AD 677591

MEMORANDUM
RM-5779-PR
OCTOBER 1968

SAMPLING METHOD: SUGGESTIONS FOR MILITARY COST ANALYSTS

G. C. Sumner

D D C
RECEIVED
NOV 14 1968
B

PREPARED FOR:

UNITED STATES AIR FORCE PROJECT RAND

The RAND Corporation
SANTA MONICA • CALIFORNIA

92

MEMORANDUM
RM-5779-PR
OCTOBER 1968

SAMPLING METHOD: SUGGESTIONS FOR MILITARY COST ANALYSTS

G. C. Sumner

This research is supported by the United States Air Force under Project RAND—Contract No. F44620-67-C-0015—monitored by the Directorate of Operational Requirements and Development Plans, Deputy Chief of Staff, Research and Development, Hq USAF. Views or conclusions contained in this study should not be interpreted as representing the official opinion or policy of the United States Air Force.

DISTRIBUTION STATEMENT
This document has been approved for public release and sale; its distribution is unlimited.

The RAND Corporation
1700 MAIN ST. • SANTA MONICA • CALIFORNIA • 90406

This Rand Memorandum is presented as a competent treatment of the subject, worthy of publication. The Rand Corporation vouches for the quality of the research, without necessarily endorsing the opinions and conclusions of the authors.

Published by The RAND Corporation

PREFACE

This Memorandum is intended primarily to help fill a need in the array of statistical tools now in common use throughout the Air Force cost analysis community. For the past five years, for example, the growth in the use of regression analysis has been very rapid. More importantly, the sophistication and understanding with which the statistical mechanics are being applied is also growing. There has, however, been very little use of probability sampling in Air Force cost analysis, although many have recognized the possible utility of sampling applications.

This Memorandum was prepared at the request of cost analysts at both Headquarters United States Air Force and major air command levels. For the most part, it represents the distillation of material from available sources (see Bibliography). The intent is to provide an introduction to sampling methods, using the most applicable features of several recognized sampling techniques. It is hoped that this document will provide some basic understanding and encouragement, leading to more widespread application in military cost analysis.

SUMMARY

In this Memorandum, various aspects of probability sampling are discussed with a view toward supplementing the tool-kit of the military cost analyst. Beginning with a discussion on the relative merits of the sample as a means of data collection, the paper moves to a simplified treatment of sampling theory, and of the more basic techniques of sample design and estimation. Attention is given throughout to the use of cost-effectiveness criteria in choosing among alternative sampling plans. Sufficient coverage is provided to guide simple survey investigations, and a bibliography is provided for further reference. The exposition assumes at least a limited familiarity with statistical theory (as, for example, might be provided in the Air Force Institute of Technology training programs); accordingly, many concepts and definitions are given only salutory treatment.

Cost analysts rely heavily on data that are often imperfectly defined. Existing data sources are often fraught with errors of observation, reporting errors, and errors of classification. Sampling method offers an approach to the data quality problem that is usually cheaper, faster, and more flexible than attempts to modify existing massive data collection systems. A sample is, of course, also subject to error in that it only represents some fraction of the total; the difference is that, with proper procedures, the magnitude of this kind of error (i.e., sampling error) can be objectively estimated from the sample itself.

The basic motive underlying the design of a sampling scheme is to minimize sampling error for a given cost, or alternatively, to minimize costs for a given allowable sampling error. In either case, the solution to the design problem depends on the particular behavior under study and the amount of prior information available. Some basic "tools" that the analyst has at his disposal are stratification, clustering, subsampling, systematic sampling, ratio and regression estimators, and sampling with unequal probabilities. Although some design problems may find optimum solutions in rather complicated combinations of these tools, such so-called "complex" samples usually sacrifice the virtue

of objectively estimable sampling error (at least with the current state-of-the-art of sampling theory). To some extent, this inadequacy also exists in using sampled data for regression analyses, although it appears that in this case the problem can be merely circumvented.

ACKNOWLEDGMENTS

Acknowledgment is due Brent Bradley for his considerable advice on the accommodation of sampling method into the cost analysis context.

CONTENTS

PREFACE iii

SUMMARY iv

ACKNOWLEDGMENTS v

Section

I. INTRODUCTION 1
  Motivations for Sampling 2

II. SOME SAMPLING CONCEPTS 8
  Toward Representative Samples 8
  Statistical Basis for Inference 13
  Simple Random Sampling 19

III. ELEMENTS OF SAMPLE DESIGN 29
  Sample Precision and Cost 29
  Basic Techniques of Sample Design 33

IV. REFINEMENTS IN THE ESTIMATOR 45
  Ratio Estimator 45
  Regression Estimator 49
  Unequal Probability Sampling 52
  Comparison of Designs 56

V. SOME REAL LIFE COMPLEXITIES 58
  Complex Designs 58
  Estimating Relationships and Sample Design 65

VI. SURVEY PROCEDURE 70
  Formulate the Problem 70
  Define the Population 70
  Specify Precision 71
  Construct a Frame 71
  Select a Sampling Plan 72
  Conduct Field Work 72
  Summary, Analysis, and Documentation 72

APPENDIX: ESTIMATORS 74

BIBLIOGRAPHY 81

I. INTRODUCTION

In 1589, Galileo Galilei tossed a couple of weights from the top of the tower at Pisa and made the remarkable observation that they both landed at the same time. He did not find it necessary to drag every movable object in Pisa to the top of the tower for similar disposition; inductive logic led him to conclude that all objects, regardless of mass, are equally accelerated by the earth's gravity.

Ten years ago, a sampling expert carefully selected a hundred oranges on a hundred different trees and successfully estimated the juice content of the entire Florida orange crop within 2½ percent. The usual method of estimation, a canvass of growers' expectations, was typically off 7½ percent.

In 1936, Literary Digest's pre-election poll predicted an easy victory for Alfred Landon over Franklin Roosevelt. Roosevelt won by a landslide, carrying 46 of 48 states, and Literary Digest soon faded from existence.

In reviewing the effects of airmen personnel policies, the Air Force relies on a survey of airmen attitudes, using a questionnaire sample of less than one percent of the total airmen.

* * * * * *

A sample survey is a vehicle for inductive reasoning; it provides for the transformation of observations of a part into conclusions regarding the whole, whether that whole be lead weights, oranges, voters or airmen attitudes; it can be a very powerful device for information or misinformation, depending on the sampler's adherence to good procedure. The intent of this document is to discuss aspects of good sampling procedure and how they might be applied in cost analysis.

The pages that follow provide a broad overview of sampling method, particularly as it might apply to cost analysis. A large number of topics will be touched upon, albeit briefly. The rather shallow depth will be complemented by references to the sampling literature that is listed in the bibliography. Simplified examples will be provided both

to illustrate points and to suggest applications to the various techniques discussed. The result is an abridged "primer" on sampling for an audience of analysts who might become involved in the actual design and implementation of a sample. There is no pretense of providing a short course in sample theory, but to draw together in summary different aspects pertinent to sampling cost data. The reader will be provided armament to tackle only the simplest sampling surveys, but perhaps he will be encouraged to peruse sampling literature of greater depth or to communicate his needs to a more experienced sampling consultant.

Since the military cost analyst is typically concerned with support of planning or programming activities, his interest in data collected is usually for input into some forecasting relationship. Nevertheless, there is often significant interest in simply assessing the state-of-the-world through the estimation of averages, totals, and ratios. Except for a small section dealing with the use of sample data in regression analysis, the emphasis of this paper is on obtaining data and estimates that reflect current fact. This orientation should not be thought of as ignoring the forecasting problem facing the analyst, but as an attempt to limit the scope to the problems of data collection, which are the same for forecasting as for estimating current totals and averages.

MOTIVATIONS FOR SAMPLING

Cost analysis is highly dependent upon large amounts of data which, ideally, are reliable, accurate, and precisely defined. The required data are both historical and current, financial and non-financial, and are often imperfectly provided by existing reporting systems. It is, of course, seldom economically and administratively expedient to suggest that the cost analyst go outside the existing reporting system for large-scale data collection. The basic premise of this document is that there are occasions where small-scale methods could alleviate the data quality problem.

Consider one important element of data quality, its reflection of the precise characteristics to be analyzed. In a general sense, the cost analyst is "end-product" oriented; the typical frame of reference is the weapon or support system and the activities and costs relating

to it. Many reporting systems, on the other hand, reflect data in organizational, functional, or commodity terms. These data, while useful for management purposes, may be of no direct use for cost analysis since they are not also coded to end item.* In the absence of more precise information, artificial analytical means (such as prorating) must be used to infer the relationships between the available data and the weapon system, program element, or other focus of interest. Sampling may provide a direct means of obtaining the relationship. A sample survey can often be used for direct observation of work in process at a limited number of sites where direct identification to end product is possible. Similarly, a sample survey may be designed which calls for personnel within an organization to keep supplemental records for a short period of time. Other samples may make use of data available at the transaction level but which are summarized out of existence in the preparation of upward moving reports. Whatever the exact content of the sample survey, the intent would always be to collect a relatively small amount of data (by weapon system or program element, etc.) from which reliable and consistent inferences about total behavior can be made.

Sample or Census

One of the obvious solutions to the cost analyst's difficulty in obtaining the required end product data is the preparation of a new report which would provide a continuing census of the data. The following five considerations are basic to the choice between sampling and complete enumeration.

(1) Flexibility. A sample survey is not permanent and may be easily modified to reflect interest in different characteristics should conditions change. By contrast, a formal

*The Resources Management Systems (RMS) concept would in part reduce the frequently large informational disparity between weapon/support system or program element and functional, commodity, and organizational management. It will be some time, however, before RMS will have an appreciable effect on the data available to the cost analyst, particularly in the operating area.

reporting system is often difficult to modify and often continues to exist after the need for the data has been obviated.

(2) Cost and Available Resources. Depending on the nature of the information source, it is usually cheaper to secure data from a fraction of the aggregate, allowing a relatively larger allocation of resources to the interpretation of results.

(3) Speed. Similarly, data often can be collected and summarized more quickly with a sample than with a complete count.

(4) Scope. Sampling may be preferable when the purpose is to study broad, aggregate characteristics. However, if accurate information is wanted for many subcategories, a complete census may be more appropriate.

(5) Accuracy. Strangely enough, a sample may actually produce more accurate results than a census. Inaccuracy in a census may stem from carelessness in handling the voluminous data, poorly trained assistants, or the necessity to use data collected by other people for other purposes. Although a sample deals with only a portion of the total, the data may be much more credible.

Flexibility and speed are important advantages when considering the application of sampling for cost analysis data. Often, the data required in support of a planning or programming study are transitory. If cost analysis is to play a role in the study, usually there will be a premium on the timeliness of the data. Hence, a new data collection and reporting system is likely to be of little use.

One useful by-product of sampling is that it helps formalize the analysis procedure. It stimulates a rational, organized process of inquiry by forcing the analyst to ask questions about objectives, scope, relevant data, and desired precision.

Sampling Computerized Data

When the existing data reporting system is computerized, the foregoing factors might seem irrelevant; with all the data on tapes

or cards, it would probably be easier to program a routine to summarize all the information than to draw off a representative sample. Even so, there may be circumstances that suggest sampling either within or outside of the existing system.*

There are two main reasons why sampling from existing computerized data might be desired:

(1) When individual data points are to be examined further for qualitative or non-reported characteristics, there may not be enough time to deal with an entire census.

(2) The data base may be too bulky or complex to handle in the aggregate even with the data-reduction capabilities of the computer. It may therefore be necessary to sample in order to determine the most useful breakout for the computer to follow in summarizing the data.

There are three reasons why sampling outside existing systems might be suggested:

(1) It may be desirable to generate a new data base when the existing system is fraught with inaccuracies.

(2) Sampled data may be useful in testing the credibility of the system, and in some cases may be used for data adjustment.

(3) There may be no existing system that provides the type of data needed.

Depending on the sample size, sampling outside existing systems (i.e., actual observation of the behavior under study) can require considerable time and expense. Expense is minimized by the proper choice of sample design, which in turn depends on many factors: allowable error of estimate, allowable budget, variability of the behavior under study, geographic scope of the study, etc. These topics will be considered later.

---

*IBM has developed some interesting ways to sample computerized information. See Fan, Muller, and Rezucha.

Sources of Error in Existing Data

The inaccuracies often found in existing reporting systems have already been briefly mentioned. It should be useful now to consider the sources of inaccuracy common in mass data collection systems; these should be considered in planning a survey. The following discussion is perhaps more speculative than objective since there is actually no available documentation of attempts to measure the extent to which reported data (cost or activity oriented) differ from fact. It is often acknowledged, however, by those "within the trade" that inadequacies do exist. So what follows is a categorization of reasons why such problems occur, with no attempt to assess the importance of any particular source.

Errors of Observation. These are errors of measurement (misread gauges, faulty calculations, etc.). They arise from improper training of the data-gatherer or inadequate instruments of measurement. Compared to other errors, they are probably not too important in cost analysis.

Reporting Errors. These are errors of omission, commission, and willful adjustment of observed information. They may arise through misinterpretation of reporting goals, or the desire to make things look different than they really are. Such manipulation is provoked, for example, by the use of performance goals and activity levels as criteria for promotions or manpower allocation. On the other hand, reporting errors may be motivated by the simple wish to avoid paperwork.

Errors of Classification and Aggregation. A classification error occurs when some resource is attributed to the wrong task, or category. Recent studies of replenishment spares consumption have shown, for example, that numerous items are misclassified by maintenance shop personnel because of carelessness or failure to use up-to-date technical manuals.

Aggregation error results when expended resources are totaled and reported at periodic intervals, rather than being attributed to the time periods in which they were consumed. Aggregation error of

another sort occurs when data for several categories are lumped together. Such data may be very appropriate for management purposes, but the cost analyst must often arbitrarily prorate the information among the categories of interest in order to accomplish his own ends.

Specious Accuracy

Data may be accurate in the sense that there have been no errors from initial observations to final reporting, yet they may not really represent the particular behavior that one supposes. Such misleading accuracy is said to be specious.

For example, Operation and Maintenance resources expended for base support on Air Training Command (ATC) bases are normally identified as "Training Support." Although such accounting may precisely reflect support costs on those bases, the "training support" label clouds the fact that the cost of support rendered to other major command tenants is also included; these costs may be largely independent of the training function.

As another example, consider maintenance data obtained from an independent sample that is designed to circumvent the problem of inflationary (or deflationary) reporting. Such data may better measure the actual maintenance needs of various equipment than does the established reporting system. However, it might be a mistake to base an estimating relationship (ER) on these data; the ER may estimate maintenance needs, but may not reflect maintenance practice.

II. SOME SAMPLING CONCEPTS

TOWARD REPRESENTATIVE SAMPLES

Section I tacitly recommends a basic distrust of data recorded by anyone but the cost analyst who will use those data. The suggestion has been for the analyst to determine at what point objectionable error occurs in the data handling process and to go to that point and make his own observations (or engage a well trained staff of observers). When data are voluminous, customized collection implies the use of sampling method. The task remains to discuss how to insure that a sample is representative of the total, for this is the necessary assumption if decisions are to be based on sample information.

For it to be representative, one might specify that the sample reflect, in proper proportion, the various attributes of the population under study. The sample need not be an exact miniature of the population to be useful; the allowable latitude in this respect depends on how sensitive the analyst's purposes are to errors in estimates. A discussion of sample representation involves terms such as population, distribution, bias, and error. These notions will be described, since their meanings as used in sampling may differ from common use.

Populations and Their Distributions

Sampling is motivated by the desire to evaluate some characteristic of interest in order to aid subsequent decisionmaking. In statistical terminology, the population is the complete set of values of that characteristic. Specification of the population requires definition in terms of:

(1) Content. What characteristic of the population is under evaluation?

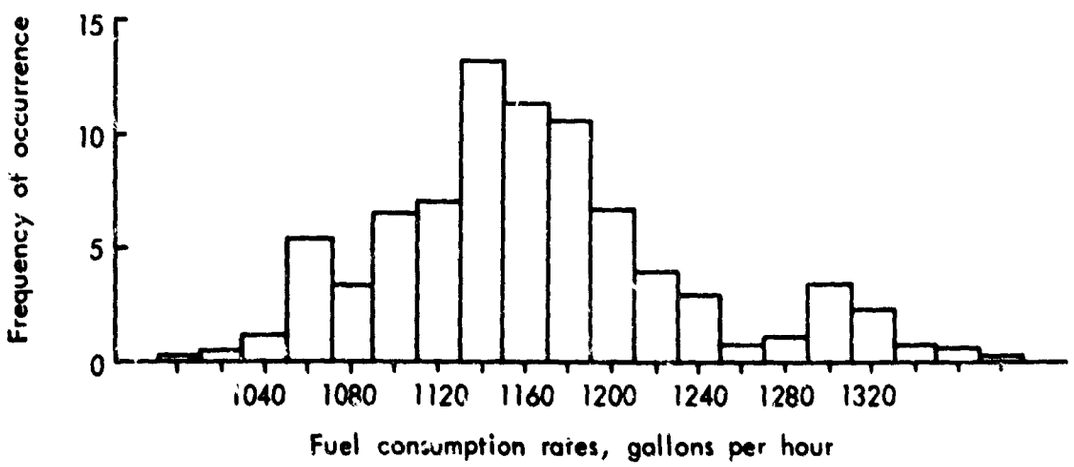
(2) Units. What are the units into which the population can be divided?

