THE JOHNS HOPKINS UNIVERSITY
SCHOOL OF HYGIENE AND PUBLIC HEALTH
INSTITUTE FOR INTERNATIONAL PROGRAMS

Industrial Sampling Plans:

Prospects for Public Health Applications

William A. Reinke, Ph.D.

March 1988

OCCASIONAL PAPER NO. 2

Industrial Sampling Plans:

Prospects for Public Health Applications

William A. Reinke, Ph.D.

March 1988

The Johns Hopkins University
School of Hygiene and Public Health
Institute for International Programs
615 North Wolfe Street
Baltimore, Maryland 21205
USA

Preface

The **Occasional Papers Series** of the Johns Hopkins Institute for International Programs was established as a means of disseminating information on advances in international health. The Institute is currently conducting research and providing technical support to child survival programs throughout the world, in cooperation with the United States Agency for International Development.

The Institute for International Programs, as part of the Johns Hopkins School of Public Health, brings together leading Hopkins experts in biomedical, environmental, social science, and health management fields. When the Institute was founded in 1985, it identified three key child survival activities -- immunization, vitamin A supplementation, and program evaluation -- and established a faculty working group to focus on each.

This distinguished group of Hopkins professionals, together with collaborators from established research centers and other major academic institutions throughout the world, provides a comprehensive approach to international development.

IIP Occasional Paper No. 2
March 1988

William A. Reinke, Ph.D., Author
Denise Harrison, Design and Production
Johns Hopkins University Publications, Cover Art
Johns Hopkins Printing Services, Printing

The research reported on in this paper was funded by the United States Agency for International Development under Cooperative Agreement #DPE-5951-A-00-5051-00.

Contents

Introduction..... 1

Essential Factors in the Decision Process..... 2
Universe Characteristics and Size..... 2
Parameters and Standards of Acceptability..... 2
Nature of Variability..... 2
Representativeness..... 3
Information Costs..... 3
Cost and Quality of Resulting Decisions..... 4

Review of Industrial Sampling Plans..... 5
Basic Lot Quality Assurance..... 5
Double and Multiple Sampling..... 6
Rectification..... 9
AOQL Plans with Continuous Production..... 11
Reduced and Tightened Inspection..... 11
Control Charts..... 12
Moving Averages and Exponential Smoothing..... 14
Factorial Designs..... 14

Specific Applications..... 15
Cluster Sampling, Lot Quality Assurance, Sentinel Surveillance..... 16
Basic Sampling Plans: Attributes versus Continuous Variables..... 19
Investigation of Breastfeeding through Double Sampling..... 20
Monitoring Services Delivery through Control Charts..... 21
Selective Supervision with Reduced and Tightened Inspection..... 22

Further Steps Recommended..... 23
Likely Field Tests..... 24
Closer Scrutiny of Patterns of Variation..... 24
Quick-and-Dirty Approaches..... 25
Efficient Cluster Designs..... 25
Pattern Recognition..... 27

INDUSTRIAL SAMPLING PLANS: PROSPECTS FOR PUBLIC HEALTH APPLICATIONS

William A. Reinke, Ph.D.

Regardless of the setting in which sampling procedures are applied, the purpose is the same: to determine at minimum cost the value(s) of specified parameter(s) in a defined universe with sufficient precision to permit appropriate decisions to be made. Sample findings, judged in comparison with predetermined desired states, serve as the basis for the decisions to be made with tolerably small risks for error.

This purpose has been pursued actively in industry throughout most of this century. We recall that the widely used "t" test in statistics was originally formulated by an employee of the Guiness Brewery in connection with his work with small samples. Efforts to develop efficient sampling schemes were intensified as a result of the military effort during World War II and have continued, notably in the field of industrial quality control, ever since.

Considering the broad range of industrial sampling plans now in use for purposes fundamentally the same as those considered in health matters, it is not unreasonable to expect that the fruits of industrial statistics might be harvested in biostatistics as well. Indeed, some cross-fertilization has already occurred, but not to the extent possible or desirable. The intent of this paper is to explore many of the more promising possibilities.

Introduction

Starting with the stated purpose of sampling, we examine in the next section the individual factors to be considered in designing a sampling plan for achieving the objectives. For example, if conclusions are to be reached concerning certain segments of a population, as well as the overall population, this must be reflected in the sample design. The concerned factors are used in the following section to form a classification framework within which relevant industrial sampling plans are presented. Practical examples of their application are given to highlight attractive features and limitations of the individual plans. We then consider in greater detail ways in which selected sampling schemes are recommended to be applied to certain important health issues. We also describe modifications to be made in existing sampling procedures to make them more applicable to the particular field conditions of interest. The final section outlines needs for additional work on the subject and presents a suggested protocol for carrying out further studies.

Essential Factors in the Decision Process

Universe Characteristics and Size

Possible actions following sampling relate to a defined universe. It is necessary, therefore, to consider the size and characteristics of the population and its components. Geopolitical or other management units that serve as the basis for action must be specified. To illustrate, are conclusions to be drawn about specific villages, or only at the district level? Are all workers under one supervisor to be considered as a unit? Other bases for stratification should also be enunciated. For example, if comparisons in behavior are to be made by level of educational attainment, it may be necessary to employ stratified sampling in order to ensure enough observations among the highly educated. Although universe size is usually not a critical factor, interest in particular segments of the population may result in groups small enough that their size should be noted in determination of sample size.

Parameters and Standards of Acceptability

Universe parameters to be investigated must obviously be defined unambiguously in advance. The definition must permit measurement or classification into mutually exclusive categories. Simultaneous interest in multiple parameters will lead to consideration of ways in which the several items of information can be obtained and analyzed most efficiently.

Considering the decision making intent of sampling, estimates derived regarding parameters should provide the basis for action. This requires comparison in one of three respects: (1) with predetermined standards of acceptability (e.g., at least 70 percent of the target population immunized); (2) among units of interest (e.g., determination that one worker's performance is significantly inferior to that of others); or (3) with earlier conditions (e.g., assessment of improvements over time).

Nature of Variability

The efficiency of the sampling procedures employed and the validity of conclusions reached depend upon patterns of variation present in the universe. In particular, if cluster sampling is to be used in deriving conclusions about a large population, systematic differences between clusters should be minimized, i.e., heterogeneity within clusters is desirable. On the other hand, if decisions are to be made about

individual subgroups, or clusters, homogeneity within subgroups is to be sought. It is desirable insofar as possible, therefore, to determine in advance relative magnitudes of variation "between" and "within" population segments, and to take cognizance of such patterns in designing sampling plans.

Statistical analyses typically assume "normal" patterns of variation. Bimodal conditions, or other departures from normality, can thus produce misleading results and should be avoided if possible.

Representativeness

Sampling necessarily produces only estimates of the universe parameters of interest. Errors in the estimates relate to the magnitude of variability just considered and can be assessed if a representative sample has been obtained, meaning that principles of randomness have been followed. Since matters of cost and administrative convenience, as well as statistical rigor, enter into sampling decisions, representativeness is sometimes sacrificed, and consequent ramifications must be contemplated. Specifically, it may not be feasible to select villages or other population subgroups randomly. Likewise, if observations within subgroups are to be made in clusters, rather than through random selection, the effects should be understood and proper allowances made in the analysis.

Information Costs

Assuming that the sampling design embodies adequate representativeness, additional information tends to generate greater precision, but at a cost. While each additional observation may incur a cost that is nearly constant, there are diminishing returns in precision. This is because precision is a function of the square root of the sample size, not the sample size itself. Thus, for example, if a random sample of 25 deliveries produces estimates of average birthweight that are within 150 grams of the universe average, an additional 25 observations will reduce the maximum error by 44 grams, but the next 25 observations will increase precision by only 19 grams.

For this reason, a commonly applied rule-of-thumb in industry is that samples of 30-50 are generally adequate; additional sampling effort produces little new information relative to the costs incurred. The rule is reasonable when the normal probability distribution is operative. Thus, if variation among clusters follows the normal curve, then merit can be seen in the selection of 30 clusters, as is frequently

done in EPI surveys. It is worth emphasizing, however, that sampling for attributes (e.g., condition of "low birthweight", rather than the weight itself) requires substantially more information to achieve acceptable precision because each observation provides less detail. Whether the cost and other advantages from sacrificing detail outweigh the corresponding increase in sample size necessary is a matter for determination in individual cases.

Cluster sampling is a well-known example of the trade-off between cost and precision. The "design" effect introduced by the departure from strict randomness increases sampling error and hence the sample size needed to achieve a specified level of precision. The reduced cost per observation with clustering must be judged against the required increase in the number of observations.

Cost and Quality of Resulting Decisions

Presumably, more accurate estimates of universe parameters lead to better decisions. This should not be accepted uncritically, however. If the same action would be taken, regardless of whether forty percent, or only ten percent, of mothers know the proper treatment for diarrhea, then increased precision is not likely to change the decision and therefore its quality.

The cost of sampling and consequent precision must be weighed against the costs of two possible types of erroneous decisions. To take "corrective" action erroneously results in so-called Type I error. The risk of Type I error is complementary to Specificity in epidemiologic terms, since we are dealing with false positives. Thus, 90 percent specificity is associated with ten percent risk of Type I error. Failure to take needed action is known as Type II error and is the complement of Sensitivity.

