Series 2
Number 55

The Prediction Approach to Finite Population Sampling Theory: Application to the Hospital Discharge Survey

DHEW Publication No. (HSM) 73-1329

U.S. DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE
Public Health Service

Health Services and Mental Health Administration
National Center for Health Statistics
Rockville, Md. April 1973

Vital and Health Statistics-Series 2-No. 55

For sale by the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402
Price 65 cents Domestic Postpaid or 45 cents GPO Bookstore

NATIONAL CENTER FOR HEALTH STATISTICS

OFFICE OF STATISTICAL METHODS

Vital and Health Statistics-Series 2-No. 55

DHEW Publication No. (HSM) 73-1329
Library of Congress Catalog Card Number 72-600134

FOREWORD

The Center contracted with the School of Public Health, Johns Hopkins University and Dr. Richard Royall to investigate the possible application to the Hospital Discharge Survey of the prediction approach to finite population sampling. This report presents the results of the research completed under these contracts.

The prediction approach is based on "super-population" probability models. It is an alternative to the conventional theory of sampling from finite populations and does not apply the conventional concept of repeated random sampling from a fixed population. Rather, it applies classical prediction theory to solve sampling problems. Viewing finite population sampling problems as prediction problems is a relatively new development and hence is probably known to only a few statisticians. Furthermore, Dr. Royall's style is throughout the report quite elegant. Therefore, we asked him to prepare a nonmathematical description of the prediction approach and indicate how it differs from the classical approach. This material is presented in the Introduction.

We commissioned this research project in anticipation of redesigning the Hospital Discharge Survey. Overall, the findings presented in this report throw a favorable light on the existing design and estimator. The findings suggest some changes for improving the design and also identify some areas for further research. We believe this report will help us to develop an improved design for the Hospital Discharge Survey.

Dr. Jay Herson worked with Dr. Royall and the Office of Information in preparing this manuscript for publication.

MONROE G. SIRKEN

CONTENTS

THE PREDICTION APPROACH TO FINITE POPULATION SAMPLING THEORY: APPLICATION TO THE HOSPITAL DISCHARGE SURVEY

Richard M. Royall, Ph.D., *Associate Professor, Department of Biostatistics, School of Public Health, Johns Hopkins University*

INTRODUCTION

The material presented is the result of an unorthodox approach to finite population sampling problems. Specifically, it describes the elements and results of an application of this approach to the Hospital Discharge Survey (HDS), a continuing sample survey of the Nation's short-stay hospitals conducted by the National Center for Health Statistics. It is not presented as a finished and polished analysis but as a basic sketch whose contents must be critically evaluated, adjusted, and refined if it is to be of real value in HDS. The mathematical model used in this work expresses plausible *initial* assumptions about certain variables of interest. With experience will come increasing knowledge concerning the HDS population and relationships among its variables. Such information must be used to alter and develop the basic model described in this report.

In this section the approach guiding the investigation will be contrasted with the conventional approach to finite population sampling problems. For purposes of illustration, an imaginary population of 50 hospitals in some relatively homogeneous geographical region will be considered. The number of beds in each hospital is known. A sample of 10 hospitals is selected, and the number of patients discharged from these 10 during some given time period is observed. The problem is to estimate the total number of discharges from all 50 hospitals (the *population total*).

In its basic, simplest version, the conventional approach treats the 50 unknown numbers of discharges as unknown constants. The only random variation in the problem is injected by the sampler, who uses a random sampling plan to decide which 10 hospitals will comprise the sample. This sampling plan specifies the probability of selection of each potential sample. A sampling and estimation *procedure* consists of a sampling plan together with an estimator or formula for calculating estimates from samples. The characteristic feature of orthodox sampling theory is that a procedure is evaluated in terms of the statistical properties of the estimator, principally its expected value and variance, under the random sampling plan chosen by the sampler. Of course other factors, e.g., costs, feasibility, and ease of estimation of variance from the sample influence the choice of a procedure. Nevertheless, the basic objective is to find, subject to limitations such as cost, a procedure whose estimator is unbiased (at least approximately) and has small variance.

For present purposes only one sampling plan and two estimators are considered. The plan calls for simple random sampling—only samples which consist of exactly 10 different hospitals are allowed, and all such samples are equally likely to be selected. Let $t_1, t_2, \ldots, t_{50}$ represent the respective numbers of discharges from the 50 hospitals, let $b_1, b_2, \ldots, b_{50}$ be their respective numbers of beds, and let s represent the set of 10 hospitals in the sample. A simple estimator of the population total, $T=\sum_1^{50} t_i$, is the product of {the average number of discharges per hospital in the sample} × {the number of hospitals in the population}, i.e.,

$$\left(\sum_s t_i/10\right)50. \qquad (1)$$

This is called the simple expansion estimator. Under the present (simple random) sampling plan it is unbiased.

Another estimator (the ratio estimator) estimates T by the product of {the average number of discharges per bed in the sample} × {the total number of beds in the population}:

$$\left(\sum_s t_i \Big/ \sum_s b_i\right) \sum_1^{50} b_i. \tag{2}$$

Under the simple random sampling plan the ratio estimator is biased.

Two observations concerning the variances of the expansion and ratio estimators are needed:

(i) both variances are defined as average values of squared errors over all samples, and

(ii) the two variances are unequal.

Such biases and variances are certainly relevant in planning surveys and choosing procedures which can be expected to produce good estimates. However, after the sample s is selected the situation is drastically changed. As indicators of uncertainty in the estimator when it is applied to a particular sample, the conventionally defined bias and variance can be quite misleading. For example, if the sample contains mostly small (few beds) hospitals, we can be confident that the expansion estimator (1) will give an underestimate of T. In this situation, to describe the estimator as "unbiased" is at best irrelevant and at worst misleading. Here it would seem accurate and informative to describe the estimator as having a negative bias, yet this is impossible—for a given sample s there is no probability distribution with respect to which bias can be defined. Similar remarks apply to samples containing a disproportionate number of large hospitals—in these samples the expansion formula tends to produce overestimates of T. In this context, the statement that the estimator is "unbiased" in the conventional sense simply means that samples containing too many small units, which tend to give underestimates of the population total, will be balanced, in a hypothetical infinite sequence of samples, by samples containing too few small units, which tend to give overestimates.

It would appear that when s contains an excess of small hospitals, an upward adjustment is required if (1) is to deserve the description "unbiased." The adjustment might be made by multiplying (1) by the factor $\left(\sum_1^{50} b_i/50\right) \Big/ \left(\sum_s b_i/10\right)$, the ratio of {the average number of beds per hospital in the population} ÷ {the average number of beds per hospital in the sample}. The effect of this factor will be to increase the estimate when the average sample hospital is small and to decrease the estimate when the average sample hospital is large. The resulting estimator,

$$\left\{\left(\sum_1^{50} b_i/50\right) \Big/ \left(\sum_s b_i/10\right)\right\} \left\{50 \sum_s t_i/10\right\}, \tag{3}$$

is the ratio estimator (2), which, according to the conventional definition, is biased. Thus in this problem a notion of bias useful for inference from a given sample s must be in direct conflict with the conventional theory; the unbiased estimator should be called biased and vice versa.

The orthodox variance (or its square root, the standard error) is not a satisfactory measure of the uncertainty in the estimator after s is fixed, although it is usually interpreted as such a measure. The two estimators (1) and (2) have different variances, yet when the sample is such that the average size of sample hospitals is equal to the average size for the whole population, the results of using (1) and (2) are identical. That is, when such samples are selected, the ratio and expansion formulas are the same and therefore equally precise, equally uncertain, equally accurate, etc. Yet orthodox theory assigns different standard errors depending on whether formula (1) or formula (2) was used.

The prediction approach recognizes that, after the sample is observed, the population total can be written

$$T = \sum_s t_i + \sum_{\tilde{s}} t_i \tag{4}$$

where $\tilde{s}$ denotes the collection of hospitals *not* in the sample. Since the first of the two sums in (4) is now *known*, the problem is to estimate the second sum, the total number of discharges in hospitals not in the sample. Any estimator of T can be written in a form comparable to (4), i.e.,

$$\hat{T} = \sum_s t_i + \left(\hat{T} - \sum_s t_i\right). \tag{5}$$

Using $\hat{T}$ to estimate T is, in effect, using $\hat{T} - \sum_s t_i$ to estimate $\sum_{\tilde{s}} t_i$. Clearly the questions of whether a particular estimator when applied to a particular sample s is good or bad, reasonable or foolish, unbiased or biased, etc. are answerable only in light of the relationship between hospitals in the selected sample and those not in the sample. An estimator $\hat{T}$ is precisely as good for estimating T as is the difference $\hat{T}-\sum_{s} t_i$ for predicting the unobserved sum $\sum_{\tilde{s}} t_i$.

The prediction approach expresses the relationship between sample and nonsample hospitals by a probability model ("super-population" model) in which the numbers of interest, $t_1, t_2, \ldots, t_{50}$, are thought of as having been produced by some probabilistic process described by a mathematical model. This process serves as a vital link between the observed and unobserved totals. What these two totals have in common and what enables us to use the observed to make inferences concerning the unobserved is that they were all produced by one underlying probabilistic process. Inferences from the sample can be made concerning certain important characteristics of the process; this information can then be used to predict the values of the totals not observed.

The simplest model describing the basic structure of the hospital problem treats t_i, the number of discharges from hospital i, as an observation on a random variable whose expected value is proportional to b_i, the number of beds. That is, the expected number of discharges is βb_i, where β is some unknown positive constant which can be estimated from the sample. If (1) and (2) are written in forms comparable to (4), then the expansion estimator is

$$\sum_{s} t_i + 40\left(\sum_{s} t_i/10\right), \tag{6}$$

and the ratio estimator is

$$\sum_{s} t_i + \left(\sum_{s} t_i \Big/ \sum_{s} b_i\right) \sum_{\tilde{s}} b_i. \tag{7}$$

Using the probability model, s can be held fixed and the statistical properties of the estimators for the *given* sample examined. Thus the second terms in (6) and (7) are actually predictors for the random total discharges from nonsample hospitals. The properties of the expansion estimator for this sample are precisely the properties of $40\left(\sum_{s} t_i/10\right)$ when it is used to predict $\sum_{\tilde{s}} t_i$. The expected value of the predictor is $40\left(\sum_{s} \beta b_i/10\right)$, while the variable predicted has expected value $\sum_{\tilde{s}} \beta b_i$. Since the expected value of the predictor is less than that of the variable predicted when the average size of sample hospitals is less than the average size of nonsample hospitals, the prediction approach describes the expansion estimator as "biased" in this context. The ratio estimator, on the other hand, is called "unbiased" for every s since the expected value of the predictor, $\left(\sum_{s} \beta b_i \Big/ \sum_{s} b_i\right) \sum_{\tilde{s}} b_i$, equals the expected value, $\beta \sum_{\tilde{s}} b_i$, of the variable predicted.

The variance used to measure uncertainty in an estimate under the prediction model is, like that used in the conventional approach, the variance of the difference $\hat{T}-T$ between the estimator and the quantity estimated. But whereas the conventional approach calculates the variance of this difference with respect to the random sampling plan (the probability distribution over all possible samples), the prediction approach calculates the variance with respect to the probability model with the sample s held fixed. Thus the conventional approach states for the ratio estimator, say, the same standard error for all samples of size 10, while the prediction approach quotes one value when the sample contains mostly large hospitals and a larger value when most of the sample hospitals are small. (See formula (3), page 15.) Both the conventional and prediction variances are unknown and must be estimated from the sample. There is theoretical and empirical evidence that the latter is the more useful measure of the uncertainty in an observed estimate [1].[1]

This simplified example suggests the inadequacy of orthodox notions of bias and variance for purposes of inference and points to the prediction approach as being more relevant and informative at the data-analysis stage. However, some of the most interesting implications of the prediction approach appear when the problem of sample selection is considered. When this approach is adopted random sampling loses its status as the one and only fundamental and indispensable component of finite population sampling theory; it

[1] Figures in brackets indicate the literature references at the end of this paper.

assumes instead the more humble role of a useful and important tool.

To apply the prediction approach to a real problem, we must first be able to produce an adequate model which is simple enough to analyze. The adequacy of a model is to some extent a matter of judgment, but mathematical investigations can help. Thus considerable attention is paid in this report to the effects of errors in the basic model and especially to the identification of samples for which the conclusions derived from the model are relatively insensitive to the most obvious sorts of departure from the model.

The models in this report are used in two ways: to generate sampling and estimation procedures having certain desirable statistical properties and to provide increased appreciation of the properties of procedures currently in use.

SUMMARY

The author has recently been studying finite population sampling problems using an approach which is based on viewing such problems as straightforward classical prediction problems rather than on applying the conventional concept of repeated random sampling from the fixed population. Previous work by Royall [1, 2] has suggested that the prediction approach, which employs super-population probability models, is a useful alternative to the conventional theory and can be of value in illuminating the strengths and weaknesses of standard procedures as well as in suggesting and providing a theoretical basis for new procedures.

Other recent studies viewing finite population sampling problems as prediction problems have been made by Ericson [3, 4], who adopts a Bayesian approach, and by Kalbfleisch and Sprott [5], whose approach is fiducial. There have also been other studies in which the classical (non-Bayesian, non-fiducial) approach is adopted, e.g., Brewer's paper [6] and parts of the paper by Scott and Smith [7], whose basic approach is Bayesian.

HDS employs a two-stage sampling plan in which hospitals are the first-stage sampling units and patient discharge records the second-stage units. Within each of four geographical regions, hospitals are stratified according to size, as measured by the number of beds (bed size) listed in the 1963 Master Facilities Inventory of Hospitals and Institutions (MFI).

For the purposes of this study, the hospitals in the four geographical regions are treated as natural, distinct populations which represent four separate instances of the same basic problem. Thus the "population" referred to in this report corresponds to the HDS population within any of the four large geographical regions, and stratification is on the bed size variable only, not on geographical region.

