TWO EMPIRICAL STUDIES OF STATISTICAL METHODS APPLIED TO DATA FROM COMPLEX SURVEYS

Myron J. Katzoff, Gretchen K. Jones, Lester R. Curtin
National Center for Health Statistics
Barry Graubard, National Institute of Health

1. INTRODUCTION

This paper is to serve these purposes:

(1) to report on our progress to date in developing software for empirically studying the effects of complex survey designs on analyses of health survey data;

(2) to present some results from a first empirical study, in which we examine the properties of the BRR and Jackknife variance estimators; and

(3) to outline our approach to a second empirical study involving the comparison of vector means for subdomains.

In our earlier paper (Katzoff, et al 1988) we listed six major survey-design and analysis tasks that we shall expect the software we develop to perform:

(i) construction of primary sampling unit (PSU) strata;
(ii) selection of PSUs;
(iii) allocation of sample size;
(iv) generation of sample weights and first order estimates;
(v) computation of variances and their sample-based estimates; and
(vi) creation of empirical distributions of estimates, especially, estimates of variances, test criteria or other statistics of interest.

At the time of our previous paper, we had completed work on one set of options for tasks (i), (ii) and (iii) only. Much work to embed more investigative options in those portions of the software remains. (For example, we expect that further development of the software will provide for the examination of the consequences of applying a probability minimum replacement sampling technique (Chromy, 1986) and the unequal probability sampling scheme of Saxena, Singh and Srivastava (1988). However, we have given highest priority to developing at least one software option for (iv), (v) and (vi) because we felt that it was necessary to demonstrate our approach while interest remains high, to further the identification of problem areas and to permit development of software options in parallel with the conduct of some empirical studies.

2. DESIGN OF GENERAL SYSTEM FOR EMPIRICAL EVALUATIONS

To reflect the structure of current survey designs, we created a first design with a large number of strata consisting of individual counties, groups of contiguous counties and large metropolitan areas (SMSAs). Since these geographical units are chosen at the first stage of sample selection, we refer to them as primary sampling units (PSUs). Some PSUs are taken with probability one; they are called self-representing (SR) PSUs. PSUs chosen with probability less than one are called non-self-representing (NSR) PSUs. NSR PSUs have been aggregated into strata from each of which two PSUs are drawn for each simulation.

Before addressing the problems of computing sample weights and first order estimates, we considered addition of an option for a self-weighting design. For initial empirical work, we had originally only allowed for allocation of the sample in proportion to the PSU populations. For strata where PSUs were not chosen with certainty, this would result in a situation where the size of the sample drawn from a NSR PSU would depend upon which other PSU was included in the sample. It was easy to add this option and we felt it was important to do so since this would permit us comparisons with results from a common survey design practice. In general, we expect that sample allocation methodology for multistage designs for situations where the allocation of sample depends upon from which clusters the observational units are to be drawn will recieve more attention in the future.

The sample weights are, of course, the inverses of the total probabilities of inclusion for each unit. They reflect a decision that, at the second stage of selection, clusters of housing units (called segments) were chosen with equal probabilities. To provide for the possibility of investigating some item and unit nonresponse patterns which depend upon membership in various subpopulations, sample weights are assigned to individual records.

Random Selection of PSUs. In the strata of NSR PSUs, the PSUs have been chosen with probabilities proportional to the amount of population in those geographical units according to Brewer's method (Brewer, 1985) as recommended by Pfefferman (1986).

In using Brewer's procedure the NSR PSU strata have been defined and the proportion of the stratum population in the I-th PSU, P_I, has been calculated for each value of the index I. Let the values of P_I be stored in a vector VECP of dimension D, equal to the number of NSR PSUs in the stratum.

These specifications result in a selection of two PSUs per stratum. The operations given in the next section must be repeated for each stratum:

(1) Calculate the entries of a vector VECP1, of the same dimension as VECP, as follows:

a. for each value of I=1,2,...,D, compute $\text{VECP1(I)}=P_I(1-P_I)/1-2P_I$ (Note that P_I=VECP(I).)

b. calculate S, the sum of the coordinates of VECP1.

c. normalize the values of VECP1(I) by dividing each value by S; in other words, replace the value of VECP1(I) computed (1) a with VECP1(I)/S. (2) Select a random number, R1, between zero and one and choose unit I if

I is the smallest index for which R1≤VECP1(1)+VECP1(2)+....+VECP1(I). Record and store the index of the unit selected, IO say.

