out-of-scope, such as, institutionalized populations, Indian reservations, and military installations.

The secondary sampling units (SSUs) were area segments, mainly consisting of blocks or combinations of neighboring blocks (generally contiguous in urban areas. In rural areas the SSUs were blocks or portions of EDs.

The measure of size (MOS) for each segment in the Southwest that was established was approxi mately equal to the sum of: 3/4 of "eligible" Hispanics aged 6 months - 19 years; 1/2 of "eligible" Hispanics aged 20 - 44 years; and, all of "eligible" Hispanics aged 45 - 74 years. The segment sizes were so arranged as to produce about 18 "eligible" Hispanics after the subsampling by age groups.

After selecting the sample segments, households were listed within each segment. Depending on the MOS of a particular segment, all or a subsample of the listed households were screened to determine whether any persons self-identifying as Mexican origin were present.

Once the eligible households were identified, every family within the household was eligible to participate in the HHANES if it contained at least one "eligible" Hispanic who was in the subsample. Every member 6 months - 74 years of age (who usually resides at the household) within an eligible family had a probability of selection since persons were subsampled across eligible families at the same age-specific sampling rates used to compute the MOS of segments.

ESTIMATION PROCEDURES

Goals of Estimation Method

Estimates for the Southwest HHANES were derived through a multistage estimation procedure which was designed to yield statistics that come close to minimizing the mean square errors of desired estimates. The procedure had four basic features and the final weight associated with an examined sample person was the product of the following four components:

1. inflation of sample person observations by the reciprocals of the probabilities of selection at each stage of the design: PSU, segment, household, and sample person;

2. adjustments for interview and examination nonresponse within homogeneous sociodemographic cells. The purpose of this adjustment was to reduce the potential bias due to nonresponse, under the assumption that within adjustment cells the characteristics of the respondents are similar to those of the nonrespondents;

3. adjustment for noncoverage within sample PSUs. The purpose of this adjustment was to reduce the potential bias due to the exclusion of BGs/EDs with few Hispanic residents; and

4. poststratified ratio adjustment by age and sex to make the final sample estimates of the population correspond to the most current Bureau of the Census estimates of the civilian noninstitutionalized target population. The ratio adjustment served two purposes. One was to reduce sampling variances, as is normally accomplished by ratio estimates. The second was to dampen any potential biases introduced by the omission of counties with small Hispanic populations.

Components 1,2, and 4 above are the three basic components that are normally included in the estimation procedures for most large scale surveys. However, component 3 which deals with the noncoverage of the "eligibile" Hispanic population residing in excluded BGs/EDs within sample PSUs was fairly unique. Although reducing the coverage rate of the eligible Hispanic population within the sample PSUs resulted in a considerable savings in screening costs, the NCHS realized that it introduced some bias in the sample. Although the number of Hispanics omitted was fairly small, an important concern was that the low Hispanic density BGs/EDs contained a disproportionate percentage of high income Hispanic households. It seemed likely that as Hispanics (as other ethnic groups) climb the socioeconomic ladder they are more likely to move out of their high ethnic concentration areas and assimilate more into the general population, in which case the sample would underrepresent high income Hispanic households.

In order to investigate the magnitude of the undercoverage of the high income Hispanic house-

holds, a comparison [2] was made of the 1979 percent distribution of Hispanic family income for all BGs, in-scope BGs, and out-of-scope BGs within each sample PSU . Most of the sample PSUs showed some difference in the income distributions of Hispanics in in-scope BGs vs. out-of-scope BGs. Therefore, a decision was made to make a noncoverage adjustment by income within each sample PSU, using the Census data on Hispanic income in in-scope and out-of-scope areas.

HHANES SOUTHWEST ESTIMATOR

The following is a detailed description of the HHANES Southwest estimator of an aggregate. The estimator that follows will reflect the complex multistage, stratified, probability cluster design of the Southwest HHANES.