While the exact costs of these errors may be difficult to quantify, their impact must be given explicit attention nevertheless in relation to information costs, both in monetary terms and with respect to time and effort. The requisite explicit attention to Type I and Type II errors is generally accorded through Operating Characteristics (OC) curves which graph the "power" of a particular sampling scheme to detect departures of specified amounts from conditions declared to be acceptable. It is assumed that failures to detect large deviations from standards is more costly than failures to detect less dramatic discrepancies, even if the magnitudes of the costs cannot be measured definitively. In the discussions that follow, we will refer to Type I and Type II error, rather than to the alternative nomenclature of Sensitivity and Specificity. We will then

use OC curves as a basis for judging and comparing various sampling plans.

Review of Industrial Sampling Plans

Industrial sampling plans fall into two general groupings, depending upon the setting in which they are used. When product is organized into discrete batches for assessment of quality, a lot acceptance procedure is applicable. This is distinguished from the monitoring of a continuous operation to make process decisions. To put this in public health terms, decisions on the acceptability of immunization levels achieved in a particular district would call for a lot acceptance sampling plan, whereas an ongoing child weighing program would be more appropriately monitored through process sampling. The two approaches will be discussed separately, and sub-classifications within each category will be identified.

Basic Lot Quality Assurance

The most fundamental lot acceptance procedure selects a sample that is large enough to distinguish between a level of quality explicitly defined to be acceptable and another level that is clearly unacceptable. Thus, two points on the OC curve are highlighted, along with tolerable risks of Type I and Type II error respectively. Conditions of quality intermediate between these two points are necessarily associated with higher risks of Type II error. The minimum single sample size needed to satisfy these requirements is then determined, along with the number of "defects" permissible for acceptance.

This approach has been adopted by health workers under the heading Lot Quality Assurance Sampling (LQAS). Proponents cite the ability to make decisions on the basis of small samples. This has nothing to do with the procedure, however, since it follows the most basic rules of probability theory. Rather, the small sample requirement comes either from the wide differences between what is defined as "acceptable" and "unacceptable", or from the relatively high risks of error that are permitted.

To illustrate, consider the case in which an immunization rate of 80 percent in a target population is considered acceptable. If in a sample of ten children, at least six are found to be immunized, one can regard the population as adequately immunized with less than five percent probability of erroneously flagging for special attention a group that, in fact, meets the 80 percent immunization target. In other words, the procedure to select a random sample of ten children, coupled with d = 4, i.e.,

permitting up to four non-immunized children in the sample, will limit to less than five percent the risk of Type I error.

Only if the immunization rate is 35 percent or below, however, will the procedure give 90 percent assurance of declaring conditions in the population to be unacceptable. For rates between 35 and 80 percent the risk of Type II error is quite high, as can be seen from the OC curve of Figure 1. If a rate as low as 60 percent is unacceptable, the sample size must be 42. To detect the difference between 70 percent and 80 percent immunization coverage requires a sample of 156.

The argument is sometimes made that a program is likely to be active in some communities and not in others; therefore, coverage will be either very high or quite low. Under these circumstances a small sample will be adequate to detect such gross differences. On the other hand, the cluster sampling procedures employed in the EPI program, among others, usually lead to analysis based upon the assumption that variation among clusters follows a continuum characterized by the normal probability curve. It is apparent that some advance knowledge of variability among subgroups is desirable when it comes to selecting an appropriate sampling plan. If differences tend to be bimodal, an LQAS approach based upon small samples should be quite adequate. Evidence of a more nearly normal pattern of variation, however, may signal the need for an alternative procedure.

Double and Multiple Sampling

In practice, of course, it may be difficult to determine in advance patterns of variation among lots, or geographic areas. Even past experience may not be a reliable indicator of conditions to be encountered at future field sites. In this case the investigator faces a dilemma. A small sample will not be definitive in identifying moderate departures from desired conditions. Yet a large sample with be unnecessarily costly in detecting gross departures. Double sampling procedures have been devised to overcome this dilemma. An initial small sample is selected to determine whether a clear-cut decision can be made. If not, the acquisition of additional information is considered to be worthwhile, and a second sample is taken.

Consider, for example, a standard industrial double sampling plan applicable to lot sizes of 500-800 under circumstances in which acceptable lot quality contains 10 percent or fewer defects. Translated into terms of interest to us, we wish to sample a target population of 500-800 to determine whether coverage is significantly below 90 percent.

The plan calls for an initial sample of 25, coupled with acceptance number 5 and rejection number 11. Thus, if five or fewer defects are found, sampling will be discontinued with adequate assurance that quality meets or exceeds minimum standards. Eleven or more defects leads to the definitive conclusion that the product is sub-standard. Results in the range of 6-10 defects are not definitive and call for a second sample of 50. If the combined sample produces at least 11 defects, the lot is to be rejected; otherwise it is to be accepted.

The procedure might be applied to the checking of blood pressure of pregnant women during their first antenatal visit to a clinic. If a random sample of 25 cases shows that the blood pressure was recorded in at least 20 instances, the investigation is discontinued. If 14 or fewer blood pressures are found, corrective action can be initiated immediately without need for further evidence. Otherwise, 50 more records would be sampled.

If the procedure were to be employed over time, the amount of sampling per application would obviously be 25 or 75 cases. As graphed in Figure 2, the average would be closer to 25 whenever quality is decidedly inferior or clearly superior. In less clear-cut situations it would be closer to 75. The average sample size might be near 50; the procedure is more efficient than one involving a single sample of 50, however, for that sample size would be inadequate in some cases and excessive in others. In short, double sampling procedures automatically focus data gathering efforts where they can be most useful, but at some cost in administrative complexity. The procedures are widely applied without difficulty in industry, however.

Considering the desirable features of double sampling, it would seem that triple sampling would be even better, and indeed such plans have been developed and are in use. The ultimate in this regard is Sequential Sampling, which provides for continuous assessment of the need for additional information before making a decision. The decision boundaries are two upward-sloping parallel lines constructed as in Figure 3. Values for their intercepts and slope are determined by the defined levels of acceptable and unacceptable quality, along with the associated risks of Type I and Type II error that are to be tolerated.

The chart on which the boundaries are plotted can also be used to plot sample results in sequence. With each observation one moves one unit along the horizontal axis. Any defect that is found causes a shift of one unit up the vertical axis. As long as the combination of samples selected and defects found places one within the parallel

bands, sampling is continued. Acceptance occurs when one moves beyond the boundary to the right, i.e., when few defects have been found in a sufficiently large sample. On the contrary, a rash of defects encountered in a small sample moves one above the upper band into "reject" territory.

Although the mathematics behind sequential sampling is complex, the resulting linear equations defining the decision boundaries are quite simple and are fixed once the definitions of acceptability and risk are established. Large differences between what is termed acceptable and unacceptable serve to move the intercepts closer together, thereby narrowing the band within which sampling is continued. Strict definitions of acceptability reduce the slope, so that a rash of defects will move one more quickly above the upper limit into the "reject" area. These effects occur automatically, of course, with the relatively simple calculation of the boundary equations.

Consider the case in which children are to be checked for BCG scars. Coverage of 80 percent is considered acceptable, whereas anything below 60 percent is unacceptable. Tolerable risks of Type I and Type II error are set at 10 and 20 percent respectively. Designating the cumulative number of samples taken by n and the number of defects (children without scars) by d, the boundary equations are:

Lower Bound $d = .29n - 1.54$;
Upper Bound $d = .29n + 2.13$.

The equations are plotted in Figure 3 and show, for example, that sampling can be terminated if the first six children selected all have scars. In the illustrative example given in Figure 3, however, the third child was found without a scar and sampling had to continue until nine children had been sampled and eight were found to have scars.

Whenever conditions are either very satisfactory or decidedly unsatisfactory, the sequential sampling approach tends to produce definitive findings promptly and at low cost. Under intermediate conditions, however, sampling could continue indefinitely. If, in the illustrative case, BCG coverage were around 70 percent, one might remain in the "Continue" range of Figure 3, even after fifteen or twenty observations had been made. Moreover, as the sample size grows, the amount of sampling error is reduced and the likelihood of moving outside the boundary lines likewise declines. Some means of truncating the sampling process is therefore desirable.

One basis for truncation is to consider the number of observations required in single sampling. Referring to the

circumstances defined as acceptable (80 percent coverage) an unacceptable (60 percent coverage), along with correspondin risks of Type I and Type II error (10 percent and 2 percent), a single sample of 22 observations would b required. If at most six individuals were unvaccinate conditions would be declared "acceptable." The resultin truncation is shown by dashed lines in Figure 3.

Rectification

Sampling is meant to be an aid to decision making. Th design of the sampling plan depends, therefore, upo available action alternatives and their costs. I particular, a finding of poor lot quality in some industria settings leads to 100 percent inspection of the product t screen out faulty material, rather than outright rejection. This option has led to development of so-called Rectificatio Plans.

Under these plans Type II error is considered to be mor serious than Type I error. Allowing poor quality to g undetected can cause major difficulties later, whereas commission of Type I error only results in making alread good quality even better, though at a cost of unnecessar inspection.

With respect to Type II error, one of two criteria is employed. Some plans specify a limit on Lot Tolerance Defective (LTD). Lots which are at this quality level or poorer are to be subject to rectification with high probability. For example, one might determine that material that is as much as 10 percent defective should have not more than ten percent probability of outright acceptance (resulting in Type II error).

The other criterion that is sometimes employed specifies an Average Outgoing Quality Limit (AOQL). The AOQL consideration arises from the fact that outgoing product will either be at the original level of quality or it will contain zero percent defective as the result of rectification. On the average the quality level will be somewhere between these extremes. Thus, lots in the previous example that are 10 percent defective will retain that quality level one time in ten and will be released with zero percent defective nine times out of ten; over all lots of this initial quality the average outgoing quality will be one percent defective.

