Introduction: In 1956, Congress passed the Health Acts for health surveys on the U.S. population. The health surveys include three distinct types: interview surveys, examination surveys, and record surveys. Any one type cannot cover all the aspects of health events. Table 1 shows NCHS surveys. Other NCHS surveys - telephone sample survey, the master facility inventory survey (not sample survey), and the sampling of vital records to construct U.S. life table - are not included in this report. The oldest survey is the National Health Interview Surveys (NHIS) which was started in 1957. NHIS has been the model for all other surveys.

Table 1 Types of Surveys at NCHS

<table>
<thead>
<tr>
<th>A. INTERVIEW</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. National Health Interview survey (NHIS)</td>
</tr>
<tr>
<td>2. National Family Growth Survey (NFGS)</td>
</tr>
<tr>
<td>3. National Medical Care Utilization and Expenditure Survey (NMHCUES)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>B. EXAMINATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>4. National Health and Nutrition Examination Survey (NHANES)</td>
</tr>
<tr>
<td>5. National Hispanic Health and Nutrition Examination Survey (NHANES)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>C. RECORD</th>
</tr>
</thead>
<tbody>
<tr>
<td>6. National Hospital Discharge Survey (NHDS)</td>
</tr>
<tr>
<td>7. National Ambulatory Medical Care Survey(NAMCS)</td>
</tr>
<tr>
<td>8. National Natality Survey (NNS) AND National Fetal Mortality Survey (NFMS)</td>
</tr>
<tr>
<td>9. National Nursing Home Survey (NNHS)</td>
</tr>
<tr>
<td>10 National Mortality Follow Back Survey (NMFS)</td>
</tr>
<tr>
<td>11 National Maternal and Infant Health Survey (NMIHS)</td>
</tr>
</tbody>
</table>

Most NCHS surveys are multistage designs; often the sampling plan includes selecting some sampling units with probability proportional to the size (pps). Such a sampling plan provides self-weighting samples from which we can obtain estimates with minimum variance among all others. However, sometimes certain subpopulations like blacks and poor people, were sampled more than others to allow for meaningful analysis of their data.

The sample estimation of population parameters is usually done by basic weighting and ratio adjustment. Variances are usually estimated by BRR methods and presented U.S. life table. A major change of new design is the introduction of four independent panels of approximately equal size, each representing the U.S. population. In case of a budget cut, one or
more panels may be dropped without damaging the sampling efficiency. Another point is the sampling of two PSUs from NSR strata, which makes it possible to derive variance between PSUs. The Table 3 highlights some of the changes made for the new design.

1.2 Estimation: As the sampling is basically pps design until 1985, in theory the probability of selection is more or less same for all sample persons (Kendall and Stuart, vol3, 1977, p 199).

1.2.1 Basic Weighting: The indexes l, I, h, j, and k are for Strata 1 = 1,..., L,
PSUs 1 = 1,..., m l,
substrata h = 1, 2, 3. (used after 1985)
segments j = 1,..., m l,
and persons k = 1,..., D l i h j.
Denote \( w_{l i h j} = M_{l i} / (m_{l i} m_{l i h}) \) for all k.
The basic population total with characteristic x before 1985 is given by
\[
X' = \sum_{l} \sum_{i} \sum_{h} \sum_{j} \sum_{k} \frac{w_{l i h j}}{W_{l i h j}} x_{l i h j} \tag{1.1}
\]
where \( x_{l i h j} = 1 \) if the \((l i h j)-th\) sample person in the sample has x characteristic, and \( = 0 \) otherwise. With differential sampling rate in the PSU after 1985, \( X' \) is
\[
X' = \sum_{l} \sum_{i} \sum_{h} \sum_{j} \sum_{k} \frac{w_{l i h j}}{W_{l i h j}} x_{l i h j} \tag{1.2}
\]
where \( w_{l i h j} = 1/(m_{l i h} f_{l i h}) \), and similarly for \( x_{l i h j} \).

