OPERATIONS RESEARCH CENTER

working paper

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

PRELIMINARY SURVEY OF CLASSICAL STATISTICAL TECHNIQUES FOR INCORPORATION INTO ADAPTIVE EVALUATION METHODOLOGY

by

Alan S. Minkoff

OR 110-81 February, 1981

CONTENTS

ABSTRACT

Evaluation of public programming currently tends toward plans that are set in advance of any sampling and adhered to throughout. Because increments in the knowledge profile during the course of an evaluation might beckon adjustment of the working procedure, fixed evaluation methodology may be cost-inefficient. It is desired to develop a methodology that is adaptive to changes in the knowledge profile. This might be most easily accomplished by borrowing ideas from some of the disciplines in which relevant problems occur. The most promising fields for this task include classical and Bayesian statistics, reliability theory, and dynamic programming. This paper reviews the techniques in classical statistics that seem most apt for handling the problem of adaptive changes in an evaluation to updated knowledge profiles, and considers the paths along which future research ought to be conducted.

I. INTRODUCTION

Evaluation methodology is the tool policy-makers use to assess the effectiveness, prospective or actual, of public programs. Those who put a program into effect require feedback on the program's impact: who it is affecting, and to what degree. In order to decide whether to continue, alter, or terminate a program, it is essential to know what it has done already, and what it is likely to do in the future. Answering these questions is the role of evaluation.

The brief description of evaluation in programming holds mainly in the utopian world. The fact of the matter is, scientific evaluation often breaks down in the face of program changes and other human concerns. Mathematical techniques for the physical sciences applied to the task of evaluation often produce cost-negligent methodology, and are sometimes abandoned for haphazard guesswork when critical assumptions prove unfounded. The science of evaluation research is still quite young, and only now are measures being developed for combating the shortcomings of evaluation technology as it stands today.

One critical drawback in evaluation procedures is that they are almost always of a rigid, totally preplanned design. The various pieces of the evaluation process are laid out before the investigation takes place, and this design is strictly adhered to (one exception: if something unplanned for occurs, who knows what will be done next!). This strategy comes up short on two accounts: it may be cost-inefficient; and changes to the environment of the program or of the evaluation might render evaluations difficult or meaningless. These possibilities beckon a more flexible approach, which we call adaptive evaluation design.

Adaptive evaluation ameliorates the weaknesses of fixed designs in two ways: it uses the data already collected, sometimes as they arrive, to guide future evaluation design; and it attempts to reduce its susceptibility to the destruction of its foundations arising from changes in the program and its environment.

Adaptive evaluation designs can always be conjured and employed. An important question to answer is, how do they perform? We want to answer this both in comparisons to other adaptive designs, and in comparisons to fixed designs. Concurrently with our development of adaptive evaluation methodology, then, we need to develop a set of performance measures that will enable us to evaluate the evaluations, so to speak. These performance measures should provide the means to demonstrate the advantages of flexibility in evaluation design. They should also be available to the evaluator when he/she must decide which evaluation design to choose.

Performance measures, it may be conjectured, will be numerical-based aspects of evaluation designs. It is important to keep in mind that strict adherence to numbers may blind evaluation researchers toward the practicality of the methodology. While an "optimal" technique, in some sense, is desirable, a simpler technique might be more attractive than a more complex one with a higher value on a performance measure. It should be recognized that not every evaluator is a statistician, decision analyst, and computer programmer all rolled into one. Thus, evaluation designs ought to be assessed through both their performance measure results and their practicality. There are other considerations, too, but work in those areas is being undertaken by Mr. Thomas Campbell, and will be bypassed here.

II. OBJECTIVES

This paper is a survey of the techniques of classical statistics that might be applied to adaptive evaluation design. The scope of the survey will be limited to those techniques applicable toward the problem of adaptivity to updated knowledge profiles. This includes the problem situations of adaptive switching of time periods in an evaluation, and adaptive allocation of scarce evaluation resources; one section shall be devoted to each situation. No attempt to formulate versions of techniques ready for simulation testing or the like shall be made. Rather, the paper is designed to present the possibilities, mull them over, and indicate those techniques that seem most fruitful for further investigation.

The task outlined above cannot be adequately performed by mere description of procedure. Discussion concerning assumptions of the technique and how they might fit in with evaluation circumstances, plus appraisal of the practicality and the performance of the techniques will accompany description. Also, suggestions for performance measures arising from exploration of the technology will be pointed out where relevant. A final section outlines the course of future research.

III. ADAPTIVE TIME PERIODS OF THE EVALUATION

Evaluation processes can often be partitioned into two or more distinct task segments in time. A simple example is an evaluation divided into baseline and experimental phases. The first time period is devoted to collection of data that describe the situation before the start of the program, and the second is used to measure changes caused through implementation of the program. Often, these time periods are fixed in duration before the collection of any data. But these fixed periods might not be the most efficient use of the resources (time, money) that can be arranged.

Let us define the problem more concisely. The goal in each segment is to measure a parameter of some probability distribution and thereby characterize the situation. Each segment may be connected to a program variation for evaluation or an assessment of the environment with no program in effect. We want to measure shifts in the parameter's value under program/no program circumstances, or between different variations of the program. Assume the whole evaluation period consists of a certain number of "days" of sampling, with "days" clustered into the segments of the evaluation. A "design" is a method for arranging the sampling days into evaluation segments.

