ETHNIC STRATIFICATION IN THE CALIFORNIA HYPERTENSION SURVEY*

Sandra Hendricks, Westat, Inc., Amnon Igra, California Department of Health Services, and
Joseph Waksberg, Westat, Inc.

1. INTRODUCTION

The California Hypertension Survey had two major aims: first, to estimate the prevalence and control of hypertension for the adult population of the State; and second, to estimate these same characteristics for four major population categories: Blacks, Hispanics, Asians, Pacific Islanders, and Whites. Within the Asian group, a secondary goal was to obtain estimates for the three major groups: Chinese, Japanese, and Filipino.

It was clear from the outset that an equal probability sample would not produce enough sample cases for the smaller ethnic groups to permit separate analyses. The proportion of the population in the various groups ranged from about 4 percent for the Asian part to 75 percent for the White population. With equal probability sampling, an unreasonably large sample would have been necessary to provide adequate data for the smaller groups. Stratification by ethnicity, with different sampling rates for the various groups, was thus decided on as the way to obtain prespecified numbers of each ethnic group. Furthermore, it was also clear that data on ethnicity from the 1970 Census, to be used in 1978–79, could serve only as a rough guide. This argued against subsampling ethnicities at different rates, using a single shot technique based on 1970 Census data. Consequently, a decision was made to obtain more current estimates of the percentage of each ethnic group in some relatively large sampling frame.

Given the financial constraints that are always present, a decision was still necessary on how extensive such a first-stage screening might need to be in order to have an adequate basis for setting sampling rates to meet ethnic targets. An 18,000 household screening level was chosen as providing a reasonable compromise between cost and the precision of the resulting data for individual race-ethnic groups.

These decisions provided the framework for most of the remaining sampling activities. The rest of this paper will concentrate on the sampling issues involved in accomplishing the required goals, and the effect on some of the field operations.

2. SAMPLE DESIGN ISSUES

General Design. The survey was conducted by Westat, Inc., for the State of California Hypertension Program. The basic framework was to be a multistage sample, with the first stage Primary Sampling Units (PSUs) consisting of fairly large area clusters, the second stage comprising smaller area segments, and the third stage consisting of a subsample of households within the second stage segments. Some information was to be obtained for all adults in the sample household. More detailed data were required for one randomly selected adult (RSA) in each unit. A multistage sample design of this sort is, of course, an efficient sample to use when face-to-face interviews with a sample of the population is required. The State stated the sample sizes for each of these stages—100 PSUs, 400 area segments, and 18,000 households to be screened, with a subsample of the 18,000 to be interviewed and to have blood pressure measured.

Fairly standard methods were used to designate these three or four stages of the sample. Most of the PSUs were defined to be groups of generally contiguous Census tracts containing an average of about 7,500 households. In the non-tracted parts of California, minor civil divisions were used instead of tracts. The area sample was only used to represent housing units built before 1970. A sample of building permits issued in the period 1970–1978 was selected to represent housing units built after 1970. For this new construction part of the sample the PSUs were the cities, towns, and counties that issue building permits.

The PSUs were stratified by geography prior to sample selection. A sample of 100 PSUs was then selected with probability proportionate to size. For the area PSUs, the measure of size was the 1970 count of housing units. For new construction, it was the number of housing units authorized by building permits.

Each sample PSU was subdivided into area segments containing an average of some 200 housing units per segment. Where Census block statistics existed, groups of neighboring blocks constituted the area segments. In other parts of California, small geographic areas with physical or political boundaries were used. For the new construction sample, sets of building permits issued in the same political jurisdiction during a group of consecutive months (or years in some cases) were used.

Four such segments were selected at random within each PSU, with probability proportionate to size. The selected segments were listed, and a sample of housing units in each segment was designated for screening. The sampling rate in each segment was determined in such a way that most households in California had an identical chance of selection for screening.

Part of California had been delineated as a "high Asian" stratum, and some of the PSUs had been selected in that stratum. All households within the Asian PSUs also had an identical chance of selection, but that probability was 3.5 times as great as the one used in the rest of California.

Interviewers chose the randomly selected adult through the use of a sampling table which was designed to rotate the sample selection among different household members.