In HDS a sample of hospitals is selected from each stratum, and a sample of discharges is drawn from each selected hospital. For each discharge in the sample a numerical characteristic of interest, or response, is observed. Sample discharges from a given hospital are used to estimate the total for all discharges from that hospital. These estimated totals for the sample hospitals are then used, along with the auxiliary variable, bed size, to construct a ratio-type estimator for the stratum total. This estimation procedure is applied independently within each stratum.

In Part I of this report complications produced by the second stage of sampling are set aside, and only single-stage sampling problems are considered. The main purpose of this part of the study is to gain an increased understanding of the simple and valuable ratio estimator. Thus we consider a range of probability models, but with more attention paid to studying the performance of the ratio estimator under such models than to describing optimal sampling and estimation strategies for each model. We see in Part I a new explanation for the success of the ratio estimator in practical applications: although real problems are not often depicted with great accuracy by the probability model under which the ratio estimator is optimal, frequently, *for the particular sample drawn*, the ratio estimator is approximately optimal under a wide range of models.

Stratification on the size variable with separate ratio estimation in the strata is examined as a technique for efficiently insuring unbiasedness. Finally, the effects of errors in the model on the performance of variance estimates are considered.

In Part II the second stage of sampling is introduced. The problem is first studied in its simplest form; later the phenomena of out-of-scope and nonresponse discharges are represented in the model.

Overall, the results throw a favorable light on the HDS design and estimator. This investigation suggests that the rule used to allocate the first-stage sample among the various strata might be improved, but that, given the rule actually used, the allocation of the second-stage sample is approximately optimal. Another suggestion is that the average bed size per hospital in each stratum's sample should be approximately equal to the average bed size per hospital in the entire stratum. It is supposed that the present method of hospital selection produces samples which satisfy this condition, but this should be verified.

Two areas in which further research with super-population models is expected to be fruitful are

analysis of the HDS variance estimator and study of the sophisticated sampling technique known as "controlled selection." The first of these is of more immediate importance since the current HDS variance estimator is an adaptation of the variance estimator conventionally used in single-stage ratio estimation problems. There are theoretical results, supported by some empirical work, which imply that this conventional variance estimator should be replaced by one suggested by superpopulation theory [1].

Controlled selection procedures are used by the HDS to select the first-stage sample. Investigation along the lines leading to defensive samples in Part I would probably increase our appreciation of precisely what these procedures accomplish and how. Such an investigation should provide theoretical support for these selection procedures.

PART I. SINGLE-STAGE SAMPLING

Description of Problem

Terminology, notation.—The population of interest consists of M units labeled 1, 2, . . ., M. Associated with unit k are two numbers (B_k, t_k) with B_k known and t_k fixed but unknown. The units might be hospitals of a certain type with B_k some measure, for instance, number of beds, of the size of hospital k, and t_k some characteristic of interest such as number of days of care provided by hospital k during a particular month. A sample consisting of m units is to be selected from the population and the t-values associated with the sample units are to be observed. The objective is to estimate the total

$$T=\sum_{k=1}^{M} t_k \tag{1}$$

and give a measure of the precision of the estimate. The set of m labels identifying the sample units is denoted by s, and the set of M-m labels of units not in the sample is denoted by $\tilde{s}$.

Probability models.—In this study the numbers $t_1, t_2, \ldots, t_M$, whose sum we must estimate, are considered to be realized values of independent random variables $T_1, T_2, \ldots, T_M$. The expected value and variance of T_k depend on the size measure B_k and are denoted by $h(B_k)$ and $\sigma^2\, v(B_k)$, respectively. Thus we can write

$$T_k=h(B_k)+\epsilon_k\sqrt{v(B_k)} \qquad k=1, \ldots, M$$

where $\epsilon_1, \ldots, \epsilon_M$ are independent random variables, each having mean zero and variance σ^2. In particular, attention is focused on models in which $h(B)$ is a polynomial, say, of order J (at most). That is,

$$h(B)=r_0\beta_0+r_1\beta_1B+r_2\beta_2B^2+\ldots+r_J\beta_JB^J$$

where the r's are zeroes and ones. If $r_j=1$, it means simply that the term β_jB^j appears in the regression function; $r_j=0$ indicates the absence of this term. When the regression function h has the above form, we refer to the probability model as $\xi(r_0, r_1, \ldots, r_J : v(B))$. For example, $\xi(0, 1 : B)$ refers to the model

$$T_k=\beta_1B_k+\epsilon_k\sqrt{B_k}\ ,$$

in which both the expected value and the variance of T_k are proportional to the size B_k. As another example, $\xi(1, 1, 0, 1 : 1)$ refers to the model

$$T_k=\beta_0+\beta_1B_k+\beta_3B_k^3+\epsilon_k.$$

Here $\mathrm{Var}\, T_k=\mathrm{Var}\, \epsilon_k=\sigma^2$, a constant.

It should be emphasized that the fundamental problem is that of estimating the sum (1) of the actual t-values. If a particular model, say $\xi(1, 1 : B)$, applies, an intermediate step in the process of estimating the sum is estimation of β_0 and β_1, but the objective is to estimate $\sum_{k=1}^{M} t_k$, not the parameters in the super-population model. It will be especially important to keep this objective in mind when seeking optimal sampling plans since the plan which is best for estimating $\sum_{k=1}^{M} t_k$ under a particular model is not generally the best plan for estimating parameters of the model.

Under probability models of the sort considered here, the problem of estimating the total $\sum_{1}^{M} t_k$ on the basis of a sample s is a version of the general problem of predicting future observations on random variables. This is evident when the total is expressed as the sum of two terms, $\sum_{s} t_k$ and $\sum_{\tilde{s}} t_k$. The first of these two is known after the sample has been observed, and estimating $\sum_{\tilde{s}} t_k$ is equivalent to predicting the sum of the unobserved random variables $\sum_{\tilde{s}} T_k$. For further discussion of this view of certain finite population sampling problems as prediction problems, the reader is referred to Royall [2].

Optimality

Best linear unbiased (BLUE) estimators.—For a given sample s and a given model ξ, an estimator $\hat{T}$ will be said to be unbiased if $E_\xi(\hat{T}-T)=0$, where the expectation is taken with respect to the probability distribution specified by the model. For example, for all s the ratio estimator,

$$\left(\sum_s t_k \Big/ \sum_s B_k\right) \sum_1^M B_k,$$

is unbiased under the model $\xi(0, 1:v(B))$ for any variance function v:

$$E_\xi\left[\left(\sum_s T_k \Big/ \sum_s B_k\right) \sum_1^M B_k - \sum_1^M T_k\right]$$

$$= \left(\sum_s \beta_1 B_k \Big/ \sum_s B_k\right) \sum_1^M B_k - \sum_1^M \beta_1 B_k = 0.$$

Under the model $\xi(1, 1 : v(B))$

$$E_\xi\left[\left(\sum_s T_k \Big/ \sum_s B_k\right) \sum_1^M B_k - \sum_1^M T_k\right]$$

$$= \frac{m\beta_0 + \beta_1 \sum_s B_k}{\sum_s B_k} \sum_1^M B_k - M\beta_0 - \beta_1 \sum_1^M B_k$$

$$= \beta_0 \left(\frac{m \sum_1^M B_k}{\sum_s B_k} - M \right).$$

Thus under this model the ratio estimator is unbiased only if $\sum_1^M B_k \ / \ M = \sum_s B_k / m$.

Only estimators which are linear functions of the t's in the sample are considered here. The determination of a best linear unbiased estimator under a given model and for a given s is quite simple. We seek among all linear unbiased estimators $\hat{T}$ one which minimizes the mean square error (MSE), $E_\xi(\hat{T}-T)^2$. The estimator $\hat{T}$ is unbiased if and only if the difference between $\hat{T}$ and the sample total $\sum_s T_k$ is an unbiased estimate of the total for nonsample units, i.e.,

$$E_\xi\left(\hat{T} - \sum_s T_k\right) = E_\xi\left(\sum_{\tilde{s}} T_k\right).$$

Thus if $\hat{T}$ is unbiased,

$$E_\xi\left(\hat{T}-T\right)^2 = E_\xi\left(\left(\hat{T} - \sum_s T_k - E_\xi \sum_{\tilde{s}} T_k\right)\right.$$

$$\left. - \left(\sum_{\tilde{s}} T_k - E_\xi \sum_{\tilde{s}} T_k\right)\right)^2$$

$$= E_\xi\left(\hat{T} - \sum_s T_k - E_\xi \sum_{\tilde{s}} T_k\right)^2$$

$$+ \operatorname{Var}\left(\sum_{\tilde{s}} T_k\right).$$

$$= \operatorname{Var}\left(\hat{T} - \sum_s T_k\right) + \operatorname{Var}\left(\sum_{\tilde{s}} T_k\right).$$

Note that linearity of $\hat{T}$ is equivalent to linearity of $\hat{T} - \sum_s T_k$. Therefore, under the model $\xi(r_0, r_1, \ldots, r_j : v(B))$, $\hat{T}$ is a BLUE estimator for T if and only if $\hat{T} - \sum_s T_k$ is a BLUE estimator for the expected value of $\sum_{\tilde{s}} T_k$, $\sum_{j=0}^{J} \left(\sum_{\tilde{s}} B_k^j\right) r_j \beta_j$.

The generalized Gauss-Markov theorem (see Rao [8]) shows that the BLUE estimator for such a linear function of the regression coefficients is obtained by straightforward application of the familiar method of weighted least-squares estimation. Thus under the present model, $\hat{T}$ is the BLUE estimator for T if

$$\hat{T} = \sum_s T_k + \sum_{j=0}^{J} \left(\sum_{\tilde{s}} B_k^j\right) r_j \hat{\beta}_j$$

where the $\hat{\beta}$'s are the weighted least-squares estimates of the regression coefficients under the specified model.

Two examples will perhaps clarify this point:

Example: Under the model $\xi(1, 1 : 1)$ the weighted least-squares estimates of β_0 and β_1 are

$$\hat{\beta}_0(1,1:1) = \left(\sum_s B_k^2 \sum_s T_k - \sum_s B_k \sum_s B_k T_k\right) \Big/ D,$$

$$\hat{\beta}_1(1,1:1) = \left(m \sum_s B_k T_k - \sum_s B_k \sum_s T_k\right) \Big/ D$$

where

$$D = m \sum_s B_k^2 - \left(\sum_s B_k\right)^2.$$

The BLUE estimator for T is thus

$$\hat{T}(1,1:1) = \sum_s T_k + (M-m)\hat{\beta}_0(1,1:1)$$

$$+\hat{\beta}_1(1,1:1)\sum_{\tilde{s}} B_k.$$

Example: Under the model $\xi(0, 1:B)$, the weighted least-squares estimator for β_1 is

$$\hat{\beta}_1(0,1:B) = \sum_s T_k \Big/ \sum_s B_k.$$

Thus the BLUE estimator for T is

$$\hat{T}(0,1:B) = \sum_s T_k + \hat{\beta}_1(0,1:B)\sum_{\tilde{s}} B_k.$$

This estimator can also be written in the more familiar form

$$\hat{T}(0,1:B) = \left(\sum_s T_k \Big/ \sum_s B_k\right)\sum_1^M B_k. \qquad (2)$$

So the BLUE estimator under the model $\xi(0,1:B)$ is the popular ratio estimator.

Optimal samples.—The model $\xi(0, 1:B)$ is of particular interest since it is under this model that the standard ratio estimator is optimal. Here the expected squared error is

$$E_\xi(\hat{T}(0,1:B)-T)^2 = \sigma^2\left(\sum_{\tilde{s}} B_k \Big/ \sum_s B_k\right)\sum_1^M B_k. \qquad (3)$$

From (3) it is apparent that in this context the optimal sample is one for which $\sum_s B_k$ attains its maximum value. This is simply the sample composed of the m units whose B values are largest. It is the sample which is optimal for use with the optimal estimator under the model $\xi(0, 1:B)$ and will be denoted by $s(0,1:B)$. (See Royall [2].)

More generally, under the model $\xi(r_0, r_1, \ldots, r_J : v(B))$, the sample for which $E_\xi(\hat{T}(r_0, r_1, \ldots, r_J : v(B)) - T)^2$ is minimized is optimal for use with the optimal estimator. This sample will be denoted by $s(r_0, r_1, \ldots, r_J : v(B))$.

Effects of Errors in the Model

We now assume that the population of interest is one for which $\xi(0,1:B)$ is a plausible model but cannot ignore the possibility that this model is inaccurate. Thus we seek strategies which are nearly optimal under $\xi(0,1:B)$ but will produce satisfactory results under various other models.

Overall ratio estimator.—Under the model $\xi(0,1:B)$ the optimal estimator is $\hat{T}(0,1:B)$, the ratio estimator, and the optimal sample for use with this estimator is $s(0,1:B)$, the sample consisting of the m units whose B-values are largest. That is, of all strategies $(s, \hat{T})$ consisting of a sample s and a linear unbiased estimator $\hat{T}$, the pair $(s(0,1:B), \hat{T}(0,1:B))$ is optimal under the model $\xi(0,1:B)$. Many questions arise at this point. How good is this strategy when $\xi(0,1:B)$ is not the correct model? If we use $\hat{T}(0,1:B)$, is $s(0,1:B)$ a good sample when the true model is $\xi(0,1:v(B))$ for some particular variance function $v(B) \neq B$? How can we find a procedure which is good under $\xi(0,1:B)$ but performs adequately under the alternative model $\xi(1,1,1:B)$? Answers to some questions of this sort are known. For example, it is well known that the unbiasedness property of BLUE estimators is not destroyed by alteration of the variance function. Thus $\hat{T}(0,1:B)$ is unbiased under the model $\xi(0,1:v(B))$ for any variance function v.