A fixed design corresponds to a clustering of sampling days into segments before any data collection is undertaken. Also, this design is not broken once it is set, so the pre-evaluation plan is strictly adhered to throughout. It may happen that the evaluators became reasonably sure of the value of the parameter before the pre-determined end of the current segment. ("Reasonably sure" will not be defined for now). Under the fixed design, they would not be able to end the current segment and begin the next one, or end the evaluation, whichever is the case. This "switch" in time periods, if it could have been performed, would have enabled more time to be spent in investigation of the parameter in later time periods, and generally effect savings in time and money.

The above argument is the main rationale for developing evaluation designs to adaptively determine switching times in an evaluation. Let us formulate the adaptive approach to the switching problem. In this instance, we again have to cluster the available sampling days into the time segments of the evaluation, but we no longer have to follow the predetermined design, if in fact we created one to begin with. At the beginning of each sampling day, we must decide whether to:

a) extend this evaluation period for at least one more sampling day, and thus continue sampling today; or

b) terminate this evaluation period and move on to the next, instituting whatever program changes are called for, and sample (or conclude all data collection, if this is the last period).

The beginning of the "day," in this formulation, connotes the point at which the decision is made to extend or switch. The remainder of the day is devoted to data collection with no such decisions. Our goal in this area is to develop a set of rules, called "switching" or "stopping" rules, to guide the decision-maker in decisions of the above type.

Criteria for making the switching decision can arise from one or both of two issues:

1) Interperiod considerations -- comparisons of costs and benefits, at present against those in subsequent periods;

2) Intraperiod considerations -- achievement of certain prescribed levels of significance (and interval width, in estimation).

In the formulation of any technique, we must see how the technique confronts these issues. Testing all hypotheses at the 5% significance level totally ignores the events and costs of future segments; indeed, it may squeeze those segments out of the evaluation process. On the other hand, some intraperiod considerations are necessary, if only to calculate certain indices to compare with those of future periods. Individual techniques might place more emphasis on one issue than the other. We must always see how these issues are represented, for neglect of one can render a procedure meaningless for practical usage.

A key concept in all this is cost, or loss. "Cost," as it will be used in this paper, does not solely refer to the expenses of observation, but also to losses incurred from the difference in value between the true parameter and the estimated or hypothesized one. As such, it can be difficult to measure, for there are many forms the loss may take. How does one measure the loss in under- or overestimation of a crime rate, for example? The answer is highly dependent on the environment of the program, which cannot be specified now, if this research is to have wide-spread application. Yet cost-minimization is a very appealing and very rational basis for making decisions. It is an easier concept for the evaluator to understand and use than, say, the utility of significance levels or confidence widths. So what we might hope for is a set of guidelines for loss structures pertinent to program evaluation to develop cost-based decision techniques.

With this concept under our belts, let us reiterate our objectives in the adaptive switching problem. We wish to develop a methodology that will adaptively guide us in determining the proper time to switch from one segment to another in an evaluation. The implication at the time of the switch is that we are sure enough of the value of the parameter to warrant commencement of study of a different aspect of the program. How "sure" we need to be depends on overall budget constraints as well as on the relative costs of inaccuracy from period to period; all this is implied in the "cost" concept detailed above. Simply put, if the degree of certainty an extra day of sampling is expected to add is not worth more to us than either the additional time we can spend in subsequent periods or the savings in costs of observation in this period, we should switch periods. This is precisely what is meant by "interperiod considerations."

A good application of this principle can be found in Willemain [16]. His paper takes a Bayesian approach to the adaptive switching problem, so his results cannot be utilized directly here. We might, however, derive techniques from it to "clean up" our classical analysis, so it is worth glancing at. Bayesian analysis treats parameters to be measured as random variables. The original, or prior, distribution of each parameter is initially constructed by the researcher. Distributions are updated based on the data that comes in. Loss functions dependent on the absolute distance between true and predicted values can then be manipulated through the assessed distribution of the parameters involved. In Willemain's work, even the impact of the program is assessed in a distribution, so that this component fits in with the overall scheme. A utility function is added, and preposterior analysis is used to assess the utility of each option in the continue/switch decision.

Classical statistics differs from Bayesian statistics in that parameters to be investigated are treated as single unknown values rather than distributions. When we make an inference about a parameter in classical statistics, we only say either that it is "this" and not "that," or only that it is "this," or that it is "somewhere in here, but one can't say where it is more likely to be." The first statement arises out of hypothesis testing, the second from the same or from point estimation, and the last from interval estimation. We do not associate probabilities, in the Bayesian sense, to these statements. We might indicate a relevant error rate or confidence interval, but the inference itself carries no weighting factor for what we say is so. Adding such weights to a statement draws the inference out of the classical realm and into Bayesian theory.

This presents a problem in designing classical methods to take on interperiod consideration. Should one project program characteristics of future periods, one must be very careful in manipulating them. If one projects a possible set of values for a period, then the assignment of probabilities each value has of occurring, no matter how small the set, is essentially Bayesian analysis. An alternative manipulation might be a minimax procedure, where the action that performs the best in the worst case of future values is chosen. This technique might be cumbersome, though, and probably too pessimistic. In the author's opinion, classical statistics will perform better in most respects if interperiod consideration are handled subtly. Just how will be outlined shortly.