1.2.2 Poststratified Ratio Estimation:
The first stage adjustment is for the 48 cells of color-residence, indexed by c = 1, ..., C
(C = 48), done only for the NSR PSUs. The second stage ratio adjustment is for 60 cells of age-sex-race categories, indexed by a = 1, ..., A (A = 60). The age-sex-race ratio adjustment is
\[
X' = \sum_{a} \frac{x_a}{y_a} \hat{y}_a \tag{1.3}
\]
where \( x_a \) and \( y_a \) are so defined as the numerator and denominator of next equation (1.4) under the summation sign. The color-residency ratio adjustment done as
\[
X' = \sum_{a} \frac{x_{aSR}}{y_{aSR}} \frac{Z_{CCEN}}{Z_{CCEN}} \tag{1.4}
\]
where \( x_{aSR} = \sum_{l} \sum_{i} \sum_{h} \sum_{j} \sum_{k} \frac{w_{l i h j}}{W_{l i h j}} x_{l i h j} \), and \( y_{aSR} = \sum_{l} \sum_{i} \sum_{h} \sum_{j} \sum_{k} \frac{w_{l i h j}}{W_{l i h j}} y_{l i h j} \) for all k.

1.3 \( \hat{v}(X') \): After 1973, the combination of Balanced Repeated Replication (BRR) and Generalized Variance Function (GVF) are used in order to present relative standard errors.
Since 1985, the same GVF has also been used to present sample variances, but the sample variances are estimated differently. The ratio estimate \( X' \) is linearized, and then the usual formula for two stage sampling variance formula is applied to these linearized variables instead. The variances so derived are now used to estimate the parameters, a and b of GVF curve. The GVF is also used, but instead of the curve, the estimator of parameters a and b are given so that one can easily calculate the relative standard error of x by \((a + b/x)^{1/2}\).

1.3.1 Variance by BRR: 149 pseudo-strata are used to generate 152 balanced half sample replications in order to estimate the variance. NHIS classified estimates into about 75 domains. Each domain uses about 100 points of varying sizes to produce the GVF curve of each domain.

1.3.2 Variance for two stage sampling: The sampling is done only in the i-th and j-th levels, the l-th level indicates the strata, the h-th level indicates different sampling rate, and the k-th level indicates the units in the segment.
The variance is now obtained by two steps. The first step is to find the variance formula of two stage sampling. The second step is to find the linear form of two stage ratio estimation.
We are interested in the variance of the basic estimate, which is the form of Brewer or Durbin estimator, and it was extended to two stage sampling as shown in Cochran (1977, p308, 11.4).
Since strata are independent, first we find the variance of each stratum and sum the variances for all the strata.
where \( t_{ihjk} = 1 \) if the \((ihjk)\) person has the \(t\) characteristic, and \(0\) otherwise. The variance and its estimator are shown in (11.42) and (11.44) (Cochran, 1977), respectively, when two PSUs are sampled in the \(i\)th level.

When units were selected with unequal probabilities with replacement. Unbiased estimator of the population total of pps sample, its variance, and the unbiased estimator of the variance are also shown in Cochran (1977, 11.31, 11.32, and 11.35).

\[
\hat{\text{var}}(t) = \sum_{i=1}^{l} \left( \frac{\pi_{11}}{\pi_{12}} - \frac{\pi_{12}}{\pi_{11}} \right) \left( t_{11} - t_{12} \right)^2
\]

(1.9)

where \( s_{i1h}^2 \) is the usual sample estimator of \( S_{i1h}^2 \). \( \pi_{11}, \pi_{12} \) and \( \pi_{112} \) is also known (Cochran, 1977, p 308, 11.44).

### 1.3.3 Linearization of \( X' \): (Woodruff, 1971)

The first step is the linearization of \( (x'_a / y'_a) \) \( y_a \) as

\[
X' = \sum_a y_a \left( x'_a - \frac{x_a}{y_a} y'_a \right),
\]

(1.10)

where \( y_a = \bar{y}_a / Y_a \), \( X_a = E(x'_a) \), \( Y_a = E(y'_a) \).