More generally, consider the estimator $\hat{T}(0,1:B)$ under the model $\xi(r_0, r_1, \ldots, r_J : v(B))$. The bias is

$$E_\xi(\hat{T}(0,1:B)-T)$$

$$= E_\xi\left(\left(\sum_s T_k \Big/ \sum_s B_k\right)\sum_{\tilde{s}} B_k - \sum_{\tilde{s}} T_k\right)$$

$$= \sum_{j=0}^{J} r_j\beta_j\left(\frac{\sum_s B_k^j}{\sum_s B_k} - \frac{\sum_{\tilde{s}} B_k^j}{\sum_{\tilde{s}} B_k}\right)\sum_{\tilde{s}} B_k.$$

Note that the summand is zero when $j=1$; the bias is not affected by the regression coefficient β_1. It is clear from this expression that $\hat{T}(0, 1:B)$ is unbiased if and only if

$$\frac{\sum_{s} B_k^j}{\sum_{s} B_k} = \frac{\sum_{\tilde{s}} B_k^j}{\sum_{\tilde{s}} B_k}$$

for all j such that the term $\beta_j B^j$ appears in the regression equation (i.e., for all j such that $r_j = 1$). It is easily shown that these conditions for unbiasedness are equivalent to

$$\frac{\frac{1}{m}\sum_{s} B_k^j}{\frac{1}{m}\sum_{s} B_k} = \frac{\frac{1}{M}\sum_{1}^{M} B_k^j}{\frac{1}{M}\sum_{1}^{M} B_k} \tag{4}$$

for all j such that $r_j = 1$. Note that (4) is always satisfied for $j = 1$. For example, $\hat{T}(0, 1:B)$ is unbiased under the model $\xi(1, 1:v(B))$ if (4) is satisfied for $j = 0$:

$$\frac{1}{m}\sum_{s} B_k = \frac{1}{M}\sum_{1}^{M} B_k. \tag{5}$$

This estimator is unbiased under the still more general model $\xi(1, 1, 1:v(B))$ if, in addition to (5), s satisfies

$$\frac{1}{m}\sum_{s} B_k^2 = \frac{1}{M}\sum_{1}^{M} B_k^2.$$

Suppose it is believed that $\xi(0, 1:B)$ is an adequate model for a given problem, but the estimator must perform reasonably well when the model is in error and the actual regression function is not a straight line through the origin. The following theorem shows that by careful choice of the sample s we can insure the optimality of the ratio estimator under polynomial regression models. For any positive integer J, let $s(J)$ denote any sample satisfying (4) for $j = 0, 1, \ldots, J$.

Theorem: If $s = s(J)$, then $\hat{T}(0, 1:B) = \hat{T}(1:1)$, and this is the BLUE estimator under the models $\xi(r_0, 1, r_2, r_3, \ldots, r_J:B)$ and $\xi(1, r_1, r_2, \ldots, r_J:1)$ for every sequence $r_0, r_1, r_2, \ldots, r_J$ of zeroes and ones.

Proof: Note that for any s, $\hat{T}(1:1) = M\sum_{s} T_k/m$, and for $s = s(J)$ this statistic is also $\hat{T}(0, 1:B)$. This estimator has already been shown to be unbiased when $s = s(J)$ under J^{th}-order polynomial regression models for any variance function. One can prove its optimality when $s = s(J)$ under all models of the form $\xi(r_0, 1, r_2, \ldots, r_J:B)$ by:

(i) finding the weighted least-squares estimates $\hat{\beta}_j(r_0, 1, r_2, \ldots, r_J:B)$ for all $j = 0, 1, 2, \ldots, J$ such that $r_j = 1$, under the model $\xi(r_0, 1, r_2, \ldots, r_J:B)$;

(ii) forming the BLUE estimator for T,

$$\hat{T}(r_0, 1, r_1, \ldots, r_J:B) = \sum_{s} T_k + \sum_{j=0}^{J}\left(\sum_{\tilde{s}} B_k^j\right)\hat{\beta}_j(r_0, 1, r_2, \ldots, r_J:B)\, r_j;$$

and

(iii) noting that when $s = s(J)$ this estimator assumes the simple form $M\sum_{s} T_k/m$.

Alternatively, we note that since $\hat{T}(r_0, 1, r_1, \ldots, r_J:B)$ is unbiased under the model $\xi(0, 1:B)$ and $\hat{T}(0, 1:B)$ is the BLUE estimator under this model, we have for all s

$$E\{(\hat{T}(0, 1:B) - T)^2 \mid \xi(0, 1:B)\} \leq E\{(\hat{T}(r_0, 1, r_2, \ldots, r_J:B) - T)^2 \mid \xi(0, 1:B)\}. \tag{6}$$

Now when $s = s(J)$, $\hat{T}(0, 1:B)$ is unbiased under $\xi(r_0, 1, r_2, \ldots, r_J:B)$. But $\hat{T}(r_0, 1, r_2, \ldots, r_J:B)$ is the BLUE estimator under this model. Thus when $s = s(J)$

$$E\{(\hat{T}(0, 1:B) - T)^2 \mid \xi(r_0, 1, r_2, \ldots, r_J:B)\} \geq E\{(\hat{T}(r_0, 1, r_2, \ldots, r_J:B) - T)^2 \mid \xi(r_0, 1, r_2, \ldots, r_J:B)\}. \tag{7}$$

Now

$$E\{(\hat{T}(r_0, 1, r_2, \ldots, r_J:B) - T)^2 \mid \xi(r_0, 1, r_2, \ldots, r_J:B)\} = E\{(\hat{T}(r_0, 1, r_2, \ldots, r_J:B) - T)^2 \mid \xi(0, 1:B)\}$$

for all s, and when $s=s(J)$ this equality holds if $\hat{T}(r_0, 1, r_2, \ldots, r_J:B)$ is replaced by $\hat{T}(0, 1:B)$. Thus (6) implies that when $s=s(J)$, equality must hold in (7).

Optimality under $\xi(1, r_1, r_2, \ldots, r_J: 1)$ can be proved by entirely analogous arguments.

Samples $s(J)$ will be referred to as *defensive samples.* Selection of a defensive sample insures that the ratio estimator retains not only its unbiasedness but also its optimality under the polynomial regression models. As noted above, one convenient feature of defensive samples is the simple form which the ratio estimator assumes. When $\sum_s B_k/m=\sum_1^M B_k/M$, the ratio estimator is simply the expansion estimator, $M \sum_s T_k/m$.

Example: Under the model $\xi(1, 1: B)$, the BLUE estimator is

$$\hat{T}(1, 1: B)=\sum_s T_k+(M-m)\hat{\beta}_0(1, 1:B)$$

$$+\hat{\beta}_1(1, 1: B) \sum_{\tilde{s}} B_k$$

where

$$\hat{\beta}_0(1, 1: B)=\Big(\sum_s \frac{T_k}{B_k}\sum_s B_k-m\sum_s T_k\Big)/D_2,$$

and

$$\hat{\beta}_1(1, 1: B)=\Big(\sum_s T_k \sum_s \frac{1}{B_k}-m\sum_s \frac{T_k}{B_k}\Big)/D_2,$$

with

$$D_2=\sum_s B_k \sum_s \frac{1}{B_k}-m^2.$$

When $\sum_s B_k/m=\sum_1^M B_k/M=\bar{B}$, we have

$$\hat{T}(1, 1: B)=\sum_s T_k+$$

$$\frac{(M-m)\Big[\sum_s \frac{T_k}{B_k}\bar{B}-\sum_s T_k+\Big(\frac{1}{m}\sum_s T_k \sum_s \frac{1}{B_k}-\sum_s \frac{T_k}{B_k}\Big)\bar{B}\Big]}{\bar{B}\sum_s \frac{1}{B_k}-m}$$

$$=\sum_s T_k+\frac{(M-m)}{m}\sum_s T_k$$

$$=M\sum_s T_k/m.$$

When a defensive sample is used, the mean square error (MSE) of the ratio estimator under $\xi(0, 1:B)$ is, from (3),

$$E(\hat{T}(0, 1:B)-T)^2=\frac{M^2}{m}\left(1-\frac{m}{M}\right)\sigma^2\bar{B}. \quad (8)$$

When the estimator is unbiased, the MSE is simply the variance, and the variance does not depend on which terms appear in the regression model. Thus we see that when s is $s(J)$, expression (8) applies under the model $\xi(r_0, r_1, \ldots, r_J:B)$ for *any* combination $r_0, r_1, \ldots, r_J$ of zeroes and ones. (Note that (8) does *not* apply under $\xi(r_0, r_1, \ldots, r_J:1)$.) It follows that when $\hat{T}(0, 1:B)$ is the chosen estimator under the model $\xi(0, 1:B)$, the ratio of the MSE when s is the optimal sample $s(0, 1: B)$ to the MSE when $s=s(J)$ for *any* $J \geqslant 1$ is

$$\min_s\Big(\sum_{\tilde{s}} B_k/\ M-m\Big)\Big/\Big(\sum_s B_k/m\Big).$$

These results may be interpreted as follows: When $\xi(0, 1:B)$ is the true model, the ratio estimator is optimal for any s. If the ratio estimator is used but the model is actually $\xi(1, 1:v(B))$, a bias is incurred. We can guard against such a bias by choosing the defensive sample $s(1)$ instead of the sample $s(0, 1:B)$. Protection against a certain type of error in the model $\xi(0, 1:B)$ is gained, and some efficiency under this model is lost. If we now decide to impose the additional conditions $\sum_s B_k^j/m=\sum_1^M B_k^j/M$, $j=2, 3, \ldots, J$, thereby insuring the unbiasedness of our ratio estimator under any model of the form $\xi(r_0, r_1, \ldots, r_J:v(B))$ (and insuring the *optimality* of our estimator under any model $\xi(r_0, 1, r_2, \ldots, r_J:B)$ or $\xi(1, r_1, r_2, \ldots, r_J:1)$), we incur no additional loss in efficiency. Protection against many types of error in the model $\xi(0, 1:B)$ has been gained at no cost in terms of *additional* loss of efficiency under $\xi(0, 1:B)$.

Some rough idea of the cost of such protection can be gained by looking at a population in which the B_k are uniformly distributed over the interval $(a, a(1+\Delta))$ for any a, $\Delta \geqslant 0$. In this case, for $m < M$,

$$\min_s \left(\sum_{\tilde{s}} B_k / M-m \right) \Big/ \left(\sum_s B_k/m \right) = 1 - \frac{\Delta}{2+\Delta\left(2-\frac{m}{M}\right)}.$$

When Δ is very small, this ratio is nearly 1. When $\Delta=1$, so that the largest B_k is twice the smallest, the ratio is between $2/3$ and $3/4$. When $\Delta=2$, the ratio is between $1/2$ and $2/3$.

Suppose now that for some known or unknown characteristic C_k of unit k, the regression function ET_k contains a term $\beta' g(C_k)$ for some arbitrary function g. The ratio estimator incurs no bias from such a term if the sample s is such that

$$\frac{\sum_s g(C_k)}{\sum_s B_k} = \frac{\sum_1^M g(C_k)}{\sum_1^M B_k}.$$

If a defensive sample is drawn so that $\sum_1^M B_k/M = \sum_s B_k/m$, then the term $\beta' g(C_k)$ contributes no bias if the sample is "representative" in the sense that $\sum_s g(C_k)/m = \sum_1^M g(C_k)/M$, i.e., if the average value of $g(C_k)$ in the sample is the same as the average value in the population.

The foregoing results provide some theoretical support for the procedure of selecting a sample at random and using either the simple expansion estimator or the ratio estimator. The average value of $\sum_s B_k^j/m$ over all $\binom{M}{m}$ samples s is $\sum_1^M B_k^j/M$ for $j=1, 2, \ldots$. In precisely the same sense that the mean of a simple random sample can be expected to be approximately equal to the population mean, a sample selected at random can be expected to approximate $s(1)$. This is true because s is $s(1)$ when $\sum_s B_k/m = \sum_1^M B_k/M$. The same reasoning applied to higher powers of B implies that simple random sampling will frequently produce a sample which is a fair approximation to $s(J)$ for some $J \geqslant 1$. Whenever this occurs, the expansion and ratio estimators are approximately the same; both are approximately unbiased under the model $\xi(1, 1, \ldots, 1 : v(B))$ and approximately optimal under this model when $v(B)=B$ or $v(B)=1$. The same argument applies to the unobservable (or simply unobserved) regressor $g(c)$. Unbiased estimation of T is possible only if the effect of $g(c)$ is negligible or if the sample is "representative" in the sense defined above.

An important role of random sampling is to provide samples which are "representative" with respect to such regressors. Of course random sampling cannot *guarantee* successful choice of a representative sample, and the probability of a successful choice depends on the unknown distribution of $g(c)$ in the population. Nevertheless, random sampling provides a basis for optimism, as shown by the Tchebycheff inequality.

The use of simple random sampling as a means of obtaining a sample approximating $s(J)$ produces samples which are not all good approximations to $s(J)$ and are, on the average, less efficient than $s(J)$. Under the model $\xi(0, 1{:}B)$, for a given sample s the MSE of the ratio estimator is $\sigma^2 \sum_{\tilde{s}} B_k \sum_1^M B_k / \sum_s B_k$. The average value of this quantity over all $\binom{M}{m}$ possible samples of the required size is $\sigma^2 \sum_1^M B_k \left(\frac{c}{m} \sum_1^M B_k - 1\right)$ where c is the average value of $1/\left(\sum_s B_k/m\right)$ over all $\binom{M}{m}$ samples. Now a well-known inequality[a] shows that c is greater than or equal to $1/\left(\sum_1^M B_k/M\right)$, with equality only in case $B_k = B_l$ for all $k, l = 1, \ldots, M$. Thus, except when all B's are equal, the average MSE over all possible samples of size m is greater than $\frac{M^2}{m}\left(1-\frac{m}{M}\right)\sigma^2 \bar{\bar{B}}$, the MSE when the sample is $s(J)$.