(3) Calculate the entries of a vector VECP2 also of dimension D as follows:

a. VECP2(I)=VECP(I) if I is not equal to IO; when I=IO, VECP2(I)=0.

b. normalize the values of VECP2(I) by dividing each by 1-VECP(IO).

(4) Select a random number, R2, between zero and one and choose unit J if J is the smallest index for which R2≤ VECP2(1)+VECP2(2)+...+VECP2(J). Record and store the index of this unit, the second unit selected as the value of JO.

(5) Compute the joint inclusion probability of units IO and JO as JOINT=VECP1(IO)*VECP2(JO) + VECP1(JO)*RATIO where RATIO=VECP(IO)/{1-VECP(JO)}.

Allocation of Sample and Selection of Segments Under Proportional Allocation to PSUs by Size. This section is concerned with the process of distributing the sample across strata, PSUs, and segments. For our purposes, a sample size of 5,000 was considered adequate.

(1) Begin by determining a sample size for each stratum according to this formula:

$$n(s) = \text{sample size for stratum } s = \frac{(\text{population of stratum}) \times 5000}{\text{universe population}}$$

(2) Next, proportionately allocate the sample size to the two PSUs selected for stratum s; that is, compute

$$n(1,s)=[P_1/(P_1+P_2)] \times n(s),$$

the sample size for the first PSU selected from stratum s, and

$$n(2,s)=[P_2/(P_1+P_2)] \times n(s),$$

the sample size for the second PSU selected from stratum s, where P_1 and P_2 are the population sizes for the first and second PSUs respectively.

(3) Finally, for i=1,2, determine the number of segments to be selected from each PSU from this expression:

$$n'(i,s)= [P_i/g_i(s)] \times n(i,s)$$

with n'(i,s) rounded to the nearest integer, where $g_i(s)$ is the number of segments in the i-th PSU chosen from stratum s.

[When $g_i(s)$ is a constant for all i and s]. It can be seen that the above allocation scheme produces sample weights which are proportional to the inverses of the relative sizes of the PSUs chosen in each stratum only. In a self weighting scheme, where the sample weights would be constant throughout, n(i,s)=n(s)/2 if the constant is just the inverse of the sample size divided by the universe size.

Self-weighting selection scheme. The purpose of this addition is to provide the user with the option of examining a self-weighting design. Using the earlier notation for SR PSUs n'(s) is unchanged. For NSR PSUs, n(i,s) = n(s)/2 and $n'(i,s)=[g_i(s)/P_i] \times n(i,s)$. Here we use the earlier definitions: (1) $g_i(s)$ equals the number of segments in the i-th PSU; and (2) P_i is the population size for the i-th PSU. Observe that since n'(i,s) here is not a function of which pair of PSUs were selected at the first stage, you should always get the same value for n'(i,s) for the i-th PSU. Therefore, the bias-correction term discussed below should equal zero for unbiased variance estimates when the self-weighting design option is selected.

Selection of segments. The procedure for selecting segments is method 1 for sequentially selecting a SRS (Sunter, 1977). The procedure is to be applied independently to each PSU.

Let M denote the number of segments in a PSU and let m be the number of segments to be selected at random (i.e., m=n'(i,s) for some i and s, in the previous notation). For clarity suppose that the M segments from which we are to draw a sample have been serially assigned index values 1,2,...,M. (In practice this is not necessary of course.) Select the first unit in the PSU with probability m/M and, thereafter, select unit j with probability

$$(m-m_j)/(M-j+1),$$

where m_j denotes the number of units selected prior to consideration of the j-th unit. The process moves sequentially from one unit to the next and terminates as soon as m_j=m.

This sequential procedure can be implemented with one pass of the file and without use of lots of computer storage. This is important in any large scale simulation.