Sampling for Race-Ethnic Groups. Double sampling was used to select the households ultimately designated for interview and blood pressure measurements. The 18,000 households were screened, essentially to obtain information on race and ethnicity. Subsamples were then selected for the detailed interview. Different subsample rates were used for the four race-ethnic groups that were the focus of the survey.

From the early planning stages, it was clear that there were likely to be problems in achieving the ethnic sample sizes desired by the State. If the screening had been followed up with a sample selected at a single rate throughout the State, the ethnic composition of the sample would have been roughly the same as in the entire State. It was estimated that such a sample would produce only about half of the number of Asians desired and that the number of Black households desired would only be attained if a very high cooperation rate were achieved (see Table 1).

Although the ability to achieve the number of Black households was somewhat uncertain, it did not appear that the discrepancy, if any, would be very large. No special provisions were therefore made for Black households. However, a different situation existed for the Asian part of the sample. It was obviously necessary to oversample Asian households in the screening operation. The only way to oversample them was to sample PSUs and segments in which Asians tend to live, at a higher rate than the rest of the State.

Of course, not all Asian households reside in such areas. Consequently, increasing the sampling rate in these areas does not reduce the sampling variances at a comparable rate. Useful reductions in the standard errors can be obtained by oversampling only up to a certain point. The level of oversampling that is efficient depends on the proportion of Asians living in areas to be oversampled and on the percentage of the population within the areas who are Asians.[1]

An intensive effort was made to identify areas in California with high concentrations of Asians. We started off by classifying as "Asian" all Census tracts (and comparable units in nontracted areas) with 10 percent or more Asians in the 1970 Census. Tracts with 10 or 15 percent Asians can hardly be considered as predominantly Asian. However, it was clear that only by using such low levels would reasonably high proportions of Asians be found. Secondly, both state demographer and county planning officials were contacted in an attempt to update the 1970 Census data. Planning offices in all counties containing a significant number of Asians in 1970 were contacted. Some of these offices, though not all, were able to indicate new areas of Asian concentration.

All Census tracts classified as Asian by either of these two sources were used to construct the Asian stratum PSUs. Special tabulations of the 1970 Census tapes indicated that 40 percent of the Asians in California lived within the Asian stratum in 1970. There was no good estimate of the comparable 1979 percentage. If 40 percent still applied, and if the Asian group constituted 4 percent of the total California households, then sampling the Asian stratum at a rate 3.5 times the rate in the rest of California would have produced the required number of Asian households. If either of these percentages was now lower than in 1970, a higher rate of oversampling in the Asian stratum was necessary. Conversely, if the percentages had increased since 1980, then the rate of oversampling could be reached.

Westat recommended, and the State agreed, to go ahead on the assumption that the rates had not changed since 1970. There were two reasons for this recommendation. The first was that the factor of 3.5 was close to the upper limit of oversampling that would result in significant reductions in the sampling variances. Oversampling beyond this level would produce more sample units, but these additional units would have only a trivial effect on the variances.

The second reason was the effect that oversampling would have on statistics for other race-ethnic groups. Obviously with a fixed total screening workload, the more cases that were allocated to the Asian stratum, the less there would be available for the rest of the sample. As mentioned earlier, even with equal probability sampling there was some uncertainty that the Black sample would be as large as desired. Oversampling in the Asian stratum made it even more difficult to achieve an equal probability sample of the Black population of the desired size.

Resulting Sample Size. At the time the screening was mostly completed but before the interviewing started, it became clear that some of the sample sizes could not be attained. The most serious problem was with the Asian sample. The main reason for the shortfall was that the Asian stratum contained only about 20 percent of the Asian households in California in contrast to the 40 percent in 1970. The oversampling was thus only about half as effective as had been hoped. A contributing factor for the entire sample, both Asian and other ethnic groups, was that the nonresponse rates turned out to be greater than expected. The total number of Asian households in California appeared

to be reasonably close to the 4 percent of all households estimated in advance by the State. This estimate did not, therefore, contribute to the problem.

There are two possible reasons for the decrease from 40 to 20 percent. The first is that there has been a reduction in the proportion of Asians who live in concentrated areas. The second is that Asians have established new areas of concentration but that the sources we approached were not sufficiently knowledgeable. If Asians are more dispersed than 10 years ago, there is no efficient way of getting a larger sample except by increasing the screening beyond the 18,000 specified. If there are many new concentrated areas, the local experts that were contacted were not aware of them. Some of the difficulties may thus be due to working with census materials almost 10 years old.

