ON THE THEORY AND PRACTICE OF OBTAINING UNBIASED AND EFFICIENT SAMPLES IN SOCIAL SURVEYS

PREPARED UNDER A GRANT FROM THE U.S. DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE

CARL N. MORRIS, JOSEPH P. NEWHOUSE,

RAE W. ARCHIBALD

R-2173-HEW
JANUARY 1980

Rand
SANTA MONICA, CA. 90406

The research reported herein was performed pursuant to Grant No. 016B-7501-P2021 from the U.S Department of Health, Education, and Welfare, Washington, D.C.

Library of Congress Cataloging in Publication Data

Morris, Carl N
On the theory and practice of obtaining unbiased and efficient samples in social surveys and experiments.

([Report] - Rand Corporation ; R-2173-HEW)
1. Social surveys. 2. Social sciences--Statistical methods. 3. Sampling (Statistics) I. Newhouse, Joseph P., joint author. II. Archibald, Rae W., joint author. III. Title. IV. Series: Rand Corporation. Rand report ; R-2173-HEW.
AS36.R3 R-2173 [HN29] 081s [300'.7'23]
ISBN 0-8330-0186-8 79-26960

Published by The Rand Corporation

ON THE THEORY AND PRACTICE OF OBTAINING UNBIASED AND EFFICIENT SAMPLES IN SOCIAL SURVEYS

PREPARED UNDER A GRANT FROM THE U.S. DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE

CARL N. MORRIS, JOSEPH P. NEWHOUSE,

RAE W. ARCHIBALD

R-2173-HEW
JANUARY 1980

PREFACE

This report was written as part of the experimental design work for the Rand Health Insurance Study, supported by a grant from the U.S. Department of Health, Education, and Welfare. It considers a number of problems that either have not been addressed in the literature on experimental design, or have been inadequately addressed. The report, which draws upon Rand experience in designing and operating the Health Insurance Study, should be of interest to analysts conducting longitudinal surveys and social experiments, as well as to those interested in experimental design.

SUMMARY

This report takes up four problems in experimental design of social experiments that either have not been addressed or, in our view, have been inadequately addressed in the literature. The first problem concerns definition of the sampling frame when repeated sampling is attempted and part of the population is transient. Sampling the transient population may not be feasible if a minimum length of participation is necessary. Even ignoring this problem, traditional sampling rules may force inappropriate analyses, e.g., of partial families if a new member cannot be included because that member would be given a second chance to participate. Because such practices can lead to analytical bias, one should consider analytical aims when formulating sampling rules. We illustrate how problems of defining a sampling frame in a longitudinal study may arise, and give some practical suggestions for dealing with them. However, our major purpose is to call attention to the sampling-frame problem, in the hope that new theory and methods will be developed to deal with it.

The second problem we address concerns optimal choice of survey samples. We consider two distinct, although related, issues--disproportionate sampling of populations of greater interest (e.g., low income) and disproportionate sampling of neighborhoods with certain characteristics (e.g., poor neighborhoods). We show that if the population of interest is described by a variable subject to measurement error (e.g., current income as a measure of permanent income), disproportionate sampling can be inefficient relative to proportionate sampling, even for the favored group. Similarly, the oversampling of neighborhoods that have many people with a desired characteristic can also be less efficient relative to proportionate sampling. Before using disproportionate sampling, the analyst should carefully review the modeling assumptions that will be necessary and the likelihood of misclassification.

The third problem we consider concerns the allocation of subjects to treatments. The theory of optimal design pertains to this problem,

although we view its recommendations skeptically. If a functional form is known, it is optimal to sample from certain portions of the distribution of a variable. For example, if a linear response to variation in income is to be estimated, subjects with extreme incomes (both high and low) provide minimum variance estimates. Because functional forms are almost never known with certainty in social science research, a design that assumes they are known may not be robust. It also is possible to exploit the dependence between a subject's characteristics and the cost of improving precision (e.g., by using the Conlisk-Watts design model [2]). However, these gains are realized only if one is willing to trust the assumptions made about the nature of responses and the functional form (e.g., ruling out certain interactions). The resulting design may provide little basis for testing these assumptions, and if they are incorrect, the resulting analysis may produce biased estimates. Finally, crossover designs have been suggested as a way to reduce sample size. If practical, crossover designs are attractive; however, we believe that crossover designs rarely can be used to advantage in social experiments.

In the final section, we discuss the problem of balancing covariates when assigning subjects to treatments in the presence of field constraints. We prove that where nonacceptances of the enrollment offer are random, the precision gains in estimating treatment contrasts by a balanced sample will be degraded approximately in proportion to the nonacceptance rate. We discuss practical reasons why avoidance of such degradation was impossible in the Health Insurance Study.

CONTENTS

I. INTRODUCTION

Classical experimental design has been developed for studies that use cross-sectional techniques, i.e., those that seek to make inferences about a population from observations made at one point in time. Such designs frequently attempt to satisfy the criteria of efficiency (minimum variance) and unbiasedness of estimates, and in doing so they often assume that certain attributes of the population can be measured costlessly, instantly, and without error.

Field experiments (or social experiments) that involve economic phenomena frequently will violate these assumptions. Time must pass to gather data (e.g., data on labor supply in income maintenance experiments, data on medical care and on electricity consumption in health insurance and peak-load pricing experiments); thus, measurements are not made on a population at one point in time. Furthermore, it is impossible in field experiments to measure individual or family attributes costlessly or instantly, and the presence of error in such measurements can substantially reduce the gains in efficiency that an optimal design purports to achieve.

As a result of these constraints, and of the practical difficulties of administering large-scale social experiments in real time, experimental design issues arise that have not been well addressed in the literature. Our purpose in raising them here is twofold: to give those who will design field experiments the benefits of our experience in designing and implementing the experimental portion of the Health Insurance Study (HIS); and to encourage the scientific community to rethink the criteria and methods needed for design in these complicated situations. Before we turn to the main issues of the report, however, a brief description of the HIS may be helpful for the reader.[1]

The HIS has several objectives, including: (1) to measure the insurance elasticity of demand for medical care services (i.e., the response to varying the portion of the expenditure that the participant

[1] A detailed description can be found in Newhouse (1974).

must pay out of pocket); (2) to determine if the insurance elasticity of demand depends on permanent income; and (3) to determine what effects on health, if any, are observed from variation in the consumption of medical care services because of differences in amount paid out of pocket. To achieve these ends, some 2,800 families have been enrolled in the experiment. The families are located in six geographic locations (Dayton, Ohio; Seattle, Washington; Fitchburg-Leominster, Massachusetts; Franklin County, Massachusetts; Charleston, South Carolina; Georgetown County, South Carolina).

Each family in the HIS is enrolled in one of fourteen health insurance plans that vary the fraction of total expenditure to be paid by the participant. The fraction is either 0, 25, 50, or 95 percent. In addition, the family's financial exposure is limited to a certain amount in any one year, an amount called the Maximum Dollar Expenditure (MDE). Generally the MDE is set as a fraction of income, but in one plan it is $150 per person. Some families are assigned to a Health Maintenance Organization (HMO) (prepaid group practice), and their care is free to them so long as it is received at the HMO. Families participate for either 3 years (70 percent) or 5 years (30 percent) to permit measurement of transitory behavior at the beginning and end of the experiment. Several years were needed to allow for transitory demand to disappear (i.e., rates of consumption that do not reflect steady-state behavior, such as restorative dentistry done on a one-time basis) and for health status effects to appear.

During the period of participation in the experiment, families do not use their own health insurance; rather, they assign the benefits of that insurance to the experiment. They are paid lump sums (not based on utilization) to ensure that they will not be worse off financially by participating in the experiment. They do not have a choice of insurance plan within the experiment but are made an all-or-nothing offer to participate in the plan to which they have been assigned.

Families were enrolled using the following procedure: (1) A screening interview was administered to determine eligibility (the aged and certain other populations are not eligible). (2) A baseline interview was administered to the eligible families to elicit certain information; in particular, information about health insurance policies. This

information, verified with the employer or insurance company, was used as the basis for the guarantee to the families that they would not be worse off by participating. (3) Following verification of the insurance information, families were selected, assigned to insurance plans (experimental treatments), and offered a chance to enroll.

The experiment is well under way. All the required families are enrolled, with about 75 percent of the ultimate number of person-years having been completed as of December 1979.

II. LONGITUDINAL SURVEYS OF NONSTATIONARY POPULATIONS

Difficulties arise in the practice of repeated interviews of nonstationary populations. These difficulties are not dealt with effectively by the existing theory and practice of survey sampling, which usually assumes a stationary target population. A description of some of these difficulties encountered in the HIS appears below. While our purpose here is only to call attention to these problems, not to solve them, we do discuss, later in this section, some practical methods for reducing the difficulties. Real progress, however, will be achieved only when new theory, methods, and standards are developed to deal directly with the complications of surveying nonstationary populations.

SAMPLING PROBLEMS

Contact with families in the HIS begins with a longitudinal (panel) survey before the experimental phase and is followed by the longitudinal experiment, lasting from 3 to 5 years. The preexperimental portion is longitudinal, i.e., it involves a reinterview of subjects because families are administered screening interviews (preliminary, 10 minutes), then baseline interviews (longer, in-depth), and finally enrollment interviews (when the insurance offer is made). Our concern here is focused primarily on problems arising from these preexperimental surveys, which in the HIS take a total of 6 to 9 months to complete, and might be expected to result in an unbiased sample for the experiment.[2]

Cross-Sectional vs. Longitudinal Sampling

The theory and practice of cross-sectional survey sampling (only one interview) is now highly developed and widely used to obtain nearly unbiased samples from specified target populations. When the target population (e.g., a specified subset of individuals in a city) can be enumerated and located, only the refusing respondents prevent

[2]We use the term "unbiased sample" loosely to mean that the probability of selection of each individual is completely known. Most frequently, this means each individual has the same selection probability.

the sample from being unbiased. If refusal rate is low, the sampling distribution can be assumed with confidence. When a human population cannot be enumerated and located readily, standard practice requires that dwelling units be listed and then sampled from that list as a basis for locating individuals. The occupants of a dwelling at the time of the "first knock" are considered to be in the sample; they are followed if they move to another dwelling before the interview is actually conducted. Hence, the sample switches from a dwelling sample to a sample of individuals at the first knock. This method works well so long as (a) almost all individuals are associated with exactly one dwelling unit at any one time, (b) individuals who move can be found, and (c) the survey period is short relative to changes in the population (due to vital events, leaving the sampling area, etc.).

