17
2-4-80

ornl

ORNL/TM-7084

OAK
RIDGE
NATIONAL
LABORATORY

UNION
CARBIDE

MASTER

Sampling Theory Methodology Applicable to Data Validation Studies

Michael R. Chernick

OPERATED BY
UNION CARBIDE CORPORATION
FOR THE UNITED STATES
DEPARTMENT OF ENERGY

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency Thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

Printed in the United States of America. Available from
National Technical Information Service
U.S. Department of Commerce
5285 Port Royal Road, Springfield, Virginia 22161
NTIS price codes—Printed Copy: A05; Microfiche A01

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States nor any agency thereof, nor any of their employees, makes any warranty, expressed or implied, or assumes any legal liability or responsibility for any third party's use or the results of such use of any information, apparatus, product or process disclosed in this report, or represents that its use by such third party would not infringe privately owned rights.

DISCLAIMER

This book was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

ORNL/TM-7084

Contract No. W-7405-eng-26

Regional and Urban Studies Section

Energy Division

SAMPLING THEORY METHODOLOGY APPLICABLE TO DATA VALIDATION STUDIES

Michael R. Chernick

NOTICE: This document contains information of a preliminary nature. It is subject to revision or correction and therefore does not represent a final report.

Sponsored by
Office of Energy Information Validation
Energy Information Administration
of the
Department of Energy

Date Published - January 1980

OAK RIDGE NATIONAL LABORATORY
Oak Ridge, Tennessee 37830
operated by
UNION CARBIDE CORPORATION
for the
DEPARTMENT OF ENERGY

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

THIS PAGE

WAS INTENTIONALLY

LEFT BLANK

CONTENTS

THIS PAGE

WAS INTENTIONALLY

LEFT BLANK

The following is a list of related reports completed at ORNL. Copies are available from the National Technical Information Service, U.S. Department of Commerce, 5825 Port Royal Road, Springfield, VA 22161.

Title	Author(s)	Document Number	Date
The Influence Function and Its Application to Data Validation	M. R. Chernick	ORNL/TM-6871	1979
A Rigorous, Systematic Approach to Automatic Data Editing and Its Statistical Basis	G. E. Liepins	ORNL/TM-7126	1979
Refinements to the Boolean Approach to Automatic Data Editing	G. E. Liepins	ORNL/TM-7156	1979

THIS PAGE

WAS INTENTIONALLY

LEFT BLANK

SAMPLING THEORY METHODOLOGY APPLICABLE TO DATA VALIDATION STUDIES

Michael R. Chernick

ABSTRACT

In data validation studies, surveys are conducted to obtain information about the data collection process and the uses of the data. In many cases standard sampling techniques can be used. Two methods, stratified random sampling and cluster sampling, were used for surveys in the Form 4 data validation study. Form 4 is a data collection system on monthly generation and consumption of fuels by electric power plants. Those applications are described in Section B.

Sometimes time and cost constraints make more sophisticated controlled sampling approaches necessary. One such approach using balanced incomplete block designs is described in Section C and an appendix surveying the existence results for these type designs is included in the report. Sequential methods which may prove to be more cost effective are discussed in Section D. Sequential approaches to the problem of determining the size of a population is also discussed in Section D.

Problems requiring further research are discussed in Section E. This includes some preliminary results on the problem of stratification with respect to more than one variable. The results were obtained for the Form 4 respondent population.

The Form 4 study indicated that standard statistical sampling methods could be useful in data validation surveys. For example, at least 30 percent of the respondents do not report net generation as the instructions define it, and only 25 percent of the state regulatory agencies use the Form 4 data. Such inferences were possible only because statistical sampling procedures were used.

A. INTRODUCTION

In data validation studies, surveys are conducted to obtain information about the data collection process and the uses of the data. These surveys are referred to as respondent and user surveys.

Often in the case of respondent surveys the population from which the sample is taken is known, and one wants to design a sampling plan from which inferences can be drawn about the population. For such cases, techniques such as pure random sampling, stratified random sampling, and one- or two-stage cluster sampling can usually be applied. Some of these methods will be described in Sect. B; they are standard and are covered thoroughly by Cochran [1977].

Sometimes additional controls are required on the sampling plan. In such cases, controlled random sampling procedures may prove useful. One method based on the use of balanced incomplete block designs is given by Avadhani and Sukhatme [1973] and is discussed in Sect. C.

In some cases, there may be advantages in using sequential sampling methods. The theory of sequential sampling in finite populations is not a well-developed area. A recent paper by Lai [1979] deals with the finite population problem. Lai's stopping rule is applicable to the problem of testing a hypothesis about a proportion in the population. His method or an extension of it could be used for a sequential test of the proportion of respondents in the population who would give a particular answer to a survey question. Lai's method and a method based on the technique of Sobel and Wald [1949] are given in Sect. D.

For some data validation surveys (in particular, user surveys) the size of the population is unknown. We may be interested in estimating the population size or determining a stopping procedure, which assures us of a high probability that we have found all the users. The first problem has been approached in several ways, while the latter problem has not received much attention in the literature. One solution to the second problem is given by Darling and Robbins [1967]. Several authors have treated the first problem. A good survey of the literature on capture-recapture methods is given by Johnson and Kotz [1977, pp. 248–58]. Solutions involving log linear models are given by Bishop, Fienberg, and Holland [1975, pp. 229–56]. Efron and Thisted [1975] use an empirical Bayes approach to the problem. Govindarajulu [1975] discusses the sequential estimation problem as treated by Samuel [1968]. Some of these methods are discussed in Sect. D.

In all cases some modeling assumptions are made about the sampling mechanism, which will not strictly hold in the actual data validation surveys. The effect of the assumptions on the estimates is a subject for further research. The author of this paper has conducted a test, using literature articles on outlier methods as the population and reference lists from chosen articles as the sampling mechanism. Results of the test and implications of the results are given in Sect. D.

This document is preliminary for several reasons. First, the author has had limited experience with data validation projects, and the methods described herein are only those which have proved useful to date. It is expected that several other sampling problems and

methods of solution will arise in the future. Second, some of the methods described in this paper require further research as to their applicability. Third, better methods of solution may be developed through further research. A discussion of possible approaches to these problems, which require further research, is given in Sect. E. Also, a stratified sampling problem for which current research holds promise for solution is given in Sect. E.

B. STANDARD SAMPLING METHODS

Many of the standard sampling techniques described in textbooks such as Cochran's [1977] are applicable to data validation surveys. Two techniques, stratified random sampling and cluster sampling, were used for surveys in the FPC form 4 data validation study. Those applications are described in this section to illustrate their usefulness for future validation efforts.

We shall first describe the basic concept of random sampling. We start with a finite population of size N. We select n members from the population, and based on these n observations we wish to draw inferences about the population. There are $\binom{N}{n}$ distinct ways of choosing n members out of the population of size N. A sample of n from the population is said to be chosen at random if the $\binom{N}{n}$ choices are equally likely.*

*This is the definition used in most survey sampling texts for a pure or simple random sample.