In order to partially compensate for the shortfall, it was decided to supplement the Asian sample within segments that contained a high proportion of Asians. An additional sample of households in these areas was drawn from units that had been listed but not previously screened. It was recognized that this sample would not reduce appreciably the sampling variances for statistics on all Asians in California. It would permit analysis based on the special part of the California Asian population living in urban and ghetto-like surroundings.

Partially because of a higher than anticipated nonresponse rate and partially due to the oversampling in Asian areas, leaving less available for the remaining sample, the Black sample was also below the target number. However, this shortfall was fairly small and was made up by bringing in Black households in the Asian stratum at a higher sampling rate than in the rest of California.

The problems in achieving the initially planned sample sizes and the higher than anticipated nonresponse rates resulted in revisions in the sampling goals. The revised target numbers and the final sample sizes actually attained are summarized in Table 2 which shows the number of interviews by race-ethnic groups.

3. ORGANIZATION OF PRESCREENING AND SUBSEQUENT SUBSAMPLING OPERATIONS

As indicated above, a double sampling procedure was used in which the sample of 18,000 households was reduced to a smaller number by taking subsamples within each of the race-ethnic groups. The prescreening operation provided the information necessary for classifying the households by race and ethnicity.

Two ways of organizing the prescreening were considered. The first was to carry out the prescreening and interviewing as a single operation. This meant that after the listing was complete, the screening sample would be selected as an office operation, and the subsampling for the detailed interviews would be carried out by the interviewers. The second way was to do the prescreening simultaneously with the listing. Then, after the prescreening was complete, subsamples could be selected.

There was a serious disadvantage to the first method. It required specifying the subsampling rates in advance. Because the actual race-ethnic composition of California was only approximately known, this would have introduced a considerable amount of uncertainty in the ability to attain the required sample sizes. The second method was therefore chosen. However, a modification was made in order to improve the efficiency of the field operations.

In order to first prescreen and then reach the exact sample sizes specified, it would have been necessary to complete virtually the entire listing and prescreening before

starting the subsampling. This seemed undesirable since it would have been costly and would have resulted in a severe delay in schedules. Furthermore, even with such procedures, some deviations from the exact sample sizes would have occurred. Part of the reason is that the prescreening operation classified households, not persons. Households containing persons from more than one group were classified as though a household was completely homogenous. Of course, during the interview phase, each person was classified individually. A second reason was mobility. The sample was viewed as a sample of addresses, not persons, and households who moved between the time of prescreening and interviewing would create some difference between the size of the intended sample and the size actually interviewed. A third reason for change in the way households are classified is errors in prescreening. A few errors are likely no matter how much quality control is imposed.

Since exact sample sizes could not be achieved regardless of the procedure used, it seemed advantageous to introduce a plan that could be carried out efficiently, even at the price of some additional minor deviations from the desired sample sizes. Consequently, the following method was used in moving from the prescreening to the interview phase of the survey.

1. For each segment, before subsampling was started, there was a cutoff on all further prescreening. At that time, all sample households were classified as follows:
   * Completed—race determined (or vacant unit);
   * No one at home, refused or other nonresponse, but conversion attempt not yet made; and
   * Household not yet approached for interview. Obviously, most households were in the first category above.

2. For households in the second and third categories, an attempt was made to obtain race and ethnicity from neighbors. When this could not be done, they were assigned tentative race codes for subsampling purposes. The race code assigned was the dominant race in the segment. A similar procedure was followed for vacant units.

3. Subsampling was then carried out for all completed cases and those with tentative race codes. All of the subsampled cases were assigned for interview. The final race code for a household was, of course, the one reported at the time of the interview, not the one tentatively assigned from prescreening. (This was also the case in the completed households where the initial household had moved or the race reported during prescreening was in error.)

4. In weighting the sample results, all households are weighted in accordance with the way they were sampled and not as they are finally classified. This, of course, means that the weights are not identical for all Hispanics or Whites. However, in practice, only a small percentage of the White or Hispanic sample had weights different from the "normal" weight for the particular race-ethnic group.