The successes of cross-sectional sampling foster expectations that longitudinal sampling should produce equally good results. This is unrealistic, except in cases of relatively stationary populations. Longitudinal surveys cannot do as well. Even the concept of the "target population" becomes ambiguous. The target population consists of those individuals about whom the survey is to make inferences (in the HIS these would be the populations in the six HIS experimental sites at the end of the experiment who satisfy certain age and other eligibility constraints). The "survey population" is the set of individuals who make up the sampling frame during the preexperimental survey period. These two populations often coincide for cross-sectional surveys (one interview), since the period of analytical interest is the sampling period, or nearly so. They cannot coincide for longitudinal surveys of nonstationary populations.

Transient Population and a Long Period

Suppose the site has a transient population and the experimental period is lengthy. Only by constantly replenishing the sample during the experimental period is it possible to maintain the matching of sample characteristics with those of the target population. This is infeasible in the HIS because the survey population is the cross-section of eligible people in each site at the time of the preexperimental

surveys and must remain fixed during the experiment. New entrants (save for newborns and adopted children) are not allowed into the sample during the experiment for two principal reasons. First, a minimum number of years of participation is required to allow long-term changes in health status to occur. Individuals who have been used to replenish the sample will not show these effects. Second, transitory behavior may occur at the outset and at the end of the HIS if the participant's own insurance differs from that provided by the experiment. For example, the experimental insurance usually is more generous in that it covers both dental and psychiatric expenses. To the extent that these are durable goods, experimental families may purchase dental and psychiatric care at the beginning and end of the experiment in greater quantity than they would if their coverage were unchanging. Thus, their behavior differs from steady-state behavior. That individuals must be enrolled for a substantial period of time also has implications for cross-over designs; these are taken up in Section IV.

Defining the Survey Population

A second difference with cross-section surveys arises because the three surveys for the HIS during the preexperimental period (screening, baseline, enrollment) make the survey population hard to define. The screening survey can, and does in the HIS, provide a "first knock" cross-sectional sample that is acceptably representative of the community by cross-sectional survey standards. The survey population at that time is the "eligible" community during the screening period; thereafter it must be modified. (In the HIS, the eligible community excludes certain families, on the basis of income exceeding $25,000 [1973 dollars], the aged, the institutionalized, and certain students. Those whose current insurance cannot be verified [e.g., held by an employer who will not cooperate with the HIS] also are excluded. In a broader context, eligibility also requires meeting certain space and time restrictions, namely, that the individual reside within the sampling area during the preexperimental surveys.)

The first new difficulties arise when the interviewer returns to administer a second survey, the baseline survey, to an eligible family that already has been screened. It may happen that

(a) The family has moved out of the sampling area.
(b) The entire family, or perhaps some of its members, has changed eligibility status since the screening period.
(c) The family has reconfigured; births, adoptions, deaths, marriages, divorces, or a member's coming of age and becoming a separate family, all act to produce family reconfigurations (which may include the formation of new families within the household).

The screening and baseline interviews are followed by the enrollment interview, which occurs several months after the baseline interview. The sampling problems occur again in this third interview. Sampling problems (a) through (c) occurred frequently in the HIS. As shown in Table 2, page 33, nonrefusal losses from the sample (moved, unable to locate, ineligible, unable to verify insurance) were about 22 percent, and the final enrollment sample in the HIS was slightly more than 105 percent of the original sample due to the discovery of new families.

Dating the Survey Population

In the presence of these events, the survey population cannot be dated to the screening period. If no effort is made to recoup moving and eligibility losses, the sample will be more stable than the eligible population at the time of each interview. Biased estimates can generally result from a standard analysis, because the sampling probabilities are modified in unknown ways. It is expensive and of negligible value to follow out-of-area movers in pursuit of an unbiased sample at the screening period. The HIS opted instead to administer both screening and baseline interviews to new families occupying those dwellings that housed out-of-area movers, and to follow only in-area movers. This partially atones for loss of the moving population, although not perfectly, because out-of-area movers are replaced by new families from both out of the area and within the area. It also moves the survey target population closer to the eligible population at the baseline period. Of course, the survey population cannot be updated entirely to the baseline period without returning to all households that, during the screening period, refused, were never at home, were vacant, or were not even contacted,

and attempting again to complete screening and baseline interviews. Further, it would be necessary to return to households occupied by ineligible persons to determine if their eligibility status had since changed. Therefore, individuals who were ineligible but became eligible will not be represented in the sample; moreover, we will not know exactly how the community's population has changed over time, and so will not have an appropriate denominator to compute sampling probabilities.

Oversampling

All these moving, eligibility, and reconfiguration difficulties cause the baseline survey to oversample the population that is stable during both the screening and baseline periods. The nonconstancy and nonpredictability of eligibility characteristics causes this. By contrast, ineligibility due to age does not cause this problem, because (ignoring deaths) future ages are totally predictable.

Family Reconfiguration in the HIS

The standard rule for treating the problem created by family reconfigurations is to ignore new persons joining the family between the screening and the baseline interviews. The HIS, however, is especially interested in families, because the family is the economic decision-making unit, and because national health insurance may well apply to the family unit. To take an example: Suppose a widower and his child are insured by the HIS, but the stepmother is not because she married into the family after the screening interview. This family's behavior will match neither that of three-person, man-woman-and-child families with all members insured (since the woman's expenditures in such families affect whether the family meets the deductible), nor that of two-person, father-and-child families (since the stepmother shares income and also is likely to influence the child's demand for health care). Furthermore, national health insurance is unlikely to exclude some family members, such as the mother. A sampling procedure that excludes new family members in order not to change their selection probability could therefore lead to biased estimates in the analysis of the HIS experimental data. On the other hand, if some sampling bias is accepted in order

to enroll families as a unit, the utilization of health services as a function of family characteristics may actually be estimated with reduced bias. In the preceding example, inclusion of the new spouse clearly would aid the analysis. In fact, an estimate of the conditional distribution of utilization from a biased sample may be unbiased. The point is this: *A sampling design must consider the combined effects of two sources of bias, sampling biases and limitations (here, the incomplete family) imposed on the analysis.* An unbiased sample may not minimize bias of inferences concerning population parameters.[3]

PRACTICAL METHODS TO REDUCE LONGITUDINAL SAMPLING PROBLEMS

As noted earlier, our primary purpose is to call attention to the increased difficulties of sampling populations longitudinally, not to resolve them. A solution will be achieved only when the research community explicitly recognizes the longitudinal problem and provides generally acceptable standards and methods for dealing with it. We think the standards should be concerned with minimizing some function of both bias and variance of estimates in relation to cost. Acceptable sampling frames must include broader concepts than a specified population at a particular point in time.

While we will not attempt to make general recommendations, certain methods for reducing the magnitude of the problem have come to our attention in the course of designing the HIS. Related ideas in the context of assigning treatments are presented in Section V. Some of the following suggestions were used in the HIS and others were not. We are not claiming that we always made the proper choices for the HIS or that we would make the same choices now. When reasons for choice are given below, they are those that applied at the time of decision.

1. *Sample Replenishment.* As attrition takes place, new members with characteristics matching those lost can be brought into the study.

---

[3]The problem is analogous to the theory of the second best. Imposing the requirement of an unbiased sampling frame when there are resulting limitations on analysis is analogous to insisting that one marginal condition be satisfied when another cannot be.

This can reduce or eliminate sampling bias, but it also can be difficult, costly, and lead to incomplete data for subjects. A minor example of sample replenishment in the HIS was replacing out-of-area movers during the preexperimental survey period by the families who moved into the vacated dwelling units.

2. *Sample Compensation.* If certain groups eventually will be underrepresented, it may be advisable to overrepresent them initially. In the HIS, those recently discharged from the military or leaving college could have been oversampled, because their cohorts will not be picked up later.[4] While data were gathered at the baseline stage on those soon to be discharged or graduated, this oversampling strategy was not followed because these categories include only a minor segment of the population, and because a proper oversampling rate was not known.

3. *Techniques for Shortening the Preenrollment Survey Period.* In the first HIS site, the screening survey preceded the baseline by several months. In all other sites, a "doorstep screener" was used. The interviewer attempted at first knock to complete the screening interview. After establishing family eligibility, a randomization table (the randomization was based on income and family size for the purpose of oversampling low-income families) was used to determine whether to administer a baseline interview. If so, the family was asked to continue at that time or for an appointment to complete the baseline interview. By shortening the survey period, and generally reducing the number of interview contacts, this strategy significantly reduced costs, fielding time, and potential biases. It required more training of interviewers to make correct eligibility and randomization decisions, because they cannot be made centrally.

Another similar tactic was used in the Los Angeles peak-load electricity pricing experiment (Manning et al., 1976) for the purpose of shortening the time to enrollment. A random two-thirds of the households

---

[4] An HIS participant who enters the military is suspended; therefore, those with military experience will be underrepresented. College dormitories were not sampled because of the mobility of students and the likelihood that students would continue to use a student health service even if national health insurance were enacted; as a result, college students will also be underrepresented.

were asked to participate in the experiment on predesignated treatments at the end of the first interview. The remaining households were assigned to balance treatment variables later, after baseline data were available for the entire sample. This method appears to be very cost-effective, although it does somewhat restrict the ability to balance variables in assignment of subjects to treatments.