Let $Y_1, Y_2, \ldots, Y_N$ denote the values in the population of a variable which would be measured for each unit in the chosen sample. The population total, $\sum_{i=1}^{N} Y_i$, population mean, $\sum_{i=1}^{N} \frac{Y_i}{N}$, and population variance, $\sum_{i=1}^{N} \frac{(Y_i - \overline{Y})^2}{N-1}$, are all parameters of the population which we may choose to estimate. The population mean is denoted by $\overline{Y}$ and the variance by S^2.

For a random sample, the sample mean and sample variance are unbiased estimates of $\overline{Y}$ and S^2 respectively. The proofs are given by Cochran [1977]. The variance of the sample mean is $\frac{S^2}{n}(1 - \frac{n}{N})$. The quantity $(1 - \frac{n}{N})$ is called the finite population correction, since the variance would be $\frac{S^2}{n}$ if the population were infinite.

Often we can divide the population into distinct categories called subpopulations. If we have L subpopulations and N_i members in the i^{th} subpopulation for i = 1, 2, ..., L, we can choose a random sample of size $n_i \leq N_i$ from the i^{th} subpopulation. A sample obtained in this manner is called a stratified random sample. For any stratified sampling plan, unbiased estimates can be obtained for the population mean and variance. Stratified random sampling can be advantageous when the variability within the subpopulations is less than the variability between the subpopulations. When this is the case, we have for a total sample of size n a smaller variance for the estimate of the mean from the stratified random sample than for the estimate we would get from a purely random sample. When the total sample size n for a stratified random sample is fixed, the choice of the sizes n_i, i = 1, 2, ..., L, for the subpopulations is called the allocation scheme (i.e., a scheme which minimizes some specified loss function).

In practice we may have some information about the variability within the subpopulations, but it would be unlikely that we would know the variances. In such cases we would choose an allocation that was not optimal but would probably give better estimates than a pure random sample of comparable size. A proportional allocation scheme (i.e., $n_i \approx \frac{nN_i}{N}$) will often satisfy this purpose when information about the variability within the subpopulations is vague. Such an allocation scheme was used for the telephone survey of form 4 respondents.

In the form 4 respondent survey it seemed plausible that the variability in response to some survey questions might be related to the size of the power company. Consequently, we decided to stratify according to company size. A study of the frequency distribution of companies according to size indicated a logical breakout into six categories. Since the sixth category represented such a small percentage of the total number of companies, we decided to combine it with category five.

An optimal allocation scheme was not attempted, since there would be many variables of potential interest. Even if enough information were available for an optimal allocation scheme with respect to one of these variables, the scheme could be far from optimal for other variables.

We chose a proportional allocation scheme, since it would be better than a pure random sample if our assumptions about the responses were correct. On the other hand, if the responses are not at all related to size class, the stratified sample would do about the same as a pure random sample.

Geographical location was another variable which could be related to the responses. Since location would be another candidate for stratification, we chose to consider the nine census regions and determine the distribution of size class for each of the regions. If the distributions were similar, we could use a stratified random sample with proportional allocation among size class and not expect to induce significant bias with respect to census region.

This statistical problem could be treated more precisely. We choose to stratify a sample one way, thus exercising some control over the sample. We would like the sample to be distributed "nicely" with respect to another possible stratification scheme without exercising further control. We need to define precisely what we mean by distributed nicely. One possible definition is that the number of samples in specific categories fall within some limits. The work of Sobel, Uppuluri, and Frankowski [1977] may give answers to the probability of occurrence for the event nicely distributed. The possibility of further research on this type of problem is discussed in Sect. E.

A sample of size 31 was chosen for the form 4 respondent survey, since a sample of about 30 was considered adequate, and 31 gave a closer approximation to the proportional allocation. Note that an exact proportional allocation would not in general give an integer allocation to each subpopulation.

The sample of size 31 was taken from a population of 1474. The stratified random sample enables us to compute unbiased estimates of the proportion of the population giving a specific response and assess the variability of those estimates.

The stratified random sample would probably yield estimates with better precision than would a pure random sample and hence narrower confidence intervals. Table 1 gives the standard deviations and approximate confidence intervals for proportions when the true proportion p ranges between 0.1 and 0.9; $\hat{p}$ is the unbiased estimate for p.

Table 1. Standard deviations and confidence intervals for proportions (p)

p	Standard deviation	Approximate 95% confidence interval[a]
0.1	0.0529	[p − 0.1058, p + 0.1058]
0.2	0.0714	[p − 0.1428, p + 0.1428]
0.3	0.0812	[p − 0.1624, p + 0.1624]
0.4	0.0871	[p − 0.1642, p + 0.1642]
0.5	0.0889	[p − 0.1778, p + 0.1778]
0.6	0.0871	[p − 0.1642, p + 0.1642]
0.7	0.0812	[p − 0.1624, p + 0.1624]
0.8	0.0714	[p − 0.1428, p + 0.1428]
0.9	0.0529	[p − 0.1058, p + 0.1058]

[a]Confidence limits are obtained using the normal approximation without the continuity correction.

These intervals are quite broad, indicating that a sample of size 30 will not give precise estimates. In order to reduce the width of

the interval from 0.3 to 0.1, we would need to increase the sample size by nearly a factor of 9. The gain in precision is probably not worth the expense of taking 280 samples.

Besides telling us about the precision of our estimates, the intervals enable us to make some valuable inferences. If, for example, we obtain an estimate for the percentage of respondents reporting negative generation of $\hat{p} = 0.8$, we can say with high confidence that the proportion in the population is greater than 0.65 and less than 0.95. From this we can conclude with near certainty that a majority of the population report negative generation. On the other hand, if $\hat{p}$ were near 0.5 we could not be confident that a majority report negative generation. We could, however, conclude that a significant proportion of the population report negative generation, since for $\hat{p} = 0.5$ we would get an interval of [0.32, 0.68].

In some sampling problems it is not easy to identify the sampling units in the population, but is rather easy to identify groups of sampling units. When groups are selected at random and all sampling units are taken within the group, the sampling procedure is called single-stage cluster sampling. When, in addition to selecting the group at random, a random subset of the units within the group is taken, the procedure is called two-stage cluster sampling.

In the FPC form 4 state users survey, a two-stage cluster sample seemed appropriate; the sampling unit was state agencies. It was known that each state has at least one regulatory agency, but a list of the agencies was not available. It was convenient to pick the

states at random, contact the selected states to get a list of their agencies, and then choose a random sample from the list.

We consider a general problem where there are N states and M agencies within each state. Time and cost considerations restrict us to a total sample size of L or less. Our sampling procedure will be to choose n of the N states at random and then for each state choose m of the M agencies at random. The allocation problem is to decide n and m, given the requirement $nm \leq L$.

One way to choose is to pick a parameter of interest and minimize the variance of an unbiased estimate of the parameter, keeping the restriction in mind.

One example of a parameter of interest from the survey would be the proportion of the population of state agencies which use form 4 data.