Replacing \( x'_a \) and \( y'_a \) the numerator and denominator of (1.4), (1.10) can be written as

\[
X' = A \sum_a y_a \left( x_{aSR} - \frac{\sum_c z_{acNSR} z_{cCCEN}}{z_c} \right)
\]

(1.11)

Second linearization under the summation over \( c \) is

\[
X' = A \sum_a y_a \left( x_{aSR} - \frac{x_a}{y_a} y_{aSR} \right) + \sum_a \frac{C}{y_a} \sum_c z_{cCCEN}
\]

(1.12)

where \( \bar{Z}_c = E(z_c), \bar{Y}_{acNSR} = E(y_{acNSR}), \bar{X}_a = E(x_{acNSR}) \).

Replacing the five variables \( x_{aSR}, y_{aSR}, \) and \( Z_c \) with (1.4), (1.5), and (1.6), and \( Y_{acNSR} \) and \( X_{acNSR} \) with similar definitions, and exchanging the summation signs of the levels \( l, i, h, j, \) and \( k \) with the summation signs of \( a \) and \( c \), the equation (1.12) can be expressed as

\[
X' = \sum_a \left( \sum_{h} \left( \sum_{j} \left( \sum_{k} \sum_{l} \frac{1}{\pi_{11}} m_{i1h} t_{ilhjk} \right) \right) \right)
\]

(1.13)

for the NSR-PSUs and

\[
X' = \sum_a \left( \sum_{h} \left( \sum_{j} \left( \sum_{k} \sum_{l} \frac{1}{\pi_{11}} m_{i1h} t_{ilhjk} \right) \right) \right)
\]

(1.14)

for self-representing PSUs.

Since the variables \( t_{ilhjk} \) of NSR-PSUs in the equation (1.14), are the weighted variables, which included the weights \( M_{ih} / (\pi_{11} m_{i1h}) \), the variance formula (1.9) can be used by replacing \( X_{ilhjk} / \pi_{11} \) with \( t_{ilhjk} \) in the first term, and

\[
M_{i1h}^2 / (\pi_{11} m_{i1h}) \) with \( \pi_{11} m_{i1h} \) in the second term.

Similarly the \( t_{ilhjk} \) of SR-PSUs of the first term in (1.14), as the sampling is done only in the \( j \)-th level in the SR-PSUs, the second term of the variance formula (1.9) can be adjusted, dropping
the first term.

New software is developed Research Triangle Institute for the two stage sampling when simple random sampling is used within each stage sampling. They also show how to use this formula for the pps sampling case.

Thus, the total variance is estimated by

$$\text{var}(X') = s^2_{SR} + s^2_{NSR}.$$  \hfill (1.15)

The use of this method instead of BRR method may reflect the actual survey design better.

2. NATIONAL FAMILY GROWTH SURVEY (NFGS)

There have been four cycles of NFGS survey since NFGS was established in 1972. The first cycle of survey was conducted in 1973, the second cycle in 1976, the third in 1982, and the fourth in 1987.

In the cycle 1, the sample plan is based on a five-stage selection of 203 PSUs, 2 replicate groups of PSUs out of 4, EDs, segments, dwelling units, and one eligible person.

The resulting data are stratified into the 12 age-race classes, and ratio-adjusted. Nonresponses were also ratio-adjusted corresponding to each stage. The weights of the inverse of selection probability in each stage were also multiplied to get the weights for each individual.

Since it may be impossible to find exact formula of the variance of an estimate from this sample, the balanced half sample replication is used to find variances under certain assumptions.

The 103 FGS PSUs were grouped into 48 strata, seven of self-representing and 41 nonself-representing. The self-representing PSU's were randomly divided into two pseudo PSUs, and the nonself-representing PSUs are paired. Taking one member of the pair, a 48 x 48 replication matrix was formed. Following the usual procedures, the variance of an estimate was obtained. The details are previously given in NHIS section.

In the cycle 2, the sampling is a five stage selection of PSUs, EDs, segments, households, and finally women of 15-44 years. Each sample person is inflated by the inverse of the selection probabilities in each of the five stages. The basic weight was adjusted by multiplying nonresponses at the segment level, and for the 12 age-race ratios only of ever married women.