Several features of this plan should be pointed out. The statistics will be unbiased because the weighting is in accordance with the probabilities of selection. There are a number of advantages to these procedures. They provided a smoother transition from the prescreening to the interview phase. They permitted more intensive followup effort because of the increased length of time available for followup. Also, housing units that refused the prescreening interview, could not be contacted in time, or were vacant were included in the subsample and given a second chance for response, thus further increasing the response rate. They reduced costs

because the effort involved in attempting to make contact with the hardcore "not at homes" in many cases only had to be done for the subsample, instead of the full 18,000 household sample.

The effectiveness of the different methods used in assigning race codes for sampling purposes was evaluated by comparing the race assigned at the time of prescreening to the race reported in the final interviews. The results are shown in Table 3. As can be seen, the method of assigning race according to the predominant race of the segment was almost as effective at predicting the race of the household as obtaining a racial classification from a neighbor was. Of those households which were assigned a race code based on the predominant race of the segment, 75.6 percent reported the same race at interviewing. Where race was reported by a neighbor, 80.2 percent of the households reported the same race. When a prescreener had been completed with a household member, 92.3 percent of the households reported the same race at interviewing.

4. ESTIMATION AND WEIGHTING

A three-stage weighting procedure was used. The first stage weights consisted of the reciprocals of the probabilities of selection of each sample person. Separate weights were necessary for tabulations of the randomly selected adults and for tabulations of all interviewed persons.

The second-stage weights were designed to adjust for nonresponse. They were carried out by grouping the segments into homogenous strata, and using the ratios of weighted total households designated for interview to the weighted number of interviewed households. The weights used were the household first-stage weights. The ratios were the nonresponse adjustments. The ratios were multiplied by the first-stage weights to create the second-stage weights.

Thirty weighting groups were established. Because so much of the analysis was to be focused on race and ethnicity, these were basic items in defining the 30 groups. Other information used to create homogeneity was geography and income. The race-ethnicity and income are based on 1970 Census data for the Block Groups (BGs) or Enumeration Districts (EDs) containing the segments. The new construction sample was not based on 1970 Census data, and advance information on race-ethnicity and income are not available. For new construction, we used geography and the year the building permit was issued to create the groups.

Segments were defined as Black or Hispanic if the race-ethnic proportion of the population of the BGs or EDs from which they were selected was 20 percent or greater. For Asians, the cutoff was 10 percent. All other segments were classified as White. Similarly, the family income of the BG or ED containing a segment was used for income stratification. Two year-built categories were used in the new construction stratification—permits issued 1970–1973 and permits issued in 1974 or later.

The first two stages of estimation produced generally very satisfactory estimates of the population of California. When the second-stage weights were applied to the sample households, they resulted in an estimate of 8,498,000 households in California. The Census Bureau has published an estimate of the number of households in California in 1978 (P–25, No. 807), amounting to 8,401,000. Updating this to 1979 would produce a figure of about 8,700,000. The survey estimate is thus within about 2 percent of the Census estimate, remarkably close for a sample survey. The weighted population estimates are almost as close to the State's own estimates of the population of California 18 years and over— within about 3 percent for randomly selected adults and 6 percent for the sample of all persons. The combined group

produced a somewhat lower population figure because no special nonresponse adjustment was made for households in which some of the persons were not interviewed.

Although the total population estimates came quite close to the State's and Census Bureau's population figures, an analysis of the survey figures by race, age, and sex indicated that some anomalies seemed to exist. For some age groups, the sex ratios did not seem reasonable. Also, comparisons with the State estimates showed fairly large differences in a few classes. A third, and final, stage of estimation was therefore introduced to reduce some of these discrepancies.

The third stage was a ratio estimation procedure carried out by post-stratifying the population into 16 race-age-sex groups. The 16 cells were defined as the crossclassification of two race groups—Black and All Other (the latter including White, Hispanic, Asian, and All Other races except Black)—two sexes, and four age groups: 18—34, 35—49, 50—64, and 65 years and over. The independent figures used were those produced by the 1976 Survey of Income and Education (SIE), updated to 1979 by multiplying the SIE data by a factor of 1.075. A number of other alternatives were considered, including establishing 32 cells with Asians and Hispanics defining separate cells, and using other sources of independent data instead of the SIE. However, analyses of the other data did not give any assurance that they were any more accurate than the results of the Hypertension Survey.