(3) Extent. What are the boundaries of the population?

(4) Time. What is the time interval during which information is relevant, and what is the time interval for which an inference is to be drawn?

If the problem is to estimate average fuel consumption of a particular model aircraft in Fiscal Year 1967, the population is the collection of fuel consumption rates for each such aircraft that was operational during FY 1967. If, on the other hand, the problem is to estimate (i.e., forecast) average fuel consumption of that aircraft during FY 1968 - FY 1973, the population is the collection of fuel consumption rates for each such aircraft operational during those five years. Lacking clairvoyance, the procedure in the latter case would be to substitute a related population, e.g., all relevant experience in the past year, and assume that the substitute population reflects the target population closely enough for practical purposes.

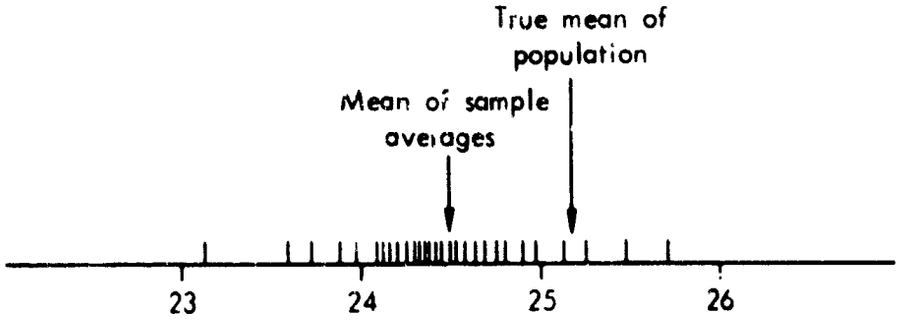
Populations can be characterized by their distributions. Suppose it is possible to categorize each unit in a population according to its value, and then prepare a graph of the frequencies with which each category is represented. The result is a frequency diagram of the distribution, which represents a visual illustration of the population. For example, if the fuel consumption rates for all operational aircraft are allocated into 20 gallon/hour categories, the result might be graphed as follows:



The choice of category width is arbitrary, and is rather a matter of visual taste; if the categories are made very small (e.g., 1 gallon/hour intervals), the graph begins to assume the appearance of a smooth curve. For simplicity, all subsequent frequency diagrams in this document will be pictured as smooth curves.

Error and Bias

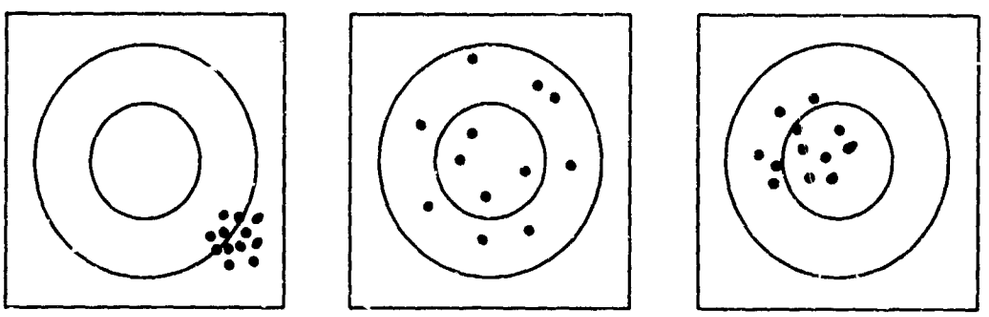
A sampling procedure is usually judged by the accuracy with which it reflects the population, or with which it provides estimates of population characteristics (such as the population average). This accuracy is composed of two factors, sampling error and bias. Consider a sampling procedure in which ten observations are taken from a population of fifty, and their average value recorded. Suppose that this procedure were repeated an infinite number of times.* There would result quite a number of sample average values, but they would tend to concentrate within some sharply defined region. This dispersion of sample averages is called sampling error. Now, it is conceivable that these sample averages might, in turn, be averaged to produce a "grand sample average," and that the latter may not coincide with the characteristic being estimated (i.e., the average of the population taken as a whole). The difference between the population average and the average of sample averages is due to bias in the sampling procedure. Suppose this example produced results graphed below.



*That is, a sample of ten is selected, recorded, and replaced, then another sample of ten is selected, recorded, and replaced, etc.

The population average is 25.2 and the average of sample averages is 24.5. Sampling error ranges within about 1.0 units, and the bias associated with this sampling procedure is equal to .7 (i.e., 25.2 - 24.5 =.7). The combined effect of sampling error and bias may or may not preclude the usefulness of the sampling procedure, depending on the accuracy required by the problem.

A helpful analogy is to consider the markmanship of three riflemen, where the riflemen are attempting to "estimate" the center of the bullseye:*



A B C

The target on the left was turned in by marksman A. He has a very steady arm, but apparently suffers from astigmatism; although his aim is precise (i.e., small sampling error), he consistently misses his mark. Marksman B has no bias in his score, but his precision is quite a bit less than A. Marksman C displays a small bias and more precision than A. Since marksman C's particular mix of precision and bias tends to consistently put him nearer to the center of the target, we would probably consider him the most accurate of the three.

In light of the previous discussion on errors in cost data, we might say that marksman A could represent estimates resulting from the use of data coming out of any existing reporting system; the estimates are consistent, but their bias tends to invalidate their usefulness. Marksmen B and C might represent two alternative sampling schemes. Scheme B is virtually free of bias but is burdened by large sampling error. Scheme C displays some bias but has the saving grace of small

*Illustration adapted from Jessen.

sampling error. Scheme C would probably be the preferred (provided the magnitude of the bias could be assessed).

Probability Samples

There are two broad approaches to a representative sample: (1) judgment sampling and (2) probability sampling. In judgment sampling, the analyst relies on his experience and skill to select a number of sample points that are "typical" of the total population under study. The judgment sample is characterized by the following comments:

(1) Accuracy may vary from sampler to sampler, but (for a given sampler) is fairly uniform as sample size is varied.
(2) There is generally some bias present.
(3) There is no objective measure of the combined effects of sampling error or bias.

Probability samples are drawn with the aid of a table of random numbers or any other device that assures that each sample point selection is independent of all others. The general characteristics of probability sampling are:

(1) Accuracy is not dependent on who is doing the sampling, but it is dependent on sample size.
(2) There is no sampling bias.
(3) Sampling error can be estimated objectively.

Sampling error can usually be estimated from a single sample, but very rarely is it possible to estimate bias. A highly experienced sampler who is intimately familiar with the subject under analysis may be able to satisfactorily convince himself that the bias in his procedure is "reasonable." But the researcher's audience is typically a skeptical one and is inclined to have less faith in his judgment. The presence of bias muddies up any objective statement of accuracy. For this reason, it is usually easier to accept a lot of sampling error rather than a little bias. This also motivates this paper's near total emphasis on probability sampling.

STATISTICAL BASIS FOR INFERENCE

The next several pages review some basic statistical principles as they relate to sampling and develop the line of reasoning that supports the sampling method as a basis for decisionmaking. For those who are already convinced of the credibility of probability sampling, this discussion will hold little interest.

In the remainder of this Memorandum, population parameters are denoted by Greek letters and sample statistics* are denoted by Roman letters:

	Population	Sample
mean	μ (mu)	$\bar{X}$
variance	σ^2 (sigma)^2	s^2

The size of the population is represented by N, and sample size is n. An unbiased estimate of a parameter is indicated by placing a "hat," ^, over the parameter symbol. Thus, to say that a sample mean is an unbiased estimate of the population mean is equivalent to the expression:

$$\hat{\mu} = \bar{X}$$

Descriptors of Populations and Samples

Recall the earlier discussion of population distributions. There are generally two characteristics of any population distribution that interest the analyst: central tendency and dispersion.

Two measures of central tendency are the median and the mean. If all units (noted as X_i) in the population are arrayed in order of size, the median is the value of the middle unit. The mean is the average of the population units:

$$\mu = \frac{\sum^{N} X_i}{N}$$

*Parameters are constants associated with the population; statistics are numbers calculated from the sample, and therefore are variable from sample to sample.

Of the two parameters, the mean is most often of interest, especially with near-symmetric distributions. The median, on the other hand, is independent of the distribution, and is often, therefore, the preferred parameter in situations where the shape of the distribution is irregular.

The most common measure of dispersion is the variance, defined as the average of squared deviations of units from the mean:

$$\sigma^2 = \frac{\sum^{N} (X_i - \mu)^2}{N}$$

The standard error (or standard deviation) is the square root of the variance:

$$\sigma = \sqrt{\sigma^2}$$

An alternative measure of dispersion is the mean deviation:

$$\frac{\sum^{N} (|X_i - \mu|)}{N}$$

The mean deviation is seldom used; the standard deviation is more popular because of its relationship to confidence intervals, to be discussed later.

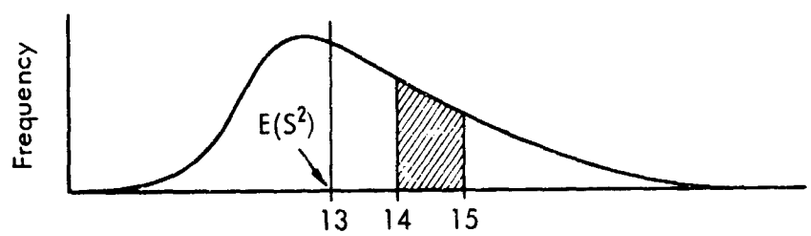
A sample is some portion of the population composed of n units. Analogous to the two population parameters, μ and σ^2, are the sample mean and the sample variance:

$$\bar{X} = \frac{\sum^{n} X_i}{n} \quad ; \quad S^2 = \frac{\sum^{n} (X_i - \bar{X})^2}{n}$$

These are called <u>statistics</u> because they are variables dependent on the particular assortment of n units chosen for the sample.

Sampling Distributions

Being a variable, a statistic also has a distribution. This distribution is called the sampling distribution, since it reflects the frequencies with which the statistic would take on different values if the sampling procedure were repeated an infinite number of times. The expected value of a statistic is defined as the mean of its sampling distribution. For example, the following might be the sampling distribution of S^2* [the expected value of S^2 is denoted as $E(S^2)$]:



Since the purpose of sampling is to obtain information about the population, we are generally concerned that our sample statistics are accurate estimates of the corresponding population parameters. The two aspects of accuracy, precision and bias, can now be characterized in terms of the sampling distribution.

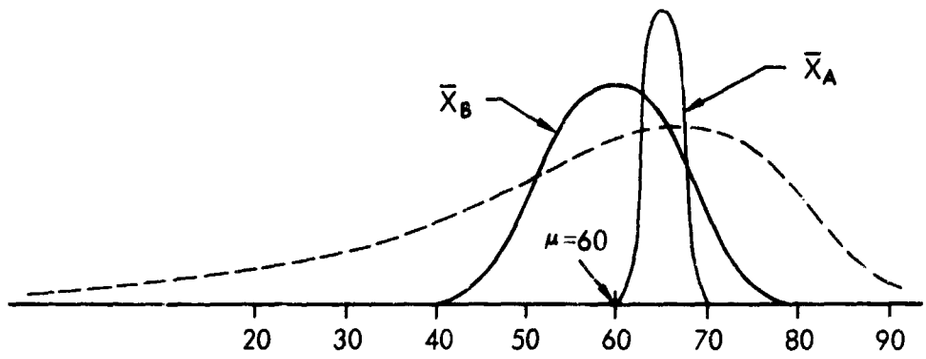
An estimator (i.e., the formulas actually used in deriving estimates) is unbiased if the expected value of the statistic is equal to the parameter it estimates.

An estimator is precise if it has a relatively narrow sampling distribution (i.e., if the sampling error is small).

The diagram below represents the sampling distribution of $\bar{X}$ for two different sampling procedures superimposed upon the distribution of the parent population (dashed lines):

---

*Since the distribution is conceptually derived from an infinite number of iterations, its graph is drawn in terms of relative frequency. For geometric interpretation, the probability (P) that S^2 will assume a value within some interval is equal to the percentage area under the curve that is bounded by that interval (e.g., $P[14.0 < S^2 < 15.0] = .10$).



Estimator B is unbiased (its expected value is equal to μ) but not very precise. Estimator A is more precise but is biased:

$$\text{Bias (A)} = 65 - 60 = 5$$

From the diagram, it appears that a precise, slightly biased estimator might be preferable to an unbiased, less precise estimator. This tradeoff is difficult to evaluate since μ is unknown. The usual practice is to follow procedures that are known to produce unbiased estimators, then select the estimator that has the greatest precision. Most of the unbiased procedures require probability sampling. The requisites for probability sampling are:

(1) Every unit of the population has a known probability of being included in the sample.

(2) The sample is drawn by some method of random selection (each selection is independently determined).

(3) Probabilities of selection are taken into account when making estimates from the sample.

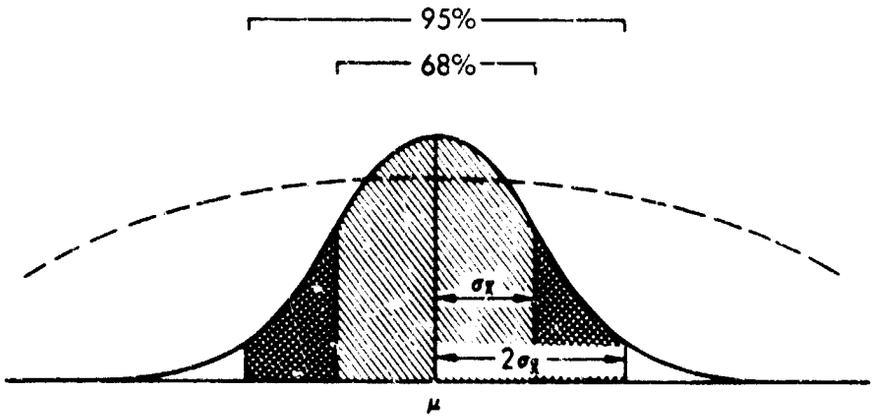
Probability sampling methods provide unbiased estimates of population parameters, or contain certain bias that can be evaluated. For example, $\bar{X}$ is always an unbiased estimate of μ if probability sampling has been employed; the sample variance, S^2, is a biased estimator of the population variance, σ^2, but the bias is corrected by a simple adjustment factor, $\frac{n}{n-1}$. Non-probability methods, such as judgment

sampling, may provide more precise estimates, but it is usually impossible to identify bias. Probability sampling also furnishes information on the sampling distributions of the estimators, and thus provides the bridge necessary to be able to draw inferences about the population, based on the sample.

Confidence Intervals

So far it has been shown that samples can provide useful estimates of population parameters; it is known that if the sampling procedure is repeated an infinite number of times, the average values of $\bar{X}$ and $(\frac{n}{n-1})S^2$ will be μ and σ^2. A question remains, however, about the inferences drawn from a single sample; how close is $\bar{X}$ to μ? This cannot be determined with certainty, but thanks to a very helpful characteristic of nature which is expressed as the central limit theorem, it is possible to specify the shape of the sampling distribution of $\bar{X}$ and thereby find the probability that the quantity $|\bar{X} - \mu|$ is within some specified tolerance level. The central limit theorem provides that, as sample size increases, sample means tend to be distributed normally regardless of how the parent population is distributed. The distribution of sample means has the same mean as the parent population, but its standard error is equal to $\frac{\sigma}{\sqrt{n}} = \sigma_{\bar{x}}$.

Pictured below are the population distribution (dashed line) and the sampling distribution of $\bar{X}$ from samples of size n=10:



Since the distribution of $\bar{X}$ is approximately normal, theory tells us that about 68 percent of the area beneath the curve lies within one standard error ($\sigma_{\bar{x}}$) of μ, and 95 percent of the area lies within two standard errors. Another interpretation is that the probability that $\bar{X}$ will fall within one standard error μ is .68. Standardized normal curve tables are available in most texts that provide this information for fractional multiples of $\sigma_{\bar{x}}$.

Suppose that a sample is drawn from the above population and the two statistics computed:

$$\bar{X} = 43, \qquad S^2 = 25.$$

Neither μ nor $\sigma_{\bar{x}}$ is known, but $\sigma^2_{\bar{x}}$ can be estimated:

$$(\hat{\sigma}^2_{\bar{x}}) = \frac{\hat{\sigma}^2}{n} = \left(\frac{n}{n-1}\right)\frac{S^2}{n} = \frac{S^2}{n-1} = \frac{25}{9}$$

A slightly biased estimate of $\sigma_{\bar{x}}$ is found by finding the square root of $(\hat{\sigma}^2_{\bar{x}})$. From the previous discussion it is known that if the sample were drawn repeatedly, 95 percent of the sample means would fall within $2\sigma_{\bar{x}}$ of the population mean. This statement is equivalent to saying that μ is within $2\sigma_{\bar{x}}$ of the sample mean 95 percent of the time. Thus it is said that the 95 percent confidence interval for μ is $43 \pm 10/3$. This does not mean that the probability that μ lies in this interval is .95; however, if one were to follow this procedure for setting confidence intervals in sample after sample, he would expect his intervals to contain μ 95 percent of the time.