We have seen that the strategy $(s(J), \hat{T}(0, 1{:}B))$ produces unbiased estimates under the model $\xi(1, 1, \ldots 1{:}v(B))$. The strategy $(s(1, 1, \ldots, 1{:}B), \hat{T}(1, 1, \ldots, 1{:}B))$, which is optimal under $\xi(1, 1, \ldots, 1{:}B)$, also produces unbiased esti-

[a]For any nonnegative random variable X, $E\frac{1}{X} \geqslant \frac{1}{EX}$. Equality holds if and only if X is a constant (with probability one).

mates under this model. The MSE for this strategy is the minimum value over all s of

$$E\{(\hat{T}(1, 1, \ldots, 1{:}B) - T)^2 \mid \xi(1, 1, \ldots, 1{:}B)\}, \tag{9}$$

which is less than the value of this expression when $s=s(J)$. But when $s=s(J)$ we know from the theorem that $\hat{T}(0, 1{:}B)=\hat{T}(1, 1, \ldots, 1{:}B)$ and thus that

$$E\{(\hat{T}(1, 1, \ldots, 1{:}B)-T)^2 \mid \xi(1, 1, \ldots, 1{:}B\}$$

$$=E\{(\hat{T}(0, 1{:}B)-T)^2 \mid \xi(1, 1, \ldots, 1{:}B)\}$$

$$=E\{(\hat{T}(0, 1{:}B)-T)^2 \mid \xi(0, 1{:}B)\}.$$

Therefore, under $\xi(0, 1{:}B)$ the MSE of the strategy $(s(1, 1, \ldots, 1{:}B), \hat{T}(1, 1, \ldots, 1{:}B))$ is less than that of the strategy $(s(J), \hat{T}(0, 1{:}B))$. Nevertheless, because of the popularity and simplicity of the ratio estimator, as well as because the current HDS estimator is of the ratio type, the remainder of Part I is devoted to situations in which the ratio estimator (or a sum of ratio estimators) is to be used.

Stratification on the size variable.—We have seen that the unbiasedness of the ratio estimator can be preserved under J^{th}-order polynomial regression models by the choice of a sample $s(J)$ which is "like" the population in the sense that

$$\sum_s B_k^j/m = \sum_1^M B_k^j/M \qquad \text{for } j=1, 2, \ldots, J.$$

An alternative means of preserving the unbiasedness property employs stratification on the size (B) variable and use of a separate ratio estimate in each stratum.

The double subscript hk denotes quantities associated with unit k in stratum h. Thus B_{hk} is the size and T_{hk} the response of unit k in stratum h. The number of units in stratum h is M_h, and T_h and B_h are the *totals* for stratum h. In this notation the grand total T is expressed as

$$T=\sum_{h=1}^{H} T_h = \sum_{h=1}^{H}\sum_{k=1}^{M_h} T_{hk}$$

where H is the number of strata.

The strata are defined as follows: the M_1 smallest units form stratum 1, the next M_2 smallest units form stratum 2, etc. Thus when $h<h'$, $B_{hk} \leq B_{h'k'}$ for all $k=1, \ldots, M_h$ and $k'=1, \ldots, M_{h'}$.

A sample s_h consisting of m_h units is chosen from stratum h, and the total T_h for that stratum is estimated by

$$\hat{T}_h = \frac{\sum_{s_h} T_{hk}}{\sum_{s_h} B_{hk}} \sum_1^{M_h} B_{hk}. \tag{10}$$

Any sample s_h for which $\sum_{s_h} B_{hk}^j / m_h = \sum_1^{M_h} B_{hk}^j / M_h$ for $j=1, 2, \ldots, J$ will be referred to as $s_h(J)$. If $s_h=s_h(J)$, then $\hat{T}_h$ is an unbiased estimate of T_h under the model

$$T_{hk} = \sum_{j=0}^{J} r_j\beta_j B_{hk}^j + \epsilon_{hk}\sqrt{v(B_{hk})}$$

$$k=1,2, \ldots, M_h \tag{11}$$

where the ϵ's are independent, each with mean zero and variance σ^2, and $r_0, r_1, \ldots, r_J$ is a sequence of zeroes and ones. The case in which (11) applies for $h=1,2, \ldots, H$ will, as previously, be denoted by

$$\xi\,(r_0, r_1, \ldots, r_J{:}v(B)).$$

From the earlier results we see that the estimator $\hat{T}=\sum_h \hat{T}_h$ is unbiased with MSE

$$\sigma^2 \sum_{h=1}^{H} B_h \frac{\sum_{\bar{s}_h} B_{hk}}{\sum_{s_h} B_{hk}} \tag{12}$$

under $\xi(0, 1{:}B)$. The estimator is unbiased and has the MSE (12) under the more general model $\xi(r_0, r_1, \ldots, r_J{:}B)$ if $s_h=s_h(J)$ for $h=1, \ldots, H$. Note that when such a sample is chosen, the estimator $\hat{T}$ becomes simply

$$\sum_{h=1}^{H} (M_h \sum_{s_h} T_{hk} / m_h)\,,$$

and the MSE is

$$\sigma^2 \sum_{h=1}^{H} \frac{M_h^2}{m_h} \left(1 - \frac{m_h}{M_h}\right) \bar{B}_h \tag{13}$$

where $\bar{B}_h = B_h / M_h$.

If a defensive sample is drawn within each stratum, i.e., if $s_h = s_h(J)$ for some $J \geqslant 1$ and every $h = 1, 2, \ldots, H$, then with proportional allocation (m_h/M_h constant) the estimator $\hat{T} = \sum_h \hat{T}_h$ is simply the overall expansion estimator

$$M \sum_{h=1}^{H} \sum_{s_h} T_{hk}/m.$$

In this case the MSE (13) becomes simply $\sigma^2 \bar{B} M(M-m)/m$, where $\bar{B}$ denotes the grand average

$$\sum_{h=1}^{H} B_h/M.$$

Optimal allocation, subject to fixed total sample size m and with defensive sampling within strata, is easily seen to require that m_h be proportional to $M_h\sqrt{\bar{B}_h} = \sqrt{M_h B_h}$ for $h = 1, \ldots, H$ (cf. Cochran [9]). With optimal allocation the MSE (13) becomes

$$\sigma^2 \left[\frac{1}{m} \left(\sum_{h=1}^{H} \sqrt{M_h B_h} \right)^2 - \sum_{h=1}^{H} B_h \right].$$

In order that proportional allocation be optimal, we must have M_h proportional to $M_h\sqrt{\bar{B}_h}$, which means that $\bar{B}_h$ must be constant. But with stratification on the size (B) variable, $\bar{B}_h$ can be constant only in the degenerate case of a population whose units are all of the same size. Thus in nontrivial cases, proportional allocation cannot be optimal.

The foregoing results establish the superiority, with respect to MSE, of the stratification procedure to the nonstratified defensive sampling procedure. We refer to these procedures as II and I, respectively, and summarize the argument establishing the superiority of the former.

Procedure I. Choose any $s(J)$ and use the estimator (2).

Procedure II. Stratify on the size variable, choose any $s_h(J)$ from the h^{th} stratum, and use the estimator $\hat{T} = \sum_{h=1}^{H} \hat{T}_h$.

(i) Both procedures (I and II) produce estimators which are linear and unbiased under $\xi(r_0, r_1, \ldots, r_J : v(B))$ for any sequence $r_0, r_1, \ldots, r_J$ of zeroes and ones.

(ii) Optimal allocation for procedure II requires that m_h be proportional to $\sqrt{M_h B_h}$.

(iii) If proportional allocation is used in procedure II ($m_h = mM_h/M$), then procedures I and II have the same MSE.

(iv) Proportional allocation cannot be optimal except in trivial cases.

From (i)–(iv) we conclude that

(v) If optimal allocation is used, then procedure II has smaller MSE than procedure I.

Note that (v) is true regardless of the number and relative sizes of the strata. The only requirement is that it be possible to use optimal allocation and defensive sampling. Now the same argument, (i)–(v), which shows procedure II to be superior to procedure I can be applied *within* any stratum to which more than one observation is allocated. It pays to substratify. This implies that when $J = 1$, the optimal number of strata is $H = m$ (and optimal allocation is $m_h = 1$, $h = 1, \ldots, m$). Of course, if we want to guarantee unbiasedness under more general models (larger J), each m_h must be greater than or equal to J, because when m_h is less than J it is impossible to select $s_h = s_h(J)$ except in highly special, degenerate cases. There are also the obvious problems encountered in selecting samples $s_h(J)$; exact satisfaction of $\sum_{s_h} B_{hk}^j / m_h = \sum_{1}^{M_h} B_{hk}^j / M_h$ for $j = 1, 2, \ldots, J$ is ordinarily impossible. When the sampling fraction is small, however, and J is small, approximate satisfaction of these conditions is frequently easy to effect.

Note that balanced sampling within strata provides an unbiased estimator under the more general model in which β_j varies from stratum to stratum. Such a model, even when J is small, say $J = 1$, is frequently a good approximation to a model containing a quite general regression function. That is, when the intervals of B-values which define strata are narrow, a straight-line approximation within each stratum can provide a close fit to a general, smooth regression function.

We see, then, that when optimal allocation (m_h proportional to $M_h\sqrt{\bar{B}_h}$) and defensive sampling within strata ($s_h = s_h(J)$ for all h) are employed, stratification produces smaller MSE's than simple defensive sampling ($s=s(J)$). The next problem is that of finding good rules for stratifying a population on the size variable. For $J=1$ the optimal number of strata is m with one-sample-unit-per-stratum allocation. How should the $m-1$ boundaries which define the m strata be chosen? We consider a slightly more general problem: For a fixed number H of strata, given that equal allocation ($m_h = c$, $h=1, \ldots, H$) and defensive sampling within strata are to be used, how should stratum boundaries be chosen? Under these conditions, the MSE under any model of the form $\xi(r_0, r_1, \ldots, r_J : B)$, and in particular under the model of most interest, $\xi(0, 1 : B)$, is

$$E_\xi(\hat{T}-T)^2=\sigma^2\sum_{h=1}^{H}\frac{M_h^2}{c}(1-\frac{c}{M_h})\,\bar{B}_h$$

$$=\sigma^2\left[\sum_{h=1}^{H}\frac{M_hB_h}{c}-M\bar{B}\right]. \tag{14}$$

Thus optimal stratification for equal allocation requires minimization of

$$\sum_{h=1}^{H} M_hB_h\,.$$

It should be noted that optimal stratification for equal allocation is not necessarily obtained by stratifying in such a way that equal allocation is optimal. Equal allocation is optimal when all $M_h(\bar{B}_h)^{1/2}$ are equal, that is, when all $M_h\ B_h$ are equal. But it is easy to produce examples in which this way of stratifying does not minimize

$$\sum_{h=1}^{H} M_hB_h\,.$$

It can be shown (a proof is contained in the appendix) that for a given stratification scheme to be optimal when equal allocation is used, it is necessary that

$$M_1 \geqslant M_2 \geqslant \ldots \geqslant M_H. \tag{15}$$

This is established by demonstrating that for any $h=1, \ldots, H-1$, whenever $M_h < M_{h+1}$, the MSE (14) is reduced if the strata are redefined so that the smallest unit in stratum $h+1$ is shifted into stratum h. It is also true that, except for one special situation, for a given stratification scheme to be optimal it is necessary that

$$B_1 \leqslant B_2 \leqslant \ldots \leqslant B_H. \tag{16}$$

(See appendix.) The exceptional situation can occur only when two adjacent strata, say the h^{th} and the $h+1^{st}$, have $M_h = M_{h+1}+1$, and all of the $2M_h - 1$ units in these two strata are of approximately the same size. Then we can have $B_h > B_{h+1}$, but if we attempt to satisfy (16) by shifting a unit from the h^{th} stratum to the $h+1^{st}$, the MSE (14) is increased. Note that in this case, shifting the unit introduces violation of the inequality $M_h \geqslant M_{h+1}$.

The inequalities (15) and (16) indicate the essential features of a good stratification scheme for use with equal allocation, defensive sampling within strata, and the estimator (10). The strata should be so constructed that there are more units in stratum h than in stratum $h+1$, but there should not be so *many* more units in stratum h that the sum of the size measures in stratum h exceeds the corresponding sum in stratum $h+1$. Three special cases in which both (15) and (16) are satisfied are:

(i) $M_1 = M_2 = \ldots = M_H$,
(ii) $B_1 = B_2 = \ldots = B_H$, and
(iii) $M_1B_1 = M_2B_2 = \ldots = M_HB_H$.

In case (i) equal allocation is proportional allocation, while in case (iii) equal allocation is optimal allocation. In an obvious sense, (i) and (ii) represent two extremes among all stratification schemes consistent with (15) and (16), with all others, for example (iii), located between these extremes. It appears that the relative efficiency of (iii), with respect to the optimal scheme, is ordinarily quite nearly one. (Cf. Cochran [10].)

Of course, if $J=1$, yet fewer than m strata are to be created, optimal allocation requires that m_h be proportional to $\sqrt{M_hB_h}$. With such allocation, optimal stratification is that which minimizes $\sum_{h=1}^{H}\sqrt{M_hB_h}$.

The relation between stratified random sampling, using separate expansion or ratio estimators within strata, and the present results concerning defensive sampling within strata is quite analogous to that, discussed earlier, between simple random sampling, using either the simple expansion or the ratio estimator, and defensive sampling. The average value of $\sum_{s_h} B^j_{hk} / m_h$ over all samples s_h of size m_h from stratum h is $\sum_1^{M_h} B^j_{hk} / M_h$. Thus we should not be surprised to find that stratified random sampling frequently produces samples in which s_h is approximately $s_h(J)$ for some $J \geqslant 1$. When this occurs the estimator is approximately (10), regardless of whether the ratio or the expansion formula is used, and is approximately unbiased under rather general models. The random sampling procedure chooses samples which are, on the average, less efficient than the nonrandom defensive strategy, as was shown to be the case when simple random sampling and defensive sampling were compared.

Estimation of Variance

A detailed study of variance estimation when using stratification has not been attempted here. In this section the stratum subscript h is dropped, and results are stated for an unstratified population. Of course, an example of such a population is an individual stratum in a stratified population. In this section, then, the bed size and the response associated with unit k are denoted by B_k and T_k, respectively.

Unbiased variance estimation.—Under a particular model $\xi(r_0, r_1, \ldots, r_J\!: v(B))$, the MSE $E_\xi(\hat{T}-T)^2$ is a measure of how much inaccuracy might be expected when $\hat{T}$ is used as an estimate of T. If $\hat{T}$ is unbiased under the model, then the MSE is simply the variance of the error $\hat{T}-T$. For any linear estimator $\hat{T}=\sum_s \ell_k T_k$, this variance is easily calculated:

$$\begin{aligned}\mathrm{Var}\left(\sum_s \ell_k T_k - \sum_1^M T_k\right) &= \mathrm{Var}\left(\sum_s (\ell_k - 1) T_k + \sum_{\tilde{s}} T_k\right)\\ &= \sigma^2 \left\{\sum_s (\ell_k - 1)^2 v(B_k) + \sum_{\tilde{s}} v(B_k)\right\}. \qquad (17)\end{aligned}$$

Unbiased estimation of this variance requires simply that an unbiased estimate of σ^2 be substituted for this unknown constant in (17) since for a given estimator, sample, and model, the rest of (17) is fixed and known.