Lots of somewhat better initial quality are likely to have an AOQ higher than one percent because they are subject to higher probability of acceptance without rectification and consequent quality improvement. Thus AOQ will be low for obvious reasons when initial quality is good, and it will be

low because of rectification in cases of poor initial quality. The highest AOQ, designated the AOQL, will occur at some intermediate level of quality. The AOQL is inevitably lower than the corresponding LTD because of the rectification intervention.

Having addressed the issue of Type II error through specification of LTD or AOQL, the matter of Type I error is considered. Although less serious, the resulting unnecessary inspection is of concern. Rectification plans are therefore designed to minimize total inspection, given the specified LTD or AOQL criterion. A small sample will cost little to begin with but will necessarily have a high risk of Type I error and consequent cost of rectification. As the sample size is increased to reduce the latter cost a point is eventually reached where further error reduction is so small that the added sampling effort is not warranted. The point of optimum balance between the two costs depends upon lot size and its effect on the cost of rectification. In the long run it also depends upon the average process quality level that is maintained. Once these factors have been defined, optimum sample sizes can be calculated and have been built into the published Rectification Plans.

To illustrate the public health applicability of these plans, consider a scheme in which community health volunteers (CHV) are to maintain a supply of drugs for local distribution. In particular, they are always to have a supply of chloroquine. The decision is made to monitor this at the subdistrict level through a sampling plan with rectification. The drug stock for a sample of the roughly 100 CHVs in each subdistrict is to be checked for availability of chloroquine. If an excessive number of CHVs is found to be out of stock, all CHVs in the subdistrict will be contacted and resupplied if necessary. Experience has shown that in general about three percent of CHVs are out of stock at any one time, but there are occasional errant subdistricts in which the requisite supervision and control are lacking. The intent is to detect such exceptions so that in-stock levels of at least 96 percent are maintained in every subdistrict. Thus the AOQL is to be 4 percent. The published tables indicate that 17 CHVs should be selected and if two or more are out-of-stock, all CHVs in the subdistrict should be investigated.

While these procedures will satisfy the AOQL requirements, the tables also reveal that the associated Lot Tolerance level is 21.5 percent. That is, occasionally an out-of-stock rate that high will go undetected. Suppose that a Lot Tolerance level as low as 10 percent is required. The Lot Tolerance set of tables indicates that the sample size must be raised to 33; this procedure would reduce the AOQL to 1.7 percent.

The maximum Lot Tolerance for which plans have been developed is 10 percent. However, tables for selected AOQL values up to 10 percent have been prepared, and these generally have associated Lot Tolerances in the 20-40 percent range. For most public health field applications, therefore, the AOQL tables are likely to be preferred. Rectification plans with double sampling procedures have also been published.

AOQL Plans with Continuous Production

Variations on the rectification theme have been devised to suit circumstances in which production takes place continuously, rather than in discrete lots. The rules are as follows:

a. begin with 100 percent inspection until n units in succession are found without defect;

b. after this happens, inspect only every fth unit until a defect is found;

c. then revert to 100 percent inspection until n consecutive units are again found without defect.

Values of n and f are related to the AOQL selected. Suppose, for example, that the process of inserting IUDs at a family planning clinic is intended to have at least 98 percent of the women leaving the clinic with the IUD properly inserted. This translates to an AOQL of 2 percent. The plan devised for such circumstances calls for n = 76 and f = 20 (i.e., 5 percent of cases). This means that the first 76 insertions will take place under supervision. If all are satisfactory, only every 20th insertion will be checked thereafter. If a faulty insertion is detected, every insertion will again be checked.

Reduced and Tightened Inspection

The procedures considered thus far deal with separate subgroups independently. When present expectations are colored by experience of the recent past, however, such trends may be worth considering. If operations have become routine with few problems encountered, a relaxation of the monitoring system may be justified. On the other hand, a period of difficulty may dictate the introduction of more stringent standards. Plans employing normal, reduced, and tightened levels of inspection have been devised for these purposes. They are based upon specified Acceptable Quality Levels (AQL) with limited risk of Type I error when quality at such levels is maintained.

We shall describe one specific plan employing an AQL of 15 percent in terms of a sweep through a population, sampling small areas to determine whether a particular service is covering 85 percent or more of the population. The plan is initiated by selecting samples of 75 and requiring at least 56 to have received the service in order to consider coverage to be acceptable. The procedure is repeated in each subarea, and a running tally is maintained of the cumulative coverage observed. If the average level found drops below 81 percent, trouble is suspected, and tightened inspection is instituted. Thus, if the first three subareas showed 82, 86, and 72 percent coverage, giving an overall average of 80, the tightened inspection procedures would be employed.

Under tightened inspection the sample size continues at 75, but coverage of at least 62 of those selected is required for acceptance. Tightened inspection continues until the overall results reach the AQL, i.e., 85 percent coverage.

If the overall average goes as high as 89 percent, reduced inspection is permitted. This involves sample sizes of 15, one-fifth the original, and requires at least 7 of the 15 to have received services for declaration of acceptability. Reduced sampling is continued as long as each successive sample gives acceptable results and the overall average exceeds the 85 percent level deemed acceptable. Otherwise a reversion to normal inspection is indicated.

Because the procedures take advantage of past experience, they can improve the efficiency of data collection. It is essential, however, that the past experience be relevant to current decisions. Specifically, homogeneity is assumed, so that findings in neighboring areas are meaningfully related to expectations for the subarea presently under investigation.

Control Charts

All of the examples thus far have dealt with dichotomous data rather than continuous measurements. This is because the classes of sampling plans considered have typically been set up to investigate attributes, though they could be readily adapted to handle continuous variables. In contrast, Control Charts, the most common means of monitoring processes (as opposed to discrete lots), have been formulated for both attributes and continuous variables. Indeed applications involving the latter are more common and will be the basis for the present discussion.

Most Control Chart procedures call for frequent selection of small samples of 5-20 observations each. This

ensures prompt signaling of trouble, permits clear portrayal of trends, and simplifies calculations. In particular, ranges are usually used as measures of variation, rather than more complicated standard deviations. It can be shown mathematically that the difference between the minimum and maximum values in a random sample of five observations from a normal distribution will average 2.326 standard deviations. Thus, control limits to serve as decision points can be calculated easily on the basis of the average range, yet have the same interpretation as if the standard error were calculated. Conventional practice in industry is to set limits three standard errors from the mean, rather than the two standard errors typically employed in health applications.

Suppose that a system of home visiting is carried out by community health workers supervised from health posts, each of which looks after five workers. Workers average 28 visits per week, and the average difference between the best and worst weekly performance at a health post is 7 visits. The district health officer would like to monitor the system in several ways. He wishes: to identify promptly any post that exhibits exceptional performance levels (good or bad) during any week; to note any post in which variability among workers is extreme; to be aware of important trends in performance at any post; and to compare average levels of performance among posts over time. Control Charts for averages and ranges along the lines of Figure 4 serve all of these purposes.

Published tables relative to samples of five indicate that individual ranges should be not more than 2.11 times the average range and that separate averages of five should deviate from the overall process average by no more than 0.58 times the average range. In the present case, then:

Upper Limit on Ranges = (2.11)(7) = 14.8
Upper Limit on Averages = 28 + (0.58)(7) = 32.1
Lower Limit on Averages = 28 - (0.58)(7) = 23.9

Any weekly result that is outside these limits indicates that the subject health post is out-of-control, i.e., its pattern of performance differs significantly from usual practice. In addition, the visual display of plotted results, as in Figure 4, makes possible the discernment of trends and comparison of performance among posts without resorting to sophisticated statistical analysis. Certain rules-of-thumb have also been devised to be used with the charts. For example, since the probability is less than .008 that seven consecutive points will fall by chance alone on the same side of an average, such a condition is treated as out-of-control, even if no single point falls outside the control limits.

In addition to Control Charts for averages and ranges, p-charts for proportions and c-charts for counts of cases are used. Monthly tallies of the proportion of mothers making post-partum visits following delivery could be tracked with the aid of a p-chart. Weekly counts of new acceptors of family planning might be plotted on a c-chart.

Moving Averages and Exponential Smoothing

Although Control Charts for averages can be very useful in depicting trends, moving averages are sometimes employed to give an even clearer picture. Because averages of, say, the most recent ten observations tend to form a smoother line than the connection of individual points, trends appear more striking in plots of moving averages. Care must be taken in interpretation, however, for the various points on the line are not independent of one another; therefore, an apparent trend may simply portray the effect of one maverick observation persisting in succeeding averages.

Exponential smoothing is a more recent innovation that has gained considerable attention, especially in areas of forecasting. The thinking behind the concept is that the dynamics of a particular situation may be such that recent experience may be more relevant than earlier observations in assessing current conditions; yet a single recent observation is likely to be an unreliable indicator of present underlying conditions. Thus exponential smoothing uses past data but reduces exponentially the weight given as one goes back in time. With a weighting factor of 0.9, for example, the most recent result is given full weight, the one prior is given a weight of 0.9, the one before that counts 0.9^2, the preceding one 0.9^3, etc.

An attractive feature of exponential smoothing is the simplicity in updating averages. Again using a weighting factor of 0.9 for illustration, the previously calculated exponential average gets that weight in determining the new average, while the latest single result gets a weight of 0.1. Thus, the resulting average is somewhat like a ten-point moving average, whereas a weighting factor of 0.95 would produce results akin to a twenty-point moving average. The weight selected, therefore, depends upon how rapidly circumstances are thought to be changing.