The only flaw in the procedure for arriving at confidence intervals has been the use of S^2 to estimate σ^2. Fortunately, this only causes problems with small samples, which are discussed on page 26.

The concept of sampling error may be a little ponderous for the decisionmaker to employ. If the decisionmaker desires a certain maximum tolerance in order to use the sample results, the sampler can estimate the odds that the tolerance will be met (absolute assurance of a

given tolerance is for most purposes impossible). Whether the odds are acceptable depends on the decisionmaker's aversion to the risk of incorrect conclusions.

As an example, suppose an on-site sample survey has been made of the number of direct depot man hours required to repair and refurbish a certain missile guidance system component. The sample yields a statement that on the average 1191 direct man hours are required for each of the components of interest. From information concerning the sample, an analyst can estimate the odds of achieving a specified tolerance. This might be stated as "the population mean is equal to 1191 direct man hours $\pm$ 12.5 hours (tolerance) at .93 confidence." If the analyst is willing to act upon estimates with this tolerance and confidence, the survey has provided useful information. A more conservative analyst might feel comfortable only when this tolerance is achieved with .99 confidence, in which case the sample procedure would need to be revamped to obtain better representation of the population.

SIMPLE RANDOM SAMPLING

The simplest probability sample is the simple random sample. The required conditions are:

(1) Independent selection of sample units.

(2) Equal probability of selection for all units in the population.

The first condition specifies that the inclusion of a particular unit in the sample is in no way dependent on the inclusion of some other unit; this is accomplished by randomizing the selection. The second condition assures that the sample will not be biased.

Both conditions are implemented by proper selection of a sampling frame. A frame is a list; a way of dividing the population into sampling units that are distinct and non-overlapping and that together constitute the whole of the population. A suitable frame allows the listing or numbering of all units in order to make a random selection (although for some sampling procedures to be discussed later the complete list is not necessary).

A table of random numbers is one way to draw the sample. Suppose, for example, that a sample of size n=10 is desired from a population of N=452. Choose some arbitrary point in a table of random numbers and read down the column of 3-digit numbers, picking out the first ten numbers that do not exceed 452. The sample consists of those sampling units that correspond to the chosen numbers (any number appearing more than once should be ignored after the first time).*

Choice of Sample Size

The choice of sample size involves a tradeoff between cost and precision; increased precision requires a larger sample size, which in turn implies higher cost. For the analyst who does not have a fixed budget, it is probably more meaningful to translate sampling cost to sampling time (assuming the preferred path to a solution is the shortest path); cost and time can be considered synonymous. The typical procedure for determining sample size is to specify some level of precision, solve for sample size required for several alternative sampling schemes, then compare costs (and possibly adjust the precision requirement if costs for all alternatives are out of line with the budget). The following steps assume imple random sampling; the rationale is the same for other sampling schemes, but the computation is more complex.

The first step is to decide how large an error can be tolerated in the estimate. This requires careful thinking about the use to be made of the estimate and about the consequences of sizable error (is

*There is nothing essential about the use of random number tables, for more simple devices such as tossed dice or numbered chips drawn from a hat will often do. Sometimes it may be assumed that the population units occur randomly in the sampling frame, so that any arbitrary selection is valid; for example, if one is sampling 40 airmen to estimate the average skill level of airmen at a particular base, the first 40 airmen listed in the base directory can probably be regarded as a random sample (since skill level is not related to surname). Care should be exercised that such devices actually do assure independent and equal probability of selection. The advantage of a random number table is that such assurances are scientifically provided.

the estimate to be very precise or just a rough estimate?). The figure arrived at may be, to some extent, arbitrary, but this is the necessary step that patterns the sample estimate to the objective of the analysis. The second step is to express the allowable error in terms of confidence limits. Suppose L is the allowable tolerance in the sample mean, and we are willing to take a 5 percent chance that the error will exceed L (we want to be "reasonably certain" that the error will not exceed L). The 95 percent confidence limits computed from a single mean are:

$$\bar{X} \pm 2\sigma_{\bar{X}} = \bar{X} \pm \frac{2\sigma}{\sqrt{n}}$$

Since the tolerance is L:

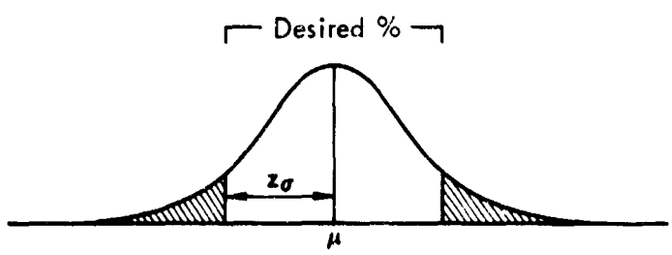
$$L = \frac{2\sigma}{\sqrt{n}}$$

$$n = \frac{4\sigma^2}{L^2}$$

The general formula is:

$$n = \frac{z^2\sigma^2}{L^2}$$

where z is the standard normal deviate, i.e., the multiple of $\sigma_{\bar{X}}$ that corresponds to the desired confidence interval.*



*The appropriate z-values can be found in tables of standard normal deviates in most statistics texts.

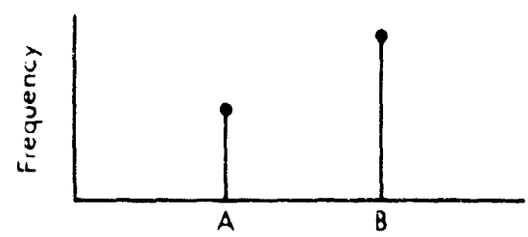
In order to use this formula, an estimate of σ is necessary. This may be accomplished by a small preliminary sample, or by examining previous samplings of similar populations. For populations of size greater than 500, a crude estimate of σ is (range)/6, the range being defined as the difference between the highest and lowest values in the population.

Having calculated the sample size required for the stated precision, the third step is to evaluate the sample cost. If the cost is high, it may be necessary to relax the precision requirement. It may even appear preferable to give up the sampling plan altogether in favor of a complete census.

Sampling for Attributes

Population characteristics can be classified as quantitative or qualitative. Quantitative characteristics (e.g., annual income) are called variates and are expressed numerically. Qualitative characteristics (e.g., sex) are called attributes and are non-numerical. Sampling of variates leads to the estimation of totals and averages; sampling of attributes leads to the estimation of proportions, or percentages. The various sampling designs generally apply in both cases, the main difference being the form of the estimators (i.e., the formulas used in deriving estimates). There has been no attempt in this survey to grant "equal time" to attributes, since the discussion and examples would simply parallel that of variates.

Consider a study to determine the proportion of overseas Air Force installations that maintain their own telephone switchboard facilities; each base selected for the sample would be classified as either (A) maintaining its own facility or (B) contracting that function out. The frequency distribution has the following form:



If A and B were each assigned a numerical value, this distribution could be handled the same as the variate case. The analyst is usually interested in determining the <u>proportion</u> of units exhibiting property A:

$$\Pi_A = \frac{N_A}{N}$$

This number is the same as μ if every sample unit exhibiting characteristic A is given a value of one (1), and all other sample units are valued at zero:

$$\Pi_A = \frac{N_A}{N} = \frac{\sum^{N} X_i}{N} = \mu_A$$

Assuming simple random sampling, the sample proportion, p, is an unbiased estimator of the population proportion, Π. The variance of the sampling distribution of p takes the following form:

$$\sigma_p^2 = \frac{\Pi(1-\Pi)}{n}$$

Its unbiased estimator is

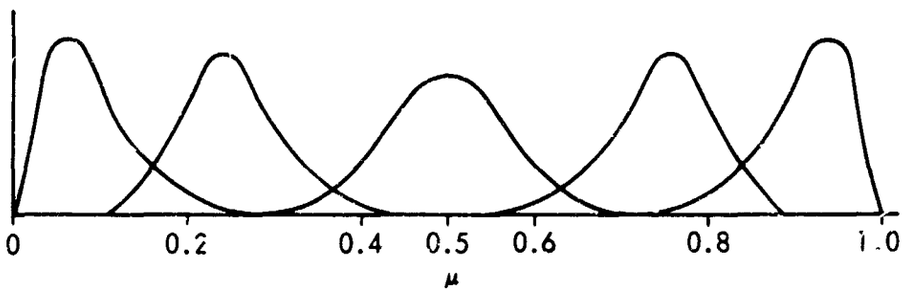
$$s_p^2 = \frac{p(1-p)}{n}$$

Sometimes the intent is to estimate N_A, the total units in the population having the desired attribute. The appropriate estimators here are:

$$\hat{N}_A = p_A N$$

$$s^2_{N_A} = N^2 \left(\frac{p(1-p)}{n}\right)$$

The sampling distribution of p (and of pN) has the desirable central limit theorem property of tending toward normal as the sample size increases. However, if either π or $(1-\pi)$ is very small, very large sample sizes may be required. This is because the sampling distribution tends to be non-symmetrical for values of π that are very high or very low:



As a rule of thumb, the following conditions should hold before relying on normal distribution properties:

$$np > 5 < nq.$$

For values of p that are very large or very small, it is much cheaper in terms of sample size to base confidence intervals on the properties of the Poisson distribution (a general class of skewed distributions) or the binomial distribution. Reference to these two distributions can be found in most basic statistics texts (e.g., Hoel).

Finite Population Correction

This paper assumes non-replacement sampling throughout. This is the general class of samples in which individual population units are not allowed to appear in the sample more than once; i.e., there is no duplication in the selecting of random numbers. When sampling with

non-replacement from finite populations, it is necessary to introduce the factor $(1 - \frac{n}{N})$ into the computation of sampling variance. Hence:

$$\sigma^2_{\bar{x}} = (1-f)\,\frac{\hat{\sigma}^2}{n}\ ;\ f = \frac{n}{N}$$

$$\hat{\sigma}_{\bar{x}} = S\sqrt{\frac{1-f}{n-1}}$$

This factor is called the finite population correction (fpc), and assures that the estimated sampling variance tends to zero as the sample size approaches the population size N. In practice, the fpc can be ignored when the sampling fraction is not greater than 5 or 10 percent. The effect of ignoring the correction is to overestimate the standard error, which generally is not as serious as underestimation.

Dissecting the Sampling Variance Estimator

It may be of interest to summarize the "anatomy" of the sampling variance. The various components are:

$\frac{\sum^{n} (X_i-\bar{X})^2}{n}$	The average of squared deviations of sample observations from the sample mean; simply a convenient descriptive measure of variability within the sample, but which is also useful because of its relationship to σ^2.
$\frac{n}{n-1}$	The factor necessary to convert the measure of sample variability into an unbiased estimate of population variability as measured by σ^2.
$\frac{1}{n}$	The factor that converts the measure of population variability into a measure of variability of the sampling distribution.
$1 - f$	The factor that makes allowance for sampling from finite populations $(f = \frac{n}{N})$.

Assembling the components gives:

$$S^2_{\bar{x}} = (1-f)\left(\frac{1}{n}\right)\left(\frac{n}{n-1}\right)\left(\frac{\sum^{n}(X_i-\bar{X})^2}{n}\right) = (1-f)\frac{S^2}{n-1}$$

which is the unbiased estimator of

$$\sigma^2_{\bar{x}} = (1-f)\frac{\sigma^2}{n}$$

Examination of this formula draws attention to the fact that the sampling error ($\sigma^2_{\bar{x}}$) depends primarily on the population variance (σ^2) and the <u>absolute</u> sample size (n). The <u>relative</u> sample size (i.e., the fraction of the population sampled) is not an important factor in large populations. For example, 50 observations from a population of 20,000 will give an estimate about as precise as 50 observations from a population of 1,000, provided that the population variances are the same.

<u>Confidence Intervals from Small Samples</u>

One problem in the determination of confidence intervals arises from the use of the following formula for determining the upper and lower limits.

$$|\bar{X} - \mu| = L = \frac{z\sigma}{\sqrt{n}}$$

The variable z is the standard normal deviate that corresponds to the degree of confidence desired. This formulation rests on the fact that

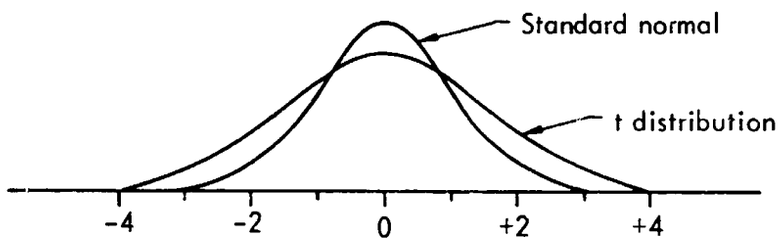
$$z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$$

has a standard normal distribution (i.e., μ=0, and σ=1). Since σ is not usually known, it is often necessary to use its sample estimator

instead. The expression

$$\frac{X-\mu}{S/\sqrt{n-1}}$$

follows what is called the t-distribution. The t-distribution is very close to normal, but has wider dispersion when the sample size is small.



For this reason it is preferable to use t-values instead of z-values when sample sizes are less than 30;* upper and lower limits are then determined from the expression:

$$|\bar{X} - \mu| = L = \frac{tS}{\sqrt{n}}$$

where t and S have been substituted for z and σ, respectively. Tables are available from which t-values can be ascertained in much the same manner as the z-values, except that the sample size must be specified. A portion of a t-table appearing in R. A. Fisher's 1934 volume of Statistical Methods for Research Workers is reproduced below. If, for example, the

degrees of freedom	level of significance				
(n - 1)	.5	.3	.1	.05	.01
13	.694	1.079	1.771	2.160	3.012
14	.692	1.076	1.761	2.145	2.977
15	.691	1.074	1.753	2.131	2.947
16	.690	1.071	1.746	2.120	2.921
17	.689	1.069	1.740	2.110	2.898

*This boundary between "small" and "large" samples is arbitrary; experience has shown that for most purposes, the z-distribution sufficiently approximates the t-distribution when sample size exceeds 25 to 30.

desired confidence level were .70 and the sample size were n=15, the appropriate t-values would be t = 1.076. The column headings refer to the area under the "tail" of the curve (e.g., a .70 confidence level implies .3 significance). The row headings refer to degrees of freedom, a rather abstruse statistical concept which for simple random sampling is one less than the sample size.

III. ELEMENTS OF SAMPLE DESIGN

Designing a sample is a matter of getting the most accuracy for your money, and is a problem apart from that of obtaining "valid" results (in the sense of being able to draw correct inferences). Validity derives from adhering to the rather well defined rules of good procedure, such as using correct estimators and maintaining independent selection of sample points. A designed sample, on the other hand, seeks to utilize prior subjective knowledge about the population in order to increase accuracy or decrease costs.

SAMPLE PRECISION AND COST

Increasing Precision

Precision is increased by decreasing the variance of the sampling distribution. There are four fundamental methods for achieving this result:

(1) Increasing sample size.
(2) Stratifying the population.
(3) Using auxiliary variables in the estimator.
(4) Using unequal probabilities of selection.

The simplest way to increase precision is to increase the sample size. This has already been discussed in connection with choosing the sample size for simple random samples.

Stratification involves dividing the population into two or more subpopulations and sampling from each. Stratification always reduces the sampling variance provided the variability within strata is less than the variability in the overall population. It is also possible to stratify after the sample has been drawn, but this is usually not as efficient.

Sometimes in sampling there is the opportunity to observe an auxiliary variable which is closely related to the main variable of interest, and which can be utilized in the estimator to give more precise estimates. Two such estimators (ratio and regression) are discussed in Section IV.

Unequal probability sampling offers another way of making good use of an auxiliary variable. As well as being used in the estimator, the auxiliary variable is used to determine the probability with which various sample points fall into the sample. Probabilities are set proportionally to the auxiliary variable, and the closer the correlation between the two variables, the more precise the final estimate.

These methods will be discussed more fully in the process of describing the several sample designs which follow in this section and in Section IV.

Cutting Survey Costs

Costs of running a survey fall naturally into four categories:

(1) costs of observation,
(2) travel costs,
(3) coding costs, and
(4) overhead costs.

Observation costs are those incurred "on-location" in recording the behavior under study. These costs vary directly with the number of sampling points, and therefore are reduced by decreasing the sample size. Any of the sample designs that offer increased precision, for a given sample size, can likewise be used to provide the same precision at less observation cost.

Travel costs are those incurred in moving between sample points and home base. These are mostly irrelevant when sampling from a centralized reporting system. The common method to reduce travel is to group the sample points into clusters so that the sampler can pick up several observations at each location rather than just one. The cluster technique is less efficient (less precise), but sometimes the reduction in travel cost may allow the sampler to recoup his precision loss by selecting more sample points. Cluster sampling will be explained later in detail.

Coding includes those administrative tasks relating to the transformation of sample information recorded by field workers into a form that is amenable to analysis. This may simply require the consolidation

of data from several worksheets, or it may involve numerical interpretation of responses recorded on sample questionnaires. The magnitude of coding costs depends on the mode of data collection; but for a given mode, they vary directly with sample size. Careful planning of sample observation procedures may lead to significant savings in data handling costs.