The method used by the peak-load pricing experiment was not applicable to the HIS because the HIS families had to have their insurance formally verified (which took at least 6 weeks) before an offer could be made. However, the HIS moved the enrollment period closer to the baseline period by selecting a portion of the enrollment sample before the baseline period was complete. Care must be taken when doing this; problems encountered are discussed in Section V.

4. *Crossover Designs with Preenrollment in Control Group Status.* In experimental situations, even though the treatments cannot be assigned immediately after completion of the baseline interview, it may be advantageous to preenroll interviewees in control group status at that time. Later they would cross over from control to treatment status. If this can be done, then transitory effects attributable solely to participating in an experiment may diminish before assignment of the treatments. While this does not alleviate many of the difficulties attributable to longitudinal sampling, it can lead to better comparisons among the treatments because some early attrition would be forced due to forms burden, and some refusals would occur before assignment of treatments. Other advantages are that participants will serve as their own control group while on the experimental measurement system, that the size of the entire control group is increased, and that analysis of those who refuse the offer of the experimental treatment is facilitated. The HIS did not use this method (except for a 2-year preenrollment group for part of the sample in South Carolina) primarily because of the additional risk of bias attributable to possible nonrandom attrition from the control group, unless quite large payments were made to the families. (By contrast, the magnitude of benefits received leads to low attrition during the experimental period.) Secondarily, it was costly to make preexperimental data available quickly for assignment of treatments.

5. *Rules for Following Movers and Reconfigured Families.* Standard survey practice is to follow individuals within the sampling area if this is possible. What if it is impossible? Under what circumstances should the person who moves into the vacated dwelling be interviewed instead? Which procedure is cost-effective? These issues become more complex if family units are to be sampled when divorce, separation, remarriage, and coming-of-age lead to reconfiguration and to creation of new family units.

Randomization, in conjunction with a change of viewpoint away from preserving a population of individuals and toward preserving the characteristics of the initial sample, offers one possibility. For example, if a husband and wife divorce and both remarry during the sampling period, then if the survey unit is the individual, both would be followed and their new spouses ignored. When the survey unit is the family, it may be preferable to follow the husband with probability one-half, and the wife with probability one-half, incorporating the new spouse, and ignoring the unselected family. Even though the new spouse has a second chance of being enrolled, certain characteristics are maintained: the wholeness of one husband-wife family is retained instead of changing the sample to include two partially enrolled families. This example is straightforward, but the matter becomes complex when one must consider the myriad combinations involving children and other members in the old and new family units, the inability to follow some members, and multiple family dwellings. We believe that much useful research could be carried out here.

6. *Connecting Cross-Sectional Surveys with the Panel Survey.* A cross-sectional survey of the site made after a panel has been selected will include two groups: those who might have been (or are) panel members because they were present and eligible during the preenrollment period; and those who could not have been because they were ineligible at the time (lived out of the area, etc.). Questions can be included on the cross-sectional survey that would provide information about which group included the interviewee. Analysis of such data would provide information about biases that might obtain because the panel was constrained to the more stable population. In particular, biases due to

selection of a nonrepresentative population along measurable dimensions (e.g., the stable population is older) can be relatively well estimated. However, any interaction between the experimental treatment and the transitory population cannot be measured by using a later cross-sectional sample.

In concluding this section, we wish to restate the main points.

Longitudinal surveys are different from cross-sectional surveys, and more difficult to carry out. A better understanding of the nature and problems of longitudinal surveys is needed by the scientific community so that proper standards can be agreed upon and suitable sampling methods developed.

Longitudinal surveys are not always more informative than cross-sectional ones; indeed, for many purposes they can be less informative. They gain because they permit estimation when experimental effects are time-dependent. Even when such models are not of interest, longitudinal experiments may be required if the number of experimental subjects is limited, if costs per year are constrained (but several years' support is available), if recall error requires frequent interview (e.g., asking a panel about consumption every month rather than asking once for annual consumption), if transitory behavior is expected, or if long-term effects are to be measured. In the absence of such conditions, a cross-sectional analysis will produce results earlier and may also be more accurate (in steady-state, a mean estimated from two positively correlated measurement periods on one subject has greater variability than a mean estimated from one measurement period on two subjects).

A goal of sampling should be to minimize a function of both variances and biases of the estimates finally generated by the analysis, subject to prescribed cost constraints. Because sampling errors are just one component of the total error, some sampling bias may be acceptable if this leads to decreased errors in the fitted analytical model, or reduced costs. Research on methods and standards for reinterview sampling should address these points.

III. CONSTRAINTS ON CHOOSING OPTIMAL SAMPLES

Because the HIS sponsors had greater interest in the effects of insurance on the low (permanent) income population than on other income groups, it was agreed that the HIS would oversample low-income families. This oversampling amounts to choosing a more efficient sample for analysis by not sampling proportionately. Since oversampling is required, survey costs can also be reduced by overselection of low-income neighborhoods. Under certain circumstances, both kinds of oversampling can be effective, but *our purpose here is to show that without strong modeling assumptions, the gains from each can be negligible or even negative.*

OPTIMIZATION WHEN CLASSIFICATION ERRORS EXIST

There is a fundamental difference between oversampling with respect to a variable that varies randomly with time, as income does over the several-year experimental period of the HIS, and one that is constant or completely predictable, such as race, sex, or age. Oversampling a group is desired for values of a variable that will occur during the experiment, but only the preexperimental values are available for this purpose. If the preexperimental variable is not perfectly correlated with the experimental value, two difficulties arise. First, there will be some regression to the mean so that the experimental value will not be oversampled as strongly as the preexperimental value, causing the result to be less efficient than desired. Second, unless researchers are in a position to assert that no latent variables (omitted variables that are partially correlated with the oversampling variable) exist, then an analysis of experimental data must account simultaneously for the separate oversampling rates due to the preexperimental and to the experimental variables. This degrades precision, possibly to the extent that proportional sampling would be more efficient than oversampling. Of course, weighted analyses are also more cumbersome to conduct because the weights must be carried throughout.

We shall illustrate these ideas with a simple example that will permit numerical evaluation of the efficiency gains and losses. Suppose

each family falls into a "low income" or "high income" category on the basis of their income for the year immediately preceding the experiment, and that each of these two categories represents one-half of all families. Similarly, the average income of each family during the life of the experiment (permanent income would be another candidate for the variable of interest) falls into "low" and "high" categories, and again each is assumed to include one-half of all families. The assumption that each category corresponds to one-half the population is made for convenience only. Although more general situations may be treated, we keep matters simple here because our purpose is only to illustrate the problem caused by classification errors.[5]

The only data available for oversampling are preexperimental incomes, so the fraction f is designated as the proportion of the sample that is to have low preexperimental income, while $\bar{f} = 1 - f$ will have high income. Let c be the probability of correct classification. Because of the symmetry in this example, c is the probability that a preexperimental low (or high) income family is low (or high) during the experiment and $\bar{c} = 1 - c$ is the probability of changing income categories during the two periods. The case c = 1 corresponds to variables like race, age, and sex (assuming no errors in measurement of these variables). Ordinarily, c is greater than one-half, with c = .5 meaning that the preexperimental and experimental classifications are independent. Figure 1 contains the assumptions used for this presentation. Note in particular that the proportion of low-income families actually experienced, f_c, is less than f (for $f > \frac{1}{2}$), since $f_c = cf + \bar{c}\bar{f}$. For convenience, we also denote $\bar{f}_c = 1 - f_c = \bar{c}f + c\bar{f}$ to be the fraction of families with high incomes during the experiment.

The population means μ_{ij} are estimated unbiasedly by $\bar{x}_{ij}$, the mean response in cell i, j. Assume that var $(\bar{x}_{ij}) = \sigma^2/n_{ij}$ and that a total

---

[5]The problem discussed in this section arises whether the classification variable is continuous (e.g., income) or discrete (e.g., employment status), so long as the preexperimental values are not perfectly correlated with the experimental values.

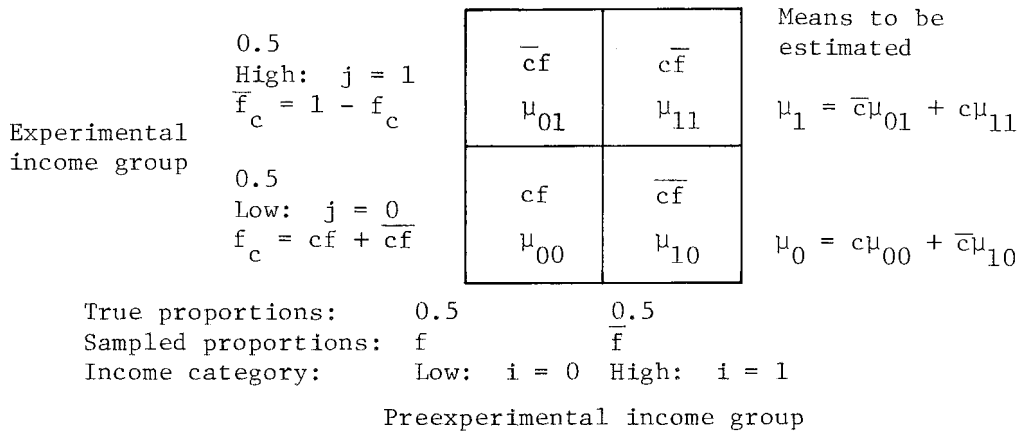

Fig. 1--Sampling proportions and responses μ_j and μ_{ij} for crossed income categories. $\bar{c} \equiv 1 - c$, $\bar{f} \equiv 1 - f$, μ_{ij} = mean of response of interest when the preexperimental income group is i and the experimental group is j; μ_j = mean response when the experimental group is j. The naturally occurring fraction is 0.5. Sampled fractions are f, 1 - f preexperimentally and f_c, 1 - f_c during the experiment.

of N families is used so that[6] $n_{00} = cfN$, $n_{10} = \bar{c}\bar{f}N$, $n_{01} = \bar{c}fN$, and $n_{11} = c\bar{f}N$. With these definitions, the mean responses during the experiment and the two quantities to be estimated are $\mu_0 \equiv c\mu_{00} + \bar{c}\mu_{10}$ for the low income group and $\mu_1 \equiv \bar{c}\mu_{01} + c\mu_{11}$ for the high income group.