Factorial Designs

Factorial designs represent a class of procedures for data collection and analysis that have been employed widely

in health, as well as industrial applications. Though they are already well known to health professionals, they are included here because two advantages associated with them are so important that they should be employed wherever possible, especially when the effects of two or more factors are to be assessed simultaneously.

To highlight those benefits, consider the situation in which use of an underfives clinic is hypothesized to be affected by travel distance and education of the mother. A distinction is to be made as to whether the residence is less that one mile, between 1 and 3 miles, or more than three miles from the clinic. Mothers are also to be distinguished according to whether or not they have had primary schooling. The dependent variable is the number of clinic visits made during the infancy of their youngest child. Hence, averages among the various combinations of circumstances are to be compared.

Considerations of tolerable risk of Type I and Type II error dictate the size of sample required. Let us suppose that 100 observations are needed from each of the three resident distances in order to achieve the requisite precision in analysis. If each group of 100 includes 50 schooled and 50 unschooled mothers, the total sample of 300 will be sufficient to analyze the importance of both factors. In other words, design efficiency is obtained in that each observation can be used for two analyses instead of one. Perhaps more important, the two analyses can be carried out without danger of confounding. In determining the importance of distance, for example,the effect of education can be ignored, since the proportion of educated and uneducated mothers within each of the "distance" categories is the same.

Apart from the benefits of its efficiency, factorial designs have the advantage of being able to test for interactions. Perhaps the more highly educated mothers utilize the clinic, regardless of travel distance, whereas the others tend to use the services only if they are readily accessible. This interaction possibility is routinely tested in the factorial analysis.

Specific Applications

To assist in the explanation of various sampling plans, brief reference has been made to a number of possible applications to public health investigations. While they were intended to be realistic, they were introduced mainly for purposes of illustration, and no attempt was made to elaborate them in detail. On the contrary, the purpose in the present section is to spell out specific recommended applications of selected sampling methods.

The list of applications is by no means exhaustive; it is meant to be comprehensive in two regards, however. First, it includes three generic topics of applicability: needs assessment; services delivery; and management. Second, the suggested applications represent a range of methodology that encompasses both lot and process sampling of attributes, as well as continuous variables.

Cluster Sampling, Lot Quality Assurance, Sentinel Surveillance

We begin the consideration of applications with reference to three approaches that are already widely used in the health field. Our purpose here is not to describe in detail the methods of Cluster Sampling, Lot Quality Assurance and Sentinel Surveillance but to offer a critical appraisal of their roles and usefulness within the comprehensive framework of sampling plans that has been presented.

Lot Quality Assurance sampling methodology is sometimes presented erroneously as an attractive alternative, or successor to Cluster Sampling. The fallacious reasoning is that LQAS can provide definitive findings about population subgroups on the basis of small samples and at the same time permit aggregation to yield estimates regarding parameters of interest in the total population. To appraise this contention critically in specific cases, we must first define clearly the target population. Then, if certain subgroups are of interest, we must establish the patterns of variation among them in order to be able to select a sampling plan that is appropriate to the particular circumstances. Finally, that plan should go as far as possible in ensuring that each member of the population or subgroup has an equal chance of being selected.

If the target population is large, the selection of clusters may be the only practicable way of obtaining a sample of reasonable size. In this case the clusters should be defined to be as heterogeneous as possible. Ideally, each cluster should be a mirror image of the total population. Short of this, variation among the clusters in the characteristic of interest should follow a normal probability distribution.

Having defined the clusters, selection of a sample of them should follow principles of randomness, with selection probability proportional to size (PPS). Selection within clusters should likewise be random. Systematic sampling is sometimes employed for convenience and to ensure that the sample is broadly representative (e.g., all areas of a village included). Care must be taken, however, to avoid any cyclical patterns in the data. In sampling medical records, for example, it would be inappropriate to select every

seventh day for investigation, since this would mean that only one day of the week would be represented.

If the population of interest turns out to be small, then clustering is unlikely to be necessary. If it consists of diverse small subgroups, it may be most appropriate to treat each subgroup as a separate population for decision making purposes. This is especially so if differences between subgroups tend to be bimodal. Thus, if an immunization program has been active in some communities and not in others, coverage levels are likely to cluster around two percentages, one high and one relatively low. The aim of sampling would be to distinguish between the two groups of communities for purposes of differential action. To aggregate findings to obtain an overall estimate of coverage would not be very meaningful and, in fact, could be misleading. One should not be trapped into the meaningless exercise of sampling from a bowl containing watermelon and grapes in order to obtain an estimate of the average diameter of fruit in the bowl.

Although the combining of information from subgroups to obtain composite estimates may be inadvisable, it may be useful to analyze the entire body of information in an attempt to discern meaningful patterns of diversity. For example, when two distinct levels of coverage exist, do these tend to vary among villages according to the district in which they are located, according to accessibility to paved roads, or following some other pattern? Such investigations will be considered more fully in the next section.

Where diverse subgroups are to be studied, and possibly compared, each should be defined to be as <u>homogeneous</u> as possible. Thus, the principles of stratification should be followed, not the principles of clustering. The LQAS approach to the sampling of these subgroups is frequently advocated, but we should recall that this approach has no magical powers; in fact, it is among the most rudimentary methods available, and in many cases will be distinctly inferior to other methods we have described. In particular, if the sampling procedures include decision rules regarding rectification, an AOQL plan is likely to be preferred.

In any event, one should not expect too much from any sampling plan. Three points in this regard are especially worth noting. First, division of the population into small subgroups serves to reduce the sample size per subgroup only slightly and, of course, increases tremendously the overall sampling required. To illustrate this, consider a sample of 50 from a population of 2,000 for purposes of estimating the prevalence of a certain condition. The estimate would have a standard error of about 7 percent. If for some reason ten groups of 200 were identified, and similar estimates were

sought for each group separately, samples of 41 would be required to obtain the same precision as before. Although the new populations are one-tenth the size of the original one, the new sample sizes required are 82 percent as large as the original.

The second cautionary note concerns comparisons. If a sample size is selected to obtain an estimate of coverage that is in error by no more than, say, ten percent in any population investigated, then estimates of the real difference between two specified populations could be in error by as much as 14 percent. This is because the estimated difference is subject to two sources of error instead of one. To obtain estimates of differences that are in error by no more than ten percent would require a doubling of the sample size in each population, thereby leading to a fourfold increase in sampling overall.

The third note of caution relates to absolute error in contrast to proportionate error. Absolute sampling error is largest when the prevalence is near fifty percent. In this case a sample of 100 is required to obtain an estimate that is within ten percent of the true prevalence. On the other hand, if prevalence is roughly twenty percent, a sample of 64 would be sufficient to obtain the same accuracy. As prevalence declines, one generally insists upon a smaller absolute error. Whereas estimates within ten percent of fifty might be tolerable, one might tolerate no more than, say, four percent error from a prevalence of twenty percent. In both cases tolerable error is twenty percent of population prevalence. When interest centers on relative error, rather than absolute error, large samples are required to investigate comparatively rare events.

Despite these limitations, the fact remains that a remarkably accurate picture of circumstances in a population can be obtained from a relatively small representative sample. It is also well established that routine records are frequently not a reflection of population conditions. They are maintained for a biased group of service users, and they are typically inaccurate and incomplete. The focusing of data collection at a few well chosen and carefully controlled sites through Sentinel Surveillance therefore has considerable appeal, especially if the information is actively sought from representative members of the population rather than passively and selectively received.

The monitoring process by means of Sentinel Surveillance is usually served better through frequent small samples instead of occasional large samples. If conditions are stable, the aggregation of results from the small samples is as informative as the one larger sample, and in the event of instability the frequent monitoring will reveal trends or

otherwise give early warning of impending trouble. In short, process Control Charts are ideally suited to Sentinel Surveillance.

Whenever Poisson probability is applicable, procedures are simplified because one need only track the number of cases periodically. Suppose, for example, that over a period of time an average of 16 cases of marasmus have been seen per month from a particular area. If the decision is made to set control limits to signal action at two standard deviations above the average (risk of Type I error is then 2.3 percent), the limit in this case is readily determined to be 24. This is because the standard deviation is simply the square root of the average, or 4, and the control limit is therefore 8 cases above the average, as shown in Figure 5. The observation of 25 or more cases in a month would call for special investigative action. The Control Chart might also portray a trend that would trigger action before the upper limit was actually reached. This visual feature of Control Charts is especially useful where seasonal patterns are anticipated or outbreaks are common.

Basic Sampling Plans: Attributes versus Continuous Variables

We move now to other suggested sampling approaches to three main areas of investigation: needs assessment, services delivery and management. Often we can be quite satisfied, of course, with a standard single sample of size large enough to limit risks of Type I and Type II error to tolerable levels. Even under these basic circumstances, the amount of sampling can be reduced through careful selection of the population parameter to be studied. This is illustrated in the first needs assessment case.

Suppose that limited resources for an intensified program of maternity care are to be concentrated in certain districts with especially high incidence of low birthweight (LBW). A rate of 10 percent is considered acceptable, and the sampling plan is to give a high degree of assurance that districts with LBW rates of 15 percent or more will be identified as "significantly high" and therefore candidates for special program attention. These stipulations could be interpreted to mean that 10 percent LBW is acceptable, and the associated sampling risk of Type I error is to be 0.05, whereas LBW of 15 percent is clearly unacceptable and is to have associated risk of Type II error limited to 0.10. The statistical requirements translate into a sample size of 363.