An attempt has been made to emphasize the nested design of the HHANES and the level, as well as the sequence, of a particular stage of sampling or adjustment by hierarchically ordering the subscripts.

Consider an X-characteristic of the r^{th} sample person in the q^{th} household unit, p^{th} segment, ℓ^{th} age-income-household size (HHS) interview nonresponse (NR) adj. cell, k^{th} age-HHS examination NR adj. cell, j^{th} income noncoverage (NC) adj. cell, i^{th} PSU, h^{th} stratum, and the g^{th} age-sex poststratification cell in the Southwest, denoted by $X_{ghijk\ell pqr}$, i.e.,

Subscript	Variable	Range
g	age-sex (10x2) poststratification cell	g=1,...,20
h	stratum	h=1,...,14
i	PSU	i=1
j	income NC adj. cell	$j=1,...,J_{hi}$
k	age-HHS (3x3) exam NR adj. cell	k=1,...,9
ℓ	age-income-HHS (3x3x2) interview NR adj. cell	ℓ=1,...,18
p	segment	$p=1,...,P_{hi}$
q	household unit	$q=1,...,Q_{hip}$
r	sample person	$r=1,...,R_{ghijk\ell pqr}$

[Note: The above range for the g subscript applies to the estimation procedures for all sample persons that were interviewed or examined. There were also special subsamples of examined persons that were used for laboratory tests (such as: glucose tolerance test (GTT), ultrasound, and pesticides); additional test-specific subsampling weights were assigned for those groups and the range for the g subscript varies according to the specific age groups for which the laboratory tests were administered.]

A. Simple Inflation Estimator

Since the data were obtained from sample persons selected through a four-stage design, a sample observation, X_{ghipqr} , must be inflated by the reciprocals of the sampling probabilities at each stage of selection. That is, the simple inflation estimator, X'_g , of a total aggregate, X_g , for the g^{th} age-sex group in the Southwest is secured as follows:

$$X'=\sum_g \sum_{hi} W_{1.hi} \sum_p W_{2.hip} \sum_q W_{3.hipq} \sum_r W_{4.hipqr} X_{ghipqr}$$

which, of course, can be equivalently written as

$$X'=\sum_g \sum_h \sum_i \sum_p \sum_q \sum_r W_{1.hi} W_{2.hip} W_{3.hipq} W_{4.hipqr} X_{ghipqr}.$$

Where

$W_{1.hi}$ = first-stage design weight = $(P_{1.hi})^{-1}$, the reciprocal of the probability of selecting the i^{th} PSU in the h^{th} stratum [Note: $W_{1.hi} = 1$, for those PSU's that were selected from self-representing (or certainty) strata.]

$W_{2.hip}$ = second-stage design weight = $(P_{2.hip})^{-1}$, the reciprocal of the probability of selecting the p^{th} segment in the i^{th} PSU and the h^{th} stratum.

$W_{3.hipq}$ = third-stage design weight = $(P_{3.hipq})^{-1}$, the reciprocal of the probability of selecting the q^{th} household unit in the p^{th} segment, i^{th} PSU, and h^{th} stratum.

$W_{4.hipqr}$ = fourth-stage design weight = $(P_{4.hipqr})^{-1}$, the reciprocal of the probability of selecting the r^{th} sample person in the q^{th} household unit, p^{th} segment, i^{th} PSU, and h^{th} stratum. This is also known as the differential age weight for subsampling sample persons within household.