3. VARIANCE ESTIMATION PROCEDURES

One standard output of the simulation software is linear estimates of finite population parameters: weighted sums of the sample values of the variables measured. Estimators of population parameters that can be expressed as differentiable functions of the linear estimates (for example, ratio estimates and subdomain means) can use these simulation outputs for studies of such estimators. Furthermore, since another standard simulation output is a covariance matrix for the linear estimates, in most cases of interest it should be easy to compile variance estimates with the help of Taylor series linearization.

Design based variance. Under the design currently programmed, the textbook formula (Cochran, 1987 p. 301-2) for the NSR component of total variance from NSR-stratum s when the sample sizes are nonrandom is given by

$$V_{1s}= \frac{\pi_{1s}\pi_{2s} - \pi_{12,s}}{\pi_{12,s}} (\tilde{Y}_{1s}-\tilde{Y}_{2s})(\tilde{Y}_{1s}-\tilde{Y}_{2s})' + A$$

where

$$A = \sum_{i=1,2} \frac{n'(i,s)\pi_{is}[N'(i,s)-n'(i,s)]}{N'(i,s)\ [n'(i,s)-1]} B_i$$

and

$$B_i= \sum_{k=1}^{n'(i,s)} (\tilde{y}_{iks}-\tilde{Y}_{is}/n'(i,s))\ (\tilde{y}_{iks}-\tilde{Y}_{is}/n'(i,s))'$$

Here, for sample PSU indices i=1,2 , $\bar{y}_{iks}$ is the weighted sum of the sample values for sample segment k of sample PSU i drawn from NSR stratum s and

$$\tilde{Y}_{is}=\Sigma_{\{1\leq k\leq n'(i,s)\}}\tilde{y}_{iks}.$$

$\pi_{is}=2P_{is}$, where P_{is} is the proportion of stratum s population that belongs to sample PSU i drawn from stratum s

$$\pi_{ij,s} = \frac{\pi_{is}\pi_{js}\{(1-\pi_{is})^{-1} + (1-\pi_{js})^{-1}\}}{\sum_{r=1}^{N_s} [\pi_{rs}(2-\pi_{rs})/(1-\pi_{rs})]}.$$

When the sample sizes at the second stage depend upon which PSUs are included in the sample, this bias correction term must be added to V_{1s} to obtain unbiased variance estimates

$$\sum_{i=1,2} \frac{[n'(i,s)]^2}{[n'(i,s)-1]} (2-\pi_{is})(\zeta_{is}-\eta_{is})C$$

where

$$C = \sum_{k=1}^{n'(i,s)} (\tilde{y}_{iks}-\tilde{Y}_{is}/n'(i,s))\cdot(\tilde{y}_{iks}-\tilde{Y}_{is}/n'(i,s))'$$

where if $A(i)=\{1,2,\ldots N_s\}\backslash\{i\}$ and N_s denotes the number of PSUs in stratum s,

$$\zeta_i = \sum_{j \text{ in } A(i)} (1/n_{is}\{i,j\}) \frac{\pi_{ij,s}}{\pi_{is}}$$

and

$$\eta_{is} = \sum_{j \text{ in } A(i)} (1/n_{is}\{i,j\}) \frac{\pi_{js}}{2-\pi_{is}}$$

In the last two formulas $n_{is}\{i,j\}$ denotes the size of the sample drawn from PSU i of stratum s if the sample contains PSUs i and j. π_{ij} and π_{js} are as defined above as is $\pi_{ij,s}$. The component of total variance that arises from the SR PSUs is the sum over s, as s runs over the SR-strata, of

$$V_{2s} = D \sum_{k=1}^{n'(s)} (\tilde{y}_{ks}-\tilde{Y}_s/n'(s))\ (\tilde{y}_{ks}-\tilde{Y}_s/n'(s))'.$$

where

$$D = \frac{n'(s)[N'(s)-n'(s)]}{N'(s)[n'(s)-1]}$$

$n'(s)$ is the number of segments drawn from the s-th SR stratum and $N'(s)$ is the total number of segments contained by that stratum. $\tilde{y}_{ks}$ and $\tilde{Y}_s$ are self-representing analogues of $\tilde{y}_{iks}$ and $\tilde{Y}_{is}$ -- the subscript i is unnecessary for the SR PSUs because they are taken with certainty and are, therefore, strata themselves.