In the course of sampling, the actual birthweight might be recorded. Then the parameter of interest would be average birthweight, rather than percentage of births below 2500 grams. If birthweights are normally distributed with

standard deviation 500 grams, a population with average 3140 grams will have 10 percent LBW, while the 15 percent LBW population would have an average birthweight of 3020 grams. Using these averages to define acceptable and unacceptable, a sample of 149 would be sufficient to ensure that the tolerable risks of statistical error are not exceeded. Thus by working with sample means instead of the LBW attribute, the sample size could be reduced by 59 percent. This is because more information would be obtained from each observation.

For the latter approach to be strictly applicable, the pattern of variation among birthweights should be "regular." In particular, it should not be bimodal. Bimodality occurs when the population is comprised of two distinctly different subgroups, and in such cases the averaging of unlike units can be misleading. In the absence of a meaningful continuous measure, sampling for attributes (eg., LBW) is sometimes recommended. The problem of combining unlike things remains, however. Moreover, the condition is difficult to detect in attribute sampling, whereas it is readily picked up in the distribution of a continuous variable. In summary, consideration should be given to the use of continuous measurements in place of attributes whenever this is feasible.

Investigation of Breastfeeding through Double Sampling

A second illustration of needs assessment concerns breastfeeding. We suppose that surveys of mothers are to be carried out in various communities to determine whether they continue to breastfeed their babies at age three months. The aim is for at least 80 percent to be breastfed at that age, and the surveys are meant to detect significant departures from that level.

To meet the stated conditions we have slightly modified one of the standard industrial sampling plans as follows. Select a random sample of 35 mothers. Consider breastfeeding to be satisfactory if at least 25 women follow the practice. Consider it unsatisfactory and initiate corrective action if 8 or fewer women breastfeed. If the survey finds that between 9 and 24 women practice breastfeeding, a second sample of 70 is to be selected to obtain more definitive results.

The OC curve for this plan is presented in Figure 6. We observe that wherever the desired level of breastfeeding in fact prevails, the sampling procedure declares "satisfaction" with 97 percent probability. On the other hand, breastfeeding rates as low as 60 percent will be detected as "significant" departures from satisfactory levels nearly 90 percent of the time.

Figure 6 also shows the average sample size required under the range of breastfeeding practice that might be encountered. As expected, a second sample rarely will be required if breastfeeding rates are at desired levels and above or at levels of 20 percent and below. Under intermediate conditions, however, the average sample size can exceed 100.

The single sampling plan with OC curve similar to that of Figure 6 specifies a sample of 47 and requires at least 33 breastfeeding women for conditions to be considered satisfactory. Thus, if conditions are stable, i.e., most communities are at desired levels of breastfeeding practice, the double sampling plan will require a sample that is 15-25 percent smaller than is required by the corresponding single sampling plan. Moreover, the occasional deviant community will be automatically investigated more thoroughly by means of a second sample and will most likely be classified correctly. On the other hand, if practices vary widely among communities, the benefits from double sampling will probably not be worth the additional administrative burdens they entail.

Monitoring Services Delivery through Control Charts

As suggested earlier, Control Charts are widely used in industry for process monitoring, and they can play a useful role in the monitoring of health services delivery as well. We have already illustrated the use of control charts in a couple of ways, but there are many variations on the theme, and the following scenario introduces yet another.

Field workers carry out home visiting for several purposes, and they submit periodic reports of volumes of various categories of service provided. In addition, an overall indicator of field activity is of interest, and weekly mileage reports are used for this purpose. Analysis of past reports has shown that week-to-week differences for the typical worker are such that a standard deviation of 15 kms. has been calculated. Thus, a worker who averages 60 kms. of travel per week, may travel as much as 90 kms. in some weeks and as few as 30 kms. in others.

In order to gain a more stable picture of individual performance, weekly averages for each month are to be calculated and plotted on a control chart. It has been established that workers who normally average less than 30 kms. per week in travel are probably not sufficiently active in home visiting and should be investigated. At the other extreme, those who average more than 100 kms. per week may be guilty of misuse of vehicles.

The data indicate that the standard error of four-week averages is 7.5. If control limits are to be based upon two standard errors, therefore, they will be set at 15 and 115, as in Figure 7. Any worker with a weekly average for the month below 15 can be considered to be operating significantly below 30. Similarly, an average above 115 is significantly greater than 100.

In addition to the two control limits, allowing for chance variation, the control chart displays (dashed lines) the extremes of long-term acceptability. Any worker who performs consistently above or below the dashed lines is likewise "out-of-control," even if no single point exceeds the outer control limits.

The approach just described is conservative in that it only flags cases that are clearly exceptional. In contrast, managers may prefer to set control limits within the dashed lines, rather than outside them. In this case a point beyond the solid control limits would indicate that the worker "could be" in violation of standards and should be checked, whereas in the earlier situation the conclusion would be that the worker "is" in violation beyond reasonable doubt.

Selective Supervision with Reduced and Tightened Inspection

The principle of Management by Exception leads to the concept of selective supervision. Some workers need more support and assistance than others, and scarce management time and effort should be focused on the exceptional cases. Although a supervisory visit is far more than an inspection, industrial quality control procedures can serve a useful purpose in connection with supervision.. Specifically, the following case illustrates the use of reduced and tightened sampling procedures in selective supervision.

We suppose that ORS was introduced into a certain region two years ago, and it is expected that by now 60 percent of diarrhea cases should be so treated. To check on this the field supervisor selects twenty cases for review during each of the monthly visits with each worker. If over a three-month period he finds that 28 or fewer of the 60 cases chosen had received ORS, he concludes that the area concerned falls significantly below standard. He investigates the cause and introduces tightened sampling procedures.

Under tightened sampling the supervisor also selects twenty cases per month. If in the next three-month period he finds at least 37 ORS-treated cases among the 60 selected, he concludes that performance is satisfactory and returns to normal inspection rules. If he finds between 31 and 36 treated cases, performance remains suspect in a **Yellow Alert**

mode. Thirty or fewer treated cases confirms earlier concerns and calls for Red Alert crisis action.

If in the course of normal inspection, the supervisor finds at least 45 ORS-treated cases out of 60, he is satisfied that performance is at or significantly above accepted standards and reduces surveillance to once every three months. If, during any subsequent quarterly visit, the twenty cases sampled reveal seven or fewer treated cases, performance becomes suspect, and the supervisor returns to normal monthly surveillance procedures.

The verbal statement of the decision rules may make them seem complex. With the aid of a graphic display, as in Figure 8, however, they are seen to be quite straightforward.

Figure 9 displays the OC curves for the three procedures. The normal sampling procedure is virtually certain to find conditions satisfactory when indeed they are (60 percent ORS treatment rate) and provides nearly 90 percent assurance that corrective action will be initiated whenever the treatment rate is as low as 40 percent. Once trouble is suspected and tightened inspection procedures introduced, concern for Type II error is heightened. The more stringent procedures reduce that risk below five percent when the treatment rate is 40 percent: necessarily, of course, the risk of Type I error is also raised slightly to about seven percent. Under the circumstances leading to reduced sampling, no difficulties are anticipated, and the decision rules are designed to avoid Type I error. In the unlikely event that ORS use declines, however, the reduced sampling may fail to detect the deterioration of performance immediately unless the treatment rate falls to 25 percent or below.

Further Steps Recommended

It is clear, in principle at least, that much of the survey experience gained in other fields is applicable to public health. Practical field tests are needed, however, first to confirm the feasibility of certain especially promising applications and second to adapt and refine existing sampling methods to suit the peculiar field conditions faced by health workers, especially as they deal with issues outside the relatively narrow confines of immunization coverage where substantial expertise already exists.

Illustrative subjects for investigation are summarized in the following paragraphs. Although the list is certainly not exhaustive, it does cover a number of issues of current importance. This is not the place to present detailed

research protocols, but each of the proposals is presented in sufficient detail to clarify the problem to be addressed and to specify the suggested approach to its solution.

Likely Field Tests

The pervasive need exists for effective monitoring procedures designed to attain uniformly high levels of services coverage across diverse areas. Limitations of LQAS were cited earlier, and in any case relatively little experience has been gained outside EPI. Where coverage is found to be unsatisfactory, application of the concept of rectification seems reasonable, and accordingly AOQL plans are attractive. Such plans should be designed and tested in connection with EPI and for some other program with a very different character.

Sentinel Surveillance is gaining increasing respect as a basis for the monitoring of health conditions. Control Charts, the most widely used means of monitoring industrial processes, appear to offer promise in connection with Sentinel Surveillance as well. They are easy to design and implement; they involve frequent selection of small samples and therefore signal trends promptly; and they provide readily interpreted visual comparisons. These considerations lead to the recommendation that field studies be carried out to determine appropriate ways to marry Control Chart and Sentinel Surveillance procedures.

Closer Scrutiny of Patterns of Variation

Numerous sampling plans have been devised to serve different objectives and to accommodate a wide range of field conditions. The efficiency and validity of a specific plan in particular circumstances greatly depend upon the nature of variation under those conditions. For example, as we have already observed, the aggregation of information from subgroups with bimodal characteristics is a far different matter from aggregation under conditions approximated by the normal distribution. Moreover, the magnitude of variation among clusters in relation to that within subgroups affects the efficiency of various sampling plans. The critical importance of variation makes this subject the appropriate focus for field operations research.

Quick-and-Dirty Approaches

Three specific aspects of variation are proposed for investigation. The first consideration stems from the usual trade-off between statistical rigor and administrative ease. In particular, statistical considerations demand representativeness, whereas administratively we prefer to sample readily accessible members of the population. The problem is that the latter are likely to exhibit more favorable status regarding the parameter of interest than is evident in the population generally.