Let $Y_{ij} = 0$ if i^{th} agency in the j^{th} state does not use form 4 data

$= 1$ otherwise.

Let $\bar{y} = \sum_{i=1}^{n} \frac{\sum_{j=1}^{m} Y_{ij}}{nm}$;

$\bar{y}$ is then the proportion in the sample which uses form 4 data, and since the states and the agencies were chosen at random, it is an unbiased estimate of the parameter.

$$\text{Let } \bar{\bar{Y}} = \sum_{i=1}^{N} \frac{\sum_{j=1}^{M} Y_{ij}}{NM}, \quad S_1^2 = \sum_{i=1}^{N} \frac{(\bar{Y}_i - \bar{\bar{Y}})^2}{N-1},$$

$$S_2^2 = \sum_{i=1}^{N} \sum_{j=1}^{M} \frac{(Y_{ij} - \bar{Y}_i)^2}{N(M-1)}, \text{ where } Y_i = \sum_{j=1}^{M} \frac{Y_{ij}}{M}.$$

$\bar{\bar{Y}}$ is the parameter of interest, and S_1^2 and S_2^2 are population parameters which affect the variance of $\bar{y}$; S_1^2 is called the variance among primary unit means, and S_2^2 is called the variance among sub-units within primary units. In our example the primary units are the agencies. The variance of $\bar{y}$ is derived by Cochran [1977, p. 277]:

$$V(\bar{y}) = \frac{(N-n)}{N} \frac{S_1^2}{n} + \frac{(M-m)}{M} \frac{S_2^2}{mn}.$$

Our goal is to minimize $V(\bar{y})$, subject to $nm \leq L$. We see that

$$V(\bar{y}) = \frac{S_1^2}{n} - \frac{S_1^2}{N} + \frac{S_2^2}{mn} - \frac{S_2^2}{Mn},$$

and since $\frac{S_1^2}{N}$ is constant, minimizing $V(\bar{y})$ is equivalent to minimizing

$$\frac{S_1^2}{n} + \frac{S_2^2}{mn} - \frac{S_2^2}{Mn},$$

subject to $nm \leq L$. If we choose $nm = L$, the term $\frac{S_2^2}{mn}$ is as small as possible, and this choice does not affect the other terms. We have thus reduced the problem to minimizing $\frac{S_1^2}{n} - \frac{S_2^2}{Mn}$, subject to $nm = L$.

We see that the solution depends on whether or not $MS_1^2 > S_2^2$. If $MS_1^2 > S_2^2$, we minimize the variance by making n as large as possible, and if $MS_1^2 < S_2^2$, the minimum occurs when m is as large as possible. So if $L < N$ and $MS_1^2 > S_2^2$, we choose $n = L$ and $m = 1$. If $L > N$ and $MS_1^2 > S_2^2$, we take $n = N$ and $m = L/N$ if L/N is an integer. When $MS_1^2 < S_2^2$, we take $n = 1$ and $m = L$ if $L \leq M$, or we take $n = L/M$ and $m = M$ if $L > M$ and L/M is an integer. In our case, $N = 50$, L was chosen to be 30, and M was believed to be 2. It turned out that M was 1 for some states and 3 or more for other states. Although this violation of the assumptions has an effect on the variance formula and the feasibility of the optimal allocation, the necessary adjustments were not considered serious.

In the case of proportion of agencies using form 4, we do not even know whether or not $MS_1^2 > S_2^2$. Assuming $N = 50$, $M = 2$, and $L = 30$, we would either want to choose $n = 30$ and $m = 1$, or $n = 15$ and $m = 2$. Answers to other questions in the survey provide estimates for other parameters of interest from the population. It is likely that $n = 15$ and $m = 2$ would be optimal for estimating some of these parameters. Note that the optimization depended only on the estimate being a sample mean and the relationship between S_1^2 and S_2^2. The choice $n = 15$ and $m = 2$ seems to be best, since it would likely be optimal for some of the estimates and would probably not increase the variance

very much for those cases where n = 30 would be optimal. We note that when m = M = 2,

$$V(\bar{y}) = \frac{(N - n)}{N} \frac{S_1^2}{n} .$$

The finite population correction $\frac{(N - n)}{N}$ is 0.7, and hence it is significant.

Since some of the chosen states had three agencies and others had only one, m was chosen to be 3 for some. A total of 30 state agencies was chosen. This change in procedure will have some effect on the variance of the estimates.

This example shows the value of using cluster sampling methods in data validation surveys. It is the first application to the author's knowledge of the use of this standard sampling method.

Unfortunately, stratified random sampling, a very commonly used sampling procedure, has not been used much in data validation surveys. It appears to the author that a proportional allocation scheme for a stratified random sample could be reasonably applied to many respondent surveys in data validation studies.

In Sects. C, D, and F, special sampling techniques are described which we feel will have some application in data validation studies. Some of these methods can be applied immediately, while others may be improved by further research. We feel, however, that the basic methods of survey sampling, of which the two described in this section are examples, can have a major and favorable impact on data validation surveys, if properly applied.

C. CONTROLLED RANDOM SAMPLING

Pure random sampling without replacement, when used to draw a sample of size n from a population of size N, gives equal probability of selection to each of the $\binom{N}{n}$ possible samples. In some sampling situations, some of the possible samples may be undesirable, possibly because the sampling units are too widespread. The selection of such a sample might then lead to excessive traveling expense or require much more time for the collection of the sample. The techniques of stratification and clustering described in the previous section can be used to exercise some control on the sample. However, even after stratification, there may be a need to control the selection of samples within the subpopulations. Cluster sampling has the drawback that it usually causes a reduction in precision of the estimate of the parameter of interest. The method given by Avadhani and Sukhatme [1973] is a simple procedure which gives the undesirable samples (called nonpreferred samples) lower probability of occurrence than for the preferred samples without loss of precision in the estimate. In some cases, some or all of the nonpreferred samples can be given zero probability of occurrence.

The method is based on the use of balanced incomplete block designs, the theory of which was developed to provide good sampling designs in experiments. It is used here in a different context. Nevertheless, all known results on the existence and construction of balanced incomplete block designs are relevant, since the more we know about the construction of these designs, the more flexibility we have

in constructing a controlled random sample. There is a vast literature on existence and construction of these designs which will be summarized in the Appendix.

In an agricultural experiment, crops may be planted in rows with an equal number of crops in each row. These rows are referred to as blocks. Several different types of fertilizer are applied to the crops. The distinct types of application of fertilizers are referred to as treatments. In order to obtain estimates of differences between treatments for all possible pairs with equal precision, the balanced incomplete block design was devised. It requires that each treatment occurs in the same number of blocks and that each possible pair of treatments occurs in the same number of blocks.

In a survey sampling problem, the blocks correspond to the samples to choose from with equal probability. That each sample unit occurs the same number of times implies that the sample mean will be an unbiased estimate of the population mean just as for a pure random sample. The condition on the appearance of pairs of sample units implies that the variance of the sample mean will be the same as for a pure random sample. So there is no loss in precision.