In some empirical work it may be necessary to have the covariance matrix of the estimators and not just an estimate of that covariance matrix. The covariance matrix of the estimators is

$$V_0 = \sum_{\substack{s \text{ in NSR} \\ \text{strata}}} Q_{1s} + \sum_{\substack{s \text{ in SR} \\ \text{strata}}} Q_{2s}$$

where, dropping the s subscript to simplify the notation,

$$Q_1 = \sum_{i=1}^{N} \pi_i^{-1} (N_i')^2 (\zeta_i - 1/N_i')S_i^2 \ +$$

$$\sum_{i=1}^{N}\sum_{j>i}^{N} (\pi_i\pi_j - \pi_{ij})(\pi_i^{-1}Y_i - \pi_j^{-1}Y_j) \cdot(\pi_i^{-1}Y_i - \pi_j^{-1}Y_j)'$$

and

$$Q_2 = N'(N'-n')S^2/(n')$$

where N is the number of PSUs in the NSR stratum

$\pi_i = 2P_i$, where (as previously described) P_i is the PSU population expressed as a fraction of total stratum population

N' or N_i' is the number of segments in an SR-PSU or the i-th NSR-PSU of an NSR-stratum respectively

n' is the sample size allocated to an SR stratum and

$$S^2 \text{ or } S_i^2 = (N_i'-1)^{-1} \sum_{k=1}^{N_i} (y_{ik}-Y_i/N_i')(y_{ik}-Y_i/N_i')'$$

for y_{ik} equal to the k-th segment total and

$$Y_i = \sum_{k=1}^{N_i'} y_{ik}$$

BRR and Jackknife Variance Estimates. The purpose of this first empirical study is to compare the distributional properties of the BRR and Jackknife estimators with those of the design-based unbiased estimates of variance. Our initial study of estimators of variance examined those for linear estimates of finite population totals. In so doing we hoped to avoid the confusion that can result from trying to deal with a lot of complexity in the very beginning. The variability and bias of the different variance estimators are of special interest to us.

Because the variance estimators are quadratic forms, we thought it would be useful to summarize the numerical results in two ways instead of just one: (1) computation of the first few central moments of the empirical distributions; and (2) finding a Satterthwaite approximation to a multiple of a chi-squared variable.

In producing the BRR and Jackknife estimates, we largely followed Rust (1985). To meet the requirements for the use of these procedures, the segments of the SR-PSUs were randomly assigned at each iteration of the simulation to one of two groups and the weighted sums of the variables for each group were formed. Similarly, for each NSR-PSU, a weighted sum of variables was calculated. For each stratum, s, we denote the sum for group or PSU i by $\tilde{Y}_{is}$. (Note that for NSR strata, $\tilde{Y}_{is}$ is as previously defined.)

There are several ways of implementing the BRR variance estimation procedure. We now describe the approach we used:

(1) compute the half-sample estimators of totals

$$\tilde{Y}_\alpha = \sum_{s=1}^{60} (2\delta_{\alpha 1,s}\ \tilde{Y}_{1s} + 2\delta_{\alpha 2,s}\ \tilde{Y}_{2s})$$

where $\delta_{\alpha 2,s}=1-\delta_{\alpha 1,s}$ and

$$\delta_{\alpha 1,s} = \begin{cases} 1, \text{ if PSU or group 1 of stratum s chosen} \\ \quad \text{for the half-sample} \\ 0, \text{ otherwise} \end{cases}$$

(2) put

$$\tilde{Y} = \sum_{s=1}^{60} (\tilde{Y}_{1s}+\tilde{Y}_{2s})$$

and calculate the variance estimate

$$v_{BRR}(\tilde{Y}) = \frac{1}{60} \sum_{\alpha=1}^{60} (\tilde{Y}_\alpha - \tilde{Y})^2$$

For the jackknife estimators we define jackknife totals

$$\tilde{Y}_{(is)} = \sum_{u\neq s}^{60} (\tilde{Y}_{1u}+\tilde{Y}_{2u}) + 2\tilde{Y}_{js}$$

for j not equal to i; i,j=1,2; and s=1,2,...,60. The jackknife estimate of variance is then

$$v_J(\tilde{Y}) = \frac{1}{2} \sum_{s=1}^{60}\sum_{i=1}^{2} (\tilde{Y}_{(is)} - \tilde{Y})^2$$