Overhead includes such items as frame construction, sample selection, calculating estimates. These costs are rather insensitive to different sample designs, since planning and design probably constitute the bulk of these costs. However, if the population is very large, the choice of sampling design can significantly affect the time necessary to construct a frame and select sample points.

Cost-Precision Tradeoff

It has been stated that the choice of sample design depends on both cost and precision of the alternative sampling schemes, but there has been no discussion of combining the two into a single measure. The usual measure is Net Relative Efficiency (NRE). The concept of NRE will be developed by means of a simple example.

Suppose two alternative sampling schemes, A and B, are available. For a sample of size 50, it is estimated that sampling variance for scheme A will be 35, and that for scheme B will be about 42. The Relative Efficiency (RE) of A to B is the inverse ratio of the variances:

$$RE(A/B) = \frac{var\ (B)}{var\ (A)} = \frac{42}{35} = 1.20$$

Scheme A is said to be 20 percent more efficient than scheme B. A 10 percent sample using scheme A would provide the same precision as a 12 percent sample using scheme B.

What about costs? The costs of the two schemes are estimated, variable costs (those proportional to sample size) are separated out, and the variable cost per sample point for each is computed. This variable component for A is $50, and for B is $40 (these costs might

as easily have been stated in terms of man-hours). The Relative Variable Cost (RVC) of A to B is the ratio of these costs:

$$RVC(A/B) = \frac{VC(A)}{VC(B)} = \frac{50}{40} = 1.25$$

Scheme A is 25 percent more costly than scheme B on a sample point basis.*

Since sampling variance is inversely proportional to sample size, and the RVC is on a sample point basis, a consistent way to combine the two criteria is to divide the Relative Efficiency by the Relative Variable Cost. The new measure is called Net Relative Efficiency:

$$NRE(A/B) = \frac{RE}{RVC} = \frac{1.20}{1.25} = .96 \quad \left[= \frac{var\ (B) \cdot VC(B)}{var\ (A) \cdot VC(A)} \right]$$

When costs are considered, scheme A is 4 percent less efficient than B; <u>equivalently</u>, scheme B is 4 percent more efficient than A ($\frac{1}{.96}$ = 1.04).

For a given level of precision, scheme B will be 4 percent cheaper; for a given budget, scheme B will provide 4 percent greater precision.

The choice of scheme B has depended on some necessarily rough "guesstimates." The feeling is, however, that these calculations lead to a best guess when performed by someone with good subjective familiarity with the behavior under study. The need for this kind of preliminary analysis illustrates the usefulness of prior sample surveys that are well documented. The concept of Net Relative Efficiency receives detailed treatment in Jessen, pages 97-103.

Again, sampling designs are motivated by the desire to re-establish the cost-precision tradeoff at a more favorable level than is obtained by simple random sampling. The most basic designs are described in the remainder of this section.

---

*Assume equal or insignificant fixed cost.

BASIC TECHNIQUES OF SAMPLE DESIGN

The most basic of the sampling techniques are those that "partition" the population so that the resulting sample will reflect some special knowledge of the manner in which the population units naturally occur. Four techniques are considered:

- Stratified sampling
- Cluster sampling
- Subsampling
- Systematic sampling

These are the foundations of the more complex schemes often required in real world applications of sample surveys.

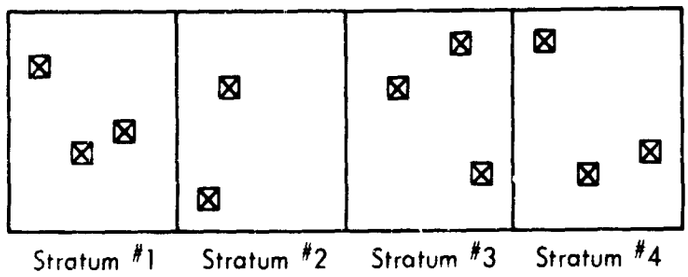
Description of the basic sample designs will include the motivation for their use: why the design is used; advantages and disadvantages; relative costs of application; and allocation of sample units. A simple illustration of each design is also included.

The formulas for estimation of population means and variances are not found in the descriptions but are given in Appendix II. This has the dual purpose of (1) smoothing the way for those who are more interested in the rationale behind different designs than the arithmetic of estimation, and (2) gathering the various formulas into a few pages for easy comparison. Most of the examples include measures of estimation and may prompt the interested reader to refer to the appendix; although the general tone of the applications and the reasoning behind the choices of designs should be apparent without having to become immersed in actual numbers, these calculations were included for those desiring to see the formulas in action.

Please note the following convention for stratification, clustering, and subsampling. The population is divided into N partitions, of which n partitions are designated for sampling; each partition consists of M data points, of which m are selected for the sample. Thus the total number of data in the population is equal to MN, and the total sample size is mn.

Stratified Sampling

In stratified random sampling, the population is divided into non-overlapping subpopulations, called strata. A simple random sample is then drawn in each stratum.



There are four principal reasons for stratifying.

First, it sometimes is desired to obtain estimates for subdivisions of the population.

Second, it may be administratively convenient to break up the population into strata of a size easier to work with.

Third, sampling problems may differ in different parts of the population. For example, in sampling long-haul communications personnel stationed on air bases, it would be practical to put SAC and ADC in a separate stratum since they administer their own communications. The data sources for these two commands, and their sampling frame, would be of a different nature than that of the other major commands, which are served by the Air Force Communication Service.

Fourth, considerable precision may be gained if it is possible to divide a heterogeneous population into strata that are internally homogenous. Differences between strata do not contribute to the stratified sampling variance. Thus, the less variability within strata, the smaller the sampling variance.

The simplest way to allocate the sample is to use proportional allocation, that is, to make the number of sample units drawn from each stratum proportional to the total number of units in that stratum. The gain in precision over simple random sampling is, in this allocation,

entirely due to that prior knowledge of the population that led to its partitioning (i.e., the knowledge that all of the variances within strata are smaller than the overall population variance).

Proportional allocation overlooks two items of information that may be at the disposal of the analyst: (1) differences in variance (σ_i^2) from stratum to stratum, and (2) differences in the cost (c_i) associated with observing a unit in each stratum. Since the dual purpose is to minimize both overall sampling variance and cost, it follows that more units should be drawn from high variance strata where sampling is inexpensive. When the stratum sample sizes (m_i) are set proportional to the respective standard deviations (σ_i) and stratum sizes (M_i), and inversely proportional to the square root of the costs (c_i), allocation is said to be optimal. The fact that some uncertainty may be attached to the knowledge of σ_i and c_i does not impair the lack of bias of the final estimate of μ. If the analyst is confident in his estimate of at least the relative magnitudes of the c_i and of the σ_i, it is better to use optimal allocation rather than proportional allocation.

Example. Suppose that an estimate is desired for the average dollar-cost of replenishment spares for a tactical fighter with the following deployment (by command):

Command	No. UE
TAC	450
TAC-CCTW*	150
PACAF	75
USAFE	75
	750

For each aircraft there is a record of all major modifications, spares consumed, and major maintenance. Each aircraft can be identified by its tail-number. Assume that the desired tolerance for estimated average cost is ± $300 at the 90 percent confidence level.

*Combat Crew Training Wing.

The first task is to get a rough estimate of the standard deviation of spares costs for the 750 aircraft. Suppose an informed individual suggests that the distribution of costs is fairly bell-shaped, but skewed to the right; furthermore, he feels that about 95 percent of the aircraft have spares costs equal to \$11,250 $\pm$ 2800. Noting that $\pm 2\sigma$ usually encompasses 95 percent, the standard deviation is estimated to be ½(2800), or \$1400.

If simple random sampling were applied to this problem, the sample size would be determined as follows:

$$300 = \left(z_{.10}\right)S_{\bar{x}} = 1.65\left(\frac{1400}{\sqrt{n}}\right)$$

$$n = 59.4 \cong 60$$

One could expect to do better by stratifying according to the four command-categories (TAC, PACAF, etc.) above, since program characteristics (flying-hour programs, etc.) are likely to affect spares consumption. Using the same sample size as above, and adopting proportional allocation, the m_i for the various strata are:

Stratum	m_i
TAC	60(450/750) = 36
TAC-CCTW	60(150/750) = 12
PACAF	60(75/750) = 6
USAFE	60(75/750) = 6
	60

The design may be improved by speculating as to the relative differences in dispersion and sampling costs among the strata by using optimal allocation:

$$M_i \text{ proportional to } \frac{M_i\sigma_i}{\sqrt{c_i}}$$

Suppose one could expect sampling costs overseas to be double those in the U.S. Furthermore, one might expect the dispersion in TAC-CCTW to

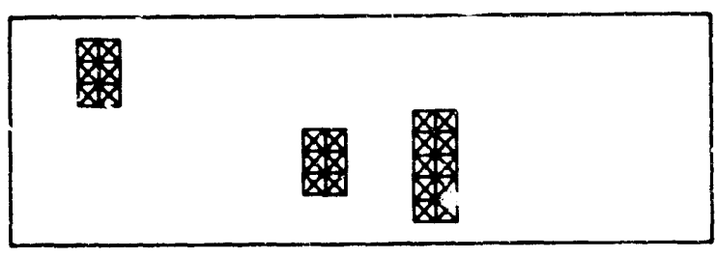
be one-half the dispersion within TAC, with the other strata somewhere in between. Accordingly, the following table lists relative costs, standard deviations, and the allocation that results:

Stratum	Relative Cost	Relative Disp.	Estimate of $M_i(\sigma_i/C_i^{\frac{1}{2}})$	m_i
TAC	1	4	$450(4/1^{\frac{1}{2}}) = 1800$	$60(1800/2420) = 44.6 = 45$
TAC-CCTW	1	2	$150(2/1^{\frac{1}{2}}) = 300$	$60(300/2420) = 7.4 = 7$
PACAF	2	3	$75(3/2^{\frac{1}{2}}) = 160$	$60(160/2420) = 4.0 = 4$
USAFE	2	3	$75(3/2^{\frac{1}{2}}) = 160$	$60(160/2420) = 4.0 = 4$
			2420	60

Notice that in the new allocation, high-cost strata are sampled less and the high-variance stratum is sampled more.

Cluster Sampling

In cluster sampling, the population is divided into groups, or clusters, of units. Several of the clusters are chosen at random, and all units in each selected cluster become part of the sample. The clusters are referred to as primaries, whereas the units contained therein are secondaries.



There are two major reasons that lead to the choice of cluster sampling.

First, there is sometimes no list of the population available on which to base a sampling frame and it is felt that such a list would

be too expensive to construct, whereas it is relatively easy to come by a list of clusters of units. Suppose it is desired to sample message lengths in a Communications Sector. The practical procedures would be to sample clusters of messages, i.e., messages received at selected installations during some specified time interval.

Second, cluster sampling may be desirable if the population is such that travel costs can be reduced by selecting adjacent units. For example, if failure rates for some item of base equipment are being sampled, it may be cheaper to select a number of bases and observe all units on those bases than to take a simple random sample.

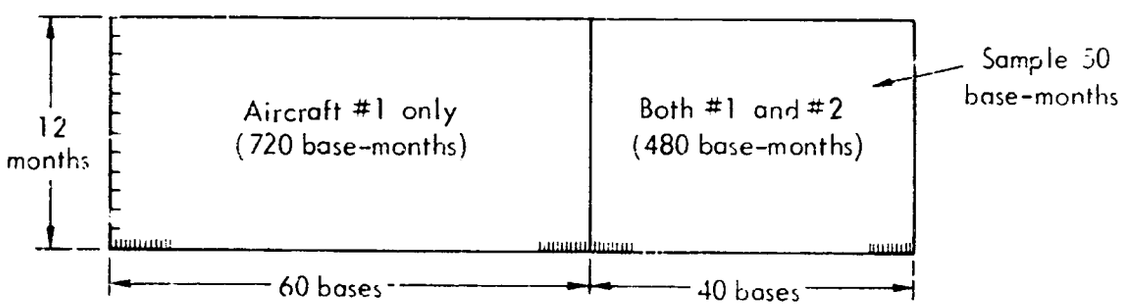
The relative cost for specified precision (and equivalently the relative variance for specified cost) is (1) proportional to the relative cost of observing one cluster, (2) proportional to the variation between clusters, and (3) inversely proportional to the relative size of the cluster. If in estimating $\bar{X}$ a choice is to be made between several different cluster sizes, it can be shown that the criterion is to choose that cluster size that minimizes the product of sampling variance times total cost (both of which vary, depending on cluster size).

When cluster sampling is chosen as a matter of convenience, the final estimate will generally be less precise than a simple random sample of the same size. Therefore the decision rests on whether the cost reduction allows the selection of a large enough sample to actually increase precision. This situation contrasts with the stratified sample, where an estimate less precise than that from simple random sampling is very unlikely, and would almost require contrived strata designed specifically for that result. Of course, if the clustering were designed so that variation <u>within</u> clusters was greater than that between clusters, then the estimate would be more precise than the simple random case. Such an arrangement is not likely; it is typically easier to partition the population into groups of homogeneous units (as in stratification) than heterogeneous units.

<u>Example</u>. A frequent problem for the cost analyst is to estimate the cost of consumption items that are common to more than one system.

Frequently, these items are centrally managed, and their consumption reported only in aggregate. Let us postulate, for example, a study of administrative/support aircraft, in which it is desired to know the annual cost of low value replenishment spares consumed. A large group of spares are common to two aircraft (aircraft #1 and aircraft #2) assigned to 100 world-wide locations. Consumption accounting is by commodity only, necessitating some external data collection for study purposes. One solution would be to request that maintenance managers at each of the 100 locations keep detailed records of the final application of the common spares in question. This would be time consuming and costly and would probably provide more detail than necessary. What follows is a cluster sampling design that would probably provide very adequate information at significantly less cost.

Suppose that aircraft #1 is stationed on all 100 bases, but aircraft #2 is only on 40 bases. Designating a one-year time period and defining a cluster to be a one-month period (12 clusters per base), the population contains 1200 clusters, 50 of which will be sampled. Since all common spares in question sent to 60 bases are consumed by aircraft #1, attention may be restricted to the remaining 40 bases:



The procedure will be to estimate the proportion of common spares by aircraft #2 in the smaller stratum, then make an adjustment to allow for the other 60 bases.

Fifty clusters are randomly chosen from the smaller stratum. The maintenance chief at each selected base is instructed to keep records regarding the disposition of all common spares during the particular

month(s) chosen. The information to be reported is the month's total consumption of common spares (M_i) and the consumption recorded for aircraft #2 (X_i). When all the information is in, the estimated proportion of common spares going to aircraft #2 for the 40 bases is:

$$P_* = \frac{\sum^{50} X_i}{\sum^{50} M_i}$$

where $\sum^{50} M_i$ is the total sample consumption and $\sum^{50} X_i$ is consumption by aircraft #2. The sampling variance for this estimator is estimated by the formula for unequal cluster sizes, substituting P_* for $\bar{X}_{cl}$:

$$S_*^2 = \frac{1 - \frac{50}{480}}{50(\bar{M})^2 49}\left[\sum^{50} M_i^2 + P_*^2 \sum^{50} M_i^2 - 2P_* \sum^{50} M_i X_i\right]$$

The proportion of common spares consumed by all bases for aircraft #2 is then estimated by weighting P_* to allow for the difference between the 40 bases and the entire 100 bases:

$$\hat{P}_2 = \left(\frac{\sum^{480} M_i}{\sum^{1200} M_i}\right) P_*$$

where $\sum^{480} M_i$ is the year's consumption at the 40 bases (representing 480 clusters), and $\sum^{1200} M_i$ is the consumption for all 100 bases (equivalent to 1200 clusters). The proportion consumed by aircraft #1 is estimated by:

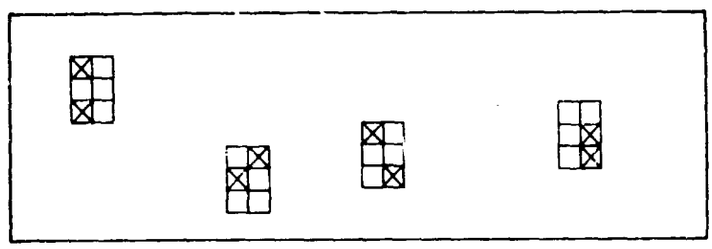
$$\hat{P}_1 = 1 - \hat{P}_2 .$$

The sampling variance is the same for both P_1 and P_2, and is estimated by weighting S_*^2:

$$S_p^2 = \left(\frac{\sum^{480} M_i}{\sum^{1200} M_i} \right)^2 S_*^2$$

Subsampling

Subsampling, or two-stage sampling, is a hybrid of cluster and stratified sampling. The population is partitioned into N primaries, and n of these primaries are randomly selected. A subsample of m secondaries is then randomly selected from each primary. This technique is sometimes extended to three or four stages. The discussion that follows will consider the case where each primary contains the same number (M) of secondaries, and the same number of secondaries (m) are sampled from each primary.



n 4
M = 6
m 2

The main advantage of subsampling over one-stage sampling is flexibility. It reduces to cluster sampling when m = M, or to stratified sampling when n = N; but in terms of the cost-precision tradeoff, a scheme that falls somewhere between these two may be preferable. The problem is to determine values of n and m such as to minimize sampling variance for a given cost (or equivalently, to minimize cost for a specified variance). Appendix II provides a method for solving this problem that requires preliminary estimates of (1) the cost of sampling associated with each cluster (c_1), (2) the cost of sampling secondaries within clusters (c_2), (3) the variance between cluster means (S_B^2), and (4) the variance of secondaries within clusters (S_w^2).