The product of the above four sampling weights is usually referred to as the basic weight, i.e., hereafter, the basic weight will be denoted by:

$$W^B = W_{1.hi} W_{2.hip} W_{3.hipq} W_{4.hipqr}.$$

B. Interview Nonresponse Adjustment

In order to adjust survey estimates for interview nonresponse, i.e., for persons that were sampled but were not interviewed, the basic weight, W^B, was multiplied by a interview nonresponse adjustment factor, $f_{hi\ell}$. That is, the interview nonresponse adjusted weight for the r^{th} sample person in the ℓ^{th} cell in the hi^{th} PSU is as

follows:

$$W' = f_{hi\ell}\, W^B$$

Where

$$f_{hi\ell} = \frac{\sum_{r \in S_\ell} W^B}{\sum_{r \in I_\ell} W^B}$$

= adjustment for interview nonresponse computed by dividing the sum of the basic weights for all sample (S) persons within the ℓ^{th} cell by the sum of the weights for all interviewed persons within the same ℓ^{th} cell. ϵ denotes element (or member) of set S or set E above. The definition of the cells for the interview nonresponse adjustment was:

Age:	1 = 6 months to 19 years
	2 = 20 to 44 years
	3 = 45 to 74 years
Household size:	1 = less than 4 persons
	2 = 4 persons or more
Income:	1 = less than $10,000
	2 = $10,000 to $19,999
	3 = $20,000 and over

Note: For sample persons (SPs) who were missing income (not obtained on the family questionnaire) the imputed value was the 1980 Census median income of the BG/ED where the sample segment was located.

C. Examination Nonresponse Adjustment

In order to adjust survey estimates for examination nonresponse, i.e., for persons that were interviewed but were not examined, the interview nonresponse adjusted weight, W', was multiplied by an examination nonresponse adjustment factor, f_{hik}. That is, the examination nonresponse adjusted weight for the r^{th} sample person in the k^{th} cell in the hi^{th} PSU is as follows:

$$W'' = f_{hik}\, W'$$

Where

$$f_{hik} = \frac{\sum_{r \in I_k} W'}{\sum_{r \in E_k} W'}$$

= adjustment for examination nonresponse computed by dividing the sum of the interview nonresponse adjusted weights, W', for all interviewed (I) persons within the k^{th} cell in the hi^{th} PSU by the sum of the W' for all examined (E) persons within the same cell. As before ϵ denotes element (or member) of set I or set E above.

The definition of the cells for the examination nonresponse adjustment was:

Age:	Household size:
1 = 6 months to 19 years	1 = 1-2 persons
2 = 20 to 44 years	2 = 3-4 persons
3 = 45 to 74 years	3 = 5 persons or more

D. Noncoverage Adjustment

As mentioned earlier, a noncoverage adjustment was deemed appropriate to partially compensate for the somewhat higher undercoverage of high income Hispanic households within sample PSUs . Since the survey coverage of the Mexican American population was different among PSUs, the noncoverage adjustment was carried out on a PSU-by-PSU basis by income. That is, the interview and examination nonresponse adjusted W'' was multiplied by f_{hij} :

$$W''' = f_{hij}\, W''$$

Where

$$f_{hij} = \frac{N_{hij}}{N'_{hij}}$$

= noncoverage adjustment factor which is the ratio of the total Spanish origin families in the j^{th} income cell in the hi^{th} PSU in the 1980 Census to the number of Spanish origin families in in-scope BG/ED's in the same hij^{th} cell.

The income cells defined for the noncoverage adjustment were:

1 = less than $5,000	5 = $20,000 to $24,999
2 = $5,000 to $9,999	6 = $25,000 to $34,999
3 = $10,000 to $14,999	7 = $35,000 to $49,999
4 = $15,000 to $19,999	8 = $50,000 and over

In three PSUs two or three of the high income cells were collapsed because the number of Hispanic families in those income cells was very small and would have resulted in a very unstable estimate of the noncoverage ratio.