To clarify these ideas we will consider a simple example. Suppose we have a population of size 6 and want to take a sample of size 3. We label the sampling units with 1, 2, ..., 6. In pure random sampling we choose from the $\binom{6}{3} = 20$ possible samples, each with probability 1/20. Now consider the following balanced incomplete block design:

124	125	135	136	146
234	236	256	345	
456				

Each of the ten samples listed would be taken with probability 1/10. We see that each unit occurs five times in the design, while each pair occurs in two blocks. We have eliminated half of the possible samples. If we had several nonpreferred samples, we could include some of them in the ten samples excluded from the design. For example, the triple 123 is not included in the design, and so we could let 123 designate an unpreferred sample, thus preventing its selection.

Let Y_i, $i = 1, 2, \ldots, 6$, denote the observations and $\overline{Y} = \sum_{i=1}^{6} \frac{Y_i}{6}$ be the population mean. Let X_1, X_2, and X_3 denote the sample values and $\overline{X} = \frac{X_1 + X_2 + X_3}{3}$. We see that $E(\overline{X}) = \overline{Y}$, since

$$E(\overline{X}) = \frac{1}{10}\left[\frac{Y_1 + Y_2 + Y_4}{3} + \frac{Y_1 + Y_2 + Y_5}{3} + \frac{Y_1 + Y_3 + Y_5}{3} + \frac{Y_1 + Y_3 + Y_6}{3}\right.$$

$$+ \frac{Y_1 + Y_4 + Y_6}{3} + \frac{Y_2 + Y_3 + Y_4}{3} + \frac{Y_2 + Y_3 + Y_6}{3} + \frac{Y_2 + Y_5 + Y_6}{3}$$

$$\left. + \frac{Y_3 + Y_4 + Y_5}{3} + \frac{Y_4 + Y_5 + Y_6}{3}\right] = \frac{1}{30}\, 5 \sum_{i=1}^{6} Y_i = \sum_{i=1}^{6} \frac{Y_i}{6} = \overline{Y}\,.$$

Also, $\text{Var}\,(\overline{X}) = \sum_{i=1}^{6} \frac{(Y_i - \overline{Y})^2}{6 \times 5}\left(1 - \frac{3}{6}\right) = \frac{1}{60} \sum_{i=1}^{6} (Y_i - \overline{Y})^2$,

the same as for a pure random sample.

If we are interested in estimating the population variance, the usual sample estimate is unbiased just as for a pure random sample. However, the variance of the estimator will be different from the variance of the corresponding estimator from a pure random sample. The author has done some calculations on simple examples which indicate that the differences can be small.

Suppose the population consisted of the following seven values: 1, 5, 6, 7, 10, 11, and 14. The population parameters are $\bar{Y} = 7.714$ and $S^2 = 18.57$. We compare a random sample of size three to the following controlled random sample:

124	135	167	236	257
347	456			

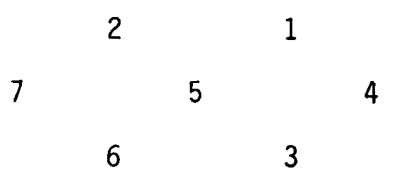
Each element appears three times and each pair once in the design. For the pure random sample the variance of the estimator of variance is 161.86, as compared with 161.91 for the controlled random sample.

When drawing a sample of size n from a population of size N, our ability to eliminate certain sample paths depends on the existence of a balanced incomplete block design with n elements in each block and less than $\binom{N}{n}$ blocks. For some values of n and N, no such design may exist, or even if it exists it may not be known. A complete characterization of balanced incomplete block designs has not been found to date.

In order to be more flexible, Avadhani and Sukhatme [1973] allow the nonpreferred samples to have smaller but nonzero probabilities. They divide the N samples into k groups containing N_1, N_2, ..., N_k units. The elements are placed in the groups at random. For each

$i = 1, 2, \ldots, k$ we choose an integer n_i such that $n_i < n_i' < N_i$, where $n = n_1 + n_2 + \ldots + n_k$, and n_i is proportional to N_i. The choice of n_i' is made so that a balanced incomplete block design exists with the parameters $(n_i', b_i, r_i, n_i, \lambda_i)$; n_i' denotes the number of treatments, b_i the number of blocks, r_i the number of replications of each treatment, n_i the block size, and λ_i the number of replications of pairs of treatments. We select a simple random sample of n_i' units from the i^{th} group and do this independently for each i. We then proceed to determine the sample for the i^{th} group just as in the simple examples described earlier. We make the correspondence between preferred samples if necessary. This makes the probability of drawing a nonpreferred sample as small as possible for this design.

Avadhani and Sukhatme [1973] give the following example: The illustration is from a survey carried out in India in 1964–1965. The objective was to obtain reliable estimates of acreage, yield rates, and total production of major fruit crops. The district under study was divided into several stratum, based on information from previous surveys. In one strata, there were 90 villages, and a sample of 9 villages was proposed to be selected. Because of the mountainous terrain and lack of good roads and transportation, it was desirable to keep the distance between the selected villages as small as possible. To show how controlled random sampling could have been used in this problem, they divided the 90 villages into three random groups of size 30; $\binom{30}{3}$ is a large number, so n_i' was taken to be 7. Suppose the 7 villages in group 1 (these 7 were chosen at random from the 30) are located as shown:



We have $\binom{7}{3} = 35$ possible combinations of size 3. They are: (1,2,3)*, (1,2,4), (1,2,5), (1,2,6)*, (1,2,7), (1,3,4), (1,3,5), (1,3,6)*, (1,3,7)*, (1,4,5), (1,4,6)*, (1,4,7)*, (1,5,6), (1,5,7), (1,6,7)*, (2,3,4)*, (2,3,5), (2,3,6)*, (2,3,7)*, (2,4,5) (2,4,6)*, (2,4,7)*, (2,5,6), (2,5,7), (2,6,7), (3,4,5), (3,4,6), (3,4,7)*, (3,5,6), (3,5,7), (3,6,7), (4,5,6) (4,5,7), (4,6,7)*, and (5,6,7).

The 14 combinations marked with the asterisk are not preferred. For a pure random sample the probability of obtaining a nonpreferred sample is $\frac{14}{35} = 0.4$. The following balanced incomplete block design contains only one nonpreferred sample:

(1,2,4)	(2,6,7)
(1,3,7)*	(3,4,6)
(1,5,6)	(4,5,7)
(2,3,5)	

So using the controlled sampling procedure the probability of obtaining a nonpreferred sample is only $\frac{1}{7} = 0.14$.

D. SEQUENTIAL SAMPLING METHODS

We shall first consider the problem of determining whether the fraction of a population which would give a yes response to a yes-no

question is greater than some specified proportion Θ_1 but less than some other proportion Θ_2. We consider this problem from the sequential point of view. We sample and decide to stop sampling based on a decision rule which leads us to one of three conclusions: (1) Θ is less than Θ_1, (2) Θ is between Θ_1 and Θ_2, (3) Θ is greater than Θ_2. When we use a sequential stopping rule, there is a probability of making an incorrect decision when using the rule. Sequential sampling procedures are chosen which control these errors.