For this purpose, these estimates do not require great precision because the sampling variance is not highly sensitive to the choice of m. It is usually easier to estimate ratios c_1/c_2 and S_w^2/S_B^2, in which case tables are available to aid the evaluation of m (see Cochran, page 282).

Example. Since the greater portion of USAF base-level reporting systems have been designed primarily for management and control purposes, the needs of the planning and programming oriented cost analyst are not always satisfied; it has generally been more expedient to put accountability on an organizational basis rather than a program basis. Certain base-support organizations provide service to a plurality of programs, and in order to allocate activity on a program basis, the cost analyst must often adopt some arbitrary pro-ration scheme.

The following example suggests how a subsampling design might be used to estimate the average daily man-hours devoted by Civil Engineering squadrons to repair and maintenance of aircraft alert facilities during a 90-day period. It is assumed that a daily record of work-orders is maintained in a general ledger, and that inspection of the ledger will provide the data needed.

Assume that there are 126 C-E squadrons overseas and in the Z.I.; each of these will be regarded as a primary cluster. Each cluster consists of 90 days of information. The procedure will be to select n squadrons at random, then select m days within each cluster. The total man-hours devoted to aircraft alert facilities maintenance during the selected squadron-days will be found by examining the appropriate ledger.

The first problem is to decide the optimum value of m. This requires "guesstimates" of S_w^2, S_B^2, c_1, and c_2. Since there are about two to 300 entries per day in each squadron's ledger, an allowance of four hours per squadron-day seems reasonable. The cost, c_2, of visiting each squadron would be in the neighborhood of one and a quarter "working" days, or 10 hours. S_w^2 and S_B^2 are considerably more elusive, but suppose that examination of ledgers from two or three representative squadrons suggests 230 and 40, respectively; the optimum subsample size (m) is then determined as:

$$m = \sqrt{\left(\frac{S_w^2}{S_B^2 - \frac{S_w^2}{M}}\right)\frac{c_1}{c_2}} = \sqrt{\left(\frac{230}{40 - \frac{230}{90}}\right)\frac{10}{4}}$$

$$= 6.3 \overset{\circ}{=} 6$$

The number of clusters selected (n) can be determined in one of two ways, depending on whether total cost or overall precision is held constant. Suppose the total time allocated to the collection of data is set at 40 workdays (320 hours):

$$C = nc_1 + nmc_2$$

$$320 = n(10) + n(6)(4)$$

$$n = 9.4 \overset{\circ}{=} 9$$

If, on the other hand, one can tolerate a sample mean variance of about 5 man-hours, the following formula is solved for n:

$$S^2_{\bar{x}_{ss}} = \frac{S_B^2}{n}(1-f_1) + \frac{S_w^2}{nm}(1-f_2)f_1 = 5$$

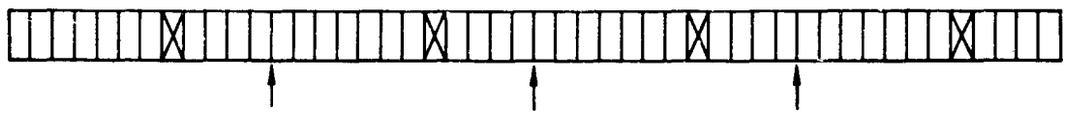
$$= \frac{40}{n}\left(1-\frac{n}{126}\right) + \frac{230}{n(6)}\left(1-\frac{6}{90}\right)\frac{n}{126}$$

$$n = 7.1 \overset{\circ}{=} 7$$

Systematic Sampling

Systematic sampling is not so much a sampling "technique" as it is a refinement in the use of random numbers. It is discussed here because it often produces the same effects as stratification or clustering, and because it is almost an indispensable device when sampling from very large frames.

The procedure begins with the decision to sample some fraction of the population, say 1/12. The population is listed and a random number is selected between 1 and 12, say 8. For the sample, the eighth unit, and every twelfth unit thereafter, are selected (i.e., #8, #20, #32, #44, etc.).



Systematic sampling has two advantages over simple random sampling. First, it is easier to draw the sample, since only one random number is required. Second, it distributes the sample more evenly over the population and therefore often provides more accurate results.

There are also two potential disadvantages. If the population contains some periodic variation, and the sampling interval coincides with that variation, the sample obtained may be badly biased. Second, evaluation of sampling variance is contingent on knowing the behavior of the population with respect to the listing.

IV. REFINEMENTS IN THE ESTIMATOR

The techniques previously discussed each dealt with some way of partitioning the population preliminary to drawing the sample. The estimator of the population mean, $\bar{X}$, was the same in all cases: the simple or weighted average of the sampled X_i.

There are many sampling situations where there exists some "auxiliary" variable which is known to correlate with the variable of interest. In such cases, sampling variance can be reduced by instituting a basic change in the estimator so as to take advantage of the information contained in the auxiliary variable. This is the case with ratio and regression estimation, which are explained in this section. A third technique, unequal probability sampling, uses the auxiliary variable in determining selection probabilities as well as in the estimator.

The format for this section is similar to that of Section III, although the more complex designs inherently require more formulations in their descriptions. A summary of the fundamental characteristics of all the sampling techniques described in this document concludes this section.

RATIO ESTIMATOR

In ratio estimation, two variables are observed on each sample unit: X_i, the variate of interest, and W_i, an auxiliary variable. The auxiliary variable is such that its population mean, μ_w, is known. The ratio estimate of the population mean of the X_i is given by:

$$\bar{X}_R = \left(\frac{\frac{\Sigma X_i}{n}}{\frac{\Sigma W_i}{n}}\right)\mu_w = \left(\frac{\bar{X}}{\bar{W}}\right)\mu_w$$

The ratio estimator is biased, except in the situation where a regression of X on W would be a straight line through the origin (i.e., the ratio X_i/W_i is approximately constant). The bias is negligible in large samples.

The sampling distribution is hard to pin down, since both X and W vary from sample to sample. However, for large samples, the distribution tends to normal and the bias in the approximate variance formula becomes negligible.

In spite of these difficulties, ratio estimation can be a very useful way to use extraneous information that is not directly of interest to the analyst. If this extra information is easily picked up with the regular sample, the gain in precision is cheap, since only the final computations are affected.

Knowledge of the exact relationship between X and W is not required, but in order for the precision of the ratio estimate to be greater than a simple sample mean, it is necessary that the following condition holds:

$$\rho_{xw} > \frac{CV_w}{2CV_x}, \qquad CV_w = \sigma_w/\mu_w$$

$$CV_x = \sigma_x/\mu_x$$

where ρ_{xw} is the correlation coefficient between X and W, and CV_x and CV_w are the coefficients of variation for X and W, respectively.

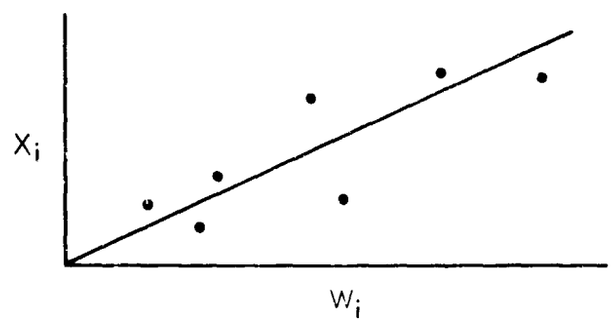
The variability of the auxiliary variate, W, is thus an important factor; if its coefficient of variation is more than twice that of X, the ratio estimate is always less precise, since ρ_{xw} cannot exceed 1. The preceding result is based on the approximate variance formula and therefore is applicable to large samples; for small samples, the condition would be more stringent, since the approximate formula is usually an underestimate.

Example

A common use of the ratio estimator occurs when there has been a complete census of the particular variable of interest in some previous time period. Suppose it is desired to estimate the current average inventory of fuel at USAF air bases, and that for purposes of the example these data are available on a base-by-base basis only as of the end of the previous year. Let X_i be the current inventory and W_i

the previous inventory at the ith base in the sample. The population average as of the end of the previous year will be indicated by μ_w.

Before applying the ratio estimator, it will be prudent to determine its usefulness compared with a simple sample mean. It is reasonable to assume that the ratio estimator will be unbiased (i.e., X_i/W_i is constant) since a force-wide adjustment in fuel inventories would probably derive from some implicit general policy change that has proportional effects on all bases. A quick check of this assumption can be made by plotting X against W, noting whether a freehand regression line passes through the origin. For example:



(If the regression line does not pass through the origin, and the sample is not large, it would be preferable to consider the regression estimator as described in later pages.) Attention is next directed to whether the ratio estimator is more precise than the simple mean, using the criterion $\rho_{xw} > \frac{1}{2}(CV_w)/(CV_x)$. In this case, CV_w and CV_x are probably the same, since W and X are essentially the same variable. So the question reduces to whether ρ_{xw} is greater than one-half, which does not seem unreasonable unless base fuel inventories fluctuate widely over time. A quick check is provided by observing whether the free-hand regression line seems to "explain" more than one-half the variation in X.

If the foregoing analysis establishes the ratio estimator as appropriate, estimates of the mean and variance proceed according to the formulas given in Appendix II. Supposing there are 150 air bases in the population from which 20 are sampled, the calculations might proceed according to the following worksheet (inventories are expressed in thousand of barrels):

Population Data		Sample Selection	Sample Data		
Item No.	W_i (Previous Inventory)		W_i	X_i (Current Inventory)	$[X_i - (\bar{X}/\bar{W})(W_i)]^2$
1	24	✓	24	22	31.4
2	49	--	--	--	--
3	56	--	--	--	--
4	19	✓	19	25	9.6
↓	↓	↓	↓	↓	↓
149	72	✓	72	81	3.2
150	51	--	--	--	--
Totals	8250		940	1080	592.3

$$\mu_w = 8250/150 = 55$$

$$\bar{W} = 940/20 = 47$$

$$\bar{X} = 1080/20 = 54$$

The ratio estimate of average fuel inventory is:

$$\bar{X}_R = \left(\frac{\bar{X}}{\bar{W}}\right)\mu_w = \left(\frac{54}{47}\right)55 = 63$$

The variance of the ratio estimate is estimated by:

$$S^2_{\bar{x}_R} \cong \frac{1-f}{n(n-1)} \sum^{n} \left[X_i - \left(\frac{\bar{X}}{\bar{W}}\right)W_i\right]^2 = \frac{1 - \frac{20}{150}}{20(19)}(592.3) = 1.4$$

This example has not included any discussion of how the data are to be collected. This simplest case would be a simple random selection of bases, but there is no reason why stratified or cluster sampling should not be used, if the characteristics of the population warrant it. In the present example, it would probably be useful to stratify by major command, since base fuel consumption should be significantly more

homogeneous within commands than between commands. Remembering that the stratified estimator of the population mean is just a weighted average of stratum means, the stratified-ratio estimator can be written as:

$$\bar{X}_R = \sum^{N}\left[\left(\frac{M_i}{N\bar{M}}\right)\left(\frac{\bar{X}_i}{\bar{W}_i}\right)\mu_{w_i}\right]$$

where $\bar{X}_i$ and $\bar{W}_i$ are simple means of bases sampled from the ith command. The estimator of variance will be:

$$S^2_{\bar{X}_R} \cong \sum^{N}\left[\left(\frac{M_i}{N\bar{M}}\right)^2\left\{\left(\frac{1-f_i}{m_i}\right)\frac{\sum^{m_i}\left(X_{ij} - \left(\frac{\bar{X}_i}{\bar{W}_i}\right)W_{ij}\right)^2}{m_i-1}\right\}\right]$$

Thus, even a simple marriage of two sampling techniques complicates estimation of variance. This problem is discussed in a general way under the heading <u>Complex Designs</u>, beginning on page 57.

REGRESSION ESTIMATOR

The regression estimator is more appropriate than the ratio estimator if the relation between X and W is linear but does not go through the origin. In this case, the estimate of the population mean is:

$$\bar{X}_r = \bar{X} + b(\mu_w - \bar{W})$$

where b is an estimate of the change in X when W is increased by 1. The reasoning is that if the sample $\bar{W}$ is below average, one could expect the sample $\bar{X}$ to also be below average by an amount $b(\mu_w - \bar{W})$. The value of b is usually estimated from the sample using the least-squares estimator:

$$b = \frac{\sum^{n}(X_i - \bar{X})(W_i - \bar{W})}{\sum^{n}(W_i - \bar{W})^2}$$

Contrary to the case in general regression analysis, it is not necessary to assume exact linearity between X and W, nor that the variance of X for a given W_i is constant (again, provided the sample size is large).

As with ratio estimates, the regression estimate is generally biased. But for large samples, the ratio of bias to standard error becomes small, making the bias negligible. Furthermore, there is no bias if an exact linear relationship exists between X and W. What constitutes a "large" sample depends on how X and W are correlated, and cannot be summarized by a rule of thumb.

For large samples, the regression estimate is more precise than the simple sample mean provided that there is some correlation between X and W; it is more precise than the ratio estimate unless the relation between X and W is a straight line through the origin. Thus, there is nothing to lose in using a regression estimator except the extra time spent in calculation.

Example

An interesting application of the regression estimator is the use of "eyeball" estimates for the auxiliary variables. For example, suppose there is a proposal to replace some training equipment at an air base, but it is first necessary to assess the salvage value of the old equipment. The analyst, or a salvage expert, would quickly survey each item of equipment, roughly estimating its approximate salvage value. Then a random sample would be selected, and the exact salvage value of each sampled item determined by close inspection. The regression estimator is then applied, labeling the individual rough estimates W_i, the average of all rough estimates μ_w, and the more thorough estimates X_i.

Supposing the population contains 120 items of equipment and a sample of 20 is to be drawn, the following analysis might result:

Population Data		Sample Data						
Item No.	W	W_i	X_i	$(X_i-\bar{X})$	$(W_i-\bar{W})$	$(X_i-\bar{X})(W_i-\bar{W})$	$(W_i-W)^2$	$[(X_i-\bar{X})-b(W_i-\bar{W})]^2$
1	190	--	--	--	--	--	--	
2	220	220	214	5	5	25	25	1.7
3	230	230	223	14	15	120	225	9.0
4	150	--	--	--	--	--	--	--
↓	↓	↓	↓	↓	↓	↓	↓	↓
119	180	180	176	-33	-35	1155	1225	54.8
120	90	--	--	--	--	--	--	--
Totals	$23760	4300	4180	0	0	10350	14083	605.8

$$\mu_w = 23{,}760/120 = 198$$

$$\bar{W} = 4300/20 = 215$$

$$\bar{X} = 4180/20 = 209$$

$$b = \frac{\Sigma(X_i-\bar{X})(W_i-\bar{W})}{\Sigma(W_i-\bar{W})^2} = 10{,}350/14{,}083 = .73$$

The regression estimate of average salvage value is given by:

$$X_r = \bar{X} + b(\mu_w-\bar{W}) = 209 + .73(198-215) = 196.6$$

The variance of this estimate is estimated by

$$s^2_{x_r} \doteq \frac{1-f}{n(n-2)} \Sigma^{20} [(X_i-\bar{X})-b(W_i-\bar{W})]^2$$

$$= \frac{1-\frac{20}{120}}{20(18)}(605.8) = 1.4$$

Although the rough estimates (W_i) are biased, one could expect the bias to be constant from item to item, except for random variation. If this random variation is not too great, the regression estimator will be unbiased for small samples. For this reason, it is important

that the same person make all the rough estimates.* It is also important that this person not know what items fall into the sample until the rough estimates have been made. Provided the latter condition holds, the consistency of rough estimates can be checked by plotting the sample X_i versus the W_i.

UNEQUAL PROBABILITY SAMPLING

This technique utilizes an auxiliary variable in determining selection probabilities as well as in a special estimator. As previously mentioned, the idea is to find a variable which is closely correlated with the particular variable of interest. Probabilities of selection are set proportional to the former, the sample is collected, and the following estimator is used:

$$\bar{X}_p = \frac{1}{nN}\sum^{n}\frac{X_i}{P_i}$$

where X_i is the variable of interest, W_i is the auxiliary variable, and $P_i = \frac{W_i}{\Sigma W_i}$ is the probability of selecting X_i.

Unequal probability sampling is a great aid in increasing precision, when an auxiliary variable with the proper characteristics is available. The technique has received much attention in the past ten years or so despite problems in application. For example, in replacement sampling the calculation of variance is straightforward. When sampling with nonreplacement, however, there are problems of controlling the P_i and estimating variance that are beyond the scope of this paper. Furthermore, the exact form of the sampling distribution is not known. Suffice it to say that gains are to be made when X and W are closely correlated, but that the complete theory of this kind of sampling is still being developed in current research literature (see

*An alternative way to arrive at rough estimates is for the analyst to develop an estimating relationship on the basis of historical information, using such parameters as original cost, age, and usage rate.

bibliography). A simplified example will be given to illustrate the power of the technique.