If we decide to employ classical statistics, we will need to choose a strategy for working with the parameters. Classical statistics provides two main themes for investigation of a parameter: hypothesis testing, and (point or internal) estimation. Which theme we choose depends on what we want to find out, and the loss structures involved. Suppose loss depends primarily on whether the parameter is a certain value, or is significantly different from it. Then hypothesis testing is probably all we need to answer the question. On the other hand, if we have no preconceptions, and loss depends critically on how accurate we are, and increases steadily with increasing error, we probably need to estimate the parameter.

The purpose of classical statistical analysis is to make inferences about the data and the population(s) they came from. We cannot make perfectly accurate statements about a population in all confidence unless we look at the whole population. This option is normally too expensive to perform. Thus, we must make concessions. We want our statements to be true, naturally, but we cannot assure perfect certainty. So we specify some level of probability of our being wrong that we are willing to tolerate. In hypothesis testing, this is the significance level. In estimation, it is the confidence level associated with a confidence interval of a certain width. Actually, this becomes a pair of tolerances, as the confidence interval can be adjusted to the confidence level, too.

It was expressed before that classical statistics adaptations to adaptive switching will probably have to concentrate on the intraperiod considerations, as far as actual mechanics are concerned. If so, we will almost certainly have to work with tolerances. And we will have to include overall evaluation constraints somewhere in the technique. These two parameters are very closely related, especially where interperiod considerations are shoved into the background. We would like to use the overall constraints, plus assessments of future behavior, to set the tolerances within periods. This point will be reiterated and inspected later in the paper, after we have looked at some sequential techniques.

Another reason for selecting intraperiod-based switching rules for study is that the bulk of classical statistical literature is geared toward them. A common application of these rules is in reliability theory. A typical example is to estimate the proportion of defective items produced by a machine. The main difficulty concerning adaptation of this sort of technique to switching rules is that the applications describe stopping rules. If the choice is not to continue, the experiment is terminated. There is little coverage of the case where, in the example, the machine is adjusted and more sampling conducted, mainly because this can be done simultaneously, with two such machines. Situations rarely arise where it is possible to perform parallel testing in evaluation situations. The subjects are mostly human, and the testing environments complex. We must therefore build on the stopping rules described in the literature to create switching rules. The first step in this process is to examine elementary techniques and their stopping rules.

One of the first advances in sequential sampling theory was made by Wald [13]. He formulated a Sequential Probability Ratio Test (SPRT) to adaptively decide which one of two hypotheses held for a particular situation. He proved several desirable characteristics of the SPRT, including optimality for certain true values of the parameter [with Wolfowitz, 14]. Although many more questions about it have remained unanswered, and doubts have been raised about its actual performance, the SPRT seems to be the backbone of classical sequential theory, at least as far as the literature goes. Its position in the theory and the wealth of material on it makes the SPRT the logical starting point of our investigation.

In its basic form, the SPRT is a test of two simple hypotheses. The quantity tested is a probability ratio, or the log of one, of the value of the density function at the point sampled under the alternative hypothesis to the value there under the null hypothesis. Bounds on the value of the cumulative product (sum if taking logs) of these ratios are set. Sampling continues until the cumulative product passes outside one of the bounds. Depending on which bound is exceeded, the null hypothesis is or is not rejected.

In order to use the SPRT, the evaluator must frame his/her questions in terms that hypothesis testing can handle. Two specific values of the parameter (three or more in variations of the technique) must be proposed, and the value decided upon will be one of these pre-specified points. In theory, this does not present a problem. The questions that must be tackled concern the nature of the program itself. It is reasonable to assume that costs of error are minimized if the value of the parameter that comes out of this analysis is equal to the true value of the parameter. But how do costs increase as the absolute difference between the true and predicted values of the parameter increases?

Bayesian statistics can cope with a wide variety of cost schemes relatively easily. Since the predicted value is expressed in a probability distribution, rather than a point, deviation from the true value of the parameter (perhaps using the expected value of the posterior) also can be embodied in a distribution. This distribution serves to weight costs based on the difference to yield an expected cost. This can be used to compare expected costs under both ends of the continue/switch decision and choose a cost-minimizing plan.

On the other hand, the discrete set of possible parameter values arising from an hypothesis test is very sensitive to the cost scheme. In the case of measurement of a normal mean, especially with a known or anticipated large variance, there is considerable room for the mean to vary. Costs increasing linearly or quadratically with the error of prediction might require more detailed specification of prediction than a small set of hypotheses would give. But if costs of error follow some sort of step function, than the multiple-hypothesis test might compare favorably with Bayesian techniques on relevant performance measures. Performance quality aside, the hypothesis test may be more intuitive to evaluators with a classical training. An SPRT-based test might be simpler to implement, for the evaluator need not bother with assessing personal opinions as distributions if, say, a minimax approach to loss is used. The point is, there may be situations, contingent on the evaluation environment, in which classical techniques are preferred to Bayesian ones, which justifies further investigation of classical techniques.