The 79 cycle 2 PSUs were regrouped into 37 super-strata, 18 self-representing strata and 19 nonself-representing strata. The 18 self-representing strata were randomly divided into two pseudo-strata, and two PSUs were selected from each of nonself-representing superstrata. Taking one of the pair, the 37 strata gives 40 replications for the application of BRR method.

In the cycle 3, the sample design is similar to that of the cycle 2. It also uses a five stage probability sampling that incorporated oversamples of black and teenage women, but the cycle 3 takes a supplementary sample of women living in college dormitories and sororities.

Cycle 4 is entirely different from the previous designs, and linked to the NHIS sample persons of the previous years. It uses the 4th quarter sample of 1986 (3 panels system), from which the national sample of black women 15-44 years was obtained, all 4 quarters of 1986 (3 panels system), from which the national sample of women 15-44 years was taken; the 1st quarter of 1987 (4 panels system), from which the national sample of black women 15-44 years was used. At most one woman was sampled from each household. The sample size was 10,694 women of which about 1/3 were black. The estimation, variance, and other details are not yet available.

3. NATIONAL MEDICAL CARE UTILIZATION AND EXPENDITURE SURVEY (NMUCES)

3.1 Sampling: NMUCES was designed to collect data about the U.S. civilian noninstitutionalized population during 1980. During the 1980, information was obtained on health, access to and use of medical services, associated charges and sources of payment, and health insurance coverage.

Two independent nationwide samples were pooled together. One was the sample taken by Research Triangle Institute (RTI), and the other by National Opinion Research Center (NORC). Both are characterized as stratified, five-stage area probability designs. The differences between two designs are mainly the type of stratification variables and the specific definitions of sampling units.

3.2 Estimation: The basic method used in NHIS estimation was applied to the NMUCES sample data. The details are presented in other publication.

3.3 Variance: The balanced half sample replication was used to find the variance of the sample estimates.

4 NATIONAL HEALTH AND NUTRITION EXAMINATION SURVEY

There have been six surveys since 1959. The first three of them (NHANES I, II, and III) are the health examination surveys, the two of them (NHANES I and II) are the health and nutrition examination surveys, and one of them is the detailed examination survey (NHANES IA). These are:


5. National Health and Nutrition Examination Survey (HANES II). 1974 - 1975 (14 months). Persons 25 - 74 years. To augment the size of the sample originally included in the sample of NHANES I, an additional sample was drawn from the same age group for detailed examinations.


4.1 Sampling: All these surveys are using the similar plan of multistage probability sampling design. Minor modifications of the self-weighting features were introduced to handle particular situations. The last three surveys took more sample from the poor, and young and older age groups. The probability of inclusion of every sample person was known over each stage of
4.2 Estimation: The subscripts i, j, k, t, - and 1 are
for the PSU (i = 1, ..., n),
for ED (j = 1, ..., Jij),
for the segment (k = 1, ..., Kij)
for poor (t = 1) and for nonpoor (t = 2),
for the person ( l = 1)

Define \( \pi_{ijkt} = W_{ijkt} \) and \( Y_{1ijkt} = 1 \)
if the \( ijkt \)-th person included in the sample
and = 0 otherwise. \( \pi_{ijkt} \) is the probability
of the \( ijkt \)-th sample unit being in the sample.
If this person belongs to the \( a \)-th age-cell,
and the poststratified ratio of this cell is
\[ Y_a / \hat{Y}_a \]
\( Y_a \) is the census cell counts of \( a \)-th age-sex
and \( \hat{Y}_a = \sum_i \sum_j \sum_k \sum_t \sum_l \pi_{ijkt} Y_{1ijkt} \).
The final weight of this person is adjusted as
\[ w_{ijkt} Y_a / \hat{Y}_a. \]

The final estimator of the population total is
\[ \hat{Y} = \sum_i \sum_j \sum_k \sum_t \sum_l \sum_a w_{ijkt} Y_a / \hat{Y}_a. \]

Actually \( w_{ijkt} \) is the inverses of the selection probabilities. Note that,
if it is a self-representing PSU, \( w_{ijkt} \) do
not include the inverse of the probability of the first stage selection for \( \pi_{i} = 1 \).