Example I

Suppose it is decided to estimate the total personnel stationed on five military installations in some remote region of Northern Canada, using a sample of three. It is expected that the average number of personnel presently stationed at each installation (variable X) is closely correlated with the average number of personnel of the previous year (variable W), the data for which are known. The procedure is to choose three random numbers between 1 and $\sum^{n} W_i$, making the selection of sample points on the basis of a cumulative list of variable W. Hence:

Base	W_i	Cumulative W_i	Random Number		X_i
1	22	22	14		25
2	36	58			37
3	21	79	62		29
4	34	113	97		34
5	11	124			12
Total	124				137

The usual estimate for simple random sampling would be:

$$N\bar{X} = \frac{N}{n}\sum^{n} X_i = \frac{N}{n}(X_1 + X_3 + X_4) = \frac{5}{3}(25 + 29 + 34) = 152$$

The unequal probability estimate is:

$$N\bar{X}_p = \frac{1}{n}\sum^{n} \frac{X_i}{P_i} = \frac{1}{n}\left(\frac{X_1}{P_1} + \frac{X_3}{P_3} + \frac{X_4}{P_4}\right)$$

$$= 1/3\left[25\left(\frac{124}{22}\right) + 29\left(\frac{124}{21}\right) + 34\left(\frac{124}{34}\right)\right] = 145$$

where $p_i = \frac{W_i}{\Sigma W_i}$ is the probability with which the sample point entered the sample.

There are a total of ten possible samples of size three that could be drawn from this population. The following table compares the simple random sample estimate with the unequal probability estimate for each case:

Bases sampled	123	124	125	134	135	145	234	235	245	345
Simple random	152	160	123	147	110	118	167	130	138	125
Unequal probability	147	131	135	145	149	133	141	145	129	143

Except for two cases, the unequal probability estimate is closer to the population value of 137. The standard error for the unequal probability estimates is 7.3, whereas that for simple random sampling is 17.8. If W and X were more closely correlated, one would expect even better results.

Example II

Sampling with unequal probabilities is often used to yield a "self-weighting" sample in cluster sampling or subsampling when the clusters are of unequal size. This is the context in which Hansen and Hurwitz first introduced the technique.

When sampling n clusters from a total population of N clusters, where the cluster size, M_i, is the same for each cluster, the population mean is estimated by averaging the cluster means:

$$\bar{X}_{cl} = \frac{1}{n}\sum^{n}\bar{X}_i$$

However, if cluster size varies from clust cluster, a weighted estimator would be more precise:

$$\bar{X}_{cl} = \frac{1}{n}\sum^{n}\left(\frac{M_i}{\bar{M}}\right)\bar{X}_i = \frac{1}{n\bar{M}}\sum^{nM_i} y_{ij}$$

where $\bar{M}$ is the average cluster size. This expression can be manipulated as follows:

$$\bar{X}_{cl} = \frac{1}{n}\sum^{n}\left(\frac{M_i}{\bar{M}}\right)\bar{X}_i = \frac{N}{nN}\sum^{n}\left(\frac{M_i}{\bar{M}}\right)X_i = \frac{1}{NP_i}\sum^{n}\left(\frac{M_i}{\bar{M}}\right)\bar{X}_i$$

$$= \sum^{n}\left[\frac{1}{NP_i}\left(\frac{M_i}{\bar{M}}\right)\bar{X}_i\right]$$

where P_i is the probability of selecting the ith cluster. So far, all of the P_i's have been the same. Suppose, however, that each p_i is made proportional to its corresponding M_i. The probability of selecting each cluster is then $n(M_i/\bar{M}N)$. Substituting this into the above formula gives:

$$\bar{X}_{cl} = \sum^{n}\left[\frac{1}{N}\left(\frac{\bar{M}N}{nM_i}\right)\left(\frac{M_i}{\bar{M}}\right)\bar{X}_i\right] = \sum^{n}\left(\frac{\bar{X}_i}{n}\right) = \frac{1}{n}\sum^{n}\bar{X}_i$$

which is the same as the simple unweighted estimator. Thus, the sample is said to be self-weighting: $\bar{X}$ is the appropriate estimator for X even though cluster sizes vary.

The technique for selecting the clusters with unequal probabilities is the same as outlined before, except that the basis for selection is now a cumulative list of cluster sizes, rather than the auxiliary variable. The worksheet for such a sample might have the following format:

Cluster	M_i	Cumulative M_i	Random Number		X_i
1	14	14	--		--
2	22	36	25		105
3	7	43	--		--
4	18	61	47		92
5	23	84	82		112
6	12	96	--		--

The estimated population mean is then:

$$\bar{X}_{c1} = \frac{1}{3}(105 + 92 + 112) = 103.$$

COMPARISON OF DESIGNS

The task of compiling some sort of quantitative comparison of the foregoing sample designs is not realistic, since so much depends on the characteristics of the particular population under study (see Des Raj, Zarkovich). It may be helpful to briefly categorize the attributes of the various designs and estimation procedures as they relate to accuracy and cost:

Simple random sample	o Simplest design.
Stratified sampling	o Nearly always more precise than simple random sample.
Cluster sampling	o Simpler frame and reduced travel costs.
	o Usually less precise than simple random sample.
Sub-sampling	o Flexibility in balancing cost-precision trade-off, especially when convenient cluster size is too small for stratification and too large for cluster sampling.
Systematic sampling	o Ease in selecting sample points.
	o May give better representation, depending on frame.
	o May be biased, depending on frame.

Ratio estimator	o Usually more precision than simple estimator. o "Significant" bias in small samples.
Regression estimator	o More precision than simple estimator. o "Significant" bias in small samples.
Unequal probabilities	o Usually more precision than equal probabilities. o Difficult to assess sampling error.

Usually, the circumstances of the analysis readily suggest the most appropriate design or combination of designs. For example, USAF-wide sampling immediately leads to the possibility of stratification by major command or some geographical classification; also, the presence of a convenient auxiliary variable makes ratio or regression estimators attractive.

On the other hand, there are often factors to consider that do not readily fit into the framework of cost-precision tradeoffs. One such factor is the need to minimize the imposition of field work on USAF personnel who have other responsibilities (e.g., maintenance chiefs or accounting clerks); the essence of sample work is loyal adherence to good procedure, and it is often more fuss than the busy serviceman can handle. More often than not, the situation will be such that there is no completely objective approach to designing the sample.

In any case, the general procedure is the same as with all problem solving: specify the objectives, survey factors related to the problem, identify alternative solutions, quantify the problems as much as possible to reduce subjective uncertainty, and make such intuitive decisions as are necessary.

V. SOME REAL LIFE COMPLEXITIES

This Section gives recognition to some topics which often arise in the application of sampling techniques and which seem particularly relevant to the forecasting nature of cost analysis. Complex sample design is discussed and an example of a sampling study recently conducted by the Cost Analysis Division, Headquarters Strategic Air Command, is described. The Section concludes with a discussion of the application of sampled data in regression analysis.

COMPLEX DESIGNS

Theory surrounding the subject of complex sample designs is generally less developed than for the basic designs, and documentation of research is highly fragmented among various journal publications and a few books.

Two kinds of complexity are worth noting: (1) compounding of design and (2) compounding of purpose.

Compounding Designs

Sometimes the characteristics of the population are such that it is convenient to compound the various basic designs. Drawing on the example in the previous section where the salvage value of some training equipment was estimated, suppose a USAF-wide estimate was desired. The most simple scheme might be to select a simple random sample of 30 bases, then apply the regression estimator within each. A more precise estimate might be achieved by designing a "complex" sample along the following lines: (1) stratify bases on a two-way scheme using major command and geography as classifications, resulting in about 15 strata; (2) select two bases for each strata with unequal probabilities, using number-of-airmen as the auxiliary variable; (3) sub-sample several items of equipment from each base; (4) estimate the total salvage value for each base with a regression estimator, using the salvage expert's "eyeball" estimates as the auxiliary variable. The total USAF estimate

would be given by:

$$Y = \frac{1}{2} \sum^{15} \left(\frac{Y_i}{P_i} + \frac{Y_j}{P_j} \right)$$

where Y_i and Y_j are the two estimated base totals from each stratum, and P_i and P_j are their respective probabilities of selection. It would be extremely difficult to estimate what the sampling variance from such a design would be. The rationale for the procedure was generated by reasoning subjectively at each stage of the design that some particular technique would contribute most to precision in the final estimate. Although an objective estimate of the sampling error is not known, an upper limit may be set by computing the error that would result from a less complex design.

There are simplified methods for computing the sampling error once the sample has been drawn. One way is to use the technique called replication. Instead of drawing the entire sample in one operation, only a fraction of the sample points are drawn, and the procedure is repeated until the total sample is drawn. The variance of the sampling distribution is then computed from the several estimates. The example above, for example, might be completed in three seperate samples, the difference being that for each sample, only one-third as many units of training equipment are selected from each base. The overall sample size is still about the same. Sampling variance is estimated as follows:

$$S^2_{\bar{X}} = \frac{1}{2} \sum^{3} (\bar{X}_i - \bar{X})^2$$

where $\bar{X}$ ($i = 1, 2, 3$) is the estimate from the ith sample and $\bar{X}$ is the average of the three samples. The use of replication in sample design is treated extensively in Deming. Discussion of other methods is found in Zarkovich.

Multi-Purpose Surveys

Survey design has been described as a process of evaluating alternative methods in terms of relative cost and precision. This is reasonably straightforward when only one characteristic is under measurement.

It often happens that the sampler takes advantage of the situation by observing several characteristics instead of one. For example, a survey of airmen may involve observations on age, training, motivation, and rank. Determination of optimum design is now considerably less objective. The proper sample size for one characteristic may provide no useful information on a second characteristic, and give superfluous precision on a third. One characteristic may be perfectly suited for a stratified design while a companion characteristic is more adaptable to something else. Objectivity requires the assessment of the relative utility of the different information sought. These problems are discussed in Kish and in Yates.

A related complexity is found in the so-called "analytic" surveys. The objective of an analytic survey generally is to make comparisons between sub-populations, where the sub-populations cannot be framed (i.e., sampling units can be identified by sub-population only after the sample is taken). Referring to the sample on page 42, (sub-sampling from Civil Engineering Squadron work-order ledgers), the purpose might well have been to compare the resources devoted to several program categories. Each squadron-day selected in the sample would consist of work-orders in one or more categories, leading to estimates of total activity devoted to each program. So, in effect, several samples are being conducted, one for each sub-population. The feature that makes this different from other procedures heretofore discussed is that the sample size from each sub-population is also a variable. Furthermore, the sample sizes are negatively correlated and cannot be treated as independent variables. Procedures for handling this situation are discussed in Yates and in Hartley (the latter reference is probably the more straightforward). The general problem of analytical statistics from complex samples is summarized by Kish (pages 582-587), including a brief description of seven approaches to computing or approximating standard errors.

<u>Example: SAC Aircraft Maintenance</u>

The example that follows is a rather detailed description of a research program recently undertaken in SAC. The project is of

direct interest here because (1) it illustrates a rather complex sample design problem, and (2) it provides a case study of data collection at a level of aggregation useful to a cost analyst.

The Cost Division at SAC Headquarters used probability sampling to collect some maintenance man-hour and material cost data on the KC-135, the B-52 (G and H series), and the UH-1F. Sampling was necessitated by the desire to obtain data from the original source documents, a procedure involving considerable effort; probability selection was chosen because of a preference for unbiased estimates and because there was no apparent basis for assuming a more precise judgment sample. Since the selection procedures are similar for all three aircraft, emphasis will be on the KC-135 sample.

Motivation. The project had three primary objectives. The first and most important was to evaluate the general behavior of maintenance requirements as an aircraft ages. Current aircraft costing models often assume (at least implicitly) that maintenance costs slope downward during the initial months following deployment into the force, then level off after "shake-down" is accomplished. SAC cost analysts hypothesize that, instead of leveling off indefinitely, costs tend to rise again as the equipment gets older. There is considerable interest in resolving the question since (1) there is some uncertainty as to when the strategic aircraft in the force will be replaced, and (2) the proposed Resources Management System has suggested changes in military management that could triple SAC's responsibility in programming and budgeting for maintenance resources.

The second objective of the sample was to explore the relationship, if any, between maintenance man-hours and other maintenance costs. It is common practice to pro-rate base maintenance costs among the various aircraft systems on the basis of man-hours. Since some systems require relatively greater parts requirements than others, the validity of such practice is questionable.

The third objective was to investigate errors in recording and reporting maintenance material consumption. There is evidence that parts data are sometimes treated in cavalier fashion from crew level on up the

line to final reporting. If sources of error can be identified and subsequently reduced, the maintenance data will have greater utility for financial planning.

The Population. The population to be sampled consisted of the documents on which maintenance personnel record their work (AFTO 210, 211 and 212). This includes parts and labor expended by the base maintenance shops (field maintenance, CA&E maintenance, etc.); bench stock items are omitted, but this is a very insignificant portion of overall maintenance. These source documents are easily identified by aircraft tail-number.

Design. Sample design was addressed primarily to the first objective of the study, and the other two were more or less regarded as by-products of the first. The general idea was to obtain estimates of maintenance labor and material costs for each of several age groups, then observe whether the estimates conform to the hypothesized curve (data on engines would be recorded separately since engines move around from aircraft to aircraft). Initial delivery dates of the KC-135s range from 1958 to 1965, providing eight yearly age groups.

From the standpoint of precision, the sample design should assure good representation over a number of variables besides age that affect maintenance. For example, some variation can probably be associated with base-to-base differences in climate and maintenance management. Differences in flying-hour programs (e.g., ready alert vs. regular status) are likely to be even more significant.

With regard to cost and selection control, the best design would cluster the maintenance documents by aircraft tail number since this is the manner in which they are filed at base level. Any other arrangement would involve the field workers in sample selection or require much additional time in constructing a sampling frame. By designating the sampling unit as all maintenance performed on a given aircraft within a given time interval, sample selection can be accomplished entirely at SAC Headquarters.

Since little was known about the magnitude of variance that could be expected, the design approach was to decide how much time could

be expected from the base personnel who would be doing the field work, then select as many aircraft as possible. The Cost Analysis Division at SAC Headquarters has at its disposal the part-time services of one man at every SAC base, and it was desired that each man be equally involved in the study. After examining the work involved in searching for the documents, copying labor and parts data, and searching catalogues for parts costs, it was decided to sample one KC-135 on each base over a two-month period (June and July, 1967). Since the number of aircraft varies from base to base, this plan necessitated sampling with unequal probabilities. The design was further complicated by the fact that the proportions of aircraft in the different age groups also vary from base to base.

The final choice was a two-stage design, with the first stage following a procedure first introduced by Goodman and Kish,* and the second stage using simple random selection.

For the first stage, primaries were designated as comprising those aircraft on a given base that belong to the same age group; thus with 31 bases and 8 age groups, there was a maximum of 248 (31 x 8 = 248) clusters. Each cluster was assigned a probability of selection that is roughly proportional to the number of aircraft therein (exact proportionality was precluded since the total aircraft per base varied). The next step was to construct 21 "acceptable" samples such that each sample contained one primary from each base and at least one primary from each age group; the samples were simultaneously assigned probabilities such that if one adds up the probabilities of all samples in which any particular primary appears, the sum will equal the probability originally assigned that primary. Finally, one of the samples was randomly chosen with probability as assigned.

In the second stage, one aircraft was chosen at random from each selected primary.

The overall effect of both stages was to select a sample that is stratified according to base and "controlled" by age group, while giving

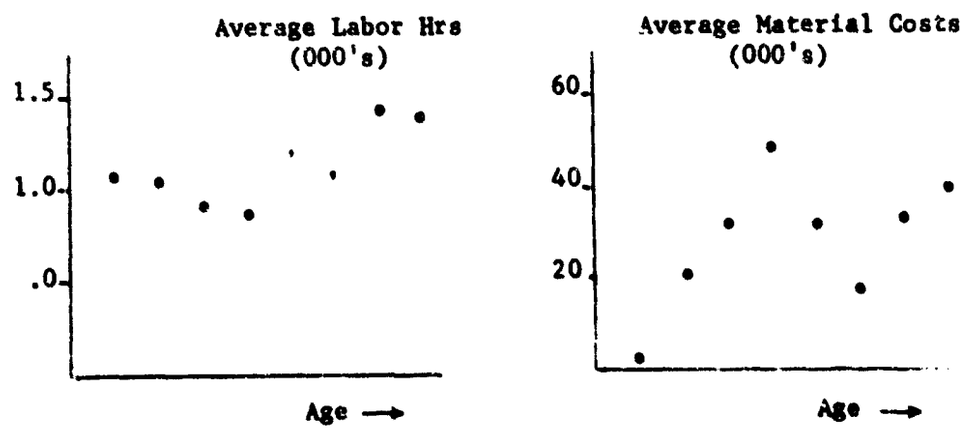
---

*Goodman and Kish, "Controlled Selection--A Technique in Probability Sampling," Journal of the American Statistical Association, Vol. 45, pp. 350-372. Also see Kish, Survey Sampling, 1965, pp. 488-496.

all aircraft approximately equal selection probabilities (the probabilities varied from about .03 to .06). The price of having such a controlled sample in an unbalanced population is that there is no unbiased estimator of sampling variance. However, a weighted estimator for variance is available that leads to overestimation, which is less objectionable than underestimation. Since any estimate of variance is itself subject to sampling error, the bias may not be too important. In any case, the bias would only be associated with the sample's first stage, from which sampling error should be small compared to that from the second stage.