Classical sequential theory extends much further than sequential hypothesis tests. In particular, techniques for sequential estimation have been described in the literature. Sequential estimation, in the classical sense, seems not to have been explored to the degree that sequential testing has. It is not clear whether it is because the questions are easier to answer, or that Bayesian techniques dominate the classical ones, or whatever. Yet a number of classical techniques for sequential estimation may be identified and investigated. These will be dealt with later in the paper. For now, let us observe how the basic SPRT is used to create more specialized techniques in sequential testing.

Wald left much unsaid concerning the SPRT. What he did was to work out the elementary theory, and illuminate it through two distributions for examples: the binomial, and the normal mean. Even in these examples, some important questions were answered by suggestions for working procedure, with little or no analysis of the procedure's behavior. For instance, a budget constraint might limit the total number of observations. Wald gives a rule for resolving the decision, but does not examine what effect this resolution has on the error rates. Truncation is a very important consideration in any practical application of the SPRT. Luckily, much has been written about the SPRT since Wald's work came out, filling in some of the gaps with useful material.

A closer look at what is bound up in the SPRT lends some idea about how difficult it would be to compose an all-encompassing work on it. One should be able to compare it to other techniques which might be cheaper to use and/or more accurate. This has been done through comparison of OC and ASN curves. The OC (Operating Characteristic) curve plots the probability that the null hypothesis will be accepted against the true value of the parameter being tested. The ASN (Average Sample Number) curve depicts the expected sample size through use of the SPRT as a function of the true value of the parameter.

The OC and ASN curves are by no means straightforward to determine, as they often involve solutions of non-closed form equations and the like. And the character of the solution varies according to the working distribution of the data, the hypothesized parameters, and the error rates projected. Often, the values on the curves can only be examined at a select number of combinations of hypotheses, error rates, and true values, for a particular distribution. Depending on the distribution, sometimes only empirical or asymptotic analyses are possible. And this does not say anything about the variance of the expected sample size for the ASN curve, which is a crucial piece of information for judging the ultimate practicality of the technique.

As was mentioned above, this problem has been somewhat alleviated through the subsequent works on the SPRT. Those articles and books generally concentrate on individual aspects of the SPRT. Some cover a distribution or two and assess expected sample sizes and their variances, operating characteristics, or other characteristics of interest. They generally select representative or wide-ranging sets of true values, hypothesized values, and error rates, and variance of the distribution for examination. Others explore variations of the SPRT. Often, these works are presented in response to undesirable behavior brought to light by works of the first type. Some are attempts to adapt the SPRT to situations it was not initially designed to handle. All works on the SPRT, as well as on other sequential techniques, may have something to offer us, and it is hoped that as many of them as possible may be surveyed in the coming months during the process of assembling adaptive evaluation methodology and performance measures. The author has been rummaging through the literature, in order to present a collection of the most likely candidates for future implementation.

The first problem that comes to mind regarding the basic SPRT is that only two hypotheses are being tested. This procedure could not even adequately handle a test of three hypotheses, perhaps standing for "low," "medium," and "high." Armitage [1] and others have described the use of the SPRT on two or three of the possible hypothesis pairs in the three-hypothesis case. The procedures have nice graphic interpretations that may be of use to evaluators, but little is known about the operating properties of these techniques. Two-sided tests of certain parameters have also been investigated. Often the number of SPRT's necessary to perform can be reduced in these tests. But again, the tests can turn out to be uncertain quantitites, as far as performance goes.

It is this uncertainty, and the general possibility that, for some parameter values, open tests of this sort may occasionally lend to intolerably high sample sizes, that has driven many potential practitioners away from the SPRT. In response, variations have been devised and studied. One desirable feature of a sequential test would be to impose a constraint on the number of observations allowed. Several papers with proposals for and analyses of truncated SPRT's have been published [7,5]. The results show significant improvement upon the SPRT in certain situations that the SPRT has proven weak in.

The major principle of truncation is to create boundaries that vary with the sample size, which "cuts off" observation at a pre-determined number. Boundaries can be tailored to curtail the sampling period to adapt to the circumstances of the evaluation. Another method, suggested by Wald, would be to operate the SPRT normally, but cease sampling after a specified number of observations, and make a final decision based on the cummulative sum of logs, if a boundary has not been exceeded. However, this method alters error rates in an unspecified fashion, and has been disregarded for the most part in later studies.

Another variation designed to improve the operating properties of the plan is the partial SPRT. In this method, a number of observations are taken before any sequential testing is performed. This number is based on the difference in hypothesis values. Billard [2] claims that this "has the advantage of providing a well-defined and mathematical structure to the scheme...before serious sequential comparison of the hypotheses is undertaken." He goes on to compute OC and ASN functions for the binomial case and to compare these with Armitage's truncated procedure. This also looks favorable for use, but the study was conducted only for the binomial case. We might require additional research to make the technique as widely applicable as possible.

There also exist techniques, some SPRT-oriented, for performing sequential t-tests, χ^2 tests (for testing variance), ANOVA through F-tests, and correlation coefficients. Further research could probably turn up more applications of this sequential theory. The point is, many tasks that might crop up in evaluation research can be handled sequentially. The characteristics of the relevant procedures are not always fully known, but it may be possible to construct assessments of particular characteristics to our satisfaction, where necessary. The drawback of the techniques mentioned above is that they are techniques for hypothesis testing. If we need to perform estimation at some point, all these techniques just will not do. We therefore need to develop a sequential estimation-oriented approach to adaptive switching to complete our arsenal of evaluation methodology.