HISTOGRAMS

Histograms of estimates of totals and jackknife estimates of their variances are presented for short stay hospital days for nonwhite and white persons in the age group 46-57. These histograms represent what one might get from approximately 1000 independent randomly selected samples from the universe previously described if they are generated according to the survey design discussed earlier; for each sample so generated, an estimate of total short stay hospital days and the estimate of its variance were calculated. The estimates of total days for white persons formed a histogram that is clearly less skewed than that for nonwhite persons. Since the portions of samples drawn from the nonwhite subdomain can be too small for central effects to show up it is not hard to understand the apparent lack of symmetry in the histogram for that group. Whether there is sufficient symmetry in the histogram of estimates of total days for white persons is open to question; a nonparametric test for distributional shape (for example, a Lillifors test) might be useful in this case. The histograms for the jackknife variance estimates for white and nonwhite persons are not very different and are strongly positively skewed. The need for comparisons with appropriate parametric distribution models is clearly indicated.

4. COMPARISON OF SUBDOMAIN MEANS

As an illustrative case, we anticipate undertaking a detailed study of the comparison of vector means for short-stay hospital days in accordance with the age breakdown: 17-25, 26-32, 33-45, 46-58 and 58+, for white and nonwhite groups. The discussion of this section describes how we plan to use the basic linear estimates and their estimated and expected variances and covariances in that study. For now there is a limit on the dimensionality or size of vectors that can be considered in multivariate comparisons. This is due to our limiting the number of variables for which we can compute a variance-covariance matrix to twenty. However, we may later endeavor to circumvent this limitation somehow by using the observation that if $\mathbf{X}$ and $\mathbf{Y}$ are vector random variables, all the necessary variances and covariances for $\mathbf{Z}=(\mathbf{X}',\mathbf{Y}')'$ can be obtained from suitable operations with $\mathbf{Var(X)}$, $\mathbf{Var(Y)}$ and $\mathbf{Var(X+Y)}$.

In what follows we assume that algorithms for calculating the basic (or linear) estimates and their variances and covariances have been developed and are available for use without revision. The outputs of those algorithms are essential inputs for operations described here. In particular, it is assumed that we can now quantify the vector of basic estimates

$$E'=(W_1,\dots,W_5,PW_1,\dots,PW_5,O_1,\dots,O_5,PO_1,\dots,PO_5)$$

and their covariance matrix, $\mathbf{V}$, where for i=1,2,...,5, W_i denotes the estimated number of short stay hospital days for white persons in age category i; PW_i denotes the number of white persons in age category i; O_i denotes the estimated number of short stay hospital days for nonwhite persons in age category i; and PO_i denotes the estimated number of nonwhite persons in age category i. We will have need of the 20x5 "multiplier" matrix for the estimates, $\mathbf{M}$, defined as four stacked diagonal matrices

$$D[(PW_1)^{-1},(PW_2)^{-1},\dots,(PW_5)^{-1}]$$

$$D[-W_1(PW_1)^{-2},\dots,-W_5(PW_5)^{-2}]$$

$$D[(PO_1)^{-1},\dots,(PO_5)^{-1}]$$

$$D[-O_1(PO_1)^{-2},\dots,-O_5(PO_5)^{-2}]$$

to make use of the Taylor series linearization method. We will also have need of a second multiplier matrix $\mathbf{M}_0$, which has the same stacked-diagonal matrix structure as $\mathbf{M}$ but in which the estimates have been replaced with population values. It is important to note that $\mathbf{M}$ is calculated at each iterate of the simulation whereas $\mathbf{M}_0$ is calculated only once for a simulation. To assess our choices for nominal asymptotic significance levels, we need a noncentrality parameter which is to be computed only once for the simulation, which, in turn, entails additional one-time computations described here. For this purpose, define and quantify:

$$\mu_W'=(W_{10}/PW_{10},W_{20}/PW_{20},\dots,W_{50}/PW_{50})$$

$$\mu_O'=(O_{10}/PO_{10},\dots,O_{50}/PO_{50})$$

where the additional zero (second) subscript indicates a population value not an estimate or sample-based value. The covariance matrix of the estimators is calculated according to the procedure for obtaining $\mathbf{V}_0$ given earlier.