If $\mu_{00} = \mu_{10}$ and $\mu_{01} = \mu_{11}$, so that the response for each income group during the experiment is independent of the preexperimental income categorization, then $\mu_0 = \mu_{00} = \mu_{10}$, $\mu_1 = \mu_{01} = \mu_{11}$, and *unweighted*[7] estimates may be used:

---

[6]We ignore the unimportant complication that the n_{ij} actually are stochastic in order to streamline the presentation. In large samples, n_{ij} will be nearly equal to its expectation. Values in Table 1 (p. 18) under f = 1.0 are then interpreted as limiting values.

[7]We say "unweighted" because (3.1) and (3.2) are the simple averages ignoring the preexperimental income category, i.e., total response divided by total number of subjects in each experimental income category.

$$\hat{\mu}_0 = \frac{n_{00}\bar{x}_{00} + n_{10}\bar{x}_{10}}{n_{00} + n_{10}} , \tag{3.1}$$

$$\hat{\mu}_1 = \frac{n_{01}\bar{x}_{01} + n_{11}\bar{x}_{11}}{n_{01} + n_{11}} . \tag{3.2}$$

These estimates are unbiased if $\mu_{00} = \mu_{10}$ and $\mu_{01} = \mu_{11}$. If not, let $\delta_0 \equiv \mu_{00} - \mu_{10}$, $\delta_1 \equiv \mu_{01} - \mu_{11}$. Then (3.1) and (3.2) are biased by the amounts

$$E\hat{\mu}_0 - \mu_0 = \frac{c\bar{c}(2f - 1)}{f_c} \delta_0, \qquad E\hat{\mu}_1 - \mu_1 = \frac{c\bar{c}(2f - 1)}{\bar{f}_c} \delta_1, \tag{3.3}$$

which are not zero unless $c = 1$ (or $c = 0$) or proportional sampling ($f = \frac{1}{2}$) has been used. Whether the estimates are biased or not, their variances are

$$\text{var } (\hat{\mu}_0) = \frac{\sigma^2/N}{f_c} , \qquad \text{var } (\hat{\mu}_1) = \frac{\sigma^2/N}{\bar{f}_c} . \tag{3.4}$$

If δ_0 and δ_1 cannot be assumed to be zero, because the preexperimental income category is partially correlated (given experimental income) with the response, then although (3.1) and (3.2) are biased, the following *weighted* estimates are unbiased for μ_0 and μ_1:

$$\hat{\hat{\mu}}_0 = c\bar{x}_{00} + (1 - c)\bar{x}_{10} , \qquad \hat{\hat{\mu}}_1 = (1 - c)\bar{x}_{01} + c\bar{x}_{11} . \tag{3.5}$$

Their variances are

$$\text{var } (\hat{\hat{\mu}}_0) = \frac{\sigma^2}{N} \frac{\bar{f}_c}{f\bar{f}} , \qquad \text{var } (\hat{\hat{\mu}}_1) = \frac{\sigma^2}{N} \frac{f_c}{f\bar{f}} . \tag{3.6}$$

The values in (3.6) are always larger than the corresponding ones in (3.4) unless c = 1 or f = ½. If f = ½, weighted and unweighted estimates are the same; hence,

$$\text{var}\ (\hat{\mu}_0) = \text{var}\ (\hat{\hat{\mu}}_0) = \frac{2\sigma^2}{N}\ , \qquad \text{var}\ (\hat{\mu}_1) = \text{var}\ (\hat{\hat{\mu}}_1) = \frac{2\sigma^2}{N}\ . \quad (3.7)$$

When f = ½, the variances (3.7) are therefore independent of the correct classification probability c, and, furthermore, unweighted estimates are unbiased [see (3.3)]. These are two strong advantages of choosing f = ½ in this case, and in more general situations, of sampling proportionally.

We now are in a position to compute the variance of the weighted estimates with f ≥ ½ from (3.6), relative to the variance in (3.7) for f = ½. These values appear in Table 1 for both income groups for the two cases c = .5 and c = 1.0. Since this ratio is linear in c, variance ratios for other values of c can be obtained by linear interpolation, as illustrated in Fig. 2.

Table 1 illustrates the major result of this section, that the variance for the unfavored group is always increased by oversampling,

Table 1

RATIO OF VARIANCES OF WEIGHTED ESTIMATES FOR VALUES OF f ≥ ½, FORMULA (3.6), RELATIVE TO THE VARIANCES WHEN f = ½, FORMULA (3.7)

Low Income (Favored) Group (f ≥ ½)			High Income (Unfavored) Group ($\bar{f} \leq$ ½)		
	Variance Ratio			Variance Ratio	
f	c = 0.5	c = 1	$\bar{f}$	c = 0.5	c = 1
0.5	1.00	1.00	0.5	1.00	1.00
0.6	1.04	0.83	0.4	1.04	1.25
0.7	1.19	0.71	0.3	1.19	1.67
0.8	1.56	0.63	0.2	1.56	2.50
0.9	2.78	0.56	0.1	2.78	5.00
1.0	∞	0.50	0	∞	∞

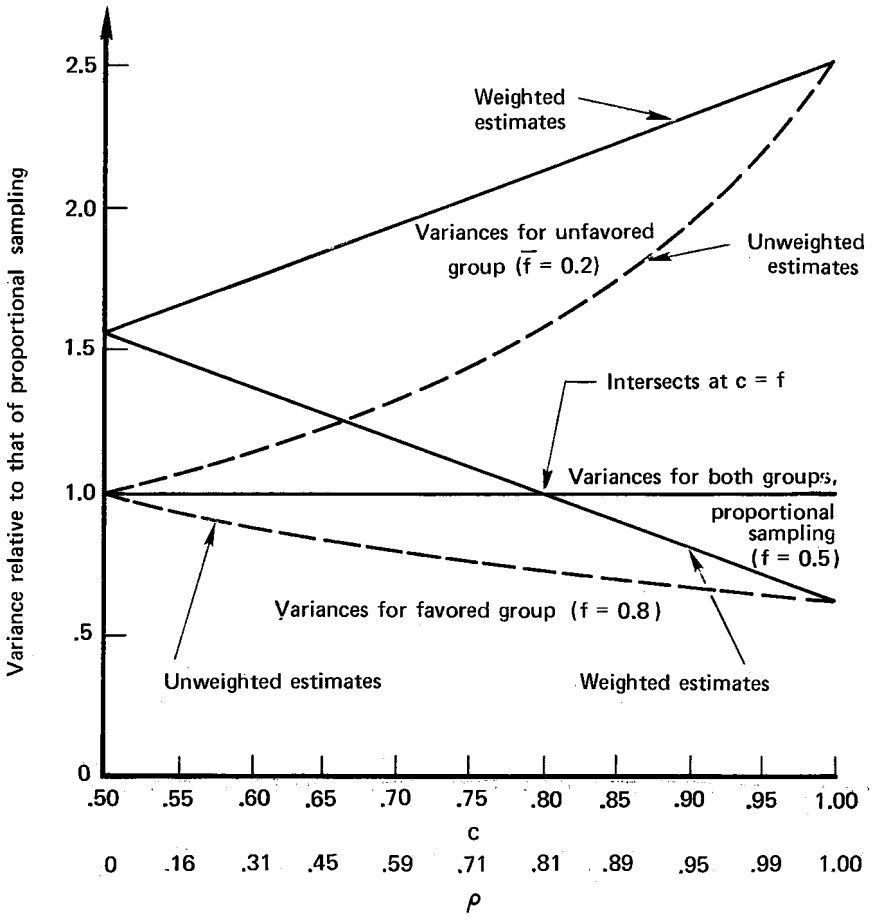

Fig. 2--Variances of weighted and unweighted estimates for the case f = 0.8, relative to variances resulting from proportional sampling. The horizontal axis is the classification probability c, or the corresponding correlation coefficient $\rho = \cos[\pi(1 - c)]$, between the preexperimental and the experimental values of the classification variable.

and for c near to 0.5, oversampling even harms estimation of the favored group. In fact, since the simple average of the two variance ratios for any value of c is $1/(4f\bar{f})$, independent of c, we may define this amount to be the overall increase in variance accepted in order to produce lower variances for the favored group. It is easy to see that even the favored group has larger variance than for proportional sampling if $c < f$. Hence: *oversampling leads to uniformly higher variances if the oversampling rate exceeds the correct classification probability.*

These ideas are illustrated for the case f = 0.8 in Fig. 2. The horizontal line going through 1.0 is the variance available if proportional sampling is used. The two solid sloped lines, which meet at 1.56 for c = 0.5 (see Table 1), are the variances from using the weighted estimates (3.5) when 80 percent of the sample is low income (preexperimentally) and 20 percent is high income (preexperimentally). If $c \leq 0.8$, even the favored group does not improve on the variance for proportional sampling; hence the lines cross there. The variances of the unweighted estimates (3.1), (3.2) are the dashed curving lines, whose two values always average in excess of 1.0. The gains for the favored group are small for c near to 0.5 because substantial misclassification produces experimental samples that nearly match the population proportions. At c = 1, the unweighted and weighted estimates are the same, and so their variances also agree. At c = 0.5, the unweighted estimates have variance independent of the oversampling rate f (since $f_{.5} = 0.5$ for all f) and therefore have relative variance equal to 1.

The horizontal axis is indexed not only by c but also by ρ, the value of the correlation coefficient that is consistent with c. It is computed from the formula $\rho = \cos[\pi(1 - c)]$, and is the correlation required between two normally distributed observations in order that the conditional probability of the second exceeding its median is c, given that the first exceeds its own median. (This formula is exact only for the case of two categories, each having probability one-half, as in the example.) Thus, in the case f = 0.8 of Fig. 2, correlations between preexperimental and experimental income less than 0.81 will lead to higher variances than for f = 0.5, *even for the favored group*.