If, however, conditions are unsatisfactory under the best of circumstances, it is perhaps unnecessary to search further. Thus if only 30 percent of families living within five kilometers of a health unit are found to utilize its services, it may not be necessary to go to the trouble of sampling remote areas to document underutilization. This is especially the case if a pattern observed among those within one, three, and five kilometers of the unit can be extrapolated to greater distances. Extrapolation can be dangerous, of course, but there are surely instances in which appropriately designed convenience samples are good enough. The proposed examination of the issue is not very sophisticated but could lead to decision rules of considerable practical value.

Efficient Cluster Designs

The second area of investigation, related to Cluster Sampling, is more complex and requires some introductory statistical explanation. We define:

m = number of clusters selected
n = number of observations in each cluster
V_o = variance within clusters
V_c = variance among clusters

The process of sampling generates a statistical measure for each cluster. The measures could be prevalence (or coverage) rates

$$P_1, P_2, P_3, \ldots, P_m$$

or they could be sample averages. For purposes of present discussion we will limit attention to coverage. We also assume that clusters are defined to be of equal size. (To assume otherwise would complicate the mathematics but would not alter the principles to be enunciated). Then

V_p = variance among P_i $(i = 1,m)$

$$= V_c + V_o/mn. \qquad (1)$$

The average of the P_i is an estimate of population coverage and is subject to sampling error measured by the variance

$$V_e = V_c/m + V_o/mn. \qquad (2)$$

Where variation within clusters follows the binomial distribution, V_o is $P_i(1 - P_i)$ and is at most (.5)(.5). For the sake of this discussion we therefore assume

$$V_o = (.5)(.5) = .25$$

We also specify that sampling error is to be limited such that two standard errors are not to exceed a specified level L, or

$$V_e = L^2/4.$$

It follows, from (2), that

$$.25L^2 = V_c/m + .25/mn$$

or

$$L^2 = 4V_c/m + 1/mn. \qquad (3)$$

The total sample size, **mn**, will be minimized if clusters of size 1 are chosen. Then

$$L^2 = (4V_c + 1)/m. \qquad (4)$$

Because of economies associated with clustering, we are willing to exceed this minimum, but we might stipulate that the total sample size is not to be more than **K**. Then (3) becomes

$$L^2 = 4V_c/m + 1/K. \qquad (5)$$

With L and K specified, the number of clusters, **m**, depends upon the magnitude of V_c.

These considerations lead to a possible algorithm to be described with the aid of a practical example.

1. Select five preliminary clusters of five observations each and obtain P_1, P_2, ..., P_5. [Suppose the results are .2, .6, .2, .8, 1.0]

2. Calculate V_p. [V_p = 0.128]

3. Determine V_C from (1). [$V_C = 0.078$]

4. Specify L and determine from (4) the minimum sample size required with simple random sampling. [We suppose that the estimate of population coverage is to be in error by 10 percent at most, so that the sample size must be minimally 132. Note that if V_C were 0 (clusters maximally heterogeneous), then the minimum sample size would be 100. We could specify K as a multiple of either of these minima.]

5. Specify K. [We select K = 264, being willing to make as many as twice the number of observations needed with a random sample.]

6. Determine from (5) the minimum number of clusters needed. [We find m = 21.2; then n = 12.5. As a practical matter, select 20 clusters of 13 observations each.]

Table 1 presents generalized findings for a range of values of V_C and K, holding L fixed at 10 percent. With V_C = 0, that is all clusters equally representative of the population as a whole, a single cluster is adequate. As V_C increases, meaning that individual clusters are more unique, a larger number of clusters is required to get adequate representation of the entire population.

The relationships that have been spelled out suggest other manipulations and inferences. The main point of interest is to explore the advantage of two-stage sampling, the purpose of the first stage being to select an efficient clustering design for use in the definitive second-stage investigation. The problem is that a small sample at Stage 1 would give an imprecise estimate of V_C and could mislead the Stage 2 design. On the other hand, an excessively large Stage 1 sample could itself prove to be inefficient. Investigation of these trade-offs under field conditions is expected to lead to flexible decision rules and consequent Cluster Sampling designs most appropriate to local circumstances.

Pattern Recognition

The preceding section has been concerned with the efficient aggregation of cluster findings in order to make overall assessments of defined populations. In contrast, sample surveys are frequently intended to support rational action in relation to population subgroups or small areas. Because sample size bears little relation to population size, this can lead to substantial data collection from a large number of individual areas, and it is important to economize on the effort in any way that is acceptable statistically.

While findings from individual surveys are to be interpreted separately, they may not be entirely independent. Results from one village may be closer to those from an adjacent community than to those from another district. Similarly, findings from a current survey may be more similar to those derived one month ago than to those of one year ago.

These considerations lead to the search for practical techniques of pattern recognition. Two techniques in particular hold promise, the first of which is exponential smoothing. Under certain conditions there could be merit in a kind of weighted average that gives full weight to findings (F_t) from the latest survey, less weight to the immediately preceding results (F_{t-1}), and still less to those earlier, so that the best estimate of overall conditions is

$$F_t + WF_{t-1} + W^2F_{t-2} + W^3F_{t-3} + \ldots + W^nF_{t-n}. \quad (6)$$

The trick, of course, is to find a suitable weight, W, between 0 and 1 that does not give excessive attention to old data, yet minimizes the effects of imprecision associated with a single recent sample estimate. We propose investigation of the matter under actual field conditions in order to determine whether practical algorithms can be derived.

The second approach to pattern recognition involves the "response surface" technique of analysis that has been used to improve yields from industrial processes. The approach can be most easily visualized in terms of an outcome (quality of product) that is a function of two factors, say temperature and pressure. Levels of the two factors are portrayed along two axes, X and Y, and the yields resulting from specific combinations of temperature and pressure are represented by a corresponding perpendicular height above the (X,Y) coordinates. Thus the process is analogous to a mountain to be climbed; the analytical problem is to locate the peak and to define the terrain that characterizes the ascent to that peak. Hypothesizing that the areal patterns of service coverage may fit this model, we propose testing the response surface approach to analysis of findings from subgroups that are related geographically, or possibly along other dimensions.

The recommended investigations are not likely to be equally feasible or of equal priority in a particular setting. Likewise, there will undoubtedly be resource or time constraints in carrying out the studies. After agreement has been reached on the subject and site of certain field trials, and the level of available resources and the sophistication of investigators has been established, realistic protocols can be prepared to meet the locally relevant objectives.

Table 1

OPTIMUM NUMBER OF CLUSTERS
FOR SELECTED V_c AND K
$(L=0.1)^c$

$S_p = \sqrt{V_p}$	V_p	V_c	Recommended No. Clusters			
			K=150	K=200	K=250	K=300
.224	.0500	0	1	1	1	1
.25	.0625	.0125	15	10	8	7
.30	.0900	.0400	48	32	27	24
.35	.1225	.0725	87	58	48	43
.40	.1600	.1100	132	88	73	66
.45	.2025	.1525	150	122	102	91
.50	.2500	.2000	150	160	133	120