When using the BLUE estimator $\hat{T}(r_0, r_1, \ldots, r_J\!: v(B))$, the usual estimator of σ^2, which is based on the weighted least-squares residuals, is unbiased:

$$\hat{\sigma}^2(r_0, r_1, \ldots, r_J\!:v(B)) = \frac{1}{m-c}\sum_s \left\{T_k - \hat{T}_k(r_0, r_1, \ldots, r_J\!:v(B))\right\}^2 / v(B_k) \qquad (18)$$

where

$$\hat{T}_k(r_0, r_1, \ldots, r_J\!:v(B)) = \sum_{j=0}^{J} r_j \hat{\beta}(r_0, r_1, \ldots, r_J\!:v(B)) B_k^j$$

and $c = \sum_{j=0}^{J} r_j$, the number of regression coefficients estimated.

Under the model $\xi(0, 1\!:\!B)$, this estimate of σ^2 is given by

$$\hat{\sigma}^2(0, 1\!:\!B) = \frac{1}{m-1}\sum_s \left\{T_k - \frac{\sum_s T_k}{\sum_s B_k} B_k\right\}^2 / B_k. \qquad (19)$$

and the MSE for a given s is estimated by

$$\hat{\sigma}^2(0, 1 : B)\, \frac{\sum_{\tilde{s}} B_k}{\sum_s B_k} \sum_1^M B_k. \qquad (20)$$

This statistic is not the estimate of the variance of the ratio estimate, which is usually used when s is selected by simple random sampling. As an indication of the inaccuracy in an observed estimate $\hat{T}$, (20) seems usually to be superior to the conventional variance estimate. For example, confidence intervals with width proportional to the square root of (20) are frequently more accurate indicators of the uncertainty in an observed estimate than are the same intervals with (20) replaced by the usual variance estimate. This point is discussed in Royall [1], where some theoretical results and empirical evidence are presented.

Effects of errors in the model.—Under the more general model with regression function $h(B)$ and variance function $v(B)$, i.e., $T_k = h(B_k) + \epsilon_k \sqrt{v(B_k)}$, the MSE of the ratio estimator is actually

$$E(\hat{T}-T)^2 = \sigma^2 \left[\sum_s v(B_k) \left(\frac{\sum_{\tilde{s}} B_k}{\sum_s B_k} \right)^2 + \sum_{\tilde{s}} v(B_k) \right] + \left[\frac{\sum_{\tilde{s}} B_k}{\sum_s B_k} \sum_s h(B_k) - \sum_{\tilde{s}} h(B_k) \right]^2. \tag{21}$$

The first term in (21) is the variance of the difference $\hat{T} - T$, and the second is the square of the bias $E(\hat{T} - T)$. Under this general model the *estimate* (20) of the MSE has expected value

$$(\sigma^2 V + D) \sum_1^M B_k \sum_{\tilde{s}} B_k / \sum_s B_k \tag{22}$$

where

$$V = \frac{1}{m-1} \left[\sum_s \frac{v(B_k)}{B_k} - \frac{\sum_s v(B_k)}{\sum_s B_k} \right]$$

and

$$D = \frac{1}{m-1} \left[\sum_s \frac{h^2(B_k)}{B_k} - \frac{\left(\sum_s h(B_k) \right)^2}{\sum_s B_k} \right].$$

Note that (22), like (21), is naturally represented as the sum of two nonnegative terms, the first depending on the variance function and σ^2 but not on the regression function, and the second depending on the regression function but not on the variance.

When $v(B) = B$ the first terms of (21) and (22) are equal, and when $h(B) = \beta B$ the second terms in both expressions vanish. In particular, under the model $\xi(0, 1 : B)$ the two expressions are equal and (20) is an unbiased estimate of $E(\hat{T}-T)^2$.

If a defensive sample $s(J)$ is chosen, the actual MSE under $\xi(r_0, r_1, \ldots, r_J : v(B))$ is

$$\sigma^2 \frac{M^2}{m} \left(1 - \frac{m}{M}\right) \left[\left(1 - \frac{m}{M}\right) \sum_s v(B_k)/m + \frac{m}{M} \sum_{\tilde{s}} v(B_k)/M - m \right],$$

which can be rewritten as

$$\sigma^2 \frac{M^2}{m} \left(1 - \frac{m}{M}\right) \left[\frac{1}{m} \sum_s v(B_k) + \frac{m}{M-m} \left(\frac{1}{M} \sum_1^M v(B_k) - \frac{1}{m} \sum_s v(B_k) \right) \right]. \tag{23}$$

In this case the estimate (20) of this MSE has expected value

$$\sigma^2 \frac{M^2}{m} \left(1 - \frac{m}{M}\right) \bar{B} \left\{ V + \frac{1}{\sigma^2} D \right\}. \tag{24}$$

Note that when $v(B)$ is a J^{th}-order polynomial and $s = s(J)$, the MSE is simply

$$\sigma^2 \frac{M^2}{m} \left(1 - \frac{m}{M}\right) \frac{1}{M} \sum_1^M v(B_k).$$

The choice of $s = s(J)$ protects the ratio estimator against bias in case a J^{th}-order polynomial regression model applies. This protection does not extend to the estimate (20) of the *variance* of the ratio estimator, whose expected value depends, through the quantity D, on the regression coefficients. If the model $\xi(0, 1 : B)$ is *correct* with respect to its specification of the variance function $v(B) = B$ but erroneous in its specification of the regression function, the variance estimate is biased by the amount ((24) minus (23) with $v(B) = B$):

$$\frac{M^2}{m} \left(1 - \frac{m}{M}\right) \bar{B} D. \tag{25}$$

The Cauchy-Schwarz inequality shows that D is nonnegative. Thus when we use the estimates (2) and (20) which are appropriate under $\xi(0, 1:B)$, we choose a defensive sample $s(J)$, and the true model is $\xi(r_0, r_1, \ldots, r_J:B)$ with $r_j = 1$ for some $j \neq 1$, we encounter a *positive* bias in our estimate of variance.

When the model $\xi(0, 1:B)$ is *correct* in its specification of the regression function but erroneous in its variance function, how does the actual MSE (21) compare to the expected value (22) of our estimate? If in fact the true model is $\xi(0, 1:v(B))$, then the expected value of the ratio of the estimated MSE to the actual MSE, the ratio of (22) to (21), can be written

$$1+\frac{e_1-e_2}{1+e_2}$$

where

$$e_1=\frac{m}{m-1}\left(\frac{\sum_s B_k}{\sum_s v(B_k)}\,\frac{1}{m}\sum_s \frac{v(B_k)}{B_k}-1\right)$$

and

$$e_2=\left\{\frac{\sum_1^M v(B_k)\,\sum_s B_k}{\sum_s v(B_k)\,\sum_1^M B_k}-1\right\}\Bigg/\left\{\frac{\sum_1^M B_k}{\sum_{\tilde{s}} B_k}-1\right\}. \tag{26}$$

When $v(B)=B^2$, the actual MSE is no less than

$$\frac{M^2}{m}\left(1-\frac{m}{M}\right)\sigma^2\left(\frac{\sum_{\tilde{s}} B_k}{M-m}\right)^2,$$

while the expected value of the estimate is no greater than

$$\frac{M^2}{m}\left(1-\frac{m}{M}\right)\sigma^2\left(\frac{\sum_{\tilde{s}} B}{M-m}\right)\left(\frac{\sum_1^M B}{M}\right).$$

The ratio of (22) to (21) is thus no greater than $\left(\sum_1^M B_k/M\right)\Big/\left(\sum_{\tilde{s}} B_k/M-m\right)$. It is equal to this value only in the degenerate case of all B_k equal. Thus when $v(B)=B^2$, with defensive sampling the variance estimator has a negative bias.

When the sample is such that

$$\left(\sum_1^M B_k/M\right)=\sum_s B_k/m$$

and $v(B)=1$, the ratio of (22) to (21) is no less than

$$\frac{1}{m-1}\left(\frac{1}{m}\sum_s B_k\sum_s\frac{1}{B_k}-1\right),$$

which is no less than 1. Thus when $s=s(J)$, and our model $\xi(0,\ 1:B)$ is erroneous in that the actual variance function is not $v(B)=B$ but instead $v(B)=1$, our estimator of the variance of the ratio estimator has a positive bias.

PART II. TWO-STAGE SAMPLING

Description of Problem

Terminology, notation.—In Part I a simple population of M units with associated size measures $B_1, \ldots, B_M$ was considered. For present purposes the units are hospitals and the size measures are their bed sizes as measured by *MFI* in 1963. The basic sampling unit in *HDS* is a patient discharge record. On this record the variable of interest, Z, is found. Thus for $k=1, \ldots, M$

- B_k is the bed size of hospital k;
- N_k is the number of discharges from hospital k during the period studied;
- $Z_{k\ell}$ is a number associated with discharge $\ell(\ell=1, 2, \ldots, N_k)$ from hospital k; and
- T_k is the sum, over all discharges from hospital k, of the Z-values:

$$T_k=\sum_{\ell=1}^{N_k} Z_{k\ell}.$$

The sample is selected in two stages. First a sample s of m hospitals is chosen; then, if hospital k was selected, N_k is observed and a sample s_k of discharges is selected from hospital k. The number of discharges in the second-stage sample is n_k. The samples s and s_k are represented as subsets of the sets $\{1, 2, \ldots, M\}$ and $\{1, 2, \ldots, N_k\}$, respectively. The expression "k in s" means that hospital k is in the sample of hospitals, and "ℓ in s_k" means that discharge ℓ is in the sample of discharges from hospital k.

The objective is to estimate the total,

$$T=\sum_{k=1}^{M}\sum_{\ell=1}^{N_k} Z_{k\ell}=\sum_{k=1}^{M} T_k,$$

which can also be expressed as

$$T=\sum_{s}\sum_{s_k} Z_{k\ell}+\sum_{s}\sum_{\tilde{s}_k} Z_{k\ell}+\sum_{\tilde{s}}\sum_{1}^{N_k} Z_{k\ell} \qquad (27)$$

where $\tilde{s}$ is the set of hospitals not in the sample, and $\tilde{s}_k$ is the set of discharges from hospital k which are not in the sample s_k. The first term in (27) is known from the sample; the second and third terms must be estimated.

Most but not all discharges from the M hospitals are within the scope of HDS. Thus a discharge record which has been selected for the sample might be found to be either (i) out of scope or (ii) in scope but nonresponding (e.g., missing from its folder or lacking necessary information). These possibilities will be considered later, but for the moment attention is confined to the simplified case of all discharges in scope and 100 percent response. In this case the analogue of the HDS estimator [11] is

$$\hat{T}=\frac{\sum_s \hat{T}_k}{\sum_s B_k}\sum_{1}^{M} B_k \qquad (28)$$

where $\hat{T}_k=N_k\sum_{s_k} Z_{k\ell}\Big/n_k$.

This estimator can also be written as

$$\hat{T}=\sum_s\sum_{s_k} Z_{k\ell}+\sum_s (N_k-n_k)\left(\sum_{s_k} Z_{k\ell}\Big/n_k\right)$$

$$+\frac{\sum_s N_k\left(\sum_{s_k} Z_{k\ell}\Big/n_k\right)}{\sum_s N_k}\,\frac{\sum_s N_k}{\sum_s B_k}\sum_{\tilde{s}} B_k. \qquad (29)$$

Here the first term is that part of T which is known, the second term estimates the sum of the unobserved discharges from sample hospitals, and the third estimates the sum for nonsample hospitals. (Compare with expression (27).)

Probability models.—The number of discharges N_k is treated as a random variable whose expected value and variance are proportional to B_k: $EN_k=\beta B_k$ and $\operatorname{Var} N_k=\sigma_N^2 B_k$, with N_k and N_j uncorrelated for $j\neq k$. For a given value of N_k, the responses $Z_{k\ell}$, $\ell=1, 2, \ldots, N_k$ are treated as exchangeable random variables. That is, all of the permutations of $Z_{k1}, Z_{k2}, \ldots, Z_{kN_k}$ have the same joint probability distribution. Thus these random variables have a common mean θ_k and variance σ_k^2; all the pairs $(Z_{k\ell}, Z_{kj})$ have the same covariance $\rho_k\sigma_k^2$.

Although ρ_k and σ_k^2 are treated here as constants (not depending on N_k), they might be more realistically represented as functions of N_k and B_k. For example, if the sum $\sum_{l}^{N_k} Z_{k\ell}$ is fixed, then exchangeability of the Z's implies that, given N_k, cov $(Z_{k\ell}, Z_{k\ell'}) = -\operatorname{Var}(Z_{k\ell})/(N_k - 1)$ for $\ell \neq \ell'$. Thus if σ_k^2 is fixed, $\rho_k = -1/(N_k - 1)$. What functions might represent the relation between σ_k^2, ρ_k and N_k, B_k with useful accuracy and whether such representations have a nonnegligible influence on the analysis are questions which call for further investigation, both theoretical and empirical.