Sobel and Wald [1949] considered this type of problem for testing the mean of a normal distribution. The method is also applicable to the binomial distribution, which would be the case here if the population were not finite. Lai [1979] considered the problem of testing $\Theta < \Theta_1$, vs $\Theta > \Theta_1$, in the case where the population is finite. Since we are considering the choice of three possible decisions instead of two, some modification of his decision rule would be necessary to apply it to the problem we are considering.

In many situations the population size is large, and the Wald-Sobel approach can be applied. Armitage [1950] gave a modification to the Wald-Sobel approach which avoids a technical difficulty, and his approach is considered to be preferable by some statisticians. When the population size is small, the Wald-Sobel approach leads to a conservative test. For some applications it may be too conservative. We shall describe the method and illustrate the conservative nature of the test for a small population size with a simple vs simple hypothesis test.

Sobel and Wald [1949] define indifference regions for θ. When $\theta \in (\theta_1^o, \theta_2^o)$, where $\theta_1^o < \theta_1 < \theta_2^o$, we do not care about differentiating between (1) and (2) but wish to reject (3). When $\theta \in (\theta_3^o, \theta_4^o)$, where $\theta_2^o < \theta_3^o < \theta_2 < \theta_4^o$, we want to reject (1) but do not care about distinguishing between (2) and (3).

For $\theta < \theta_1^o$ we wish to accept (1) and reject (2) and (3). For $\theta_2^o < \theta < \theta_3^o$ we wish to accept (2) and reject (1) and (3). When $\theta > \theta_4^o$ we wish to accept (3) and reject (1) and (2). Once we have specified θ_1^o, θ_2^o, θ_3^o, and θ_4^o, we have defined what constitutes a wrong decision for each θ in the parameter space (i.e., $0 \leq \theta \leq 1$).

We then do two sequential probability ratio tests simultaneously: θ_1^o, θ_2^o, θ_3^o, and θ_4^o are chosen so that

$$\theta_1^o < \theta_1 < \theta_2^o < \theta_3^o < \theta_2 < \theta_4^o,$$

$$\theta_1^o + \theta_2^o = 2\theta_1, \quad \theta_3^o + \theta_4^o = 2\theta_2,$$

$$\theta_2^o - \theta_1^o = \theta_4^o - \theta_3^o = \Delta.$$

The two likelihood ratios are

$$L_{1n} = \frac{P[X_1, X_2, \ldots, X_n \mid \theta = \theta_2^o]}{P[X_1, X_2, \ldots, X_n \mid \theta = \theta_1^o]},$$

$$L_{2n} = \frac{P[X_1, X_2, \ldots, X_n \mid \theta = \theta_4^o]}{P[X_1, X_2, \ldots, X_n \mid \theta = \theta_3^o]},$$

where $X_1, X_2, \ldots, X_n$ is the observed sequence of n observations. We stop sampling when both tests reach a conclusion. The stopping time T is the maximum of (T_1, T_2), where $T_1 = \inf (n: L_{1n} > A_1 \text{ or } L_{1n} < B_1)$,

and $T_2 = \inf (n: L_{2n} > A_2 \text{ or } L_{2n} < B_2)$. Let R denote this stopping rule and $L(\Theta|R)$ be the probability of a correct decision when Θ is the true value of the parameter and the rule R is used. For infinite populations, $L(\Theta|R)$ has the properties given by Sobel and Wald [1949], which are needed to keep the probability of a wrong decision less than some specified value γ for all Θ; γ will depend on the chosen A_1, A_2, B_1, and B_2.

To illustrate the conservative nature of the test in finite populations, we consider a simple vs simple hypothesis test: $H_0: \Theta = \Theta_1^{\,o}$ vs $H_1: \Theta = \Theta_2^{\,o}$ with $\Theta_2^{\,o} > \Theta_1^{\,o}$. The likelihood ratio statistic after the n^{th} observation would then be

$$L_{1n} = \prod_{i=0}^{n-1} \left[\frac{N\Theta_2 - S_i}{N\Theta_1 - S_i}\right]^{X_{i+1}} \left[\frac{N(1-\Theta_2) - i + S_i}{N(1-\Theta_1) - i + S_i}\right]^{1 - X_{i+1}},$$

where $X_1, X_2, \ldots, X_n$ is the observed sequence of zeroes and ones, $X_0 = 0$, and $S_i = \sum_{j=0}^{i} X_j$. N is the size of the population.

For the stopping time $T = \inf (n: L_{1n} > A \text{ or } L_{1n} < B)$ for $A > 1 > B > 0$, L_{1T} is a martingale with expected value 1 under H_0, and $1/L_{1T}$ is a martingale with expectation 1 under H_1. This implies that $\alpha = P$ [reject H_0 when H_0 is true] and $\beta = P$ [reject H_1 when H_1 is true] to a reasonable approximation satisfy

$$1 = \alpha A + (1 - \alpha)B ,$$

$$1 = \frac{\beta}{B} + \frac{(1 - \beta)}{A} .$$

Solving these equations yields $A = \frac{1-\beta}{\alpha}$ and $B = \frac{\beta}{1-\alpha}$. So for any specified α and β, A and B are determined to good approximations. For finite population size the theory does not go through, and the approximating equations do not hold. For some sample paths we are certain at time T which of H_0 or H_1 is true, and accordingly the likelihood ratio is either zero or $+\infty$. This leads one to suspect that the test would be conservative. This is borne out by Lai's work [1979], which shows that a reasonable procedure would be to apply a finite population correction to the stopping boundary.

To indicate what is going on, here is an example of a simple vs simple hypothesis test. We let $\theta_1^0 = 1/4$, $N = 8$, and $\theta_2^0 = 3/4$. Under both H_0 and H_1, there are 28 possible sequences of zeroes and ones which are equally probable. We let $\alpha^* = \beta^* = 0.1$, and $A = \frac{1-\beta^*}{\alpha^*}$ and $B = \frac{\beta^*}{1-\alpha^*}$; so $A = 9$ and $B = \frac{1}{9} \approx 0.11$. Under H_0 we reject H_0 for one sequence and accept it in the 27 other cases; so the true value of $\alpha = \frac{1}{28} \approx 0.036$. Similarly, under H_1, H_1 is rejected in one case, and hence $\beta = \frac{1}{28}$.

Under both H_0 and H_1 the average sample number is 3. The usual method for computing the average sample number is by Wald's lemma, which is not applicable in the finite population case, since the observations are not independent or identically distributed. Wald's lemma would give

$$E_{\theta_1}(T) = \frac{\frac{9}{10}\log B + \frac{1}{10}\log A}{-1/2 \log 3} = \frac{16}{5} = 3.2 \ , \ E_{\theta_2}(T) = \frac{16}{5} = 3.2 \ .$$

This is not too bad an approximation.