With these definitions the noncentrality parameter is

$$\delta^2=(\mu_W-\mu_O)'(\mathbf{M}_0{}'\mathbf{V}_0\mathbf{M}_0)^{-1}(\mu_W-\mu_O) \ .$$

For each iterate the approach is to quantify the vector of basic estimates E, the matrix **M** and the estimated covariance matrix for the linear estimates, **V**. The next step is to quantify the vector $\Delta'=(D_1,D_2,\ldots,D_5)$ where $D_i=W_i/PW_i-O_i/PO_i$, and compute the value of the statistic $t^2=\Delta'(\mathbf{M}'\mathbf{VM})^{-1}\Delta$. Finally we form a histogram of the quantities t^2 and tabulate the 10%, 5%, 2.5% and 1% critical values; and compute the mean and variance of this empirical distribution of t^2 values. It should be informative to compute from an available SAS routine the 10%, 5%, 2.5% and 1% upper values for a chi-squared distribution with five degrees of freedom and with the noncentrality parameter δ^2 defined above and compare these values to those obtained from the empirical distribution function.

REFERENCES

1. Brewer, K.R.W.(1975). "A Simple Procedure for Sampling πpswor". Australian Journal of Statistics, v.17, no.3, pp.166-72.
2. Chromy, James R.(1981). "Variance Estimators for a Sequential Sample Selection Procedure" in Current Topics in Survey Sampling, Krewski, Platek and Rao, eds. Academic Press.
3 Cochran, Wm. G.(1977). Sampling Techniques, 3rd Edition. Wiley, New York.
4. Katzoff,M.J.;Jones,G.K. and Curtin,L.R.(1988). "A General System for the Empirical Evaluation of Statistical Methods for Data from Complex Surveys". ASA Proceedings for the Section on Survey Methods Research.
5. Rust, Keith (1985). "Variance Estimation for Complex Estimators in Sample Surveys". Journal of Official Statistics, v.1, no.4, pp.381-97.
6. Saxena,R.R.;Singh,Padam;Srivastava,A.K. (1986). "An Unequal Probability Sampling Scheme". Biometrika, v.73, no.3, pp.761-763.
7. A.B. Sunter "List Sequential Sampling with Equal or Unequal Probabilities without Replacement" by in Applied Statistics, 1977, pp. 261-268.

APPENDIX

Derivation of the Bias Correction Term

It is enough to consider the case of one stratum from which PSUs i and j are drawn by Brewer's method. The problem is appropriately viewed as finding a term that will yield an unbiased estimate of variance when it is added to

$$\hat{Q}_1=\frac{\pi_i\pi_j-\pi_{ij}}{\pi_{ij}}\left(\frac{\hat{Y}_i}{\pi_i}-\frac{\hat{Y}_j}{\pi_j}\right)^2+ \pi_i{}^{-1}(N_i{}')^2\left(\frac{1}{n_i'}-\frac{1}{N_i'}\right)\hat{S}_i^2+ \pi_j{}^{-1}(N_j{}')^2\left(\frac{1}{n_j'}-\frac{1}{N_j'}\right)\hat{S}_j^2$$

where, dropping unnecessary subscripts, $\hat{Y}_i$ and $\hat{Y}_j$ are the unbiased estimators of Y_i and Y_j (defined at the end of section 2), $n_i{}'$ and $n_j{}'$ are functions of i and j, and $\hat{S}_i^2$ and $\hat{S}_j^2$ are the unbiased estimators of S_i^2 and S_j^2 (also defined at the end of section 2).