E. Poststratification Adjustment

The interview and examination nonresponse and noncoverage adjusted estimator for the g^{th} age-sex group in the Southwest is:

$$X''_g = \sum_h \sum_i \sum_p \sum_q \sum_r W''' X_{ghijkpqr} .$$

The last adjustment that was made in the HHANES estimator was a poststratification ratio adjustment within the g^{th} age-sex group in the Southwest, that is,

$$X'''_g = \frac{Y_g}{Y'_g} X''_g$$

Where

Y_g = updated Census population count in the g^{th} age-sex group in the Southwest

$$Y'_g = \sum_h \sum_i \sum_p \sum_q \sum_r W'''_{ghijkℓpqr} Y_{ghijkℓpqr}$$

= nonresponse and noncoverage adjusted estimate of the population count in the g^{th} group.

$$Y_{ghijkℓpqr} = \begin{cases} 1, \text{ if the } r^{th} \text{ person in the } q^{th} \text{ household, } p^{th} \text{ segment, } i^{th} \text{ PSU, } h^{th} \text{ stratum, falls in the age-sex group } g \\ 0, \text{ otherwise} \end{cases}$$

F. The Complete Southwest HHANES Estimator

The complete nonresponse and noncoverage adjusted poststratified estimator, $\hat{X}$, of a total aggregate, X , in the Southwest HHANES is secured as follows:

$$\hat{X} = \sum_g \frac{Y_g}{Y'_g} \sum_h \sum_i \sum_p \sum_q \sum_r W'''_{ghijkℓpqr} X_{ghijkℓpqr} .$$

REFERENCES

1. Gonzalez, Joe F., White, Andrew A., and Ezzati, Trena: Sample Design for the Hispanic Health and Nutrition Examination Survey, 1982-84. Proceedings of the Survey Research Methods Section, American Statistical Association, Aug. 1984.

2. Gonzalez, Joe F., Ezzati, Trena, and White, Andrew A.: Sample Design and Estimation Issues in the Hispanic Health and Nutrition Examination Survey, 1982-84. Invited Papers to the 1984 National Center for Health Statistics Data Use Conference on Small Area Statistics, Snowbird, Utah, pp. 18-23.

3. National Center for Health Statistics: Plan and Operation of the Health and Nutrition Examination Survey, United States, 1971-1973. By H. W. Miller. Vital and Health Statistics. DHEW Pub. No. (PHS) 79-1310. Series 1, No. 10A and 10B. Public Health Service. Washington, D.C. U.S. Government Printing Office, Dec. 1978.

4. National Center for Health Statistics: Plan and Operation of the Second National Health and Nutrition Examination Survey, 1976-1980. By A. McDowell, A. Engel, J. T. Massey and K. Maurer. Vital and Health Statistics. DHHS Pub. No. (PHS) 81-1317. Series 1, No. 15. Public Health Service. Washington, D.C. U.S. Government Printing Office, July, 1981.

5. Massey, J. T. and Ezzati, T. M.: HHANES Universe and Estimated Coverage. Unpublished Memorandum. National Center for Health Statistics, Aug. 1981.

6. Gonzalez, J. F. and White, A. A.: Stratifying Primary Sampling Units with the SAS Cluster Procedure. Proceedings of the Social Statistics Section, American Statistical Association, Aug. 1982.

7. SAS Institute, Inc.: SAS Users' Guide, 1979 edition. Raleigh, N.C.

8. Johnson S.C.: Hierarchical clustering schemes. Psychometrica. XXXII:241-54, 1967.

9. Goodman R., and Kish L: Controlled selection - a technique in probability sampling. JASA 45:350-72, 1950.

10. Kish L.: Survey Sampling. New York. John Wiley and Sons, Inc., 1965.

11. National Center for Health Statistics: Sample design and estimation procedures for a National Health Examination Survey of children. By E.E. Bryant, J.T. Baird, and H.W. Miller. Vital and Health Statistics. DHEW Pub. no. (HSM)72-1005. Series 2, no.43. Health Services and Mental Health Administration. Washington, D.C. U.S. Government Printing Office. Aug. 1971.

ACKNOWLEDGEMENT

The authors would like to express their gratitude to Mrs. Janice Melvin and Ms. Freda Stevens for their expert assistance in typing the manuscript for publication. Also, the authors would like to thank Ms. Linda Tompkins for her assistance in the preparation of this manuscript.