The expected values $\theta_1, \theta_2, \ldots, \theta_M$ associated with M hospitals are themselves treated as realized values of random variables $\Theta_1, \Theta_2, \ldots, \Theta_M$, which are uncorrelated and have a common mean value θ and variance τ^2. The random variables Θ_k and N_j are uncorrelated for all $k, j = 1, 2, \ldots, m.$[b]

Optimality Considerations

If all the N_k were observed and if θ_k, for k in s, and θ were known, then the best unbiased estimator of T would clearly be

$$\sum_{s}\sum_{s_k} Z_{k\ell} + \sum_{s} (N_k - n_k)\theta_k + \theta \sum_{\tilde{s}} N_k.$$

If all N_k were observed but the θ_k and θ were unknown, then the best linear unbiased estimator of T would be

$$\sum_{s}\sum_{s_k} Z_{k\ell} + \sum_{s} (N_k - n_k)\hat{\theta}_k + \hat{\theta} \sum_{\tilde{s}} N_k \qquad (30)$$

where

$$\hat{\theta}_k = \frac{\tau^2 \sum_{s_k} Z_{k\ell}/n_k + \hat{\theta}\sigma_k^2(1 + (n_k - 1)\rho_k)/n_k}{\tau^2 + \sigma_k^2(1 + (n_k - 1)\rho_k)/n_k}$$

[b] That the Θ_k have the same mean is an assumption whose plausibility is specific to a particular characteristic Z under consideration. For a characteristic such as length of stay, the expected value of Θ_k is probably dependent on B_k and is thus not the same for all hospitals. Sensitivity of subsequent results to deviation from the assumption of a common mean for the Θ_k requires further investigation. At present it is uncertain as to how much deviation from this assumption can be safely disregarded.

and

$$\hat{\theta} = \sum_{s} \frac{\sum_{s_k} Z_{kl}/n_k}{\tau^2 + \sigma_k^2(1 + (n_k - 1)\rho_k)/n_k} \Bigg/ \sum_{s} \frac{1}{\tau^2 + \sigma_k^2(1 + (n_k - 1)\rho_k)/n_k}.$$

This estimator was obtained in a Bayesian analysis for the case $\rho = 0$ by Scott and Smith [7]. They showed that, given the observations, (30) is the expected value of T when all the distributions concerned are normal and θ is itself given a uniform distribution over the entire real line. They also showed that under the present model (θ fixed), (30) is the best among linear estimators whose mean square errors are bounded functions of θ.

It might seem objectionable to estimate the parameter θ_k for a sample hospital, not by the mean $\sum_{s_k} Z_{kl}/n_k$ of the sample from that hospital, but instead by a weighted average of this statistic and $\hat{\theta}$, where $\hat{\theta}$ depends on the samples drawn from other hospitals. However, considering the case of θ known and n_k small will make it clear that such an estimator is quite reasonable under the present model.

In (29) the expression $\left(\sum_{s} N_k \Big/ \sum_{s} B_k\right) \sum_{\tilde{s}} B_k$ estimates $\sum_{\tilde{s}} N_k$. Thus in the case of N_k *known* for all $k = 1, 2, \ldots, M$, the analogue of the HDS estimator (29) is

$$\frac{\sum_{s} \hat{T}_k}{\sum_{s} N_k} \sum_{1}^{M} N_k =$$

$$\sum_{s}\sum_{s_k} Z_{kl} + \sum_{s} \left(N_k - n_k\right) \left(\sum_{s_k} Z_{kl}/n_k\right)$$

$$+ \frac{\sum_{s} N_k \left(\sum_{s_k} Z_{kl}/n_k\right)}{\sum_{s} N_k} \sum_{\tilde{s}} N_k. \qquad (31)$$

If the three conditions (i) $\rho_k=0$ for all k in s, (ii) $\tau^2=0$ (no variability among the expected values $\theta_1, \ldots, \theta_M$), and (iii) n_k/N_k=constant for all k in s (proportional allocation) are met, then (30) and (31) are the same—in this simplified problem the analogue of the HDS estimator is the BLUE estimator. If $\sigma^2=0$ the two estimators differ only in their third terms. Even if the two formulas differ, they produce the same estimate when the sample is such that the sample means $\sum_{s_k} Z_{kl}/n_k$, k in s, are all equal; they produce approximately the same estimate when the means are approximately equal. The analogue (31) of the HDS estimator is thus approximately optimal when the hospital sample means show little variability, as well as when (i)–(iii) are satisfied.

When, as is the case in practice, the N_{hk}, for k not in s, are unknown, the estimator obtained by replacing $\sum_{\tilde{s}} N_k$ in (30) by its BLUE estimator

$$\sum_{\tilde{s}} B_k \sum_{s} N_k \Big/ \sum_{s} B_k \text{ is}$$

$$\sum_{s}\sum_{s_k} Z_{kl}+\sum_{s} (N_k-n_k)\hat{\theta}_k+\hat{\theta}\, \frac{\sum_{s} N_k}{\sum_{s} B_k} \sum_{\tilde{s}} B_k. \quad (32)$$

Using the same estimate for $\sum_{\tilde{s}} N_k$ in (31) gives (29), the analogue of the HDS estimator for this case. The conditions for equivalence of (32) and (29) are again (i)–(iii), and, as before, the two formulas produce the same estimate when the within-hospital sample means are all equal.

The estimator (32) is calculable only when all of the ratios $n_k\tau^2/\sigma_k^2$, k in s, are known. For some response variables it may be known that these ratios are all quite large (or small), in which case an approximately optimal estimator can be calculated. For general values of the ratios when the n_k and m are large, an approximately optimal estimator can be obtained by substituting estimates of the ratios for their actual values. This approach is not developed here. Instead the HDS-type estimator is considered, and questions of unbiasedness, stratification, and allocation are studied.

The HDS Estimator

Case of all discharges in scope and 100 percent response.—The HDS design is stratified, and the actual estimator is the sum of estimates of the form (28). The stratification variable is bed size. Suppose the hospitals are divided according to bed size into H strata (H can be 1), and let M_h denote the number of hospitals in stratum h. Now for $h=1, 2, \ldots, H$ and $k=1, 2, \ldots, M_h$

B_{hk} is the bed size of hospital k, stratum h;

N_{hk} is the total number of discharges from hospital k, stratum h, for $l=1, 2, \ldots, N_k$;

Z_{hkl} is the response variable associated with discharge l from hospital k in stratum h;

$T_{hk}=\sum_{l=1}^{N_h} Z_{hkl}$ is the total for hospital k, stratum h; and

$T_h=\sum_{k=1}^{M_h} T_{hk}$ is the total for stratum h.

The underlying model is as before, except for obvious notational changes to indicate strata.

For the present, attention is confined to the simplified problem with all discharges in scope and 100 percent response. The HDS estimator for this case is

$$\sum_{h=1}^{H}\left(\sum_{k=1}^{M_h} B_{hk}\right)\left(\sum_{k\epsilon s_h} \hat{T}_{hk} \Big/ \sum_{k\epsilon s_h} B_{hk}\right). \quad (33)$$

Where $\hat{T}_{hk}=N_{hk}\sum_{s_{hk}} Z_{hkl}/n_{hk}$, s_h is the sample of m_h hospitals from stratum h, and s_{hk} is the sample of n_{hk} discharges from hospital k in stratum h. This estimator has the form $\sum_{h}^{H} \hat{T}_h$ where

$$\hat{T}_h=\left(\sum_{k=1}^{M_h} B_{hk}\right)\left(\sum_{s_h} \hat{T}_{hk} \Big/ \sum_{s_h} B_{hk}\right)$$

is a ratio-type estimator for the stratum h total T_h.

a. *Condition for unbiasedness.*—The condition for unbiasedness, $E(\hat{T}-T)=0$, applied to the estimator (33) is equivalent to

$$\sum_{h=1}^{H}\left(\sum_{1}^{M_h} B_{hk}\right)\left(\sum_{s_h} E(\hat{T}_{hk}) \Big/ \sum_{s_h} B_{hk}\right)$$

$$=\sum_{h=1}^{H}\sum_{1}^{M_h} E(T_{hk}). \qquad (34)$$

Suppose $E(\hat{T}_{hk})=E(T_{hk})$ for all k in s_h. Under any model for which this is true, (34) is an unbiased estimator of the grand total T if within each stratum the sample is "representative" in the sense that the ratio of the total expected value $E\left(\sum_{s_h} T_{hk}\right)$ to total beds $\sum_{s_h} B_{hk}$ in the sample is the same as the corresponding ratio for the entire stratum.

If $E(T_{hk})$ is a J^{th}-degree polynomial in B_{hk}, the earlier results regarding defensive sampling apply. The estimator (33) is unbiased if a defensive first-stage sample $s_h(J)$ is chosen for $h=1, 2, \ldots, H$, and $E\hat{T}_{hk}=ET_{hk}$ for all k in $s_h(J)$. If $E\hat{T}_{hk}=ET_{hk}=\beta_h B_{hk}$ for some constants β_h, then (33) is unbiased for any choice of the first-stage samples $s_1, s_2, \ldots, s_H$. This result applies to the present model since $ET_{hk}=E\sum_{1}^{N_{hk}} Z_{hkl}=EN_{hk}\Theta_{hk}=\theta\beta B_{hk}$ and $E\hat{T}_{hk}=E\left(N_{hk}\sum_{s_{hk}} Z_{hkl}\Big/ n_{hk}\right)=EN_{hk}\Theta_{hk}=\theta\beta B_{hk}$.

b. *Variance.*—Under the present model the HDS estimator (33) is unbiased with

$$\operatorname{Var}\hat{T}=\sum_{h=1}^{H}\operatorname{Var}(T_h).$$

Using only the conditions that: (i) given N_{hk} and θ_{hk}, the variables Z_{hkl} $\ell=1, \ldots, N_{hk}$ are exchangeable and (ii) $N_{hk}\Theta_{hk}$ $k=1, \ldots, M_h$ are uncorrelated, it can be shown that the error variance for stratum h is

$$\operatorname{Var}(\hat{T}_h-T_h)=\sum_{\tilde{s}_h}\operatorname{Var}(T_{hk})$$

$$+\left(\frac{\sum_{\tilde{s}} B_{hk}}{\sum_{s_h} B_{hk}}\right)^2 \sum_{s_h}\operatorname{Var}(T_{hk})$$

$$+\left(\frac{\sum_{1}^{M_h} B_{hk}}{\sum_{s_h} B_{hk}}\right)^2 \sum_{s_h} E[\operatorname{Var}(\hat{T}_{hk}-T_{hk}\mid N_{hk},\Theta_{hk})]. \qquad (35)$$

The sum of the first two terms in this expression is the variance of $\hat{T}_h-T_h$ if T_{hk} were *observed* for k in s_h. The third term is the increase in error variance caused by estimation of T_{hk} by $\hat{T}_{hk}$ for k in s_h and can be written in a more explicit form determined by the relation

$$E[\operatorname{Var}(\hat{T}_{hk}-T_{hk}\mid N_{hk},\Theta_{hk})]$$

$$=E\left[\frac{N_{hk}^2}{n_{hk}}\left(1-\frac{n_{hk}}{N_{hk}}\right)\sigma_{hk}^2(1-\rho_{hk})\right]. \qquad (36)$$

Expressions (35) and (36) are derived in the appendix.

c. *Design of survey. Allocation of second-stage samples within strata.*—From (36) it is easily shown that, for a given sample of hospitals s_h and a fixed total number of discharges n_h to be sampled from stratum h, the error variance is minimized when

$$n_{hk}=n_h N_{hk}[(1-\rho_{hk})\sigma_{hk}^2]^{1/2} \Big/ \sum_{s_h} N_{hk}[(1-\rho_{hk})\sigma_{hk}^2]^{1/2}.$$

If the quantities $(1-\rho_{hk})\,\sigma_{hk}^2$, k in s_h, are approximately equal, then optimal allocation is proportional allocation, $n_{hk}/N_{hk}=n_h \Big/ \sum_{s_h} N_{hk}$, for all k in s_h. Here the constant of proportionality is $n_h \Big/ \sum_{s_h} N_{hk}$. If the constant must be chosen before the denominator of this ratio is known, then the total number of observations n_h is random. Nevertheless, if the $\sigma_{hk}^2(1-\rho_{hk})$, k in s_h, are all equal, then no other scheme for allocating the n_h observations can produce a lower error variance.

Choice of hospitals within strata.—When $\sigma_{hk}^2(1-\rho_{hk})$, $k=1, \ldots, M_h$, are all equal and proportional allocation is used, the third term in (35) is a decreasing function of $\sum_{s_k} B_{hk}$. This implies that if a first-stage sample s_h for which $\sum_{s_h} B_{hk}$ is a maximum is optimal for estimating T_h in the single-stage problem (T_k *observed* for k in s_h), then it is also optimal for the two-stage problem. As in the single-stage problem, the choice of a suboptimal sample satisfying $\sum_{s_h} B_{hk}/m_h = \sum_{1}^{M_h} B_{hk}/M_h$ might be justified on the grounds that it affords protection against the errors in certain aspects of the regression model.

Allocation of second-stage samples among strata.—For a given first-stage sample and proportional allocation of the second-stage sample among sample hospitals within each stratum, the third term of the error variance is

$$\sum_{h=1}^{H}\left(\frac{\sum_{1}^{M_h} B_{hk}}{\sum_{s_h} B_{hk}}\right)^2 \sum_{s_h} E\left[N_{hk}\left(\frac{1}{\gamma_h}-1\right)(\sigma_{hk}^2(1-\rho_{hk}))\right] \tag{37}$$

Here γ_h is the sampling rate applied within sample hospitals in stratum h, i.e., $n_{hk}/N_{hk}=\gamma_h$ for all k in s_h. If the total expected number of second-stage units in the sample is fixed, say

$$E\sum_{h=1}^{H}\sum_{s_h}^{*} n_{hk} = E\sum_{h=1}^{H}\sum_{s_h}\gamma_h N_{hk} = n,$$

then what are the optimal rates γ_1^*, γ_2^*, . . ., γ_H^*? The answer is easily shown from (37) to be

$$\gamma_h^* = \lambda\frac{\sum_{1}^{M_h} B_{hk}}{\sum_{s_h} B_{hk}}\left[\frac{\sum_{s_h} E(N_{hk}\sigma_{hk}^2(1-\rho_{hk}))}{\sum_{s_h} EN_{hk}}\right]^{1/2}$$

$$h=1, 2, \ldots, H$$

where λ is a constant determined by the restriction

$$\sum_{1}^{H}\gamma_h^* E\left(\sum_{s_h} N_{hk}\right)=n.$$

If the $\sigma_{hk}^2(1-\rho_{hk})$, k in s_h, are all equal, then the γ_h^* are such that

$$\gamma_h^*\sum_{s_h} B_{hk}\Big/\sum_{1}^{M_h} B_{hk} \qquad h=1, 2, \ldots, H$$

are all equal. Note that when s_h is such that

$$\sum_{s_h} B_{hk}/m_h = \sum_{1}^{M_h} B_{hk}/M_h$$

for all h, the optimal sampling rates are determined by $\gamma_h m_h/M_h=$ constant. In other words, the optimal second-stage sampling rates are such that the overall sampling rate is the same in all strata. This is, in fact, the rule which is used in the HDS.