Define \( x_{ijkt} = 1 \) if the \( ijkt \)-th sample
person has \( x \) characteristic, and = 0 otherwise.

4.3 Variance: Table 5 shows the method used
for variance estimations.

Table 5 Variance Estimation for 6 Surveys.

<table>
<thead>
<tr>
<th>Survey</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>HES I</td>
<td>BRR (19 pseudo-strata x 20 repl.)</td>
</tr>
<tr>
<td>HES II</td>
<td>BRR (same)</td>
</tr>
<tr>
<td>HES III</td>
<td>BRR (same)</td>
</tr>
<tr>
<td>HANES I</td>
<td>BRR (40 pseudo-strata x 40 repl.)</td>
</tr>
<tr>
<td>HANES IA</td>
<td>BRR (40 pseudo-strata x 40 repl.)</td>
</tr>
<tr>
<td>1-65 stds</td>
<td>TSR (35 pseudo-strata x 81 repl.)</td>
</tr>
<tr>
<td>66-100 stds</td>
<td>BRR (19 pseudo-strata x 20 repl.)</td>
</tr>
<tr>
<td>1-100 stds</td>
<td>TSR (35 pseudo-strata x 81 repl.)</td>
</tr>
</tbody>
</table>

All the surveys used the BRR except the two
cases, that is, third sample replications
(Bryant, 1975), using 100 stands (or strata) in
HANES IA and SESUDDAN in HANES II. The formula (1.9)
presented in Section 1 can be adapted to this
situation.
psu's out of 1924 PSU's from 3141 counties, county equivalents and independent cities of the United States, Second systematic selection of office based physicians from the list of MD's from the selected PSU's such that we obtain a self-weighting sample of 4,681 physicians. They were randomly assigned to one of the 52 week periods. Third systematic selection of only ten visits each day, giving about 30 patient records during the week in each office sampled physicians.

The basic weights were obtained by the inverse of the selection probability of individual sample units. These basic weights were further adjusted by multistage post-stratified ratios to estimate the total of the target population.

The variables most frequently used in the report were general-family-practitioner, internal medicine, pediatrics, general surgical, ob-gyn, psychiatry on one side of two way table, and age, sex, and race on the other side in the publication. Sample variance was estimated by BRR method with 48 balanced half-samples from 48 pseudo-strata.

8. THE NATIONAL NATALITY SURVEY (NNS) AND THE 1980 NATIONAL FETAL MORTALITY SURVEY (NFMS)

The NNSs were conducted by NCHS based on live births in 1963, 1964-66, 1967-69, 1972, and 1980. The 1980 NNS is based on the probability sample of 9,941 live birth certificates that occurred in the U.S.A. during 1980. The files of birth certificates in the 50 states, the district of columbia, and the independent registration area of New York City constituted the target areas. The certificates were registered for an estimated 99.3 percent of all live births.

In each area, a sequential file number is assigned to each birth certificate received from the beginning to the end of each year. 105 of every 1,000 certificates were randomly selected starting from the most recent month completed. Next 25 of every 105 certificates were randomly selected. From the remaining 80, only those less than 2,500 grams were taken as sample. Finally, it was necessary to exclude from the sample some additional certificates in the State of Washington where it was required to obtain the permission of married mothers, and Idaho where it was needed to get the consent of unmarried mothers.

The total of national live births was the weighted sum of the post-stratified ratios of the number of births to U.S. residents in 1980 to the number of sample births in NNS according to the 50 cells of birth weights, marital status, race, and ages.

The variance was estimated by the BRR method from the 20 half sample replicates, based 20 strata. The relative standard error was equated to the square root of \((A + B/x)\), where \(A\) and \(B\) are the weighted least square estimators.

REFERENCES


Cypress B.K. (1983). Patterns of Ambulatory Care in General and Family Practice: The National Ambulatory Medical Care Survey, NCHS Series 13, No. 73


Acknowledgement

Author thanks to Dr Lester R. Curtin whose comments made this paper more readable.