Figure 1

OPERATING CHARACTERISTICS
TWO SAMPLING PLANS

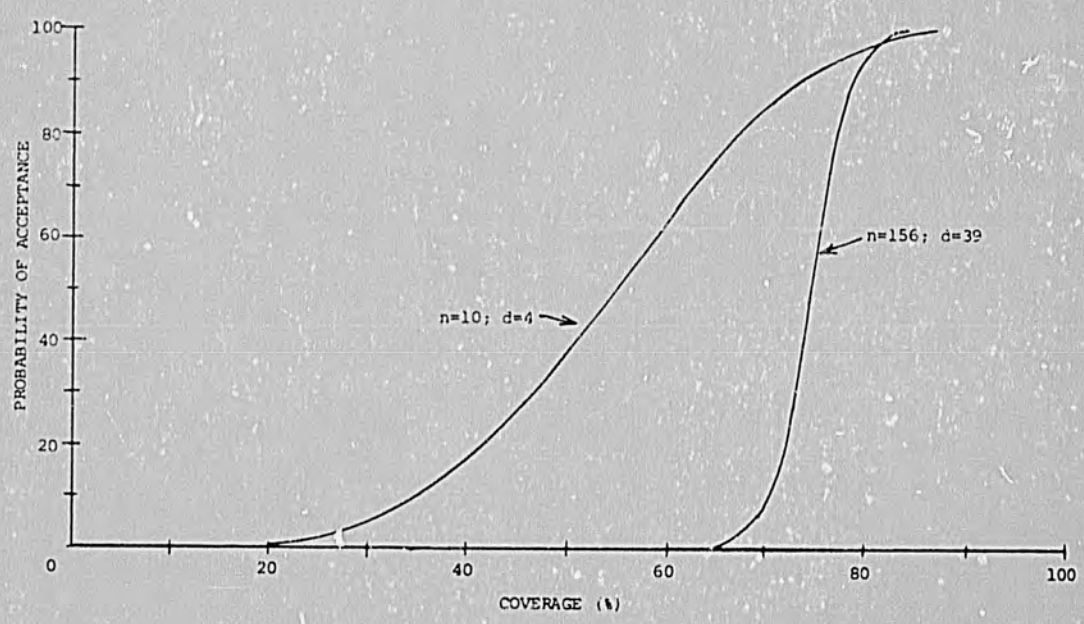


Figure 2

DOUBLE SAMPLING PLAN
AVERAGE SAMPLE SIZE

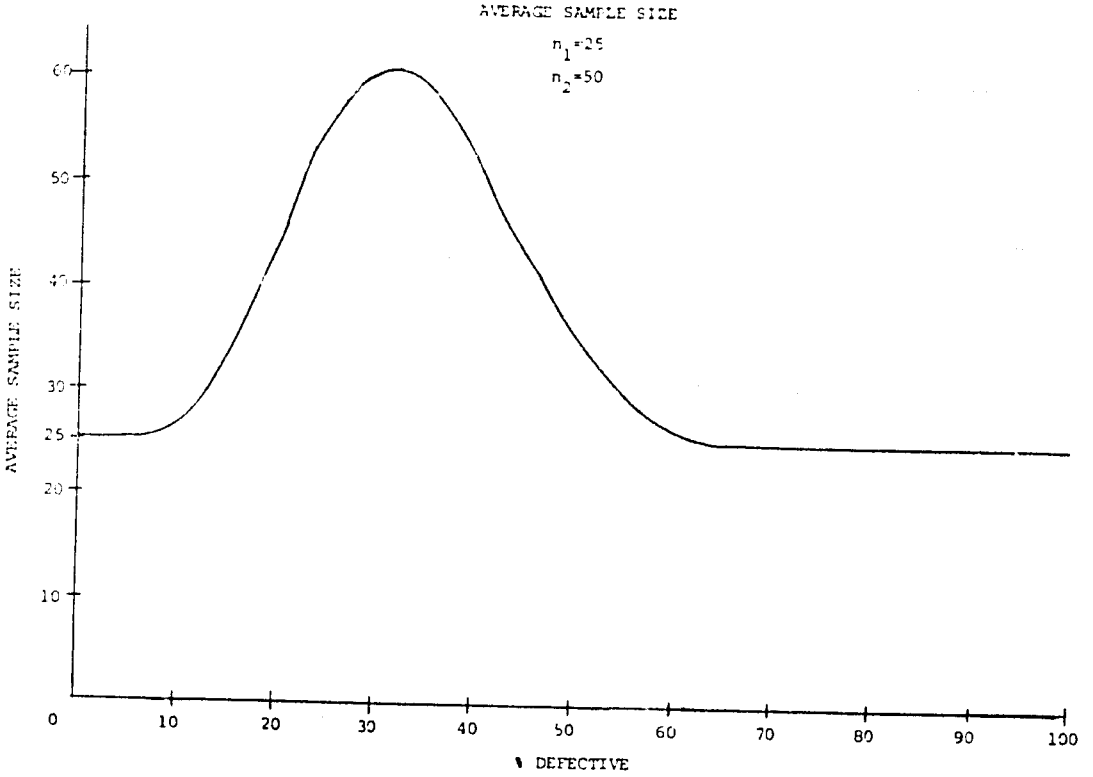


Figure 3

SEQUENTIAL SAMPLING
BCG VACCINATION

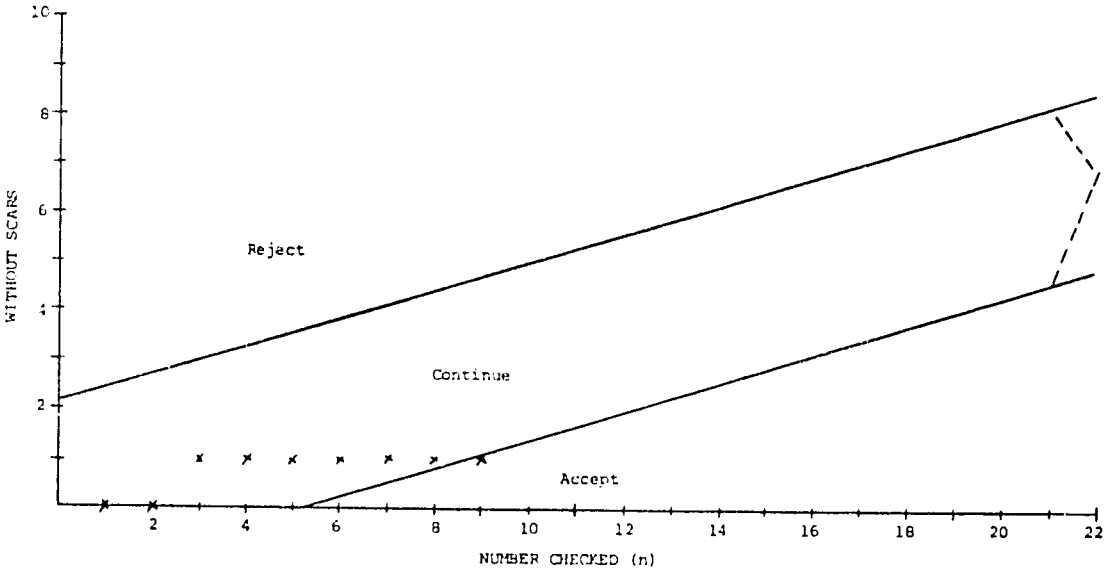


Figure 4

CONTROL CHARTS FOR HOME VISITING

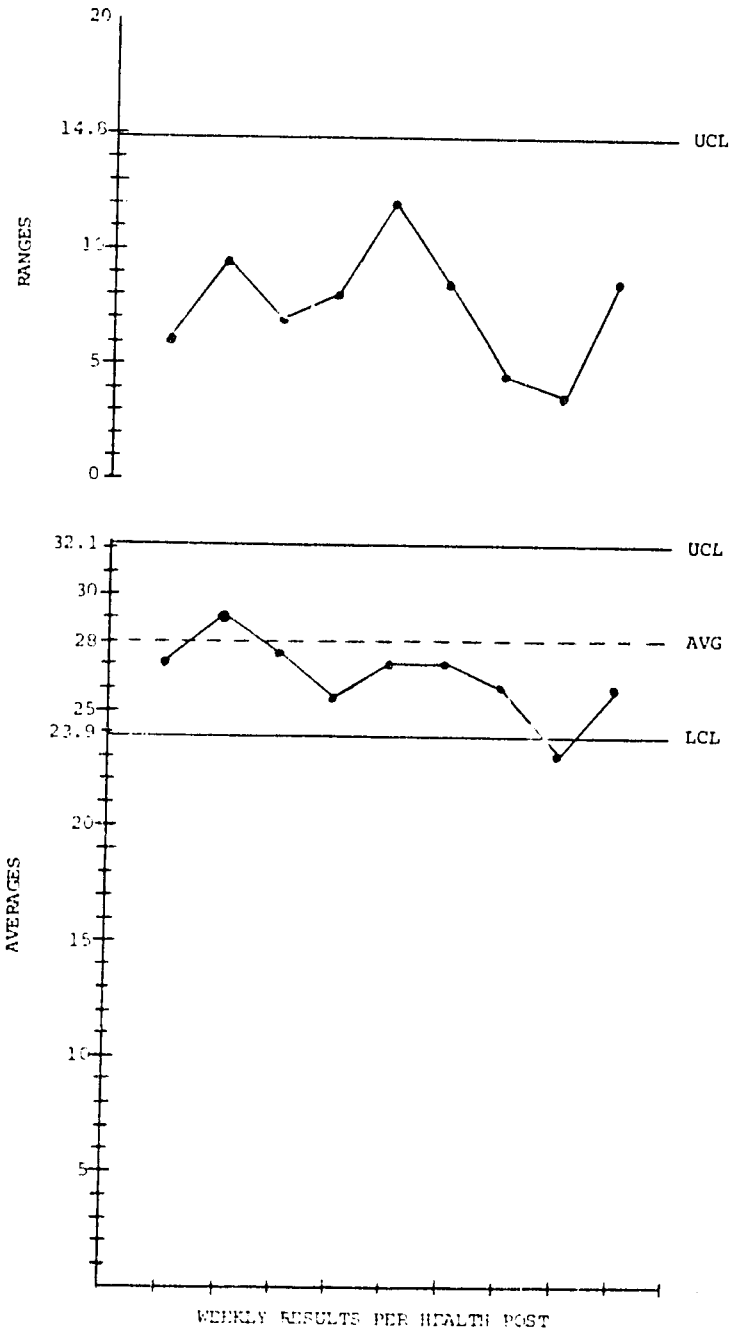


Figure 5

CONTROL CHART FOR MARASMUS