Using the expression for the variance of $\hat{Y}_i$ when $n_i{}'$ is fixed, we have that

$$E(\hat{Y}_i^2 \mid \text{choice of PSUs})= Y_i^2+(N_i')^2\left(\frac{1}{n_i{}'}-\frac{1}{N_i{}'}\right)S_i^2$$

Since sampling within PSUs is independent, we also have that

$$E(\hat{Y}_i\ \hat{Y}_j \mid \text{choice of PSUs}) = Y_i\ Y_j \ .$$

Using these facts and collecting terms one can see that

$E(\hat{Q}_1 \mid \text{choice of PSUs})=$

$$\sum_{i=1}^{N}\sum_{j>i}^{N} t_i t_j\left(\frac{\pi_i\pi_j-\pi_{ij}}{\pi_{ij}}\right)\left(\frac{Y_i}{\pi_i}-\frac{Y_j}{\pi_j}\right)^2+ \sum_{i=1}^{N}\sum_{j\neq i}^{N} t_i t_j(N_i{}')^2\left(\frac{\pi_i\pi_j-(1-\pi_i)\pi_{ij}}{\pi_i{}^2\pi_{ij}}\right) \cdot\left(\frac{1}{n_i(i,j)}-\frac{1}{N_i{}'}\right)S_i^2$$

where

$$t_i = \begin{cases} 1, & \text{if PSU i is in the sample} \\ 0, & \text{if otherwise} \end{cases}$$

The bias correction term will now follow in the form

$$\sum_{k=i,j\ \text{only}}(N_k{}')^2\left(\frac{2-\pi_k}{\pi_k{}^2}\right)(\zeta_k-\eta_k)S_k^2$$

if one now uses the facts that

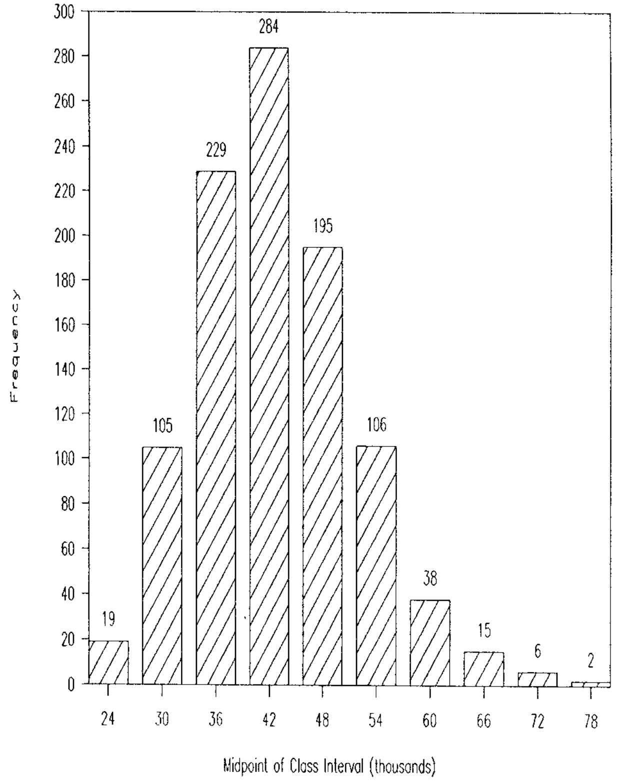
$$\Sigma\pi_i=2, \quad \sum_{j\neq i}\pi_{ij}=\pi_i$$

and the definitions for ζ_i and η_i. The specific expression shown in the text follows when the unweighted y_{ik} are replaced by the weighted values