The value ∞ for f = 1 and c = 0.5 in Table 1 raises an interesting point. If low incomes are the only ones of interest to the study, and if $c < 1$, then, to obtain unbiased estimates, it is necessary either to postulate that $\mu_{00} = \mu_{10}$ (that there are no latent variables partially correlated with preexperimental income) or to sample both preexperimental categories. That is, if $\delta_0 \neq 0$ in (3.3), any attempt to study low income preexperimental families by restricting the sample to low preexperimental incomes would be biased, because families with high preexperimental income, whose incomes drop later, would be excluded.

In summary, strong assumptions can eliminate the need to use weights based on the preexperimental variables. In that case, oversampling can be efficient if the probability of correct classification is high. If a weighted analysis is needed, the oversampling may degrade precision relative to proportional sampling, even for fairly high probabilities of correct classification. We have illustrated this in the simple case of estimating means, but this fact can hold for linear and other more complicated models. An additional reason to avoid oversampling is that a weighted analysis is more cumbersome to conduct than an unweighted one. This is quite important if many different analyses and dependent variables are being considered, because the weights must be used each time, even for those dependent variables whose prediction is not improved by the oversampling.

PROBLEMS DERIVED FROM OVERSAMPLING NEIGHBORHOODS

If one decides to oversample certain categories, possibly because classification errors are small, it may appear to be efficient to obtain the sample by oversampling neighborhoods that are abundant with the desired characteristics. This strategy reduces cost by decreasing the number of screening interviews needed to produce the desired sample proportions. If the analytical models legitimately can ignore the neighborhood effects, the savings are real. Even when this simplifying assumption cannot be made with confidence, such a sampling procedure may be required and some bias accepted to keep costs within reasonable bounds for studies that focus on individuals with uncommon characteristics.

But oversampling to reduce screening costs may not be cost effective. If the study must oversample low income families, the income distributions in census tracts can be estimated from census data. This information could be used to reduce the number of screening interviews substantially if the between-census-tract variation in income is large relative to the within-tract variation. However, a biased sample will result if low income families in low income tracts differ systematically in their responses from low income families in high income tracts. For example, low income families in high income neighborhoods are often

young, single people living with their high income parents. These people probably differ in their responses from low income families in low income neighborhoods having different age, family size, and employment characteristics. Therefore, a weighted analysis would be required to avoid bias, and as stated earlier, this increases variances of estimates. We have constructed many models, similar to those presented above, where the fractional increases in variance resulting from this kind of strategy substantially exceeds the fractional decrease in costs. In such cases, these strategies are not cost effective for a given budget. Furthermore, oversampling of neighborhoods has other disadvantages: Derivation of a proper sampling scheme is costly and quite difficult. If done improperly, it can lead to biased estimates. In addition, weighted analyses are more difficult to conduct.

To summarize this section, two examples were considered where oversampling of a population in which the analyst has greater interest may be undesirable. Oversampling is common in these situations. More discussion of the problems of oversampling, involving treatment assignment, appears in Section V. Our view is that oversampling should be used only after reviewing both the additional modeling assumptions required to make the exercise advantageous and the likelihood of classification error.

IV. DIFFICULTIES WITH OPTIMAL ALLOCATION OF SUBJECTS TO TREATMENTS IN SOCIAL EXPERIMENTS

In the preceding two sections we discussed the difficulties of getting unbiased samples of transitory populations and of oversampling subpopulations in the presence of classification errors. In experimental situations there also are opportunities to control the manner in which subjects are allocated treatments, the purpose being to produce a more cost-effective experiment. In this section we briefly consider three of these options: (a) choosing the sample to be nonrepresentative of the survey population; (b) making unbalanced assignments of subjects to treatments; and (c) using a crossover design. In each of these cases extra modeling assumptions are required to avoid bias, assumptions we were unwilling to make in the context of the Health Insurance Study.

NONREPRESENTATIVE SAMPLES

Certain experimental subjects may be more valuable for the estimation of parameters or less costly than others, and may therefore be preferred for allocation to treatments. A simple example of this occurs if a linear regression function of one independent variable is to be estimated, in which case subjects with extremely low or high values on this variable (e.g., income) are much more informative than those with intermediate values. When more subjects are available for selection than are needed for assignment to treatments, the preferred subjects may be assigned to treatments and the others excluded from the sample. Unbiased estimates may be derived from such assignments if the assumed parametric distribution is correct, or, in the absence of such assumptions, if every member of the target population has positive probability of being assigned to the treatments and, in the analysis, subjects are weighted inversely to their selection probability. Of course, the weighted analysis may be inefficient.

More hazardous yet are samples with some subjects having no possibility of selection. This occurs in the simple linear response example. The exclusion of the intermediate subjects makes the estimation of a

nonlinear model very difficult, and no weighted analysis can rectify the situation. The HIS did not follow this procedure by excluding middle income families, for example. While such a procedure would reduce variances for estimating a linear effect of income on the response, we have no assurances that responses are linear in income, and there is a strong likelihood that middle income families differ importantly along other critical dimensions that are correlated with income.

Note that the problem here is one of determining a priori the appropriate parametric density function. Unfortunately, economic theory rarely gives much help with respect to functional form. Conlisk (1973) has proposed a decision-theoretic approach to this problem in which the analyst assigns probabilities to functional forms and then minimizes expected loss. While attractive conceptually, it is not clear how one proceeds in practice to assign zero probability to functional forms one will not consider, nor how one should think about appropriate probabilities (in some cases information from other studies can be helpful). In the absence of information concerning functional form, a self-weighting (representative) sample appears to us to be the best choice.

UNBALANCED ASSIGNMENTS OF SUBJECTS TO TREATMENTS

If the cost of including a subject in the experiment depends jointly on his characteristics and on the treatment to which he is assigned, then it may be possible to increase the sample size and precision of the experiment by exploiting this relationship. The allocation model of Conlisk and Watts (1969) provides a means for computing the optimal sample allocation based on such a cost function, assuming a specified functional form.

For example, the cost of the Health Insurance Study could have been reduced by assigning those who had generous insurance before the experiment to the generous experimental insurance plans, and those with little or no insurance to the less generous plans. This would have eliminated the "worst-case" participation incentives paid to compensate those families whose experimental insurance was inferior in any way to their pre-experimental insurance. It also would have been unwise. Families in poor health tend to purchase better insurance than they would were they

in good health (Phelps, 1976). Generous insurance plans therefore would be overrepresented by unhealthy families and, as a result, comparisons between treatments would be biased unless the exact amount of overrepresentation were known.

The New Jersey Negative Income Tax Experiment provides a second example. The designers of this experiment found it less costly to put higher income families on generous plans and lower income families on less generous plans, and so generosity of the treatment and family income are correlated. Of course the modeling assumption that income does not interact with treatment generosity and that responses are linear with income provides one way to unravel these effects. Presumably, such assumptions were made during the design of the New Jersey experiment when it was decided that unbalanced assignment of incomes would be cost effective. The alternative to making these modeling assumptions is to use weighted estimates. But this will yield less efficient estimates than would have resulted from balanced assignments of subjects to treatments.

CROSSOVER DESIGNS

Thus far, we have considered designs in which subjects are exposed to one treatment only. In a long experiment it may be possible to expose subjects to several treatments, one at a time--a crossover design. If valid responses can be elicited in this fashion, then increased precision for estimating treatment effects can be expected because each subject acts as his own control. That is, comparisons between treatments can be made directly by observing the changed responses of each subject as he changes treatments. This eliminates accounting for the differences between subjects. Response variances are reduced by a factor $1 - R^2$ if every subject is exposed to every treatment for an equal length of time, where R^2 is the portion of the total variance of the response of subjects attributable to differences between subjects. Hall (1975) has argued that a crossover design could have reduced the necessary sample size by a factor of six in the New Jersey Negative Income Tax Experiment. We doubt this. The responses from crossover experimentation will be invalid if there is transitory behavior at the

beginning or end of each treatment period that is not accounted for. Transitory behavior may exist for a variety of reasons. Learning effects produce transitory behavior if it takes time for the subject to become familiar with all the benefits of the treatment and to develop ways to take advantage of them (e.g., to find dentists or psychiatrists in the HIS). Put another way, the actual response to a treatment may be delayed. Or if the treatment benefits permit subjects to purchase durable goods, they may engage in transitory behavior in crossover situations by waiting to purchase these goods until they are on a generous treatment (e.g., elective surgery, dental work, and psychiatric care are likely to be purchased in greater quantity per unit time by a subject who is on a generous health insurance plan for only a short period than by one who has generous insurance for a long time). Changes in behavior resulting from the subject's heightened awareness of the experiment are more likely if the treatment is changed frequently.

In the face of time-varying experimental responses, the design must either (1) allow sufficient time within each treatment for transitory effects to disappear--this amount of time is likely not to be known a priori, and additional calendar time may be necessary to estimate it; or (2) estimate the rate of change in behavior and extrapolate to a steady-state value. The first solution in practice will defer the results in time, and therefore a larger sample with no crossover (but earlier results) may well be preferred; the second solution requires strong assumptions.

Unfortunately, time-varying experimental responses appear to be common. In peak-load pricing experiments for electricity, it takes time to adjust the household's stock of appliances; in income maintenance experiments, it takes time to search for a new job; in housing demand experiments, it takes time to locate new housing. Further, all involve aspects of durable goods. Thus, while crossover designs can be very useful, we doubt that they will prove to be desirable as a general design for the types of social experiments considered here.

V. OPTIMAL BALANCE OF ASSIGNMENTS TO TREATMENTS AND THE DIFFICULTIES OF ACHIEVING IT WITH FIELD CONSTRAINTS

For reasons just mentioned, imbalances among the treatment assignments and the target population, or imbalances among treatments, can lead to poor estimates of the treatment effects. Suppose then that it is desired to balance subjects across treatments as carefully as possible. The time-honored and simplest method for doing this is by simple random sampling (SRS): The available subjects are assigned at random to treatments in accord with the required sample size for each treatment, *without* regard to any preexperimental measurements made on the subjects. Of course some imbalance of the preexperimental measurements still occurs, but it is random; and if large samples are assigned to each treatment, only minor increases in variance are experienced relative to a perfectly balanced assignment. If sample sizes are not large, however, then the imbalances from SRS can be substantial, and more control over the sample is desired.