SPRT theory outdistances sequential estimation methodology in terms of the sheer number of articles on each subject. But this may be due to the number of gaps left in the SPRT theory. Sequential estimation is not very different from fixed sample size estimation, with respect to the inferences made about the results. Results are usually best expressed as a point estimate with associated standard error, or an interval estimate. We can discard contemplation of how the behavior depends on the hypotheses, because there are no hypotheses. We need only concern ourselves with the estimator, the stopping rule, and the method for making point or interval estimates.

Another possible reason for the comparatively small volume of work in classical sequential estimation theory is the technological proximity of it to Bayesian theory. The Bayesian attack also produces a point estimate. This one is designed to minimize expected loss. If cost minimization is our primary concern, how different can classical sequential estimation be? According to Wetherill [15, p. 115],

> "Both point and interval estimation can be based on either of two approaches, that through 'sampling distributions,' or that through 'likelihood.'

The likelihood approach consists of Bayesian, Bayesian decision-theoretic, and "pure" likelihood procedures. Thus, for the purposes of this paper, we shall be concerned with sampling distribution-based sequential estimation.

The first step in techniques taking this approach is to state a cost function, for the goal here is to minimize costs. Then an estimator must be chosen. Optimally, we would prefer an estimator that yields minimum expected cost over the distribution of possible samples and the cost associated with each sample. This is not possible to do uniformly; certain restrictions on the properties of the estimator must be specified. One selects an estimator after imposing these conditions, that will minimize the expected loss. Next, a stopping or switching rule must be devised. It might be based directly on costs in subsequent periods (in either a projected or a minimax sense), or indirectly on these through the specification of tolerances in interval width and the associated confidence level, or in the sample variance. A procedure for making inferences must also be formulated. It will be based on the estimator, the stopping rule, and the character of the distribution.

The last paragraph is really a compression of several different classical approaches to sequential estimation. It summarizes the major avenues for development of sequential estimation methodology. Research exists which explores these avenues. But, as mentioned before, Bayesian methods find more favor in the eyes of the researchers, so the emphasis in the literature is on Bayesian techniques. To utilize classical methodology in evaluation research, we might be called upon to undertake some original research. References in classical sequential estimation work are among those listed at the end of this paper.

There is one procedure that can be applied toward both hypothesis testing and estimation, and is appealing because of its relative simplicity. It is called double or two-stage sampling. The idea of this procedure is to take one sample of prescribed size, and following this, take another sample whose size may be dependent on the first sample. For hypothesis testing, this may result in a savings, because the second sample might be by-passed if results from the first sample are conclusive enough. Or, the total sample size might be smaller than that in the corresponding fixed sample size test. In estimation, this same possibility for savings holds. This may be classified as a partially sequential technique, but it can result in considerable conservation of resources. Also, the theory is more tractable than most fully sequential techniques. It has been researched, both generally and in particular cases.

While double sampling has all this to offer, there are some noticeable shortcomings. Several particular methods do not make use of all the information collected. Having already looked at fully sequential techniques, one might not feel satisfied with the one sequential-oriented decision made at the end of the first sample. And not as much work has been done in double sampling as in, say, the SPRT. But each of these drawbacks can be mitigated to a degree. Methods can be found, or developed, which can make use of all the data. The reduction in the number of decisions may actually be a benefit, contingent on the evaluation process and the evaluators involved. And, once again, the SPRT was originally put forth in a rather vague manner, leaving many holes to be filled. The double sampling technique, in its simplicity, can be covered with less work. The biggest question is, is it adequate to handle the problems that arise in the evaluation of public programs? This cannot be answered firmly until a thorough survey of those problem situations can be conducted and studied. The author is inclined to think that it might prove very useful.

The final topic in this section covers tolerances. Tolerances have been brought up on occasion in this paper, only to be pushed aside till later. Yet the tolerance ought to be the heart of the stopping rule. If we are attempting to estimate an interval, sequentially, we can make an estimate after the first observation. It would probably be quite large, though. Our goal would naturally be to sample until the width of the confidence interval falls below some critical value. However, we could also reduce the confidence level to shrink the interval width. Thus, we need combinations of tolerances in confidence level and interval width to define meaningful stopping rules. In hypothesis testing, the tolerance manifests itself in type I and type II error rates which are incorporated into the determination of the boundaries. Our development of adaptive evaluation methodology, if not based on strict comparisons between costs in present and in future periods, must provide for a rational basis for setting tolerances.

The simple-minded approach for tolerance generation might be to enforce the same tolerances in every period; i.e., design stopping rules to yield equal significance levels or confidence levels and interval widths in each period. This would be the easiest way to go about it, but probably not the most efficient way. We can detect ways to improve tolerance assignments by looking at disadvantages in the simple-minded approach:

*The cost of sampling might be much higher in one period than in another. If that is the case, it would make sense to settle for less stringent tolerances where sampling cost is high.

*Different assessed variances for future periods may make tolerance assignment by period, based in some way on the projected variance for the period, a more loss-minimizing approach. This is especially true when overall loss depends on the difference between estimated and true values of a parameter.