$$\tilde{y}_{ik}=\frac{N_i{}'}{\pi_i n_i}, y_{ik}$$

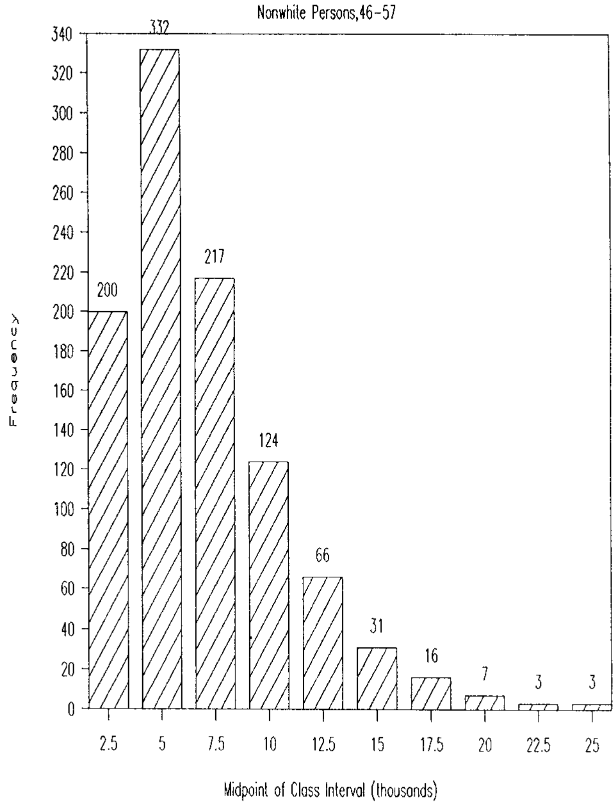
TOTAL SHORT STAY DAYS

White Persons,46-57



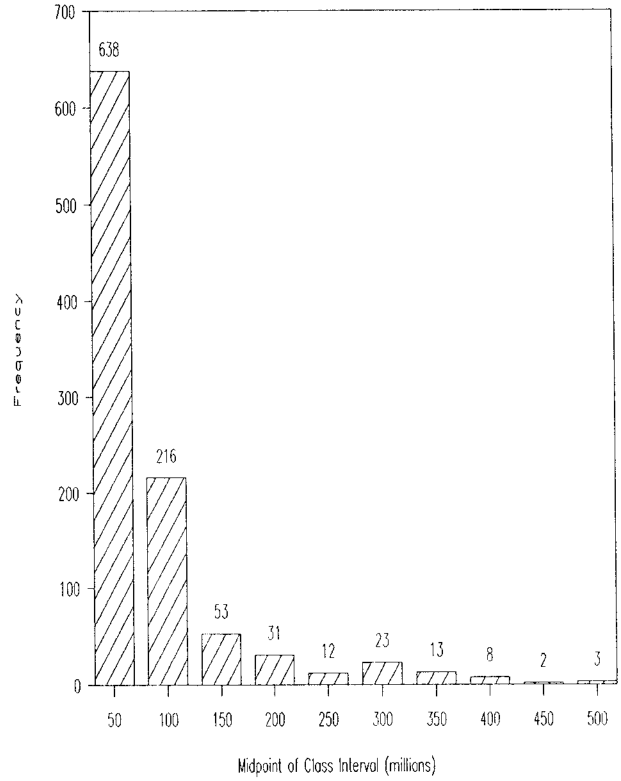
TOTAL SHORT STAY DAYS

Nonwhite Persons,46-57



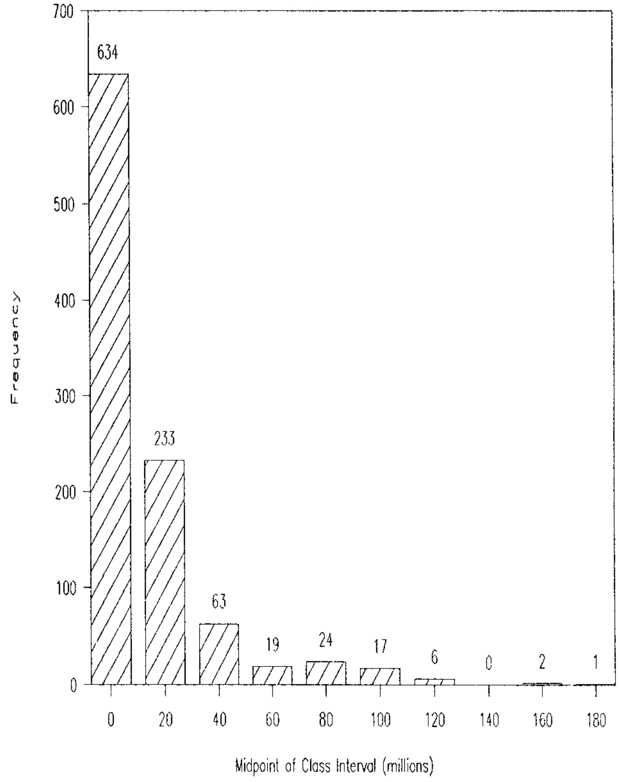
JACKKNIFE VARIANCE ESTIMATES

Short Stay Days,White,46-57



JACKKNIFE VARIANCE ESTIMATES

Short Stay Days,Nonwhite,46-57


Frequency
Midpoint of Class Interval (millions)