Classical ways to use preexperimental measurements to improve balance over SRS include proportional stratification and blocking. In the Health Insurance Study, the "Finite Selection Model" (FSM) was used as an alternative because it can handle more independent variables than the classical methods and does not require converting continuous variables, like income, to categorical variables (Morris, 1975). The method of proportional stratification involves dividing the sample into k strata on the basis of the preexperimental variables and then assigning subjects to treatments separately from each stratum, in proportion to the treatment size, using simple random sampling. Blocking, a special case of proportional stratification, can be used only if all the treatments are the same size, or if all treatment sizes are small integer multiples of the smallest treatment. The entire sample then would be broken into homogeneous groups of the proper size, and each group assigned to the treatments in proportion to their size by simple random sampling.

PROPORTIONAL STRATIFICATION TO IMPROVE BALANCE

Because the calculations can be carried out conveniently for proportional stratification, that case is treated in our examples here. Proportional stratification includes blocking as a special case, while it in turn is a special case of the FSM.

The notation needed is given in Fig. 3, below. Subjects are classified into k strata S_1, ..., S_k on the basis of their preexperimental measurements and are to be assigned to treatments T_1, ..., T_t, with N_i subjects assigned to treatment T_i. Hence $\Sigma N_i = N_+$ is the total number of subjects in the pool. The proportion of subjects available for selection in stratum j is p_j, so $\Sigma p_j = 1$. Because there may be refusals of the offer or other failures to enroll assigned subjects in treatments, n_i (with $n_i \leq N_i$) is the number actually enrolled in T_i.

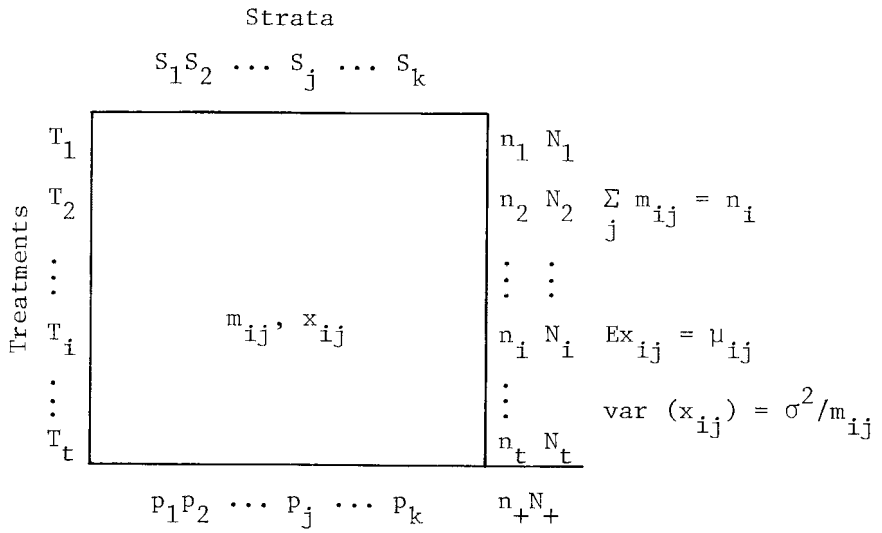

Fig. 3--Notation for assignment of subjects from k strata to t treatments.

The number of subjects actually enrolled in T_i from stratum j is m_{ij}, $\Sigma_j m_{ij} = n_i$. The true mean response to T_i in S_j is μ_{ij}, which is to be estimated by the sample mean x_{ij} of the m_{ij} observations in that cell. We assume

$$Ex_{ij} = \mu_{ij}, \qquad \text{var } (x_{ij}) = \frac{\sigma^2}{m_{ij}} \quad . \tag{5.1}$$

The average response to T_i among these subjects is

$$\mu_i = \sum_j p_j \mu_{ij} \ , \tag{5.2}$$

which is estimated unbiasedly by

$$\hat{\mu}_i = \sum_j p_j x_{ij} \tag{5.3}$$

with variance

$$\text{var } (\hat{\mu}_i) = \sigma^2 \sum_j \frac{p_j^2}{m_{ij}} \ . \tag{5.4}$$

The variances of contrasts between treatments are determined by formula (5.4). For example, the difference in effects of T_1 and T_2 is estimated by $\hat{\mu}_1 - \hat{\mu}_2$ with variance var $(\hat{\mu}_1)$ + var $(\hat{\mu}_2)$.

Suppose, first, that all assigned subjects actually enroll, so that $n_i = N_i$. Assuming that $(n_1, \ldots, n_t)$ is given, the uniformly best assignment scheme for all treatments--that which minimizes (5.4) for every i, subject to $\Sigma_j \ m_{ij} = n_i$--is proportional stratification:

$$m_{ij} = n_i p_j \tag{5.5}$$

for all i, j, in which case

$$\text{var}_{\text{PROP}} \ (\hat{\mu}_i) = \frac{\sigma^2}{n_i} \tag{5.6}$$

is the optimal variance. This is proved by minimizing (5.4) subject to $\Sigma_j \ m_{ij} = n_i$. Of course an exact identity for (5.5) may be impossible for any treatment because $n_i p_j$ may not be an integer. We ignore this complication.

Now suppose that some selected subjects fail to participate. If these subjects can be replaced by others from the same stratum,

perfectly proportional samples still will be achieved. This sequential procedure was infeasible in the context of the HIS for reasons to be described later. Instead it was necessary to select N_i subjects, in excess of the n_i desired for each treatment ($N_i > n_i$), permitting the imperfect acceptance rate $\pi < 1$ to yield approximately the desired number n_i. Our purpose is to demonstrate how much the variance (5.6) increases when this happens. As a first step, the expected precision from simple random sampling will be determined.

We are supposing that the $\{p_j\}$ are defined by the sample, based only on the N_+ observations, and not on the entire universe. The variances realized from SRS will therefore be reduced because of finite sampling, as we shall see. Comparisons between SRS and the results of proportional stratification will also be made conditionally on the value of $\{n_i\}$, since the random mechanism that operates to produce these values is the same for either design method.

The sample is assumed to have been constructed as follows: The universe is assumed to be large (e.g., a city) and a simple random sample of N_+ subjects, eligible for treatment assignment, is obtained. The numbers $\{N_i\}$ are fixed in advance; $\Sigma\, N_i = N_+$ subjects are assigned to T_i either at random or by the method of proportional stratification. Let m^*_{ij} be the number of subjects assigned to T_i from S_j at this step, with $\Sigma_j\, m^*_{ij} = N_i$. If SRS is used, m^*_{ij} is random, whereas $m^*_{ij} = N_i p_j$ if proportional stratification is used. These subjects are enrolled with probability π, so, given m^*_{ij}, the number that actually accept on T_i from S_j are $m_{ij} \sim$ binomial (m^*_{ij}, π). [This notation means that the random variable m_{ij} has the binomial distribution with mean m^*_{ij}, π and variance m^*_{ij}, $\pi(1 - \pi)$.] Of course, it is m_{ij}, not m^*_{ij}, that affects variances in (5.4). We wish to compute the expected value of (5.4), i.e., $\sigma^2 E\, \Sigma\, p_j^2/m_{ij}$, conditionally on this sampling scheme, on $\{n_i\}$, and on the sample of N_+ subjects, for both simple random sampling and proportional sampling. Formally, these expectations do not exist unless m_{ij} cannot be zero, but we shall interpret the expectations as conditional on the event that $m_{ij} \geq 1$. The approximate method used to obtain these expectations takes care of this problem. The results that follow are proved in the appendix.

Theorem 5.1. *The expected value of* (5.4) *under random sampling is approximately*

$$E_{SRS} \text{ var } (\hat{\mu}_i) \doteq \frac{\sigma^2}{n_i}\left[1 + \frac{k-1}{n_i}\left(\frac{N_+ - n_i}{N_+ - 1}\right)\right]. \tag{5.7}$$

Since σ^2/n_i is given by (5.6) as the variance for proportional stratification, the factor

$$1 + \frac{k-1}{n_i}\left(1 - \frac{n_i - 1}{N_+ - 1}\right) \tag{5.8}$$

is the expected fractional increase in variance from SRS relative to proportional stratification. When N_+ is large in relation to n_i, (5.8) simplifies approximately to

$$1 + \frac{(k-1)}{n_i}\,. \tag{5.9}$$

Note that (5.9) is the SRS variance for sampling from the universe, whereas (5.8) is smaller, because sampling without replacement from a universe of size N_+ is more efficient than from one of infinite size. Formulas (5.8) and (5.9) illustrate the well-known fact that SRS becomes less efficient as the number of subjects per treatment per stratum, n_i/k, decreases. SRS is asymptotically optimal as n_i/k becomes very large. But we believe that as n_+ increases in experiments as costly as the HIS, more treatments and more strata will be created, so that n_i/k ordinarily would not be large.

Turning to proportional stratification with $m^*_{ij} = N_i p_j$, formula (5.6) no longer will obtain because of random nonacceptances. Instead, conditional on n_i (the number of acceptances of T_i), the expected variance would be given by the following theorem.