In analyzing the data, separate estimates were made for each age group, and the resulting group means were subjected to regression analysis using age in years as the independent variable. The use of group means instead of the raw data was necessary in order to (1) give each age group equal weight in the regression (sample aircraft were unevenly allocated among age groups) and (2) to accommodate the stratification and probability aspects of the sampling--that is, to help dampen other sources of variability and reveal any age-related behavior. Since the age-group means were derived from samples of various sizes, the usual assumption of equal variance along the regression line was clearly violated; the implication is loss in efficiency in obtaining the least-squares fit.

Some initial results are shown below:



The labor hours data conformed to the general hypothesis that maintenance increases with aircraft age and, when fitted in least-squares fashion to a parabolic curve, survived an F-test at .95 confidence (the F-test was a useful bench-mark despite violation of some underlying assumptions); analysis of material costs did not fare well at this level of aggregation. Subsequent examination entailed a closer look at individual sample aircraft and a distribution of labor hours and material according to maintenance shops. The results were presented at the March 1968 OSD Cost Research Symposium.*

The good performance of the labor hours regression is curiously inconsistent with the very large variability of the raw data within age-groups. At least part of this contradiction can be explained by the efficiency of the sample design; the design should have provided broader representation than would be expected from, say, simple random sampling, which is the usual data collection technique for regression analysis. This illustrates one of the several considerations surrounding data collection that are discussed in the next section on estimating relationships.

ESTIMATING RELATIONSHIPS AND SAMPLE DESIGN

Very little attention in statistical literature is addressed explicitly to the use of sampled data in regression analysis, a technique often used to derive estimating relationships for military cost analysis. There exists, in estimating relationship studies, the implied assumption that the data base constitutes a sample of some larger population (unless the regression is simply intended to describe a particular set of points), and the main concern is whether that sample is representative. Moreover, it is simple random sampling that is implied; the more complicated designs (stratification, unequal probabilities, etc.) are ignored because they are not generally used to build data bases. The intent is now to suggest how these designs might be so used in connection with least squares simple linear regression. The presentation can be made clearer by establishing a conceptual scheme within which data collection can be described. Accordingly, the data collection process will be divided into three phases:

---

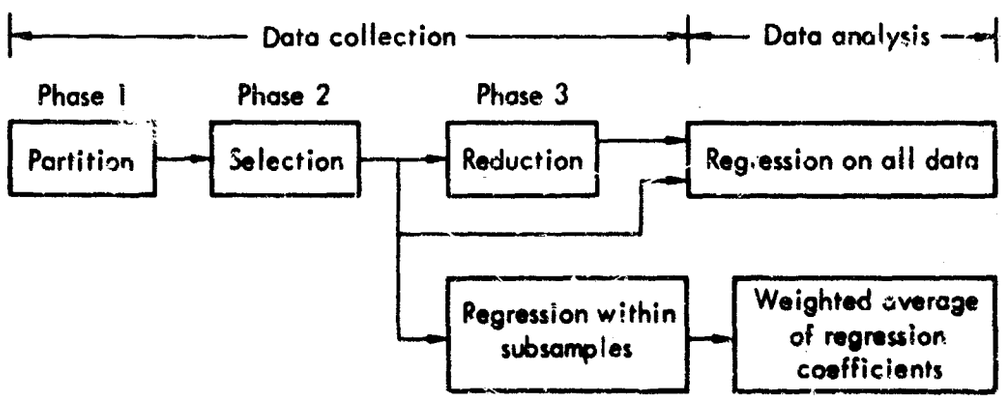
*Jean Mullery, Aircraft Maintenance Cost Research, KC-135, Directorate of Budget, Headquarters Strategic Air Command.

(1) Partitioning the population. The population is divided into sub-populations that may be treated either as strata or clusters.

(2) Data selection. This phase includes any methods of determining what data will fall into the sample. The data from any given sub-population will comprise a sub-sample. In clustering, each sub-sample would thus include the entire sub-population, whereas only a portion would be included with stratification.

(3) Data reduction. The data in each sub-sample are reduced to a single mean value, using some estimator (simple mean, ratio estimator, etc.): the data base now consists of one value per stratum, or one value per cluster.

Regression can be performed on the data base either after the second phase (eliminating reduction) or after the third phase. If regression is performed after the second phase, there are two alternatives available: (1) a single regression on all data, or (2) weighted averages of regression coefficients calculated separately from each sub-sample. The following flow chart characterizes the total process of data collection and subsequent data analysis:



There appear to be two basic motives for using reduced data rather than the original sample: (1) to adjust for unequal sample sizes in the various sub-populations, and (2) to utilize the special estimators* for increased accuracy.

---

*The ratio and regression estimators will be referred to collectively as the "special" estimators so as to avoid confusion with the use of regression to develop forecasting relationships.

If sample sizes within sub-populations are unequal, and it is desired that all sub-populations be of equal importance in determining the regression line, a sort of "weighted" regression is produced by using reduced data; each sub-population is represented by the same number of data points, namely, one. However, if it is preferable to weight on the basis of individual observations rather than sub-populations, the data should not be reduced.

If the opportunity should present itself, it would seem prudent to use one of the special estimators of the unequal probability estimator to reduce the data within each sub-population. These estimators would provide more accurate estimates of the true sub-population means, μ_i, hence lead to more accurate regression estimates. However, this accuracy is gained at the cost of using some auxiliary variable, and it might be preferable to use this variable as a second independent variable in the regression equation. The objective side of deciding which way the extra variable should be used involves the usual cost-precision trade-off (which use will provide greater precision for a given cost?). On the subjective side, the decision might be governed by whether the extra variable can appropriately be specified in the estimating relationship; a variable might be closely correlated with the independent variable but still be ruled out of the regression model because there is no logical causal relationship, or because its future behavior is as doubtful as the dependent variable. In either of these cases, the extra variable could be suitably used as an auxiliary variable in a special estimator or unequal probability estimator. When using the special estimators, each sample point will contain three kinds of observations: one each for the dependent variable, the independent variable, and the auxiliary variable. In unequal probability sampling, only the independent and dependent variable will be observed since values for the auxiliary variable are known prior to sample selection.

The sampling techniques that have been discussed in this paper can be categorized into the three phases as follows:*

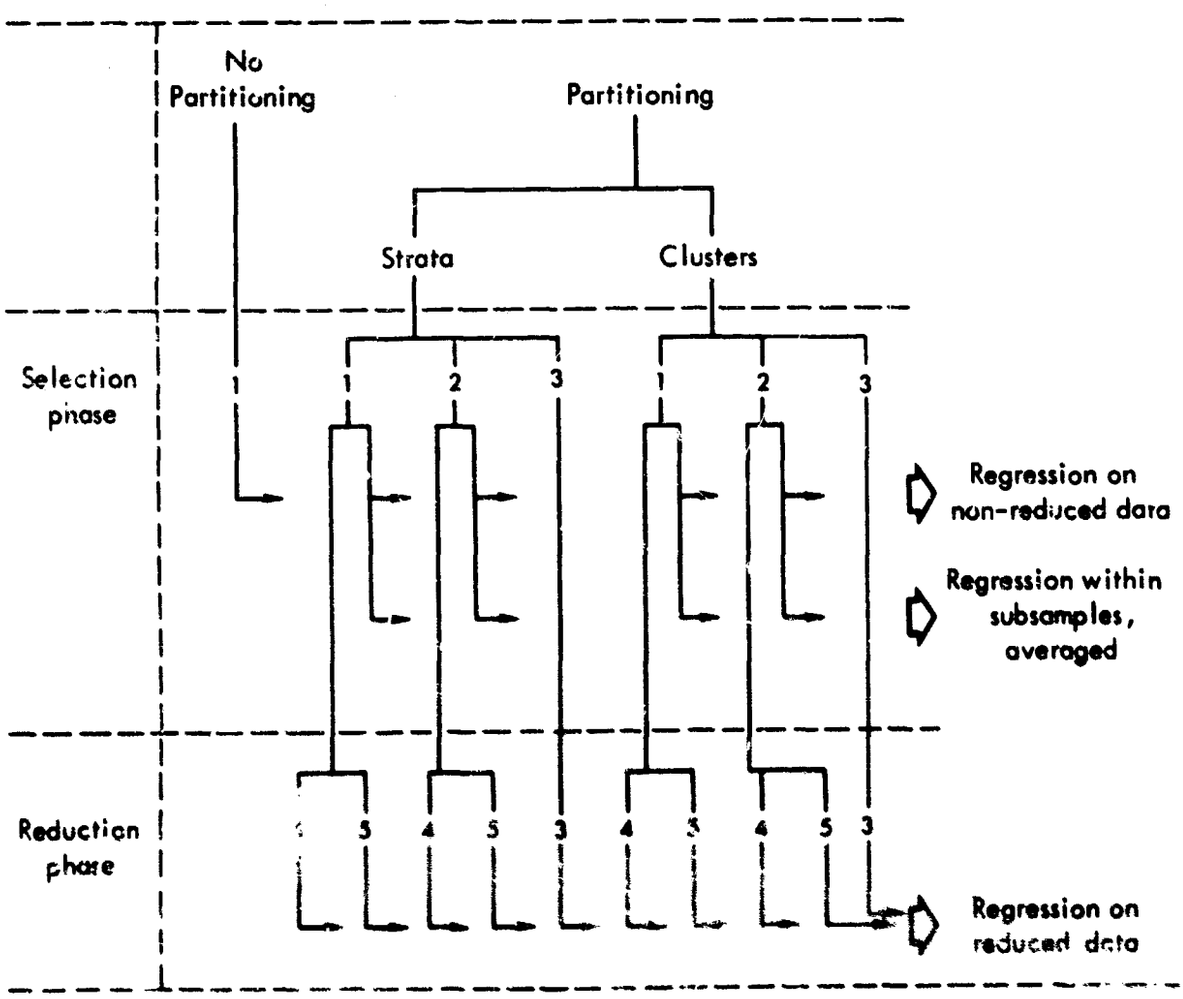
---

*Note that in using this manner of classification, the techniques of stratification and clustering include only the act of population partitioning; the functions of sample point selection and estimation of the population mean fall into the second and third phases.

Partitioning	Stratification Clustering
Selection	Simple random sampling Systematic random sampling Sampling with unequal probabilities
Reduction	Simple mean Special estimators (ratio and regression)

Sampling with unequal probabilities falls into two categories because selection according to this procedure requires the subsequent use of the unequal probability estimator. Sub-sampling was omitted from the list since it is really a hybrid of the other designs.

Below is a schematic diagram of the full set of feasible designs that can be put together from these techniques.



The selection and reduction techniques are numbered for simplicity.

(1) Simple random sampling.

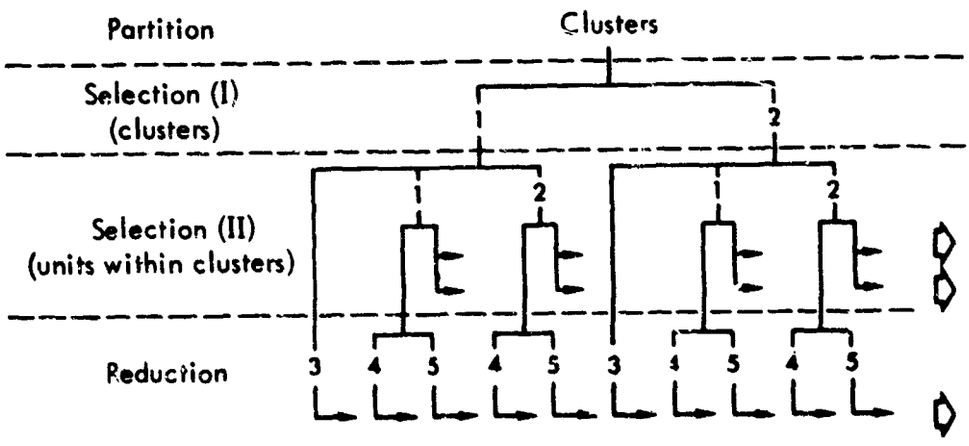
(2) Systematic random sampling.

(3) Unequal probabilities.

(4) Simple mean.

(5) Special estimators.

With no partitioning, there is no alternative to simple random sampling and regre[illegible] on the non-reduced data. Partitioning, on the other hand, allows 18 different basic designs, i.e., there are 18 paths by which the final regression analysis can be reached. Any other design would essentially be an extension of those above. For example, a schematic for collection procedures based on sub-sampling indicates that there is simply a replication of the selection phase:



The foregoing has provided a rather cursory treatment of the preparation of estimating relationships from sampled data. If data are gathered by simple random sampling, subsequent regression analysis is straightforward. More complex schemes lead to difficulties in interpreting regression results. For example, unequal probability sampling will usually lead to underestimated prediction intervals. Stratification will sometimes produce biased regression coefficients. These problems fall into the general area of analytic surveys and are currently being addressed in a peripheral way by such men as Hartley, Kish, and Konijn.

VI. SURVEY PROCEDURE

The preceding pages have provided a summary of those aspects of probability sampling that relate to the design of sample selection procedures. Some peripheral topics, such as questionnaire design and training of field workers, have been ignored as outside the scope of the paper but can be found in such texts as Stephan and McCarthy; quota sampling, a widely used non-probability method, is discussed also in Stephan and McCarthy, and in Kish.

The following generalized chronology of a sample survey is intended to "wrap things up." These steps amount to formalization of the typical decisionmaking process, but conscious observance of them is essential to the mechanics of a valid survey, thereby forcing a rational approach to the analysis.

FORMULATE THE PROBLEM

The first and most important step is to identify the objectives in a rather formalized way so that any subsequent planning alternative can be clearly evaluated with respect to its contribution to those objectives. The analyst is not merely seeking information; he is seeking information that will eventually become part of the basis for some specific decision or class of decisions. It would be well to itemize the objectives and, as far as possible, to model the eventual decision process. Where the survey is part of a group effort, it is equally important to clarify each person's role and to establish a consensus of group objectives. Having defined the problem, the analyst should refer back to it often to avoid becoming over-engrossed on the details of planning and administration.

DEFINE THE POPULATION

The objectives of the investigation determine the population from which information is desired--the target population. The target population is often different from that actually sampled. Although careful planning will tend to eliminate this difference, there are some situations

where the discrepancy simply cannot be physically or economically resolved. The obvious example is the use of historical data in planning for the future. Another example is the exclusion of portions of the population that are too inconvenient to sample. Upper limits can sometimes be computed for the bias introduced, but usually some judgment must be exercised to evaluate the extent to which the sampled population mirrors the target population. This judgment should be documented in the form of a list of assumptions, disclaimers, and uses for which the survey results are appropriate.

SPECIFY PRECISION

The specification of desired precision is an important first step in the design of a survey, although this specification may be simply to obtain the greatest precision for a given budget. In any case, it would seem prudent to examine the survey objectives with respect to the accuracy required in the estimates. If the estimates are to be the bases for comparisons, it makes little sense for them to have greater precision than the standards against which they are compared. Some studies are so heavily burdened with non-statistical uncertainty (e.g., poorly documented data or requirements uncertainty) that high precision may be superfluous.

In complex efforts, such as large models that require partitioning into several sub-models, it would be well for the analysts concerned to discuss together the precision of the various components with respect to (1) the ultimate use of the model, (2) the interrelationship of estimates within the model, (3) the maximum attainable precision for the various estimates, and (4) budget and time constraints. The logical time for such discussion would be after the model has been designed and preliminary investigation of the various subject areas has been accomplished.

CONSTRUCT A FRAME

To construct a sampling frame is to divide the population into sampling units (cluster, strata, and/or simple population units), such that every element of the population belongs to one and only one unit.

The frame is an ordering scheme, or list, that facilitates consistent and unbiased selection of sampling units from the population. If more than one sample design is under consideration, the frame must be flexible so as to suit any one of them (e.g., the population might be divided into strata and the population units grouped into clusters).

SELECT A SAMPLING PLAN

After a preliminary investigation, the analyst should be able to identify characteristics of the population that can be used to design a sampling procedure that is more efficient (less variance for a given sample size) than a simple random sample. These characteristics should suggest several alternatives. An available auxiliary variable may lend itself either to regression estimation or to sampling with unequal probabilities. The alternatives can usually be narrowed down to one or two by making a priori assumptions about the different sampling variances and costs. It may be necessary to make the final choice on the basis of a pre-test that would try out the various plans on a small scale.

CONDUCT FIELD WORK

There is little to be said here, provided the planning has been carefully done. However, if the analyst is not doing his own field work, there should be provisions made to check the quality of the data as soon as it starts coming in. In any case, the analyst should do his own sample point selection; the field worker should be concerned only with collection. There should also be a procedure drawn up to handle non-response, the failure of some selected sample point to be available for sampling.

SUMMARY, ANALYSIS, AND DOCUMENTATION

The data should be examined for erroneous observations, and the estimates calculated. The sampling error should also be calculated, and the sampling procedure summarized.