*In the two-hypotheses case, we might express expected loss as:

$$E(\ell|\alpha,\beta) = \alpha e_1 + \beta e_2 + c_0 \; E\;(n|\theta_1,\theta_2,\alpha,\beta)$$

where α is the probability of deciding for the alternative hypothesis when the null hypotheis is true, e_1 the loss associated with that wrong decision, β the probability of deciding for the null when the alternative holds, e_2 the loss in that decision, c_0 the cost per observation, n the number of observations, θ_1 the null hypothesis, and θ_2 the alternative. The expected sample size is almost always a function of the hypotheses and error rates. The expected loss-minimizing choices for α and β depend on the other costs involved, and probably are not equal, so uniform calibration of error rates across the whole evaluation is simply too naive an approach.

*Assessments of costs, variances, etc. themselves may change during the course of an evaluation, so a sequential scheme for setting tolerances may be desirable.

This last point brings us to a theory that looks very much like dynamic programming. The study of dynamic programming is beyond the scope of this paper, but it looks as though it will prove very useful in adaptive evaluation. Dynamic programming might also help us in deciding how to sequence segments of the evaluation when that is not forced by the circumstances of the problem. Tolerances could be interpreted in information theory terms to produce the best design. This is all conjecture, but enlightened conjecture. Ideas such as these, and other prospects such as taking tolerance assignments as states for use in Markovian decision processes, will require much inspection before models for testing may be set up. The configuration of the sets and the goals involved may be incompatible with the domains of the dynamic theories. They are brought to mind in order to make this survey more thought-provoking and more exhaustive.

This concludes our preliminary survey of candidates for methodology to manage the problem of adaptive switching of time periods in an evaluation. While the literature seems to lean towards Bayesian methods more and more nowadays, classical theories should not be dismissed without a fair trial. When we have produced a set of performance measures, we will be able to compare techniques. But as the calculation of performance measures is not likely to be simple for these techniques, the most promising ones ought to be tapped for further research.

IV. ADAPTIVE ALLOCATION OF SCARCE RESOURCES

At any particular time during the course of an evaluation, information of interest might be coming in from a variety of sources. It does not simply walk in and present itself. It must be actively gathered, requiring the consumption of scarce information-gathering resources. We desire a scheme for adaptively allocating these resources to bring in the most needed and most substantial information we can acquire.

The adaptive allocation of resources problem is strongly reminiscent of the adaptive switching problem. It looks much as if the latter problem had been "turned over on its side." Accordingly, we might think to bend what we have produced for handling adaptive switching to fit the adaptive allocation problem. However, the difference is more pronounced than that, operationally. In the former case, only one segment may be measured directly; the rest must be projected. In the latter, all segments are under investigation at the time of the decision. In effect, we are performing sequential estimation or testing on all segments at the same time. While it may seem we must start all over again in tackling it, the problem does bear a close resemblence to some situations that well-established statistical techniques have been designed to handle, possibly making matters much easier for us.

After looking at the nature of the adaptive allocation problem, the first idea that comes to mind is to call it a Markovian decision process (MDP). The basic principle of this theory is to portray the experimental process as one of entering and exiting various states. One receives a certain reward upon entering a particular state. One then chooses an action, exits the state, and enters a new one randomly according to a probability distribution that is dependent only on the previous state and the action chosen. For our purposes, the interpretation of "state" would be our state of knowledge collected, the "action" would involve the selection of a sampling plan, and the "transition" would be the next set of data to come in, updating the knowledge profile.

The classification of "reward" in the MDP is the troublesome spot. One of MDP's most common applications is the "two-armed bandit" problem. A gambler plays a two-armed slot machine, and can pull only one of the arms on any trial. Winnings are based on same distribution whose underlying parameter is unknown, and are dependent on which arm is chosen. The gambler wants to maximize his winnings over the term of the game. This type of problem, generalized to K arms, has broad applications in fields such as medicine. MDP theory may be used in the screening of a set of drugs to find the most effective one.

Markovian decision theory falls short for the adaptive allocation problem as it now stands. There are two reasons for this. First, the theory is aimed at selecting only one segment to examine on a particular trial. Second, it is unclear what we would be trying to find the "best" of. So, we will leave Markovian decision theory alone for now, and shift gears completely.

Perhaps the most attractive theory for adaptation to the adaptive allocation of evaluation resources problem is stratified sampling theory. Stratified sampling is designed to minimize the sample variance of a global mean estimate through the division of the sampling population into strata. These strata essentially represent the variations in some characteristic of the population. There should be little variation in the characteristic within each stratum, and lots of variation between strata. Substantial reduction in the standard error of the overall estimate can be achieved through judicious selection of strata. The author believes that, through careful interpretation and priority setting, stratified sampling theory can provide all the basic tools for the adaptive allocation problem.

Strata selection and strata sample size assignment are the two preliminary tasks which must be performed in stratified sampling. As such, they are the most frequent topics treated in the literature. If we were to implement stratified sampling into adaptive evaluation methodology, we need only concern ourselves with the second question. A "stratum" may be interpreted as one of the information sources or groups. For instance, if teachers, administrators, students, and other citizens are being interviewed during an evaluation, each set of interviewers may be considered a stratum. This removes the task of organizing the population into homogeneous groups.