Theorem 5.2. *The expected value of* (5.4), *assuming random acceptance of proportionally stratified offers* ($m^*_{ij} = N_i p_j$ *made to* T_i *from* S_j), *is approximately*

$$E_{PROP} \text{ var } (\hat{\mu}_i) \doteq \frac{\sigma^2}{n_i}\left[1 + \frac{k-1}{n_i}\left(\frac{N_i - n_i}{N_i - 1}\right)\right] . \qquad (5.10)$$

To interpret Theorem 5.2, define

$$\psi_i \equiv \frac{n_i - 1}{N_i - 1} \cdot \frac{N_+ - N_i}{N_+ - n_i} , \qquad (5.11)$$

which depends on the acceptance rate n_i/N_i, and on N_i and N_+, but not on k or σ^2. If $n_i = N_i$, then $\psi_i = 1$; if $n_i = 1$, $\psi_i = 0$; and, in most instances, ψ_i is slightly less than, but fairly close to, the acceptance rate π. We may rewrite (5.10) in terms of ψ_i and (5.7) as

$$E_{PROP} \text{ var } (\hat{\mu}_i) \doteq (1 - \psi_i)E_{SRS} \text{ var } (\hat{\mu}_i) + \psi_i \text{ var}_{PROP} (\hat{\mu}), \qquad (5.12)$$

where $\text{var}_{PROP} (\hat{\mu}) \equiv \sigma^2/n_i$ is the variance achieved by proportional sampling (5.6). The quantity ψ_i therefore shows how much a random acceptance rate would be expected to cut into the gains due to proportional stratification. Roughly, the improvement over SRS would be reduced by a factor equal to the nonacceptance rate, $1 - \pi$.

To illustrate this improvement, the acceptance rate was about 0.58 in the HIS, with losses caused by refusal of the offer, refusal of the final interview, and other circumstances such as families having moved or become ineligible. A detailed breakdown of the HIS enrollment experience is shown in Table 2. Note that one-half of the 42-percent sample loss is due to nonrefusal attrition, and the other half is due to refusal of the enrollment interview or the actual offer of enrollment.

If $N_+ = 800$ families, a typical value, then for values of N_i at 10, 40 (where ψ_i is maximum), 75, 120, and 180, the corresponding values of ψ_i are 0.531, 0.557, 0.550, 0.537, and 0.515. Taking 0.54 as a typical value here, slightly less than the 58 percent acceptance rate, we would expect to have gained only about 54 percent of the improvement over SRS nominally available had proportional stratification been

Table 2

DISPOSITION OF THE HEALTH INSURANCE STUDY ENROLLMENT SAMPLE

HIS Enrollment	Dayton		Seattle[a]		Massachusetts[b]		South Carolina[c]		Total	
	No.	%	No.	%	No.	%	No.	%	No.	%
Families who completed baselines and were assigned to treatments	528	--	2054	--	1068	--	1232	--	4882	--
Families added to sample at enrollment interview[d]	(e)	--	101	5.2	46	4.5	89	7.8	236[f]	5.1[f]
Attrition due to nonrefusals[g]	106	20.1	414	20.2	264	24.7	245	19.9	1029	21.1
Enrollment refusals	32	6.1	404	19.7	238	22.3	356	28.9	1030	21.1
Enrolled	390	73.9	1236	60.2	566	60.0	631	51.2	2823	57.8
Families interviewed for enrollment	422	79.9	1640	79.8	804	75.3	987	80.1	3853	78.9
Refused interview	4	0.9	147	9.0	108	13.4	125	12.7	384	10.0
Refused offer to enroll	28	6.6	257	15.7	130	16.2	231	23.4	646	16.8
Enrolled	390	92.4	1236	75.4	566	70.7	631	63.9	2823	73.3

[a]Includes a fee-for-service sample, a health maintenance organization sample, and a control group.

[b]Fitchburg-Leominster and Franklin counties.

[c]Charleston and Georgetown counties (includes a preexperimental control group sample).

[d]Line 2 is a subset of line 1.

[e]Not available.

[f]Excluding Dayton. The percentage figure is calculated as 236/(4882-236).

[g]Families moved, could not be located, were ineligible, or were unable to verify insurance.

used in the HIS. In fact, proportional stratification was not used because it was infeasible with the large number of covariates considered in the HIS design, and the FSM was used instead. We believe that 54 percent of the nominal gain also provides a rough percentage for the actual improvement over SRS provided by the FSM after sample reduction in the field.

Two points need to be made in relation to (5.12). First, while the expected variance of $\hat{\mu}_i$ is given by (5.7), the actual amount varies randomly. When many treatments are involved, it is likely that some treatment effects will be estimated with much larger variance than (5.7) suggests. This is an additional argument against SRS. However, proportional stratification, even in the context of random nonacceptances, not only reduces the expected variance of the treatment effects, but it also reduces the variability of the actual precision from that expected. This is highly desirable, especially so in view of criticism that would likely occur if, after the sample was selected, it was observed that the actual assignments to treatments from strata were unbalanced (i.e., if m_{ij}/n_i were substantially different from p_j for some treatment stratum combinations). A second point is that the variances in (5.12) obtain with respect to the strata defined on the basis of the preexperimental observations. The gains from proportional stratification with respect to the experimental observations would be diminished if the correlation between the two sets of observations were imperfect. Of course, SRS does not suffer from this phenomenon.

If the variance of x_{ij} is not given by (5.1), but instead by v_{ij}/m_{ij} (i.e., depends on stratum and treatment), or if the objective function (5.4) is replaced by a more general one, such as

$$\text{var } (\hat{\mu}_i) = \sum_j W_{ij} \text{ var } (x_{ij}) \ , \qquad (5.13)$$

with $W_{ij} \geq 0$ and fixed, then proportional stratification is not optimal. However, proportional stratification still is uniformly better than SRS, and formula (5.12) holds for the objective function (5.13). That is, even in this more general situation, the variance afforded by proportional sampling is a $(\psi_i, 1 - \psi_i)$ mixture of the proportional sampling variance resulting from no refusals, and of the expected SRS variance.

The discussion thus far has treated the n_i as given. Let n_i^* be the desired value for n_i, i.e., the desired number of subjects to be assigned to treatment T_i. If π is the acceptance rate, then $N_i = n_i^*/\pi$ selections would be made. Suppose the actual number of acceptances are random, so that n_i has a binomial distribution with parameters N_i and π. Then the mean of n_i is n_i^*; but since it may differ considerably from this value, some loss of overall precision would be expected. Suppose the n_i^* were derived by minimizing a weighted sum of the variances of $\hat{\mu}$, each given by (5.6), subject to a cost constraint which assumes that c_i is the cost of observing each subject on T_i. It is proved in the appendix, under the assumptions just made, that the weighted variance achieved for the optimal n_i^* is increased approximately by the factor

$$1 + \frac{1 - \pi}{E_c n^*}, \qquad E_c n^* \equiv \frac{\Sigma\, n_i^* c_i}{\Sigma\, c_i} \quad . \tag{5.14}$$

This ordinarily is fairly small. For example, with $\pi = 0.58$ and $E_c n^* = 35$ being the approximate average number of families enrolled on each HIS insurance plan, a 1.2-percent increase in variance results.

FIELD CONSTRAINTS IN THE HIS

We conclude this section by describing field constraints that acted in the Health Insurance Study to diminish the gains provided by the FSM. We have just showed that an imperfect acceptance rate reduces precision gains, unless the nonaccepting subjects are replaced by others from the same stratum. The sequential procedure needed to replace losses by subjects from the same stratum requires close contact with field operations. Although it is worth some effort to implement, this is more difficult than it may seem for the following five reasons.

First, the enrollment process is more efficient if enrollers have many cases to work at any one time rather than a few, with replacements arriving only as cases are closed. Thus, the majority of the cases to be assigned (N_+ in Fig. 3) should be assigned at the beginning of the enrollment period.

Second, at any particular time, families selected in the field are in one of four categories: (1) they have dropped out of the sample because of ineligibility, etc.; (2) they have accepted; (3) they have refused; (4) they are in an indeterminate state because the interviewers have not attempted to contact them, or they have been hard to contact, or their eligibility is being verified, or there has been a change in family composition since the baseline interview, or they are pondering the offer, and so on. The indeterminate category tends to be large through most of the enrollment period, and until it diminishes significantly, backup selections are not useful.

Third, a long interval, sometimes 6 weeks, elapsed between the time the field staff declared that a family could not be enrolled and the time a replacement selection could be fielded. Although quicker communications could have been designed, some of the clerical and data processing procedures necessary to maintain control in administering such a large survey would have had to be circumvented. Processing data about the refusal, making the new selection, preparing legal documents for the newly selected family, and integrating the new selections into the field schedule involved coordination among Rand and two subcontractors that required careful control procedures and, inevitably, time.

Fourth, the end of the field enrollment period in the HIS came not long after most families had achieved a final enrollment/nonenrollment status, and by that time it was too late for the field to enroll new families with enrollment complications. Replacing families near the end of the enrollment period could create a bias, because unless enrollers are given sufficient time to work on families with enrollment complications (e.g., replacing families who moved out, or are changing in composition, or are contemplating the offer), the experiment would be overloaded with problem-free families, and the acceptance rate would be lowered. Obviously, the responses of problem-free families to the treatments may well differ from those of the target population.

Fifth, the number of available backup families tends to be small in any stratum, unless large amounts are spent generating baseline interviews to purchase protection. Even if the true nonacceptance probability is the same for every stratum, random differences in the actual

acceptance rate would deplete some strata of the necessary reserves, forcing some nonproportional stratification at the backup stage.

We will not consider the case in which the acceptance rate varies by stratum, or worse yet, varies on the basis of some unmeasured variable. The former problem can be corrected by using the true acceptance probabilities (although defining and estimating them may be difficult); the unmeasured variable problem is much harder.

The full benefits of proportional stratification in the HIS were reduced further by the need to make selections in "bursts." Burst is a term used in the HIS to refer to selections being made not one at one time, but in stages. The baseline period ended close to the beginning of the enrollment period so that we could reduce the number of transitory changes in the sample, reduce cost, and get the experimental data as early as possible. Because of the time lag between administration of the baseline and availability of machine-readable data for selection (often several months), we had to make selections before all baseline data were available. Thus, selections were made in stages, with each stage called a "burst." The number of bursts ranged from one to nine in the sites. Families that were harder to reach or whose insurance took longer to verify tended to appear in the later bursts. Because they are likely to differ systematically from other families (e.g., families living in multiple family units tended to be associated with later bursts), the burst itself was considered as a stratum. That is, each family was identified with a burst, and the assignment of families to treatments was made proportionally from each burst. Bursts degrade the efficiency gains of proportional sampling relative to random sampling because they reduce the number of options available for balancing the sample.