Optimal stratification and allocation of first-stage sample.—Suppose that for $h=1, 2, \ldots, H$:

(i) $\sum_{s_k} B_{hk}/m_h = \sum_{1}^{M_h} B_{hk}/M_h$;

(ii) a fixed sampling fraction γ_h is to be used within all sample hospitals in stratum h; and

(iii) the rates γ_1, . . ., γ_H are chosen so that the overall sampling rate is constant, i.e., $\gamma_h m_h/M_h=$ constant.

Note that the previous analysis provides some justification for (i)–(iii). Under these conditions, if the variance of T_{hk} is proportional to B_{hk} (as is true when $\tau=0$ and $\rho=0$), then the variance (35) is of the form $c_1\sum_{1}^{H} M_h B_h/m_h + c_2$ for some constants c_1 and c_2. Therefore, the problem of optimal stratification and optimal allocation of the first-stage sample (choice of m_1, m_2, . . ., m_H) is the same as in the single-stage sampling problem considered in Part I. Thus when the above conditions are approximately satisfied and it is not required that $s_h = s_h(J)$ for $J>1$ (condition

(i) means $s_h = s_h$ (1)), the optimal number of strata is m, and optimal stratification must satisfy inequalities (15) and (16). If fewer strata ($H < m$) are to be created, then optimal allocation is given by the familiar rule $m_h/(M_hB_h)^{1/2}$ = constant. Using this allocation rule, optimal stratification is achieved when $\sum_1^H (M_hB_h)^{1/2}$ is minimized.

The allocation rule used in HDS is m_h/B_h = constant. Both allocation rules, $m_h/(M_hB_h)^{1/2}$ = constant and m_h/B_h = constant, imply that the larger the average bed size B_h/M_h, the larger should be the first-stage sampling rate m_h/M_h. However, with the former rule this rate is proportional to $(B_h/M_h)^{1/2}$, while with the latter rule the rate is proportional to B_h/M_h. Thus the former rule yields a more nearly constant first-stage allocation rate than does the latter. Note that with the latter, optimal stratification requires minimization of $\sum_1^H M_h \sum_1^H B_h/m$, which does not depend on the way in which strata are formed. Thus when this rule is used, the choice of stratum boundaries appears to have little effect on the performance of the overall estimator.

Effects of out-of-scope and nonresponse discharges.—In this section it is recognized that some of the discharges from which the sample is drawn might be outside the scope of the HDS study. A two-valued variable δ is used to indicate whether or not a discharge is in scope: $\delta_{hk\ell} = 1$ if discharge ℓ from hospital k in stratum h is in scope, and $\delta_{hk\ell} = 0$ otherwise. The variables $\delta_{hk\ell}$ $\ell = 1, 2, \ldots, N_{hk}$ are treated as realized values of independent random variables, and π_{hk} is the probability that $\delta_{hk\ell} = 1$.

The response $Z_{hk\ell}$ can now be represented as the product of $\delta_{hk\ell}$ and a random variable $X_{hk\ell}$. Then the X-value is the characteristic of interest and $T_{hk} = \sum^{N_{hk}} Z_{hk\ell} = \sum_1^{N_{hk}} \delta_{hk\ell} X_{hk\ell}$. Given the number N_{hk} of total discharges, $\sum_1^{N_{hk}} \delta_{hk\ell}$ represents the random number of in scope discharges from hospital hk.

The expected response of each in-scope discharge from hospital hk is denoted by μ_{hk}; i.e., μ_{hk} is the expected value of $X_{hk\ell}$, given that $\delta_{hk\ell} = 1$. Thus $EZ_{hk\ell} = \theta_{hk} = E\delta_{hk\ell}X_{hk\ell} = \pi_{hk}\mu_{hk}$. The variances and covariances and responses of in-scope discharges are $\sigma^2_{X_{hk}}$ and $\rho_{X_{hk}}\, \sigma^2_{X_{hk}}$. Thus

$$\operatorname{Var} Z_{hk\ell} = \sigma^2_{hk} = \operatorname{Var}(\delta_{hk\ell}X_{hk\ell})$$

$$= \pi_{hk}\sigma^2_{X_{hk}} + \mu^2_{hk}\pi_{hk}(1-\pi_{hk})$$

and

$$\operatorname{Cov}(Z_{hk\ell}, Z_{hk\ell'}) = \rho_{hk}\sigma^2_{hk} = \pi^2_{hk}\rho_{X_{hk}}\sigma^2_{X_{hk}}.$$

The most direct estimator for this case, which is the analogue of the HDS estimator, is $\hat{T} = \sum_1^H \hat{T}_h$, in which

$$\hat{T}_h = \frac{\sum_{s_h} \hat{T}_{hk}}{\sum_{s_h} B_{hk}} \sum_1^{M_h} B_{hk} = \sum_{s_h}\sum_{s_{hk}} \delta_{hk\ell}X_{hk\ell} + \sum_{s_h} (N_{hk} - n_{hk})\hat{\pi}_{hk}\hat{\mu}_{hk} + \frac{\sum_{s_h} N_{hk}\hat{\pi}_{hk}\hat{\mu}_{hk}}{\sum_{s_h} B_{hk}} \sum_{\bar{s}_h} B_{hk}$$

where

$$\hat{T}_{hk} = N_{hk} \sum_{s_{hk}} \delta_{hk\ell}X_{hk\ell}/n_{hk},\ \hat{\pi}_{hk} = \sum_{s_{hk}} \delta_{hk\ell}/n_{hk},$$

and

$$\hat{\mu}_{hk} = \sum_{s_{hk}} \delta_{hk\ell}X_{hk\ell} \Big/ \sum_{s_{hk}} \delta_{hk\ell}.$$

Note that this is the same estimator as (33). The response $Z_{hk\ell}$ has simply been expressed as the product $\delta_{hk\ell}X_{hk\ell}$. Similarly, the previous analysis and results remain valid; the parameters θ_{hk}, σ^2_{hk}, and ρ_{hk} in the previously stated formulas are simply recognized as functions of the parameters in the distributions of the more fundamental random variables $\delta_{hk\ell}$ and $X_{hk\ell}$.

Thus the previously derived error variance (35) can be expressed in terms of the parameters in the present, more detailed model, as follows:

$$\mathrm{Var}\,(\hat{T}_h - T) = \sum_{\bar{s}_h} \mathrm{Var}\,(T_{hk})$$

$$+\left(\frac{\sum_{\bar{s}_h} B_{hk}}{\sum_{s_h} B_{hk}}\right)^2 \sum_{s_h} \mathrm{Var}\,(T_{hk})$$

$$+\left(\frac{\sum_1^{M_h} B_{hk}}{\sum_{s_h} B_{hk}}\right)^2 \sum_{s_h} E\left[\left\{\frac{N^2}{n}\left(1-\frac{n}{N}\right)\left(\pi\sigma_X^2\right.\right.\right.$$

$$\left.\left.\left. + \pi(1-\pi)\,\mu^2 - \pi^2\,\rho_X\sigma_X^2\right)\right\}_{hk}\right]. \quad (38)$$

The subscript hk is placed outside the braces in lieu of its being used repeatedly with N, n, π, σ_X^2, and ρ_X inside the braces. In this variance expression

$$\mathrm{Var}\,(T_{hk}) = \mathrm{Var}\,(N_{hk}\pi_{hk}\mu_{hk}) + E\,[\{N\,(\pi\sigma_X^2 + \pi\,(1-\pi)\,\mu^2) + N\,(N-1)\,\pi^2\rho_X\,\sigma_X^2\}_{hk}]$$

for all $k = 1, 2, \ldots, M_h$.

The situation is complicated by the introduction of nonresponse. A second indicator variable is employed to denote response status: for ℓ in s_{hk}, $\zeta_{hkl} = 1$ indicates response, and $\zeta_{hkl} = 0$ indicates nonresponse. Thus for ℓ in s_{hk}, $Z_{hkl} = \zeta_{hkl}\delta_{hkl}X_{hkl}$ is observable, and the problem is to estimate $\sum_1^H \sum_1^{M_h} \sum_1^{N_{hk}} \delta_{hkl}X_{hkl}$, the sum of X-values over all in-scope discharges. The response indicator variables are treated as random, independent, and independent of all the other random variables present. For an in-scope discharge from hospital hk, the response probability $Pr(\zeta_{hkl} = 1)$ is denoted by φ_{hk}.

It is assumed that each selected discharge can be classified as in scope or out of scope, even if it is nonresponding. That is, of the n_{hk} discharges selected in the sample from hospital hk, the number of in-scope discharges $\sum_{s_{hk}} \delta_{hkl}$ is *observable.* Of the $n''_{hk} = \sum_{s_{hk}} \delta_{hkl}$ in-scope discharges, only a random number, $n'_{hk} = \sum_{s_{hk}} \zeta_{hkl}\delta_{hkl}$, will respond. A direct estimate of $T_{hk} = \sum_1^{N_{hk}} \delta_{hkl}X_{hkl}$ is clearly

$$\hat{\hat{T}}_{hk} = \sum_{s_{hk}} \zeta_{hkl}\delta_{hkl}X_{hkl} + (n''_{hk} - n'_{hk})\hat{\hat{\mu}}_{hk} + (N_{hk} - n_{hk})\hat{\hat{\pi}}_{hk}\hat{\hat{\mu}}_{hk} \quad (39)$$

where

$$\hat{\hat{\mu}}_{hk} = \sum_{s_{hk}} \zeta_{hkl}\delta_{hkl}X_{hkl} \Big/ n'_{hk} \text{ and } \hat{\hat{\pi}}_{hk} = n''_{hk} \Big/ n_{hk}.$$

The first term in (39) is the observed sum of the in-scope, responding sample discharges. The second term estimates the sum of X-values for the $(n''_{hk} - n'_{hk})$ discharges known to be in scope but unobserved because of their nonresponse. The third term is the product of $(N_{hk} - n_{hk})\hat{\hat{\pi}}_{hk}$, the estimated number in scope among the $N_{hk} - n_{hk}$ nonsample discharges, and the estimated average response $\hat{\hat{\mu}}_{hk}$. The estimate can also be written in the more compact form

$$\hat{\hat{T}}_{hk} = N_{hk}\,\frac{\dfrac{n''_{hk}}{n'_{hk}} \sum_{s_{hk}} \zeta_{hkl}\delta_{hkl}X_{hkl}}{n_{hk}} = N_{hk}\hat{\hat{\pi}}_{hk}\hat{\hat{\mu}}_{hk}.$$

If the stratum total T_h is estimated by a ratio-type statistic employing these estimated hospital totals,

$$\hat{\hat{T}}_h = \sum_{s_h} \hat{\hat{T}}_{hk} + \frac{\sum_{s_h} \hat{\hat{T}}_{hk}}{\sum_{s_h} B_{hk}} \sum_{\bar{s}_h} B_{hk} = \frac{\sum_{s_h} \hat{\hat{T}}_{hk}}{\sum_{s_h} B_{hk}} \sum_1^{M_h} B_{hk},$$

and if the grand total is estimated by the sum of these estimated stratum totals, $\hat{\hat{T}} = \sum_1^H \hat{\hat{T}}_h$, then the resulting estimator is that used in HDS.

Note that it is possible to have no responding in-scope discharges in the selected sample s_{hk}, i.e., $n'_{hk} = 0$. In such a case if the simplest natural course is taken and $\hat{\hat{T}}_{hk}$ is defined as zero, a small bias appears. Given the values of N_{hk}, π_{hk}, μ_{hk}, and φ_{hk}, the expected value of T_{hk} is simply $N_{hk}\pi_{hk}\mu_{hk}$, while the expected value of $\hat{\hat{T}}_{hk}$ is

$$N_{hk}\pi_{hk}\mu_{hk}\,[1 - (1-\varphi_{hk})\,(1-\pi_{hk}\varphi_{hk})^{n_{hk}-1}].$$

For all except extremely small values of π_{hk} and φ_{hk} and small n_{hk}, the bias is clearly negligible.

The error variance of $\hat{\hat{T}}$ can be shown, by tedious but straightforward calculations, to be given approximately by an expression of the same form as (35):

$$\operatorname{Var}(\hat{\hat{T}}-T) \doteq \sum_{1}^{H}\left[\sum_{\tilde{s}h} \operatorname{Var}(T_{hk}) + \left(\frac{\sum_{\tilde{s}h} B_{hk}}{\sum_{sh} B_{hk}}\right)^2 \sum_{sh} \operatorname{Var}(T_{hk})\right] + \sum_{1}^{H}\left(\frac{\sum_{1}^{Mh} B_{hk}}{\sum_{sh} B_{hk}}\right)^2 \sum_{sh} E\,[\operatorname{Var}(\hat{\hat{T}}_{hk}-T_{hk}|N_{hk}, \Theta_{hk})].$$

The error incurred in using this approximation for the true error variance arises from the slight bias in $\hat{T}$ and is negligible whenever the bias is. Similarly, the last term in this variance is given approximately by

$$\sum_{1}^{H}\left(\frac{\sum_{1}^{Mh} B_{hk}}{\sum_{sh} B_{hk}}\right)^2 \sum_{sh} E\left[\left\{\frac{N^2}{n}\left(1-\frac{n}{N}\right)(\pi\sigma_x^2 + \pi(1-\pi)\,\mu^2 - \pi^2\rho_x\sigma_x^2)\right\}_{hk}\right]$$

when the nonresponse probabilities $1-\varphi_{hk}$ are all small. Thus the earlier results concerning allocation remain relevant when a small probability of nonresponse is present at the second stage of sampling. For the case of sizable nonresponse probabilities, the estimator should be reexamined and various alternate estimators considered in which π_{hk} and μ_{hk} are estimated by linear functions $\sum_{sh} \ell_{hk}\hat{\pi}_{hk}$ and $\sum_{sh} c_{hk}\hat{\mu}_{hk}$.