The second problem is almost as easily solved in the general theory, but there are several catches involved. With respect to certain knowledge and priorities, there exists an optimal scheme for selecting strata sample sizes. If the size of stratum h is N_h sampling units, S_h the square root of the true variance in h, and C_h the cost per observation sampled there, then the number of units from stratum h to be sampled should be proportional to:

$$\frac{N_h \; S_h}{\sqrt{C_h}} \; .$$

A design of this type will minimize the variance of the overall estimate for a specified cost, and minimize the cost for a specified variance [3].

This is a very nice result, and very easy to use, but as indicated before, there are catches. The first is of concern to any who use stratified sampling. What does one use for true variance, if this is not known? If some sampling has been conducted already, one might try to use the sample variances. But not all sampling theorists condone this, especially if only a small number of observations have been taken. Some prefer an approach where sample sizes are taken more proportional to stratum sizes. Double sampling can also be applied here, in a sense. The first sample would establish working sample variance estimates for use in determining the second sample sizes.

Observe that stratified sampling is not necessarily a sequential technique.

The sample sizes selected are intended to be one-time samples. However, it seems plausible to extend the double sampling principle of the preceeding paragraph to a K-stage sampling procedure. Notice, also, that the optimal sample size described above is specified only in "proportional to" terms. A scaling factor must be injected into the selection of sample sizes. If there is an overall budget constraint imposed on the evaluation, the constraint maybe apportioned over the time segments of the evaluation process. At each decision point, we would work with the budget constraint for the next period, plus the current knowledge profile, to determine sample sizes for the next period. The method of apportionment can range from simple to complex: equal distribution over all periods; distribution according to the "importance" of the measurement, or some other weighting factor, in each period; even a dynamic programming-based allocation, possibly with sequential updating. The literature does not go much further than double sampling insofar as sequential techniques go. Procedures of this sort will probably demand original research.

Another area in which "importance" becomes significant is in stratum weights. In a school program evaluation, there might be one hundred times as many citizens as teachers who might be sampled. Yet we would probably not want to interview one hundred times as many citizens as teachers, assuming other factors were equal. The teachers' opinions are probably worth more to us than those of the citizens. Accordingly, there must be some adjustment to stratum sizes for the importance of the information coming out of the stratum. This can be incorporated into the stratum sizes themselves. In fact, N_h is often expressed as W_h, the stratum weight, in the optimal allocation formula. There is no quantitative difference here, for this re-expression can be derived by dividing all N_h's by N, the total population size. We might think to take "stratum weight" more literally, as a measure of importance. Weights might be assessed by the stratum's participation in the program, its loss associated with acceptance or rejection of the program, or some other relevant factor. It would probably be contingent on the program and the environment. If such a transformation of stratum weight were made, one must then be very careful with the meaning of "sample size" and "cost per observation," as these quantities are likely to change.

Transformations of this type also bring to mind the final quirk in our adaptation. The object of stratified sampling is to estimate some parameter in an entire population. Strata are employed mainly to reduce the variance of the estimate. In contrast, we might want to examine each stratum as a distinct entity. Strata are not introduced by choice, but by the nature of the population of interest. If this is the case, we must look at the priorities stratified sampling assumes and see whether the approach is valid. We must articulate our priorities if stratified sampling does not use them, and design techniques that will be aware of them. We have to identify these aspects of the analysis we are trying to control or minimize. This might involve setting tolerances in error magnitudes and rates. Again, the problem of tolerance assignment is a critical one, and a key to developing effective adaptive techniques using classical machinery.

The foregoing arguments are not intended to dismiss any statistical technique other than stratified sampling from consideration for the adaptive allocation of resources problem. The characteristics of the problem do seem to match well the conditions under which stratified sampling operates, however. As stated before, Markovian decision theory is designed to select the best out of a set of alternatives, rather than to collect information on all groups. That is the main reason it has been bypassed thus far. Yet the MDP techniques might be put to use in this problem, if we alter our definitions of the components. Suppose the alternatives consisted of the sampling plans, and the states are all the possible combinations of numbers to be sampled from each stratum during the next period. Markovian decision theory might then provide us with a "best sampling plan" for the next period. Although stratified sampling might be our basic tool for this problem, let us not limit ourselves in our set of accessories.

To conclude this section, let us backtrack to the adaptive switching problem. What if we were to look at it as the adaptive allocation problem, "turned over on its side?" We might be able to apply assessed variances, costs, and weighting factors to future periods and synthesize an "N" for the coming sampling day. Of course, we can only sample from this stratum at this time, where "stratum" no longer means "sampling group" but "sampling period." This is probably no answer to the adaptive switching problem, but we should be open to ideas that we generate through formulation of adaptive allocation techniques.

V. PROPOSAL FOR FUTURE RESEARCH

The generation of ideas for devising solutions to a problem can be an unending process; no progress is made until the ideas are fitted to the physical aspect of the problem, the ill-fitting ones discarded or altered, the promising ones measured and tailored to meet the needs of the user. That is the current situation of the classical approach to the formulation of adaptive evaluation methodology. Many ideas have been presented for handling the problems specified. What is needed now is one or both of two responses: identification of the more promising techniques for further development; and/or provision of representative evaluation cases that can supply a clearer picture of the requirements that the methodology must measure up to.