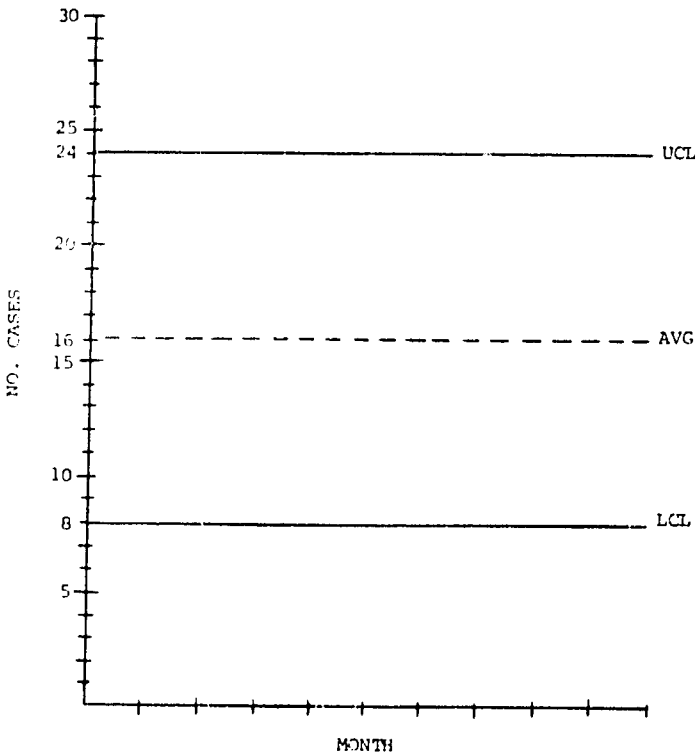


Figure 6

DOUBLE SAMPLING PLAN
FOR BREASTFEEDING

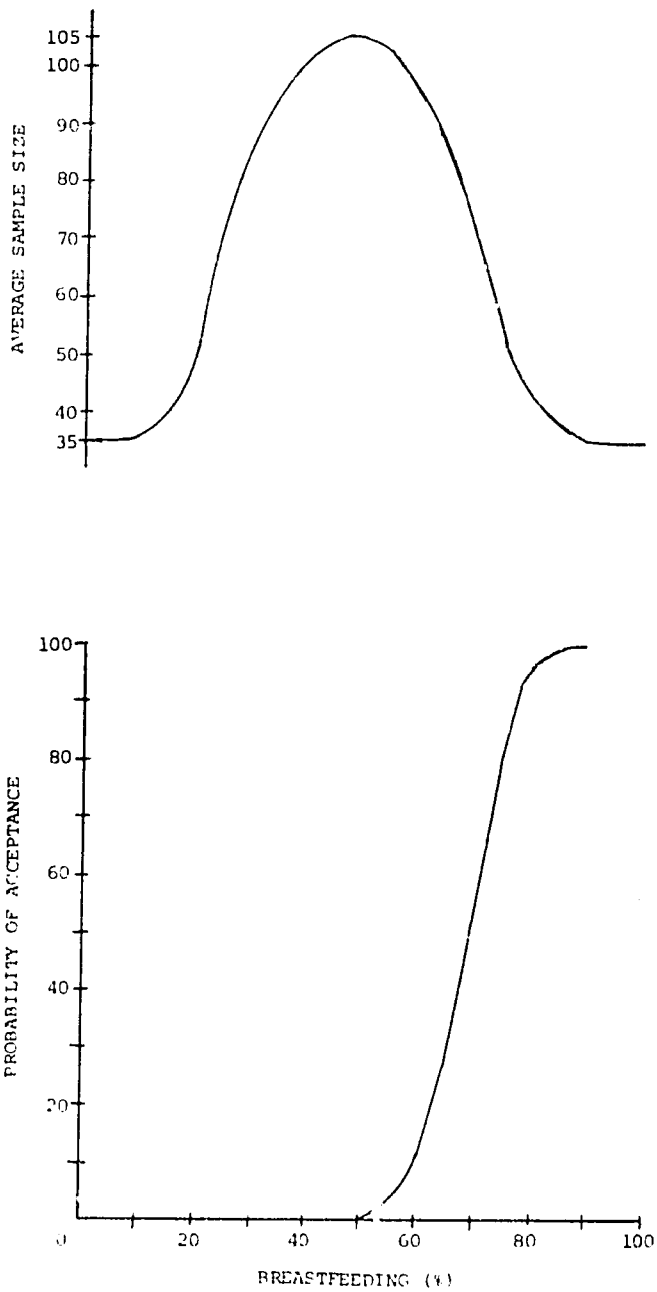


Figure 7

MILEAGE CONTROL CHART

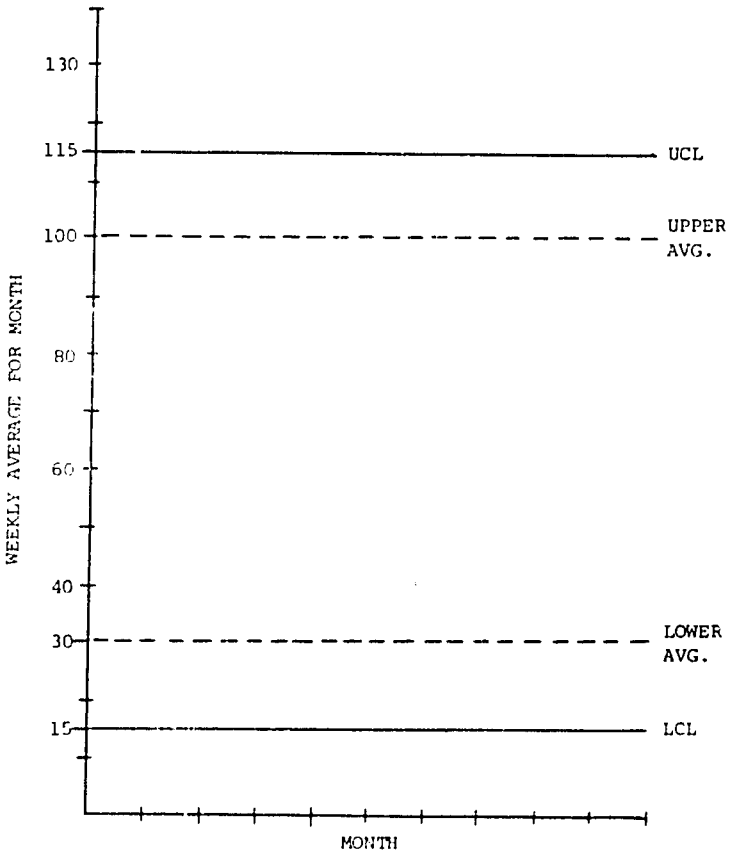


Figure 8

REDUCED-NORMAL-TIGHTENED
SAMPLING SCHEME

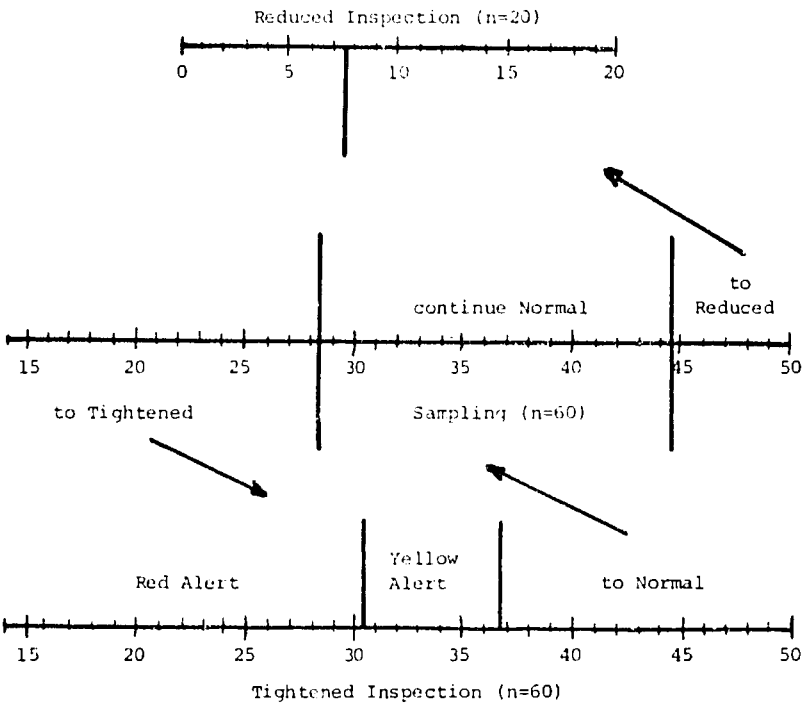
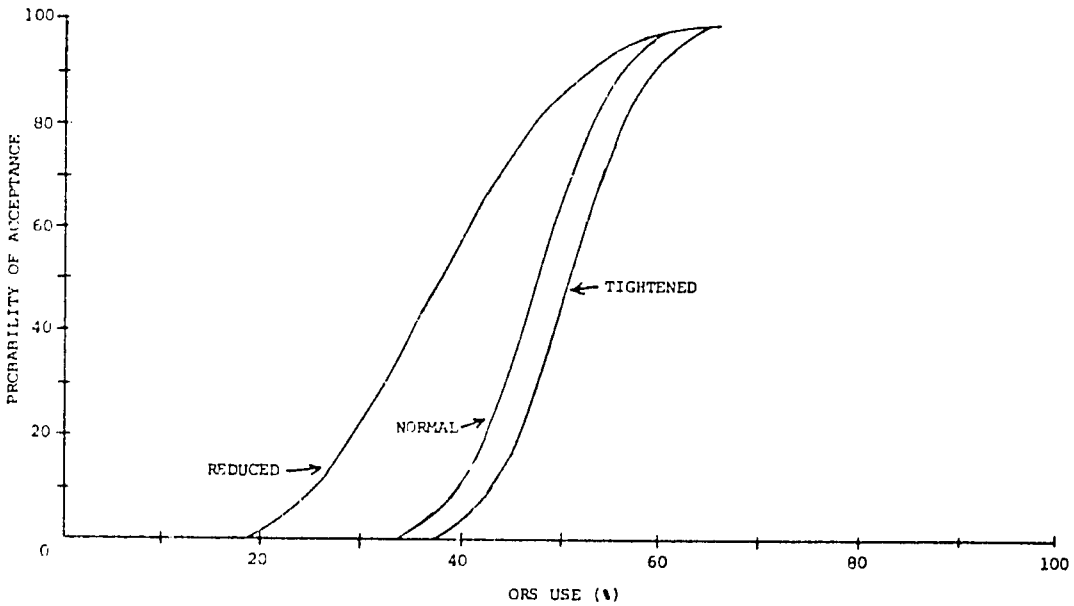


Figure 9

OPERATING CHARACTERISTICS
REDUCED, NORMAL, TIGHTENED SAMPLING PROCEDURES