REFERENCES

[1] Royall, R. M.: Linear Regression Models in Finite Population Sampling Theory, in V. P. Godambe and D. A. Sprott, eds., *Foundations of Statistical Inference*, Holt, Rinehart, and Winston, Ltd., of Canada, 1971.

[2] Royall, R. M.: On finite population sampling theory under certain linear regression models. *Biometrika* 57(2):377–87, Aug. 1970.

[3] Ericson, W. A.: Subjective Bayesian models in sampling finite populations (with discussion). *Journal of the Royal Statistical Society*, Series B, 31(2):195–224, 1969.

[4] Ericson, W. A.: Subjective Bayesian models in sampling finite populations: stratification, in N. L. Johnson and H. Smith, Jr., eds., *New Developments in Survey Sampling*. New York. Wiley-Interscience, 1969, pp. 326–357.

[5] Kalbfleisch, J. D., and Sprott, D. A.: Applications of likelihood and fiducial probability to sampling finite populations, in N. L. Johnson and H. Smith, Jr., eds., *New Developments in Survey Sampling*. New York. Wiley-Interscience, 1969. pp. 358–89.

[6] Brewer, K. R. W.: Ratio estimation and finite populations: some results deducible from the assumption of an underlying stochastic process. *Australian Journal of Statistics* 5(3):93–105, Nov. 1963.

[7] Scott, A., and Smith, T. M. F.: Estimation in multi-stage surveys. *J.Am.Statist.A.* 64(327): 830–40, Sept. 1969.

[8] Rao, C. R.: *Linear Statistical Inference and its Applications*. New York. Wiley, 1965. Chapter 4.

[9] Cochran, William G.: *Sampling Techniques*. New York. Wiley, 1963. p. 174.

[10] Op. cit., p. 131.

[11] National Center for Health Statistics: Design of the NCHS Hospital Discharge Survey, by Simmons, W. R., and Schnack, G. A. Rockville, Maryland, 1969. Unpublished report.

APPENDIX

Derivations of Conditions on Optimal Stratification with Equal Allocation and Defensive Sampling

$M_1 \geqslant M_2 \geqslant \ldots \geqslant M_H$ *at optimum.*—It will be shown that whenever $M_h < M_{h+1}$ for any $h=1,2,\ldots, H-1$, the MSE (12) is reduced if the smallest unit in stratum $h+1$ is shifted into stratum h. The desired result follows directly from this fact.

No generality is lost if attention is restricted to the case of $h=1$ and $H=2$. Let $B^{(1)} \leqslant \ldots \leqslant B^{(M_1+M_2)}$ be the size measures B_{hk}, $k=1, 2, \ldots, M_h$ $h=1, 2$ arranged in nondecreasing order. Then $B^{(1)}, B^{(2)}, \ldots, B^{(M_1)}$ are the sizes of units in stratum 1 and $B^{(M_1+1)}, B^{(M_1+2)}, \ldots, B^{(M_1+M_2)}$ are the sizes of units in stratum 2. $\sum_1^{M_1} B_{1k} = \sum_1^{M_1} B^{(k)}$ and $\sum_1^{M_2} B_{2k} = \sum_{M_1+1}^{M_1+M_2} B^{(k)}$. The MSE is

$$\sigma^2 \left[\frac{M_1 - m_1}{m_1} \sum_1^{M_1} B^{(k)} + \frac{M_2 - m_2}{m_2} \sum_{M_1+1}^{M_1+M_2} B^{(k)}\right]$$

$$= \frac{\sigma^2}{m_1} \left[M_1 \sum_1^{M_1} B^{(k)} + M_2 \sum_{M_1+1}^{M_1+M_2} B^{(k)}\right] - \sigma^2 \sum_1^{M_1+M_2} B^{(k)} \tag{40}$$

because $m_1 = m_2$. Now if the smallest unit in stratum 2 is shifted into stratum 1, the new MSE is

$$\frac{\sigma^2}{m_1} \left[(M_1+1) \sum_1^{M_1+1} B^{(k)} + (M_2 - 1) \sum_{M_1+2}^{M_1+M_2} B^{(k)}\right] - \sigma^2 \sum_1^{M_1+M_2} B^{(k)}. \tag{41}$$

The difference of (40) minus (41) is proportional to

$$(M_2 - M_1 - 2)B^{(M_1+1)} - \sum_1^{M_1} B^{(k)} + \sum_{M_1+1}^{M_1+M_2} B^{(k)},$$

which is $\geqslant (M_2 - M_1 - 2) \; B^{(M_1+1)} - M_1 B^{(M_1+1)} + M_2 B^{(M_1+1)} \geqslant 0$ since $M_1 < M_2$. The first inequality is strict unless all the $B^{(k)}$ are equal.

$B_1 \leqslant B_2 \leqslant \ldots \leqslant B_H$ *at optimum.*—Under the assumption that the necessary condition $M_1 \geqslant M_2 \geqslant \ldots \geqslant M_H$ is satisfied, it will be shown that whenever $B_h > B_{h+1}$ for any $h=1, 2, \ldots, H-1$, the MSE is reduced if the largest unit in stratum h is shifted into stratum $h+1$ *unless* such a shift forces violation of $M_h \geqslant M_{h+1}$. As before, no generality is lost by restricting attention to the case of two strata. By the same basic argument used before, it can be shown that shifting the largest unit in stratum 1 into stratum 2 reduces the MSE by the factor

$$\sum_1^{M_1} B^{(k)} + (M_1 - 1)B^{(M_1)} - \sum_{M_1+1}^{M_1+M_2} B^{(k)} - (M_2 + 1)B^{(M_1)},$$

which is positive when $(M_1 - M_2 - 2)B^{(M_1)}$

$$> \sum_{M_1+1}^{M_1+M_2} B^{(k)} - \sum_1^{M_1} B^{(k)} = B_2 - B_1.$$

Since by assumption $B_2 - B_1 < 0$, the shift results in a positive reduction in the MSE if $M_1 - 1 \geqslant M_2 + 1$. Note that $M_1 - 1$ and $M_2 + 1$ are the sizes of the new strata.

Derivation of Expressions (35) and (36) for Variance

For convenience, the conditional expectation and variance of a statistic Y, given Θ_{hk}, N_{hk} $k=1, \ldots, M_h$, are denoted by $E(Y|C)$ and Var $(Y|C)$, respectively. Then

$$\text{Var}\,(\hat{T}_h - T_h) = \text{Var}\,[E(\hat{T}_h - T_h|C)] + E[\text{Var}\,(\hat{T}_h - T_h|C)]. \tag{42}$$

Since $\hat{T}_h - T_h = \sum_1^{M_h} B_{hk} \sum_{s_h} \hat{T}_{hk} / \sum_{s_h} B_{hk} - \sum_{s_h} T_{hk} - \sum_{\bar{s}_h} T_{hk}$, and, given C, T_{hk}, and $\hat{T}_{hk}$ are independent of $T_{hk'}$, and $\hat{T}_{hk'}$, for $k \neq k'$, the second term in (42) is equal to

$$E\left[\sum_{\tilde{s}_h} \mathrm{Var}(T_{hk} \mid C) + \sum_{s_h} \mathrm{Var}(T_{hk} \mid C) - 2\frac{\sum_1^{M_h} B_{hk}}{\sum_{s_h} B_{hk}} \sum_{s_h} \mathrm{Cov}(\hat{T}_{hk}, T_{hk} \mid C) + \left(\frac{\sum_1^{M_h} B_{hk}}{\sum_{s_h} B_{hk}}\right)^2 \sum_{s_h} \mathrm{Var}(\hat{T}_{hk} \mid C)\right]. \quad (43)$$

But from the exchangeability of the Z's, given C, it is easily shown that

$$\mathrm{Cov}(\hat{T}_{hk}, T_{hk} \mid C) = \mathrm{Var}(T_{hk} \mid C) .$$

Therefore, (43) can be rewritten

$$E\left[\sum_{\tilde{s}_h} \mathrm{Var}(T_{hk} \mid C) + \sum_{s_h} \mathrm{Var}(T_{hk} \mid C) - 2\frac{\sum_1^{M_h} B_{hk}}{\sum_{s_h} B_{hk}} \sum_{s_h} \mathrm{Var}(T_{hk} \mid C) + \left(\frac{\sum_1^{M_h} B_{hk}}{\sum_{s_h} B_{hk}}\right)^2 \sum_{s_h} \mathrm{Var}(T_{hk} \mid C) + \left(\frac{\sum_1^{M_h} B_{hk}}{\sum_{s_h} B_{hk}}\right)^2 \sum_{s_h} (\mathrm{Var}(\hat{T}_{hk} \mid C) - \mathrm{Var}(T_{hk} \mid C))\right]$$

$$= E\left[\sum_{\tilde{s}_h} \mathrm{Var}(T_{hk} \mid C) + \left(1 - \frac{\sum_1^{M_h} B_{hk}}{\sum_{s_h} B_{hk}}\right)^2 \sum_{s_h} \mathrm{Var}(T_{hk} \mid C)\right] + \left(\frac{\sum_1^{M_h} B_{hk}}{\sum_{s_h} B_{hk}}\right)^2 \sum_{s_h} E\ [\mathrm{Var}(\hat{T}_{hk} \mid C) - \mathrm{Var}(T_{hk} \mid C)]. \quad (44)$$

If the relation $\mathrm{Cov}(\hat{T}_{hk}, T_{hk} \mid C) = \mathrm{Var}(T_{hk} \mid C)$ is applied to the final sum, then after some rearrangement (44) can be rewritten as

$$E\left[\mathrm{Var}\left(\frac{\sum_{s_h} T_{hk}}{\sum_{s_h} B_{hk}} \sum_{\tilde{s}_h} B_{hk} - \sum_{\tilde{s}_h} T_{hk} \mid C\right)\right] + \left(\frac{\sum_1^{M_h} B_{hk}}{\sum_{s_h} B_{hk}}\right)^2 \sum_{s_h} E\ [\mathrm{Var}(\hat{T}_{hk} - T_{hk} \mid C)]. \quad (45)$$

The first term in expression (42) for $\mathrm{Var}\ (\hat{T}_h - T_h)$ is

$$\mathrm{Var}\ [E(\hat{T}_h - T_h \mid C)] = \mathrm{Var}\left\{\frac{\sum_{\tilde{s}_h} B_{hk}}{\sum_{s_h} B_{hk}} \sum_{s_h} N_{hk}\Theta_{hk} - \sum_{\tilde{s}_h} N_{hk}\Theta_{hk}\right\}$$

$$= \mathrm{Var}\left[\mathrm{E}\left\{\frac{\sum_{s_h} T_{hk}}{\sum_{s_h} B_{hk}} \sum_{\tilde{s}_h} B_{hk} - \sum_{\tilde{s}_h} T_{hk} \mid C\right\}\right]. \quad (46)$$

Adding (45) and (46) yields (35).

Now

$$\operatorname{Var}(\hat{T}_{hk}-T_{hk}|C)=E[(\hat{T}_{hk}-T_{hk})^2|C]$$

$$=E\left[\left(N_{hk}\sum_{s_{hk}} Z_{hkl}/n_{hk}-\sum_{1}^{N_{hk}} Z_{hkl}\right)^2\Big|C\right],$$

which is, by exchangeability, the same for all samples s_{hk} containing n_{hk} units. This quantity is thus the same as

$$\frac{1}{\binom{N_{hk}}{n_{hk}}}\Sigma^* E\left[\left(N_{hk}\sum_{s_{hk}} Z_{hkl}/n_{hk}-\sum_{1}^{N_{hk}} Z_{hkl}\right)^2\Big|C\right]$$

where Σ^* indicates summation over all the $\binom{N_{hk}}{n_{hk}}$ samples s_{hk} of size n_{hk}.

Interchanging the order of summation and expectation in this last expression establishes (36).

☆ U. S. GOVERNMENT PRINTING OFFICE : 1973 515-212/59

VITAL AND HEALTH STATISTICS PUBLICATION SERIES

Originally Public Health Service Publication No. 1000

Series 1. Programs and collection procedures.—Reports which describe the general programs of the National Center for Health Statistics and its offices and divisions, data collection methods used, definitions, and other material necessary for understanding the data.

Series 2. Data evaluation and methods research.—Studies of new statistical methodology including: experimental tests of new survey methods, studies of vital statistics collection methods, new analytical techniques, objective evaluations of reliability of collected data, contributions to statistical theory.

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

Series 4. Documents and committee reports.—Final reports of major committees concerned with vital and health statistics, and documents such as recommended model vital registration laws and revised birth and death certificates.

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

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

Series 12. Data from the Institutional Population Surveys.—Statistics relating to the health characteristics of persons in institutions, and their medical, nursing, and personal care received, based on national samples of establishments providing these services and samples of the residents or patients.

Series 13. Data from the Hospital Discharge Survey.—Statistics relating to discharged patients in short-stay hospitals, based on a sample of patient records in a national sample of hospitals.

Series 14. Data on health resources: manpower and facilities.—Statistics on the numbers, geographic distribution, and characteristics of health resources including physicians, dentists, nurses, other health occupations, hospitals, nursing homes, and outpatient facilities.

Series 20. Data on mortality.—Various statistics on mortality other than as included in regular annual or monthly reports—special analyses by cause of death, age, and other demographic variables, also geographic and time series analyses.

Series 21. Data on natality, marriage, and divorce.—Various statistics on natality, marriage, and divorce other than as included in regular annual or monthly reports—special analyses by demographic variables, also geographic and time series analyses, studies of fertility.

Series 22. Data from the National Natality and Mortality Surveys.—Statistics on characteristics of births and deaths not available from the vital records, based on sample surveys stemming from these records, including such topics as mortality by socioeconomic class, hospital experience in the last year of life, medical care during pregnancy, health insurance coverage, etc.

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

Office of Information
National Center for Health Statistics
Public Health Service, HSMHA
Rockville, Md. 20852