We note that when N is small we obtain lower error probabilities than we controlled for. A test procedure such as Lai's would not be so conservative and hence might have a smaller average sample number.

Lai [1979] formulates the problem as follows. He tests H: $\Theta \leq \frac{1}{2}(1 - \Theta_N)$ vs the alternative K: $\Theta > \frac{1}{2}(1 + \Theta_N)$, where N is the population size and the interval $[\frac{1}{2}(1 - \Theta_N), \frac{1}{2}(1 + \Theta_N)]$ is an indifference zone for the experimenter. For $0 < \alpha < \frac{1}{2}$, he defines

$$C_N = \min\left[\max\left(\{\log[(1-\alpha)/\alpha]\}/\{\log[(1+\Theta_N)/(1-\Theta_N)]\},\ 1\right),\ N(1-\Theta_N)+1\right].$$

He proposes the stopping rule

$$\tau = \inf\{n \geq 1:\ |S_n| \geq C_N - (n-1)(C_N - 1)/[N(1-\Theta_N)]\},$$

where $S_n = 2X_n - n$, and X_n is the number of yes votes in the first n samples. The term subtracted from C_N can be viewed as a finite population correction. He proves the following asymptotic result:

Theorem [Lai, 1979]: As $N \to \infty$ and $\Theta_N \to 0$ such that $N\Theta_N \to \infty$,

$$P_H[(\tau,\delta) \text{ rejects } H] = P_K[(\tau,\delta) \text{ rejects } K] \to \alpha .$$

Lai presents some results which show that the approximation can be reasonable even when $N = 200$ and $N\Theta_N = 6$.

Another problem which can be treated sequentially is the estimation of the size of a population. Solutions to this problem may have application to user surveys.

There has been much work done on this problem, using capture-recapture methodology. Other methods used are the empirical Bayes approach and stochastic modeling.

The capture-recapture method can be described by the following urn model. The urn contains an unknown number, N, of white balls and no others. An estimate of N is desired, based on a procedure of drawing balls at random one at a time and coloring the white balls black before returning them. Black balls are returned unchanged. Let w_i, b_i denote the number of white and black balls respectively observed in the first i draws.

Samuel [1968] considers this formulation of the problem. Others have considered the problem in the more general setup of sampling n_i balls at the i^{th} draw. Samuel considers the following stopping rules:

Rule A. Let $A > 0$ be a fixed integer $t_A = A$.

Rule B. Let $B > 0$ be a fixed integer $t_B = \inf (i | b_i = B)$.

Rule C. Let $C > 0$ be fixed,

$$t_C = \inf (i | b_i \geq Cw_i) = \inf [i | i \geq (C + 1)w_i].$$

Rule D. Let $-\infty < D < \infty$ be fixed,

$$t_D = \inf [i | b_i \geq \max (1, w_i \log w_i + w_i D)]$$

$$= \inf \{i | i \geq \max [w_i + 1, w_i \log w_i + w_i (D + 1)]\}.$$

Rule E. Let $\{D_j\}$ be such that $\lim D_j = \infty$,

$$t_E = \inf \{i | b_i > \max [1, w_i \log w_i + w_i (D_{w_i})]\}.$$

Rule D was considered by Darling and Robbins [1967]. They show that for $0 < \alpha < 1$ and a suitable choice of D, one obtains the probability that $W_D = N$ is greater than $1 - \alpha$ for any N.

For user surveys, one may start with an initial incomplete list of users and sample from this list. The users sampled will then be asked to name other users. The extended list could then be used to interview additional users and to continue to add to the list. A sequential rule would be used to decide when to stop interviewing. If the capture-recapture methodology is to be used to decide when a "good" estimate of the population size is available, we must determine that the urn model on which it is based is a good approximate model for the actual sampling mechanism. One way to test this would be to conduct a sample on a population of known size and see how good the estimates are. The author chose to use a bibliography of outlier papers. The bibliography was large, containing 419 references. The population was reduced from 419 to 141 so that only references available in the Mathematics and Statistics Research Department were included. This was done to expedite the identification of references.

The sampling procedure was as follows: Thirty reference papers were selected at random to form the initial list. One paper was chosen at random from the list, and the references listed in that paper were added to the list. If any of the references were on the initial list, they were noted as repeats. The procedure was conducted for 24 iterations, with estimates of the population size computed at each stage. The nearly unbiased estimator given by Johnson and Kotz [1977, p. 253] was used. This estimator has small bias when the urn model is

appropriate. However, in the outlier paper example, the model is not appropriate, and the estimator seemed to be highly underestimating the population size. Table 2 gives the results of the test, with the estimate of population size at each stage.

Let $N_0 = 30$ and N_i = number of papers on the i^{th} reference list. Let T_i = number of papers on the i^{th} list which are already on the population list, and let $U_i = N_i - T_i$ be the number of new articles in the list. The estimator was

$$\hat{f}_i^* = \frac{\left[\left(\sum_{j=0}^{i-1} U_j\right) + 1\right](N_i + 1)}{N_i - U_i + 1} - 1 .$$

The final estimate of 72.8 is only about half the size of the population. There is some dependence structure in the reference population which is causing this severe underestimation. One aspect of this dependence is that authors can only reference articles which preceded their paper. This type of temporal dependence may not be as severe for data form users, since they are asked to list contemporary users, and their knowledge of users may not be as dependent on the time at which they became users.

The example illustrates the potential difficulties that could arise from using capture-recapture methodology on user surveys. Although the method might not be as bad for user surveys as it was in this population size test, it may be necessary to test it on known user populations or to find other methods which would prove to be better in such tests.

Table 2. Population size test

Sample No.	N_i	U_i	T_i	f_i^*
0	30	30		
1	9	4	5	50.7
2	1	1	0	69
3	2	0	2	35
4	7	3	4	56.6
5	3	2	1	77
6	6	5	1	142.5
7	3	1	2	60.3
8	2	1	1	69.5
9	1	1	0	95
10	4	1	3	60.25
11	5	0	5	49
12	4	3	1	124
13	4	2	2	87.3
14	11	2	9	65
15	3	1	2	75
16	6	6	6	57
17	2	0	2	57
18	1	0	1	57
19	1	0	1	57
20	2	1	1	86
21	1	1	0	115
22	5	2	3	89
23	9	2	7	76.5
24	14	2	12	72.8

E. PROBLEMS FOR FURTHER RESEARCH

Although many of the sampling problems in data validation can be handled reasonably well with existing techniques, some problems occur which need special treatment. The idea of reducing survey costs by sequential testing has led the authors to tests such as Lai's solution to the hypergeometric sampling problem. Further research may lead to a modification of Lai's results which would be applicable to the three-decision problem.

That user populations for various data forms are unknown has led us to consider the sequential problems of estimating population size and of determining when the entire population has been identified with high probability. A preliminary test, using reference articles, has indicated that the various procedures need to be tested against an example where the population size is known and the population and sampling mechanism are very much like they are for a user survey. Methods using birth-death or other types of stochastic models should be studied and tested.