The final constraint we will mention here is that the overall acceptance rate π is not known until after enrollment is complete, and one must work with an estimate $\hat{\pi}$. In the HIS, π differed from site to site, so that we were operating under uncertainty in every case. Furthermore, in the early bursts, one cannot know for sure how many eligible families, which we shall call N_E, eventually will become available.

These two uncertainties make uncertain what portion of families should be assigned to treatments in early bursts, and what portion should be withheld. Substantial penalties are paid for significant errors in either direction. Two few assignments cause field schedules to slip and lead to low field morale because enrollers, ready to work, don't have enough to do. Eventually they become overworked near the end of the enrollment period, with the result that a full effort will not be made on the families selected with the final bursts. In this case some bias will result from undersampling those families. Of course, when it was realized later that more families should have been assigned from the early bursts, these selections would be made, but it is the release of these families that overburdens the field. On the other hand, too many selections from an early burst leads to overenrollment of that group, and if the enrollment targets (and budget) are not to be exceeded, the situation cannot be corrected later. If π and N_E are known, and if n_+ is the number of acceptances desired in the site, then the fraction of the burst that should be assigned is

$$\frac{(n_+/\pi)}{N_E}, \tag{5.15}$$

since n_+/π is the number of selections that eventually would have to be made to get n_+ acceptances, and N_E is the number of families eventually eligible for selection. Therefore the product

$$\pi N_E \tag{5.16}$$

must be estimated. The conservative approach would choose a high value for this product, making undersampling of the burst much more likely than oversampling, and compensating later. This means the field period must be extended enough to provide time to pursue the replacement families from the early bursts as the actual number needed becomes known later.

Our recommendation is to protect against this situation by earmarking a balanced fraction of the enrollments (we used 20 percent in

the last sites) and instructing the field office to set them aside until the other enrollment selections have been fielded and followed up. This keeps options open until near the end of the fielding period and eliminates the lag time in the likely event that some of the remaining selections must be used.

To summarize this section, blocking, proportional stratification, and the Finite Selection Model all can be used to improve the balance and precision of the estimates from an experiment. But the gains that they provide relative to simple random sampling are reduced by random nonacceptances, approximately in proportion to the nonacceptance rate. We discussed the principal reasons why, in the context of the HIS, a sequential procedure designed to replace nonaccepting families with others from the same stratum was infeasible, even though, if successful, such a procedure would reclaim the precision losses due to nonacceptances.

APPENDIX

Carl Morris

Definition: Given k, n_+, and $(N_1, \ldots, N_k)$, the random vector $(n_1, \ldots, n_k)$ of nonnegative integers with $\Sigma\, n_i = n_+$ is distributed as the *Multivariate Hypergeometric Distribution*, $MHG_k\,(n_+; N_1, \ldots, N_k)$, if

$$P(n_1 = n_1^*, \ldots, n_k = n_k^*) = \frac{\binom{N_1}{n_1^*} \cdots \binom{N_k}{n_k^*}}{\binom{\Sigma\, N_i}{n_+}} . \qquad (A.1)$$

The first two moments of $MHG_k\,(n_+; N_1, \ldots, N_k)$ are

$$En_i = n_+ p_i , \qquad \text{var}\,(n_i) = n_+ p_i (1 - p_i) \frac{N_+ - n_+}{N_+ - 1} , \qquad (A.2)$$

where $N_+ \equiv \Sigma\, N_i$, $p_i \equiv N_i/N_+$. The expected value of the reciprocal of the ith coordinate of MHG_k does not exist if $P(n_i = 0) > 0$, but

$$E\left(\frac{1}{n_i} \middle| n_i \geq 1\right) \doteq (1 + \gamma_i^2) \frac{1}{En_i} , \qquad (A.3)$$

with γ_i^2 being the squared coefficient of variation of n_i given by

$$\gamma_i^2 = \frac{\text{var}\,(n_i)}{E^2 n_i} = \frac{1 - p_i}{n_+ p_i} \frac{N_+ - n_+}{N_+ - 1} . \qquad (A.4)$$

Formula (A.3) is determined by noting that since

$$\frac{\mu}{X} = 1 - \left(\frac{X - \mu}{\mu}\right) + \left(\frac{X - \mu}{\mu}\right)^2 - \left(\frac{X - \mu}{\mu}\right)^3 \frac{\mu}{X} ,$$

for any random variable X with mean μ,

$$E\,\frac{\mu}{X} = 1 + \frac{\text{var }(X)}{\mu^2} - E\left(\frac{X-\mu}{X}\right)\left(\frac{X-\mu}{\mu}\right)^2 .$$

The remaining term $E[(X-\mu)/X][(X-\mu)/\mu)^2]$ is ignored, which is legitimate for the expectation in (A.3) involving the hypergeometric distribution, provided $P(n_i = 0)$ is small and γ_i is small. In the examples we have looked at, satisfying $n_+p_i \geq 4$ and $\pi \geq 0.5$ [note that $1 - \pi \doteq (N_+ - n_+)/(N_+ - 1)$ in (A.4)], the right-hand side of (A.3) is smaller than the left-hand side, but by less than 4 percent, with diminishing error as n_+p_i increases. In Theorems (5.1) and (5.2), these errors are averaged over all strata, with smaller weights for small strata, thereby improving further the approximations in those theorems.

We turn to the proof of Theorems (5.1) and (5.2). Under the assumtion of Theorem (5.1), N_i subjects are chosen at random for treatment T_i from the N_+ available. Then n_i accept at random from the N_i. Hence, the accepting subjects also are a random sample of size n_i from the N_+ available. Since the stratum sizes available are $(p_1N_+, \ldots, p_kN_+)$, the conditional distribution of the numbers selected for T_i, i.e., of $(m_{i1}, \ldots, m_{ik})$, subject to $\Sigma_j\, m_{ij} = n_i$ with n_i given, is

$$(m_{i1}, m_{i2}, \ldots, m_{ik}) \sim \text{MHG}_k(n_i;\, p_1N_+, \ldots, p_kN_+) . \qquad \text{(A.5)}$$

It follows from (A.2), (A.3), and (A.4) that

$$\sigma^2 E_{SRS}\, \Sigma\, p_j^2/m_{ij} \doteq \sigma^2\, \Sigma\, p_j^2\, \frac{1}{n_ip_j}\left(1 + \frac{1-p_j}{n_ip_j} \cdot \frac{N_+ - n_i}{N_+ - 1}\right), \qquad \text{(A.6)}$$

and this reduces to (5.7).

Under the assumption of n_i random acceptances from N_i offers, where proportional stratification requires $m_{ij}^* = N_ip_j$ assignments to be made from S_j to T_i,

$$(m_{i1}, \ldots, m_{ik}) \sim \mathrm{MHG}_k(n_i;\ p_1 N_i, \ldots, p_k N_i) \ . \qquad (A.7)$$

This is formally equivalent to (A.5), with N_i replacing N_+. Theorem (5.2) therefore follows from Theorem (5.1), with N_+ replaced by N_i.

Finally, (5.14) needs proof. If n_i^* minimizes $\sigma^2 \Sigma\, w_i/n_i$ subject to the budget constraint $\Sigma\, c_i n_i = C$, then

$$n_i^* \propto (w_i/c_i)^{\frac{1}{2}} \ . \qquad (A.8)$$

With $n_i \sim$ binomial $(n_i^*/\pi,\ \pi)$, then $En_i = n_i^*$, var $(n_i) = n_i^*(1 - \pi)$. The approximation (A.3) also holds for the binomial distribution; therefore,

$$E\ 1/n_i \doteq [1 + (1 - \pi)/n_i^*]/n_i^* \ .$$

It follows that

$$\sigma^2 E\ \Sigma\, w_i/n_i \doteq \sigma^2\ \Sigma\, w_i/n_i^*[1 + (1 - \pi)/n_i^*]$$

$$= \sigma^2\ \Sigma\, w_i/n_i^* \left[1 + (1 - \pi)\ \frac{\Sigma\, w_i/n_i^{*2}}{\Sigma\, w_i n_i^*/n_i^{*2}}\right]$$

$$= \sigma^2\ \Sigma\, w_i/n_i^*[1 + (1 - \pi)/E_c n^*] \ , \qquad (A.9)$$

since $w_i/n_i^{*2} \propto c_i$. Formula (5.14) follows.

REFERENCES

1. Conlisk, John, "Choice of Response Functional Form in Designing Subsidy Experiments," *Econometrica*, Vol. 41, No. 4, July 1973, pp. 643-656.

2. Conlisk, John, and Harold Watts, "A Model for Optimizing Experimental Designs for Estimating Response Surfaces," *Proceedings of the Social Statistics Section*, American Statistical Association, 1969, pp. 150-156.

3. Hall, Robert E., "Effects of the Experimental Negative Income Tax on Labor Supply," in Joseph Peckman and P. Michael Timpane (eds.), *Work Incentives and Income Guarantees*, Washington: Brookings, 1975, pp. 115-147.

4. Manning, Willard G., Bridger M. Mitchell, and Jan P. Acton, "Design of the Los Angeles Peak-Load Pricing Experiment for Electricity," The Rand Corporation, R-1955-DWP, November 1976.

5. Morris, Carl, "A Finite Selection Model for Experimental Design of the Health Insurance Study," *Proceedings of the Social Statistics Section*, American Statistical Association, 1975, pp. 78-85; also published in the *Journal of Econometrics*, Vol. II, No. 1, September 1979, pp. 43-61.

6. Newhouse, Joseph P., "A Design for a Health Insurance Experiment," *Inquiry*, Vol. 11, No. 1, March 1974, pp. 5-27.

7. Phelps, Charles E., "Demand for Reimbursement Insurance," in Richard Rosett (ed.), *The Role of Health Insurance in the Health Sciences Sector*, New York: National Bureau of Economic Research, 1976.