As an aid to future surveys, it is useful to make a detailed summary of the sampling procedure, including costs that were encountered,

peculiar sampling problems, and characteristics of the population, such as within-strata variances. These might help later surveys by giving more confidence to a priori assumptions and eliminating the need for pre-tests.

Appendix

ESTIMATORS

This appendix gathers together the estimators of means and variances along with formulas for determining sample allocation. The various designs are treated in the same order as in Sections III and IV.

STRATIFIED SAMPLING

The population mean is estimated by $\bar{X}_{st}$, the weighted average of the stratum means:

$$\bar{X}_{st} = \sum^{N}\left(\frac{M_i}{N\bar{M}}\right)\bar{X}_i$$

N = number of strata
M_i = sub-population size of ith stratum
$\bar{X}_i$ = sample mean from i^{th} stratum
$\bar{M}$ = average sub-population size

The estimate for sampling variance of $\bar{X}_{st}$ is also a "weighted" average:

$$S^2_{\bar{x}_{st}} = \sum^{N}\left(\frac{M_i}{N\bar{M}}\right)^2 S^2_{\bar{x}_i} \quad , \quad S^2_{\bar{x}_i} = \text{sampling variance from } i^{th} \text{ stratum}$$

$$= \left(1-f_i\right)\frac{\sum^{m_i}\left(X_{ij}-\bar{X}_i\right)^2}{m_i\left(m_i-1\right)} \quad , \quad f_i = \frac{m_i}{M_i}$$

m_i = sample size from i^{th} stratum.

The total sample may be divided among strata according to the proportional allocation scheme:

$$m_i \propto M_i$$

If estimates are available for the variances within each stratum (σ_i^2) and the costs of sampling from each stratum (c_i), optimal allocation may be used:

$$m_i \propto \frac{M_i \sigma_i}{\sqrt{c_i}}$$

CLUSTER SAMPLING

Given that n primaries are selected, each containing M_i secondaries, the estimator of the population mean of secondaries is given by:

$$\bar{X}_{cl} = \frac{\sum\sum^{nM} X_{ij}}{nM} = \frac{\sum^{n} \bar{X}_i}{n}$$

$\sum\sum^{n M_i} X_{ij}$ = sum of all secondaries sampled

$\bar{X}_i$ = mean of the ith cluster.

The variance of this estimate is generated solely by differences between clusters. The estimate of sampling variance is

$$S^2_{\bar{x}_{cl}} = \frac{S^2_B}{n}(1-f) \; ; \; f = \frac{n}{N}$$

where S^2_B is the estimator for variance between the cluster means, $\bar{X}_i$:

$$S^2_B = \frac{(\bar{X}_i - \bar{X}_{cl})^2}{n-1}$$

The general criterion for choosing cluster size is to minimize the product of sampling variance times total cost (both of which vary, depending on cluster size:

$$\text{minimize } (V)(C) = \frac{S^2_B}{n}(n \cdot c) = S^2_B \cdot c \quad \text{(ignoring fpc)}$$

S^2_B = variance among cluster means

c = cost of observing one cluster

Unequal Cluster Sizes

In some situations the number of secondaries per cluster, M_i, may itself be variable. This complicates the theory and the estimators and

introduces bias of the type associated with ratio estimators. The estimator for the population mean is:

$$\bar{X}_{cl} = \frac{\sum^{n} \sum^{M_i} X_{ij}}{\sum^{n} M_i}$$

Sampling variance is estimated by

$$S^2_{\bar{x}_{cl}} = \frac{1-f}{n(\bar{M})^2(n-1)} \quad \sum^{n} X_i^2 + X_{cl}^2 \sum^{n} M_i^2 - 2\bar{X}_{cl} \sum^{n} M_i X_i$$

$$f = \frac{n}{N}$$

$$X_i = \sum^{M_i} X_{ij} = \text{total of } i^{th} \text{ cluster}$$

$$\bar{M} = \text{average cluster size}$$

Since $\bar{X}$ is based on the ratio of two variables (X_i and M_i), these estimators have a bias that increases as the variability of the M_i increases. As a rule of thumb, the bias may be overlooked when the coefficient of variation for M_i is less than .2; i.e.,

$$\frac{\sigma^2_{M_i}}{\bar{M}} < .2$$

Otherwise, the sample size should be large (see Kish, page 276).

SUB-SAMPLING

The population mean is estimated as the average of primary means:

$$\bar{X}_{ss} = \frac{1}{n} \sum^{n} \bar{X}_i$$

The sampling variance has two components, one representing variation

<u>between</u> primaries and one representing variation among secondaries <u>within</u> primaries. The estimator is:

$$S^2_{\bar{X}_{ss}} = \frac{S^2_B}{n}(1-f_1) + \frac{S^2_w}{nm}(1-f_2)f_1$$

$$S^2_B = \frac{1}{n-1}\,\Sigma(\bar{X}_i-\bar{X})^2$$

$$S^2_w = \frac{1}{n(m-1)}\,\Sigma\Sigma^{nm}(X_{ij}-\bar{X}_i)^2$$

$$f_1 = \frac{n}{N}\ ,\ f_2 = \frac{m}{M}$$

X_{ij} is the jth sample unit in the ith cluster.

The optimal number of secondaries to select from each primary is given by:

$$m = \sqrt{\frac{S^2_w}{S^2_B - \frac{S^2_w}{M}} \cdot \frac{c_1}{c_2}}$$

c_1 = cost of sampling one primary ("fixed" cost)

c_2 = cost of sampling each secondary ("variable" cost)

If the value computed is equal to 1 or less, then m = 1; if the value is greater than M, one-stage (cluster) sampling should be used.

The determination of n, the number of primaries selected, depends on whether total cost or precision is to be held constant. In the latter case, n is found by solving the sampling variance formula. If total cost, C, is fixed, the following formula is solved for n:

$$C = nc_1 + nmc_2$$

The use of the foregoing methods for determining n and m requires preliminary estimates of c_1, c_2, S_B^2, and S_w^2. For this purpose, these estimates do not require great precision because the sampling variance is not highly sensitive to the choice of m. It is usually easier to estimate ratios c_1/c_2 and S_w^2/S_B^2, in which case tables are available to aid the evaluation of m (see Cochran, page 282).

Unequal Cluster Sizes

As in simple cluster sampling, sub-sampling becomes more difficult if cluster size, M_i, is variable. An estimator of the population mean is:

$$\bar{X}_{ss} = \frac{\sum^{n} M_i \bar{X}_i}{\sum^{n} M_i}$$

An estimator of the sampling variance is

$$S^2_{\bar{X}_{ss}} = \left(\frac{1-f_1}{n}\right)\sum^{n}\left[\left(\frac{M_i}{\bar{M}}\right)^2 \frac{\left(\bar{X}_i-\bar{X}\right)^2}{n-1}\right] + \frac{f_1}{n^2}\sum^{n}\left[\left(\frac{M_i}{\bar{M}}\right)^2 \frac{\sum^{m_i}\left(X_{ij}-\bar{X}_i\right)^2}{m_i\left(m_i-1\right)}\left(1-f_{2i}\right)\right]$$

$$f_1 = \frac{n}{N}, \quad f_{2i} = \frac{m_i}{M_i}$$

$\bar{M}$ = average cluster size.

The bias in these estimators again relates to the variability of M_i, and can be made negligible by making n large (see Cochran, page 300).

SYSTEMATIC SAMPLING

The estimator of $\bar{X}$ in systematic sampling is the same as for simple random sampling:

$$\bar{X} = \frac{\sum^{n} X_i}{n}$$

There is no single reliable method of estimating the sampling variance because so much depends on the way the population is listed. Variance formulas for specific kinds of populations can be found in sampling texts.

RATIO ESTIMATOR

The ratio estimate of the population mean of the X is given by:

$$\bar{X}_R = \left(\frac{\sum^n X_i}{\sum^n W_i}\right)\mu_w = \left(\frac{\bar{X}}{\bar{W}}\right)\mu_w$$

For large sample sizes, the approximate sampling variance is estimated by:

$$S^2_{\bar{X}_R} \doteq \frac{1-f}{n}\left[\frac{\sum^n\left[X_i - \frac{\bar{X}}{\bar{W}} W_i\right]^2}{n-1}\right]$$

REGRESSION ESTIMATOR

The regression estimate of the population mean of the X variable is given by:

$$\bar{X}_r = \bar{X} + b(\mu_w - \bar{W})$$

The least-squares estimator for b is:

$$b = \frac{\sum^n (X_i-\bar{X})(W_i-\bar{W})}{\sum^n (W_i-\bar{W})^2}$$

The sampling variance for large samples is estimated by:

$$S^2_{\bar{X}_r} = \frac{1-f}{n(n-2)}\sum^n [(X_i-\bar{X})-b(W_i-\bar{W})]^2$$

UNEQUAL PROBABILITY SAMPLING

The estimate of the population mean is provided by:

$$\bar{X}_P = \frac{1}{nN} \sum^{n} \frac{X_i}{P_i}$$

where P_i is the probability of selecting X_i. The variance for replacement sampling is estimated by:

$$S^2_{\bar{X}_{PR}} = \frac{1}{n(n-1)} \sum^{n} \left(\frac{X_i}{P_i} - \frac{1}{n} \sum \frac{X_i}{P_i}\right)^2$$

The variance for non-replacement sampling can be estimated by:

$$S^2_{\bar{X}_{NPR}} = \sum_{i>j} \left(P_i P_j - P_{ij}\right)\left(\frac{X_i}{P_i} - \frac{X_j}{P_j}\right)^2$$

where P_i and P_j are the respective probabilities with which Y_i and Y_j were included in the sample, and P_{ij} is the joint probability with which both Y_i and Y_j were included. This estimator is unbiased if P_{ij} is non-zero for all i and j.

Non-replacement sampling offers the same kind of efficiency advantages for unequal probability sampling as in sampling with equal probabilities, and is therefore widely used. However, while the benefits with equal probabilities are reflected in the finite population correction factor (page 23), the corresponding theory for unequal probabilities finds no such simple expression. A "unified," or all-inclusive theory has not been forthcoming; current literature is generally focused on various aspects of three related problems: (1) the control of the P_i through selection procedures, (2) the control of the P_{ij} through selection procedures, and (3) the conditions under which estimates behave according to the central limit theorem (i.e., tend to be normally distributed).

BIBLIOGRAPHY

The following list of references is more representative than comprehensive. The brief appraisals may be helpful in guiding the reader to further study. An extensive bibliography is available in Kish.

BOOKS

Cochran, W. G., Sampling Techniques, 2d ed., John Wiley and Sons, New York, 1963.

Cochran develops sampling theory with considerable rigor and completeness. The book is well documented, with references listed at the end of each chapter, but it is not particularly concerned with the operational aspects of conducting a survey.

Deming, W. E., Sample Design in Business Research, John Wiley and Sons, New York, 1960.

Deming is a consultant in statistical surveys with considerable experience in both business and government. His book is, accordingly, strongly oriented toward implementation of survey designs. There is abundant illustrative material, although his emphasis on replicated designs (see page 59 of this Memorandum) may lead to some initial confusion for someone accustomed to the more traditional textbook approach.

Hoel, P. G., Elementary Statistics, John Wiley and Sons, New York, 1954.

_____, Introduction to Mathematical Statistics, 3d ed., John Wiley and Sons, New York, 1962.

Hoel's books are basic statistics texts. The first listed is concise and relatively painless. The second is more rigorous, mathematically.

Jessen, R. J., Statistical Survey Techniques, (unpublished manuscript).

This is an easy-reading exposition of sampling theory and the practical problems of application, but it is still in process and lacks an index. The mimeographed draft is available at the UCLA Book Store in Los Angeles.

Johnston, J., Econometric Methods, McGraw-Hill Book Company, Inc., New York, 1963.

This text is mentioned here as a useful reference for regression analysis.

Kish, L., Survey Sampling, John Wiley and Sons, New York, 1965.

This is a very comprehensive book. Besides general theory and general problems of implementation, there is frequent discussion of special topics such as sources of bias, complex samples, and sampling from imperfect frames. The bibliography is large.

Malinvaud, E., Statistical Methods of Econometrics, Rand McNally, Chicago, 1966.

This text is mentioned here because of its brief treatment of regression computations from grouped data, beginning on page 242.

Slonim, M. J., Sampling in a Nutshell, Simon and Schuster, New York, 1960.

Slonim writes a lively exposition of a usually unentertaining subject. He gives general coverage to the several sampling designs, sacrificing jargon and detail for readability. The book is fast reading and provides a good introduction for someone interested in the possibilities of sampling.

Snedecor, G. W., Statistical Methods, 5th ed., Iowa State College Press, Ames, 1956.

This book is more experimental design than sample theory. However, there is one chapter (written by Cochran) on sampling that offers a concise statement of basic sampling design.

Stephan, F. S., and P. J. McCarthy, Sampling Opinions, John Wiley and Sons, New York, 1958.

This book includes recommendations on questionnaire design, training of field workers, and the conduct of quota sampling.

Yates, F., Sampling Methods for Censuses and Surveys, 3d ed., Charles Griffin and Company, London, 1960.

Yates offers a useful sampling text, including a discussion of analytic surveys, a topic that probably should have received more attention in this Memorandum (see page 59).

ARTICLES ON UNEQUAL PROBABILITY SAMPLING WITH NON-REPLACEMENT

Hansen, M. H. and W. N. Hurwitz, "On the Theory of Sampling From Finite Populations," Annals of Mathematical Statistics, 1943, p. 333.

Hartley, H. E., "Systematic Sampling with Unequal Probabilities and Without Replacement," Journal of the American Statistical Association, 1966, p. 739.

Rao, J. N. K., H. O. Hartley, and W. G. Cochran, "On a Simple Procedure of Unequal Probability Sampling Without Replacement," Journal of the Royal Statistical Society, Series B, 1962, p. 482.

Yates, F., and P. M. Grundy, "Selection Without Replacement From Within Strata With Probability Proportional to Size," Journal of the Royal Statistical Society, Series B, 1953, p. 253.

ARTICLES RELATED TO REGRESSION ON SAMPLE SURVEY DATA

Cramer, J. S., "Efficient Grouping, Regression and Correlation in Engel Curve Analysis," Journal of the American Statistical Association, 1966, p. 391.

Konijn, H. S., "Regression Analysis in Sample Surveys," Journal of the American Statistical Society, 1963, p. 590.

Prais, S. J., and J. Aitchinson, "The Grouping of Observations in Regression Analysis," Review of the International Statistical Institute, 1954, p. 1.

MISCELLANEOUS ARTICLES

Hartley, H. O., "Analytic Studies of Survey Data," Proceedings, Social Statistics Section, American Statistical Association, 1959, p. 146.

Fan, C. T., M. E. Muller, and I. Rezucha, "Development of Sampling Plans by Using Sequential (Item by Item) Selection Techniques and Digital Computers," Journal of the American Statistical Association, 1962, p. 387.

Godambe, V. P., "A Review of Contributions Towards a Unified Theory of Sampling From Finite Populations, " Review of the International Statistical Institute, 1953, p. 242.

McCarthy, P. J. "Replication, An Approach to the Analysis of Data from Complex Surveys," USDHEW, Public Health Service, National Center for Health Statistics, Series 2, Number 14, 1966.

Raj. D., "On the Relative Accuracy of Some Sampling Techniques," Journal of the American Statistical Association, 1958, p. 98.

United Nations, Statistical Office, "Recommendations for the Preparation of Sample Survey Reports (Provisional Issue)," Statistical Papers, Series C, Number 1, Revision 2, 1964.

Zarkovich, S. S., "On the Efficiency of Sampling with Varying Probabilities and the Selection of Units with Replacement," Metrika, 1960, p. 53.

DOCUMENT CONTROL DATA

1. ORIGINATING ACTIVITY THE RAND CORPORATION	2a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED 2b. GROUP
3. REPORT TITLE SAMPLING METHOD: SUGGESTIONS FOR MILITARY COST ANALYSTS	
4. AUTHOR(S) (Last name, first name, initial) Sumner, G. C.	
5. REPORT DATE October 1968	6a. TOTAL No. OF PAGES 92 / 6b. No. OF REFS. ---
7. CONTRACT OR GRANT No. F44620-67-C-0045	8. ORIGINATOR'S REPORT No. RM-5779-PR
9a. AVAILABILITY/LIMITATION NOTICES DDC-1	9b. SPONSORING AGENCY United States Air Force Project RAND

10. ABSTRACT

An examination of aspects of probability sampling and application of the sampling method to cost analysis. The relative merits of the sample as a means of data collection, sampling theory, and the more basic techniques of sample design and estimation are discussed. Sampling offers an approach to the data quality problem that is usually cheaper, faster, and more flexible than attempts to modify existing massive data collection systems. Basic tools that the analyst has at his disposal are stratification, clustering, subsampling, systematic sampling, ratio and regression estimators, and sampling with unequal probabilities. The study was prepared to provide an introdution to sampling methods: There is sufficient coverage to guide simple survey investigations, and attention is given throughout to the use of cost-effectiveness criteria in choosing among alternative sampling plans.

11. KEY WORDS

Cost analysis
Data processing
Statistical methods and processes
Curve fitting
Probability
Bibliography