There are several main directions in which future research may head. One of the more favorable ones at this point is in the direction of adaptive allocation of evaluation resources. The opportunities for methodology development seem more clearcut in this area, and the possibility was raised that techniques implemented here might aid us in the development of methodology in adaptive switching times of an evaluation. Yet it is difficult to evaluate what will work in practice after the theory is developed. We need some sense of what is actually going to be out there when it comes time to use the techniques. What is being measured (in terms of parameters and the distributions they arise from)? What are the risks (loss functions)? Where do our priorities lie (weights, objects of the evaluation)? If cases cannot be supplied, we should at least have some guidelines for handling the questions asked above. Only then can effective, practical techniques be developed.

Of course, the same questions should be answered even if we were to proceed right to the adaptive switching problem. But more is needed here. The proliferation of techniques, with all their pros and cons, makes it difficult to move forward.

Some thought must be lent to the approaches and the procedures touched on in this report. It would be desirable to have an evaluation-oriented appraisal of the subject matter presented, with the top candidates earmarked for further study. In particular, feedback on the hypothesis test/estimation question, including comment on the associated loss structures, is necessary. Perhaps we should continue to look at both themes. The field of classical sequential theory is simply too wide and too inconclusive for something to appear that is uniquely well-suited to the problem.

Research in the coming months is thus likely to include a closer look at the promising techniques and their assumptions, assessed in an evaluation-oriented context. Also, noting that the development of a set of performance measures for comparison and selection of techniques is a short-term goal of this project, characteristics of the techniques will be examined broadly for possible adaptation to performance measures. At this point, two characteristics which have some potential for this purpose are the Operating Characteristic and Average Sample Number curves described in Section III. Capsulization of these curves into a value or set of values for ease of comparison will also be explored. On a more economic scale, expected loss comes across as a particularly meaningful candidate, but may require more input than the classical approach normally allows for. Other ideas for performance measures shall also be invited and investigated.

One tool that might be employed to assess some of the techniques proposed is the computer. A small project that has been started would produce a computer demonstration of the inefficiency of the "always test at 5%" approach, through comparison of this and loss-minimizing approaches, and/or simulation. The prospect of simulation also arises for testing some of the other techniques in sequential estimation and adaptive allocation. Computer programs would allow for the user to specify the parameters of the problem, and provide a wide variety of results. Computers may also be used to supplement the development of performance measures.

The author looks forward to working with Mr. John VandeVate in the coming months. Mr. VandeVate is exploring Bayesian approaches to the problems handled in this paper. It is hoped that Bayesian and classical methods may be compared head-to-head, which may weed out some of the inferior designs and save us from pointless, time-consuming analysis. Also anticipated from Mr. VandeVate is a mini-case study that might help to crystallize the salient concerns of adaptive evaluation research.

REFERENCES

1. Armitage, P., (1957) "Restricted Sequential Procedures," Biometrika, 44, 9-26.

2. Billard, L., (1977) "Optimum Partial Sequential Tests for Two-Sided Tests of the Binomial Parameter," Journal of the American Statistical Association, 72, pp. 197-201.

3. Cochran, W.G., (1977) Sampling Techniques, John Wiley & Sons, New York.

4. DeGroot, M.H., (1970) Optimal Statistical Decisions, McGraw-Hill, New York.

5. Genizi, A., (1965) "On the Performance of the Truncated Sequential Probability Ratio Test," JASA, 60, pp. 979-984.

6. Hald, A., (1975) "Optimum Double Sampling Tests of Given Strength, I. The Normal Distribution," JASA, 70, pp. 451-456.

7. Madsen, R., (1974) "A Procedure for Truncating SPRT's" JASA, 69, pp. 403-410.

8. Read, C.B., (1971) "The Partial Sequential Probability Ratio Test," JASA, 66, pp. 646-650.

9. Solomon, H. and Zacks, S., (1970) "Optimal Design of Sampling from Finite Populations: A Critical Review and Indication of New Research Areas," JASA, 65, pp. 653-677.

10. Spurrier, T.D., and Hewett, J.E., (1975) "Double Sample Tests for the Mean of a Normal Population," JASA, 70, pp. 448-450.

11. Stein, C., (1945) "A Two-Sample Test For a Linear Hypothesis Whose Power is Independent of the Variance," Annals of Mathematical Statistics, 16, 243-258.

12. Suktahme, B.V., and Tang, Victor K.T., (1975) "Allocation in Stratified Sampling Subsequent to Preliminary Test of Significance," JASA, 70, pp. 175-179.

13. Wald, A., (1947) Sequential Analysis, John Wiley & Sons, New York.

14. Wald, A. and Wolfowitz, T., (1948) "Optimum Character of the Sequential Probability Ratio Test," Annals of Mathematical Statistics, 19, 326-339.

15. Wetherill, G.B., (1966) Sequential Methods in Statistics, John Wiley and Sons, Inc., New York.

16. Willemain, Thomas R., (1978) "Analysis of a Contingent Experimental Design: A Before-and-After Experiment With a Baseline Period of Random Duration," working paper, Operations Research Center, MIT.

17. Wolde-Tsadik, G., (1976) "A Generalization of an SPRT for the Correlation Coefficient," JASA, 71, pp. 709-710.

18. Wolfe, D.A. (1977) "On A Class of Partially Sequential Two-Sample Test Procedures," JASA, 72, pp. 202-205.