For simple stratified random sampling schemes, there are often two or more variables which are reasonable candidates for stratification. Many questions arise from this problem which require further research. For example, which variable would be the best single variable to stratify with? Cochran [1977, pp. 124–26] gives a method for two-way stratification. Is it desirable to control for more than one variable at a time? If we control for only one variable, what effect will this have on the distribution of the sample with respect

to another variable? The work of Sobel, Uppuluri, and Frankowski [1977] may be helpful in answering this last question.

At this time, some preliminary work of this type has been accomplished. We considered the data from the form 4 respondent survey given in the following table:

Distribution of respondents by census region and production capacity

Class size	Census region	1	2	3	4	5	6	7	8	9
1	126	5	7	15	42	11	2	15	22	7
2	423	19	13	50	228	20	2	34	34	23
3	570	37	35	121	109	83	23	69	44	49
4	211	18	11	24	38	28	9	35	26	22
5	144	9	16	27	12	23	11	20	10	16
Total	1474	88	82	237	429	165	47	173	136	117

The stratified random sample used in the form 4 study was as follows:

Class size	No. of samples
1	3
2	9
3	12
4	4
5	3

One important question which could be asked about this scheme is what is the probability that certain census regions will not be represented. Questions of this type for committee problems have been addressed by Walter [1976].

We computed these probabilities for each of the census regions by using the computer package for the multivariate hypergeometric distribution available from V. R. R. Uppuluri of the Mathematics and Statistics Research Department of Union Carbide Nuclear Division. The results are given in the following table:

Census region number	Probability of no observations
1	0.1483691
2	0.1658927
3	0.0039392
4	0.0000065
5	0.0241125
6	0.3655248
7	0.0203849
8	0.0467072
9	0.0764983

It is not surprising that census region 4 should have the lowest probability (a very low probability), since region 4 has the most companies of any region and has over half the total number of plants in class size 2. Census region 6 has the highest probability, since it has the fewest companies in each class size except number 5. The fact that this probability is as high as 0.366 is significant and indicates that if we wanted to be sure that this census region is represented, further controls on the sample would be necessary. In the actual sample, one member of census region 6 was selected. We now know that it could have easily happened that no samples were chosen from region 6.

Other questions such as the probability that at least one region is not represented or that the minimum number of representatives in any region is at least two (or any specified number) can be answered

through the use of the program, and further research to answer some of the more difficult questions looks promising.

F. CONCLUSIONS

Our experience in data validation studies has shown that survey sampling methods such as those given by Cochran [1977] can be useful in respondent and user surveys.

Sequential methods may prove to be useful for some problems, and further research is needed, particularly in the population size problem.

Sometimes cost and time constraints are such that certain samples must be excluded. The methods of Avadhani and Sukhatme [1973] enable one to exclude such samples in some instances without sacrificing accuracy of estimates of population parameters. Even when it is not possible to exclude these undesirable samples, it may still be possible to make their probability of occurrence very small. Assessment of the usefulness of this method and/or modifications to it awaits some practical applications.

G. REFERENCES

Armitage, P., 1950. "Sequential Analysis with More Than Two Alternative Hypotheses, and Its Relation to Discriminant Function Analysis," J. R. Stat. Soc. Suppl. 9: 250–63.

Avadhani, M. S., and Sukhatme, B. V., 1973. "Controlled Sampling with Equal Probabilities and without Replacement," Int. Stat. Rev. 41: 175–82.

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W., 1975. Discrete Multivariate Analysis, Theory and Practice, MIT Press.

Cochran, W. G., 1977. Sampling Techniques, 3d ed., Wiley, New York.

Darling, D. A., and Robbins, H., 1967. "Finding the Size of a Finite Population," Ann. Math. Stat. 38: 1392–97.

Efron, B., and Thisted, R., 1975. "Estimating the Number of Unseen Species (How Many Words Did Shakespeare Know?)," Department of Statistics, Stanford University, Technical Report No. 70.

Govindarajulu, Z., 1975. Sequential Statistical Procedures, Academic Press, New York.

Johnson, N. L., and Kotz, S., 1977. Urn Models and Their Application, Wiley, New York.

Lai, T. L., 1979. "Sequential Tests for Hypergeometric Distributions and Finite Populations," Ann. Stat. 7: 46–59.

Samuel, E., 1968. "Sequential Maximum Likelihood Estimation of the Size of a Population," Ann. Math. Stat. 39: 1057–68.

Sobel, M., and Wald, A., 1949. "A Sequential Decision Procedure for Choosing One of Three Hypotheses Concerning the Unknown Mean of a Normal Distribution," Ann. Math. Stat. 20: 502–22.

Sobel, M., Uppuluri, V. R. R., and Frankowski, K., 1977. "Dirichlet Distribution, Type I," Selected Tables in Mathematical Statistics Volume IV.

Walter, S. D., 1976. "A Generalization of a Matrix Occupancy Problem," Biometrics 32: 471–75.

APPENDIX

Existence results for balanced incomplete block designs

We denote the number of treatments (also referred to as varieties in the literature) with the letter v and the number of blocks with b. No variety appears more than once in a block, and no two blocks will be identical. We assume $v > 1$ and $b > 1$. Every block must contain the same number, say k, of varieties, and each variety must appear in the same number, say r, of blocks. We assume, moreover, that any pair of varieties appears together in the same number, say λ, of blocks. This then defines a balanced incomplete block design.

It follows that (1) $bk = vr$ and that (2) $r(k-1) = \lambda(v-1)$.

When $b = v$ and hence $r = k$, the balanced incomplete design is called symmetric.

A balanced incomplete block design is called resolvable if its blocks can be grouped into sets of equal sizes, so that the blocks of each set contain between them all the varieties once. If in addition any two blocks from different sets have the same number of varieties in common, then the design is called affine resolvable.

A natural question to ask is for what values of v, b, r, k, and λ do balanced incomplete block designs exist and how these designs can be constructed.

There are several necessary conditions known for the parameters v, b, r, k, and λ. From (1) and (2) we also get that $b\binom{k}{2} = \binom{v}{2}$. It was first proved by Fisher [1940] that $b \geq v$. Later, Bose [1949] gave a simpler proof which used the concept of an incidence matrix.

The theory of projective and euclidean geometries can be used to construct various types of balanced incomplete block designs, and results are given in Mann [1949] and Vajda [1967].

For resolvable designs $b \geq v + r - 1$, is a stronger inequality than $b \geq v$. A proof is given in Vajda [1967].

A balanced incomplete blocked design is said to be of the linked block type if two of the blocks have the same number of varieties in common. Vajda [1967, p. 12] proves that symmetry is equivalent to being of the linked block type.

Equations (1) and (2) imply that $r = \lambda(v - 1)/(k - 1)$ and $b = \lambda v(v - 1)/k(k - 1)$, both must be integers. These necessary conditions are also sufficient when $\lambda = 1$ and $k = 2$, $\lambda = 1$ and $k = 3$, $k = 3$ or 4 with any value of λ, and also for $k = 5$ when $\lambda = 1$, 4, or 20; see Hanani [1961] for proofs. When $k = 6$ and $\lambda = 1$ the conditions are not sufficient. This result was given by Tarry [1900].