The final ratio-estimate factors are shown in Table 4, together with implied factors for the two race groups and for the total. The much higher factors for Blacks can be due to several reasons: (a) sampling error; (b) nonresponse rates not fully compensated for in the nonresponse adjustment procedure; and (c) undercoverage of some Black persons in the survey. (Such undercoverage is quite common in household surveys.)

The ratio estimates automatically made the Black survey estimates consistent with the updated SIE. They also made the total of Asians, Hispanics, and White and Others consistent, but not each race-ethnic group by itself. It is interesting to compare the final survey estimates with the updated SIE. Table 5 compares the population estimates based on RSAs with the updated SIE. As can be seen, the survey shows a slightly lower White population, but considerably more Hispanics and Asians. As mentioned earlier, the survey was not forced to agree with SIE because the data looked at least as plausible. Better information will be available after the 1980 Census data are tabulated.

---

1 See "The Effect of Stratification with Differential Sampling Rates on Attributes of Subsets of the Population" by Joseph Waksberg, in Proceeding of the Social Statistics Section of the 1973 Annual Meeting of the American Statistical Association

TABLE 1. Sample sizes in a self-weighting sample of 18,000 households in California compared to desired sizes

Type of Household	State's Estimate of Percent in California	Number in 18,000 Household Sample	Number Desired for Survey (Initially)
Asian	4%	720	1,250
Black	8%	1,440	1,250
Hispanic	16%	2,880	1,250
White and Other	72%	12,960	3,125

TABLE 2. Number of interviewed households and persons in California Hypertension Survey, by race-ethnicity*

Race— Ethnicity	Number of Persons			
	Randomly Selected Adult		Other Household Members	
	Target No.	Actual	Target No.	Actual
Total	4,025	4,068	3,622	3,706
Black	850	825	765	580
Hispanic	850	800	765	872
Asian	600	518	540	630
White and Other	1,725	1,925	1,552	1,624

* These tables do not include households brought into the sample through the Asian Supplement as these cases contributed little to the lowering of variances.

TABLE 3. Accuracy of racial designation for alternate method of assigning race

Method of Assigning Race	Number of Households	Percent in Which Assigned Race Equals Race Obtained in Interview
Response to screener	3,660	92.2%
From neighbor	252	80.2%
Imputed from predominant race of segment	677	75.6%

TABLE 4: Third-stage ratio estimate factors for California Hypertension Survey

Race	Age	Sex	Factor RSA	Factor All Adults
Black	18–34	Male	1.55	1.35
		Female	1.29	1.30
	35–49	Male	.93	1.13
		Female	.97	1.10
	50–64	Male	1.41	1.40
		Female	.96	1.00
	65+	Male	.90	.89
		Female	1.02	1.11
Other	18–34	Male	.93	1.00
		Female	1.03	.95
	35–49	Male	1.03	1.08
		Female	.91	.96
	50–64	Male	1.19	1.16
		Female	.97	1.08
	65+	Male	1.09	1.18
		Female	1.11	1.12
Implied all Blacks			1.17	1.20
Implied all Others			1.02	1.05
Implied Total			1.03	1.06

TABLE 5: Population estimates based on Randomly Selected Adults (RSAs) compared to SIE estimates for Hispanics, Asians, and White and Other

Age—Sex		Ratio-Estimate Factor	Survey Estimates (000s) White and Other	Survey Estimates (000s) Asian	Survey Estimates (000s) Hispanic	Ratio to SIE White and Other	Ratio to SIE Asian	Ratio to SIE Hispanic
18–34	Male	.93	2,217	119	677	.92	1.09	1.35
	Female	1.03	2,267	201	647	.93	1.52	1.20
35–49	Male	1.03	1,360	109	285	.99	1.49	.94
	Female	.91	1,323	123	286	.97	1.04	1.14
50–64	Male	1.19	1,178	86	229	.90	1.62	1.71
	Female	.97	1,318	80	202	.93	1.63	1.44
65+	Male	1.09	736	31	69	1.03	.63	.93
	Female	1.11	1,082	25	81	1.01	.63	1.05
Total		xx	11,477	774	2,476	.95	1.24	1.23

Note: SIE estimates used:
1. Ratio estimates to national figures by age, sex, and race (Black and Other).
2. The State estimates were then adjusted to agree with independently derived estimates of state populations for three age categories.

The SIE was conducted in 1976, and the data were multiplied by a factor of 1.075 to project data to 1979.