We shall now state three existence theorems and give some consequences.

Theorem 1. If $v = b$ is an even integer, then $r - \lambda = k - \lambda$ must be a square. As a consequence the designs $v = b = 22$, $k = r = 7$, $\lambda = 2$; and $v = b = 46$, $k = r = 10$, $\lambda = 2$ are not possible.

Theorem 2. If v, k, and λ are parameters of a symmetric balanced incomplete block design and v is odd, then $x^2 = (k - \lambda)\, y^2 + (-1)^{(v-1)/2} \lambda z^2$ has an integer solution, where x, y, and z are not all zero. The proof is given in Vajda [1967]. The theorem shows that the existence of certain symmetric designs is dependent on the existence of nontrivial solutions to certain Diophantine equations.

Theorem 3. If in an affine resolvable design with parameters v, b, r, k, and λ the number v of varieties is odd, then if r is odd, k must be a square; and if r is even, v must be a square. This result is proved in Vajda [1967].

A theorem due to Chowla and Ryser [1950] is also given in Vajda [1967], and it can be used to show that a design with $v = b = 29$, $r = k = 8$, and $\lambda = 2$ cannot exist.

A complete characterization of balanced incomplete block designs is not yet known. In any event, existing results indicate that such a characterization would be complicated. Many designs have been constructed, and some results can be used to show that designs of certain specified orders cannot exist.

The literature on balanced incomplete block designs is vast and scattered. This summary of results is intended to give the reader some idea of the types of results available and is not intended to enumerate all known results. If the reader is interested in details, the monograph of Vajda [1967] is a good starting point. Cochran and Cox [1957] list several types of designs for practical use.

References for the Appendix

Bose, R. C., 1949. "A Note of Fisher's Inequality for Balanced Incomplete Block Designs," Ann. Math. Stat. 20: 619–20.

Chowla, S., and Ryser, H. J., 1950. "Combinatorial Problems," Can. J. Math. 2: 92–99.

Cochran, W. F., and Cox, G. M., 1957. Experimental Designs, 2nd ed., Wiley, New York.

Fisher, R. A., 1940. "An Examination of the Different Possible Solutions of a Problem in Incomplete Blocks," Ann. Eugen., 10: 52–75.

Hanani, H., 1961. "The Existence and Construction of Balanced Incomplete Block Designs," Ann. Math. Stat. 32: 361–86.

Mann, H. B., 1949. Analysis and Design of Experiments, Dover.

Tarry, G. "Le Probleme des 36 Officers," C. R. Acad. Franc. pour l'Avancement de Science Naturel 1 (1900): 122–23; 2 (1901): 170–203.

Vajda, S., 1967. The Mathematics of Experimental Design, Hafner.

THIS PAGE

WAS INTENTIONALLY

LEFT BLANK

ORNL/TM-7084

INTERNAL DISTRIBUTION

1. J. T. Arehart
2. S. I. Auerbach
3. B. H. Bronfman
4. R. L. Burgess
5. R. S. Carlsmith
6-20. M. R. Chernick
21. R. M. Davis
22. W. Fulkerson
23. S. V. Kaye
24. W. E. Lever
25. G. E. Liepins
26-35. A. S. Loebl
36. D. C. Parzyck
37. J. W. Sims
38. V. R. R. Uppuluri
39. G. W. Westley
40. T. J. Wilbanks
41. T. Wright
42-43. Central Research Library
44. Document Reference Section
45-46. Laboratory Records
47. Laboratory Records (RC)
48. ORNL Patent Office

EXTERNAL DISTRIBUTION

49. Assistant Manager, Office of Energy Research and Development, DOE-ORO
50. Richard Beckman, Los Alamos Scientific Laboratory, MS-600, P. O. Box 1663, Los Alamos, NM 87545
51. Martha Bentley, Librarian, Florida State Energy Office, 301 Bryant Building, Tallahassee, FL 32301
52. Brian Berry, Gund Hall, Room 333, Harvard University, Cambridge, MA 01720
53. Larry Bray, Tennessee Valley Authority, 270LB, Knoxville, TN 37902
54. Paul Cho, Regional Assessment Division, E-201, Department of Energy, Washington, D.C. 20585
55. Craig H. Cranston, Office of Energy Information Validation, Department of Energy, 12th and Pennsylvania Avenue, N.W., Washington, D.C. 20461
56. Gary de Mik, MERT Division, Oak Ridge Associated Universities, P. O. Box 117, Oak Ridge, TN 37830
57. Dr. Asil Gezen, Transportation Research and Economics Association, Suite 888, 1901 N. Fort Myer Drive, Arlington, VA 22209
58. Dr. Thomas B. Griswold, Department of Energy, Capital Plaza Tower, Frankfort, KY 40601
59. Mildred Kikta, Energy Information Administration, 12th and Pennsylvania Avenue, N.W., Washington, D.C. 20461
60. Claire E. Leslie, Science Applications, Inc., P. O. Box 843, Oak Ridge, TN 37830
61. Gus Whiton, Evaluation Research Corporation, 8330 Old Courthouse Road, Vienna, VA 22180

62. Nancy C. Lewis, Arkansas Energy Conservation and Policy Office, 960 Plaza West Bldg., Little Rock, AR 72205
63. Oficina de Energia, Apartado 41089, Estacion Minillas, Santurce, Puerto Rico 00090
64. Barbara Otto, Regional Energy Information Center, P. O. Box 35228, Dallas, TX 75235
65. Walter Page, Department of Economics, West Virginia University, Morgantown, WV 26505
66. Louis Gordon, Energy Information Administration, Department of Energy, 12th and Pennsylvania Avenue, Washington, D.C. 20461
67. Lincoln Moses, Energy Information Administration, Department of Energy, 12th and Pennsylvania Avenue, Washington, D.C. 20461
68-72. Brian Poole, Energy Information Administration, Department of Energy, 12th and Pennsylvania Avenue, Washington, D.C. 20461
73. Steve Reynolds, Energy Information Administration, Department of Energy, 12th and Pennsylvania Avenue, Washington, D.C. 20461
74. Prof. David J. Rose, Dept. of Nuclear Engineering, Room 24-210, Massachusetts Insititute of Technology, Cambridge, MA 02139
75. Dr. Kerry Smith, Resources for the Future, 1755 Massachusetts Avenue, N.W., Washington, D.C. 20036
76. David Sircle, 226 Capitol Blvd. Building, Suite 707, Nashville, TN 37219
77. Richard Stevens, Texas Energy Advisory Council, 7703 North Lamar, Austin, TX 78752
78. Lester Taylor, Department of Economics, University of Arizona, Tucson, AZ 85721
79. Ronald Wyzga, Electric Power Research Institute, 3412 Hillview Avenue, P.O. Box 10412, Palo Alto, CA 94303
80-106. Technical Information Center, P. O. Box 62, Oak Ridge, TN 37830